Journal of Information Hiding and Multimedia Signal Processing ©2016 ISSN 2073-4212
Ubiquitous International Volume **7**, Number **6**, November **2016**

Image Retrieval Using Macro- and Micro-Based Visual Vocabulary

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

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

Received March, 2016; revised June, 2016

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. Because images contain very diverse contents, training limited visual words for representing various contents with high reliability is difficult. In this work, we will introduce a new scheme that divides the visual words into two types based on the analysis of visual word contents. 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. The two types of visual words are appropriately further combined to describe an image effectively. We also apply the new approach 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.** There are growing computer applications and the Internet-based services in the past two decades. A variety of download offerings and subscription services that make huge amounts of digital contents are available. Effective solutions on image indexing, retrieving and categorization have become urgent to allow users to access the information through the internet. As in conventional text files, the early solutions had been focused on developing text-based image retrieval. In these approaches [1], images are first manually annotated by texts, and then the users can retrieval images through using management system of image database. The approaches are usually heuristic and familiar for most users; however, the annotating process for large database is very time-consuming. Additionally, the subjectivity of human perception may result in a great difference in annotations for the same image. In recent years, an increasing level of research interest has concentrated on the content-based image retrieval (CBIR) techniques [2-5], which were introduced to meet with higher success. Typically, CBIR techniques are based on global features of images. Image contents, such as color, texture, and shape, are automatically extracted from the image and then be used to represent the characteristics of the image. However, the existence of semantic gap between low-level and high-level features usually leads to some uncertainty in determining the users' demand.

Recently, visual vocabulary (or bag-of-visual words) representation approach has been successfully applied to many multimedia and vision applications, such as visual recognition [3] [7], image retrieval [9] and scene modeling/categorization [5] [8], because its richness of local information and robustness to occlusions, geometric deformations and illumination variations [5]-[9]. To construct a visual vocabulary, a set of selected image samples are fist trained. The training samples can be obtained by point-based method [3] or block-based method [4][9]. Each feature points or blocks are described by a feature vector [3]-[9]. 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.

The paper is organized as follows. The problems about visual vocabulary are formulated in Section 2. In section 3, the proposed visual vocabulary construction method is discussed. The performance evaluation is presented in Section 4, and finally the conclusion is drawn in Section 5.

2. **Brief Analysis.** 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.

Besides the two types of visual word, the construction of visual vocabulary is a key procedure to make the image retrieval efficient. Some of the concerns regarding the two types of visual word are as follows. Typically, the point-based visual word represents the low level characteristics such as the magnitude or direction of gradient in local regions. The block-based visual word characterizes the real image content to represent images; this allows the visual words to preserve more high-level image characteristics inside. We can conclude that the point-based method is effective for the image details, which are usually with significant variation, and the block-based method is more suitable for the flat regions, which are usually with smooth and slow variation.

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.

3. **Construction of Point-Based Visual Vocabulary.** In general, an image can be partitioned into foreground and background as shown in Fig. 1. Because foregrounds usually contain significant variation, their fine details can be considered as micro sense feature. On the other hand, the backgrounds contain much smoother regions, which are considered as global or macro features of image. In order to increase the accuracy, we consider both micro and macro features in image simultaneously.

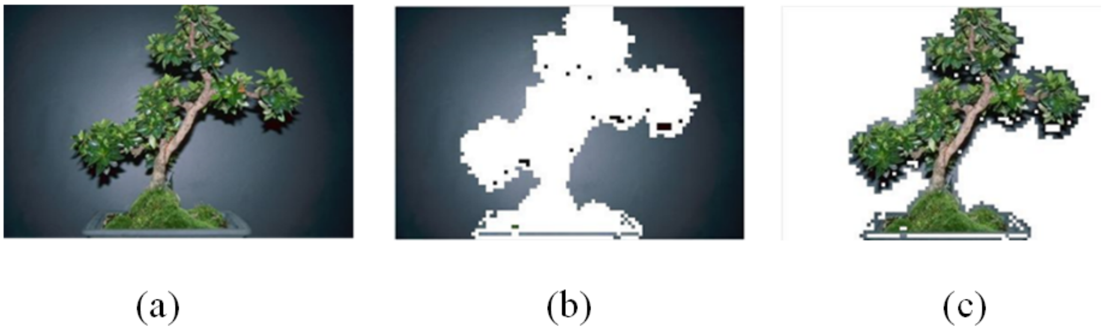

Figure 1. Example for foreground and background (a) training image, (b) background of the training image, (c) foreground of the training image

According to the analysis in Section 2, the point-based and block-based visual words are suitable for describing micro and macro features, respectively. 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. 2. It includes two major components: visual vocabulary construction, image retrieval.

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. 3 is an example to illustrate the concept.

For all blocks b, we sort their RGB values in an ascending order, and then calculate the difference vectors $(V_R(d_i), V_G(d_i), V_B(d_i))$ using Eq. (1)

$$\begin{aligned} V_R\left(d_i\right) &= \sqrt{\left(\left[\frac{d_i^{R8}+d_i^{R9}}{2}\right]-\left[\frac{d_i^{R1}+d_i^{R2}}{2}\right]\right)^2+\left(\left[\frac{d_i^{R15}+d_i^{R16}}{2}\right]-\left[\frac{d_i^{R8}+d_i^{R9}}{2}\right]\right)^2} \\ V_G\left(d_i\right) &= \sqrt{\left(\left[\frac{d_i^{G8}+d_i^{G9}}{2}\right]-\left[\frac{d_i^{G1}+d_i^{G2}}{2}\right]\right)^2+\left(\left[\frac{d_i^{G15}+d_i^{G16}}{2}\right]-\left[\frac{d_i^{G8}+d_i^{G9}}{2}\right]\right)^2} \\ V_B\left(d_i\right) &= \sqrt{\left(\left[\frac{d_i^{B8}+d_i^{B9}}{2}\right]-\left[\frac{d_i^{B1}+d_i^{B2}}{2}\right]\right)^2+\left(\left[\frac{d_i^{B15}+d_i^{B16}}{2}\right]-\left[\frac{d_i^{B8}+d_i^{B9}}{2}\right]\right)^2} \end{aligned} \tag{1}$$

The macro or micro sense block content are then determined by

$$b_i \in \begin{cases} B^{Mac},\ if\ max\left(V_R\left(d_i\right), V_G\left(d_i\right), V_B\left(d_i\right)\right) \le T \\ \qquad\qquad B^{mic},\ else \end{cases} \tag{2}$$

where B^{MAC} and B^{Mic} , are macro sense and micro sense blocks, respectively. The max() is a function that selects maximum element in (); and T is a threshold, which is determined by extensive experiments. Fig. 4 is an example, which shows the classified results with T=15. The blocks labeled with macro-sense are replaced by white blocks, and the blocks labeled with micro-sense are not changed. Obviously, it achieves satisfactory results.

3.1. **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.

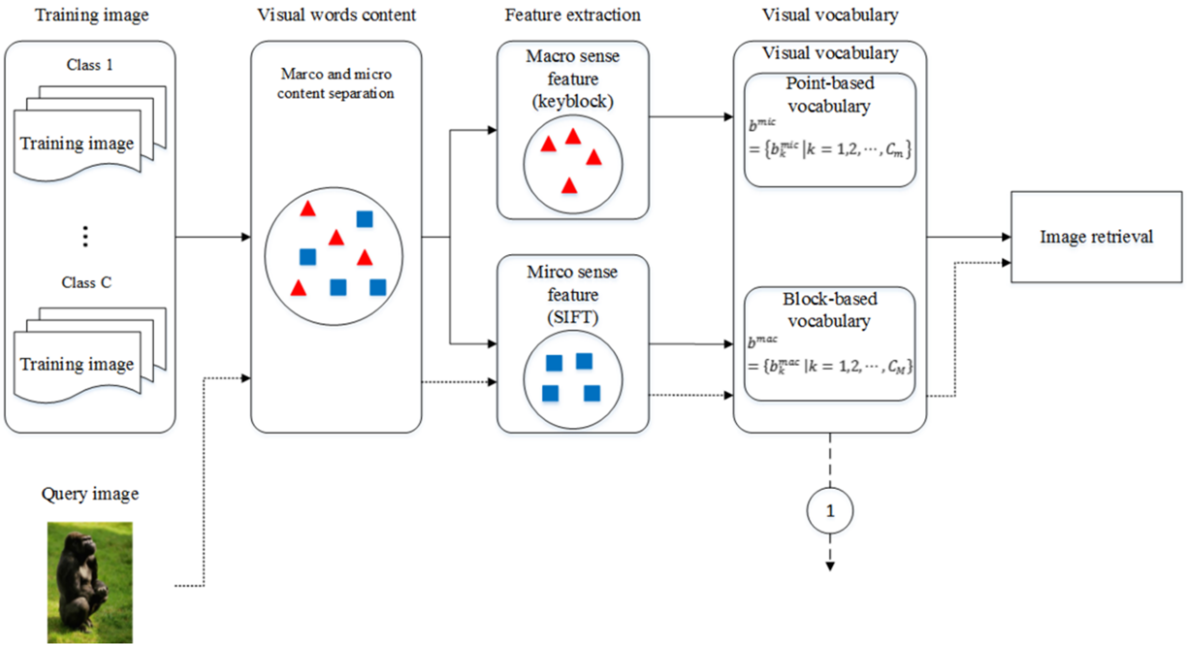

Figure 2. The proposed architecture of visual vocabulary construction and image retrieval

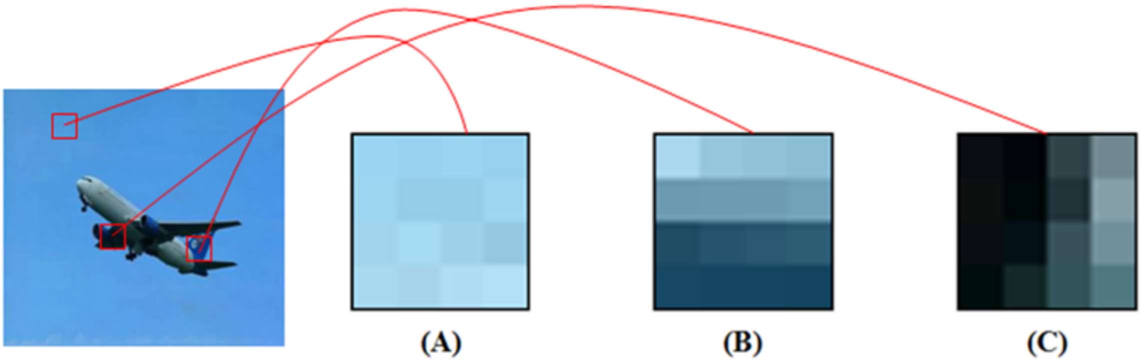

Figure 3. Sample image, (A) Macro sense block, (B) and (C) Micro sense blocks

In conventional SIFT descriptor, as shown in Fig. 5(a), 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. See Fig. 5(b). Therefore, the dimension of the proposed descriptor is reduced to 8, but the statistics are more convincing.

The descriptor is defined as follows. We first define the surrounding region as

$$M\left[u_s\left(i,j\right)\right] = \left\{I\left(x,y\right)\left|x \in \left(i-4 \leq i < i+4\right), y \in \left(j-4 \leq j \leq j=4\right)\ and\ i,j \neq 0\right.\right\} \tag{3}$$

where $u_s(i,j)$ is the feature point in image s, and $M[u_s(i,j)]$ is the surrounding region at center (i,j). The gradient vector for this surrounding region is as

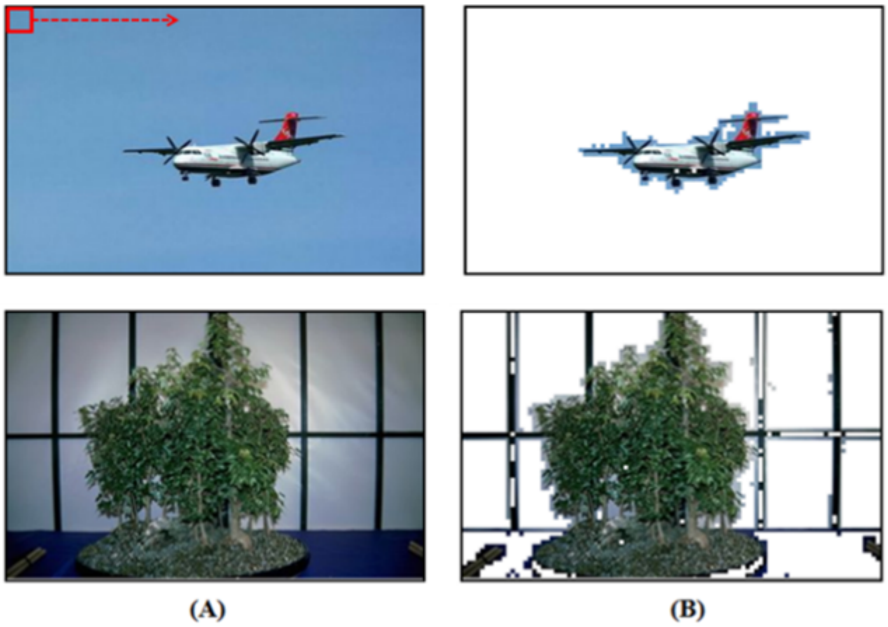

FIGURE 4. Examples of macro and micro content separation (threshold value T=15)

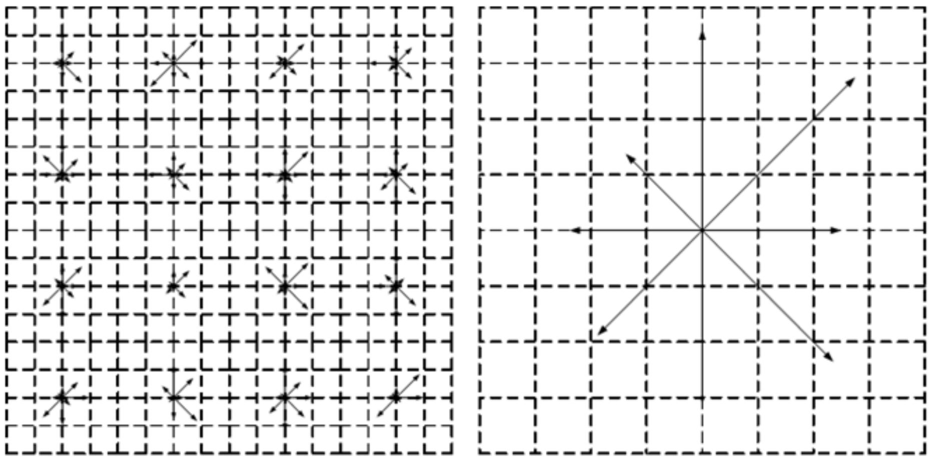

FIGURE 5. (a) Original 16×16 descriptor (b) Proposed 8×8 descriptor

$$\nabla M\left[u_s(i,j)\right] = \nabla\left\{I(x,y)\right\},\ I(x,y) \in M\left[u_s(i,j)\right] \tag{4}$$

where $\nabla\left\{I(x,y)\right\}$ is the gradient of all pixels in $M\left[u_s(i,j)\right]$. The directional histogram is expressed as

$$h\left(u_s(i,j)\right) = h^1\left(u_s(i,j)\right), h^2\left(u_s(i,j)\right), h^3\left(u_s(i,j)\right), \cdots, h^8\left(u_s(i,j)\right) \tag{5}$$

$$h^\theta\left(u_s(i,j)\right) = \sum_{\theta=1}^{8} \delta\left(\frac{\angle\nabla I(x,y)}{\frac{\pi}{4}} - \theta\right), (x,y) \in M\left[u_s(i,j)\right] \tag{6}$$

where $\nabla\left\{I(x,y)\right\}$ is the direction of the gradient vector in Eq. (4) and $\lceil . \rceil$ is the ceiling function. An example of the directional histogram in Eq. (5) is shown in Fig. 6.

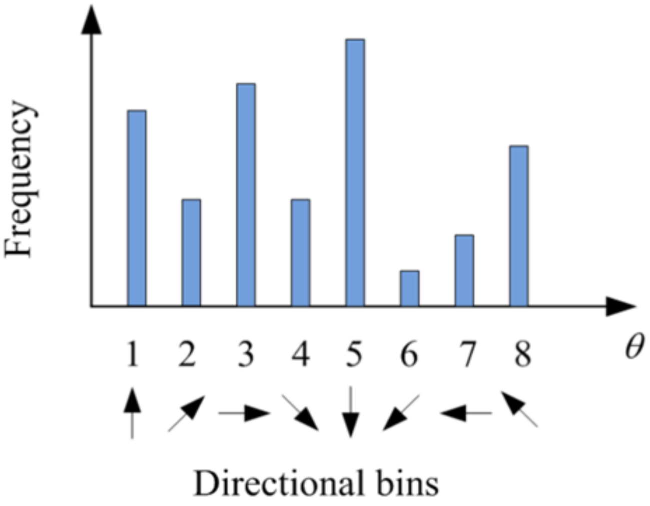

Figure 6. Directional histogram

We define the maximum direction of feature point as the main direction,

$$V_{Max}\left(u_s\left(i,j\right)\right) = Max\left(h^{\theta}\left(u_s\left(i,j\right)\right)\right)\ \ \theta = 1,2,\cdots,8 \tag{7}$$

where the $V_{Max}\left(u_s\left(i,j\right)\right)$ is the main direction of $u_s\left(i,j\right)$. Then we rotate the region counter-clockwise by an angle θ.

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 $M\left[u_s\left(i,j\right)\right]$ located on the main direction. In order to reduce the noise interference, the total number of the bins is quantized to 32. Thus, the color feature is expressed as

$$H_{RGB}\left(u_s\left(i,j\right)\right) = h_{RGB}^{(0)}\left(u_s\left(i,j\right)\right), h_{RGB}^{(1)}\left(u_s\left(i,j\right)\right), \cdots, h_{RGB}^{(7)}\left(u_s\left(i,j\right)\right) \tag{8}$$

$$h_{RGB}^{\left(\theta_{V_{Max}}\right)}\left(k\right) = \frac{1}{N}\sum_{0}^{7}\delta\left(\frac{RGB\left(I\left(x,y\right)\right)}{32} - k\right) \tag{9}$$

$$\left[RGB\left(I\left(x,y\right)\right)\right] = \left[I_R\left(x,y\right), I_G\left(x,y\right), I_B\left(x,y\right)\right], \left(x,y\right) \in M\left[u_s\left(i,j\right)\right] \tag{10}$$

where $H_{RGB}^{\left(\theta_{V_{Max}}\right)}\left(u_s\left(i,j\right)\right)$ is the color histogram of the pixels in main direction, the $h_{RGB}^{k}\left(u_s\left(i,j\right)\right)$ is the k^{th} bin of the quantized color, and $\lfloor . \rfloor$ is the floor function.

3.2. **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 C_M; that is

$$L\left(B_s^c\left(n\right)\right) = k = \underset{k}{\operatorname{argmin}}\left(B_s^c - d_k\right) k = 1,\ldots,C_M, n = 1,\ldots,N \tag{11}$$

where L(.) is labeling function, $B_s^c\left(n\right)$ is the input image block, and d_k is the k^{th} 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).

$$\begin{aligned} h_s^c &= \ h_s^c\left(1\right), \ h_s^c\left(2\right), \ldots, \ h_s^c\left(C_M\right) \\ h_s^c\left(k\right) &= \tfrac{1}{N}\sum_{n=1}^{N} \delta(L(B_s^c\left(n\right)) - k), \ k = 1, \ldots, C_M \end{aligned} \tag{12}$$

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. (11).

3.3. **Constructing macro- and micro-based visual vocabulary.** The training images of each class contain various variations including luminance change and contrast change. In our work, the training images are represented as $T = I_s^c | s = 1, ..., S, c = 1, ..., C$, 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. 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, we design a merging procedure to obtain a class vocabulary. The procedure is illustrated in Fig. 7.

For micro-based visual vocabulary, in our work we use the difference of main direction between two feature points to measure their similarity. For directional histogram, the difference is given by

$$\mathrm{S}\left(V_{Max}\left(u_s\left(i,j\right)\right), V_{Max}\left(u_s\left(k,l\right)\right)\right) = \begin{cases} 1 \ if \ V_{Max}\left(u_s\left(i,j\right)\right) - V_{Max}\left(u_s\left(k,l\right)\right) \leq T \\ 0 \ else \end{cases} \tag{13}$$

where T is the threshold, the values 1 and 0 use to represent similar or dissimilar, respectively. The color similarity is defined as

$$S\left(H_{RGB}\left(u_s\left(i,j\right)\right), H_{RGB}\left(u_s\left(k,l\right)\right)\right) = \begin{cases} 1 \ if \ \frac{1}{8}\sum\limits_{N=0}^{7} h_{RGB1}^{(N)}\left(u_s\left(i,j\right) - h_{RGB2}^{(N)}\left(u_s\left(k,l\right)\right)\right) \leq T \\ 0 \ else \end{cases} \tag{14}$$

Where $H_{RGB}\left(u_s\left(i,j\right)\right)$ and $H_{RGB}\left(u_s\left(k,l\right)\right)$ represent the color histograms of two feature points. The similarity of two feature points is defined as

$$SIM\left(u_s\left(i,j\right), u_s\left(k,l\right)\right) = \begin{cases} 1 \ \begin{matrix} S\left(V_{Max}\left(u_s\left(i,j\right)\right), V_{Max}\left(u_s\left(k,l\right)\right)\right) = 1 \\ and \ S\left(H_{RGB}\left(u_s\left(i,j\right)\right), V_{RGB}\left(u_s\left(k,l\right)\right)\right) = 1 \end{matrix} \\ 0 \ else \end{cases} \tag{15}$$

If two feature-point are similar, we merge them into a new descriptor as

$$h^{(\theta)new}\left(u_s\left(i,j\right)\right) = \frac{h^{(\theta)}\left(u_s\left(i,j\right) + h^{(\theta)}u_s\left(k,l\right)\right)}{M} \ \theta = 1, 2, \ldots 8 \tag{16}$$

$$h_{RGB}^{new(N)}\left(u_s\left(i,j\right)\right) = \frac{\left(h_{RGB}^{(N)}\left(u_s\left(i,j\right)\right) + h_{RGB}^{(N)}\left(u_s\left(k,l\right)\right)\right)}{M} \ N = 0, 1, \ldots, 7 \tag{17}$$

where $h^{(\theta)new}$ new is new directional histogram, and $h_{RGB}^{(N)new}\left(u_s\left(i,j\right)\right)$ is the new color histogram. M is the number of merged feature points. According to extensive simulations, the T is set to 0.15 in our work.

After all feature points have been tested and merged, the visual words are collected into primitive visual vocabulary. Because the feature points are from all training images, so

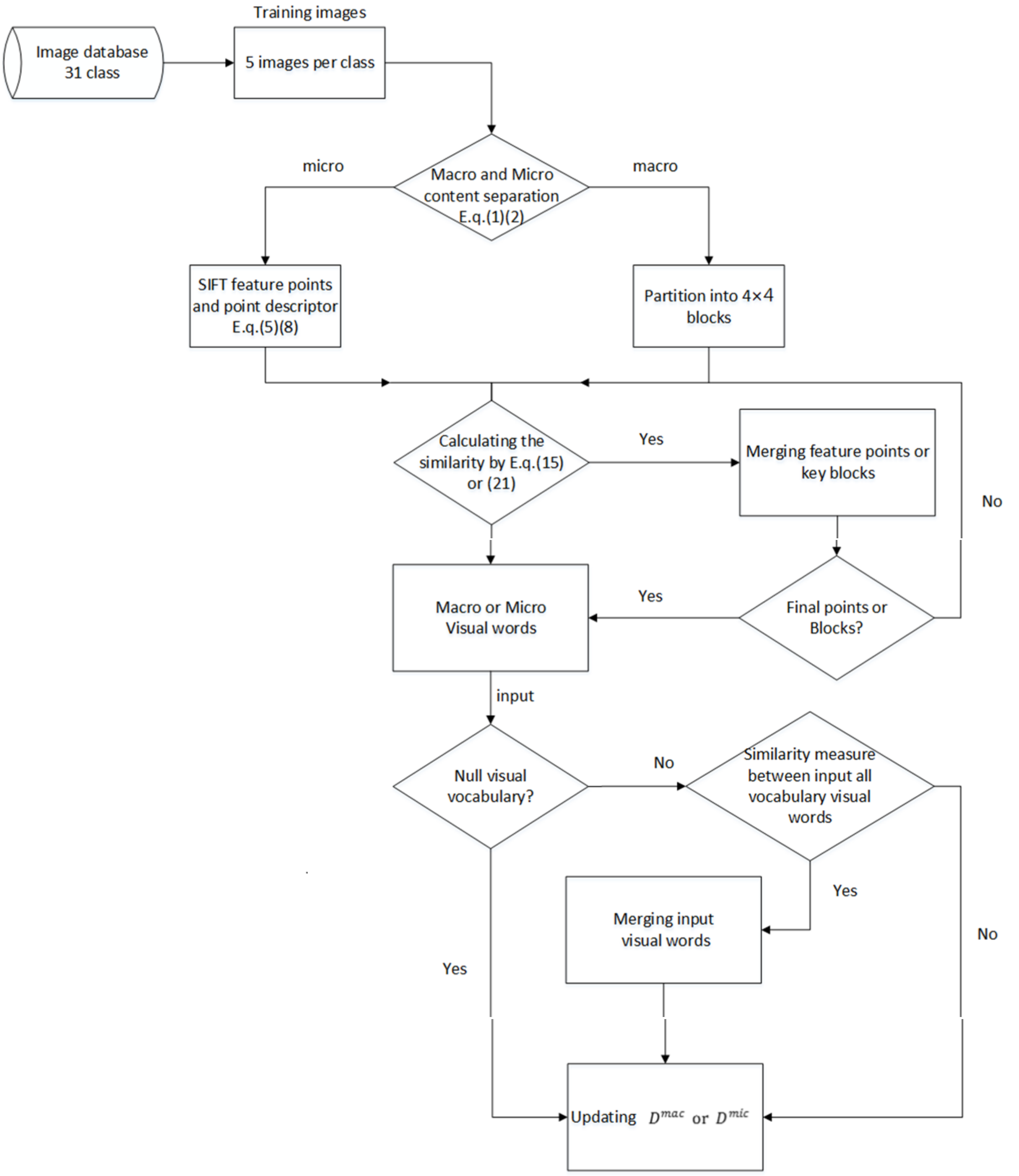
Training images
Image database 31 class
5 images per class
micro
macro
Macro and Micro content separation E.q.(1)(2)
SIFT feature points and point descriptor E.q.(5)(8)
Partition into 4×4 blocks
Calculating the similarity by E.q.(15) or (21)
Yes
Merging feature points or key blocks
No
Macro or Micro Visual words
Yes
Final points or Blocks?
input
Null visual vocabulary?
No
Similarity measure between input all vocabulary visual words
Yes
Merging input visual words
Yes
No
Updating D^{mac} or D^{mic}

Figure 7. The block diagram of visual vocabulary construction

the size of primitive visual vocabulary is large. Therefore, we should carefully construct micro-based visual vocabulary. To describe the micro-based visual vocabulary, the image description based on primitive visual vocabulary is used. The feature points extracted from micro content of training images are labeled by the index of nearest visual words in the primitive vocabulary:

$$L^{mic}\left(u_s\left(i,j\right)\right) = \min_k u_s\left(i,j\right) - d_k^{premitive} \tag{18}$$

where $L^{mic}\left(u_s\left(i,j\right)\right)$ is the label function for the feature point in micro content using primitive vocabulary. Then the histogram for all training images can be calculated as

$$\begin{array}{l} h_s = h_s\left(1\right), \ldots, h_s\left(M\right) \\ h_s\left(k\right) = \frac{1}{N}\sum\limits_{n=0}^{N} \delta\left(L\left(u_s\left(i,j\right)\right) - k\right) \;,\; k = 1, 2, \ldots, M \end{array} \tag{19}$$

where M is the size of the primitive vocabulary.

We calculate the label histogram for all training images, and then sort the usage frequency of visual words. In our work, the top P words, which are selected from most frequently used visual words, will be used to form micro sense vocabulary. In our work, considering the effectiveness and efficiency, the *P=256* is selected for simulations.

The procedure of construction of macro sense vocabulary is the same as micro sense visual vocabulary except that the visual words are blocks. The scheme is summarized as follows.

(1) Partition the macro content of training image into N blocks, each block size is $4times4$, and the block samples are denoted as $B_s^c\left(n\right)$, n=1,2,,N.

(2) The similarity measure of two blocks is calculated by their Euclidean distance as

$$\begin{array}{c} Dis\left(B_S^C\left(i\right) B_S^C\left(j\right)\right) = \frac{1}{16}\times \\ \times\sum\limits_{m=0}^{3}\sum\limits_{n=0}^{3}\sqrt{\left(B_S^C\left(i\right)_{mn}^{R} - B_S^C\left(j\right)_{mn}^{R}\right)^2 + \left(B_S^C\left(i\right)_{mn}^{G} - B_S^C\left(j\right)_{mn}^{B}\right)^2 + \left(B_S^C\left(i\right)_{mn}^{B} - B_S^C\left(j\right)_{mn}^{B}\right)^2} \end{array} \tag{20}$$

The similarity of two blocks is defined as

$$SIM\left(B_S^C\left(i\right), B_S^C\left(j\right)\right) = \left\{ \begin{array}{l} 1,\; D\left(B_S^C\left(i\right), B_S^C\left(j\right)\right) < T \\ 0, else \end{array} \right. \tag{21}$$

where T is a threshold.

1 . If two block-based visual words are similar, then we merge them into a new word as shown in Fig. 7.

2 . The top P words will be selected to form the macro sense vocabulary. In our work, the *P=64* as selected in micro sense visual vocabulary.

3.4. **Image description and similarity measure.**

The image description scheme is shown in Fig. 8. The input block is first categorized as macro-sense or micro-sense by using Eq. (2). 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 $H_s^c\left(q\right) = \left(\; H_i^{MAC}\left(q\right),\; H_j^{mic}\left(q\right)\;\right), where\, i = 1, .., N_{MAC},\; j = 1, \ldots N_{mic}$. We can define the similarity of images as

$$\begin{array}{c} S^{Mac(or\;\; mic)}\left(q,l\right) = \sum\limits_{i=1(or\;\; j=1)}^{N_{Mac(or\;\; mic)}} \left(1 - \left|H_{i(or\;\; j)}^{Max(or\;\; Mic)}\left(q\right) - H_{i(or\;\; j)}^{Max(or\;\; Mic)}\left(l\right)\right|\right)\times \\ \times min\left(H_{i(or\;\; j)}^{Max(or\;\; Mic)}\left(q\right) - H_{i(or\;\; j)}^{Max(or\;\; Mic)}\left(l\right)\right) \end{array} \tag{22}$$

$$\mathrm{S}\left(q,l\right) = S^{Mac}\left(q,l\right)\times w_1 + S^{Mic}\left(q,l\right)\times w_2 \tag{23}$$

where $S^{Mic}\left(q,l\right)$ is the similarity of image q and l in macro sense histogram, and $S^{Mic}\left(q,l\right)$ is the micro sense similarity; $w_(1,)w_2$ are the weighting value, and $S\left(q,l\right)$ is the overall similarity. In our work, the histogram similarity measure of visual words is

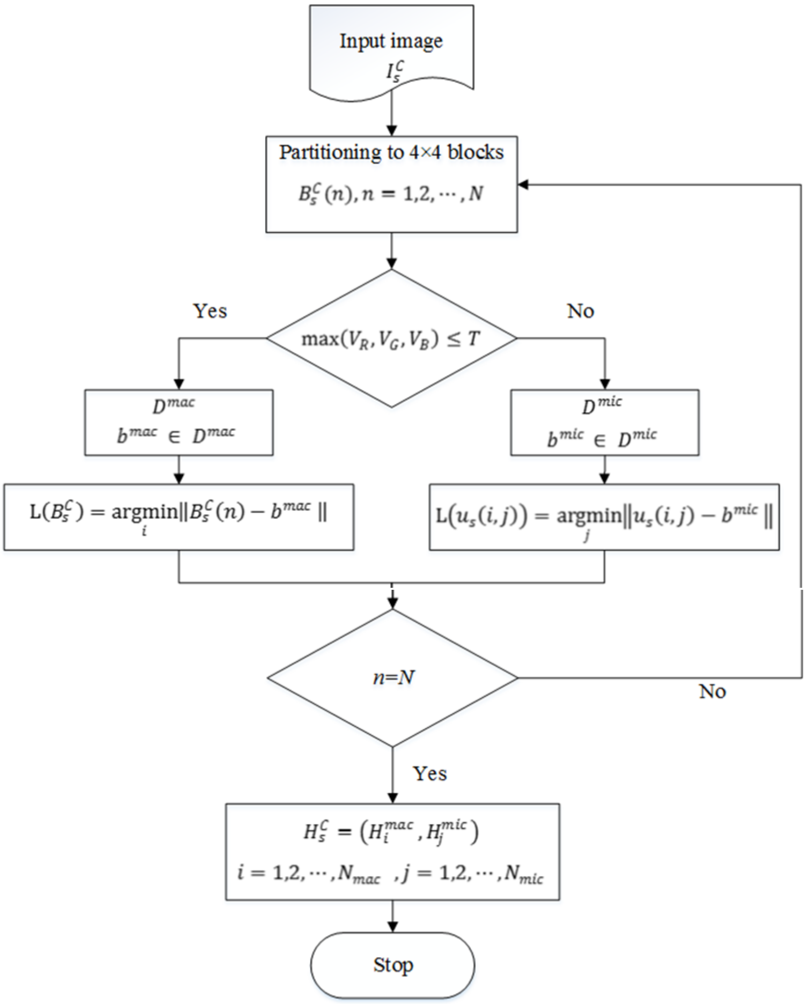
Input image
I_s^C
Partitioning to 4×4 blocks
$B_s^C(n), n = 1,2,\cdots,N$
Yes
No
$\max(V_R, V_G, V_B) \le T$
D^{mac}
$b^{mac} \in D^{mac}$
D^{mic}
$b^{mic} \in D^{mic}$
$\mathrm{L}(B_s^C) = \underset{i}{\operatorname{argmin}} \| B_s^C(n) - b^{mac} \|$
$\mathrm{L}(u_s(i,j)) = \underset{j}{\operatorname{argmin}} \| u_s(i,j) - b^{mic} \|$
$n=N$
No
Yes
$H_s^C = (H_i^{mac}, H_j^{mic})$
$i = 1,2,\cdots,N_{mac}\ , j = 1,2,\cdots,N_{mic}$
Stop

Figure 8. Block diagram of image description

obtained by modifying our previous work [2], which has been verified superior to state-of-the-art approaches for image retrieval by extensive simulations.

4. **Experimental Results.** We use a database (31 classes, 3901 images) from Corels photo to test the performance of the proposed method. The database has a variety of images, please see [9] for details. To evaluate the performance of image retrieval, two popular performance indexes ARR [2] (Average Retrieval Rate) and ANMRR [2] (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.

The similarity retrieval is based on the weighted measure $S = S^{Mac} \times w_1 + S^{Mic} \times w_2$. We denote the ratio of macro and micro as $(w1 : w2)$=(1:0),(0.7:0.3),(0.5:0.5),(0.3:0.7),(0:1). Therefore, as shown in Table1, the weighted value $(0.7 : 0.3)$ achieves the best performance for most classes. Therefore, in our work, the weighting value $(0.7 : 0.3)$ is chosen for all simulations.

In the following, the comparison of the proposed and some typical methods is listed for evaluating the performance and effectiveness. Table 2 shows the performance ARR for conventional SIFT descriptor; SIFT descriptor combined with color histogram, MPEG-7

Table 1. Comparison of different weighted values on ARR performance

		ARR(ANMRR)									
	Category	**1:00**		**0.7:0.3**		**0.5:0.5**		**0.3:0.7**		**0:01**	
1	Orangutans	0.20541	0.73156	0.27976	0.63818	0.2663	0.654	0.25829	0.66286	0.11616	0.84369
2	Chinese painting birds	0.34641	0.55678	0.35345	0.54191	0.3574	0.53931	0.35604	0.53819	0.23691	0.70041
3	Pot plant	0.13951	0.82208	0.19958	0.74777	0.19183	0.75715	0.18809	0.7625	0.19501	0.75352
4	Card	0.73073	0.15221	0.81143	0.0965	0.82406	0.08989	0.82726	0.08784	0.24023	0.6941
5	Cloud	0.0978	0.87305	0.18171	0.75906	0.16145	0.78526	0.15123	0.79753	0.21547	0.71175
6	Sunset	0.09175	0.87687	0.16857	0.77054	0.15256	0.79475	0.1422	0.80708	0.21262	0.71825
7	Pumpkin	0.14307	0.80971	0.27737	0.65216	0.26962	0.65987	0.26859	0.66675	0.06714	0.90384
8	Cake and cookie	0.12755	0.83318	0.23448	0.69577	0.21673	0.71425	0.20653	0.72614	0.1053	0.85816
9	Dinosaur	0.7902	0.08682	0.7664	0.08953	0.7555	0.09717	0.7509	0.10116	0.5616	0.34107
10	Wheel and dolphin	0.18865	0.76133	0.22212	0.71294	0.21885	0.71777	0.21661	0.72056	0.18253	0.77106
11	Elephant	0.16682	0.78167	0.21373	0.71939	0.20529	0.72969	0.20097	0.7362	0.1139	0.84562
12	Firework	0.75221	0.13369	0.82638	0.08935	0.81884	0.09208	0.81463	0.09419	0.37146	0.53964
13	Flower	0.11898	0.83154	0.19504	0.74463	0.18791	0.75285	0.18275	0.75797	0.10281	0.85622
14	Vegetable and fruit	0.1229	0.8431	0.17605	0.77112	0.16456	0.78415	0.15958	0.79128	0.1957	0.74721
15	Ceramic duck	0.4148	0.47481	0.5281	0.34242	0.4993	0.36978	0.4858	0.3851	0.1938	0.74649
16	Leopard	0.24175	0.68356	0.36006	0.53661	0.35612	0.53929	0.35401	0.5413	0.29547	0.62222
17	Leaf	0.27223	0.61945	0.35418	0.53297	0.36082	0.52526	0.36311	0.52228	0.27786	0.64719
18	Car	0.07122	0.90584	0.13342	0.82204	0.12117	0.83728	0.11411	0.8457	0.14868	0.79898
19	Cactus	0.16667	0.76876	0.21081	0.70915	0.21437	0.70417	0.216	0.70192	0.09708	0.86629
20	Airplane	0.30929	0.6124	0.19469	0.74131	0.18484	0.75421	0.18104	0.76167	0.18843	0.74037
21	Mural	0.19636	0.74866	0.29725	0.62825	0.30257	0.62501	0.30434	0.62413	0.20912	0.73362
22	Sea animal	0.08072	0.89177	0.12878	0.82662	0.11792	0.83992	0.11222	0.84703	0.09478	0.87312
23	Horse	0.07456	0.89832	0.1048	0.85913	0.10354	0.8603	0.10501	0.85989	0.06049	0.91936
24	Helicopter	0.11697	0.83787	0.14594	0.79852	0.14222	0.80387	0.14013	0.80744	0.11017	0.84575
25	Ship	0.09451	0.87243	0.10797	0.84795	0.10518	0.85094	0.1043	0.85301	0.07342	0.89811
26	Snow	0.20365	0.71734	0.24053	0.6734	0.24635	0.66605	0.24868	0.66268	0.12212	0.83721
27	Hot air balloon	0.12774	0.82757	0.16997	0.77752	0.15924	0.79125	0.15457	0.79848	0.13946	0.80368
28	Waterfall	0.16593	0.77719	0.20162	0.72522	0.19653	0.7322	0.19326	0.73659	0.10224	0.8595
29	Classical architecture	0.11127	0.84868	0.148	0.79815	0.15046	0.79547	0.15128	0.79461	0.09316	0.87212
30	Sports field	0.38791	0.52654	0.51342	0.38229	0.51084	0.3861	0.50929	0.38878	0.46487	0.45254
31	Person	0.10365	0.86537	0.17268	0.76271	0.15631	0.786	0.149	0.79706	0.27547	0.62299
	Average	**0.231**	**0.7087**	**0.28768**	**0.63848**	**0.28124**	**0.64629**	**0.2773**	**0.6509**	**0.18914**	**0.75561**

DCD and color descriptor [2]. It can be seen that the conventional SIFT descriptor is the worst one due to the lack of color information; however, its ARR will be improved significantly when the color information is considered. The simulation results indicate that the proposed method achieves the highest ARR because it considers the background information. Table 2 also lists the performance index of ANMRR, which is similar with ARR.

5. **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.

REFERENCES

[1] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter, On feature distributional clustering for text categorization, *Proc. Assoc. Comput. Machinery Special Interest Group Informat. Retrieval (SIGIR)*, pp. 146153, 2001.

[2] 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.

[3] D. G. Lowe, Distinctive Image Features from Scale-invariant Keypoints, *International Journal Of Computer Vision*, vol. 60, no. 2, pp.91-110, 2004.

[4] L. Zhu, A. Zhang, A. Rao, and R. S. Cedar, Keyblock: an Approach for Content-Based Image Retrieval, *ACM Multimedia*, pp.157-166. 2000.

[5] S. Xu, T. Fang, D. Li, and S. Wang, Object Classification of Aerial Images with Bag-of-Visual Words, *IEEE Geoscience And Remote Sensing Letters*, vol. 7, no. 2, pp.366-370, 2010.

[6] Y.G. Jiang, J. Yang, C.W. Ngo and A. G. Hauptmann, Representations of Keypoint-Based Semantic Concept Detection: A Comprehensive Study, *IEEE Transactions On Multimedia*, vol. 12, no. 1,pp.42-53, 2010.

[7] 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.

[8] A. Bolovinou, I.Pratikakis and S.Perantonis, Bag of spatio-visual words for context inference in scene classification, *Pattern Recognition*, vol. 46, pp.10391053, 2012.

[9] C.M. Kuo, C.H. Hsieh, N.C. Yang, C.-M. Kuo, C.K. Chang, Y.M. Chen, Constructing a discriminative visual vocabulary with macro and micro sense of visual words, *Multimedia Tools and Applications*, 2015.

TABLE 2. Comparison of different methods on ARR performance

ARR/ANMRR											
	Category	**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	Chinese painting for 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	0.8859	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	0.0346	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	0.0129	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	0.0504	0.9367	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	0.0764	0.8948	0.10224	0.8595	0.161511	0.776893	0.180808	0.765502	0.19653	0.7322
29	Classical 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