A Video Summarization Method Based on Key Frames Extracted by TMOF

Xiaohua He[1] and Jian Ling

School of Information Engineering, Wuhan University of Technology, Wuhan, China

[2]*School of Electronics and Information, Zhejiang University of Media and Communications, Hangzhou, China*

Email: hxh8848@126.com and lingjian99@yeah.net

***Abstract*—In this paper, we propose a video summarization method based on the Temporally Maximum Occurrence Frame (TMOF). First, the key frames are extracted from the video and then they are clustered by calculating the distance between their feature vectors; the TMOF is constructed in the clustered collection. Finally, the video summarization is formed by the frames with the smallest distance from the TMOF. Taking a news video as example, the experiment result shows that the algorithm of video summarization meets the video semantic well.**

***Keywords*—TMOF; key frame; video summarization; image clustering**

I. INTRODUCTION

Video summarization is an effective tool to realize video retrieval and it can help people find the part that they are the most interested among a large amount of video programs. Usually, people will see the catalog when they read a book and find the chapter they are interested in, and then turn to the page they want to read. In video retrieval, people also hope to find a summarization for a long video. The difference is that, a video summarization does not only contain words, but can also contain the audio and video information that people are likely to understand. So, we can define a video summarization with video clips or/and images extracted from the original video in order to facilitate browsing for videos. Also, the sequence retains the main content of the video and omits a lot of detail [1, 2]. People can browse the video, and decide whether or not to watch the video in detail.

II. MAIN FORMS OF VIDEO SUMMARIZATION

Video summarization includes static and dynamic video summarization. Static video summarization expresses the video with a group of key frames that can represent the content of video. But, dynamic video summarization expresses the video with a short video collection that represents the original video. Compared with the abbreviated video or slide, the still pictures with limited exhibition space of a static summarization have been found by a user survey to be more acceptable [3]. So, this paper uses static video summarization.

The main forms of static video summarization are: the title, the poster and the storyboard.

1. The title describes the video with a short sentence. This is the most compact form of video summarization, but it doesn't make full use of video multimedia features because of less intuition. So, it is difficult to interact with users.
2. The poster, also called a video thumbnail, is a kind of static summarization constituted by a few pictures extracted from video with some information about the video. They can be characters, or video content abstract, etc. Compared with the title, the poster provides video pictures so it can give users very intuitive visual information. Making a poster is simple, as long as you have pictures and text.
3. The storyboard cuts the video into shots first, and then extracts all key frames, and the key frames are combined in chronological order. The storyboard shows the structure of the video. Also, users can easily find the interesting part of the video in the process of browsing.

This paper proposes a video summarization method of news video based on TMOF. First, the key frames are extracted from the news video. Then, the key frames are clustered according to the similarity distance of key frames feature. Finally, the TMOF is constructed in each cluster and the video summarization is formed by the frame with the smallest distance from the TMOF. Although the method in this paper, the poster and the storyboard all extract video pictures, the poster only extracts a few static pictures, so it can't describe the whole video. Compared with the poster, the storyboard extracts all key frames of the video and contains more video semantics. It constitutes the general framework of the video, so it can provide more detailed information of the video. But the number of key frames will be a lot, and there will be redundancy between key frames. The method in this paper keeps only a frame in the similar key frames, so, the number of key frames is greatly curtailed. That is to say, the video summarization is constructed by "key frame of key frames".

978-1-4673-2546-2/12/$31.00 ©2012 IEEE

III. Video Frame Feature Extraction

Color, texture and shape are the most basic visual characteristics of an image. The color is the most simple and effective characteristic of an image [4]. It has good stability and is not sensitive in the direction and size, so it is one of the first choices in image recognition characteristics. Common color models have RGB, HSV, or CMY, etc. The RGB color model is the most commonly used in image processing, but it has a certain gap between visual perception with people. The HSV color model is oriented to visual perception, so it is a more intuitive color representation to users. HSV has the three elements that correspond directly to human eye in observing color vision characteristics: hue, saturation and value. Therefore, this paper uses HSV color model.

Considering an image generally contains many colors, if we calculate the histogram after quantizing the HSV, it can save storage space and reduce the computational complexity. Because it is easy to calculate, the H, S, V 3D characteristic vectors with different weights can be combined into 1D characteristic vector in the practical operation. In the three vectors, compared with saturation and value, the human eye is more sensitive to hue. So, we can do non-uniform quantification for the three components, and divide H into more quantitative levels: H divided into 8 levels; S and V are divided into three levels. The color space is divided into 72 colors after quantification.

The methods of color features description include color histogram, color correlogram, and coherence vector, etc.; color histogram is the most commonly used. Color histogram is a very important and common feature to describe the color characteristic of the image. The basic idea is to divide the color space into some subspace, and then classify the number of pixels belonging to each subspace of the image. That is to say, the number of pixels of each color subspace in the image is calculated, then with all the color values as the horizontal axis and the number of pixels of each color value as the vertical axis, the corresponding color histogram of the color image is established. This paper divides color space into 72 kinds of colors, and calculates the probability of color pixel of the image, and gets a characteristic vector containing 72 components.

The color histogram contains only the frequency of a color value, but loses the pace information of the pixels. Therefore, in order to get the spatial distribution information of the color, the image is usually divided into appropriate blocks. If the blocks of the image are too small, it cannot be partitioned. More image blocks can improve the spatial resolution of the image, but it also increases the storage space of the image characteristics, and may make the image too broken or color information not rich enough, to make the retrieval precision of image decline. Therefore, the actual block number should be a compromise, and the number of blocks can be adjusted according to the image frame size. Because each block represents different important degrees of space information, according to the position of the block in the picture, different weight coefficients are set, and the closer to the center, the greater the weight will be.

After dividing the image into blocks, we compare the similarity of the corresponding blocks between the two images; the commonly used methods of similarity measure are Euclidean distance, quadratic form distance, and absolute distance, etc. In this paper, we use the Euclidean distance, as shown in Formula 1:

$$D(A,B)=\left[\sum_{i=1}^{72}\left|H_A(i)-H_B(i)\right|^2\right]^{\frac{1}{2}} \tag{1}$$

D(A,B) is defined by comparing the similarity of the corresponding blocks between the two images of A and B. $H_A(i)$ and $H_B(i)$ are defined by the ith component of the characteristic value of corresponding blocks. After calculating the distance between each corresponding blocks, the total distance of the two images can be obtained by multiplying each distance by weight and adding them together, as shown in Formula 2:

$$L(A,B)=\sum_{i=1}^{n}\left[D_i(A,B)\times W_i\right] \tag{2}$$

We can judge the similarity of the two images according to the total distance L. The smaller the distance, the greater the similarity will be. Conversely, the greater the distance, the smaller the similarity will be. According to the similarity between the key frames, the key frames are clustered, and similar key frames are gathered into one class [4].

IV. TMOF Construction

The significant frame among a video frame collection should be one with the smallest distance from TMOF. Where the TMOF is defined as follows: 1) Frame image size is consistent with video picture size. 2) Every pixel gray value of TMOF is determined by the gray value of the highest probability appearance in the corresponding position of all frames.

The constructing method of TMOF is shown in Figure 1. First of all, the time gray histogram is structured with the pixel gray value in the corresponding pixel of all frames in the sub-shot. In order to reduce noise, the gray histogram is imposed on a Gaussian smoothing. The pixel gray value of TMOF is the gray value of corresponding pixel in the smooth histogram that appears in the highest probability. Calculation methods are as follows:

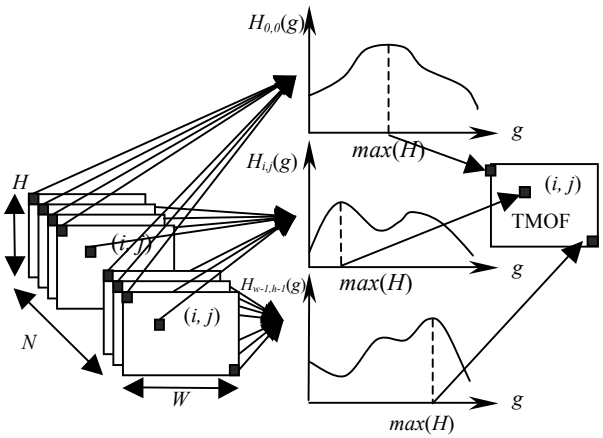

Figure 1. Derivation of TMOF

1. Time gray histogram is calculated according to Formula 3:

$$H_{ij}(g) = \sum_{n=0}^{N-1} \delta[I_n(i,j) - g] \tag{3}$$

$$(0 \le I \le 255; 0 \le i \le W-1; 0 \le j \le H-1)$$

Where H_{ij} is the time histogram of pixels in position (i,j), N is the total number of frames in the sub-shot, $I_n(i, j)$ is the pixel gray value in position (i, j) of the nth frame, and W and H are the length and height of the video image, respectively. Delta function is defined by the formula as follows:

$$\delta(x) = \begin{cases} 1 & \text{当} x = 0 \\ 0 & \text{当} x \neq 0 \end{cases} \tag{4}$$

2. In order to reduce the noise effect, each histogram data calculated by Formula (3) are smoothed by a convolution with Gaussian function:

$$H'_{ij}(g) = H_{ij}(g) * Guass(\sigma) \tag{5}$$

where Guass(σ) is the one-dimensional Gaussian function and * expresses convolution.

3. The pixel gray value of TMOF is the gray value of corresponding pixel in the histogram which appears in the highest probability:

$$G(i,j) = \max[H'_{ij}(g)] \tag{6}$$

After obtaining TMOF of the sub-shot, characteristic vectors of all frames are calculated using the methods referred in the second section and the distance between TMOF and the frames of the sub-shot is calculated. The frame with the smallest distance from TMOF is regarded as the key frame of the sub-shot.

V. News Video Summarization

The key to video summarization is how to use the minimum amount of elements to cover as much video information as possible, but generally, the number of key frames will be more. To match the commentary, television director often cuts a long shot into some sub-shots, and intersperses the broadcast with the other shots. That is, the similar shots may appear in the different locations of the news stories. So, some key frames will be similar [5] and the key frames need to be further processed, and the video summarization is formed by a fewer number of key frames. Generally speaking, in a news video, the important shots last longer, and therefore the key frames of the shots that last a short time can be removed.

The algorithm of video summarization:

(1) Remove the key frames of the shot that last shorter than 4 seconds.

(2) Key frames are clustered according to their similarity:
   a. Put the first key frame of the key frames collection into the first set;
   b. Take out the next key frame, if key frame collection is empty, turn to exit, or go to Step 3;
   c. Calculate the similarity between the current key frame and all the existing collection. The similarity between key frame and collection is defined as: the similarity between current key frame and the centroid of the collection; here the similarity algorithm uses the method mentioned in Section III.
   d. Compare the largest similarity with the predefined threshold, if over the threshold the key frame will be included in the largest similarity collection, otherwise produces a new collection and put the key frame in the new collection, return to Step 2;

(3) Structure the TMOF in each collection; every pixel gray value of the TMOF is determined by the gray value of the highest probability appearance in the corresponding position of all frames. Calculate the distance between each key frame and TMOF; determine the key frame of the collection according the distance.

Through the above three steps, key frames can be further reduced. Greatly different key frames in the collection can be kept, and some key frames belonging to less important shots or similar with another can be moved. The video summarization can be constructed with a few representative key frames.

Take a Hangzhou news broadcast as an example. We extracted twenty-two (22) key frames using the method of key frame extraction, as shown in Figure 2. Ignore the key frames of the shots that last less than 4 seconds, and cluster the key frames according to the similarity. So, we get eight collections, as shown in Figure 3.The video summarization is formed by selecting a suitable

frame in every collection by the TMOF algorithm, as shown in Figure 4.

Figure 2. All key frames

(a) (b) (c) (d)

(e)

(f)

(g)

(h)

Figure 3. The clustered key frames collections

In the example above, we reduced twenty-two (22) key frames to eight and the news is now described by these 8 key frames, as shown in Figure 4. Similar shots that appeared in different locations were greatly eliminated, in order to eliminate the redundant information.

Figure 4. The video summarization

VI. Conclusion

Different applications have different requirements for video summarization. This paper proposes an algorithm of static video summarization formed by image frames through further processing the video key frames. The algorithm has low time complexity and less data, and is suitable for processing massive videos automatically. However, the algorithm is not suitable for the computer to understand and retrieve the results of the summarization, because the summarization is composed entirely of images. If video summarization is combined with semantic information such as voice or text that is extracted from the original video or key frames, it will be able to effectively improve the expression ability of the video summarization.

References

[1] W. Sack, and M. Davis. Assembling video sequences from story plans and content annotations, in *Proc. Intl. Conf. on Multimedia Computing and Systems*, 1994, pp. 30-36.

[2] A. Hanjalic, and H. Zhang. "An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 9, no. 8, pp.1280-1289, 1999.

[3] A. Komlodi, and G. Marchionini. "Key frame preview techniques for video brow sing," in *Proceedings of the 3rd ACM Conference on Digital Libraries*, Pittsburgh, 1998, pp. 118-125.

[4] Y. Wang, and B. C. Li. "Video abstraction based on K-L transform and clustering," *Application Research of Computers,* pp. 3585-3587, 2010.

[5] X. D. Luan, Y. X. Xie, L. Ying, L. D. Wu, and P. Xiao. "Research on news video summarization based on EDU model," *Journal of System Simulation*, pp. 3770-3774, 2007.