The Algorithm of Scene Summary synthesis Based on Semantics

Yaling Zhu1,*, Xiaodi Wei2, Xiangwei Li3, Gang Zheng4 and Juan Wang5
1,2,3,4,5College of Computer and Artificial Intelligence, Lanzhou Institute of Technology, Lanzhou, Gansu, China

*Corresponding author e-mail: 463421579@qq.com

Abstract. According to the key technique of semantic-based video retrieval, the paper presents directly extracting I-frames in the compressed domain, and using the DC coefficient of I frame to establish the model of information system. Then the reduction of I Frame by the theory of attribute reduction based on RS, to produce A key frame that reflects the main content of the lens. Finally, the key frames ‘s DC coefficient has constructed the mathematical modeling, and has classified the shot according to the differences of frames. So the polymerization of I frames of similar lens, we could obtain the scenario summary.

1. Introduction

With the rapid development of electronic technology and information processing, the amount of digital video data is increasing day by day, such as TV programs, video surveillance information, video data electronic library, video on demand etc. Image, video, audio and other multimedia information gradually become the primary form of information media in the field of information processing. According to the scientific statistics, at present in the massive information gained by human beings, text information is maybe less than 20%, while multimedia information maybe occupies more than 80% of the total results [1]. Therefore, it is an urgent task to study the browsing and retrieval of video database. In the process of video query and video browsing, one of the main problems is how to exhibit the video information. If the user needs to watch all of the videos in order to find the video you want , the workload will be unimaginable. Therefore, for the video search and video browsing, especially on the network under the condition of limited bandwidth, with less amount of data to represent the video stream, make users can quickly understand the content of the video is particularly important.

2. The Standard of MPEG

The MPEG is the abbreviation of Moving Picture Experts Group. It has been established by the international organization for standardization in 1988. MPEG organization was originally authorized to develop a variety of standards for the activity image coding, followed by the expansion of its accompanying audio and its combination coding. The paper presents directly extracting I-frames in the compressed domain based on the analysis of MPEG-2 coding structure. Then on the basis of I frame, it is proposed an algorithm for synthesizing the scene.
2.1. The Standard of MPEG-2

The main goal of MPEG-2 design is as a universal audio and video coding standard, with advanced industrial standard image quality and higher transmission rate, to adapt to a wider range of applications. Such as various forms of digital storage, standard digital TV, high definition television, as well as high quality video communications, etc. The MPEG-2 standard includes four parts, that is system stream, video stream, audio stream and their detection and test.

MPEG-2 stream is divided into three layers: basic stream, packet basic stream, multiplexing transport stream and program stream. The basic stream consists of the video elementary bit stream (VES) encoded by the video compression standard and the audio elementary bit stream(AES) encoded by the audio compression standard [2,3].

2.2. The Structure of VES

A MPEG-2 video file can contain the plurality of video sequences, each video sequence is composed of the plurality of picture of group, and sequence header contains start code and sequence parameters, such as grade, level, color image format, frame field selection and so on. Each picture of group contains a number of images, and the header of the picture of group contains the start code, GOP signs, such as video tape recorder time, control code, B frame processing code, etc.. Each image is divided into a number of slices, and the header information of the image has a start code, P logo, such as time, reference frame number, image type, MV, classification, etc.. The slice is the smallest unit of synchronous, contains a number of the information macroblocks, the header of the slice has start code, address, quantization step etc.. Each macroblock consists of several blocks, the macroblock’s header includes macroblock address, macroblock type, motion vector etc.. The sequence layer, the picture of group layer, the picture layer and the slice layer all have a different starting code of four bytes, which can be used to identify the data layer.

2.3. The Frame of MPEG-2

In general, because the video is a continuous broadcast of the image sequence, in the same frame and the between two frames contain a lot of statistical redundancy and subjective redundancy. The ultimate goal of video coding is to reduce the bit rate required for storing and transmitting video information by mining statistical redundancy and subjective redundancy. In order to ensure the quality of the image without decreasing and can obtain high compression ratio, MPEG-2 defines three types of frames:

On the basis of intra coding, I-frame DCT coding is adopted by using the correlation of image itself, but other frames are not needed as reference [4,5]. I frame provides the most advanced random access function, edit simplified and the best ability to prevent transmission error expansion. P frame is the interframe coding image, uses a previous or subsequent adjacent to the image, or at the same time as the previous or subsequent adjacent to the image as a reference frame to predict motion compensation, and then encoding the image. The frame of B is encoding on the current frame and the difference between the front and the rear of the I frame or P frame and prediction of a next P frame, and removes time redundancy. B frame compression ratio can reach 200:1, the file of the compressed size is generally I-frame 15%, less than half the size of P frame compression. In the process of decompression, the decoder must access the past and future reference frame.

3. The Design and Implementation of the scene summary Synthesis

3.1. the scene summary

The scene summary of video can have a variety of media form. It can be a text, an image or image combination, can also is a video or the form of the multimedia document of a variety of media combinations[2]. At present, the main form of video summarization has five forms, that is the title, key frame, scene, video, video skimming and multimedia video summary.

Video data can be divided into four hierarchies according to the order from the coarse to fine: the video, the scene, the shot and the image frame, as shown in figure 1.
From the content point of view, video has a very strong logic structure. It is often through a number of consecutive frames to describe events, characters and actions in specific time and space environment to express a specific concept of information, but it don’t uses symbolic text but a more vivid audio-visual language [3]. Usually a video data can be divided into several scenes, each scene also contains one or a number of shots [4]. The shots is the basic unit of video data, which represents a continuous movement in time and space in a scene, and it is a video image produced by a video camera. The scene is a set of semantically related and temporally adjacent continuous shot sequences, which is the smallest semantic unit of video information.

The basic features of the scene are two points: one is the content of the scene is happened in the same place, that is the scene is composed of a series of shots at the same venue. Therefore, these lenses have similarities in picture content; the other is a scene transition makes use of fade in or fade out lens convergence in a video editing process.

So, the scene abstract is extracting meaningful parts from the original video based on automatic analysis of video structure and content, they will somehow merge into the smallest video summary of compact, fully showing the video semantic content.

3.2. The Idea of algorithm

Any video summarization algorithm follows a principle of "first separate and afterwards combine". So to carry out video content analysis and understanding, firstly the video must be divided into reasonable basic unit. These basic units include scenes, shots, frames, etc.

According to MPEG standards for video compression, video transmission uses a binary stream. In the binary video stream, we can directly extracted DCT coefficients that are obtained by pretreatment of DC coefficients. Based on the DCT transformation, the DC coefficient is the main information carrier frame image that represents the average luminance of the image for I-frame[5] in allowable range of error for shot segmentation, key frame extraction and so on. On the other hand, in the visual range, there is a part of the information is not sensitive [5], so we think that the DC coefficient of information is sufficient for shot detection, key frame extraction and the scene summary. Or, the information provided by the DC coefficient has been enough to meet the needs of our video retrieval.

We can build a two-dimensional information system that the row is the DC coefficient and the column is I-frame. In the information system, to compare the DC coefficients of the corresponding blocks of two adjacent frames can obtain the sum of the absolute values of the difference of the DC coefficients. That is, first of all I frame is segmented into shots based on video sequences, then according to the attribute reduction theory of RS theory to finish the reduction of I-frame to produce the core of information system understood as the relative I-frame without redundancy, namely key frames that reflects the main content of the shot. Because the shot is a basic unit of analysis and the composition of the scene, so to determine the degree of correlation between the shots is the key of scene detection, so the DC coefficient of key frames in each shot is composed further the information system, to classify according to the difference of the degree of the frame, so as to obtain the corresponding shot classification. According to the classification results of the shots, will be divided into a class of I-frames in the shots to obtain the scene Abstract polymerization.

![Figure 1. Video data logic hierarchy](image-url)
The mathematical model of the average value of the difference of the DCT direct current coefficient is constructed as follows:

\[ D(I_i, I_{i+1}) = \frac{1}{n} \sum_{t=1}^{n} \text{abs}(dct - dct') \]

The \( I_i \) and \( I_{i+1} \) represents the \( i \)th and \((i+1)\)th I-frame, \( dct \) and \( dct' \) is the DCT coefficient corresponding blocks of between two adjacent frames, \( n \) indicates the number of blocks in a frame. In order to describe convenience, the above formula is called the difference between the two adjacent frames.

3.3. The Realization Of Algorithm

From the basic idea of the algorithm, the scenario summary synthesis based on compressed MPEG stream includes the following steps.

Input: MPEG video stream.
Output: the number of scenario summary and synthetic scenario summary.

Step 1: The extraction of I frame from video streams;
Step 2: The extraction of DCT coefficients of I frame in the video stream;
Step 3: The extraction of DC coefficients from the DCT coefficients of I frame;
Step 4: The construction of information system model using DC coefficients, and the pretreatment DC coefficients of I-frame to store as a line of information system;
Step 5: To compute the average difference between the two adjacent frames, and then is compared with a given threshold called shotreference, which divides the video sequence into a set of shots;
Step 6: The attribute reduction of the information system using RS theory to obtain the information system's core. To compute the difference the two adjacent frames of a shot, then to compare the difference with the given threshold called keyframereference, if greater than a given threshold, then as a key frame to retain, if less than a given threshold, then give up the frame;
Step 7: To remodel the information system using key frames set obtained in the step6. To compare the average difference of two adjacent frames with a given threshold called scenereference, if greater than a given threshold, then the key frame is the scene segmentation point;
Step 8: According to the scene segmentation point obtained in the step7, to aggregate the I-frames between every two adjacent scene segmentation points can get the scene abstract.

/*IDClist is the information system model*/
LinkedList<LinkedList> IDClist = new LinkedList();
LinkedList<LinkedList> KeyFrameDClist = new LinkedList();
LinkedList<Integer> IDCrow = new LinkedList();
readIDCFromDB_CreateInfoModal(IDClist, IDCrow);
/* To compute the average difference between the two adjacent frames, and to compare with a given threshold called shotreference.*/

splitShot(IDClist, shotreference);
/* To compute the difference the two adjacent frames of a shot and to compare the difference with the given threshold called keyframereference, then to extract the key frames and to reestablish the information system model based on key frames.*/
estractKeyFrame_createKeyFrameInfoModal(IDClist, keyframereference, KeyFrameDClist);
findSceneSplitPoint(KeyFrameDClist, scenereference);
/* To aggregate the I-frames between every two adjacent scene segmentation points to get the scene abstract. */
getIfromScenePoint(IDClist, KeyFrameDClist);

4. Experimental Results

We build the video training library with kinds of sports, scenery, animation, story and news and so on. In order to test the effectiveness and feasibility of the technology, different MPEG video sequences are selected to provide the performance of the method. It can be seen that the key frame...
representation in the video sequence can get a concise representation of the content, and these frames can effectively represent the content of the original video.

The algorithm is implemented using lots of trial. First of all I frames extracted from the videos are stored in the database called DB_IDCT, and then set up the information system by using DC coefficients of I frames, the attribute reduction of rough set theory to get the shot, key frame, further DC coefficient on the key frame to establish information system, attribute reduction to obtain scene segmentation. Some experimental data are shown in Table 1.

Table 1 the experiment data result

| Video Type | Video Size | Numbers of Frames | Numbers of I-Frames | Numbers of Shots | Numbers of Scene Summary |
|------------|------------|-------------------|---------------------|------------------|-------------------------|
| Sport      | 19.2M      | 1270              | 90                  | 20               | 7                       |
| Animation  | 61.7M      | 4665              | 359                 | 62               | 22                      |
| View       | 294M       | 21488             | 1632                | 287              | 86                      |

5. Conclusion

In the compressed domain of video structure, the paper presents the scene synthesis algorithm based on semantics, and the algorithm uses the DC coefficient of I-frame to construct the information system model, then through attribute reduction, to finish the shot segmentation, key frame extraction, classification of lens, the final synthesis of scene abstract. Allows the user to view the scene summary and quickly understand the content of the videos.

6. Acknowledgement

This research was supported by Project of College level Youth Science and technology innovation project of Lanzhou Institute of Technology(18K-021), the ‘Kaiwu’ Innovation Team of Lanzhou Institute of Technology (2018KW-02), Project of the China Postdoctoral Science Foundation (2015M582722), Project of the China National Social Science Foundation (15xmx035), Project of Lanzhou Science and Technology Development Project (2015-3-30), Project of Lanzhou Science and Technology (2018-4-56), Project of Gansu Natural Science Foundation (18JR3RA228), Project of the Education Ministry's Cooperative Education between Industry and University in 2018 (201801037035), Project of the Education Ministry's Cooperative Education between Industry and University in 2016 (201602010072).

References

[1] Ke Shen and Edward J.Delp.(1996) A Fast Algorithm for Video Parsing Using MPEG Compressed Sequences[J]. Proceedings of International Conference on Image Processing(ICIP'96), Lausann.,
[2] Internation Standard.ISO/IEC13818-2 Generic coding of moving pictures and associated audio information:video second edition 2000.
[3] Di Zhong, S.F.Chang(2004). Real-time view recognition and event detection for sports video[J]. J.Vis.Commun.Image R. (15):330-347.
[4] N.Babaguchi, Y.Kawai. Y.Yasugi and T.Kitahashi(2000). Linking Live and Replay Scenes in Broadcasted Sports Video[J]. Proc. ACM International Workshop on Multimedia Information Retrieval, (11):216-227.
[5] Hari Sundaram. Segmentatio, structure detection and summarization of multimedia sequences[D]. University of Columbia, 2002.
[6] Dan Tao.The Research and Implementation of KeyFrame Extracion Methods in Content-Based Video Retrieval System[D]. University of Columbia, 2004.