A Study on Shot Segmentation and Indexing of Language Education Videos by Content-based Visual Feature Analysis

교육용 어학 영상의 내용 기반 특징 분석에 의한 샷 구분 및 색인에 대한 연구

Heejun Han (한희준)*

ABSTRACT

As IT technology develops rapidly and the personal dissemination of smart devices increases, video material is especially used as a medium of information transmission among audiovisual materials. Video as an information service content has become an indispensable element, and it has been used in various ways such as unidirectional delivery through TV, interactive service through the Internet, and audiovisual library borrowing. Especially, in the Internet environment, the information provider tries to reduce the effort and cost for the processing of the provided information in view of the video service through the smart device. In addition, users want to utilize only the desired parts because of the burden on excessive network usage, time and space constraints. Therefore, it is necessary to enhance the usability of the video by automatically classifying, summarizing, and indexing similar parts of the contents. In this paper, we propose a method of automatically segmenting the shots that make up videos by analyzing the contents and characteristics of language education videos and indexing the detailed contents information of the linguistic videos by combining visual features. The accuracy of the semantic based shot segmentation is high, and it can be effectively applied to the summary service of language education videos.

초 록

IT기술이 급속히 발달하고 스마트 기기의 개인보급이 늘어나면서 정보의 전달 매체로 시청각 자료 중에서도 특히 영상 자료가 많이 활용된다. 문헌정보서비스 콘텐츠로서 영상자료는 필수 요소가 되었으며, TV를 통한 단방향 전달, 인터넷을 통한 양방향 서비스, 도서관 시청각 자료 대출 등 다양한 방법으로 활용되고 있다. 특히 인터넷 환경에서 스마트 기기를 통한 영상서비스 관점에서 정보 제공자는 제공 정보에 대한 가공에 적은 노력과 비용을 들이고자 하고, 또한 사용자는 과도한 데이터 사용량에 대한 부담과 시간, 공간적인 제약으로 인해 원하는 부분만을 효율적으로 이용하고자 한다. 따라서 영상에 대한 내용을 유사한 부분끼리 자동으로 구분하고 요약, 색인하여 이용 편의성을 높일 필요가 있다. 본 논문에서는 교육용 어학 영상의 내용과 그 특성을 분석하여 영상을 이루는 샷을 자동으로 구분하고 비주얼 특징을 조합하여 어학 영상의 세분화된 내용 정보를 결정하고 색인하는 방법을 제안한다. 외국어 강의 영상을 이용한 실험에 의해 의미기반의 샷 결정에 높은 정확률을 보였으며, 교육용 어학 영상의 요약 서비스에 효율적으로 적용 가능함을 확인하였다.

Keywords: shot segmentation, keyframe extraction, content-based summary, language education video, shot indexing
샷 구분, 키프레임 추출, 내용기반 요약, 교육용 어학 영상, 샷 색인

* 경기대학교 대학원 문헌정보학과 박사과정(hfireguy@gmail.com)

■ 논문접수일자 : 2017년 2월 21일 ■ 최초심사일자 : 2017년 3월 6일 ■ 게재확정일자 : 2017년 3월 13일
■ 정보관리학회지, 34(1), 219-239, 2017. 〔http://dx.doi.org/10.3743/KOSIM.2017.34.1.219〕

1. Introduction

The media industry is seeking ways to converge its service with internet networking services, shifting from passive one-way streaming to two-way interaction that can fulfill the needs of viewers. The two-way interaction service should be able to offer programs, content summaries, and program highlights that are customized to preferences and characteristics of the viewers. And consumer should be able to process and modify their media content. Also most education language video services, as well as Korea's representative language education site, 'e4u.ybmsisa.com' or 'www.ebs.co.kr/language', provide easy access to grammatical, vocabulary and expression areas in a single video. The growing demand for content that can satisfy diverse preferences has generated much research on summarizing, browsing, and indexing methods for media content. In addition, as the amount of media content increases, service providers are in need of video summarization techniques to manage content efficiently and build a database.

Several studies have been conducted on video summarization; however, most of them discuss content-based summarization methods where lower-level visual features are extracted from videos and analyzed to determine higher-level events.

Zhong and Chang (1997) used color features to segment frames and affine motion to track image segments. Sudhir, Lee, Jain (1998) explained automatic classification of tennis video developing a court line detection algorithm; and a robust player tracking algorithm to track the tennis players over the image sequence. Lee, Kim, Nam, Kang, Ro (2002) analyzed the golf-video contents in the shot with multiple content features using multiple MPEG-7 and proposed a content-based summary generation method using MPEG-7 metadata. In addition, many methods have been studied for detecting shots using various visual features, clustering scenes, and summarizing video (Basu, Yu, & Zimmermann, 2016; Divakaran, Peker, Radhakrishnan, Xiong, & Cabasson, 2003; Fei, Jiang, & Mao, 2017; Hu, Xie, Li, Zeng, & Maybank, 2011; Mundur, Rao, & Yesha, 2006; Ngo, Ma, & Zhang, 2005; Peng & Xiaolin, 2010; Thakre, Rajurkar, & Manthalkar, 2016). Sports videos or news, in particular, have well-structured content; therefore, they were the subject of a great deal of research with a focus on content-based summarization using colors, edge, and motion information. Meanwhile, despite the fact that language education videos, which are increasing in number, also have a concise and standardized structure, not many studies have been done about video summarization using multiple visual features for these videos.

This study therefore discusses an automatic summarization method for language education videos, which are showing rapid growth due to multichannel environment and heightened enthusiasm for education. The language education video industry will be able to provide efficient service and manage the huge volume of content effectively by using this automatic summarization method. This study identified properties of visual features of language education videos, extracted content-based features, combined the extracted features to generate segmented content information auto-

matically, and produced a meaningful summarization result. Some of the visual features that were used are internationally standardized MPEG-7 descriptors. These features were used in consideration of reusability and compatibility.

The rest of this paper is organized as follows: Section 2 discusses content analysis of language education videos; Section 3 explains properties of the visual features and a method to combine the features, and discusses the proposed summarization method which incorporates a segmentation algorithm of content information necessary for summarization; Section 4 describes the experiment on foreign language education videos and discusses the summarization result to verify the effectiveness of the proposed method. Finally, Section 5 presents conclusions and directions for future research.

2. Video Analysis

A meaningful content information analysis should be performed prior to summarizing a video. After reviewing about 20 contents of instructional language, most of the videos have lecturer explaining the contents and grammar of the lecture, showing the contents again with fingerprints, and there are conversation screens of various speakers how they are actually used. That is, language education videos usually consist of an explanation segment where a host gives a lecture, a dialog segment where speakers of the target language make conversation, a text-based segment that is composed of text information, and a remaining segment that contains other information as shown in <Table 1>.

Each segment comprises elements that have meanings. As shown in <Figure 1>, in the explanation segment, the host gives a lecture at a specific place such as a studio. The dialog segment contains the conversations among foreign speakers that take place at locations other than the place where the host is, either indoor or outdoor. The text segment is mainly composed of the text information appearing on a screen.

The explanation segment contains educational content about speaking, grammar, and pronunciation; the dialog segment has content useful for listening training. The text-based segment is important for learning grammar and vocabulary. The remaining segment includes the title of the program or information on the producer, or shows scene changes. This segment is not considered necessary for the summarization

〈Table 1〉 Content based Segmentation of Languages Education Video

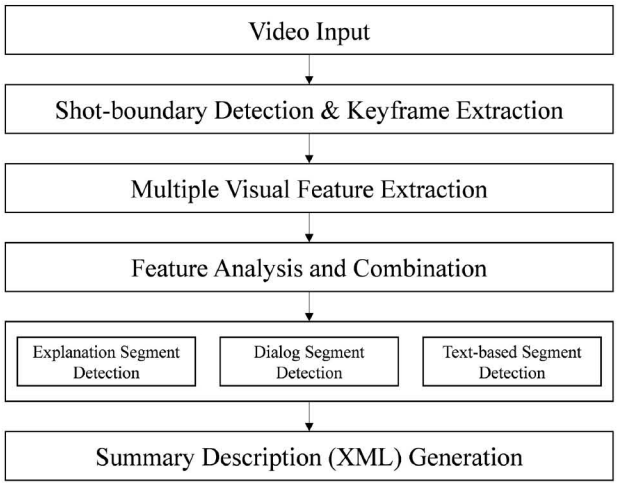
partition		Features
Languages Education Video	Explanation segment	The part that the moderator or instructor explains
	Dialog segment	The part where two or more speakers talk to each other
	Text-based segment	The part expressing the vocabulary and grammar of the sentence related to the language mainly on the fingerprint
	Remain segment	The rest of the above three parts

〈Figure 1〉 Example Frames of Each Segment

of language education videos because the segment not only does not contain important educational information, but also fails to meet the consumer's needs. A distinctive characteristic of educational content is that consumer preference for the content that can address their weaknesses is much higher than that for other content. Therefore, this study aims to detect the explanation segment, the dialog segment, and the text-based segment in language education videos and create content-based summary information automatically to offer content that suits the preferences of viewers.

3. Proposed Method

This section discusses an algorithm to generate summary information by detecting the segments of language education videos as defined in Section 2. <Figure 2> presents the structure of a summarization system for language education videos. It first detects shot boundaries from the input video then extracts key frames that represent each shot. It extracts visual features needed for the summarization from the extracted key frames and creates the pre-defined segmented content information such as the explanation segment, the dialog segment, and the text-based segment. The final generated content information is presented in XML document format in accordance with a hierarchical summary structure specified in MPEG-7 Multimedia Description Scheme (Manjunath, Salembier, & Sikora, 2002; Salembier & Smith, 2001).



〈Figure 2〉 Content based Video Segmentation Process

3.1 Detecting Shot Boundaries and Extracting Key Frames

After the segments of a language education video are defined, shots and key frames, which are basic units for visual feature extraction, are extracted. Along with color information, texture and edge information of the shots and frames that make up a video are important visual features that describe image information. The texture and edge information show structure, directionality, and roughness of an image and can be utilized as crucial features to extract shot boundaries for creating a content-based summary of the video data (Cieplinski, Jeannin, Kim, & Ohm, 2000; Ro, Yoo, Kim, & Kim, 1999; Yamada, Pickering, Jeannin, & Jens, 2001).

The shot boundary detection module used in this algorithm is an improved version of the shot boundary detection routine of the Hierarchical Summary DS, which was separated from the MPEG-7 reference software eXperimental Model (Walker & Sull, 1999). Shot boundaries are detected based on feature information that share similar distribution such as color, texture, and edge. Therefore, for easier calculation, a center frame, among all other frames that make up each shot, is determined as a key frame that represents the shot. The routine to obtain the key frame from each shot is as follows:

$$for \quad n=1 \quad through \quad TotalShotNum \quad do$$
$$\{\ keyframe_n = (Startframe_n - Endframe_n)/2\ \} \qquad (1)$$

In this routine, *TotalShotNum* denotes the number of shots detected from the input video, *Startframe* means the frame number of a beginning frame of each shot, and *Endframe* means the frame number of an end frame of each shot.

$$L = \{s_1, s_2, s_3, ..., s_i\}, \quad i = shot \quad number,$$
$$k_i = keyframe \quad of \quad s_i \qquad (2)$$

When L is the language education video and s is a shot, L is represented as a set of s, k is the key frame that represents the shot.

3.2 The Applied Multiple Visual Features

3.2.1 Kurtosis of Hue Histogram Distribution

Kurtosis of hue histogram distribution is used to detect the same background color in the text-based segment. The hue histogram exhibits hue distribution of an image. Hue values are computed from RIB values of the image, and data distribution for every 360th bin is obtained (Cieplinski, Kim, Ohm, Pickering, & Yamada, 2001). Kurtosis indicates pointiness of a distribution. Formula (3) is a method of calculating kurtosis of the hue histogram of key frames that represent each shot.

$$H_{k_i} = \{h_i^1, h_i^2, h_i^3, ..., h_i^n\}, \quad n = 360$$

$$Kur_{k_i} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{p=1}^{n} \left(\frac{h_i^p - \bar{h}_i}{SD_i} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

$$where \quad SD_{k_i} = \sqrt{\frac{n\sum_{q=1}^{n}(h_i^q)^2 - \left(\sum_{q=1}^{n} h_i^q\right)^2}{n^2}} \qquad (3)$$

H_{k_i} is a set of hue histogram data h_i values extracted from the i[th] key frame k_i. SD_{k_i} is standard deviation of h_i, and Kur_{k_i} is a kurtosis value of the hue histogram distribution of k_i.

3.2.2 30-channel Energy Standard Deviation of Homogeneous Texture Descriptor

Texture information defines a pattern that indicates homogeneity of an image and determines image homogeneity of frames that make up the text-based segment of a language education video. To obtain texture feature information, Radon transform is applied on images, and one-dimension Fourier transform is applied on the projected data. Frequency domain of the images is then divided into six directional components and five size components and the energy of the 30 channels is calculated (Cieplinski et al., 2001). Formula (4) is used to calculate standard deviation of the 30 channel energy of key frames in order to extract texture information.

$$E_{k_i} = \{e_i^1, e_i^2, e_i^3, ..., e_i^m\}, \quad m = 30 \text{ and}$$

$$SD_{k_i\_ChannelEnergy} = \sqrt{\frac{m\sum_{r=1}^{m}(e_i^r)^2 - \left(\sum_{r=1}^{m} e_i^r\right)^2}{m^2}} \qquad (4)$$

E_{k_i} is a set of the channel energy e_i extracted from k_i, and $SD_{k_i\_ChannelEnergy}$ is standard deviation of E_{k_i} which is an element of e_i.

3.2.3 Edge Histogram Descriptor

The visual feature specified by the edge histogram descriptor is used to characterize an image of the host who appears in the explanation segment. In order to represent edge information of an image, the image is divided into 16 sub-blocks and five edge components - vertical, horizontal, 45 degree, 135 degree, and non-directional edges - are denoted for each sub-blocks. A total of 80 local edge histograms are computed from one image by combining the edge components of each block and 40 semi-global edge histograms and 5 global edge histograms

are computed by combining the local edge histograms of the 16 sub-blocks. A feature vector of the edge histogram descriptor and a similarity distance value calculated using the feature vector are as follows:

$$\overline{ED} = \begin{cases} f_{local\_1}, f_{local\_2}, \cdots, f_{local\_80} \\ f_{semi-global\_1}, f_{semi-global\_2}, \cdots, f_{semi-global\_40} \\ f_{global\_1}, f_{global\_2}, \cdots, f_{global\_5} \end{cases} \quad (5)$$

$$dist_{ED} = \sum_i \left| \overline{ED}_{keyframe}(i) - \overline{ED}_{keyframe'}(i) \right| \quad (6)$$

$\overline{ED}$ is a feature vector of the edge histogram descriptor. f_{local_s} indicates the s[th] local edge histogram bin, $f_{semi-global\_t}$ denotes the t[th] semi-global edge histogram bin, and f_{global_u} is the u[th] global edge histogram bin (Cieplinski et al., 2001). $dist_{ED}$ is distance calculated using the edge histogram feature that are used to measure similarity between key frames. $keyframe$ and $keyframe'$ are frames that are different from each other.

3.2.4 Scalable Color Descriptor

The visual feature defined by scalable color descriptor is used to separate the color feature of the explanation segment from the color feature of the text-based and dialog segments. This feature is represented as a histogram that shows distribution of colors in an image. RBG values of the image are transformed to HSV values nonlinearly, and a feature vector is computed by dividing the HSV color space into a total of 256 bins and counting the number of pixels in each bin (Cieplinski et al., 2001). The distance values calculated using the color descriptor feature are used to measure similarity between frames. The distance values are calculated as in Formula (7).

$$dist_{CD} = \sum_i \left| \overline{CD}_{keyframe}(i) - \overline{CD}_{keyframe'}(i) \right| \quad (7)$$

$dist_{CD}$ is similarity distance between frames computed using the color descriptor feature. $keyframe$ and $keyframe'$ are frames that are different from each other.

3.2.5 Intensity of Motion Activity

Motion intensity is a feature that expresses the degree of motions of an object in video sequence within a certain range. The motion intensity is standard deviation of motion vector magnitude values, which are normalized and quantized by frame resolution, obtained from a macro bloc of each frame (Peker, Divakaran, & Papathomas, 2001). This feature value is used to distinguish the dialog segment where objects show high degree of motions from the explanation segment where objects show low degree of motions. Unlike other features, the motion intensity is computed by shot unit. The method to compute motion intensity from the shots is as following:

$$mv_{mag} = \sqrt{mv_x^{\,2} + mv_y^{\,2}},$$

$$Intensity_{motion} = \sqrt{\frac{\sum_{u=1}^{w \times h \times n} (mv_{mag})^2}{w \times h \times n} - \left(\frac{\sum_{u=1}^{w \times h \times n} mv_{mag}}{w \times h \times n}\right)^2} \quad (8)$$

mv_{mag} is motion vector magnitude and mv_x indicates a horizontal motion vector and mv_y is indicates a vertical motion vector. w and h are width and height of a frame and n is the number of frames that constitute shot. The motion intensity $Intensity_{motion}$ is generated

by calculating standard deviation of mv_{mag}.

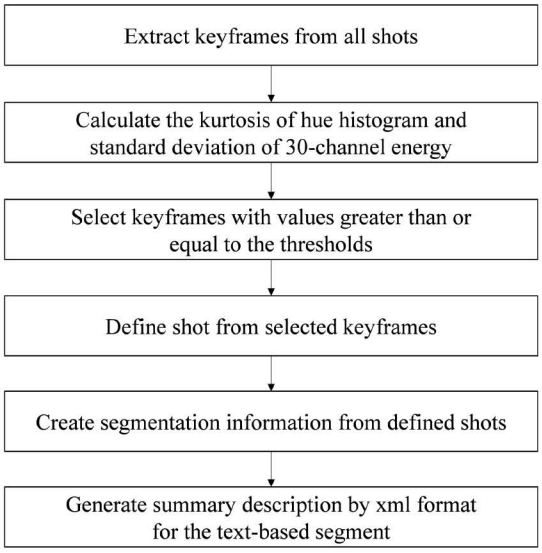
3.3 Detecting Segmented Content of Videos

This section discusses a method to detect the explanation segment, the dialog segment, and the text-based segment of a language education video using the multiple visual features described in the previous section. Shots and key frames, which are the basic unit of visual feature extraction, from the input video are detected and shots and key frames that make up the explanation segment, the dialog segment, and the text-based segment are classified using the extracted features. After segment information for the explanation, dialog segment, and text-based segments are generated, an XML document that represents summary information of the language education video is produced in accordance to the hierarchical summary structure of MPEG-7.

3.3.1 Method to Detect the Text-Based Segment

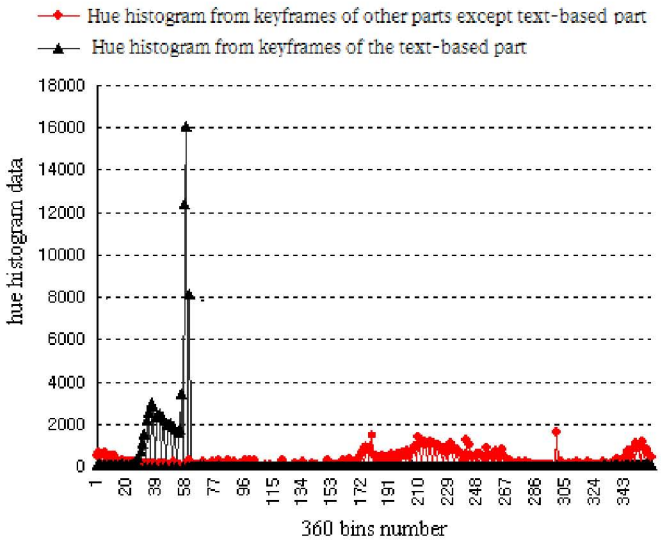
In order to detect the text-based segment, kurtosis of hue histogram of the key frames that represent each shot and the 30 channel energy standard deviation specified in the MPEG-7 homogeneous texture descriptor are used. <Figure 3> is a simple flowchart of the process to detect the text-based segment and produce summary information. If a visual feature value to detect the text-based segment satisfies threshold conditions, the corresponding key frames that have the visual feature make up the shots that constitute the text-based segment. Then summary information for the segment is generated using segment information of the shots.

Frames that represent the shots that constitute the text-based segment have background of almost the same color, on which text information appears.

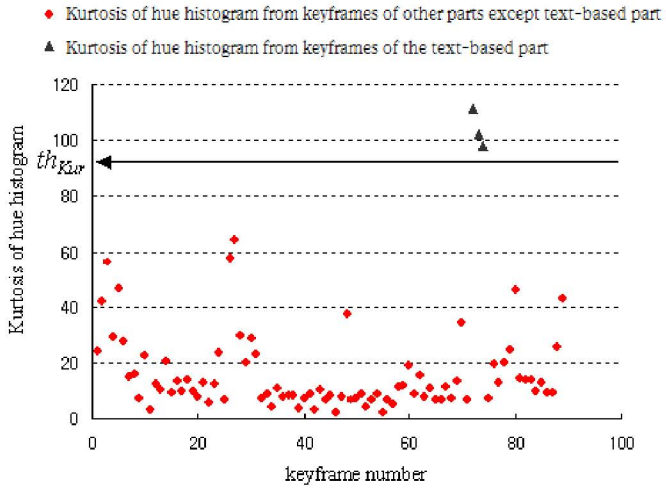


〈Figure 3〉 Flow Chart of Detecting the Text-Based Segment

Unlike the hue histograms generated from key frames of the explanation segment and dialog segment, hue histogram data of key frames of the text-based segment are concentrated in one area as shown in <Figure 4> because of the same background color of the key frames and the distribution tends to be pointier. Therefore, kurtosis values of the hue histogram are large positive numbers. <Figure 5> shows the kurtosis values of the hue histogram distribution of each key frame that represents shots from the video.



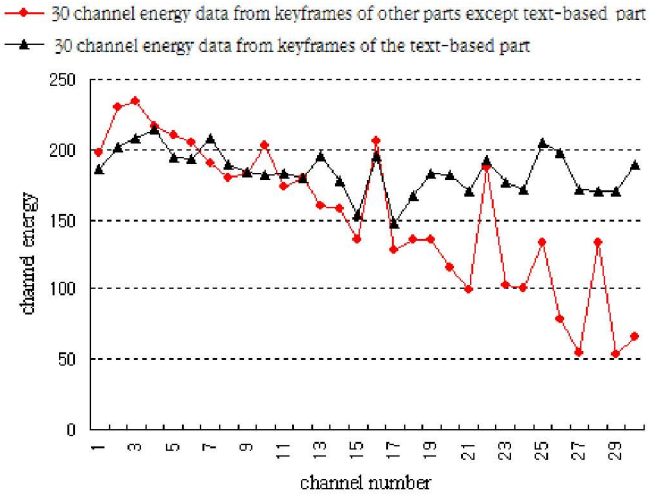
〈Figure 4〉 Hue Histogram from Keyframes



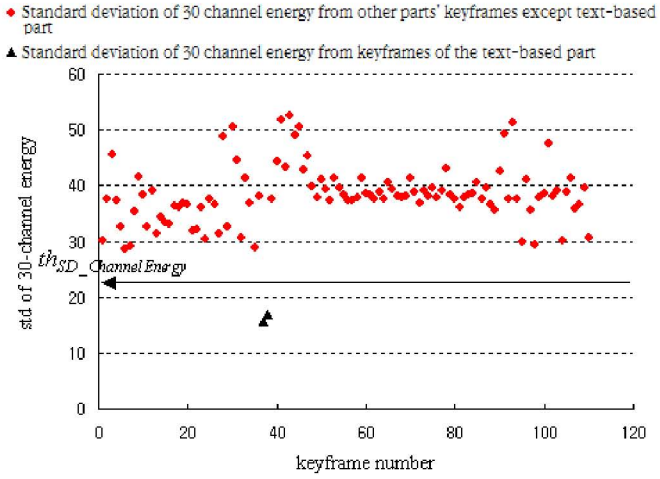
〈Figure 5〉 Kurtosis of Hue Histogram from Keyframes

As exhibited in <Figure 5>, kurtosis values of key frames that represent segments other than the text-based segment are numbers between 0 and 40 while those of key frames of the text-based segment are larger than 95. Therefore, the key frames for which kurtosis is higher than the kurtosis threshold are determined to be the key frames that represent shots of the text-based segment.

In addition, since the key frames of the text-based segment are consisted of almost the same color information, the 30 channel energy data specified by the texture descriptor are scattered within the small range between 150 and 200 as shown in <Figure 6>. On the other hand, the channel energy data obtained



〈Figure 6〉 30-channel Energy of Homogeneous Texture Descriptor



〈Figure 7〉 30-channel Energy Standard Deviation of Homogeneous Texture Descriptor

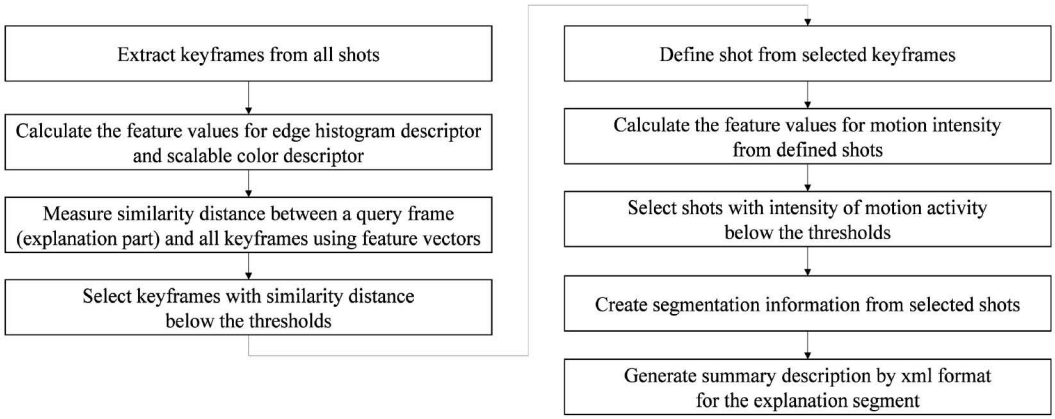
from key frames of segments other than the text-based segment are distributed over a wider range. Therefore, as shown in <Figure 7>, the standard deviation of the key frames that represent the text segment has smaller values than that of the explanation or dialog segments where a variety of colors are used. The frames that represent each shot have one channel energy standard deviation value as shown in Formula (4). If this value is smaller than the threshold $th_{SD\_ChannelEnergy}$, the corresponding frames are determined to be obtained from shots that constitute the text segment. Then, lastly, key frames that satisfy the two threshold conditions th_{kur} and $th_{SD\_ChannelEnergy}$ are detected by computing AND computation of the two conditions and the corresponding shots are generated as segments that constitute the text-based segment.

3.3.2 Method to Detect the Explanation Segment

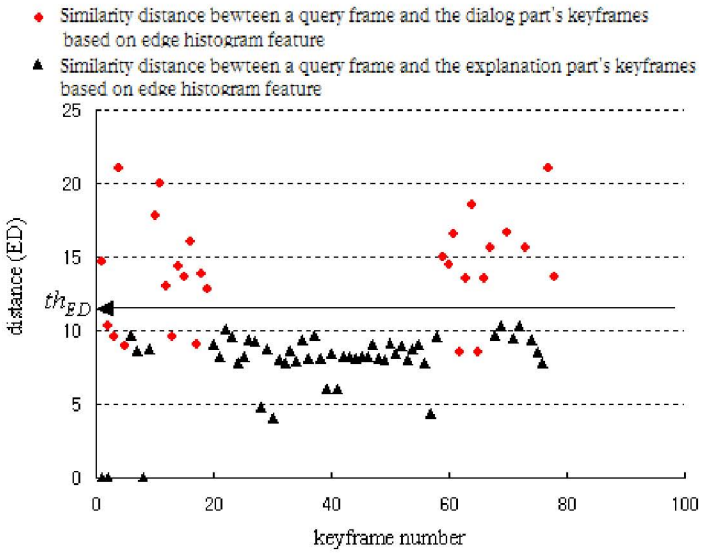
The process to detect the explanation segment is as shown in <Figure 8>. The three features described in the section 3.2.3 to 3.2.5 were used to detect the explanation segment. A feature analysis is conducted to determine shots that correspond to the explanation segment and segment information is generated.

After detecting shots and extracting key frames, edge histogram and scalable color descriptor are generated from each key frame. Then similarity between features of the frame that appears after playing the video for one minute, which is the 1800th frame that belongs to the explanation segment and also a query frame to measure similarity, and features of all the other key frames is measured using Formula (6) and (7).

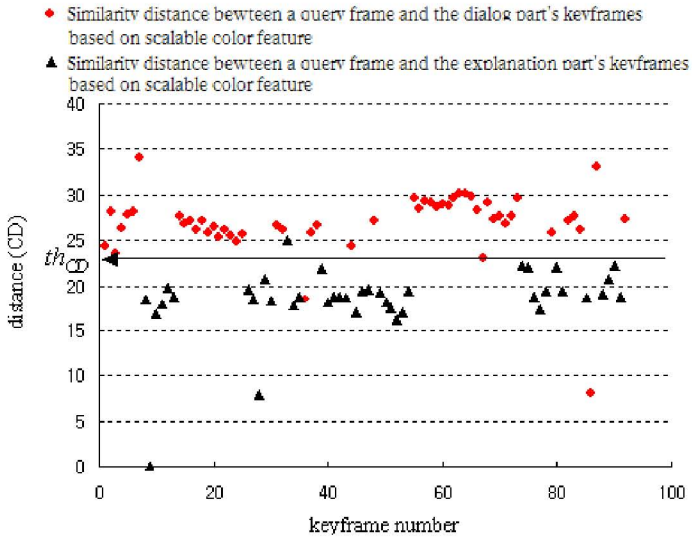
<Figure 9> and <Figure 10> show the similarity measures. As shown in <Figure 9> and <Figure 10>, if the similarity value satisfies conditions below thED and thCD at the same time by computing the AND operation, the corresponding key frame is determined to represent shots that constitute the explanation segment.



〈Figure 8〉 Flow Chart for the Process of Detecting the Explanation Segment

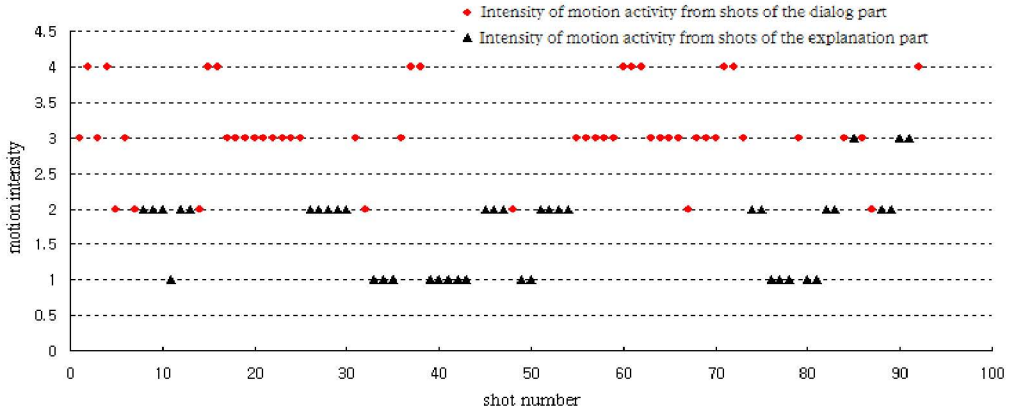


〈Figure 9〉 Similarity Distance among Features by Edge Histogram Descriptor



〈Figure 10〉 Similarity Distance among Features by Scalable Color Descriptor

Next, motion intensity is calculated from the shots that are determined to belong to the explanation segment. <Figure 11> is an example of motion intensity feature of the explanation segment and the dialog segment. Since the explanation segment consists of shots where an object shows almost no motion, motion intensity is relatively low; therefore, excluding shots of which motion intensity exceeds a certain level can detect the shots that make up the explanation segment.

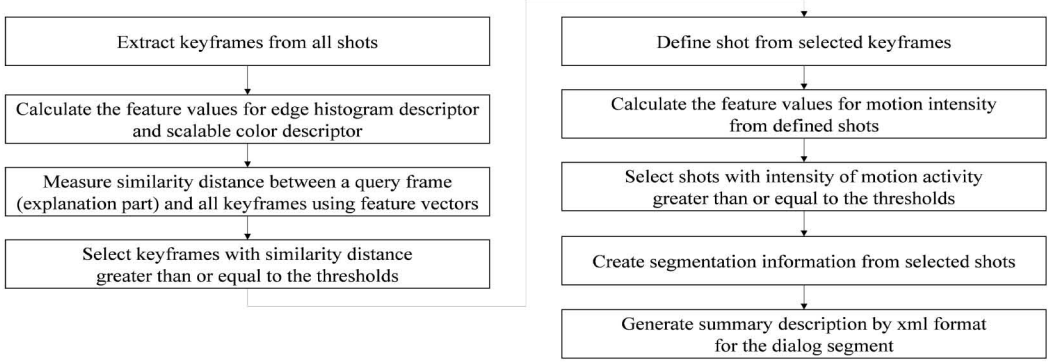


〈Figure 11〉 Intensity of Motion Activity from Dialog Segment and Explanation Segment

3.3.3 Method to Detect the Dialog Segment

Shots that constitute the dialog segment should be detected to generate the dialog segment. As in the process of detecting the explanation segment, edge histogram, scalable color descriptor, and motion intensity feature values are used for this process. As shown in <Figure 9 and 10>, if the similarity value between features of a key frame and features of the query frame that belongs to the explanation segment exceeds th_{ED} and th_{CD}, the key frame is determined to constitute the dialog segment.

Then a process to detect shots with high motion intensity from the shots that are determined to belong to the explanation segment is repeated. As shown in <Figure 11>, objects in the shots of the dialog segment show bigger motions, thus the motion intensity is larger than that of the shots from the explanation segment. Therefore, the shots with bigger motion intensity are determined to constitute the dialog segment and segment information is generated. <Figure 12> is a simple flow chart of the process of detecting the dialog segment and generating summary information.



〈Figure 12〉 Flow Chart of Detecting the Dialog Segment

4. Results of the Experiment

I conducted an experiment on foreign language education video to verify the effectiveness of the proposed automatic summarization method. I used a total of 15 videos having 20-minutes duration in MPEG-2 format, which included videos on learning conversational Chinese, English, French, German, and Japanese. <Table 2> is the result of extracting shot boundaries and key frames from the 15 videos and each number indicates the number of shots. These numbers indicate the answer set identified by eye-checking. Since each shot has one key frame, the number of shots and the number of key frames of a video are the same.

I first obtained hue histogram data from the extracted key frames and calculated kurtosis values to detect the text-based segment, then computed standard deviation of the 30 channel energy of the key frames by applying the texture descriptor. Next, I detected key frames that have feature values that meet the given threshold values. The shots represented by the detected key frames constitute segment information for the text-based segment.

<Table 3> is the result of detecting the text-based segment by applying kurtosis values of hue histograms and <Table 4> is the result using the 30 channel energy standard deviation. Precision rate is low when a single feature is applied, however, the result of detecting the text-based segment by applying multiple

〈Table 2〉 Shot Composition of Language Education Videos

	Explanation segment	Dialog segment	Text-based segment	Remain segment	Total
Chinese Education Video	147	118	9	42	316
English Education Video	140	124	4	32	300
French Education Video	144	120	7	41	312
German Education Video	137	117	11	28	293
Japanese Education Video	150	99	4	30	283
Total no. of shot	718	578	35	173	1504

〈Table 3〉 The Result of Detecting the Text-Based Segment Using Kurtosis Values of Hue Histograms

	Answer	Correct	Miss	False	Recall(%)	Precision(%)
Chinese	9	8	1	23	88.89	25.81
English	4	4	0	4	100	50.00
French	7	7	0	8	100	46.67
German	11	10	1	11	90.91	47.62
Japanese	4	4	0	0	100	100
Total	35	33	2	46	94.29	41.77

combination of kurtosis values of hue histograms and the 30 channel energy standard deviation shows the precision rate as high as 90% as in <Table 5>. This is due to the fact that the text-based segment has distinctive features compared to the explanation or dialog segments when the hue histogram and the 30 channel energy standard deviation are applied.

The videos used for the experiment contained fairly large of amount of the explanation and dialog segments, thus these segments have more shots than the text-based segment. What is important is to apply an effective combination of the three features - edge histogram, scalable color, and motion intensity - to generate efficient summary information for the explanation and dialog segments. Detecting the text-based segment using each of the single features does not yield high precision rate; therefore, I applied the combining technique described in sections 3.3.2 and 3.3.3 and improved the precision rate. <Table 6> is the result of applying the feature defined in edge histogram descriptor only in order to detect the explanation segment and <Table 7> is the result of detecting the explanation segment using a scalable color feature.

<Table 8> is the result of detecting the explanation segment by applying edge histogram and the scalable color feature at the same time. For better precision rate, I applied the motion intensity feature to the result in <Table 8>. <Table 9> is the result of detecting the explanation segment by applying the combination of the three features, which shows very high precision rates.

〈Table 4〉 The Result of Detecting the Text-Based Segment using the 30 channel Energy Standard Deviation of the Texture Descriptor

	Answer	Correct	Miss	False	Recall(%)	Precision(%)
Chinese	9	8	1	8	88.89	50.00
English	4	4	0	17	100	19.05
French	7	6	1	22	85.71	21.49
German	11	10	1	48	90.91	17.24
Japanese	4	4	0	0	100	100
Total	35	32	3	95	91.43	25.20

〈Table 5〉 The Result of Detecting the Text-Based Segment using Multiple Features (the Kurtosis of Hue Histogram and 30-channel Energy Standard Deviation of Homogeneous Texture Descriptor)

	Answer	Correct	Miss	False	Recall(%)	Precision(%)
Chinese	9	9	0	3	100	75.00
English	4	4	0	1	100	80.00
French	7	6	1	2	85.71	75.00
German	11	10	1	2	90.91	83.33
Japanese	4	4	0	0	100	100
Total	35	33	2	8	94.29	80.49

〈Table 6〉 The Result of Detecting the Explanation Segment Using Edge Histogram Feature Values

	Answer	Correct	Miss	False	Recall(%)	Precision(%)
Chinese	147	118	29	36	80.27	76.62
English	140	123	17	87	87.86	58.57
French	144	142	2	33	98.61	81.14
German	137	123	14	118	89.78	51.04
Japanese	150	141	9	30	94.00	82.94
Total	718	647	71	304	90.11	68.03

〈Table 7〉 The Result of Detecting the Explanation Segment Using Scalable Color Feature Values

	Answer	Correct	Miss	False	Recall(%)	Precision(%)
Chinese	147	138	9	109	93.88	55.87
English	140	135	5	16	96.43	89.40
French	144	130	14	75	90.28	63.41
German	137	122	15	18	89.05	87.14
Japanese	150	142	8	54	94.67	72.45
Total	718	667	51	272	92.89	71.03

〈Table 8〉 The Result of Detecting the Explanation Segment using Multiple Features (Edge Histogram Descriptor and Scalable Color Descriptor)

	Answer	Correct	Miss	False	Recall(%)	Precision(%)
Chinese	147	128	19	20	87.07	86.49
English	140	133	7	16	95.00	89.26
French	144	130	14	30	90.28	81.25
German	137	122	15	9	89.05	93.13
Japanese	150	141	9	28	94.00	83.43
Total	718	654	64	103	91.09	86.39

〈Table 9〉 The Result of Detecting the Explanation Segment using Multiple Features (Edge Histogram Descriptor, Scalable Color Descriptor and Intensity of Motion Activity)

	Answer	Correct	Miss	False	Recall(%)	Precision(%)
Chinese	147	128	19	15	87.07	89.51
English	140	130	10	10	92.86	92.86
French	144	129	15	25	89.58	83.77
German	137	122	15	6	89.05	95.31
Japanese	150	138	11	3	92.00	97.87
Total	718	647	71	59	90.11	91.64

As explained in section 3.3.3, key frames of which similarity value of edge histogram and a scalable color feature, which are the values that measure similarity between the corresponding key frames and the key frames that constitute the explanation segment, exceeds a given threshold represent the shots that make up the dialog segment. In addition, as the shots that constitute the dialog segment generally have more motion information, the precision rate of detecting can be improved by applying the motion intensity feature. <Table 10>, <Table 11>, and <Table 12> is the result of detecting the dialog segment by each visual feature. And <Table 13> is the result of detecting the dialog segment by combining the visual features, and <Figure 13> is an example of detecting the text-based segment. These are the key frames of shots that form the dialog segment of a conversational Japanese education video.

〈Table 10〉 The Result of Detecting the Dialog Segment Using Edge Histogram Feature Values

	Answer	Correct	Miss	False	Recall(%)	Precision(%)
Chinese	118	105	13	28	88.98	78.95
English	124	109	15	57	87.90	65.66
French	120	106	14	40	88.33	72.60
German	117	98	19	77	83.76	56.00
Japanese	99	75	24	25	75.76	75.00
Total	578	493	85	227	85.29	68.47

〈Table 11〉 The Result of Detecting the Dialog Segment Using Scalable Color Feature Values

	Answer	Correct	Miss	False	Recall(%)	Precision(%)
Chinese	118	106	12	45	89.83	70.20
English	124	118	6	33	95.16	78.15
French	120	112	8	28	93.33	80.00
German	117	99	18	50	84.62	66.44
Japanese	99	80	19	10	80.81	88.89
Total	578	515	63	166	89.10	75.62

〈Table 12〉 The Result of Detecting the Dialog Segment using Multiple Features (Edge Histogram Descriptor and Scalable Color Descriptor)

	Answer	Correct	Miss	False	Recall(%)	Precision(%)
Chinese	118	108	10	28	91.53	79.41
English	124	118	6	15	95.16	88.72
French	120	112	8	23	93.33	82.96
German	117	106	11	15	90.60	87.60
Japanese	99	82	17	12	82.83	87.23
Total	578	526	52	93	91.00	84.98

〈Table 13〉 The result of Detecting the Dialog Segment using Multiple Features (Edge Histogram Descriptor, Scalable Color Descriptor and Intensity of Motion Activity)

	Answer	Correct	Miss	False	Recall(%)	Precision(%)
Chinese	118	106	12	19	89.83	84.80
English	124	114	10	10	91.94	91.94
French	120	98	22	15	81.67	86.73
German	117	114	3	15	97.44	88.37
Japanese	99	97	2	11	97.98	89.81
Total	578	529	49	71	91.52	88.17

〈Figure 13〉 A Segment of the Key Frames Detected As a Dialog Segment of a Conversational Japanese Education Video

Finally, when the explanation, dialog, and text segments of the videos are detected, I generated an XML document in accordance to the hierarchical summary structure as defined in MPEG-7 MDS by using segment information of the shots that constitute each segment. As show in <Figure 14>, when the shots of the input video are determined to constitute each segment, the shots are chronologically connected to become one video clip. The frame number of a beginning frame of the video clip and that of an ending frame are used as media time information of the key video clip components specified in the hierarchical summary structure. The figure below is a segment of an XML document that represents summary information of a conversational English education video, which contains segment information generated as a result of detecting explanation, dialog, and text-based segments.

In <Figure 14>, an attribute 'mediaTimeUnit' is expressed by a form of 'PnDTnHnMnSnNnF' and represents a duration in time using a lexical representation of days (nD), time duration and a fraction specification (TnHnMnSnN) including the specification of the number of fractions of one second (nF). So 'PT1N30F' is a typical examples of 1N which is 1/30 of a second according to 30F and that means

```
<?xml version="1.0" encoding="UTF-8"?>
<Mpeg7 xmlns="urn:mpeg:mpeg7:schema:2001" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:mpeg7="urn:mpeg:mpeg7:schema:2001" xsi:schemaLocation="urn:mpeg:mpeg7:schema:2001 ./Mpeg7-2001.xsd">
    <Description xsi:type="SummaryDescriptionType">
        <Summarization>
            <Summary xsi:type="HierarchicalSummaryType" hierarchy="independent">
                <SourceInformation href="../english.mpg"/>
                <SummaryThemeList>
                    <SummaryTheme id="Explanation_Part">Explanation Part</SummaryTheme>
                    <SummaryTheme id="Dialog_Part">Dialog Part</SummaryTheme>
                    <SummaryTheme id="Text_based_Part">Text-based Part</SummaryTheme>
                </SummaryThemeList>
                <!-- ********************************************************** -->
                <!--                    themeIDs: Explanation_Part              -->
                <!-- ********************************************************** -->
                <SummarySegmentGroup level="0" themeIDs="Explanation_Part">
                    <SummarySegment id="Explanation_1">
                        <KeyVideoClip>
                            <MediaTime>
                                <MediaRelIncrTimePoint mediaTimeUnit="PT1N30F">672</MediaRelIncrTimePoint>
                                <MediaIncrDuration mediaTimeUnit="PT1N30F">5082</MediaIncrDuration>
                            </MediaTime>
                        </KeyVideoClip>
                    </SummarySegment>
                    <SummarySegment id="Explanation_2">
                        <KeyVideoClip>
                            <MediaTime>
                                <MediaRelIncrTimePoint mediaTimeUnit="PT1N30F">8042</MediaRelIncrTimePoint>
                                <MediaIncrDuration mediaTimeUnit="PT1N30F">2230</MediaIncrDuration>
                            </MediaTime>
                        </KeyVideoClip>
                    </SummarySegment>
                    <SummarySegment id="Explanation_3">
                        <KeyVideoClip>
                            <MediaTime>
                                <MediaRelIncrTimePoint mediaTimeUnit="PT1N30F">10507</MediaRelIncrTimePoint>
                                <MediaIncrDuration mediaTimeUnit="PT1N30F">2812</MediaIncrDuration>
                            </MediaTime>
                        </KeyVideoClip>
                    </SummarySegment>
                    <SummarySegment id="Explanation_4">
```

〈Figure 14〉 An XML Document that Represents Summary Information

the exact sampling rate of NTSC of 29.97Hz. The value of 'MediaRelIncrTimePoint' element is a start frame number of segment, and the value of 'MediaIncrDuration' element is the number of frames of segment. For example, '672' is the start frame number of explanation_1 segment and '5082' is the number of frames of explanation_1 segment.

Service providers can use this method to offer summarized content that viewers want. In particular, users can easily summarize their videos using mobile apps using this algorithm.

5. Conclusion and Future Studies

This paper extracted content-based visual features from language education videos using visual features of the videos and some descriptors provided by the MPEG-7 international standard, detected the explanation, dialog, and text-based segments of the videos by using efficient combination of the features, and automatically generated summary information. Then, it created a summary that describes structural

content information of the videos. The amount of content of educational videos is growing and the viewers' demand for it is also increasing. The language education videos, in particular, are the content with higher demand for summary information. The proposed method in this paper provides accurate summary of language education videos and could be utilized in two-way broadcasting service, web service and mobile service to create content that are customized to viewers' preferences in the future. XML documents that represent summary information, which is the final product of the proposed method, can be used by the viewers to review efficiently and quickly and use the language education content. In addition, this method allows the viewers who want only segments of the language education content to browse the segment they want in an easy, accurate, and convenient manner.

More research on general summarization methods need to be done using more variety of language education content, as well as studies that use information about the viewers' preferences.

References

Basu, S., Yu, Y., & Zimmermann, R. (2016). Fuzzy clustering of lecture videos based on topic modeling. Proceedings of the 2016 14th International Workshop on Content-Based Multimedia Indexing (CBMI), 1-6. https://doi.org/10.1109/cbmi.2016.7500264

Cieplinski, L., Jeannin, S., Kim, M., & Ohm, J. R. (2000). Visual working draft 4.0. ISO/IEC JTC1/SC29/WG11 N, 3399.

Cieplinski, L., Kim, M., Ohm, J. R., Pickering, M., & Yamada, A. (2001). Text of ISO/IEC 15938-3/FCD information technology-multimedia content description interface-part 3 visual. ISO/IEC JTC1/SC29/WG11 N, 4062, 30-53.

Divakaran, A., Peker, K. A., Radhakrishnan, R., Xiong, Z., & Cabasson, R. (2003). Video summarization using mpeg-7 motion activity and audio descriptors. Video Mining, 6, 91-121. https://doi.org/10.1007/978-1-4757-6928-9_4

Fei, M., Jiang, W., & Mao, W. (2017). Memorable and rich video summarization. Journal of Visual Communication and Image Representation, 42, 207-217. https://doi.org/10.1016/j.jvcir.2016.12.001

Hu, W., Xie, N., Li, L., Zeng, X., & Maybank, S. (2011). A survey on visual content-based video indexing and retrieval. IEEE Transactions on Systems, Man, and Cybernetics, Part C, 41(6), 797-819. https://doi.org/10.1109/TSMCC.2011.2109710

Lee, H. K., Kim, C. S., Nam, J. H., Kang, K. O., & Ro, Y. M. (2002). Video contents summary using

the combination of multiple MPEG-7 metadata. Proceedings of the Korean Institute of Broadcast and Media Engineers Conference, 227-232.

Manjunath, B. S., Salembier, P., & Sikora, T. (2002). Introduction to MPEG-7: multimedia content description interface. Chichester; New York: John Wiley & Sons.

Mundur, P., Rao, Y., & Yesha, Y. (2006). Keyframe-based video summarization using delaunay clustering. International Journal on Digital Libraries, 6(2), 219-232. https://doi.org/10.1007/s00799-005-0129-9

Ngo, C. W., Ma, Y. F., & Zhang, H. J. (2005). Video summarization and scene detection by graph modeling. IEEE Transactions on Circuits and Systems for Video Technology, 15(2), 296-305. https://doi.org/10.1109/TCSVT.2004.841694

Peker, K. A., Divakaran, A., & Papathomas, T. V. (2001). Automatic measurement of intensity of motion activity of video segments. Progress in Biomedical Optics and Imaging, (4315), 341-351.

Peng, J., & Xiaolin, Q. (2010). Keyframe-based video summary using visual attention clues. IEEE MultiMedia, 17(2), 64-73. https://doi.org/10.1109/MMUL.2009.65

Ro, Y. M., Yoo, K. W., Kim, M. C., & Kim, J. W. (1999). Texture description using radon transform. ISO/IEC JTC1 SC29 WG11 (MPEG).

Salembier, P., & Smith, J. R. (2001). MPEG-7 multimedia description schemes. IEEE transactions on circuits and systems for video technology, 11(6), 748-759. https://doi.org/10.1109/76.927435

Sudhir, G., Lee, J. C. M., & Jain, A. K. (1998). Automatic classification of tennis video for high-level content-based retrieval. Proceedings of the 1998 IEEE International Workshop on Content-Based Access of Image and Video Database, 81-90. https://doi.org/10.1109/caivd.1998.646036

Thakre, K. S., Rajurkar, A. M., & Manthalkar, R. R. (2016). Video partitioning and secured keyframe extraction of MPEG video. Procedia Computer Science, 78, 790-798. http://dx.doi.org/10.1016/j.procs.2016.02.058

Walker, T., & Sull, S. (1999). Proposal for a video summary description scheme. Proceedings of the 2000 IEEE International Conference, 3, 1559-1562.

Yamada, A., Pickering, M., Jeannin, S., & Jens, L. C. (2001). MPEG-7 visual part of experimentation model. Version 9.0-Part 3 Dominant Color, ISO/IEC JTC1/SC29/WG11/N3914.

Zhong, D., & Chang, S. F. (1997). Video object model and segmentation for content-based video indexing. Proceedings of the 1997 IEEE International Symposium on Circuits and Systems, 2, 1492-1495. https://doi.org/10.1109/iscas.1997.622202