지역 컨텍스트 및 전역 컨텍스트 정보를 이용한 비디오 장면 경계 검출

(Detection of Video Scene Boundaries based on the Local and Global Context Information)

강 행 봉 †

(Hang-Bong Kang)

요 약 장면 경계 검출은 비디오 데이타에서 의미적인 구조를 이해하는데 있어서 매우 중요한 역할을 한다. 하지만, 장면 경계 검출은 의미적인 일관성을 갖는 장면을 추출하여야 하므로 셧 경계 검출에 비해 매우 까다로운 작업이다. 본 논문에서는 비디오 데이타에 존재하는 의미적인 정보를 사용하기 위해 비디오 셧의 지역 및 전역 컨텍스트 정보를 추출하여 이를 바탕으로 장면 경계를 검출하는 방식을 제안한다. 비디오 셧의 지역 컨텍스트 정보는 셧 자체에 존재하는 컨텍스트 정보로서 전경 객체(foreground object), 배경(background) 및 움직임 정보들로 정의한다. 전역 컨텍스트 정보는 주어진 비디오 셧이 주위에 존재하는 다른 비디오 셧들과의 관계로부터 발생하는 다양한 컨텍스트로서 셧들간의 유사성, 상호 작용 및 셧들의 지속 시간 패턴으로 정의한다. 이런 컨텍스트 정보를 바탕으로 연결 작업, 연결 검증 작업 및 조정 작업등의 3단계 과정을 거쳐 장면을 검출한다. 제안된 방식을 TV 드라마 및 영화에 적용하여 80% 이상의 검출 정확도를 얻었다.

키워드 : 장면 검출, 컨텍스트 정보, 세만틱스

Abstract *Scene boundary detection is important in the understanding of semantic structure from video data. However, it is more difficult than shot change detection because scene boundary detection needs to understand semantics in video data well. In this paper, we propose a new approach to scene segmentation using contextual information in video data. The contextual information is divided into two categories: local and global contextual information. The local contextual information refers to the foreground regions' information, background and shot activity. The global contextual information refers to the video shot's environment or its relationship with other video shots. Coherence, interaction and the tempo of video shots are computed as global contextual information. Using the proposed contextual information, we detect scene boundaries. Our proposed approach consists of three consecutive steps: linking, verification, and adjusting. We experimented the proposed approach using TV dramas and movies. The detection accuracy of correct scene boundaries is over than 80 %.*

Key words : Scene segmentation, contextual information, semantics

1. 서 론

비디오 데이타는 대용량이고 비구조적이기 때문에, 원하는 비디오 클립을 신속하고 정확하게 인덱싱 및 검색하기 위해서는 비디오 데이타에 존재하는 셧-장면-비디오로 이루어지는 구조를 추출하는 것이 바람직하다[1, 2]. 셧이란 비디오 데이타의 기본 단위로서, 한 대의 카메라로 끊기지 않고 촬영한 연속된 프레임 시퀀스를 뜻하며, 인지적인 연속성을 가지고 있다. 종래의 많은 연구들은 주로 비디오 셧 검출에 집중되어 왔다. 장면은 연속적인 셧으로 구성되어 있으며, 의미적인 일관성을 갖는 단위이다. 따라서, 정확한 장면 추출 작업은 의미를 기반으로 한 비디오 데이타의 인덱싱 및 검색에 매우 유용하다.

장면 경계 추출에는 다양한 방법들이 사용되어 왔다. 이러한 방식들은 대체적으로 두 가지 방식으로 분류할 수 있다. 한 방식은 비디오 셧을 처리하는데 있어서 이

· 본 연구는 2001년 가톨릭대학교 교비연구비 지원으로 이루어졌음.

† 종신회원 : 가톨릭대학교 컴퓨터전자공학부 교수
hbkang@catholic.ac.kr

논문접수 : 2002년 4월 9일
심사완료 : 2002년 8월 20일

산적인 처리 방식이고[3, 4], 또 다른 방식은 비디오 셧의 연속성을 기반으로 한 처리 방식이다[5, 6]. 이산적인 처리 방식에서는 장면 경계를 검출하기 위해 비디오 셧을 하나의 기호로 치환하여, 궁극적으로 비디오 데이타를 연속된 기호 열로 표현한 다음, 연속된 기호 열로부터 일정한 시간 안에 존재하는 유사한 셧들의 군집화(clustering)를 수행하여 얻는다. 연속성을 기반으로 한 처리 방식은 비디오 셧의 길이 및 시간적인 연속성을 고려한 것으로서 비디오 셧간의 응집도(coherence)가 지역적으로 최소가 되는 점을 찾아 장면 경계로 검출하는 방식이다.

이러한 장면 경계 검출 방식이 인간과 같은 성능을 얻지 못하는 이유중의 하나는 의미 정보가 장면 경계 검출 방식에 제대로 반영되지 못하기 때문이다. 즉, 신호적인 유사성이나 연속성이 아닌 인간이 인지하는 의미적인 연속성을 찾는 것이 장면 경계 검출에 있어서 중요한 역할을 한다. 비디오 데이타로부터 인간이 인지하는 것과 유사한 의미 정보 추출은 매우 까다로운 작업이지만, 의미 정보를 반영하기 위해서 비디오 셧에 존재하는 다양한 컨텍스트 정보를 계산하여 이를 바탕으로 장면 경계를 검출하는 것도 바람직한 방법 중의 하나이다.

본 논문에서는 장면 경계 검출을 위해 컨텍스트 정보를 이용한 방법을 제안한다. 컨텍스트 정보는 비디오 셧에 존재하는 지역 및 전역 컨텍스트 정보로 구분하여 계산한다. 컨텍스트 정보를 바탕으로 유사한 셧들의 연결 작업, 연결 검증 작업 및 조정 작업등의 3단계 과정을 거쳐 장면 경계를 검출한다. 본 논문의 구성은 다음과 같다. 제 2장에서는 비디오 데이타의 장면 검출에 관련된 기존 연구에 대해 기술하며, 제 3장에서는 컨텍스트 정보에 대해 상세히 설명한다. 제 4장에서는 컨텍스트 정보를 이용한 장면 검출 방법에 대해 기술하며, 제 5장에서는 실험 결과를 분석한다.

2. 관련 연구

비디오 데이타에서 장면 경계를 검출하는 연구는 크게 비디오 셧을 이산적으로 취급하는 처리 방식과 연속성을 기반으로 처리하는 방식으로 구분할 수 있다. 이산적인 처리 방식은 비디오 데이타를 연속된 기호 열로 표현하고, 이 기호 열로부터 반복된 패턴을 인식하여 처리하는 방식이다[3, 4]. Yeung et al.은 셧으로 구성된 비디오 기호 열을 장면 전환 그래프를 사용하여 모델링하였다[3]. 장면 전환 그래프로부터 시각 정보의 유사성 및 시간 축 상의 윈도우를 통하여 군집화를 수행하였다. 장면 분할은 전체 그래프에서 닫혀진 서브 그래프를 찾아 이것을 바탕으로 그래프를 분할하여 분할된 곳을 장면 경계로 검출하였다. Hanjalic et al.은 영화에 존재하는 시각 정보의 유사성 및 시간적인 변화를 계산하여 장면 분할을 시도하였다[4]. 셧들간의 유사도를 계산하기 위해 적응적인 임계 함수를 사용하였으며, 장면 경계는 유사한 셧을 연결하는 오버랩 연결 정보로부터 결정하였다.

이러한 이산 처리 방식이 장면 경계 분할에 있어서 좋은 성능을 얻기 위해서는 여러 가지 문제점을 해결해야 한다. 첫째는, 시간 축 상의 윈도우 크기를 정하는 문제이다. 윈도우 크기에 따라 장면 분할의 결과가 달라지므로, 데이타에, 알맞는 윈도우 크기를 정하는 것이 중요하다. 둘째로, 셧의 연관성을 측정하는 과정에 있어서 비디오 셧의 길이나 셧간의 간격 등이 똑 같게 취급되는 문제이다. 셧의 길이나 셧들간의 시간 간격에 따라 의미의 연관성이 실제적으로는 다르기 때문에 이것을 반영하기 어렵다. 끝으로, 계산량이 많이 요구되는 것도 단점이다.

연속적인 처리 방식은 비디오 셧들간의 연관성을 계산하는 데 있어서 셧의 길이나 시간적인 간격을 고려한 방식이다. Kender와 Yeo는 무한 메모리 모델(infinite memory model)을 사용하여 비디오 셧의 연관성을 계산하였다[5]. Sundaram과 Chang은 Kender와 Yeo의 모델을 가역적 FIFO(First-in-First- Out)의 성질을 갖는 단기 메모리 모델(short-term memory model)로 단순화하여 비디오 셧들간의 연관성을 계산하는 모델로 사용하였다[6]. 장면 경계의 검출은 연관성이 지역적으로 최소화되는 점으로 정하였다.

이러한 연속적인 처리 방식에 있어서 문제점은 메모리 버퍼의 크기를 어떻게 정하느냐에 따라 성능이 좌우될 수 있다. 또 하나는 비디오 셧들이 FIFO 규칙에 따라 컨텐트에 상관없이 메모리로부터 순차적으로 제거되는 것보다, 의미 정보에 따라 제거되는 것이 비디오 셧의 연관성을 계산하는데 바람직하다.

따라서, 본 논문은 의미 정보를 반영한 비디오 장면 검출 방식을 제안하는 것으로서, 비디오 셧에 존재하는 컨텍스트 정보를 계산하여, 이를 이산적인 처리 방식에 적용한 장면 경계 검출 방식을 제안한다.

3. 지역 컨텍스트 정보 및 전역 컨텍스트 정보 추출

비디오 데이타에 존재하는 세만틱 정보를 추출하기 위해서 본 논문에서는 비디오 셧에 존재하는 컨텍스트

정보를 이용한다. 비디오 셧을 특징짓는 컨텍스트 정보를 두 가지-지역 컨텍스트 정보(local contextual infomation)와 전역 컨텍스트 정보(global contextual information)-로 구분한다. 지역 컨텍스트 정보는 비디오 셧 자체에 존재하는 컨텍스트 정보라고 정의한다. 예를 들어, 비디오 셧은 전경 객체(foreground object), 배경(background) 및 움직임 정보들로 구성되어 있다. 이러한 요소들의 조합에 의해 다양한 비디오 셧의 컨텍스트를 표현한다. 그림 1은 비디오 셧의 지역 컨텍스트 정보를 보여주고 있다. 비디오 셧으로부터 이 셧을 대표하는 대표 프레임을 선택하고 대표 프레임에서 전경 객체와 배경 및 모션 정보를 추출하여 이 셧에 존재하는 지역 컨텍스트로 표현한다. 전역 컨텍스트 정보는 주어진 비디오 셧이 주위에 존재하는 다른 비디오 셧들과의 관계로부터 발생하는 여러 가지 컨텍스트로 정의한다. 예를 들어, 비디오 셧들간의 유사성을 추출하여 컨텍스트의 연속성을 판단할 수 있고, 비디오 셧들간의 상호 작용으로 부터 특정 비디오 셧들간의 대화 컨텍스트를 검출할 수 있다. 더욱이, 비디오 셧들의 지속 시간에 대한 발생 패턴으로부터 컨텍스트의 진행 템포를 알 수 있다. 그림 2는 전역 컨텍스트 정보를 보여주고 있다. 그림 2(a)에서 비디오 셧은 군집화 과정을 거쳐 유사한 셧들끼리 그룹을 이루어 같은 기호로 치환되는 것을 보여주고 있다. 그림 2(b)는 셧들간의 유사성을 이용하여 시간축 상에서 연결된 셧들의 발생 패턴으로부터 셧들간의 상호 작용을 알 수 있다. 여기서는 셧 "A"와 셧 "C"가 번갈아 가며 발생하므로 상호 작용 관계에 있음을 알 수 있다. 그림 2(c)는 각 셧들의 지속 시간을 보여주고 있다. 셧들의 지속 시간 패턴으로부터 셧의 진행 템포를 알 수 있다. 이러한 정보들은 셧의 전역 컨텍스트로 계산한다. 이 장에서는 이러한 비디오 셧의 지역 및 전역 컨텍스트에 대해 기술한다.

3.1 지역 컨텍스트 정보

비디오 셧의 지역 컨텍스트 정보는 전경 객체 정보, 배경, 그리고 비디오 셧에 관련 된 모션 정보로 정의한다. 즉, 비디오 셧에서 전경 객체 및 배경을 찾고, 전경 객체가 어떠한 움직임을 하였는지 또는 카메라가 어떻게 움직였는지를 계산하여 비디오 셧의 지역 컨텍스트로 정한다(그림 1 참조). 전경 객체는 비디오 셧을 특징짓는 중요한 요소로서, 비디오 셧의 대표 프레임으로부터 모션 및 칼라 정보를 이용하여 반자동적으로 추출한다[8, 9]. 대표 프레임 추출 방식으로 다양한 방법들이 제안되어 있으나[10-12], 본 논문에서는 칼라 정보와 카메라 모션에 따라 대표 프레임을 선택한다[10, 11].

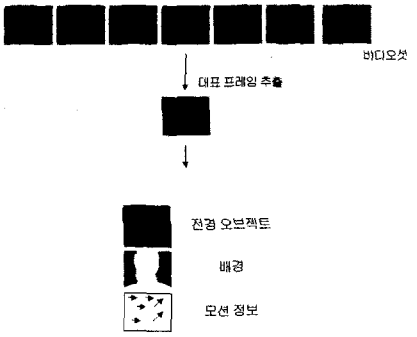
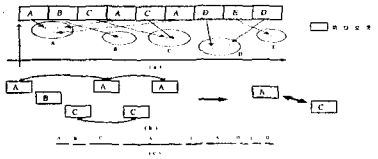
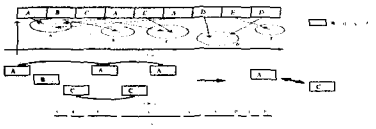


그림 1 비디오 셧의 지역 컨텍스트



(a) 셧들간의 유사성



(b) 셧들간의 상호작용

(c) 셧의 지속 시간의 패턴

그림 2 비디오 셧의 전역 컨텍스트

카메라 모션을 "팬" "틸트", "줌" 및 "카메라모션 없음"으로 구분하고, "팬" 및 "틸트"에서는 첫 번째 프레임을 기준이 되는 대표 프레임으로 선택하고 연속된 프레임들로부터 기준 프레임과의 칼라 히스토그램의 차이를 계산하여 히스토그램 차이가 임계값보다 크면, 해당 프레임을 대표 프레임으로 선택한다. 새로 선택된 대표 프레임을 기준 프레임으로 하여 같은 작업을 셧의 마지막 프레임까지 반복한다. "줌"에서는 처음과 마지막 프레임이 카메라 감독의 의도가 담겨 있는 프레임이므로 대표 프레임으로 선택한다. 카메라 모션이 없는 경우는 프레임 가운데 위치한 영역이 큰 프레임을 대표 프레임으로 선택한다.

대표 프레임으로부터 전경 객체를 추출하기 위해 우선 셧의 카메라 모션을 계산한다. 먼저, 각각의 프레임에서 모션을 계산하고, 이를 바탕으로 셧 전체에 존재하는 모션을 검출한다. 프레임에서 모션을 구하기 위해 영역으로 분할된 프레임 데이타를 사용한다. 왜냐하면, 각각의 프레임은 중요한 영역들로 특징지울 수 있고, 또 이런 영역들에는 보통 같은 방향의 옵티칼 플로우가 존재하고 있으므로[13], 각 영역들의 모션을 계산하여 한 방향으로 향하는 모션을 구하는 것이 프레임 전체 영역을 계산하는 것 보다 간단하고 정확하게 움직임을 계산할 수 있다. 즉, 프레임을 영역으로 분할하기 위해서 모폴로지컬 필터(morphological filter)를 사용하여 영상을 단순화하였다. 단순화된 영상으로부터 편평한 영역(flat region)을 구하고, 이런 편평한 영역으로부터 분수령 알고리즘(watershed algorithm)을 이용하여 영역의 경계를 검출한다[14]. 각각의 프레임에서 분할된 영역 중에서 5개의 커다란 영역을 주 영역(dominant region)으로 정하고, 각각의 주 영역에서 대표 모션을 계산한다. 대표 모션을 구하기 위해 옵티칼 플로우 벡터를 8개의 방향으로 양자화하여, 각 영역에서 8개 방향에 대한 옵티칼 플로우 벡터의 개수를 계산한다. 만약 방향 θ를 갖는 옵티칼 플로우 벡터의 개수가 전체 옵티칼 플로우 벡터의 개수의 합의 반 이상이 될 경우, 이 방향(θ)을 영역의 대표 모션 벡터의 방향(Motion Phase)으로 간주한다. 또, 모션 벡터의 크기(Motion Intensity)는 대표 모션 벡터와 같은 방향의 옵티칼 플로우 벡터의 크기를 평균하여 계산한다. 각 영역의 모션 벡터 방향중 특정 방향으로 향하는 벡터의 수가 다른 방향의 모션 개수보다 적어도 두 배이상 클 때, "팬"이나 "틸트"로 분류한다. 이런 방향을 정할 수 없을 때, "줌"을 테스트한다[11, 14]. 카메라 모션이 있는 경우에는 영역으로 분할된 대표 프레임으로부터 가운데 존재하고 있는 영역을 전경 객체로 선택한다. 카메라 모션이 없는 경우에는 움직이는 물체가 전경 객체가 될 가능성이 많으므로, 움직이는 물체를 찾는다. 일반적으로 움직이는 물체의 밝기 변화는 배경보다 변화가 심하므로 통계학적인 가정을 바탕으로 변화 마스크를 이용하여 추출한다[8]. 정확한 전경 객체 추출은 매우 어렵기 때문에 실험에서는 추출된 전경 객체에 대해서 사용자와의 상호 작용을 통한 바람직한 전경 객체를 추출할 수 있도록 반자동적인 전경 객체 추출 방식을 사용한다[9]. 이러한 반자동적인 추출은 유용한 지역 컨텍스트 정보를 구하기 위한 목적에 필요한 작업이다. 추출된 전경 객체는 비디오 셧을 대표하고, 또 비디오 셧들간의 유사성을 계산하는 데 사용한다.

배경 정보는 비디오 셧에서 전경 객체가 존재하는 여러 가지 환경에 관한 정보를 가지고 있다. 특히, 전경 객체의 크기가 대표 프레임에서 차지하는 비중이 작을 경우 배경 정보는 셧을 특징짓는 중요한 요소가 되기 때문에, 본 논문에서는 전경 객체의 크기가 20% 이하일 경우 전경 객체를 무시하고 배경으로 처리한다. 배경 정보로는 칼라 히스토그램 분포도 뿐만 아니라 배경 칼라가 주는 느낌이나 칼라 콘트라스트(color contrast)를 계산한다. 본 논문에서는 셧을 대표하는 대표 프레임에서 주된 칼라(dominant color)를 구한 다음, 주된 칼라가 빨강, 노랑, 오렌지 색 계열이라면 따뜻한 느낌을 주는 셧으로 해석하고, 파랑 계열의 색이라면 차가운 느낌을 주는 셧으로 배경 칼라를 해석한다[16]. 이러한 주된 칼라의 배열이 검정과 하얀색 계통으로 구성되어 있거나 따뜻한 색과 차가운 색으로 구성되어 있을 경우 콘트라스트가 있다고 계산한다[16].

비디오 셧의 움직임은 카메라 모션과 객체 모션으로 구성되어 있으며, 이러한 비디오 셧의 움직임도 셧의 컨텍스트를 나타내는 중요한 요소 중의 하나이다. 카메라 모션은 옵티칼 플로우 방식을 이용하여 "팬", "틸트", "줌" 및 "카메라 모션 없음"의 4가지 방식으로 구분한다. 연속된 프레임의 같은 방향의 모션으로부터 셧의 모션을 검출하고, 객체 모션은 전경 객체 영역의 모션을 사용한다.

3.2 전역 컨텍스트 정보

비디오 셧의 전역 컨텍스트는 비디오 셧의 주위 환경이나 다른 비디오 셧들과의 관계에서 발생하는 정보로 정의한다. 본 논문에서는 전역 컨텍스트로 셧들간의 상호 작용, 셧들간의 연관성 및 연속된 셧들의 지속 시간에 대한 패턴 등으로 표현한다. 그림 2는 전역 컨텍스트를 보여주고 있다. 셧들간의 연관성은 유사한 셧들이 지역적으로 그룹을 형성하는 컨텍스트를 나타내며, 셧들간의 상호 작용은 여러 셧들 중에서 두 개 또는 그 이상의 셧들이 반복되어 나타나는 컨텍스트를 뜻하고, 셧의 지속 시간에 대한 패턴은 시간 축 상에서 셧의 템포를 의미한다.

이러한 요소들은 전체적인 비디오 데이타에서 각 셧들로부터 발생하는 다양한 컨텍스트를 계산하는데 여러 가지 유용한 점들을 가지고 있다. 첫째로, 비디오 셧들간의 연관성은 비디오 데이타 흐름에 있어서 셧들간의 유사한 그룹을 찾는 것으로서, 적절한 유사도를 측정하는 것이 중요하다. 셧들 사이의 유사도는 셧이 가지고 있는 지역 컨텍스트의 유사도로 계산할 수 있다. 즉, 전

경 객체들 사이의 유사도, 배경의 유사도 및 셧에 존재하는 움직임의 유사도를 이용하여 셧들간의 연관성을 추출한다. 전경 객체의 유사도는 같은 객체를 갖는 그룹을 찾을 수 있어 유용하고, 배경의 유사도에 의한 연관성은 유사한 장소 또는 분위기 속에서의 이야기 전개를 감지할 수 있으며, 셧에 존재하는 카메라 움직임의 유사도는 이야기 전개 방식에 대한 정보를 추출할 수 있다.

둘째로, 비디오 셧들간의 상호 작용은 비디오 데이타에서 대화 장면들을 추출하는데 이용된다. 일반적으로 대화 장면들은 비디오에 있어서 중요한 메시지를 표현하는 방식으로 정확한 추출이 요구되어 진다. 대화 장면 검출을 위해 여러 가지 방법이 제안되어 있지만[3, 7], 본 논문에서는 동시 발생 패턴(co-occurrence pattern)을 계산하여 검출한다. 먼저, 비디오 셧을 군집화 과정을 통해 기호로 치환하고, 기호로 표시된 비디오 셧의 시퀀스로부터 각 기호의 발생 빈도를 조사한다. 다음에 가장 발생 빈도수가 큰 기호로부터 동시 발생 패턴 사이의 변위 길이가 d를 갖는 동시 발생 패턴 $P(i,j)$를 계산한다. 여기서 i 및 j는 기호 열에서 동시에 발생하는 기호의 쌍이고, d는 두 기호 사이의 거리로서 두 기호 사이에 존재하는 셧의 개수이다. 본 논문에서는 동시 발생 패턴을 계산하는 데 있어서 변위 길이 d를 0 또는 1로 정한다. 예를 들어, 비디오 데이타에서 군집화 과정을 거쳐 기호로 치환한 기호 열이 그림 3처럼 "ABCA CADED"인 경우, 빈도수가 가장 큰 기호는 "A"이다. 따라서, 셧 "A"를 기준으로 동시 발생 패턴 $P(i,j)$를 계산한다. 여기서 i는 "A"이고, j는 "B", "C", "D" 또는 "E" 이다. 한 쌍의 기호 패턴 사이의 거리 d가 0 일 경우에 가능한 동시 발생 패턴은 "AB", "AC", "AD" 또는 "AE"이다. d 가 1일 경우 가능한 동시 발생 패턴은 "AXB", "AXC"... 등 다양하다. 여기서 기호 "X"는 동시 발생 패턴을 계산할 때 무시해도 되는 셧의 기호를 의미한다. 그림 3에서 보면 가장 빈도수가 높은 동시 발생 패턴은 "AC" 또는 "AXC"로서 빈도수가 2 이다. 따라서, 두 셧 "A"와 "C" 사이에 상호 작용이 있다는 것을 알 수 있다.

끝으로, 비디오 셧의 지속 시간에 관한 정보도 셧의 전체 흐름에 관한 컨텍스트를 나타내는 유용한 정보이다. 비디오 셧은 평균 셧의 길이와 비교하여, 평균 길이의 1.5배보다 클 때 긴 셧(long shot), 평균 길이의 반보다 작을 때 짧은 셧(short shot) 및 나머지를 중간 셧으로 구분한다. 이러한 비디오 셧의 길이 패턴은 비디오 셧의 템포를 나타내는 것으로, 일반적으로 비디오 전체에 있어서 전달하려는 메시지의 강도를 나타내고 있다. 예를 들어, 짧은 셧의 반복은 긴장이나 두려움 등의 감정을 나타낼 수 있고, 긴 셧은 일반적으로 편안한 감을 표현할 때 사용된다.

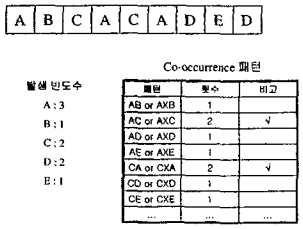


그림 3 비디오 셧의 상호작용 추출 방법

4. 컨텍스트 정보를 이용한 장면 분할

장면이란 의미적인 일관성을 갖는 단위로서 비디오 데이타를 구조적으로 해석할 때 매우 중요한 요소이다. 이 장에서는 3장에서 기술한 컨텍스트 정보를 이용하여 비디오 데이타의 장면 분할 방법을 설명한다. 본 논문에서 제안한 장면 분할 방법은 이산적인 방식으로서 세 가지 모듈에 의해 수행된다. 첫째, 셧들간의 연관성을 바탕으로 한 연결 작업, 둘째, 셧들간의 컨텍스트 연속성을 고려한 연결 검증 작업, 마지막으로 연결되지 않은 셧들에 대한 조정 작업등이다.

4.1 연결 작업

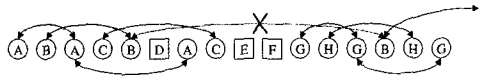
유사한 셧들 간의 연결 작업은 비디오 데이타에서 연관성이 있는 셧들의 그룹을 추출하는 것이다. 유사한 셧들의 연결을 위해, 주어진 셧들의 대표 프레임으로부터 지역 컨텍스트를 계산한다. 다음에, 계산된 지역 컨텍스트를 바탕으로 셧들 간의 유사성을 검출한다. 셧들간의 유사성을 측정하기 위해 본 논문에서는 전경 객체와 배경의 정보를 분리하여, 각각의 유사성을 선형 결합하여 다음과 같이 계산한다.

$$\begin{aligned}\text{Similarity}(A,B) &= 1-\text{Dissim}(A,B)\\ &= 1-(w_1 * \text{Fg\_dissim}(A,B)+w_2\\ &\quad * \text{Bg\_dissim}(A,B)) \qquad (1)\end{aligned}$$

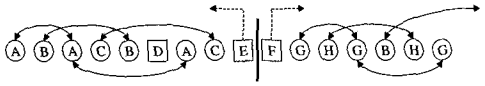
여기서, Fg_dissim(A,B)는 두 셧 A 및 B 사이의 전경 객체간의 비 유사도이고, Bg_dissim(A,B)는 두 셧 A 및 B의 배경에 대한 비 유사도이며, w_1과 w_2는 유사도 계산에 있어서 전경 객체와 배경에 대한 기여도를 조정하는 가중치이다. 일반적으로 전경 객체의 유사도가

배경의 유사도보다 중요하므로, w_1을 w_2보다 크게 정한다. 비 유사도는 칼라 히스토그램의 차이의 제곱으로 계산한다. 인간의 지각에 부합되는 칼라 히스토그램 차이를 계산하기 위해 RGB 칼라 스페이스에서 HSV 칼라 스페이스로 변환한 다음 11개의 컬쳐 칼라(culture color)로 변환한다[17, 18]. 즉, 칼라 정보를 HSV의 값에 따라 빨강, 노랑, 연두색, 파랑색, 갈색, 보라색, 분홍색, 오렌지색, 회색, 검정색 및 흰색의 11개 칼라로 매핑한다. 다음에 전경 객체 및 배경에서 각각의 칼라의 히스토그램 차이를 계산한다. 전경 객체 및 배경의 크기는 프레임마다 서로 다르므로 이들을 각각의 크기로 나누어 정규화한 값을 사용하여 유사도를 계산한다. 두 셧들의 대표 프레임에서 유사도가 임계값 이상이면 두 셧은 유사하다고 연결한다. 이러한 연결 작업을 마지막 셧까지 수행하여, 그림 4(a)와 같은 유사한 셧들 간의 연결을 얻는다.

(a) 연결 작업



(b) 연결 검증 작업

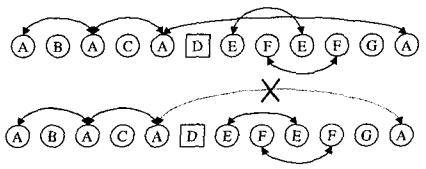


(c) 조정 작업

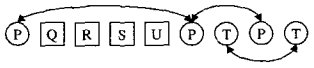
그림 4 장면 분할 작업

4.2 연결 검증 작업

연결 모듈에서 유사한 셧들간의 연결은 단지 셧들간의 유사성에 관련 된 컨텍스트만을 고려한 연결이므로, 바람직하지 않은 연결들이 발생할 수 있다. 그림 4(b)에서 보면 두 번째 셧 "B"와 세 번째 셧 "B"의 연결은 이 연결에 포함된 셧들의 개수가 많고, 이 연결을 해지하였을 경우 배경으로 이루어진 셧들을 경계로 유사성 연결이 분리되는 컨텍스트를 갖고 있다는 것을 알 수 있다. 이러한 경우 연결을 해지하는 것이다. 예를 들어, 그림 5(a)와 같은 경우 세 번째 셧 "A"와 네 번째 셧 "A"의 연결은 두 셧들 사이에 많은 셧들이 포함되어 있어서, 이 연결을 해지하였을 경우 연결이 분리되는 것을 알 수 있으므로, 연결을 해지한다. 하지만, 그림 5(b)와 같이 첫 번째 셧 "P"와 두 번째 셧 "P" 사이에 많은 셧들이 포함되어 있어도 포함된 셧들이 단순히 배경 셧들일 경우, 이 비디오 셧들의 컨텍스트는 전경 객체의 연속된 흐름이다. 이러한 경우 연결을 해지하지 않는 것이 더 바람직하다. 단순한 유사도에 의한 연결로부터 전역 컨텍스트를 생각하여 잘못된 연결을 해지하는 방식은 유사한 셧들 간의 지역성을 고려하기 위해 시간 축 상의 윈도우를 이용하는 방식보다 더 효과적이다[3, 4]. 왜냐하면, 시간 축 상의 윈도우를 이용하면 단지 셧들간의 컨텍스트에 관계없이 윈도우 안에 위치한 셧들간의 유사도만 고려하기 때문에, 윈도우 크기에 따라 성능이 차이가 나는 단점이 있다. 하지만, 본 논문에서는 연결된 셧들 사이에서 컨텍스트를 계산하므로 보다 적절하게 셧들간의 연결을 해지할 수 있다.



(a)



(b)

그림 5 비디오 셧 연결의 검증 작업

4.3 조정 작업

연결 검증 작업에서 전역 컨텍스트 정보에 의해 잘못 연결된 셧들의 연결이 해지되면, 연결되지 않은 셧들이 남을 수 있는데 이들 셧들을 주위의 유사한 셧들의 그룹에 연결함으로써 장면 경계를 최종적으로 검출한다. 일반적으로 배경으로 이루어진 셧들이 남는 경우가 많은데 배경의 칼라 히스토그램의 차이를 이용한 유사도를 계산하여 연결하는 작업을 수행한다. 그림 4(c)는 연결되지 않은 셧 "E" 및 "F"를 주위의 셧에 연결한 것을 보여준다. 단, 남겨진 셧들이 이웃한 셧들과의 연결 상태가 서로 엇갈리게 되는 경우에는 유사도가 큰 쪽으로 연결을 수행한다. 한편, 배경 칼라들의 유사도가 양쪽 그룹에 대해 20% 이하 정도로 매우 낮을 경우에는 배경 칼라의 느낌을 계산한다. 즉, 배경 칼라가 빨강, 노랑, 오렌지 색들의 따뜻한 계열의 색으로 구성되어 있는지, 아니면 파란색 계열의 차가운 색으로 구성되어 있는지를 계산하여 유사한 분위기를 나타내는 쪽으로 연결

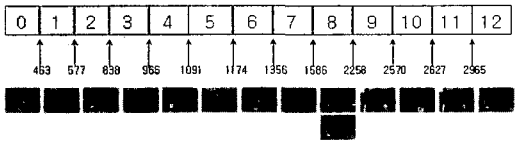
을 수행한다. 조정 작업을 수행한 후, 그림 4(c)처럼 장면 경계를 추출한다.

5. 실험 결과

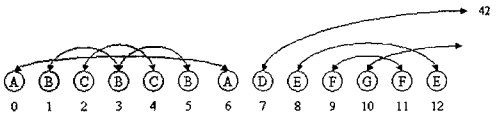
제안된 장면 경계 검출 방법을 실험하기 위해 20분짜리 TV 드라마 2편과 25분 분량의 영화 데이타 2편을 이용하였다. 먼저, MPEG 비디오 데이타로부터 칼라 히스토그램의 변화를 이용하여 셋을 구분하였다[1, 11]. 각각의 셋으로부터 칼라 정보와 모션 정보를 이용하여 3.1절에서 기술한 방식에 따라 대표 프레임을 선택하였다. 대표 프레임으로부터 전경 객체 및 배경을 분리하기 위해 각 셋에 존재하는 카메라 모션을 옵티칼 플로우를 이용하여 계산하였다. 모션 방향은 8개로 양자화 하였으며, 10% 이하의 작은 모션 크기(motion intensity)를 갖는 모션은 무시하였다. 구해진 각 프레임의 모션 정보로부터 셋의 모션 정보를 계산하였다. 실험 결과 연속된 3장의 프레임이 잘못 검출될 확률이 매우 적으므로, 5장의 연속된 프레임으로부터 셋의 모션을 검출하였다. 구해진 카메라 모션은 "팬", "틸트", "줌" 및 "카메라 모션 없음"으로 구분하였으며, 각각의 모션에 대해 전경 객체와 배경을 사용자와의 상호작용을 통해 반자동적으로 분리하였다. 구해진 전경 객체, 배경 및 움직임 정보 등으로부터 셋의 지역 컨텍스트를 계산하였으며, 비디오 셋들 간의 군집화 과정을 통해 셋들간의 유사성 및 상호 작용등의 전역 컨텍스트를 계산하였다.

컨텍스트 정보를 바탕으로 장면 경계를 검출하기 위해 연결 모듈, 연결 검증 모듈 및 조정 모듈등의 3단계 과정을 수행하였다. 연결 모듈에서는 셋들간의 유사성을 구하기 위해 식(1)을 이용하여 계산하였다. 본 논문에서는 실험을 통해 식 (1)의 가중치 w_1과 w_2를 3과 1로 각각 정하였다. 전경 객체의 가중치가 배경에 비해 큰 이유는 셋의 유사도가 전경 객체에 많이 좌우되기 때문이다. 이러한 현상은 전경 객체가 화면의 많은 부분을 차지하는 드라마나 영화에서 뚜렷하게 나타나고 있다. 전경 객체 크기가 프레임 전체 크기의 20% 미만일 때는 전경 객체를 무시하고 배경 프레임으로 간주하였다. 그림 6은 지역 컨텍스트의 유사성에 의해 연결된 상태를 보여 주고 있다. 이때 유사도는 70% 이상일 때 바람직한 결과를 얻었다. 유사도 연결 과정에서 바람직하지 않은 유사도의 연결이 발생하므로 정확한 장면의 경계를 검출하기 위해서는 연결 검증 작업이 필요하다.

연결 검증 모듈에서는 연결된 비디오 셋들 사이에 포함된 다른 셋들의 개수가 많을 때 단순히 연결을 해지하는 것이 아니라, 컨텍스트 정보를 이용하여 연결의 타당성을 검증하였다. 본 논문에서는 연결된 두 셋 사이에 존재하는 다른 셋들의 개수가 4개 이상일 때, 연결을 해지하는 것이 타당한 지를 계산하였다. 즉, 분리된 유사성 연결이 나타나는 지, 또는 포함된 셋들이 단순히 배경 셋들로만 구성되어 있는 지를 계산하여 연결 해지를 판단하였다. 그림 7(a)는 그림 6의 유사도 연결 작업의 결과에서 셋 "D"의 연결이 셋 번호 7번과 셋 번호 42번 사이의 연결이어서, 두 연결된 셋 사이에 존재하는 다른 셋들의 개수가 4개보다 많고, 두 셋의 연결을 해지하였을 경우 서로 분리된 유사성 연결이 나타나므로 연결을 해지하였다.

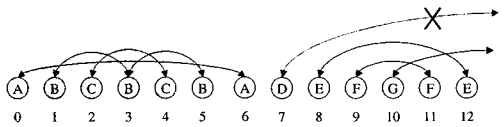


(a) 비디오 데이타

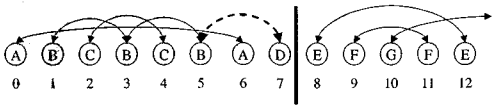


(b) 유사한 셋의 연결도

그림 6 연결된 비디오 셋



(a) 잘못된 연결 해지 작업



(b) 조정 작업

그림 7 비디오 장면 분할

끝으로 조정작업을 수행하였다. 그림 7(a)와 같이 연결을 해지하였을 경우 연결이 되지 않은 셋들은 양쪽의 그룹에 유사성을 계산하여 연결을 수행한 후 장면 경계를 구한다. 그림 7(b)에서는 조정 작업이 필요한 셋은 "D"이다. 셋 "D"에 대해 양쪽 그룹에 대한 유사도를 계산하였다. 여기서는 셋 "D"가 셋 "B"와 유사하므로 서

로 연결하고 셧 "D"와 셧 "E" 사이를 장면 경계로 분할하였다. 본 논문에서 제안한 방식으로 얻은 결과 중 1 단계 장면 경계 검출 결과는 표 1에 나타나 있다. 표 1에서 N은 바람직한 장면 개수이고, D(Detected) 는 검출된 장면 개수이다. M(Misses)은 장면 경계인데 검출이 되지 않은 횟수를 나타내고, F(False alarms)는 장면 경계가 아닌데 장면 경계로 검출 된 횟수이다. Recall과 Precision은 다음과 같이 구한다.

$$Recall = \frac{D}{D+M} \quad (2)$$

$$Precision = \frac{D}{D+F} \quad (3)$$

연결 검증 작업 후 조정 작업을 통해 보다 향상된 결과를 얻었다. 표 2는 제안된 방식의 최종 검출 결과이다. 검출 정확도는 80% 이상으로 바람직한 결과를 얻을 수 있었다. 또, 표 3은 본 논문에서 제안한 방식과 기존의 Yeung et al.에 의해 제안된 시각 정보의 유사성 및 시간 축 상의 윈도우를 이용한 이산적인 방식과 비교 결과이다[3]. 본 논문에서 제안한 방식이 약간 우수한 결과를 얻었는데, 이는 비디오에 존재하는 지역 및 전역 컨텍스트 정보를 장면 검출에 반영하였기 때문이다. 더욱이, 각 셧에 대한 다양한 컨텍스트 정보를 알고 있기 때문에 다양한 검색 및 인덱싱이 가능하게 되었다.

6. 결 론

본 논문에서는 비디오 셧에 존재하는 컨텍스트 정보를 바탕으로 장면 경계 검출 방법을 제안하였다. 비디오 셧의 컨텍스트 정보로 지역 컨텍스트 및 전역 컨텍스트 정보를 사용하였다. 비디오 셧의 지역 컨텍스트는 전경 객체 정보, 배경, 그리고 비디오 셧에 관련 된 모션 정보로 정의하였고, 셧의 전역 컨텍스트는 비디오 셧의 주위 환경이나 다른 비디오 셧들과의 관계에서 발생하는 정보로 셧들간의 상호 작용, 셧들간의 연관성 및 연속된 셧들의 길이에 대한 패턴등으로 표현하였다. 이러한 컨텍스트 정보를 바탕으로 장면 경계는 3단계 과정으로 수행하였다. 먼저, 연결 과정에서는 유사도를 바탕으로 셧들간의 연결을 수행하였고, 연결 검증 과정에서는 바람직하지 않은 연결을 컨텍스트 정보를 바탕으로 해지하였으며 조정 과정을 거쳐 최종적으로 장면 경계를 검출하였다. 컨텍스트 정보를 이용한 장면 검출 방식을 TV 드라마 및 영화에 적용하여 80% 이상의 검출 정확도를 얻었다.

좀 더 바람직한 장면 경계 검출 결과를 얻기 위해서는 시각 정보이외에도 오디오 정보 및 텍스트 정보를 결합하는 것이 필요하다. 아울러, 비디오에 존재하는 다양한 의미를 적절한 학습 방식을 이용하여 구현한다면,

표 1 제 1단계 장면 검출 결과

비디오데이타	N (장면개수)	D (Detected)	M (Misses)	F (False alarms)	Recall	Precision
드라마 A (20분: 사랑은 아무나 하나)	11	8	3	2	0.73	0.80
드라마 B (18분: 비밀)	14	10	4	3	0.71	0.77
영화 A (28분: Basic Instinct)	18	12	6	3	0.67	0.80
영화 B (29분: Autumn in New York)	16	11	5	5	0.69	0.68

표 2 최종 장면 검출 결과

비디오데이타	N (장면개수)	D (Detected)	M (Misses)	F (False alarms)	Recall	Precision
드라마 A (20분: 사랑은 아무나 하나)	11	9	2	2	0.81	0.81
드라마 B (18분: 비밀)	14	11	3	2	0.78	0.84
영화 A (28분: Basic Instinct)	18	14	4	3	0.77	0.82
영화 B (29분: Autumn in New York)	16	12	4	3	0.75	0.80

표 3 기존 방식과의 비교

비디오데이타	기존의 이산적인 처리 방식[3] Recall	기존의 이산적인 처리 방식[3] Precision	제안한 방식 Recall	제안한 방식 Precision
드라마 A (사랑은 아무나 하나)	0.73	0.80	0.81	0.81
영화 B (Autumn in New York)	0.69	0.73	0.75	0.80

보다 바람직한 내용 기반의 비디오 인덱싱 및 검색이 가능한 시스템의 구현이 가능할 것이다.

참 고 문 헌

[1] W. Grosky, R. Jain and R. Mehrotra, *The Handbook of Multimedia Information Management,* Prentice Hall PTR, 1997.

[2] S. Chang and H. Sundaram "Structural and Semantic Analysis of Video," *Proc. ICME'00,* Aug. 2000.

[3] M. Yeung, B. Yeo and B. Liu, "Segmentation of Video by Clustering and Graph Analysis," *Computer Vision and Image Understanding,* Vol. 71, No. 1, pp. 94-109, 1998.

[4] A. Hanjalic, R. Lagendijk and J. Biemond, "Automated High-Level Movie Segmentation for Advanced Video-Retrieval Systems," *IEEE Trans. Cir. and Sys. for Video Tech.*, Vol. 9, No. 4, pp. 580-588, June 1999.

[6] J. Kender and B. Yeo, "Video Scene Segmentation Via Continuous Video Coherence," *Proc. CVPR'98,* June 1998.

[7] H. Sundaram and S. Chang, "Determining Computable Scenes in Films and their Structures using Audio-Visual Memory Models," *Proc. ACM Multimedia'00,* 2000.

[8] M. Kim, J. Choi, D. Kim, H. Lee, C. Ahn and Y. Ho, "A VOP Generation Tool: Automatic Segmentation of Moving Objects in Image Sequences Based on Spatio-Temporal Information," *IEEE Trans. Cir. Sys. for Video Tech.*, Vol. 9, No. 8, pp. 1216-1226, Dec. 1999.

[9] S. Cooray, N. O'Connor, S. Marlow, N. Murphy, and T. Curran, "Hierarchical Semi-Automatic Video Object Segmentation for Multimedia Applications," *Proc. SPIE Internet Multimedia Management Systems II,* pp.10-19, 2001.

[10] H. Zhang, J. Wu, D. Zhong and S. Smoliar, "An Integrated System for Content-based Video Retrieval and Browsing," *Pattern Recognition,* 30(4), pp. 643-658, 1997.

[11] 강행봉, "비디오 셧으로부터 영역, 모션 및 퍼지 이론을 이용한 계층적 대표 프레임 선택", 정보과학회 논문지, 제 27권 5호, pp. 510-520, 2000.

[12] W. Wolf, "Key Frame Selection by Motion Analysis," *Proc. ICASSP' 96,* pp. 1228-1231, 1996.

[13] B. Lucas and T. Kanade, "An Iterative Techinque of Image Registration and Its Application to Stereo," *Proc. IJACI,* pp. 674-679, 1981.

[14] L. Vincent and P. Soille, "Watersheds in Digital Spaces: An Efficient Algorithms based on Immersion Simulation." *IEEE Trans. PAMI,* Vol. 13, No. 6, pp. 583-598, Jun. 1991.

[15] V. Kobla and D. Doermann, "Compressed domain video indexing techniques using DCT and motion vector information in MPEG video," *Proc. of SPIE,* 1997.

[16] J. Corridoni, A. Bimbo, and P. Pala, "Image Retrieval by Color Semantics," *ACM Multimedia Systems Journal,* Vol. 7, No. 5, pp. 359-368, Sept. 1999.

[17] E. Goldstein, *Sensation and perception,* Brooks/Cole, 1999.

[18] E. Chang, B. Li and C. Li, "Toward Perception-Based Image Retrieval," *Proc. IEEE Workshop on Content-Based Access of Image and Video Libraries,* pp. 101-105, Jun. 2000.

강 행 봉

1980년 한양대학교 전자공학과 졸업(학사). 1986년 한양대학교 대학원 전자공학과 졸업(석사). 1989년 미국 Ohio State University 컴퓨터공학(석사). 1993년 미국 Rensselaer Polytechnic Institute 컴퓨터공학 박사. 1994년 ~ 1997년 삼성종합기술원 수석연구원. 1997년 ~ 현재 가톨릭대학교 컴퓨터 전자공학부 부교수. 관심분야는 컴퓨터 비젼, 멀티미디어 시스템, 인공지능, 생체인식 및 Bioinformatics 등