Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 29 (2012) 2578 – 2582

Procedia Engineering

www.elsevier.com/locate/procedia

2012 International Workshop on Information and Electronics Engineering (IWIEE)

Design of Video Retrieval System Using MPEG-7 Descriptors

Bin Liang*, Wenbing Xiao, Xiang Liu

College of Computer Science and Technology, Taiyuan University of Technology, No.79 West Yingze Street, Taiyuan 030024, China

Abstract

The paper proposes a content-based video retrieval system designed using MPEG-7 (multimedia content description interface), which provides a standard description for a video. The system consists of three parts: shot boundary detection, feature extraction and similarity measurement. In shot boundary detection, cut and dissolve can be detected using the histogram difference and skipping image difference, respectively. In feature extraction part, two MPEG-7 visual descriptors, Color Structure Descriptor (CSD) and Edge Histogram Descriptor (EHD), are used to represent the color feature and edge feature of the key frames. Lastly, the similarity between key frames is calculated using dynamic-weighted feature similarity calculation. The proposed system is tested on three kinds of videos. Promising results are obtained in terms of both effectiveness and efficiency.

© 2011 Published by Elsevier Ltd. Selection and/or peer-review under responsibility of Harbin University of Science and Technology

Keywords: video retrieval, shot boundary detection, feature extraction, MPEG-7 descriptor

1. Introduction

With both the rapid increase in the amount of generated video data and wide range of video applications, an efficient and effective management of video records is much demanding. In other word, many of these record data are currently hardly usable, and this is mainly because of lack of appropriate techniques, which can make the video content more accessible. Moreover, in automatic video management, it is not practicable to use keywords to describe every video sequence because this

* Corresponding author. Tel.: +86 0351 6960312
E-mail *address*: bin.liang.ty@gmail.com. (B. Liang)

1877-7058 © 2011 Published by Elsevier Ltd.
doi:10.1016/j.proeng.2012.01.354

annotation procedure normally requires tremendous manpower. Furthermore, keywords are more or less subjective. Therefore, retrieval and indexing based on video content is a solution.

Content-based video retrieval system is an active field of research. These systems typically include three steps: video segmentation, feature extraction and feature grouping. Video segmentation algorithms try to divide the video sequences into meaningful subgroups called shots. Over the years, a number of techniques, varying from color histogram to block based approaches with motion compensation have been proposed for this purpose [1]. Most of the presented methods work on the key frames of the shots to measure the similarity. Dimitrova *et al.* regarded the average distance of corresponding frames between two videos as the similarity measure [2]. Zhang *et al.* defined another similarity according to the sum of the most similar pairs of key frames [3]. A color histogram comparison routine is used in [4] to parse video sequence into scenes. A hierarchical video stream model which uses template or histogram matching technique to identify scene changes in a video segment is proposed in [5]. These methods, however, are useful in specific domains, and therefore not readily applicable in the development of general-purpose video data indexing and retrieval systems.

In this paper, a video retrieval system using MPEG-7 [6] descriptors is proposed. The development of the prototype and results of a usability survey using the movie are described. The system includes shot boundary detection, feature extraction and similarity measurement. MPEG-7 is a standard for describing the content of different types of multimedia data. As the first standard of the Moving Picture Experts Group to focus not on comparison, but rather on metadata or descriptions for the multimedia content, it offers richer semantics as compared with other existing audiovisual metadata like Dublin Core [7] and TV-Anytime [8]. MPEG-7 documents can be defined and modified with the help of the Description Definition Language (DDL), which is based on XML Schema with extensions to support array, matrix and some temporal data types.

In section2 we describe briefly the shot boundary detection, including both the cut and dissolve. We then focus on the feature extraction in section3. The similarity measurement using dynamic-weights is presented in section4.

2. Shot boundary detection

Shot boundary detection is an essential elementary component of video analysis. Generally, there are two kinds of shot boundaries: cut and dissolve. A cut is an abrupt transition between shots which is naturally formed by the video capturing process. A dissolve is a gradual transition between shots, which is an effect added by video editors where two adjacent shots are partly overlapped, while the frame intensities of the first shot are decreased to zero and the frame intensities of the second shot are increased from zero.

The image difference between two adjacent frames can be a good cue for detecting cuts. There are many ways to compute the difference between two images. We use histogram difference to represent two images. We can think of the image difference as a distance between two vectors in a feature space. So the features are in histogram space, with each histogram being a component of each feature. This method can be computed by the Euclidean distance or any other defined histogram distance in the histogram space. Our experiments support our analysis that the method, histogram difference, is appropriate for evaluating image difference. In our implementation we compute the histogram difference as

$$D_{RGB}(X,Y) = \sum_i \left(\left| h_x^r(i) - h_y^r(i) \right| + \left| h_x^g(i) - h_y^g(i) \right| + \left| h_x^b(i) - h_y^b(i) \right| \right) \qquad (1)$$

in which h^r, h^g and h^b are the histogram of red, green and blue, respectively.

In dissolve detection, we cannot use the image difference between two adjacent frames, because it is small during the dissolve. But there will be a large image difference between two frames if we skip an interval (such as a 2.5 frame interval). This skipping image difference can be used as a cue to dissolve detection. We adopt a method for dissolve detection which combines two criteria. Suppose we currently have a sequence of w images which is a segment in the video sequences starting with $I(t)$ and ending with $I(t+w-1)$. We define the current skipping image difference value as

$$D(t)=\left|I(t+w-1)-I(t)\right| \tag{2}$$

We define the current normalized linear error along this part of video as

$$LE(t)=\frac{\sum_{i=0}^{i=w-1}\left|I(t+1)-((1-\frac{i}{w-1})I(t)+\frac{i}{w-1}I(t+w-1))\right|}{\left|I(t+w-1)-I(t)\right|} \tag{3}$$

We detect a dissolve by the simultaneous presence of a peak of $D(t)$ and a valley of $LE(t)$. In Eq.3, we use the linearity assumption in RGB color space.

3. Feature extraction

3.1 Color feature

We use color histogram to represent the color feature. Color feature is one of the low-level features, but there are still a lot of video retrieval applications using color feature and the color feature has achieved good results. It is easy and fast to extract the color histogram feature, but the drawback is that the spatial characteristics of the image are lost, which can not represent the spatial relationship among the parts of the image. On the other hand, this ensures the invariance of the image transformation and rotation. We choose the RGB color space, as the HSV color space is insensitive to the brightness changes. Thus, the RGB color space is widely used in the literatures. In order to reduce the dimension of the color histogram, the number of bin for each color component is set as 16 and the dimension of each histogram is $16^3=4096$.

3.2 Edge feature

As the spatial relationships are lost in color histogram, we choose the MPEG-7 edge histogram to describe the features of the image in order to get the spatial relationships. Edge of the image is an important feature to represent the content of an image. One edge histogram represents the direction and frequency of the brightness changes in image, which is the unique feature and does not repeat the color histogram. The Edge Histogram Descriptor (EHD), which is used to represent the edge feature, is defined in MPEG-7. The EHD mainly represents five kinds of edge distribution for each local region (sub-image). Sub-image is one non-overlapping block of the image which is divided into 4×4 pixels. Each edge histogram of the sub-image contains five kinds of edge: vertical, horizontal, 45° diagonal, 135° diagonal and non-directional edges. Thus, each histogram of the sub-image expresses the occurring frequency of five kinds of edge. Therefore, the number of bin for each sub-image edge histogram is 5 and there are 16 sub-images in one image. The number of bin for one edge histogram is $5\times16=80$. [9] proposed the method for MPEG-7 edge histogram.

3.2 Used MPEG-7 descriptors

The following color and edge descriptors are used in our system to represent the features of the key frames:

Color Structure Descriptor (CSD) is a color feature descriptor that represents an image by both color distribution (similar to color histogram) and the local spatial structure of the color. An 8×8 element is used to embed color structure information into the descriptor. CSD uses the $l_1 - norm$ for matching as the similarity measure.

Edge Histogram Descriptor (EHD) calculates spatial distribution of five types of edge. The image is divided into 4×4 sub-images, and then the edges in 16 sub-images are categorized into five types. For matching edge histograms, global ($h^g - 5$ bins) and semi-global ($h^s - 65$ bins) edge distributions are calculated from the local histogram bins ($h - 80$ bins). We use $l_1 - norm$ for similarity matching.

4. Similarity measurement

Selecting representative frames, namely key frames, is a common approach for reducing the amount of video data to store and index for efficient content-based retrieval. A common approach for visual feature weighting is to assign fixed weights to each visual feature, as used in [10]. Our solution is to use dynamic-weighted feature similarity calculation based on the success rate of the visual similarities of different features. Inspired from interest point matching techniques, where two closest matches are compared to each other, we define the success rate (weight) of a descriptor as the ratio of similarity values of most similar match to the 5th one. So the weights for each visual feature are calculated for each query key frame separately.

The *j*th most similar visual feature to f_q^i is found as $f_{r,j}^i$ in the reference database of features f_R^i by k-nearest neighbor retrieval. We calculate the dynamic-weights of each visual feature *i* (i.e. CSD and EHD) for the query key frame *q* with:

$$w_i(f_q) = 1 - \frac{D_i(f_q^i, f_{r,1}^i)}{D_i\,(f_q^i, f_{r,5}^i)} \tag{4}$$

By using the dynamic-weights, we find the most similar reference key frame r_m by minimizing the combined dissimilarity:

$$r_m = \arg\min_r \frac{\sum_i w_i(f_q) \times (1 - D_i(f_q, f_r))}{\sum_i w_i(f_q)} \tag{5}$$

Key frame based similarities are calculated with dynamic-weighted MPEG-7 visual features. The most similar and most voted matching reference videos are reported as the retrieval candidates.

5. Experiment results

In the previous section, three main parts of the system have been presented. In this section, we verify the performance of the system through experiments. The experiments were implemented on an Intel P8700 2.53GHz personal computer with 2GB RAM running on Windows Server 2003.

The experimental data consists of the collection of 3 different types of videos. The evaluation was investigated in terms of two main measurements, namely precision and recall. Precision represents the ratio for the cardinality of correctly returned video clips over the cardinality of the resulting video clips. Recall indicates the ratio for the cardinality of correctly returned video clips over the cardinality of the relevant video clips. The experimental results are shown in Table 1. The experiment results demonstrate the effectiveness and efficiency of the proposed system.

Table 1. Evaluation of video retrieval system

Video clips	Precision	Recall
movie clip	0.72	0.74
news clip	0.81	0.79
sports clip	0.77	0.75

6. Conclusion

We propose a system for content-based video retrieval using MPEG-7. The proposed system consists of three parts: shot boundary detection, feature extraction and similarity measurement. The experimental results show that the proposed method performs well, in terms of both effectiveness and efficiency. It is clear that the system already achieves high correct retrieval rates; however, there is still some potential for improvement to combine more features of the video.

Our future extensions will focus on how to constrain the range to specific situation, such as face recognition in video. It is one branch of the content-based video retrieval and it needs more techniques about face detection and face recognition. We may need to improve the method for face recognition in video, and consider the illumination effects to human faces.

Acknowledgements

This work is supported by the Key Scientific and Technological Projects of Shanxi Province (No.20080322008). The authors would like to thank the reviewers for their valuable comments.

References

[1] S.V. Porter, Video segmentation and indexing using motion estimation, Ph.D. Thesis, University of Bristol, Bristol, 2003.

[2] Dimitrova, N., Abdel-Mottaled, M., 1998. Content-based video retrieval by example video clip. SPIE 3022, 50–70.

[3] Zhang, H.J., Zhong, D., Smoliar, S.W., 1997. An integrated system for content-based video retrieval and browsing. Pattern Recognition. 30 (4), 643–658.

[4] Nagasaka, A., Tanaka, Y., 1991. Automatic video indexing and full video search for object appearances. In: Proc. 2nd Working Conf. Visual Database Systems, IFIP WG2.6, pp. 119–133

[5] Swanberg, D., Shu, C.-F., Jain, R., 1993. Knowledge guided parsing in video databases. Storage and Retrieval for Image and Video Dat-abases, In: Proc. SPIE'93, San Jose, CA, pp. 13–24

[6] Jose´ M. Martı´nez, MPEG-7 Overview (version 8), ISO/IECJTC1/SC29/WG11-N4980, Klangenfurt, 2002. http://www.mpeg-industry.com/mp7a/w4980_mp7_Overview1.html

[7] Diane Hillmann, Using Dublin Core, Dublin Core Metadata Initiative (DCMI), 2005. <http://dublincore.org/documents/usageguide/>.

[8] The TV-Anytime Forum, Specification Series: S-3 On: Metadata, 2003. <http://www.tv-anytime.org/>.

[9] Chee S W, Dong K P, Soo J P. Efficient use of MPEG-7 edge histogram descriptor. ETRI Journal, 2002, 24(1): 23-30.

[10] M. Bertini, A.D. Bimbo, W. Nunziati, Video clip matching using MPEG-7 descriptors and edit distance, Lecture Notes in Computer Science 4071 (2006) 133–142.