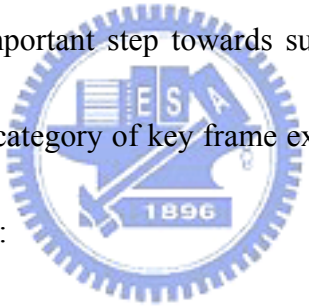
CHAPTER 3 KEY FRAME EXTRACTION

Efficient access of correct video clips from a video database is not an easy task due to the huge amount of unstructured data inherently contained in every video. Related issues such as video abstraction/summarization [16]、 feature description [17]、 and the design of matching metrics [18], have been extensively explored in the past decade. In order to reduce redundant information in a video, we have to segment a video stream into shots since a shot is an unbroken sequence of frames recorded from a single camera, which forms the building block of a video. In order to access videos in an efficient way, people working in this research field did propose several mechanisms to cut down the size of every shot. The guideline of cutting down the size of a video shot is that the topology of the shot cannot be changed too much. Under these circumstances, previous researchers proposed to either represent the shot by its most significant key frames or by a most meaningful video clip. In this work, we intended to take the former one to represent a shot. We extract the key frames from each shot based on the noticeable visual content in the shot. These extracted key frames will be further used for video annotation [4] and video retrieval [4] in our approach.

In this section, we will focus on the issue of key frame extraction based on the color feature. Depending on the complexity of the visual content in a video, one or more key frames will be extracted from a shot. Before introducing our approach, the related work using color feature for efficient database retrieval conducted in the past decade will be introduced.

3-1 Related key frame extraction techniques

Video key frames not only represent the salient visual content of shots, but also greatly reduce the amount of data required in video retrieval. Therefore, the process of key frame extraction is an important step towards successful video retrieval. Some related work that fall into the category of key frame extraction will be introduced and classified into four approaches:



- **The first/last frame approach [**19**]**

After the video stream is segmented into shots, a natural and easy way of key frame extraction is to use the first or the last frame as the key frame of each shot. Although this approach is very simple and fast, the number of key frames for each shot is limited to one, regardless of the visual complexity of the shot. Furthermore, the first or the last frame normally is not stable and does not capture the major visual content of a video shot.

- **Visual content based approach**

Zhang et al. [20] proposed to use multiple visual criteria to extract key frames. The first frame in a shot is always selected as the first key frame. Then they use the following two criteria to determine whether more than one key frame needs to be chosen.

-Color feature based criterion:

After the first key frame is selected, following frames in the shot will be compared against the last key frame based on the similarities defined by color histogram or moments. If a significant content change occurs between the current frame and the previous key frame, the current frame will be selected as a new key frame.

-Motion based criteria:

Global motions resulting from camera operations are important source of content change, thus they are also important clues for key frame selection. However, color histogram and moments representations often cannot capture such motions quantitatively. Therefore, they use two new criteria below to extract other key frames in the shot. For a panning-like sequence, if a frame is shifted by 30% in any direction from the previous key frame, a new key frame should be chosen. For a zooming-like sequence,

the first and the last frames are selected as key frames, since one can represent a global, and the other can represent a more focused view.

Although this approach is simple, it is possible that the number of key frames would be more than one due to the complexity of the video content. The major drawback of the above mentioned approach is that it does not effectively capture the major or significant content of the video shot, since the first frame is not necessarily a key frame.

- **Motion analysis based approach**

Wolf [21] proposed a motion-based approach to extract key frames. The basic idea of his approach is the stillness in a shot emphasizes the image for viewers. Therefore, he first computes the optical flow for each frame, and then computes the sum of the magnitudes of the components of optical flow at each pixel as a motion metric for each frame. Finally, he analyzes the motion metric and select key frames at the local minima of motion. This technique allows us to identify both gestures which are emphasized by momentary pauses and camera motion which links together several images in one shot. Moreover, this approach is more sophisticated due to its analysis on motion. However, it is computationally expensive and its underlying assumption of local minima of motion is not necessarily correct.

- **clustering based approach**

In [22], Zhaung et al. proposed a key frame extraction approach based on unsupervised clustering. First, they cluster all frames of a given shot into several clusters. After clusters are formed, the next step is to select key frames by two strategies. Only the clusters which are big enough are considered as key clusters, and a representative frame is extracted from the cluster as the key frame of the shot. This approach might effectively capture the major visual content of the video shots and might be efficient to compute. However, it is well known that the threshold parameter δ which controls the density of clustering is hard to be determined.

Ideally, the key frames of a shot should be able to characterize the semantics of a shot. However, at current stage, most of the existing approaches are not advanced enough to automatically generate such kind of key frames. Instead, we have to select key frames based on low-level visual features, such as color, texture, and shape of the salient object in a shot. In the next subsection, we will present our proposed approach that is both efficient and effective in key frame extraction.

3-2 Our approach

Before doing key frame extraction, we have to decompose a video sequence into several shots. The approach of our shot boundary detection is that we calculate the histogram distance measure of successive frames by the following criterion.

$HDM(f_t, f_{t+1}) = \frac{1}{M \times N} \sum_j |H_t(j) - H_{t+1}(j)|$, where H_t and H_{t+1} are the color histograms of two successive frames f_t and f_{t+1}. If $HDM(f_t, f_{t+1})$ is larger than a predefined threshold, a shot boundary is identified. After all shot boundaries are detected, a video sequence is segmented into several shots.

In order to represent a shot in a more efficient manner and at the same time maintain the semantic of its visual content, a new key frame selection method is proposed in this subsection. As is known to all, it is hard to define the semantic content of a shot due to its complexity. However, a good movie director will not waste films on meaningless scene or objects, thus it is a reasonable assumption that a scene or an object that lasts for a long period of time is most meaningful to the audience. The above concept tells us that we can make use of the dominant colors of a shot to capture the semantic meaning of a shot. As a result, the frames whose dominant colors are very similar to the dominant colors of the shot will be selected as key frames. The details of the proposed key frame extraction approach will be introduced in the rest of this section.

First, we calculate the RGB color histogram for each frame. The original RGB color domain is quantized into 16 values for R and G, 8 values for B to avoid the sensitivity caused by noises and also to reduce the storage space. We choose K dominant colors based on the significance ranked in the color histogram. The significance here means the proportion a color occupies in the color histogram. Then for each color, we calculate the number of times that it is selected as the dominant color in the shot. The more the color is selected as the dominant color, the more important it is considered in the shot. The procedure is described in **Figure 3**.

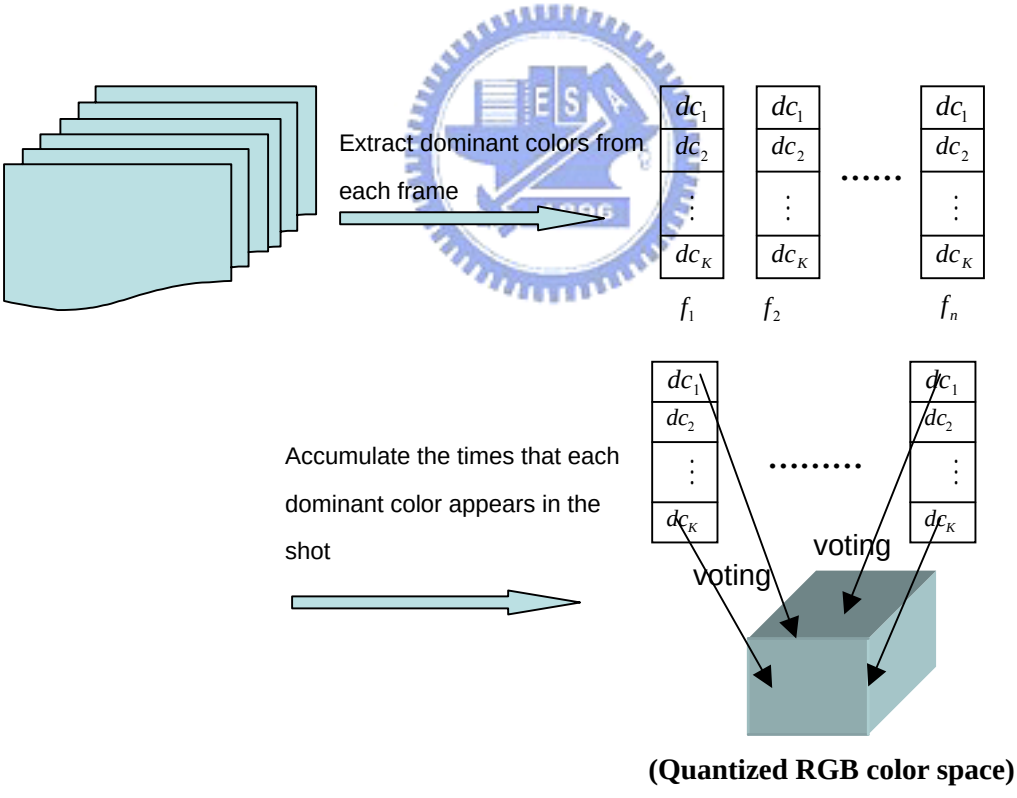


Figure 3

Now, we have the information about the proportion of each dominant color in each frame and the times that a color is selected as a dominant color in the shot. For each frame *f* in the shot *s*, we define the correlation between *f* and *s* as follows:

$$Cor\ (f, s) = \sum_{i=1}^{K} Ratio\ (dc_i) \times times\ (dc_i),$$

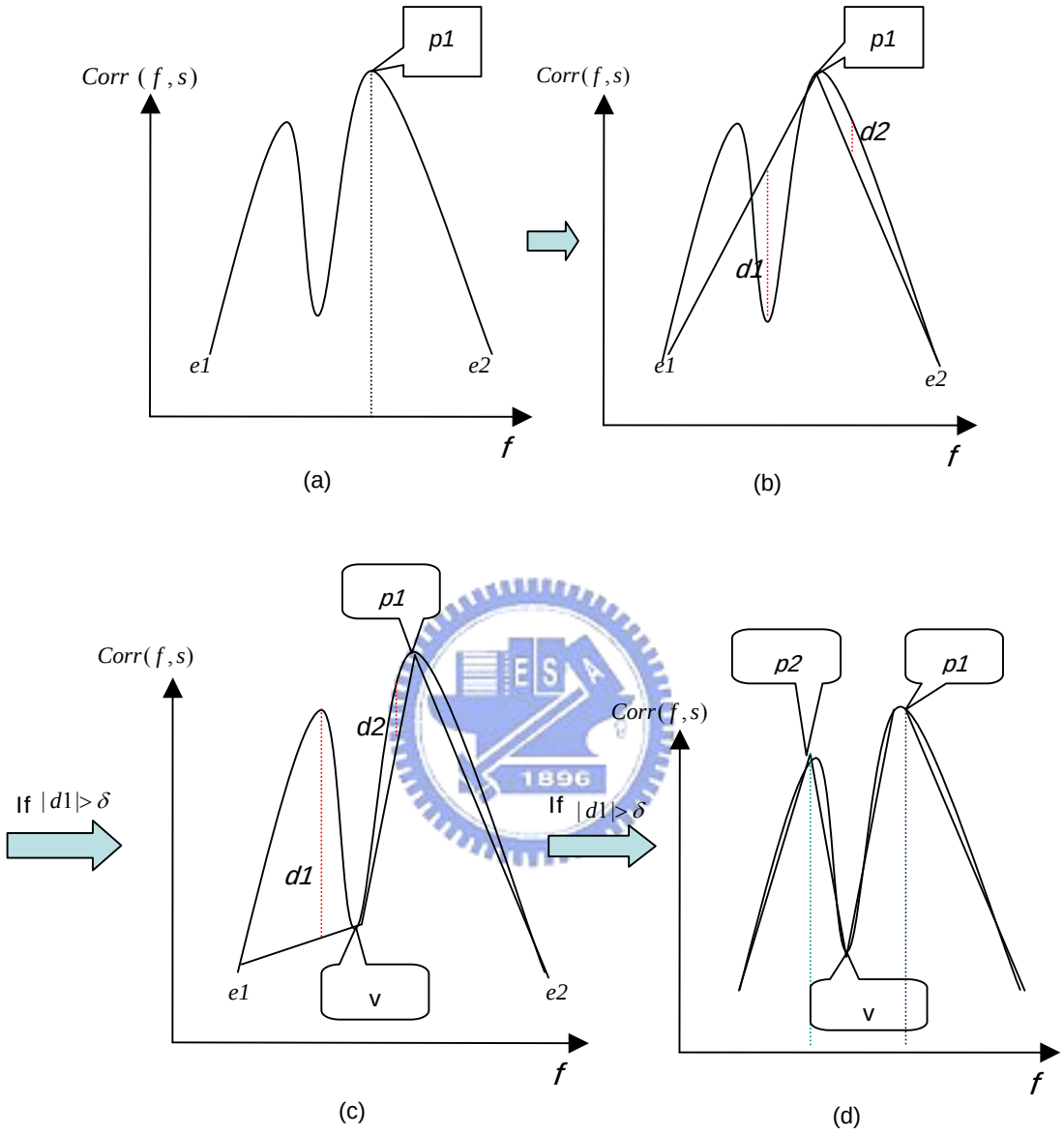
where $Ratio(dc_i)$ denotes the proportion of the ith dominant color dc_i of f, and $times(dc_i)$ denotes the number of times that dc_i is selected as a dominant color in s.

After we calculate the correlation value of each frame, the next step is to select the key frames from the shot. From the constituent frames in a shot, we choose those frames that are most correlated with the shot as key frames of the shot. Since the correlation values of all frames in a shot represent the color variation within the shot, we shall extract more than one key frame to capture the content changes within the shot if the variation in large. Here, we propose to take all local maxima of the curve formed by all correlation values as key frames. The algorithm we use to extract key frames can be summarized as follows:

Suppose there are *N* frames within a shot, and (i, c_i) denotes the ***i***th frame with correlation value c_i to the shot.

1. Using Gaussian filter to smooth the curve formed by linking all of the correlation values of all frames.
2. Choosing the frame *k* whose correlation value is the highest as a key frame of the shot. Then one can use two lines to approximate this curve, one is from $(1, c_1)$ to (k, c_k); the other one is from (k, c_k) to (N, c_N). Whether one more key frame should be selected is determined by the following step.
3. For each line from (i, c_i) to (j, c_j), we can find an approximate correlation value L_h for every frame *h* located in between frame *i* and frame *j*. Suppose there is one frame, say *x*, located in between frame *i* and frame *j*, such that $\|c_x - L_x\|$ is the largest difference that can be found. On the other hand, (x, c_x) is the highest peak or the lowest valley on the curve from (i, c_i) to (j, c_j). If $\|c_x - L_x\|$ is smaller than the predefined threshold δ ,we stop the process. Otherwise, frame *x* with the highest peak will be chosen as a key frame of the shot, and we use two sub-lines to replace the original line for approximating the curve, where the first one is from (i, c_i) to (x, c_x), the second one is from (x, c_x) to (j, c_j). Then, execute step3 again for further selection of key frames until nothing can be found.

We use a simple example to illustrate our key frame extraction algorithm and show it in **Figure 4**.



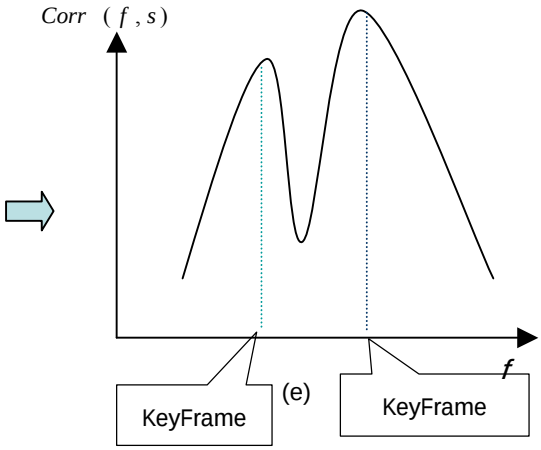


Figure 4

As shown in **Figure 4**(a), we choose the highest peak *p1* on the curve. The frame whose correlation value is *p1* will be selected as the first key frame. Then, we use two line segments to approximate this curve, as indicated in **Figure 4**(b). Suppose *d1* is the distance between the line segment $\overline{e1\,p1}$ and the lowest valley on the curve. Because *d1*>δ,the line segment from *e1* to *p1* cannot precisely approximate the curve from *e1* to *p1*. Since *d2* ≤δ, we do not have to further split $\overline{p1e2}$. The split of $\overline{e1\,p1}$ is as follows. First, we divide $\overline{e1\,p1}$ into two line segments as indicated in **Figure 4**(c). Then we compute two distance *d1* and *d2* by the same way described above. Since *d1*>δ, we divide $\overline{e1v}$ into two line segments as shown in **Figure 4**(d). It is obvious that the four line segments shown in **Figure 4**(d) can better approximate the total curve. From **Figure 4**(d), a new peak *p2* appears on the curve and its corresponding frame can be chosen as the second key frame in the shot as indicated in **Figure 4**(e).

In order to represent a shot more efficiently and also maintain the visual semantic content of the shot, we propose a new key frame selection method. Our algorithm is very effective and efficient since it actually selects the key frames that contain the semantic scene or objects in a shot. As an example, given a video clip with two shots that introduces the scenes of country site. We shall show how our proposed key frame extraction approach works. **Figure 5(a)** shows the correlation value of each frame and each black straight line refers to each peak. The extracted key frames are shown in **Figure 5(b)**. We can find that the first shot has an obvious change, thus the first shot selects more than one key frame. Moreover, all extracted key frames are most correlated with the shot.

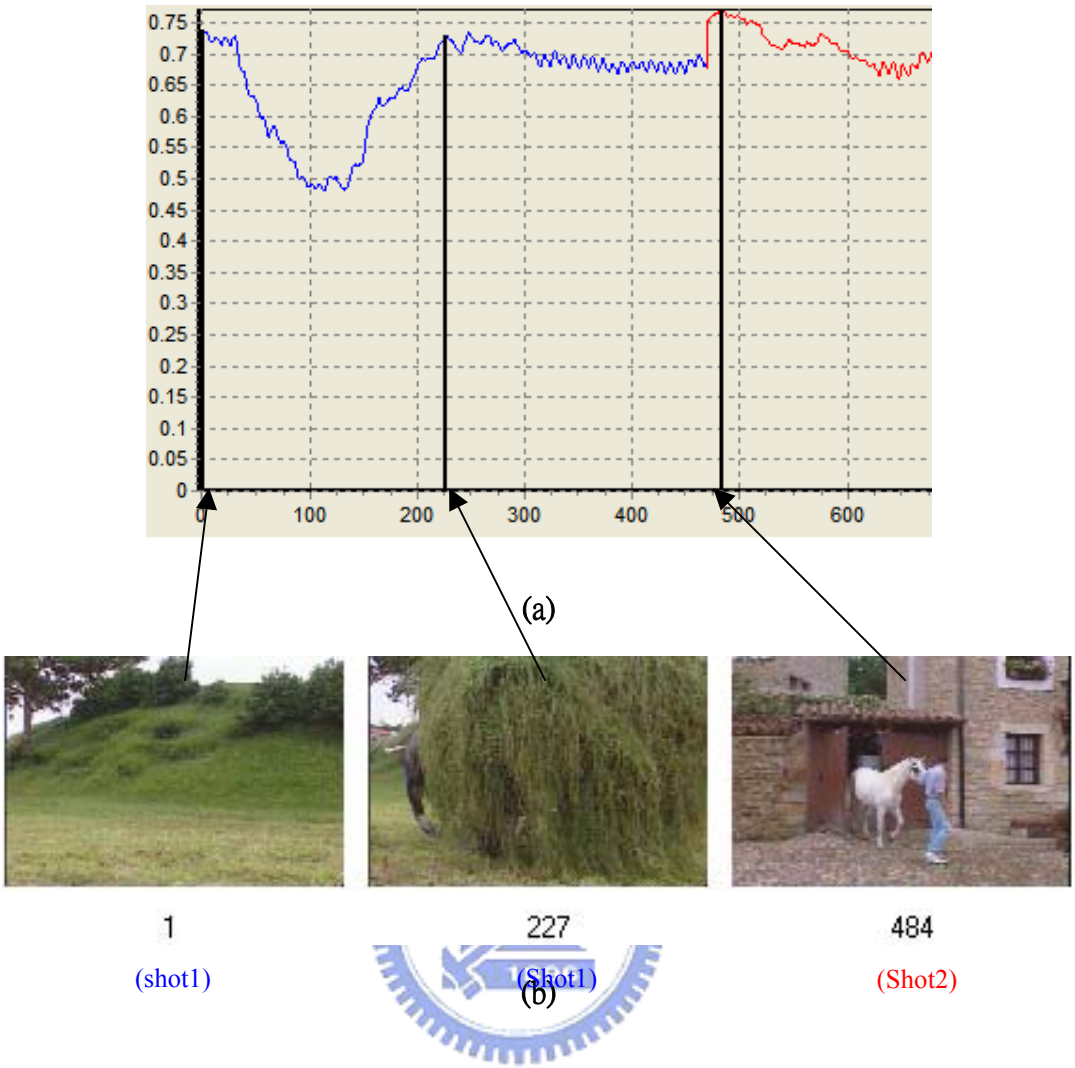

0.75
0.7
0.65
0.6
0.55
0.5
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
100
200
300
400
500
600
(a)
1
227
484
(shot1)
(Shot1)
(Shot2)
(b)

Figure 5