특집논문-06-11-1-03

교육용 비디오의 ToC 자동 생성 방법

이 광 국[a], 강 정 원[b] , 김 재 곤[b], 김 회 율[a]‡

A Method of Generating Table-of-Contents for Educational Video

Gwang-Gook Lee[a], Jung Won Kang[b], Jae-Gon Kim[b], Whoi-Yul Kim[a]‡

요 약

양방향 맞춤형 방송의 실현으로 인해 비디오의 내용을 자동으로 분석하여 그 구조를 기술하거나 요약을 생성하는 등의 내용 기반 비디오 분석 기술의 필요성이 요구되고 있다. 본 논문에서는 온라인에서 수요가 높고 특히 맞춤형 방송에 적합한 방송 콘텐츠인 교육용 비디오의 ToC를 자동으로 생성하기 위한 방법을 제안한다. 제안한 ToC 생성 방법은 씬 분할과 씬 서술의 두 단계로 이루어져 있다. 씬 분할 단계에서는 샷 분할을 수행한 후 샷 간의 연결관계 분석을 통해 입력 영상을 씬 단위로 분할하게 된다. 씬 서술 단계에서는 분할된 각 씬이 장면 분류, 자막 검출, 화자 인식 등에 의해 그 내용이 자동으로 서술된다. 제안된 방법을 통해 생성된 ToC는 씬과 샷의 계층 구조를 통해 비디오의 구성을 표현하고, 검출된 여러 특징을 이용해 각 씬과 샷의 내용을 서술함으로써 사용자가 비디오의 내용을 한눈에 알아볼 수 있고 원하는 내용에 손쉽게 접근할 수 있도록 도와줄 수 있다. 또 보다 상세한 ToC가 요구되는 경우에는 유용한 정보들이 포함되어 있는 초기 형태의 ToC로써 이용되어 수작업에 의한 ToC 생성에 필요한 시간을 효과적으로 줄이는 것이 가능하다. 실험을 통해 제안한 방법으로 여러개의 교육용 비디오에서 ToC를 효과적으로 생성될 수 있음을 확인하였다.

Abstract

Due to the rapid development of multimedia appliances, the increasing amount of multimedia data enforces the development of automatic video analysis techniques. In this paper, a method of ToC generation is proposed for educational video contents. The proposed method consists of two parts: scene segmentation followed by scene annotation. First, video sequence is divided into scenes by the proposed scene segmentation algorithm utilizing the characteristics of educational video. Then each shot in the scene is annotated in terms of scene type, existence of enclosed caption and main speaker of the shot. The ToC generated by the proposed method represents the structure of a video by the hierarchy of scenes and shots and gives description of each scene and shot by extracted features. Hence the generated ToC can help users to perceive the content of a video at a glance and to access a desired position of a video easily. Also, the generated ToC automatically by the system can be further edited manually for the refinement to effectively reduce the required time achieving more detailed description of the video content. The experimental result showed that the proposed method can generate ToC for educational video with high accuracy.

Keyword: Multimedia Ontology, Semantic Retrieval, Semantic Integration

a) 한양대학교 전자통신전파공학과
Division of Electrical and Computer Engineering, Hanyang University
b) 한국전자통신연구원 디지털방송연구단 방송시스템연구그룹
Broadcasting Media Research Group, Digital Broadcasting Research Division, ETRI
‡ 교신저자 : 김회율(wykim@hanyang.ac.kr)

I. 서 론

멀티미디어기기 발달에 따른 디지털 데이터양의 급격한

증가로 인해 자동화된 비디오 분석에 대한 관심이 날로 증가하고 있으며, 최근에는 기존의 장면전환 검출과 같은 기본적인 비디오 분석을 벗어나 비디오의 특징, 구조를 기술할 수 있는 비디오 요약/인덱싱과 같은 내용 기반 비디오 분석에 대해 많은 연구가 진행되고 있다. 특히 셋탑박스와 PVR등의 인터넷 연동과 인터렉티브 TV등의 개발로 인해 차세대 방송 콘텐츠는 일방적인 정보전달에서 벗어나 사용자 대화형(user interactive)방송으로 발전하고 있다. 그러나 비디오 데이터는 시간적 미디어(temporal media)라는 그 특성상 비디오의 편집 및 분석에 많은 노력이 요구되므로, 이러한 맞춤형 방송 서비스를 제공하기 위해서는 멀티미디어 데이터의 내용을 자동으로 요약해주고 주요 장면을 선별해주는 내용 기반 비디오 분석 기술이 필수적으로 요구된다.

자동화된 비디오 분석을 위한 초기의 많은 연구들은 주로 비디오를 샷 단위로 분할하는 것을 목적으로 하였다. 그러나 샷으로 비디오를 분할하는 것은 비디오 분석단계에서 의미 있는 전처리 과정이기는 하나, 샷 그 자체는 많은 정보를 포함하고 있지 않으며 또 샷의 경계는 그 발생빈도가 높아서 비디오의 내용을 이해하는데 사용자에게 직접적인 도움을 주기 어려운 문제가 있다. 이에 기존의 많은 연구에서는 비디오를 동일한 의미를 가지고 있는 샷들의 집합인 씬 단위로 분할하기 위한 노력을 기울였으며, 이러한 연구는 대부분 동일한 씬 내에 포함된 샷들은 유사한 영상을 포함하고 있으며 씬 내에서 동일 샷의 반복이 자주 발생한다는 사실에 기반하고 있다[3][4][5][8]. 이러한 씬 분할 연구는 주로 영화나 드라마 등 스토리를 포함하고 있는 일반적인 비디오 영상을 대상으로 하고 있다. 이러한 일반적인 영상을 대상으로 한 씬 분할 연구 이외에 특정한 종류의 영상만을 대상으로 하여 내용 기반 분석을 수행하고자 하는 많은 노력 역시 있었다. 이러한 연구들은 해당 도메인의 사전 지식을 이용하여 비디오의 구조나 비디오 내에서 발생하는 사건(event)의 모델을 설정하고 이를 이용하여 비디오 분석을 하였다. 이러한 연구로는 뉴스 비디오를 분할하여 구조를 표현하는 연구가 있었으며, 여기에서는 뉴스 비디오에 나타나는 특정 패턴을 이용하여 비디오의 구조를 분석하였다[9]. 또 스포츠 비디오의 색인(index)을 생성하거나 주요 장면을 추출하기 위한 많은 연구가 있어왔으며, 이러한 연구에서도 마찬가지로 해당 도메인의 알려진 사전 지식이 비디오의 내용 분석에 이용되었다. 예를 들면, 움직임 벡터와 색상들을 이용하여 농구 경기 영상의 주요 이벤트에 대한 색인을 생성하기 위한 연구가 있었다[15]. 또한 축구 경기 영상에서 주요 장면이 갖는 일반적인 패턴을 이용하여 주요 장면만을 추출하기 위한 연구도 있었다[11].

본 논문에서는 여러 비디오 가운데 특히 교육용 비디오를 목표로 하여 이에 대한 ToC(Table of Contents)를 자동으로 생성하는 방법을 제안한다. 교육용 비디오는 온라인상에서 그 수요가 높으며, 대화형/맞춤형 방송 제공 시에 학습자의 능력에 맞는 선택적 학습이 가능해지므로 그 효용성이 큰 콘텐츠 가운데 하나이다. 제안한 방법은 씬 분할을 통해 비디오를 서로 연관된 장면의 집합인 씬들로 구분하며, 이벤트 검출을 통해 각각의 씬들의 이벤트를 서술함으로써 교육용 비디오의 ToC를 생성한다.

본 논문의 구성은 다음과 같다. 2장에서는 시스템의 전체 구성을 설명하며, 3장에서는 비디오를 씬으로 분할하는 방법을, 4장에서는 분할된 각 씬의 서술(description)의 생성 방법이 설명된다. 5장에서는 실험 결과를 보이며, 마지막으로 6장에서 결론을 맺는다.

Ⅱ. System Overview

이 절에서는 본 논문의 관심 대상인 교육용 비디오의 특징을 설명하고 제안한 방법을 개략적으로 소개하며, 제안한 방법을 통해 생성되는 ToC의 구성을 소개한다. 그림 1은 일반적인 교육용 비디오의 구조를 보여주고 있다.

그림 1에 나타난 바과 같이 교육용 비디오는 다음의 몇 가지 종류의 장면으로 구분될 수 있다.

- **도입 장면** : 도입 장면은 프로그램의 시작, 끝 부분 혹은 서로 다른 장면이나 주제의 전환부분에 주로 나타난다.

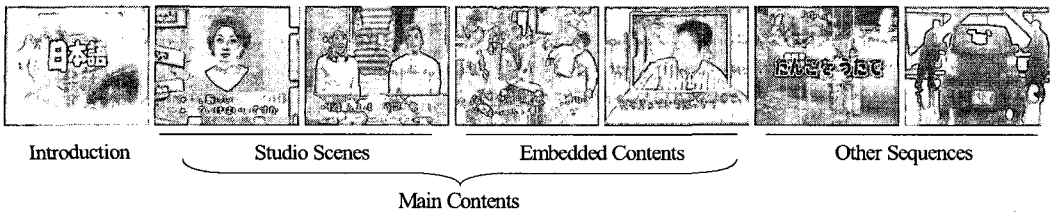


그림 1. 일반적인 교육용 비디오의 구성
Fig. 1. Structure of a general education video

이 장면은 일반적으로 그래픽 효과에 의해 표현되어 장면의 변화가 급격하게 일어나는 특징이 있다.

- **스튜디오 장면** : 교육용 비디오의 대부분의 장면은 스튜디오 내에서 얻어진다. 이 장면들은 일반적으로 고정된 카메라 위치에서 촬영되어 정적인 특징을 가지며, 전체 비디오 내에서 가장 많은 시간을 차지하게 된다.
- **기타 장면** : 교육용 비디오에는 또한 사용자의 이해를 돕기 위해 다른 비디오 시퀀스가 중간에 삽입되기도 한다. 이러한 비디오 시퀀스는 스튜디오 장면과 교차되어 나타나는 경우도 있고, 독립적으로 나타나는 경우도 있다. 또 이러한 장면들은 시각적인 특성이 굉장히 다양하여 일반화하기 어렵다.

비디오의 종류에 따라서는 더 많은 종류의 장면을 특징짓는 것이 가능하며, 이렇게 더욱 많은 사전 지식을 이용하게 되면 보다 많은 정보를 포함하고 있는 ToC를 생성하는 것이 가능할 것이다. 그러나, 본 논문에서는 제안한 방법이 특정 종류의 교육용 비디오만을 대상으로 하지 않도록 하기 위해 일반적인 교육용 비디오에 적용 가능하다고 판단되는 장면들만을 장면 모델로 선정하였다. 예를 들면, 특정 비디오에서는 오직 한명의 강사만이 등장하여 스튜디오 내에서 비디오 전체를 진행하기도 하며, 이러한 경우에 비디오 내에는 강사가 등장하는 장면과 보조 장면(칠판 또는 그래픽 장면)만이 존재하기도 한다. 앞에서 설명된 교육용 비디오의 구성 모델은 이러한 형태의 비디오를 포함하여 다양한 형태의 비디오에 적용이 가능하다.

그림 2는 제안한 시스템의 블록도를 나타내고 있다. 그림 2에 나타난 것과 같이 제안한 방법은 씬 분할과 씬 서술의 두 부분으로 나누어져 있다. 입력 영상은 우선 장면 전환 검출을 통해 샷 단위로 분할되며, 분할된 샷 사이의 관계를 분석하여 씬 단위로 비디오를 분할하며, 이를 통해 비디오 전체의 구조를 사용자에게 전달할 수 있게 된다. 씬 서술에서는 장면 분류, 자막 검출, 화자 인식 등을 통하여 분할된 각 씬을 자동적으로 서술하며 이를 통해 전체 비디오에 대한 ToC를 생성한다.

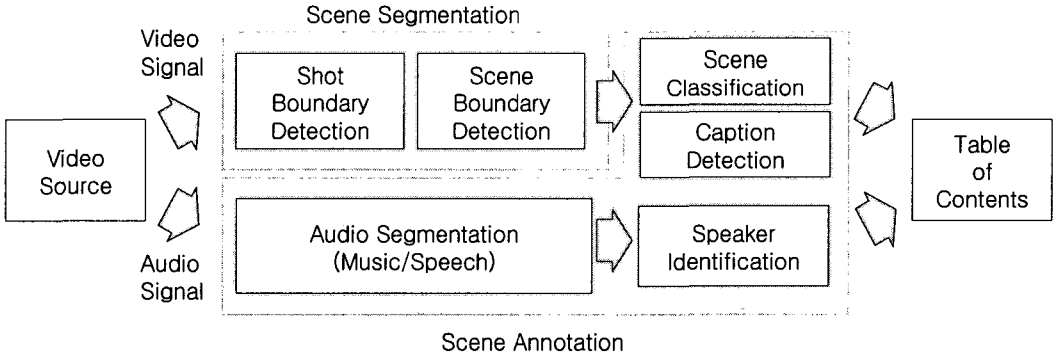


그림 2. 제안된 시스템의 블록도
Fig. 2. Block diagram of the proposed method

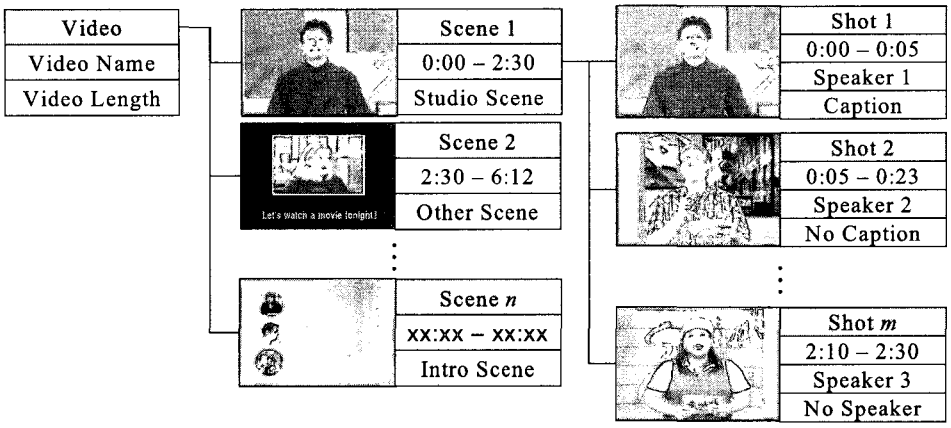


그림 3. 생성된 ToC의 예
Fig. 3. An example of the generated ToC

그림 3은 본 논문에서 제안된 방법에 의해서 생성된 ToC의 예를 나타낸다. 그림에 나타난 것처럼 생성된 ToC에는 입력된 비디오 영상의 구조가 씬과 샷에 의해서 표현되며, 각 씬과 샷은 대표 이미지와 시간 정보등의 기본적인 정보 이외에 씬의 종류, 화자, 자막 정보 등의 특징이 기술된다.

이렇게 생성된 ToC는 그 자체로도 비디오의 대략적인 구조를 기술함으로써 사용자가 비디오 콘텐츠의 구성을 한눈에 파악하고 특정 부분에 쉽게 접근할 수 있도록 할 수 있다. 뿐만 아니라 제안된 방법에 의해 생성된 ToC는 수작업으로 ToC를 생성하기 위한 시작점으로 이용될 수도 있다. 비디오를 수작업으로 분석하고 그 정보를 편집하는 것은 매우 시간을 요구하는 작업이므로, 편집에 필요한 기본적인 정보를 추출해 놓은 간략한 형태의 ToC를 제공함으로써 사용자에 의한 수동 편집에 소요되는 시간과 노력을 줄이는 것이 가능하다. 예를 들면, 비디오의 구조가 이미 씬과 샷으로 표현되어 있고 각 씬에 대해서는 스튜디오 장면과 도입 장면이 표시되어 있으므로 수작업으로 ToC를 편집하는 과정에서 매우 손쉽게 ToC의 구조를 설정할 수 있다. 또 ToC에 주요 학습 내용을 입력하고자 할 때 자막 검출 결과는 자막이 발생한 시간적 위치를 지정해줌으로써 입력을 손쉽게 도와준다. 화자 인식 결과는 ToC에 입력된 대사가 누구에 의해서 발생하는지를 표시해 줄 수 있으며, 또 특정 화자의 장면만에 접근하고자 할 때 도움을 줄 수 있다.

Ⅲ. 씬 분할

비디오는 순차적인 매체(sequential media)라는 점에서 책과 유사하다. 즉 사용자는 일반적으로 책의 내용을 한눈에 살펴보기 힘들며, 원하는 내용을 찾기 위해서는 순차 검색(linear search)이 요구된다. 그러나 ToC 혹은 인덱스를 사용함으로써 사용자는 그 내용을 한눈에 쉽게 파악할 수 있으며 원하는 내용에 쉽게 접근하는 것이 가능해진다. 마찬가지로 비디오에 있어서도 ToC와 같은 구조화된 표현은 사용자가 비디오의 내용을 한눈에 파악하고 원하는 지점으로 이동할 수 있도록 할 수 있다.

본 논문에서는 씬과 샷에 의해 비디오의 구조가 표현된다. 샷은 동일한 카메라에 의해 촬영된 장면전환을 포함하지 않는 연속된 프레임의 집합이며, 씬은 동일한 의미를 갖는 샷들의 집합이다. 사용된 샷과 씬 분할 방법은 다음 절에서 설명된다.

1. 샷 단위 분할

입력 비디오를 동일한 주제를 갖는 의미적 단위인 씬으로 분할하기 위한 전처리로 우선 장면 전환 검출이 수행되

었다. 비디오는 동일한 카메라로 촬영된 연속된 프레임의 집합인 샷들로 이루어져 있다. 이러한 샷과 샷 사이의 전환점을 검출하는 장면 전환 검출은 대부분의 동영상 처리에서 의미적인 분석을 위한 초기 분할로써 널리 이용되고 있다. 장면 전환의 종류에는 장면과 장면 사이가 두 영 특수한 효과 없이 갑자기 바뀌는 급격한 장면 전환과, 장면전환 과정에서 디졸브 혹은 와이프 등이 발생하는 점진적 장면 전환이 있다.

본 논문에서는 급격한 장면전환 검출을 위해서 Yusoff등에 의해 제안된 적응 임계화 방법[1]을 이용하였으며, 점진적 장면 전환의 검출을 위해서는 Bescos등에 의해 제안된 방법을 이용하였다[2]. Yusoff의 방법은 지역적 통계치를 이용하여 임계값을 동적으로 설정하기 때문에 동영상의 특성에 관계없이 적절한 임계값이 자동으로 선택되는 장점이 있으며, Bescos에 의해 제안된 방법은 점진적 장면 전환의 모델이 아니라 프레임간 거리의 변화에 기반하고 있기 때문에 특정 점진적 장면 전환의 형태에 관계없이 검출 가능한 장점이 있다.

장면전환 검출에 사용되는 프레임간 거리(distance) 계산에는 64x64 크기로 축소된 영상에서 화소값 차이의 평균이 사용되었다. 또 교육용 비디오에서는 자막이 자주 발생하는데, 이러한 자막의 발생은 장면전환 검출이나 장면 비교에 잘못된 결과를 초래할 수 있으므로 관심영역을 영상의 상위 2/3 부분만으로 제한하여 축소된 영상을 생성하였다.

2. 씬 단위 분할

2.1 씬 분할 방법

앞 절에서 설명된 장면 전환 검출에 의해 샷 단위로 분할된 비디오를 얻을 수 있으나, 샷은 단지 물리적 특징에 의해 구분된 단위로 대부분의 경우에서 많은 의미를 지니고 있지는 않다. 따라서 분할된 샷들을 동일한 주제를 갖는 이야기의 단위인 씬으로 재구성할 필요가 있다. Yeung등은 샷 간의 연결관계를 나타내는 씬 전환 그래프(STG, Scene Transition Graph)를 구성하여 비디오를 씬 단위로 분할하는 알고리즘을 제안하였다[3]. 이 방법은 동일한 씬 내에 포함된 샷 들은 동일 장소에서 촬영되었거나 동일한 물체가 나타나므로 유사한 샷이 반복적으로 발생한다는 관찰에 기초한 것으로 영화나 드라마와 같이 일반적인 비디오를 효과적으로 씬 단위로 분할할 수 있었다. Hanjalic등은 유사한 샷 사이에 링크를 연결하고 더 이상 링크가 연결되지 않는 지점에 씬 경계를 설정하는 방법을 제안하였으나[4], 이 방법 또한 상기의 유사 씬 반복이라는 가정에 기초한 것이므로 Yeung등이 사용했던 STG 방법의 greedy approach로 생각될 수 있다. Tavananpong등은 Hanjalic에 의해 제안된 것과 유사한 방법을 제안하였으나, 샷 간의 비교 방법에 비디오의 특징을 고려하여 프레임을 몇 개의 영역별으로 나누어 비교하였다[5].

앞에서 설명된 기존의 씬 분할 방법은 1) 같은 씬에 속한 샷들은 시각적으로 서로 유사하며 2) 비디오 시퀀스 내에서 시간적으로 인접해 있다는 두 가지 가정에 기반하고 있다. 이러한 가정은 동일한 씬에 속한 샷들은 동일한 배경이나 분위기를 갖기 때문에 시각적으로 유사하며, 반면 시각적으로 유사한 샷들이라 하더라도 비디오 내에서 시간적으로 멀리 떨어져 있는 샷들은 다른 씬에 포함된 장면일 가능성이 높다는 관찰에 기반하고 있다. 이러한 가정은 영화나 드라마와 같이 일반적인 비디오에 대해서는 매우 적합하지만, 본 논문에서 관심의 대상인 교육용 비디오에서는 종종 잘못된 결과를 야기하기도 한다. 교육용 비디오는 일반적인 영화나 드라마와는 달리 일반적으로 스튜디오와 같은 제한된 환경에서 촬영되며, 따라서 시각적으로 유사한 장면은 비디오에서 시간적으로 멀리 떨어져 있다고 하더라도 같은 장면일 가능성이 높다. 또한 교육용 비디오에서는 동일 씬 내에 속한 유사 장면의 반복도 영화나 드라마와 같은 일반적인 영상에 비해 짧게 나타나는 경우가 많다. 예를 들면, 두 진행자 사이의 대화 장면이 일반적인 영화 내에서의 대화 장면과는 달리 단지 두 개의 샷만을 포함하고 있는 경우도 종종 나타난다.

그림 4는 씬 분할의 예를 보이고 있으며, 그림에는 기존의 일반적 영상을 대상으로 하는 씬 분할 방법에 의해 잘못된 분할 결과가 나타나 있다. 그림 4에는 비디오 시퀀스 내에서 시간적으로 멀리 떨어져 있는 3 씬이 나타나 있으며, 씬 내의 샷들 가운데 시각적으로 유사한 샷들은 같은 문자로 표현되어 있다. 그림에 나타난 예에서 앞에서 설명된 일반적인 비디오를 대상으로 하는 기존의 씬 분할 방법은 지

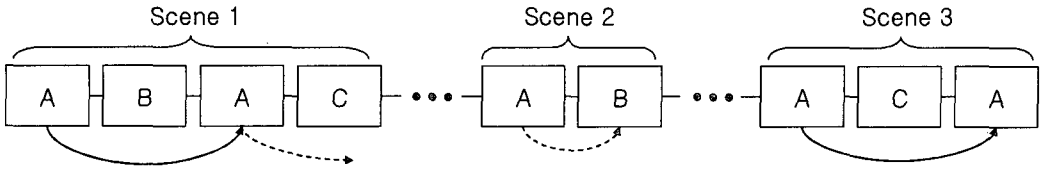


그림 4. 잘못된 씬 분할의 예
Fig. 4. An example of scene sequences that leads miss-segmentation

역적 비교에 기반하고 있기 때문에 첫 번째 씬(Scene 1)의 마지막 샷(Shot C)은 첫 번째 씬에 포함되지 않은 잘못 분할된 결과가 된다. 그러나, Scene 3의 분할 결과에서 Shot A와 Shot C가 같은 씬에 포함된다는 사실을 이용하면 이러한 잘못된 분할 결과를 피할 수 있다. Scene 2에는 샷 사이의 반복이 충분하지 않아 Shot A와 Shot B가 동일한 씬으로 분할되지 않았으나, Scene 1의 분할 결과를 이용하면 이러한 잘못된 결과 또한 피할 수 있다.

따라서 본 논문에서는 교육용 비디오의 특성을 고려하여 샷 간의 연결관계 분석을 시간적 인접성에 기반하지 않고 전역적으로 수행하여 씬을 분할하였으며 구체적인 알고리즘은 다음과 같이 설명할 수 있다.

- N 개의 샷 $s_1, s_2, \cdots, s_N$ 을 정점(node)으로 하는 그래프 상의 모든 정점이 연결되어 있는 완전연결 그래프 G를 생성한다.
- G의 두 정점 s_i와 s_j 사이의 유사도가 임계값 T_{shot}이하인 모든 정점 사이의 간선(edge)을 제거한다.
- G 내의 모든 정점에 대하여 2)를 반복된 이후, M개의 부그래프(sub-graph)가 발생하며, 각각의 부그래프는 유사한 샷들만을 포함하고 있는 샷 군집이 된다.
- M 개의 샷 군집 $c_1, c_2, \cdots, c_M$을 정점으로 하는 완전연결 그래프 H를 생성한다.
- H 의 두 정점 c_i와 c_j 사이에 연결관계가 발생하는지 조사하여 연결관계가 존재하지 않는 정점 사이의 간선을 제거한다. 두 샷 군집 사이의 연결관계는 다음의 수식 (1)과 같이 정의된다.

  수식에서 s_l은 c_i내의 임의의 샷이며 s_m은 c_j내의 임의의 샷이다. 또 index(s_l)은 샷 s_l의 샷 분할 결과에서의 인덱스를 나타낸다. 수식에 나타난 바와 같이 c_i내의 샷이 c_j내의 인접한 두 샷 내에 위치하며, 이 때 c_j의 인접한 두 샷 사이의 시간적 거리가 임계값 k보다 낮으면 두 샷 군집 사이에 연결관계가 있다고 판단한다.
- H 의 내의 모든 정점에 대하여 5)의 과정이 반복된 이후에는 l 개의 부그래프가 발생하며, 이들 각각은 동일한 씬에 포함된 샷 군집을 포함하고 있으므로 씬 군집이 된다.
- 최종적으로 샷 시퀀스에서 인접한 두 샷이 서로 다른 씬 군집에 포함되어 있으면 씬 경계(scene boundary)를 설정한다.

위 방법에 의한 씬 분할 과정은 그림 5에 나타나 있다. 장면전환 검출에 의해 비디오를 샷 단위로 분할하며, 분할된 각 샷들의 유사도를 비교하여 유사한 샷들을 동일한 클러스터로 병합한다. 이 때 샷간의 유사도 계산 방법으로 여러가지 방법을 이용할 수 있지만, 본 논문에서는 다음 절에서 설명되는 샷 비교 방법을 이용하였다. 이후 생성된 각 샷 클러스터들 간의 연결 관계를 분석하여 연결이 존재하는 클러스터내의 샷들을 동일 씬에 포함된 것으로 판단하여 씬 경계를 설정하게 된다.

2.2 샷 비교 방법

앞 절에서 설명한 씬 분할을 위해서는 샷 간의 유사도

$$Link(c_i, c_j) = \begin{cases} TRUE \quad \text{if } \{ index(s_l) < index(s_m) < index(s_{l+1}) \\ \qquad\qquad AND \;\; index(s_l) - index(s_{l+1}) < k \; \} \\ \qquad\qquad OR \;\; vise\,versa \\ FALSE \;\; otherwise \end{cases} \tag{1}$$

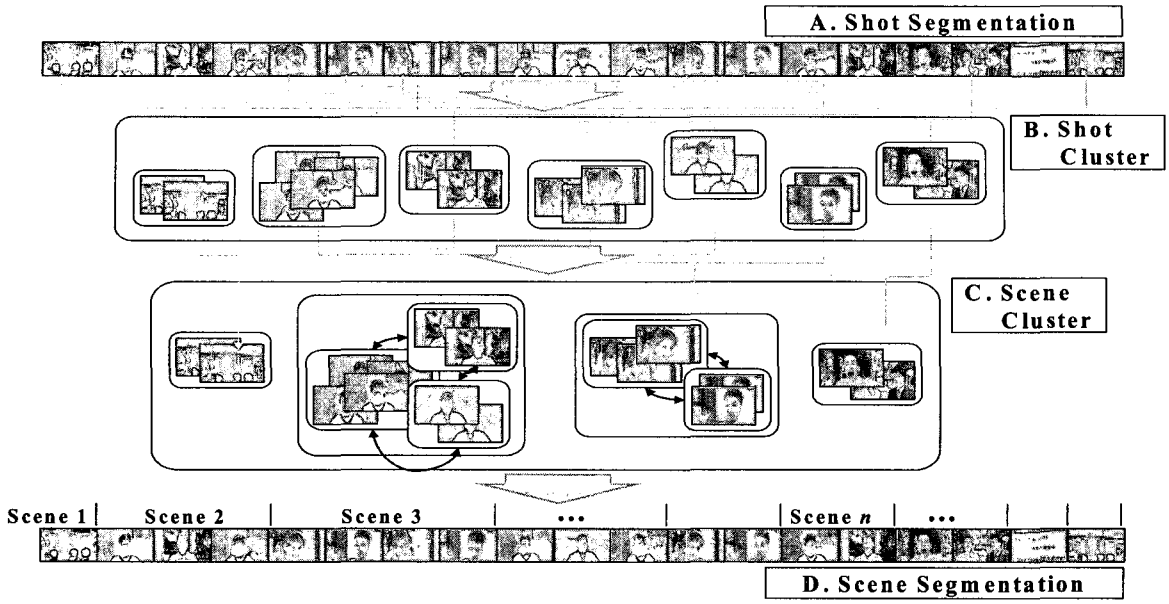


그림 5. 제안된 씬 분할 방법
Fig. 5. Proposed scene segmentation method

측정이 필요하며, 본 절에서는 제안한 방법에서 사용한 샷 간의 유사도 비교 방법을 설명한다. 샷을 기술하기 위해 먼저 샷에서 키 프레임(key frame)을 추출하였으며, 선택된 각 키 프레임의 특징으로는 MPEG-7 컬러 레이아웃 서술자를 이용하였다[6]. 샷의 키 프레임은 비균일(non-uniform) 샘플링에 의해 선택하였다[7]. 이는 가장 최근의 키 프레임과 현재 프레임을 비교하여 그 차이가 임계값 이상이 되면 새로운 키 프레임을 추출하는 것으로써, 장면의 변화가 큰 샷에서는 많은 수의 키 프레임이 추출되고 변화가 적은 샷에서는 작은 수의 키 프레임이 추출되어 비디오를 효율적으로 표현할 수 있다. 이러한 방법에 의해 전체 비디오의 2%에서 5%에 해당하는 키프레임 만으로 전체 영상을 효과적으로 표현할 수 있다. 샷 간의 유사도는 다음의 수식 (1)과 같이 샷 내의 모든 키 프레임간의 거리(distance)의 평균값으로 계산된다. M, N은 각각 샷 s_i와 s_j에 포함된 키 프레임의 개수를 나타내며, kl과 km은 샷 s_i와 s_j의 키 프레임을 나타낸다. 키 프레임의 거리(distance)는 컬러 레이아웃 서술자의 거리 비교 방법에 의해 얻어진다. 샷 간의 유사도 계산 시 점진적 장면전환 구간 내에 포함되어 있는 키 프레임은 유사도 계산에서 제외되었다.

$$dist(s_i, s_j) = \frac{1}{MN}\sum_{l}\sum_{m} d(k_l, k_m) \qquad (2)$$

Ⅳ. 씬 서술

앞 절에서 설명된 방법에 의해 비디오의 구조가 씬과 샷으로 표현된 이후에는 사용자가 비디오의 내용을 쉽게 파악하는 것을 돕기 위해 각 씬의 특징이 서술된다. 씬 서술 단계에서는 각각의 씬 내에 포함되어 있는 씬의 종류, 자막의 존재 여부, 그리고 화자 인식 등에 의해 씬의 내용을 표현한다.

1. 씬 분류

분할된 비디오의 각 씬으로부터 관심 장면을 검출하는 씬 분류는 사용자가 관심 장면에 선택적으로 접근할 수 있도록 하기 때문에, 많은 비디오 분석 시스템에서 널리 활용

되고 있다. Sundaram등은 일반적인 비디오에서 샷의 반복성을 이용하여 대화 장면을 검출하는 방법을 제안하였다[8]. Hsu등은 뉴스 비디오에서 분할된 세그먼트들을 광고검출을 통해 뉴스 장면과 비뉴스 장면으로 분류하는 방법을 제안하였다[9]. Girgensohn등은 비디오 프레임의 DCT 계수와 그 통계적 모델을 이용하여 비디오 프레임을 분류하는 방법을 제안하였고, 이를 이용하여 비디오 내에서 발표장면을 분류하였다[10]. 또 Ekin 등은 축구 비디오에서 주요 색상의 비율을 이용하여 경기 장면과 비경기 장면을 구분하는 방법을 제안하였다[11].

본 논문에서는 씬 서술을 위하여 교육용 비디오에서 가장 주가 되는 스튜디오 장면과 코너의 시작 부분에 흔히 사용되는 도입 장면을 검출하였다. 본 논문에서 사용된 씬 분류 방법은 다음과 같이 교육용 비디오에서 일반적으로 관찰될 수 있는 특징에 기반하고 있다.

- 일반적인 어학 학습용 교육 비디오에서는 진행자가 학습 내용을 설명해주는 부분이 가장 주가 되며, 따라서 전체 비디오 시퀀스 가운데 시간적으로 가장 많은 부분을 차지한다. 이 장면은 주로 스튜디오 내의 고정된 위치에서 고정된 카메라에 의해 촬영되는 경우가 많기 때문에 샷 내에서 영상의 변화가 적다.
- 각 코너의 시작부분에 삽입되는 도입 장면은 많은 그래픽 효과의 사용으로 샷 내의 움직임이 크다. 그러나 이러한 장면은 상대적으로 지속시간이 짧으며 또 비디오 내에서 반복되지 않아 전체 비디오 시퀀스에서 적은 부분만을 차지한다.

이러한 관찰에 기반하여 모든 씬 가운데 가장 길이가 길고 씬 내에서 영상의 변화가 적은 장면이 스튜디오 장면으로 선택되었다. 이는 다음의 수식 (3)에 의해 표현된다. 수식 (3)에서 S_i는 비디오 내의 임의의 씬을 나타내며, s_j는 씬 Si 내에 포함된 임의의 샷을 의미하고 N은 씬 S_i내의 샷의 개수이다. 또, $len(s_j)$는 샷 s_j의 길이이며, $activity(s_j)$는 샷 s_j의 활동도를 나타낸다. 활동도는 샷 내에서 영상의 변화 정도를 표현하며, 따라서 활동도의 값이 큰 샷은 샷 내에서 영상의 변화가 크다. 수식 (3)에 의해 씬의 길이가 길고 씬 내에서 영상의 변화가 작은 씬은 큰 값을 갖게된다. 수식 (3)에 나타난 것과 같이 스튜디오 장면의 선택에는 씬의 길이와 씬의 활동도의 두 척도가 이용되는데, 이들은 그 값의 범위가 다르며 그 영향을 조절하기 위한 가중치로 w_1과 w_2가 이용되었다.

$$P_{main}(S_i) = w_1 \sum_{j}^{N} len(s_j) + w_2 \frac{N}{\sum_{j}^{N} activity(s_j)} \quad (3)$$

이때 샷의 활동도 activity(s_j)는 수식 (4)와 같이 정의된다. k_l은 샷 s_j에 포함된 키 프레임을 나타내며, time(k_l)은 키 프레임의 비디오 내에서의 시간을 나타낸다. 또, M은 샷 내의 키 프레임의 개수를 나타낸다. 즉 샷의 활동도는 장면 내의 키 프레임간의 시간적 간격의 역수의 평균값으로 표현된다. 장면 내에 영상의 변화가 클수록 키 프레임 사이의 시간적 간격이 짧아지기 때문에 영상의 변화가 클수록 연속한 키 프레임 시간의 차는 작아지며 그 역수는 커지게 된다. 따라서, 키 프레임의 시간의 역수는 장면 내의 영상의 변화에 대한 간단한 측정 방법으로 이용될 수 있다.

$$activity(s_j) = \frac{1}{M} \sum_{l \in s_j} \frac{1}{time(k_{l-1}) - time(k_l)} \quad (4)$$

2: 자막 검출

비디오에 삽입되어 있는 자막 정보는 많은 비디오에서 시청자에게 정보 전달을 위한 중요한 수단이며, 자동화된 비디오 분석에 있어서 중요한 단서가 된다. 특히 교육용 비디오에서는 사용자에게 교육의 효과를 높이기 위해서 자막이 매우 빈번히 중요하게 사용된다. 따라서 동영상에서 자막의 위치는 그 자체로 사용자에게 중요한 정보로 이용될 수 있으며, 수동 편집에 의한 ToC의 보정에 큰 도움이 될 수 있다.

본 논문에서는 영상에서 자막 발생을 검출하기 위해 다

음과 같이 텍스트 영역에서 큰 응답을 보이는 필터를 이용하였다[12]. 이 필터는 문자 영역에는 수직 방향의 획 성분(vertical stroke)이 다수 존재한다는 특성을 이용하며, 수식 (3)에 나타난 것과 같이 표현된다. 수식에서 $I(x,y)$는 영상의 명암도이며, t는 수직 방향 획성분을 누적시키기 위한 윈도우의 크기를 나타낸다. 수식 (5)에 나타난 것과 같이 영상을 x 방향으로 미분하여 그 값을 가로 방향으로 누적함으로써 문자 영역의 경계 성분에서 높은 응답을 얻게 된다.

$$A(x,y) = \sum_{i=-t}^{t} \frac{\partial I}{\partial x}(x+i,y) \qquad (5)$$

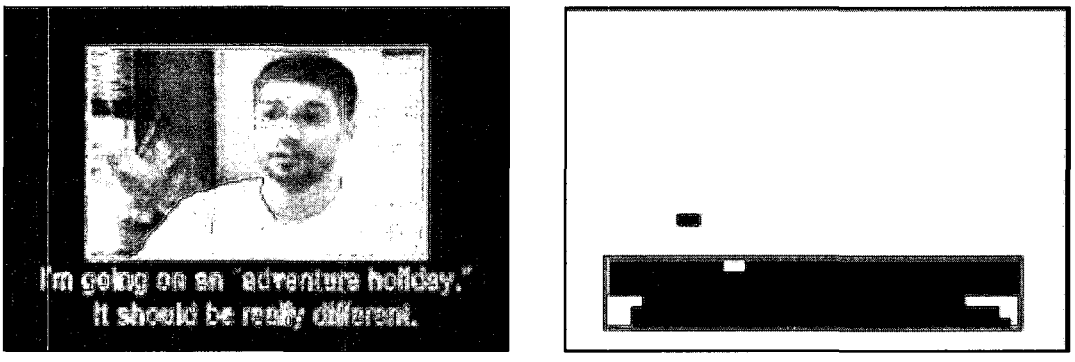
그림 6은 자막 영역의 검출 결과를 보여준다. 그림 6의 (a)는 자막이 존재하는 원 영상이며, 그림 6의 (b)는 수식 (3)에 의해 나타난 필터를 이진화한 결과로 흰색 부분은 자막의 영역일 가능성이 높은 부분을 의미한다. 이렇게 얻어진 이진 영상내의 각 영역 가운데 자막이라고 생각되는 부분을 Bayesian 결정 방법을 이용하여 선택하며, 이때 특징값으로는 각 영역의 크기와 가로, 세로의 비율이 이용되었다.

3. 화자 인식

자동화된 동영상 분석에 있어서 음성 신호는 영상 신호가 제공하는 이외에 부가적인 많은 정보를 제공하며, 이에 따라 오디오 신호를 동영상 분석에 이용하려는 많은 노력이 있어왔다. H. Sundaram은 영상 신호와 오디오 신호를 모두 이용하여 씬을 검출하려는 시도를 하였다[8]. 또 축구 비디오에서 오디오 신호만을 이용하여 중요 이벤트를 검출하려는 시도도 있었다[13].

본 논문에서도 사용자에게 유용한 정보를 전달할 수 있는 ToC를 생성하기 위해 영상 신호 이외에도 오디오신호를 이용하였다. 교육용 비디오는 영화 등의 일반적인 비디오와는 달리 주요 등장인물이 극히 제한되어 있는 특징을 갖는다. 콘텐츠에 따라 여러 명의 등장 인물이 있을 수는 있으나 스튜디오 내에서 학습 내용을 전달하게 되는 주요 진행자의 수는 세 명을 넘지 않는 경우가 대부분이다. 따라서 주요 화자의 수가 제한되어 있는 이러한 상황에서는 음성 인식을 통한 화자 구분이 상대적으로 용이한 장점이 있다.

따라서 본 논문에서는 화자 인식을 통해 각 씬과 샷의 화자가 특징으로써 서술되었으며, 본 논문에서 사용된 방법의 블록도는 그림 7에 나타난 것과 같다. 입력된 오디오 신호는 우선 프레임 단위로 분할되며, 각 프레임은 해밍(Hamming) 윈도우와 고역 강조(pre-emphasis)의 전처리를 거친다. 전처리된 각 프레임에서 특징 벡터가 추출되며, 추출된 특징 벡터에 의해 각 프레임이 분류된다. 이때 사용된 특징으로는 영교차율(ZCR, Zero-Crossing Rate), MFCC



(a) (b)

그림 6. 자막 검출의 예
Fig. 6. An example of caption detection

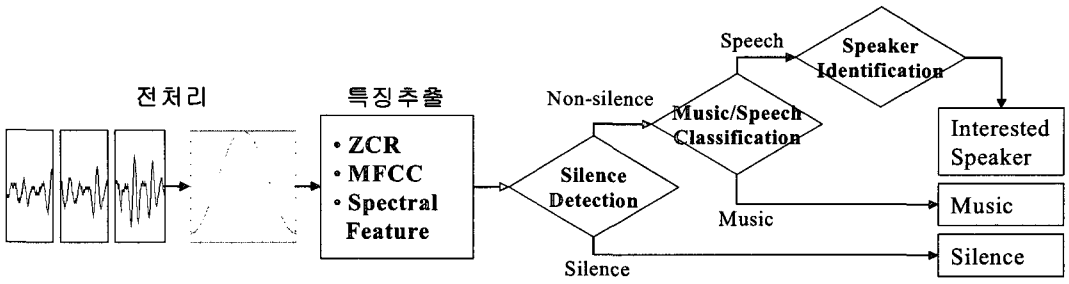


그림 7. 오디오 신호처리 블록도
Fig. 7. Flow of audio processing

(Mel-Frequency Cepstral Coefficients), 주파수 특징 등이 있다.

본 논문에서는 오디오 신호의 분류를 위하여 Reynolds에 의해 제안된 방법을 이용하였다[14]. Reynolds에 의해 제안된 방법은 각 클래스의 특징 벡터의 확률 분포를 GMM (Gaussian Mixture Model)로 모델링하며, 모델링된 확률 밀도를 이용하여 최대 우도 판단(Maximum Likelihood estimation)에 의해 클래스를 구분하는 방법이다. 음악과 음성을 구분하기 위해서는 영교차율과 주파수 특징이 이용되었으며, 화자 구분을 위해서는 MFCC가 사용되었다.

영교차율은 프레임 내에서 연속된 신호가 다른 부호를 갖는 비율로 정의되며, 이는 오디오 신호의 고주파 특성에 대한 간단한 평가로 이용될 수 있다. MFCC는 Mel 필터를 이용하여 계산된 캡스트럼 계수로 인간의 청각 특성을 반영할 수 있기 때문에 많은 응용에 이용되고 있다. Mel 주파수는 지수적으로 증가하며, Mel 주파수가 두 배 차이가 나는 신호는 인간의 귀에도 두 배의 높이 차이로 들리는 특성이 있다. 주파수 특징으로는 푸리에 변환의 결과에 선형 필터 뱅크를 적용하여 각각의 필터의 계수값을 특징으로 이용하였다.

각 프레임으로부터 특징 벡터를 추출한 이후에는 묵음(silence) 검출, 음악/음성 구분, 화자 인식등이 수행된다. 음성 신호의 분류는 다음의 과정에 의해서 이루어졌다.

- 각 프레임의 오디오 신호 에너지가 임계값보다 낮으면 해당 프레임이 묵음으로 판단된다. 이 때 과분할을 방지하기 위하여 길이가 0.5초 이하인 연속된 묵음 신호는 무시되었으며, 사용된 임계값은 비디오에서 임의로 추출된 1,000개의 샘플 프레임의 평균값이 이용되었다.
- 묵음 이외의 신호 부분을 음악과 음성 부분으로 구분한다. 이때의 특징 벡터로는 영교차율과 주파수 특징이 이용되었으며, 앞서 설명한 GMM 모델에 의한 최대 우도 판단이 이용되었다.
- 음성으로 구분된 각 오디오 세그먼트(segment)에 대하여 미리 훈련된 화자 모델을 이용하여 화자 인식이 수행된다. 화자 인식에는 MFCC가 특징으로 이용되었으며 음악 구분과 마찬가지로 EM(Expectation-Maximization) 알고리즘을 이용하여 얻어진 GMM모델에 의해 각각의 세그먼트에 대해 최대 우도 판단이 이용되었다. 이때 화자 인식은 등록된 화자와 등록되지 않은 화자를 구분하는 문제를 피하기 위해, 앞 절에서 설명된 씬 분류 결과에서 스튜디오 장면으로 구분된 씬 내의 샷에 대해서만 이루어졌다.

V. 실험 결과

제안한 방법의 성능을 평가하기 위해 실제 방송에서 얻어진 비디오 영상이 이용되었으며, 본 절에서는 실험 결과에 대해 설명한다. 실험에는 EBS사의 "TV English"와 "Survival English"의 두 종류의 비디오 시퀀스가 이

용되었으며, 각각에 대하여 10편 씩 총 20편의 비디오가 사용되었다. 각각의 비디오는 20분 분량이었으며, 앞에서 설명되었던 교육용 비디오의 일반적인 구성을 따르고 있었다. 사용된 영상 가운데 30%는 제안된 방법의 훈련에 사용되었으며, 나머지 70%만이 실험에 이용되었다.

표 1은 제안한 방법에 의한 씬 분할의 정확도를 보여준다. 씬 분할의 성능 측정에는 검출률(recall)과 정확도(precision)이 측정되었다. 검출률과 정확도는 다음의 수식 (6)에 의해서 표현되며, 여기서 N_c는 정확하게 검출한 씬 경계의 개수, N_m은 검출하지 못한 씬 경계의 개수 나타내며 N_f는 잘못 검출된 씬 경계의 개수를 나타낸다.

$$recall = \frac{N_c}{N_c + N_m} \times 100\% \qquad precision = \frac{N_c}{N_c + N_f} \times 100\% \tag{6}$$

표 1. 씬 분할의 정확도
Table 1. Result of scene segmentation

	TV English	Survival English	Average
검출률	89.42%	85.06%	84.71%
정확도	58.33%	54.52%	56.42%

씬 검출 결과 약 85%의 검출률과 55%의 정확도를 얻을 수 있었다. 이러한 씬 분할 결과는 샷 분할 결과에 비해 매우 낮은것처럼 보이나, 씬의 경우는 샷과는 달리 물리적인 경계가 아니라 의미적인 경계에 의해 나뉘기 때문에 상대적으로 검출이 어렵고, 또 그 개수가 적기 때문에 단지 몇 개의 잘못된 검출 결과만에 의해서도 성능이 낮게 나오게 된다. 실험에 의해 정확도가 검출률에 의해 상대적으로 낮은 것을 확인할 수 있었는데, ToC의 이용에 있어서 잘못된 경계는 사용자에 의해 손쉽게 제거될 수 있으며 또한 정보를 잃는 것은 아니기 때문에 씬의 경계를 검출하지 못하는 것보다는 잘못된 경계를 검출하는 것이 선호되기 때문에 검출률이 더 중요하다.

표 2는 씬 분류 결과를 보여준다. 씬 분류의 성능 측정을 위해서는 마찬가지로 검출률과 정확도가 이용되었다. 표 2에 나타난 것과 같이 스튜디오 장면의 분류에는 약 90%의 검출률과 정확도를 얻을 수 있었으며, 도입 장면에 대해서는 90%의 검출률과 60%의 정확도를 얻을 수 있었다. 도입 장면의 검출에 있어서 상대적으로 정확도가 낮음을 알 수 있었는데, 이는 주로 많은 움직임을 포함하고 있는 독립된 장면들이 씬으로 분류되는 경우에 의해 발생하였다.

표 2. 씬 분류 결과
Table 2. Result of scene classification

	TV English		Survival English		평균	
	스튜디오 장면	도입장면	스튜디오 장면	도입장면	스튜디오 장면	도입장면
검출률	91.36%	85.23%	83.24%	95.32%	87.39%	90.28%
정확도	98.17%	55.66%	86.54%	81.24%	92.35%	68.45%

표 3은 자막 검출 실험 결과를 나타낸다. 마찬가지로 검출율과 정확도가 성능 평가 방법으로 이용되었다. 자막 검출은 정확도와 검출율 면에서 모두 90% 이상의 성능을 나타내었다. 비자막 영상의 경우 검출률이 자막 영상에 비해 상대적으로 낮게 나타난 것을 알 수 있었으며, 그림 8은 이러한 경우의 예를 보여준다. 그림 8의 (a)에 나타난 것과 같이 영상 내엣 삽입된 그래픽의 일부분이 자막 영역과 유사한 특성을 갖거나, (b)에 나타난 것과 같이 자연 영상에서 자막의 특성을 갖는 부분이 존재하는 경우에 자막 검출이 실패하는 것을 볼 수 있었다.

표 3. 자막 검출 결과
Table 3. Result of caption detection

	TV English		Survival English		평균	
	자막	비자막	자막	비자막	자막	비자막
검출률	97.32%	90.67%	94.36%	91.83%	95.84%	91.25%
정확도	92.24%	93.55%	94.56%	94.55%	93.40%	94.05%

(a) (b)

그림 8. 잘못된 자막 검출 결과의 예
Fig. 8. Example of false detections in caption detection

표 4는 화자 인식 성능을 보여준다. 화자 인식을 위한 분류기(classifier)는 하나의 비디오 시퀀스에서 샘플을 추출하여 훈련하였으며, 총 세 개의 비디오에서 샘플 데이터를 추출하여 두 종류의 비디오 시퀀스에 대하여 각각 3개의 분류기가 훈련되었다. 성능 측정 방법으로는 수작업에 의해 미리 분류한 샘플 데이터에 대하여 각각의 분류기의 성능을 측정하는 방법을 이용하였다. 이는 비디오에서 장면의 경계가 음성 신호의 경계와 일치하지 않는 경우가 많고 또 하나의 장면 내에 여러 화자가 말하는 경우가 다수 발생하였기 때문이며, 이러한 이유로 음성 신호만을 독립적으로 추출하여 실험에 이용하였다. 또 분류기의 입력으로는 1, 10, 30, 50개 오디오 프레임의 특징 벡터를 이용하였을 때의 성능을 각각 측정하였다.

표 4. 화자 인식 결과
Table 4. Result of speaker identification

버퍼 길이	TV English		Survival English		
	Class 1	Class 2	Class 1	Class 2	Class 3
1	69.73%	76.62%	64.72%	66.75%	48.42%
10	78.42%	78.51%	72.65%	75.38%	52.36%
30	83.21%	87.11%	79.03%	83.83%	53.73%
50	85.46%	88.23%	82.46%	86.21%	54.31%

표 3에 나타난 것과 같이 성능 측정 결과 분류기의 입력을 사용한 특징 벡터의 개수가 많을수록 인식 성능이 증가하였으며, 최대 85% 정도의 성능을 얻을 수 있음을 확인하였다. "TV English" 에는 두 명의 강사가, "Survival English"에는 세 명의 강사가 강의를 진행하였다. "Survival Ensligh"의 세 번째 강사에 대해서는 다른 강사들에 비해 상대적으로 낮은 인식 성능을 얻을 수 있었는데, 이는 세 번째 강사의 경우 다른 강사들에 비해 상대적으로 영상 내에서 말하는 부분이 적게 발생하기 때문인 것으로 해석된다. 세 번째 강사의 등장 부분이 상대적으로 적기 때문에 훈련에 사용된 샘플의 개수가 적어서 분류기의 훈련이 제대로 이루어지지 못한 것으로 생각된다.

Ⅳ. 결 론

최근 양방향 데이터 방송의 본격화됨에 따라 비디오의 내용을 분석하여 인덱스를 생성하는 자동화된 비디오 요약 방법에 대한 관심이 높아지고 있다. 본 논문에서는 특히 교육용 방송 비디오에 적합한 ToC 생성 방법을 제안하였으며, 제안한 방법은 샷 간의 연결 관계 분석에 의한 씬 분할을 통해 비디오의 구조를 표현하고, 씬 서술을 통해 비디오의 내용을 사용자가 쉽게 파악할 수 있도록 한다. 실험 결과 제안한 방법은 실제 방송용 교육용 비디오에서 ToC를 효과적으로 생성할 수 있음이 확인 되었다.

그러나 또한 실험을 통해서 본 논문에서 제안한 방법은 각 단계 간의 의존성에 의해 잘못된 결과가 발생할 수 있

음이 확인되었다. 예를 들면, 씬 분류에는 씬의 길이가 중요한 특징으로 사용되기 때문에 씬 분할이 잘못 이루어진 경우에는 이에 의해 씬 분류의 결과가 잘못 나타날 수 있다. 또한 잘못된 씬 분류 결과는 씬 서술 단계에 영향을 끼치게 된다. 이러한 각 구성 단계간의 의존성을 해결하기 위해서는 각 단계의 결과가 해당 단계에서 바로 결정되는 것이 아니라 이전 또는 다음 단계와 영향을 주고받음으로써 결과를 수정할 수 있도록 하는 연구가 필요할 것으로 생각된다.

또 본 논문에서는 ToC 생성을 위해 영상 신호 이외에 오디오 신호 분석을 통한 화자 인식을 이용하였다. 그러나 제안된 방법에서 오디오 신호는 샷의 내용을 기술하기 위해서 비디오 분석과는 독립적으로 사용되고 있기 때문에 그 이용이 제한적이라고 할 수 있다. 따라서 향후 오디오 신호를 영상 신호와 함께 이용하여 비디오 분석에 이용하는 방법에 대한 연구가 진행되어야 할 것이다. 예를 들면, 씬 분할에 있어서 영상 신호 뿐 아니라 오디오 신호를 함께 이용함으로써 더욱 정확한 씬 경계를 검출 할 수 있을 것이다. 또한 등록된 화자와 등록되지 않은 화자를 구분함으로써 스튜디오 장면 구분의 정확도를 높이는 것도 가능할 것으로 생각된다.

참 고 문 헌

[1] Y. Yusoff, W. Christmas, and J. Kittler, "Video Shot Cut Detection Using Adaptive Thresholding," in Proceedings of the 11th British Machine Vision Conference, pp. 362-372, 2000.

[2] J. Bescos, J. M. Menendez, G. Cisneros, J. Cabrera, and J. M. Martinez, "A Unified Approach to Gradual Shot Transition Detection", in Proceedings of International Conference on Image Processing, Vol. III, pp. 949-952, 2000

[3] M. Yeung and B. L. Yeo, "Time-constrained clustering for segmentation of video into story units," in Proceedings of ICPR, Vol. C, Vienna, Austria, Aug. 1996, pp. 375-380.

[4] A. Hanjalic, R. L. Legendijk, and J. Biemond, "Automated High-Level Movie Segmenation for Advanced Video-Retirieval Systems", in IEEE Transactions of Circuits and Systems for Video Technology, Vol. 9, No. 4, June 1999.

[5] W. Tavananpong, "Shot Clustering Techniques for Video Browsing," in IEEE Transactions on Multimedia, Vol. 6, No. 5, August 2004.

[6] "MPEG-7 Visual part of experimentation Model Version 10.0, "ISO/IEC JTC1/SC29/WG11, N4063, Singapore, Mar. 2001.

[7] B. L. Yeo and B. Liu, "Rapid Scene Analysis on Compressed Videos," in IEEE Transactions on Circuits and Systems for Video Technology, 5(6): 533-544, Dec. 1995.

[8] H. Sundaram, S.-F. Chang, "Computable scenes and structures in films," in IEEE Transactions on Multimedia, Volume: 4 , Issue: 4 , Dec. 2002, Pages:482 - 491

[9] Winston H.-M. Hsu, L. Kennedy, C.-W. Huang, S.-F. Chang, C.-Y. Lin, G. Iyengar, "News Video Story Segmentation using Fusion of Multi-Level Multi-modal Features in TRECVID 2003," in Proceedings of International Conference on Acoustics, Speech, and Signal Processing, Montreal, Canada, May 17-21, 2004.

[10] A. Girgensohn and J. Foote, "Video Classification using Transform Coefficients," in Proceedings of International Conference on Acoustics, Speech, and Signal, vol. 6, pp. 3045-3048, 1999., March 15, 1999

[11] A. Ekin, A. M. Tekalp and R. Mehrotra, "Automatic Soccer Video Analysis and Summarization," in IEEE Transactions on Image Processing, Vol. 12, No. 7, July 2003.

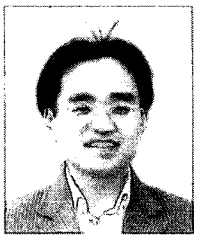
[12] C. Wolf, J.-M. Jolion, F. Chassaing, "Text Localization, Enhancement and Binarization in Multimedia Document" in Proceedings of 16th International Conference on Pattern Recognition, Volume: 2 , 11-15 Aug. 2002

[13] M. Xu, N. C. Maddage, C. Xu, M. Kankanhali and Q. Tian, "Creating Audio Keywords for Event Detection in Soccer Video," in Proceedings of International Conference on Multimedia and Expo, pp. 281-284, 2003.

[14] D. A. Reynolds.: A Gaussian Mixture Modeling Approach to Text-Independent Speaker Identification. PhD thesis. Electrical Engineering Department, Georgia Institute of Technology,. 2000

[15] W.Zhou, A.Vellaikal, and C. J. Kuo, "Rule-based Video Classification System for Basketball Video Indexing," in Proceedings of ACM Multimedia 2000 workshops, 2000

저 자 소 개



이 광 국

- 2002년 : 한양대학교 전자전기공학부 학사
- 2004년 : 한양대학교 전자통신전파공학과 석사
- 2004년~현재 : 한양대학교 전자통신전파공학과 박사과정
- 주관심분야 : 내용기반 멀티미디어 분석

강 정 원

- 1993년 : 한국항공대학교 항공전자공학과 (학사)
- 1995년 : 한국항공대학교 항공전자공학과 신호처리전공 (석사)
- 2003년 : Georgia Institute of Technology ECE (공학박사)
- 2003년~현재 : 한국전자통신연구원 방송미디어연구그룹 선임연구원
- 주관심분야 : 비디오 신호처리, 비디오분석, MPEG-7, MPEG-21

김 재 곤

- 1990년 2월 : 경북대학교 전자공학과 (학사)
- 1992년 2월 : KAIST 전기 및 전자공학과 (석사)
- 2005년 2월 : KAIST 전기 및 전자공학과 (박사)
- 2001년 9월~2002년 11월 : 뉴욕 콜롬비아대학교 연구원
- 1992년~현재 : ETRI 방송미디어연구그룹 방통융합미디어연구팀장/선임연구원
- 주관심분야 : 영상통신, 비디오 신호처리, 디지털 방송, 멀티미디어 프레임워크, TV-Anytime/MPEG-7/MPEG-21

김 회 율

- 1980년 : 한양대학교 전자공학과 졸업 (공학사)
- 1983년 : Pensylvania State University 전기공학과 졸업 (공학석사)
- 1989년 : Purdue University 전기공학과 졸업 (공학박사)
- 1989년 9월~1994년 2월 : University of Texas 조교수
- 1994년~현재 : 한양대학교 전자통신컴퓨터공학부 정교수
- 주관심분야: 영상처리, 컴퓨터비전, 패턴인식, 머신비전, MPEG-7등