샷 경계 검출을 이용한 영상 클립 생성
(Generation of Video Clips Utilizing Shot Boundary Detection)

김 혁 만 † 조 성 길 ††
(Hyeokman Kim) (Sungkil Cho)

요 약 대용량 영상을 다루는 디지털 비디오 라이브러리나 웹 방송에서는 영상 색인이 매우 중요한 역할을 하며, 이는 영상을 내용 단위로 분할하는 알고리즘에 기반한다. 본 논문에서 구현된 V2Web Studio는 영상 색인을 지원하는 시스템으로서, 샷 경계 검출 알고리즘을 이용한 영상 클립 생성 시스템이다. V2Web Studio는 영상 클립 생성 과정을 1) 영상 신호를 분석하여 샷 경계를 자동 검출하는 단계, 2) 검출된 결과에 포함될 수 있는 오류를 수작업으로 제거하는 단계, 3) 물리적인 샷 경계를 논리적인 계층구조로 모델링하는 단계, 4) 계층구조로 모델링된 각 모델링 인스턴스를 다양한 표준 압축 포맷으로 생성하는 단계로 구분하고, 각 단계에 해당하는 작업은 샷 검출기, 샷 검증기, 영상 모델기, 클립 생성기라는 독립적인 소프트웨어 도구로 구현하였다.

Abstract Video indexing plays an important role in the applications such as digital video libraries or web VOD which archive large volume of digital videos. Video indexing is usually based on video segmentation. In this paper, we propose a software tool called V2Web Studio which can generate video clips utilizing shot boundary detection algorithm. With the V2Web Studio, the process of clip generation consists of the following four steps: 1) Automatic detection of shot boundaries by parsing the video, 2) Elimination of errors by manually verifying the results of the detection, 3) Building a modeling structure of logical hierarchy using the verified shots, and 4) Generating multiple video clips corresponding to each logically modeled segment. The aforementioned steps are performed by shot detector, shot verifier, video modeler and clip generator in the V2Web Studio respectively.

1. 서 론

인터넷의 대역폭이 증가함에 따라 인터넷에서 영상을 제공하는 VOD 형태의 웹 서비스가 급속히 확산되고 있다. 대규모 웹 VOD 서비스에서는 아날로그 영상에서 압축 영상 클립과 메타 정보를 빠르고 정확하게 생성하는 과정이 중요하다.

매일 만들어지는 뉴스를 제공하는 웹 VOD 사이트의 경우, 일반적으로 뉴스 프로그램의 각 뉴스 아이템을 압축 포맷으로 변환해 저장한다. 일반적으로 각 뉴스 아이템은 뉴스 앵커가 내용을 요약 설명하는 앵커 샷과, 그 뒤를 이어 내용과 관련된 여러 배경 샷들이 연속해서 나타난다. 만일 아날로그 형태의 뉴스에서 뉴스 아이템들을 각각 독립적인 MPEG 영상 클립으로 만든다면, 이 클립은 앵커 샷의 첫 번째 프레임에서 시작해서 다음 뉴스 아이템의 앵커 샷 직전 프레임에서 끝나야한다. 그러나 현재의 MPEG 압축기들은 빨리감기, 되감기 등의 VCR 기능을 이용해 아날로그 영상을 제어하므로, 프레임 단위로 정확히 압축할 프레임의 시작과 끝을 설정하기 매우 힘들다. 따라서 작업자의 숙련도에 따라 몇 개 혹은 몇 십개 프레임의 오차가 발생할 수 있다.

그림 1은 어떤 뉴스 프로그램의 일부를 나타내고 있다. 뉴스 아이템 n의 압축 파일을 만들 경우, 이상적인 형태는 그림의 첫 번째 구간만을 압축하는 것이다. 그러나 VCR 기능을 이용해 수작업으로 만들 경우 이와 같은 완벽한 구간을 선택하기 힘들므로, 실제로는 두 번째나 세 번째와 같은 오차 구간이 포함되어 압축된다. 두 번째 클립의 경우 끝 부분에 다음 뉴스 아이템의 앵커 샷 일부가 포함되며, 세 번째 클립에서는 시작 부분에서

† 정 회 원 : 국민대학교 컴퓨터학부 교수
hmkim@cs.kookmin.ac.kr
†† 학생회원 국민대학교 전산학과
makeway@cs.kookmin.ac.kr
논문접수 : 2001년 4월 9일
심사완료 : 2001년 8월 25일

직전 뉴스 아이템의 배경 샷이 일부 포함된다. 따라서 이런 클립을 재생시키면 처음 혹은 마지막 부분에 불필요한 내용이 보여 어색한 재생이 이루어진다. 클립을 만드는 거의 대부분의 작업자들은 이를 방지하기 위해 네 번째와 같이 처음과 끝이 약간 제외된 상태가 될 때까지 압축 및 재생을 반복하여 가장 잘된 클립을 선택하고 있다.

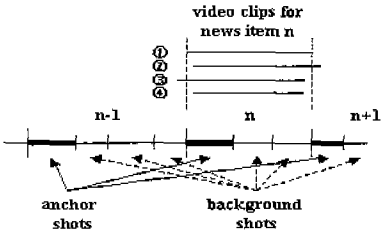

그림 1 영상 클립의 생성

실제로 수십개의 뉴스 아이템으로 이루어진 하나의 뉴스 프로그램을 뉴스 아이템별로 독립적인 압축 영상 클립으로 만들 때 지금까지는 다음의 두 가지 방법을 사용하고 있다. 첫 번째 방법은 각 뉴스 아이템 별로 독립적으로 압축시킨다. 즉 VCR 기능을 이용해 아날로그 영상을 제어하면서 각 뉴스 아이템별로 압축시킨다. 이 방법에서 생성되는 뉴스 클립은 프레임 단위로 정확하게 원하는 구간이 포함되지 않는 경우가 대부분이다. 뿐만 아니라 1시간 정도의 뉴스에서 뉴스 아이템별로 압축 영상 클립으로 만드는데는 그 몇 배의 시간이 소요된다.

두 번째 방법은 뉴스 프로그램 전체를 한번에 압축한 후에, 하드웨어 혹은 소프트웨어 영상 편집기를 이용해 압축된 영상을 뉴스 아이템별로 쪼개어 독립적인 여러 개의 압축 영상 클립을 생성하는 것이다. 이 방법은 현재 일반적으로 선호되는 방법이다. 그러나, 만일 같은 구간을 두 개 이상의 서로 다른 포맷의 클립으로 만들 경우에는 작업 시간이 매우 많이 소요되며, 동일한 반복적인 일을 작업자가 수동으로 계속해야 하는 단점이 있다. 또한 사용하는 편집기가 프레임 단위로 정확하게 자르는 작업을 지원하지 못하는 경우, 생성되는 영상 클립이 정확한 구간을 포함하지 못하게된다. 특히 웹 VOD 사이트에서 제공하는 저대역폭 압축 포맷을 위한 편집기들은 프레임 단위로 정확히 편집하는 기능을 제공하지 못하는 경우가 대부분이다.

스트리밍 영상을 보내는 대부분의 인터넷 사이트는 마이크로소프트의 ASF 포맷과 리얼비디오의 RV 포맷, 애플의 MOV 포맷 영상을 다양한 네트워크 대역폭에 따라 제공한다. CNN 뉴스를 제공하는 CNN.com 사이트의 경우 각각의 뉴스 아이템을 28.8Kbps와 80Kbps 대역폭으로 각각 ASF 및 RV 포맷, 그리고 56Kbps의 MOV 포맷으로 제공한다. 즉 하나의 뉴스 아이템마다 동일한 내용을 담는 5개 파일이 인코딩되어야 한다. 이런 응용의 경우 첫 번째 방법으로 압축 영상 클립을 만드는 것은 거의 불가능하며, 두 번째 방법으로 영상 클립을 생성하더라도 반복적인 많은 수작업이 필요하며 작업 시간도 많이 걸린다. 뿐만 아니라 생성되는 영상 클립의 내용도 작업자의 숙련도에 따라 약간씩 차이가 날 수 있다.

따라서 원하는 구간을 정확히 포함하면서, 영상 클립 생성에 필요한 작업 시간을 최대한 단축시키는 자동화된 소프트웨어 도구가 필요하다. 이런 도구는 매일 꾸준히 몇 시간씩의 영상을 처리하여 저장하는 대규모 영상 데이타베이스 및 이를 이용한 웹 VOD와 같은 다양한 응용 시스템에서는 필수적이다.

한편 방송이 디지털화 하면서 방송 편집도 디지털화되고 있다. 디지털 방송 편집에서는 주로 화질이 우수한 M-JPEG(Motion JPEG) 포맷을 사용하며, 편집이 끝난 후의 송출 및 내부 저장용으로는 MPEG 포맷을 사용한다. 또한 이들을 인터넷으로 스트리밍시킬 때에는 저대역폭의 ASF, RV, MOV 혹은 H.263/G.723을 사용한다.

본 논문은 현재 개발중인 V2Web (Videos to the Web) digital video library 시스템의 주요 서브시스템인 V2Web Studio에 대하여 기술한다. V2Web 시스템은 디지털 방송 및 웹 VOD 서비스 개발에 필수적인 소프트웨어들로 이루어진 개발 플랫폼이다. V2Web Studio는 디지털 편집이 완료된 M-JPEG 포맷의 영상(뉴스 프로그램)으로부터 임의의 짧은 단위의(뉴스 아이템별로) MPEG1 영상과 H.263/G.723 영상 클립을 생성하는 자동화된 도구이다. 즉 V2Web Studio에서는 M-JPEG 포맷의 영상을 입력으로 하여, 대역폭이 높은 인트라넷에서는 1 Mbps급 고대역폭 고화질 MPEG1 영상, 대역폭이 작은 인터넷에서는 28.8 Kbps 모뎀을 위한 저대역폭 영상인 H.263/G.723 영상을 제공한다. 또한 영상의 내용을 스트리밍하여 보기 전에 영상의 내용을 브라우징하기 위해 주요 장면의 대표화면(key frame)으로 추출하는 기능도 제공한다. V2Web Studio는 MPEG1 및 H.263/G.723 영상 클립과 대표화면 생성 과정을 영상 해석(video parsing) 알고리즘을 이용해 최대한 자동화한 도구이다.

본고의 구성은 다음과 같다. 2장에서는 기본적인 용어

정의와 개념을 살펴본다. 3장에서는 V2Web Studio의 주요 구성 요소인 샷 검출기, 샷 검증기, 영상 모델기, 블립 생성기에 대해 설명한다. 마지막으로 4장에서는 결론 및 앞으로의 연구 방향을 서술한다.

2. 영상 모델링과 영상 해석

2.1 영상 모델링

영상 구성의 기본 단위는 샷(shot)이다. 샷이란 촬영 시에 카메라가 멈춤 없이 한 번에 기록한 연속적인 프레임이다[2]. 일반적으로 샷은 1~10초 정도이다. 샷은 물리적인 특성을 나타내는 것으로서, 시청자의 입장에서 논리적으로 인식되는 단위로 볼 수는 없다[1]. 시청자의 입장에서는 샷보다 상위의 논리적 인식 단위가 필요하다. 본고에서는 영상을 세그먼트(segment), 장면(scene), 샷(shot)의 세 단계 모델링 단위를 사용해 모델링한다.

임의의 영상 V는 다음과 같은 세그먼트 인스턴스(instance) SE_i $(1 \le i \le N)$의 연속이다.

$$V = (SE_1, SE_2, \ldots, SE_N)$$

세그먼트 인스턴스 SE_i는 장면 인스턴스 SC_{ij} $(1 \le j \le N_i)$의 연속이다.

$$SE_i = (SC_{i1}, SC_{i2}, \ldots, SC_{iN_i})$$

장면 인스턴스 SC_{ij}는 샷 인스턴스 SH_{ijk} $(1 \le k \le N_{ij})$의 연속이다.

$$SC_{ij} = (SH_{ij1}, SH_{ij2}, \ldots, SH_{ijN_{ij}})$$

샷 인스턴스 SH_{ijk}는 프레임 F_{ijkl} $(1 \le l \le N_{ijk})$의 연속이다.

$$SH_{ijk} = (F_{ijk1}, F_{ijk2}, \ldots, F_{ijkN_{ijk}})$$

즉 N_P개의 프레임으로 구성된 임의의 영상 V는 N개의 연속적인 세그먼트로 볼 수 있으며, 각 세그먼트 SE_i는 N_i개의 연속적인 장면으로 볼 수 있다. 또 특정 장면 SC_{ij}는 N_{ij}개의 연속적인 샷으로 볼 수 있으며, 특정 샷 SH_{ijk}는 N_{ijk}개의 연속적인 프레임으로 구성된다.

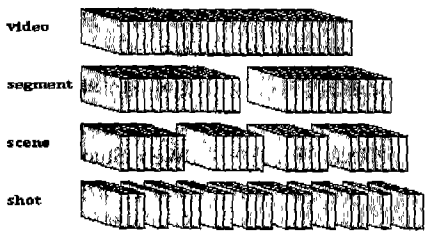

그림 2 모델링 인스턴스간의 계층 구조

$$N_P = \sum_{i=1}^{N} \sum_{j=1}^{N_i} \sum_{k=1}^{N_{ij}} N_{ijk}$$

설명한 샷, 장면, 세그먼트는 영상의 물리적, 논리적 경계를 구분하는 모델링 단위이다. 특정 영상에서 모델링 단위에 의해 분할된 프레임의 연속은 모델링 인스턴스이다. 정의에 따라 모델링 인스턴스들은 계층 구조를 이룬다. 그림 2는 모델링 인스턴스들간의 계층 구조를 나타낸다.

2.2 영상 해석

영상은 촬영한 원시 테이프에 있는 샷들을 선택하여 컷(cut), 와이프(wipe), 디졸브(dissolve) 등의 편집 효과를 사용하여 연결하는 편집 과정을 통해 제작된다[2]. 영상 해석은 편집된 영상의 신호를 분석하여 역으로 편집 단위, 편집 효과, 카메라 효과 등을 자동적으로 찾아내는 것을 말한다. 영상 해석의 주요 분야인 샷 및 장면 경계 검출이란 편집 효과를 검출하여 편집시 연결해 놓은 샷과 장면의 경계를 찾아내는 것이다.

앞에서 정의한 모델링 단위 중 샷은 물리적 특성에 기인하나, 장면은 물리적 및 논리적 성격을 모두 갖는다. 장면은 동일한 배경, 동일한 객체 혹은 사람, 동일한 주제를 갖는 인접한 샷들의 묶음(cluster)으로 정의할 수 있다[1]. 따라서 같은 장면에 속하는 샷들은 시각적으로 비슷한 특성을 갖을 수 있다. 그러나 경우에 따라서는 시각적 특성이 비슷하지 않은 샷들도 같은 장면으로 묶어야 하는 경우가 있으며, 이 점이 장면 경계 검출 자동화를 어렵게 만드는 요소이다. 반면에 세그먼트는 논리적 의미의 단위(story unit)로서, 독립적으로 내용을 전달할 수 있어야 한다. 일반적으로 시청자는 영상을 장면, 혹은 세그먼트 단위로 인식한다. 만일 영상을 책과 비교한다면 프레임에 나타나는 객체들은 단어, 프레임은 문장, 샷은 문단, 장면은 비슷한 내용을 다루는 몇 개의 연속적인 문단, 세그먼트는 장 혹은 절에 해당할 수 있다.

그림 3의 뉴스 세그먼트를 살펴보자. 이 뉴스는 공중전화에서 휴대전화로 전화할 때의 요금에 관한 총 90초 분량의 내용으로, 29개 샷으로 구성되어 있다. 그 중 첫 번째 앵커 샷이 약 30초이므로, 나머지 샷들은 평균 2.14초(64.2 프레임)의 분량이다. 그림 3에서는 각 샷의 대표화면을 작은 축소화면(thumbnail)으로 나타내었다. 이 세그먼트의 샷들을 하나의 뉴스 앵커 샷, 5개의 공중전화 관련 샷, 13개의 휴대전화 관련 샷, 10개의 휴대전화 제공회사 관련 샷으로 묶어보면 4개의 장면으로 해석이 가능하다. 이러한 뉴스 세그먼트들이 모이면 하나의 뉴스 프로그램이 구성된다. 이 뉴스 세그먼트를 재생하면, 빠르게 지나가는 샷들을 무의식중에 모두 인식하

기는 거의 불가능하다. 따라서 시청자들은 단지 설명한 4개의 장면 혹은 하나의 뉴스 아이템(세그먼트)으로 인식하게 될 것이다.

뉴스는 다른 영상에 비해 비교적 구조화가 잘 되어있다. 그러나 영화와 같이 구조화가 잘 되어있지 않은 영상에 대해서도 특정 구간에 대한 색인을 구성하려면 이와 같은 모델링의 도움이 필요하다. 본고에서의 영상 색인이란 설명한 모델링 단위의 어느 레벨을 이용해서도 주어진 영상의 특정 구간을 임의 접근할 수 있는 방법을 의미한다. 영상 색인을 위해 샷 경계를 검출(segmentation)하고, 이들을 내용에 따라 장면 혹은 세그먼트 단위로 묶는(clustering) 과정이 필요하다.

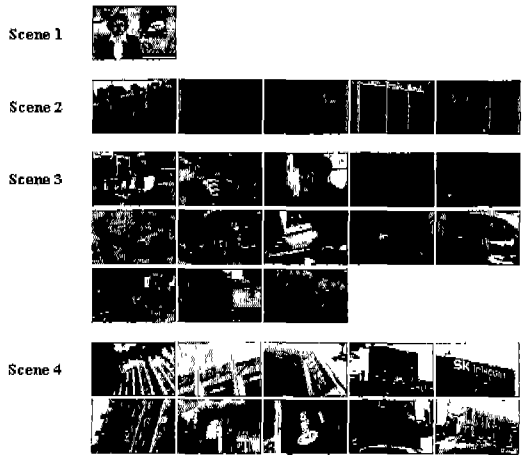

그림 3 뉴스 세그먼트의 예

3. 영상 클립 생성

V2Web Studio는 영상 해석 알고리즘과 모델링 개념을 결합하여 입력 영상으로부터 임의의 구간별로 여러 개의 압축 영상 클립, 영상의 계층 구조 정보, 그리고 대표화면을 최대한 자동으로 생성하는 시스템이다. 클립으로 생성할 단위는 샷, 장면, 세그먼트 중 어느 모델링 단위도 사용할 수 있다. 그림 4는 V2Web Studio에서 생성하는 정보를 나타낸 것이다. 클립으로 생성할 구간(뉴스 프로그램의 경우 뉴스 아이템)들이 정해지면, 각 구간의 영상을 MPEG1, H.263, G.723 포맷으로 압축하여 독립적인 파일로 만든다. 그리고 그 구간에서 정의된 몇 개의 대표화면을 JPEG 파일로 생성한다. 또 영상의 계층 구조 정보를 텍스트 파일로 생성한다. 이 시스템은 스튜디오 품질의 영상을 대상으로 설계되었기 때문에 직접 M-JPEG으로 디지털 편집된 영상을 입력으로 한다. 만일 Betacam 혹은 VHS 아날로그 영상이 입력이라면 이를 M-JPEG(Motion JPEG) 영상으로 디지털화한 후, M-JPEG 영상을 분석하여 최종 압축 파일을 생성하도록 설계되었다.

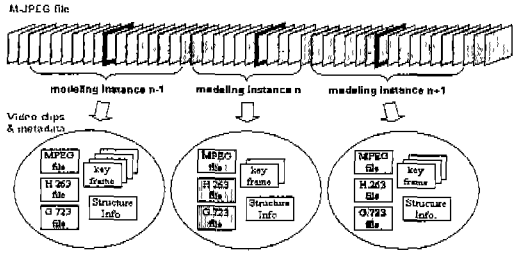

그림 4 V2Web Studio에서의 영상 클립 생성

V2Web 시스템에서는 영상 클립 생성 과정을 다음과 같은 단계로 나누어, 각 단계별로 독립적인 소프트웨어를 개발하였다: 1) 영상 신호를 분석하여 샷 경계를 자동 검출하는 샷 검출기(shot detector), 2) 검출된 결과에 포함될 수 있는 오류를 수작업으로 제거하는 샷 검증기(shot verifier), 3) 물리적인 샷 경계를 장면, 세그먼트 단위의 논리적인 계층구조로 모델링하는 영상 모델기(video modeler), 4) 계층구조로 모델링된 각 비디오 세그먼트를 다양한 표준 압축 포맷으로 생성하는 클립 생성기(clip generator). 각각의 도구들은 편리한 GUI를 사용하여 배우고 쓰기 편하도록 구현하였다. 그림 5는 시스템의 구조를 입출력 데이타 흐름을 중심으로 표현한 것이다. 여기서는 구현된 도구들을 차례로 설명한다.

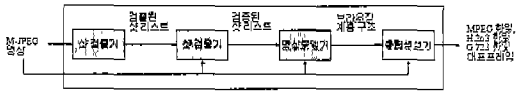

그림 5 V2Web Studio 시스템 구성도

3.1 샷 검출기

샷 검출기는 샷 경계 검출 알고리즘을 이용해 영상 신호를 분석하여 기본적인 구조 정보를 파악한다. 샷 경계 검출을 위한 다양한 연구가 진행되었다[2, 5, 6, 8, 10]. 그러나 대부분의 연구가 컷 검출에 집중되고 있으며, wipe과 dissolve 같은 점진적 전환은 거의 검출하지 못하고 있다. 또한 컷 검출마저도 미검출(missing shot)

과 오검출(false positive shot)을 상당히 내포하고 있어 실생활에서 사용 가능한 수준으로 실용화되지 못하고 있는 실정이다.

본 논문에서는 시각율동(visual rhythm)이라는 독특한 이미지를 사용하여 샷 경계를 검출한다[3]. 시각율동은 영상의 각 프레임의 대각선에 있는 화소들을 샘플링하여 수직으로 세우고, 이들을 수평(시간축)으로 누적시킴으로써 생성된다. 시각율동은 3차원 영상의 내용을 한 장의 2차원 이미지로 요약한 것이다. 그림 6은 시각율동의 개념을 나타낸 것이다.

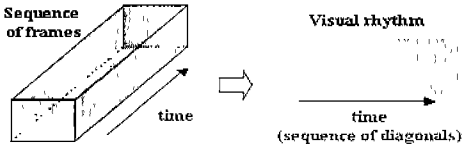

그림 6 시각율동

동일한 샷에 속하는 프레임에서 샘플링한 대각선 화소들은 거의 비슷한 시각적 특성을 지니므로, 시각율동의 샷 경계 부근에서는 두드러진 시각적 변화가 나타나게 된다. 즉 컷은 수직선으로 나타난다. 새로운 샷이 이전 샷을 화면에 수평한 방향으로 밀어내는 수평 와이프는 사선으로 나타나고, 새로운 샷이 화면의 중앙에서 나타나 이전 샷을 주변으로 밀어내는 확장 와이프는 곡선 혹은 방향이 반대인 두 사선이 시각율동의 중간에서 만나는 형태로 나타난다. 디졸브는 컷과 와이프같이 시각적으로 인식 가능한 선의 형태로 나타나지는 않으나, 색상이 점진적으로 꾸준히 변하는 모습을 띠게 된다. 그림 7은 설명한 컷, 와이프, 디졸브가 실제 영상으로부터 생성한 시각율동에서 어떻게 나타나는지를 보여주고 있다. 샷 검출기는 이러한 시각적 패턴을 검출하는 알고리즘을 이용하여 샷 경계를 찾아내는 모듈이다. 시각율동의 수학적 정의 및 생성 방법, 시각율동을 이용한 샷 경계 검출 방법 등은 [3, 4]를 참조하기 바란다.

3.2 샷 검증기

미검출과 오검출이 전혀 없는 완벽한 샷 경계 검출 알고리즘은 현재까지 존재하지 않는다. 시각율동을 이용한 알고리즘이 컷과 와이프, 그리고 모션이 많지 않은 디졸브의 경우는 기존 알고리즘 보다 매우 우수한 결과를 보여주고 있지만, 아직도 약간의 미검출 및 오검출을 내포하고 있다[3, 7]. 따라서 완벽한 결과를 얻기 위해서는 알고리즘의 결과를 사람의 눈으로 직접 검증하는 단계가 필요하다. 샷 검증기는 미검출 및 오검출된 샷을 쉽고 빠르게 찾아내어, 수작업으로 오검출된 샷을 제거하고 미검출된 샷을 검출된 리스트에 추가하는 도구이다[4].

그림 8은 구현한 샷 검증기의 사용자 인터페이스이다. 이 인터페이스는 크게 시각 율동부, 축소화면부, 제어부로 나눌 수 있다. 인터페이스 상단에 있는 시각 율동부는 생성한 시각 율동과 샷 검출기에서 찾아낸 샷 경계를 함께 디스플레이한다. 그림에서 알 수 있듯이 검출한 샷 경계는 시각 율동의 해당 위치 상단에 조그만 삼각형으로 표시한다. 시각 율동의 우측 상단에 있는 버튼은 시각 율동을 시간축으로 확대/축소하는 기능을 제공한다. 시각 율동 위의 커서는 축소화면부의 중앙에서 보고 있는 프레임의 위치를 나타낸다.

인터페이스 중앙 부분은 커서가 위치하고 있는 프레임, 그리고 전후로 인접한 각각의 열 개 프레임을 축소화면으로 나타낸다. 축소화면부의 중앙 좌측에 흘로 있는 축소화면이 현재 커서가 위치한 프레임이고, 그 위와 아래의 열 개 프레임이 커서가 위치한 프레임에 전후로 인접한 프레임들이다.

인터페이스 하단에 있는 제어부는 왼쪽의 영상 제어부와 오른쪽의 프레임 제어부로 나눌 수 있다. 프레임

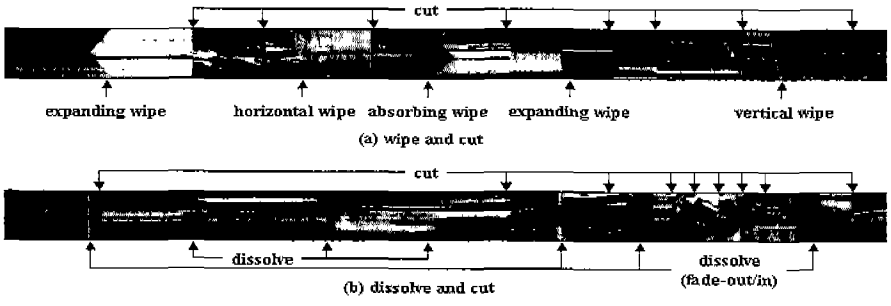

그림 7 실제 영상에서 생성된 시각율동

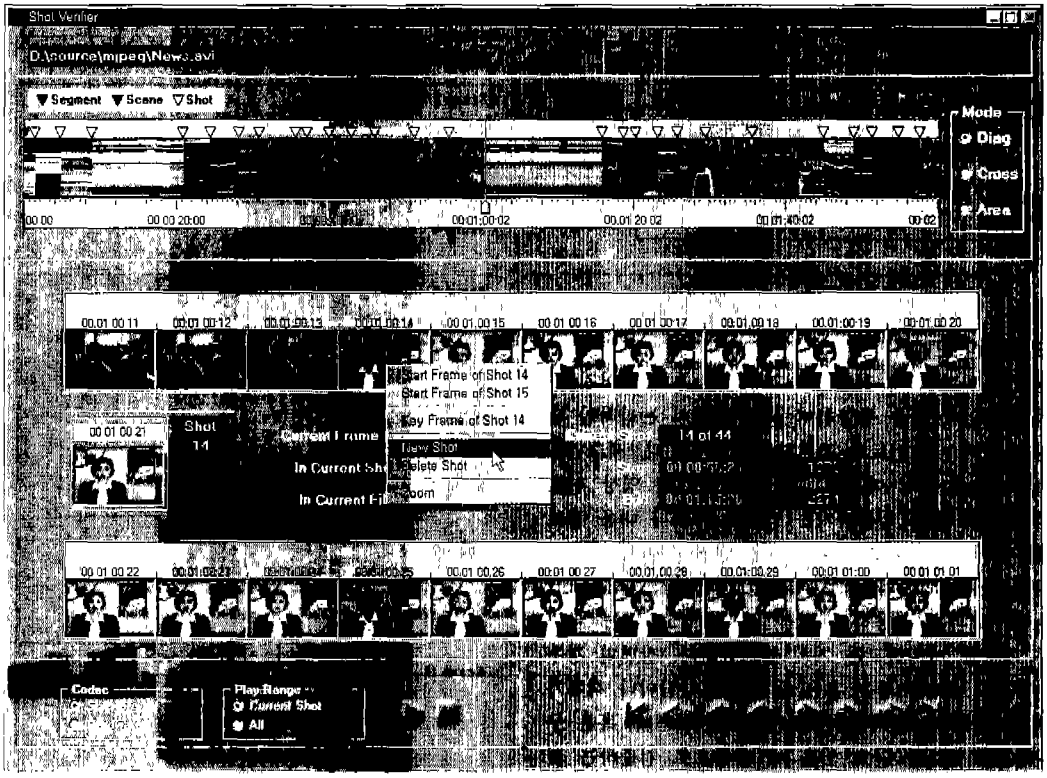

그림 8 샷 검증기의 사용자 인터페이스

제어부는 커서의 위치를 프레임 단위로 제어하는 기능을 제공한다. 즉 현재의 커서 위치에서 전후로 1, 5, 30 프레임 단위로 세밀하게 이동하는 기능을 제공한다. 만일 임의의 위치로 이동하려면 시각 율동부의 커서를 직접 끌어 붙이기(drag & drop)할 수도 있다. 커서의 위치가 바뀌면 이와 대응하여 축소화면부에도 새로운 축소화면이 디스플레이된다. 영상 제어부는 샷 단위의 영상 재생 기능을 제공한다. 재생 버튼을 누르면 새로운 창이 생성되어 영상이 재생된다.

샷 검증기를 이용한 검증 과정은 다음과 같다. 시각 율동부의 시각 율동을 살펴보다 의심스러운 부분이 있으면 커서를 그 부분으로 이동시킨다. 그림 7에서는 시각 율동부에서 컷으로 의심되는 직선 위에 컷 표시가 없으므로 커서를 해당 위치로 옮겼다. 그러면 축소화면부에 해당 프레임과 전후의 20개 프레임이 나타난다. 이들을 살펴봄으로써 샷 경계 여부를 거의 대부분 판별할 수 있다. 그림 8의 경우 축소화면부의 네 번째 프레임부터 새로운 샷이 시작됨을 알 수 있다. 만일 그 이상의 프레임들을 보거나 영상을 재생하려면 제어부의 기능을 이용한다. 미검출 및 오검출된 샷 경계를 찾아내면 축소화면부의 해당 프레임 위에서 마우스를 클릭한다. 그림 8의 경우 네 번째 프레임 위에서 마우스를 클릭한다. 그러면 그림 8과 같은 작은 메뉴 창이 나타난다. 이 창에서는 미검출된 샷을 새로운 샷으로 정의하고, 오검출된 샷을 제거하는 기능이 제공된다. 미검출된 샷을 정의하면 시각 율동의 해당 위치에 작은 삼각형 표시가 나타나고, 축소화면부의 해당 프레임 위에 샷 시작 표시가 나타난다. 오검출된 샷을 제거하면 시각 율동의 해당 위치에 있는 삼각형 표시가 사라지고, 축소화면부의 해당 프레임 위에 있는 샷 시작 표시가 사라진다.

시각 율동을 사용하면 영상을 재생시키지 않고도 대부분의 미검출 및 오검출된 샷 경계를 교정할 수 있다. 실제로 이 도구를 사용한 결과, 한 시간의 영상 내용을 불과 몇 분만에 검증하는 것이 가능하였다.

그림 9는 샷 검증기의 구조를 나타낸 것이다. 시각율동 관리자는 M-JPEG 복원기를 이용해 시각율동을 생성하고 제어하면서, 샷 검증에 필요한 화면을 제어하는 기능을 갖는다. 샷 검증이 완료되면 검증된 샷 리스트를

생성하며, 이 정보는 영상 모델기의 입력으로 사용된다.

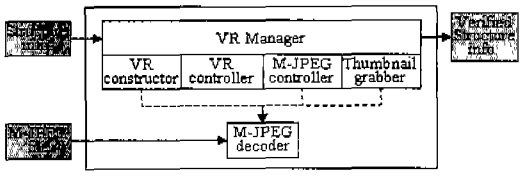

그림 9 샷 검증기의 구조

3.3 영상 모델기

영상 모델기는 샷 검출기와 검증기를 통해 얻어진 물리적 경계인 샷들을 통합하여 상위 레벨의 논리적인 단위인 장면과 세그먼트로 조직화하는 도구이다. 세그먼트는 story unit이기 때문에 그들간의 경계를 영상 신호 분석을 통해 자동 검출하기는 영상의 구조나 내용 정보(예를 들어, 뉴스 세그먼트의 경우 앵커 샷이 도입부에 나타남, 어떤 영화의 경우 세그먼트 전환시 fade-out/fade-in 효과를 사용하고 있음 등)를 알고 있는 극히 일부 경우를 제외하고는 거의 불가능하다. 장면의 경우는 배경이 동일한 경우 혹은 등장하는 사람이나 객체가 전체 화면에서 상당히 많은 부분을 차지하거나, 혹은 구조나 내용 정보를 알고 있을 때 자동으로 검출하는 것이 가능할 수도 있다. 이를 위한 장면 경계 검출 알고리즘이 시도되고 있다[9, 11]. 그러나 일반적인 영상에서 장면을 자동 검출하는 것은 샷 경계 검출에 비해 매우 많은 오류를 내포하게 된다.

그림 10은 그림3의 샷들을 장면 경계 검출 알고리즘을 사용하여 비슷한 내용의 샷들로 묶은 결과이다. 그림에서 알 수 있듯이 알고리즘은 29개의 샷을 13개의 장면으로 묶고 있으나, 이 결과는 그림 3과는 많이 다름을 알 수 있다. 예를 들어 그림 3의 장면 2는 그림 10에서 세 개의 장면 (장면 2-4)으로 묶여짐을 알 수 있다. 또 그림 3의 장면 3은 그림 10의 여섯 개 장면(장면 5-10)으로 묶여짐을 알 수 있다. 그림 3의 장면 3은 임의의 휴대전화를 사용하는 13개 샷으로 구성되어 있으므로, 신호를 분석하여 이들을 같은 부류로 묶기는 매우 힘듬을 알 수 있다. 반면에 그림 3의 장면 4는 그림 10에서 장면 11-13으로 묶여지는데, 이는 구성되는 샷들이 대부분 비슷한 배경과 비슷한 건물색으로 나타나 신호를 분석하면 비슷한 결과를 얻을 수 있으므로 동일한 장면으로 자동 검출이 가능할 수 있다.

이렇게 장면 검출이 어느 정도 가능한 경우도 있고 그렇지 못한 경우도 발생하는데, 이는 영상의 신택스(영상 신호)와 시멘틱스(사람에게 인지되는 구분)의 차이에 의한 것이다. 즉, 장면 경계 분석 알고리즘의 결과는 현재의 기술로는 영상 신호의 차이, 즉 신택스만을 분석하여 구한다. 반면에 사람들은 영상의 내용 전개, 즉 시멘틱에 의하여 영상의 내용을 구분짓는다. 따라서 이와 같은 차이를 극복하려면, 수작업에 의한 장면 경계 검출 단계가 필요하며, 이를 편하게 하기 위해서는 이런 수작업을 도와주는 도구가 필수적이다. 제안하는 영상 모델기에서는 자동 검출한 샷 경계를 이용해 장면과 세그먼트 경계를 사람이 직접 설정하도록 구현하였다.

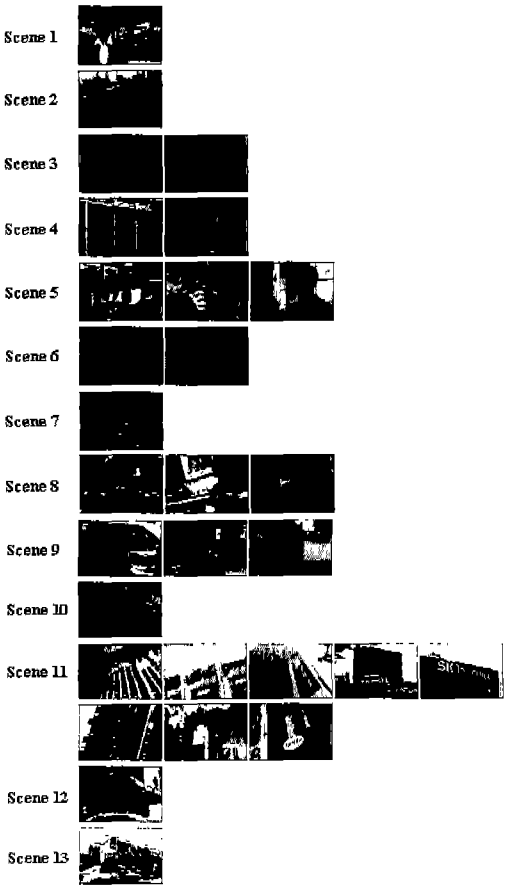

그림 10 자동 검출된 뉴스 세그먼트의 예

검출된 샷으로부터 장면을 정의하려면 먼저 장면의 시작 샷과 끝나는 샷을 설정하고, 설정된 구간내의 샷들로부터 그 장면을 대표하는 한 개의 대표 샷(representative shot)을 설정한다. 장면의 대표화면은 대표 샷의 대표화면을 사용한다. 정의된 장면을 이용해 세그먼트를

정의하는 것도 마찬가지로 시작과 끝나는 장면, 그리고 설정된 구간내의 장면들로부터 한 개의 대표 장면(representative scene)을 설정한다.

그림 11은 구현한 영상 모델기의 사용자 인터페이스이다. 이 인터페이스는 대표화면부와 모델연산부로 나눌 수 있다. 대표화면부는 각각 10개의 대표화면을 갖는 3개의 대표화면 리스트로 구성된다. 세 리스트는 맨 아래에서부터 각각 샷, 장면, 세그먼트 리스트를 나타낸다. 대표화면들은 각각 해당 모델링 단위에 속하는 인스턴스들의 대표화면이다. 임의의 대표화면 위에 커서를 이동하면, 그 모델링 인스턴스에 속하는 하위 단계 모델링 인스턴스가 있을 경우, 그 인스턴스들의 대표화면들이 디스플레이된다. 모델링 중인 그림 11의 2번째 세그먼트(그림 3의 뉴스 세그먼트)는 현재 3개의 장면으로 모델링되어 있고, 그들중 3번째 장면은 23개 샷으로 구성되어 있다. 3번째 장면의 대표 샷을 23개 샷중 1번째 샷으로 정의하면, 이 샷의 대표화면이 장면의 대표화면으로 디스플레이된다. 또 2번째 세그먼트의 대표 장면을 1번째 장면으로 정의하면, 이 장면의 대표화면이 세그먼트의 대표화면으로 디스플레이된다.

설명한 대표화면부는 계층 브라우저(hirerarchical browser)와 유사하다. 실제로 영상 모델기는 계층 브라우저로도 사용할 수 있다. 그러나 영상 모델기가 계층 브라우저와 근본적으로 다른 점은 모델연산부에서 제공하는 논리적 구조 정의 기능 때문이다. 모델연산부는 장면과 세그먼트 경계를 설정할 수 있는 여섯 개 모델링 연산이 아이콘 형태로 나타난다. 제공하는 모델링 연산에는 새로운 모델링 인스턴스의 위치를 지정하는 연산, 새로운 모델링 인스턴스의 구간을 정의하기 위하여 하위 레벨의 모델링 인스턴스의 시작과 끝을 설정하는 연산, 두 개의 인접한 모델링 인스턴스를 합치는 연산, 대표 샷 혹은 장면을 설정하는 연산이 있다.

영상 모델기를 이용해 영상의 내용을 계층구조화하는 과정은 다음과 같다. 처음에 대상 영상을 오픈하면 검출된 모든 샷이 샷 리스트에 나타난다. 초기에는 샷 리스트의 모든 샷을 하나의 장면에 모두 포함하고, 이 장면의 대표 샷은 묵시적으로 첫 번째 샷으로 정의한다. 따라서 장면 리스트에는 샷 리스트의 첫 번째 샷의 대표화면만 나타난다. 마찬가지로 초기의 세그먼트는 초기값으로 정의된 하나의 장면만을 갖는다. 따라서 세그먼트 리스트에도 샷 리스트의 첫 번째 샷의 대표화면만 나타난다. 이후 작업자는 모델링 연산을 이용해 원하는 구간과 그 구간을 대표하는 모델링 인스턴스를 설정하여 자유롭게 계층 구조를 정의할 수 있다.

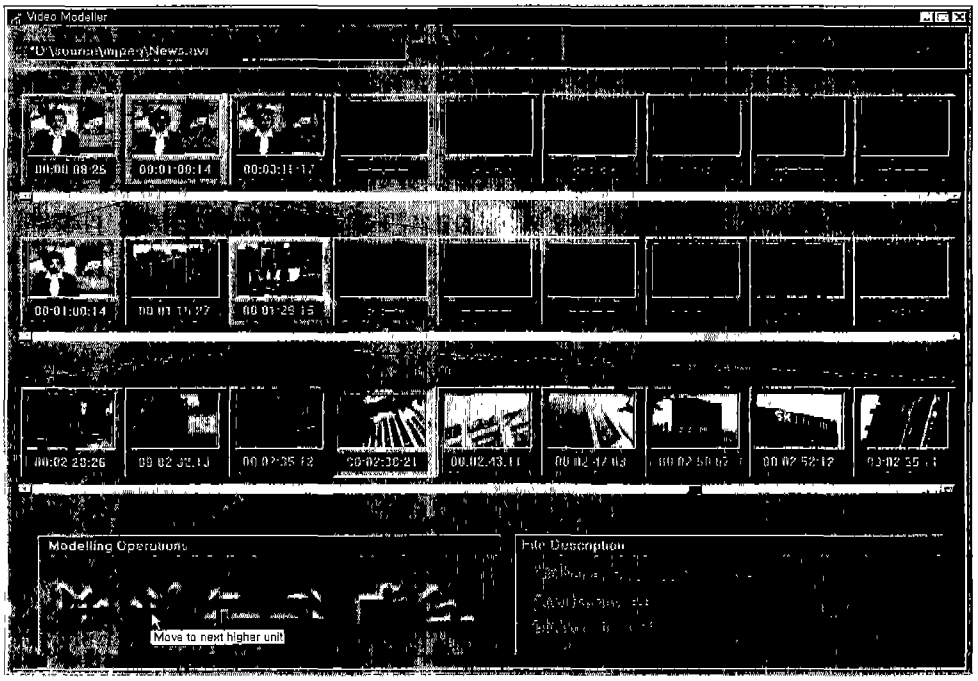

그림 11 영상 모델기의 사용자 인터페이스

그림 11의 장면 리스트에서 새로운 장면을 정의하는 과정을 살펴보자. 장면 리스트의 세 번째 장면을 선택한 후 모델연산부의 다섯 번째 연산을 선택하면, 장면 리스트의 네 번째 자리에 새로운 장면이 삽입될 공간이 생성된다. 샷 리스트의 14번째 샷을 선택한 후(그림의 샷 리스트에서 보면 4번째 샷) 모델연산부의 두 번째 연산을 선택하면, 이 샷을 시작으로 하여 마지막 23번째 샷까지를 구간으로하는 새로운 장면이 설정된다. 또 설정된 장면의 첫 번째 샷이 새로운 장면의 대표 샷이되어, 장면 리스트의 빈 공간에는 이 샷의 대표화면이 디스플레이된다.

그림 12는 영상 모델기의 구조를 나타낸 것이다. 구조관리자는 M-JPEG 복원기를 이용해 필요한 대표화면을 생성하고 구간 재생 기능을 제공하면서, 영상 모델링에 필요한 화면을 제어하는 기능을 갖는다. 영상 모델링 기능과 계층 브라우징 기능은 각각 모델링 연산부와 계층 브라우저에서 제공된다. 모델링이 완료되면 트리 구조의 구조 정보가 생성되며, 이 정보는 클립 생성기의 입력으로 사용된다.

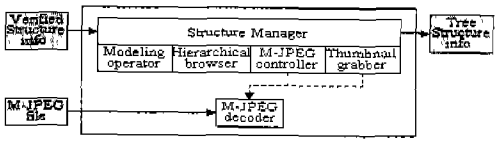

그림 12 영상 모델기의 구조

3.4 클립 생성기

클립 생성기는 영상 모델기로 정의한 논리적 구조를 이용해 원하는 모델링 인스턴스에 대한 압축 영상 클립과 대표화면을 파일로 생성한다. 생성하는 압축 영상 포맷은 MPEG1과 H.263/G.723 혹은 둘 모두 가능하며, 대표화면의 포맷은 JPEG이다. M-JPEG 입력 영상에서 소프트웨어만으로 MPEG1이나 H.263/G.723 영상을 생성하려면, M-JPEG-to-MPEG1 및 M-JPEG-to-H.263/G.723 포맷변환기(transcoder)가 필요하다. 이 소프트웨어를 개발하는데는 매우 많은 노력이 필요할 뿐만 아니라, 좋은 컴퓨팅 파워를 사용하더라도 소프트웨어 포맷 변환에는 많은 시간이 소요된다.

클립 생성기에서는 클립으로 생성할 구간의 영상을 M-JPEG 복원 보드로 복원하여 아날로그 신호로 만든 다음, 이 신호를 MPEG1 압축 보드에 연결하여 하드웨어 압축을 한다. H.263/G.723은 소프트웨어만으로도 실시간 압축이 가능하다. 따라서 복원된 아날로그 신호를 AV 캡춰 보드(capture board)에 연결하여 H.263 및 G.723 소프트웨어 압축기를 이용해 압축한다. 관련된 하드웨어 보드들은 보드 제어 모듈을 통해 제어한다. 그림 13은 클립 생성기의 구조를 나타낸 것이다.

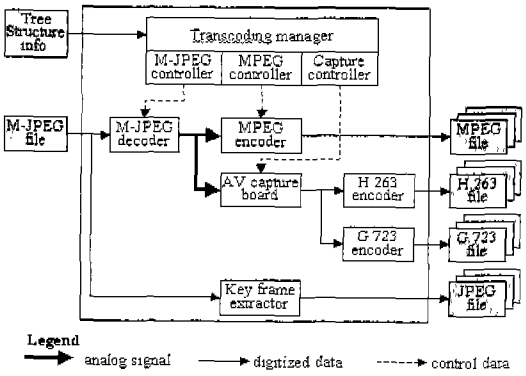

그림 13 클립 생성기의 구조

그림 14는 클립 생성기의 사용자 인터페이스를 보여준다. 이 인터페이스는 구조 제어부, 영상 디스플레이부, 파일 디스플레이부, 압축 제어부로 나눌 수 있다. 인터페이스 좌측 상단의 구조 제어부는 영상 모델기에서 정의한 논리적 구조를 바탕으로 클립으로 생성하려는 모델링 인스턴스들의 리스트를 작성한다. 작성시 대표화면의 JPEG 이미지 생성 여부를 선택한다. Add 버튼을 누르면 정의된 논리적 계층 구조를 보여주고, 그들 중 원하는 모델링 인스턴스를 선택할 수 있는 화면이 나타난다. 리스트 작성을 완료하면 이 화면은 사라지고, 작성된 리스트가 그림처럼 구조 제어부에 나타난다. 좌측 하단의 영상 디스플레이부는 작성한 리스트의 모델링 인스턴스들의 내용확인을 위한 영상 재생 기능을 제공한다. 또한 클립 생성 중에는 현재 복원되고 있는 M-JPEG 영상의 내용을 자동으로 보여준다. 우측 상단의 파일 디스플레이부는 클립 생성시 생성되는 클립의 파일명을 디스플레이한다. 우측 하단의 압축 제어부는 압축에 필요한 파라메타 설정 기능을 제공한다. 압축시 관련된 하드웨어 오류가 발생할 수 있기 때문에 초기화/시작/중단/지우기와 같은 제어를 제공한다.

클립 생성 과정은 먼저 구조 제어부를 이용해 원하는 모델링 인스턴스들을 선택하여 리스트를 만들고, 압축에 필요한 파라미터를 설정한 후, 압축 제어부의 시작 버튼을 선택하면 작성한 리스트의 모델링 인스턴스에 해당하는 영상 클립과 대표화면들이 차례대로 자동적으로 압축되어 각각 파일로 저장된다.

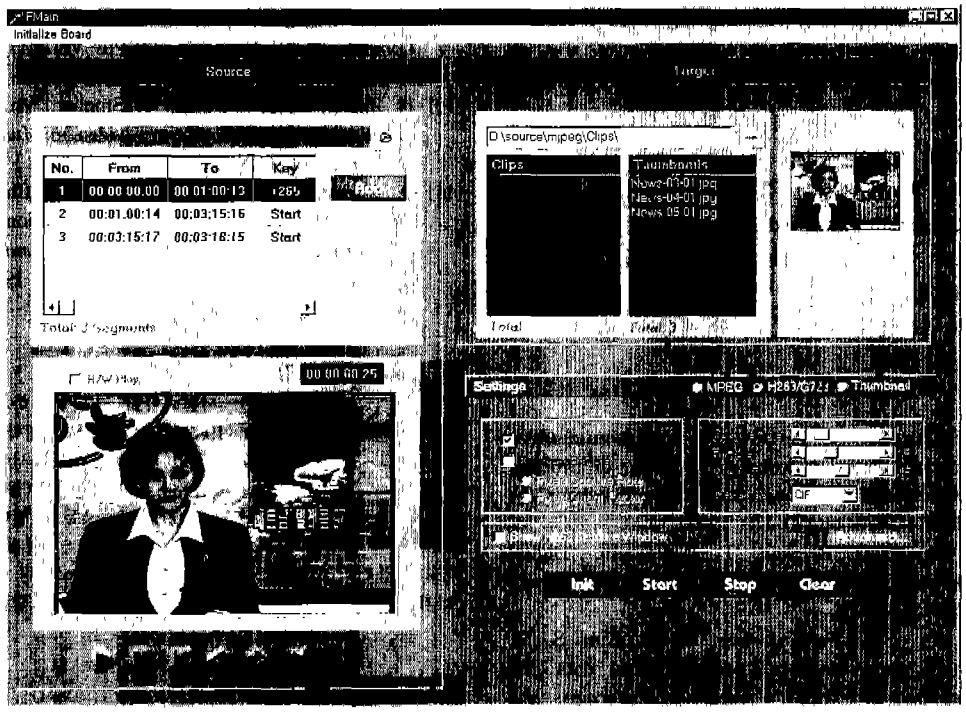

그림 14 클립 생성기의 사용자 인터페이스

4. 결 론

여러 개의 뉴스 아이템으로 구성되는 뉴스 프로그램과 같은 영상 컨텐트의 부분적 내용들을 각각 별도의 압축 영상 클립으로 생성하는 작업은 수작업으로 할 경우 매우 많은 시간이 소요되는 작업이다. 뿐만 아니라 생성된 압축 영상에 오류 구간이 포함되어 부자연스런 재생이 불가피하다.

본고에서는 샷 경계 검출 알고리즘을 이용한 영상 클립 생성 시스템인 V2Web Studio을 구현하였다. V2Web Studio는 영상 클립 생성 과정을 1) 영상 신호를 분석하여 샷 경계를 자동 검출하는 단계, 2) 검출된 결과에 포함될 수 있는 오류를 수작업으로 제거하는 단계, 3) 물리적인 샷 경계를 논리적인 계층구조로 모델링하는 단계, 4) 계층구조로 모델링된 각 모델링 인스턴스를 다양한 표준 압축 포맷으로 생성하는 단계로 구분하였다. 각 단계에 해당하는 작업은 샷 검출기, 샷 검증기, 영상 모델기, 클립 생성기의 독립적인 소프트웨어 도구로 구현하였다. 각각의 도구들은 편리한 GUI를 사용하여 배우고 쓰기 편하도록 구현하였다.

V2Web Studio를 사용하여 1시간 정도의 뉴스 프로그램에서 각 뉴스 아이템을 영상 클립으로 만들 경우, 작업자는 불과 20~30분 정도만 작업하면 시스템이 자동적으로 작업자가 정의한 클립들을 생성한다. 이는 기존의 7~10 시간 걸리던 작업 시간을 크게 단축할 뿐만 아니라, 생성되는 클립에도 오류 구간이 전혀 포함되지 않기 때문에 질 좋은 클립을 얻을 수 있다.

본 시스템은 인터넷용 스트리밍 비디오를 생성하는 응용에 적용되도록 확장할 수 있다. 현재는 저대역폭 스트리밍 영상으로 H.263/G.723을 사용하고 있지만, 클립 생성기 부분을 보완하여 ASF, RV, MOV 포맷을 제공하도록 확장할 예정이다. 동일한 내용의 클립을 여러 가지 비디오 포맷으로 제공하는 응용에서는 V2Web Studio와 같은 도구의 도움은 작업 시간을 크게 단축시킬 뿐만 아니라, 생성되는 클립들의 내용도 정확히 동일하게 만들 수 있다. 특히 후자의 특징은 차후에 MPEG-7과 같은 영상의 내용을 서술하는 메타데이타를 생성할 때, 복수로 생성된 파일에 대해 단 하나의 메타데이타를 유지하게 하는 잇점을 제공한다.

본 시스템은 샷 경계 검출 알고리즘을 적용한 응용 시스템이다. 현재 이 시스템은 자동적인 소프트웨어 분

석과 사람의 수작업이 병행되는 반자동 시스템이다. 사람의 수작업 부분을 자동화하려면 샷 및 장면 경계 검출 알고리즘을 더욱 개선해야 한다. 샷 경계 검출 결과가 정확하면 샷 검증 과정을 생략할 수 있다. 장면 경계 검출이 실제 적용할 수준이 된다면 모델링 과정이 단순화될 수 있다. 또 제한된 경우이지만 비디오의 구조나 내용 정보(domain knowledge)을 미리 알 수 있다면, 세그먼트의 자동 검출도 가능할 수 있다. 앞으로는 정교한 샷 및 장면 경계 검출 알고리즘 개발 및 비디오의 내용이나 구조 정보 사용에 주력할 예정이다.

참 고 문 헌

[1] J. S. Boreczky, L. A. Rowe, "Comparison of video boundary detection techniques," Proc. of *Storage and Retrieval for Image and Video Database IV*, SPIE Vol.2670, pp.170-179, 1996.

[2] A. Hampapur, R. Jain, T. Weymouth, "Digital video segmentation," Proc. Of *ACM Multimedia*, pp.357-564, 1994.

[3] H. Kim, S.-J. Park, J. Lee, W. M. Kim, S. M. Song, "Processing of partial video data for detection of wipes," Proc. of *Storage and Retrieval for Image and Video Databases VII*, SPIE Vol.3656, pp.280-289, San Jose, California, January 1999.

[4] H. Kim, J. Lee, J.-H. Yang, W. M. Kim, S. M. Song, "Visual rhythm and shot verification," To appear in *Multimedia Tools and Applications*, Kluwer Academic Publishers, Vol.15, No.3, November 2001.

[5] J. Meng, Y. Juan, S. Chang, "Scene change detection in a MPEG compressed video sequence," Proc. of *Digital Video Compression: Algorithms and Technologies*, SPIE Vol. 2419, pp.14-25, 1995.

[6] Y. Nakajima, K. Ujihara, A. Yoneyama, "Universal scene change detection on MPEG-coded data domain," Proc. of *Visual Communication and Image Processing*, SPIE Vol.3024, pp.992-1003, 1997.

[7] S. M. Song, T.-H. Kwon, W. M. Kim. H. Kim, B.-D. Rhee, "On detection of gradual scene changes for parsing of video data," Proc. of *Storage and Retrieval for Image and Video Database VI*, SPIE Vol.3312, pp.404-413, 1998.

[8] B.-L. Yeo, B. Liu, "Rapid scene analysis on compressed video," *IEEE Transactions on Circuits and Systems for Video Technology*, Vol.5, No.6, pp.533-544, Dec. 1995.

[9] M. M. Yeung, B.-L. Yeo, W. Wolf, B. Liu, "Video browsing using clustering and scene transitions on compressed sequences," Proc. of *Multimedia Computing and Networking*, SPIE Vol.2417, pp.399-413, 1995.

[10] H. Zhang, K. Kankanhalli, S. Smoliar, "Automatic partitioning of full-motion video," *Multimedia Systems*, Vol. 1, No. 1, pp.10-28, 1993.

[11] D. Zhong, H. Zhang, S.-F. Chang, "Clustering methods for video browsing and annotation," Proc. of *Storage and Retrieval for Image and Video Database VI*, SPIE Vol.2670, pp.239-246, 1996.

김 혁 만

1985년 서울대학교 컴퓨터공학과 졸업(학사). 1987년 서울대학교 대학원 컴퓨터공학과 졸업(석사). 1996년 서울대학교 대학원 컴퓨터공학과 졸업(공학박사). 1987년 ~ 1999년 한국통신 멀티미디어 연구소 연구원. 1999년 ~ 현재 국민대학교 컴퓨터학부 교수. 관심분야는 비디오 모델링 및 비디오 데이터베이스, 비디오 저작도구, 웹 방송 시스템 등임.

조 성 길

2000년 국민대학교 컴퓨터학부 졸업(학사). 2000년 ~ 현재 국민대학교 대학원 전산학과 재학중. 관심분야는 멀티미디어, 컴퓨터 비젼, 데이터베이스 등임.