CHAPTER 4 FEATURE EXTRACTION

After a set of key frames are selected from each shot of a video, we have to extract features from these key frames and then use them to represent the video for future retrieval purpose. Here, we propose to directly use a number of indexing techniques that were commonly adopted in the field of image processing to execute feature extraction. In image indexing research, color, texture, and shape are three main features that people usually use to represent the visual content of an image. We shall review the literature of this part in the following paragraph.

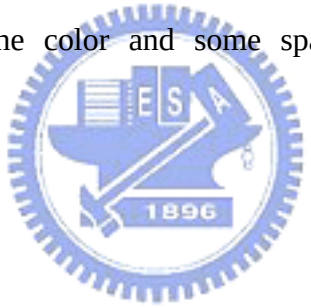
Color is the most widely used feature in the context of image indexing and retrieval because it is relatively robust to background complication and independent of image size and orientation. Moreover, the result of color-based image retrieval is considered more close to the function of human visual system. Very often the representation of color content of an image is by using color histogram because (a) it is very easy to compute; and (b) it is robust with respect to rotation and small camera viewpoint variations. However, most of color histogram based approaches are statistics-based. Thus, the structural information which is sometimes very crucial in the retrieval process is not fully used. To solve this problem, there are many color feature descriptors that incorporate spatial information have been proposed in recent years.

Texture is also an important feature of a visible surface where repetition of fundamental pattern occurs. On the other hand, it contains important information about the structural arrangement of surfaces and their relationship to the surrounding environment. Texture features are usually used to provide measures of properties such as smoothness, coarseness and regularity. However, these properties generally cannot present the presence of any particular color. The three commonly adopted approaches for describing the texture of a region are statistical, structural and spectral. The major drawback of using texture features is the task of texture segmentation, which remains a challenge and computationally intensive task. In addition, texture based techniques usually lack of robust texture models which can be used for characterization. In addition, most of the texture-based techniques are not correlated with human perception.

Shape is an important feature to represent the profile and physic structure of an object. In image retrieval applications, shape features can be classified into global and local features. Global features are the properties derived from the entire shape such as roundness, central moments and eccentricity. Local features are the properties derived by partial processing of a shape including size and orientation of consecutive boundary segments, points of curvature, corners and turning angle. Fourier Descriptors and

Moment invariants are two principal schemes of shape representation. Shape information is generally used to describe local objects instead of global information. Moreover, retrieval by shape similarity is a difficult problem because of the lack of exact definition of shape similarity, which accounts for the various semantic qualities that humans assign to shapes.

In this thesis, we shall adopt the color feature in the process of video retrieval due to its robustness, effectiveness and efficiency in image retrieval process. In addition, its close relation to the human visual system also provides a strong support for us to use it. We shall integrate the color and some spatial information together for describing color features.



4-1 Previous work

Color histogram is widely used as a descriptor of color features because (a) it is easy to compute; and (b) it is robust with respect to rotation and small camera viewpoint variations. However, classical color histogram-based approaches do not consider spatial information about pixels arrangement. As a result, very different images might have similar color histograms.

Stricker and Dimai [23] proposed another color descriptor that does consider the effect of spatial information. They first divide an image into an elliptical central region and four corners. Then they calculate the color histogram of each region and compute the first three moments (average color, variance and skewness) from each color histogram. They use the above features to represent the color information of each region. Finally, they collect the comparison results from each of the above mentioned sub-images, by setting more weight to the central region. However, this approach is a strictly domain-dependent solution: it could be effective for an archive of photographs, but it might not work well in other applications.

Pass et al. [24] present a histogram-based approach named color coherence vector (CCV) for comparing images that incorporates spatial information. They classify each pixel of a quantized image as either coherent or incoherent, based on whether or not it is part of a large similarly colored region (a region is determined as large if its size exceeds a user-set value). For each color c_i, CCV stores the number of coherent pixels, α_{c_i} and the number of incoherent pixels, β_{c_i}. Thus, each entry in the CCV is a pair (α_{c_i}, β_{c_i}) and the whole coherence vector is defined as:

$$CCV(I) = \{(\alpha_{c_1}, \beta_{c_1}), (\alpha_{c_{21}}, \beta_{c_2}), \ldots..(\alpha_{c_n}, \beta_{c_n})\}.$$

By separating coherent pixels from incoherent pixels, this method provides a finer distinction between images than color histogram-based approach. However, this approach does not contain the information about the location of each similarly colored region. As a result, different images shown in **Figure 6** may have the same CCV.

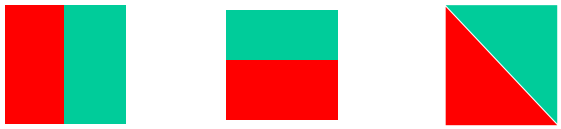


Figure 6

Chinque et al. [25] proposed another approach named spatial chromatic histogram (SCH) that records spatial information of each color such as the location of pixels of similar color and their arrangement in the image. For each color in the quantized image, the proportion of the color is calculated, and the spatial information includes the coordinate of the center of their spatial distribution and the corresponding standard deviation from the center. The distance between two SCH, *H* and *H'*, is defined as:

$$D(H,H') = \sum_i \min(h_{H(i)}, h_{H'(i)}) \left(\frac{\sqrt{2} - d(b_H(i), b_{H'}(i))}{\sqrt{2}} + \frac{\min(\sigma_H(i), \sigma_{H'}(i))}{\max(\sigma_H(i), \sigma_{H'}(i))} \right)$$

Consider an image that contains two distinct regions with similar color. Viewers prefer to get the spatial information of each region, while SCH stores the information of two regions together.

Huang et al. [26] proposed a new color feature descriptor for image indexing called color correlogram that expresses how the spatial correlation of colors changes with distance. In their work, the color correlogram is a set of values, $\gamma_{c_i,c_j}^{(k)}$, that is the probability of a pixel with color c_i at distance *k* from a pixel with color c_j:

$\gamma_{c_i,c_j}^{(k)} = \underset{p_1 \in I_{c_i}, p_2 \in I}{pr}[p_2 \in I_{c_j} \mid |p_1 - p_2| = k]$, where p_1, p_2 are two pixels of image *I*, and I_{c_i}, I_{c_j} are the set of pixels of colors c_i and c_j.

This feature descriptor can tolerate significant change in the appearance of the same scene caused by the changes in viewing position and background or camera zooms.

Wynne et al. [27] proposed an integrated color-spatial approach based on maximum entropy discretization method. They first calculate two sets of representative colors, one is selected from global image histogram and the other is selected from a color histogram of a predefined central window. Then they extract spatial information based on maximum entropy discretization with event covering method. The color-spatial information of an image is a list of two tuples $\{(c_1, r_1), ..., (c_n, r_n)\}$, where c_i denotes a chosen color and r_i is a list of cluster regions of color c_i.

4-2 Our approach

In this thesis, we propose a new color-based approach to conduct video retrieval. First, all the frames in a video are mapped to the feature space (H, S, I, X, Y). Then, we use *K*-means clustering to generate *K* clusters in the feature space. We calculate the mean vector of every cluster as the color-spatial information of each frame. These computed mean vectors are used to represent the visual content of frames. Instead of using RGB color space, we use the HIS space by taking the human visual factor into account. The detail of our approach is introduced in the rest of this section.

We first transfer the RGB color space to the HSI (Hue, Intensity-value, Saturation) space. $I = \frac{1}{3}(R+G+B)$ is defined as the intensity, where R, G and B have been normalized to [0,1], and we project the RGB cube on the plane corresponding to V. On the plane, $H = \cos^{-1}\left\{\frac{\frac{1}{2}[(R-G)+(R-B)]}{[(R-G)^2+(R-B)(G-B)]^{1/2}}\right\}$ is an angle corresponding to the value of hue, and $S = 1 - \frac{3}{(R+G+B)}[\min(R,G,B)]$ denotes the ratio of saturation, as shown in **Figure 7**.

Each pixel in an image is represented by a feature vector $V = (w_h H, w_s S, w_i I, w_x X, w_y Y)$ consisting of color features *H*, *S*, *I* and the image coordinates *X*, *Y*. In our experiment, we set $w_h = w_s = w_i = 1$, and $w_x = w_y = 1/4$. We consider the position and weight and use them to separate the irrelevant tokens and to merge the similar tokens.

By using the *K*-means clustering algorithm in the 5-*D* feature space, we generate *K* clusters in the feature space. The procedure is as follows:

Step 1. Choose *K* initial cluster centers $V_1^{(0)}, V_2^{(0)}, ..., V_K^{(0)}$. They are selected randomly from the given image.

Step 2. At the *t* th iterative step, we distribute each feature vector sample V among the *K* cluster domains by using the Euclidean distance relation,

$V \in C_j(t)$, if $\left\|V - V_j^{(t)}\right\| < \left\|V - V_i^{(t)}\right\|$ for all *i*=1,2,…,*K*, $i \neq j$, where $C_j(t)$ denotes the set of feature vector samples whose cluster center is $V_j^{(k)}$

Step 3. From the result of Step 2, we calculate the new cluster centers $V_j^{(k+1)}$, *j*=1,2,…,*K*, such that the sum of the Euclidean distance from all pixels in $C_j(t)$ to the new cluster center is minimized. The new cluster center is defined as

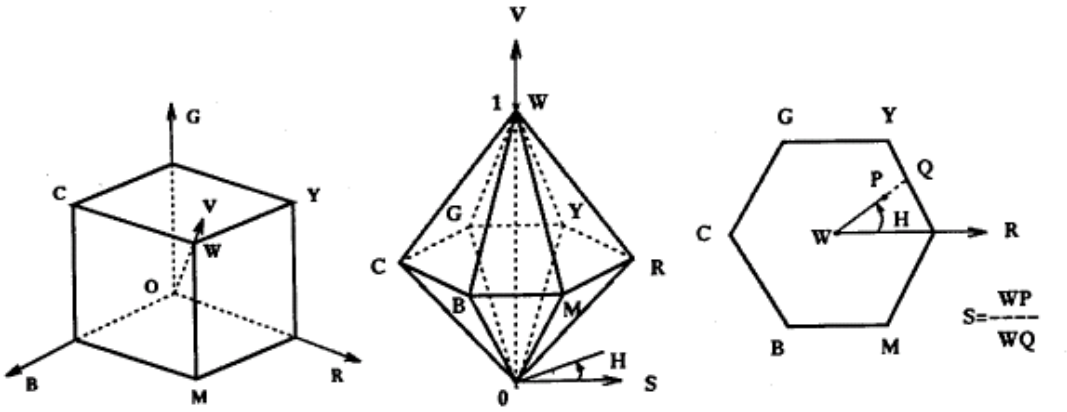
$V_j^{(t+1)} = \frac{1}{N_j} \sum_{V \in C_j(t)} V$, *j*=1,2,…*K*, where N_j is the number of feature vector samples in $V_j^{(k)}$.

Step 4. If $\left\|V_j^{(t+1)} - V_j^{(t)}\right\| < \delta$ for j=1,2,…,K, the algorithm has converged and the procedure is terminated. Otherwise go to Step 2.

After executing the K-means clustering algorithm, the tokens of similar color and position form a single cluster in the 5-D feature space. **Figure 8** shows the input images, and **Figure 9** shows the images after executing the K-means clustering algorithm on the input images. Then, we calculate the mean vector and covariance matrix of each cluster as the color description of the given image. On the other hand, any image can be represented by K mean vectors. The degree of similarity between two images, I_1 and I_2, is measured by the equation:

$$S(I_1, I_2) = \sum_{i=1}^{K} \frac{\min\left(Size(C_i), Size(C_i')\right)}{D(C_i, C_i')},$$

where K is the cluster numbers, C_i is the ith cluster in I_1, and C_i' is the corresponding cluster of C_i in I_2. $D(C_i, C_i')$ is the Bhattacharyya distance between the mean vectors of C_i and C_i'. For each cluster C_i, the corresponding cluster C_i' is decided by the following criterion: $\left\|M_i - M_i'\right\| = \min_{j=1}^{K} \left\|M_i - M_j'\right\|$, where M_i is the mean vector of C_i,and M_i' is the mean vector of C_i'.

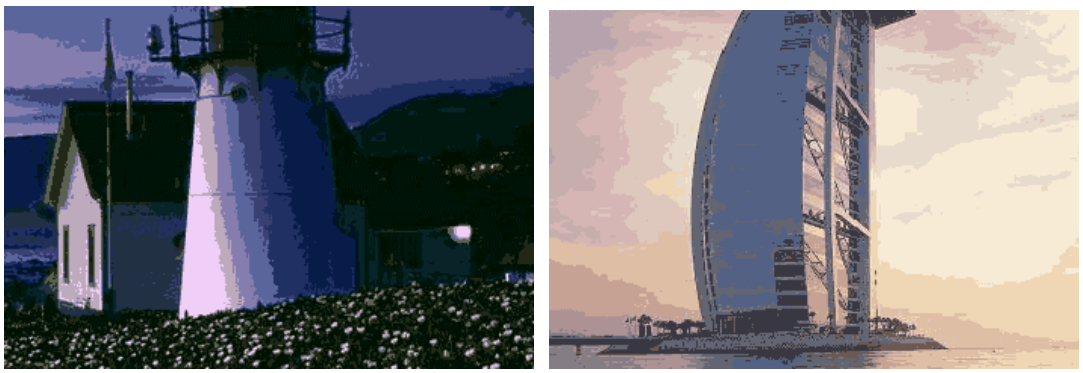


(a) *RGB* Space (b) *HVS* Space (c) Definition of H and S

Figure 7

(a) (b)

Figure 8



(a) **(b)**

Figure 9