대화형 방송 환경을 위한 동적 비디오 요약

김재곤, 장현성, 김진웅

한국전자통신연구원 방송기술연구부

Dynamic Video Abstraction for Interactive Broadcasting Applications

Jae-Gon Kim, Hyun-Sung Chang and Jinwoong Kim

Electronics and Telecommunications Research Institute
E-mail : {jgkim,chs,jwkim}@video.etri.re.kr

요 약

방송 환경의 디지털화와 더불어 단순히 단방향 방송 정보만을 시청하던 기존의 방식을 넘어 사용자의 다양한 욕구를 수용할 수 있는 대화형 방송 서비스(interactive broadcasting service)가 시작되고 있다. 대화형 방송 환경에서는 특히 사용자 측에 제공된 방대한 양의 디지털 멀티미디어 자료에 대한 효과적인 접근이 요구되는데, 본 논문에서는 이를 위하여 방송 비디오를 효과적으로 브라우징(browsing) 및 검색하고 전체의 내용을 짧은 시간 내에 개관할 수 있도록 하는 동적 비디오 요약(dynamic video abstraction) 기법에 관하여 고찰한다. 동적 비디오 요약에 의한 요약 비디오(skim video)는 전체 비디오를 내용에 기반하여 효과적으로 표현할 수 있도록 동영상 내의 주요 구간만으로 구성된 것으로, 대화형 방송에서 새로운 형태의 프로그램 안내 및 사용자 저장 자료에 대한 브라우징 도구 등으로써 매우 유용하게 사용할 수 있다. 본 논문에서는 자동으로 비디오 요약을 구현하기 위한 접근 방법과 전체 기능 구성 및 각 기능들의 구현 방법에 대하여 기술한다.

1. 서론

방송 환경의 디지털화가 급진전됨에 따라, 방송 서비스는 기존 단방향의 단순 시청형 방송 서비스에서 방송 정보에 대한 사용자의 다양한 욕구를 수용하는 새로운 형태의 대화형 서비스로 변모하고 있다. 현재 대화형 방송은 인터넷 또는 전화망 등 저속의 interaction 채널을 이용한 양방향 방송 및 수신기의 대용량 저장 장치를 이용한 국부 대화형 서비스 형태를 띠고 있다. DAVIC(Digital Audio Visual Council) [1]에서 추진 중인 'TV Anytime', 'TV Anywhere' 등은 사용자측의 대용량 저장 장치를 이용한 push/pull 형태의 국부 대화형 서비스와 고속 인터넷을 통하여 전세계 어디라도 원하는 방송 프로그램을 제공하는 대화형 방송 서비스의 한 전형으로 볼 수 있다. 또한, 인터넷도 인터넷 방송이나 interaction 채널 등의 형태로 방송과 결합하여 대화형 방송 환경을 구성하고 있다. 이러한 대화형 방송의 환경에서 제공되는 방대한 양의 멀티미디어 자료를 사용자가 효율적으로 이용하기 위해서는 이들 자료에 대한 효과적인 접근 및 검색 기법이 요구된다. 이에 대한 노력의 일환으로 현재 MPEG-7[2]에서 멀티미디어 자료의 표현 규격에 대한 작업이 활발히 진행 중이다.

본 논문에서는 대화형 방송 환경에서 방대한 양의 방송 비디오를 효과적으로 브라우징하여 원하는 비디오를 검색하거나 프로그램 전반의 내용을 짧은 시간 내에 개관할 수 있는 요약 비디오를 생성하는 동적 비디오 요약 기법에 관하여 고찰한다. 동적 비디오 요약에 의한 요약 비디오는 전체 비디오를 내용에 기반하여 효과적으로 표현할 수 있도록 주요 비디오 구

간만으로 구성된 짧은 시간의 비디오이다. 대화형 방송 환경에서 요약 비디오는 새로운 형태의 프로그램 안내[3]나 사용자 저장 자료에 대한 효과적인 브라우징 도구 등의 역할을 담당할 수 있는데, 본 논문에서는 요약 비디오 생성의 자동화를 위한 접근 방법과 전체적인 기능 구성 및 각 기능의 구현을 위한 요소 기술에 관하여 기술한다.

2. 동적 비디오 요약

2.1 개요

동영상 자료에 대한 브라우징 및 검색 과정에서 내용에 대하여 요약된 형태의 정보를 가지고 있다면 'fast-forward'나 'playback' 등과 같이 많은 시간과 노력을 요하는 작업을 덜어 줌으로써 효율성을 극대화할 수 있다. 비디오 색인에 관하여 수행된 연구의 상당수가 비디오 자료의 구조화 및 요약 작업을 위한 샷(shot) 경계의 검출[4], 대표 프레임의 추출[5,6], 씬(scene)의 분할[7,8] 등 요약에 관련한 연구가 집중되었다.

비디오 요약은 그 형식에 따라 다음과 같이 정적인 형태와 동적인 형태로 분류할 수 있다.

- 정적 요약
  Story board –
  scene transition graph[7], table of content[8]
  Video mosaic[9]
- 동적 요약
  Video skim[10] (movie trailer)

현재까지의 연구는 정적 요약에 주로 초점이 맞추어져 왔는데, 이는 대개 샷을 검출하고, 각 샷을 표현하기 위한 소수의 대표 프레임들을 선별하여 비디오의 내용을 요약하는 방식을 취한다. 반면, 요약 비디오는 전체 비디오 내용을 효과적으로 표현할 수 있는 주요 구간들을 선별하는 방식의 요약 기법으로 오디오를 함께 포함하므로 사용자의 이해도에 있어 더 유리한 요약이라고 할 수 있다.

본 논문에서는 이들 두 형태의 요약 기법을 적절히 활용한 복합형 요약 비디오 표현 구조(description scheme)를 제안한다.

2.2 동적 비디오 요약 구성도

그림 1 은 비디오 요약을 위한 전체 기능 구성도이다. 크게 다음의 세 기능부로 구성된다.

- 비디오 특징 추출
  내용 기반의 동적 요약을 생성하기 위하여 샷, 씬 등의 비디오 구조 정보와 카메라 움직임, 얼굴 영역, 자막(caption) 영역 등의 비쥬얼 특징 그리고 오디오 정보로부터 추출한 주요 단어 및 주요 문장 특징을 사용한다.
- 요약 비디오 생성
  실제 요약 비디오를 생성하는 과정은 그림 3 과 같이 요약 규칙(rule)에 의하여 추출된 특징 간의 문맥을 분석하여 사건(event)을 검출하는 부분과 이러한 정보를 받아서 최종적으로 요약 비디오에 포함될 구간을 확정하는 부분으로 구성된다.
- 요약 비디오의 발생 및 재생
  요약 생성에서 요약 비디오에 포함된 비디오 구간을 편집하여 실제 요약 비디오 비트열을 발생하고 표현 구조에 따라 요약 비디오를 기술한다. 응용에 따라서 실제적인 요약 비디오 비트열을 발생하지 않고 원 비디오 비트열에서 요약 비디오의 구간만 접근하여 재생하도록 구성할 수도 있다.

2.3 요약 비디오 표현 구조

요약 비디오는 동적인 형태의 요약으로 사

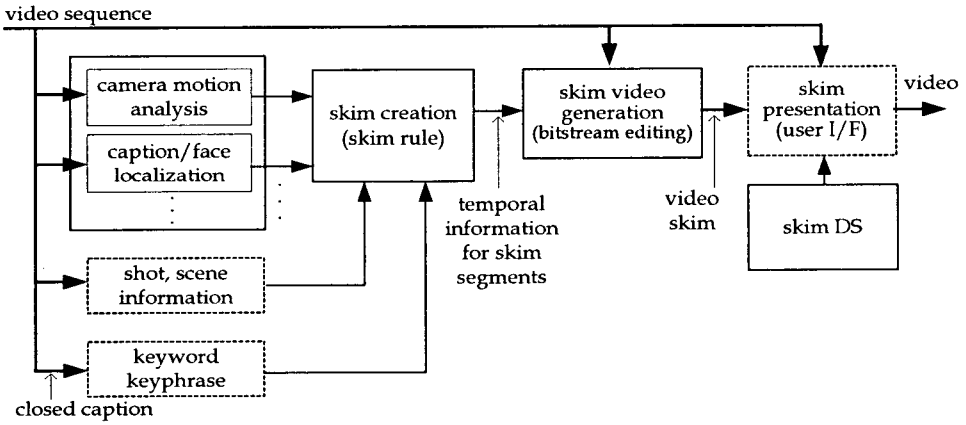

그림 1. 비디오 요약 시스템 전체 기능 구성도

용자로 하여금 효과적으로 대상 비디오의 전반적인 내용을 파악하고 원하는 내용을 쉽게 검색할 수 있도록 하는 하나의 브라우징 도구이다. 이를 위해서 단순히 동적 요약을 제공하기 보다는 효과적인 브라우징과 검색을 가능케 하는 비디오 표현 구조를 갖는 것이 필요하다. 그림 2 와 같은 형태로 구조화된 요약 비디오가 제공될 때 사용자는 다음의 시나리오에 의해 보다 효율적으로 브라우징 및 검색을 수행할 수 있다.

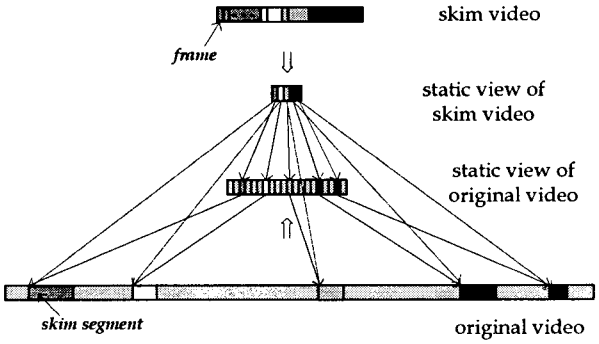

그림 2. 요약 비디오의 표현 구조

- 요약 비디오로 전체의 내용을 개관한다.
- 요약 비디오에서 관심있는 부분이 있으면 요약 비디오의 정적 요약에서 그 구간을 확인하고 정적 요약을 통하여 그 부분의 원 비디오를 브라우징한다.
- 이 때, 그 요약 구간에 해당하는 원 비디오의 양이 많을 경우 계층적 표현 구조를 통하여 그 구간 부근의 원 비디오의 정적 요약을 보고 그 중에서 원하는 샷만을 접근하는 더 효과적인 브라우징을 한다.
- 요약의 압축율이 높아 요약 비디오에서 직접 관련 부분을 접근하기 힘들 때는 조금이라도 관련이 있는 요약 구간을 중심으로 그 근처의 정적 요약을 통하여 브라우징한다.

3. 비디오 특징 추출

3.1 샷 경계의 검출

샷은 비디오 촬영시 카메라 작업이 연속성을 갖는 구간으로 정의되며[4], 통상적으로 비디오 내용 분석의 최소 단위로서 인식된다. 따라서 내용 기반의 비디오 분석은 샷 경계의 검출로 시작되는 경우가 많다. 샷의 경계는 그 연결방식에 따라 두 개의 샷이 바로 연속하여 연결되는 급진적인 장면전환(abrupt change)과 두 개의 연속한 샷이 한 프레임 이상의 간격을 두고 서서히 진행되는 디졸브(dissolve), 와이프(wipe) 등의 점진적인 장면전환(gradual transition)으로 분류된다.

내용 기반의 비디오 분석과 관련한 초기 연구의 상당수는 샷 경계를 검출하기 위한 노력으로[4], 대개 이웃한 프레임 간의 시각적 특징 차이가 샷 내부에서보다 샷의 경계에서 특히 크다는 사실을 이용하고 있다. 현재, 급진적인 장면전환에 있어서는 만족할 만한 결과들이 보고되고 있으며, 점진적인 장면전환의 경우에도 유망한 기법들이 제안되고 있다. 특히, 점진적인 장면전환은 최근 스포츠 비디오에서 재재생부(replay) 등에 와이프와 같은 특수효과 등이 많이 이용되고 있다는 점에서 주요한 검출 대상의 하나로 인식된다.

3.2. 비쥬얼 특징 추출

(1) 자막 및 얼굴영역 검출

비디오에서 발생하는 자막영역은 주로 화면의 내용을 직접적으로 묘사하거나 관련정보를 전달하기 위하여 사용되며, 의미적으로 중요한 부분에서만 발견되는 특징을 가진다. 따라서, 자막영역의 검출은 비디오 요약의 관점에서 매우 유용하게 이용될 수 있다. 표 1 에 실제 실험영상(축구)에서 발생한 자막영역의 실례를 나타내었다.

자막영역이 갖는 공간적인 고대비(high contrast)성[10,11]이나 발생과 소멸시점에 발생하는 영상의 국부적 변화에 관련된 특징[12]을 이용한 검출방법 등이 기존에 제안된 바 있으며, 이들은 검출과정에서 발생하는 잡음적인 요소를 제거하기 위하여 자막영역의 기하정보를 함께 고려하기도 한다[10-12].

표 1. 실험영상(축구)에서의 자막영역의 실례

식별번호	구간 (프레임)		특징
	발생	소멸	
1	381	483	슛 장면 이후 해당선수 이름
2	1898	1982	골인 이후 해당선수 이름
3	2559	2586	골인 이후 해당선수 이름
4	4761	4862	슛 장면 이후 해당선수 이름
5	6043	6112	슛 장면 이후 해당선수 이름
6	7743	7854	슛 장면 이후 해당선수 이름

얼굴영역은 자막과 함께 비디오 색인을 위한 중요한 객체의 하나로 이용될 수 있다[10,13]. 얼굴 영역은 뉴스의 앵커 샷 혹은 드라마의 대화 장면 등과 직접 연관되며, 그 영역의 크기 정보는 상위의 의미를 유추하는 단서를 제공할 수 있다.

(2) 카메라 움직임 분석

카메라 움직임은 샷 또는 부 샷(sub-shot) 단위로 비디오를 색인하기 위한 특징으로 많이 연구되었다[14,15]. 또한 카메라 움직임 분석을 통한 전역 움직임 정보를 이용한 비디오 모자이크[9]는 비디오의 정적 시각화 기법으로 연구되기도 하였다.

카메라 움직임은 방송 자료에 따라서 다른 특징을 갖는 것을 확인할 수 있었는데, 드라마에서는 비교적 카메라 움직임이 미미하여 상대적으로 카메라 움직임이 존재하는 구간 자체가 우선 고려될 수 있고, 특히 인물의 클로즈업(close-up)은 중요한 장면을 포함하는 경우가 많다. 스포츠에서는 상당히 전형적인 카메라 움직임 패턴을 갖는데, 골프에서의 퍼팅 후의 줌(zoom)이 그 일례이다.

카메라 움직임 분석을 위하여 affine 이동 모형을 사용한다. MPEG 비트열에 포함된 이동 벡터로부터 affine 이동 모형의 계수를 추정하고, 그 매핑 계수의 분석을 통하여 카메라 움직임을 분류한다. 블록정합에 의한 MPEG 의 이동 벡터는 예측 부호화 오차의 최소화만을 고려한 것으로 많은 잡음을 포함하고 있으므로 신뢰할 만한 이동 벡터만을 선택하는 전처리를 한다.

(3) 카메라 초점길이 분석 및 샷 패턴 분류

특히, 스포츠를 대상으로 한 비디오는 몇 대의 고정된 카메라에 의하여 번갈아 촬영된 장면으로 구성되는 경우가 빈번하고, 이때 각각의 카메라는 특정 패턴의 장면을 비교적 일관된 초점길이로 촬영한다. 여러 실험영상의 관찰 결과, 화면의 원근을 1 – 4 정도의 양자화된 준위로 분류하고 각 샷의 패턴을 대략적으로 분류하는 작업이 비디오의 내용 파악에 상당히 유용하다.

(4) 느린 움직임 검출 및 활성도(activity) 추정

비디오에서, 느린 움직임(slow motion)을 통한 재생은 주요한 장면이 발생된 경우 이를 시청자에게 보다 자세히 전달하기 위한 목적으로 자주 이용된다. 또한, 의미적으로 중요한 부분의 활성도가 그렇지 않은 부분의 활성보다 크다는 사실도 비디오의 주요부분을 신호차원에서 찾아내는 데 효과적인 단서로 이용될 수 있다[16].

4. 요약 비디오 생성

요약 비디오의 생성은 추출한 비디오 특징 정보에 기반하여 요약 생성 규칙을 설정하고 그 규칙에 따라서 요약 비디오에 포함되는 비디오 구간을 선별한다. 그림 3 에서 보는 바와 같이 요약 생성 규칙은 비디오 요약을 위하여 검출할 사건의 집합을 규정하고, 실제 검출된 사건을 비디오 요약에 포함시킬 때 사용자의 이해 환경을 고려한 구간 선택 규칙을 정의한다. 요약 생성 규칙은 적용 범위에 따라서 일반적인 규칙과 대상 자료에 제한적인 규칙을 고려할 수 있다.

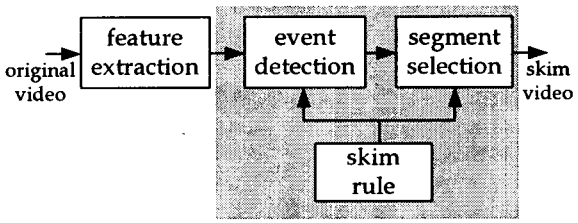

그림 3. 비디오 요약 생성 기능 블록도

4.1 사건 검출

영상의 내용 관점에서 중요한 부분이란 대체로 임의의 사건이 발생한 구간을 의미하는 경우가 많다. 축구 영상에서 골 또는 슛 등의 사건이 발생한 장면 등이 그 일례이다. 사건은 대개 상위의 의미적인 요소를 반영하므로 시각적인 특징에 의하여 직접적으로 검출하기가 어렵다. 대신, 여러 시각적 특징을 종합하여 그 문맥에 의하여 미루어 추정할 수 있는 경우가 종종 발견되는데, 각각의 개별적 특징은 사건 검출의 기반 정보로 이용되며, 실제로 비디오에서 그러한 특징들이 검출되었을 때, 사건의 발생확률은 증가한다.

스포츠 비디오에서 재재생부는 대개 내용을 자세히 전달하기 위하여 느린 움직임으로 구성되는 경우가 많으며, 시간적으로 역순임을 분명히 하기 위하여 샷의 경계가 디졸브나 와이프 등의 점진적인 전이 형태를 사용하여 편집되는 경우가 많다. 따라서, 이러한 샷 경계의 시각적 특징들을 통하여 재재생되는 샷으로 판정될 수 있다. 슛 등의 다른 사건들도 그 해당 샷이나 부근의 시각적 특징을 종합한 문맥에 의하여 검출될 수 있다. 비디오 특징을 통한 축구에 대한 사건검출의 예를 그림 4 에 나타내었다.

4.2 일반적인 요약 규칙

요약 비디오는 결국 편집을 통하여 얻을 수 있는 하나의 새로운 비디오 시퀀스이므로 이를 구성하는 각 구간들 간의 연결은 자연스럽게 재생되도록 하여야 한다. 이를 위하여 요약 구

간은 가급적 샷의 경계를 벗어나지 않도록 하여 시각적으로 거슬림을 방지한다. 또한, 한 구간의 길이에 대한 하한값을 설정하여 시청각 정보의 의미가 충분히 전달됨에 무리가 없도록 한다.

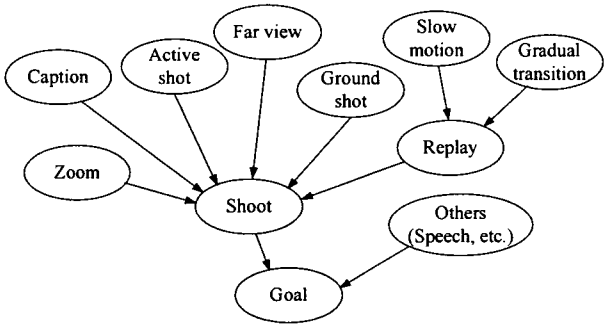

그림 4. 비디오 특징을 통한 사건검출 예(축구)

4.3 영역 제한적 규칙

드라마, 스포츠, 뉴스 방송 비디오를 대상으로 한 영역 제한적 규칙의 예를 기술한다.

(1) 드라마

드라마는 방송대본[17]의 형식에서 보듯이 씬의 단위로 줄거리가 전개된다. 드라마에서는 중요한 사건을 객관적으로 정의하기가 어려우며, 이야기 형식의 사건은 시각적 특징들만으로 유추할 수 없는 형태가 대다수이다. 따라서, 자동으로 드라마의 줄거리를 요약하기는 상당히 어려우므로, 줄거리를 통한 이야기의 흐름을 전달하기 보다는 전반적인 내용을 빠지지 않게 골고루 포함하여 전체적인 내용을 표현할 수 있도록 요약한다.

구체적으로 씬의 분할이 선행되어[7,8] 각 씬에서 가급적 하나 이상의 구간을 선택하며, 구간 선택은 자주 재현되는 장면이나 얼굴의 클로즈업, 활발한 카메라와 객체의 움직임이 있는 등 시각적 특징에서 유력한 구간을 취한다.

(2) 스포츠

스포츠의 경우는 드라마와 달리 중요한 부분이 비교적 분명하며, 시각적 특징을 바탕으로 한 사건 검출의 가능성도 높다. 예를 들면, 골프의 경우 카메라 움직임 패턴에 의하여 인식되는 샷 동작이나 퍼팅을, 축구는 그림 4 의 방식에 의한 골 또는 슛 장면을 주요장면으로서 채택할 수 있다. 사용자의 이해를 위하여 주요 장면 이외에도 자막이 검출되는 구간 등 이들을 묘사하는 부분을 적절히 요약 비디오에 포함되어야 한다.

(3) 뉴스

뉴스는 각 기사별로 구성되며, 대부분의 경우 앵커 샷이 각 뉴스 항목의 서두에 반복적으로 위치하는 구성을 취한다. 뉴스 항목의 제목 및 주요 정보가 자막으로 재현되는 점을 이용하여 자막이 나타나는 구간을 주요구간으로 설정하고, 앵커 샷, 인터뷰 샷, 카메라 플래쉬가 존재하는 구간[12] 등을 포함하는 형태의 요약 비디오를 구성할 수 있다.

4.4. 모의실험

5 분 길이의 축구 비디오에 대하여 요약 규칙을 수작업의 형태로 구현하여 적용한 모의실험 결과를 제시한다. 앞에서 기술한 비디오 특징 추출이 자동화될 수 있음을 가정하고, 다음과 같은 요약 규칙을 적용하여 7.5:1 정도로 압축한 40 초 내외의 요약 비디오를 구현하였다.

축구 비디오에 대하여 그림 5 와 같은 분석을 통하여 시각적 특징의 문맥에 의한 사건검출의 가능성에 관한 실험도 병행하였는데, 우선 앞서 기술한 비디오 특징 추출에 의하여 다음과 같은 결과를 얻을 수 있음을 가정하였다.

- 샷 경계 검출 및 각 샷의 패턴을 색상이나 질감 등에 의하여 운동장 샷(G), 벤치 샷(B), 응원석 샷(S)의 3 부류로 분류
- 자막영역의 발생 및 소멸시점 검출
- 원경에서 근경의 카메라 초점거리를 1 – 4 (0.5 단위)의 양자화된 준위로 분석
- 재재생부의 느린 움직임 검출

그림 5 는 슛이나 골 장면은 원경(카메라 초점거리 1-2)의 운동장 샷(G)에서 주로 나타나고 있으며, 그 직후 카메라 초점거리가 상당히 짧아지고 자막과 재재생부가 검출됨을 알 수 있다. 검출된 자막과 재재생부로부터 이에 상응하는 원경의 운동장 샷을 찾으면 슛 또는 골 등을 포함하여 경기상 중요한 사건이 검출된다 또한 그 사건에 관련된 내용의 경계를 구분지을 수 있으며, 그 경계의 길이나 청각 등의 기타 다른 특징들로부터 지지도를 분석하여 골, 슛 등을 구분하는 것이 가능함을 알 수 있다.

비디오 요약에 대한 성능 판단 기준으로 기능성(functionality), 사용자 친밀도(user intimacy), 신축성(scalability), 효율성(efficiency) 등이 고려될 수 있다. 이 중, 기능성은 요약 비디오가 원비디오의 주요부분을 충분히 포함하면서 동시에 중요하지 않은 부분을 배제하여 압축성을 달성하여야 함을 나타내며, 이는 각 사건의 검출율에 의하여 정량적으로 측정할 수 있다.

5. 결론 및 검토

본 논문에서는 대화형 방송 환경에서 비디오 프로그램에 대한 효과적인 브라우징 및 검색을 위한 동적 비디오 요약 기법을 기술하였다. 요약 비디오 생성을 위한 동적 비디오 요약의 전체적인 시스템 구조와 각 요소기술을 기술하였으며, 또한 정적 요약을 포함한 복합 형태의 요약 비디오 표현구조를 통하여 저장된 방송 자료에 대한 효과적인 검색 및 요약 프로그램 안내 등의 응용을 제시하였다. 방송 자료의 관찰 및 모의실험 결과, 비디오 특징 분석 및 요약 규칙 개발에 의한 요약 비디오 자동 생성의 가능성을 확인하였다.

감사의 글

본 연구는 정보통신부 출연 "MPEG-7 표준화 기술 개발" 과제로 수행되었으며, 연구에 많은 도움을 주신 실감 AV 팀원들께 감사 드립니다.

참고문헌

[1] DAVIC' s 1.5 series Application Scenarios, available at http://www.davic.org/
[2] ISO/IEC JTC1/SC29/WG11, MPEG-7 Requirements Document v.8, N2727, Mar. 1999.
[3] Gist EPG, http://www.tv1.com/tv
[4] B. -L. Yeo and B. Liu, Rapid Scene Analysis on Compressed Video, *IEEE Trans. CSVT*, vol.5, no.6, Dec. 1995.
[5] M. Yeung and B. Liu, Efficient Matching and Clustering of Video Shots, in *Proc. ICIP'95*, vol.1, Oct. 1995.
[6] H. S. Chang, S. Sull and S. U. Lee, Set-theoretic Approach to Video Key Frame Extraction, in *Proc. ITC-CSCC'98*, vol.1, July 1998.
[7] M. Yeung and B. -L. Yeo, Time-constrained Clustering for Segmentation of Video into Story Units, in *Proc. ICPR'96*, vol.C, Aug. 1996.
[8] Y. Rui, T. S. Huang and S. Mehrotra, Exploring Video Structure beyond the Shots, in *Proc. ICMCS'98*, June 1998.
[9] D. Yow, B. -L. Yeo, M. Yeung and B. Liu, Analysis and Presentation of Soccer Highlights from Digital Video, in *Proc. ACCV'95*, Dec. 1995.
[10] M. Smith and T. Kanade, Video Skimming and Characterization through the Combination of Image and Language Understanding Techniques, in *Proc. CVPR'97*, June 1997.
[11] B. T. Chun *et al.*, Video Caption Extraction Method using Topographical Features of Characters, A proposal for MPEG-7, Feb. 1999.
[12] B. -L. Yeo, Efficient Processing of Compressed Images and Video, Ph.D. dissertation, Jan. 1996.
[13] H. Wang and S. -F. Chang, A Highly Efficient System for Automatic Face Region Detection in MPEG Video, *IEEE Trans. CSVT*, vol.7, no.4, Aug. 1997.
[14] M. V. Srinivasan, Qualitative Estimation of Camera Motion Parameters from Video Sequences, *Pattern Recognition*, 1997.
[15] M. Pilu, On Using Raw MPEG Motion Vectors to Determine Global Camera Motion with Applications to Image Mosaicing, in *Proc. VCIP'98*, vol.SPIE-3309, Jan 1998.
[16] D. Manoranjan and V. V. Vinod, Video Segment Activity, A descriptor proposal for MPEG-7, Feb. 1999.
[17] 방송대본, http://www.mbc.co.kr/drama/script

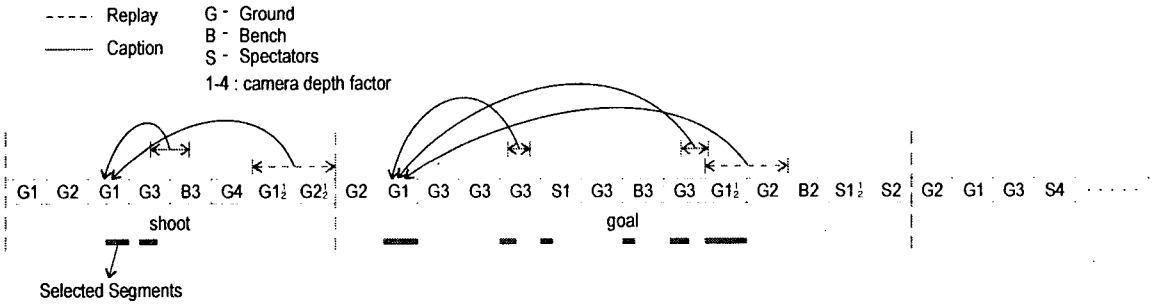

그림 5. 시각적 특징의 문맥에 의한 사건 검출 실례 (축구)