Retrieval of News Video using Video Sequence Matching

Young-tae, Kim
Broadcasting Media Research Group, ETRI,
161 Gajeong-dong, Yuseong-gu, Daejeon,
Korea
kytae@etri.re.kr

Tat-Seng, Chua
School of computing, National University of
Singapore, Singapore 117543
chuats@comp.nus.edu.sg

Abstract

In this paper, we propose a new algorithm to find video clips with different temporal durations and some spatial variations. We adopt a longest common sub-sequence (LCS) matching technique for measuring the temporal similarity between video clips. Based on the measure we propose 3 techniques to improve the retrieval effectiveness. First, we use a few coefficients in the low frequency region of DCT block as the basis to represent spatial features. Second, we heuristically determine a suitable quantization step-size for visual features to better tolerate spatial variations of similar video clips and propose a paired quantizer method. Third, we incorporate the compactness and/or continuity of matched common sub-sequences in the LCS measure to better reflect temporal characteristics of video. The performance of the proposed algorithm shows an improvement of 63.5% in terms of MAP (mean average precision) as compared to an existing algorithm. The results show that our approach is effective for news video retrieval.

Keywords: video retrieval, sequence matching, longest common sub-sequence, similarity measure.

1. Introduction

With the availability of huge amounts of multimedia data, especially video, there has been active research to process and manipulate multimedia information. Active research area include compression, indexing and retrieval [1~3], customization [4] and streaming [5], etc. This paper focuses on video retrieval. With the arrival of digital broadcasting era, it is expected that new requirements on video retrieval appear. For example, user may want to retrieve interesting scenes among the video data saved in the digital TV receiver [6]. Therefore video retrieval and indexing will become one of the most important functions in such environments.

Many algorithms for video matching and retrieval have been developed [7-15]. The techniques can broadly be broken down into 2 parts. One deals with the choice of feature vectors [7-8], while the other concerns with the similarity measure [9-12]. The features used include low level features such as the color, edge, texture, motion, and mid level features such as the moving object's trajectory and object's color, etc. As for similarity measure, there are typically two approaches. One is key-frame-based [9] and the other is sequence-based matching methods [10-12]. For the former, temporal information is reflected into shot boundary detection and key-frame selection process. In video sequence matching, the factors we should consider had been presented in [11].

One of the common characteristics of video matching algorithms is that they focus on specific target applications rather than general applications. For example, in video surveillance application, the target is to find video clips with moving objects [13]. The algorithm proposed in [14] has been developed to find commercials with multiple versions with different duration and order. In [15], the video matching algorithm is used to find actions in sports such as diving and jumping. Therefore feature vectors to be extracted and similarity measure to be used are determined according to its application objective.

News is one popular video genre that has been actively researched for video retrieval. News video has unique characteristics different from general video. One of them is that it has very structured and regular patterns [16]. For example it consists of sequence of stories. Each story is composed of shots of type anchor, reporter and interview, etc. Such characteristics can be utilized for efficient indexing and retrieval. Multi-modal approach has been used popularly to retrieve news video [17~18]. In the process, video, text and audio are utilized simultaneously. In the retrieval process, text information such as the embedded video text, closed caption, speech recognition output is dominantly used because news has semantic information.

In general, user tries to find interesting news by submitting text-based queries but sometimes user may want to find visually similar videos by submitting video clips as queries. Some interesting scenes are broadcast repeatedly with different temporal and/or spatial editions. In these cases the proposed approach can be a useful tool for news video retrieval. The objective of our approach is to find visually similar video clips from the stored news video when a query video clip is given. In other words, it aims to find relevant news stories possibly with different editions which mean clips with different temporal durations and spatial variations. As a typical example in the headline news, the introductory stories are broadcast in brief followed by detailed version in the main news section. The introductory stories in the headline news typically contain some spatial and/or sound effects of the tailor version to present some impression to a TV viewer. Moreover some interesting scenes may be broadcast for several days.

In this research, we aim to analyze video contents encoded in DCT based representation, such as MPEG, directly in compressed domain without decompression. For this purpose, we first propose a new variant of content features based on DC and a few low-frequency AC coefficients in the compressed domain. The reason of our choice is that these coefficients are sufficient to represent the main features of the block such as intensity, edge with fidelity. Second, we heuristically find a suitable quantization step size which is employed for string representation of spatial features. It is designed in order to better tolerate spatial variations between frames. Additionally, we propose the method using paired quantizer to avoid similar feature value being mapped to different bin. Finally we incorporate the compactness and/or continuity of matched common sub-sequence in the process of measuring a similarity to better reflect the temporally continuity of video.

This paper is organized as follows. Section 2 outlines the system architecture for video retrieval. Section 3 details the proposed algorithm, while Section 4 presents the experimental results. Finally in Section 5, we conclude this paper.

2. System architecture for video retrieval

The overall system architecture for content-based video retrieval is shown in Figure 1. The overall framework is based on video sequence matching using the longest common subsequence or edit distance as a similarity measure. In this approach, the content of each frame is encoded as a set of symbols, each based on a selected feature, and the temporal content of a video clip is modeled as a string of symbols. Video matching is then regarded as a matching of video feature trajectories in a multi-dimensional feature space. Here, the longest common sub-string matching algorithm is used to find similar video clips.

The system architecture is mainly divided into two parts. The first is analysis, while the second is retrieval. In the analysis part, visual features are extracted for both the query and video sequences in the database. The feature extraction procedure is the same for both the query and database videos. During retrieval, the extracted features are compared to find videos similar to the video clip queried by the user. The analysis is further divided into 4 sub-modules. The first is the pre-processing module, where the shot boundary is detected. For this task we adopt the algorithms proposed in [19]. In our research, shot is the basic unit for comparing video clip.

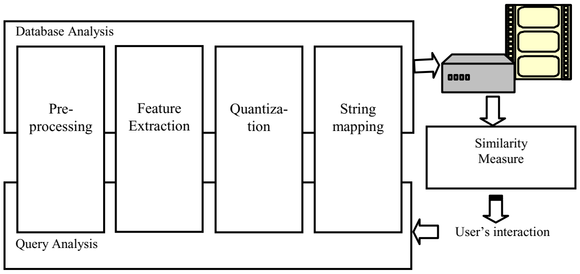


Figure 1. The overall system architecture

The second module extracts visual features such as color, edge and texture. The third module quantizes each set of continuous feature values into a smaller set of discrete feature values. The last module maps each quantized result to a symbol, one per video frame.

During retrieval, part of video clips with similar frame sequences is identified. Video clips are represented as a trajectory in a multi-dimensional feature space with one frame as a point in the space. To compare the similarity of two video clips, one obvious method is to perform multi-dimensional curve matching which is computationally expensive. One of the efficient approaches is to count the number of similar frames within two video clips in a fixed temporal order, but ignoring all the in-between dissimilar frames. This can be achieved by finding the longest common sub-sequence (LCS) within the two frame sequences. LCS of two input video clips can be computed by dynamic programming algorithm. In the similarity measure, the length of LCS rather than LCS itself is used. Let two shots be S_q, S_d, and its longest common sub-sequence be LCS *(S_q, S_d)*, the similarity between two shots can be expressed in a normalized form as shown in Equation (1):

$$Sim(S_q, S_d) = \frac{LCS(S_q, S_d)}{|S_q|} \quad (1)$$

Several features may be used in measuring the similarity between two shots. The overall similarity between the two shots can be expressed as the weighted sum of different features as shown in Equation (2):

$$Sim(S_q, S_d) = \sum_{i \in \{FeatureSet\}} w_i \cdot Sim_i(S_q, S_d) \quad (2)$$

where $w_i > 0$, $\sum w_i = 1$

3. The proposed algorithm

Based on the framework of video retrieval with LCS measure, we propose 3 techniques to improve the effectiveness of news video retrieval. First, in the feature extraction step, in order to express spatial features of the frames, we select a set of new features based on DC and a few AC values in low frequency region of the DCT Block. In the DCT-based compression scheme, they are important coefficients to capture the contents of blocks, and thus they are finely quantized. They represent the main features of the blocks such as the intensity and edge with fidelity. Second, in the quantization process we adjust the quantization step size, which determines whether two frames to be compared are matched or not. Here we try to adjust the quantization step size to "maximize" the tolerance of spatial variations by the algorithm. In the quantization step, although the frames may have very similar feature values, if they are located near the boundary of the quantization decision level, they may be mapped to different bins. In order to avoid this problem, we develop a new method using paired quantization. Third, in the similarity measure module we consider the continued length of longest matched common sub-sequence to better reflect temporal continuity of video. The existing measure is designed originally for measuring the similarity between strings. So it does not take into consideration the continuous characteristics of video. Actually it is biased towards matching longer shots. In this approach, we try to nullify the bias that long shots has.

3.1 Feature extraction

In our approach, we employ the DC and several low frequency AC coefficients in the DCT domain to develop our new features. The values of AC coefficients represent the energy of the corresponding frequency region. In particular most energy is concentrated in the low frequency region. The low frequencies we selected are *DC(0, 0)*, *AC(0, 1)*, *AC(1, 0)* and *AC(1, 1)*. DC represents average intensity of the block, whereas AC values represent directional edge components. *AC(0, 1)* and AC(1, 0) respectively represent the vertical and horizontal edge components in the low frequency region of the block [20]. We extract five features from the low-frequency components of the DCT block. The first two features are the first and second moment of DC component of the frame composed of M x N blocks as shown in Equations (3-4). The other three features are based on the averages of the absolute value of *AC (1, 0), AC (0, 1)* and *AC (1, 1)* as shown in Equation (5).

$$\overline{DC(0,0)} = \frac{1}{M * N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} DC_{(i,j)}(0,0) \quad (3)$$

$$Var(DC) = \frac{1}{M * N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} (DC_{(i,j)}(0,0) - \overline{DC(0,0)})^2 \quad (4)$$

$$|\overline{AC(p,q)}| = \frac{1}{M * N} \sum_{i=0}^{M-1} \sum_{j=0}^{N-1} |AC_{(i,j)}(p,q)| \quad (5)$$

3.2 Quantization

Quantization means the mapping of continuous feature values into a smaller finite set of values. In our scheme, we perform uniform quantization. We represent each feature component of a frame as a character. Because we use 5 features, there are 5 feature values for each frame. The quantization step size determines the degree of tolerance to spatial variation. If the step size is too big, it can tolerate much spatial variations. However, it decreases the differentiating power of the feature. If the step size is too small, it can tolerate only minor spatial variations. Therefore the choice of quantization step size is a very important factor to determine the performance of our algorithm. Let the maximum value of a feature be Max and the corresponding minimum value be Min. The dynamic range of the feature space is Max-Min. if we divide the range into *N* bins which are numbered from *0* to *N-1*. The step-size of a bin is *(Max – Min)/N*. The range of the *k*-th bin, R_k is expressed as shown in Equation (6)

$$(k-1)\cdot\frac{Max-Min}{N} < R_k \le k\cdot\frac{Max-Min}{N} \quad (6)$$

Because all feature values within a bin are mapped to the same character, the ratio of maximally tolerable variation to the dynamic range is *1/N*. The range of tolerable variation, *t*, is shown in Equation (7) as:

$$0 < t < \frac{Max-Min}{N} \quad (7)$$

In this research we heuristically determine the best quantization step-size to achieve the tolerable ratio based on a set of training video clips.

By using only a single set of quantization levels, two feature values that are very similar to each other may be mapped to different quantization bins, if they are located at or near the edge of decision level for quantization. It means that very similar features may be mapped to different characters. This cause the performance of LCS based video matching to be degraded. To overcome this problem we propose a method using paired quantizer, Q_1, Q_2.

In this scheme, a feature for a frame is expressed as a pair of characters from the paired quantizer. We set the step size of the two quantizers to be the same but with different decision level. Decision level of one quantizer Q_1 is different from the other quantizer Q_2 by an off-set of $\Delta S/2$. Here ΔS denotes the step size of paired quantizer. Thus the range of *k*-th bin of one quantizer, Q_1, is expressed as in Equation (8) and the *k*-th bin of the other quantizer, Q_2 is expressed as shown in Equation (9).

$$(k-1)\cdot\Delta S \le R_{k\ of\ Q_1} < k\cdot\Delta S \quad (8)$$

$$(k-1)\cdot\Delta S + \frac{\Delta S}{2} \le R_{k\ fQ_2} < k\cdot\Delta + \frac{\Delta S}{2} \quad (9)$$

During retrieval, we compare the pair of characters to determine whether frames to be compared are matched. If at least one of the two characters is the same, we say that two frames compared are matched. Our scheme guarantees that if the distance between two feature values is smaller than $\Delta s/2$, they will be a match between them.

3.3 Similarity measure

Edit distance and longest common subsequence are two similarity measures commonly used to match two strings. Both measures can be computed by a kind of dynamic programming algorithm. The difference between the two measures is that edit distance consider “switches” between adjacent characters, whereas LCS do not.

As video has temporally unidirectional characteristics, LCS seems to be a natural similarity measure to use between video clips for video retrieval. In our approach, we apply LCS to video matching adaptively. In other words we consider continuity of matched common sub-sequences resulting from the LCS measure. In general video is composed of a sequence of shots, where each shot is a physical unit of one camera operation such as a cut, pan, and zoom in/out. Therefore the frames in a shot exhibit visually similar characteristics. In order to ensure that the matched LCS sequence covers a significant portion of the consecutive sequence of the video to be matched, we ignore those matched common sub-sequence whose continued lengths are too short.

Let the two strings to be matched be s_q and s_d, where s_q represents the query string and s_d represents the string in the database. Let s_d be composed of *n* characters. If the *k*-th character in s_d is matched with a character in s_q, we mark it as “1”; otherwise we mark it as “0” as shown in Equation (10):

$$m(k) = \begin{cases} 1, & if\ k-th\ character\ is\ matched \\ 0, & if\ k-th\ character\ is\ unmatched \end{cases} \quad (10)$$

The length of LCS is expressed as shown in Equation (11):

$$LCS = \sum_{k=1}^{n} m(k) \quad (11)$$

We denote the length of continuous matched accumulative character sequence as *LC*. It is expressed by using *m(k)* as shown in Equation (12):

$$LC(k) = \begin{cases} LC(k-1)+m(k), & if\ m(k)=1 \\ m(k), & if\ m(k)=0 \end{cases} \quad (12)$$
$$k = 1, 2, ..., n$$

The length of the continuously matched sub-string separate from unmatched strings, SLC, is expressed as shown in Equation (13):

$$SLC(k) = \begin{cases} LC(k), & if \ m(k+1) = 0 \\ 0, & if \ m(k+1) = 1 \end{cases} \quad (13)$$

$$k = 1, 2, ..., n-1$$

Therefore a length of LCS is re-expressed by means of SLC as:

$$LCS = \sum_{k=1}^{N} SLC(k) \quad (14)$$

In our approach, we do not count the matched common sub-sequence if the continuous length of a matched common sub-sequence is smaller than the threshold (*th)* which is fixed empirically. Let it be CLCS (Compacted-LCS):

$$CLCS = \sum_{k=1}^{n} MLC(k) \begin{Bmatrix} MLC(k) = SLC(k), \ if \ SLC(k) > th \\ MLC(k) = 0, if \ SLC(k) < th \end{Bmatrix} \quad (15)$$

In Equation (15), the threshold is decided in proportion to the length of database string as shown in Equation (16).

$$th = k \cdot | s_d |, \quad 0 < k < 1 \quad (16)$$

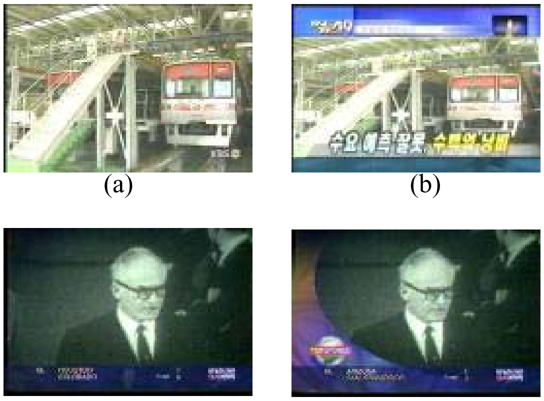
This CLCS measure is used as the measure for video matching.

For comparison purpose we also derive the corresponding measure based on the popularly used edit-distance measure for matching string similarity.

4. Experimental results

We test the effectiveness of our video matching algorithm by using new video selected form CNN and KBS (Korean Broadcasting System). Our test video set consists of 27 minutes of news video from CNN with 277 shots; and 37 minutes of news video from KBS with 626 shots. We choose 4 video clips from headline section and use that as test queries to retrieve similar video clips in the main news sections.

Figures 2 (a)-(b) show examples of key-frames of test video clips from KBS with some spatial variations, while Figures 2(c)-(d) give the corresponding examples from CNN. In general the video clips from the headline news contain some special spatial effects as shown in Figures 2(b)-(d). The length of video clips represented in Figure 2(b) or (d) broadcast in the headline news is much shorter than those broadcast in the main news section (Figure 2(a) or (c)). In other words, these video clips are examples with different temporal durations and spatial variations.



(a) (b)

(c) (d)

Figure 2. Example images of test video

In general, there are two methods to measure the similarity between two strings. One is LCS and the other is edit distance (ED). The difference of two measures is that LCS does not permit the exchange of neighboring characters while ED permits some exchanges of neighboring characters in the process of computing the similarity. We compare the performance of the two measures as presented in Figure 3. The Figure shows the mean average precision (MAP). On average, the result of LCS is 2.8% higher than that of ED in terms of MAP. LCS, which does not allow the switching of neighboring characters, is more suitable to video matching because video has stronger temporal coherence and continuity.

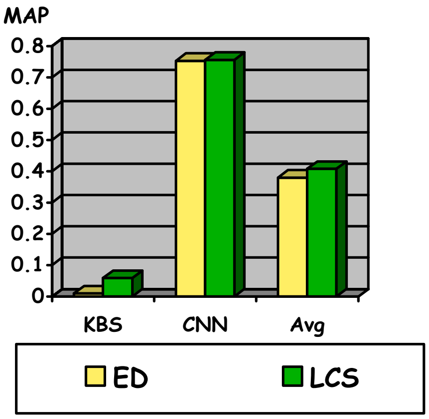


Figure 3. Performance comparison of ED and LCS

We compare the proposed algorithm with an existing algorithm [14]. The existing algorithm was developed to find video clip with different temporal duration and order. It focuses on measuring some temporal variations without spatial variations, and thus it does not tolerate spatial variations very well.

In our approach, we try to find a suitable quantization step size in order to tolerate some spatial variations from the sample set. A suitable step size is found by empirically doubling the step size of the existing quantizer until the "best" retrieval performance is achieved. The step sizes of quantization for the two algorithms are listed in Table 1.

Table 1. Step-size of quantization

	Avg.	Var.	Edge	ACs
The existing method [14].	5	3	3	-
The proposed Method	10	6	-	10

Based on the selected quantization step size, the performance of the algorithms is shown Figure 4 in terms of mean average precision. We carry out several sets of test to evaluate the performance of the algorithms with variations in parameter. Figure 4(a) shows the results of the existing algorithm, while Figure 4(b) gives the results of using our proposed approach. The results show that our proposed approach is able to improve the performance by about 4.3% over the existing method (Figure 4(a)).

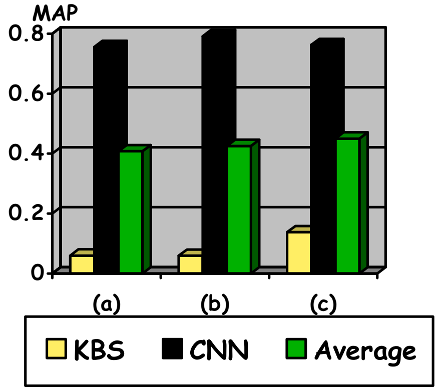


Figure 4. The experimental results
(a) Existing algorithm (b) Modified step size
(c) Paired quantizer

Next, we test the method using paired quantizer. The results are presented in Figure 4(c), which shows an increase in performance by 10.3% over Figure 4(a). Although the method is slightly more complex and requires more storage space, it provides better results on average. This is because when the range of feature values for frames composing a shot goes across the decision level of quantizer, it causes significant misleading results. By using this scheme, we can eliminate the possibility of this kind of errors.

In Figure 5, we evaluate the effect of using new feature (Figure 5(a)) and CLCS as a similarity measure (Figure 5(b)) and both of them (Figure 5(c)). We found that the use of "suitable" quantization step size is one of the major factors that influence the performance of using new feature and CLCS. Thus when we apply the new feature, CLCS and both of them, the quantization with a tuned step size is used. So we compare performance of these algorithms with that of Figure 4(b) using tuned quantizer.

In Figure 5(a), we employ DC and the 3 low frequency AC coefficients as basis to extract new features instead of using the edge feature in [14]. We employ a similarity measure given in Equation (15) to combine the new feature values. In the process, between AC (0, 1) and AC (1, 0), we select the feature with the higher similarity because one of them may be dissimilar due to spatial variations. The weights are the same for all the features used.

$$Sim(S_q, S_d) = \{ \sum_{i \in \{DC, Var(DC), AC(1,1)\}} Sim_i (S_q, S_d) + \\ Max(Sim_{AC(0,1)}(S_q, S_d), Sim_{AC(1,0)}(S_q, S_d))\} / 4 \quad (15)$$

Its performance is presented in Figure 5(a). It results in a further increase in performance of 26.7% as compared to Figure 4 (b).

In Figure 5(b), we present the result of using the compacted LCS measure to better reflect the continuity characteristics of video in the similarity measure as discussed in Section 3.3. The results show an increase in performance of 24.2% as compared to Figure 4(b).

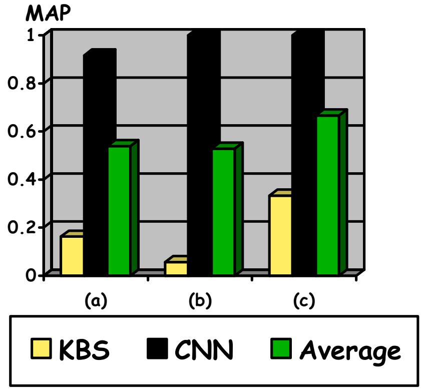


Figure 5. The experimental results
(a) New features. (b) Compacted LCS
(c) New quantizer and Compacted LCS

In Figure 5(c), we represent the results of combining the new features and compacted-LCS simultaneously. The resulting increase in performance is 56.7 % as compared to Figure 4(b). In other words, three methods we propose are together applied in Figure 5(c). Finally when it is compared to the existing algorithm of Figure 4(a), performance of the proposed algorithm is increased by 63.5%.

In case of CNN news, the effect of compacted LCS and quantizer with tuned step-size is significant while the effect of using new feature is insignificant. In case of KBS news, we found that the effect of using new feature and compacted-LCS is significant rather than that of using tuned quantizer. When we apply the proposed algorithm to KBS, we can find out that there is synergy effect by applying compacted-LCS and new features simultaneously. On average, the performance of KBS news is lower than that of CNN. This is because KBS news has relatively higher spatial variations to that of CNN. Thus further research needs to be done to improve the performance for video clips that have relatively high spatial variations.

5. Conclusions

We proposed new algorithm to find visually similar video clips with different temporal durations and spatial variations. Our contributions are as follows. First, in the feature extraction process we developed a new set of features based on DC and a few low-frequency AC coefficients in the MPEG compressed domain. The new features provide more precise representation of the main component of the image. Second, we adjusted the quantization step size of each feature space in order to find a step size that best tolerates spatial variations. A suitable step-size was determined heuristically. In the quantization step, we proposed the method using paired quantizer. In this scheme, it can eliminate significant misleading cause that happens when the range of feature values for frames composing a shot go across the decision level of quantizer. Third, in the process of measuring similarity between video clips, we considered the compactness and/or continuity of matched common sub-sequence to reflect the temporal characteristics of video. As a result, the performance of the proposed algorithms is increased by 38.4% in terms of precision as compared to the existing algorithm.

We employ this algorithm to retrieve news video. Although algorithms using text such as embedded text, closed caption may be popularly used to retrieve news video, users may sometimes want to find visually similar video clips. Interesting scenes are often broadcast repeatedly with different temporal and/or spatial editions for several days. In such cases our proposed approach can be a useful tool for news video retrieval. Another advantage of using visual features as compared to the text matching is that it can be applied to all the news independent to the language used. Finally we expect the fusion of video matching using the proposed algorithm and text matching to be an effective approach for news video retrieval. This, together with news story detection and tracking, will be the next theme of our research.

6. References

[1] T. S. Chua and L. Q. Ruan, "A Video Retrieval and Sequencing System," *ACM Trans. Information Syst*., vol. 13, no. 4, pp. 373~407, Oct. 1995.

[2] Y. A. Aslandogan and C. T. Yu, "Techniques and Systems for Image and Video Retrieval," *IEEE Trans. Knowledge and Data Eng*., vol. 11, no. 1, pp. 56~63, Jan. 1999.

[3] H. S. Chang, S. Sull, and S. U. Lee, "Effective Video Indexing Scheme for Content-based Retrieval," *IEEE Trans. Circuits Syst. Video Technol*., vol. 9, no. 8, pp. 1269-1279, Dec. 1999.

[4] J. Magalhaes and F. Pereira, "Using MPEG Standards for Multimedia Customization," *Signal Processing: Image commu.*, pp. 1-20, Elsevier, 2004.

[5] C-W. Lin, et. al., "MPEG Video Streaming with VCR Functionality," *IEEE Trans. Circuits Syst. Video Technol*., vol. 11, no. 3, pp. 415-425, Mar. 2001.

[6] TV-Anytime Specification Series: S-3 on Metadata (Normative), SP003v1.1, Aug. 2001.

[7] K. Kashino, T. Kurozumi, and H. Murase, "A Quick Search Method for Audio and Video Signals Based on Histogram Pruning," *IEEE Trans. on Multimedia*, vol. 5, no. 3, pp. 348-357, Sep. 2003.

[8] T. Lin, C-W. Ngo, H-J. Zhang, and Q-Y Shi, "Integrating Color and Spatial Features for Content-based Video Retrieval," in Proc. *IEEE ICIP'01*, Oct. 2001, vol. 3, pp. 592-595.

[9] S. H. Kim and R.-H. Park, "An Efficient Algorithm for Video Sequence Matching Using the Modified Hausdroff Distance and the Directed Divergence," *IEEE Trans. Circuits Syst. Video Technol*., vol. 12, no. 7, pp. 592-596, July 2002.

[10] X. Liu, Y. Zhuang, and Y. Pan, "A New Approach to Retrieve Video by Example Video Clip," in Proc. *ACM Mutimedia'99*, pp. 41-44.

[11] D. A. Adjeroh, M. C. Lee, and I. King, "A Distance Measure for Video Sequences Similarity Matching," in Proc. *Multi-Media Database Management Systems*, Aug. 1998, pp. 72-79.

[12] S. Santini and R. Jain, "Similarity Measures," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, no. 9, pp. 871-883, Sep. 1999.

[13] W. You, et. al., "Content-based Video Retrieval by Indexing Object's Motion Trajectory," in Proc. *ICCE'01*, June 2001, pp. 352-353.

[14] L. Chen and T. S. Chua, "A Match and Tiling Approach to Content-based Video Retrieval," in Proc. *IEEE ICME'01*, Aug. 2001, pp. 301-304.

[15] R. Mohan, "Video Sequence Matching," in Proc. *IEEE ICASSP'98*, May 1998, vol. 6, pp. 3697-3700.

[16] Y. T. Kim, et. al., "Content-based News Video Retrieval with Closed Captions and Time Alignment," in Proc. *IEEE PCM'01*, Oct. 2001.

[17] M. Bertini, A. D. Bimbo, and P. Pala, "Content Based Annotation and Retrieval of News Videos," in Proc. *IEEE ICME'00*, July 2000, pp. 483-486.

[18] Q. Huang, et. al., "Automated Generation of News Content Hierarchy by Integrating Audio, Video, and Text Information," in Proc. *IEEE ICASSP'99*, Mar. 1999, vol. 6, pp. 3025-3028.

[19] T. S. Chua, M. Kankanhalli, and Y. Lin, "A General Framework for Video Segmentation Based on Temporal Multi-resolution Analysis," in Proc. *Int'l Workshop on Advanced Image Technology, 2000*, pp. 119-124.

[20] J. W. Kim and S. U. Lee, "A Transform Domain Classified Vector Quantizer for Image Coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 2, pp. 3-14, Mar. 1992.