준자동 비디오 모델링 기법

조성길 김혁만
국민대학교 전산과학과 멀티미디어 데이터베이스 연구실
{makeway,hmkim}@cs.kookmin.ac.kr

Semi-automatic video modeling

SeongGil Cho, Hyeokman Kim
Multimedia Database lab, Dept. of Computer Science, Kookmin University

요 약

디지털 비디오의 급속한 사용으로 인해 비디오를 좀더 효과적으로 구조화하여 브라우징할 필요성이 대두되고 있다. 비디오를 효과적으로 브라우징하기 위한 구조로 트리 형태의 계층구조가 주로 사용된다. 트리 형태로 비디오를 계층구조화 시키기 위한 여러 가지 방법이 제안되었지만 비디오의 컨텐트가 너무 다양하기 때문에 이를 완전하게 자동화 한다는 것은 거의 불가능하다. 본 논문에서는 내용기반 이미지 검색 기법을 이용한 클러스터링을 통해 3단계 계층구조를 자동적으로 생성한 후, 이 구조를 사용자가 수작업을 통해 원하는 형태로 전환시키는 기법을 제안한다. 또한 생성된 계층구조를 MPEG-7 메타데이타 표준으로 표현한다.

1. 서론

비디오의 내용을 재생, 빨리감기, 되감기와 같은 전통적인 VCR 제어 기능으로 파악하는 데는 많은 시간이 걸린다. 디지털 비디오의 경우에는 VCR 제어 기능 외에 특정 시간대로 임의 접근이 가능하나 이 역시 전체 내용을 파악하기 위해서는 거의 대부분을 재생시켜 보아야 한다. 따라서 비디오의 내용을 신속하게 파악하는 다양한 브라우징 기법이 개발되었다.

기존 비디오 브라우징 기법들은 크게 시간 연속 구조를 보여주는 스토리보드 기법[1], 트리 구조를 이용한 계층 브라우징 기법[2], 그리고 좀더 복잡한 관계의 표현이 가능한 그래프 브라우징 기법[3]으로 나눌 수 있다. 스토리보드 기법은 시간 순으로의 나열에 불과하므로 비디오의 평면적인 구조만 제공가능하며, 그래프 브라우징 기법은 복잡한 내용을 표현할 수 있지만 생성이 힘들고 숙달된 사용자가 아니면 내용을 직관적으로 파악하기 힘들다. 따라서 단순하면서도 직관적인 해석이 가능하고, 어느 정도 복잡한 내용의 표현이 가능하며, 사용자의 관심에 따라서 간단한 요약에서부터 특정 부분으로 점진적으로 자세히 브라우징해 들어갈 수 있는 계층 브라우징 기법이 가장 보편적으로 사용되는 추세이다.

계층 브라우징을 위해서는 비디오의 내용을 트리 형태의 계층구조로 표현해야 하며, 이를 위한 자동화된 알고리즘들이 제안되었다.[2,4] 하지만 비디오의 내용이 다양하므로 자동화된 기법만으로는 사용자가 원하는 형태의 계층 구조를 생성하기 힘들다. 본 논문에서는 자동화된 알고리즘과 사용자의 수작업을 같이하여, 수작업을 최소화하면서 쉽고 빠르게 비디오의 계층구조를 생성하는 기법을 제안한다. 또한 생성된 계층구조를 MPEG-7 메타데이타 형식[5]으로 표현한다.

2. 계층 비디오 모델링

비디오 모델링의 기본 단위는 세그먼트(segment)이다. 세그먼트는 연속된 프레임의 집합으로, 각 세그먼트는 몇 개의 부세그먼트(subsegment)로 구성될 수 있다. 각 세그먼트 및 그것의 부세그먼트들의 종속 관계를 표현하면 트리 형태가 되고, 특정 비디오에 대해 이런 트리 구조를 만드는 과정을 계층 비디오 모델링(hierarchical video modeling)이라 한다. 비디오는 편집될 때 이미 편집의 단위인 샷(shot)이 정의되므로, 세그먼트의 최소 단위로는 일반적으로 샷이 사용된다. 즉 일반적으로 계층 비디오 모델링으로 구성되는 트리의 단말노드는 샷을 의미한다.

그림 1은 15개의 샷으로 구성된 뉴스 비디오를 계층구조로 표현한 것이다. 그림에서 1부터 15까지의 노드로 표현된 세그먼트는 15개의 샷을 나타내며, 루트 세그먼트 21은 뉴스 비디오 전체를 의미한다. 즉, 루트 세그먼트 21은 1부터 15까지의 시간적으로 연속적인 부세그먼트들로 구성되어 있음을 알 수 있다. 이런 2단계 계층구조는 임의의 샷 경계 검출 알고리즘을 사용하여 자동적으로 검출한 샷들의 연속구조(sequential structure)를 2단계 계층구조로 전환하여 생성할 수 있다.

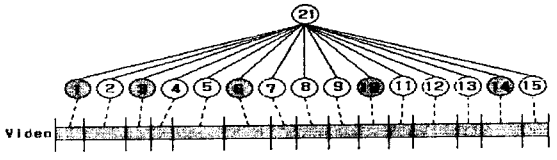

그림 1. 2단계 계층구조 (초기구조)

그림 1의 샷들중 1, 3, 6, 10, 14는 앵커샷이라 하자. 일반적으로 각각의 뉴스 아이템은 도입부에 앵커샷이 나온 후

그 뉴스 아이템에 관계된 샷들이 나오게 된다. 따라서 앵커 샷을 기준으로 샷들을 묶으면(clustering) 뉴스 아이템별로 세그먼트를 정의할 수 있다. 이를 위해 각 샷의 대표화면(key frame)을 추출한 후 이 대표화면들에 대해 임의의 내용기반 이미지 검색 알고리즘을 적용해 모든 앵커샷들을 찾아내고, 찾아낸 앵커 샷을 이용해 관련샷들을 묶으면 그림 2와 같은 3단계 계층구조를 자동적으로 생성할 수 있다. 그림에서 세그먼트 31부터 35는 이미지 검색 알고리즘에 의해 자동으로 정의된 세그먼트이다.

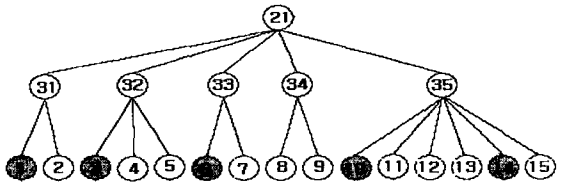

그림 2. 3단계 계층구조

그러나 그림 2에서 샷 6-9는 하나의 세그먼트로 정의되어야 하나, 실제로는 두개의 세그먼트(세그먼트 33과 34)로 정의되었다. 이는 사용한 이미지 검색 알고리즘이 샷 8의 대표화면을 앵커샷으로 잘못 인식하였기 때문에 발생한 것이다. 또한 샷 10-13과 샷 14-15는 두개의 세그먼트로 정의되어야 하나, 실제로는 하나의 세그먼트(세그먼트 35)로 정의되었다. 이는 사용한 이미지 검색 알고리즘이 샷 14의 대표화면을 앵커샷으로 인식하지 못하였기 때문에 발생한 것이다.

내용기반 이미지 검색 알고리즘의 성능이 많이 개선되었지만 미검출 및 오검출이 완벽히 배제된 알고리즘이 개발되기란 역시 거의 불가능하여, 그림 2와 같은 결과를 얻을 개연성은 항상 존재한다. 따라서 이와 같은 잘못된 계층 구조를 상용자가 수작업으로 수정하고, 경우에 따라서는 원하는 구조로 전환시킬 필요가 있다.

그림 3은 그림 2의 계층구조를 사용자가 자신이 원하는 형태로 수작업을 통해 전환시킨 예이다. 그림 2의 세그먼트 33과 34는 합하여 세그먼트 42를 만들었다. 세그먼트 35는 세그먼트 43과 44로 분할한 후, 이 둘을 묶는 새로운 세그먼트 45를 만들었다. 또 세그먼트 31과 32는 묶어서 새로운 세그먼트 41을 만들었다. 이 결과 그림 3과 같은 4단계 계층구조를 얻을 수 있다. 만일 세그먼트 41, 42, 45에 각각 정치, 스포츠, 문화라는 타이틀을 단다면 정치에 관련된 뉴스 아이템이 2개, 스포츠에 관해 1개, 그리고 문화에 관해 2개의 뉴스 아이템이 존재함을 알 수 있다.

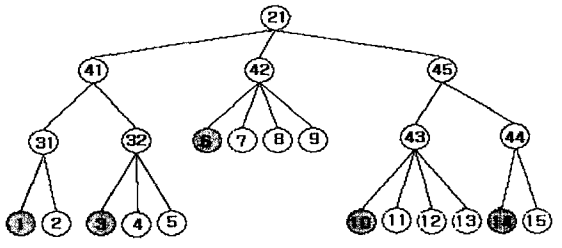

그림 3. 4단계 계층구조 (최종결과)

3. 준자동 비디오 모델링

설명한 계층 비디오 모델링 과정은 그림 4와 같은 단계로 이루어진다. 먼저 단계 1에서 입력 비디오에 임의의 샷 경계 검출 알고리즘 및 대표화면 선택 알고리즘을 적용하고, 단계 2에서 그 결과를 계층구조로 표현한다. 그 결과 그림 1과 같은 2단계 계층구조를 얻을 수 있다. 사용자가 자동 클러스터링을 원하면(단계 3), 단계 1에서 구한 대표화면에서 질의로 사용할 대표화면을 선택한다(단계 4). 예를 들어 뉴스 프로그램에서는 앵커샷의 대표화면을 질의로 선택한다. 이 대표화면을 입력으로 하여 임의의 내용기반 이미지 검색 알고리즘으로 이와 비슷한 대표화면을 갖는 샷들을 검색하여(단계 5), 검색된 유사한 샷들을 시간순으로 정렬시킨다(단계 6). 단계 7과 8에서 정렬된 샷 리스트에서 두 샷 사이에 속한 모든 샷들을 하나의 세그먼트로 묶으며, 이 과정을 모든 검색된 샷들에 대해 반복한다. 이 결과로 그림 2와 같은 3단계 계층구조를 얻을 수 있다. 단계 9에서는 수작업으로 계층구조를 변경하는 모델링 연산자를 임의의 순서로 적용하면서 3단계 계층구조를 사용자가 원하는 형태로 수동으로 전환한다. 만일 단계 3에서 사용자가 자동 클러스터링을 원하지 않으면, 단계 9에서 모델링 연산자만을 이용해 수동으로 클러스터링을 수행한다.

```
1. Detect shot boundaries and select key frames
2. Group all shots into a single segment
3. If (automatic clustering) {
4.     Select a query frame from the key frames of
       the detected shots.
5.     Search for shots having similar key frame with
       the query frame.
6.     Arrange the search results in temporal order.
7.     For each shot in the ordered list {
8.         Group the shot and all intermediate shots
           between the shot and the next one in the
           ordered list into a single segment.
       }
   }
9. Transform the current structure into a desirable one
   with a proper combination of modeling operators.
```

그림 4. 준자동 비디오 모델링 알고리즘

그림 4의 알고리즘은 샷 경계 검출, 대표화면 선택, 그리고 내용기반 이미지 검색 알고리즘을 필요하며, 이에 관한 기존의 어떤 알고리즘을 사용하여도 구현이 가능하다. 그림 4의 알고리즘은 위의 세가지 알고리즘을 결합하여 계층 비디오 모델링 과정을 체계적으로 최대한 자동화하였다. 단계 4에서 사용자의 선택이 한번 필요한 것 외에는 그림 8까지는 자동적으로 동작할 수 있다. 그러나 단계 9에서는 어떤 모델링 연산자를 계층구조의 어느 부분에 적용할 것인지 사용자가 결정하여야 한다.

4. MPEG-7 메타데이타로의 변환

생성된 계층구조에 속한 각각의 세그먼트에는 타이틀, 주석, 대표화면의 저장위치, 시작시각 및 재생시간과 같은 다양한 메타데이타가 부여될 수 있다. 이러한 메타데이타는 멀티미디어 내용 기술 표준인 MPEG-7으로 표현할 수 있다. MPEG-7은 XML Schema를 기반으로 되어 있으므로 웹을 포함한 모든 시스템 혹은 프로그램간에 데이터 이식이 가능하다. 여기서는 MPEG-7의 Summarization DS (Description Schene)를

이용해 계층구조를 표현한다. 만일 내용 기술에 대한 정밀한 메타데이타가 요구되면 Segment DS로 표현하는 것이 바람직하다.

그림 5는 MPEG-7 Summarization DS의 데이터 타입간의 집단화(aggregation) 관계를 그림으로 표현한 것이다. 그림에서 사각형은 각각의 데이터 타입을 나타내고, 링크는 데이터 타입간의 part-of 관계를 나타내며, 링크의 라벨은 XML 태깅 시 사용하는 엘리먼트(element)의 이름을 나타낸다. 그림에서 알 수 있듯이 SummarySegmentGropType은 Name이라는 이름의 TextualType, SummarySegment라는 이름의 SummarySegmentType, 그리고 SummarySegmentGroup이라는 이름으로 자기 자신의 타입을 구성 요소로 가질 수 있다.

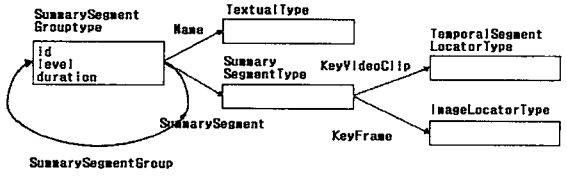

그림 5. MPEG-7 Summarization DS

그림 5의 Summarization DS를 이용하여 먼저 계층구조의 단말 세그먼트, 즉 샷의 메타데이타는 그림 6과 같이 표현할 수 있다. 그림 6에서는 그림 3의 세그먼트 6의 메타데이타를 표현하고 있다. 이 메타데이타는 야구경기 샷이 23분 45초에서 10초간 지속되고, 대표화면은 23분 47초에서 뽑아내었으며, 6.jpg라는 이름으로 디스크의 /news1/kframes에 저장되어 있음을 나타낸다. 그외에 주석, 대표오디오 등 다양한 정보를 표현할 수 있다.

```
<SummarySegmentGroup id="seg_6" level="2"
                     duration="PT10S000N">
 <Name>야구경기</Name>
 <SummarySegment>
  <KeyVideoClip>
    <MediaTime>
      <MediaTimePoint>T00:23:45</MediaTimePoint>
      <MediaDuration>PT10S000N</MediaDuration>
    </MediaTime>
  </KeyVideoClip>
  <KeyFrame>
    <MediaUri>/news1/kframes/6.jpg</MediaUri>
    <MediaTimePoint>T00:23:47</MediaTimePoint>
  </KeyFrame>
 </SummarySegment>
</SummarySegmentGroup>
```

그림 6. MPEG-7을 이용한 샷의 표현

그림 6과 같은 형태로 샷을 표현하면, 이를 이용해 상위 레벨의 세그먼트를 표현할 수 있다. 그림 7은 그림 3의 세그먼트 42를 MPEG-7으로 표현한 것이다. 그림 7은 스포츠 경기를 담고있는 세그먼트 42의 지속시간은 1분35초이고, 4개의 부세그먼트로 이루어져 있음을 나타낸다.

```
<SummarySegmentGroup id="seg_42" level="1"
                     duration="PT01M35S000N">
```

그림 7. MPEG-7을 이용한 계층구조의 표현

이렇게 각각의 세그먼트에 대한 메타데이타를 서술하므로써 연속된 샷들을 하나의 논리적 단위로 묶어 의미를 부여할 수 있다. 예를 들어 그림 7의 1번은 야구경기, 2와 3번은 골프경기, 4번은 축구경기를 하는 샷이라고 할 때 이들을 묶어 스포츠 중계라고 표현할 수 있다.

5. 결론과 향후 과제

본 논문에서는 수동 비디오 모델링과 자동 비디오 모델링의 한계를 극복하기 위해 준자동 비디오 모델링 알고리즘을 제안하고, 이 알고리즘에 의해 생성되는 계층구조를 MPEG-7 표준으로 표현하였다. 이 알고리즘을 통해 사용자가 원하는 임의의 어떤 형태의 계층구조도 쉽고 편하게 생성이 가능하다. 따라서 비디오의 내용에 따라 빠르게 브라우징 할 수 있도록 하였다. 앞으로는 수작업을 최대한 줄이며 편하게 작업할 수 있는 모델링 연산자를 개발하고, 제안한 알고리즘을 사용하는 자동화된 도구를 개발할 예정이다.

참고 문헌

1. F. Arman, et al., " Content-based browsing of video sequences" , Proc. ACM Multimedia, pp.97-103, Oct. 1994.
2. D. Zhong, et al., " Clustering methods for video browsing and annotation" , Proc. Storage and retrieval for image and video databases IV, pp.239-246, Jan. 1996.
3. B.-L. Yeo, M. M. Yeung, " Classification, simplification and dynamic visualization of scene transition graphs for video browsing" , Proc. Storage and retrieval for image and video databases VI, pp.60-70, Jan. 1998.
4. H. Zhang, et al., " Video parsing and browsing using compressed data" , Multimedia tools and applications, Vol.1, No.1, pp.89-111, March 1995.
5. ISO/IEC FDIS 15938-5, " Multimedia content description interface (MPEG-7) - Part 5: Multimedia description schemes (MDS), Oct. 2001.