2018 3rd International Seminar on Social Science and Humanistic Education (SSHE 2018)
ISBN: 978-1-60595-581-0

The Algorithm of Scenario Summary Synthesis Based on Semantics

Yaling Zhu*, Xiangwei Li & Jundi Wang
College of Software Engineering, Lanzhou Institute of Technology, Lanzhou, Gansu, China

ABSTRACT: The MPEG video data contains three types of frames, which are I frame, P frame and B frame. However, the I frame records the main information of video data, and the P frame and B frame are just as motion compensation of the I frame. The paper presents the method that analyzes the MPEG video stream in the compressed domain, and then 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 reduce I Frame by the theory of attribute reduction based on RS, and produce a key frame that reflects the main content of the shots. Finally, the key frame's DC coefficient has constructed the mathematical modeling, and has classified the shot according to the differences of frames. So by the polymerization of I frames of similar shots, we could obtain the scenario summary. Experiment indicated that this method can be realized in the compressed MPEG video automatically, and can lay the foundation for the video processing in the future.

Keywords: MPEG; key frame; DCT; scene summary

1 INTRODUCTION

With the development of electronic and information processing, the digital video data sharply increases every day, such as TV programs, video surveillance information, electronic library of video data, and video on demand, etc. Multimedia information of image, video, and audio has gradually become the main form of information media in the field of information processing. According to the scientific statistics, at present in the massive information obtained from human beings, text information is maybe less than 20 percent, but multimedia information maybe occupies more than 80 percent of the total outcomes [1]. However, in the face of a huge ocean of multimedia information, it is not easy to accurately find the required information, which often needs to spend a lot of time and energy. This is the so-called "data explosion but lack of knowledge" [1], namely the problems faced by people is no longer the lack of multimedia content, but how to effectively search information of interest from these massive data according to the characteristics of multimedia data. In the face of such a vast information, how to effectively organize and manage data, how to efficiently get the needed information, these problems have brought new challenges to the research field of traditional databases, information retrieval, computer vision, artificial intelligence, etc.

So the research of browsing and searching the video database is pressing. In the process of video query and browsing, one of the main problems is how to exhibit the video information. If the users need to watch all videos in order to find what they want, the workload will be unimaginable. Therefore, for the video search and browsing, especially on the network under the condition of limited bandwidth, with fewer amounts of data representing the video stream, making users can quickly understand the content of the video is particularly important.

2 VIDEO DATA MINING

2.1 *Overview of video data mining technology*

Video data is a kind of special data, which exists in the form of data stream. It is vivid, and can carry a lot of information, which is easy to be perceived. Therefore, there has been a great deal of interest in video

*Corresponding author: 463421579@qq.com

data. In recent years, with the popularity of video capture devices and the decline in mass storage equipment prices, available video data is increasing rapidly. There is a large number of video surveillance equipment, such as transportation and bank video surveillance systems which generate vast video data every day; we can get mass entertainment and sports videos from the Internet; digital TV and network TV make it easier for the TV programs to be turned into videos, which can be played on the computer and applications.

Faced with a large number of video data, we hope to establish an effective mechanism to manage and access them, so that users can quickly retrieve their desired video information. We also expect that the computer will be able to handle the tedious surveillance videos automatically, and find the exception, thus reducing a lot of human labor. Even the computer is best able to find some of the hidden video information, which is difficult to find and interrelate. Based on the above requirements, data mining technology is introduced into the field of video data processing.

Through the comprehensive analysis of characteristics of audio-visual video data, time structure, event relation and semantic information, we can find a hidden valuable and understandable video mode, where the video shows the trend and related events, and improve the degree of intelligence video information management. However, due to the non-structural nature of video data, the traditional data mining technology based on relational database and transaction database can hardly be directly applied to the field of video data mining. The difficulty of video data mining is mainly reflected in the gap between video data and semantics that human can understand. Traditional data mining objects are generally numerical data and text, and for video data, computer understanding is obviously much more difficult. Generally speaking, the time cost of video data processing is much larger than numerical and text data. In order to solve these problems, a variety of methods and techniques for video data mining are proposed. Computer vision, digital image processing technology and traditional data mining technology are combined into the field of video mining.

2.2 *Classification of video data mining technology*

Video data mining research begins in the beginning of this century, so the time is relatively short. As one of the research directions in data mining, the technology is not mature, and there is no classical and accepted classification theory. At present, video mining technology can be divided into the following five categories.

2.2.1 *The classification based on the field*

We can classify video data mining according to the relevance of the target video, so it can be divided into traffic video mining, medical video mining, entertainment video mining, etc. On the first face, this classification is of no practical significance, but the facts suggest otherwise. A field usually has its own characteristics, which may determine the purpose and means of video mining. For example, traffic video is surveillance video, and its screen background is usually unchanged, which helps to simplify the mining process [2,3], sports and entertainment video mining may be concerned with the scene (scene semantic), such as semantic shooting foul, and the traffic video may be concerned with the object or mobile objects, such as speeding vehicles, etc. Therefore, the research method of video data based on domain can not only greatly reduce the difficulty of research, but also has a larger application background in real life.

2.2.2 *The classification based on mining objects*

The object of traditional data mining is numerical data or text data. However, the video data mining, which does not directly bear on many large video files, usually cuts and clips the original video file as the direct object. Thus, we can divide the video data mining into two categories. The first type is the basic unit of data mining based on [4,5] shot or scene. A shot is a set of closely related frames (frame) in a video sequence. The scene is a set of semantically connected and contiguous shots. A single frame usually does not show obvious semantics, and the shot or scene can carry the most basic semantic information; in some video format, some adjacent frame data are compressed together to separate, such as MPEG4 and RM, so we usually don't use frames as the mining unit. Usually the length of a shot is a few seconds [6]. The second type is regarding the object as the basic unit of data mining [7,8]. An object is a meaningful object in a video screen. It may be relatively fixed and does not change over time in the relative position of the screen, such as subtitles; it may also be moving, and we need to track it in multi-frame images, such as vehicles. The object is the object of data mining in order to highlight important information and discard irrelevant or secondary information. It is not difficult to see that for the first case, and we need to segment the original video in the process of data preprocessing. For the second category, we need to do video space segmentation.

2.2.3 *The classification based on mining purposes*

Literature [9,10] gave a classification of video data mining. It summed up three types: special pattern detection, video clustering and classification, video association mining. Among them, special pattern detection is a kind of special model (usually some events) in the video. Clustering and classification are grouped according to the subject of the video (the subject of the classification is determined in advance, and the theme of the cluster is not). Video association mining is the use of association mining technology to find hidden information in video. This classification

method can cover the vast majority of video mining work.

2.2.4 *The classification based on mining technology*

According to the video mining technology, we can roughly divide video mining into three categories. The first is the traditional video data mining technology, such as regarding text information or other annotation information as mining object, and using numerical traditional database mining technology or numerical simulation database mining. The second category is the use of digital image processing, computer vision and other multimedia related technologies based on content video data mining. This method reflects the essential characteristics of the video data, and overcomes the subjectivity and randomness of the pervious artificial annotation. However, because of the non-uniqueness of the underlying features of the video, it is difficult to determine the query video examples, so the video data mining technology based on semantic appears. On account of the special structure of video data and imperfect mathematical theories, the semantic-based video mining technology is very difficult, so the technology will become a hot research topic in video data mining based on semantic.

2.2.5 *The classification based on information sources*

In video data mining, the main source of information is video, but some other ancillary data will also be used. According to the information sources used in excavation, we can divide the existing work into the following three categories: (1) Use video data only, which is the most important information in the video stream. Video stream is the reaction of the most essential characteristics, and because the video can be regarded as a set of static frame sequence connected with each other to launch, the video information is the ultimate response to static image feature information. (2) Use video and audio data. The audio information is also very important information in the current video stream, which can supplement and explain video information to a certain degree, so it is a very meaningful work. (3) Use text data. In addition to video and audio information, there still exists very important information in video files, which is text information, such as the text in movies, animation, and television, which is the most relevant to video content, Therefore, mining the text information based on video stream is also one of the important contents of video mining.

2.3 *Video coding technology*

The continuous analog video signal is not processed by computer, but also cannot be transmitted in various digital systems and storage. The continuous analog video signals must be converted into discrete digital signal, which is called digital video. In order to facilitate the transmission and storage, the digital signal is coded to reduce the redundancy in time and space.

2.3.1 *Digital video*

Video digitization includes two aspects: sampling and quantization. The discretization of the video signal in space or time is called sampling, which uses the limited sampling point to replace the infinite coordinate value. Obviously, the smaller the sampling interval, the more sampling points, the greater the amount of data. Conversely, the smaller the number of sampling points, the smaller the amount of data.

When the analog video signal is sampled in space and time, it is dispersed into a series of display frames, and the corresponding value of each pixel represents the light energy at the sampling point. Obviously, it is a continuous variable with infinite value. It is necessary to convert the continuous quantity into discrete values and give different code words to truly become digital video, and then processed by computer or other digital devices. This transformation is called quantization, which can be used both in the space domain and time domain, and in the transform domain (e.g. frequency).

2.3.2 *Video coding*

Whatever format of the digital video information is sampled in any way, the amount of data is very alarming, so no data compression for transmission, especially the real-time transmission of video information is almost impossible. Therefore, the video encoding in digital video processing and transmission, occupies a very important position in storage, and the most important is the application of digital video. The representation of image and video signal requires a large amount of data, but these data are often highly correlated, which will cause information redundancy. Therefore, the redundant information can be removed by video data compression. A major goal of still image compression is to guarantee the reconstruction quality, and try to remove the spatial redundant data image by itself. The video signal is compressed, which removes redundant data in space at the same time, as well as temporal redundancy and other redundancy, so as to achieve high compression ratio and low bit rate. The purpose of video coding is to reduce the amount of data in the video sequence, so that it is easier to transmit real-time video information on a given channel or to store more video sequences in a given capacity memory, such as using telephone line to transmit video.

3 THE MPEG STANDARD

The MPEG (Moving Picture Experts Group) is established by the international organization in 1988, which is committed to draw up the international standard of the moving image compression coding. MPEG organ-

ization was originally authorized to develop a variety of standards for the activity image coding, followed by the expansion of its accompanying audio and video compression coding. Later, for different application needs, the lifting of the restrictions for digital storage media has now become the development of active image and audio coding standards of the organization. So far, in the field of video compression, MPEG is the most popular, which includes MPEG-1, MPEG-2, MPEG-4, MPEG-7, MPEG-21 and other several series. Here is the main analysis of MPEG-2 video compression coding standard [11,12].

The paper directly presents the extracting I-frames in the compressed domain based on the analysis of MPEG-2 coding structure. Then on the basis of I frame, it proposes an algorithm for synthesizing the scene.

3.1 *The standard of MPEG-2*

The MPEG-2 introduced after MPEG-1 stands for the standard of the generic coding for moving pictures and audio information, which aims at being designed 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, and high quality video communications. The MPEG-2 standard includes system, video, audio, the detection and testing of stream.

The MPEG-2 stream is divided into three layers which are the basic stream, the packet elementary stream and the multiplexed transport streams, and the program stream. The basic stream is composed of video elementary stream (VES, Video Elementary Bit Stream) encoded by video compression standard, and audio elementary stream (AES, Audio Elementary Bit Stream) encoded by audio compression standard [12, 13].

3.2 *VES structure*

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, etc. Each picture of group contains a number of images, of which the header contains the start code, and 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 its header information has a start code, and P logo, such as time, reference frame number, image type, MV, classification, etc. The slice is the smallest unit of synchronous, which contains a number of the information macro blocks, and its header has start code, address, quantization step, etc. Each macro block consists of several blocks, whose header includes macro block address, macro block type, motion vector, etc. The sequence layer, the picture of group layer, the picture layer and the slice layer all have different starting codes of four bytes, which can be used to identify the data layer [2, 13]. MPEG-2 video structure is shown in Figure 1.

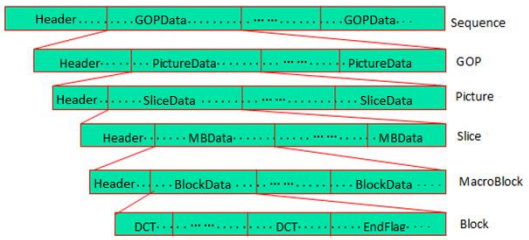

Figure 1. MPEG-2 video structure.

3.3 *The frame of MPEG-2*

In general, because the video is a continuous broadcast of the image sequence, in the same frame and between two frames it contains 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 obtain high compression ratio, MPEG-2 defines three types of frames: I frame, P frame, and B frame, using different compression encoding, as shown in Figure 2.

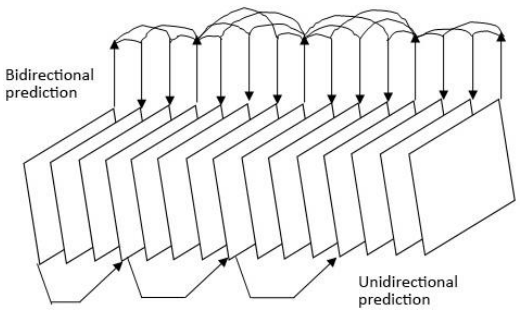

Figure 2. The different compression encoding of three types of frames.

I frame is adopted DCT to encode by using the correlation of image itself based on intra frame coding, but not need other frames as a reference[14,15], that is, there is no need to refer to images that have been encoded in the past, nor do they need to refer to subsequent images that have not been coded. So, the intra frame coding image does not consider the correlation of time. I frame recording the main information is the foundation frame in the MPEG international standard, and refers to the other types of image. I frame can only remove the spatial redundancy at the compression ratio of 6:l without any perceptible fuzzy phe-

nomenon, which is the reference point to generate the P frame and B frame for subsequent motion estimation [16]. It provides the most advanced random access function, the simplicity of editing, and the best ability to prevent transmission error expansion.

P frame is the inter frame coding image. When P frames are encoded, we need to use previously or later neighboring images, or at the same time, we need to predict motion compensation and encode images by the previous and subsequent adjacent image. That is, it takes the motion characteristics into account, provides inter frame coding, and removes time redundancy. Its encoding takes image macro block as the basic encoding unit, and each color input frame of the video sequence is divided into multiple non overlapping blocks [17,18]. The P frame uses two types of parameters to represent the difference between the macro block of the image and the reference image, and the motion vector of the macro block.

The B frame is encoding on the current frame and the difference between the front and the rear of I frame or P frame and next P frame by removing time redundancy. The compression ratio of B frame can reach 200:1 and the file of the compressed size is generally 15% of I-frame, which is less than half the size of P frame compression. In the process of decompression, the decoder must access the past and future reference frame. Therefore, the coding frame should be rearranged before transmission, and then the decoder can reconstruct and display the frames in the appropriate sequence. After using bidirectional prediction, the content that cannot be predicted in the previous frame is well predicted in the latter frame, and it is very effective to reduce the influence of noise on the prediction of the average [16,17].

4 THE DESIGN AND IMPLEMENTATION OF THE SCENE SUMMARY SYNTHESIS

4.1 *The scene summary*

For extracting meaningful parts from the original video, it is very important to analyze the structure and content of the video by automatic or semi-automatic way. Then combine meaningful parts extracted in some way to form a compact. The compact full representation of video semantic content is called the scene summary of video stream. In the process of video query and video browsing, one of the main problems is how to exhibit the video information. If users need to watch all videos for a period of time to find what they want, the workload will be difficult to imagine. Therefore, for the video search and browsing, especially on the network under the condition of limited bandwidth, with fewer amounts of data to represent the video stream, a user can quickly understand the content of the video is particularly important.

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

Video data can be divided into four hierarchies from coarse to fine: the video, the scene, the shot, and the image frame, as shown in Figure 3.

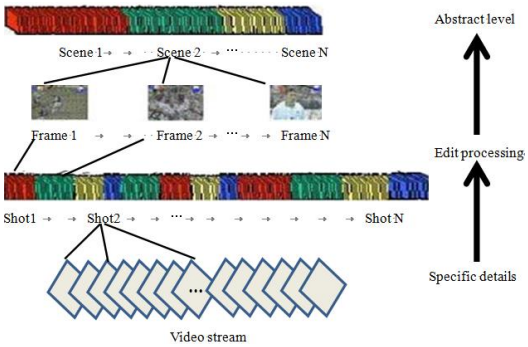

Figure 3. Video data logic hierarchy.

From the content point of view, video has 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 used a more vivid audio-visual language instead of symbolic text [13]. Usually a video data can be divided into several scenes; each scene also contains one or a number of shots [14]. The shots are 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.

There are two basic features of the scene: one is that the content of the scene happens in the same place, and the scene is composed of a series of shots at the same venue, and therefore these shots have similarities in picture content; the other is that is a scene transition makes use of fade in or fade out of lens in video editing process.

So, the scene abstract extracts the meaningful parts from the original video based on automatic analysis of video structure and content, which will be somehow merged into the smallest video summary of compact, fully showing the semantic content of video.

4.2 *The idea of algorithm*

Any video abstract algorithm follows the principle of "separate first and then combine". So, for the analysis and understanding of video content, the video must be

firstly divided into reasonable basic units, which include scenes, shots, frames, etc.

According to MPEG standards for video compression, video transmission uses a binary stream; in the binary video stream can be directly extracted DCT coefficients, the DCT coefficients 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 [15] 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 [15], 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 with the row being the DC coefficient and the column being I-frame. In the system, comparing 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, 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 the basic unit of analysis and the composition of the scene, it is the key of scene detection to determine the degree of correlation between the shots, and the DC coefficient of key frames in each shot composed the information system further, 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, to polymerize I-frames divided into a class of lenses will obtain the scene abstract.

The mathematical model of the average value of the difference of the DCT direct current coefficient is constructed as follows:

$$D(I_i, I_{i+1}) = (\sum_{1}^{n} abs(dct - dct')) / n$$

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

4.3 *The realization of algorithm*

From the main idea of algorithm, the scene summary synthesis from compressed MPEG stream consists of the following steps.

Input: A MPEG video stream.
Output: Synthetic scenario summary and the number of 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: Compute the average difference between the two adjacent frames, and then compare 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 uses RS theory to obtain the information system's core. Compute the difference between two adjacent frames of a shot, then compare the difference with the given threshold called keyframereference, if greater than a given threshold, it retains as a key frame; if less than a given threshold, give up the frame;
Step 7: Remodel the information system by using the key frame sets obtained in Step 6. Compare the average difference of two adjacent frames with a given threshold called scenereference, if it is greater than a given threshold, the key frame is the scene segmentation point;
Step 8: According to the scene segmentation point obtained in the Step 7, aggregate the I-frames between every two adjacent scene segmentation points, and the scene abstract can be obtained.

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

5 EXPERIMENTAL RESULTS

We build the video training library with kinds of sports, scenery, animation, story and news, etc. In order to verify the validity and feasibility of the technology, the various MPEG video sequences are selected to examine the performance of the proposed method. We can obviously see that the condensed and succinct representations of the content are obtained by key frame representation from video sequences, and these frames can effective represent the contents of original video.

The algorithm is implemented in a console with VC++6.0 and SQLServer2000 after many trials [17]. 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, using 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 of the experimental interfaces are shown in Figure 4(a), (b), (c), and the experimental data results are shown in Table 1.

(a)

(b)

(c)

Figure 4. The experimental interface.

Table 1. The experiment data result.

Video type	Video size	The number				
		frame	I frame	shot	key frame	scene summary
Sport	19.2M	1270	90	20	20	7
Animation	61.7M	4665	359	62	65	22
View	294M	21488	1632	287	523	86

6 CONCLUSION

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

ACKNOWLEDGMENTS

This paper is sponsored by Colleges and Universities Scientific Research Project of Gansu Province “Research on Key Technologies of medical image fusion based on transform domain” (2016A-097); National Natural Science Foundation of China “compressed domain massive video concentrated key technology research” (61462057).

REFERENCES

[1] Liu J, Bhanu B. Learning semantic visual concepts from video. 2002, In: *IEEE Proceedings of 16th International Conference on Pattern Recognition*. Quebec, Canada, 20(2): 1051-1064.

[2] Ke Shen and Edward J. Delp. 1996. A fast algorithm for video parsing using MPEG compressed sequences. In: *Proceedings of International Conference on Image Processing (ICIP'96)*, Lausanne.

[3] Internation Standard. ISO/IEC13818-2 Generic coding of moving pictures and associated audio information: video second edition 2000.

[4] Di Zhong, S.F. Chang. 2004. Real-time view recognition and event detection for sports video. *Journal of Visual Communication and* Image *Representation*, 15: 330-347.

[5] N. Babaguchi, Y. Kawai. Y. Yasugi, T. Kitahashi. 2000. Linking live and replay scenes in broadcasted sports video. In: *Proc. ACM International Workshop on Multimedia Information Retrieval*, 11: 216-227.

[6] Hari Sundaram. 2002. *Segmentation, Structure Detection and Summarization of Multimedia Sequences*. University of Columbia.

[7] Dan Tao. 2004. *The Research and Implementation of Key Frame Extraction Methods in Content-based Video Retrieval System*. University of Columbia.

[8] Yizhen Zhang, Tao Liu. 2002. *Visual C++ Implementation of MPEG/JPEG Encoding and Decoding Technology*. Beijing: Posts and Telecommunications Press.

[9] John S. Boreczky, Lawrence A. Rowe. 1996. Comparison of Video Databases IV. In: Proc., SPIE 2670(1996), pp: 170-179.
[10] Arun Hampapur. 1995. *Design Video Data Management Systems*. The University of Michigan.
[11] M.M. Yeung, B. Yeo, and B. Liu. 1996. Extracting story units from long programs for video browsing and navigation. In: *Proc. IEEE Multimedia Computing and Systems*.
[12] A Hanjali, H J Zhan. 1999. An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. In: *IEEE Transactions on Circuits and Systems for Video Technology*.
[13] Vasanth Tovinkere Richard J. Qian. 2001. Detecting semantic events in soccer games towards a complete solution. In: *IEEE International Conference on Multimedia and Expo*, pp: 1040-1043.
[14] Xiaoguang Li, Lansun Shen. 2000. Video content analysis and abstract extraction technology in compressed domain. *Measurement and Control Technology*, 25(5): 17-19.
[15] Yong Fang, Feihu Qi. 2000. A new method of video shot boundary detection and key frame extraction. *Journal of South China University of Technology*, 32(11): 18-21.
[16] Anne Brink, Sherry Marcus, VS Subrahmanian. 1995. Heterogeneous multimedia reasoning. *IEEE Computer*, 28(9): 33-39.
[17] Boon-Lock Yeo, Bede Liu.1995. Rapid scene analysis on compressed video. *IEEE Transactions on Circuits & Systems for Video Technology*, 5(6): 533-544.
[18] R. Leonardi, P. Migllorati. 2002. Semantic indexing of multimedia documents. *IEEE Multimedia*, 9: 44-51.