論文
2-11-11

I 프레임에 기반한 MPEG 압축영상에서의 자막 탐지 (Localization of captions in MPEG compression images based on I frame)

유 태 웅*
(Tae-Woong Yoo)

요 약

실시간 자막 탐지는 비디오 인덱싱, 텍스트의 이해, 자동 자막 탐지시스템 등 수많은 응용 분야에서 요구된다. 본 논문은 I 프레임을 기반으로 MPEG 압축 동영상에서 자막을 탐지하는 알고리즘을 제안한다. 제안한 알고리즘은 자막 텍스춰 정보와 색체 정보를 사용하여 배경 영상으로부터 자막을 정확히 분리한다. 기존 알고리즘들은 압축 동영상으로부터 텍스트 영역을 추출하기 전에 압축을 먼저 해제하는데 제안한 알고리즘은 DCT 압축 도메인에서 직접 자막 텍스트 영역을 탐지한다.

ABSTRACT

For the applications like video indexing, text understanding, and automatic captions localization system, real-time localization of captions is an essential task. This paper presents a algorithm for localization of captions in MPEG compression images based on I frame. In this algorithm, caption text regions are segmented from background images using their distinguishing texture characteristics and chrominance information. Unlike previously published algorithms which fully decompress the video sequence before extracting the text regions, this algorithm locates candidate caption text region directly in the DCT compressed domain.

1. 서론

최근 들어 동영상을 대상으로 자막 탐지 및 인식, 자동 주석 생성 등 자막 탐지 및 추출에 관한 연구가 활발히 이루어지고 있다. 동영상의 자막은 음성과 더불어 동영상의 프레임들을 요약하는 수단이며 동시에 사용자에게 중요한 정보의 주제를 전달할 수 있어 동영상에서 중요한 요인이다. 또한, 비디오 및 영상 내의 문자 데이터는 영상 데이터베이스 인덱싱과 문서 이해 등과 같은 수많은 응용에서 중요하다.

이러한 자막 탐지 문제는 자막 내 문자들의 폰트, 크기, 스페이싱, 정렬, 방향, 컬러, 그리고 텍스춰의 다양성 때문에 매우 어려운 문제이다. 더욱이, 문자들은 종종 영상의 복잡한 배경에 내포되어있다.

* 정회원 : 서해대학 컴퓨터정보기술계열 교수

논문접수 : 2001. 11. 7.
심사완료 : 2001. 11. 15.

자막 추출에 관한 연구는 다음과 같이 분류될 수 있다. 압축된(compressed) 비디오를 대상으로 하는 방법과 비압축(uncompressed) 비디오를 대상으로 하는 방법, 그리고 텍스춰화된 물체(textured object)로서 자막을 간주하는 텍스춰 기반의 분석 방법과 연결 요소 분석(connected component analysis)을 사용하는 방법으로 크게 분류할 수 있다.

먼저 텍스춰 기반 분석 방법은 자막 영역 문자들의 텍스춰 성질을 이용하는 방법으로 Gabor 필터, 웨이블릿(wevelet), 공간 편차(spatial variance) 등을 사용하여 문자 영역 위치를 탐지한다. 텍스춰 기반 분석 방법은 문자의 폰트, 크기 등에 민감한 영향을 받으며, 일반적으로 처리 속도 측면과 정확성 측면에서 비효율적이다. 때문에 텍스춰 필터 디자인에 필요한 노력을 줄이기 위해 자동생성에 관한 연구가 수행 중이다.

연결 요소 분석 방법은 상향식 방법으로 작은 영역에서 점차 큰 영역으로 합쳐 문자 영역과 비문자 영역으로 분리하는 방법으로 구현 측면에서 용이한 반면, 문자 크기와 문자간의 거리등에 대한 사전지식이 필요하다. 또한 높은 처리 속도와 정확한 문자 위치 탐지가 가능하지만 이진 영상(binary image)에서만 사용 가능하다.

위 대부분의 기존 연구들은 문자들은 한정된 크기를 갖으며, 한 텍스트 라인은 수평으로 배열된 문자들의 클러스터, 그리고 텍스트는 배경으로부터 비교되는 대비를 갖는다는 텍스트의 특성 및 사전 지식을 기반으로 비압축 비디오 및 영상을 대상으로 실험한 연구 결과들이다. 현재뿐만 아니라 앞으로는 멀티미디어 데이터의 사용이 광범위하게 사용될 것이며, 특히 압축된 형태의 정지 영상 및 동영상 데이터가 많이 사용될 것이다. 대부분의 동영상은 압축된 형태로 저장되고 전송되므로 압축된 상태에서 자막 탐지 및 인식을 할 수 있는 방법에 대한 연구가 이루어져야한다. 또한, 압축된 비디오를 대상으로 비디오 내의 자막의 크기에 무관한 텍스춰 필터 설계, 그리고 실시간에 비디오 자막을 검출 할 수 있는 알고리즘이 개발되어야 한다. 그러므로 본 연구에서는 웹 검색, 컬러 영상 인덱싱, 영상 데이터베이스 구성, 자동 주석, 그리고 비디오 인덱싱 등에서 중요한 역할을 하는 자막을 실시간 자동 탐지하는 알고리즘에 대한 연구를 목적으로 한다. 특히, 압축된 비디오를 대상으로 효율적인 동영상 분석을 위하여 I 프레임에 기반한 MPEG 압축영상 내 자막(caption)의 위치를 자동 탐지하는 알고리즘을 제안한다.

텔레비전 프로그램(뉴스 방송, 광고), 영화 등의 비디오 클립을 대상으로 펜티엄 800MHz 중앙처리장치, 256MB의 주기억장치 등을 가지고 실험한다. 효율적인 비디오 자막 탐지 알고리즘의 구현 프로그래밍 언어로는 웹기반 프로그램으로의 확장성을 고려하여 자바(Java)를 사용한다.

본 논문 내용은 다음과 같다. 먼저 2 장에서는 자막 탐지 및 인식에 관련된 기존 알고리즘들을 설명하고, 3 장에서는 압축 동영상으로 많이 사용되는 MPEG 형태의 동영상에 대하여 간략하게 기술한다. 그리고 4 장에서는 본 연구에서 제안한 자막 탐지 방법에 대하여 실험 결과와 함께 기술한다. 마지막으로 5 장에서 결론 및 향후 연구과제에 대하여 기술한다.

2. 관련 연구

기존 자막 탐지에 관한 연구는 텍스춰화된 물체로서 자막을 간주하는 텍스춰 기반의 분석 방법과 상향식 방법으로 작은 영역으로부터 점차 큰 영역으로 합쳐가면서 문자영역과 비문자 영역으로 나누는 연결 요소 분석 방법을 사용하는 방법으로 크게 분류할 수 있다. 먼저 텍스춰 기반 분석 방법은 자막 영역 문자들의 텍스춰 성질을 이용하는 방법으로 Gabor 필터, 웨이블릿(wevelet), 공간 편차(spatial variance) 등을 사용하여 문자 영역 위치를 탐지한다. 텍스춰 기반 분석 방법은 문자의 폰트, 크기 등에 민감한 영향을 받으며, 일반적으로 처리 속도 측면과 정확성 측면에서 비효율적이다. 때문에 텍스춰 필터 디자인에 필요한 노력을 줄이기 위해 자동생성에 관한 연구가 수행 중이다. 연결 요소 분석 방법은 상향식 방법으로 작은 영역에서 점차 큰 영역으로 합쳐 문자 영역과 비문자 영역으로 분리하는 방법으로 구현 측면에서 용이한 반면, 문자 크기와 문자간의 거리등에 대한 사전지식이 필요하다. 또한 높은 처리 속도와 정확한 문자 위치 탐지가 가능하지만 이진 영상(binary image)에서만 사용 가능하다.

[1]은 뉴스 비디오를 대상으로 낮은 해상도 문자와 복잡한 배경에 대한 문제 제기와 이러한 문제를 해결하기 위하여 보간 필터(interpolation filter), 다중 프레임 통합(multi-frame integration)과 4가지 필터의 조합을 적용하여 자막 탐지 및 세그멘테이션 방법을 제안하였다. [2]는 복잡한 컬러 영상으로부터 텍스트 위치를 자동으로 탐지하기 위하여 두 방법을 제안하였다. 먼저 동일한 컬러를 가진 부분을 세그멘테이션하고 텍스트를 포함한 영역을 선택하기 위하여 크기, 정렬, 근접성 등의 휴리스틱을 사용한다. 두 번째는 명암 영상을 대상으로 국부 공간 편차(horizontal spatial variation)를 계산하여 높은 편차를 나타내는 영역을 텍스트 위치로 탐지한다. 즉, 텍스춰 기반 분석 방법과 연결 요소 분석 방법을 혼합하여 사용하였다. [3][4]는 압축 동영상인 MPEG 비디오와 JPEG의 I 프레임(Intraframe-coded picture)을 대상으로 DCT(Discrete Cosine Transform) 영역 내에서 인코드된 휘도 편차(intensity variation) 정보를 사용하여 직접 자막을 탐지하는 방법을 제안하였다. [5]는 뉴스 비디오를 대상으로 얼굴과 자막을 추출하였다. 비디오 내 얼굴은 스킨 컬러를 사용하였고 자막 추출은 [1]이 제안한 알고리즘을 사용하여 자막내 이름을 검출하였다. 뉴스 비디오를 대상으로 낮은 해상도 문자와 복잡한 배경에 대한 문제 제기와 이러한 문제를 해결하기 위하여 미분 필터, 평활화 필터, 보간 필터(interpolation filter), 다중 프레임 통합(multi-frame integration) 등을 조합 적용하여 자막 추출 및 세그멘테이션을 하였다.

[6]은 연결 요소 분석 방법을 기반으로 종이 신문 광고의 웹사이트 자동 변환을 위한 텍스트 탐지, 멀티미디어 검색 엔진과 비디오 인덱싱을 위한 텍스트 탐지 방법을 제안하였다. 제안된 방법은 비디오 프레임들을 다른 컬러의 부영상들로 분해하고 부영상들이 갖는 텍스트 요소들이 미리 명세된 휴리스틱을 만족하는가 실험하였다. [7]은 이진 영상과 웹 영상은 분해(decomposition)을 먼저하고 분해에 의한 전경을 선택하여 전경 영상을 출력한다. 다음으로 전경 영상들을 대상으로 연결 요소 분석 방법을 사용하고 마지막으로 텍스트 확인을 한다. 컬러 영상, 그리고 비디오 등의 데이터는 먼저 컬러 공간 축소를 하고 분해 처리를 한 다음 앞의 방법과 동일하게 처리하였다. [8]은 텍스와 그래픽 영상이 포함된 지도와 같은 영상을 대상으로 연결 요소 분석 방법과 피라미드 사용에 의해 텍스트를 탐지하는 방법을 제안하였다. 제안한 알고리즘은 먼저 영상의 디지털화를 수행하고 임계값에 의해 이진 영상으로 변환한다. 다음으로 연결 요소들을 생성하고 커다란 그래픽 연결 요소들은 제거한다. 마지막으로 피라미드 생성과 순회에 의해 텍스트 영상을 결과로 출력한다. 그러나 제안한 방법은 복잡한 배경을 갖는 데이터에서는 적용 불가능하다.

[9]는 신경망을 이용한 텍스춰 기반의 문자 탐지 방법을 제안하였다. 먼저 학습 샘플을 이용해 구성된 신경망을 이용하여 입력 영상 내의 모든 화소들을 문자 화소와 비문자 화소로 분류하고 이 분류 결과로 얻은 영상을 평활화하고 히스토그램 분석을 통해 문자 영역 탐지한다. [10][11]은 TFs(Topographical Features)와 isodata 컬러 클러스터링을 사용하여 자막 영역을 탐지하는 방법을 제안하였다. 먼저 문자의 지형학적 특징을 이용하여 TFs 포인트를 추출하고 비슷한 컬러를 클러스터링하는데 isodata 클러스터링 방법을 사용하여 후보 영역을 검출한다. 다음으로 각각의 클러스터화된 그룹들을 블록화화하는데 포인트라인영역 방법을 사용하고 구성된 블럭들은 수정되고 후보 영역들이 결정된다. 마지막으로 자막 영역 몇 개의 조건 테스팅과 후보 영역 수정에 의해 결정되어진다. [12]는 비디오 프레임내의 자막 영역이 가지고 있는 텍스춰 특성을 분석하여 자막 영역을 분할하고 프레임간의 자막 연속성을 이용하여 자막 프레임구간과 대표 자막 영역 및 색상 추출 방법을 제안하였다.

위 대부분의 기존 연구들은 문자들은 한정된 크기를 갖으며, 한 자막의 텍스트 라인은 수평으로 배열된 문자들의 클러스터, 그리고 텍스트는 배경으로부터 비교되는 대비를 갖는다는 텍스트의 특성 및 사전 지식을 기반으로 비압축 비디오 및 영상을 대상으로 실험한 연구 결과들이다. 그러므로 압축된 영상과 영상 내 자막의 크기에 무관한 텍스춰 필터 설계, 그리고 실시간에 비디오 자막을 탐지할 수 있는 알고리즘이 개발되어야 한다.

3. MPEG(Moving Picture Expert Group)

영상 압축기법으로는 손실압축(lossy compression)과 비손실(lossless)압축, 또는 대칭압축기법과 디코딩보다 인코딩에 더 많은 노력이 소요되는 비대칭 압축으로 분류된다. 기초적인 압축기법으로는 엔트로피 인코딩으로 런랭스 코딩(run-legth coding), 허프만(Huffman)코딩, 소스 인코딩으로 예측(prediction) 방식, 변환방식, 다단계코딩방식, 벡터 양자화 등이 있으며, MPEG, JPEG(Joint Photographic Expert Group) 등은 엔트로피 인코딩 방식과 소스 인코딩 방식을 혼합한 방식이다.

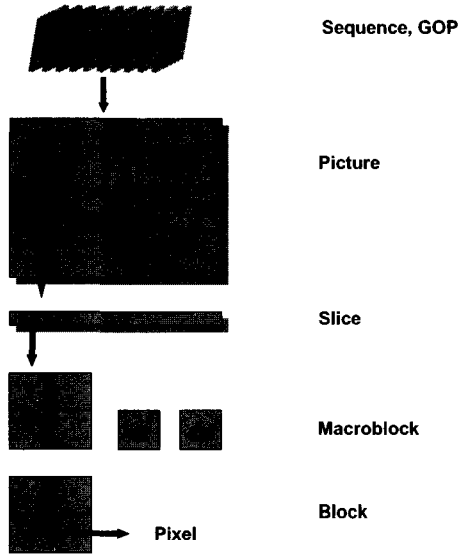
MPEG의 정식명칭은 동화상전문가그룹(Moving Picture Experts Group)이며1988년 설립된 MPEG은 시간에 따라 연속적으로 변화하는 동영상 압축에 관하여 연구하는 그룹 명칭이다. MPEG은 모션 비디오와 오디오 코딩을 위해서 ISO(International Organizations for Standardization)/IEC(International Electro mechanical Commission) JTC1/SC 산하의 29/WG 11에 의해서 만들어 졌으며 미국의 AT&T, 영국의 BT, 일본의 NTT 등의 통신업체 및 후지쓰, 미쓰비시, 픽처텔, 비디오텔리컴 등 영상회의 장비업체들이 소속되어 있다.

MPEG 종류로는 MPEG-1, MPEG-2, MPEG-3, MPEG-4, MPEG-7, 그리고 MPEG-21이 있다. MPEG-1은 비디오 CD와 MP3의 표준이 되었고, MPEG-2는 디지털 TV와 DVD의 오디오/비디오 압축 표준으로 채택되었으며 MPEG-4의 표준도 완료되어 그 응용분야가 점차 확대되어가고 있다. "Multimedia Content Description Interface"를 위한 MPEG-7 표준화가 2001년 9월 국제 표준 제정을 목표로 추진되고 있으며, "Multimedia Framework"이라 일컬어지는 MPEG-21 표준화를 1999년 10월부터 시작하여 진행하고 있다.

MPEG 표준은 동영상을 압축하는 기법으로 MPEG보다 먼저 개발된 JPEG과 H.261을 기반으로 연구 개발되고 있으며 동영상은 정지영상의 연속적인 배열이라고 볼 수 있기 때문에 MPEG에서도 JPEG의 압축과 H.261 기술을 사용한다. 또한 MPEG은 대칭압축방식과 비대칭압축 방식 모두에 적합하다.

3.1 계층구조

MPEG은 계층화된 구조를 가지며 헤더에 의해 비디오 신호와 오디오 신호를 분류되며 비디오 신호는 또 다른 헤더에 의해 각 계층으로 구분된다. 다음 [그림 1]은 MPEG의 각 계층을 나타낸다.



[그림 1] MPEG의 계층 구조
[Fig. 1] MPEG hierarchy

MPEG 계층 구조 중 가장 상위 계층으로 시퀀스는 비디오 신호를 나타내는 헤더로 시작하여 영상의 수직/수평 크기, 수평/수직 비율, 초당 프레임 수, 비트율, 디코더에 내장된 양자화 행렬을 사용할 것 인지의 여부, 사용자 데이터나 확장 데이터가 있는지의 여부를 등의 정보를 포함하고 있다. GOP(Group of Picture)와 픽쳐(Picture)는 앞 [그림 1]에 나타난 것처럼 시퀀스의 다음 계층으로 픽쳐의 그룹이며 동영상은 서로 다른 여러 장의 정지영상들로 구성된다. 픽쳐(Picture)란 정지영상을 말하며 MPEG에서는 네 가지 종류의 픽쳐를 정의한다. 다음은 MPEG가 갖는 네 종류의 픽쳐(프레임)를 설명한다.

Intra-coded picture(I picture 또는 I frame)는 다른 프레임과 중복되는 정보를 제거하지 않는 인코딩 방법으로 다른 프레임의 참조없이 재구성이 가능하다. 또한 시간적인 정보를 전혀 가지고 있지 않으며 공

간적인 정보만을 가지고 인코딩을 수행한다. 즉, JPEG 압축기법을 사용하여 압축을 실행하므로 이동 벡터에 대한 정보를 포함하지 않는다. I 프레임은 다른 픽쳐, 즉 P 또는 B 프레임이 이동 벡터를 생성하는데 기준이 되며 공간적인 정보만을 이용하여 압축을 실행하므로 압축율이 다른 픽쳐에 비교하여 가장 비효율적이다.

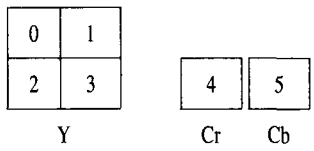
Predictive-coded picture(P picture 또는 P frame)는 시간적인 정보와 공간적인 정보 모두를 사용하여 생성된다. I 프레임 또는 최근의 다른 P 프레임을 기준으로 전향 예측(forward predict), 즉 전향 이동 벡터를 검출하고 검출된 이동벡터를 이용하여 P 프레임을 구성한다. 그러나 전향 이동벡터를 검출할 수 없는 경우에는 JPEG 압축기법처럼 공간적인 정보만을 이용 P 프레임을 생성한다. 그러므로 다른 프레임(I, P)의 데이터 없이 재구성은 불가능하다.

Bidirectionally predicted-coded picture(B picture 또는 B frame)은 시간적으로 전후의 I, P 프레임에서 양쪽 또는 한쪽으로부터 움직임을 보상하는 방법으로 구성된다. 즉, I 프레임과 P 프레임을 사용하여 전향 이동벡터와 후향 이동벡터를 동시에 검출하고 이를 이용하여 B 프레임을 구성한다. 그러므로 재구성 시 두 프레임이 반드시 필요하다. 또한 전향 및 후향 이동벡터 모두 사용하므로 다른 프레임들 보다 데이터량은 가장 작다.

DC-picture(D picture 또는 D 프레임)는 재생한 영상의 전 화소 값들이 DCT(Discrete Cosine Transform)의 DC계수 값만으로 구성된 프레임으로 빠른 검색 또는 Fast Forward Mode에서 사용되어 지는 프레임이다.

GOP 계층은 적어도 하나 이상의 I 프레임을 포함하여야 한다. 그 이유는 I 프레임을 기준으로 디코딩이 이루어진다. 각 영상의 인코딩 순서는 I 프레임, P 프레임, 그리고 I 프레임과 P 프레임 사이에 B 프레임 순서로 인코딩된다. P 프레임이 먼저 인코딩되는 이유는 B 프레임이 이전과 후의 이동 벡터를 모두 사용하므로 먼저 P 프레임을 인코딩하고 여기서 획득한 이동 벡터를 B 프레임 인코딩시 사용한다.

한 프레임은 슬라이스로 구성되며 슬라이스는 한 라인의 매크로블록으로 구성되어진다. 한 프레임 내의 각 슬라이스 크기는 가변적인 크기를 가질 수 있으며 처음 슬라이스는 영상의 왼쪽 위의 매크로블록에서 시작하고 마지막은 오른쪽 아래의 매크로블록으로 완성된다. 각 화면은 일반적으로 352×240 크기의 MPEG SIF(Source Input Format)의 영상을 슬라이스 단위로 분할하여 인코딩한다. 슬라이스는 16×16 화소로 구성되는 매크로블록으로 구성되며 매크로블록은 휘도(luminance) 정보를 갖는 4개의 블록과 색도 정보를 갖는 2개의 블록으로 구성된다. 다음 [그림 2]는 매크로블록의 구성을 나타낸다.



[그림 2] 매크로블록
[Fig. 2] Macro block

위 [그림 2]에서 각 블록은 8×8 화소 크기를 가지며 번호는 인코딩 또는 디코딩 시의 처리 순서를 나타낸다. 매크로블록은 영상간 인코딩 시 이동보상을 하는 단위로서 사용된다. 즉, 이동벡터를 계산하는데 사용되어진다. 최 하위 계층으로 한 개의 블럭은 8×8의 화소로 구성되며 블럭은 DCT 연산과 엔트로피 부호화의 기본 단위가 된다.

3.2 인코딩

동영상 압축기법은 영상간의 중복성을 없애기 위한 방법들을 많이 사용하며 영상에 내포된 중복성은 크게 세 종류로 분류된다. 먼저 24 fps 또는 30 fps의 프레임이 발생했을 때 이웃하는 두 장의 프레임은 매우 비슷하다. 두 프레임 내의 객체들이 정지한 경우 두 프레임은 완전히 같고 프레임 내의 객체 움직임이 있더라도 그 부분을 제외하면 배경은 같다. 이러한 프레임과 프레임간에 존재하는 중복성을 시간적 중복성이라한다. 두 번째, 한 프레임 내에서 이웃하는 화소간의 값들이 매우 비슷한데 이것이 화소와 화소 사이에 존재하는 공간적 중복성이다. 시간적 중복성과 공간적 중복성을 없애기 위해 손실부호화를 사용하는데 이동보상(motion compensation) 예측기법, DCT(Discrete Cosine Transform) 그리고 양

자화가 있다. 양자화된 이동보상 DCT 계수들은 통계적으로 어떤 값들은 자주 나타나고 어떤 값들은 희박하게 나타난다. 이러한 중복성이 통계적 중복성이다. 통계적 중복성을 없애기 위하여 무손실 인코딩 방법을 사용하는데 MPEG에서는 허프만 인코딩을 사용한다.

동영상은 종종 프레임 내에 회전이나 파도와 같은 비규칙 패턴을 가진다. 이러한 강한 움직임의 불규칙한 패턴을 가진 영역은 I 프레임 인코딩 비율과 비슷한 정도로 줄여질 수 있다. 시간적 예측의 사용은 방대한 양의 정보와 영상을 위한 저장장치가 필요하다. 그러므로 이 요구량과 가능한 압축률 사이에 균형이 필요하다. 대부분의 경우 이 예측 인코딩은 전체 영상이 아니라 영상의 부분들에 대해서 더 합리적이다. 따라서, 각 영상들은 매크로 블록이라 불리는 작은 영역으로 나누어진다. 각 매크로 블록은 16×16의 휘도 요소와 8×8의 색도(chrominance) 요소로 분할이 되며 이 블록을 단위로 압축한다.

MPEG에서는 프레임 코딩 방법으로 4 개의 방법이 있다. 이는 효율적인 코딩과 빠른 접근을 위해서다. 높은 압축률을 획득하기 위해서는 후속프레임들의 시간적 중복성을 이용해야한다. 즉 인터프레임(interframe) 코딩이 필요하고, 빠른 접근(access)를 위해서는 인트라프레임(intraframe) 코딩이 필요하다. 이러한 요구를 만족하기 위하여 MPEG에서는 4 가지 프레임 타입을 제안한다. I 프레임은 정지영상으로 간주되며 다른 프레임에 대한 참조없이 코딩된 프레임이다. MPEG에서는 I 프레임을 위해서 JPEG 압축기법을 사용한다. 그러나, 압축이 실시간에 실행되어야 하기 때문에 압축률이 MPEG 내에서 가장 낮다. 대신 I 프레임은 MPEG 스트림에서 무작위 접근(random access)를 위해서 사용된다. I 프레임들은 매크로 블록에서 정의된 8×8 블록들을 사용하며, 매크로 블록에도 두 가지가 있는데 하나는 단지 인코딩된 데이터를 포함하고 다른 하나는 비례축소에 사용되는 파라미터만을 가진다.

P 프레임은 인코딩과 디코딩에 이전 I 프레임 또는 이전 모든 P 프레임들을 필요로 한다. 실제로 연속되는 프레임들은 대부분의 경우 변화하지 않고 변화하더라도 조금만 변화된다. P 프레임은 이러한 성질을 이용하여 현재 블록과 가장 비슷한 이전 P 또는 I 프레임을 매칭(matching) 알고리즘을 사용하여 결정한다. 이러한 방법으로 결정된 매크로 블록과 이동벡터(매크로 블록의 공간적 위치간 차)가 인코딩된다.

B 프레임은 이전 프레임들과 다음 프레임들의 정보를 이용하여 코딩하며 가장 높은 압축률을 획득할 수 있다. B 프레임은 이전 영상의 예측(prediction)과 다음 P 프레임 또는 I 프레임의 차이에 의해서 정의된다. 이러한 이유로 B 프레임은 직접 액세스될 수 없다. 같은 배경에 공이 날아가는 프레임 경우 다음 프레임에서 예측할 수 있다. 즉, 역방향 이동벡터(reverse direction motion vector)도 사용될 수 있으며 양쪽 매크로 블록의 정합만을 사용하는 보간(interpolation)이동보상도 사용된다. D 프레임은 인트라프레임이며 fast-forward 또는 fast rewind mode에 사용된다. 따라서, D 프레임은 영상의 lowest frequency로 구성되어 있다. 위와 같은 각 프레임의 특성에 따라서 몇 가지 규칙이 정해진다. 기본적으로 일정한 주기마다 I 프레임이 필요하며 효율적인 압축률을 위해서 되도록 많은 B 프레임이 생성될수록 효율적이나 B 프레임 주변에 I, P 프레임이 필요하기 때문에 무작정 많은 B 프레임을 생성할 수는 없다. "IBBPBBPBBIBBPBBPBB..."와 같은 시퀀스가 실용적인 응용에서 적합하다.

4. 비디오 자막 탐지

본 4장에서는 자막 영역을 탐지하기 위해서는 MPEG 형식으로 되어있는 동영상을 디코딩 해야한다. 다음은 MPEG의 디코딩 과정을 간략하게 기술한다. MPEG 디코딩 과정은 인코딩의 역순으로 먼저 가변길이 디코딩을 하고 역스캔을 한다. 이 때 역스캔은 ZigZag 스캔 또는 대체 스캔(MPEG-2 사용) 중 하나를 사용한다. 다음으로 역양자화와 IDCT(역이산여현변환)하면 디코딩된 표본이 생성되고 움직임 보상과 프레임 재순서화 절차를 거쳐 디코딩된 영상들이 나타난다. 실험에서는 DCT 계수들을 대상으로 하기 때문에 DCT 계수 블록까지만 디코딩하여 실험에 사용한다.

4.1 텍스춰 특징 추출

자막 영역 위치를 탐지하기 위하여 MPEG 동영상을 디코딩 하는데 먼저 가변길이 디코딩을 하고 ZigZag 역스캔을 한다. 다음으로 역양자화하여 DCT 계수를 갖는 8×8 블록들을 대상으로 실험한다. 다음 $F(u, v)$는 DCT 수식을 나타내고 $f(i, j)$는 IDCT 수식을 나타낸다.

$$F(u,v)=\frac{1}{4}C(u)C(v)\sum_{i=0}^{7}\sum_{j=0}^{7}f(i,j)\cos\left(\frac{(2i+1)u\pi}{16}\right)\cos\left(\frac{(2j+1)v\pi}{16}\right)$$

$$f(i,j)=\sum_{u=0}^{7}\sum_{v=0}^{7}F(u,v)\frac{1}{4}C(u)C(v)\cos\left(\frac{(2i+1)u\pi}{16}\right)\cos\left(\frac{(2j+1)v\pi}{16}\right)$$

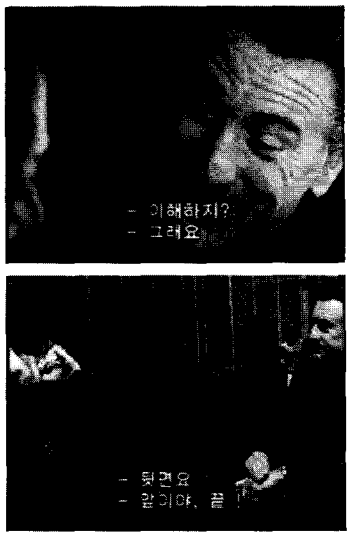
$$with,\ C(u)=\left\{\frac{1}{\sqrt{2}}\ \text{if}\ u=0\ ,\ 1\ \text{if}\ u>0\right\},$$
$$C(v)=\left\{\frac{1}{\sqrt{2}}\ \text{if}\ v=0\ ,\ 1\ \text{if}\ v>0\right\}$$

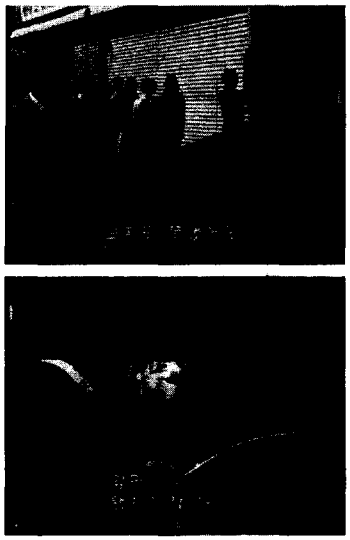
위 식에서 $f(i, j)$는 표본 영역에서의 좌표 i, j의 화소 값, $F(u, v)$는 변환 영역에서 좌표 u, v의 값을 나타낸다. DCT는 8×8 영상의 값들을 8×8의 주파수 계수의 행렬로 변환한다.

본 연구에서는 압축 영상에서 텍스춰 정보와 색체 정보를 이용하여 자막 영역을 탐지한다. 자막 영역은 다음과 같은 일정한 텍스춰와 색체 특징을 갖는다. 각 자막 영역은 자막 라인간에 대략 같은 간격을 갖고 같은 방향을 갖는다. 또한 자막 라인은 같은 크기의 문자들로 구성된다. 이러한 비디오 영상내의 자막 특징은 공간 영상의 국부 주기성과 방향성을 갖는다. 또한 자막은 인위적으로 해당 프레임들을 설명하거나 번역을 해놓은 것이기 때문에 비슷한 색체 정보를 갖는다. 다음 [그림 3]과 [그림 4]는 각각 뉴스 비디오 영상들과 영화 비디오 영상들을 나타낸다.

[그림 3] 뉴스 자막 영상

[Fig. 3] News caption images





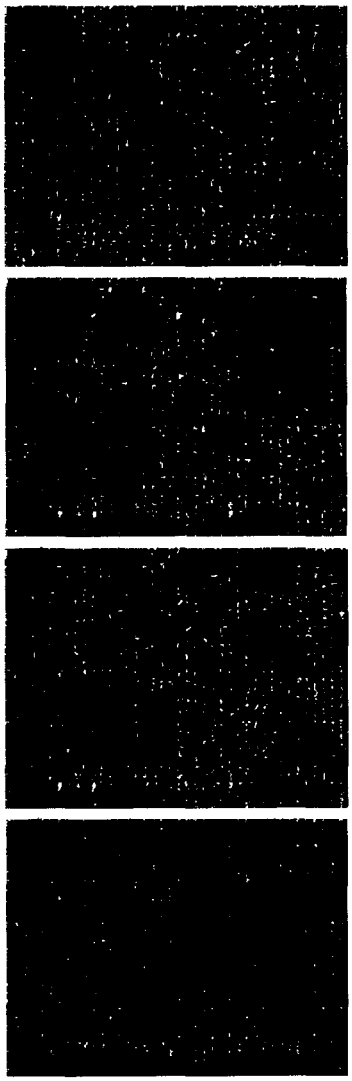
[그림 4] 영화 자막 영상
[Fig. 4] Movie caption images

8×8 블록을 대상으로 계산되는 DCT 계수들은 영상내의 국부 특징을 가지며, DCT 계수는 8×8 블록에 포함된 2차원 공간 주파수를 표현하므로 공간 주기성과 방향성의 측정 방법으로 사용될 수 있다. 또한 양자화된 DCT 계수들은 비디오 스트림 또는 JPEG 데이터로부터 쉽게 추출될 수 있으며, DCT 계수들이 양자화되어도 계수 정보는 유지되고 다른 디코딩 처리없이 텍스춰 특징을 계산할 수 있다. 압축된 도메인에서 DCT 계수 값들을 사용하여 공간 영상의 국부 주기성과 방향성을 획득할 수 있다. 다음은 이러한 근거에 의해 자막 위치를 탐지한다.

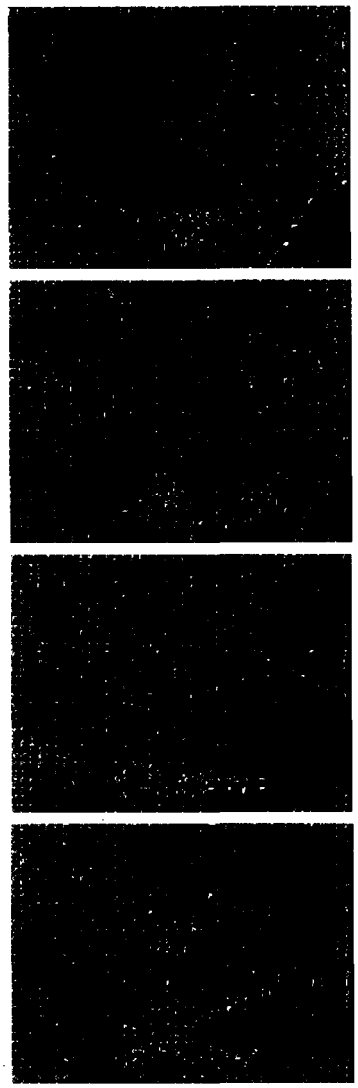
제안된 알고리즘은 자막 라인내의 문자들이 빠른 휘도 및 색도 변화를 보이기 때문에 수평적으로 강한 반응을 보이는데 근거한다. 또한 자막 라인 간에 빠른 휘도 및 색도 변화를 보여 수직적으로도 강한 반응을 보인다. AC 계수들을 자막 라인 내 문자들의 수평 휘도 및 색도 변화와 자막 라인간의 수직 휘도 및 색도 변화를 획득하기 의하여 사용한다.

4.2 압축 영상에서의 자막 탐지

압축된 영상에서 후보 자막 영역을 획득하기 위하여 I 프레임내의 8×8 DCT 블록을 연산 기본 단위로 사용하여 수평 휘도 변화 및 수평 색도 변화를 계산한다. 자막 후보 영역을 획득하기 위하여 수평 휘도 및 수평 색도 변화가 적은 부분은 임계값에 의해 제거한다. 다음으로 고립된 잡은 블록은 제거하고 연결되지 않은 블록 영역은 병합한다. 마지막으로 자막 영역을 수직 휘도 및 수직 색도 변화에 따라 수정하고 필터링한다. 다음 [그림 5]과 [그림 6]는 앞 [그림 3]과 [그림 4]의 MPEG 영상들의 DCT 영상을 나타낸다.



[그림 5] DCT 영상
[Fig. 5] DCT images



[그림 6] DCT 영상
[Fig. 6] DCT images

각 8×8 DCT 블록(i, j)를 대상으로 수평 고주파의 절대값 합에 의해 수평 자막 에너지 $E(i, j)$를 계산한다.

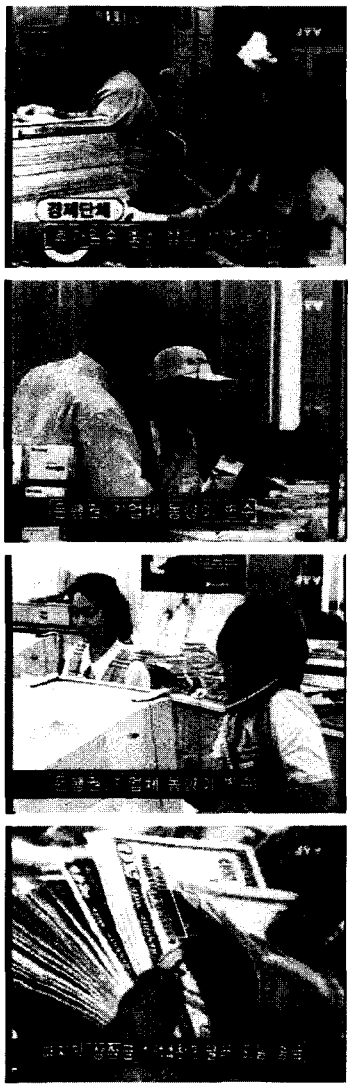
$$E(i,j) = \sum_{v1 < v < v2} |C_{ov}(i,j)|$$

위 식에서 v_1과 v_2는 문자 크기에 근거한다. 본 실험에서는 $v_1 = 2$, $v_2 = 6$을 사용한다. 수평 자막 에너지 값은 수평 휘도 및 수평 색도 변화에 의해 한 블록의 수평 자막 에너지가 임계값보다 크면 자막 후보이다. 단순한 임계값에 의하여 대부분의 자막 블록이 인증된다. 그러나 높은 수직 휘도 및 수직 색도 변화를 갖는 비자막 블록을 선택하기도 하였다. 더욱이 문자간 단어간 간격에 의존하므로 넓은 간격, 낮은 대비, 커다란 폰트 등에 의하여 검출된 자막 후보 블록들은 분리되고 비연결 되었다. 그러므로 단순한 임계 절차를 사용한 결과는 수정되어야한다. 본 논문에서 제안한 알고리즘은 먼저 I 프레임의 각 DCT 블럭의 수평/수직 휘도 및 색도 변화를 계산한다. 다음으로 잠정 자막 블록을 선택하고 잡음 블록 제거 및 비연결 자막 블록을 병합한다. 마지막으로 자막 영역 수정 및 수직 휘도 및 수직 색도에 의한 필터링하고 자막 영역을 탐지한다.

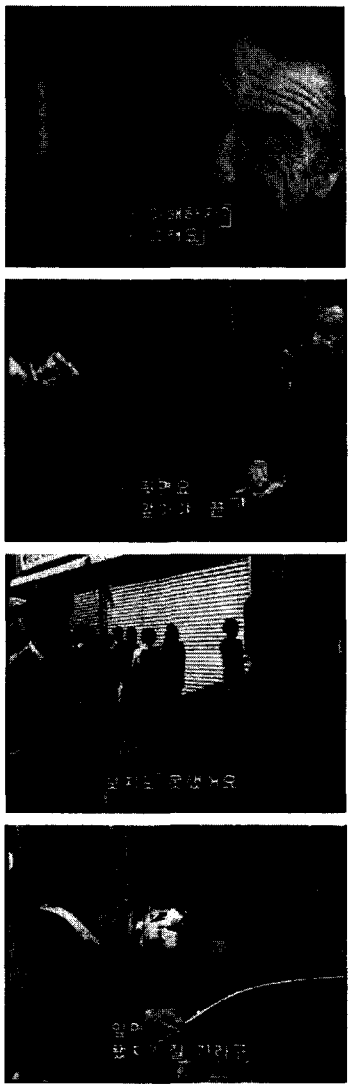
위 알고리즘의 실험 환경으로 펜티엄 800MHz 중앙처리장치, 256MB의 주기억장치, 그리고 MPEG 동영상 클립 6개를 가지고 실험하였다. 비디오 자막 탐지 알고리즘의 구현 프로그래밍 언어로는 웹기반 프로그램으로의 확장성을 고려하여 자바(Java)를 사용하였다.

4.3 실험 결과

제안된 방법을 실험하기 위하여 MPEG 압축영상으로 앞 [그림 3]과 [그림 4]를 사용하였다. 비교적 영상의 질이 뉴스 영상 보다 영화 영상이 더 양호하여 자막 탐지에도 영향을 나타내었다. 다음 [그림 7]와 [그림 8]은 자막 탐지 영상들을 나타낸다. [그림 7]의 우측 하단의 영상에서는 돈 부분에 나타난 문자들을 자막으로 오탐지하였다. 또한 [그림 8]의 좌측 하단 영상의 특수문자 등도 문자들을 탐지하는데 있어 영향을 주었다.



[그림 7] 뉴스 영상(자막 탐지)
[Fig. 7] News images(caption localization)



[그림 8] 영화 영상(자막 탐지)
[Fig. 8] Movie images(caption localization)

5. 결론 및 향후 연구과제

본 논문에서는 MPEG 압축영상에서 직접 자막 영역을 탐지할 수 있는 알고리즘을 제안하였다. I 프레임 내의 DCT 계수들을 대상으로 텍스춰 정보와 색체 정보를 이용하여 자막 텍스트 영역을 탐지하였다. MPEG의 I 프레임 내에 직접 적용 가능한 알고리즘으로 자막 텍스트의 텍스춰 정보와 I 프레임의 각 DCT 블럭의 수평/수직 휘도 및 색도 변화를 계산하여 자막 영역을 탐지하므로 자막 탐지가 정확하고 처리 속도가 효율적이다. MPEG 압축영상은 YUV 컬러 공간을 기반으로 압축을 실행하므로 Y 요소의 영향을 많이 받으므로 HSI 컬러 공간과 같은 휘도 요소에 영향을 덜 받는 공간에서의 자막 탐지 방법에 대한 연구도 필요하다. 또한 다양한 형태의 압축 영상에 적용하려면 다른 압축 영상들의 인코딩 방법과 디코딩 방법 등을 분석하고 인코딩/디코딩 시에 사용되는 다양한 알고리즘을 분석하고 적합한 텍스춰 특징 추출 및 적용에 관한 연구가 필요하다.

※ 참고문헌

[1] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video OCR for digital news archive," Content-Based Access of Image and Video Database, Proceedings., 1998 IEEE International Workshop on , pp. 52-60, 1998.

[2] Y. Zhong, K. Karu, and A. K. Jain, "Locating text in complex color images," Pattern Recognition, Vol. 28, No. 10, pp. 1523-1535, 1995.

[3] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic caption localization in compressed video," Image Processing, ICIP 99. Proceedings. 1999 International Conference on , Vol. 2, pp. 96-100, 1999.

[4] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic caption localization in compressed video", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 4, pp. 385-392, April 2000.

[5] S. Satoh, Y. Nakamura, and T. Kanade, "Name-It: naming and detecting faces in news videos," IEEE Multimedia , Vol. 6, pp. 22-35, Jan.-March 1999.

[6] A. K. Jain and Y. Zhong, "Page segmentation using texture analysis," Pattern Recognition, Vol. 29, No. 5, pp. 743-770, 1996.

[7] A. K. Jain and B. Yu, "Automatic text location in images and video frames," Pattern Recognition, Vol. 31, No. 12, pp. 2055-2076, 1998.

[8] C. L. Tan and P. O. Ng, "Text extraction using pyramid," Pattern Recognition, Vol. 31, No. 1, pp. 63-72, 1998.

[9] 정기철, 김항준, "텍스춰를 이용한 영상내의 문자 추출", 제4회 문자인식 워크샵, pp. 49-58, 2000.

[10] B. T. Chun, Y. L. Bae, and T. Y. Kim, "Caption segmentation method in videos using isodata clustering of topographical features," TENCON 99. Proceedings of the IEEE Region 10 Conference, Vol. 2 , pp. 915-918, 1999.

[11] B. T. Chun, Y. L. Bae, and T. Y. Kim, "A method for original image recovery for caption areas in video," Systems, Man, and Cybernetics, 1999. IEEE SMC '99 Conference Proceedings. 1999 IEEE International Conference on, Vol. 2, pp. 930-935, 1999.

유 태 웅

1991년 전북대학교 수학과
학사
1993년 전북대학교
컴퓨터과학과 석사
1998년 전북대학교
컴퓨터과학과 박사
1999년 - 현재 서해대학
컴퓨터정보기술계열 조교수
2001년 - 현재 서해대학
테크노정보지원센터 소장
관심분야 : 컴퓨터비전,
무선원격감시시스템, VM,
컴퓨터그래픽스 등