4. Visual-Word-Based Image Feature

Region-based approach has been widely adopted for image representation recently. An image is considered to be a set of regions, and each of regions can be described by visual features such as color, texture, shape. Some researchers adopted the model of visual words, as described in Section 2.3, to build their region-based image representation. This chapter describes our proposed visual-word-based image features that categorize visual features as words for image representation. We first describe the model of visual words, and present the generation of visual words for images analysis. Next, we propose the smoothed visual words to design the visual-word-based image features for image representation.

4.1. Visual Words

The related works of visual words have been described in Section 2.3. Here we first present the basic model of visual words used in image analysis and then describe the construction flow of visual words.

4.1.1. The model of visual words

In the model of visual words, an image can be regarded as a bag of visual words. Figure 4-1, which is captured from [ICCV 2005 Courses], illustrates the idea that an image is considered the compound of visual words which are extracted from the image. This idea likes that a document consists of many textual words.

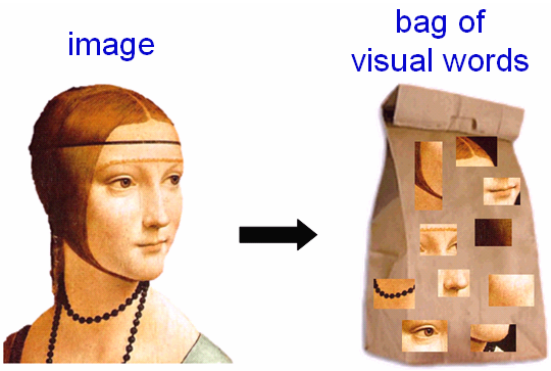
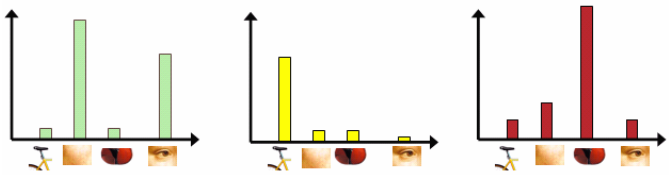


Figure 4-1. The bag of visual words.

(a). Three original images.

(b). Region features of the three images.



(c). Distribution of visual words for the three images.

Figure 4-2. Image representation using visual words.

Hence, regions of all images are categorized as the predefined visual words of vocabulary and the visual words are used for representing images. An example of image representation based on visual words is shown in Figure 4-2 that is also taken from [ICCV 2005 Courses]. In this example, region features in Figure 4-2(b) are

extracted from the three images in Figure 4-2(a). Four visual words are then generated from the region features and used to represent the images shown in Figure 4-2 (c) by use of the distribution of visual words for each image.

4.1.2. Generation of visual words

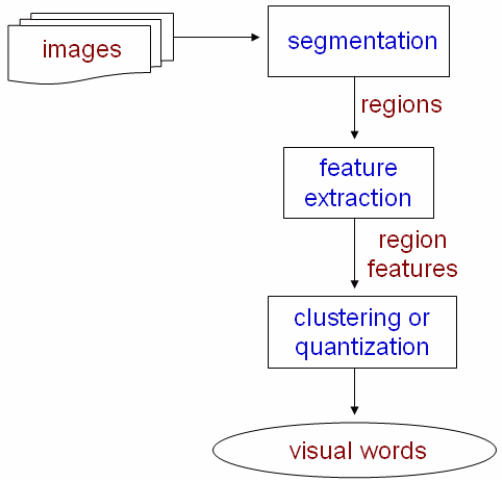


Figure 4-3. Generation of visual words.

Figure 4-3 shows the flow of generating visual words for images. Each of images can be first segmented into a set of regions, and then visual features are extracted from these regions. Thus, region features of all images are quantized to a pre-defined vocabulary of visual words. That is to say, all region features of images are collected and the feature space is quantized to describe the representation of image regions. There are three major issues in the generation of visual words: (i) image segmentation, (ii) visual feature extraction, and (iii) quantization. We discuss these issues as the follows.

The aim of image segmentation is to partition an image into several parts that can well represent the image. However, there is still no perfect method for image

segmentation. Several related works of image segmentation have been presented in Section 2.3. In this thesis, we adopted watershed segmentation to partition images according to the intensity of signals, also described in Section 3.1.

The choice of visual feature is also important because different types of features can characterize different contents of images. No matter what types of features we choose, images can be transformed to vectors in the chosen feature spaces. Hence, the selection of visual feature is independent of visual words at most cases.

Quantization of feature vectors is the major part in the generation of visual words. Most researchers employed *K*-means clustering to quantize the feature space into *K* areas. These *K* parts will form *K* visual words, and an image can be represented solely by a histogram of visual words.

As mentioned above, visual words are still constructed in visual feature spaces, not in a semantic-level space. However, we can say the visual word is a middle-level feature that is between low-level visual features and high-level semantic concepts for image representation. The main reason is that many types of semantic information of images are learned based on visual words. For example, most approaches to solving image annotation modeled the relationship between labels and visual words, which will be described in the next chapter in details.

4.2. Smoothed Visual Words

Visual words generated by *K*-means clustering is a 0-1 assignment, that is to say, any region will fall in only one visual word in the feature space. One drawback of *K*-means clustering is that some neighboring feature vectors may be assigned to different visual words. Fuzzy *K*-means clustering is a potential method for quantizing

the feature space to overcome the problem of 0-1 assignment. However, the storage complexity of fuzzy *K*-means clustering is very high. In this section, we propose the smoothed the smoothed visual words for image representation to overcome the drawback of 0-1 assignment.

4.2.1. Generation of smoothed visual words

The proposed smoothed visual words are based on the standard *K*-means clustering and design a smoothed approach to avoid the solid quantization of clusters. The process of smoothed visual words is stated as the follows. Note that these steps are performed only in a given feature space F.

1. Set a constant value *K*, and apply the standard *K*-means clustering to all region features $\{F(R_i)\}$ to generate *K* clusters. This yields *K* centroids (means) denoted as $\{C_j \mid 1 \le j \le K\}$ of these clusters in the feature space. To simplify the notation, we use these *K* centroids instead of clusters.
2. Compute the pairwise similarity measure, $sim(C_p, C_q)$, between any two centroids C_p and C_p using the following equation:

$$sim(C_p, C_q) = \frac{\exp(-d(C_p, C_q)/\sigma)}{\sum_{k=1}^{K} \exp(-d(C_p, C_k)/\sigma)}. \tag{4.1}$$

   In this equation, σ is a smoothed parameter: the larger the parameter σ, the more smooth the visual words. In addition, $d(C_p, C_q)$ means the Euclidean distance between the centroids of the two clusters C_p and C_q.
3. The similarity measure $sim(C_p, C_q)$ should be reflexive and symmetric, i.e., $sim(C_p, C_p) = 1$ and $sim(C_p, C_q) = sim(C_q, C_p)$. Thus, the two following steps

are performed:

(3.a) Reflexive. For all *p* and *q*, similarity measures $sim(C_p, C_q)$ are normalized by

$$sim(C_p, C_q) \leftarrow sim(C_p, C_q) / sim(C_p, C_p). \quad (4.2)$$

Then $sim(C_p, C_p) = 1$.

(3.b) Symmetric. After been identical, we then perform the following equation for all *p* and *q*:

$$sim(C_p, C_q) = (sim(C_p, C_q) + sim(C_q, C_p)) / 2. \quad (4.3)$$

Thus, the proposed smoothed visual words consist of two parts: (i) the *K* visual words that are the *K* centroids of clusters, $\{C_j \mid 1 \le j \le K\}$, generated by *K*-means clustering, and (ii) the pairwise similarity $sim(C_p, C_p)$ of *K* visual words, where $1 \leqq p \leqq K$ and $1 \leqq p \leqq K$. The pairwise similarity of *K* visual words is the smoothed factor to dissolve the solid boundaries of visual words.

4.2.2. Region representation

Given a region in feature space F, we then present the region representation by use of the proposed smoothed visual words. The procedure of representing a region is stated as the follows.

1. For each region R, extract its feature vector in the space F and find cluster C_r to which region R belongs. Note that region R can either be in the original data in *K*-means clustering or be a new region.

2. For each cluster C_j $(1 \le j \le K)$

$$w(R, C_j) \leftarrow sim(C_r, C_j). \tag{4.4}$$

3. The output of the region representation is $\mathbf{w}_R$,

$$\mathbf{w}_R = \{w(R, C_1), ..., w(R, C_K)\}. \tag{4.5}$$

$w(R_i, C_j)$ means the confidence value that region R can be represented by the centroid C_j. Thus, we collect $w(R_i, C_j)$, for all C_j, to be a visual-word-based vector for the region. Next, we can extract the visual-word-based image feature according to these region features in Section 4.3.

4.2.3. Complexity

Now, let us analyze the time and storage complexity of the proposed smoothed *K*-means clustering. For the time complexity, it is obvious that the smoothed *K*-means clustering also needs to perform *K*-means clustering in the beginning. The additional time complexity compared with *K*-means visual words is $O(K^2)$ for computing the pairwise distances of K centroids. Note that, in general, $O(K^2) << O(N_R{}^2)$ for $K << N_R$ where N_R means the number of all regions.

Moreover, we discuss the storage complexity of the proposed visual-word-based region feature. The system does not need to store all correspondence of N_R images and K centroids, but it does require two tables: (i) one for pairwise similarity of any two clusters with complexity $O(K^2)$, and (ii) the other one for the cluster assignment of each region with complexity $O(N_R)$. Hence the space complexity of our proposed

method is $O(K^2 + N_R)$. On the other hand, fuzzy *K*-means clustering needs the storage with complexity $O(K \cdot N_R)$ that records the correspondence of N_R regions and *K* centroids, and $O(K^2 + N_R) << O(K \cdot N_R)$ for $K<<N_R$.

4.3. Image Representation

This section presents the scheme of image representation by designing a visual-word-based image feature that is based on the proposed smoothed visual words.

4.3.1. Visual-word-based image feature

Suppose that there are *n* regions in an image *I*, which, without loss of generality, are denoted as $\{R_i \,|\, 1 \le i \le n\}$. Table 4-1 shows the algorithm to extract the visual-word-based image feature, which is denoted as $\mathbf{v}_{VW} = \{v_1, ..., v_K\}$, based on the smoothed visual words. The visual-word-based image feature of an image *I* is to accumulate all values of $size(R_i) \cdot w(R_i, C_j)$ for each region R_i in image *I* to visual words C_j, where $size(R_i)$ means the size percentage of region R_i in image *I*, and $w(R_i, C_j)$ is the confidence part of the smoothed visual words that is described in the previous section.

Table 4-1. The algorithm for region-based image representation.

Input:

$\{R_i \mid 1 \le i \le n\}$ // the set of regions in the image *I*.

$\{C_j \mid 1 \le j \le K\}$ // the visual words.

$sim(C_p, C_p), 1 \le p, q \le K$ //pairwise similarity of visual words.

$\{\mathbf{w}_{R_1}, ..., \mathbf{w}_{R_n}\}$ //VW-based region features for all regions.

Output:

$\mathbf{v}_{VW} = \{v_j \mid 1 \le j \le K\}$ // *K*-D vector for image representation.

Initialization:

$v_j = 0$ for all j

1. for (*i*=1; *i*<=*n*; *i*++)
2. for (*j*=1; *j*<=*K*; *j*++)
3. $v_j = v_j + size(R_i) \cdot w(R_i, C_j)$

   // $\mathbf{w}_{R_i} = \{w(R_i, C_1), ..., w(R_i, C_K)\}$

   //compute VW-based region features and accumulate them.

   // $Size(R_i)$: the size percentage of the region R_i .

4.3.2. An example

Now we illustrate our proposed visual-word-based image feature in Figure 4-4. Assume that, shown in Figure 4-4(a), an image contains three regions A, B, and C with size percentages 0.2, 0.3, and 0.5, and each of regions belongs to visual words C2, C1, and C1, respectively. We also assume that the scheme of the smoothed visual words is shown in Figure 4-4(b) that contains three visual words (i.e., *K*=3) and their pairwise similarity measures.

Then, the three regions of the image shown in Figure 4-4(a) can be represented by the only three visual words shown in Figure 4-4(c) and by three visual words with

pairwise confidence values shown in Figure 4-4(d). The case in Figure 4-4(c) represents regions based on the visual words by 0-1 assignment, and the case in Figure 4-4(d) represents that by the smoothed assignment proposed in this section. Therefore, the accumulated feature of all regions with the size percentages are shown in Figure 4-4(e) and (f). Comparing the two image features in Figure 4-4(e) and (f), the latter is more smoothed than the former for applying the smoothed visual word.

region	A	B	C
words	C2	C1	C1
size	0.2	0.3	0.5

(a). The compound regions in an image.

	C1	C2	C3
C1	1	0.335	0.174
C2	0.335	1	0.107
C3	0.174	0.107	1

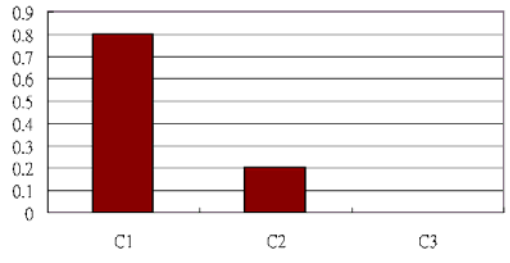
(b). Visual words and their pairwise similarity measures.

	A	B	C	**v**
C1	0	1	1	0.8
C2	1	0	0	0.2
C3	0	0	0	0
Size	0.2	0.3	0.5	

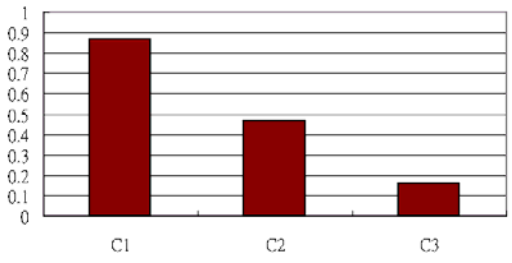
(c). Region and image features using 0-1 assignment, $\mathbf{v} = \{0.8, 0.2, 0\}$.

	A	B	C	$\mathbf{v}_{VW}$
C1	0.335	1	1	0.867
C2	1	0.335	0.335	0.468
C3	0.107	0.174	0.174	0.160
Size	0.2	0.3	0.5	

(d). Region and image features using the smoothed assignment, $\mathbf{v}_{VW} = \{0.867, 0.468, 0.160\}$.



(e). Image representation for (c).



(f). Image representation for (d).

Figure 4-4. Illustration of the proposed visual-word-based image representation.