A New Image Retrieval Algorithm Based on Sparse Coding

R.X. Wang
School of Natural and Applied Sciences
Northwestern Polytechnical University
Xi'an, China

G.H. Peng
School of Natural and Applied Sciences,
Northwestern Polytechnical University,
Xi'an, China

H.C. Zheng
School of Natural and Applied Sciences,
Northwestern Polytechnical University
Xi'an,China

***Abstract*—The Bag-of-visual-words (BOVW) model discards image spatial information, and the computing cost is expensive on spatial pyramid matching(SPM) model. Due to sparse coding approach exhibit super performance in information retrieval, hence, we propose a new sparse coding image retrieval algorithm. Using l_2 norm replace l_0 norm in SPM vector quantization. The local information was incorporated into sparse term by local adapter. Sparse coding was transformed into least square convex optimization proplem. Each block was encoded by the k nearest neighbor (KNN) approach, and the coding coefficients were integrated by the max pooling function. Each block required different weight according to the image itself information. Euclidean distance and the cosine theorem were combined with the similarity calculation. Our method is evaluated on the two datasets—Caltech-101 and Corel-1000. Comparing with the BOVW and SPM, the results are shown that the new approach greatly improves the image retrieval accuracy.**

Keywords-bag of visual words; spatial pyramid matching; sparse coding; image retrieval

I. INTRODUCTION

BOVW[1] represents images using local feature histograms, it is stable to the spatial translation invariance. Due to only one visual word represents a feature, which discards feature multilayered semantic information. And the feature spatial arrangement information was neglected during histogram quantification. BOVW model can not capture object shape and location. SPM model[2] divides the image into some blocks, and statisticses features histogram of each block. So it can retain spatial information. SPM model is a valid extension to disorder BOVW model [3].

Using vector quantization (VQ) solves constrained least squares fitting problem on traditional SPM, l_0 norm is as conditional constraint, image reconstruction error is greater. The multilayer semantic information of each image block can not be obtained, and the similar features are encoded as completely different codes [4] . Considering quantization loss, sparse coding (SC) [5] insteads of VQ method in this paper. In place of conditional constraint l_0 norm, we use sparse regularization term. Sparse coding can acquire lower image reconstruction error and capture more significant features. However, the norm sparse regularization term is not smooth, and codebook is often over complete [6].In order to facilitate sparsity , similar features may be encoded as totally different sparse codes. However, K. Yu, and T. Zhang show that locality is more essential than sparsity[7], locality must lead to sparsity and vice versa. In order to ensure that similar features have similar sparse codes, the local information was incorporated into sparse term by local adapter. The sparse term and the regularation term merge into a whole entry. Sparse coding was transformed into least square convex optimization proplem.It can get analytical solution and saves computing cost.The image was divided to a series of blocks by the SPM.Each block was encoded by the KNN approach, and the coding coefficients were integrated by the max pooling function.Local smooth sparsity can be reserved.

On the similarity, each block gives different weight according to the image itself information. Euclidean distance and the cosine theorem are combined with the similarity calculation.The experiments are based on the tow datasets—Caltech-101 and Corel-1000. The results are shown that proposed new approach greatly improves the image retrieval accuracy.

The remainder of the paper is organized as follows: Section 1 introduces the theory and basic idea of the sparse coding of SPM model; Section 2 explains the optimization problem of sparse regularization term of sparse coding; Section 3 states the similarity calculation; Section 4 is experiment; and Finally Section 5 is conclusions and future research issues.

II. THE VECTOR QUANTIZATION OF THE SPM MODEL

SPM model is the improvement of BOVW model. If only one layer, SPM is consistent with BOVW. SPM partitions the image more and more finer subregions. Calculate the local feature histograms from each subregion, each subregion is referred as a block, typically three layers are $1\times1, 2\times2, 4\times4$. If the pyramid layer is too large , it will

lead to the curse of dimensionality, therefore, the paper uses a typical three-layers spatial pyramid, three layer weights are $1/4$, $1/4$, $1/2$ respectively[2].

X_i is a d dimensional local descriptor (such as sift feature), $X = [X_1, X_2, \cdots,$

$X_N] \in R^{d \times N}$ is the feature representation, $B = [B_1, B_2, \cdots, B_M] \in R^{d \times M}$ is visual dictionary (codebook). Each local descriptor X_i is encoded into an M dimensional vector by the codebook B.

VQ method is that solve the formula (1):

$$\min_{B} \sum_{i=1}^{N} \min_{j=1,2,\ldots,M} \left\| X_i - B_j \right\|^2 \quad (1)$$

VQ problem is that find the optimal coefficient $C = [C_1, C_2, \cdots, C_N] \in R^{M \times N}$ of satisfying (2):

$$\arg\min_{C} \sum_{i=1}^{N} \left\| X_i - BC_i \right\|^2 \quad (2)$$

s.t. $\left\| C_i \right\|_{l^0} = 1, \left\| C_i \right\|_{l^1} = 1, C_i \geq 0, \forall i$.

The conditional constraint in formula (2) shows each descriptor X_i is represented by only one non-zero coefficient, the rest coefficients are all zero. Although encoding coefficients are sparse, the local spatial information is lost. Therefore, SPM method leads larger encoding error.

III. The Regularization Term Optimization of Sparse Coding

In order to reduce the loss of vector quantization information, The VQ is converted a standard sparse coding problem.

A. Sparse Coding (SC)

Formula (3) is a standard sparse coding problem:

$$\min_{C} \sum_{i=1}^{N} \left\| X_i - BC_i \right\|^2 + \lambda \left\| C_i \right\|_{l^1} \quad (3)$$

The original VQ is the same as BOVW model, only using one code word represents a local feature descriptor. The multilayer semantic information of a feature is neglected. The constraint $\left\| C_i \right\|_{l^0} = 1$ is canceled in the formula (3), we can use several codewords encoding a feature through the KNN approach, which can remain the multilayer semantic information.The KNN selects the K nearest neighbor codewords for reconstruction.This reduces the reconstruction error, and increases the image semantic information, here we select K = 5.

B. The Optimization of Sparse Regularization Term

In the formula (3), the sparse regularization term — l_1 norm, is not smooth, and codebook is often over complete($M >> d$). To facilitate sparsity, similar features could be encoded as totally different codes, the correlations are lost between the features. Reference[7] shows that locality is more essential than sparsity. For ensuring the similar sparse codes, local smooth sparsity was achieved by incorporating local adapter as locality constraint into sparse term, and using l_2 norm replace l_1 norm. Then sparse coding was converted into least square convex optimization proplem.It can get analytical solution and saves computing.

The l_1 norm in formula (3) alters into the l_2 norm in formula (4):

$$\min_{C} \sum_{i=1}^{N} \left\| X_i - BC_i \right\|^2 + \lambda \left\| d_i \cdot C_i \right\|^2 \quad (4)$$

where: $d_i = \exp(\frac{dist(X_i, B)}{\sigma})$, $dist(X_i, B) = [dist(X_i, B_1), \cdots, dist(X_i, B_M)]^T$, $dist(X_i, B_j)$ is the euclidean distance, σ is the local adapter to adjust the weights of the decay rate.

d_i is normalized to [0,1], and is the distance between the local feature descriptor X_i and each codeword. Its physical significance is that the more farther the distance between feature and codeword ,the more smaller reconstructed coefficient is. Extreme case is to use the nearest neighbor codeword to reconstruct, and formula (4) has analytical solution [4].

The SPM divides the image into a series of blocks. Each block is seen as a local image whole in the SPM Model, and attains a set of encoding coefficients by formula (4). In order to capture more significant features, all encoding coefficients in every block are merged using max pooling function.

Max pooling function is formula (5):

$$Z_i = \left\{ \max\left\{ \left| C_{i_{11}} \right|, \cdots, \left| C_{i_{1p}} \right| \right\}, \max\left\{ \left| C_{i_{21}} \right|, \cdots, \left| C_{i_{2p}} \right| \right\}, \max\left\{ \left| C_{i_{M1}} \right|, \cdots, \left| C_{i_{Mp}} \right| \right\} \right\} \quad (5)$$

where $\boldsymbol{Z_i}$ represents encoding of the i block, $C_{i_{jk}}$ are the encoding coefficients, it represents the j encoding coefficient of the k feature in the i block.

IV. Similarity Calculation

Similarity calculation usually have histogram matching, euclidean distance, mahalanobis distance, minkowski distance, cosine theorem, pearson correlation coefficient. Histogram matching and euclidean distance apply more widely in image retrieval [8]. The Figure 1 (a) shows histogram matching can match successfully two very different images. On the left part of Figure 1 (a),The image (1) is the query image, (2), (3) and (4) are retrieval results. On the right part of Figure 1 (a),the images are their corresponding gray histograms. Four images have similar histograms. According to the histogram matching (here is BOVW model), the similarities in (2) and (3) are higher than

(4). However,they are wrong retrieval images. In fact, only (4) is the correct retrieval result.

A variety of similarity calculation methods combination are indispensable. Euclidean distance and the cosine theorem were combined with the similarity calculation in the paper. Three layers pyramid decomposition obtain 21 blocks, the weights of 0-layer and 1-layer are $\frac{1}{4}$,the weight of 2-layer is $\frac{1}{2}$. According to the photographing habit, the 16 blocks of the 2-layer are given different weights following the structure of query image . The image center is the theme of the whole image, so the weight is relatively high. Here we adopt two kinds of structures to give the right value, as shown in Figure1 (b), where the total weight is 0.3 in 1 region , it is 0.1 in 2 region, it is 0.1 in 3 and 4 regions.According to the query image structure information, the subregions remaining weights may be zero, except the total weight of 1 region is 0.3. Similarity calculation is an important issue in image retrieval, it directly affects the retrieval result.If image attribute is different, similarity calculation is also diverse. Any similarity calculation methods inevitably affect image retrieval accuracy. Hence, the combination of euclidean distance and cosine theorem can narrow this gap.

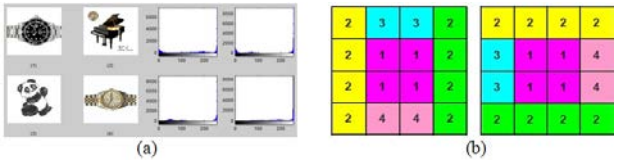

FIGURE I. RETRIVAL IMAGES HISTOGRAMS AND REGION PARTITIONS OF THE 16 BLOCKS.

V. EXPERIMENTS

Our method is evaluated on the tow datasets—Caltech-101 and Corel-1000.

A. *Caltech-101*

The Caltech-101 dataset contains 9144 images in 101 classes and one background classe,including animals, vehicles, flowers, etc, with significant variance in shape. The number of images per category varies from 31 to 800. We trained a codebook with 1024 bases, and used $1\times1, 2\times2, 4\times4$ subregions for three layers SPM.

FIGURE II. THE RETRIEVAL RESULTS OF VARIOUS METHODS,(A):BOVW;(B):SPM-BOVW; (C): SPMSC ;(D): NEW METHOD.

In the experiments, the images were resized to be no larger than 300×300 pixels with preserved aspect ratio.Compared with BOVW,SPM-BOVW and SPMSC, the results show sparse coding is more effective than BOVW. Specially, image retrieval accuracy is above 90% for face, car_side, airplanes,watches. We return advanced 20 retrieval images, shown as Figure 2.

Based on limited space, we can not list all retrieval results. We only put a small part of retrieval results list in Figure 3.

FIGURE III. THE RETRIEVAL IMAGES OF THE NEW METHOD ON CALTECH-101.

B. *Corel-1000*

The Corel-1000 dataset contains 1000 images in 10 classes.The sizes of all images are 384×256. We return partial retrieval results, shown as Figure 4.

FIGURE IV. THE RETRIEVAL IMAGES OF THE NEW METHOD ON COREL-1000.

For complex images and more background features images, such as elephants , landscapes, buildings and people in Figure 4, the proposed similarity calculation method gets better result in this paper.

C. *Evaluation criteria*

Image retrieval used precision(P) and recall(R) as evaluation criteria.

$$P = \frac{NC}{NR}, R = \frac{NC}{NA}, PR = \frac{2PR}{P+R}$$

NC is the number of retrieval correct images, NR is the total number of retrieval images,NA is the number of relevant images in database.The experiment of several classes images in Caltech-101 is shown in Table 1.

Table 1: TABLE I. RETRIEVAL ACCURACIES OF SEVERAL ALGORITHMS.

methods	average precision	error
BOVW	29.1688%	0.017921
SPM-BOVW	38.2758%	0.032998
SPMSC	52.4636%	0.018919
New method	61.6026%	0.027385

VI. CONCLUSIONS

There are three main improvements: First, we use modified spatial pyramid matching model to encode features; Second, a image block is encoded by KNN method and encoding coefficients are integrated using the max pooling function;Third, on the similarity calculation,we combined with euclidean distance and cosine theorem to give different weights according to divided image blocks. The results show that the algorithm greatly improves the image retrieval accuracy.

REFERENCES

[1] Josef Sivic &Andrew Zisserman, Video google: a text retrieval approach to object matching in videos. *Proc. of the 9th IEEE Int. Conf. On Computer Vision*, pp.1470-1477, 2003.

[2] Svetlana Lazebnik, Cordelia Schmid & Jean Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. *Proc. of the 19th IEEE Int. Conf. On Computer Vision and Pattern Recognition*, pp. 2169-2178, 2006.

[3] Jinjun Wang, Jianchao Yang & Kai Yu, etc, Locality-constrained linear coding for image classification. *Proc. of the 23th IEEE Int. Conf. On Computer Vision and Pattern Recognition*, pp.3360-3367, 2010.

[4] Jianchao Yang, Kai Yu & Yihong Gong, etc, Linear spatial pyramid matching using sparse coding for image classification. *Proc. of the 22th IEEE Int. Conf. On Computer Vision and Pattern Recognition*,pp.1794-1801, 2009.

[5] Jianchao Yang, Jiangping Wang & Thomas Huang, Learning the sparse representation for classification. *Proc. of the 12th IEEE Int. Conf. On Multimedia and Expo*, pp.1-6, 2011.

[6] Shenghua Gao, Ivor Wai-Hung Tsang & Liang-Tien Chia, Laplacian sparse coding, hypergraph laplacian sparse coding, and applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **1(35)**, pp.92-104, 2013.

[7] Kai Yu, Tong Zhang, & Yihong Gong, Nonlinear learning using local coordinate coding. *Proc. of the 23rd Annual Conf.* On Neural Information Processing Systems, pp.2223-2231, 2009.

[8] Lijia Zhi, Shaomin Zhang & Dazhe Zhao,etc, Combining similarity measures in content-based image retrieval guided by mutual information. *Image and Graphics*,**16(10)**,pp.1850-1857, 2011.