A new approach for content-based video retrieval

Nac-Woo Kim
Optical Communication Research Center
ETRI, Gwangju, 500-480, Korea

Byung-Tak Lee
Optical Communication Research Center
ETRI, Gwangju, 500-480, Korea

Jai-Sang Koh
Optical Communication Research Center
ETRI, Gwangju, 500-480, Korea

Ho-Young Song
Network Research Department, Broadcasting & Telecommunications Convergence Research Laboratory
ETRI, Daejeon, 305-700, Korea

ABSTRACT

In this paper, we propose a new approach for content-based video retrieval using non-parametric based motion classification in the shot-based video indexing structure. Our system proposed in this paper has supported the real-time video retrieval using spatio-temporal feature comparison by measuring the similarity between visual features and between motion features, respectively, after extracting representative frame and non-parametric motion information from shot-based video clips segmented by scene change detection method. The extraction of non-parametric based motion features, after the normalized motion vectors are created from an MPEG-compressed stream, is effectively fulfilled by discretizing each normalized motion vector into various angle bins, and by considering the mean, variance, and direction of motion vectors in these bins. To obtain visual feature in representative frame, we use the edge-based spatial descriptor. Experimental results show that our approach is superior to conventional methods with regard to the performance for video indexing and retrieval.

Keyword: *motion classification, representative frame, video retrieval, non-parametric*

1. INTRODUCTION

Recently, interest in multimedia information with the popularization of the internet user and development of network technology has largely increased. But it is still so difficult to retrieve wanted information from enormous sources stored in remote places. So, to easily utilize a retrieval service for multimedia information, an effective and flexible video retrieval system is strongly needed.

In this paper, after obtaining the representative frame (R-frame) and motion information (M-info) in the video sequence by shot-based video analysis, we propose a method that embodies an effective content-based video retrieval system by applying the acquired R-frame and M-info to an edge-based spatial descriptor and proposed non-parametric based motion classification method. First, our method parses the exact video sequence with a shot change detection algorithm, and extracts R-frame and M-info to present substances of segmented shots. To detect the exact shot boundary, we use the effective shot change detection method proposed in [1]; it also uses the reconstruction method of motion-compensated DC images of the DCT DC coefficient to realize significant computation savings by operating directly on compressed data with a partial decoding process. Then, our approach extracts the edge-based visual features from R-frames, and obtains shot-based camera motions by applying non-parametric based motion classification algorithms to the M-info. The extracted information is indexed, and finally we fulfill the development of an effective video retrieval system supporting similarity measurement between feature vectors.

* *Corresponding author. E-mail : nwkim@etri.re.kr*
Manuscript received May. 19, 2008 ; accepted Jun. 23, 2008

2. MOTION CLASSIFICATION METHOD

The acquisition of motion features in videos is generally performed by the feature analysis of the MVs pattern in each frame, and so, using this result, we can estimate the camera movement in a frame. The camera operation consists of an axis-fixed type, such as panning, tilting, and zooming; and an axis-free type, such as tracking, booming, and dollying. In this paper, we analyze such camera motions for each frame on videos and index them by shot unit, that is, a bundle of frames that have successively the same camera operation.

Due to similarities in MV patterns, it is not easy to distinguish between panning in the axis-fixed camera operation and tracking in the axis-free camera operation. So, we regard mutually corresponding camera operations-both an axis-fixed and axis-free camera operation-as one motion pattern. Accordingly, various camera operations and symbols with regard to corresponding camera operations are shown in Table I. Here, we briefly represent various camera operations by the use of simple symbols.

Table 1. Various camera operations and symbols.

Camera motion		Symbol	
Stationary		S	
Pan (Track)	Pan_left	P	P_L
	Pan_right		P_R
Tilt (Boom)	Tilt_up	T	T_U
	Tilt_down		T_D
Zoom (Dolly)	Zoom_in	Z	Z_I
	Zoom_out		Z_O
Rotation	Rotation	R	R

2.1 Proposed non-parametric based motion descriptor

To analyze a camera movement in each frame, many research studies make use of the parametric-based motion classification method[2]. But, because these parametric-based motion classification methods have still some problem, to solve such problems, we propose a non-parametric based motion descriptor.

First, before estimating the motion type for each frame, the procedure that converts MVs for a macroblock in an MPEG-compressed domain to a uniform set, independent of the frame type and the direction of prediction, has to be preceded. We normalize MVs in each coded-frame by using our motion flow estimation method proposed in [3]. The noise of normalized MVs is removed by peer group filtering[4], and whether the frame is a stationary frame or not, it is decided by the number of effective normalized-MVs[3]. In the next step, we quantize the angle of each MV in a frame into several bins (B_T). From Eq. (1), we obtain the angle of MVs at (x, y) position in i^{th} frame, and quantize the angle by the maximum bin value, B_T, as shown in Eq. (2). Here, $\tilde{\alpha}$ is the quantization value for α.

$$\alpha(x,y)^i = tan^{-1}\left(\frac{u(x,y)^i}{v(x,y)^i}\right) \quad (1)$$

$$1 \le \tilde{\alpha}(x,y)^i \le B_T \quad (2)$$

Let us consider a histogram for angle bins, $H^i(k), k \in \{1,2,...,B_T\}$. We can obtain an average, E^i and a variance, σ^i for the angle of MVs at i^{th} frame for the histogram H from Eq. (3) and Eq. (4).

$$E^i = \frac{1}{B_T}\sum_{k=1}^{B_T} H^i(k) \quad (3)$$

$$\sigma^i = \left(\frac{1}{B_T}\sum_{k=1}^{B_T}\left(E^i - H^i(k)\right)^2\right)^{1/2} \quad (4)$$

Here, we classify the input frame as Z/R or P/T frame on the basis of σ^i derived from Eq. (4). In case of Z or R, $H^i(k)$ has a uniform value over all bins, but in the case of P or T, $H^i(k)$ has a very large value in any one bin of all bins.

2.1.1 Motion classification method in Z or R frame

The frame recognized as Z or R is again classified as one of three motion frames, Z_I, Z_O or R. Let us consider the Z frame first. The Z frame has generally a target object around the center point of the frame. Because we have to consider a correlation between an object center point (OCP) and MVs placed on the outside position of the object, we extract an OCP before classifying the frame in detail. The algorithm for extraction of the OCP is as follows. First, a 3x3 sliding window mask covers the normalized MV field with peer group filtering, as mentioned above. So, because the MV size of the OCP in the Z or P shot has to naturally have a very small value, if a maximum value of a set of MV within a 3x3 window mask area, $\{\overrightarrow{MV_1}, \overrightarrow{MV_2}, ..., \overrightarrow{MV_9}\}$, does not exceed the threshold value, $Thres\_D$, we assign the center point of the mask as the possible point of OCP, $\hat{R}$. The numerical formula is expressed in Eq. (5).

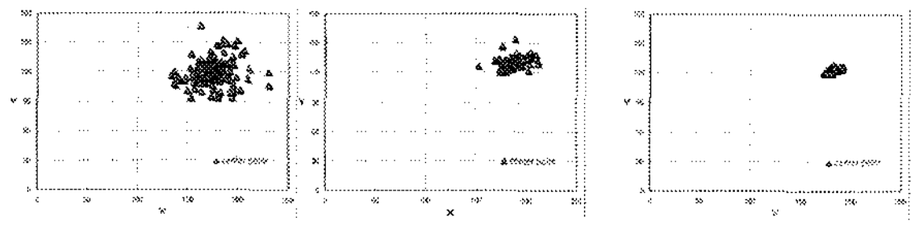
$$\hat{R} = \begin{cases} 1 & if \ \underset{i=1\sim 9}{Max}\left\lfloor \overrightarrow{MV_i} \right\rfloor < Thres\_D \\ 0 & otherwise \end{cases} \quad (5)$$

Then, we apply an 8-neighbor labeling algorithm to the points of $\hat{R}$, extracted by Eq. (5), and select one of several labeled regions as the center region. The center region is selected by criterion order of a bigger region, a closer region from the image center, and a lower elongatedness. Fig. 1 shows the result to indicate OCPs extracted from successive frames identified as Z shot, after $Thres\_D$ by 1, 3, 5 is applied to the video sequence with zoom, respectively. As shown in Fig. 1, we know that OCP in the Z frame is accurately extracted when $Thres\_D$ is 1.

Using the extracted OCP, in the next step, we classify the Z frame in two frame types. By calculating the angle (θ_z) between the extracted OCP and outside MVs from center region in the Z frame, we can distinguish whether the frame is Z_O or Z_I (see Fig. 2). The MVs in the Z_I frame head outward on the basis of the OCP, and the MVs in the Z_O frame head inward, contrarily. Accordingly, the angle, θ_z, is closer to 0 or 180 degrees. If most of the MVs in a frame exist in $|\theta_z| \le \frac{\pi}{6}$, the frame is classified as Z_O, and if most of the

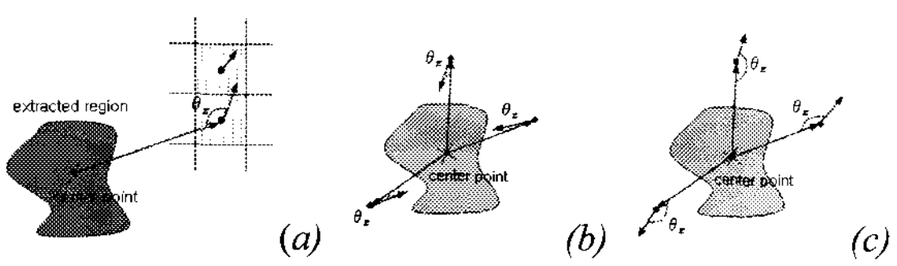
MVs in a frame exist in $\frac{5}{6}\pi \le |\theta_z| \le \pi$, the frame is classified as Z_I, tentatively. To reconfirm the classified Z_I, Z_O frames, we consider a moving direction of MV belonging to each quadrant.

The R frame is computed by an angle (θ_r) between a center point of the frame, not OCP, and MVs placed on the outside position of the frame. In this case, the angle, θ_r, is closer to $\left|\frac{\pi}{2}\right|$. So, we discriminate it as R frame when most of the MVs in a frame exist in $\frac{5}{12}\pi \le |\theta_r| \le \frac{7}{12}\pi$.



(a) Thres_D =5 (b) Thres_D =3 (c) Thres_D =1

Fig. 1. Extraction of OCP by the variation *Thres_D* in Z shot.



(a) (b) (c)

Fig. 2. Discrimination of Z frame: (a) the calculation of θ_z (b) θ_z in Z_O (c) θ_z in Z_I.

2.1.2. Motion classification method in P or T frame

A motion classification in P or T is very simple. From the preprocessing method using Eq. (4), we divide again the MVs quantized by B_T bins into four directional bins by θ_t for the frame classified as P/T frame, and estimate P_R, P_L, T_D, T_U frame from bin n ($n \in m$) satisfying Eq. (6).

$$H(n) > \left(\sum_{m=0}^{3} H(m)\right) - H(n) \tag{6}$$

3. FEATURE EXTRACTION AND MEASUREMENT

We define the frame with the least motion within a shot duration as R-frame, and apply an edge-based spatial descriptor (ESD), proposed in [5], as a descriptor of the visual feature extracted in the R-frame. ESD is an effective visual feature descriptor using an edge correlogram[6] and color coherence vector (CCV)[7]. To lessen the effect of illumination, we perform beforehand the task of classifying the pixels into smooth or edge pixels in the R-frame by using a pixel classification system based on a color vector angle, after applying a 3 x 3 window to every pixel on the whole image, and detect a color edge by using the center pixel and its neighboring pixel making the maximum color vector angle. If the center pixel in a 3 x 3 window is an edge pixel, the global distribution of the gray pairs in the edges is represented by the edge correlogram based on colors quantized in the *RGB* color space. Conversely, if the center pixel in a 3 x 3 window is a smooth pixel, the color distribution is represented by CCV. The augmented feature map, which consists of both the edge correlogram and CCV, is then used as the ESD. Since the edge correlogram uses edge pixels, it can effectively represent the adjacency between colors in an image and provide robustness to substantial appearance changes. The CCV method in the spatial area can also effectively represent the global color distribution of smooth pixels in an image. The segmented shot has one of eight motion indexes, that is {S, P_L, P_R, T_U, T_D, Z_I, Z_D, R}, from our motion classification method proposed in Section II, and we define it as the motion feature for the shot.

To index feature descriptors, we use the R*-trees structure to largely improve the effectiveness of other R-tree variants.

4. SIMULATION RESULTS

In this chapter, we construct a DB for simulation from various kinds of video sequence compressed by MPEG, and evaluate the performance of the proposed retrieval method. A simulation DB consists of various videos, such as natural clips, drama clips, music videos, and educational videos etc, and has about 780 shots acquired from 24 videos. Fig. 3 shows our retrieval system.

Table 2. Comparison of motion classification method.

method	precision	recall	FET (ms)
parametric-based [2]	0.44	0.26	0.254
proposed	0.49	0.51	0.113

The retrieval accuracy is measured in terms of the recall, precision, and ANMRR. Note that Recall, Precision, and ANMRR will always be in the [0.0~1.0] range-the higher for the precision and recall, the better; the lower for the ANMRR, the better.

Table 2 shows a simulation result that estimates the performance of the motion classification algorithm. The conventional parametric-based method in contradiction to our approach has a very low detection performance for the Z, R frame, and also, it is difficult to define a correct threshold value for the extraction of the P, T frame. In particular, the Recall value is so low by reason that many Z, R frames are not retrieved. The proposed method, contrary to the conventional method addressed in this paper, is capable of more diverse motion classifications and more accurate motion measurements. Nevertheless, because of the false detection problem in the retrieval of the R frame, the Precision value in our method is not so high. Table II shows the comparison of the feature extraction time (FET) for each method. Because a parametric-based classification method uses a linear least square method, it needs much more FET than ours.

Finally, by combining the proposed motion classification algorithm with other visual feature descriptors, we compare

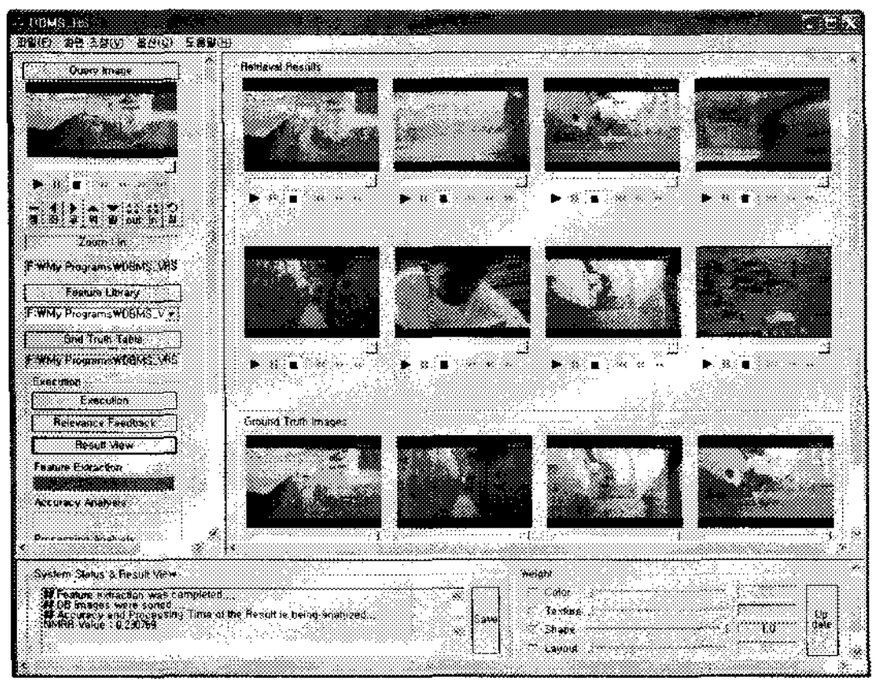


Fig. 3. Video retrieval system.

the mutual performance of the proposed video retrieval system in Table 3. As shown in Table 3, combining the proposed non-parametric based motion descriptor with the ESD visual feature rather than a feature descriptor, like CCV or a correlogram, represents improved performance in video retrieval. This is because the ESD method divides the R-frame into two parts by the pixel's frequency and applies the CCV and the correlogram to the smooth area and the edge area, respectively. Since the histogram for the spatial area in a natural image is generally not consistent with that of the edge area, considering such a gap in the areas, the retrieval performance can be largely improved by splitting the feature extraction method into the CCV method in the spatial domain and the correlogram method in the edge domain. Fig. 4 shows the retrieval results for a query and its relevant video clips, including a camera zoom and change of viewing position. Our experiment in Fig. 5 compares retrieval performance based on Recall and Precision. Fig. 5 shows the average value for an overall performance comparison. The graph of effectiveness measurement shown in Fig. 5 indicates that our method using the ESD method for video retrieval is superior to the method using the CCV or correlogram.

5. CONCLUSIONS

In this paper, we propose an effective video retrieval method in an MPEG-compressed stream using spatio-temporal features obtained by the shot change detection and motion classification method. We extract visual features by using the R-frame and classify motion features by the extraction of moving objects in shots segmented by the shot change detection method.

In this analysis, the ESD method is used as a visual feature descriptor. In addition, a non-parametric based motion classification method is used as the motion feature descriptor. After applying the motion analysis to normalize the MVs in the MPEG-compressed domain, our non-parametric based motion classification method extracts a center point of a moving object and an angle between the normalized MVs, and classifies the motion feature for the shot extracted by using a method like discretizing MVs into various angle bins, etc, as eight motion indexes. The simulation results show that the proposed method is very effective in point of video indexing and retrieval.

Table 3. Comparison of content-based retrieval method using various feature descriptors.

method	precision	recall	ANMRR	FET (ms)
CCV	0.32	0.78	0.33	27
correlogram	0.56	0.80	0.28	42
ESD	0.77	0.81	0.22	53

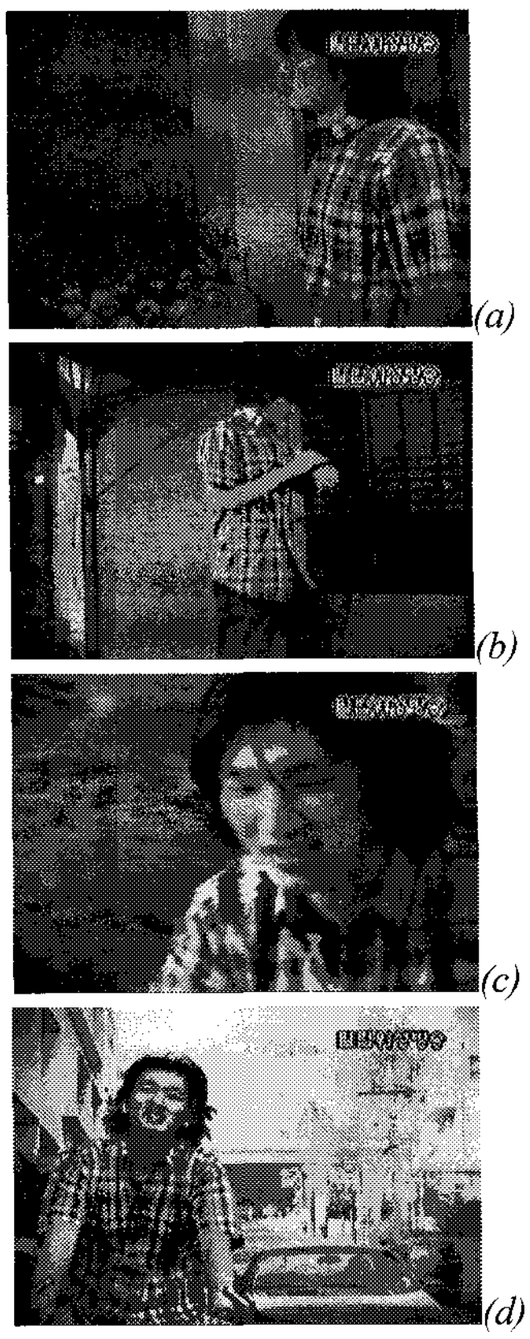


Fig. 4. Retrieval results with rank: (a) Correlogram (CG) : rank 1, CCV (CV): rank 1, Proposed (P): rank 1 (b) CG : rank 3, CV: rank 7, P: rank 3 (c) CG: rank 5, CV: rank 4, P: rank 5 (d) CG: rank 12, CV: rank 11, P: rank 6.

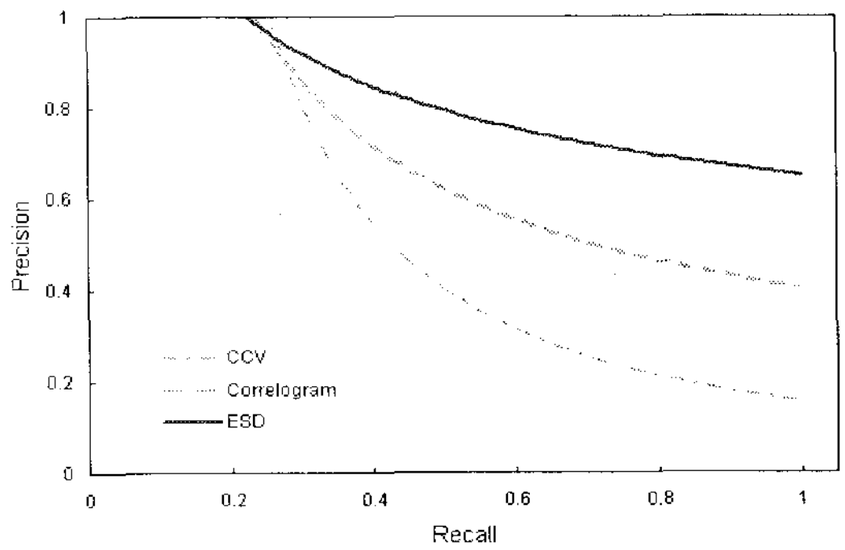


Fig. 5. Recall and precision.

REFERENCE

[1] N.W. Kim, E.K. Kang, et al., *"Scene change detection and classification algorithm on compressed video streams,"* Proc. of the ITC-CSCC 2001, vol. 1, 2001, pp. 279-282.

[2] R. Wang R., T. Huang, *"Fast camera motion analysis in MPEG domain,"* International Conference on Image Processing, vol. 3, 1999, pp. 691-694.
[3] N.W. Kim, T.Y. Kim, and J.S. Choi, *"Probability-based motion analysis using bi-directional prediction-independent framework in compressed domain,"* Optical engineering, vol. 44, no. 6, 067008.1-067008.13, 2005.
[4] Y. Deng, C. Kenney, M.S. Moore, and B.S. Manjunath, *"Peer group filtering and perceptual color image quantization,"* Proc. of IEEE Intl. Symposium on Circuits and Systems, vol. 4, 1999, pp. 21-24..
[5] N.W. Kim, T.Y. Kim, and J.S. Choi, *"Edge-based spatial descriptor using color vector angle for effective image retrieval,"* LNAI, vol. 3558, 2005, pp. 365-375.
[6] J. Huang, S. R. Kumar, M. Mitra, W. J. Zhu, and R. Zabih, *"Image indexing using color correlograms,"* CVPR, 1997, pp. 762-768.
[7] G. Pass and R. Zabih, *"Histogram refinement for content-based image retrieval,"* IEEE WACV, 1996, pp. 96-102.

Nac-Woo Kim
received the BS degree in control and instrumentation engineering from Chung-Ang University, Seoul, Korea, in 1997 and he received his MS degree and PhD degree in image engineering form Chung-Ang University, Seoul, Korea, in 2002 and 2006. He currently works as a senior member at ETRI. His research interests include video coding, image and video retrieval, IPTV and 3DTV.

Byung-Tak Lee
received the BS degree from YonSei University in 1992, MS and PhD degree from Korea Advanced Institute of Science and Technology (KAIST) in 1994, 2000. He worked as a principal R&D engineer at LG electronics from 2000 to 2004 in the area of 1.6Tbps long-haul DWDM system. From 2005 up to date, he has worked as a team leader at Electronics and Telecommunications Research Institute (ETRI). His current research topics include fiber-to-the-home, passive optical networks, and IPTV.

Ho-Young Song
received his B.S. degree in Computer Science from Hongik University in 1983 and M.S. degree in Computer Science from Chung Bok National University Korea in 1996. He received his PhD degree from Chungbuk University in 2008. He is currently a Principal Research Staff at ETRI (Electronics and Telecommunications Research Institute). His research interests are in the field of the FTTH Solutions and Killer Applications including PON and Personalized IPTV technologies.

Jai-Sang Koh
received his B.S. degree in Industrial Engineering from Korea University in 1980, and M.S. degree and PhD degree in Chonnam University in 1985, 1997. He is currently a Senior Vice President of Optical Communications Research Center at ETRI (Electronics and Telecommunications Research Institute).