Image Retrieval Using Point- and Block-Based Visual Vocabulary

Chang-Ming Kuo, Nai-Chung Yang, Chung-Ming Kuo*, Liang-Kang Huang

Department of Information Engineering, I-Shou University Dashu, Kaohsiung,Taiwan
*corresponding author, e-mail: kuocm@isu.edu.tw

Abstract

With the rapidly evolving computer technologies, the multimedia and vision applications, such as visual recognition, scene modeling, image retrieval, and image categorization attract significant attention. The visual words, a collection of local features of images, can be used to represent image information. By considering the content homogeneity of visual words, we design a visual vocabulary which contains macro-based and micro-based corresponding to feature points and key blocks visual words, respectively. We also apply the proposed visual vocabulary and description scheme to construct an image retrieving system. The performance evaluation of the systems indicates that the proposed visual vocabulary achieves promising results.

Keywords: Visual words, Image retrieval, Macro-based, Micro-based

1. Introduction

Recently, visual vocabulary representation approach has been successfully applied to many multimedia and vision applications, such as visual recognition [1][5], image retrieval and scene modeling/categorization [2][4]-[6], because its richness of local information and robustness to occlusions, geometric deformations and illumination variations. To construct a visual vocabulary, a set of selected image samples are fist trained. The training samples can be obtained by point-based method or block-based method. Each feature points or blocks are described by a feature vector. Then all the feature vectors grouped into a number of clusters by using a clustering algorithm such as *K*-means. Once the feature vectors are clustered, the visual words are defined as the cluster centers to represent image feature. The feature vectors falling in the same cluster are considered as the same visual word. Therefore, the representation feature extracted from the training images, in analogy to text file, is known as visual words. This strategy achieves high accuracy since large number of local information can be well-defined so that it can effectively describe the images.

The purpose of this work is to construct a new visual vocabulary for the applications of image retrieval. The main idea of proposed method is taking into account the inhomogeneous and incomplete content of visual words, we develop a new approach that combines feature points and key blocks visual words to precisely describe macro and micro semantics in images. We will also briefly discuss the advantages and disadvantages of different types of visual words and then construct a visual vocabulary accordingly.

2. Construction of Visual Vocabulary for Image Description and Categorization

Generally, the point-based and block-based visual words that are two different words types are most widely used to construct visual vocabulary. The point-based method contains three steps: 1) extract feature points; 2) define local descriptor for each feature point; and 3) construct visual vocabulary. Another simpler approach is the block-based method, which equally partition image into small non-overlapping blocks, i.e., samples of visual words, and does not need extraction procedure.

In this paper, a novel feature point and key block visual words representation approach will be developed for the applications of image retrieval. For describing macro content, the block-based visual words are used to represent the whole and global sense of visual perception. On the other hand, for describing micro content, point-based visual words are used to represent the detail of image content. Since the proposed method can represent an image according to its content, it achieves high performance in retrieving. In our work, a new image description scheme that integrates the advantages of point- and block-based approaches will be introduced. The overall system architecture structure of visual vocabulary is illustrated in Fig. 1.

To construct macro and micro-based visual vocabulary, the background and foreground should be properly separated. For simplicity, we scan the image block by block to identify each block belonging to macro or micro sense. The macro block is characterized with smooth content, and the micro block is with high activity content such as edges or obvious textures. Fig. 2 is an example to illustrate the concept.

A. Micro-based visual description

In order to capture the characteristics in micro image content, we select the SIFT to extract the feature points, and then define the corresponding feature descriptor. Conventional SIFT descriptor has two disadvantages need to be addressed. First, the dimension of feature descriptor is very high; second, it has no color information. Therefore, we will propose a new scheme to improve it.

In conventional SIFT descriptor, a 16×16 surrounding region is defined for each feature point, and then the 16×16 region is divided into sixteen 4×4 sub-region. In each sub-region, the directional histogram includes 8 different directions, thus the descriptor's dimension is 8×4×4=128. The descriptor is used to present the local feature for feature point in image. To reduce the dimension, we set the surrounding region to 8×8 block, and its directional histogram includes 8 different directions as the SIFT descriptor. Therefore, the dimension of the proposed descriptor is reduced to 8, but the statistics are more convincing.

Since the SIFT descriptor only considers the low level physical characteristics such as the magnitude or direction of gradient in local region, the results cannot match the perception of human visual system well due to the lack of color information. In our work, the color feature will be considered in the proposed descriptor. According to the main direction of feature point, we calculate the color histogram of those pixels in surrounding region located on the main direction.

B. Marco-based visual description

For the blocks belonging to macro sense content, they are suitable to be described by block-based visual words. The macro content of an input image is partitioned into N blocks; each block is labeled by the index of nearest visual words in the macro sense visual vocabulary with size of CM; that is

978-1-4799-4208-4/15/$31.00 ©2015 IEEE

$$L\left(B_s^c(n)\right)=k=\arg\min_k\left(\left\|B_s^c-d_k\right\|\right)_{k=1...C_M,n=1...N} \quad (1)$$

where $L(\bullet)$ is labeling function, $B_s^c(n)$ is the input image block, and d_k is the kth visual word in macro sense vocabulary. After labeling, we can calculate the histogram of the labels of the input image by Eq. (5) and (6).

$$h_s^c=\left\langle h_s^c(1),h_s^c(2),...h_s^c(C_M)\right\rangle$$

$$h_s^c=1/N\sum_{n=1}^{N}\delta\left(L\left(B_s^c(n)\right)-k\right),k=1...C_M \quad (2)$$

Since each label corresponds to a visual word, the macro content of an image can be reconstructed with visual words corresponding to the labels obtained from Eq. (1).

C. Building macro and micro-based visual vocabulary

We first collect a large amount of training samples to construct the representative visual vocabulary for each class. The training images of each class contain various variations including luminance change and contrast change. In our work, the training images are represented as $T=\{\ I_s^c\ |s=1,\ldots,S,\ c=1,\ldots,C\}$, where s and c represent image sample and image class, respectively. Therefore the number of total training images are $T=C\times S$. As shown in Fig. 1, each training image I_s^c is separated into macro and micro content. Then the macro- and micro-based visual vocabularies are constructed accordingly.

For micro- and macro-based visual vocabulary, the SIFT and block partitioning are applied to extract the visual words candidates, respectively. Therefore, huge visual words candidates are extracted from training images in which high redundancy exists in them. To reduce the redundancy for obtaining a good visual vocabulary, the merging procedure is used to obtain a representative visual vocabulary.

D. Image description and similarity measure

The input block is first categorized as macro-sense or micro-sense by checking their content variation. If the input block belongs to macro-sense, it is labeled with macro vocabulary; otherwise with micro sense vocabulary.

The image description is the combination of macro sense and micro sense histogram, and can be expressed as

$$H_s^c(q)=\left(H_i^{MAC}(q),H_i^{mic}(q)\right) \quad (3)$$

where $i=1,..,N_{MAC}$, $j=1,\ldots N_{mic}$. We can define the similarity of images as

$$S^{MAC(or\ mic)}(q,l)$$

$$=\sum_{i=1(or\ j=1)}^{N_{MAC(or\ mic)}}\left(1-\left|H_{i(or\ j)}^{MAC(or\ mic)}(q)-H_{i(or\ j)}^{MAC(or\ mic)}(l)\right|\right) \quad (4)$$

$$\times\min\left(H_{i(or\ j)}^{MAC(or\ mic)}(q)-H_{i(or\ j)}^{MAC(or\ mic)}(l)\right)$$

$$S(q,l)=w_1\times S^{MAC}(q,l)+w_2\times S^{mic}(q,l) \quad (5)$$

where $S^{MAC}(q,l)$ is the similarity of image q and l in macro sense histogram, and $S^{mic}(q,l)$ is the micro sense similarity; w_1,w_2 are the weighting value, and $S(q,l)$ is the overall similarity.

In our work, the histogram similarity measure of visual words is obtained by modifying our previous work [7], which has been verified superior to state-of-the-art approaches for image retrieval by extensive simulations.

Experimental Results

We use a database (31 classes, 3901 images) from Corel's photo to test the performance of the proposed method. To evaluate the performance of image retrieval, two popular performance indexes ARR (Average Retrieval Rate) and ANMRR (Average Normalized Modified Retrieval Rank) were selected as quality measure. An ideal performance will consist of ARR values equal to 1 for all values of recall. A high ARR value represents a good performance for retrieval rate, and a low ANMRR value indicates a good performance for retrieval rank.

In the following, the comparison of the proposed and some typical methods is listed for evaluating the performance and effectiveness. Table 1 shows the performance ARR/ANMRR for conventional SIFT descriptor; SIFT descriptor combined with color histogram, MPEG-7 DCD and color descriptor. It can be seen that the conventional SIFT descriptor is the worst one due to the lack of color information; however, its ARR/ANMRR will be improved significantly when the color information is considered. The simulation results indicate that the proposed method achieves the best ARR/ANMRR because it considers the background information.

Conclusion

In this paper, we have proposed a systematical approach that constructs a discriminative visual vocabulary with macro and micro sense of visual words. We also present an effective image description method based on the macro and micro visual vocabulary. In order to evaluate the performance of proposed visual vocabulary, the image retrieval is extensively simulated. The experiments indicate the visual vocabulary achieves promising results for retrieval. Therefore, we can conclude that the proposed visual vocabulary can effectively extract the visual features from images. In the future, advanced image categorization methods based on the proposed visual vocabulary will be further studied.

Acknowledgement

This work was supported by the National Science Counsel Granted NSC 102-2221-E-214-040-

Reference

[1] L. Wu, S. C. H. Hoi, and N. Yu, "Semantics-Preserving Bag-of-Words Models and Applications," *IEEE Transactions On Image Processing,* Vol. 19, No. 7,pp.1908-1920, July 2010.

[2] F. Perronnin, "Universal and Adapted Vocabularies for Generic Visual Categorization," *IEEE Transactions On Pattern Analysis And Machine Intelligence,* Vol. 30, No. 7, pp.1243-1256, July 2008.

[3] R. Ren and J. Collomosse, "Visual Sentences for Pose Retrieval over Low-Resolution Cross-Media Dance Collections," *IEEE Transactions On Multimedia,* Vol. 14, No. 6, pp.1652-1661, December 2012.

[4] J. Qin and N. C. Yung, "Scene categorization via contextual visual words," *Pattern Recognition 43 (2010)*,pp.1874–1888, November 2009.

[5] R.J. López-Sastre, T. Tuytelaars, F.J. A. RodrÍguez and S. M.Bascón, "Towards a more discriminative and semantic visual vocabulary," *Computer Vision And Image Understanding 115 (2011)*, pp.415–425, November 2010.

[6] A. Bolovinou, I.Pratikakis and S.Perantonis, "Bag of spatio-visual words for context inference in scene classification," *Pattern Recognition 46 (2013)*, pp.1039–1053, September 2012.

[7] N. C. Yang, W. H. Chang, C. M. Kuo, and T. H Li, "A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval," *Journal of Visual Communication and Image Representation*, Vol. 19, pp. 92-105, February. 2008.

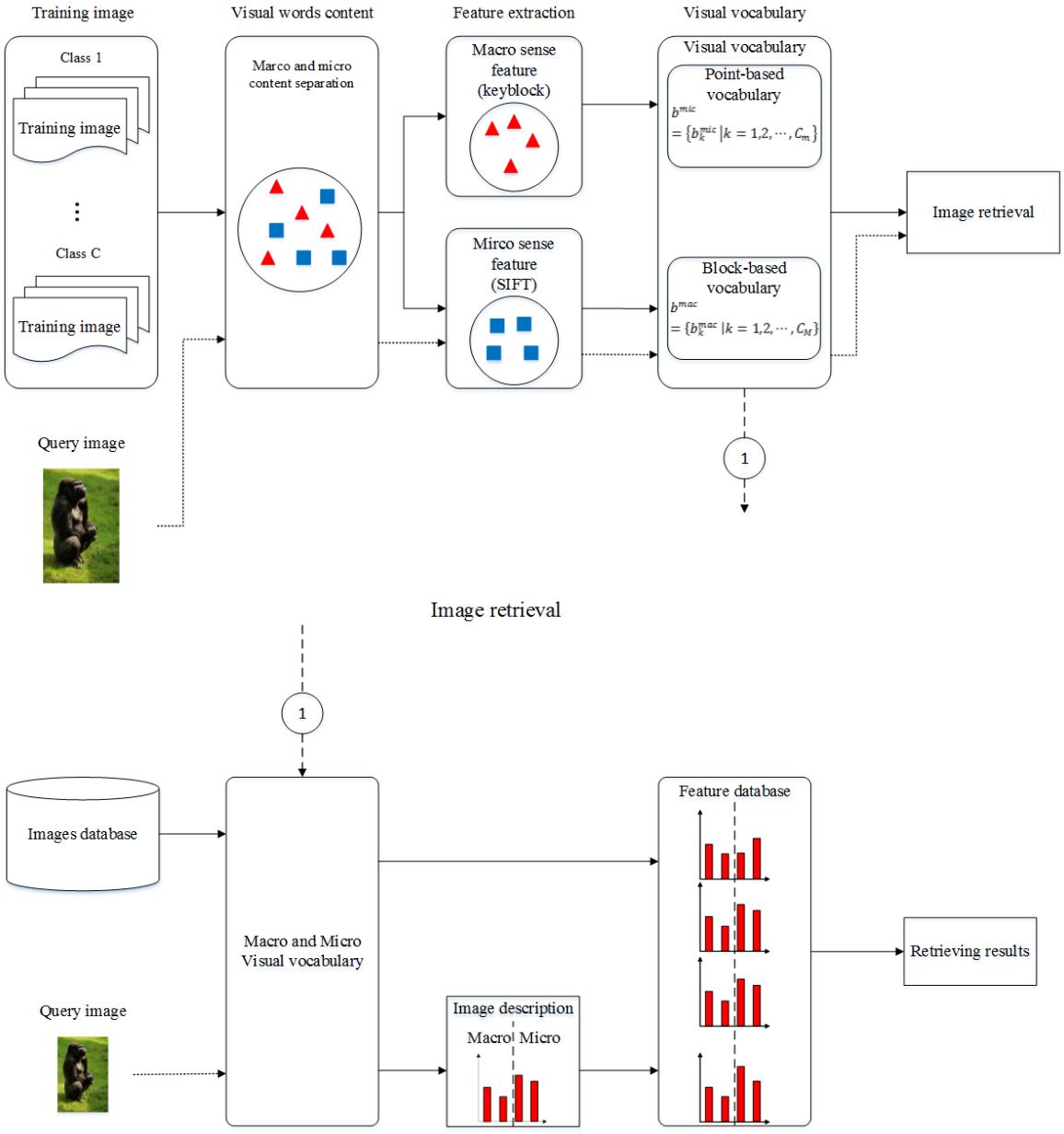

Fig. 1 The proposed architecture of visual vocabulary construction and image retrieval

(a) (b) (c)

Fig. 2 Example for foreground and background (a) original image, (b) background of the image, (c) foreground of the image.

Table 1 Comparison of different methods on ARR/ANMRR performance

Method (index)	ARR/ANMRR

Image Categories		SIFT		SIFT+Color		DCD		LBA[3]		Proposed	
1	Orangutans	0.02932	0.95839	0.11616	0.84369	0.172408	0.771103	0.205576	0.720271	0.2663	0.654
2	painting of birds	0.06283	0.91793	0.23691	0.70041	0.267037	0.63575	0.441235	0.43212	0.3574	0.53931
3	Pot plant	0.12069	0.84291	0.19501	0.75352	0.126505	0.827006	0.167612	0.786425	0.19183	0.75715
4	Card	0.07475	0.88590	0.24023	0.6941	0.431848	0.462517	0.73015	0.155727	0.82406	0.08989
5	Cloud	0.22202	0.71384	0.21547	0.71175	0.150849	0.797512	0.142361	0.810044	0.16145	0.78526
6	Sunset	0.21185	0.71894	0.21262	0.71825	0.18107	0.755743	0.146624	0.79876	0.15256	0.79475
7	Pumpkin	0.03460	0.95458	0.06714	0.90384	0.069731	0.905154	0.135331	0.817084	0.26962	0.65987
8	Cake and cookie	0.02428	0.96762	0.1053	0.85816	0.142653	0.803189	0.168571	0.771761	0.21673	0.71425
9	Dinosaur	0.01290	0.98271	0.5616	0.34107	0.3825	0.516499	0.7598	0.095861	0.7555	0.09717
10	Wheel and dolphin	0.22589	0.72396	0.18253	0.77106	0.208652	0.725244	0.236507	0.699198	0.21885	0.71777
11	Elephant	0.04409	0.94091	0.1139	0.84562	0.094014	0.862039	0.164571	0.781522	0.20529	0.72969
12	Firework	0.09885	0.87515	0.37146	0.53964	0.736101	0.161975	0.891131	0.069976	0.81884	0.09208
13	Flower	0.04484	0.92364	0.10281	0.85622	0.174423	0.775077	0.117788	0.837276	0.18791	0.75285
14	Vegetable and fruit	0.23155	0.70176	0.1957	0.74721	0.099516	0.860052	0.141453	0.815388	0.16456	0.78415
15	Ceramic duck	0.04539	0.92833	0.1938	0.74649	0.2344	0.682917	0.438	0.46208	0.4993	0.36978
16	Leopard	0.17178	0.76508	0.29547	0.62222	0.223009	0.690686	0.234425	0.687509	0.35612	0.53929
17	Leaf	0.15851	0.78628	0.27786	0.64719	0.306179	0.593291	0.275206	0.621669	0.36082	0.52526
18	Car	0.01977	0.97147	0.14868	0.79898	0.082668	0.88202	0.097468	0.860129	0.12117	0.83728
19	Cactus	0.05339	0.92351	0.09708	0.86629	0.184947	0.740643	0.193857	0.738161	0.21437	0.70417
20	Airplane	0.14301	0.78526	0.18843	0.74037	0.132485	0.826501	0.156629	0.79235	0.18484	0.75421
21	Mural	0.06722	0.90796	0.20912	0.73362	0.275402	0.642684	0.358932	0.554504	0.30257	0.62501
22	Sea animal	0.01949	0.97056	0.09478	0.87312	0.093361	0.871631	0.104219	0.857358	0.11792	0.83992
23	Horse	0.05040	0.93670	0.06049	0.91936	0.086326	0.88007	0.081285	0.886098	0.10354	0.8603
24	Helicopter	0.05301	0.91545	0.11017	0.84575	0.144105	0.796993	0.146238	0.798364	0.14222	0.80387
25	Ship	0.05395	0.92525	0.07342	0.89811	0.108269	0.844651	0.126974	0.823867	0.10518	0.85094
26	Snow	0.10375	0.85706	0.12212	0.83721	0.217592	0.69956	0.210697	0.709595	0.24635	0.66605
27	Hot air balloon	0.06058	0.89184	0.13946	0.80368	0.087353	0.882675	0.113339	0.85063	0.15924	0.79125
28	Waterfall	0.07640	0.89480	0.10224	0.8595	0.161511	0.776893	0.180808	0.765502	0.19653	0.7322
29	architecture	0.06437	0.91382	0.09316	0.87212	0.10412	0.852389	0.11821	0.831196	0.15046	0.79547
30	Sports field	0.28925	0.64323	0.46487	0.45254	0.173037	0.766169	0.394112	0.510708	0.51084	0.3861
31	Person	0.19769	0.70311	0.27547	0.62299	0.138227	0.806578	0.162618	0.773891	0.15631	0.786
	Average	**0.09891**	**0.86541**	**0.18914**	**0.75561**	**0.193235**	**0.745007**	**0.252959**	**0.68113**	**0.28124**	**0.64629**