Journal of Korea Multimedia Society Vol. 7, No. 10, October 2004(pp. 1452-1463)

다중 비주얼 특징을 이용한 어학 교육 비디오의 자동 요약 방법

한희준[†], 김천석[††], 추진호[†††], 노용만[††††]

요 약

양방향 방송 서비스로의 전환을 맞아 다양한 사용자 요구 및 기호에 적합한 컨텐츠를 제공하고, 증가하는 방송 컨텐츠를 효율적으로 관리, 이용하기 위해 비디오의 자동 요약에 대한 요구가 증가하고 있다. 본 논문에서는 내용 구성이 잘 갖추어진 어학 교육 비디오의 자동 요약에 대한 방법을 제안한다. 내용 기반 요약을 자동으로 생성하기 위해 먼저 디지털 비디오로부터 샷 경계를 검출한 후, 각 샷을 대표하는 키프레임으로부터 비주얼 특징들을 추출한다. 그리고 추출된 다중 비주얼 특징을 이용해 어학 교육 비디오의 세분화된 내용 정보를 결정한다. 마지막으로, 결정된 내용 정보를 기술하는 요약문을 MPEG-7 MDS(Multimedia Description Scheme)에 정의된 계층적 요약 (Hierarchical Summary) 구조에 맞추어 XML 문서로 생성한다. 외국어 회화 비디오에 대해 실험하여 제안한 자동 요약 방법의 효율성을 검증하였으며, 제안한 방법이 교육 방송용 컨텐츠의 다양한 서비스 제공 및 관리를 위한 비디오 요약 시스템에 효율적으로 적용 가능함을 확인하였다.

Automatic Summary Method of Linguistic Educational Video Using Multiple Visual Features

Hee-Jun Han[†], Cheon-Seog Kim[††], Jin-Ho Choo[†††], Yong-Man Ro[††††]

ABSTRACT

The requirement of automatic video summary is increasing as bi-directional broadcasting contents and various user requests and preferences for the bi-directional broadcast environment are increasing. Automatic video summary is needed for an efficient management and usage of many contents in service provider as well. In this paper, we propose a method to generate a content-based summary of linguistic educational videos automatically. First, shot-boundaries and keyframes are generated from linguistic educational video and then multiple(low-level) visual features are extracted. Next, the semantic parts (Explanation part, Dialog part, Text-based part) of the linguistic educational video are generated using extracted visual features. Lastly the XML document describing summary information is made based on Hierarchical Summary architecture of MPEG-7 MDS(Multimedia Description Scheme). Experimental results show that our proposed algorithm provides reasonable performance for automatic summary of linguistic educational videos. We verified that the proposed method is useful for video summary system to provide various services as well as management of educational contents.

Key words: Automatic Summarization(자동 요약), MPEG-7, Video Indexing(비디오 색인)

※ 교신저자(Corresponding Author) : 한희준, 주소 : 대전광역시 유성구 어은동 52번지(305-600), 전화 : 042)828-5181, FAX : 042)869-0919, E-mail : hhj@kisti.re.kr
접수일 : 2003년 11월 28일, 완료일 : 2004년 3월 23일
[†] 준회원, 한국과학기술정보연구원(KISTI) 정보시스템부 근무
[††] 준회원, 한국정보통신대학교 대학원 공학부 박사과정
(E-mail : cheonseog@icu.ac.kr)
[†††] 삼성전자 디지털 미디어 총괄 근무
(E-mail : jinho.choo@samsung.com)
[††††] 종신회원, 한국정보통신대학교 공학부 부교수
(E-mail : yro@icu.ac.kr)
※ 본 연구는 정보통신부의 지능형 통합정보방송 (SmarTV) 기술 개발 사업의 일환으로 수행된 연구결과 입니다.

1. 서 론

현재 방송 서비스는 기존의 수동적인 단방향 방송에서 벗어나 통신 서비스와의 융합을 모색하고 있으며, 소비자의 요구 사항을 만족시키는 양방향 방송 환경으로의 전환을 맞이하고 있다. 양방향 방송 서비스는 소비자의 기호 및 성향에 적합한 프로그램, 요약된 컨텐츠 및 하이라이트 장면 등을 효율적으로 제공할 수 있어야 한다. 시청자의 다양한 취향에 적합한 컨텐츠 제공에 대한 요구가 증가하면서 방송 컨텐츠에 대한 요약, 검색 및 색인 기술 연구가 활발히 수행되고 있다. 또한 방송 컨텐츠의 양이 증가함에 따라 서비스 제공자는 효율적인 컨텐츠 관리와 데이터 베이스 구축을 위해 비디오 요약 기술을 필요로 하게 되었다.

지금까지 비디오 요약에 관련된 많은 연구가 진행되어 왔으며, 대부분 비디오로부터 하위 수준의 비주얼 특징들을 추출, 분석하여 상위 수준의 이벤트를 결정함으로써, 내용 기반으로 비디오를 요약하는 방법이 주류를 이룬다. 특히, 스포츠 비디오나 뉴스는 내용구조가 잘갖추어져 있기 때문에 칼라 및 에지, 움직임 정보 등을 이용한 내용 기반 요약에 대한 많은 연구가 실행되어 왔다[1-8]. 반면에, 증가하는 어학 비디오는 그 구조가 간결하고 정형화 되어 있지만, 비디오 요약에 대한 연구는 거의 미흡한 실정이다.

따라서 본 논문에서는 다채널 환경 및 교육열 고취와 더불어 그 양이 급속히 증가하는 어학 교육 비디오에 대한 자동 요약 방법을 논한다. 어학 교육 비디오는 자동 요약을 통해 효율적인 서비스를 제공할 수 있고, 방대한 컨텐츠를 효과적으로 관리할 수 있다. 본 논문에서는 어학 교육 비디오의 비주얼 특성을 파악하여 내용 기반 특징들을 추출하고, 추출된 특징을 조합하여 어학 비디오의 세분화된 내용 정보를 자동 생성한 후 의미있는 요약 결과를 도출하였다. 이용한 비주얼 특징 중에서 일부는 국제 표준으로 정의된 MPEG-7 기술자로써, 향후 재사용성 및 호환성을 고려하였다.

논문의 구성은 다음과 같다. 2절에서는 어학 교육 비디오의 내용 분석에 대해, 3절에서는 사용된 비주얼 특징들과 조합 방법, 요약에 필요한 비디오 내용 정보의 세분화 알고리즘을 포함하는 제안된 요약 방법에 대해 설명하고, 4절에서는 제안된 방법의 유효성 검증을 위해 MPEG-2 형식의 외국어 회화 비디오를 이용한 실험 및 요약 결과를 기술한다. 마지막으로 5절에서는 결론 및 향후 과제에 대해 논한다.

2. 비디오 분석

비디오 요약을 위해 의미 있는 내용 정보 분석이 선행되어야 한다. 어학 비디오는 일반적으로 그림 1과 같이 사회자가 일반 강의를 이루는 설명 부분(Explanation part), 외국인이 서로 대화하는 대화 부분(Dialog part), 텍스트 정보로 이루어진 지문 부분(Text-based part)과 그 밖의 기타 정보를 담고 있는 잔여 부분(Remain part)으로 구성된다. 각 부분은 의미를 지닌 요소로 이루어지는데 그림 2와 같이 설명 부분은 일정한 장소(스튜디오)에 위치한 사회자가 강의를 한다. 그리고 대화 부분은 사회자가 위치한 장소가 아닌 실외 및 실내의 다른 장소에서 외국인들끼리의 대화를 나타내며, 지문 부분은 비디오 화면에 텍스트 정보가 주류를 이룬다.

설명 부분은 말하기, 문법 및 발음에 대한 어학 학습내용을 담고 있으며, 대화 부분은 주로 듣기 학습에 유용한 내용으로 구성되어 있다. 또한 지문 부분은 어학 학습에 있어서 문법 및 단어학습에 중요성을 가진다. 잔여 부분은 프로그램 타이틀이나 제작자 정보를 포함하고, 또는 화면 전환 부분을 나타내는데 이는 중요 교육 정보를 담고 있지 않을 뿐만 아니라, 사용자의 선호에 부합하는 교육 내용이 아니기 때문에 어학 교육 비디오 요약을 위한 생성 부분으로 간주하지 않는다. 이런 교육용 컨텐츠는 특히 소비자의 어학 학습에 있어서 부족한 영역에 대한 선호 정도가 상대적으로 크게 발생하는 특징을 지닌다. 따라서 본

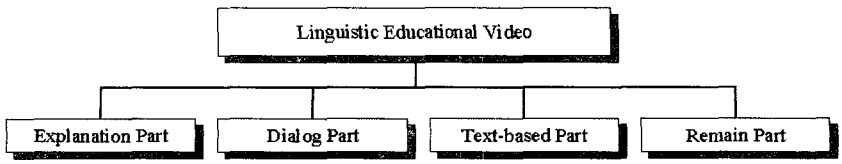


그림 1. 교육용 어학 비디오의 내용 세분화

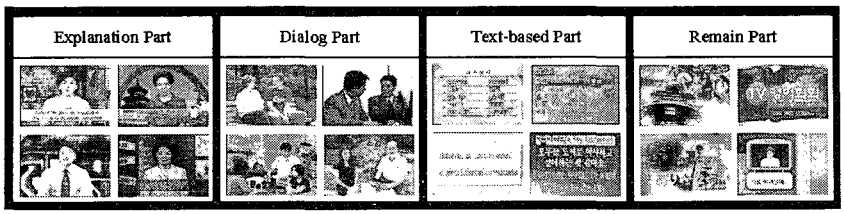


그림 2. 각 세분화된 내용에 해당하는 프레임의 예

논문은 사용자 선호에 적합한 요약된 컨텐츠를 제공하기 위하여 어학 비디오로부터 설명 부분, 대화 부분, 지문 부분을 검출해 내용 기반의 요약 정보를 자동으로 생성하는 것을 목표로 한다.

3. 세부 알고리즘

본 절에서는 2절에서 정의한 어학 교육 비디오의 세부 내용 검출을 통하여 요약 정보를 생성하기 위한 알고리즘에 대해 논한다. 그림 3은 교육용 어학 비디오 요약 시스템 구조를 보여준다. 입력된 비디오로부터 먼저 샷 경계(Shot boundary)를 검출하고 각 샷을 대표하는 키프레임(Key frame)을 추출한다. 추출된 키프레임으로부터 요약에 필요한 비주얼 특징들을 추출한 후 미리 정의한 세분화된 내용 정보(설명 부분, 대화 부분, 지문 부분)를 생성한다. 최종적으로 생성된 내용 정보는 MPEG-7 MDS에 정의된 계층적 요약 (Hierarchical Summary) 구조에 맞추어 XML 문서 형식으로 표현된다[9].

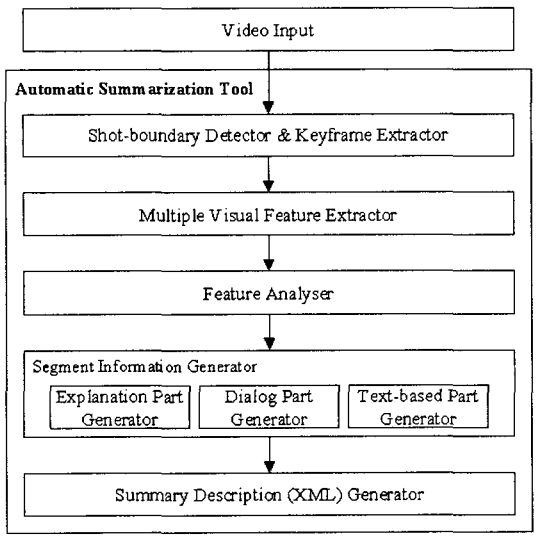


그림 3. 요약 시스템 구조

3.1 샷 경계 검출 및 키프레임 추출

어학 교육 비디오의 세분화된 내용 정의가 선행된 후에, 입력 비디오로부터 비주얼 특징 추출의 기본 단위인 샷과 키프레임이 추출된다. 비디오를 이루는 샷, 프레임의 질감과 에지 정보는 색상 정보와 함께 영상 정보를 기술하는 중요한 시각적인 특징이다. 이 질감과 에지 정보는 영상의 구조, 방향성, 거친 정도 등을 나타내고 비디오 데이터를 내용 기반 요약하기 위한 샷 경계 검출에 있어서 중요한 특징으로 이용될 수 있다[10-12].

본 알고리즘에서 사용한 샷 경계 검출 모듈은 MPEG-7 참조 소프트웨어 XM(eXperiment Model)에서 분리한 Hierarchical Summary DS 안의 샷 경계 검출 루틴을 개선하여 사용하였다[13,14]. 샷 경계 검출은 유사한 분포를 가진 특징 정보, 예를 들면 색상, 질감, 에지 등에 의해 검출되므로, 계산 복잡도를 줄이기 위해, 각 샷을 이루는 프레임들 중에서 중간 프레임을 해당 샷을 대표하는 키프레임으로 결정한다. 각 샷으로부터 키프레임을 구하는 루틴은 다음과 같다.

$$for \quad n=1 \quad through \quad TotalShotNum \quad do \quad \{ \\ keyframe_n = (Startframe_n - Endframe_n)/2 \quad \} \qquad (1)$$

여기서 $TotalShotNum$은 입력된 비디오로부터 검출된 샷의 개수, $Startframe$은 샷의 시작 프레임 번호를 나타내고 $Endframe$은 샷의 끝 프레임 번호를 의미한다.

$$L = \{s_1, s_2, s_3, ..., s_i\}, i = shot \quad number \\ k_i = keyframe \quad of \quad s_i \qquad (2)$$

교육용 어학 비디오를 L이라 하고 s를 샷이라 하면, 식 (2)와 같이 L은 s의 집합으로 표현되며 k는 해당 샷을 대표하는 키프레임이 된다.

3.2 적용되는 다중 비주얼 특징

3.2.1 색도 히스토그램 분포의 첨도 (Kurtosis of hue histogram distribution)

색도 히스토그램 분포의 첨도는 지문 부분의 동일한 배경색을 검출하기 위해 사용된다. 색도 히스토그램은 이미지의 색도 분포를 나타낸다. 이미지의 RGB 값으로부터 hue 값을 구한 후 360 빈마다의 데이터 분포 정도를 얻게 된다[15]. 첨도는 분포의 뾰족한 정도를 나타내는 지표이며, 식 (3)은 각 샷을 대표하는 키프레임으로부터 색도 히스토그램 분포의 첨도를 구하는 방법이다.

$$H_{k_i} = \{h_i^1, h_i^2, h_i^3, \ldots, h_i^n\}, \quad n = 360$$

$$Kur_{k_i} = \left\{ \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{p=1}^{n} \left(\frac{h_i^p - \bar{h}_i}{SD_i} \right)^4 \right\} - \frac{3(n-1)^2}{(n-2)(n-3)}$$

$$where \quad SD_{k_i} = \sqrt{\frac{n\sum_{q=1}^{n}(h_i^q)^2 - \left(\sum_{q=1}^{n} h_i^q\right)^2}{n^2}} \tag{3}$$

여기서 H_{k_i}는 i번째 키프레임 k_i로부터 추출한 색도 히스토그램 데이터 h_i 값들의 집합이다. SD_{k_i}는 h_i로부터의 표준편차이며, Kur_{k_i}는 k_i의 색도 히스토그램 분포의 첨도값이다.

3.2.2 질감 기술자의 채널 에너지 표준편차(30-channel energy Standard deviation of HomogeneousTexture Descriptor)

질감 정보는 이미지의 균질성을 나타내는 패턴을 정의하며, 어학 비디오 지문 부분을 이루는 프레임들의 이미지 균질성을 특징짓는다. 질감 특징 정보 표현을 위해 이미지를 라돈 변환하고, 투영된 데이터에 대해 1차원 푸리에 변환 (Fourier transform)을 거친다. 그리고 이미지의 주파수 영역을 6개 방향 성분과 5개 크기 성분으로 나누어 30개 채널에 대한 에너지를 구한다[15]. 식 (4)는 질감 정보 추출을 위해 키프레임들의 30개 채널 에너지 표준편차를 구하는 식이다.

$$E_{k_i} = \{e_i^1, e_i^2, e_i^3, \ldots, e_i^m\}, \quad m = 30$$

$$SD_{k_i\_ChannelEnergy} = \sqrt{\frac{m\sum_{r=1}^{m}(e_i^r)^2 - \left(\sum_{r=1}^{m} e_i^r\right)^2}{m^2}} \tag{4}$$

여기서 E_{k_i}는 k_i로부터 추출한 채널 에너지 e_i의 집합이며, $SD_{k_i_ChannelEnergy}$는 E_{k_i}의 원소 e_i로부터의 표준편차이다.

3.2.3 에지 히스토그램 기술자(Edge Histogram Descriptor)

에지 히스토그램 기술자에 정의된 비주얼 특징은 설명 부분의 사회자의 모습을 특징짓기 위해 사용된다. 이미지의 에지 정보를 나타내기 위해 먼저 이미지를 16개의 서브 블록으로 나눈 후, 각 블록에 대해 모두 5개의 에지 성분인 수직(vertical), 수평(horizontal), 45°, 135°, 무방향성(non-directional)을 기술한다. 각 블록의 에지 성분을 조합하여 한 이미지로부터 모두 80개의 로컬 에지 히스토그램(local edge histogram)을 얻으며, 16개 하위 블록의 로컬 에지 히스토그램을 조합하여 각각 40개의 세미-글로벌 에지 히스토그램(semi-global edge histogram)과 5개의 글로벌 에지 히스토그램(global edge histogram)을 구성한다. 에지 히스토그램 기술자의 특징 벡터와 특징벡터를 이용한 유사도 거리값은 아래와 같다.

$$\overline{ED} = \begin{cases} f_{local\_1}, f_{local\_2}, \cdots, f_{local\_80} \\ f_{semi-global\_1}, f_{semi-global\_2}, \cdots, f_{semi-global\_40} \\ f_{global\_1}, f_{global\_2}, \cdots, f_{global\_5} \end{cases} \tag{5}$$

$$dist_{ED} = \sum_i \left| \overline{ED}_{keyframe}(i) - \overline{ED}_{keyframe'}(i) \right| \tag{6}$$

여기서 는 에지 히스토그램 기술자의 특징 벡터이다. f_{local_s}는 s번째 로컬 에지 히스토그램 빈을 나타내고, $f_{semi-global_t}$는 t번째 세미-글로벌 에지 히스토그램 빈을, f_{global_u}는 u번째 글로벌 에지 히스토그램 빈을 나타낸다[15]. 그리고 $dist_{ED}$는 각 키프레임들간의 유사도를 측정하는데 이용되는 에지 히스토그램 특징을 이용한 거리이며, $keyframe$과 $keyframe'$는 서로 다른 키프레임을 의미한다.

3.2.4 스케일러블 칼라 기술자(Scalable Color Descriptor)

스케일러블 칼라 기술자에 정의된 비주얼 특징은 지문 부분 및 대화 부분의 칼라 특징으로부터 설명 부분의 칼라 특징을 구분하기 위해 사용된다. 이 특징은 이미지내의 칼라의 분포를 나타내는 히스토그

램으로 표현된다. 이미지의 RGB 값들은 HSV 값으로 비선형 변환되며, HSV 칼라 공간을 모두 256개의 빈으로 나누고, 각 빈에 속하는 픽셀의 수를 측정함으로써 특징벡터를 구성한다[15]. 칼라 기술자 특징을 이용한 거리값은 각 프레임들간의 유사도를 측정하는데 이용되고 식 (7)에 의해 계산된다.

$$dist_{CD} = \sum_{i} \left| \overline{CD}_{keyframe}(i) - \overline{CD}_{keyframe'}(i) \right| \quad (7)$$

여기서 $dist_{CD}$는 칼라 기술자 특징값을 이용한 프레임간의 유사도 거리이며, *keyframe* 과 *keyframe'* 는 서로 다른 키프레임을 의미한다.

3.2.5 움직임 강도(Intensity of Motion Activity)

움직임 강도(Motion Intensity)는 비디오 시퀀스 내 객체의 움직임 정도를 일정 범위에 걸쳐 표현해 주는 특징이다. 움직임 강도는 각 프레임의 매크로블록으로부터 구해진 움직임 벡터들의 크기(Motion vector magnitude)를 프레임 해상도(Frame resolution)로 적절히 정규화하고 양자화시킨 값들의 표준편차이다[15]. 이 특징값은 비교적 객체의 움직임이 큰 대화 부분과, 샷을 구성하는 객체의 움직임이 상대적으로 거의 없는 설명 부분을 구분하기 위하여 사용된다. 다른 비주얼 특징 추출과는 달리 이것은 샷 단위로 연산되며, 비디오의 샷들로부터 움직임 강도를 구하는 방법은 아래와 같다.

$$mv_{mag} = \sqrt{mv_x^{\,2} + mv_y^{\,2}}$$

$$Intensity_{motion} = \sqrt{\frac{\sum_{u=1}^{w \times h \times n} (mv_{mag})^2}{w \times h \times n} - \left(\frac{\sum_{u=1}^{w \times h \times n} mv_{mag}}{w \times h \times n}\right)^2} \quad (8)$$

여기서 mv_{mag}는 움직임 벡터의 크기이며 는 수평방향 움직임 벡터, mv_y는 수직방향 움직임 벡터를 나타낸다. w와 h는 각각 프레임의 폭과 높이, n은 샷을 구성하는 프레임의 개수이고 움직임 강도 $Intensity_{motion}$는 mv_{mag}의 표준편차를 구하여 얻어진다.

3.3 비디오의 세분화된 내용 검출

본 절에서는 앞에서 설명한 다중 비주얼 특징을 이용하여 어학 교육 비디오의 설명 부분, 대화 부분, 지문 부분을 검출하는 방법을 논한다. 입력 비디오로부터 비주얼 특징 추출의 기본 단위가 되는 샷 및 키프레임을 검출한 후, 추출된 특징을 이용하여 설명 부분, 대화 부분 및 지문 부분을 나타내는 샷과 키프레임을 결정한다. 설명 부분, 대화 부분 및 지문 부분에 해당하는 각각의 세그먼트 정보가 생성되면 최종적으로 MPEG-7 계층적 요약 구조에 맞추어 교육용 어학 비디오에 대한 요약 정보를 표현하는 XML 문서를 생성한다.

3.3.1 지문 부분 검출 방법

지문 부분을 검출하기 위하여 각 샷을 대표하는 키프레임들로부터 색도 히스토그램의 첨도와 MPEG-7 질감 기술자(Homogeneous Texture Descriptor)에 정의된 30채널 에너지 표준 편차를 이용한다. 그림 4는 지문 부분을 검출하여 요약 정보를 생성하기 위한 흐름을 간단히 보여준다. 지문 부분을 검출하기 위한 비주얼 특징값이 임계값 조건을 만족하면 해당 특징을 가지는 키프레임으로부터 지문 부분을 구성하는 샷을 구성하게 된다. 그리고 샷의 세그먼트 정보를 이용해 지문 부분에 해당하는 요약 정보를 생성한다.

지문 부분을 구성하는 샷들을 대표하는 프레임들은 거의 동일한 색상의 배경을 가지며, 그 위에 텍스트 정보가 나타난다. 설명 부분이나 대화 부분의 키프레임으로부터 얻은 색도 히스토그램과는 달리 지문 부분의 키프레임은 동일한 배경색 때문에 그림 5의 좌측처럼 색도 히스토그램 데이터는 일정 부분에 편중되어 존재하고, 분포는 뾰족한 성향을 띈다. 따라서 색도 히스토그램 분포의 첨도값은 큰 양수를 가진다. 그림 5의 우측은 어학 비디오로부터 구해진 샷들을 대표하는 각각의 키프레임들이 가지는 색도

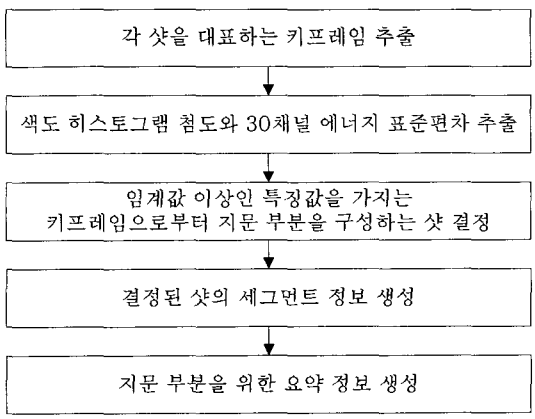


그림 4. 지문 부분 검출 흐름도

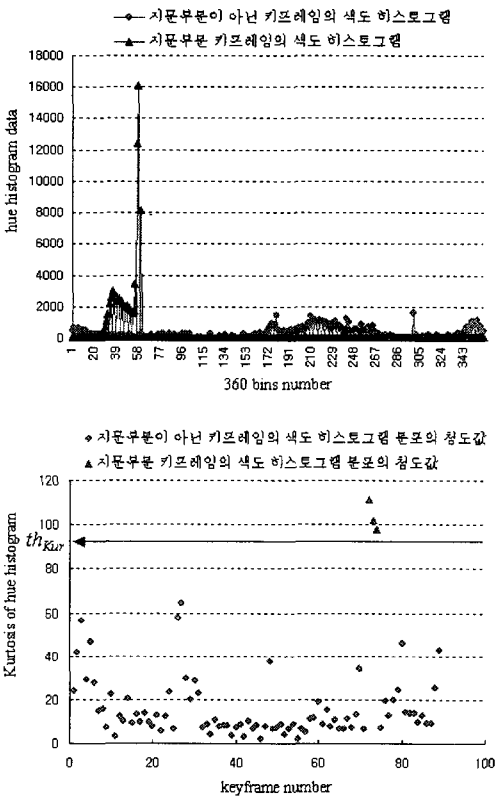


그림 5. 색도 히스토그램 분포 및 첨도값

히스토그램 분포의 첨도값을 보여준다. 그림에서처럼 지문 부분이 아닌 부분을 대표하는 키프레임이 가지는 첨도는 대부분 0~40 사이의 값을 가지며, 지문 부분의 키프레임은 약 95 이상의 첨도값을 가진다. 따라서 첨도 임계값 th_{Kur} 이상의 값을 가지는 키프레임을 지문 부분의 샷을 대표하는 키프레임으로 결정하게 된다.

또한 지문 부분의 키프레임은 거의 동일한 색상 정보로 이루어져 있기 때문에, 질감 기술자에 정의된 30채널 에너지 데이터는 그림 6의 좌측처럼 150~200 사이의 좁은 범위를 가진다. 반면 지문 부분이 아닌 키프레임에서 얻은 채널 에너지 데이터는 비교적 분포 범위가 넓다. 따라서 그림 6의 우측에서와 같이 지문 부분을 대표하는 키프레임의 채널 에너지 표준편차는 다양한 색상으로 이루어진 설명 부분이나 대화 부분에 비해 상대적으로 작은 값을 가지게 된다. 각 샷을 대표하는 프레임들은 식 (4)에 의해 채널 에너지 표준편차 값을 각각 하나씩 가지는데, 이 값이 임계값 $th_{SD\_ChannelEnergy}$ 이하의 값을 가지게 되면 해당 키프레임을 지문 부분을 구성하는 샷으로부터 얻어진 것으로 결정한다. 최종적으로 동시에 두 가지 임계값 th_{Kur}, $th_{SD\_ChannelEnergy}$ 을 만족(두 조건의 AND 연산)하는 키프레임들이 검출되고 그에 해당하는 샷들은 지문 부분을 구성하는 세그먼트로 생성된다.

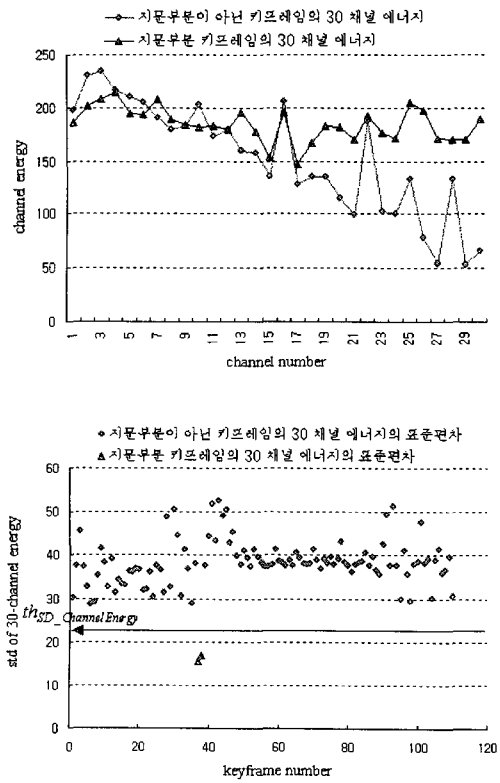


그림 6. 질감 기술자의 30채널 에너지와 에너지 표준편차

3.3.2 설명 부분 검출 방법

어학 교육 비디오로부터 설명 부분을 검출하는 과정은 그림 7과 같은 흐름을 가진다. 설명 부분을 검출하기 위해서 3.2.3~3.2.5 절에서 설명한 세 가지 특징이 사용되며, 특징 분석에 의해 설명 부분에 해당하는 샷을 결정하고 세그먼트 정보를 생성한다.

샷 검출 및 키프레임 추출 후에 각각의 키프레임들로부터 먼저 에지 히스토그램과 스케일러블 칼라 기술자를 생성한다. 그리고 어학 비디오의 1분 재생 후의 프레임(설명 부분에 해당하는 프레임으로써 약

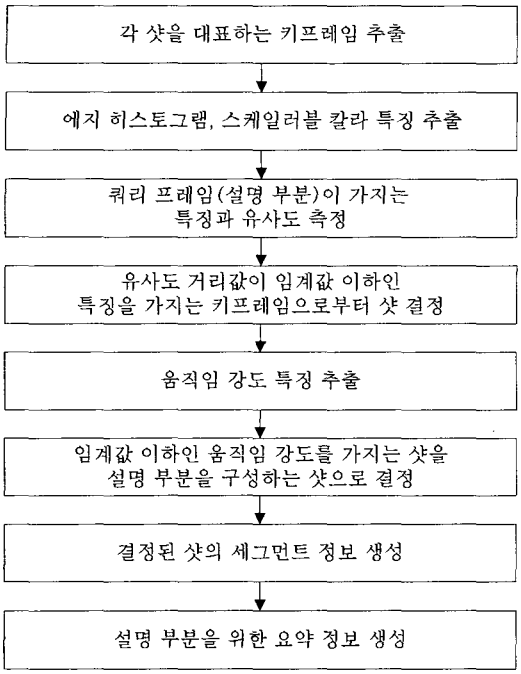


그림 7. 설명 부분 검출 흐름도

1800^{th} 프레임; 유사도 측정을 위한 쿼리 프레임)이 가지는 특징과 다른 모든 키프레임들이 가지는 특징과의 유사도 측정을 식 (6), (7)에 의해 행한다. 그림 8은 유사도 측정 결과를 보여준다. 그림과 같이 유사도 측정값이 일정값 th_{ED}, th_{CD}이하의 조건을 동시에 만족(두 조건의 AND 연산)하면, 해당 키프레임은 설명 부분을 구성하는 샷을 대표하는 것으로 결정한다.

그 다음으로, 설명 부분으로 결정된 샷들로부터 움직임 강도를 구한다. 그림 9는 실험용 어학 비디오의 설명 부분과 대화 부분의 움직임 강도 특징의 예를 보여준다. 설명 부분은 객체의 움직임이 거의 없는 샷들로 구성되기 때문에 움직임 강도는 비교적 작고, 따라서 움직임 강도가 일정값 이상인 샷들을 제외시키면 설명 부분을 구성하는 샷들을 검출하게 된다.

3.3.3 대화 부분 검출 방법

대화 부분을 생성하기 위하여 대화 부분을 구성하는 샷을 검출해야 한다. 설명 부분 검출 과정과 마찬가지로 에지 히스토그램, 스케일러블 칼라 기술자와 움직임 강도 특징값을 이용한다. 그림 8에서 보는 바와 같이 설명 부분에 해당하는 쿼리 프레임과의 특징 유사도 값이 일정값(th_{ED}, th_{CD}) 이상이면, 해당 키프

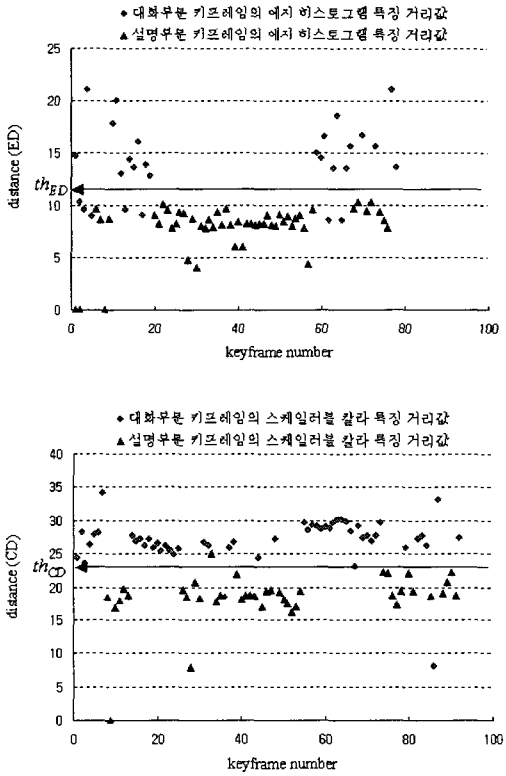


그림 8. 에지 히스토그램, 스케일러블 칼라 특징값에 의한 유사도 측정결과

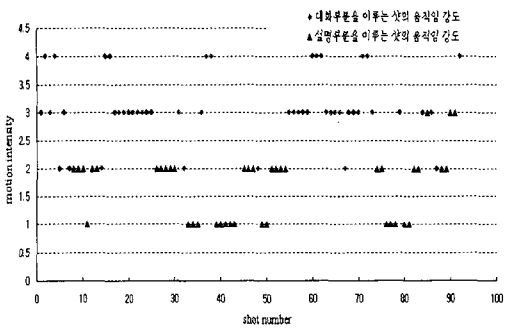


그림 9. 설명 부분과 대화 부분의 움직임 강도 특징

레임은 대화 부분을 구성하는 샷으로 결정한다.

그리고 나서 설명 부분으로 결정된 샷들중에서 큰 움직임 강도를 가지는 샷을 다시 검출하는 과정을 거친다. 그림 9에서와 같이 대화 부분을 구성하는 샷은 객체의 움직임이 상대적으로 크고, 움직임 강도는 설명 부분의 샷보다 큰 값을 가진다. 따라서 움직임 강도가 큰 샷들을 대화 부분으로 결정하여 세그먼트

정보를 생성한다. 그림 10은 앞에서 설명한 것처럼 대화 부분을 검출하여 요약 정보를 생성하는 흐름을 간단히 보여준다.

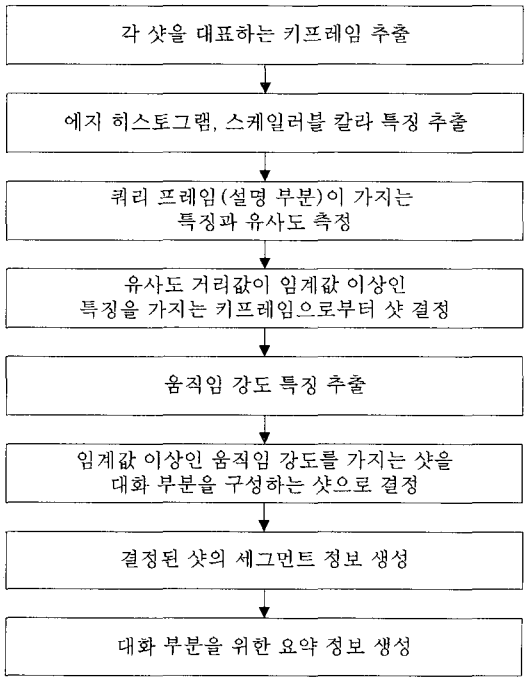


그림 10. 대화 부분 검출 흐름도

4. 실험 결과

제안한 어학 교육 비디오 자동 요약 방법의 유효성을 검증하기 위하여 외국어 회화 비디오를 이용하여 실험하였다. MPEG-2 형식의 비디오는 모두 15개로 각각 20분 분량이며, 중국어(Chinese) 회화, 영어(English) 회화, 프랑스어(French) 회화, 독일어(German) 회화, 일본어(Japanese) 회화를 위한 교육 비디오들이다. 표 1은 실험에 사용된 15가지 비디오의 샷 경계 검출 및 키프레임 추출 결과로써 각각의 숫자는 샷의 개수를 나타낸다. 여기서 각 샷은 하나씩의 키프레임을 가지므로 임의의 어학 비디오로부터 얻어진 샷과 키프레임의 개수는 같다.

먼저 어학 교육 비디오의 지문 부분을 검출하기 위하여 추출된 키프레임들로부터 색도 히스토그램 데이터를 구하여 첨도값를 얻었으며, 질감 기술자를 적용해 키프레임들의 30채널 에너지로부터 표준편차를 측정하였다. 그 다음 미리 정의된 임계값들을 적용하여 조건에 만족하는 특징값을 가지는 키프레임을 검출하였다. 최종적으로 검출된 키프레임이 대표하는 샷들은 지문 부분을 위한 세그먼트 정보를 이루게 된다.

표 1. 실험에 사용된 교육용 어학 비디오 샷의 이벤트 분포

	설명부	대화부	지문부	잔여부	전 체
Chinese	147	118	9	42	316
English	140	124	4	32	300
French	144	120	7	41	312
German	137	117	11	28	293
Japanese	150	99	4	30	283
합 계	718	578	35	173	1504

표 2는 색도 히스토그램 첨도값을 적용, 표 3은 30 채널 에너지 표준편차를 적용하여 지문 부분을 검출한 결과이다. 단일 특징만 적용하였을 때는 precision 률이 낮지만, 색도 히스토그램 첨도값과 30 채널 에너지 표준편차의 다중 조합에 의한 지문 부분 최종 검출결과는 표 4에서 보는 바와 같이 90% 이상의 높은 정확도를 보이며, 이는 지문 부분이 색도 히스토그램과 30 채널 에너지 표준편차 특성에 의해 설명 부분이나 대화 부분으로부터 투렷히 구분되는 특성을 보이기 때문이다.

표 2. 색도 히스토그램 첨도값을 이용한 지문 부분 검출결과

	Total (개)	Correct (개)	Miss (개)	False (개)	Recall (%)	Precision (%)
Chinese	9	8	1	23	88.89	25.81
English	4	4	0	4	100	50.00
French	7	7	0	8	100	46.67
German	11	10	1	11	90.91	47.62
Japanese	4	4	0	0	100	100
합 계	35	33	2	46	94.29	41.77

표 3. 질감 기술자의 30 채널 에너지 표준편차를 이용한 지문 부분 검출결과

	Total (개)	Correct (개)	Miss (개)	False (개)	Recall (%)	Precision (%)
Chinese	9	8	1	8	88.89	50.00
English	4	4	0	17	100	19.05
French	7	6	1	22	85.71	21.49
German	11	10	1	48	90.91	17.24
Japanese	4	4	0	0	100	100
합 계	35	32	3	95	91.43	25.20

표 4. 다중 조합에 의한 지문 부분 검출결과

	Total (개)	Correct (개)	Miss (개)	False (개)	Recall (%)	Precision (%)
Chinese	9	9	0	3	100	75.00
English	4	4	0	1	100	80.00
French	7	6	1	2	85.71	75.00
German	11	10	1	2	90.91	83.33
Japanese	4	4	0	0	100	100
합 계	35	33	2	8	94.29	80.49

실험에 사용한 비디오들은 설명 부분과 대화 부분을 비교적 많이 포함하고 있으며, 따라서 해당 샷의 개수는 지문 부분보다 많다. 설명 부분과 대화 부분을 위한 효율적인 요약 정보를 생성하기 위하여 적용할 에지 히스토그램, 스케일러블 칼라 및 움직임 강도, 세 가지 특징들의 효율적인 조합이 중요하다. 각각의 단일 특징을 이용한 설명 부분 검출은 높은 정확도를 가지지 못하므로, 3.3.2절과 3.3.3절에서 설명한 조합 방법을 적용하였으며 다중 조합에 의해 precision 률 향상을 얻는다.

표 5는 어학 비디오로부터 설명 부분을 검출하기 위하여 에지 히스토그램 기술자에 정의된 특징만을 이용한 결과이며, 표 6은 스케일러블 칼라 특징을 이용한 설명 부분 검출결과이다. 그리고 표 7은 에지 히스토그램과 스케일러블 칼라 특징을 동시에 적용하여 설명 부분을 검출한 결과이다. 보다 높은 정확도를 얻기 위하여 표 7로부터 얻은 결과에 다시 움직임 강도 특징을 적용한다. 표 8은 최종적으로 앞에서 논한 세 가지 특징들을 조합하여 얻은 어학 비디오의 설명 부분 검출결과를 나타내며, 높은 검출 정확도를 보여준다.

3.3.3 절에서 설명하였듯이 설명 부분의 키프레임과의 에지 히스토그램, 스케일러블 칼라 특징 유사도가 임계값 이상인 키프레임들은 대화 부분을 구성하는 샷들을 대표한다. 또한, 대화 부분을 구성하는 샷들은 움직임 정보가 많은 특성을 지니므로 움직임 강도 특징의 적용을 통해 검출 정확도를 높일 수 있다. 표 9는 비주얼 특징들을 조합하여 얻은 대화 부분 검출결과이며, 그림 11은 지문 부분 검출결과의 예로써 일본어 회화에서 검출된 대화 부분을 구성하는 샷들의 키프레임이다.

최종적으로 어학 교육 비디오의 설명 부분, 대화 부분, 지문 부분이 검출되면, 각 부분을 이루는 샷에 대한 세그먼트 정보를 이용하여 MPEG-7 MDS에 정의된 계층적 요약 (Hierarchical Summary) 구조

표 5. 에지 히스토그램 특징값을 이용한 설명 부분 검출결과

	Total (개)	Correct (개)	Miss (개)	False (개)	Recall (%)	Precision (%)
Chinese	147	118	29	36	80.27	76.62
English	140	123	17	87	87.86	58.57
French	144	142	2	33	98.61	81.14
German	137	123	14	118	89.78	51.04
Japanese	150	141	9	30	94.00	82.94
합 계	718	647	71	304	90.11	68.03

표 6. 스케일러블 칼라 특징값을 이용한 설명 부분 검출결과

	Total (개)	Correct (개)	Miss (개)	False (개)	Recall (%)	Precision (%)
Chinese	147	138	9	109	93.88	55.87
English	140	135	5	16	96.43	89.40
French	144	130	14	75	90.28	63.41
German	137	122	15	18	89.05	87.14
Japanese	150	142	8	54	94.67	72.45
합 계	718	667	51	272	92.89	71.03

표 7. 에지 히스토그램, 스케일러블 칼라 특징값을 동시에 이용한 설명 부분 검출결과

	Total (개)	Correct (개)	Miss (개)	False (개)	Recall (%)	Precision (%)
Chinese	147	128	19	20	87.07	84.29
English	140	133	7	16	95.00	88.49
French	144	130	14	30	90.28	81.25
German	137	122	15	9	89.05	93.13
Japanese	150	141	9	28	94.00	83.43
합 계	718	654	64	103	91.09	86.21

표 8. 세 가지 특징의 다중 조합에 의한 설명 부분 검출결과

	Total (개)	Correct (개)	Miss (개)	False (개)	Recall (%)	Precision (%)
Chinese	147	128	19	15	87.07	89.51
English	140	130	10	10	92.86	92.86
French	144	129	15	25	89.58	83.44
German	137	122	15	6	89.05	95.31
Japanese	150	138	11	3	92.00	97.87
합 계	718	647	71	59	90.11	91.64

표 9. 대화 부분 검출결과

	Total (개)	Correct (개)	Miss (개)	False (개)	Recall (%)	Precision (%)
Chinese	118	106	12	19	87.18	82.93
English	124	114	10	10	97.56	97.56
French	120	98	22	15	91.18	93.94
German	117	114	3	15	89.74	92.11
Japanese	99	97	2	11	94.74	94.74
합 계	578	529	49	71	91.52	88.17

에 맞추어 XML 문서를 생성한다. 즉, 그림 12에서 보는 바와 같이 입력 비디오의 샷들이 각각의 해당 부분을 이루는 것으로 검출되면, 시간적으로 연결된 샷들은 하나의 비디오 클립이 되고, 그 비디오 클립의 첫 프레임 번호와 마지막 프레임 번호는 계층적 요약 구조에 정의된 <KeyVideoClip> 요소의 <MediaTime> 정보로 쓰인다. 아래 그림은 영어 회화 비디오에 대한 요약 정보를 나타내는 XML 문서의 일부이며, 설명 부분, 대화 부분 및 지문 부분 검출 결과로 생성된 세그먼트 정보를 담고 있다. 이것은 방송 제공자가 시청자가 원하는 요약 컨텐츠를 제공하는데 이용될 수 있다. 그림 13은 제안된 요약 방법에 의해 생성된 결과를 이용하여 요약 비디오 서비스 제공에 응용한 예이다.

그림 12. 요약 정보를 나타내는 XML 문서

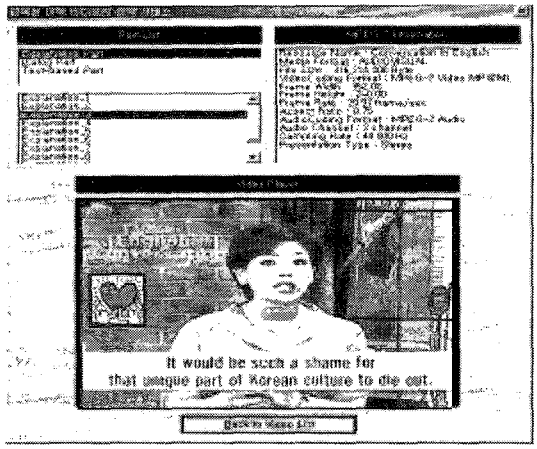


그림 13. 요약 정보를 이용한 비디오 서비스 응용의 예

5. 결론 및 향후 과제

본 논문에서는 비디오의 비주얼 특성과 MPEG-7 국제 표준에서 제공하는 기술자 일부를 이용하여 어학 교육 비디오의 내용 기반 비주얼 특징들을 추출하고, 이들을 효율적으로 조합하여 어학 교육 비디오의 설명 부분, 대화 부분, 지문 부분을 검출한 후 요약 정보를 자동으로 생성하였다. 그리고 비디오의 구조적 내용 정보를 기술하는 요약문을 생성하였다. 교육 방송용 비디오에 대한 컨텐츠는 증가하고 있으며, 방

그림 11. 일본어 회화의 대화 부분으로 검출된 키프레임의 일부

송 서비스상에서 시청자의 요구가 늘어가고 있다. 특히 어학 교육 비디오는 요약된 정보를 제공할 필요성이 높은 컨텐츠이다. 본 논문에서 제안한 방법은 어학 교육 비디오에 대한 정확한 요약 결과를 제공하며, 향후 양방향 방송 서비스상에서 소비자의 기호에 적합한 컨텐츠를 제공하는데 이용될 것이다. 제안한 방법에 의해 얻어진 최종 결과인 요약 정보를 표현하는 XML 문서는 시청자가 어학 교육 컨텐츠의 내용을 신속하고 효율적으로 파악 및 이용하는데 쓰일 수 있다. 또한 시청자가 많은 어학 교육 컨텐츠 중에서 선호하는 부분을 원할 때, 편의성을 제공하며 빠르고 정확한 검색도 제공할 수 있다.

향후 다양한 어학 교육 컨텐츠를 이용한 일반적인 요약 방법에 대한 연구가 요구되며, 실제 방송 환경을 고려한 요약 시스템 및 시청자의 선호도 정보를 이용한 요약 방법에 대한 연구가 이루어져야 할 것이다.

참 고 문 헌

[1] Stefan Eickeler and Stefan Muller, "Content-Based Video Indexing of TV Broadcast News using Hidden Markov Models," ICASSP '99 Proceedings, IEEE, Vol. 6, pp. 2997-3000, March 1999.

[2] Dalong Li and Hanqing Lu, "Model Based Video Segmentation," SiPS 2000 IEEE, pp. 120-129, Oct. 2000.

[3] Hari Sundaram and Shih-Fu Chang, "Video Scene Segmentation Using Video and Audio Features," IEEE International Conference on Multimedia and Expo, ICME 2000, Vol 2, pp. 1145-1148, 2000.

[4] Yan Liu and John R. Kender, "Video Frame Categorization Using Sort-Merge Feature Selection," Motion and Video Computing IEEE, pp. 72-77, Dec. 2002.

[5] Hee Kyung Lee, Cheon Seog Kim, Yong Ju Jung, Je Ho Nam, Kyeong Ok Kang, Yong Man Ro, "Video contents summary using the combination of multiple MPEG-7 metadata," SPIE Electronic Imaging, Vol. 4664, pp.1-12, 2002.

[6] Cheon Seog Kim and Yong Man Ro, "Semantic Event Detection using MPEG-7," SPIE Vol. 5021, pp. 372-379, 2003.

[7] Ichiro Ide, Koji Yamamoto and Hidehiko Tanaka, "Automatic Video Indexing Based on Shot Classification," AMCP'98, LNCS 1554, pp. 87-102, 1999.

[8] Nguyen Ngoc Thanh, Truong Cong Thang, Tae Meon Bae, Yong Man Ro, "Soccer Video Summarization System Based on Hidden Markov Model with Multiple MPEG-7 Descriptors," CISST 2003, Vol. 2, pp. 673-678, June 2003.

[9] Multimedia Descirption Schemes(MDS) Group, "Text of ISO/IEC 15938-5 FCD Information technology-Part 5 Description Schemes," March 2001.

[10] Y. M. Ro, K.W. Yoo, M.C. Kim, and J.W. Kim, "Texture Description using Radon transform," *ISO/MPEG*, m4703, Vancouver, 1999.

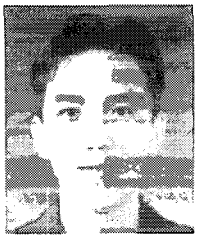
[11] MPEG-7 Visual part of eXperimentation Model Version 8.0, *ISO/MPEG*, w3673, La Baule, Oct 2000.

[12] Visual Working Draft 4.0, *ISO/MPEG*, w3522, Beijing, July 2000.

[13] Toby Walker and Sanghoon Sull, "Proposal for a Video Summary Description Scheme," July 1999.

[14] Sang-Heun Shim, Seung-Ji Yang, Jeong-Hyun Yoon, Yong-Man Ro, "Real-time Shot Boundary Detection Based on Digital Video Cameras using the MPEG-7 Descriptor," 2001년도 한국방송 공학회, p.193-198, 2001.

[15] Video Group, "Text of ISO/IEC 15938-1 FCD Information technology-Part 3 Visual," March 2001.



한 희 준

2002년 전북대학교 정보통신공학과, 학사
2002년~2003년 ㈜인터정보 기술연구소 인턴연구원
2004년 한국정보통신대학교 공학부, 석사
2004년~현재 한국과학기술정보연구원 (KISTI) 정보시스템부 근무
관심 분야 : 영상/비디오 신호 처리, MPEG-7/21

김 천 석

1981년 홍익대학교 전기공학과, 학사
1983년 고려대학교 전기공학과, 석사
2001년~현재 한국정보통신대학교 공학부 박사과정
관심 분야 : 영상 처리, MPEG-7, MPEG-21

추 진 호

2002년 홍익대학교 전자전기제어공학부, 학사
2002년~2003년 한국전자통신연구원 위촉연구원
2004년 한국정보통신대학교 공학부, 석사
2004년~현재 삼성전자 디지털미디어 총괄 근무
관심 분야 : MPEG-7/21, 비디오 인덱싱, 데이타 방송

노 용 만

1985년 연세대학교 전자공학과, 학사
1987년 한국과학기술원 전자공학부, 석사
1987년 컬럼비아대학 연구원
1992년 UC 어바인대학 초빙 연구원
1992년 한국과학기술원 전자공학부, 박사
1996년 UC 버클리대학 연구원
1997년~현재 한국정보통신대학교 공학부 부교수
관심 분야 : 이미지/비디오 처리 및 분석, MPEG-7, 특징 인식, 이미지/비디오 인덱싱