Template을 이용한 동영상 비정규 구조 색인 방법

권순용, 오상욱, 윤자천, 설상훈
고려대학교 전자컴퓨터공학과
e-mail : {sykwon, osu, jcyoon, sull}@mpeg.korea.ac.kr

Indexing for irregular structured video using template

Soonyong Kwon, Sangwook Oh, Ja-Cheon Yoon, Sanghoon Sull
Dept. of Electronics and Computer Engineering, Korea University

요 약

동영상 데이터가 급증하면서 뉴스 검색과 같은 검색 서비스를 위한 내용 기반의 색인 방법에 대한 관심이 높아지고 있다. 이에 따른 최근의 동영상 색인 기술들은 특히 용량이 큰 동영상에서 자동 또는 반자동으로 색인하기 위한 방법들을 소개하고 있으며, 대부분 저수준 특징(low-level feature)들을 이용하여 특정 구간을 검출하는 방법을 사용하고 있다. 그러나 저수준 특징만을 이용할 경우 고수준 특징(high-level feature)이 나타내는 인지적 측면에서의 동영상 내용을 다룰수 없어, 사용자가 일일이 내용에 따른 구분 작업을 병행해야 하는 단점을 갖고 있다. 본 논문에서는 뉴스와 같이 시간적으로 비정규적이지만 내용에 따라 구조를 지니고 있는 동영상에 대해서, 저수준 특징만을 사용하여 고수준 특징을 이용한 것과 같은 성능으로 사용자가 수행해야 하는 작업을 최소화하는 내용 기반 구간 검출 방법에 의한 색인 방법을 제안한다. 특히 저수준 특징을 이용한 프로그램 구간 검출은 빠른 처리 및 효과적인 검출 성능 보이고 있어, 본 논문에서 제안한 비정규 구조 색인 방법이 내용 기반 색인에 매우 효과적임을 보여주고 있다.

1. 서론

지상파, 케이블 및 위성 방송 등 방송 인프라가 확대되고, 디지털화 되면서 동영상 데이터가 급증하고 있어, 점점 많아지는 동영상에서 원하는 것을 쉽게 찾기 위한 동영상 검색 서비스가 최근 주목 받고 있다. 특히, 뉴스 검색과 같은 서비스를 위하여 동영상 색인의 필요성이 늘고 있으며, 이에 따른 대용량 동영상을 자동 또는 반자동으로 색인하기 위한 관련 기술들이 최근 많이 소개되고 있다.[1-2] 이와 같은 기술은 대부분 저수준 특징(low-level feature)에 기반한 특정 구간 검출 방법을 사용하고 있다. 그러나 저수준 특징을 이용하는 방법은 사람들이 이해하는 동영상의 내용과 관련 있는 고수준 특징(high-level feature)을 다루지 못하기 때문에 사용자가 일일히 내용을 분석하고 구분해야 하는 작업을 병행해야 한다. 본 논문에서는 뉴스와 같이 시간적으로 비정규적이지만 내용에 따라 구조를 지니고 있는 동영상 색인 시, 사용자가 수행해야 하는 작업을 최소화할 수 있는 내용 기반 구간 검출을 이용한 색인 방법을 제안 한다. 본 논문에서 제안하는 방법은 크게 1) 내용 구조의 대표 화면을 선택하는 템플릿 설정(template registration), 2) 설정된 템플릿의 저수준 특징을 이용한 프로그램 구간 검출, 3) 검출된 구간에 대한 검증 단계로 구성되어 있다. 특히 저수준 특징을 이용하여 동영상 내용의 기본 구조를 파악하기 때문에 고수준 특징만을 이용한 경우 발생하는 구현 과정의 어려움을 피할 수 있으며, 저수준 특징의 빠른 처리 및 효과적인 검출 성능을 사용할 수 있어 본 논문에서 제안하는 방법이 동영상 데이터를 색인하는데 매우 효과적임을 알 수 있다. 실험 결과는 본 논문에서 제안한 색인 방법이 매우 효과적임을 보여주고 있다.

본 논문의 구성은 다음과 같다. 2 장에서는 앵커화면을 검출하기 위한 새로운 템플릿 설정 방식을 제안

하고, 3 장에서는 제안된 템플릿과 비주얼 패턴을 이용한 빠른 앵커화면 검출 방법을 보인다. 그리고 4 장에서는 그 실험 결과를 보이고 5 장에서 결론을 맺는다.

2. 템플릿 설정(Template registration)

일반적으로 내용상 구조를 가지고 있는 동영상은 각 구간의 시작점에 비슷한 화면(타이틀)을 구성하는 경우가 많다. 예를 들어, 뉴스 동영상의 경우 뉴스 앵커가 뉴스에 대한 해드라인을 먼저 설명하고 자세한 내용은 기자의 설명과 함께 상황을 찍은 내용으로 보여주고 있다. 이와 같은 경우에 앵커는 실내인 스튜디오에 있으므로 앵커의 장면은 비슷한 특징을 가지게 된다. 따라서, 뉴스 동영상은 앵커가 존재하는 화면(앵커샷)이 내용 구간을 나눌 수 있는 기준이 될 수 있다. 이에 따라, 뉴스 동영상 색인에 중요한 역할을 하는 앵커샷 검출을 위한 여러 방법들이 소개되고 있으나, 그 기반 기술은 대부분 앵커들의 모습을 검출하는 데 그 목적을 두고 있다.[3-4] 이와 같은 기술은 하나의 동영상에서 앵커를 검출하는데는 유용하지만, 매일 방영되는 뉴스 앵커의 모습은 때에 따라 변할 수 있기 때문에 동일한 프로그램을 여러 동영상에 동시에 적용하기 힘든 단점이 있다. 이와는 반대로 앵커의 배경화면인 뉴스 스튜디오는 그 변화가 적고 움직임이 없어, 앵커의 모습을 검출하여 앵커샷을 찾는 방법보다 뉴스 스튜디오의 특징을 검출하여 앵커샷을 찾는 방법이 더 효율적이 될 수 있다. 즉, 정확한 얼굴 검출은 매우 어려운 기술에 속하며, 더욱이 앵커는 말을 하는 동안 조금씩 움직임이기 때문에 그 구현에 있어서 고비용을 요구한다. 반면, 스튜디오 배경은 거의 고정적이며 간단한 특징 검출 방법으로 그 존재 여부를 검출할 수 있어 실제 구현시 많은 성능 차이를 낼 수 있다.

앵커 화면의 구성을 보면 화면 중앙 또는 옆으로 치우쳐진 공간에 앵커의 모습이 존재하며, 오른쪽 또는 왼쪽 상위 공간에 간단한 뉴스 대표 화면 및 뉴스 제목이 있고, 그 외 부분에는 배경화면이 차지하게 된다. 이 때 배경 화면으로는 스튜디오가 사용될 수 도 있으며, 또는 다른 특수 화면이 사용될 수 도 있다. 그림 1 은 국내 방송의 뉴스 및 미국 CNN 방송의 뉴스에서의 앵커화면의 예를 나타내고 있다. 그림 1 (a)의 앵커 화면은 앵커와 뉴스 대표화면, 그리고 건물 영상을 배경으로 하는 배경화면으로 구성되어 있으며, 그림 1 (b)는 앵커와 뉴스 대표화면, 그리고 특수 효과와 스튜디오를 배경으로하는 배경화면으로 구성되어 있음을 알 수 있다.

배경화면으로부터 앵커화면의 템플릿을 구성하기 위하여 배경화면에서 고정된 영역을 선택하는 과정이 필요하다. 앵커 화면의 구성 중 고정된 영역을 검출하기 위한 방법으로 간단히 앵커화면에서 사용자가 수동으로 고정 영역을 선택하는 방법이 있다. 이 경우 일반적으로 고정 영역을 정사각형, 직사각형과 같은 일정 형태로 대부분 설정하게 되는데, 선택된 영역이 전체 화면에서 차지하는 비율이 작아지는 한계가 있다. 비교되는 선택 영역이 작으면 저수준 특징이 효과적으로 앵커화면을 대표하지 못하고 검출 실패(missing) 또는 잘못된 검출(false)을 발생할 수 있게 된다.

(a) KBS 뉴스의 앵커화면 (b) CNN 뉴스의 앵커화면

그림 1. 뉴스에서의 앵커 화면의 예들

본 논문에서는 앵커화면의 템플릿을 구성하기 위하여 자동 템플릿 설정 방법(template registration)을 제안한다. 자동 템플릿 설정은 사용자가 선택한 2 장 이상의 앵커화면을 이용하여 화면의 고정영역을 추출하는 것으로 다음 과정을 거친다.

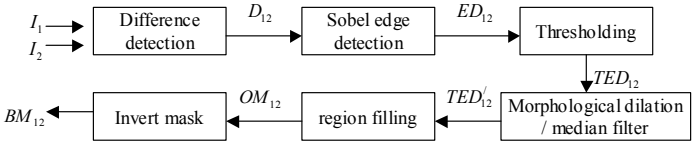


그림 2. 템플릿 설정 과정을 나타내는 도식

영상 1 과 영상 2 를 각각 I_1, I_2 이라고 하고 영상에서의 픽셀값을 p 라고 하면 차영상 D_{12} 은 다음과 같이 정의할 수 있다.

$$D_{12} = \{p \mid \quad p \mid I_1(p) - I_2(p) \mid\} \qquad (1)$$

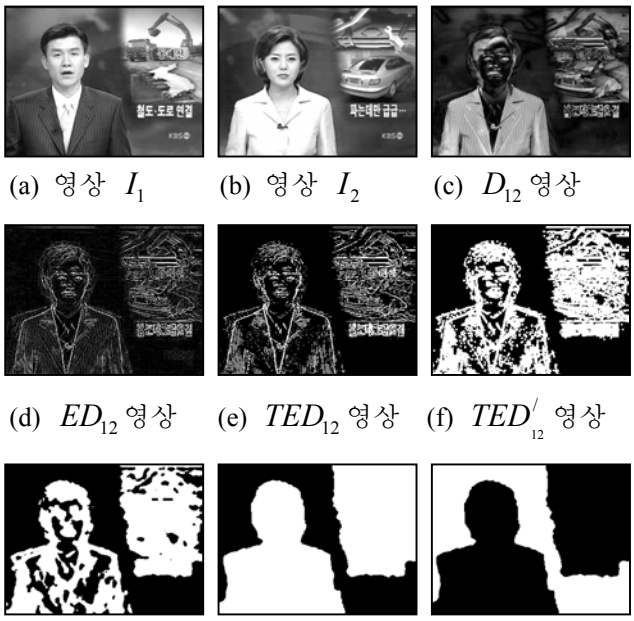
그림 3 (c)에서 보여지듯이 차영상에서 배경 영역과 앵커 등 객체 영역 모두에서 차이값이 발생함을 알수 있다. 배경 영역에서 발생하는 차이값들은 조명, 카메라 위치, 카메라 노이즈 등으로 인해 발생하는데, 이 차이값들은 객체 영역에서의 차이값보다 작다는 것을 알 수 있다. 하지만, 차영상에 그대로 임계치(threshold)을 적용하여 구분한다면 배경 영역과 객체 영역을 적절히 구분할 수 없다. 따라서, 본 논문에서는 차영상의 에지(edge)를 구한 후, 임계치를 주어 배경 영역에서의 에지와 객체 영역에서의 에지를 구분한다. 그림 3 (d)의 ED_{12} 는 소벨(Sobel) 에지 검출을 한 후의 그림이다. 에지 영상 ED_{12} 및 객체 영역에서의 에지만을 나타내는 TED_{12} 는 다음과 같이 나타낸다.

$$ED_{12} = \{p \mid \quad p = Sobel(D_{12}(p))\} \qquad (2)$$

$$TED_{12} = \{p \mid \quad p = ED_{12}(p) > Th_e\} \qquad (3)$$

이때 TED_{12} 를 구하기 위한 경계값 Th_e 는 에지 분포도를 이용하여 구하는데, 그림 4 과 같이 분포도가 두 개로 나뉘어져 있으며, 낮은 값을 갖는 분포도는 배경 영역에서 0 인 값과 약한 에지를 나타낸다는 것을 알 수 있다. 또한 위에서 언급했듯이 배경 영역에

서의 차이는 카메라 노이즈 등에 영향을 받기 때문에 카메라의 특성을 따르고 있음을 알 수 있다. 따라서 Th_e는 두 개의 분포도를 나눌 수 있는 값에서 결정할 수 있다. 이렇게 구한 TED_{12}를 몰폴로지 확장(morphological dilation), 미디언 필터(median filter) 및 영역 채우기(region filling) 기법을 통하여 객체 영역을 구하게 된다. 이렇게 구해진 객체 영역 마스크 OM_{12}을 반대로 전환하여 배경 마스크 BM_{12}을 구한다. 이렇게 구해진 BM_{12}은 영상 I_1에 적용하여 템플릿을 추출하게 된다.



(a) 영상 I_1 (b) 영상 I_2 (c) D_{12} 영상

(d) ED_{12} 영상 (e) TED_{12} 영상 (f) TED'_{12} 영상

(g) TED'_{12} 영상 (h) OM_{12} 마스크 (i) BM_{12} 마스크

그림 3. 앵커화면에서의 템플릿 설정 과정 예

그림 4. ED_{12}에서의 에지 분포도

3. 비주얼 패턴을 이용한 앵커화면 검출 방법

앵커샷을 검출하기 위하여 앞절에서 보인 템플릿을 마스크로 이용하여 추출된 고정영역을 얻은 후 저수준 특징을 추출한다. 추출된 특징값의 유사도는 앵커샷이 존재하는 부분을 결정하게 한다. 이때 모든 프레임마다 프레임 전체 영역에 대한 저수준 특징을 추출하고 유사도를 결정하는 것은 시간적으로 고비용 작업이다. 따라서, 본 논문에서는 원 영상의 특징을 잘 포함하고 있는 축소영상에서의 비주얼 패턴, 즉 [5]에서 제시한 Visual Rhythm (VR)을 사용하여 전체 시스템의 성능을 향상시킨다.

VR 은 비디오 V 에서 시간 t 상의 한 프레임 좌표 (x, y)에 위치한 화소 값에 대한 함수 $f_V(x, y, t)$ 로 표현된다. 이때 비디오 V의 r 비율 만큼 축소된 영상 V_R은 $f_{V_R}(u, v, t)$ 로 표현되고 다음 식에 의해 비디오 V에 의해서 구해진다.

$$f_{V_R} = f_V(ru + k_u, rv + k_v, t), \quad u, v, t \in \{0,1,2,...\} \tag{4}$$
$$0 \le k_u, k_v \le r - 1$$

식(4)에서 k_u 와 k_v는 픽셀단위의 오프셋 값이고 r 은 축소비율이다. 따라서 축소된 비디오에서의 VR 은 다음과 같이 구해진다.

$$VR = \{f_{VR}(z, t)\} = \{f_{V_R}(u(z), v(z), t)\} \tag{5}$$

식 (5)의 $u(z)$ 와 $v(z)$ 는 독립변수 z 의 1 차 함수이다. VR 은 함수 $u(z)$ 와 $v(z)$ 에 따라 얻어지는 방식이 달라진다. 예를 들어 대각선 방향의 샘플링은 상수 δ 로 이루어진 함수 $(u(z), v(z)) = (\delta z, z)$ 로 구해질 수 있다. 그림 3 은 VR 을 구성하는 방법과 추출된 VR 의 특징을 나타낸다.

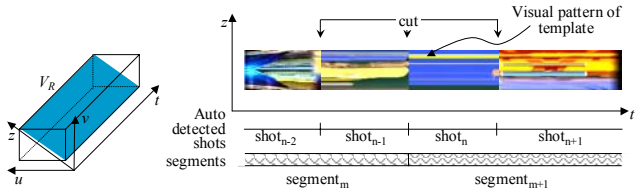


그림 5. Visual Rhythm 구성 방법 및 추출된 VR 예

비디오 V 에 대하여 구해진 VR_V을 통하여 쉽게 컷을 찾을 수 있을 뿐만 아니라 템플릿의 비주얼 패턴 VR_T을 이용하여 앵커샷을 쉽게 결정할 수 있다.

템플릿의 비주얼 패턴 VT_T과 비디오 V로부터 얻어진 비주얼 패턴 VR_V 의 비교 시, 그림 3 (i)의 BM_{12} 마스크 영상의 비주얼 패턴 VR_M 을 이용한다. 다음 식은 비디오 V 상의 임의의 시간 t 에서 두 패턴의 유사도 S^t를 나타내는 식이다.

$$S^t = \sum_{h=0}^{k} \{(\mathrm{VR_V}, t+h) - \mathrm{VR_T}\} \otimes \mathrm{VR_M} \tag{6}$$

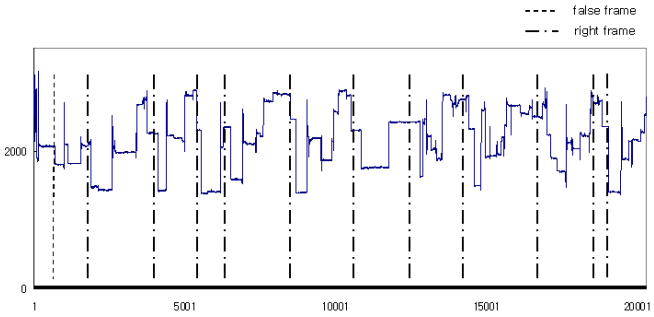
식 (6)에서 k 는 비교할 프레임의 수, 즉 템플릿과 비교하는 시간을 나타내며, 연산자 $\otimes$는 마스크 연산자로서 두 비주얼 패턴 비교시 앞절에서 설명한 템플릿의 고정영역만을 취하기 위함이다.

따라서 앵커샷은 시간 t 에서 유사도 S^t가 임의의 임계치 TH_S 보다 작을 때 또는 시간 t에 대한 S^t의 변화량 $DS = \left|S^t - S^{t-1}\right|$가 TH_{DS} 보다 클 때 결정되며, 이때 시간 t 에서 가장 가까운 곳에 위치한 컷이 앵커샷의 시작 위치로 검출된다.

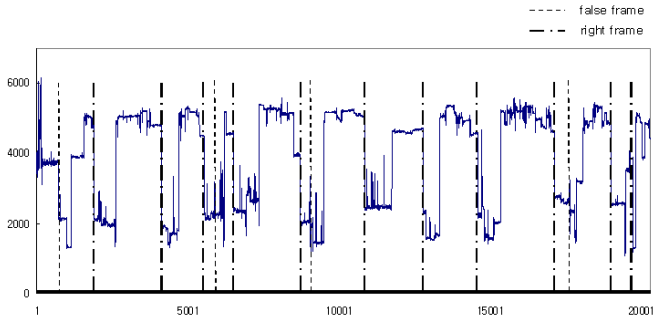
4. 실험 및 논의

실험은 주기적으로 방송되고 있는 뉴스 프르그램과 고 3 수능 강좌 프로그램을 이용하여 본 논문에서 제안하는 방법을 실험하였다. 각 프로그램은 템플릿 설정을 하기 위해 두개의 동영상에서 템플릿 화면을 각 2 개씩 선택하여 사용하였다.

그림 6 는 VR 에서 템플릿과 동영상 화면과의 저수준 특징값들을 비교하여 유사도값을 구했을 때, 시간에 따른 유사도 값 변화를 나타내는 그래프이다. 그림에서 보여주듯이 템플릿 화면인 경우 유사도가 매우 높게(그림에서는 distance 값) 나옴을 알 수 있다. 특히 앵커샷이 있는 부분에서 변화량이 매우 크게 변한다는 것을 알 수 있다. 그림 6 에서 수직선은 VR 을 이용하여 검출된 후보 집합에 대한 시간적인 위치를 나타내고 있다. 그림 6 (a)는 전체 화면에서 VR 을 구하고 앵커 샷을 검출했을 때의 유사도 변화량을 나타낸 그래프이며, 그림 6 (b)는 DC 영상에서의 유사도 변화량을 나타낸 그래프이다. 그림 6 (a)을 보면 총 11 개의 실제 앵커샷을 갖는 동영상에서 총 12 개의 샷을 검출하였고, 그 중 false 샷은 1 개임을 나타내고 있다. 따라서 본 논문에서 제안하고 있는 방법이 매우 효과적으로 앵커샷을 검출하고 있음을 보여주고 있다. 또한 그림 6 (b)에서는 총 15 개를 검출하였으며, DC 영상임에도 효과적으로 제안된 방법을 적용할 수 있음을 나타내고 있다.



(a) 전체 화면에서의 검출시 유사도 변화 그래프



(b) DC 영상에서의 검출시 유사도 변화 그래프

그림 6. 동영상 재생 시간에 따른 VR 에서의 유사도 값의 변화 및 검출된 후보 화면의 시간 위치

표 1 은 몇가지 동영상에 대한 검출된 결과를 나타내고 있다. 동영상 5 인 경우를 보면 실제 존재하는 앵커샷은 총 10 개이며, 검출된 샷은 14 개이다. 이 중 빠진 샷은 없으며 false 샷은 4 개이내임을 알 수 있다.

표 1. 다양한 동영상에 대한 검출 결과

	Decoding Level	Number of anchor frames	Detected frames	Falsely detected frames	Missed frames	Correctly detected frames
Video 1	Full	13	23	10	0	13
	DC	13	19	6	0	13
Video 2	Full	10	23	14	1	9
	DC	10	24	14	0	10
Video 3	Full	13	24	11	0	13
	DC	13	35	22	0	13
Video 4	Full	12	19	7	0	12
	DC	12	50	38	0	12
Video 5	Full	10	14	4	0	10
	DC	10	44	34	0	10

5. 결론

본 논문에서는 효과적으로 동영상을 내용 기반 색인할 수 있는 방법을 제안하고 있다. 기존의 방법은 앵커샷 내에 존재하는 객체를 검출하고 검출된 객체의 특징들을 이용하여 앵커샷을 검출하지만, 움직이는 객체를 판단하는 것은 매우 고비용이며, 객체의 특성이 바뀌기 때문에 문제가 발생할 확률이 높게 된다. 따라서 본 논문에서는 특징이 고정적이며 상대적으로 객체보다 추출하기 쉬운 고정 영역을 이용함으로써 기존 방법에서의 문제점들을 보완하고 있다. 또한 제안된 방법은 뉴스, 동영상 강의와 같이 주기적으로 방송되는 프로그램에도 효과적으로 적용할 수 있어 사용자의 작업을 최소화하면서 저비용으로 동영상을 색인할 수 있다. 실험 결과에서는 제안된 방법이 많은 비용이 필요한 작업을 최소화하여 전체 시스템의 성능을 향상시키고 있음을 보여주고 있다.

참고문헌

[1] Ardizzo, E.; La Cascia, M.; Di Gesu, V.; Valenti, C., *"*Content-based indexing of image and video databases by global and shape features,*"* in *Proc.* ICPR 96, IEEE International Conference on, vol. 3, pp. 140 – 144. 1996.

[2] Xinbo Gao; Hong Xin; Hongbing Ji, *"*A study of intelligent video indexing system,*"* in *Proc.* Intelligent Control and Automation, vol. 3, pp. 2122 – 2126, 2002.

[3] Jin, H.; Yoshitomo, Y.; Sakauchi, M., "Detection of information relating to building object in news video", in *Proc.* ICIP 99, International Conference on, vol.3, pp. 329 – 333, 1999.

[4] O'Connor, N.; Czirjek, C.; Deasy, S.; Marlow, S.; Murphy, N.; Smeaton, A., "News story segmentation in the fishlar video indexing system", in *Proc*. Image Processing, International Conference on, vol. 3, pp. 418 – 421, 2001.

[5] H. Kim, J. Lee, J. Yang, S. Sull, W. Kim and S. M. Song, "Visual rhythm and shot verification," Multimedia Tools and Applications, vol. 15, pp. 227-245, 2001.