멀티미디어 검색을 위한 shot 경계 및 대표 프레임 추출

Shot boundary Frame Detection and Key Frame Detection for Multimedia Retrieval

강대성, 김영호
Dae-Seong Kang, Young Ho Kim

요약

본 논문에서는 MPEG 비디오 스트림을 분석하여 DCT DC 계수를 추출하고 이들로 구성된 DC 이미지로부터 제안하는 robust feature를 이용하여 shot 검출을 수행한 후 각 feature들의 통계적 특성을 이용하여 스트림의 특징에 따라 weight를 부가하여 구해진 characterizing value의 시간 변화량을 구한다. 구해진 변화량의 local maxima와 local minima는 비디오 스트림에서 각각 가장 특징적인 frame과 평균적인 frame을 나타낸다. 이 순간의 shot을 구함으로서 효과적이고 빠른 시간 내에 key frame을 추출한다. 추출되어진 key frame에 대하여 원영상을 복원한 후, 색인을 위하여 다수의 parameter를 구하고, 사용자가 질의한 영상에 대해서 이들 파라메터를 구하여 key frame들과 가장 유사한 대표영상들을 검색한다. 실험결과 일반적인 방법보다 더 나은 결과를 보였고, 높은 검색율을 보였다.

Abstract

This paper suggests a new feature for shot detection, using the proposed robust feature from the DC image constructed by DCT DC coefficients in the MPEG video stream, and proposes the characterizing value that reflects the characteristic of kind of video (movie, drama, news, music video etc.). The key frames are pulled out from many frames by using the local minima and maxima of differential of the value. After original frame(not dc image) are reconstructed for key frame, indexing process is performed through computing parameters. Key frames that are similar to user's query image are retrieved through computing parameters. It is proved that the proposed methods are better than conventional method from experiments. The retrieval accuracy rate is so high in experiments.

Keywords : Shot boundary frame detection, key frame detection

Ⅰ. 서론

최근 통신 기술의 발달로 많은 정보들이 비디오 데이터로 전송 또는 저장되고 있다. 이에 많은 비디오 스트림들의 데이터베이스화를 위해 고용량의 비디오 데이터를 효과적으로 색인하고 검색[4,7]할 수 있는 기술이 연구되어 지고 있다[1-6]. 비디오 데이터를 내용기반의 검색을 하기 위해서는 먼저 비디오 데이터를 계층적으로 분할해야 한다. 그림 1은 비디오 데이터 계층 표시를 나타낸다.

비디오를 shot으로 구분하는 작업[3,6]을 비디오 분할(video segmentation)이라고 하며, 비디오 분할을 위해 장면의 전환점인 shot boundary frame을 추출하는 작업을 cut 추출이라고 한다. 비디오는 연속된 프레임의 집합이므로 연속된 장면에서는 인접한 프레임 사이의 유사성이 강하고 장면의 전환이 이루어지는 부분에서는 프레임 사이의 유사성이 상대적으로 약하다. 따라서 cut을 추출하기 위해서는 비디오 요소의 프레임간의 차이를 이용하여 그 요소의 연속성을 계산하고 불연속 지점을 cut으로 간주한다. 지금까지 cut 추출을 위한 다양한 알고리즘이 연구되었다.

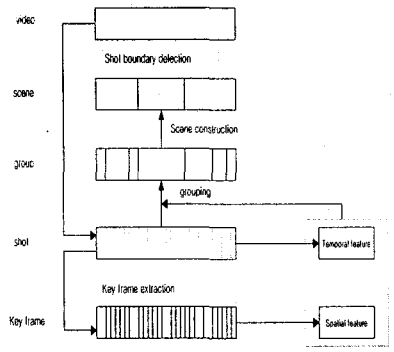

그림 1. 비디오 데이터 계층 표시
Figure 1. A hierarchical video representation

자동으로 cut을 추출하기 위한 방법으로는 히스토그램의 차이 비교, 화소간의 차이 비교, 에지 변화 비교, 압축 상관 계수 비교, 유사율 측정법, 그리고 움직임 벡터 비교[2-5] 등이 있다. 히스토그램 기반의 방법은 같은 장면으로 분류해야 할 프레임들의 색상 분포는 거의 비슷하다는 성질을 이용하여 각 프레임간의 히스토그램 차이를 계산해 정해진 임계값을 넘을 경우 cut으로 판단한다. 화소간의 차이 비교방법은 화면을 구성하는 화소들은 히스토그램과 마찬가지로 동일한 장면 내에서 변화가 적다는 성질을 이용하여 각 프레임의 화소들을 비교해 차이가 임계값을 초과하면 그 프레임간에는 장면 전환이 있다고 본다. 유사율 측정법은 개개의 화소를 비교하는 것이 아니라 연속된 프레임들에서 대응하는 일정 영역의 통계치를 비교하는 방법이며, 에지 기반의 방법에서는 주요 성분 에지를 파악하고 프레임간에 에지 단위의 비교를 수행함으로써 특정 객체가 다음 프레임에 포함되어 있는지 또는 변화가 어떻게 이루어지는지를 판단한다. 본 논문에서는 shot boundary frame 추출 즉, cut 추출을 위해 4개의 feature를 조합하여 사용하였다. 이전 영상과의 pixel간의 차분값, 히스토그램에 대한 chi-square값, 히스토그램 분산의 차분값, 새로이 제안하는 양자화한 영상의 히스토그램의 각 bin 값들의 열과 행의 위치에 대한 분산의 chi-square값이다. 본 논문의 구성은 다음과 같다. Ⅰ장 서론에 이어 Ⅱ장은 shot boundary 및 key frame 추출, Ⅲ장에서는 key frame 색인 및 검색을 한다. Ⅳ장에서는 실험한 결과와 이에 대한 고찰을 하였고 Ⅴ장에서 결론을 맺었다.

Ⅱ. shot 및 key frame 추출

1. MPEG 비디오 스트림에서의 DC 이미지 추출

MPEG 에서는 I(Intra), P(Predictive), B(Bidirectionally predictive) picture가 있다. I picture는 inter frame 예측을 사용하지 않고 해당 화면 정보만으로 부호화하는 화면으로서 모든 MB type은 intra frame이며, 채널 전환 시의 원 영상 복구와 오류의 전파를 막기 위해 GOP(Group of Picture)내에 최저 한 장의 I picture를 필요로 한다. P picture는 I 혹은 P picture로 부터의 순방향 움직임 보상 예측 수행으로 생기는 화면으로서 MB(Macroblock) type은 intra frame, forward inter frame, backward inter frame, interpolative inter frame 예측부호화가 수행된다. B picture가 삽입됨으로써 화면 처리 순서가 원 화면 순서와 달라져서 부호기에서는 B picture를 건너 뛰어 다음의 I, P picture를 우선 부호화하고, 그 후 사이에 있는 B picture를 부호화한다. 복호기에서는 I, P picture 사이의 B picture를 먼저 처리한 후 I와 P picture를 표시한다. I picture는 다음 예측에 사용되므로 양자화 스텝을 세밀히 하여 고화질을 유지하고 B picture는 다음 예측에 사용되지 않으므로 양자화 스텝을 크게 한다. 본 연구에서는 shot을 추출하기 위하여 MPEG에서 I picture의 계수만으로 영상을 구성한 DC image을 사용함으로서 video 스트림의 decoding과정을 생략할 수 있었으므로 처리시간을 대폭 감소시켰다.

2. Shot boundary frame 추출

shot boundary frame를 추출하기 위한 feature로서 4개의 feature[8] 조합하여 사용한다. 첫 번째 feature는 이전 영상과의 pixel간의 차분값으로서 서로 다른 영상의 유사도를 측정하는데 가장 기본이 되며, 전체적인 휘도 변화를 나타낸다. 두 번째 feature는 shot 검출을 위하여 일반적으로 사용되어 지고 있는 DC image의 히스토그램에 대한 chi-square 값이며, 이전 영상에 대한 히스토그램의 변화를 나타낸다. 세 번째 feature는 이전 DC image와 현재 DC image와의 히스토그램 분산의 차분값이다. 이 feature는 히스토그램의 전체적인 분포에 대한 변화를 나타낸다. 네 번째 feature는 본 논문에서 새로이 제안하는 feature로서 양자화한 영상의 히스토그램의 각 bin값들의 열과 행의 위치에 대한 분산의

chi-square 값이다. 이 feature는 히스토그램의 bin 값들을 이용함으로써 객체의 움직임에 강인하고, 각 bin 값들의 열과 행의 위치에 대한 분산의 chi-square 값을 구함으로서 칼라의 변화에 둔감하다. 이상과 같이 구해진 DC image의 feature들로부터 아래와 같은 단계로 shot boundary frame을 추출한다.

Step 1. 각 파라메터들의 전체 프레임에 대한 평균을 구한다.

$$\overline{DiffImg},\ \overline{X},\ \overline{DiffD},\ \overline{PX},\ \overline{PY}$$

Step 2. DC image의 각 feature 값을 실험을 통해 구한 값보다 큰 경우 shot boundary frame으로 추출한다.

(※ 각 상수 $\alpha, \beta, \gamma, \delta$는 실험을 통해서 2.0, 2.0, 2.0, 2.0으로 검출)

그림 2와 3은 shot boundary frame에 의해 구해진 shot들이다.

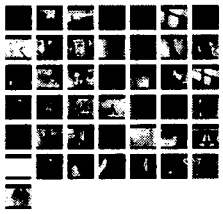

그림 2. 추출된 shot들
Figure 2. Shots of detected

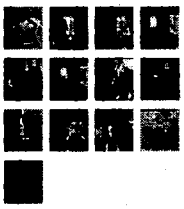

그림 3. 추출된 shot들
Figure 3. Shots of detected

3. Key frame 추출 알고리즘

key frame 추출은 구해진 shot들 중 그 video 스트림을 가장 잘 표현할 수 있는 대표 frame을 찾는 과정이다. 본 논문에서는 각 video 스트림마다 종류에 따라 다른 특징을 갖고 있고 key frame이 그 video 스트림의 특징에 따라 파라메터에 대해 민감도가 다른 것을 반영하기 위하여 전체 shot의 각 파라메터들의 통계적 특성에 따라 weight를 부가한 characterizing value를 구함으로서 보다 적합한 key frame 추출을 수행한다. 다음은 key frame 추출을 위한 shot의 파라메터들이다.

① $f_1 =$

$$AveShot_i = \frac{\sum_{x=0,y=0}^{M-1,N-1} Shot_i(x,y)}{MN}$$

휘도의 평균으로 shot 전체 pixel에 대한 휘도의 평균이다.

② $f_2 = DisShot_i^2 = E(|H_i(n) - \overline{H_i(n)}|^2)$

히스토그램의 분산값으로 shot의 히스토그램의 분포 특성을 나타낸다.

③ $f_{3x} = AveDisX_i = \frac{\sum_{k=0}^{n_1-1} \rho X_i(k)}{n_1}$

$$f_{3y} = AveDisY_i = \frac{\sum_{k=0}^{n_2-1} \rho Y_i(k)}{n_2}$$

양자화한 shot 히스토그램의 각 bin값들의 열과 행의 위치에 대한 분산의 평균으로 히스토그램과 위치정보의 조합을 나타낸다.

④ $f_4 =$

$$DiffShot_i = \frac{\sum_{x=0,y=0}^{m-1,n-1} |Shot_i(x,y) - Shot_{i-1}(x,y)|}{MN}$$

이전 shot과의 휘도 차 변화량을 나타낸다

⑤ $f_5 = AccDiffShot_i = \frac{\sum_{k=0}^{i-1} DiffShot_k}{DiffShot_i}$

휘도 차의 누적에 대한 f_4의 비율로 전체적인 휘도 변화율에 대한 상대값을 나타낸다.

위의 파라메터들을 각 검출된 shot에 대해서 구하고 난 후 key frame을 추출하기 위하여 다음과 같은 단계를 거쳐 characterizing value를 구한다.

Step 1. shot으로 추출된 모든 프레임에 대한 각 feature들의 분산을 구한다.

$$\rho_{f_n}{}^2 = E(|F_n - \overline{F_n}|^2)$$

Step 2. 각 feature들의 전체 평균에 대한 차를 구한다.

$$f_n{}' = |f_n - \overline{f_n}|$$

Step 3. 각 feature들에 대해 분산에 대한 비율만큼 weight를 부가하여 각 프레임의 특징 값 (C_m)을 구한다.

$$C_m = \omega_1 f_1{}' + \omega_2 f_2{}' + \cdots + \omega_n f_n{}'$$

여기서 m은 shot으로 추출된 프레임의 개수, 가중치 ω는 아래와 같다.

$$\omega_n = \frac{\rho_n}{\sum_{i=1}^{n} \rho_i}$$

위의 단계를 거쳐 계산되어진 characterizing value의 시간에 대한 변화량을 구하여 local maxima와 local minima를 구한다. 각 local maxima와 local minima는 그 비디오 스트림의 가장 특징적인 frame과 평균적인 frame을 나타내게 되며 이 순간의 frame을 key frame으로 추출할 수 있다. 그림 4와 5는 shot boundary frame들 중 추출된 key frame이다.

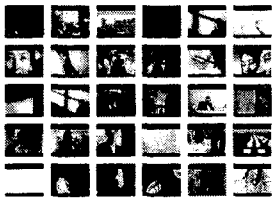

그림 4. 추출된 key frame들

Figure 4. Key frames of detected

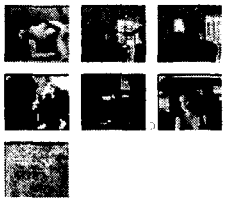

그림 5. 추출된 key frame들

Figure 5. Key frames of detected

Ⅲ. key frame 색인 및 검색

보다 정확한 색인을 위하여 key frame으로 추출된 DC image를 원 영상으로 복원한다. 그리고 key frame의 색인을 위하여 파라메터들을 구한다. 원 영상에 대한 파라메터를 구함으로 보다 정확한 색인이 가능하고, 검색 시 입력되는 질의 영상과 좀 더 객관적인 유사도 측정이 가능하다. 색인과 검색알고리즘의 단계는 다음과 같다.

Step 1. key frame으로 추출 된 frame을 원 이미지로 복원한다.

Step 2. frame의 색인을 위한 파라메터를 구한다.

① 영상의 휘도 평균(p_1)

② 히스토그램의 분산(p_2)

Step 3. Step2에서의 색인 값과 질의영상의 색인 파라메터 값의 차이가 적은 순으로 유사도 평가를 한다.

$$Sim_i = \frac{\sum_{x=0,y=0}^{M-1,N-1} |Img_{query}(x,y) - Img_i(x,y)|}{NM}$$

Step 4. Sim_i의 값이 일정 임계치(S) 이하가 될 때까지 Step 3을 반복한다.

Step 5. Sim_i값이 작은 순으로 최대 10개까지의 응답영상을 출력한다.

그림 6은 질의 영상과 그에 대한 응답 영상이다.

그림 6. 질의영상과 응답영상

Figure 6. Query image and Retrieval image

Ⅳ. 실험 및 고찰

본 연구에서는 테스트 영상으로 384 × 288 크기를 갖는 1618 프레임의 뮤직비디오와 746 프레임의 영화를 사용하였다. 표 1은 shot 검출 결과를 나타낸다.

표 1. shot 검출 결과

Table 1. Experiment results of shot detection

	correct	miss	extra detection	total
Music video f1, f2	28 / 39 (72%)	11	2	41
Music video f1, f2, f3, f4	34 / 39 (87%)	5	4	43
Movie f1, f2	8 / 11 (73%)	3	2	13
Movie f1, f2, f3, f4	10 / 11 (91%)	1	2	13

표1 에서 f1, f2, f3, f4는 shot boundary frame를 추출하기 위한 feature이다. correct는 사람의 지각으로 뽑은 shot과 정확히 일치하는 shot을 추출한 경우이고, miss 항목은 사람이 shot으로 지각하였으나 추출하지 못한 경우이고, extra detection 항목은 실험에서 구한 total 개수에서 correct 개수와 miss개수를 뺀 값이다. extra detection의 경우는 shot의 근처를 추출하는 경우가 대부분이라 문제가 되지 않으나, miss와 같이 shot인 부분을 추출하지 못하는 것은 다음 과정인 key frame 추출에 잘못된 결과를 초래할 수

도 있다. 표1 에서 보는 바와 같이 영화의 데이터로 실험한 경우는 shot의 경계에 비교적 특수효과가 작고 shot의 변화가 명확함으로 shot의 추출율이 높고, 뮤직비디오의 경우는 shot의 경계에 특수효과가 많음으로 추출율이 떨어지는 것을 볼 수 있다. 그러나, 전체적으로 기존의 feature를 사용하는 경우보다 제안한 feature를 사용함으로서 보다 향상된 결과를 보임을 알 수 있다.

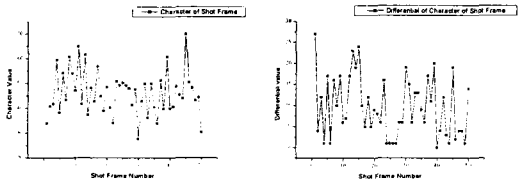

그림 7. Characterizing value의 미분값

Figure 7. Differential value of characterizing value

그림 7은 구해진 characterizing value의 시간에 대한 미분값이다. 계산되어진 characterizing value의 시간에 대한 미분값을 구하여 local maxima와 local minima를 구해보면, 각 local maxima와 local minima는 그 비디오 스트림의 가장 특징적인 frame과 평균적인 frame을 나타내고 있다. 이 순간의 frame을 key frame으로 추출하였다. 특히 28~31번까지의 shot frame의 시간에 대한 미분치가 거의 동일하다는 것을 볼 수 있을 것이다. 이것은 video stream의 중요한 부분에서 shot의 급격한 변환으로 유사한 여러 frame이 shot으로 추출됨으로써 발생한 것으로 실제 video 스트림에서 대부분 가장 대표적인 frame으로 추출될 수 있다. 표 2는 제안한 방법으로 색인하여 150개(테스트 영상 중 임의로 선택한 프레임 개수)의 영상 중 1순위에 정확한 검색이 이루어진 경우는 92%로서 신뢰도 높은 검색이 가능하다.

표 2. 검색결과

Table 2. Experiment results of retrieval

	1st out	2nd out	3rd out	others
accuracy	138/150	5/150	3/150	4

V. 결론

본 연구에서는 객체의 움직임에 강인하면서 shot내에서의 칼라의 변화에 둔감한 새로운 feature를 제안하고, 이를 이용하여 shot을 추출하였다. shot에서 구한 각 feature들의 통계적 특성을 이용하여 video 스트림의 특징에 따라 weight를 부가하는 characterizing value를 제안하고, 이를 적용하여 key frame을 추출하였다. 실험 결과 일반적으로 사용되어지는 feature와 함께 제안한 feature를 적용함으로서 보다 정확한 shot추출이 가능하였다. 또한 제안된 characterizing value의 시간에 대한 미분의 local maxima와 local minima에서 video 스트림의 중요한 대표 frame을 key frame으로 추출하였고, 특히 가장 대표되는 스트림의 중요 부분에서 뚜렷한 변별력을 보였다. 색인과 검색에 있어서는 key frame을 색인하고 질의영상 입력 시 질의영상의 색인 파라메터를 구해 파라메터 값에 따라 순차적으로 유사도를 비교함으로서 보다 빠르고 효율적인 검색이 가능하게 하였다. 실험에서는 3개의 비디오 데이터에서 shot과 key frame을 추출하고 추출된 key frame을 색인하여 데이터베이스를 구성하고 검색을 실시함으로서 검색시스템의 효율성을 실험하였다. 실험 결과 shot boundary frame추출과 key frame추출에서 제안된 알고리즘의 우수성을 증명하였고, 검색 시에 높은 검색율을 보였다. 그러나 뮤직비디오와 같은 특수효과가 많이 삽입된 비디오 데이터에 대해서는 기존의 방법보다 우수하긴 하나 만족할 만한 결과를 추출하지 못하였다.

본 연구에 유용한 여러 종류의 video 스트림을 구하기 어려워 다양한 실험을 하지 못하였다. 앞으로 여러 종류의 video 스트림에 대해 다양한 실험을 수행하여 제안한 알고리즘을 보완할 것이다.

접수일자 : 2000. 9. 10 수정완료 : 2000. 11. 7

본 논문은 정보통신부의 2000년도 대학기초연구지원사업의 지원으로 수행되었음

참고문헌

[1] K. R. Rao, J.J.Hwang. "Techniques and Standards for Image · Video and Audio Coding," Prentice Hall. 1996.

[2] Yunis S. Avrithis, Anastasios D. Doulamis, Nikolaos D. Doulamis and Stefanos D. Kollias, "A Stochastic Framework for Optimal Key Frame Extraction from MPEG Video Databases," Computer Vision and Image Understanding Vol.75, Nos. 1/2, July/August, pp. 3~24, 1999.

[3] J. Meng, Y. Juan, and S.F. Chang, "Scene change detection in a mpeg compressed video sequence," in Proc. SPIE-Digital Video Compression: Algorithms and Tech., (San Joce, CA), Feb. 1995.

[4] M. Abdel-Mottaleb and R. Desai, "Image Retrieval by Local Color Features," The 4th IEEE Symposium on Computers and Communications, Egypt, July 1999.

[5] D. Zhong, S. F. Chang, "Spatio-Temporal Video Search using the Object-Based Video Representation," Proc. ICIP'97, Vol1, pp. 21-24, Oct 1997, Santa Babara, CA.

[6] Wei Xiong and Chung-Mong Lee, "Efficient Scene Change Detection and Camera Motion Annotation for Video Classification," Computer Vision and Image Understanding Vol.71, Nos 2/2, August, pp. 166-181, 1998.

[7] Ramin Zabih, Justin Miller, Kevin Mai, "A feature-based algorithm for detecting and classifying production effects," Multimedia Systems 7 pp. 119-128, 1999.

[8] 한국 신호처리 · 시스템 학회 2000년 하계 종합 학술대회 논문집 pp. 297-300

강대성(Dae-Seong Kang)
正會員
1984년 경북대학교 전자공학과 학사.
1991년 Texas A&M Univ., Electric Eng. 석사
1994년 Texas A&M Univ., Electric Eng. 박사
1984년-1989년 국방과학연구소 연구원
1994년-1995년 한국전자통신연구소 선임연구원
1995년-현재 동아대학교 전기전자컴퓨터 공학부 조교수
관심분야 : 영상처리, 패턴인식, 영상코딩 통신시스템 등.

김영호(Young Ho Kim)
準會員
1999년 동아대학교 전자공학과 학사
2000년 동아대학교 전자공학과 석사과정 재학 중
관심분야 : 영상처리, 멀티미디어 색인 및 검색, 영상통신 등.