

Spatial string matching for image classification

Yunqiang Liu

Image Group
Barcelona Media – Innovation Center
Barcelona, Spain
e-mail: liuyunq@yahoo.com

Vicent Caselles

Department of Technology
University of Pompeu Fabra
Barcelona, Spain
e-mail: vicent.caselles@upf.edu

Abstract—This paper presents a spatial string matching method to incorporate spatial information into the bag-of-words model, which represents an image as an unordered distribution of local features. Spatial constraints among neighboring features are explored in order to achieve better discrimination power for image classification. The features from neighboring points are combined together and taken as a spatial string, and then our method matches the images according to the similarity of string pairs. The categorization problem can be formulated using KNN or SVM classifier based on the spatial string matching kernel. The proposed method is able to capture spatial dependencies across the neighboring features. Experiment results show promising performance for image classification tasks.

Keywords—Bag-of-words, Spatial string matching, image classification

I. INTRODUCTION

The bag-of-words model has been widely employed in object detection and scene classification tasks. It describes images as sets of elementary local features called visual words. Based on keypoints as salient image points and extracting on each of them a descriptor, an image can be represented as a bag of visual words [1]. This method consists of several basic steps: (I) keypoints (which represent salient image patches that contain local information of the image) are detected using various detectors, (II) keypoints are represented using local descriptors such as SIFT [2], (III) descriptors are vector quantized into a fixed-size codebook so that those which are similar are assigned into the same cluster, and (IV) mapping keypoints into visual words, an image is represented as a vector whose coordinates are the counts of each visual word in the image. This feature vector can be used as an input to a classifier in the final classification step.

The bag-of-words model uses only the local appearance information to represent an image, and characterizes the statistics of local patch appearance

without involving any geometric information. This model is efficient and has received a lot of attention owing to its simplicity, robustness, and good practical performance. However, this representation ignores spatial relationships between the local features; and is missing some discriminative power since the spatial layout of the features may be almost as important as the features themselves.

The importance of spatial layout has been recognized in object detection and classification tasks. The part-based approach in [3] is based on the construction of a dictionary of composite semi-local parts to model spatial configurations for texture and object recognition. Moreels et al. [4] represent objects as flexible constellation of rigid parts. Li et al. [5] combine feature representations with some spatial ordering constraints to resolve local visual ambiguities. Lazebnik et al. [6] propose to repeatedly divide an image into sub-regions and compute histograms of local features in each sub-region, and build-up a spatial pyramid kernel; however the representation is not invariant to geometrical transformations. In [7],[8],[9], higher order features are constructed by encoding spatial relationships between an specified number of local features. However, higher-order features have very high computational complexity as they significantly increase the dimension of feature space.

This paper presents a spatial string matching method to incorporate spatial information into the bag-of-words model, which represents an image as an unordered distribution of local features. Spatial constraints among neighboring features are explored in order to achieve better discrimination power for image classification. The features from neighboring points are combined together and taken as a string which contains the spatial layout of these features. Based on such a representation, the image similarity is measured by computing the distances of the spatial strings. The categorization problem can be formulated using a KNN (K nearest neighbor) or SVM (support vector machine) classifier based on a spatial

string matching kernel. The proposed method can capture spatial dependencies across neighboring features. Experimental results show that the proposed method provides a significant improvement on image classification tasks.

II. SPATIAL STRING MATCHING

We describe in this Section the spatial string matching kernel of two images to capture the spatial dependency across visual words.

A. Image representation

In this work, each image is divided into equivalent blocks on a regular grid with spacing d . We take as keypoints the set of grid points, each with a circular support area of radius r . Each support area can be taken as a local patch. The patches are overlapped when $d < 2r$. Each patch is described by a descriptor like SIFT (scale-invariant feature transform). Then a visual vocabulary is built-up by vector quantizing the descriptors using a clustering algorithm such as K-means. Each resulting cluster corresponds to a visual word. With the vocabulary, each descriptor is assigned to its nearest visual word in the visual vocabulary.

Following this procedure, each image can be represented as a 2D sequence of visual words, and each word is associated with a block. It should be noted that dense image feature sampling has shown promising performance in many computer vision applications [10].

B. Image matching based on spatial strings

Similar to the bag-of-words method, each image is first represented as a 2D sequence of visual words. In this situation, visual words come from vector quantization and each visual word corresponds to a group of local features. Note that the quantization error inevitably introduces ambiguity of the visual word representation. In fact, since objects of different categories may share similar local appearances, a visual word may appear in images from different categories. Now, if local features (we can fix ideas and consider SIFT descriptors as local features) coming from different categories share the same visual word, it is very difficult to distinguish their categories. However, these features seldom share similar neighborhoods. On the other hand, if local features represented by the same word in images of the same category really match each other, they often share similar neighborhoods. One expects that the spatial dependencies in the neighboring region will increase the discriminative power.

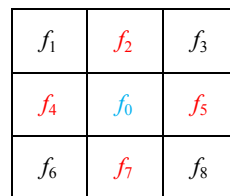


Fig.1. Neighbors relationship of features

In order to capture the spatial dependencies in images, we consider the combination of the visual words in a neighboring region. The basic idea is that we combine together the visual words of neighborhoods and take the combined words as a string. This string contains the spatial layout of these visual words. Based on such a representation, the image similarity is measured by computing the distances of the spatial strings.

A spatial string is defined as a group of visual words which are arranged orderly. We give an example in Fig.1. Let f_0 denote a local feature, $f_1 \sim f_8$ are its eight spatial neighbors. And $w_0 \sim w_8$ are the corresponding visual words related with $f_0 \sim f_8$, respectively. We can combine the ordered words w_0, w_2, w_4, w_5, w_7 as a spatial string in the case of 4-nearest neighbors (4-NN), or $w_0 \sim w_8$ in the case of 8- nearest neighbors. This spatial string is represented as $s(f_0)$, and consists of the visual words of the neighborhood of feature f_0 . The string $s(f_0)$ can capture the spatial dependency among the features around f_0 .

We used the Levenshtein distance (LD) [11] to measure the strings difference. LD distance, also called edit distance, is defined as the least number of deletions, insertions, or substitutions required to transform one string into another.

Based on the LD distance, we can define the similarity between two images. Let $I_1 = [f_1^1, f_2^1, \dots, f_{N_1}^1]$ and $I_2 = [f_1^2, f_2^2, \dots, f_{N_2}^2]$ denote the set of features of two images, where N_1 and N_2 are the number of features in image I_1 and I_2 , respectively. Suppose that we build-up a visual vocabulary $W = [w_1, w_2, \dots, w_V]$, where V is the size of the vocabulary, that is, the number of visual words.

Computing all possible string pairs between two images has a very high computational cost, and it is not necessary. We only match the strings whose center features correspond to the same visual word. First, for each word w_i in the vocabulary, we collect the set of features in images I_1 and I_2 corresponding to the same visual word. That is, we define $F_j(w_i), j=1,2$, as the set of features f in image I_j corresponding to the word w_i .

Based on LD distance, we can compute the following similarity measure between images I_1 and I_2 :

$$D(I_1, I_2) = \sum_{i=1}^V \sum_{f_1 \in F_1(w_i), f_2 \in F_2(w_i)} \max \{M - LD(s(f_1), s(f_2))\} \quad (1)$$

where M is the length of spatial string, here $M=5$ for 4-NN. According to the definition of LD distance, M will never be smaller than $LD(s(f_1), s(f_2))$, so $D(I_1, I_2)$ is a non-negative quantity.

C. Spatial string matching kernel

In this Section, we describe a spatial string matching kernel for two images based on the spatial string similarity measure. Note that the similarity measure is not symmetric, *i.e.*, $D(I_1, I_2)$ is not always equal to $D(I_2, I_1)$. In order to get a symmetric quantity, we change it into:

$$K(I_1, I_2) = \sum_{i=1}^V \min \{d(w_i^1, w_i^2), d(w_i^2, w_i^1)\} \quad (2)$$

where

$$d(w_i^1, w_i^2) = \sum_{f_1 \in F_1(w_i), f_2 \in F_2(w_i)} \max \{M - LD(s(f_1), s(f_2))\} \quad (3)$$

This defines the spatial string matching kernel (SSMK). The SSMK coincides with a histogram intersection kernel (HIK) [12] if we do not consider the visual words from neighboring keypoints, that is, if a string only consists of the word of the central feature. From this point view, the SSMK kernel can be explained as a weighted histogram intersection kernel, where the words having similar neighbors between two images will be given bigger weights.

III. EXPERIMENTS

We evaluated our spatial string matching kernel for an image classification task on two public datasets: Caltech-101 [13] and MSRC-2 [7]. We first describe the implementation setup. Then we give the comparison results.

A. Experimental setup

For the two datasets, we use only the grey level information in all the experiments, although there may be room for further improvement by including color information. The keypoints are obtained using dense sampling, specifically, we compute keypoints on a dense grid with spacing $d=7$ both in horizontal and vertical directions. SIFT descriptors are computed at each patch over a circular support area with radius $r=5$. Under this configuration, the patches are overlapped.

For each keypoint, we use its 4-NN to form a spatial string with 5 visual words.

Two datasets are used in the experiment: Caltech-101 and MSRC-2. We set the visual vocabulary size as 100 for both datasets. We compare our method with the bag-of-words model using KNN and SVM classifiers on both datasets. For the bag-of-words model, a histogram intersection kernel is used as a similarity measure in the classifiers.

B. Experimental results

B.1. Caltech-101

In the experiments with this dataset, we choose eight categories out of 102 categories: {brain, butterfly, camera, chair, lamp, laptop, pyramid and stop_sign}. Moreover, images within each category are randomly divided into two subsets of the same size to form a training set and a test set. We repeat each experiment five times over different splits, and report the average results. In terms of the bag-of-words (BOW) model, a histogram intersection kernel [12] is used as similarity function for both KNN and SVM. We test the KNN method with different parameters k , and we report the best result. The classification results for KNN and SVM are shown in Table I. Some misclassified images are shown in Fig.2.

TABLE I. CLASSIFICATION RESULTS (%) ON CATECH-101

BOW		SSMK	
KNN	SVM	KNN	SVM
79.04	86.92	87.30	90.77

B.2. MSRC-2

There are 20 categories, and 30 images per category in this dataset. We choose six categories out of them: {tree, cow, face, car, bike, and book}. We randomly select 15 images per category for training, and the rest for testing. Table II shows the results for both KNN and SVM.

TABLE II. CLASSIFICATION RESULTS (%) ON MSRC-2

BOW		SSMK	
KNN	SVM	KNN	SVM
82.22	91.11	86.67	94.44

TABLE III. CONFUSION MATRIX FOR SSMK WITH SVM CLASSIFIER

	tree	cow	face	car	bike	book
tree	14	1	0	0	0	0
cow	0	15	0	0	0	0
face	0	1	13	0	0	1
car	1	0	0	14	0	0
bike	0	0	0	0	15	0
book	0	0	0	0	0	15

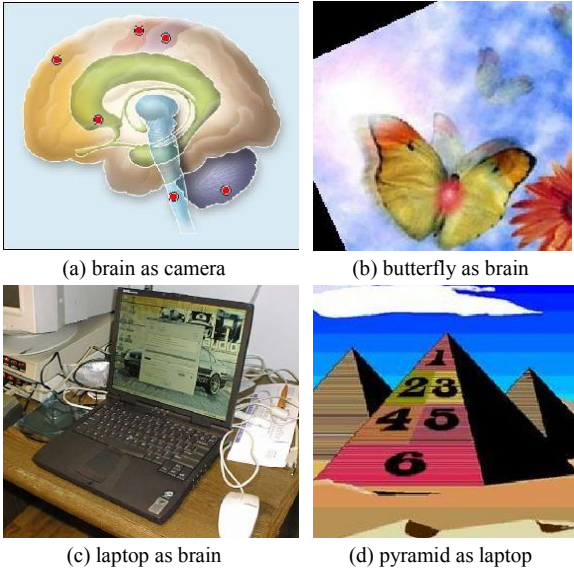


Fig.2. Some misclassified images on Catech-101.



Fig.3. Some misclassified images on MSRC-2.

The confusion matrix for SSMK with the SVM classifier is presented in Table III to give more details on the categorization of each category. The first column contains the true labels and the first row lists the referred labels. The numbers of correctly classified images for each category are shown in the diagonal.

Moreover, Fig.3 shows some misclassified images.

IV. CONCLUSIONS

In this paper we presented a spatial string matching method to incorporate spatial information into the bag-of-words model. Spatial dependencies among neighboring features are explored for image classification. We build-up a spatial string matching kernel to measure the similarity between images. By considering spatial dependencies across neighboring features, the proposed method can get higher discriminative power compared with individual feature. Experiment results show promising performance for

image classification tasks. In future work, we will address the problem of scale and rotation invariance of the features configuration.

ACKNOWLEDGMENT

This work was partially funded by Mediapro through the Spanish project CENIT-2007-1012 i3media and by the Centro para el Desarrollo Tecnológico Industrial (CDTI). Y. Liu acknowledges partial support from the Torres Quevedo Program from the Ministry of Science and Innovation, funded by the European Social Fund. V. Caselles also acknowledges partial support by MICINN project, reference MTM2009-08171, by GRC reference 2009 SGR 773 and by "ICREA Acadèmia" prize for excellence in research funded both by the Generalitat de Catalunya.

REFERENCES

- [1] C. Dance, J. Willamowski, L. Fan, C. Bray, and G. Csurka. Visual categorization with bags of keypoints. In *Proc. ECCV*, 2004.
- [2] G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, vol. 60, No.2, pp.91–110, 2004.
- [3] S. Lazebnik, C. Schmid, and J. Ponce. A maximum entropy framework for part-based texture and object recognition. In *Proc. ICCV*, vol. 1, pp. 832–838, 2005.
- [4] P. Moreels, M. Maire, and P. Perona. Recognition by probabilistic hypothesis construction. In *Proc. ECCV*, pp. 55–68, 2004.
- [5] Y. Li, Y. Tsin, Y. Fenc, and T. Kanade. Object Detection Using 2D Spatial Ordering Constraints, In *Proc. CVPR*, pp.1188–1195, 2005.
- [6] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. CVPR*, 2006.
- [7] S. Savarese, J. Winn, and A. Criminisi. Discriminative object class models of appearance and shape by correlators. In *Proc. CVPR*, 2006.
- [8] L. Yang, P. Meer, and D. Foran. Multiple class segmentation using a unified framework over mean-shift patches. In *Proc. CVPR*, 2007.
- [9] D. Liu, G. Hua, P. Viola, and T. Chen. Integrated feature selection and higher-order spatial feature extraction for object categorization. In *Proc. CVPR*, 2008.
- [10] A. Bosch, A. Zisserman, and X. Munoz. Scene classification via pLSA. In *Proc. ECCV*, pp. 517–530, 2006.
- [11] G. Navarro. A guided tour to approximate string matching. *ACM Computing Surveys*, vol.33, No.1, pp.31–88, 2001.
- [12] S. Maji, A. Berg, and J. Malik. Classification using intersection kernel support vector machine is efficient. In *Proc. CVPR*, 2008.
- [13] L. Fei-Fei, R. Fergus and P. Perona. Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In *Proc. CVPR, Workshop on Generative-Model Based Vision*, 2004.