# 비쥬얼 리듬을 이용한 효율적 비데오 인덱싱 기법

論 文
10-3-4

## An Efficient Video Indexing Scheme Exploiting Visual Rhythm

정 지 문, 김 정 길*
Ji-Moon Chung and Cheong-Ghil Kim

## Abstract

With the growing popularity of digital video applications, those areas of the efficient transmit, storage management, and search technology for video data are emerging as an important core technology. To be an effective video indexing system, users need to be able to find the video segments that they want. Unfortunately, video data is difficult to manage because of its unstructured data type and large volume with linear forms. This paper proposes a shot verification using visual rhythm and video retrieval system. The shot verification is designed to detect a segment from video easily and quickly, known as shot boundaries, just by changing the visual rhythm without playing the image. Therefore, this can decrease the false detected shots and generate the unidentified shots and key frames. The retrieval system is constructed in terms of visual descriptor from the list of MPEG-7. The effectiveness of the proposed shot verification process and video retrieval system is demonstrated.

Keywords : video retrieval, shot verification, visual rhythm. MPEG-7

## I. Introduction

As the growing popularity of digital video applications, those areas of the efficient transmit, storage management, and search technology for video data are emerging as an important core technology. Especially, the interest for video retrieval technology is increasing rapidly because we hope that video will be as easy to search as text. Unfortunately, it is still difficult to find relevant video contents as text because video is a flat sequence of frames with no other indications.

For advanced video applications, an interpretation level above the raw video information is needed, for instance to parallel those found in text applications, such as sentences, paragraphs, chapter titles and table of contents. Video indexing is a mechanism that provides such syntax and structure supplying users with entry points in video. The user is then able to quickly access a particular part of the video. In addition, the indexing mechanism can be exploited to provide a summary of the entire video so that the user may become quickly acquainted with the video content [1].

Video is composed of multiple media such as images, text, and audio as the useful tool which can include a wide variety of real world events. Therefore, video content can be approached at different levels: raw data, low-level visual content, and semantic content. The raw video data consists of elementary video units together with some general video attributes such as format, frame rate etc. Low level visual content is characterized by visual fea-

접수일자 : 2011년 07월 20일
심사일자 : 2011년 08월 13일
최종완료 : 2011년 09월 12일
*교신저자, E-mail : cgkim@nsu.ac.kr

tures such as color, shapes, textures etc. Semantic content contains high-level concepts such as objects and events. The semantic content can be presented through many different visual presentations using different sets of raw data. It is obvious that requirements for the extraction of these contents are different [2].

There have been many researches with regard to video retrieval systems. A simple method is query by example approaches, which can be useful when a user has a same or similar image at hand. However, this method could not perform well with images from different angle or with different scale. One type is the retrieval mechanisms for video data through various multimedia surrogates including titles, storyboards, and skims [3]. Extensive research efforts have been made with regard to the retrieval of video and image data based on their visual content such as color distribution, texture and shape [4]. Similarity measurement is utilized in those works; for example, VisualSEEk [5], Photobook [6], and Blobworld [7]. Unfortunately, those approaches are not satisfactory because of its unstructured data type and large volume with linear forms.

In general, the first step in video retrieval is segmentation which detects shot and scene changes in the video. The video content indexing becomes possible only after accurate shots have been detected to represent the entire video with certain syntax for efficient user access [2]. After that key-frame extraction is required. Therefore, the accuracy of the shot change detection algorithm will have a great influence on the accuracy and effectiveness of the video retrieval systems.

For shot boundary detection, we use the visual rhythm. The visual rhythm is a single image, a sub-sampled version of a full video in which the sampling is performed in a pre-determined and systematic way. It is basically a representation of the video, which includes the overall content of the video. But most importantly, a visual rhythm includes visual features that allow the operator to distinguish and classify many different types of video effects [2].

The following are the summaries of this paper. First, in order to establish video retrieval system, the shot boundary of verification of movie images should be undertaken in advance. The problem in classifying the shot boundary of movie image data is solved by using visual rhythm. This process is unable to solve all the problems of unidentified and false detection, which leads to the exertion of much time and effort. It is difficult to expect perfect results because of editing effects such as cutting, wiping, and dissolving used in the production of an image. Therefore, it is definitely essential to work on the verification and the modification manually in order to achieve the exact shot boundary. The system suggested in this paper is designed to detect the parts easily and quickly, which are assumed as shot boundaries, just by changing the visual rhythm without playing the image. Therefore, this enables to delete the false detected shot and to generate the unidentified shot and key frame.

Second, the retrieval system is constructed in terms of visual descriptor from the list of MPEG-7. The characteristics of the key frame for every shot constructing the video image can be extracted and used upon being retrieved by storing the Meta data in the form of XML according to the standard of MPEG-7. The Meta data constructed by the standard of MPEG-7 can easily be extended when needed. The advantage to be shared at the level of similar application is on the expansion of Meta data instead of on the cost of constructing the separate database.

Third, during the retrieving process, a thumbnail and keyword method of inquiry is possible and the user is able to put some more priorities on one part than the other between the color and shape. As a result, the corresponding shot or scene is displayed. However, in the case of not finding the preferred shot, the key picture frame of similar shot is supplied and can be used in the further inquiry of the next scene. The shot verification and video retrieval system suggested in this paper could be applied to the index work of video image for video retrieval system and broadcasting TV.

The paper is organized as follows. Section 2 reviews the current state-of-the-art shot detection

methods, and leads to the importance of shot verification. Section 3 formally introduces the concept of visual rhythm and shows its features and characterizations for different video effects. Our design of a shot verifier is presented in Section 4 with some of its major functions and user interfaces. Lastly, Section 5 concludes the paper with certain merits and limitations of the proposed shot verifier.

## II. Background

### 1. Visual Rhythm

Prior to the detailed introduction on visual rhythm, the following definitions of terms are necessary. A shot is an unbroken sequence of frames, which present a continuous action; it usually results from a single operation of the camera. In other words, it is a sequence of frames that is generated by the camera from the time it begins recording until the time it stops. A scene is defined as a collection of one or more adjoining shots that focus on an object or objects of interest. In general, shots have duration of 1 - 10 seconds, but the duration of scenes varies greatly depending on the genre of videos [2].

Visual rhythm was defined to be the vertical or horizontal sides of the video icon in the context of video icons [8]. Fig. 1 shows several different sampling strategies in the same image: diagonal, cross, and local sampling.

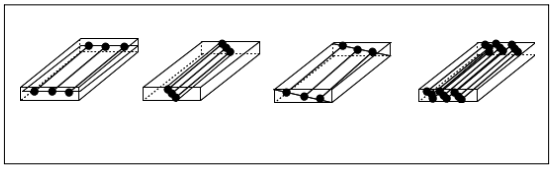

**Fig. 1 Examples of visual rhythm**

Let FV (x, y, t) be the pixel value at location (x, y) and time t of an arbitrary video V. Then, the video V may be represented as:

$$V = F_v(x,y,t), x,y,t \in \{0,1,2 \ldots\} \quad (1)$$

Let $F_{Thumbnail}$ (x, y, t) be the representation of a reduced frame of a spatially reduced video VThumbnail of the original video V. Each reduced frame, or thumbnail, is a (horizontally and vertically) reduced image of its corresponding frame in the video V by a factor. Thus, the spatially reduced video or sequence of thumbnails may be expressed as:

$$V_{thumbnail} = \{F_{thumcnail}(x,y,t)\}, x,y,t \in \{0,1,2 \ldots\} \quad (2)$$

The relationship between the video V and its spatially reduced video VThumbnail can be represented using their pixel correspondences as follows

$$F_{thumcnail}(x,y,t) = F_v(rx+k_x, ry+k_v, x,y,t \in \{0,1,2 \ldots\}, k_x k_y \in \{0,1,2 \ldots, r-1\}) \quad (3)$$

where kx and ky are offsets in pixel units. Using the spatially reduced video, we define the visual rhythm, VS, of the video V as follows:

$$VS = \{F_{vs}(z,t)\} = \{F_{thumbnail}(x(z), y(z), t)\} \quad (4)$$

where x(z) and y(z) are one-dimensional functions of the independent variable z. Thus, the visual rhythm is a two-dimensional image consisting of pixels sampled from a three-dimensional data (video). That is, the visual rhythm is constructed by sampling a certain group of pixels in each thumbnail and by temporally accumulating the samples along time. Thus, the visual rhythm is a two-dimensional abstraction of the entire three-dimensional video content.

In order to make visual rhythm, reduced image must be generated in advance. However, the digital video is usually compressed by using the discrete cosine transform (DCT) for intra-frame encoding, fast generation of the visual rhythm becomes possible. Here, the frames partitioned into 8×8 blocks are processed. Each block is transformed to the frequency domain representation resulting in one DC and 63 AC coefficients. Thus, the extraction of the DC coefficients can be performed fast, without fully decoding the compressed video. Thus, the construction of the visual rhythm is possible without the inverse DCT. We simply extract the DC coefficients by sampling diagonally the DC sequence. Fig. 2 shows the examples of visual rhythm, in which cut, wipe, and dissolve are characterized by the gradual changes of vertical, diagonal, and circle line and colors.

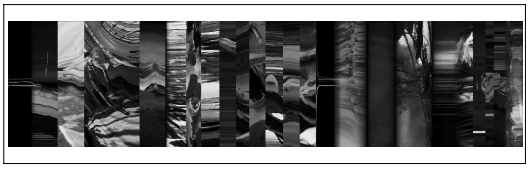

Fig. 2 Visual rhythm from Harry Potter

## 2. MPEG-7

MPEG-7 is an ISO/IEC standard developed by MPEG (Moving Picture Experts Group). The MPEG-7 standard, formally named "Multimedia Content Description Interface", provides a rich set of standardized tools to describe multimedia content. Both human users and automatic systems that process audiovisual information are within the scope of MPEG-7 [9].

MPEG-7 offers a comprehensive set of audiovisual Description Tools (the metadata elements and their structure and relationships, that are defined by the standard in the form of Descriptors and Description Schemes) to create descriptions (i.e., a set of instantiated Description Schemes and their corresponding Descriptors at the users will), which will form the basis for applications enabling the needed effective and efficient access (search, filtering and browsing) to multimedia content. This is a challenging task given the broad spectrum of requirements and targeted multimedia applications, and the broad number of audiovisual features of importance in such context. Fig. 3 shows the scope of standardization and system architecture.

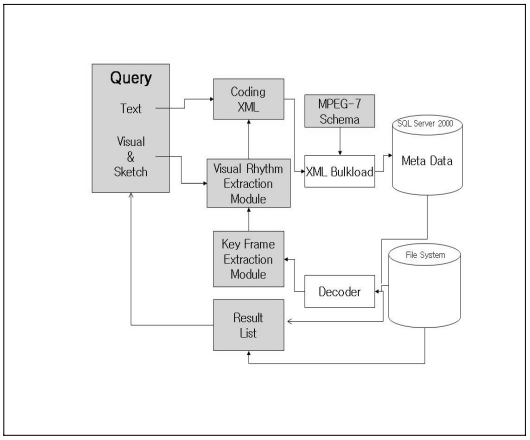


Fig. 3 MPEG-7 application system scope

# III. Proposed Video Indexing System

## 1. System Architecture

The overall structure of the proposed video indexing system using visual rhythm is depicted in Fig. 4. In this figure, the gray color areas are proposed modules. MS SQL Server 2000 is used for the Meta data storage. The proposed system is the modified version of [10]. We replace the time slice scheme in the original system with visual rhythm in the propose system. The major modules are key frame extraction module, visual rhythm extraction module, XML 9 [11,12] coding module, query module, and result list management module.

Key frame extraction module handles with moving picture data; input data is from decoder. That is, it does not process video data directly, which provides the compatibility and effectiveness. This module extracts visual rhythm; and then executes pre-processing such as filtering and thinning. After that the result is transferred to visual rhythm extraction module. Visual rhythm extraction module extracts dominant color information, color histogram, and shape information, and then transfers them to XML coding module. For edge detection, color information is converted to brightness information; after that edge information is extracted using Canny filter. The result goes through dilation, erosion, and thinning procedure. XML coding module converts the data from visual rhythm extraction module to MPEG-7 standard description, XML document. After that MPEG-7 schema verify its effectiveness and they are stored in DB. Query module captures thumbnail or displaying images and combines the contour shape information from visual rhythm extraction module with input keyword; and then converts them to XML. This module has the functionality of assigning weight. Finally, query result is managed by result list management module.

## 2. Shot detection and visual thythm

Fist shot edges are detected through several image processing procedures such as filtering and thinning.

Visual rhythm locates the diagonal information in a frame into adjacent lines vertical lines. The proposed method compares the pixel values of each vertical line and calculates the difference. If it exceeds the critical point of overall average, standard deviation to visual rhythm, and local visual rhythm deviation, the proposed system regards as a cut. That is, a cut can be defined as the place where brightness is distributed discontinuously along with vertical lines in visual rhythm I(x,t). In order to fine a cut, execute differentiate according to time as following equation:

$$d(z,t) = I(z,t) - I(z,t-1) \quad (5)$$

d(z,t) is the differentiated image of visual rhythm regarding time and direction; the discontinuity of brightness can be found at peak pixel values. To find the peak point effectively, d(z,t) is converted to the one-dimensional function as follow:

$$d_{\mu} = \frac{1}{n}\sum_{8=0}^{n-1} d(z,t) \quad (6)$$

where n is the total number of pixels of z-axis. After that find the average and distribution around t as followings:

$$\mu(t) = \frac{1}{N(B)}\sum_{k=B} d(t+k)$$

$$\sigma^2(t) = \frac{1}{N(B)-1}\sum_{k=B}\left(d_{\mu}(t,k) - \mu(t)\right)^2 \quad (7)$$

Where B signifies how may pixels will be included around in an area.

Suppose B={-16,-15,....-1,0,1,.....15,16}, N(B) becomes 32 (N(B)=32). Consequently, the extraction result can be as bellowing:

$$bu(t) = 0$$
$$= 1 \quad (8)$$

Here, if bu(t) is 1, a peak value exists at t. If is zero, any peak value does not exit at t. Normally, m is between 3 and 8.

## 3. System interface

Fig. 4 shows the user interface of the proposed system. Cuts are displayed as diamond at the top together with histograms with regard to visual rhythm. Once video data is read, the system automatically displays the basic information of video to be extracted such as file path, file name, file format, number of frames, resolution, and so on. After that we can input additional information on that video. Fig. 5 shows the extracted visual rhythm and additional information. Here, the result comes from the 1,000 frames and key frames. Fig. 6 shows the visual information of the sample video, Spider man,

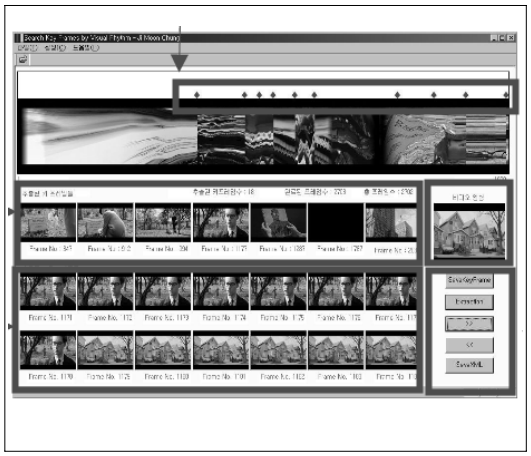

**Fig. 4 System interface**

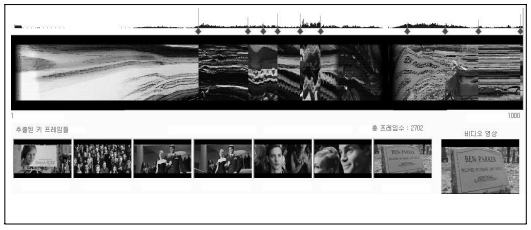

**Fig. 5 Extracted visual rhythm band and additional information**

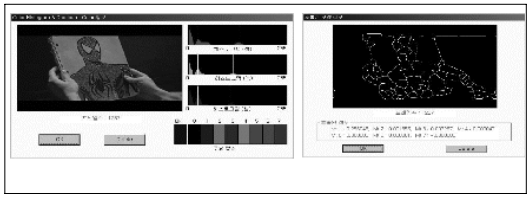

**Fig. 6 Frame information from Spider man**

## 4. Shot Verification and Visual Rhythm

The shot verification process incorporates the visual rhythm for either verifying detected shots quickly or finding new ones. It is expected to run after the automatic shot detector returns with a list of detected shots. However, it may also be used by itself to

find shots manually. Basically, this process includes frame search and addition or deletion of shot detection. Fig. 7 shows the deletion process of shot detection from user feedback. Finally, frame information is stored in DB with XML. It stores frame number, frame location in video, file path, output path, and shape and color information.

**Fig. 7 Shot detection process from user feedback**

## IV. Simulation

For simulation, we developed the modules for generating visual rhythm, extracting shots, and processing query and verification. The results are compared with other system. Simulation was executed with Pentium-IV 2.4GHz on Window XP Sever. Visual C++ 6.0 was used for programming. DBMS was ms SQL Server 2005. We used 150 sample moving pictures and DirectX 9.0 SDK for the result display. As for sample videos, they were selected as the following scenarios: rapid scene change, gradual scene change, complex lightening, rapid brightness change, color change of dominant light source, series of scenes with similar color distribution, series of scenes with similar color distribution under dart light, and rapid camera changes of zoom, pan/tilt, and zoom with pan/tilt. Furthermore, we simulated the situation of intentional editing which can destroy the regularity and connection between frames. This could be obstacles for correct cut detections. Query process takes two methods either thumbnail or keyword scheme.

Fig. 8 shows the result of executing shot detections based on visual rhythm. Here, some shots have not been detected. Fig. 9 displays insert some undetected shots artificially. Therefore, the proposed system can improve indexing performance because it enables to insert undetected shots prior to indexing stage. Fig. 10 shows the query result based on color information. If necessary, any keyword can be combined for better retrieval.

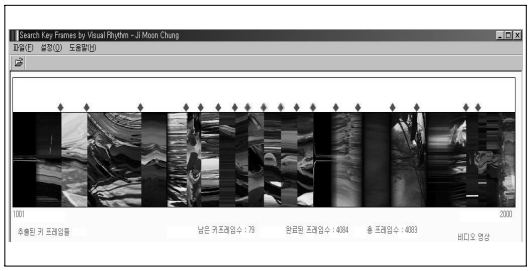

**Fig. 8. Visual rhythm and shot detection**

**Fig. 9. Visual rhythm and shot detection with the artificial insertion of shot detections**

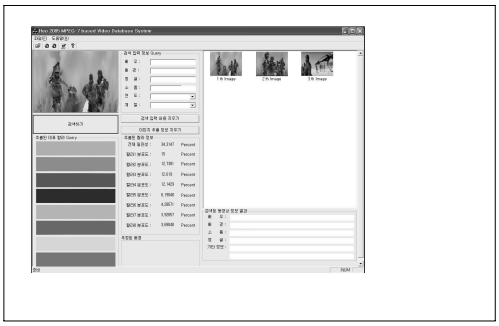

**Fig. 10. Query results with dominant color**

Table 1 tabulates the shot detection result in the propose system. In every case, the correctness is over 90%. As for key frame, the ratio of undetected shot is 7% ~ 10%. These figures are counted manually for better correctness. This results show that the proposed system is superior to other systems. Furthermore, the propose system can provide the compatibility of Meta data which cannot be found in other systems.

**Table 1. Shot detect result**

| Video | frames | key frames | correctness | udetected shots |
|---|---|---|---|---|
| D-War | 1850 | 32 | 98.2% | 3 |
| Sider man | 2703 | 48 | 92.7% | 4 |
| Harry Potter | 4083 | 73 | 94.1% | 12 |
| Transformer | 2304 | 28 | 93.5% | 6 |
| Car | 882 | 13 | 96.0% | 7 |
| Crink | 924 | 10 | 97.3% | 2 |

## V. Conclusion

This paper proposes a shot verification using visual rhythm and video retrieval system. The shot verification is designed to detect a segment from video easily and quickly, known as shot boundaries, just by changing the visual rhythm without playing the image. Therefore, this can decrease the false detected shot and generate the unidentified shot and key frame. The retrieval system is constructed in terms of visual descriptor from the list of MPEG-7. The effectiveness of the proposed shot verification process and video retrieval system is demonstrated.

## [ References ]


[1] H. M. Kim, J. H. Lee, J. H. Yang, S. H. Sull, S. Song, "Visual Rhythm and Shot Verification," Multimedia Tools and Applications, vol. 15, no. 3, p. 227-245, 2001.

[2] M. Petkovic, W. Jonker, Content-Based Video Retrieval: A Database Perspective, Kluwer Academic Publishers, Boston, Monograph, 2003.

[3] Carnegie Mellon University Informedia Digital Video Library, 2002, http://www.informedia.cs.cmu.edu.

[4] P. Aigrain, H. Zhang, and D. Petkovic, "Content based Representation and Retrieval of Visual Media: A State-of-the-Art Review," *Multimedia Tools and Applications*, Kluwer Academic Publishers, 3(3), pp. 179-202, 1996,

[5] J. R. Smith, S-F. Chang, "VisualSEEk: A Fully Automated Content-Based Image Query System," *in Proceedings ACM Multimedia Conference*, Boston, MA, November 1996.

[6] A. Pentland, R. W. Picard, and S. Sclaroff, "Photobook: Content-Based Manipulation of Image Databases," *International Journal of Computer Vision*, vol. 18, no. 3, pp. 233-254.

[7] C. Carson, M. Thomas, S. Belongie, J. M. Hellerstein, and J. Malik, "Blobworld: A System for Region-Based Image Indexing and Retrieval," *In Third International Conference on Visual Information Systems*, 1999, pp. 509-516.

[8] H. Zhang and S. W. Smoliar, "Developing power tools for video indexing and retrieval," *Proceedings of SPIE Storage and Retrieval for Image and Video Database II*, Vol. 2185, pp. 140-149, 1994.

[9] S. F. Chang, "Overview of the MPEG-7 Standard," *IEEE Transactions On Circuits and Systems for video Technology*, Vol. 11, No. 6, pp. 688-695, Jun. 2001.

[10] H. Kuyng, "The Research of Meta Data in Multimedia Contents," Osan College, No. 25, 2005.

[11] eXtensible Markup Language(XML) 1.0(second edition), W3C Recommendation, http;//www.w3.org/TR/xml, Oct. 2000.

[12] "XML Schema," W3C Recommendation, http://www.w3.org/TR/ xmlschema, May. 2001.


## Biography


**Ji Moon Chung**

1989년 연세대학교 공학대학원(공학석사)
2008년 충북대학교 컴퓨터공학과(공학박사)
1987년 한국국방연구원 총괄팀장
1989년~1994년 혜천대학 전자계산과
1994년~현재 남서울대학교 컴퓨터학과

<관심분야> Database. Cloud Computing

<e-mail> jmchung@nsu.ac.kr



**Cheong Ghil Kim**

2003년 연세대학교 컴퓨터과학과(공학석사)
2006년 연세대학교 컴퓨터과학과(공학박사)
2006년~2007년 연세대학교 박사후 연구원
2007년~2008년 연세대학교 연구교수
2006년~현재 남서울대학교

<관심분야> Multimedia Embedded Systems

<e-mail> cgkim@nsu.ac.kr