2단계 신경망과 계층적 프레임 탐색 방법을 이용한 MPEG 비디오 분할

(MPEG Video Segmentation using Two-stage Neural Networks and Hierarchical Frame Search)

김 주 민 † 최 영 우 †† 정 규 식 †††
(Joomin Kim) (Yeongwoo Choi) (Kusik Chung)

요 약 본 논문에서는 MPEG 비디오 데이타의 컷(cut)과 디졸브(dissolve)를 검출하여 샷(shot) 단위로 분할하고 각 샷의 카메라 동작 또는 객체 움직임의 형태를 분류하는 방법을 제안하고자 한다. 정확한 샷의 위치와 카메라, 객체의 세분화된 동작을 구별하기 위한 전단계의 연구에서[1] 우선 MPEG 데이타의 I(Intra) 프레임의 DC(Direct Current) 계수를 분석하여 픽처 그룹을 Shot(장면이 바뀐 경우), Move(카메라 동작 또는 객체가 움직인 경우), Static(영상의 변화가 거의 없는 경우)으로 세분화하여 분류하였다. 이 과정에서 2단계 구조의 신경망을 구성하고 여러 종류의 특징을 서로 다른 해상도에서 추출하여 결합시키는 방법을 제안하였다. 다음 단계로 Shot 또는 Move로 분류된 픽처 그룹의 P(Predicted), B(Bi-directional) 프레임을 선별적, 계층적으로 탐색하여 컷의 정확한 발생 위치와 카메라 동작 또는 객체 움직임의 종류를 결정하는 방법을 제안한다. P, B 프레임의 매크로 블록의 종류별 분포를 통계적으로 이용하여 컷의 발생 위치를 검출하며, P, B 프레임의 매크로 블록 종류와 움직임 벡터를 동시에 사용하는 신경망을 구성하여 디졸브, 카메라 동작, 객체 움직임의 종류를 검출한다. 본 논문에서 제안하는 방법은 MPEG 데이타의 압축을 풀지 않은 상태에서 I 프레임의 DC 계수만을 사용하여 픽처 그룹을 분류하며, 분류된 픽처 그룹 내에서 일부의 P, B 프레임만을 계층적으로 선택하여 탐색함으로서 처리 시간을 감소시키고자 하였다. 세 종류의 서로 다른 비디오 데이타를 사용한 실험에서 93.9~100.0%로 픽처 그룹을, 96.1~100.0%로 컷을 검출하였다. 또한 두 종류의 비디오 데이타를 사용한 실험에서 90.13% 및 89.28%의 정확성으로 카메라 동작 또는 객체 움직임을 분류하였다.

키워드 : MPEG, 비디오 분할, 샷 검출, 카메라 동작 검출

Abstract In this paper, we are proposing a hierarchical segmentation method that first segments the video data into units of shots by detecting cut and dissolve, and then decides types of camera operations or object movements in each shot. In our previous work[1], each picture group is divided into one of the three detailed categories, Shot(in case of scene change), Move(in case of camera operation or object movement) and Static(in case of almost no change between images), by analysing DC(Direct Current) component of I(Intra) frame. In this process, we have designed two-stage hierarchical neural network with inputs of various multiple features combined. Then, the system detects the accurate shot position, types of camera operations or object movements by searching P(Predicted), B(Bi-directional) frames of the current picture group selectively and hierarchically. Also, the statistical distributions of macro block types in P or B frames are used for the accurate detection of cut position, and another neural network with inputs of macro block types and motion vectors method can reduce the processing time by using only DC coefficients of I frames without decoding and by searching P, B frames selectively and hierarchically. The proposed method classified the picture groups in the accuracy of 93.9~100.0% and the cuts in the accuracy of 96.1~100.0% with three different

· 본 연구는 KISTEP 여자대학 연구기반 확충사업 지원에 의해 수행되었음.
† 비 회 원 : LG전자기술원 정보기술연구소 연구원
jmkim7@LG-Elte.com
†† 정 회 원 : 숙명여자대학교 정보과학부 교수
ywchoi@sookmyung.ac.kr
††† 종신회원 : 숭실대학교 정보통신전자공학부 교수
kchung@q.soongsil.ac.kr
논문접수 : 2000년 11월 22일
심사완료 : 2001년 11월 12일

together is used to detect dissolve, types of camera operations and object movements. The proposed types of video data. Also, it classified the types of camera movements or object movements in the accuracy of 90.13% and 89.28% with two different types of video data.

Key words : MPEG, Video Segmentation, Shot Detection, Camera Motion Detection

1. 서 론

최근에 멀티미디어 기술의 발달과 활용에 디지털 비디오의 유용성이 증가하고 있다. 따라서 비디오 데이타를 효율적으로 검색하고 브라우징하기 위한 소프트웨어의 개발이 필요하게 되었다. 비디오 분할은 비디오 프레임들을 샷(shot) 단위로 분할하는 것으로서, 내용 기반 툴 개발에 필요한 중요한 연구 분야이다. 샷은 급진적 장면 전환과 점진적 장면 전환으로 나뉘어 지며, 급진적 전환의 샷에는 컷(cut)이 있으며, 점진적 전환의 샷에는 디졸브(dissolve), 페이드(fade), 와이프(wipe) 등이 있다. 본 연구에서는 샷을 하나의 카메라로부터 프레임들의 연속된 시퀀스(sequence)로 정의하고 컷과 디졸브의 검출만을 고려한다.

비디오 데이타의 샷을 검출하기 위한 다양한 연구가 수행되고 있으며, 이러한 연구들은 압축되지 않은 형태의 데이타에서 샷을 검출하는 연구와[2-7] 압축된 형태의 데이타에서 샷을 검출하는 연구로[1,8-14] 나뉘어진다. 압축되지 않은 비디오 데이타를 분할하는 방법으로서 화소 단위로 분할하는 방법[2, 3], 부분 영역 단위로 분할하는 방법과[4, 5], 전체 프레임 단위로 분할하는 방법으로[6, 7] 나눌 수 있다. 이러한 방법들은 픽셀 또는 히스토그램 차이 값 등을 분할 특징으로 사용하였다. 대부분의 비디오 데이타는 MPEG 형식으로 압축되어 처리되기 때문에 압축된 상태에서 비디오 데이타를 분할하는 방법이 필요하며, 표 1에 이러한 연구를 소개하였다. Yeo[8] 등은 압축된 데이타로부터 먼저 DC 영상을 추출한 후 이전 DC 영상과의 히스토그램 분포를 비교하여 샷을 검출하였다. 이 방법은 I(Intra), P (Predicted), B(Bi-directional)의 모든 프레임으로부터 DC 성분을 계산해야 하는 단점이 있다. Zhang[9] 등은 I 프레임의 DC 계수 값만을 이용해서 샷을 검출하였지만, 적절한 임계값을 설정하는데 어려움이 있었다. Liu [10] 등은 P, B 프레임의 움직임 벡터를 이용해서 샷을 검출하였으며, 이 방법은 모든 P, B 프레임을 검사하기 때문에 처리 시간이 오래 걸리는 단점이 있다. Gamaz [11] 등이 제안한 skip 알고리즘은 먼저 I 프레임만을 이용해서 픽처 그룹(GOP: Group of Pictures)을 샷과 샷이 아닌 non-shot으로 구분하고, P와 B 프레임의 움직임 벡터를 이용하여 정확한 샷의 위치를 추정하였다. 이 방법은 I 프레임의 DC 영상에 적용할 적절한 임계값을 찾는데 어려움이 있다. Arman[12], 김가현[13], 이충훈[14]의 연구는 표 1을 참조하기 바란다.

본 논문은 MPEG 비디오 데이타를 압축된 상태에서 컷과 디졸브를 검출하여 샷 단위로 분할하고, 각 샷의 카메라 동작 또는 객체 움직임의 종류를 판별하는 방법을 제안한다. 제안하는 방법은 MPEG 데이타의 I, P, B 프레임을 계층적으로 탐색한다. 먼저 I 프레임의 DC 계수로부터 다양한 종류의 전역적, 지역적 이미지 특징을 먼저 추출하여 2단계 구조의 신경망에 통과시켜 샷이 발생한 장면, 카메라 또는 객체가 움직인 장면 및 변화가 거의 없는 장면으로 픽처 그룹을 세분화시킨다. 분류된 각 픽처 그룹의 P, B 프레임을 선택적, 계층적으로 탐색하여 컷, 디졸브, 카메라 동작 및 객체 움직임의 발생 위치를 검출하며, 매크로블록의 종류에 따른 개수와 움직임 벡터 성분을 특징으로 사용한다. 제안하는 방법은 픽처 그룹 분류 과정에서 서로 다른 관점의 다중 특징들을 결합함으로써 픽셀 차이 값 또는 히스토그램 분포 값만을 사용한 기존의 방법들보다 분할 오류를 감소시킬 수 있다. 또한 상세 분류된 픽처 그룹으로부터 선택적, 계층적으로 P, B 프레임을 탐색함으로서 처리 시간도 단축할 수 있다.

본 논문의 2장에서는 제안하는 방법의 픽처 그룹 분류 과정, 컷 검출 과정, 디졸브, 카메라 동작 및 객체 움직임 검출 과정에 대해서 설명한다. 2.1절의 픽처 그룹 분류 과정은 본 연구의 전단계 연구를 참조하여 기술한다[1]. 3장에서는 실험 과정과 결과를 분석하며, 4장에서는 결론과 향후 연구 방향을 논한다.

2. 제안하는 방법

MPEG 비디오 스트림은 픽처 그룹(GOP: Group of Pictures)들로 이루어져 있으며, 각 픽처 그룹들은 그림 1과 같이 I(Intra) 프레임으로 시작하여 다수의 P

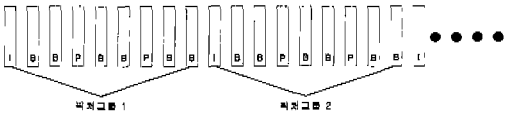

그림 1 MPEG 비디오 스트림의 구조

표 1 기존 연구 고찰

(FMB: 전방향(forward) 매크로블록, BMB: 후방향(backward) 매크로블록, IMB: 인트라(Intra) 매크로블록)

비교 항목 / 기존 연구	특 징	방 법	장, 단점	기타 사항
Yeo 외[8]	DC 계수	두 프레임 사이의 DC 영상의 픽셀 차이 값을 이용	P, B 프레임의 DC 영상을 얻기 위한 복호화 과정이 필요함	I, P, B 프레임의 DC 영상 이용
Zhang 외[9]	DCT 계수	두 프레임에서 블록 간의 계수 값의 차이를 이용	프레임만 적용 가능하여 정확한 위치 검출이 어려움	I 프레임에 대해서만 적용
Liu 외[10]	모션 벡터	-P 프레임의 FMB의 영 에너지 값 이용 -B 프레임의 FMB와 BMB 개수 차이 값을 이용	-처리 시간의 단축 급격한 카메라 동작, 빠른 객체 움직임에 민감함	P, B 프레임만 검사
Gamaz 외[11]	모션 벡터/ DC 계수	-연속된 I 프레임 사이의 DC 계수 차이 값 이용 -P 프레임에서 FMB와 IMB의 개수 차이 값 이용 -B 프레임에서 BMB와 IMB의 개수 차이 값 이용	-Skip 알고리즘으로 처리 시간 단축 -점진적인 장면 전환 검출이 어려움	픽처 그룹 단위의 검색과 픽처 그룹 내에서 정확한 발생 위치 검출
Arman 외[12]	DCT 계수	두 프레임 간 DCT 계수 값의 내적 이용	프레임만 적용 가능하여 정확한 위치 검출이 어려움	JPEG 비디오에 적용
김가현 외[13]	모션 벡터/ DC 계수	-I 프레임의 DC 계수 차이 값 이용 -P 프레임에서 IMB와 FMB 개수의 비율 이용 -B 프레임에서 BMB와 FMB 개수의 비율 이용	-I,P,B 픽처 구성에 관계없이 적용 가능 -빠른 처리 시간 -급격한 카메라 동작에 민감함	매크로블록 종류의 개수를 이용한 디졸브 검출 시도
이충훈 외[14]	DC 계수	연속된 I 프레임의 DC 영상의 픽셀 차이 값과 신경망 이용	-경험적 임계값 설정의 어려움을 해결 P, B 프레임의 DC 영상을 얻기 위한 복호화 과정 필요 점진적 장면 전환 검출이 어려움	뉴스, 광고, 영화 세 그룹으로 나누어 학습 및 테스트 수행

(Predicted) 프레임과 B(Bi-directional) 프레임으로 구성된다[1]. 본 논문에서는 각 픽처 그룹의 첫 번째 프레임인 I 프레임을 분석하여 픽처 그룹을 상세 분류하며, 해당 그룹의 P, B 프레임을 계층적으로 탐색하여 컷, 디졸브, 카메라 동작 및 객체 움직임 위치를 검출한다.

제안하는 비디오 분할 방법은 그림 2와 같이 요약된다. 1단계인 픽처 그룹 분류에서는 현재와 이전 I 프레임의 DC 영상을 비교하여 픽처 그룹을 Shot, Move, Static의 세 종류로 상세 분류한다. 픽처 그룹이 Shot으로 판정되면 해당 픽처 그룹 내의 P, B 프레임을 계층적으로 검사하여 컷 발생 위치를 정확히 검출한다. 또한, Shot에서 컷이 없거나 찾지 못한 경우 및 픽처 그룹이 Move로 판정되면 해당 픽처 그룹내의 P, B 프레임을 검사하여 디졸브, 카메라 동작, 객체 움직임을 검출한다. 이 과정에서 Shot 경계의 일종인 디졸브는 분류의 정확성을 높이기 위하여 본 논문에서는 Move의 한 종류로 설정하였다. 픽처 그룹이 Static으로 판정되면 해당 픽처 그룹 내의 모든 프레임을 Static으로 판정한다. 이와 같이 픽처 그룹을 상세 분류한 후 선별적, 계층적으로 프레임을 탐색함으로서 모든 프레임을 검사하는 기존의 방법보다 처리 시간을 단축시킬 수 있다.

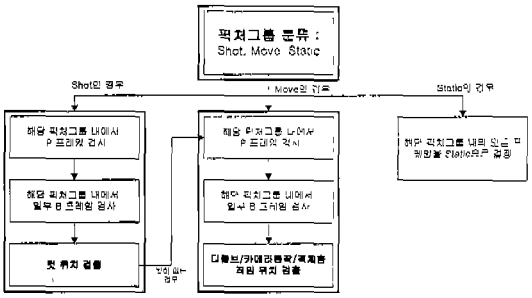

그림 2 제안하는 분할 방법

2.1 픽처 그룹 분류

비디오 데이타를 분석해 보면 장면이 바뀌는 부분, 카메라 또는 객체가 움직이는 부분과 영상의 변화가 거의 없는 부분으로 나뉘어 진다. 기존 연구에서는 장면이 바뀌는 부분과 바뀌지 않는 부분으로만 분류했지만[11], 본 연구에서는 I 프레임의 DC 영상을 Shot, Move, Static으로 세분화하여 분류한다. 장면이 바뀌는 부분은 Shot으로 정의하고, 카메라 또는 객체가 움직이는 부분은 Move로 정의하며, 영상의 변화가 거의 없는 부분은 Static으로 정의한다.

픽처 그룹을 분류하는 방법은 그림 3과 같다. 현재와 이전 I 프레임의 DC 영상에서 전역적인 이미지 특징과 지역적인 이미지 특징을 각각 추출하여 2단계 구조의 신경망에 입력한다. 전역적 특징 벡터는 1단계의 신경망인

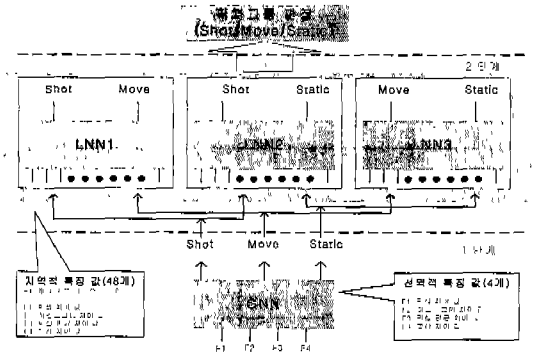

그림 3 픽처 그룹 분류를 위한 계층적 신경망 구조

GNN(Global Neural Network)에 입력되어 현재 프레임을 Shot, Move, Static 가운데 하나로 결정한다. 지역적 특징 벡터는 2단계 신경망의 LNN1(Local Neural Network), LNN2, LNN3 가운데 선택된 2개의 신경망에 입력된다. 두 개의 LNN 신경망의 선택은 GNN의 결과에 따라 결정된다. 결과가 Shot이면 LNN1과 LNN2가 선택되며, Move이면 LNN1과 LNN3, Static이면 LNN2와 LNN3이 각각 선택되어 GNN의 결과를 검증한다. 계층적 구조의 신경망을 구성하여 사용한 이유는 전역적인 특징 값을 이용한 GNN의 결과를 지역적인 특징 값을 이용한 세 개의 LNN으로 검증하기 위함이다.

픽처 그룹 분류에 사용된 특징은 표 1과 같이 현재와

표 2 특징의 종류 및 해설

종 류	수 식	해 설
픽셀 차이 값	$PD_n=\sum_{i}^{M}\sum_{j}^{N}\frac{\lvert F_n(i,j)-F_{n-1}(i,j)\rvert}{M\times N}$ $F_n(i,j)$: n번째 프레임의 (i,j) 위치에서의 픽셀 값 M: DC 영상의 가로 크기 N: DC 영상의 세로 크기	두 DC 영상에서 대응되는 픽셀의 차이 값의 누적
히스토그램 차이 값	$HD_n=\sum_{i}^{K}\frac{\lvert H_n(i)-H_{n-1}(i)\rvert}{M\times N}$ $H_n(i)$: n번째 프레임에서 명도 값 i의 히스토그램	두 DC 영상에서 명도 값의 분포를 구한 후 명도 값에 대응되는 분포의 차를 누적
픽셀 평균 차이 값	$MD_n=\lvert M_n-M_{n-1}\rvert$ $M_n=\frac{\sum_{i=1}^{K}H_n(i)*i}{M\times N}$ M_n: n번째 프레임의 평균 명도 값	두 DC 영상에서 각 영상의 명도 값의 평균을 구한 후 평균값의 차이
픽셀 분산 차이 값	$VD_n=\lvert V_n-V_{n-1}\rvert$ $V_n=\frac{\sum_{i}^{K}H(i)*(i-m)^2}{M\times N}$ m: n번째 프레임의 평균 픽셀 값	두 DC 영상에서 각 영상의 명도 값의 분산을 구한 후 그 값의 차이

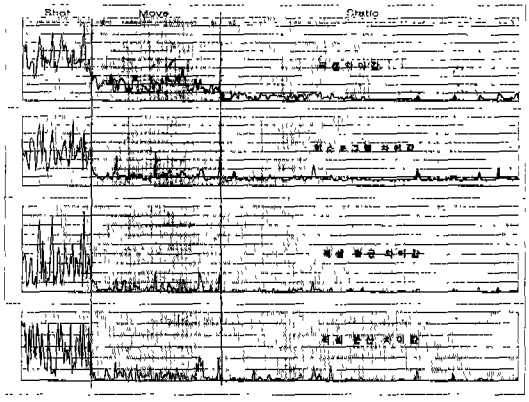

그림 4 특징 종류에 따른 분류 변별력

이전 I 프레임의 DC 영상으로부터 추출된 픽셀 차이 값, 히스토그램 차이 값, 픽셀 평균 차이 값 및 픽셀 분산 차이 값이며, 이 특징들을 영상 전체에서 추출하여 전역적인 특징 값을 구성한다. 또한 DC 영상을 가로와 세로를 각각 3과 4로 분할한 후 분할된 각 영역에서 위의 특징 값들을 추출하여 지역적인 특징 값들을 구성한다. 따라서 전역적인 특징 벡터의 차원은 4가 되며, 지역적인 특징 벡터의 차원은 48이 된다. 전체 DC 영상으로부터 추출한 전역적인 특징은 장면을 Shot, Move, Static의 세 부류로 분할하는데 적합하며, 각각의 부분 영상에서 추출한 지역적 특징은 Shot/Move, Shot/Static 및 Move/Static의 Pairwise 검증에 적합하다는 것을 실험에서 확인할 수 있었다.

그림 4는 각 특징의 분류 변별력을 보여준다. 그림의 x축은 현재의 I 프레임을 나타내며, y축은 이전 I 프레임의 DC 영상과의 특징 차이 값을 나타낸다. 픽셀 차이 값 특징이 장면을 세 부류로 분류하는데 가장 높은 변별력을 보이지만 전체 영상에 나타나는 객체의 움직임이나 카메라의 동작에 민감한 단점이 있다. 반면, 히스토그램 차이 값 특징은 전체 영상에 나타나는 객체의 움직임이나 카메라의 동작에 둔감한 장점을 지닌다. 픽셀 평균 차이 값과 분산 차이 값 특징들은 객체의 움직임을 잘 표현하는 장점을 갖는다. 따라서 본 연구에서는 비디오 데이타의 다양하고 복합적인 장면 변화에 적응하기 위하여 네 종류의 다중 특징들을 결합하여 사용하였다.

신경망 결과에 따른 픽처 그룹 결정 및 검출 내용은 표 3에 설명되며, 최종 결정은 GNN의 결과와 LNN1, LNN2, LNN3의 결과들을 조합하여 결정한다. Disregard로 표시된 부분은 현재 특징으로는 나타날 수 없는

표 3 픽처 그룹 결정 및 이후 검출 내용

GNN 결과	LNN 결과 1		LNN 결과 2		픽처 그룹 최종 결정	최종 결정에 따라 픽처 그룹 내에서 검출 할 내용
	분류 종류	결과	분류 종류	결과		
Shot	Shot/Move	Shot	Shot/Static	Shot	Shot	컷
		Shot		Static	Disregard	
		Move		Shot	Shot	컷
		Move		Disregard	Disregard	
Move	Move/Shot	Move	Move/Static	Move	Move	디졸브, 객체 움직임, 카메라 동작
		Move		Static	Move	디졸브, 객체 움직임, 카메라 동작
		Shot		Move	Shot	컷
		Shot		Static	Disregard	
Static	Shot/Static	Static	Move/Static	Static	Static	픽처 그룹 내의 모든 프레임을 Static으로 결정
		Static		Move	Move	디졸브, 객체 움직임, 카메라 동작
		Shot		Static	Disregard	
		Shot		Move	Disregard	

경우들이며, 실제 실험에서도 이와 같은 결과는 나타나지 않아서 무시함을 의미한다. 픽처 그룹이 Shot으로 분류되면 픽처 그룹 내에서 컷을 검출하고 만약 컷이 없다면 디졸브, 카메라 동작 및 객체 움직임을 검출한다. 픽처 그룹이 Move로 분류되면 픽처 그룹 내에서 디졸브, 카메라 동작 및 객체 움직임을 검출한다. 픽처 그룹이 Static으로 분류되면 해당 픽처 그룹의 모든 프레임을 Static으로 결정한다.

2.2 컷 위치 검출

픽처 그룹에서 컷의 발생 위치는 그림 5와 같이 P 프레임에서 발생하는 경우와 P 프레임 사이의 두 개의 B 프레임에서 각각 발생하는 경우로 분류된다. 현재의 픽처 그룹이 샷으로 분류되면 그림 6과 같이 픽처 그룹 내의 P, B 프레임을 선별적, 단계적으로 검사하여 컷의 발생 위치를 검출한다.

그림 5, 6의 Case 1은 B1에서 컷이 발생한 경우로서 B1, B2 프레임 모두 후방향(Backward) 참조 P 프레임과 밀접한 관계가 있기 때문에 B1, B2 프레임 모두 후방향 매크로블록의 개수가 많아진다. 또한, 후방향 참조 프레임인 P 프레임의 인트라(Intra) 매크로블록의 개수도 많아진다. Case 2는 B2에서 컷이 발생한 경우로서 B2는 후방향 참조 P 프레임과 밀접한 관계가 있기 때문에 후방향 매크로블록의 개수가 많아지지만, B1은 전방향(Forward) 매크로블록의 개수가 많아진다. 또한, B2의 후방향 참조 P 프레임의 인트라 매크로블록의 개수도 많아진다. Case 3의 경우는 P 프레임에서 컷이 발생한 경우로서 P 프레임의 인트라 매크로블록의 개수가

그림 5 컷의 발생 위치

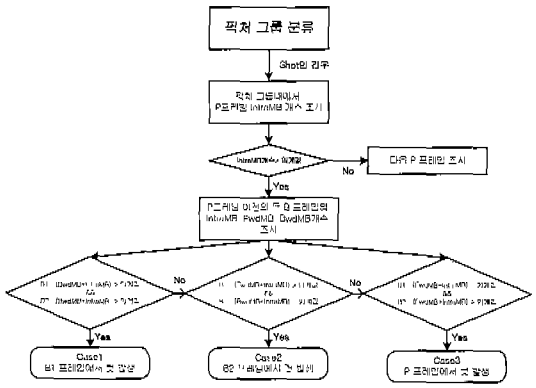

그림 6 컷 검출 과정

많아진다. 또한, B1, B2 프레임 모두 전방향 참조 P 프레임과 밀접한 관계가 있기 때문에 전방향 매크로블록의 개수도 많아진다. 만약 I 프레임에서 컷이 발생한다면 I 프레임 이전의 B1, B2 프레임의 후방향 매크로블록의 개수가 많아진다.

컷의 발생 위치에 따라서 P, B1, B2 프레임의 매크로블록 타입의 개수가 일정한 특징을 갖기 때문에 컷 위치를 검출하기 위해서 통계적인 방법을 사용한다. P 프레임

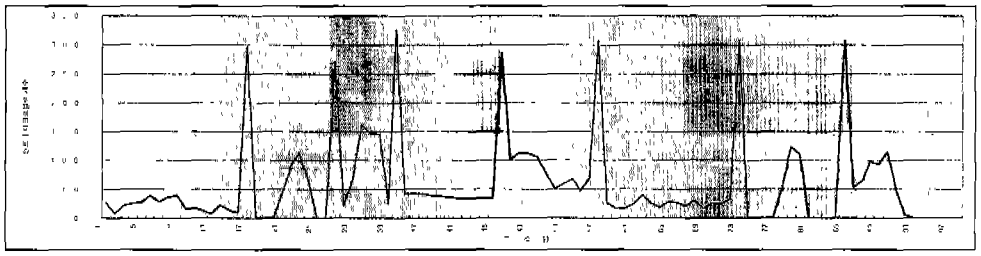

그림 7 P 프레임의 인트라 매크로블록 개수

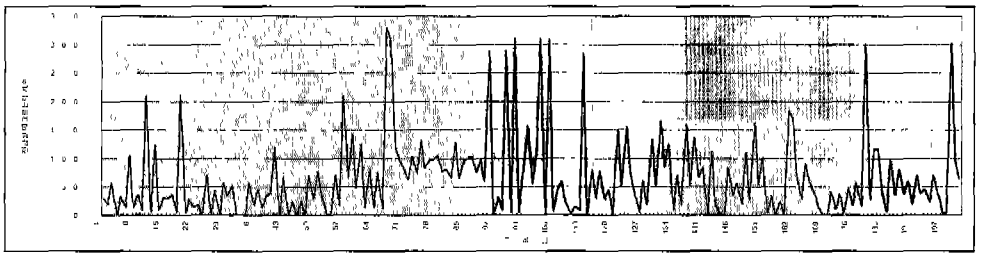

그림 8 B 프레임의 전방향 매크로블록 개수

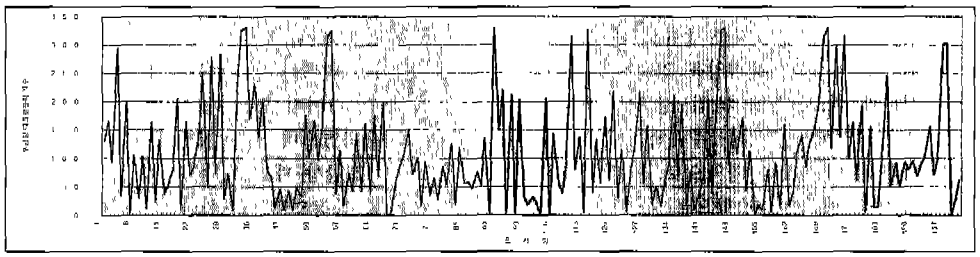

그림 9 B 프레임의 후방향 매크로블록 개수

에서는 그림 7과 같이 인트라 매크로블록의 개수를 특징으로 사용하며, B 프레임에서는 그림 8, 그림 9과 같이 전방향 매크로블록과 후방향 매크로블록의 개수를 특징으로 사용한다. 각 그림의 x축은 프레임 번호를 나타내며 y축은 해당 프레임의 매크로블록의 개수를 나타낸다.

컷 위치를 검출하기 위해서 우선 해당 픽처 그룹의 P 프레임의 인트라 매크로블록 개수를 조사한다. P 프레임의 인트라 매크로블록의 개수가 설정된 임계값보다 크면 P 프레임 이전의 두 개의 B 프레임의 전방향 매크로블록과 후방향 매크로블록의 개수를 조사하여 컷의 발생 위치를 결정한다. 각 경우의 매크로블록 타입의 특징을 표 4에 요약했다. 이와 같은 방법으로 컷의 발생 위치를 정확히 검출할 수 있음을 실험에서 확인할 수 있었다.

표 4 컷 위치에 따른 B 프레임의 매크로블록 타입 개수 분포

경우	B1 프레임	B2 프레임
Case 1	후방향 매크로블록의 개수가 많다.	후방향 매크로블록의 개수가 많다.
Case 2	전방향 매크로블록의 개수가 많다.	후방향 매크로블록의 개수가 많다.
Case 3	전방향 매크로블록의 개수가 많다.	전방향 매크로블록의 개수가 많다.

2.3 디졸브/카메라동작/객체움직임 검출

해당 픽처 그룹이 Move로 분류되면 디졸브, 카메라 동작 및 객체 움직임을 검출한다. 디졸브, 카메라 동작, 객체 움직임은 DC 영상에서 추출된 이미지 특징이나 매크로블록 종류의 분포만을 이용하여 검출하는데는 어려움이 있다. 따라서 본 논문에서는 이러한 어려움을 부분적으로 해결하기 위해서 움직임 벡터와 매크로블록 타입을 특징으로 사용하는 신경망을 구성하였다.

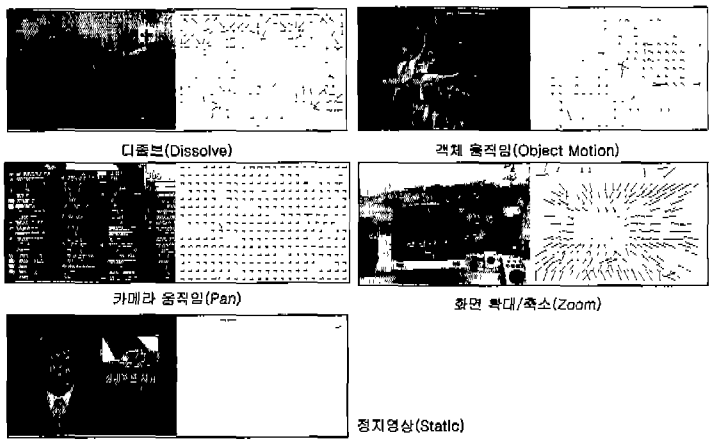

그림 10 컷을 제외한 5 종류의 비디오 데이타 분류 및 움직임 벡터 분포

카메라 동작은 카메라 움직임과 화면의 확대/축소로 구분된다. 카메라 움직임은 카메라의 수평, 수직 및 대각선 움직임을 포함한다. 따라서 그림 10과 같이 비디오 데이타에서 컷을 제외하면 디졸브, 객체 움직임, 카메라 움직임, 화면 확대/축소 및 정지 영상의 다섯 종류로 분류된다.

디졸브, 카메라 동작, 객체 움직임을 검출하기 위해서 움직임 벡터 성분과 매크로블록 타입 정보를 특징으로 사용한다. P, B 프레임에서 특징을 추출하기 위해서 프레임을 그림 11과 같이 4개의 영역으로 분할한 후 각 부분 영역의 움직임 벡터를 8방향으로 나누어 검출된 각 방향의 개수를 특징으로 사용한다. 이와 같이 네 개의 영역으로 나눈 이유는 카메라 동작과 객체 움직임의 각 경우들이 부분 영역에서 분류되는 특성을 보이기 때문이다. 디졸브는 영역 전체에 대한 변화를 갖는 반면에 객체 움직임은 전체 영역보다는 부분 영역에 대한 변화를 갖는다. 또한, 카메라 움직임은 전체 영역에서 같은 방향을 갖는 움직임 벡터 특징을 갖는 반면에 화면 확대/축소는 각 부분 영역에서 동일한 움직임 벡터 특징을 갖는다. P, B 프레임의 참조 특성을 나타내기 위해서 매크로블록 타입 개수를 특징으로 함께 사용한다. P 프레임에서는 인트라 매크로블록의 개수, 전방향 매크로블록의 개수, 나머지 모든 매크로블록 타입의 개수를 특징으로 사용하였으며, B 프레임에서는 인트라 매크로블록의 개수, 전방향 매크로블록의 개수, 후방향 매크로블록의 개수를 특징으로 사용한다.

디졸브, 카메라 동작, 객체 움직임을 검출하기 위해서 움직임 벡터와 매크로블록 타입 개수를 입력 특징으로 하는 신경망을 그림 12와 같이 구성하였다. P, B 프레임 각각의 신경망은 움직임 벡터와 매크로블록 타입 개수로 구성된 35개의 입력 벡터를 갖고, 디졸브, 객체 움

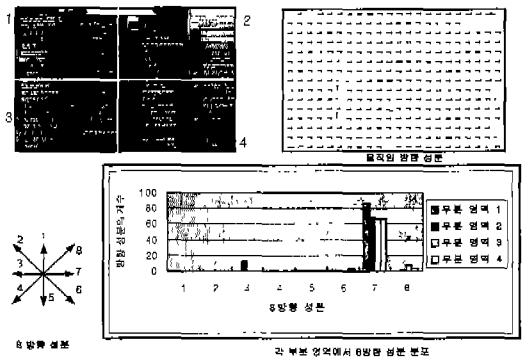

그림 11 영역 분할과 8방향 성분 분포

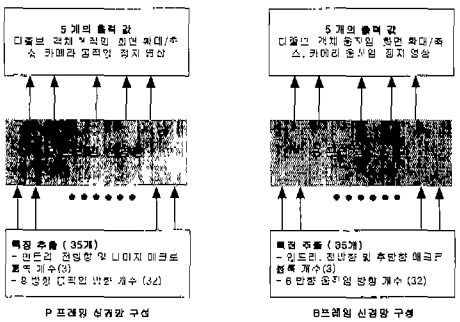

그림 12 디졸브/카메라동작/객체움직임을 검출하기 위한 P, B 프레임의 신경망 구조

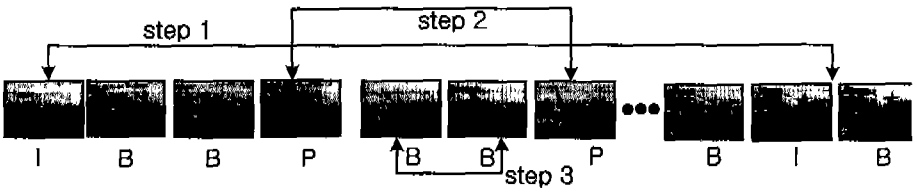

그림 14 화면 확대/축소(Zoom) 발생 위치 검출 예

직임, 화면 확대/축소, 카메라움직임 및 정지 영상의 5개의 출력 값을 갖는다. P, B 프레임 각각에 대해 신경망을 구성한 이유는 P 프레임은 전방향 프레임만 참조하는 반면에 B 프레임은 전방향, 후방향 프레임 모두를 참조하기 때문이다.

디졸브/카메라동작/객체움직임의 발생 위치를 검출하기 위한 P, B 프레임의 계층적인 검사 방법은 그림 13과 같다. 우선 픽처 그룹이 Move로 결정되면 해당 픽처 그룹 내의 P 프레임을 검사한다. 이전 P 프레임과 현재 P 프레임을 검사한 신경망 결과 값이 틀린 경우에만 P 프레임 사이의 두 개의 B 프레임을 검사한다. 이와 같이 해당 픽처 그룹 내에서 선별적, 계층적으로 프레임을 검사하여 처리 시간을 단축시키고자 하였다.

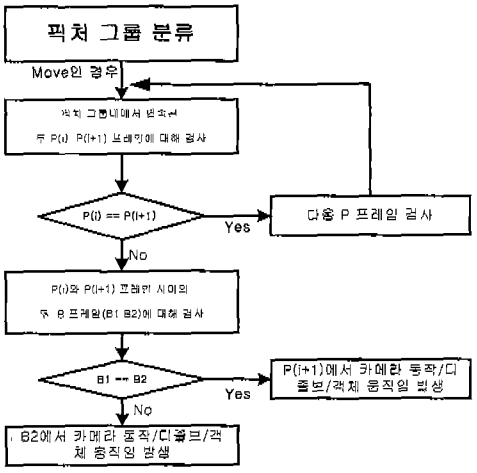

그림 13 디졸브/카메라동작/객체움직임 검출 과정

그림 14는 위의 과정을 적용하여 카메라 동작인 화면의 확대/축소가 발생된 위치를 검출하는 예를 보여준다. Step1은 I 프레임을 이용한 픽처 그룹 분류로서 Move로 판정하였으며, step 2는 연속된 두 개의 P 프레임을 검사한다. 이전 P 프레임은 정지영상(Static)으로 신경망에 의해 판정되었으며, 현재 P 프레임은 화면 확대/축소(Zoom)로 판정되었다. 판정된 결과가 다르기 때문에 step 3에서는 P 프레임들 사이의 B 프레임을 검사한다. 앞에 있는 B 프레임은 정지 영상으로 판정되었으며, 뒤에 있는 B 프레임은 화면 확대/축소로 판정되었다. 따라서 뒤에 있는 B 프레임에서 화면의 확대/축소가 발생한 것으로 최종 판정한다.

3. 실험 및 분석

제안하는 방법을 Visual C++ 6.0을 사용하여 Pentium III 450MHz PC에서 구현하였다. 실험에 사용한 비디오 데이타는 MPEG 1으로 압축된 뉴스와 뮤직 비디오를 사용하였다. 원 영상의 M(P 프레임 사이의 간격)은 3이며, N(I 프레임 사이의 간격)은 12이고, 영상의 크기는 352×240이다. MPEG으로 압축하기 위해서 상용 소프트웨어를 사용하였으며 비디오의 압축 비트 율은 1.15Mbps이다.

픽처 그룹 분류의 정확성을 측정하기 위해서 I 프레임의 DC 영상을 사용하였다. 이 영상의 크기는 44×33이며 명도(gray) 영상이다. 신경망의 학습 데이타를 만들기 위해서 이전과 현재의 I 프레임을 개발자가 확인하면서 분류하였으며, 테스트 데이타도 같은 방법으로 분류하였다. 본 논문에서 사용한 신경망 구조는 은닉층이 하나인 간단한 다층 신경망 구조로서, GNN은 4개의 입력노드, 3개의 은닉노드, 3개의 출력노드로 구성하였으며, LNN은 48개의 입력노드, 25개의 은닉노드, 2개의 출력노드로 구성하였다. 디졸브/카메라동작/객체움직임을 검출하기 위한 신경망은 35개의 입력노드, 19개의 은닉노드, 5개의 출력노드로 구성하였다. 학습 및 테스트 데이타를 구축하기 위한 Ground Truth 작업은 이전과 현재의 P, B 프레임을 개발자가 확인하면서 결정하였다. 신경망의 학습은 가장 보편적으로 사용되는 역전파학습(back-propagation learning) 방법을 사용하였다[16]. 신경망의 학습에는 대체로 많은 시간이 소요되지만, 학습된 정보를 저장하여 인식하는 과정은 은닉층이 하나인

표 5 픽처 그룹 분류의 각 신경망 결과(I 프레임을 이용한 분류)

비디오 데이타	신경망		총 I 프레임 개수	인식 결과(%)	오 류			
					Truth(학습)	인식 결과	개수	총 개수
뮤직 비디오 (학습 데이타)	GNN		594	98.1	Shot	Move	1	11
					Move	Shot	10	
	LNN 1/2/3	Shot/Move	469	100.0				0
		Shot/Static	273	100.0				
		Static/Move	439	100.0				
뮤직 비디오 (테스트데이타)	GNN		344	95.6	Shot	Move	11	15
					Move	Shot	4	
	LNN 1/2/3	Shot/Move	197	93.9	Shot	Move	8	13
					Move	Shot	4	
		Shot/Static	222	100.0			0	
		Static/Move	267	99.6	Move	Static	1	
뉴스 (학습 데이타)	GNN		357	98.3	Shot	Move	2	6
					Move	Shot	2	
					Move	Static	2	
	LNN 1/2/3	Shot/Move	125	100.0			0	0
		Shot/Static	261	100.0			0	
		Static/Move	306	100.0			0	
뉴스 (테스트 데이타)	GNN		185	95.6	Shot	Move	3	8
					Move	Shot	3	
					Move	Static	2	
	LNN 1/2/3	Shot/Move	130	96.9	Shot	Move	2	5
					Move	Shot	2	
		Shot/Static	96	100.0			0	
		Static/Move	140	99.3	Move	Static	1	
드라마 (학습 데이타)	GNN		600	98.1	Shot	Move	4	11
					Move	Shot	5	
					Move	Static	2	
	LNN 1/2/3	Shot/Move	353	99.4	Move	Shot	2	2
		Shot/Static	387	100.0			0	
		Static/Move	460	100.0			0	
드라마 (테스트 데이타)	GNN		462	95.2	Shot	Move	12	21
					Move	Shot	9	
					Move	Static	3	
	LNN 1/2/3	Shot/Move	265	96.4	Shot	Move	2	11
					Move	Shot	4	
		Shot/Static	299	100.0				
		Static/Move	360	98.6	Move	Static	5	

표 6 픽처 그룹 분류 상세 결과(T: 전체 I 프레임 개수(Total), CD: 정확히 검출한 I 프레임 개수(Correct Detection), FD: 틀리게 검출한 I 프레임 개수(False Detection), MD: 검출하지 못한 개수(Missed Detection))

		Shot				Move				Static			
		T	CD	FD	MD	T	CD	FD	MD	T	CD	FD	MD
뉴스 (185개)	GNN	43	41	2	2	89	86	2	3	53	53	1	0
	제안한 방법	43	43	1	0	89	88	0	1	53	53	0	0
뮤직 비디오 (344개)	GNN	77	66	12	11	121	111	9	10	146	146	0	0
	제안한 방법	77	75	4	2	121	117	2	4	146	146	0	0
드라마 (462개)	GNN	102	90	9	12	163	153	9	10	197	197	3	0
	제안한 방법	102	97	4	5	163	157	6	6	197	197	1	0

간단한 다층 신경망을 사용하여 빠르게 처리할 수 있다.

I 프레임을 이용한 픽처 그룹 분류를 실험하기 위한 각 신경망의 학습 및 테스트 데이타의 구성과 분류 결과가 표 5와 표 6에 실려있다. 실험에 사용된 비디오 데이타는 뮤직 비디오, 뉴스, 드라마의 세 종류로서 1,551개의 I 프레임을 학습 데이타로 사용하였으며 991개의 I 프레임을 테스트 데이타로 사용하였다. 표 5에서의 인식률은 테스트 데이타 가운데 맞게 분류한 개수만을 비율로 계산한 것이다. 또한, 표 6은 표 5의 결과를 상세하게 분류하여 Precision 및 Recall[15]을 계산할 수 있도록 하였다. 비디오 데이타의 종류마다 특성이 다르기 때문에 하나의 범용 신경망을 구성하는 것보다는 각 데이타의 종류에 적합한 신경망을 구성하여 분류 정확성을 높이고자 하였다. 실험 결과를 보면 뉴스, 뮤직 비디오, 드라마의 학습 및 테스트 데이타 모두에서 예상과 같이 전역적인 특징과 GNN 신경망을 사용한 것보다는 전역적인 특징과 지역적인 특징을 함께 사용하는 2단계의 구조의 계층적인 신경망을 사용함으로서 인식률이 개선된 것을 확인할 수 있다. 뮤직 비디오 데이타가 뉴스와 드라마 데이타보다 인식률이 낮은 이유는 카메라 동작 또는 카메라동작/객체움직임이 동시에 발생하는 경우가 많았기 때문이다.

그림 15는 2단계 신경망을 사용하여 오류가 개선된 예를 보여준다. A, B 영상 모두 비슷한 명도 분포에서 Shot이 발생한 경우이다. 전역적인 특징만을 사용한 GNN에서는 A, B를 모두 Move로 분류하였지만, 지역적인 특징을 이용한 LNN1은 A, B를 모두 Shot으로 분류하고, LNN2는 Move로 결정하였기 때문에 최종적으로 Shot으로 결정된 경우를 보여준다. 전역적인 특징과 GNN만을 이용해서 픽처 그룹을 분류한다면 A, B를 모두 Move로 분류하겠지만, 본 논문에서 제안한 2단계 계층적 신경망을 이용하여 오류 발생이 줄어 든 것을 확인하였다.

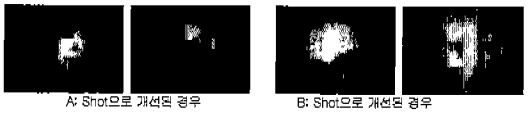

그림 15 계층적 신경망을 이용하여 인식 결과가 개선된 예

뉴스와 뮤직 비디오에서 픽처 그룹이 Shot으로 분류되었을 때의 컷 검출 결과를 표 7에 정리하였다. 픽처 그룹이 Shot으로 결정된 경우에는 모든 컷이 정확히 검출되는 것을 확인할 수 있었지만, 픽처 그룹이 Shot을 Move로 잘못 분류한 경우에는 해당 픽처 그룹 내에서 컷 검출이 실행되지 않아서 컷을 검출하지 못하는 경우도 발생했다. 그러나 픽처 그룹 분류에서 Shot으로 잘못 검출된 오류가 수정되어 틀리게 검출한 컷이 없는 것을 표 7에서 확인할 수 있다.

표 7 컷 검출 결과

	전체 컷 개수(T)	정확히 검출한 개수(CD)	틀리게 검출한 개수(FD)	검출하지 못한 개수(MD)
뉴스	40	40	0	0
뮤직 비디오	77	74	0	3
드라마	102	100	0	2

디졸브, 카메라 동작, 객체 움직임 검출을 위한 실험 데이타는 뉴스와 뮤직 비디오를 함께 편집하여 구성하였다. 디졸브(Dissove), 객체 움직임(ObjectMotion), 화면 확대/축소(Zoom), 카메라 움직임(Pan), 정지 영상(Static)의 5개 클래스에 대해서 학습과 테스트를 수행한 결과는 표 8과 같다.

표 8 P, B 프레임 신경망 인식 결과

프레임 종류	데이타 종류	총 프레임 개수 (P 또는 B)	인식률(%)
P 프레임	학습 데이타	800	93.74
	테스트 데이타	400	90.13
B 프레임	학습 데이타	800	93.13
	테스트 데이타	400	89.28

테스트 데이타에 대해 각 클래스의 인식 결과를 그림 16에 정리하였다. 카메라 움직임, 화면 확대/축소, 정지 영상은 디졸브나 객체 움직임에 비해 상대적으로 높은 인식률을 나타냈지만, 디졸브, 객체 움직임에 대한 인식률이 낮음을 알 수 있다. 디졸브, 객체 움직임, 카메라동

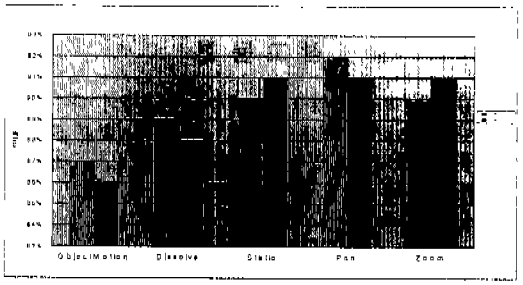

그림 16 테스트 데이타의 각 클래스별 인식 결과

작에 대한 검출 결과를 표 9에 나타냈다. 큰 물체가 움직인 경우에는 카메라 움직임과 비슷한 특징을 갖기 때문에 카메라 움직임으로 잘못 검출하는 오류가 발생하였다. 또한, 영상 전체에서 발생하는 객체들의 움직임을 디졸브로 잘못 검출하는 경우가 발생하였다.

표 9 디졸브, 객체 움직임, 카메라 동작 검출 결과

종류	총 개수(T)	정확히 검출한 개수(CD)	특이게 검출한 개수(FD)	검출하지 못한 개수(MD)
디졸브	14	12	1	2
객체 움직임	23	21	3	2
화면 확대/축소	17	16	1	1
카메라 움직임	30	27	3	3

4. 결 론

본 논문은 MPEG 비디오 데이타에서 컷과 디졸브의 발생 위치를 검출하여 샷 단위로 분할하며, 각 샷에서 카메라 동작, 객체 움직임을 검출하는 방법을 제안하였다. 제안한 방법은 I 프레임의 DC 영상을 분석하여 픽처 그룹을 Shot, Move, Static으로 세분화하여 분류하였으며, 각 픽처 그룹의 P, B 프레임을 선별적, 계층적으로 탐색하여 컷의 정확한 발생 위치와 카메라 동작 또는 객체 움직임의 종류를 결정하였다. 픽처 그룹을 세분화하여 분류하기 위해서 2단계 계층적 구조의 신경망을 제안하였으며, 다양한 특징 정보의 필요성도 확인하였다. 또한 컷 발생 위치를 정확하게 검출하기 위해서 P, B 프레임의 매크로 블록 타입의 분포를 통계적으로 이용하였으며, 디졸브, 카메라 동작, 객체 움직임의 종류를 검출하기 위해서는 P, B 프레임의 매크로 블록 타입과 움직임 벡터를 입력으로 사용한 신경망을 구성하였다.

향후 연구로는 페이드(fade)와 와이프(wipe)를 포함하는 점진적인 장면 전환에서도 정확하고 효율적으로 샷을 추출 방법을 개발하는 것이다.

참 고 문 헌

[1] 김주민, 최영우, 정규식, "계층적 신경망과 다중특징을 이용한 MPEG 비디오 분할." *한국 통신학회 하계학술발표회 논문집(상)*, pp. 52-55, 2000.

[2] H. Zhang, A. Kankanhalli, and S. Smoliar, "Automatic Partitioning of Full-Motion Video," *Multimedia Systems*, Vol. 1, No. 1, pp. 10-28, 1993.

[3] M. Cherfaoui and C. Bertin. "Two-state Strategy for Indexing and Presenting Video," *Proc. of SPIE-Storage and Retrieval for Image and Video Databases II.*, Vol. 2185, pp. 174-184. 1994.

[4] B. Shahraray, "Scene Change Detection and Content-Based Sampling of Video Sequences," *Proc. of SPIE Digital Video Compression: Algorithms and Technologies*, Vol. 2419, pp. 2-13, 1995.

[5] D. Swanberg. C. F. Shu and R. Jain, "Knowledge Guided Parsing in Video Databases," Proc. of SPIE'93 - Storage and Retrieval for Image and Video Databases, Vol. 1908, pp. 13-24, 1993.

[6] A. Hanjalic, R. Lagendijk, and J. Biemond, "A New Key-Frame Allocation Method for Representing Stored Video-Streams," *Proc. of 1st Inter. Workshop on Image Databases and Multi-Media Search*, pp. 67-74, Netherlands, 1996.

[7] H. Zhang and S. W. Smoliar, "Developing Power Tools for Video Indexing and Retrieval," Proc. of SPIE'94 - Storage and Retrieval for Image and Video Databases II, Vol. 2185, pp. 140-149, 1994.

[8] B. Yeo and B. Liu, "A Unified Approach to Temporal Segmentation of Motion JPEG and MPEG Compressed Video," *Proc. of Inter. Conf. on Multimedia Computing and Systems*, pp. 81-88, 1995.

[9] H. Zhang, C. Low, and S. Smoliar, "Video Parsing and Browsing using Compressed Data," *Multimedia Tools and Applications*, Vol. 1, No. 1, pp. 89-111, 1995.

[10] H. Liu and G. Zick, "Scene Decomposition of MPEG Compressed Video," *Proc. of SPIE Digital Video Compression: Algorithms and Technologies*, Vol. 2419, pp. 26-37, 1995.

[11] N. Gamaz, X. Huang, and S. Panchanathan, "Scene Change Detection in MPEG Domain," *Proc. of IEEE Southwest Symposium on Image Analysis and Interpretation*, pp. 12-17, 1998.

[12] F. Arman, A. Hsu and M-Y, Chiu, "Feature Management for Large Video Databases," *Proc. of SPIE Storage and Retrieval for Image and Video Databases*, Vol. 1908, pp. 2-12. 1993.

[13] 김가현, 문형식, "MPEG 압축된 비디오의 자동 분할 기법," *한국 정보처리학회 논문지*, 제6권 제4호, 1999.

[14] 이홍훈, 이홍규, "패턴인식을 이용한 MPEG 비디오 스트림상에서의 장면 전환 검출," *한국정보과학회 봄 학술발표 논문집*, 제 13권 제 1호, pp. 619-621, 1998.

[15] J. S. Boreczky and L. A. Rowe, "Comparison of Video Shot Boundary Detection Techniques," IS&T/SPIE, Vol. 2670, pp. 170-179, Feb. 1996.

[16] 대우전자 영상연구소, MPEG 비디오, 연암출판사, 서울, 1995.

[17] 김상운, 패턴인식 입문, 홍릉과학출판사, 서울, 1992.

김 주 민

1999년 숭실대학교 전자공학과 졸업(학사). 2001년 숭실대학교 전자공학과 졸업(석사). 2001년 2월 ~ LG 전자기술원 정보기술연구소 연구원. 관심분야는 영상처리, 가상현실 등

최 영 우

1985년 연세대학교 전자공학과 학사. 1986년 University of Southern California 컴퓨터공학과 석사. 1994년 University of Southern California 컴퓨터공학과 박사. 1994년 10월 ~ 1997년 2월 LG전자기술원 선임연구원. 1997년 3월 ~ 현재 숙명여자대학교 정보과학부 조교수, 부교수. 관심분야는 영상처리, 패턴인식, 문자인식 등

정 규 식

1979년 서울대학교 전자공학과 졸업(학사). 1981년 한국과학기술원 전산학과 졸업(석사). 1981년 2월 ~ 1984년 7월 금성사 중앙연구소 선임연구원. 1984년 9월 ~ 1990년 8월 University of Southern Cali(ornia 컴퓨터공학과 졸업(석사, 박사). 1993년 12월 ~ 1994년 3월 IBM Watson 연구소 방문연구원. 1998년 3월 ~ 1999년 2월 IBM Almaden 연구소 방문연구원. 1990년 9월 ~ 현 재 숭실대학교 정보통신전자공학부 부교수. 관심분야는 인터넷컴퓨팅, 인공지능, 멀티미디어 등