ISSN(Print): 2287-5905

http://dx.doi.org/10.3745/KTSDE.2012.1.2.115

Detection of Artificial Caption using Temporal and Spatial Information in Video

SungIl Joo† · SunHee Weon†† · HyungIl Choi†††

ABSTRACT

The artificial captions appearing in videos include information that relates to the videos. In order to obtain the information carried by captions, many methods for caption extraction from videos have been studied. Most traditional methods of detecting caption region have used one frame. However video include not only spatial information but also temporal information. So we propose a method of detection caption region using temporal and spatial information. First, we make improved Text-Appearance-Map and detect continuous candidate regions through matching between candidate-regions. Second, we detect disappearing captions using disappearance test in candidate regions. In case of captions disappear, the caption regions are decided by a merging process which use temporal and spatial information. Final, we decide final caption regions through ANNs using edge direction histograms for verification. Our proposed method was experienced on many kinds of captions with a variety of sizes, shapes, positions and the experiment result was evaluated through Recall and Precision.

Keywords : Artificial Caption, Text Appearance Map, Caption Detection

시·공간 정보를 이용한 동영상의 인공 캡션 검출

주 성 일† · 원 선 희†† · 최 형 일†††

요 약

동영상에 포함되는 인공 캡션은 영상과 관계있는 의미정보를 포함한다. 이러한 영상을 표현하는 정보를 이용하기 위해 캡션을 추출하는 연구는 근래에 들어 활발히 진행되고 있다. 기존 방법들은 대부분 정지영상에서 캡션을 검출하였다. 하지만 동영상의 경우에는 유용한 시간정보가 있다. 따라서 본 연구는 이러한 시간정보를 사용한 캡션영역 검출방법을 제안한다. 먼저, 캡션후보영역 검출을 위해 문자출현맵을 생성하고, 후보영역 매칭 과정에서 지속후보영역을 검출한다. 검출된 지속후보영역의 소멸성 검사를 통해 캡션의 소멸 여부를 검출하고 소멸된 캡션 일 경우 시·공간정보에 의한 병합과정을 통해 캡션후보영역을 결정한다. 마지막으로 결정된 캡션후보영역을 검증하기 위하여 에지 방향 히스토그램을 이용한 신경망 인식기를 통하여 최종캡션영역을 검출한다. 실험을 위해 다양한 크기와 형태, 위치의 캡션을 포함하는 동영상에 대해 영역검출의 성능을 평가하고자 Recall과 Precision을 이용하여 제안하는 방법의 영역검출에 대한 효율성을 입증한다.

키워드 : 인공 캡션, 문자출현맵, 캡션 검출

1. 서 론

최근 IPTV의 도입으로 인하여 대용량 멀티미디어 콘텐츠를 시간에 구애받지 않으며 시청할 수 있게 되었다. 이러한 이유로 사용자는 수많은 대용량 멀티미디어 콘텐츠 중에서 원하는 콘텐츠를 시청하기 위하여 검색 기능을 요구하게 되었고, 이러한 요구에 따라 기존 주석기반의 색인 방식의 문제점을 극복하기 위해 멀티미디어 데이터의 특성을 자동으로 추출하고 이를 기반으로 검색을 하는 내용기반검색이 중요한 이슈로 떠오르게 되었다. 대용량 멀티미디어 콘텐츠에 내용기반검색을 위한 가장 신뢰성 있는 정보는 캡션이다. 대부분의 캡션의 경우 동영상에서 그 시점에 대한 상황을 부가적으로 설명하거나 중요한 정보를 시청자에게 적극적으로 전달하기 위함이기 때문이다. 이러한 이유로 캡션검출 및 인식에 관한 많은 연구가 진행되어왔다.

캡션검출연구는 크게 세 가지 방법으로 구분할 수 있다. 첫 번째로, 영상의 구성 요소에 기초하여 텍스트 영역을 검출하는 방법이 있다[1~4]. 이 방법은 영상을 색상 양자화와 연결요소 분석을 통하여 영역을 추출한다. 추출된 영역마다

※ 이 논문은 서울시 산학연 협력사업(SS110013)의 지원을 받아 수행된 연구임.
† 준 회 원 : 숭실대학교 미디어학과 박사과정
†† 정 회 원 : 숭실대학교 미디어학과 Post Doc
††† 종신회원 : 숭실대학교 미디어학과 교수
논문접수 : 2012년 5월 31일
심사완료 : 2012년 8월 2일
* Corresponding Author : HyungIl Choi(hic@ssu.ac.kr)

크기, 모양, 배치 정보 등으로 조건에 맞는 영역들을 추출하여 텍스트 영역을 검출한다. 이 방법은 복잡하지 않은 영상에서 효과적이지만 텍스트 영역이 유사한 색상으로 이루어져야한다는 제약조건이 있다. 다음은 텍스트의 텍스쳐를 이용한 방법이다. 이 방법은 텍스트 영역을 특정한 타입의 특징으로 이루어진 것으로 간주하고 공간 분산(Spatial Variance)[5], 웨이블릿[6,7], DCT[8] 등으로 특징을 추출하여 텍스트 영역을 추출한다. 이러한 방법은 구성요소 기반 검출 방법보다 좀 더 고수준의 특징으로 좋은 성능을 보이나 배경이 복잡하여 텍스트와 유사한 형태를 갖는 영역의 경우 문제가 발생한다. 또한 복잡한 연산으로 인하여 동영상과 같은 멀티미디어에 적용하기에는 효과적이지 않다. 세 번째는 에지를 이용한 방법이다[9~11]. 문자들은 많은 직선 형태의 영역으로 이루어져 있으며 시각적인 전달을 목적으로 함으로 뚜렷하고 많은 에지를 가지고 있다. 일반적으로 소벨(Sobel) 필터를 이용하여 수평 방향과 수직 방향의 에지를 검출한 후, 임계값 이상의 화소만을 구한다. 이러한 에지를 이용하여 에지 밀도, 에지 크기 또는 형태학적 연산(Morphology Operation)등을 이용하여 텍스트 영역을 검출한다.

대부분의 캡션 검출에 관한 연구는 뉴스와 같은 정형화된 문자영역을 추출하거나 자막이 아닌 정지 영상에서 텍스트 영역을 검출하는 연구가 대부분이다. 즉, 동영상의 유용한 특징인 시간정보를 사용하지 않고 공간적인 특징 또는 형태적인 특징만을 이용하여 검출하는 방법이 대부분이었다. 그러나 본 연구와 유사하게 시간정보를 이용한 연구도 있었다. 캡션은 임의의 시점에 나타나 일정시간 존재하다 동시에 사라지는 특성을 이용하여 문자출현맵(Text Appearance Map)을 구성하고 구성된 문자출현맵의 값을 이용하여 캡션을 검출하는 방법이다[12]. 하지만 이 방법 역시 뉴스 영상에만 적용하여 실험하였으며 장면전환을 기준으로 문자출현맵을 생성하므로 장면전환이 오검출 될 경우에 문제점이 발생한다. 또한 장면전환과 동시에 캡션이 나타날 경우에는 미검출되는 문제점이 있다.

본 연구에서는 이러한 문제점들을 해결하고자 개선된 문자출현맵을 구성하고, 동영상에서의 시.공간 정보를 이용하여 캡션영역의 검출과 병합을 통해 최종적인 인공 캡션영역 검출 방법을 제안하고자 한다[1].

2. 관련 연구

문자출현맵을 이용한 캡션검출의 기본 개념은 화소값의 변화를 찾아 변화 후 일정시간동안 지속성을 유지하는 화소를 검출하는 것이다[12]. Table 1은 임의의 위치에 있는 화소의 캡션영역 포함여부를 판단하기 위한 기준이다.

기본적으로 차영상과 장면전환, 이전 시점의 메모리의 값을 판단기준으로 하여 비캡션영역과 후보영역을 분류한다. $|FD|$는 프레임 간 차이의 절대값이며, L은 사전 정의된 수, TH_{FD}는 변화 탐지를 위한 임계값이다. 또한 MEM은 캡션 후보영역과 비캡션영역을 판단하기 위한 메모리이다. 예를

Table 1. The condition of caption region[12]

No.	장면 전환	프레임 차이	이전 메모리	현재 메모리	비고
1	No	$\|FD\| \le TH_{FD}$	$MEM=0$	$MEM=0$	비캡션 영역
2	No	$\|FD\| > TH_{FD}$	$MEM \le L$	$MEM=1$	후보 영역
3	Yes	$\|FD\| > TH_{FD}$	$MEM \le L$	$MEM=0$	비캡션 영역
4	Any	$\|FD\| \le TH_{FD}$	$MEM>0$	$MEM+=1$	후보 영역
5	No	$\|FD\| > TH_{FD}$	$MEM>L$	$MEM=1$	캡션 등록
6	Yes	$\|FD\| > TH_{FD}$	$MEM>L$	$MEM=0$	캡션 등록

들어 Table 1에서 두 번째 경우는 장면전환이 일어나지 않은 시점에 어떤 화소의 위치에서 차연산 결과가 TH_{FD}이상이며 MEM이 L보다 작거나 같으면 MEM의 화소 위치에 대응되는 공간에 1로 설정하는 것이다.

이렇게 매 프레임마다 MEM을 획득하여 문자로 등록된 총 화소들의 개수를 구한다. 만약 등록된 화소의 합이 T_r을 초과하면 문자출현맵(TAM)에 대응되는 위치에 복사된다. 이렇게 TAM에는 캡션으로 등록된 화소들의 정보를 가진다. 기존 방법[12]에서는 TAM을 이용하여 프로젝션 프로파일(Projection Profile)과 지역 분해 방법을 사용하여 최종 캡션영역을 검출한다.

Fig. 1은 시간에 따른 다양한 캡션의 형태를 보여준다. 하나의 장면 안에 포함되는 캡션, 장면전환과 동시에 소멸되는 캡션, 장면과 장면 간에 연속되는 캡션 등 다양한 형태로 존재한다. Fig. 1의 숫자는 Table 1의 판단 기준의 경우의 수를 나타낸다.

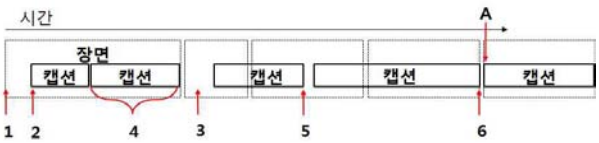

Fig. 1. The various type of caption

대부분의 경우에 Table 1의 판단 기준으로 검출이 가능하지만 Fig. 1의 A의 경우에는 캡션을 검출하지 못하는 상황이 발생한다. A의 시점은 장면전환과 동시에 캡션이 나타나는 경우이다. A 시점의 캡션위치의 화소는 캡션이 나타났기 때문에 차연산이 임계값(TH_{FD})보다 클 것이며 장면전환 시점이므로 3번과 6번의 경우에 해당한다.

그러나 두 경우 모두 MEM 값을 0으로 설정하므로 이후 변화가 없어도 4번 조건에 포함되지 않아 해당 캡션은 검출하지 못한다. 또한 캡션등록 조건에 해당하는 L값보다 캡션의 지속 프레임수가 작을 경우에도 검출하지 못한다. 뉴스와 같이 비교적 중요한 정보전달을 목적으로 하는 경우에는

다른 동영상에 비하여 캡션의 지속횟수가 길지만, 지속횟수가 상대적으로 짧은 캡션을 포함하는 동영상의 경우에는 정확한 검출을 하기 어려우며 L값을 작게 하면 오검출의 비율이 높아져 이 방법만으로는 정확한 검출이 어렵다. 본 논문에서는 이러한 단점을 보완하기 위해 캡션 판단 기준을 개선하고 추가적으로 캡션후보영역을 판단하는 단계와 시간과 공간적인 제약조건으로 두 번의 병합과정, 신경망을 이용한 캡션 검증과정을 통해 다양한 형태의 인공 캡션을 검출하는 방법에 대하여 제안하고자 한다.

3. 제안하는 방법

본 장에서는 기존 문자출현맵의 문제점을 해결하기위해 판단기준을 개선하고 개선함으로써 발생되는 오검출을 제거하여 보다 효율적이며 다양한 캡션을 검출할 수 있는 알고리즘에 대하여 제안한다.

Fig. 2은 제안한 인공 캡션검출 방법의 흐름도이다. 먼저 동영상에서 프레임을 추출하고 가우시안 마스크를 이용한 블러링을 통해 잡음을 제거하는 전처리 단계를 수행한다. 전처리된 프레임을 이전 프레임과 차이를 계산하여 차영상을 획득하고 개선된 판단기준을 이용하여 문자출현맵을 갱신한다. 다음으로 문자출현맵을 연결요소분석(Connected Component Analysis)과 에지 검사과정을 통하여 1차 후보영역을 검출하고 후보영역 매칭 단계로 입력된다. 입력된 1차 후보영역과 이전 시점에서 획득한 지속후보영역과 매칭을 통하여 갱신, 추가, 제거 작업을 수행한다. 이전시점에 획득한 지속후보영역 중에서 캡션소멸성 검사를 통하여 캡션후보영역을 판단하고 캡션후보영역으로 판단된 영역은 영역정보와 함께 캡션후보영역 저장 공간에 기억된다. 마지막으로 캡션후보영역으로 판단된 영역들을 시간정보의 유사성을 이용하여 1차적으로 군집화하고, 군집화 된 각각의 집합을 공간정보의 유사성을 검토하여 최종병합한다. 병합된 캡션영역은 에지 방향 히스토그램을 이용하여 특징벡터를 추출하고 신경망을 통한 검증단계를 거쳐 최종캡션영역으로 검출된다.

3.1 전처리

프레임이 입력되면 먼저 가우시안 마스크를 이용하여 블러링을 수행한 후 이전 프레임과의 차영상을 획득한다. 획득한 차영상은 형태학적 연산(Morphology Operation)의 닫힘(Closing) 연산과 열림(Opening) 연산을 차례로 수행한다. 이는 동영상이 대부분 압축 도메인이기 때문에 단순한 차연산 결과를 보정하고 천천히 나타나는 캡션의 경우 화소별 영향을 최소화하기 위함이다. 하지만 이러한 형태학적 연산으로 인해 실제 변화가 없는 부분에서도 차연산이 발생하게 되는데 이러한 문제는 다음단계에서 해결한다.

3.2 개선된 문자출현맵

전처리를 수행한 입력 프레임과 차영상이 주어지면 문자출현맵을 획득할 수 있다. 다음 Table 2은 기존의 자막 테스트 판단 기준을 개선한 문자출현맵 생성 기준이다.

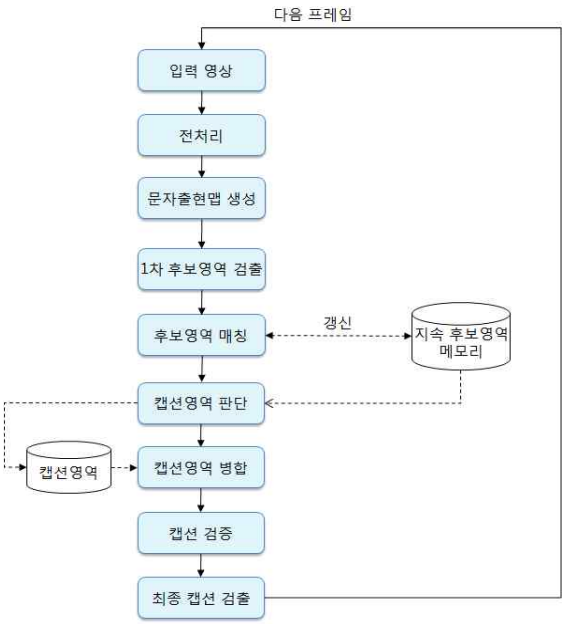

Fig. 2. The flowchart of artificial caption detection

Table 2. The condition of improved Text Appearance Map(TAM)

번호	프레임차이	변화누적 영상차이	이전 문자출현맵	현재 문자출현맵
1	$\|FD\| \leq TH_{FD}$	Any	$TAM=0$	$TAM=0$
2	$\|FD\| \leq TH_{FD}$	Any	$TAM\neq 0$	$TAM+=1$
3	$\|FD\| > TH_{FD}$	$\|CD\| \leq TH_{CD}$	$TAM\neq 0$	$TAM+=1$
4	$\|FD\| > TH_{FD}$	$\|CD\| > TH_{CD}$	Any	$TAM=1$

Table 2의 $|FD|$는 전 프레임과 현재 프레임 차이 값이며, $|CD|$는 변화누적영상과 현재 프레임의 차이값이다. 변화누적영상이란 차영상에서 발생한 화소들과 현재 프레임에서의 해당 화소들과의 차이값으로 생성한 영상이다. 이 영상을 통해 이전 시점의 누적된 프레임과 현재 프레임에서의 해당 화소의 변화정도를 판단할 수 있다. TH_{FD}와 TH_{CD}는 차연산의 임계값이다. TAM은 문자출현맵의 메모리이며 초기에는 모두 0으로 설정한다. 문자출현맵은 이전 프레임과 현재 프레임의 차이와 현재 프레임과 변화누적영상의 차이가 모두 임계값 이상이면 무조건 1로 설정되며 둘 중 하나라도 임계값을 넘지 못하고 TAM의 값이 0이 아니라면 1을 더한다. Table 2의 생성 기준은 장면전환을 고려하지 않기 때문에 장면전환검출 오차에 영향을 받지 않고 캡션 영역 검출이 가능하다. 하지만 너무 많은 영역들이 캡션영역으로 검출 될 수 있기 때문에 본 연구에서는 캡션후보영역 판단단계와 검증단계를 추가함으로써 오검출을 감소시켰다. 또한 Table 2의 3번과 4번 경우에는 현재 프레임과 변화누적영상의 차이를 재검사하게 된다. 이것은 전처리 단계에서의 형태학적 연산으로 인해 발생되는 오차를 보정하기 위함이다.

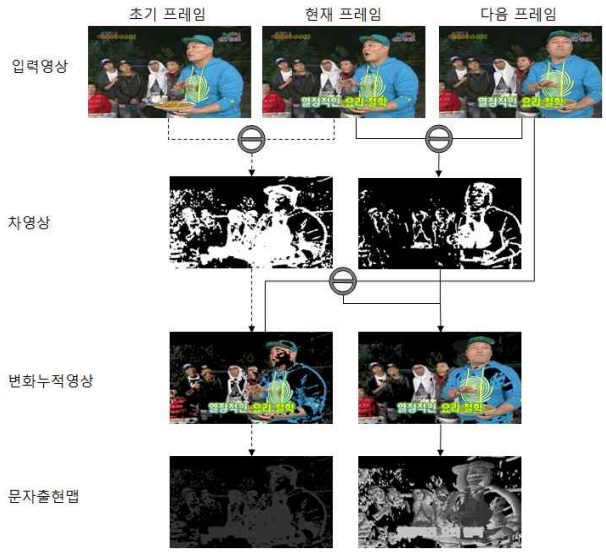

Fig. 3. The process for generating TAM

Fig. 3는 문자출현맵의 변화를 시간의 변화에 따라 영상으로 표현한 것이다. 초기 프레임에서의 차영상에서 움직임이 나타난 부분은 문자출현맵 생성 기준의 4번 경우에 의하여 $TAM=1$로 설정된다. 이후 일정 시간이 지나는 동안 캡션영역이 지속되어 문자출현맵의 값이 누적되었으므로 현재 프레임의 캡션영역이 밝은 것을 알 수 있다. 그러나 압축동영상에서 차연산을 수행하면 캡션영역에서도 간혹 임계값 이상의 변화가 일어난다. 또한 전처리 과정에서 형태학적 연산으로 차영상이 변형되었기 때문에 차연산이 일어나더라도 변화누적영상과 현재 프레임의 차이를 부가적으로 계산함으로써 이러한 문제점을 해결할 수 있다.

$$CC(x,y)=\begin{cases} Case\,4 & I_t(x,y) \\ Case\,2,3 & CC(x,y)*(1-\alpha)+I_t(x,y)*\alpha \end{cases} \quad (1)$$

식(1)은 변화누적영상의 생성 방법을 수식화한 것이다. 문자출현맵 생성 기준에 따라 (x,y)의 위치에서 2번과 3번 경우가 발생하면 변화누적영상 $(CC(x,y))$값의 $(1-\alpha)$배와 현재 프레임$(I(x,y))$의 α배의 합으로 갱신된다. 또한 4번 경우에는 캡션이 나타남을 의미하므로 변화누적영상의 화소값은 현재 프레임의 화소값으로 대체된다. 이렇게 생성된 변화누적영상은 매 프레임마다 문자출현맵을 갱신하는데 사용된다. 즉, 전처리과정에서 형태학적 연산을 통해서 변형이 이루어져 변화 없는 화소가 변화 있는 화소로 인지되는 오차가 발생할 경우 Table 2의 3번조건과 같이 현재 영상과 변화누적영상과의 차이를 추가적으로 계산함으로써 문제점을 해결한다.

3.3 에지정보를 이용한 1차 후보영역 검출

본 절에서는 문자출현맵을 이용하여 영역단위로 분리하고 특정 조건을 통하여 캡션 후보영역을 분류하는 1차 후보영역 검출 과정에 대해 설명한다. 1차 후보영역을 검출하기 위하여 연결요소분석 방법인 레이블링을 통하여 먼저 영역별로 그룹화한다. 이는 문자출현맵의 화소 값은 일정시점에 나타나 화소 값만큼 변화가 없었음을 의미하기 때문이다. 따라서 하나의 의미 정보를 갖는 캡션일 경우 동시에 나타났으며 사라지는 시점까지 변화가 없기 때문에 이상적으로는 문자출현맵에서 같은 화소 값을 갖는다.

$$TH_E=(FR^i_{Mer_W}+FR^i_{Mer_H})*r \quad (2)$$
$$Cnt^i_E=\sum_{v}^{v\in FR^i} Edge(v)$$
$$FR^i=\begin{cases} True & \text{if } Cnt^i_E \geq TH_E \\ False & otherwise \end{cases}$$

식(2)는 1차 후보영역 검출을 위한 조건이다. FR^i는 레이블링으로 연결된 i번째 영역이며 $FR^i_{Mer_W}$, $FR^i_{Mer_H}$은 i번째 영역의 최소인접사각형(Minimum Enclosed Rectangle)의 너비와 높이를 의미한다. Cnt^i_E는 i번째 영역에 속하는 에지 화소의 수이고, r은 상수로서 에지의 양을 결정하는 파라미터이다. 즉, 임계치 TH_E보다 Cnt^i_E의 값이 크다면 FR^i는 1차 후보영역으로 선택되고, 그렇지 않다면 제거된다.

Fig. 4과 같이 레이블링으로 연결된 영역 안에서 해당영역만을 분리하고 팽창연산을 수행한다. 그리고 현재 프레임에서 획득한 에지영상과 겹치는 영역의 화소들만 식(2)를 만족하는 1차 후보영역으로 검출한다. 문자출현맵에서 분리된 영역에 팽창연산을 수행하는 이유는 에지위치가 정확하게 일치되지 않을 수 있기 때문에 이러한 오차를 보정해주기 위함이다. 검출된 1차 후보영역은 다음단계인 후보영역 매칭단계에서 지속후보영역으로 추가되거나 지속후보영역의 정보를 갱신하는데 사용된다.

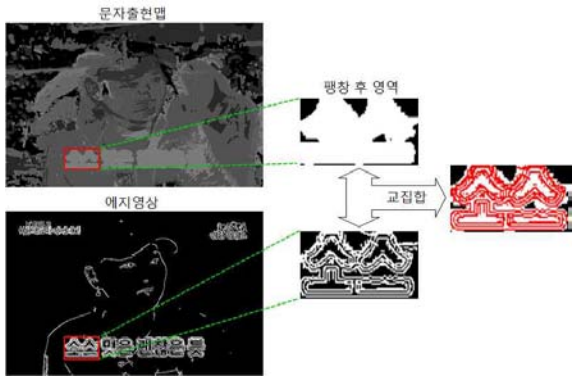

Fig. 4. The decision process of 1st candidate region

3.4 시·공간 정보를 이용한 후보영역 매칭

이 절에서는 검출된 1차 후보영역을 지속후보영역과 매칭 과정을 수행하여 지속후보영역으로 추가하거나 갱신 또는 매칭에 실패한 지속후보영역을 삭제하는 단계를 설명한다. 이 단계의 목적은 기존 문자출현맵에서는 Table 1의 자막 테스트 판단 기준만으로 캡션이 사라짐을 검출하였지만 본

연구에서는 시간 축으로 연결된 영역 단위로 사라짐을 검출하기 위함이다. 영역단위로 수행하기 때문에 화소 단위 연산보다 정확하게 캡션의 소멸성을 검사할 수 있다. 매칭과정은 1차 후보영역과 지속후보영역을 매칭하는데 초기에는 지속후보영역이 존재하지 않는다. 따라서 검출된 1차 후보영역은 모두 매칭에 실패하게 된다. 매칭에 실패한 1차 후보영역들은 T+1시점의 매칭을 위해 지속후보영역에 추가된다. 만약 지속후보영역 중에서 1차 후보영역과 매칭에 실패한 영역이 있다면 제거되며 매칭에 성공한 지속후보영역은 대응되는 1차 후보영역의 정보로 갱신된다.

$$FR^{i}_{TAM} - 1 = CR^{j}_{TAM} \tag{3}$$

$$FR^{i}_{MER} \cap CR^{j}_{MER} = FR^{i}_{MER} \tag{4}$$

식(3)과 식(4)는 매칭조건을 나타낸다. 먼저 식(3)은 시간정보를 이용한 매칭이다. FR^{i}_{TAM} 는 i번째 1차 후보영역의 지속횟수이고, CR^{j}_{TAM} 는 j번째 지속후보영역에서의 지속횟수를 의미한다. 지속횟수란 문자출현맵의 화소 값을 의미한다. 즉, 문자출현맵으로 레이블링한 영역들이 1차 후보영역이므로 연결요소분석에 사용한 값이다. 만약 제안한 알고리즘과 같이 차연산 결과에 따라서 어떤 임의의 화소가 변화가 없다면 1이 증가될 것이므로 이전에 갱신 또는 추가된 지속후보영역과 현재 프레임에서 검출된 1차 후보영역의 지속횟수는 1의 차이가 있을 것이다. 따라서 식(3)과 같이 시간적 정보가 일치하는 영역 간 매칭을 수행한다. 식(4)는 공간상의 정보를 이용한 매칭 조건식이다. FR^{i}_{MER} , CR^{j}_{MER} 는 i번째 1차 후보영역의 최소인접사각형과 j번째 지속후보영역의 최소인접사각형을 의미한다. 이 식(4)는 1차 후보영역이 지속후보영역에 포함되어야만 매칭을 수행하는 것이다. 왜냐하면 제안한 Table 2의 문자출현맵 생성기준으로는 영역이 시간에 지남에 따라 동일하거나 감소할 뿐 증가할 수는 없기 때문이다. 매칭에서 성공하여 갱신되거나 추가된 지속후보영역들은 다음 프레임 연산 과정에서 캡션의 소멸여부를 판단할 때 필요한 정보를 제공한다.

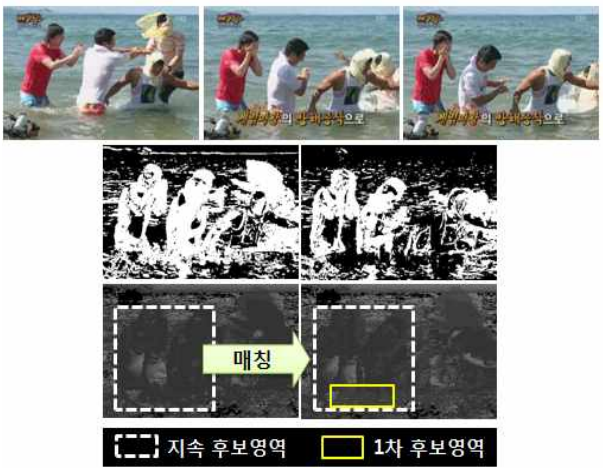

Fig. 5. Matching for continuedly candidate region and 1st candidate region

Fig. 5는 식(3)과 식(4)를 이용하여 매칭하는 과정을 나타낸다. 매칭에 성공한다면 1차 후보영역은 지속후보영역에 포함되며 지속횟수는 1의 차이가 발생한다. 이렇게 매칭에 성공하게 되면 지속후보영역은 매칭된 1차 후보영역의 영역정보, 지속횟수와 같은 정보들로 갱신되며 갱신된 지속후보영역은 다음 프레임에서의 매칭 과정에 사용된다.

3.5 지속후보영역의 캡션영역 판단

지속후보영역 중 캡션 여부를 판단하기 위해서는 하나의 가정이 필요하다. 지속횟수가 임계값 TH_{TAM} 이상인 영역에 대하여 다음 조건을 만족하면 캡션영역으로 판단하고 이후 병합 과정을 거쳐 최종 검증단계를 수행하게 된다.

$$IR^{i}_{cnt} = \sum_{v}^{v \in CR^{i}} Sub(v) \tag{5}$$

$$IR^{i}_{cnt} / CR^{i}_{cnt} > TH_{fo} \tag{6}$$

$$CR^{i}_{TAM} > TH_{TAM} \tag{7}$$

위의 식들은 캡션영역 여부를 판단하기 위한 조건으로 캡션의 소멸성 검사식이다. IR^{i}_{cnt} 는 i번째 지속후보영역에 포함된 영역 중 차연산이 발생한 화소의 수이고, CR^{i}_{cnt} 는 i번째 지속후보영역에 포함되는 화소의 수이며, TH_{fo}는 0~1 사이의 소멸성 검사를 위한 임계값이다. 또한 CR^{i}_{TAM}은 i번째 지속후보영역의 지속횟수를 의미한다.

Fig. 6는 위의 식들을 이용하여 캡션영역 판단 방법에 대하여 표현한 것이다. i번째 지속후보영역은 이전 프레임의 정보를 가지고 있기 때문에 이전 프레임의 정보를 이용하여 영역을 구성하고 현재 시점의 차영상을 이용하여 교차하는 화소의 수 IR^{i}_{cnt} 를 계산한다. 계산된 값을 이용하여 식(6)과 식(7)을 만족하는지 검사를 한다. 캡션은 정보전달용이므로 적어도 정보전달을 위한 시간동안은 유지되기 때문에 식(7)을 이용하여 너무 빨리 사라지는 영역을 제거한다. 이때 캡션영역이라 판단된 영역들은 3.4절에서 설명한 후보영역 매칭 대상에서 제외된다. 또한 캡션의 소멸성을 검사하여 사라졌다고 판단하였기에 최종캡션영역 검출을 위해 병합단계로 입력된다.

3.6 캡션영역 병합

캡션영역을 검출하였으므로 최종적으로 병합을 통해 하나의 의미가 있는 정보 단위로 영역을 분류하여야 한다. 의미정보로서 본 연구에서는 캡션 출현의 동시성을 사용하였다. 캡션에 따라 하나의 문자 단위로 몇 프레임동안 서서히 나타나거나, 하나의 의미 정보를 표현하기 위해서 문장 단위로 다른 시간에 나타나는 경우도 있다. 따라서 시간정보의 평균을 이용하여 일정 임계값 이하의 시간차이는 허용하도록 한다.

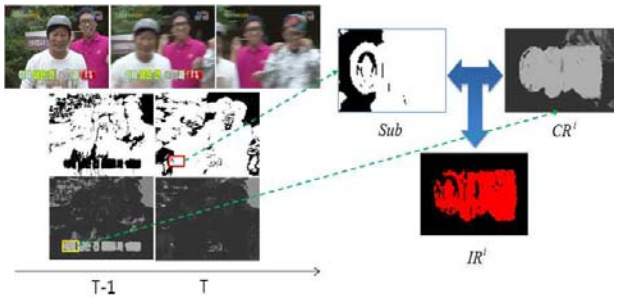

Fig. 6. The process of decision step for caption candidate region

1. 다음 조건을 만족하는 CR^j 선택

$$CR^j \notin Clust^i \qquad (i=0,\ldots,NC-1)$$

2. 새로운 집합 $Clust^{NC}$ 생성 후, $Clust^{NC}$ 에 CR^j 를 추가
3. $Clust^{NC}$ 의 평균지속횟수 계산

$$Clust^{NC}_{mean} = \frac{1}{N^{NC}} \sum_{j}^{CR^j \in Clust^{NC}} CR^j_{TAM}$$

4. 다음 조건 계산 후 만족할 경우 $Clust^{NC}$에 추가 후 3단계를 수행하고, 모든 CR^j에 대하여 다음 조건을 만족하지 않을 경우 5단계 수행

$$|Clust^{NC}_{mean} - CR^j_{TAM}| < TH_{MTAM} \qquad CR^j \notin Clust^{NC}$$

5. 다음 조건을 만족하면 종료, 만족하지 않으면 1단계부터 반복 수행

$$S = \bigcup_{i=0}^{NC-1} Clust^i$$

$$CR^j \in S \qquad \forall j$$

Fig. 7. Clustering algorithm with temporal information

Fig. 7는 시간정보를 이용한 군집화 알고리즘을 나타낸다. NC 는 현재 집합의 개수이며, $Clust^i$는 i번째 집합이다. N^i은 i번째 집합에 포함된 캡션후보영역의 수이다. 또한 CR^j_{TAM}, TH_{MTAM} 은 각각 j번째 지속후보영역의 지속횟수와 지속횟수 차이 허용을 위한 임계값을 의미한다.

먼저 1단계에서는 캡션후보영역으로 판단된 영역 중 어떤 집합에도 포함되지 않는 하나를 선택한다. 다음 새로운 집합을 하나 생성하고 그 집합에 선택된 캡션후보영역을 추가한다. 3단계에서는 현재 새로 생성된 집합에 포함되는 캡션후보영역들의 평균지속횟수($Clust^{NC}_{mean}$)를 계산한다. 계산된 평균지속횟수를 이용하여 4단계에서 현재까지 집합들에 포함되지 않은 캡션후보영역 중 현재 집합에 포함여부를 검사한다. 포함여부는 집합의 평균지속횟수와 차이를 계산하여 임계값(TH_{MTAM}) 이하 일 경우 집합에 포함시키고 3단계부터 다시 수행하며, 모든 캡션후보영역에 대하여 만족하지 않는다면 5단계를 수행한다. 5단계는 종료조건이다. 즉, 캡션후보영역으로 분류된 영역들은 모두 임의의 집합에 포함되어야 한다. 만약 하나의 캡션영역이라도 집합에 포함되지 않으면 1단계부터 다시 수행하여 새로운 집합을 생성하고 유사한 영역들을 포함시켜 군집화한다. 따라서 이 알고리즘은 군집화되는 집합의 수를 정의하지 않으며 캡션후보영역들에 따라서 하나의 집합 또는 다수의 집합으로 구성될 수 있다.

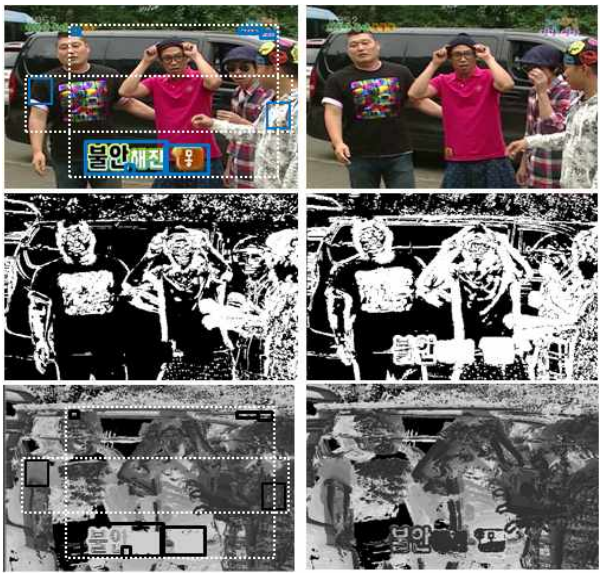

Fig. 8. The result of merging step for candidate region of caption using temporal information

Fig. 8은 시간정보를 이용하여 군집화하여 영역들을 병합한 결과를 나타낸다. 직선으로 표시된 사각형은 집합에 포함된 캡션후보영역을 나타내며 점선으로 표시된 사각형은 하나의 집합을 나타낸다. 그림과 같이 동시에 나타난 캡션은 하나의 집합에 포함된 것을 알 수 있다. 하지만 실제 캡션이 아닌 오검출부분도 함께 하나의 집합에 포함되어 있다. 이러한 문제점을 제거하기위해 2차적으로 공간정보를 이용하는 최종병합과정을 수행하게 된다.

이전 단계에서 시간정보만을 이용하여 병합을 수행하였다. 하지만 Fig. 8과 같이 의미적으로 전혀 무관한 영역 또한 같은 집합에 포함되었고, 이러한 문제점 해결을 위하여 다음과 같은 방법으로 최종병합과정을 수행한다.

Fig. 9은 공간정보를 이용한 병합 판단 기준을 나타낸다. U_{hw}, U_{hh}는 병합된 사각형 가로 길이의 반, 세로 길이의 반을 나타내며, C_{hw}, C_{hh}는 검사할 캡션후보영역 최소인접

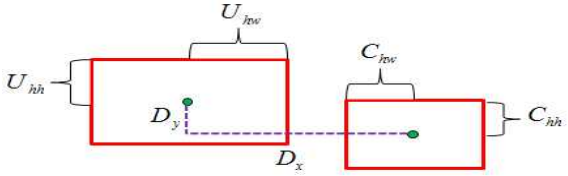

Fig. 9. Condition of merge for spatial information

사각형 가로 길이의 반, 세로 길이의 반을 나타낸다. D_x , D_y는 두 사각형의 중점 간의 거리를 나타낸다.

$$D_x - (C_{hw} + U_{hw}) < TH_w \tag{8}$$

$$D_y - (C_{hh} + U_{hh}) < TH_h \tag{9}$$

식(8)과 식(9)은 병합 조건이다. 두 식의 좌변이 절대값이 아닌 이유는 두 사각형이 겹칠 경우에 병합하기 위함이다. 만약 두 개의 사각형이 겹친다면 식(8)과 식(9)의 좌변은 음수가 되며 서로 떨어져있는 경우 x축, y축으로 근접하는 사각형의 두 변 사이의 거리가 나온다.

Fig. 10는 공간적인 정보를 이용하여 최종 병합하는 알고리즘이다. 이 알고리즘은 이전 단계에서 구한 집합들에 대해서 수행하는데 하나의 집합마다 적용한다. 즉, 하나의 집합이 입력되면 입력된 집합에 대해서 새로운 부분집합을 생성하여 최종 병합된 캡션영역을 검출한다. NSC는 공간정보를 이용하여 병합되는 $SClust$의 개수이며, $SClust^i$는 i번째 최종병합집합을 나타낸다. 알고리즘을 설명하자면 먼저 이전 단계에서 구성된 하나의 1차집합을 선택하고 그 집합 중에서 어떤 최종병합집합에도 포함되지 않는 하나의 캡션 후보영역을 선택한다. 2단계에서는 새로운 최종병합집합 $SClust^{NSC}$를 생성하고 선택된 캡션후보영역을 추가한다. 그리고 3단계에서는 최종병합집합에 포함되어 있는 캡션후보영역을 모두 병합하여 하나의 병합사각형을 생성한다. 이후 아직 어떤 최종병합집합에도 포함되지 않는 캡션후보영역을 선택하고 식(8)과 식(9)를 이용하여 병합 여부를 검사한다.

1. 다음 조건을 만족하는 CR^j 선택

$$CR^j \in Clust$$
$$CR^j \notin SClust^i \qquad (i = 0, \ldots, NSC-1)$$

2. 새로운 집합 $SClust^{NSC}$를 생성, $SClust^{NSC}$에 CR^j를 추가

3. $SClust^{NSC}$ 에 포함된 캡션후보영역들 병합

$$SClust^{NSC}_{Mer} = \bigcup_{j}^{CR^j \in SClust^{NSC}} CR^j_{Mer}$$

4. 다음 조건을 만족할 경우 $SClust^{NSC}$에 추가 후 3단계를 수행하고, 어떤 집합에도 포함되지 않는 모든 CR^j에 대하여 다음 조건을 만족하지 않을 경우 5단계 수행

$$D_x - (CR^j_{Mer_{hw}} + SClust^{NSC}_{Mer_{hw}}) < TH_w$$
$$D_y - (CR^j_{Mer_{hh}} + SClust^{NSC}_{Mer_{hh}}) < TH_h$$

5, 조건을 만족하면 종료, 만족하지 않으면 1단계부터 수행

$$S = \bigcup_{i=0}^{NSC-1} SClust^i$$

$$Clust \cap S = Clust$$

Fig. 10. The final merge algorithm using spatial information

만약 검사 조건에 만족하는 캡션후보영역이 있다면 최종병합집합에 추가하고 3단계부터 다시 진행하고, 어떤 최종병합집합에도 포함되지 않은 CR^j 중에서 조건을 만족하는 영역이 없다면 5단계를 수행한다. 마지막 5단계는 종료 조건으로서 현재 최종병합을 수행하고 있는 1차집합에 포함된 모든 캡션후보영역들이 최종병합집합에 포함되어있는지를 검사하여 모두 완료하였을 경우 종료한다. 만약 하나라도 최종병합집합에 포함되어있지 않으면 새로운 최종병합집합을 생성하여 모든 캡션후보영역이 포함 될 때까지 반복한다. 이 과정이 완료되면 아직 처리하지 않은 1차집합에 대하여 전체과정을 수행하며 모든 1차집합에 대해서 완료되면 종료한다.

Fig. 11. The final result of merging with spatial information

Fig. 11은 공간정보를 이용한 최종병합결과를 보여준다. 좌측 영상은 시간정보를 이용한 1차 병합된 하나의 집합이며, 우측 영상은 이 집합을 공간정보를 이용한 최종병합과정 후 결과를 보여준다.

3.7 캡션영역 검증

최종캡션영역이 검출되면 마지막으로 검증단계가 필요하다. 이전단계에서는 갑자기 나타나 일정시간 이상 변화가 없는 영역 중 에지의 수가 임계값 이상인 영역만을 검출하였다. 하지만 실제 캡션영역 뿐만 아니라 복잡한 배경과 다양한 상황이 발생하는 경우 다른 영역들도 검출 될 수 있다. 이러한 오검출을 막기 위하여 신경망을 이용한 검증단계를 수행한다.

검출되는 캡션영역과 비캡션영역의 가장 큰 차이는 에지의 방향과 크기이다. 비캡션영역의 경우 에지의 방향은 일정한 규칙 없이 나타나는 반면 캡션영역의 경우 대부분이 문자를 포함하고 있기 때문에 가로 또는 세로 방향의 에지가 다량 존재한다. 또한 캡션영역의 경우 에지의 크기가 비캡션영역의 에지보다 크다. 이는 인공적으로 삽입되었으며 시각적인 정보전달을 목적으로 하기 때문에 배경과 대조적으로 뚜렷한 경계를 포함하기 때문이다. 따라서 에지 방향 특징을 검출하고 신경망을 이용한 기계학습 과정을 수행하여 최종캡션영역 검증단계를 수행한다.

$$m(x,y) = \sqrt{(I(x,y+1) - I(x,y-1))^2 + (I(x+1,y) - I(x-1,y))^2} \tag{10}$$

$$\theta(x,y) = \tan^{-1}\frac{I(x,y+1) - I(x,y-1)}{I(x+1,y) - I(x-1,y)} \tag{11}$$

위 식(10)과 식(11)을 캡션후보영역의 특징추출을 위한 방법으로 $I(x,y)$는 캡션이 사라지기 바로 전 시점 영상의 화소값이다. 특징은 전경과 배경에서 각각 에지방향 히스토그램을 생성하여 추출한다. 먼저 병합과정으로 나온 영역 내에서 전경과 배경을 구분한다. 전경은 캡션후보영역의 최소인접사각형에서 캡션으로 판단된 화소이며 배경은 나머지 화소이다. 전경과 배경이 구분되면 먼저 전경에 대하여 식(11)을 이용하여 에지의 방향을 계산하고 계산된 값을 16등분하여 히스토그램 빈(bin)의 위치를 정한다. 히스토그램 빈의 위치가 정해지면 식(10)에서 구한 에지의 크기가 일정 임계값이상 일 경우에만 해당 빈에 누적한다. 이는 에지의 크기가 작은 방향성은 무의미하기 때문이다.

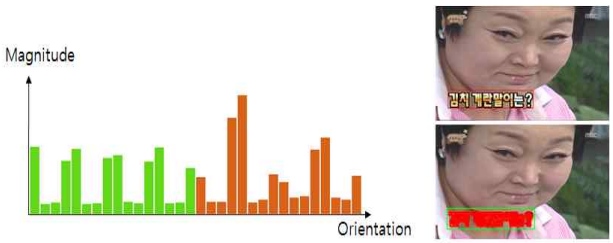

Fig. 12. Feature extraction using edge orientation histogram

Fig. 12은 위에서 설명한 방법을 이용하여 추출된 특징을 보여준다. 히스토그램의 좌측 16개의 빈에 해당하는 값은 전경에서 추출된 값이며 우측 16개의 빈은 배경에서 추출하여 누적된 값이다. Fig. 12에서 우측 하단 영상의 캡션 위에 덧칠해진 영역이 전경이며 사각형 내부에서 덧칠해진 영역을 제외한 나머지 영역이 배경이다.

이렇게 생성된 히스토그램은 식(12)를 통하여 보간을 수행하고 결과 히스토그램은 히스토그램의 합이 1이 되도록 정규화한다.

$$Histo[i] = Histo[i-1]*0.25 + Histo[i]*0.5 + Histo[i+1]*0.25 \quad (12)$$

Fig. 13은 오류역전파 신경망[13]을 이용한 학습기를 도식화한 것이다. 그림과 같이 이전 단계에서 추출한 특징을 입력으로 하여 신경망을 구성하였다. 전경과 배경에서의 방향 히스토그램을 하나의 입력으로 학습하였으며, 인식 또한 같은 방법으로 적용하였다. 본 연구에서는 입력층 32, 은닉층 6, 출력층 2개를 사용하여 인식기를 구성하였다.

4. 실험 결과

본 연구에서는 XVID로 압축된 640 * 352 크기의 뉴스 및 예능 프로그램을 입력으로 하여 캡션검출 방법을 평가하였다.

Table 3과 같이 5개의 동영상에서 무작위로 약 10분정도를 선택하여 분석하였다. Table 3의 연산프레임 수는 분석

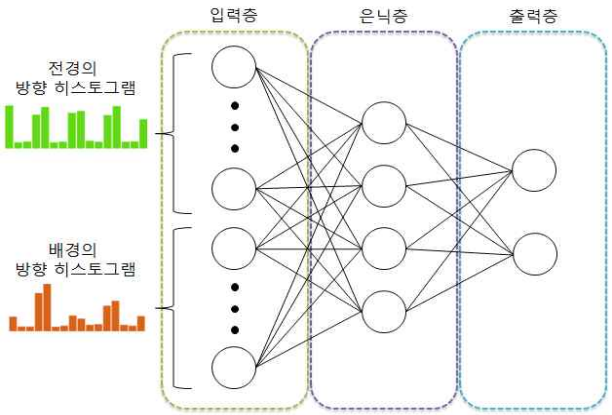

Fig. 13. The error back propergation NN

Table 3. Test video set

동영상 종류	총 연산프레임	초당 연산프레임	캡션영역 수
뉴스데스크	2461	4	54
스타부부쇼	4996	8	182
세바퀴	4872	8	225
1박2일	5012	8	194
무한도전	4929	8	278

Fig. 14. The example of caption region

시간동안 실제 연산에 사용된 프레임 수를 의미하여 초당 연산프레임은 초당 샘플링 한 프레임 수를 의미하며, 캡션영역의 수는 동시에 나타나 동시에 사라지는 캡션영역을 하나의 캡션영역으로 간주하여 계산한 수이다. 캡션영역은 Fig. 14과 같다.

본 실험의 성능평가는 크게 캡션영역검출과 캡션검증에 대해 수행한다. 첫 번째로는 캡션검증과정 이전까지의 성능이다. 이는 실제 캡션을 모두 검출하는 역할을 수행하여야 한다. 즉, False Positive Rate보다 False Negative Rate가 최소가 되도록 캡션후보영역을 검출하여야한다. 두 번째로 캡션검증단계에서의 성능은 이전 단계에서 캡션후보영역으로 검출된 영역을 캡션영역과 비캡션영역을 얼마나 잘 분류하는지에 대한 성능을 나타낸다.

4.1 캡션후보영역 검출 성능

캡션후보영역 검출은 캡션검증단계를 거치기 전의 출력 데이터로 실제캡션영역을 얼마나 많이 포함하고 있는가를

평가한다. Table 4는 캡션후보영역 검출 성능을 나타낸다. 캡션검출 비율은 실제 캡션 중 제안한 방법으로 검출된 실제 캡션의 비율을 나타내며 정밀도는 검출된 실제 캡션영역 중 정확한 영역으로 검출된 영역의 비율을 의미한다.

Table. 4 Performance for detection of caption candidate region

동영상의 종류	총 연산프레임	캡션검출 비율	정밀도
MBC 뉴스데스크	2461	90%	75%
스타부부쇼	4996	89%	82%
세바퀴	4872	85%	80%
1박2일	5012	95%	89%
무한도전	4929	92%	85%

1박 2일의 경우 가장 높은 캡션검출 비율을 보였는데 이는 대부분의 캡션이 뚜렷하며 특이한 패턴의 캡션이 거의 없었다. 이에 반해 '세바퀴'는 그림이 포함된 캡션, 움직이는 캡션 등 다양한 패턴의 캡션을 포함하고 있어 비교적 낮은 캡션검출 비율을 보였다. Fig. 15는 정확하게 검출된 캡션후보영역과 부정확하게 검출된 캡션후보영역의 결과를 보여준다. 부정확한 캡션후보영역의 기준은 다음 식(13)으로 정의하였다. 식(13)에서 MER^{rc}_{Area} 는 실제 캡션영역 최소인접사각형의 넓이이며 MER^{dc}_{Area} 는 검출된 캡션후보영역 최소인접사각형의 넓이이다. 만약 $TH_{AreaRatio}$가 0.5 이상이면 부정확한 캡션영역이며 -0.5 이하이면 미검출 캡션영역으로 분류하였다.

$$TH_{AreaRatio} = \frac{MER^{dc}_{Area} - MER^{rc}_{Area}}{MER^{rc}_{Area}} \quad (13)$$

Fig. 15. The detection result of caption candidate region

4.2 캡션후보영역 검증 성능

캡션후보영역을 검출과정에서는 최대한 많은 실제 캡션을 검출하기 위해서 몇 개의 조건만으로 캡션후보영역을 검출하였다. 따라서 실제 캡션영역 뿐만 아니라 많은 비캡션영역도 캡션후보영역에 포함되었다. 그래서 본 연구에서는 캡션영역과 비캡션영역을 분류하기 위해 신경망을 이용한 캡션검증단계를 추가하였다. 캡션검증단계의 성능은 이전단계에서 검출된 캡션후보영역에 대하여 실험하였다.

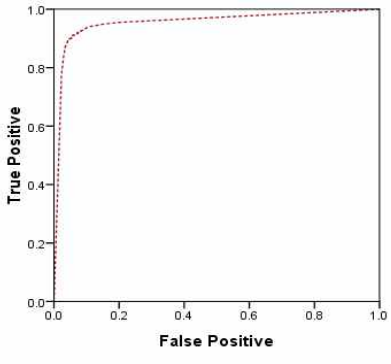

Fig. 16. ROC curve

Fig. 17. The result of recognition for caption region

Fig. 16은 제안한 특징을 이용하여 구성한 신경망 인식기의 ROC곡선을 보여준다[14, 15]. 이 곡선은 모든 동영상에 대하여 인식기를 수행하고 나온 캡션일 가능성을 나타내는

노드의 결과 값을 기준으로 생성하였다. 또한 곡선 내부의 면적은 0.955를 얻었다. 일반적으로 ROC곡선 내부 면적이 0.9 이상이면 매우 좋은 결과를 나타낸다.

Fig. 17는 캡션영역 인식 결과를 나타낸다. 먼저 참 긍정(True Positive)의 예를 보면 뉴스 캡션과 약간의 그림이 포함된 예능 프로그램의 캡션이 정확히 검출된 결과를 볼 수 있다. 좌측 영상의 경우 캡션을 포함하는 배경 영역은 검출되지 않았지만 실제 문자영역이 검출되어 병합과정을 통해 하나의 캡션으로 검출 된 결과를 볼 수 있다. 이러한 이유는 뉴스의 경우 간혹 문자를 포함하는 배경영역이 매우 투명하여 장면전환 시 차연산이 발생하여 검출하지 못하기 때문이다. 참 부정(True Negative)의 경우를 보면 캡션영역 주위에서 나타난 영역을 비캡션영역으로 정확하게 판단한 것을 볼 수 있다. 이는 병합된 하나의 영역에서 특징을 추출하지 않고 배경과 전경을 분리하여 특징을 추출하는 방법을 사용함으로써 가능하였다.

Table 5. The result of recognition in test video set

Th	TP	FP	TN	FN	Recall	Preci-sion	Accu-racy
뉴스 프로그램 : MBC 뉴스데스크							
0.5	46	8	171	3	0.94	0.85	0.95
0.6	46	7	172	3	0.94	0.87	0.96
0.7	45	7	172	4	0.92	0.87	0.95
0.8	45	7	172	4	0.92	0.87	0.95
0.9	44	5	174	5	0.90	0.90	0.96
예능 프로그램 : 무한도전							
0.5	217	40	797	39	0.85	0.93	0.95
0.6	217	39	798	39	0.85	0.93	0.95
0.7	215	33	804	41	0.84	0.93	0.96
0.8	215	28	809	41	0.84	0.94	0.97
0.9	213	25	812	43	0.83	0.94	0.97
예능 프로그램 : 세바퀴							
0.5	179	49	718	13	0.93	0.79	0.94
0.6	178	49	718	14	0.93	0.78	0.93
0.7	177	48	719	15	0.92	0.79	0.93
0.8	177	44	723	15	0.92	0.80	0.94
0.9	175	40	727	17	0.91	0.81	0.94
예능 프로그램 : 스타부부쇼 자기야							
0.5	156	48	583	6	0.96	0.76	0.93
0.6	156	42	589	6	0.96	0.79	0.94
0.7	155	39	592	7	0.96	0.80	0.94
0.8	154	37	594	8	0.95	0.81	0.94
0.9	153	31	600	9	0.94	0.83	0.95
예능 프로그램 : 해피선데이 1박 2일							
0.5	172	14	253	14	0.92	0.92	0.95
0.6	169	14	253	17	0.91	0.92	0.95
0.7	168	14	253	18	0.90	0.92	0.95
0.8	168	13	254	18	0.90	0.93	0.95
0.9	167	10	257	19	0.90	0.94	0.96
최종결과							
0.5	770	159	2522	75	0.91	0.83	0.93
0.6	760	151	2530	79	0.91	0.84	0.93
0.7	760	141	2540	85	0.90	0.84	0.94
0.8	759	129	2552	86	0.90	0.85	0.94
0.9	752	111	2570	93	0.89	0.87	0.94

거짓 긍정(False Positive)의 영상들을 보면 대부분 에지가 많은 영역이 캡션이라 판단되었다. 하지만 이것은 대부분 캡션과 유사하게 규칙적인 에지방향 히스토그램을 보였기 때문에 캡션영역으로 인식되었다. 거짓 부정(False Negative)의 경우에는 그림과 같은 특수한 상황에 많이 발생하였다. 거짓 부정의 좌측 영상과 가운데 영상을 보면 캡션 주위에 풍선 그림으로 테두리가 진하게 나타나있음을 알 수 있다. 이 테두리는 주위와 경계가 뚜렷하여 에지 방향 히스토그램에 영향을 주지만 규칙적이지 않아서 인식기에서 잘못 분류되었다. 또한 우측 영상은 3차원 지도가 캡션으로 삽입된 영상이다. 이 또한 규칙적인 에지방향 히스토그램을 가지고 있지 않아 오인식되었다.

Table 5은 각 동영상 별 캡션영역 인식 결과를 보여준다. 표의 Th는 신경망의 출력층에서 캡션일 경우 1로 설정되는 노드의 출력값을 기준으로 나눈 결과이다. 임계값 0.5를 기준으로 최종 평균 Recall을 보면 0.91로 캡션후보영역 중 실제 캡션 중 91%를 검출하였음을 알 수 있다. '무한도전' 같은 경우는 다른 동영상과 비교하여 Recall이 낮은 것을 볼 수 있는데 무한도전의 캡션은 다양한 패턴이 많았으며 영상 안에 나뭇잎이나 밭과 같은 에지가 많은 영역이 다량 존재한다. 따라서 캡션후보영역에 비캡션후보영역이 많이 포함되어 있어서 비교적 성능이 낮음을 알 수 있다.

실험에 사용한 동영상들은 XVID로 손실 압축되어 인위적으로 삽입된 캡션영역도 시간에 따라 약간의 변화가 있었으며 페이드-인(fade-in) 효과로 나타나는 문제들이 있었다. 이것은 연산프레임 샘플링 수를 설정하여 해결하려했으나 너무 천천히 나타나거나 이동하면서 나타나는 캡션의 경우에는 후보영역으로도 검출하지 못하는 문제가 있었다. 하지만 일반적인 캡션의 경우 좋은 성능을 확인하였다.

5. 결 론

본 논문에서는 캡션의 위치, 색상, 크기, 서체에 대한 제약조건 없이 동영상으로부터 인공 캡션영역을 추출하는 방법을 제안하였다. 기존의 시간적 특성만을 이용한 화소 단위 연산 방법을 개선하여 영역 단위로 연산하고 병합과정을 통하여 False Negative Rate를 최소화하였고 신경망을 이용한 캡션영역검증 과정을 통하여 False Positive Rate를 감소시켰다. 또한 천천히 나타나거나 이동하면서 나타나는 캡션에 강건하도록 시간과 공간 정보를 이용한 2단계 병합과정을 사용하여 다양한 패턴의 캡션영역을 검출할 수 있었다. 그러나 텍스트 정보가 없는 캡션의 경우는 캡션검증단계 이전까지는 검출되었으나 검증단계에서 제안한 특징과 차이가 있어서 검출하지 못하는 경우도 있었다. 하지만 이 또한 비캡션영역과 그림과 같은 캡션의 특성을 분석하여 분류 가능한 특징을 추가하고 학습하여 인식한다면 보완 가능할 것으로 생각된다. 또한 캡션이 작은 경우 캡션 검출 실패는 좀 더 고해상도 동영상으로 실험하여 해결 가능할 것으로 생각된다. 향후 잡음의 영향으로 부정확한 캡션후보영역에 대하여 개선방법과 다양한 패턴의 캡션에 대하여 인식률을 높이는 연구가 필요할 것이다.

참 고 문 헌

[1] S.I.Joo, "Unstructured Caption Detection Using Temporal and Spatial Information in Video", Master degree Thesis, Soonsil Univ., 2010.

[2] R.Lienhart and F.Stuber, "Automatic Text Recognition in Digital Videos", In Proceedings of SPIE Image and Video Processing IV, Vol.2666, pp.180-188, September, 1996.

[3] R.Lienhart and W.Effelsberg, "Automatic Text Segmentation and Text Recognition for Video Indexing", Multimedia System, Vol.8, pp.69-81, January, 2000.

[4] C.M.Lee and A.Kankanhalli, "Automatic extraction of characters in complex scene images", International Journal of Pattern Recognition Artificial Intelligence, Vol.9, No.1, pp. 67-82, 1995.

[5] Y.Zhong, K.Karu and A.KJain, "Locating Text in Complex Color Images", Pattern Recognition, Vol.28, No.10, pp. 1523-1536, October, 1995.

[6] H.PLi and D.Doermann, "Automatic Text Detection and Tracking in Digital Video", IEEE Transactions on Image Processing, Vol.9, No.1, January, 2000.

[7] P.Shivakumara, T.Q.Phan and C.L.Tan, "New Wavelet and Color Features for Text Detection in Video," 20th International Conference on Pattern Recognition, pp.3996-3999, 2010.

[8] S.I.Hwang "A Study on Text Detection using DCT Coefficients in I-frame", Master degree Thesis, Yonsei Univ., 2002.

[9] J.Xi, X.H.Hua, X.R.Chen, L.Wenyin, H.J.Zhang, "A Video Text Detection and Recognition System", IEEE International Conference on Multimedia and Expo, pp.1080-1083, August 22-25, 2001.

[10] Q.X.Ye, Q.M.Huang, W.Gao and D.B.Zhao, "Fast and robust text detection in images and video frames", Image and Vision Computing, pp.565-576, Vol.23, No.6, 2005.

[11] P.Shivakumara, W.Huang and C.L.Tan, "Efficient video text detection using edge features", in Proc. ICPR, pp.1-4, 2008.

[12] C.H.Kwon, C.H.Shin, S.Y.Kim and S.H.Park, "Caption Detection Algorithm Using Temporal Information in Video", The Transactions of The Korean Institute of Electrical Engineers, Vol.53, No.8, pp.606-610, 2004.

[13] R.O.Duda, P.E.Hart and D.G.Stork, "*Pattern Classfication*", Wiley Interscience, Second Edition, Chapter 7.

[14] J.Davis and M.Goadrich, "The Relationship Between Precision Recall and ROC curves", Proceedings of the 23rd International Conference on Machine Learning(ICML 2006), pp.233-240, 2006.

[15] T.Fawcett, "ROC graphs : Notes and practical considerations for researchers." http://www.hpl.hp.com/personal/*Tom_Fawcett/papers/index.html*, 2003.

주 성 일

e-mail : sijoo82@ssu.ac.kr
2008년 한국산업기술대학교(공학사)
2010년 숭실대학교 미디어학과(공학석사)
2010년~현 재 숭실대학교 미디어학과 박사과정
관심분야 : 영상처리, 컴퓨터비전, 패턴인식 등

원 선 희

e-mail : nifty12@ssu.ac.kr
2005년 한경대학교 컴퓨터공학과(공학사)
2007년 숭실대학교 컴퓨터학과(공학석사)
2012년 숭실대학교 미디어학과(공학박사)
2012년~현 재 숭실대학교 미디어학과 Post Doc
관심분야 : 영상처리, 컴퓨터비전, 3D 모델링, 패턴인식 등

최 형 일

e-mail : hic@ssu.ac.kr
1972년 연세대학교 전자공학과(공학사)
1982년 미시간대학교 전자공학과 (공학석사)
1987년 미시간대학교 전자공학과 (공학박사)
1995년~1997년 퍼지 및 지능시스템학회 이사
1996년~1998년 정보과학회 컴퓨터비전 및 패턴인식 연구회 위원장
1997년 IBM Waston Lab 방문연구원
2005년~2006년 한국정보과학회 이사
1987년~현 재 숭실대학교 미디어학과 교수
관심분야 : 컴퓨터비전, 퍼지 및 신경망 이론, 패턴인식, 지식기반 시스템 등