

Retrieval Method for Video Content in Different Format Based on Spatiotemporal Features

Xuefeng Pan^{1,2}, Jintao Li¹, Yongdong Zhang¹, Sheng Tang¹, and Juan Cao^{1,2}

¹ Institute of Computing Technology, Chinese Academy of Sciences,
Beijing 100080, China

{xfpan, jtli, zhyd, ts, caojuan}@ict.ac.cn

² Graduate School of Chinese Academy of Sciences,
Beijing 100080, China

Abstract. In this paper a robust video content retrieval method based on spatiotemporal features is proposed. To date, most video retrieval methods are using the character of video key frames. This kind of frame based methods is not robust enough for different video format. With our method, the temporal variation of visual information is presented using spatiotemporal slice. Then the DCT is used to extract feature of slice. With this kind of feature, a robust video content retrieval algorithm is developed. The experiment results show that the proposed feature is robust for variant video format.

1 Introduction

With the advances in multimedia and Internet applications, techniques for video retrieval are in increasing demand. The existing approaches in clip-based retrieval with visual cues are mainly based on image retrieval techniques. The most common of these are (1) to use color histogram of key frames (2) to match key frames using color, texture combined with motion information and (3) to use correlation between frames using ordinal measure [1]. As pointed out in [2], ordinal measures based techniques give better performance in comparing with color or motion based methods. But the number of partitions is critical when there existing display format changed contents [3].

It is known that two video clips with same content but compressed in different formats may have distinct color or texture characteristics. The color or texture based features used in image matching are no longer fit for retrieval the same video content in different format. Due to this consideration, we present a retrieval method based on spatiotemporal slice for matching video clips with same content but in different format in this paper. A spatiotemporal slice is a collection of scans in the same position of every frame which indicates the coherency of the video [5]. Ngo used the color and texture of spatiotemporal slice for video clustering and retrieval in [4]. But the retrieval for video content in different format was not discussed in Ngo's paper. In this paper, we will develop a method using the low-frequency AC DCT coefficients of spatiotemporal slice blocks to match video clips with the same content but compressed in different formats.

2 Clips Retrieval

The horizontal slice which is the collection of scans at the middle row of every frame is used in our method. The slice is converted into gray image and the height of slice is resized to $N \times N$. Then the resized gray slice is segmented into N blocks. By analyzing these blocks with discrete cosine transform (DCT), the variation information of video clip is generated. A sample of grayed slice and blocks are shown in Figure 1. In this way we can transform clip retrieval problem into block

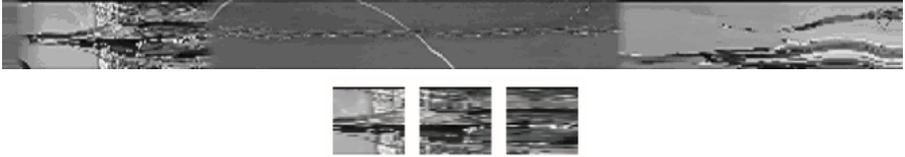


Fig. 1. Grayed slice and blocks from a video clip (n=32)

sequence matching problem. Because most energy of DCT is compacted on low-frequency AC DCT coefficients, the first P low-frequency AC DCT coefficients of each block are gathered to form a feature vector for the blocks and video content. Let $C_Q = (Q_1, Q_2, \dots, Q_m)$ and $C_T = (T_1, T_2, \dots, T_n) (m \leq n)$ denote feature vector group of the query clip Q and the target clip T . $Q_i = \langle Q_i[0], \dots, Q_i[P-1] \rangle$ denotes the feature vector of the i th block. We use Video Feature Vector Similarity (VFVS) to locate Q in T . The matching process is as the following steps.

- (1) Shift C_Q along C_T from the beginning of C_T . Let $C_T^i = (Y_i, Y_{i+1}, \dots, Y_{i+m-1})$ denote some part of C_T that has the same block number as C_Q starting at block $i (i \leq n - m + 1)$.
- (2) Compute the VFVS between C_Q and C_T^i using:

$$VFVS(C_T^i, C_Q) = 1 - \frac{\sum_{i=1}^L d(C_T^i, C_Q)}{\sum_{i=1}^L (abs(C_T^i) + abs(C_Q))} \quad (1)$$

Where $d(\cdot)$ is the distance metric defined on feature vector (here L_1 distance is applied), and $abs(\cdot)$ is the sum of the absolute value of elements of feature vector.

- (3) The local maximums of the $VFVS(C_Q, C_T^i) (i = 0, \dots, n - m + 1)$ values which above a certain threshold are taken as matches.

3 Experiments and Discussion

In this part we present the effectiveness of the proposed algorithm with experimental results.

Firstly, a 30 minutes MPEG-1 video is divided into 24 equal clips and the boundaries of each query clip are the multiple of block size N . Then each clip is taken as query clip Q . The algorithm proposed in section 2 is used to locate Q in original video T . In the experiments, all clips are correctly located. The similarity between Q and sub sequences of T is presented in Table 1.

Table 1. Query clips locating with aligned block boundary

Similarity	Mean	Stdev.
S_{same}	1.0	0.0
S_{diff}	0.232157	0.041292

Note: S_{same} : similarity between the query clips and the located clips having the same content; S_{diff} : similarity between the query clips and other clips having different content; Mean is the average value of the similarity; Standard Deviation. Stdev. is the standard deviation of the similarity.

Table 2. Query clips locating with unaligned block boundary

Similarity	Mean	Stdev.
S_{simi}	0.573746	0.169228
S_{diff}	0.226036	0.033999

Note: S_{simi} : the larger one of similarity value between the query clips and clips start at frame $32m$ and $32(m+1)$; S_{diff} : similarity between the query clips and other clips.

Because the feature used in this paper is slice block based, in practice, the block boundaries of query clips may not aligned with the block boundaries of target video. In experiments, the block size is set as $N = 32$. We chose the sub sequences of target clip T starting at frame $32m + i$, ($i = 0, 1, 2, \dots, 31$) as query clips Q . So the block boundary of query clips Q and target clip T are not aligned. The similarity between Q and sub sequences of T is presented in Table 2. Furthermore, the minimum of S_{same} is 0.40695. This is larger than $\text{Mean} + 4 \times \text{Stdev}$ of S_{diff} in Table 1. The similarity between Q and sub sequences of T is assumed to obey Gaussian distribution. According to the properties of Gaussian distribution, the proposed method can tell the same content as Q from the different content of Q with the probability larger than 0.9999. Four query clips Q are transcoded into 3 different formats. The average similarity between Q and reformatted clips Q' are presented in Table 3. The threshold of similarity for clip matching is set at 0.4 to locate the position of query clip Q in target clip T . In experiment we locate all the 12 reformatted clips in T with no false positive.

Table 3. Similarity between original clips and reformatted clips

Reformatted clip	Similarity
352×288 in AVI	0.88623
320×180 in AVI	0.87302
320×180 in MPEG1	0.90397

4 Conclusion and Future Work

A robust video content retrieval method is proposed in this paper. The proposed method adopts a novel feature extraction scheme for video content representation. It differs from most existing video retrieval methods in that it is no longer using the character of every single frame. The spatiotemporal features extracted with slice DCT are sensitive to clip content variation and robust to video format changing. With this kind of feature, we develop a robust video clip retrieval algorithm. The experiment results show that the proposed feature is robust for variant video format.

There are still things to do with our method. For example, the feature can be more compact, the feature vector matching method can be more efficient. The next target is to revise the method in this paper for large scale video data.

Acknowledgements. This work is supported by the Key Project of Beijing Natural Science Foundation (4051004), and Beijing Science and Technology Planning Program of China (D0106008040291, Z0004024040231).

References

1. Xian-Sheng Hua, Xian Chen, Hong-Jiung Zhnng: Robust Video Signature Based on Ordinal Measure, International Conference on Image Processing (2004). Page(s):685-688
2. Arun Hampapur, Ki-Ho Hyun, Ruud Bolle: Comparison of Sequence Matching Techniques for Video Copy Detection. Proc. Storage and Retrieval for Media Databases, Jan. 2002, Page(s): 194-201
3. Changick Kim, Bhaskaran Vasudev: Spatiotemporal Sequence Matching for Efficient Video Copy Detection, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 15, No. 1, January 2005, Page(s):127-132
4. Chong-Wah Ngo, Ting-Chuen Pong, Hong-Jiang Zhang: On Clustering and Retrieval of Video Shots through Temporal Slices Analysis, IEEE Transactions on Multimedia, Vol. 4, No. 4, December 2002, Page(s):446-458
5. Peng. S. L, Medioni. G, Interpretation of image sequences by spatio-temporal analysis, Workshop on Visual Motion, March 1989. Page(s):344-351