

Effective Multi-level Image Representation for Image Categorization

Hao Li and Yuxin Peng

Institute of Computer Science and Technology
Peking University
Beijing, China
{lihao, pengyuxin}@icst.pku.edu.cn

Abstract—this paper proposes a novel approach for image categorization based on effective multi-level image representation (MLIR). On one hand, to exploit fully the information of segmented regions at different levels in the image, we recursively segment the image into a hierarchical structure. On the other hand, to represent the information at different levels in a uniform manner, we construct a visual vocabulary based on the image regions of the hierarchical structure by a random sampling strategy. And the intermediate feature mapping is adopted to form a multi-level image representation, which encodes the information of the image at different levels, and can be very useful for distinguishing images from different categories. Experimental results on the widely used COREL data set have shown our proposed approach can achieve significant improvement compared with the state-of-the-art methods.

Keywords—multi-level image representation; image hierarchical structure; image categorization

I. INTRODUCTION

Image categorization refers to the labeling of images into one of some predefined categories, which is very useful for the content-based image analysis, retrieval, and other applications. In this issue, the representation of image is crucial because it is the primary step for the image categorization. Generally speaking, we can classify the information contained in the image into two types: the global information and the local information. Both of these two kinds of information are important for distinguishing images from different categories.

Early works on image categorization use global features extracted from the whole image. In [1], Szummer and Picard apply k-nearest neighbor classifier based on color histograms to discriminate the images of “indoor” and “outdoor” categories. The work of Oliva and Torralba [2] has shown that statistical properties of the scene considered in a holistic fashion, without any analysis of its constituent objects, yield a rich set of cues to its semantic category. However, these works neglect the local information in the image, which can also be very useful for identifying the category of image.

More recently, the intermediate representations are introduced to exploit local statistics in images. On one hand, in the work of Csurka et al. [3], the bag of keypoints method is proposed based on vector quantization of affine invariant

descriptors of image patches. On the other hand, Chen and Wang [6] represent an image based on regions, where a collection of instance prototypes is learned according to a diverse density (DD) function and then image features are obtained by non-linear feature mapping.

Later works demonstrate the utility of integrating global spatial information into local statistics. For local patches, Lazebnik et al. [4] partition the image into increasingly fine sub-regions and computing histograms of local patches found inside each sub-region. And Tirilly et al. [5] propose a new representation of images: visual sentences, which allow us to “read” visual words in a certain order. For regions, in the work of Gökalp and Aksoy [7], both “bag of individual regions” and “bag of region pairs” representations are used for scene classification. Specifically, the “bag of region pairs” representation is designed to capture the particular spatial relationships of regions. These methods perform well by integrating the global and local information together.

In this paper, we seek to find a uniform image representation method, which directly describes the hierarchical structure of an image, and could not only cover contents of image regions at different levels, but also imply the relationships between these regions. As is shown in Fig. 1, an image is parsed into a hierarchical structure. A proper representation of the hierarchical structure can benefit in two aspects. On one hand, it can reflect fully the rich content of an image, which contains the semantic regions at different levels and their relationships. On the other hand, it is robust to under-segmentation and over-segmentation, and can represent the image content precisely.

In our approach, images are first segmented into regions of different levels by multi-level image segmentation. A visual vocabulary is then constructed based on these regions. In particular, we use a random sampling strategy for visual vocabulary construction due to its high efficiency and good effectiveness. By mapping an image into the intermediate feature space according to the visual vocabulary, we can get its multi-level representation. Finally, given the representation, image categorization is done using SVMs. Experiment results show that exploiting image contents at different levels can significantly improve the performance of image categorization. Specifically, our proposed MLIR-

based image categorization approach outperforms the state-of-the-art methods on the COREL data set.

The rest of this paper will be organized as follows: Section 2 describes our method to segment the image into a hierarchical structure, and section 3 represents the visual vocabulary construction and intermediate feature mapping. Section 4 describes the classification method. Section 5 shows the experimental results and Section 6 concludes this paper.

II. MULTI-LEVEL IMAGE SEGMENTATION

This paper proposes a novel image categorization approach based on exploiting image contents at different levels. The key idea is the multi-level image segmentation (MLIS) algorithm. The major difference between MLIS and traditional image segmentation (TIS) algorithm is their goal. The TIS algorithm aims to parse the image into several non-overlapping regions, while MLIS algorithm aims to obtain image regions at different levels.

In our work, the image is parsed under a top-down framework. First, a single group is initialized by all the pixels in the image. Then, the initial group is partitioned into two groups according to pixels' color and location. And all the obtained groups are repartitioned in the same way until the stop criterion is reached. As a result, a tree structure will be obtained for the image. In TIS, only the regions at leaf nodes are kept. While in MLIS, all the regions are used to represent the content of image.

In particular, we employ Normalized Cuts algorithm [10], which aims to maximize the total dissimilarity between the different groups and the total similarity within the groups, to recursively partition the pixel groups. For the stop criterion, we assume that the region whose area is less than a certain threshold is not qualified to be treated as a meaningful semantic unit. So when the area of a pixel group is smaller than a predefined threshold, this pixel group is removed and its repartition is ended.

Fig. 1 shows an example for regions obtained by MLIS. And Fig. 2 shows an example for regions obtained by TIS. From these two examples, we can see that image regions obtained by MLIS are much more informative than the regions obtained by TIS. First, image contents at different levels are well exploited by MLIS. Second, the relationships between different regions are reflected and represented implicitly in the MLIS region contents. Moreover, MLIS is robust to under-segmentation and over-segmentation, and can represent the image content precisely.

To describe the content of each region, three types of visual feature (color, texture and shape) are used. We combine the extracted features together and normalize them to zero mean and unit standard deviation.

a) CH (Color Histogram) [11]: color representation as an 81-dimensional histogram in HSV color space ($9h \times 3s \times 3v$).

b) LBP (Local Binary Pattern) [12]: texture representation as a 59-dimensional feature vector ($LBP_{8,2}^{u2}$).

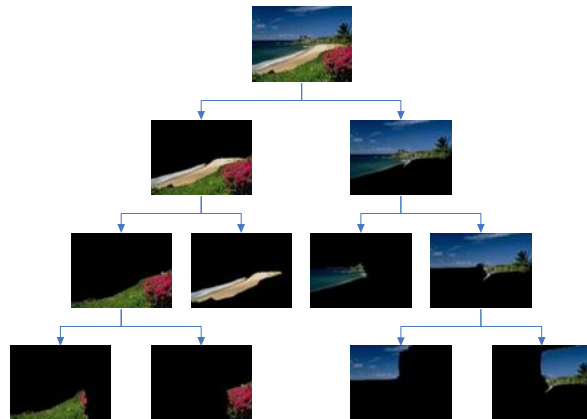


Figure 1. An example for regions obtained by MLIS

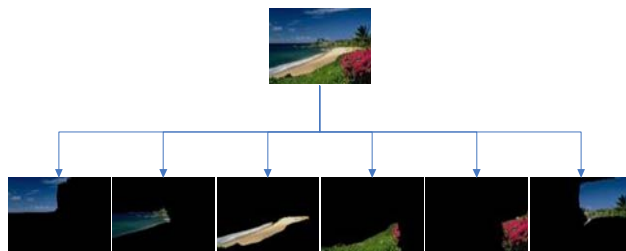


Figure 2. An example for regions obtained by TIS

c) NI (Normalized Inertia) [6]: shape representation as a 3-dimensional feature vector.

III. VISUAL VOCABULARY CONSTRUCTION AND INTERMEDIATE FEATURE MAPPING

Both instance selection methods [6] [8] [9] and clustering algorithms [3] [7] have been used for visual vocabulary construction. However, these methods are quite time-consuming even impractical for large number of instance and big size of visual vocabulary. We apply a random sampling strategy for visual vocabulary construction, since the random sampling strategy is efficient and can also get the good effectiveness. We just randomly select p regions from the training set, and use them all as visual vocabulary. The methods such as principle component analysis (PCA) can be used to optimize the visual vocabulary based intermediate feature space. However, we use the random sampling results directly for simplicity purpose.

After the visual vocabulary $V = \{W_1, W_2, \dots, W_p\}$ is constructed, we can form the intermediate feature for image according to the following equations.

$$F_i = \{F_{i,1}, F_{i,2}, \dots, F_{i,p}\} \quad (1)$$

$$F_{i,j} = \max_{k=1,\dots,n_i} \text{sim}(x_{i,k}, W_j) \quad (2)$$

$$\text{sim}(x_{i,k}, W_j) = \exp(-\|x_{i,k} - W_j\|^2 / \sigma^2) \quad (3)$$

Here, $x_{i,k}$ is the visual feature of k^{th} region in I_i , n_i is the number of regions in I_i , and σ is a smooth factor. According to the definition of $F_{i,j}$, the value of $F_{i,j}$ is positively related to the probability of finding the visual word W_j in the image I_i .

IV. CLASSIFICATION

After we get the intermediate features for images, the SVM classifier is applied to discriminate images from different categories. Here, we solve the multi-class problem using the one-against-the-rest strategy. In the training stage, an SVM classifier is trained for each category to separate that category from all the other categories. In the testing stage, the final predicted class label is decided by the winner of all SVM classifiers.

V. EXPERIMENTS

The proposed MLIR-based image categorization approach is tested on the COREL data set [6]. The data set contains 2,000 images from 20 different categories, with 100 images in each category. In [6], the image is first partitioned into non-overlapping blocks of size 4×4 . And image segmentation is based on the features of these blocks. Similarly, we only use the image thumbnails in which the image is 1/16 size of the original image.

The proposed MLIR-based image categorization approach and the existing methods [6] [8] [9] [14] are different in pixel clustering algorithm, region feature descriptor, and image classification method. To demonstrate clearly the effectiveness of describing the hierarchical structure of the image, we replace the MLIS used in our MLIR approach with the corresponding TIS, and give its experiment result. As shown in Fig. 2, the regions obtained by TIS are at a single level. We denote this approach as SLIR-based image categorization, where SLIR stands for “single level image representation”. The SLIR-based image categorization approach is the same as MLIR-based image categorization approach in terms of pixels clustering algorithm, region feature descriptor, and image classification method. However, it only uses the regions at the leaf nodes of the hierarchical structure just as the existing methods in [6] [8] [9] [14]. Since the time-consuming stage for MLIR is the pixel clustering, using the same pixel clustering algorithm, MLIR and SLIR cost nearly the same time.

In the experiments, we set the threshold of region size for segmentation stop criterion to be 0.05 times of the image size. The proper vocabulary size is related to the variety of images, and there is a trade-off between performance and speed. Here, we construct a visual vocabulary of 3000 words.

TABLE I. PERFORMANCE COMPARISON BETWEEN THE PROPOSED APPROACHES AND EXISTING METHODS

Algorithm	1000 images	2000 images
MLIR-based	85.2 : [83.9, 86.5]	74.5 : [73.8, 75.1]
SLIR-based	83.7: [82.1, 85.4]	72.4: [71.2, 73.6]
IS-MIL [9]	83.8: [82.6, 85.0]	69.3: [68.1, 70.5]
MILES [8]	82.6: [81.4, 83.7]	68.7: [67.3, 70.1]
DD-SVM [6]	81.5: [78.5, 84.5]	67.5: [66.1, 68.9]
MI-SVM [14]	74.7: [74.1, 75.3]	54.6: [53.1, 56.1]

The smooth factor for intermediate feature mapping is automatically computed, and is equal to 2 times as the average Euclidean distance between visual words. We use LIBSVM [13] to train the models, and select the RBF kernel with the default parameters in all experiments for the fair comparison.

In addition to using the complete data set with all 20 categories, we also adopt a subset of the 2000 images, which contains 1000 images from the first 10 categories. In the experiment, each category of image are randomly divided into a training set and a testing set, and each set has 50 images. We repeat each experiment for 5 random splits, and report the average of the categorization accuracy obtained over 5 different test sets together with the 95% confidence interval. The performance comparison between the proposed approaches and existing methods is presented in Table 1.

From the table, we can see that the performance of the proposed MLIR-based approach is better than that of the SLIR-based approach in the image categorization task. Specifically, the MLIR-based approach outperforms the state-of-the-art methods on the COREL data set. This is mainly because the MLIR-based approach can exploit image contents at different levels and their relationships, and yields a much richer set of cues to its semantic category. In addition, the random sampling strategy used for visual vocabulary construction in MLIR is more efficient in computation than the instance selection methods [6] [8] [9] and clustering algorithms.

We further compare the MLIR-based and the SLIR-based image categorization approaches over different visual vocabulary size. The size of visual vocabulary is set to be 500, 1000, 2000, and 3000 respectively. The experiments are carried out based on both 1000-image data set and 2000-image data set. The mean classification accuracy on five random test sets is illustrated in Fig. 3. The top two curves illustrate the results on the 1000-image data set. And the bottom two curves illustrate the results on the 2000-image data set. The experiment results show that: 1) On both data sets, MLIR-based approach outperforms the SLIR-based approach over different visual vocabulary size; 2) The advantage of MLIR-based approach becomes more significant when the data set gets more complicated from 1000 images to 2000 images; 3) Both MLIR-based approach

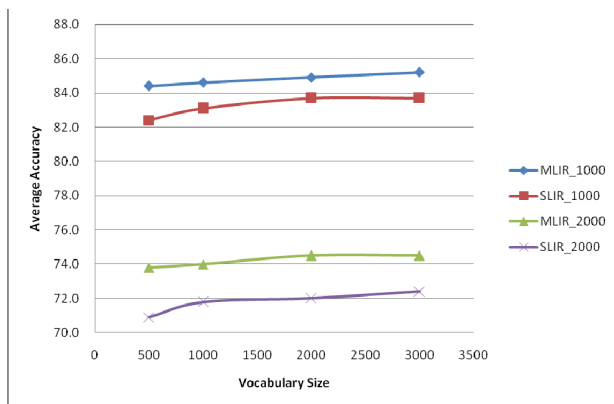


Figure 3. Mean categorization accuracy for MLIR-based approach and SLIR-based approach over five random test sets.

and SLIR-based approach perform better when the size of visual vocabulary increases. The efficiency of exploiting image contents of different levels is further confirmed by the results in Fig. 3.

VI. CONCLUSIONS AND FUTURE WORK

In this paper, we propose a novel image categorization algorithm based on exploiting image contents at different levels. Experiment results demonstrate the effectiveness of the proposed MLIR-based image categorization approach. In addition, we employ a random sampling strategy, which makes the visual vocabulary construction computationally efficient and can also get the good effectiveness.

In the future, we will further study the method of parsing and describing image hierarchical structure to improve the performance of image categorization. Since the pixel clustering in image hierarchical structure parsing is the time-consuming stage of the MLIR, our further work also includes finding the corresponding speed-up strategies.

ACKNOWLEDGMENTS

The work described in this paper was fully supported by the National Natural Science Foundation of China under Grant No. 60873154, the Beijing Natural Science Foundation of China under Grant No. 4082015, the Program for New Century Excellent Talents in University under Grant No. NCET-06-0009, and the National Development and Reform

Commission High-tech Program of China under Grant No. 2008-2441.

REFERENCES

- [1] M. Szummer and R. W. Picard. Indoor-outdoor image classification. IEEE International Workshop on Content-Based Access of Image and Video Database, 42-51, January 1998.
- [2] A. Oliva and A. B. Torralba. Modeling the shape of the scene: a holistic representation of the spatial envelop. International Journal of Computer Visual, 42(3):145-175, May 2001.
- [3] G. Csurka, C. Bray, C. Dance, and L. Fan. Visual categorization with bags of keypoints. European Conference on Computer Vision, 59-74, May 2004.
- [4] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: spatial pyramid matching for recognizing natural scene categorization. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2169-2178, June 2006.
- [5] P. Tirilly, V. Clavean, and P. Gros. Language modeling for bag-of-visual words image categorization. ACM International Conference on Image and Video Retrieval, 249-258, July 2008.
- [6] Y.-X Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. Journal of Machine Learning Research, 5:913-939, August 2004.
- [7] D. Gökalp and S. Aksoy. Scene classification using bag-of-regions representation. IEEE Conference on Computer Vision and Pattern Recognition, 1-8, June 2007.
- [8] Y.-X Chen, J.-B Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(12):1931-1947, December 2006.
- [9] Z.-Y Fu and A. Robels-Kelly. An instance selection approach to multiple instance learning. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 911-918, June 2009.
- [10] J.-B Shi and J. Malik. Normalized cuts and image segmentation. IEEE Conference on Computer Vision and Pattern Recognition, 731-737, June 1997.
- [11] J. R. Smith and S.-F. Chang. Tools and techniques for color image retrieval. Storage and Retrieval for Image and Video Databases, 2670:426-437, January 1996.
- [12] T. Ojala, M. Pietikäinen and T. Mäenpää. Multi-resolution gray-scale and rotation invariant texture classification with Local Binary Patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence, 24(7):971-987, July 2002.
- [13] C.-C Chang and C.-J Lin. LIBSVM : a library for support vector machines, 2001. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>
- [14] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. Advances in Neural Information Processing Systems, 561-568, December 2002.