영화 비디오를 위한 클러스터링 기반의 계층적 장면 구조 구축

(Clustering-based Hierarchical Scene Structure Construction for Movie Videos)

최 익 원 † 변 혜 란 ††

(Ickwon Choi) (Hyeran Byun)

요 약 최근 들어 멀티 미디어 정보의 사용이 급격히 증가하면서, 여러 미디어 형태 중 비디오가 많은 각광을 받으며, 다른 타입의 모든 미디어 정보를 하나의 자료 흐름으로 묶고 있다. 디지털 비디오의 실용 가능성은 크게 증대되고 있으나 비디오의 방대한 길이와 비구조적 형식 때문에 효과적인 비디오의 접근은 어려운 실정이다. 따라서 최근에 개발되는 영상과 비디오 정보 관리 시스템은 본 논문에서 제안하는 사용자의 최소 상호 작용과 비디오 구조의 명확한 정의를 필요로 한다.

본 논문에서는 사용자가 쉽게 비디오 내용을 요약한 형태로 보고, 임의로 접근 할 수 있도록 클러스터링 기반 비디오 계층 구조 구축 시스템을 제시한다. 제안된 시스템은 크게 샷 경계면 검출과 계층 구조 구축 단계로 이루어진다. 샷 경계면 검출 단계에서는 복수 특징들을 추출하고, 이웃한 프레임 쌍들에 대한 상호관계를 고려한 시간 적응적 필터링 기법을 이용하여 오판될 수 있는 왜곡 성분을 제거함으로써 성능을 향상시켰다. 처리된 복수 특징들은 임계치를 필요로 하지 않는 k-means 클러스터링의 입력으로 사용되어 샷 경계면을 검출한다. 결과인 순차적인 샷 리스트는 시간 지역성과 장면 구조를 효과적으로 모델링하는 특성을 가진 지능적 비감독 클러스터링 기법에 의해 계층 구조로 표현된다. 실험은 정적 영화 비디오와 동적 영화 비디오를 대상으로 수행하였으며, 샷 경계면 검출에서는 평균적으로 95%의 정확성을 보였으며 장면 경계면 검출을 하는 비디오 계층 구조 구축에서도 어느 정도 정확한 장면 경계면 검출 결과를 보였다.

Abstract Recent years, the use of multimedia information is rapidly increasing, and the video media is the most rising one than any others, and this field integrates all the media into a single data stream. Though the availability of digital video is raised largely, it is very difficult for users to make the effective video access, due to its length and unstructured video format. Thus, the minimal interaction of users and the explicit definition of video structure is a key requirement in the lately-developing image and video management systems. This paper defines the terms and hierarchical video structure, and presents the system which construct the clustering-based video hierarchy, which facilitate users by browsing the summary and do a random access to the video content. Instead of using a single feature and domain-specific thresholds, we use multiple features that have complementary relationship for each other and clustering-based methods that use normalization so as to interact with users minimally. The stage of shot boundary detection extracts multiple features, performs the adaptive filtering process for each features to enhance the performance by eliminating the false factors, and does k-means clustering with two classes. The shot list of a result after the proposed procedure is represented as the video hierarchy by the intelligent unsupervised clustering technique.

We experimented the static and the dynamic movie videos that represent characteristics of various video types. In the result of shot boundary detection, we had almost more than 95% good performance, and had also good result in the video hierarchy.

† 학생회원 : 연세대학교 컴퓨터과학과
torpedo@csai.yonsei.ac.kr
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
hrbyun@csai.yonsei.ac.kr

논문접수 : 1999년 10월 15일
심사완료 : 2000년 3월 28일

1. 서 론

최근 들어 멀티 미디어 정보의 사용이 급격히 증가하고 있는데, 텍스트, 영상, 그래픽, 음성, 비디오 등의 여러 미디어 형태 중에서 비디오가 많은 각광을 받고 있으며, 다른 타입의 모든 미디어 정보를 하나의 자료 흐름으로 묶고 있다. 저장 장치의 가격 하락과 고속의 전송율, 그리고 향상된 압축 기술에 의해 디지털 비디오의 실용 가능성은 크게 증대되고 있으나, 비디오의 방대한 길이와 비구조적 형식으로 인해 효과적인 비디오의 접근은 어려운 실정이다.

비디오에 대한 기존의 내용 기반 검색 시스템에서는 제일 기초가 되는 비디오 구조의 정의가 명확하지 않은 관계로, 이에 수반되는 하위 작업인 비디오 구조에 대한 분석 및 파싱과 색인 작업이 효과적으로 처리될 수 없었다. 검색을 위해 사용자에게 제공되는 질의 방법이나 도구들은 상위 작업인 비디오 파싱과 색인 부분에 파급적으로 영향을 주게 되고, 이것은 또한 원초적인 비디오 구조에 변화를 주게 된다. 본 논문에서의 비디오 구조는 비디오를 구성하는 단위들의 순차적 또는 계층적인 형태뿐만 아니라 이들을 표현하거나 묘사하는 방법을 모두 포함한다.

내용 기반 검색의 결과를 브라우징하는 비디오 표현과 추상화, 의미 차원의 시각적 요약(목차 구성), 비순차적이고 비선형적 임의 접근, 비디오 장면(이야기 단위) 검출 등에서 잘 정의된 비디오 구조의 특성이 많이 반영됨으로 이에 대한 연구가 매우 중요하다.

브라우징이나 검색의 관점에서 볼 때 비디오는 책과 유사하며 책의 전체 내용은 책의 의미 있는 구조를 나타내는 잘 설계된 목차에 의해 추상화 될 수 있다. 기존의 비디오에서는 브라우징이나 검색 등의 작업이 선형적으로 처리되었고, 사용자가 관심을 두고 있는 특정 내용에 대한 접근은 시간을 많이 소비하는 순차적인 빠른 감기와 빠른 되감기에 의해 처리됨으로써 사용자의 접근이 비효율적이었다. 통신, 멀티미디어, 컴퓨터 기술의 진보는 사용자가 영상과 비디오 정보에 대해서 전문적으로 자주 접할 수 있게 하였으며, 방대한 영상과 비디오 자료를 관리하는 시스템을 출현하게 했다[1][2]. 따라서 잘 정의된 비디오 구조와 이에 대해 간결하며 중복되지 않은 내용 분석, 표현, 그리고 적절한 묘사의 중요성이 대두되고 있다.

본 논문에서는 위의 사항을 고려하여 사용자에게 의미 차원의 장면 구조를 효율적으로 추상화할 수 있는 클러스터링 기반 계층적 비디오 구조 구축 시스템을 제시한다. 제안된 시스템은 샷 경계면 분할에 있어서 복잡한 내용 표현으로써 복수 특징을 추출하고, 이 특징들은 향상된 성능을 위해 일련의 신호로 모델링 되어 시간 적응적 필터링을 하게 된다. 비디오에 대한 시간적 분할의 마지막 단계로 비감독(unsupervised) k-means 클러스터링 기법을 이용한다. 계층적 비디오 구성에서는 일정한 윈도우 크기로 인해 발생하는 시간적 불연속성 문제를 시간 적응적인 방법으로 모델링함으로써 해결하고, 비디오에 대한 포괄적인 시공간 정보와 시간 지역성 정보를 취함으로써 더욱 정확한 장면(이야기 단위) 구조를 구축한다.

본 논문의 나머지 부분은 다음과 같이 구성된다. 2장에서는 연구 배경으로서 비디오 계층 구조와 용어를 정의하고, 계층별 관점에서 비디오의 구조를 구성하는 기존의 연구를 관찰해보고 평가한다. 3장에서는 본 논문에서 제안하는 비디오 계층 구조 시스템을 제시한다. 4, 5, 6장에서는 시스템의 각 단계에 대해 자세히 설명하며, 7장에서는 실험결과 및 실험결과 분석을 하고, 마지막으로 8장에서는 결론 및 향후 연구방향에 대해 논의한다.

2. 연구 배경

2.1 비디오 계층 구조와 용어 정의

먼저 의미 차원의 비디오 구조를 정의하기 위해 장면이라는 용어가 사용되는데, 이는 의미적으로 관련이 있고 시간적으로도 이웃되어 있는 샷들의 모임으로 정의한다.

초기의 비디오 파싱에 대한 연구자들은 샷 경계면과 장면 경계면의 의미를 동일한 의미로 잘못 사용하였다[3][4]. 본 논문에서 샷 경계면 검출(Shot boundary detection)은 물리적인 샷 경계면 검출로, 장면 경계면 검출(Scene boundary detection)은 의미적인 장면 경계면 검출로 사용한다.

비디오는 그림 1과 같이 5가지 레벨(비디오, 장면, 그룹, 샷, 대표 프레임)로 이루어진 계층적인 구조로 표현될 수 있다[3][4].

그림 1의 비디오 파싱에서 사용된 각 층의 용어를 정의하면 다음과 같다.

- **비디오 샷** : 하나의 카메라에 의해 기록된 연속적인 일련의 프레임들, 이것은 비디오의 물리적인 기본 단위로서 샷 경계면에 의해 경계되어 진다.
- **대표 프레임** : 한 샷에서 가장 내용을 잘 표현하는 하나의 프레임을 말한다. 샷의 내용 복잡도에 따라서 하나 이상의 대표 프레임이 추출될 수 있다.
- **비디오 그룹** : 물리적 샷과 의미적 장면 사이의 중간

다리 역할을 하며, 그룹의 예로서는 시간적으로 이웃한 샷들의 모임 또는 시각적으로 유사한 샷들의 모임이다.

- **비디오 장면** : 상위 레벨의 개념과 이야기 단위를 다루는 것으로, 의미적으로 관련되어 지고 시간적으로 이웃한 샷들의 모임으로 정의된다. 샷은 비디오의 기본 처리 단위인 반면, 장면은 사용자에게 비디오의 의미 있는 내용을 전달한다. 장면 경계면 검출은 주관적 성격 때문에 샷 경계면 검출에 비해 더 어렵다[5].

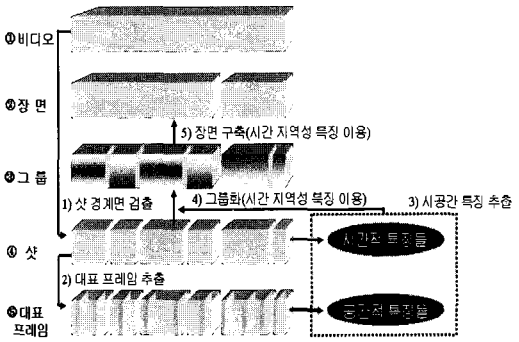


그림 1 비디오 계층 구조

장면과 그룹과의 차이는 장면은 시간의 연속적 배열이 중요하고, 그룹은 단지 장면을 구성하기 위해 거쳐가야 할 하나의 단계로써 시간의 연속적 배열은 상관 없지만 한 장면에 속한 그룹들의 샷들은 영화를 연속적 순서를 만들 수 있도록 모든 번호가 포함되어야 한다.

2.2 기존 연구

비디오 구조에 대한 분석 및 파싱은 주로 브라우징 관점에서 여러 계층 기반으로 연구되어 왔다. 비디오 브라우징은 비디오 구조에 의해 표현 및 추상화가 되어짐으로, 기존 연구도 비디오 구조 관점에서 고찰되어야 할 것이다. 이 절에서는 최근까지 제안되었던 일반적인 방법들에 대해 간략히 관찰하며 평가한다.

- **순차 기반 구조 방법**

이 접근 방법에서는 원시 비디오 자료가 일련의 샷 리스트로 분할되며, 분할 된 샷에서 대표 프레임이 추출된다. 파싱된 비디오 내용은 일련의 대표 프레임들로 추상화되는데, 사용자는 순차적인 대표 프레임들의 브라우징으로 비디오에 접근할 수 있다.

(1) 샷 경계면 검출 기법:

일반적으로 자동 샷 경계면 검출 기법은 다음과 같이 5가지 범주로 분류될 수 있다. 화소 기반 접근 방법은 화소 간의 차이 값을 이용하는 것으로써 잡음에 민감하다[6]. 이 문제를 해결하기 위해 샷 경계면 측정으로써 화소 값의 통계(평균, 분산)를 이용하는 방법이 제시되었으며[7], 샷 경계면을 빠른 속도로 검출하기 위해 압축 도메인에서 DCT 계수를 이용하는 방법도 제안되었다. 변환 기반 기법은 MPEG 스트림에 이미 포함되어 있는 움직임 벡터를 이용하는 방법[8]이 있지만, 압축 효과를 높이기 위해 변형되어 압축됨으로 내재된 움직임 벡터를 이용하는 것은 비효율적이다. 다른 각도에서 본 특징 기반 기법으로 각 프레임에서 에지 특징이 추출된 후에 에지 차이를 비교함으로써 경계면을 검출하는 방법도 제안되었다[9][10].

(2) 대표 프레임 추출 기법:

샷 경계면이 검출된 후에는 해당 샷에 대한 내용을 가장 잘 표현 할 수 있는 대표 프레임을 추출하는데, 크게 3가지 방법으로 분류된다. 첫 번째 경계면 기반 방법은 가장 간단한 방법으로 각 샷의 첫번째 또는 마지막 프레임을 선택한다. 두 번째 시각적 내용 기반 방법은 복수의 시각적 특징을 이용하는 것으로, 다시 1) 첫 번째 프레임과 다른 판단 기준들을 이용하여 하나 이상의 프레임을 추출하는 샷 기반 판단 기준(criteria), 2) 기준이 되는 프레임을 하나 선택하고 이에 대한 특징 값을 추출하여 처음부터 비교해가며 상당한 변화가 발생할 때마다 새로운 대표 프레임으로 선택하는 색상 특징 기반 판단 기준, 3) 카메라 효과를 가진 샷에서 대표 프레임을 선정하는 움직임 기반 판단 기준의 3가지 방법으로 분류된다[11]. 앞의 두 가지 방법은 상대적으로 빠르지만 샷의 시각적 내용을 효과적으로 선택하지 못한다. 세 번째 움직임 기반 방법은 비록 움직임 분석을 포함한 방법이지만 계산양이 많고, 또한 지역 극소값(local minima)을 이용함으로써 그 결과가 항상 정확한 것은 아니다. 최근에 계산이 간단하며 시각적 내용을 적응적으로 표현하는 클러스터링 기법이 많이 사용되고 있다[12].

최근에는 압축 기술의 발전에 따라 압축된 형태로 비디오가 보급되고 있어 압축 형식을 해제하여 처리하는 것은 많은 시간을 낭비하게 된다. 그러므로 MPEG의 DCT기반에서 DC 계수만을 추출해서 처리함으로써 시간 성능을 향상시키는 방법들이 존재한다[13][14][15].

이러한 순차 비디오 구조 접근 방법은 브라우징 응용에서 전체 대표 프레임의 수가 그리 많지 않을 때는 효과적이지만, 길이가 긴 비디오 클립에서는 간단한 대표 프레임의 순차적 디스플레이는 정확한 내용의 전달이 어렵기 때문에 거의 의미가 없다.

- **클러스터 기반 구조 방법**

상위 층의 비디오 계층 구조를 얻기 위해 관련이 있

는 샷들을 그룹들로 합병하고, 이 합병된 클러스터를 기반으로 브라우징 트리를 구성하여 비디오 구조를 만들 수 있다[16]. [6]은 전체 비디오 스트림을 다수의 비디오 세그먼트로 분할했는데, 각 세그먼트는 연속된 같은 수의 샷들로 이루어져 있다. 각 세그먼트는 다시 하위 세그먼트로 분할될 수 있다. 이것은 비디오 내용에 대한 계층 구조를 구성하며 브라우징에서 사용된다. 이 접근 방법은 브라우징 계층 구조를 구성하기 위해 오직 시간 요소만을 고려했고 시각적 내용은 고려하지 않았다. 반대로 [17]은 클러스터 기반 비디오 계층 구조를 제안했는데, 여기서 샷들은 시간 정보를 고려하지 않고 시각적 내용에 의해서만 클러스터링이 된다.

이 두 방법[6][17]의 단점은 비디오 구조가 의미 차원의 구조를 보여 주지 못한다는 것이다. 비록 이 그룹 기반 접근 방법이 샷과 프레임 기반의 비디오 구조보다는 더 나은 해결책을 제공하지만, 여전히 비디오의 의미 개념이나 이야기 형식을 전달하기에는 부족하다.

• 장면(이야기) 기반 구조 방법

사용자에게 편리한 접근을 제공하기 위해 의미 차원의 비디오 계층 구조 구성이 필요하다. 장면 계층의 비디오 구조를 가지는 기존의 접근 방법은 다음의 두 가지로 분류된다.

(1) 모델 기반 접근 방법 : 특정 응용 프로그램이나 도메인에 대한 모델이 사전에 미리 구성된다. 사전에 미리 구성된 이 모델은 장면 경계 특성을 기술하며, 이 모델에 기초하여 비구조적 스트림이 정형화된 비디오 구조로 추상화된다. 이 모델 기반 접근 방법은 뉴스 비디오나 TV 축구 프로그램 파싱에서 성공적으로 사용되었다[18]. 이 방법은 특정 응용프로그램 모델에 기반을 두고 있으므로 큰 정확도를 얻을 수 있으나, 파싱 처리를 하기 전에 개개의 응용 프로그램마다 사전 모델이 구성되어 있어야만 한다. 이러한 모델 기반 처리는 시간, 도메인 지식, 경험 등과 같은 요소들을 필요로 한다는 단점이 있다.

(2) 범용 기반 접근 방법 : 명확한 도메인을 필요로 하지 않는 일반화된 모델을 따르는 것으로, 비디오 스트림을 여러 샷들로 분할하고, 다음에 시각적으로 유사하고 시간적으로 이웃한 샷들을 클러스터로 구성하는 시간 제약 조건 클러스터링(time-constrained clustering)을 사용한다. 이 클러스터에 기반하여 장면 전이 그래프(scene transition graph)가 구성되며, 장면 구조를 구성하기 위해 장면들의 흐름(cutting edge)을 표시한다[19][20][21].

순차 기반과 클러스터 기반의 비디오 구조 접근 방법은 하위 레벨의 입장에서 계산 처리를 함으로 시간적이고 의미적인 불연속이 발생할 수 있다. 즉 장면은 사건과 화제가 발생하는 장소의 위치를 중심으로 하는 이야기 단위로 표현됨으로 시간의 연속성이 중요한 요소로 작용한다.

3. 제안된 비디오 계층 구조 구축 시스템

비디오 계층 구조를 생성하기 위한 시스템 처리 구성은 그림 2와 같고, 비디오 구조의 표현과 분석 과정의 입장에서 본 시스템의 프레임워크는 그림 3과 같다.

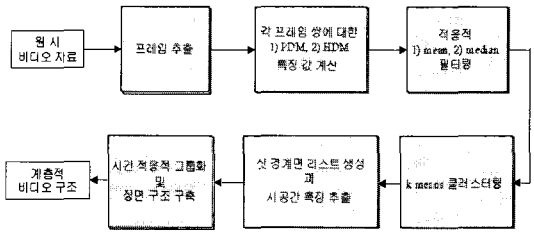


그림 2 시스템 처리 구성

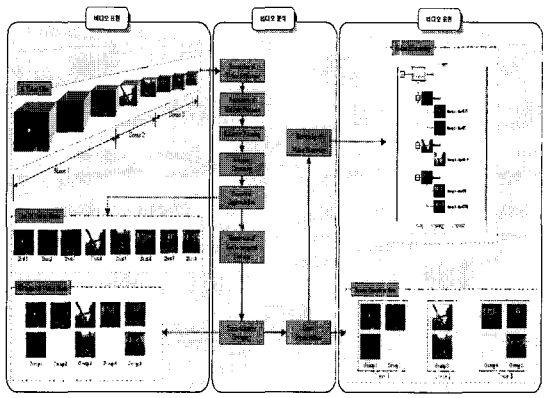


그림 3 시스템 프레임워크

4. 제안된 샷 경계면 검출과 대표 프레임 추출 방법

g비디오 시퀀스에 대한 샷 경계면 분할 기법은 주로 프레임 간의 차이 측정(frame difference)에 의존한다. 이 단계에서의 주요 가정은 단일 샷 내에서의 프레임 간의 변이는 낮은 수치 값을 가지며, 결과적으로 서서히 변화하는 신호와 같다는 것이다. 이는 샷 내의 시간적 분산(temporal variance)이 작다는 것을 의미한다. 샷 경계면에서는 전체적 프레임 특성의 급격한 변화가 발

생하는 날카로운 피크가 관찰되고, 이러한 피크들의 차이 신호는 임계치나 클러스터링을 통해 검출될 수 있다. 검출 성능에 강하게 영향을 주는 요소들은 선택된 특징들, 잡음, 카메라와 객체의 움직임 등이며, 1차원 신호 차원에서 생각했을 때 비디오 시퀀스의 시간적 움직임 변화 레벨은 일반적으로 카메라와 객체의 움직임, 급격한 변화의 효과에 의해 변화한다. 이 때 샷 경계면은 아니지만 프레임 간의 강한 움직임 변화는 잘못된 피크를 나타낼 수 있으며, 이는 임계치나 클러스터링 기반 방법에서는 경계면이 아닌데 경계면으로 잘못 판단하는 오판 경보(false alarm)를 야기할 수 있다.

본 논문의 샷 경계면 검출 단계에서는 실제의 샷 경계면에 심오한 영향을 주거나 오판하여 제거되는 것 없이 객체 또는 카메라의 강한 움직임으로 인한 샷 내의 잡음을 감소시키기 위해 프레임 차이 신호를 시간 적응적 필터링한다[22]. 그리고 최종 단계인 샷 경계면 여부를 결정하는 k-means 클러스터링의 입력으로 복수 특징이 이용되는데[23], 이는 앞에 설명한 시간 적응적 필터링 과정에 의해 전처리 된다. 샷 경계면 검출에 있어 비디오 편집 효과를 고려해야 하는데, 이는 [24]에서 비디오 생산자 입장에서 잘 정의 되어 있다. 비디오 편집 효과는 컷(cut), 이동, 페이징, 페이드, 디졸브, 와이프, 몰핑 등과 같은 것으로 이루어져 있는데, 본 논문의 방법에서는 컷에서의 성능이 좋았다.

4.1 복수 특징 추출

클러스터링 기반의 샷 경계면 검출의 큰 특징 중의 하나는 검출 성능을 향상시키기 위해 복수 특징을 동시에 사용할 수 있다는 것이다. 서로 보완 관계에 있는 프레임 쌍의 전체적인 내용을 반영하는 히스토그램 차이 방법(HDM)과 지역적 특성을 잘 반영하는 화소 차이 방법(PDM)의 복수 메트릭을 이용함으로써 비디오 내용 변화를 최적에 가깝게 표현할 수 있다. 여러 가지 색상 모델과 다차원 히스토그램에 대해 실험한 결과, RGB 색상 모델의 각 채널에 대한 히스토그램 차이를 구해 일련으로 각각을 더하는 선형 기법의 성능이 가장 좋았으므로 본 논문에서는 이 색상 모델과 히스토그램 기법을 사용한다.

프레임 크기 M×N의 k번째 쌍의 PD(f_k,f_{k+1}), HD(f_k,f_{k+1})를 각각 화소 차이와 히스토그램 차이 메트릭이라 표시하면 다음 식과 같이 표현된다.

$$d_{i,j}(f_k,f_{k+1})=\begin{cases}1 & \text{if } |t_{i,j}(f_k)-t_{i,j}(f_{k+1})|>0\\ 0 & \text{Otherwise}\end{cases}$$
$$\mathrm{PD}(f_k,f_{k+1})=\frac{1}{M\times N}\sum_{i=1}^{M}\sum_{j=1}^{N}d_{i,j}(f_k,f_{k+1}) \quad (1)$$

PDM에서 식(1)과 같이 해당 화소가 다를 때만 1로 선택한 이유는 화소의 작은 변화에 민감하지 않게 하기 위함이며, 화소 차이의 잡음을 피하기 위해 각각의 채널은 32개의 bin으로 균일하게 양자화 되었다. HDM에 대한 식은 다음과 같다. 아래에서 768은 R,G,B의 각 bin수의 합이다.

$$HD_{i,j}(f_k,f_{k+1})=\frac{1}{M\times N}\sum_{j=1}^{768}|\mathrm{H}_i(j)-\mathrm{H}_{i+1}(j)| \quad (2)$$

유동적인 이 두 특징 값을 0.0과 1.0의 사이 값으로 정규화함으로써 k-means 클러스터링에서 고정된 임계치 없이 비디오 종류와 내용에 적응적으로 대처할 수 있다.

클러스터링 기반의 기법에서 이 두 복수 특징을 이용하는 것은 화소 차이 방법의 민감성 때문에 많은 오판 경보(false alarm)를 생성하게 되는데, 이는 시간 적응적 필터링 단계에서 제거된다.

4.2 시간 적응적 필터링

위에서 추출된 두 특징 값에서 객체나 카메라의 움직임에 의한 특정 샷 내의 프레임 간의 강한 변화는 오판될 수 있는 피크를 나타낼 수 있으므로 이를 제거해 주는 단계가 필요하다.

프레임 차이 기반 샷 경계면 검출은 영상 처리에서 에지를 검출하는 응용 프로그램과 매우 유사하다. 다시 말해, 프레임 차이 신호는 일련의 1차원 화소 차이값들과 같으며, 영상에서 화소값의 불연속인 점이 에지를 형성하듯이 프레임 차이 신호 내의 샷 경계면을 나타내는 피크는 고주파 성분(충격잡음)으로 나타나고 특정 샷 내의 프레임 간의 강한 변화는 저주파(양자화 잡음)형태로 볼 수 있다. 그러므로 오판될 소지가 있는 이 양자화 잡음인 저주파 성분을 제거함으로써 샷 내의 시간적 분산(temporal variance)을 줄여, 비감독 k-means 클러스터링에서 더 좋은 성능을 얻을 수 있다.

이 두 특징 값들을 아래와 같은 신호 모델 $f[n]$으로 보고 5개의 윈도우 크기 중 median 필터링의 경우 최대 분산, mean 필터링의 경우는 최소 분산을 가지는 시간 적응적 필터링을 통해 오판 경보(false alarm)를 제거 하였다.

$$f[n]=h[n]+\eta_{impulsive}, \quad (3)$$

$f[n]$: 특징값의 변화 신호,
$h[n]$: 프레임간의 차이 신호(샷 내의 변이, 양자화 잡음),
$\eta_{impulsive}$: 충격 잡음(샷 경계면 급격한 불연속).

이 신호 모델은 $f[n]$에서 $\eta_{impulsive}$ 를 추출하기 위해 mean과 median 필터링을 이용하는데 mean 필터링은 에지를 어느 정도 손상시키면서 충격 잡음을 제거하고, median 필터링은 에지 성분의 손상 없이 충격 잡음을 제거하는 기법으로 널리 알려져 있다. 시간적 필터링 식은 다음과 같이 정의된다.

$$d[n] = \hat{\eta}_{impulsive} = f[n] - \Im_{median}\{f[n]\} = f[n] - \Im_{mean}\{f[n]\}$$
$$g[n] = \begin{cases} d[n], & \text{if } d[n] > 0 \\ 0, & \text{Otherwise} \end{cases} \quad (4)$$

위의 식에서와 같이 특징 값의 변화 신호에 대해 mean, median 필터링을 하게 되면 샷 경계면을 나타내는 고주파 신호는 저주파 패스 필터링을 통해 제거되고, 카메라나 객체의 움직임으로 인한 샷 내에 강한 변화는 남게 된다. 그러므로 원 신호에서 이 저주파 신호를 제하면 샷 경계면 신호의 추정치를 얻을 수 있다. 본 논문에서는 더 좋은 성능을 얻고자 이 두 mean, median 필터링 기법을 조합하여 사용한다. 실험 결과 mean 필터링한 후, median 필터링을 하였을 경우 샷 내의 특징 값의 편차가 작게 나왔다.

이 기법에서 검출 성능에 직접적으로 영향을 주는 파라메터는 필터 탭 수로, 길이가 긴 필터는 recall과 precision의 성능을 악화 시킬 수 있으므로 길이가 약간 짧은 필터가 좋은 결과를 얻는다. 그 이유는 길이가 긴 필터는 시간적으로 높은 움직임 변화도에 있는 샷 내의 샘플의 값을 짧은 필터보다 크게 줄일 수 없기 때문이다. 신호는 시간에 대해 변화하는 특성을 가지고 있으므로 전체 신호에 대해 동일한 필터 크기를 사용해서는 최적의 결과를 얻기 힘들다. 필터 크기는 지역적 분산의 함수로써 시간 적응적으로 정의되고 수정될 수 있다.

본 논문에서는 적절한 필터 크기의 선택과 구현을 다음과 같은 방법에 의해 수행하였다.

i) 각 표본(sample)에 대해 N개의 윈도우 크기(S_i, i=1,·,N)마다 지역적 분산이 계산된다.

ii) mean 필터링의 경우, 주어진 표본에 대한 필터 크기로 가장 큰 표본 분산을 가지는 윈도우 크기를 선택한다.

iii) median 필터링의 경우, 주어진 표본에 대한 필터 크기로 가장 작은 표본 분산을 가지는 윈도우 크기를 선택한다.

mean 필터링에서 가장 큰 표본 분산을 가지는 윈도우를 이용함으로써 높은 움직임 변화도를 가지는 샷 내에서 오판될 수 있는 피크, 즉 카메라나 객체의 움직임에 의한 활동 효과를 크게 억제 할 수 있다. 낮은 분산은 윈도우 내의 신호 값이 균일하게 분포함을 나타내는데, 이 특성은 median 필터링의 지역적 특성을 잘 반영하여 좋은 성능을 얻게 한다. 비록 이 기법은 처리를 하기 전에 사용될 윈도우 수와 크기를 정의하여야 하는 단점이 있지만, 분산과 관련된 매개변수나 임계치의 정의를 필요치 않게 한다. 본 논문의 실험에서는 필터 크기로 15, 25, 39, 59, 83의 5가지 윈도우를 사용하였다. 여기서 필터의 수는 샷의 경계면과 샷의 길이와 밀접한 관계가 있으므로 샷의 평균 길이를 넘지 않는 한도 내에서 샷을 포함할 수 있도록 설계하였다.

4.3 비감독(unsupervised) k-means 클러스터링

필터링된 두 특징 값들은 샷 경계면 여부를 분류하기 위해 2-means 클러스터링을 이용하게 되는데, 이 기법의 장점은 복수 특징들을 이용할 수 있을 뿐만 아니라 고정된 임계치 없이 도메인에 독립적으로 적용될 수 있기 때문이다.

사용된 k-means 알고리즘은 샷 경계면과 샷 경계면이 아닌 두 클러스터로 시작하며, 프레임 차이 값은 거리 측정(distance measure) 방법이므로 두 클래스의 초기 중심점(centroid)은 샷 경계면의 경우 1.0으로 선택하고, 샷 경계면이 아닌 경우는 0.0으로 선택할 수 있다. 그리고 클러스터의 중심점이 변하지 않을 때까지 자료를 입력하며 유클리디안 거리(euclidean distance)를 구해 제일 가까운 쪽으로 자료 항목을 클러스터링하고 중심점 값을 갱신한다. 결과로 나온 중심점을 이용하여 샷 경계면 검출 유무를 결정하게 된다. 그러나 이상적인 계산 시간을 얻기 위해서 모든 가능한 클러스터를 고려하지 않을 수도 있다. 이 기법의 단점으로는 클러스터링 결과가 시작 중심점과 자료 항목의 입력 순서에 따라 변할 수 있다는 것이다.

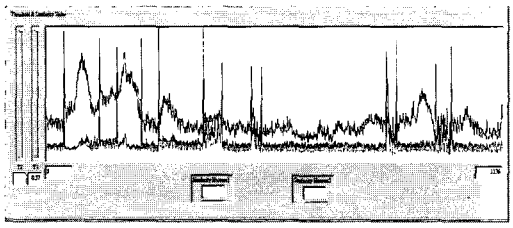
대표 프레임 추출은 빠른 처리 속도를 위해 샷의 처음과 마지막 프레임을 선택하였다.

5. 샷 경계면 검출 결과

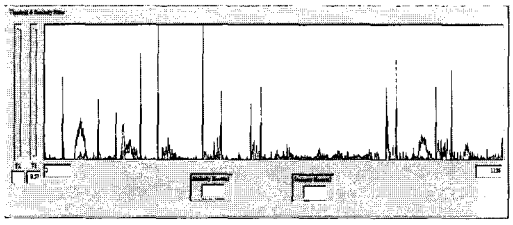
그림 4의 (a), (b), (c)는 차례대로 특징값 PDM과 HDM의 원신호와 경계면 검출의 성능 향상을 위한 mean과 median 필터링한 결과를 나타낸다. 처리된 신호는 입력 신호와 매우 유사했으며, 샷 경계면의 신호 값에서의 작은 손실과 샷 내의 오판 경보를 나타낼 수 있는 작은 변화를 제거하는 미묘한 차이를 보였다. 특징 값들에 대한 필터링 효과와 클러스터링 결과는 그림 5에서 보여지는 scatter plot에서 더 잘 관찰된다.

필터링 처리는 특징 공간의 표본 분포에 영향을 주며, 처리된 신호의 0 클리핑(clipping) 연산은 입력 신호의

동적 범위에 변화를 주는 것이 아니라 신호의 편차를 줄임으로써 성능에 영향을 준다. 즉, 모든 음의 값이 제거됨으로써 처리된 신호의 주요 표본들은 0 또는 0에 가까운 값을 가진다. 이것은 샷 비경계면(non-boundary) 클러스터의 표본의 수를 증가 시키고, 중심점의 위치를 원점으로 옮기는 역할을 한다. 결과적으로 특징 공간에서 두 클러스터의 분리 가능성을 증가시킨다. 비록 그림 5의 (a)에서는 두 클러스터의 명확한 경계가 없지만 필터링 처리 후의 결과 (b), (c)에서는 확연히 드러나는 것을 볼 수 있다. 또한 (c)는 필터링된 후 k-means 클러스터링 결과를 보여주는데, 원점에 가까운 샷 비경계면 클러스터의 중심점은 작은 청색 사각형으로 나타내었고, 경계면 클러스터의 중심점은 원점에서 멀리 떨어진 작은 적색 사각형으로 표시됨을 볼 수 있다.



(a) 영화 비디오 0~1136의 프레임 쌍에 대한 특징값 PDM(청색), HDM(적색)의 원신호



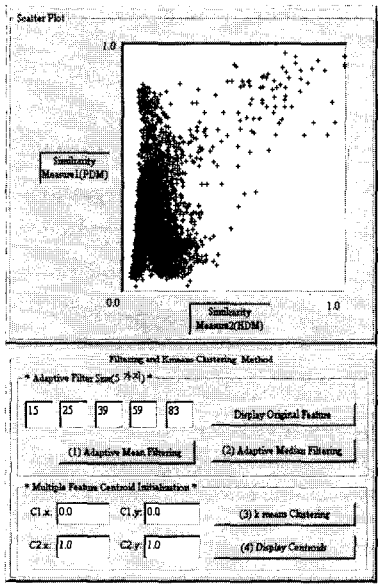
(b) (a) 결과에 대한 mean 필터링 결과

(c) (b) 결과에 대한 median 필터링 결과

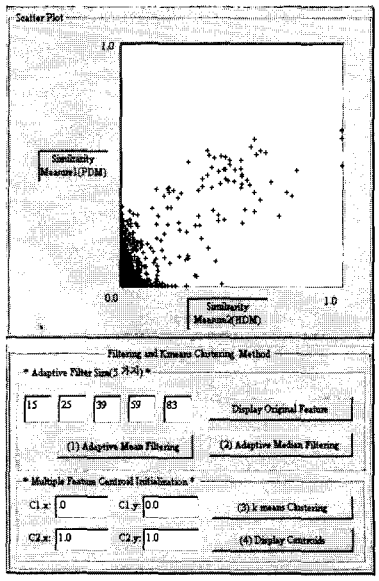
그림 4 필터링 처리 결과

6. 장면 구조 구축

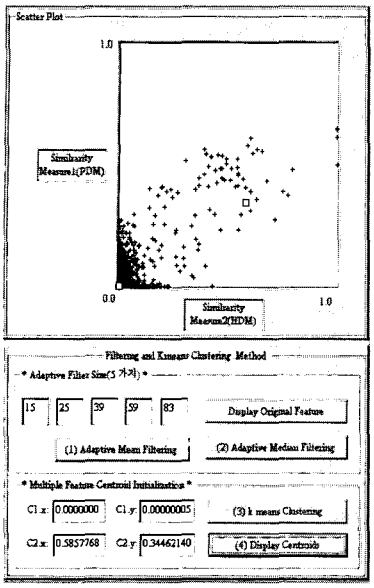
위 5장의 샷 경계면 검출에서 얻은 결과로 순차적인 일련의 샷 리스트들을 생성하고, 이 리스트들을 구성하는 샷들은 내용을 대표할 수 있는 기술자(descriptor), 즉 여러 가지 시공간 정보들로 표현되어 샷 리스트에 저장된다.



(a) 두 특징의 원신호



(b) (a)에 시간 적응적 mean 필터링을 한 신호



(c) (b)에 시간 적응적 median 필터링을 한 후 k-means 클러스터링한 결과

그림 5 2차원 특징 공간에서 프레임 비교 측정에 대한 scatter plot

6.1 시공간 특징 추출

샷은 다음과 같은 시공간 특징들로 표현된다.

$$shot_i = shot_i(b_i, e_i, Act_i, Hist(b_i), Hist(e_i)) \quad (5)$$

공간적 특징으로,

b_i : 샷의 처음 프레임 번호,

e_i : 샷의 마지막 프레임 번호,

$Hist(b_i)$: b_i 프레임의 히스토그램,

$Hist(e_i)$: e_i 프레임의 히스토그램,

시간적 특징으로,

$$Act_i = \frac{1}{N_i - 1}\sum_{k=1}^{N_i-1} diff_{k,k-1}$$

$$diff_{k,k-1} = \mathrm{Dist}(Hist(k), Hist(k-1)) \quad (6)$$

위의 (6)식에서 $Hist(k)$와 $Hist(k-1)$는 각각 프레임 k와 k-1에서의 히스토그램을 나타내고, $Dist()$은 거리 측정(distance measure)을 이용한 히스토그램 인터섹션을 나타낸다. Act_i는 조명의 영향을 제거하기 위해 RGB 색상 모델에서 HSI 색상 모델로 변환[25][26] 한 후 조명의 영향에 민감한 성분인 명도값 I(intensity)를 제외한 HS의 2차원 히스토그램 인터섹션(histogram intersection)의 평균을 사용하였다. 즉 샷의 시간적 특징 정보로는 위의 히스토그램 인터섹션 평균으로 대표했다.

6.2 시간 적응적 그룹화

샷들을 장면 구조로 구성하기 전에 유사한 샷들을 그룹화하면 편리한데 샷들이 유사(그룹화)하게 되기 위한 특성은 시각적으로 유사해야 하며, 또한 시간적으로 서로 이웃한 위치에 있어야 하는 시간 지역성을 만족해야 한다. 이를 위해 앞에서 기술한 시공간 특징들과 시간 지역성을 만족시키는 그룹화를 이 단계에서 하게 되며, 유사도는 다음과 같이 정의된다.

$$ShotSim_{i,j} = W_C \cdot ShotColorSim_{i,j} + W_A \cdot ShotActSim_{i,j}, \quad (7)$$

$ShotColorSim_{i,j}$: 샷에 대한 색상 유사도
(공간 정보 +시간의지역성 정보)

$ShotActSim_{i,j}$: 샷 내의 움직임 변화에 대한 유사도
(시간 정보 +시간의 지역성 정보)

W_C, W_A : 가중 매개 변수

W_C, W_A는 색상과 움직임 정보에 대한 가중 매개변수로 식 (8)과 같이 이전 샷과의 통계적 분산을 이용하여 적응적으로 자동 선택된다.

$$w_C = \frac{\sigma_C}{\sigma_C + \sigma_A}, \quad w_A = \frac{\sigma_A}{\sigma_C + \sigma_A} \quad (8)$$

σ_C 와 σ_A은 현재 샷과 현재 샷과 MULTIPLE(시간 인력 상수)수 만큼의 이전 샷들의 각 특징들, 칼라 히스토그램과 Act_i의 표준 편차를 나타낸다.

샷 색상 유사도(ShotColorSim$_{i,j}$)는 아래 식 (9)와 같이 정의되는데, 1)공간 정보로서 프레임 색상 유사도(2차원 히스토그램 인터섹션)와 2)시간 지역성 정보로서 시간 인력(temporal attraction)등, 이 두 가지 정보를 이용한다. 또한 시간 적응적(temporal adaptive) 그룹화를 하기 위해 4가지 중 가장 큰 유사도를 선택하는 시간 적응적 유사도를 구하게 된다.

$$ShotColorSim_{i,j} = Max(FrameColorSim'_{b_j,e_i}, FrameColorSim'_{e_j,e_i}, FrameColorSim'_{e_j,b_i}, FrameColorSim'_{b_j,b_i}) \quad (9)$$

위 식에서

$$FrameColorSim_{b_i,e_i} = Attr_{b_i,e_i} \cdot FrameColorSim_{b_i,e_i}. \quad (10)$$

$$Attr_{b_j,e_i} = \mathrm{Max}(0,1-\frac{(b_j-e_i)}{MULTIPLE \cdot 샷의\ 평균\ 길이}), \quad (11)$$

$$FrameColorSim_{b_j,e_i} = 1 - diff_{b_j,e_i} \quad (12)$$

4개의 조합 중 하나만을 표시하면, 아래의 식과 같고, 여기서 $MULTIPLE$은 시간 인력이 얼마나 빠르게 0으로 감소할 것인지를 조절하는 상수로서 본 논문의 실험 결과에서는 7이 좋은 결과를 나타내었다.

샷 색상 유사도의 한 요소인 $FrameColorSim'_{bj,ei}$은 거리 측정(distance measure)을 유사도로 전환하는 히스토그램 인터섹션의 반수(reciprocal)인 공간 정보($FrameColorSim_{bj,ei}$)와 일정한 시간의 지역성 안에 없으면 0으로 하고 아니면 가까운 쪽에 더 많은 인력정보를 주는 시간 지역성 정보($Attr_{bj,ei}$)로 표현된다.

샷 움직임 변화 유사도는 식 (13)와 같다. 식의 의미는 위와 유사하고 시간 인력은 비교하려는 두 샷의 중심거리를 이용한다.

$$ShotActSim_{i,j} = Attr_{center} \cdot |Act_i - Act_j|, \quad (13)$$

$$Attr_{center} = \mathrm{Max}(0,1-\frac{(b_j+e_j)/2-(b_i+e_i)/2}{MULTIPLE \cdot 샷의평균길이}) \quad (14)$$

위의 식에서 의미하는 것과 같이 샷에 대한 시공간 정보와 시간 적응적 지역성 정보를 함께 이용하여 그룹화와 장면 구조를 구성하게 된다.

6.3 계층적 장면 구조 구성

위의 식들을 이용한 클러스터링 기반의 시간 적응적 그룹화를 통해 계층적 장면 구조가 생성하게 되는데, 다음과 같은 두 단계의 처리에 의해 수행 될 수 있다.

단계 1 : 시간 적응적 그룹화를 이용하여 유사한 샷들을 그룹화한다.

단계 2 : 의미적으로 관계 있는 그룹들을 하나의 장면으로 합병한다.

이에 대한 전체 알고리즘은 그림 6와 같다. 이 알고리즘의 입력은 비구조적 비디오 스트림이 사용되고, 출력은 장면, 그룹, 샷, 대표 프레임으로 이루어진 계층적 비디오 구조가 생성된다. 그 예는 그림 7과 같고, 이 알고리즘의 입출력 구조는 그림 8과 같다. 이를 바탕으로 전체적인 구조를 추상화 처리하여 압축 요약한 형태인 목차 형식을 얻을 수 있는데 사용자는 브라우징된 목차 형식을 보고 전체 구조를 이해한다.

알고리즘에서 초기화로 시작하여 현재 샷이 삽입될 그룹을 찾을 때 그룹 유사도가 가장 큰 쪽으로 병합이 된다. 그룹 유사도는 그림 9에서 처럼 현재 샷과 그룹에서 가장 최근 입력된 샷의 정보를 가지고 구하게 된다. 여기서 그룹을 대표하기 위해 가장 최근에 입력된 샷을 선택했는데 그 이유는 같은 그룹에 위치한 샷들은 시공간적으로 유사하고 가장 최근에 입력된 샷이 현재 샷과 시간의 지역성이 크다고 생각했기 때문이다. 현재 샷의 장면 유사도는 현재 샷과 해당 장면에 있는 모든 그룹들과의 그룹 유사도를 구해 그 평균을 구함으로써 대표할 수 있다. 그림 9과 같이 샷 리스트의 현재 샷에 대한 그룹 유사도와 장면 유사도를 구해 사전에 정의된 임계치를 이용하여 합병하게 되는데, 그룹의 경우는 시간의 연결성이 만족하지 않아도 상관 없지만, 장면의 경우는 의미차원의 이야기 형식이므로 시간의 연속성은 중요한 조건이다. 그러므로 반복적으로 위의 알고리즘을 수행하며 시간의 연속성이 만족되지 않을 경우는 두 장면을 합병한다. 이 알고리즘에서 사용된 그룹과 장면의 임계치는 정규화 된 값으로 비교를 함으로 각 특징 값과 비디오 장르에 대한 의존성을 어느 정도 피할 수 있다. 다시 말하면 한번 정해진 임계치는 자신이 속한 도메인 내에서 동일 하게 같이 사용할 수 있다. 본 논문에서는 이 임계치(범위:0.0~1.0)를 각 장르마다 실험을 통해 구하였다.

이 알고리즘의 장점은 다음과 같다.

i) [21]의 시간 제약 클러스터링 기법은 시간 윈도우 크기 T를 사용함으로써 윈도우 효과에 의한 시간적 불연속성이 발생할 수 있다는 문제점이 있다. 이 문제점은 시간 적응적 지역성을 이용하여 샷에서 나타나는 시간에 대한 불연속적 분할을 해결할 수 있다.

ii) 시간적 해상도(temporal resolution)를 고려해서 샷과 다수의 샷으로 이루어진 그룹과의 유사도, 샷과 다수의 그룹으로 이루어진 장면과의 유사도를 구한다.

본 논문에서 사용된 방법과 유사한 계층별 비교에 대한 설명과 기법이 [27],[28]에 소개되었다. 이를 이용한 비디오 장르별 유사도 계산도 가능하다[29].

7. 실험 결과

7.1 실험 환경 및 비디오 자료

실세계에는 영화, 뉴스, 시트콤, 상업 광고, 스포츠, 다큐멘터리 비디오를 포함하는 여러 가지의 비디오 타입이 존재하며, 이중 영화 비디오는 이야기 전개 방식으로 어떤 흐름을 가지고 있는 반면에 스포츠와 같은 타입은 그런 흐름이 없다. 본 논문에서는 이야기 흐름을 가지고

❑ 비디오 계층 구조 생성 절차 ❑

Input : S = { shot 1, shot 2, shot 3,… , shot n}

Output : 샷, 그룹, 장면으로 구성된 계층적 장면 구조

1. 초기화 : 샷 0 을 그룹 0과 장면 0에 포함
2. For(NumOfGroup =1; NumOfscene=1; S != EMPTY; Next Shot;){
   FindGroupSim(); // 현재 샷과 구성된 그룹들과의 유사도를 구한다.
   MaxGSim = MaxGroupSim(); // 위에서 구한 유사도중 가장 큰 유사도를 구한다.
   If(MaxGSim > GroupThreshold){ // 현재 샷이 구성된 그룹에 합병될 수 있는지를 검사.
   MergeCurShotMaxGroup (); // 현재 샷을 최대 유사도를 가진 그룹에 합병.
   UpdateGroupScene (); // 장면과 장면의 시간적 불연속성이 발생하면 장면 구조 갱신.
   continue;
   } // end of if
   else{
   CreateNewGroupForCurShot (); // 현재 샷을 포함하는 그룹 생성.
   NumOfGroup ++;
   } // end of else
   FindSceneSim(); // 현재 샷과 구성된 장면들과의 유사도를 구한다.
   MaxSSim = MaxSceneGroupSim(); // 위에서 구한 유사도중 가장 큰 유사도를 구한다.
   If(MaxSSim > SceneThreshold){ // 현재 샷이 구성된 장면에 합병될 수 있는지를 검사.
   MergeCurShotMaxSceneGroup (); // 현재 샷을 최대 유사도를 가진 장면에 합병.
   UpdateScene (); // 장면과 장면의 시간적 불연속성이 발생하면 장면 구조 갱신.
   continue;
   } // end of if
   else{
   CreateNewSceneForCurShot (); // 현재 샷을 포함하는 장면 생성.
   NumOfScene ++;
   } // end of else
} // end of for

그림 6 비디오 계층 구조 생성 절차

❑ 비디오계층 구조 예 ❑

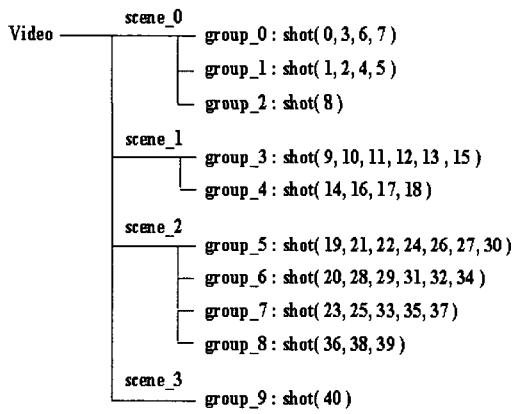


그림 7 비디오 계층 구조 예

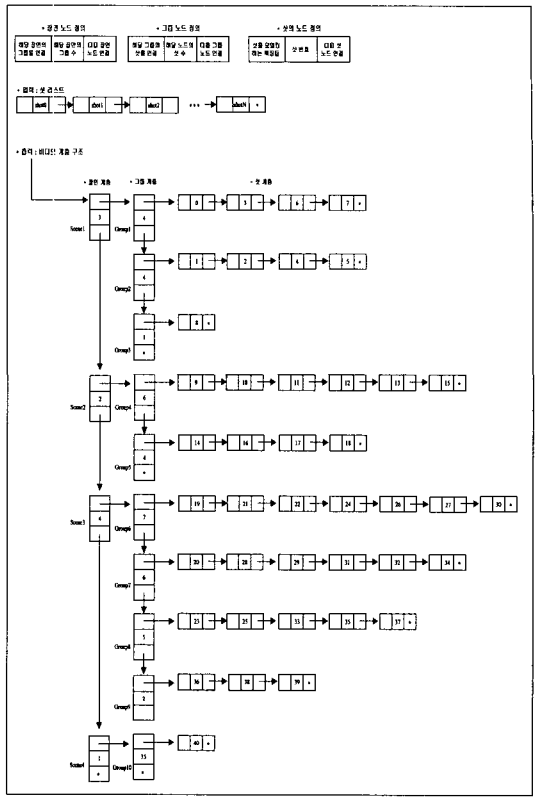


그림 8 비디오 입출력 구조

```
// 첫번째 loop는 장면의 노드를 쫓아간다.
// 두번째 loop는 그룹의 노드를 쫓아간다.
// 세번째 loop는 주어진 그룹에서 맨 마지막 샷을 찾는다.
for(pScene=m_pVideoTOC; pScene!=NULL;pScene=pScene->m_pNext)
      for( pGroup=pScene->m_pGroupNode;pGroup!=NULL;pGroup=pGroup->m_pNext){
            pShot =pGroup->m_pShotNode; //비교할 샷 노드를 찾는다.
                  while(pShot->m_pNext!=NULL)
                        pShot = pShot->m_pNext;
                  // 마지막 샷 노드를 찾았으면 현재 샷 노드와 비교하기 위해
                  // 두 샷의 유사도를 구한다.
                  Sim = ShotSim(pCurShotNode,pShot);
                  if(Sim>MaxSim){
                        *pSceneMax = pScene;
                        *pGroupMax = pGroup;
                        MaxSim = Sim;
                  }
            }
            return MaxSim;
}
```

그림 9 그룹 유사도 코드

있는 비디오 타입에 대해서 중점을 두었다. 실험 비디오 자료로는 영화 비디오 중 성능 평가에 크게 영향을 주는 15 frame/sec의 크기 240*180인 정적인 영화, 동적인 영화, 애니매이션 영화 비디오의 세 가지 형태에서 선정하여 Pentium-II 350 PC에서 실험하였고 실험 비디오 자료는 표 1과 같다.

표 1 실험 비디오 자료

종 류	비디오 제목	분 량	프레임 수
정적 영화 비디오	시네마천국	6분	5564개
동적 영화 비디오	토탈리콜	10분	9009개
애니매이션 영화 비디오	뮬 란	5분	4495개

7.1 실험 결과 및 분석

장면의 기준이 주관적이므로 본 논문에서는 검색 성능 평가를 위해 타 논문에서 자주 사용되는 recall과 precision을 사용하였다. 이 평가 방법은 각 기준에 대한 절대 평가를 할 수 있다. 장면의 기준은 실험을 하기 전에 미리 어느 정도의 객관성을 얻기 위해 여러 사람을 통해 정해지고, recall과 precision의 평가 방법에 의해 성능을 측정한다. 성능은 측정 값이 절대 단위 값 1에 접근할수록 좋은 방법임을 나타내므로 타 논문과도 비교를 할 수 있다. 이에 대한 정의는 다음과 같다.

위에서 Correct는 정확히 검출된 경계 수, False Negative는 검출하지 못한 경계 수, False Positive는 잘못된 경계 수를 나타낸다. 검색 환경에서 recall은 실제 원하는 결과에 대한 검색 결과에 대한 관련 비율을 나타내며, precision은 검색된 정보 또는 선택된 정보의 결과가 얼마나 관련되었는가를 측정한다. 두 파라메터의 값이 단위 값 1에 접근할수록 성공적이고 일관적인 기법으로 볼 수 있다.

위의 비디오 자료로 실험한 결과인 샷 경계면 검출과 계층적 장면 구조 구성 결과는 각각 표 2와 표 3과 같다.

표 2 샷 경계면 검출 결과

비디오 제목	실제 샷수	Correct	False-	False+	Recall	Precision
시네마천국	66개	61쌍	4	1	0.938462	0.983871
토탈리콜	158개	154쌍	3	10	0.980892	0.939024
뮬 란	98개	88쌍	9	3	0.907216	0.967033

표 3 계층적 장면 구성 결과

비디오 제목	그룹 수	검출된 장면 수	False-	False+
시네마천국	14개	5개	2	2
토탈리콜	23개	8개	4	3
뮬 란	24개	6개	1	3

최근까지 기존 연구에서 제시한 알고리즘과 그에 대한 성능 분석은 주로 임계치에 의존했기 때문에, 해당 논문에서 밝힌 결과는 해당 알고리즘에 최적의 성능이 출력되도록 임계치를 임의로 고정시키고 실험한 경우가 많았다. 따라서 출력 결과를 그 기법의 절대적으로 평가하는 것은 일관성이 없게 되어 버린다. 또한 임계치는 추출된 특징이나 비디오 타입에 따라서 유동적으로 변한다. 그러므로 본 논문에서는 다른 연구의 알고리즘 성능과 비교 없이 절대적 평가 방법인 recall과 precision으로 성능을 평가하였다.

샷 경계면은 카메라의 물리적인 경계면이기 때문에 실제 경계면을 사전에 검사하여 쉽게 구분하고 가록할 수 있지만, 장면 경계면인 경우는 의미적인 개념에 의해 구분되는 것이므로 사용자가 어느 사건이나 객체에 중점을 두는 가에 따라 주관적으로 달라질 수 있다[21][5]. 위의 표에서 보는 것과 같이 샷 경계면 검출의 결과는 성능이 평균 95%로 우수했고, 계층적 장면 구조 구성의 결과인 장면 경계면 검출의 경우도 어느 정도 만족할 만한 결과를 보였다.

최종적으로 계층적 비디오 구조가 구성되면 샷의 대표 프레임에 의해 의미 차원의 목차 형식으로 비디오 구조를 추상화 하고 브라우징할 수 있는데, 이에 대한 장면 계층에 대한 시각적 브라우징 결과가 그림 10과 같다. 그림에서 처럼 5564개 프레임의 시네마 천국 비디오 클립에서 5개의 장면이 생성 되었다. 비디오 생산업자 측에서는 각 장면에 주석을 달아 편집할 수 있으며, 사용자는 이 비디오 목차 구조의 대표 프레임과 주석에 의해 비디오 클립의 전체 내용을 쉽게 이해할 수 있다.

또한 사용자는 자세한 내부 구조를 이해하기 위해 장면 구조를 그룹과 샷들로 이루어진 세세한 하부 계층 구조로 확장할 수 있는데 그림 11과 같다.

그림 11에서 처럼 계층적 비디오 구조를 부라우징한 후에 사용자는 비디오 클립의 전체 내용을 시각적으로 이해하며, 빠른 감기나 되감기의 지루한 동작 없이 흥미있는 비디오의 특정 부분을 클릭하여 비선형적인 임의의 접근이 가능하다. 이에 대한 예가 그림 12에 나타나 있다.

8. 결론 및 향후 연구 방향

본 논문에서는 클러스터링 기반의 샷 경계면 검출과 더불어 지능적 클러스터링 기반의 비디오 계층 구조 구축 시스템을 제시하였다. 제안된 시스템은 비구조적 비디오에서 여러 차원(샷, 그룹, 장면)으로 비디오 파싱하여 시공간 정보와 시간 지역성 정보를 추출하였고, 또한

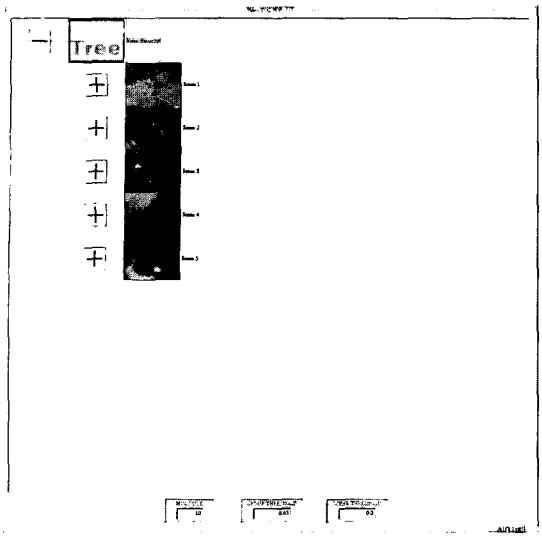


그림 10 시네마 천국 바디오 중 처음 6분 동안의 비디오 목차(장면 계층)

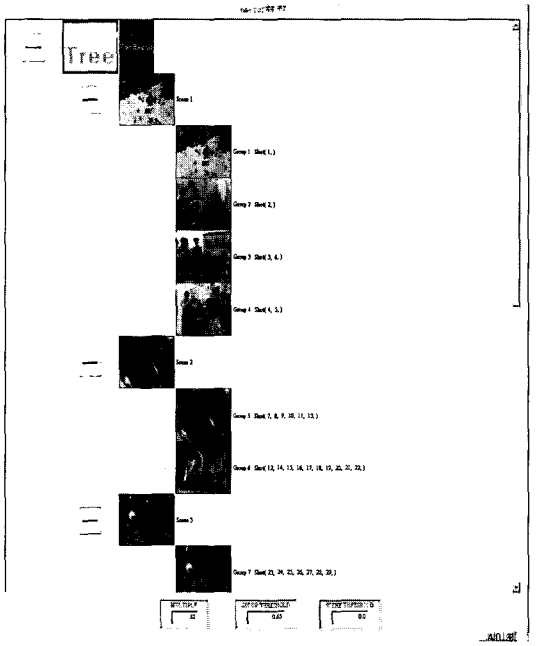


그림 11 그림 에 대한 비디오 그룹 계층

여러 시간적 해상도(temporal resolution)에 대해 유사도를 구하고 클러스터링하여 계층적 비디오 구조를 구성하였다.

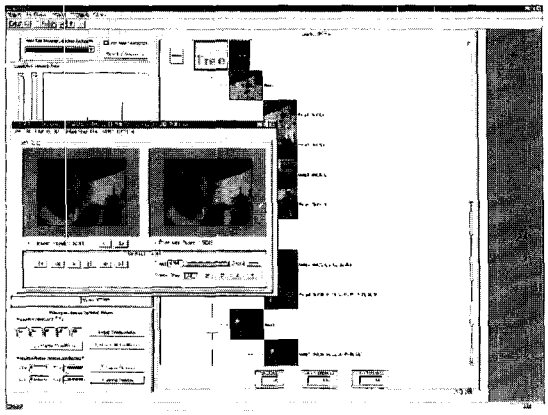


그림 12 비디오 클립의 비선형 임의 접근

또한 비디오 검색 시스템은 이 비디오 구조에 색인된 특징값들을 바탕으로 시각적 질의어와의 유사도를 구해 찾고자하는 영상이나 동영상을 탐색하는데 사용할 수 있고, 아울러 시각적 내용 기반 비디오 브라우징 응용 프로그램에서 표현과 추상화 그리고 위에서 예로 보인 의미 차원에서의 시각적 요약 목차 구성등에서 비순차적, 비선형 임의 접근이 가능하도록 사용될 수 있음을 보였다.

향후 연구 방향으로는 샷 경계면 추출에서는 영화의 전 구간에 대해 클러스터링하는 것은 부적절함으로 PDM과 HDM의 분포적 특징과 두 클러스터의 경계면 근처에 있는 데이타에 대해서 세밀하게 고려한 알고리즘의 사용으로 더 정확한 경계면을 추출할 수 있을 것이다. 또한 간단한 하위 레벨의 시공간 정보의 사용으로는 샷과 샷 또는 장면과의 의미를 정확하게 파악할 수 없으므로 전체적으로 음성 및 음향 정보, 텍스트 정보, 자막 정보와 같은 다양한 시각적 멀티미디어 특징 정보를 추가 및 통합함으로써 더 정확한 비디오 계층 구조를 구성할 수 있고[30], 사용자는 여러 형태의 질의와 검색의 정확성을 얻을 수 있을 것이다. 더 중요한 것은 처리 속도도 또한 간과해서는 안될 것이다. 사용자 입장에서 보았을 때, 비디오 질의어는 크게 시각적 내용 기반 질의와 개념(concept) 기반 질의의 두 가지로 대변된다. 시각적 질의는 주어진 예와 유사한 비디오 샷을 찾을 때, 개념 질의는 사용자가 관심 있어 하는 특정 객체나 사건이 발생하는 샷을 찾을 때 사용된다. 이 때 시각적 질의는 앞에서 설명한 멀티미디어 정보에 의해 더욱 정확한 비디오 구조를 구성 함으로써 높은 적중율 검색이 가능하며, 개념 질의의 경우는 객체에 대한 검

출, 추적, 인식 기술에 크게 의존함으로 이에 대한 비디오 객체 모델의 통합은 비디오 검색에 관련된 응용 프로그램에 큰 밑바탕이 될 것이다[16][31]

참 고 문 헌

[1] Shih-Fu Chang, William Chen, Horace J.Meng, Hari Sundaram and Di Zhong,"VideoQ: an automated content based video search system using visual cues,"*ACM Multimedia* 1997.

[2] Thomas S. Huang and Yong Rui, "Image Retrieval: Past, Present, and Future," invited paper in *Int Symposium on Multimedia Information Processing* , Dec 11-13, 1997, Taipei, Taiwan.

[3] Thomas S. Huang, Yong Rui, Trausti Kristjansson, Milind Naphade, and Yueting Zhuang, "Video Analysis and Representation," *ISO/IEC JTC1/SC29/WG11 M3110, MPEG98.*

[4] Yong Rui, Thomas S. Huang, and Sharad Mehrotra, "Constructing Table-of-Content for Videos," to appear in *ACM Multimedia Systems Journal*, Special Issue Multimedia Systems on Video Libraries, Sept, 1999.

[5] Ruud M. Bolle, Boon-Lock Yeo, Minerva M. Yeung, "Video Query: Beyond the keywords," Technical report, *IBM Research Report*, Oct 17 1996.

[6] H.J. Zhang, A. Kankanhalli, and S.W. Smoliar. "Automatic Partitioning of Full-motion Video," *Multimedia Systems*, 1(1):10-28, 1993.

[7] I.K. Sethi and Nilesh V. Patel, "A Statistical Approach to Scene Change Detection," *IS&T SPIE Proceedings:Storage and Retrieval for Image and Video Databases III*, Vol. 2420, pp. 329-339, Feb.1995, San Jose, California.

[8] B. -L. Yeo, "Efficiency processing of compressed image and video," *Technical report, PhD thesis, Princeton University*, 1996.

[9] R. Zabih, J. Miller, and K. Mai. "A feature-based algorithm for detecting and classifying scene breaks," *ACM International Conference on Multimedia*, pages 189-200, 1995.

[10] J.S. Boreczky and L.A. Rowe. "Comparison of Video Shot Boundary Detection Techniques," I.K. Sethi and R.C. Jain, editors, *Proceedings of IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases IV Vol. SPIE 2670*, pages 170-179, 1996.

[11] Wayne Wolf. "Key frame selection by motion analysis," In *Proc. IEEE Int. Conf. Acoust., Speech, and Signal Proc.*, 1996.

[12] Yueting Zhuang, Yong Rui, Thomas S. Huang, and Sharad Mehrotra, "Adaptive key frame extraction using unsupervised clustering," In *Proc. IEEE int. Cont. On Image Proc.*, 1998.

[13] B. Yeo and B. Liu. "On the Extraction of DC Sequence from MPEG Compressed Video," *Proceedings of ICIP*, 2:260-264, 1995.

[14] B. -L. Yeo and B. Liu. "Rapid Scene Analysis on Compressed Video," *IEEE Transactions on Circuits and Systems for Video technology*, 5(6), 1995.

[15] H.J. Zhang et al. "Video Parsing using Compressed Data," *SPIE Symposium on Electronic Imaging Science and Technology: Image and Video Processing II*, pages 142-149, 1994.

[16] Di Zhong and Shih-Fu Chang, "Video Object Model and Segmentation for Content Based Video Indexing," *ISCAS'97*, 11 June, Hong Kong.

[17] Di Zhong, H.J. Zhang and Shih-Fu Chang, "Clustering methods for video browsing and annotation," *Storage and Retrieval for Still Image and Video Databases IV, IS&T/SPIE's Electronic Imaging: Science & Technology 96* [2670-38].

[18] H.J. Zhang, Yihong Gong, S. W. Smoliar, and Shuang Yeo Tan, "Automatic Partitioning of news video," In *Proc. IEEE Int. Conf. on Mutimedia,Computing and Systems, 1994.*

[19] B.-L. Yeo, M.M. Yeung, IBM T. J. Research Center. "Classification, Simplification and Dynamic Visualization of Scene Transition Graphs for Video Browsing," *IS&T/SPIE Electronic Imaging98 :Storage and Image retrieval for Image and Video Databases VI*, 1998.

[20] M.M. Yeung, B.-L. Yeo, W. Wolf, and B. Liu. "Video Browsing using Clustering and Scene Transitions on Compressed Sequences," *IS& T/SPIE Multimedia Computing and Networking*, 1995.

[21] M. Yeung, B.-L. Yeo, and B. Liu. "Extracting Story Units from Long Programs for Video Browsing and Navigation," *International Conference on Multimedia Computing and Systems*, June 1996.

[22] A. M. Ferman, "Efficient filtering and clustering methods for temporal video segmentation and visual summarization," *J. Vis. Comm. and Image Rep.*, vol. 9, no. 4 (special issue), pp. 336-351, Dec. 1998.

[23] M. Naphade, R. Mehrotra, A. M. Ferman, J. Warnick, and T. S. Huang "A high performance algorithm for shot boundary detection using multiple cues," *Proc. IEEE Int. Conf. Image Proc.*, Chicago, IL, Oct. 1998.

[24] A. Hampapur, R. Jain, and T. Weymouth. "Digital Video Segmentation," In *Second Annual ACM MultiMedia Conference and Exposition*, 1994.

[25] Rafael C. Gonzales and Richard E. Woods. "Digital Image Processing," *Addison Wesley Publishing Company, Reading, Massachussetts*, 1993.

[26] Ioannis Pitas, *Digital Image Processing Algorithms*, Cambridge, Prentice-Hall, 1993.
[27] Rainer Lienhart, Silvia Pfeiffer and Wolfgang Effelsberg, "Video Abstracting," In *Communications of ACM*, pp. xx-yy, Dec. 1997.
[28] Rainer Lienhart, Wolfgang Effelsberg and Ramesh Jain, "Visual GREP:A ystematic method to compare and retrieve video sequence," *SPIE Vol . 3312, storage and Retrieval for image and Vidoe Databases VI*, 1998.
[29] Stephan Fischer, Rainer Lienhart and Wolfgang Effelsberg, "Automatic Recogition of Film Genres," In *Proc. ACM Multimedia 95*, San Francisco, CA, Nov. 1995, pp. 295-304.
[30] Rainer Lienhart and Frank Stuber, "Automatic text recognition in digital videos," *University of Mannheim, Department of Computer Science, Technical Report TR-95-036*, Dec. 1995.
[31] Di Zhong and Shih-Fu Chang, "AMOS: AN ACTIVE SYSTEM FOR MPEG-4 VIDEO OBJECT SEGMENTATION," *1998 International Conference on Image Processing*, October 4-7, 1998, Chicago, Illinois, USA.

최 익 원
1996년 연세대학교 전산학과(학사). 1999년 연세대학교 컴퓨터과학과(석사). 관심분야는 인공지능, 패턴인식, 영상처리

변 혜 란
정보과학회논문지 : 소프트웨어 및 응용 제 27 권 제 1 호 참조