2003년도 대한전기학회 하계학술대회 논문집 2003.7.21-23

시간적 정보를 이용한 비디오에서의 효과적인 캡션 검출 알고리즘

김 수 연, 신 청 호, 권 철 현, 박 상 희
연세대학교 전기전자공학과

Efficient Caption Detection Algorithm Using Temporal Information in Video

Su-Yeon Kim, Chung-Ho Shin, Chul-Hyun Kwon, Sang-Hui Park
Dept. Of Electrical & Electronic Eng. , Yonsei Univ

Abstract - 이 논문은 연속적인 비디오 영상에서 시간적인 정보를 최대한 이용하는 새로운 캡션검출과 인식 알고리즘을 제안하였다. 누적된 차영상 정보로부터 비디오에서 캡션의 시공간적인 위치를 찾아내기 위하여 구문 등록 기술을 이용하였다. 그리고 복잡한 배경 영상의 문제를 해결하기 위하여 새로운 다중 프레임 인티그레이션 방법을 이용하였다. 기존 논문과는 달리 빠른 속도의 수행을 위하여 복잡한 계산 과정을 포함하지 않는다. 본 논문에서 제안한 방법은 다양한 뉴스 데이터 영상에서 적용되었고, 그 결과는 아주 정확하고 효과적이었다.

1. 서 론

최근에 디지털 비디오의 사용이 빠르게 증가함으로써, 비디오에서의 내용을 기반으로한 색인에 대한 연구가 활발히 진행되고 있다. 비디오 프레임위에 나타나는 캡션은 어떤 영상에 대한 내용을 직접적으로 표현하기 때문에 비디온 색인 및 검색을 향상시키는데 유용한 정보를 제공한다. 예를 들면, 뉴스에서의 캡션은 지명, 관계된 사람의 이름, 뉴스 내용이 내포하는 주제 등을 포함하고 있다. 비디오 캡션 검출 방법을 연구한 많은 논문들이 있었지만, 대부분 접근하는 방법이 시간적인 정보 없이 단순히 각각 비디오 프레임에서 공간적인 정보를 이용하여 캡션을 검출하였다[1-4]. 그 중에는 시간적 정보를 이용한 것도 있었지만, 단지 다중-프레임 인티그레이션 (multi-frame integration) 방법은 캡션을 강화시키는데에만 사용되었다[2]. Tang et al은 프레임간의 차이(frame difference)로부터 추출된 특징을 이용하여 캡션의 변위가 일어나는 프레임을 찾아냈다. 그러나 캡션이 나타나는 것을 찾기 위한 주요 기준이 프레임 차이이기 때문에 비디오 시퀀스에서 대상이 움직임이 크거나, 카메라의 이동이 빠르면 캡션을 찾는데 실패할 확률이 높다.

우리는 이 논문에서 Tang et al의 접근 방법과는 다르게 캡션이 나타나고 사라지기까지의 시간적인 연속성에 초점을 두었다. 첫째로, 구문이 나타나는 맵(text appreance map)을 만들기 위하여 구문 등록 기술(text registration technique)을 사용하였다. 그 다음으로 시공간적인 캡션의 위치는 기존 방법을 사용하여 맵으로부터 추출하였다. 마지막으로 다중 프레임 인티그레이션 방법으로 캡션을 강화시키고, 캡션을 이진화시키는 방법을 이용하였다.

2.1 캡션 검출

2.1 캡션 등록 기술

캡션 검출 알고리즘의 기본적인 아이디어는 변화를 찾아내는 것이다. 그러나 Tang et al과는 다르게 우리는 캡션 검출을 위한 판단 기준을 두 개의 연속적인 프레임의 차이에 두지 않았다. 대신에 비디오 시퀀스로부터 캡션이 나타난 픽셀을 등록하고 그 캡션이 얼마나 오랫동안 지속적으로 나타나는지에 대한 정보를 구성하였다. 그림 1에서 나타내듯이 어떤 픽셀이 이전 프레임과 큰 차이를 가지며, 일정 기간의 프레임 동안 그 값이 시간적으로 지속되면 캡션을 나타내는 픽셀이라고 가정한다. 이러한 정보는 기존 방법보다 확실하며, 대상의 움직임이나 카메라의 이동에 덜 민감하다. 이를 위해, 알고리즘 안에 메모리를 구성하였다. 이 메모리는 각각의 픽셀에 대하여 이전 프레임들로부터 캡션이라고 가정된 횟수를 나타낸다. 이 메모리는 프레임 차이에 대한 히스토리(history)로 구성 및 갱신된다.

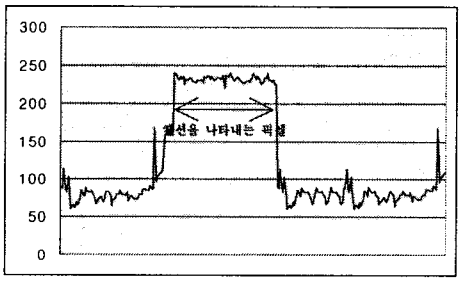


그림 1 . 매 프레임에서 고정된 임의의 픽셀값

표 1은 위에서 말한 캡션 검출을 위한 기준을 나타낸 목록이다. 표에서 MEM_k 는 k번째 프레임에서의 메모리를 의미하고, |FD|는 이전 프레임과 현재 프레임 차이에 대한 절대값을, L은 캡션이 나타나서 사라지기까지 지속되는 값에 대한 문턱값을, TH_{FD}는 변화를 찾아내기 위한 픽셀에 대한 문턱값을 의미한다. 문턱값을 구하기 위하여 T. Aach의 방법을 사용했다[6]. 초기에 MEM_0의 모든 값은 0으로 만들어준다. 다음으로 기존 방법[5]을 이용하여 각각의 프레임에서 장면 전환이 일어나는지 아닌지를 검사한다. 장면이 전환하면 많은 프레임 차이가 생기지만 이러한 프레임 차이는 캡션으로 인한 것이 아니다. 그러나 장면이 전환할때에도 캡션이 나타난 곳의 차이값은 변하지 않기 때문에 장면 전환이 일어나는 경우 시간적인 연속성만 고려해 주면 된다. (3에서 5번까지의 경우)

표 1 캡션 지역 결정 기준

Case	Shot Boundary	Frame Difference	Previous Memory	Current Memory	Description
1	No	$\|FD\| \le TH_{FD}$	MEM=0	MEM=0	Not Text
2	No	$\|FD\| > TH_{FD}$	MEM≤L	MEM=1	Candidate
3	YES	$\|FD\| > TH_{FD}$	MEM≤L	MEM=0	Not Text
4	ANY	$\|FD\| \le TH_{FD}$	MEM>0	MEM+=1	Candidate
5	ANY	$\|FD\| > TH_{FD}$	MEM>L	MEM=0	Registration

표 1에서 1의 경우, 프레임 차이에서 픽셀값에 주요한 변화가 생기지 않고 MEM이 가지는 값이 0이면, 변화는 생기지 않는다. 2의 경우는 프레임 차이가 TH_{FD}의 문턱

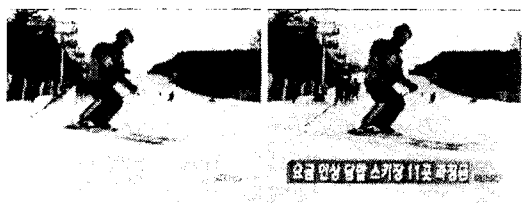
값보다는 높지만, 일정한 기간동안 지속되야 할 시간적인 연속성이 L보다 적기 때문에 픽셀은 초기의 후보(candidate)로 재설정되거나 초기값(=0)으로 설정된다. 즉, 장면 전환이 이루어지지 않은 프레임에 있는 픽셀이 프레임 차이에서 변화가 생긴 것으로 나타나고 MEM에서 나타내는 값이 L보다 작다면, MEM을 1로 만든다. 3의 경우는 장면 전환에 대한 조건을 제외하면 2의 경우와 유사하다. 2와는 다르게 프레임 차이가 장면 전환으로부터 생긴것이기 때문에 MEM의 값은 0으로 초기화한다. 4의 경우, 후보 상태였던 픽셀값의 차이가 현재 프레임에서도 작다면 MEM의 값은 1씩 증가한다. 5의 경우에는 위에서 설명한 모든 상태를 만족한 것으로 캡션을 나타내는 픽셀로 등록된다.

그 다음으로, 계산상의 복잡성을 줄이기 위하여 각각의 프레임에서 캡션으로 등록된 픽셀값의 수를 센다. 만약 하나의 프레임에서 등록된 픽셀의 수가 정해진 문턱값보다 클 경우 그 정보는 아랫식을 이용하여 Text Apperance Map(TAM)으로써 저장된다.

$$TAM_{k-MEM_k}(i,j) = MEM_k(i,j)$$

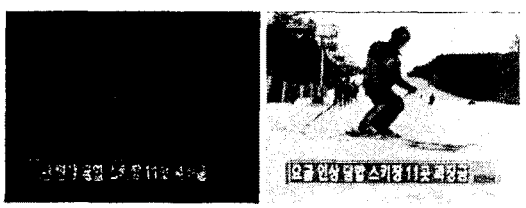
TAM_k는 k번째 프레임에서 등록된 캡션에 속해 있는 픽셀에 대한 정보를 나타낸다. TAM은 메모리에 저장된 픽셀의 값이 얼마나 많은 프레임동안 캡션으로 지속되는지를 나타낸다. 캡션에 등록되지 않은 것에 대해서는 TAM의 값을 0으로 준다. 그림 2는 캡션으로 등록된 픽셀 값들에 대한 결과를 보여준다. 그림 2의 (d)는, 프레임간의 차이에서 주요한 변화가 발생하였을 경우를 하얀색 픽셀로 표현한 이진맵이다. 그림 2의 (e)는 TAM에서 캡션이 나타나는 부분을 표시한 것이다.(224개의 프레임 동안 캡션이 지속됨)

만약 등록된 픽셀수가 전체 프레임의 1/3이 넘어가면 못찾은 장면 전환 때문이므로 그 프레임은 버린다.



(a) 원영상(709th Frame) (b) 원영상(711th Frame)

(c)프레임 차이 (d) 주요 변화 맵



(e) TAM at k=711 (f) 결과 위치

그림 2. 캡션 검출 방법의 실험결과

2.2 캡션 라인 위치 추정

이 단계에서는 이전의 섹션으로부터 얻은 TAM으로부터 캡션의 시공간적인 위치를 찾아낸다. 조악한 프로젝션 프로파일(Coarse Projection Profile) 방법을 사용한 후에 [3]의 지역 분해 방법을 적용하여 TAM을 캡션 박스로 만든다. 그 후에, 각각의 캡션 박스에 대해서 TAM이 0인 것을 제외하고 히스토그램을 계산한다. 이 히스토그램은 캡션 박스가 정확하다면 캡션 박스에 등록된 대부분의 픽셀이 동시에 사라질 것이기 때문에 캡션 박스에서 캡션이 사라질 때를 발견하고 캡션 박스가 아닌 것들을 제거하기 위해 계산된다. H(n)은 히스토그램의 최대값으로 가정한다. 만약 히스토그램의 H(n)의 비율이 문턱값보다 크다면, 캡션은 'n' 프레임 이후에 사라진다고 가정한다. 그렇지 않으면 그 캡션 박스는 버려진다. 그림 2의 (f)는 캡션 위치 추정의 한 예이다.

3 캡션 강화와 인식

비디오 영상에서는 낮은 해상도와 복잡한 배경 문제 때문에 캡션 인식이 어렵다. 따라서 이를 해결하기 위해 다중 프레임의 평균을 취하거나 최소(혹은 최대) 픽셀값을 찾는 방법이 일반적으로 사용된다. 이 두 방법의 키 포인트는 같은 캡션을 갖는 캡션 블록을 최대한 찾아내는 것이다. 제안된 방법은 같은 캡션을 가지고 있는 캡션 라인을 최대한 찾아낼 수 있다. 그림 3(b)는 다중 프레임 평균 방법을 사용하여 캡션을 강화시킨 예이다.

낮은 해상도에 대해서는 우리는 서브 픽셀 보간 기술[2]을 이용하여 캡션 라인에 대한 낮은 해상도를 확장했다. 그리고 등록된 픽셀의 평균값을 고정된 문턱값으로 이용하여 보간 이미지를 이진화했다(그림 3.(c)). 그림 3(d)는 그림3(c)를 이용하여 인식한 결과를 보여준다.

(a) 2번째 캡션 라인

(b) 프레임 평균값

(c) 보간후 이진화한 문자

요금 인상 담합 스키장 1 1곳 과징금

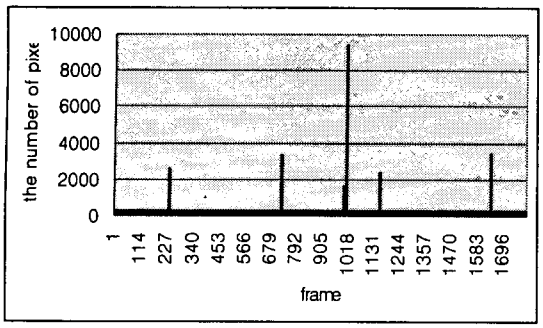
(d) 문자 인식 결과

그림 3. 캡션 강화와 이진화의 예

4. 실험과 결과

제안한 방법의 정확성을 검토하기 위하여 실험을 수행하였다. 실험 데이터는 한국의 KBS와 일본의 NHK 뉴스 프로그램을 사용하였다. 전체 5시간의 뉴스 프로그램은 MPEG-1으로 코딩한 320×240의 해상도를 가지는 영상을 사용했다. 이 비디오 데이터는 2039개의 캡션라인을 포함하고 있다. 또한 모든 프레임을 사용한 반면에 처리시간을 줄이고, 점진적으로 나타났다가 사라지는 캡션을 찾아내기 위하여 매 2프레임당 1개의 프레임을 사용하였다. 제안한 캡션 검출 방법의 결과는 표 2와 같으

며, 캡션이 나타난 부분에 대한 그래프는 그림 4와 같다.



제안된 방법을 이용하여 전체 캡션라인의 97%이상을 찾아냈고 찾아내지 못한 캡션은 2%에 불과했다. 제안된 방법은 다른 방법과는 다르게 점진적으로 나타나고 사라지는 캡션과 카메라 플러시의 빛에 대해서도 정교하게 다룰 수 있다. 대부분 못찾은 캡션라인은 움직이거나 확대나 축소와 같은 특별한 그래픽 효과가 나타난 지역에서 발생하였다.

표 2. 캡션 라인 검출의 결과

	Total	Detected	Fals Alarm	Precision	Recall
Caption Line Detection	2039	1981	41	0.98	0.97

인식 실험은 792개의 문자를 포함하는 100개의 캡션라인을 매뉴얼하게 선택하여 사용하였다. 문자인식에는 ARMI PRO(ver 6.0), OCR 패키지등을 사용하였다. 우리의 접근방법을 사용하면 이진 문자에 대한 84%의 평균 인식 비율을 얻을 수 있다.

3. 결 론

이 논문은 비디오 시퀀스로부터 인티그래이트한 시간적 정보를 이용하여 캡션을 검출하는 알고리즘을 제안했다. 캡션이 나타나서 사라질 동안 캡션의 프레임 차이가 작다는 점을 이용하여 구문 등록 기술을 발전시켰으며, 프레임 차이에 대한 히스토리를 가지고 동적 메모리를 갱신하였다. 캡션 라인은 TAM으로부터 쉽게 찾을 수 있었으며, 실험 결과는 정확하고 효과적이었다.

[참 고 문 헌]

[1] H. Li, D.Doermann, and O. Kia, "Automatic Text Detection and Tracking in Digital Video", IEEE Trans. on Image Processing, Vol. 9, pp. 147-256, Jan. 2000
[2] T. Sato, T. Kanade, E. K. Hughes, and M. A. Smith, "Video OCR for Digital News Archives", in Proc. IEEE Int. Workshop on Content-Based Access of Image and Video Database(CAVID'98), pp. 52-60, 1998
[3] X. S. Hua, X. R. Chen, L. Wenyin, H. J. Zhang, "Automatic Location of Text in Video Frames," Proceeding of ACM Multimedia 2001 Workshops: MIR2001, pp. 24-27, Ottawa, Canada, October 5, 2001.
[4] Y. Zhong, H. Zhang, and A. K. Jain, "Automatic Caption Localization in Compressed Video", IEEE Trans. on PAMI, Vol. 22, No. 4, pp. 385-392, April 2000.
[5] X. Tang, X. Gao, J. Liu, and H. J. Zhang, "A Spatial-Temporal Approach for Video Caption Detection and Recognition", IEEE Trans. on Neural Network, Vol. 13, No. 4, pp. 961-971, July 2002
[6] T. Aach, A. Kaup, and R. Mester, "Statistical model-based change detection in moving video", Signal Processing, Vol. 31, pp. 165-180, Mar. 1993.