International Journal of Signal Processing, Image Processing and Pattern Recognition
Vol.9, No.7 (2016), pp.263-274
http://dx.doi.org/10.14257/ijsip.2016.9.7.23

An Image Retrieval Method Based on Visual Dictionary and Saliency Region

Xu Gongwen[a]*, Xu Lina[a], Li Xiaomei[b], Tang Xiaobing[a],
Wang Xiaoyan[c] and Xu Chunxiu[c]

[a] *School of Computer Science and Technology Shandong Jianzhu University*
[b] *Cancer Center of the Second Hospital Shandong University*
[c] *School of Computer Science and Technology Shandong Xiehe College*
xugongwen@sdjzu.edu.cn

Abstract

An image retrieval method based on the combination of visual dictionary and region of saliency was proposed in this paper, which aims to increase the accuracy of image retrieval. The image is divided into sampled blocks and the low-level features are extracted from these image blocks. Then a variety of features vectors are taken as the input vector for learning its corresponding visual dictionary respectively by non-negative sparse coding. Spatial information is added into the sparse representations of image by proposing the saliency polling method, and the similarity measure between sparse representation vectors is defined as SED (Squared Euclidean Distance), which considering the same nonzero entries and Euclidean distance of vectors at the same time. Results of experiment carried on Corel and Caltech datasets showed that this method can effectively improve the accuracy of image retrieval compared with the methods of single visual dictionary.

Keywords: *Sampled blocks, Non-negative sparse code, Saliency region, SED*

1. Introduction

With the rapid development of internet technology and the rising demand for multimedia, how to let users analyze, manage and retrieve images efficiently has become one of the hotspots for multimedia processing. At this stage, text-based image retrieval and content-based image retrieval method have been widely used. But there are still many difficulties in some researching fields, such as the “semantic gap”, which exists between low-level features and high-level semantic. The description of image that only has underlying features cannot understand the image semantic better, which lead to the low accuracy of image retrieval [1].

Therefore, researchers had put forward a series of methods of image representation and image retrieval, which could be divided into the following two aspects.

1) Learning image sparse representation

Gao *et al.* [2] adopted the combination of nuclear sparse representation and spatial pyramid matching to generate the discriminative sparse coding, which reduced the loss in feature quantization process and had a good performance. Yang *et al.* [3] proposed a new dictionary learning method that imposed the Fisher discriminant criterion on sparse coefficient, to make within class scatter smaller and between class scatter larger, so that method is more discriminative.

According to the rated visual mechanism of human brain, high-level features of image are the combination of underlying features. The feature representation from the bottom to the top is a gradually abstract process [4]. Sparse coding method can abstract high-level image features on the basis of the underlying features, which realize the mapping between high-level semantic and underlying features of image. Traditional sparse coding method randomly select a number of image sub-blocks from natural images to compose a training set, which is used to learn a series of base vectors (*i.e.*, visual dictionary), and sparse decomposition algorithm is taken to create the sparse representation of image [5-6]. Shi *et al.* [7] extracted the Sift feature of image before sparse coding with the features to achieve further abstract of that. The sparse coding coefficients can be positive or negative. Because the latter may produce negative effects to cancel each other out, non-negative sparse coding is more suitable to structure the visual dictionary [8] and is used to structure the visual dictionary in this paper. Most image retrieval methods only extract a single low-level feature of images at the feature extraction stage, which cannot meet the users search need well. Therefore, this paper mixes together a variety of underlying image features to representing images. The "semantic gap" between underlying features of image and high-level semantic can be reduced to a certain extent.

2) Image feature space quantization

Penatti *et al.* [9] proposed a spatial arrangement method of bag of words model to coding the spatial location of visual word by dividing the image into four quadrants, the spatial information was added effectively. Li *et al.* [10] proposed a new multi-level structure of the image compression coding method, semantic and spatial levels were introduced in it.

Saliency detection is an image analysis method based on global and local features, such as image brightness, color, contrast, *etc.* Saliency detection is widely used in image processing and is a hotspot of computer vision research and can effectively identify the region of interest of image and improve the accuracy of image retrieval.

Therefore, this paper combines the non-negative sparse coding and saliency analysis of image. A saliency weighted image retrieval method based on multi-visual dictionary is proposed and a new similarity measurement of sparse representation is taken in the image retrieval. Firstly, the image feature is extracted from overlap blocks and feature dictionary are learned though non-negative sparse coding. Then, saliency and spatial information is added into the representation by saliency pooling, the sparse representation vector of the whole image on multi-scale is calculated. Finally, the similarity measurement proposed in this paper is used in image retrieval.

2. Non-Negative Sparse Dictionary

Above all, training images are divided into n overlap image block, and p dimensional features can be extracted from each image block. Assumption $\boldsymbol{X} \in \Re^{n\times p}$ is the $n \times p$ dimension feature matrix of image block, which $\boldsymbol{x}_i$ is the i th column of the feature matrix $\boldsymbol{X}$(*i.e.*, the feature vector of the i th block). By minimizing the objective function in Formula (1) sparse vector set of image feature $\boldsymbol{Z} = \{\boldsymbol{z}_1, \boldsymbol{z}_2, ..., \boldsymbol{z}_n\}$ and a set of d dimensional visual dictionary $\boldsymbol{D}$ can be obtained.

$$\min_{\boldsymbol{D},\boldsymbol{Z}} \sum_{i=1}^{n} \| \boldsymbol{D}\boldsymbol{z}_i - \boldsymbol{x}_i \|_2 + \lambda \| \boldsymbol{z}_i \|_1 \quad s.t. \boldsymbol{z}_i \geq 0 \tag{1}$$

Among which, $\| \boldsymbol{D}\boldsymbol{z}_i - \boldsymbol{x}_i \|_2^2$ is the error term generated by reconstructing the feature of image blocks with visual dictionary and sparse vector. $\lambda \| \boldsymbol{z}_i \|_1$ is sparsity penalty term to ensure the sparsity of sparse coding coefficients $\boldsymbol{z}_i$, $\boldsymbol{z}_i \geq 0$ is taken to ensure sparse vector non-negative. Regularization coefficient λ is used for balancing the error term and sparse penalty terms.

When fix the coefficient $\boldsymbol{Z}$, Formula (1) is convex with respect to the variable $\boldsymbol{D}$. When fix the dictionary $\boldsymbol{D}$, Formula (1) is convex with respect to a variable $\boldsymbol{Z}$, but formula cannot optimize both at the same time. So, $\boldsymbol{Z}$ or $\boldsymbol{D}$ can be selected repeatedly and iteratively to be fixed, and optimizing another variable. Each iteration can be divided into the following two Steps [11]:

1) When the dictionary $\boldsymbol{D}$ is fixed, the objective function is minimized by adjusting the coefficient $\boldsymbol{z}_i$ (*i.e.*, solving LASSO problem).

2) When the coefficient $\boldsymbol{z}_i$ is fixed, the objective function is minimized by adjusting the dictionary $\boldsymbol{D}$ (*i.e.*, solving convex QP problem).

When a new image block is given, its sparse vector can be obtained by solving a LASSO problem with the visual dictionary optimized above, which is the sparse vector that image blocks correspond.

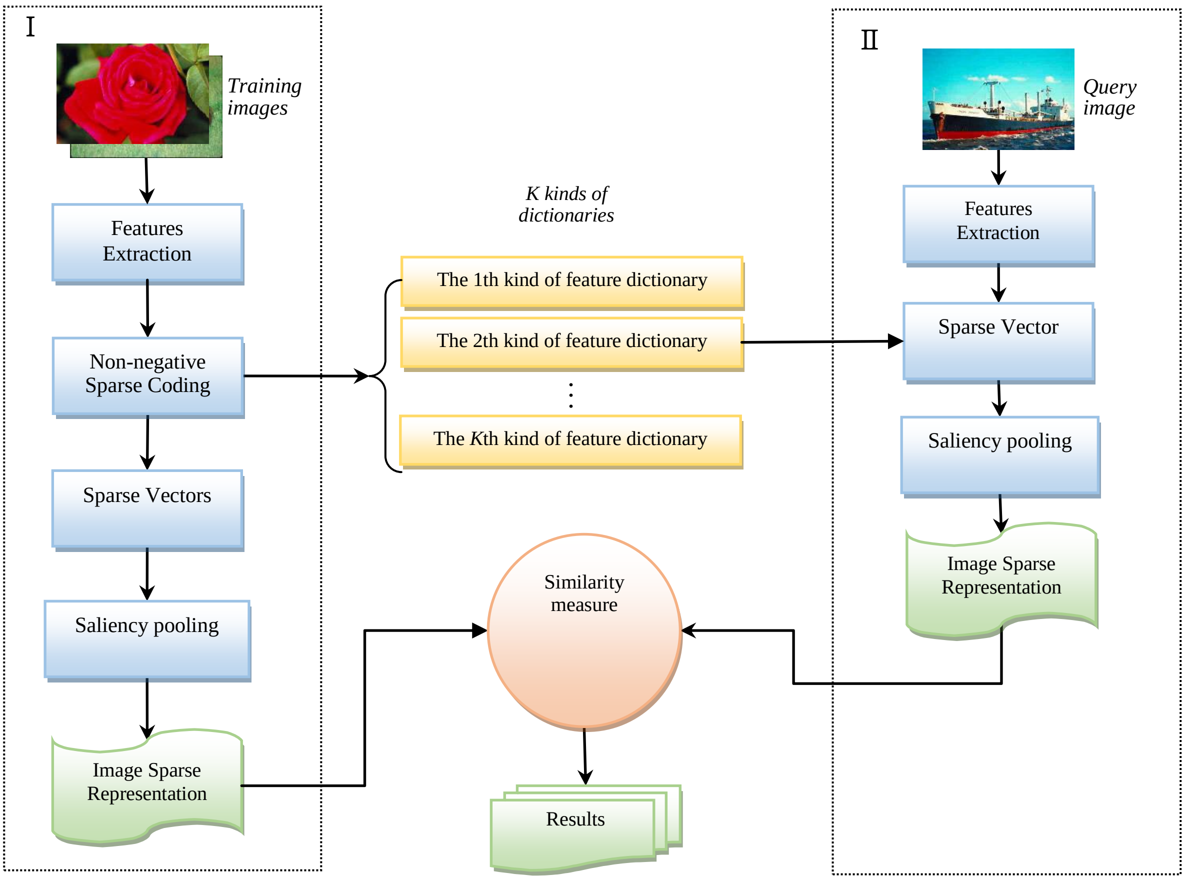


Figure 1. Procedure of the Image Retrieval Method

3. Multiple Visual Dictionary of Saliency Weighted Image Retrieval

The image retrieval method proposed in this paper includes learning visual feature dictionary and retrieving the query image, the specific process is shown in Figure 1.

In the visual dictionary learning phase, various features are extracted from the image blocks of training dataset. Each kind of features extracted is treated as the input of non-negative sparse coding in turn to learn multiple visual dictionary. Then the saliency pooling method is used to quantize the sparse vectors of each kind feature. Finally, the sparse vectors of various features are combined to be the sparse representation of training image.

In image retrieval phase, various features are extracted from the image blocks of query image. For different features of each image blocks the sparse vector corresponding to visual dictionary is obtained though the multiple visual dictionary learned before. Then the saliency pooling method is used to quantize the sparse vectors of each kind features of query image. Various sparse vectors of features are combined to be the sparse representation of query image. At last, the similarity between the query image and the training images is calculated and retrieval results are returned according to the descending order of similarity.

3.1. Saliency Pooling

In the construction of the whole image's sparse representation, if the sparse vectors of all image sub-blocks are simply linked together, the dimension of the image sparse representation is hundreds or even thousands of times of its image block sparse vector dimension, which brings a great deal of computational overhead for the similarity measure and makes the sparse representation obtained more sensitive to the changes of image scale, direction, *etc.* By the pooling method, sparse vectors of all image blocks could be summarized at multiple scales and added its spatial distribution information. The traditional pooling methods include Average-pooling and Max-pooling, *etc.* [12]. Average-pooling methods only select the means of features to represent the region of image. Max-pooling methods select the Maximum value of features to represent the region of image, which are more robust than the Average-pooling methods. But the Max-pooling methods do not consider the saliency of image, and cannot filter out the noise of background or non-objected region.

Therefore, the image sparse representation method based on the saliency weight pooling on multi-scales is proposed in this paper. Firstly, the sparse vector of each image blocks is obtained. Then, the image is divided into l^2 non-overlapping blocks by the scale l and the coefficients of blocks are weighted by the saliency value. Finally, the new vectors on all scales are combined to form image sparse representation. The formula is as follows:

$$\boldsymbol{z}_l\,' = \sum_{i=1}^{l^2}\sum_{j=1}^{d}\left(\frac{sal}{SAL_i+\varepsilon}\right)\cdot \boldsymbol{z}_{ij} \tag{2}$$

Among which, sal is the saliency value of image blocks divided in features extracting phase, SAL_i shows the saliency value of image blocks divided in pooling phase, $sal/(SAL_i+\varepsilon)$ is the contribution degree of the first term to the second term, and ε is a smoothing factor for ensuring continuous of saliency weight. Saliency detection algorithm is given by Cheng *et al.* [13]. $\boldsymbol{z}_i$ is the sparse vector of the i th image block, namely it is the j th dimension of d dimension sparse vector.

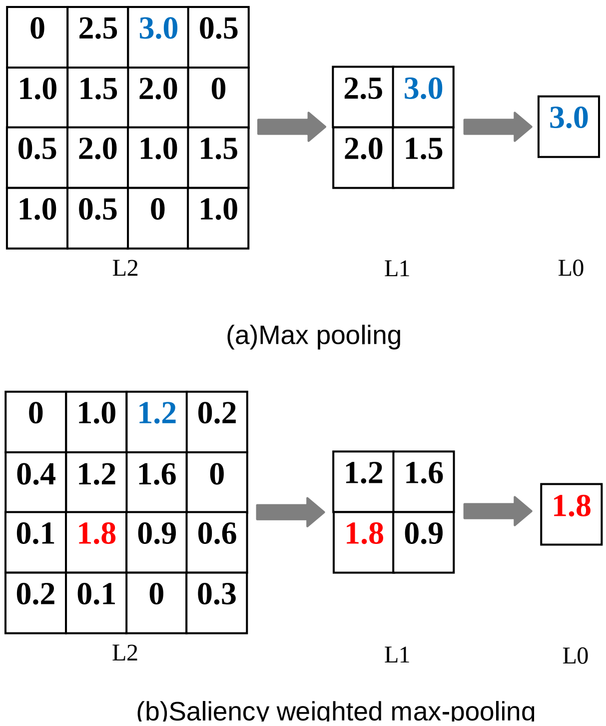


Figure 2. The Method of the Saliency Weighted Max-Pooling

Figure 2(a) shows the conventional Max-pooling method, Figure 2(b) shows the method this paper proposed. Among them, the blue one is the feature value without saliency weighted, the red one is the feature with saliency weighting. The features that exceeding the original maximum after saliency weighted are selected by saliency pooling.

As is shown in Formula (3), in order to introduce the multi-level spatial information, the sparse representation of entire image is made up by sparse vectors obtained in different scales.

$$\boldsymbol{Z}' = \sum_{l} \boldsymbol{z}'_l \tag{3}$$

3.2. Multiple Visual Dictionary Image Representation

Researches show that representing image using single feature exist the problem of poor discrimination and descriptive, and the accuracy of image representation method using a fusion of variety of features is higher than that using a single feature. Therefore, an image method based on multiple sparse visual dictionary and feature fusion is proposed in this paper [14-17].

The m kinds of different low-level features extracted in image blocks are taken to learn its corresponding visual dictionary. Thus, an image can be represented by a multiple sparse vector of features extracted in image blocks. On the assumption that $\boldsymbol{Z}'_k$ is the sparse representation of k th kind of feature of image x, which is obtained though saliency pooling with its corresponding vector, $\boldsymbol{w}_k$ is the weight of k th kind of feature sparse representation. Then, the multiple features sparse representation of image x is expressed as:

$$\boldsymbol{I}(x)=\sum_{k=1}^{m}\boldsymbol{w}_k\boldsymbol{Z}'_k \tag{4}$$

3.3. Similarity Measurement

In terms of similarity measure in image retrieval, the common similarity measures of image representation have Normalized dictionary distance (NDD) [18], Fast compression distance (FCD) [19] and *etc.* This paper proposes a method considering the position and difference between non-zero elements of sparse representation based on those measurements.

On the assumption that x is the query image, then the corresponding multi-feature sparse representation is $\boldsymbol{I}(x)$, and he sparse representation set of training image set is $\boldsymbol{Y}=\{\boldsymbol{I}(1),\boldsymbol{I}(2),...,\boldsymbol{I}(y)\}$. The formula that calculating the similarity between query image and training image set is defined as follows:

$$SED=(1-\frac{C(x\cap y)}{\max\{C(x),C(y)\}})\times\|\boldsymbol{I}(x)-\boldsymbol{I}(y)\|_2 \tag{5}$$

Among them, $C(x)$ and $C(y)$ represent the number of non-zero entries of binary sparse vector according to threshold τ. $C(x\cap y)$ is the number of elements which is not zero at the same time in the two binary sparse vector. $\|\boldsymbol{I}(x)-\boldsymbol{I}(y)\|_2$ is the L2 norm (*i.e.* the Euclidean distance).

4. Experiment

This part describes the details of the proposed method and the experiment results on several common image dataset. Corel10K and Caltech256 image dataset was taken in the experiment. 10 classes of images were chosen randomly, among which, 60 images per class for training, 10 images of each classes for testing, a total of 700 images. The configuration of experimental platform is Intel (R) Core ™ CPU 3.3GHZ, RAM 2.0G, MATLAB 2013a.

4.1. Sub-Block Division and Feature Extraction

The training images were divided into image blocks by a sliding window which was 16*16 pixels, the moving step each time was 8 pixels, and the low-level features were extracted from each image block.

The HSV (Hue, Saturation, Value) color features, Gabor texture features and Sift features were extracted in this paper. The 72-dimensional HSV feature vector was obtained by dividing the H component into 8 parts, the S component into 3 parts, and the V component into 3 parts ($\boldsymbol{L}_{hsv}=8\boldsymbol{H}+3\boldsymbol{S}+3\boldsymbol{V}$). The 48-dimensional Gabor texture features vector was the mean and variance on each orientation gradient when Gabor filters filtered the image blocks with 4 scale, 6 direction. The SIFT features have invariance to image variations, such as scale, rotate, illumination changes, and strong distinguishability [20]. The 128-dimensional dense Sift features was formed by 8 orientation gradient each small region, and the image blocks were divided into 16 small regions. Each feature vector was internal normalized to solve the problem of order of magnitude inconsistent of feature vector.

4.2. Experiment Settings

The experiment setting is shown as follows:

1) In sparse coding, the partial image blocks of each training image were selected randomly to decrease the computation burden and the low-level features of the selected image blocks were regarded as the input of visual dictionary training. Firstly, a set of d dimensional feature vectors was selected randomly to be the initial visual dictionary. Then, sparse vector $\boldsymbol{Z}$ and visual dictionary $\boldsymbol{D}$ was trained iteratively by minimizing the objective function in Formula (1), wherein the value of the regularization parameter used to balance the errors item and sparse item λ was set to 0.0001 while the highest accuracy was obtained. The experiment set the times of iteration to 50 and dimension of single dictionary to 400. Experiments show that the increasing of the dimension of dictionary would increase the accuracy of image retrieval to a certain extent, but the growth was smaller.

2) In saliency pooling, the saliency detection algorithm proposed by Cheng *et al.* [13] was taken in experiment directly. In the saliency map obtained after Grabcut, the region of saliency that less than 5% of the total image pixels were not considered in order to remove noise. The saliency value of image blocks was defined as the proportion of the region of saliency in the image block, which range is [0, 1]. The pooling scales were set to 1, 2, 4, and then the dimensions sparse vector of image after pooling is 8400. Experiments show that while the weight of the image sparse representation was [0.45, 0.23, 0.32], the image retrieval method had the best performance.

3) In similarity measurement, the binaryzation threshold of sparse vector τ is set to 0.005 and default experimental comparison method was Euclidean distance.

4.3. Corel10K Dataset

In this paper, the common performance indicators of image retrieval-recall and precision were used to detect the effectiveness of the proposed method, which were defined as follows:

$$Precision = \frac{T_M}{M},\ Recall = \frac{T_M}{N} \tag{6}$$

T_M is the number of correct Images in the image retrieval results; M is the total number of results that image retrieval returned. N is the number of images that has same semantic with the query image in the training set, here N=60.

Table 1. Performance Comparison of Different Methods in Several Dictionary Dimensions While M=30

Method		H.G.S+SC (k=1)	HSV+SC (k=1)	Gabor+SC (k=1)	Sift+SC[7] (k=1)	MSC+ED (k=3)	This paper (k=3)
d= 400	*Precision*	0.7089	0.7126	0.6732	0.5615	0.7746	**0.7938**
	Recall	0.3527	0.3615	0.3385	0.2912	0.3819	**0.4083**
d= 600	*Precision*	0.7154	0.7198	0.6853	0.5782	0.7821	**0.8132**
	Recall	0.3602	0.3593	0.3517	0.2923	0.4001	**0.4215**
d= 800	*Precision*	0.7247	0.7294	0.6897	0.5812	0.7887	**0.8365**
	Recall	0.3643	0.3712	0.3571	0.2910	0. 3965	**0.4251**

The Table 1 reflects the precision and recall of the method this paper proposed (k=3, k is the number of dictionaries) on the dataset had been enhanced respect to the method of single visual dictionary (k=1) and was increased with increasing of dictionary dimension.

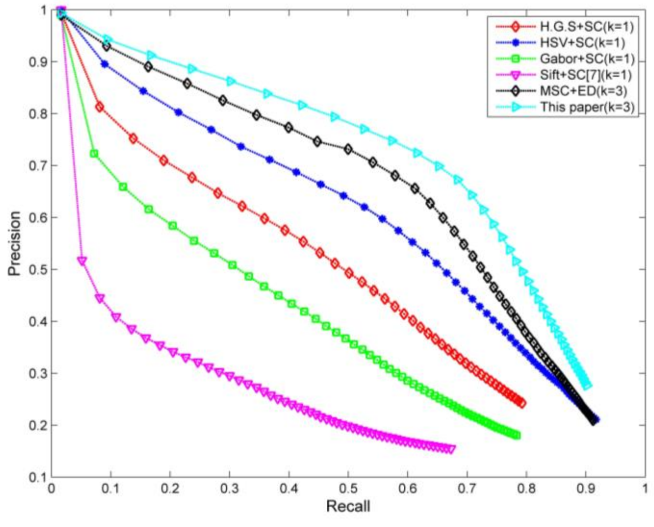


Figure 3. Precision-Recall Curves of Image Retrieval When d=400

Figure 3 shows the precision-recall curve of different image retrieval methods when the dimension of dictionary is 400. Among which, the PR curves of single feature visual dictionaries, such as HSV, Gabor and Sift dictionary, and multiple features visual dictionary method (H.G.S+SC) in Figure 3 are lower than the proposed method. In Figure 3, the proposed SDD distance could improve the performance of image retrieval at the same time.

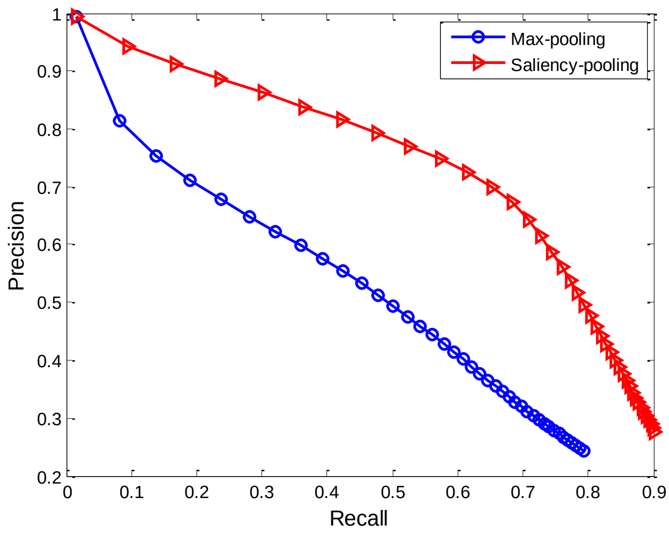


Figure 4. Retrieval Result Comparison of Saliency Weight Max-Pooling and Max-Pooling

Figure 4 reflects that the Saliency-pooling has obvious effect on increasing the accuracy of retrieval compared with the Max-pooling.

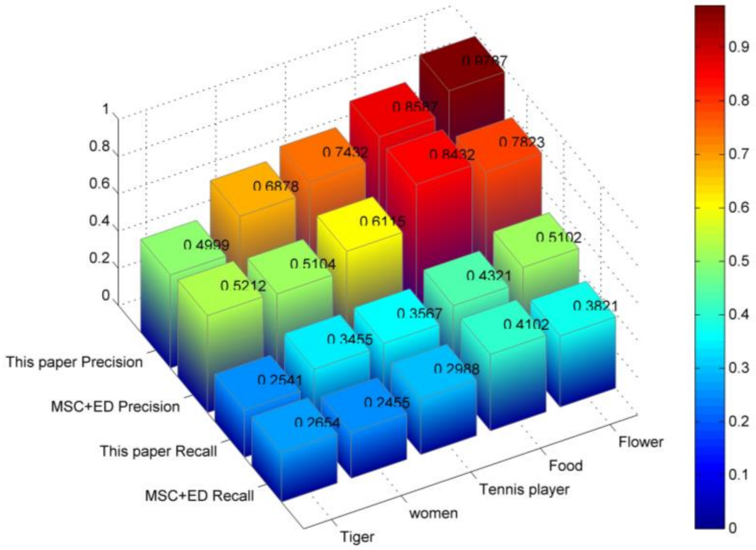


Figure 5. Retrieval Result Comparison of Several Image Sematic Classes While M=30

Figure 5 lists the average precision and recall of part of image classes in image retrieval. Among them, the retrieval results of flower and food is good, the retrieval results of tiger are poor.

The part of experimental results of image retrieval a shown in Figure 6, the leftmost image is query image, and the images behind query image are retrieval results. This figure can be seen that the precision and recall of most image classes is relatively high.

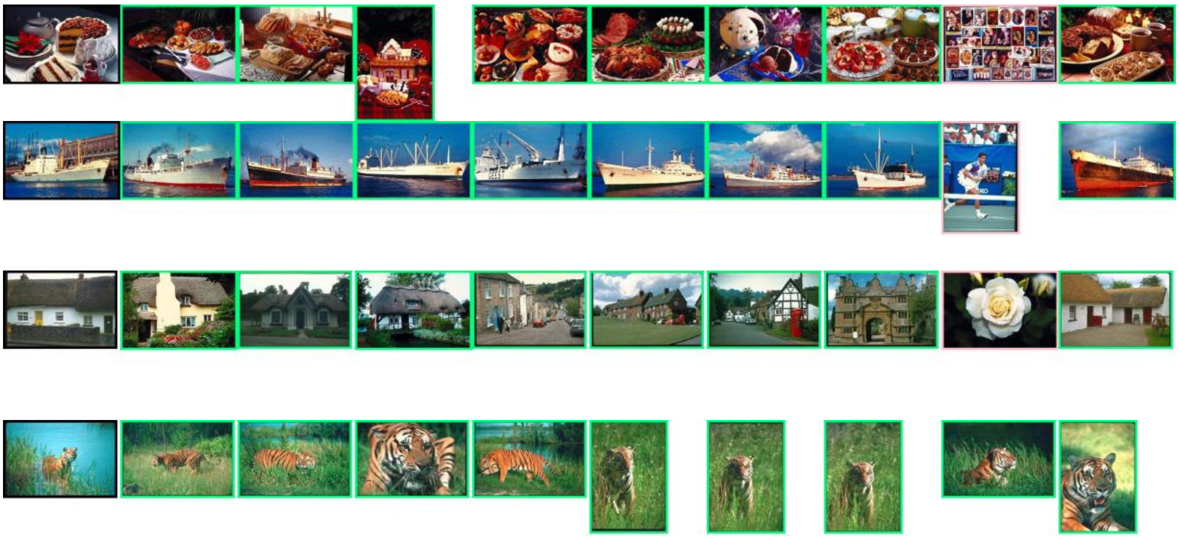


Figure 6. Parts of the Image Retrieval Results

4.4. Caltech Dataset

On the Caltech dataset, experiment used the same settings with Corel10K dataset. The experiment results show that the proposed method also showed the good performance on Caltech dataset and the specific precision rate and recall rate is shown in Table 2.

Table 2. Performance Comparison of Different Methods in Several Dictionary Dimensions on Caltech

Method		H.G.S+SC (k=1)	HSV+SC (k=1)	Gabor+SC (k=1)	Sift+SC[7] (k=1)	MSC+ED (k=3)	This paper (k=3)
d=400	*Precision*	0.6683	0.6214	0.5982	0.6754	0.6962	**0.7252**
	Recall	0.3219	0.3643	0.2823	0.3485	0.3675	**0.3951**
d=500	*Precision*	0.6731	0.6462	0.6132	0.6893	0.7051	**0.7346**
	Recall	0.3341	0.3782	0.2987	0.3607	0.3702	**0.4015**

5. Conclusion

In this paper the saliency weighted image representation method based on multiple visual dictionary is proposed by combining the non-negative sparse coding and image saliency detection. The saliency and spatial information are added into the representation by saliency pooling method and multiple visual dictionaries of image blocks' features are used to learn the sparse vector respectively by the non-negative sparse coding. The results of experiment on the Corel and Caltech dataset show that the proposed method has good performance in image retrieval. But the method only takes the combination of three features dictionaries and experiments were done on the small-scale dataset, so there will be the more future work on image retrieval.

Acknowledgements

This work is partially supported by A Project of Shandong Province Higher Educational Science and Technology Programs (J12LN31, J13LN11, J14LN59), the Shandong Xiehe College School Fund (XHXY201431), the Scientific Research Fund of the Second Hospital of Shandong University (S2015010004), the Development Projects of Science and Technology of Shandong Province (2012GGX27073, 2014GGX101011, 2015GGX101018, 2016GGE27402).

References

[1] G. Xu, Z. Zhang, W. Qi, M. Liao, L. Xu and H. Zhao, "Image Automatic Annotation Based on the Similarity of Regions", Journal of Computational Information Systems, vol. 10, 21, **(2014).**

[2] S. Gao, W. Tsang and L. Chia, "Kernel sparse representation for image classification and face recognition. The 11th European Conference on Computer Vision, Springer, Greece, (**2010)**

[3] M. Yang, L. Zhang and X. Feng, "Fisher discrimination dictionary learning for sparse representation. International Conference on Computer Vision, IEEE. (**2011**), Barcelona, Spain.

[4] D. Hubel and T. Wiesel, "Receptive fields binocular interaction and functional architecture in the cat's visual cortex", The Journal of physiology, vol. 1, no. 160, **(1962)**.

[5] M. Yang, L. Zhang and J. Yang, "Robust sparse coding for face recognition. Conference on Computer Vision and Pattern Recognition", IEEE, (**2011**), Providence, US.

[6] K. Yu, Y. Lin and J. Lafferty, "Learning image representations from the pixel level via hierarchical sparse coding. Conference on Computer Vision and Pattern Recognition", IEEE, Providence, US, (**2011**).

[7] J. Shi, Z. Jiang and H. Feng, "SIFT-based Elastic sparse coding for image retrieval", The 19th International Conference on Image Processing, IEEE, Orlando, US, (**2012**).

[8] A. Jiang, C. Wang and B. Xiao, "An Image Feature Extraction Method Based on GaborSIFT+NNScSPM", Acta Automatica Sinica, vol. 10, no. 37, **(2011)**.

[9] O. Penatti, F. Silva and E. Valle, "Visual word spatial arrangement for image retrieval and classification", Pattern Recognition, vol. 2, no. 47, **(2014)**.

[10] L. Li, J. Zhu and H. Su, "Multi-level structured image coding on high-dimensional image representation", The 11th Asian Conference on Computer Vision, Springer, Korea, **(2013)**.

[11] J. Yang, K. Yu and Y. Gong, "Linear spatial pyramid matching using sparse coding for image classification", Conference on Computer Vision and Pattern Recognition, IEEE, Miami, US, **(2009)**.

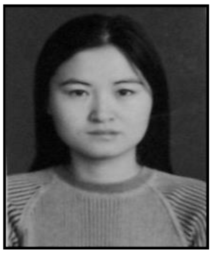
[12] J. Yang and M. Yang, "Learning hierarchical image representation with sparsity, saliency and locality. Proceedings of the British Machine Vision Conference, Dundee, British, **(2011)**.

[13] M. Cheng, G. Zhang and N. Mitra, "Global contrast based salient region detection", Conference on Computer Vision and Pattern Recognition, Providence, US, **(2011)**.
[14] Y. Fan, "Semantic annotation and storage for tourism information", Review of computer engineering studies, vol. 2, no. 2, **(2015)**.
[15] L. Sun, Y. Tang and H. Zhang, "An effective image retrieval method based on multi-features", Journal of Software, vol. 4, no. 9, **(2014)**.
[16] S. Liu, S. Li and J. Feng, "Remote sensing image classification based on adaptive fusion of multiple features", Journal of Data Acquisition and Processing, vol. 1, no. 29, **(2014)**.
[17] C. Wang and S. Zhu, "A design of fpga-based system for image processing", Review of computer engineering studies, vol. 1, no. 2, **(2015)**.
[18] A. Macedonas, D. Besiris and G. Economou, "Dictionary based color image retrieval", Journal of Visual Communication and Image Representation, vol. 19, **(2008)**.
[19] D. Cerra and M. Datcu, "A fast compression-based similarity measure with applications to content-based image retrieval", Journal of Visual Communication and Image Representation, vol. 23, **(2012)**.
[20] D. Zhang, M. Islam and G. Lu, "A review on automatic image annotation techniques", Pattern Recognition, vol. 45, **(2012)**.

Authors

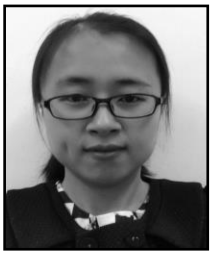
Xu Gongwen, He received his Master's degree in 2005 from Shandong University. Now he is pursuing the Ph. D. degree in Shandong Normal University. He works in School of Computer Science and Technology Shandong Jianzhu University. His current research interests include image processing, spam detection, machine learning.

Xu Lina, She received her Master's degree in 2006 from Shandong University. She works in School of Computer Science and Technology Shandong Jianzhu University. Her current research interests include spam detection, statistical sparse learning.



Li Xiaomei, She received her Master's degree in 2004, Ph. D. degree in 2014 from Shandong University. She is a member of Cancer Center of the Second Hospital, Shandong University. Her research interests include medical image processing and machine learning.

Tang Xiaobing, He received her Master's degree in 2005 from Shandong University. He works in School of Computer Science and Technology Shandong Jianzhu University. His current research interests include image retrieval, machine learning.



Wang Xiaoyan, She works in School of Computer Science and Technology Shandong Xiehe College. Her current research interests include image retrieval, machine learning.

Xu Chunxiu. She works in School of Computer Science and Technology Shandong Xiehe College. Her current research interests include image retrieval, information processing.