영상과 음성 정보를 이용한 비디오 편집 및 검색 시스템

조현칠, 윤인구, 김우생
광운대학교 전자계산학과

Video Editing and Retrieval System Using Speech Recognition Technique

Hyunchul Cho, Inku Yoon , Woosaeng Kim
Department of Computer Science, Kwangwoon University

요약

동영상 데이터가 갖는 복잡하고 다양한 관계성 때문에 기존의 키워드 기반 정보 검색 방법에는 한계가 있으며 비디오 내용에 기반해 검색을 하는 내용기반 검색기법이 요구된다. 현재 MPEG-7에서도 비디오 내용 표현 방식에 관한 국제 표준화 작업이 시작 되고 있다. 본 논문에서는 영상정보와 음성정보를 사용해 비디오의 원하는 부분을 내용에 기반해 검색할 수 있는 비디오 편집및 검색 시스템을 개발하였다.

1. 서론

컴퓨터와 통신 그리고 데이터 압축기술의 발달로 동영상 데이터를 이용한 다양한 서비스가 가능하게 되었다. 동영상 데이터는 멀티미디어 데이터의 가장 대표적인 데이터로 영상정보뿐만이 아니라 음성정보, 문자정보 및 각종 외미 있는 정보들을 포함하고 있다. 이렇게 복합적인 정보를 갖는 동영상 데이터로부터 사용자가 필요로 하는 정보를 찾기 위해서는 기존의 키워드 기반의 검색은 한계에 도달한 상황이기 때문에 사용자가 원하는 정보를 내용에 기반하여 검색할 수 있는 방법이 요구되고 있다. 이러한 최근의 기술 발전 추세 및 시장 요구를 바탕으로 하여, 국제 표준화 기구인 ISO와 IEC의 연합기술위원회 산하의 MPEG에서는 MPEG-7이라는 이름으로 멀티미디어 데이터의 내용기반 검색을 위한 내용 표현 방식에 관한 국제 표준화 작업을 시작하였다. 기존에 표준화되었거나 표준화가 진행되고 있는 MPEG-1/2/4이 오디오, 비쥬얼 데이터의 압축을 목표로 한 것과는 달리 MPEG-7은 데이터의 내용에 대한 표현 방법을 다루게 되는 것이다. 이를 메타데이터(Metadata), 또는 'Bits about bits'라고 하기도 한다.[1]

동영상에 대한 내용기반 검색을 위해서는 동영상 데이터를 색인하기 위한 비디오 파싱기법과 사용자가 원하는 데이터를 쉽게 검색할 수 있는 사용자 인터페이스 뿐만 아니라 내용기반 비디오를 효율적으로 저장하기 위한 비디오 데이터 압축 및 저장 방법 등의 기술들이 필요하다. 동영상 데이터를 색인하기 위하여 자동 색인기으로 사용할 수 있는 정보로 영상정보이다. 영상정보는 주로 비디오를 장면 분할할 때에 사용되며 이를 통하여 구조적인 비디오 브라우징을 할 수 있다. 비디오를 구성하는 최소단위는 하나의 영상을 나타내는 프레임이다. 비디오에서 장면의 전환이 이루어지는 부분을 컷(cut)이라고 하고, 컷으로 구분되어 하나의 카메라 조작에 의해 촬영된 짧은 비디오 단위를 샷(shot), 논리적인 내용이 같은 연속된 샷으로 이루어진 단위를 에피소드(episode)라 한다. 비디오를 샷으로 구분하는 작업을 비디오 분할(video segmentation)이라고 하며, 비디오 분할을 위해 장면의 전환점인 컷을 검출하는 작업을 컷검출(cut detection)이라고 한다. 다음으로 동영상의 내용검색을 위해서 사용될 수 있는 정보는 영상 내에 있는 문자정보이며, 마지막으로 동영상의 내용 검색을 위해서 사용될 수 있는 음성정보는 영상 내에 있는 오디오 정보이다. 오디오에 있는 음향정보나 음성 등을 제대로 인식하면 구조적인 비디오와 같은 분할을 할 수 있을 뿐만 아니라 오디오내의 음성정보를 대응하는 문자 정보로 변환한 후 인덱스를 만들어 내용기반 검색에 사용할 수 있다. 따라서 동영상 데이터에 대한 검색 시스템은 동영상에서 어떤 정보들을 사용했는가에 의존하게 된다. 만약 동영상에서 영상 정보만을 사용하였다면 영상분할을 통한 브라우저나 비슷한 색상이나 형태 등의 내용을 찾는 하위 레벨의 내용기반 질의를 할 수 있는 반면, 문자정보나 음성정보를 인식 기술과 함께 사용한다면 자동 인덱스를 만들 수 있어 자신이나 음성 등의 상위 레벨의 내용기반 질의를 가능케 해준다.[2] 근래에 국내외에서 동영상에 대한 내용기반 검색 기법에 관한 많은 연구가 있는데 영상의 정보만을 사용한 검색[3][4], 문자정보 또는 영상과 문자정보를 함께 사용한 검색[5][6], 오디오 정보 또는 영상과 오디오 정보를 함께 사용한 검색[7][8], 영상, 문자, 오디오 정보를 모두 사용한 검색[9]등이 있다.

본 논문에서는 인덱스된 동영상 라이브러리 구성을 위하여 동영상을 몇 개의 의미 있는 단편으로 구성할 수 있는 편집시스템과 위하는 비디오 단편을 내용에 기반한 방법으로 검색 또는 브라우징 할 수 있는 비디오 검색 시스템을 개발하였다. 특히 인덱스된 비디오 라이브러리를 만들기 위하여 영상정보와 함께 음성 정보를 이용하였다.

본 논문의 구성은 2장에서 비디오 검색 시스템의 전체적인 개요, 3장은 구현과 기능에 대해서 설명하고 마지막으로 4장에서 결론을 맺는다.

2. 비디오 검색 시스템

2.1 전체 시스템 구조

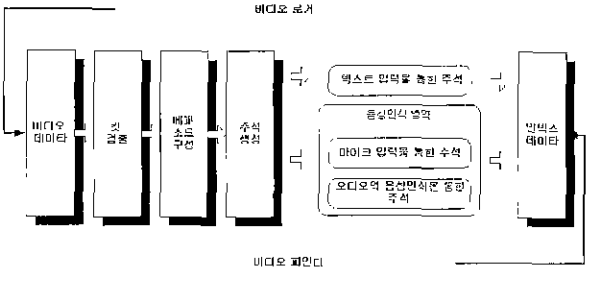

그림 1 전체 시스템 구조

전체 시스템은 비디오 로거(Video Logger)와 비디오 파인더(Video Finder)로 구성 되어있다. 비디오 로거는 비디오를 편집하는 시스템으로 원하는 비디오의 단편에 주석을 덧붙여서 하나의 에피소드를 만들 수 있다. 비디오 파인더는 사용자가 원하는 비디오 단편을 내용에 기반한 사전이 질의로 찾을 수 있고 원한 비디오 내용을 브라우징 해볼 수 있는 검색 브라우징 시스템이다.

* 이 논문은 한국전자통신연구소 위탁과제의 일환으로 수행되었음.

2.2 비디오 로거

사용자는 비디오 로거를 통하여 비디오를 편집하여 인덱싱된 비디오 라이브러리로 만들 수 있다. 이를 위하여 사용자는 원하는 비디오 단편을 구성하는 첫 번째 프레임과 마지막 프레임을 지정할 수 있고 관련된 내용의 주석과 리포터나 장소의 정보를 입력하여 하나의 에피소드를 만들 수 있다. 비디오 라이브러리를 만들기 위해 다음과 같은 과정을 거친다.

1) 비디오 샷 생성

사용자가 비디오를 재생하면서 원하는 부분을 지정하여 에피소드로 만들 수 있지만 비디오 전체의 내용을 일일이 보면서 편집해야 하므로 많은 시간이 소요된다. 이러한 비디오 편집 시간을 단축하기 위해 비디오 로거에서는 자동으로 샷들을 찾아 내주며 사용자는 인터페이스에 나타난 샷을 대표하는 프레임들 중에서 원하는 비디오의 단편들을 선택하여 에피소드로 만들어 줄 수 있다. 로거에서는 샷을 2가지 방법으로 생성할 수 있다. 하나는 히스토그램을 기반으로한 샷 검출 방법이고 다른 하나는 지정된 수만큼의 샷을 균등 분할해 주는 방법이다. 히스토그램에 기반해 샷을 만드는 경우는 영상의 장면 변화가 심한 경우 샷의 개수가 많이 생성될 수가 있다. 반면에 상해신 수만큼의 샷을 회면에 요청한 경우는 균등하게 비디오를 분할하여 샷을 만드나 장면 변화와 무관하게 샷이 만들어질 수 있다는 단점이 있다.

2) 에피소드 편집

검색이나 구조적인 비디오 브라우징을 하기 위해 사용자는 같은 내용의 샷들을 모아 하나의 에피소드를 만들 필요가 있다. 사용자는 에피소드 주석을 설정한 후 이에 대한 주석을 입력 할 수 있는데, 키보드를 통한 텍스트 입력방법과 음성인식을 이용한 입력방법이다. 음성인식에 의한 주석 생성 방법에는 두 가지 방법을 제공한다. 첫 번째는 마이크(MIC)를 사용하여 주석을 입력하는 것으로 사용자가 주석에 해당하는 내용을 발음하면 이것을 인식하여 주석이 생성되는 방법이다. 이 방법은 화자 종속 인식 방법으로 특정 화자만을 대상으로 인식하는 방법이기 때문에 인식률이 높다. 두 번째 방법은 비디오 데이터의 오디오 영역의 음성(대사)을 인식하여 자동으로 주석을 만드는 방법이다. 이 방법은 화자의 특성을 고려하지 않고 인식하는 화자 독립 방법으로, 자동으로 주석을 생성할 수 있다는 장점은 있지만 인식률이 떨어진다.

3) 인덱싱된 비디오 라이브러리 생성

생성된 에피소드는 인덱스된 비디오 데이터베이스에 저장되며 에피소드의 주석과 관련 정보를 위하여 비디오와 에피소드에 관한 데이터를 사용하였다. 비디오 테이블에는 동영상 파일에 관한 정보가 저장되는데 분류를 위한 비디오ID와 비디오제목, 생성날짜, 파일 이름이 사용된다. 에피소드 테이블에는 에피소드의 관련 비디오를 연관시키는 정보가 저장되는데 에피소드ID, 에피소드가 포함된 비디오ID, 에피소드의 이름과 구간의 키워드와 비디오 프레임 번호, 리포터, 장소가 사용된다. 음성인식 또는 텍스트 입력을 통해 생성된 주석은 비디오 인덱서에서 효율적인 검색을 위해서 인덱싱 되어 저장되는데, 인덱스 파일의 생성과정은 다음과 같은 3단계로 구성된다.

1 파싱

사용자에 의해 수동 혹은 자동으로 입력된 주석의 문장을 분석한다. 파싱의 목적은 다이어리 정보이나 그 외의 특수 문자들은 자동으로 걸러지게 된다. 파싱의 결과로서 주석의 문장에 포함된 모든 단어들을 얻게 된다.

2 키워드 추출

파싱을 통하여 얻어진 단어들 중에서 인덱스에 저장될 키워드들을 찾아내기 위하여 미리 정의된 사전과의 비교작업이 수행된다. 만약 찾은 단어가 사전에 있을 경우 그 단어는 키워드로 선택되어 진다.

3 인덱스 생성

위의 과정에서 생성된 키워드와 관련된 에피소드의 정보, 즉 에피소드 데이블의 레코드 위치를 포함하는 인덱스 파일이 생성된다.

2.3 비디오 파인더

비디오 파인더는 자연어 기반의 검색이나 구조적인 비디오 브라우징을 할 수 있다. 검색시 사용자가 단어나 문장을 입력하면 그것을 분석하여 적절한 질의어를 생성하여 검색결과를 제공하며 AND/OR검색을 사용하여 비디오 ID, 제목, 날짜, 리포터, 장소 필드들과도 조합하여 원하는 검색결과를 얻을 수 있다.

1) 질의 검색

사용자가 입력한 질의를 비디오 파인더는 검색을 위하여 적절한 형태로 파싱을 하게 되며 파싱의 결과로써 하나이상의 단어가 추출되면 에피소드 주석에 의해 만들어진 인덱스 파일에 포함된 단어들과 비교하게 되는데, 단어를 포함하는 에피소드들은 하나의 비디오에서 추출될 수도 있고 여러 개의 비디오에서 추출될 수도 있다. 사용자가 주석이나 리포터, 장소를 가지고 질의를 할 경우 그러한 정보를 포함하고 있는 에피소드들을 찾아낸 후 하나 에피소드들을 화면에 띄워주면 되나 그러나 만약 사용자가 비디오 ID나 제목 날짜만으로 질의를 할 경우, 시스템은 해당되는 비디오를 먼저 찾아낸 후 해당 비디오가 에피소드를 포함하고 있는지를 먼저 확인한 다음 에피소드들을 포함하고 있으면 해당 에피소드들이 검색되고 에피소드를 포함하고 있지 않다면 시스템은 그 비디오의 전체를 하나의 에피소드로 간주하여 비디오 전체의 중간 프레임을 에피소드의 대표 프레임으로 선정하여 이를 화면에 띄워준다. 에피소드들을 아이콘이나 테이블 형태로 보여줄 수 있다. 아이콘 형태로 보여줄 때는 각 에피소드마다 대표하는 프레임을 선정하여 화면에 보여주는데, 본 논문에서는 에피소드를 대표하는 프레임으로 에피소드 구간의 중간 프레임을 선정하였다.

2) 브라우징

사용자는 해당 에피소드의 내용을 좀더 자세히 보기 위해 에피소드를 구성하고 있는 샷들을 디스플레이해 보거나 에피소드의 비디오 단편을 재생해 볼 수 있다. 본 논문에서는 하나의 에피소드 구간을 7 등분하여 7개의 해당 샷을 보여준다. 또한 해당 에피소드의 첫 번째 프레임부터 마지막 프레임까지의 동영상 전체 혹은 일부를 재생해 볼 수 있다. 비디오를 재생할 경우 fast forward, rewind, stop, resume등의 기능을 사용하여 비디오 재생을 조작 할 수 있도록 하였다.

3. 구현및 기능

전체적인 인터페이스와 데이터베이스 관리를 위해 마이크로소프트 비주얼 베이직과 제트엔진을 사용하였고 비디오 데이터의 조작 및 재생을 위해 관련 OCX를 사용하였다. 비디오 데이터의 영상 처리를 위해서는 C++과 Windows API를 사용하였고 음성인식을 위해 CMU의 Sphinx II엔진의 변종된 형태인 MS(MicroSoft)의 Speech Recognition Engine과 SAPI 4.0 SDK를 사용하였다. 동영상은 AVI 파일 오디오는 wave 파일을 사용하였다.

3.1 비디오 로거 인터페이스

1) 비디오 관리 및 재생

비디오 정보의 관리를 위해 데이터 억세스 개체(DAO)를 사용하였으며 이를 통해서 사용자는 원하는 비디오를 추가 삭제 검색을 할 수 있으며, 비디오 ID, 제목, 날짜 등도 마찬가지로 쉽게 관리할 수 있다. 또한, 비디오관련 OCX를 이용하여 play, pause, rewind, fast forward 기능을 구현하여 비디오를 보거나 탐색할 수 있으며, 사용자가 원하는 프레임구간을 지정하여 부분적인 재생을 할 수 있다.

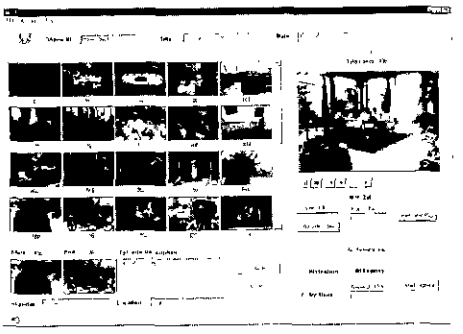

그림 2 비디오 로더

2) 샷 생성

입력돼면 비디오를 에피소드 단위로 편집하기 위해서 샷을 생성하여 준다. 샷 생성은 디폴트로 40프레임의 균등분할로 지정되어 있지만 사용자는 임의의 프레임 수를 지정할 수도 있고, 히스토그램을 이용한 컷 검출 기법으로 변화된 프레임만을 검출할 수 있다. 샷을 생성한 후 그 결과는 첫을 대표하는 프레임들의 리스트형태로 출력되어진다.

3) 에피소드 생성

사용자는 생성된 샷의 대표 프레임들의 리스트 혹은 재생되고 있는 비디오 창에서 적절 원하는 프레임들을 Drag & Drop 방식을 사용하여 에피소드로 구성할 수 있다. 또한 사용자는 해당 에피소드에 대한 주석, 장소 리포터 등의 정보를 함께 입력할 수 있다. 에피소드의 주석의 경우는 음성인식 기술을 사용하였는데, 음성인식을 위해 MS의 음성인식 엔진인 Whisper와 SAPI를 사용하였다. 마이크 입력을 통한 화자종속 음성인식의 경우는 Speech Dictation Control을 통해서 제어가 가능하게 했고, 비디오 데이터의 오디오 영역의 화자독립 음성인식의 경우는 Speech Recognition Control을 통해서 제어를 하였다. 인식률은 테스트 환경에 의존하지만 개략 실지의 경우는 75%, 후지의 경우는 35%정도의 성능을 보였다.

3.2 비디오 파인더 인터페이스

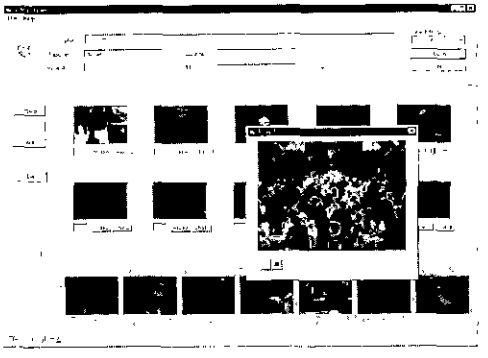

그림 3 비디오 파인더

1) 질의 검색기능

키워드, 장면의 내용을 묘사하는 리포터, 장소, 비디오 ID, 비디오 제목, 날짜 등의 정보로 질의할 수 있다. 사용자는 또한 AND OR 연산자를 이용하여 일의 질의조건에 대한 복합 검색을 사용할 수 있다.

2) 에피소드 정보제공

사용자가 검색된 에피소드의 정보를 보는 방식은 Tabular 형태와 Icon 형태의 두 가지가 제공된다. Tabular 형태는 사용자에게 검색된 모든 에피소드들의 정보를 테이블을 통해서 한눈에 보여준다. Icon 형태는 검색된 특정 에피소드에 관한 정보를 Icon 형태로 제공하는데, Icon에 나타나는 화면은 에피소드의 중간 프레임이다. 사용자가 Icon 위로 마우스를 가지다 놓으면 툴팁 형태로 비디오 ID, 리포터, 제목 등의 해당 에피소드의 정보가 나타나도록 하였다.

3) 샷 정보와 동영상 재생 기능

검색의 결과로 얻어진 에피소드의 Icon에 붙어있는 Shot 버튼을 클릭 하면 에피소드를 구성하고 있는 샷들을 하나의 그림상자 리스트에 보여주며, 'Video' 버튼을 클릭 하면 해당 에피소드의 동영상이 동작으로 생성된 창을 통해 보여지게 된다. 이를 통해서 사용자는 에피소드를 구성하고 있는 화면들을 좀더 구체적으로 분석할 수 있다.

4. 결론

복합적인 정보를 갖는 동영상 데이터로부터 사용자에게 원하는 정보를 제공하기 위해서는 내용기반의 검색이 필요하다. 멀티미디어 신호의 효율적인 검색을 지원하기 위하여 새로 시작된 MPEG-7 국제 표준화의 원래 전자 상황과 이의 관련된 내용 기반 정보 검색 기술의 연구 분량을 비추어 볼 때 디지털 AV 신호 처리, 컴퓨터 비전 자인이 처리, 멀티미디어, 음성 인식 및 데이터베이스 분야를 포함한 다양한 영역의 지식들이 복합적으로 활용될 것으로 보이기 때문에 이를 이용한 연구가 시급한 실정이다.

본 연구에서는 비디오 데이터의 내용기반 검색을 위한 비디오 편집 및 검색시스템인 비디오 로거/파인더에 대하여 설명했다. 국내의 동영상 데이터의 내용기반 검색 연구가 주로 영상정보와 문자 정보만을 사용한 반면 본 연구에서는 오디오 정보도 활용했다는 점이 차이라고 할 수 있다. 현재 착수중에 있는 비디오의 자동 분할을 통한 내용에 기반한 검색 시스템을 개발하기 위해서 영상, 음성, 문자, 음향 정보등을 이용한 비디오 검색 시스템을 개발중에 있다.

5. 참고 문헌

[1] ISO/IEC JTC1/SC29/WG11, "MPEG-7 Requirements Document V.5", MPEG98/N2208, Tokyo, March 1998

[2] 위문철, 김진웅, 김우생, MPEG-7 표준화 및 내용기반 검색, 대한전자공학회지 1998 8

[3] 권경복, 조영우, 김영모, "구조화된 논리적 비디오를 이용한 비디오 브라우징 및 검색 시스템", 한국정보처리학회 논문지 제4권 6호, 1997

[4] Shih Fu Chang et al, "VideoQ An Automated Content Based Video Search System Using Visual Cues" ACM Multimedia 97, 1997

[5] R Lienhart, "Automatic Text Recognition for Video Indexing", ACM Multimedia 96, 1996

[6] 이미숙, 방건, 한영규 용일기 김우식, 이성환, "내용기반 색인 및 검색을 위한 실시간 뉴스 비디오 파서의 설계 및 구현" 한국정보과학회 봄 학술 발표논문집 Vol24, No.1, 1997

[7] K Minami, A Akutsu, H Hamada Y Tonomura "Enhanced Video Handling based on Audio Analysis", Proceedings of the International Conference on Multimedia Computing and System, 1997

[8] Y-L Chang et al, "Integrated Image and Speech Analysis for Content-Based Video Indexing", International Conference on Multimedia Computing and Systems 1996

[9] A and M Witbrock, "Informedia News-On-Demand Using Speech Recognition to Create a Digital Video Library", Symposium on Intelligent Integration and Use of Text, Image, Video and Audio Corpora, 1997