퍼지 및 지능시스템학회 논문지 2007, Vol. 17, No. 7, pp. 924-929

# 신경망을 이용한 자막 크기에 무관한 연결 객체 기반의 자막 추출

# Connected Component-Based and Size-Independent Caption Extraction with Neural Networks

정제희 · 윤태복 · 김동문 · 이지형

Je-Hee Jung · Tae Bok Yoon · Dong-Moon Kim · Jee-Hyong Lee

성균관대학교 전자전기컴퓨터공학과

요 약

영상에 나타나는 자막은 영상과 관계가 있는 정보를 포함한다. 이러한 영상과 관련 있는 정보를 이용하기 위해 영상으로부터 자막을 추출하는 연구는 근래에 들어 활발히 진행되고 있다. 기존의 연구는 일정한 높이의 자막이나 획의 두께를 지닌 자막에서만 정상적인 작동을 한다. 본 논문에서는 일정 크기 이상의 자막에 대해서 적용할 수 있는 크기에 무관한 자막 추출 방법을 제안한다. 먼저, 자막 연결 객체의 패턴 추출을 위해서 자막이 포함된 영상을 수집하고, 신경망을 이용해서 자막의 패턴을 분석한다. 그 후로는 사전에 추출한 패턴을 이용하여 입력 영상에서 자막을 추출한다. 실험에 사용된 영상은 뉴스, 다큐멘터리, 쇼 프로그램과 같은 대중 방송에서 수집하였다. 실험 결과는 다양한 크기의 자막을 포함한 영상을 사용하여 실험하였고, 자막 추출의 결과는 찾아진 연결객체 중에 자막의 비율과 자막 중에 찾아진 자막의 비율로 분석하였다. 실험 결과를 보면 제안한 방법에 의해 다양한 크기의 자막을 추출할 수 있음을 보여준다.

키워드 : 자막 추출, 신경망, 자막, 연결 객체

Abstract

Captions which appear in images include information that relates to the images. In order to obtain the information carried by captions, the methods for text extraction from images have been developed. However, most existing methods can be applied to captions with fixed height or stroke's width. We propose a method which can be applied to various caption size. Our method is based on connected components. And then the edge pixels are detected and grouped into connected components. We analyze the properties of connected components and build a neural network which discriminates connected components which include captions from ones which do not. Experimental data is collected from broadcast programs such as news, documentaries, and show programs which include various height caption. Experimental result is evaluated by two criteria : recall and precision. Recall is the ratio of the identified captions in all the captions in images and the precision is the ratio of the captions in the objects identified as captions. The experiment shows that the proposed method can efficiently extract captions various in size.

Key Words : Caption extraction, Neural network, Caption, Connected component, Text extraction

## 1. 서 론

지식을 전달하는 역할을 하는 콘텐츠는 문서, 영상, 음악, 동영상 등의 다양한 종류가 존재한다. 특히, 영상은 동영상을 이루는 요소이며 시각적으로 의미를 전달이 가능하기 때문에 중요한 콘텐츠라 할 수 있다. 하지만, 영상 콘텐츠는 다른 콘텐츠들에 비해 비교적으로 용량이 크기 때문에 영상 콘텐츠를 효율적으로 저장하고 검색하기 위한 연구가 진행되어 왔다.

특히, 자막은 영상에서 정보를 부가적으로 전달하기 위해서 작성된다는 특징을 갖고 있기 때문에, 영상과 밀접한 관계를 가지고 있다. 이러한 자막의 특징을 이용하기 위해서 영상에서 자막을 추출하는 연구는 예전부터 진행되어 왔다.

[1]에서는 다양한 크기의 자막을 추출하기 위해서 해상도를 조절한 영상에서 동일한 연산을 반복 수행하는 방법을 개선한 방법을 제안하였다. 제안한 방법은 각 해상도에서 이미 추출한 자막은 배제하고 수행하는 방법이다. 하지만, 해상도 조정과 동일한 과정을 중복 수행하는 과정이 필요하다.

본 논문에서는 자막의 크기에 무관하게 자막을 추출하기

접수일자 : 2007년 11월 10일
완료일자 : 2007년 12월 3일
감사의 글 : 본 연구는 21세기 프론티어 연구개발 사업의 일환으로 추진되고 있는 정보통신부의 유비쿼터스컴퓨팅 및 네트워크원천기반기술 개발사업의 지원을 받았습니다.

위해 다음과 같은 과정을 수행한다. 먼저, 자막을 이루는 경계를 이루는 에지를 추출하고, 에지를 기반으로 연결 객체를 생성한다. 생성한 연결 객체 중에서 확실히 자막이 아닌 연결 객체를 제거한 후, 연결 객체의 패턴을 신경망을 통해 자막과 비 자막의 패턴으로 분류한다. 분석된 패턴을 이용하여 연결 객체를 분류하고, 남겨진 연결 객체의 위치를 찾아낸다. 마지막으로 찾아진 위치에서 자막 픽셀을 분리하는 작업을 수행하여 자막을 추출하는 방법을 제안한다.

논문의 흐름은 2장에서 기존의 논문들에 대한 분석을 할 것이다. 다음으로 3장에서 제안하는 방법을 서술하고, 4장에서 실험 결과를 정리할 것이며, 마지막으로 5장에서 결론을 내린다.

## 2. 관련 연구

영상에 포함된 자막은 다양한 특징을 갖기 때문에 어떠한 경우에도 성립되는 특징을 이용해서 자막을 추출하는 연구는 어렵다고 할 수 있다. 이러한 특징으로는 자막의 크기, 정렬 상태, 자간, 색상, 움직임, 경계등으로 정리할 수 있다[3]. 모든 경우에 적용이 가능한 특징의 결정은 어려운 일이기 때문에 자막 추출 연구는 크기, 색상, 움직임과 같은 일정한 특징을 기준으로 다양한 형태의 자막을 추출하는 연구가 계속 진행되어 왔다.

대부분의 자막 추출은 자막 추출을 위해서 자막일 가능성이 적은 픽셀들의 정보를 제거하면서 접근하는 방법을 제안하였다. 이러한 연구는 주로 3단계로 이루어진 자막 추출 방법을 제안하였다[1], [2]. 먼저 영상에서 자막이 존재하는가를 판단하는 자막 찾기, 다음으로 자막이 존재하는 위치를 파악하는 자막 위치 찾기를 수행하였다. 마지막으로 자막이 존재하는 위치에서 자막 픽셀과 비 자막 픽셀을 분리하는 자막 분리를 수행하였다. 이러한 과정은 점차적으로 자막 아닌 것으로 추측되는 픽셀들을 제거하는 과정을 통해 검색의 범위를 축소해 가는 과정으로 자막 추출의 속도를 향상시키기 위해 제안되었다.

[4], [5]에서는 자막을 추출하기 위해 라인 단위의 에지를 추출하거나 각 색상 영역에서 신경망 학습을 사용한 자막 후보들을 추출하여 하나의 자막 존재 지역으로 결합하는 방법을 제안하였다. 자막이 존재하는 영역을 추출하기 위해서 블록의 크기가 고정적으로 결정되어 있는 한계가 존재하였다. 이러한 방법은 블록 크기 의존적인 추출 알고리즘이며, 다양한 상황에서 정상적으로 작동한다고 보장하기가 어렵다는 단점이 존재한다고 할 수 있다.

[6-8]은 모폴로지 방법을 기반으로 자막을 추출하는 방법을 제안하였다. 추출한 자막의 영역 중에서 확실하게 자막이 존재하지 않은 영역을 제거하는 방법을 수행하였다. 이러한 방법은 자막 추출을 위해서는 자막의 형태가 특정한 형태로 존재한다는 가정이 필요하기 때문에 자막을 추출하기 위해선 자막이 특정한 조건을 만족해야만 한다는 한계가 존재한다.

전술한 방법들은 특정한 형태에 적응 가능한 방법에 대해서만 정상적인 작동을 보장한다. 하지만, 영상에서의 자막은 다양한 형태로 존재한다. 특히, 다양한 크기의 자막을 추출하기 위해서 [1]에서는 순차적인 다중 해상도 기법을 사용하였다. 제안한 방법은 한번 추출한 자막은 재검색을 하지 않는다. 하지만, 해상도의 조절이 필요하고, 동일한 연산을 반복 수행하는 낭비가 발생한다. 또한, [2]에서는 획 기반 자막 추출 방법을 제안하였으나, 획의 간격으로 정한 범위를 벗어난 자막의 획은 자막으로 인식하지 못하였으며, 자막의 획이 아니더라도 획 간격에 포함된 에지는 자막으로 추출하였다..

본 논문에서는 연결 객체를 기반으로 하여 자막의 크기에 무관하게 자막을 추출하는 방법을 제안하고자 한다. 입력은 색상으로 포함된 자연 영상이며, 결과로는 자막과 비 자막 픽셀이 이진화된 영상이 생성한다.

## 3. 자막 크기에 무관한 자막 추출 방법

### 3.1 자막의 특성

자막은 다양한 특징을 포함하기 때문에 모든 상황에서 만족하는 자막의 특징을 결정하기는 어려운 일이다. 그래서 본 논문에서는 추출 대상 자막을 다음과 같이 가정한다. 첫째, 자막은 일정한 색상으로 이루어지며, 배경과 일정 수준이상의 대비를 지닌다. 둘째, 자막을 이루는 획의 경계인 에지 픽셀은 반대 방향 성분을 포함한 에지 픽셀이 존재하고, 본 논문에서는 이러한 에지들을 대응 에지 쌍 이라고 말하겠다.

자막을 이루는 픽셀들은 하나의 픽셀만으로 자막을 표현할 수 없으며, 자막은 여러 개의 픽셀로 구성된 그룹인 연결 객체로 구성되어 있다고 할 수 있다. 또한, 자막의 경계에 존재하는 에지들은 이러한 자막 연결 객체를 둘러쌓기 때문에 폐곡선의 형태로 존재한다. 자막 연결 객체의 패턴을 분석하기 위해서, 먼저, 자막에 경계에 존재하는 인접한 픽셀들을 하나의 연결 객체로 구성한 후, 연결 객체의 패턴을 분석하였다.

연결 객체를 자막과 비 자막으로 분류하기 위해서 신경망 학습을 이용하였다. 연결 객체 데이터는 총 8,800개이며, 자막은 5,900개, 비 자막은 1,900개로 구성되었다. 학습 데이터는 자막 5,000개와 비 자막 2,500개로 구성되어 총 7,500개이며, 테스트 데이터는 자막 900개와 비 자막 400개로 총 1,300개로 구성하였다. 입력 특징은 총 10개로 구성하였고, 자세한 서술은 3.2.1절에서 언급하겠다.

### 3.2 자막 추출 알고리즘

제안하는 자막 추출 알고리즘은 다음의 3단계로 이루어진다. 먼저 자막의 연결 객체를 찾는 자막 찾기 단계와 찾아진 연결 객체의 위치를 확인하는 자막 위치 찾기 단계, 찾아진 위치에서 자막과 비 자막의 픽셀을 분리하는 자만 분리 단계로 구성된다.

#### 3.2.1 자막 찾기 단계

자막 찾기 단계에서는 자막을 이루는 연결 객체를 찾기 위한 단계이다. 입력으로 들어온 영상에서 자막을 추출하기 위해서 영상에서 자막으로 추측되는 픽셀들을 탐색한다.

입력으로 들어온 컬러 영상은 자막 추출 과정의 복잡성을 낮추기 위해 (1)을 사용하여 256단계의 밝기를 가진 명암도 영상 G로 변환한다[9].

$$G(x,y) = 0.3^* C_R(x,y) + 0.59^* C_G(x,y) + 0.11^* C_B(x,y) \quad (1)$$

위의 C는 RGB의 각 색상 성분을 표현한 하나의 이미지이며, R은 빨간색, G는 녹색, B는 파란색을 의미한다. x, y는 각 이미지에서 픽셀의 위치를 나타낸다.

변환한 G에는 잡음 성분이 포함되어 있다. 이러한 잡음 성분은 에지를 추출하는데 불필요하게 많은 에지를 생성하게 할 수 있다. 그러므로 잡음을 제거하는 역할을 하는 저역 필

터링(2)을 사용하여 필터링된 명암도 이미지 $G_F$를 생성한다.

$$\begin{aligned} G_F(x,y) = (&G(x-1,y+1)+G(x,y+1)+G(x+1,y+1)+\\ &G(x-1,y)+G(x,y)+G(x+1,y)+\\ &G(x-1,y-1)+G(x,y-1)+G(x+1,y-1))/9 \end{aligned} \quad (2)$$

자막과 비 자막은 대비를 이루기 때문에 $G_F$에서 자막과 비 자막의 경계지점에 존재하는 에지를 조사한다. 에지 픽셀은 방향성을 갖는 픽셀로서 기준 픽셀과 비교 대상 픽셀간의 변화 정도를 나타낸다. 그러므로 자막의 경계를 이루는 에지를 연결한 연결 객체는 자막의 획의 경계의 픽셀들의 연결 객체라고 할 수 있다.

필터링한 명암도 영상 $G_F$에서 로버츠 에지 추출 방법 (3)을 사용하여 각 방향(좌상, 상, 우상, 좌, 우, 좌하, 하, 우하)의 에지를 추출하고, 각 방향에 대한 에지의 값을 갖는 방향 에지 영상 $E_{tl}$, $E_t$, $E_{tr}$, $E_l$, $E_r$, $E_b$, $E_{bl}$, $E_{br}$을 만들어 낸다. 각 방향은 11시, 12시, 1시, 9시, 3시, 7시, 6시, 5시 방향으로 구성된다. 이때 각 영상들은 실험적인 결과에 구해진 임계치를 사용하여 이진화를 수행한다.

$$\begin{aligned} E_{tl}(x,y) &= G_F(x,y) - G_F(x-1,y+1)\\ E_t(x,y) &= G_F(x,y) - G_F(x,y+1)\\ E_{tr}(x,y) &= G_F(x,y) - G_F(x+1,y+1)\\ E_l(x,y) &= G_F(x,y) - G_F(x-1,y)\\ E_r(x,y) &= G_F(x,y) - G_F(x+1,y)\\ E_{bl}(x,y) &= G_F(x,y) - G_F(x-1,y-1)\\ E_b(x,y) &= G_F(x,y) - G_F(x,y-1)\\ E_{br}(x,y) &= G_F(x,y) - G_F(x+1,y-1) \end{aligned} \quad (3)$$

각 방향 에지 영상들은 한 방향의 에지들의 정보만을 표현하기 때문에, 자막의 경계에 존재하는 모든 에지들의 연결 객체를 생성하기 위해서 모든 방향의 에지를 하나로 통합한 모든 방향 에지 영상 $E_{all}$을 생성한다. $E_{all}$은 8개의 방향 에지 영상의 검색 픽셀의 위치에 픽셀이 하나라도 존재하면 Eall에 에지가 존재하는 영상이다. $E_{all}$는 자막의 획을 이루는 모든 에지들을 표현한 영상이다.

모든 방향 에지 영상의 픽셀들은 자막의 경계를 구성하는 에지라고 할 수 있다. 연결 객체의 패턴을 분석하여 자막과 비 자막 연결객체를 분류하는 작업을 하기 위해 영상에 포함된 인접한 픽셀들로 구성된 연결 객체를 생성한다. 연결 객체는 4-이웃 연결 객체 방법을 사용하여 생성하며 수식(4)을 만족하고, 각 라벨은 고유한 값을 갖는 1차 연결 객체 영상 $CC_{first}$를 생성한다.

$$CC_{first}(x,y) = \begin{cases} E_{all}(x,y) \neq 0 \wedge \\ \quad CC_{first}(x,y-1) \neq 0, & CC_{first}(x,y-1) \\ E_{all}(x,y) \neq 0 \wedge \\ \quad CC_{first}(x-1,y) \neq 0, & CC_{first}(x-1,y) \\ else & 0 \end{cases} \quad (4)$$

1차 라벨링 수행이 끝난 후에 $CC_{first}$에는 자막에 포함되지 않은 에지도 포함되어 있다. 자막을 이루는 획은 대응 에지 쌍들로 결합되어 있기 때문에 같은 연결 객체 안에서 서로 대응하는 방향(예를 들면, 1시와 5시)의 에지가 존재하지 않는 에지들은 제거한다. 1차 라벨링에서 대응 에지 쌍이 아닌 에지를 제거한 영상은 자막이 아닌 않은 에지들이 제거되어 있다. 다시 라벨링을 수행하면 대응 에지 쌍이 아닌 에지가 제거된 에지들로 2차 라벨링 영상$CC_{second}$이 생성된다.

$CC_{second}$는 비 자막을 이루는 연결 객체가 포함되어 있다. 비 자막 연결 객체들은 일정한 색상을 갖지 않을 확률이 높기 때문에 중간에 에지가 누락되는 경우가 발생한다. 반면에 일반적인 자막의 경우엔 자막과 비 자막에 대비가 존재하기 때문에 반드시 경계가 일정하게 연결된다. 만약 연결 객체가 8방향 중에서 하나의 성분도 포함하지 않는다면 자막 연결 객체가 아니라 판단하고, 연결 객체를 제거한다.

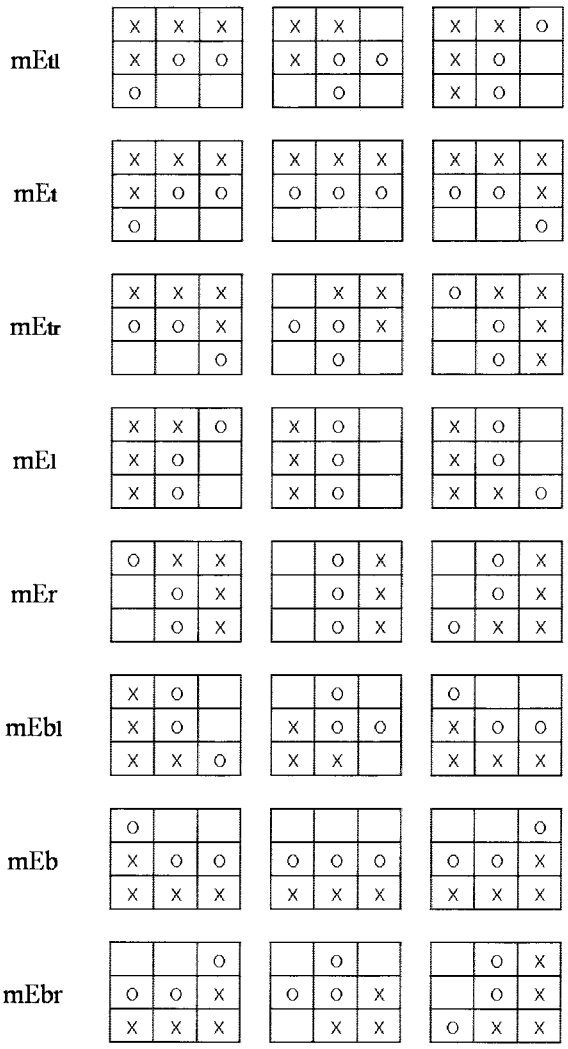



그림 1. 방향 정제 마스크

$CC_{second}$에 포함된 연결 객체들의 에지는 하나의 에지에도 다양한 방향성을 지닌다. 다양한 방향성을 지닌 에지는 방향 정보의 애매함을 제공하여 연결 객체의 패턴 분석에 어려움의 이유가 된다. 신경망 학습을 하기에 앞서 이러한 애매한 방향성을 제거하기 위해 그림 1의 방향 에지 정제 마스크 ($mE_{tl}$ $mE_t$, $mE_{tr}$, $mE_l$, $mE_r$, $mE_{bl}$, $mE_b$, $mE_{br}$)를 사용하여 에지의 애매한 방향성을 제거한다. 마스크는 현재 검색 픽셀과 3 * 3 마스크의 중앙 점을 대응하여 만약 'O'인 점에 에지가 존재하고, 'X'인 점에 에지가 존재하지 않는다면 그 점은 그 방향의 에지가 존재한다고 표시한다. 이렇게 정제한 에지 방향 정제 영상($RE_{tl}$ $RE_t$, $RE_{tr}$, $RE_l$, $RE_r$, $RE_{bl}$, $RE_b$, $RE_{br}$)를 생성한다.

에지 방향 정제 영상과 2차 연결 객체 영상의 정보는 신경망 학습에 사용되어 연결 객체를 분류한다. 사용되는 정보는 아래의 10가지를 사용하며, 각 특징 값은 하나의 연결 객체에 대해서만 조사한다. 추출한 특징 값들은 크기가 변해도 일정하게 유지가 되는 특징 값을 사용하였다.

1. $RE_{tl}$의 픽셀의 수 / $E_{all}$의 픽셀의 수
2. $RE_{t}$의 픽셀의 수 / $E_{all}$의 픽셀의 수
3. $RE_{tr}$의 픽셀의 수 / $E_{all}$의 픽셀의 수
4. $RE_{l}$의 픽셀의 수 / $E_{all}$의 픽셀의 수
5. $RE_{r}$의 픽셀의 수 / $E_{all}$의 픽셀의 수
6. $RE_{bl}$의 픽셀의 수 / $E_{all}$의 픽셀의 수
7. $RE_{b}$의 픽셀의 수 / $E_{all}$의 픽셀의 수
8. $RE_{br}$의 픽셀의 수 / $E_{all}$의 픽셀의 수
9. 각 수평으로 대응되는 에지의 중앙 픽셀의 밝기들의 표준 편차
10. 각 수직으로 대응되는 에지의 중앙 픽셀의 밝기들의 표준 편차

위의 10개의 특징 값은 다음과 같다. 특징 값 1~8는 연결 객체에 포함된 전체 에지 픽셀의 개수에서 각 방향으로 속하는 에지 픽셀의 비율이다. 9, 10은 획의 대응 에지 쌍의 중간 픽셀의 밝기 값이다. 이 값은 하나의 자막 글자에서 크게 다르지 않기 때문에 밝기의 표준 편차를 사용한다. 만약 한 픽셀이 가로와 세로로 대응되는 픽셀의 중간에 동시에 만족하면 짧은 방향의 정보를 이용한다. 여기서 사용된 각 연결 객체의 10가지 특징 값은 폰트가 바뀌어도 특별한 폰트가 아닌 경우엔 크기가 변하더라도 일정한 비율을 지닌다. 이러한 특징 값으로 신경망을 사용하여 자막과 비 자막 연결 객체의 패턴을 분류하였다. 분류된 비 자막 연결 객체는 제거하고 자막 탐색 결과 영상DR을 생성한다.

제안한 방법에서 사용한 신경망은 사전에 학습을 수행하였고, 자막과 비 자막 연결 객체의 패턴을 분석하기 위해서 사전에 자막이 존재하는 영상과 워드에서 사용하는 폰트를 사용하여 학습을 수행하였다.

#### 3.2.2 자막 위치 찾기 단계

자막 위치 찾기 단계는 자막이 존재하는 위치를 찾는 단계이다. 이 단계에서는 자막의 위치를 찾기 위해 [1]의 논문에서 제안한 coarse-to-fine localization방법을 사용한다.

이 방법은 초기에 DR의 에지를 대상으로 전체의 영상을 수평으로 프로젝션을 수행한다. 이때 자막이 존재하는 영역은 에지가 높은 분포로 존재하기 때문에 이러한 영역을 자막 존재 영역으로 분할한다. 수평 프로젝션으로 분할된 각 영역에서 에지를 대상으로 수직 프로젝션을 수행한다. 수평 프로젝션과 유사하게 에지가 존재하는 영역은 자막 존재 영역으로 분할한다. 전술한 방법은 재귀적인 방법으로 분할을 시도하며 각 자막 존재 영역이 더 이상 분할되지 않을 때 까지 반복 수행한다. 단, 프로젝션 과정에서 두 존재 영역에서 일정 간격을 포함하지 않은 존재 영역은 분할하지 않는다.

수평 프로젝션의 경우엔 폰트의 최소 높이보다 작은 높이의 자막 존재 영역은 제거하였다. 수직 프로젝션의 경우엔 이전에 수행한 수평 프로젝션에서 분할되지 않고, 높이와 최소 글자의 비율에 1.5를 곱한 값보다 너비가 좁은 영역은 제거한다. 이러한 조건들은 자막의 최소 높이와 너비를 고려하여 자막이 존재할 수 없는 영역을 제거하는 역할을 수행한다.

#### 3.2.3 자막 분리 단계

자막 분리 단계에서는 자막 존재 영역에서 자막과 비 자막을 이루는 픽셀들을 이진화 과정을 거쳐서 자막을 이루는 픽셀의 정보를 남기는 것을 목적으로 한다. 이 단계에서는 [2]에서 제안한 획 기반 이진화 방법을 사용하였다. 자막과 비 자막은 충분히 대비를 갖고 있기 때문에 자막과 비 자막의 획에 존재하는 경계 에지의 밝기는 자막과 배경의 이진화에 사용할 임계치로 사용이 가능하다.

획 기반 이진화 방법은 자막 존재 영역에 존재하는 에지의 픽셀들과 대응하는 G의 픽셀들의 밝기 값의 평균값을 이용하여 자막과 비 자막 픽셀의 임계치를 결정하고, 이진화를 수행한다. 결과로 생성한 이진화 영상은 자막과 비 자막의 픽셀이 분리되어 있다.

## 4. 실험 결과

실험 데이터는 쇼 프로그램이나 뉴스와 같은 대중 방송에서 얻어진 동영상에서 영상을 수집하였다. 수집한 영상은 자막을 포함하고 있었으며, 뉴스 및 다큐멘터리의 영상은 10개, 쇼 영상은 21개의 영상으로 전체 영상에는 총 603의 글자가 자막으로 존재하였다. 영상에 포함된 자막은 15pixel에서 45pixel의 높이로 다양한 높이의 글자 크기로 존재하였다.

표 1은 대중 방송으로부터 얻은 동영상에서 실험 데이터인 영상을 추출하여 제안하는 방법을 적용하여 얻은 자막 추출의 결과다. 영상에서 자막 추출의 평가 방법은 모든 자막을 찾는 것뿐만 아니라 정확하게 찾는 것도 중요한 평가 요소이기 때문에, Recall과 Precision을 사용하였다. Recall이란 실제 전체 자막에서 옳게 찾아진 자막의 비율이며, Precision이란 찾아진 자막 중에서 옳게 찾아진 자막의 비율이다.

표 1. 자막 추출 결과(Caption extraction result)

| | Recall(%) | Precision(%) |
|---|---|---|
| 쇼 | 70% | 89% |
| 뉴스 | 69% | 83% |

기존의 관련 연구들은 70%~90%의 Recall과 Precision을 보이며 대부분 영어 자막을 추출하는 실험이었다. 영어는 하나의 획으로 이루어지기 때문에 각 글자가 하나의 연결 객체로 구성된다. 한글은 하나의 글자가 다수의 획으로 구성되며, 다양한 조합이 사용된다. 이러한 다수의 획으로 구성되는 경우가 존재하기 때문에 한글 자막의 추출은 영어 자막의 추출보다는 어렵다고 할 수 있다. 또한, 자막 추출에 사용하는 영상들은 성능을 파악하기 위해 정해진 포맷 및 실험 데이터의 표준이 없기 때문에, 실험 데이터에 따라 자막 추출의 성능은 많은 영향을 받는다.

제안하는 방법의 실험 결과에서 precision은 recall에 비해 높은 수치를 보였다. 이러한 결과는 찾아진 연결 객체는 대부분이 자막이었다는 것을 확인시켜준다. 반면에 Recall은 낮은 수치를 보여주었는데 가장 큰 이유는 에지 검출 방법에서 실험적인 이진화에 의해 결정한 임계치가 작은 글자의 경우엔 에지를 누락시키는 경우가 발생하였기 때문이었다. 특히, 뉴스와 같이 작은 범위 안에 다수의 자막이 존재하는 경우엔 자막의 획의 에지를 구분하기에 충분하지 않기에 크기가 일정 수준 이하의 자막을 추출하기 위해선 에지를 부분적 이진화를 수행하는 알고리즘의 개선이 필요하다고 할 수 있다.

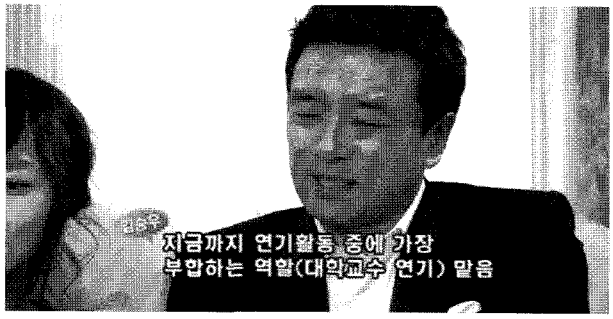



그림 2. 입력 영상1(Input Image1)

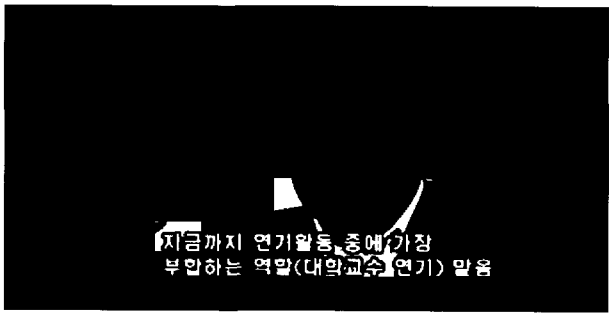



그림 3. 결과 영상1(Result Image1)

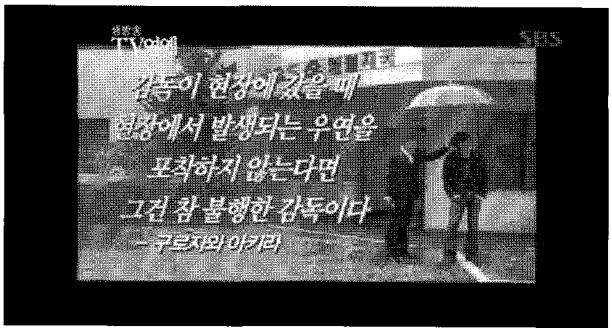



그림 4. 입력 영상2(Input Image2)

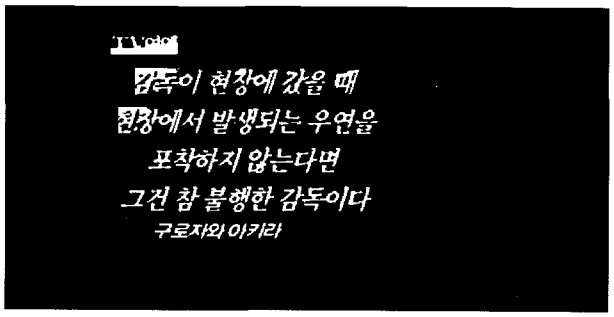



그림 5. 결과 영상2(Result Image2)

그림 2, 4, 6는 입력으로 사용한 영상이다. 그림 3, 5, 7은 각 입력 영상에 대한 제안하는 방법을 적용하여 결과 값으로 생성한 영상이다. 그림 3, 5, 7에서는 자막이 아닌 지역에서도 자막이 존재하는 오작동이 발생하였는데 이는 신경망 학습을 사용한 패턴의 분석은 100%를 보장하지 완벽하지 않다는 것을 보여준다. 각 영상에서 등장한 자막의 크기는 그림 2의 자막은 약 20픽셀 정도였고, 그림 4의 자막은 약 17픽셀과 35픽셀이었으며, 그림 6의 자막 약 18픽셀, 22픽셀과 25픽셀 이었다. 그림 3와 5에서는 자막 근처에 자막과 유사한 색상을 갖는 배경이 존재하기 때문에 자막 존재 지역 내부에서 배경까지 자막으로 이진화되는 단점을 보였다. 그림 7에서는 일부 주변 물체의 연결 객체의 에지 정보를 신경망으로 분석하지 못해서 잘못된 결과가 포함되어 있고, 연결 객체가 우측 하단의 작은 글자가 제대로 추출되지 않는 결과를 보여주고 있다. 그 이유는 한글의 한 글자에는 가로획이나 세로획이 많은 관계로 일반적으로 20픽셀 이하로 표현한 한글 한 글자는 획과 획의 간격이 좁기 때문에 자막과 비 자막이 충분한 대비를 갖지 못하는 경우가 존재하기 때문에 작은 글자의 획은 정상적인 추출이 되지 않는 경우가 존재한다.

그러나 실험을 통하여 제안하는 방법이 자막의 높이가 20픽셀 이상인 경우는 자막의 크기와 상관없이 자막을 추출하였음을 알 수 있었다. 20픽셀 이하의 글자에 대한 처리는 향후에 개선될 부분이다.

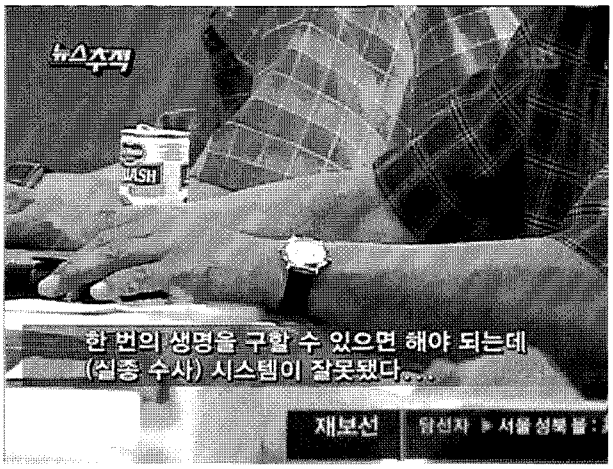



그림 6. 입력 영상3(Input Image3)

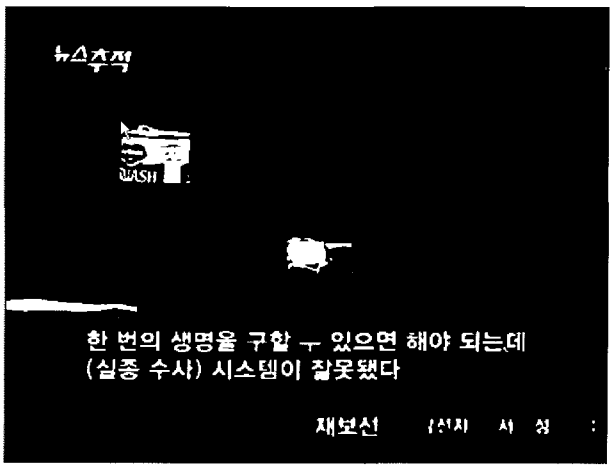



그림 7. 결과 영상3(Result Image3)

## 5. 결 론

본 논문에서는 영상에서 크기에 무관하게 자막을 추출하는 신경망을 이용한 연결 객체 기반의 자막 추출 방법을 제

안하였다. 기존의 자막 추출 방법은 다중 해상도 방법을 사용하여, 다양한 크기의 자막을 추출하였으나, 이러한 방법은 동일한 작업을 반복하는 부담이 존재한다. 제안하는 방법에서는 동일한 작업을 반복하지 않고도 다양한 크기의 자막이 추출 가능하였다.

차후에 수행할 연구는 작은 글자의 에지도 추출이 가능한 방법과 신경망 학습을 통한 자막 연결 객체 구분 방법은 100%를 보장할 수 없기 때문에 자막이 아니면서 자막으로 판별된 연결 객체를 제거하고, 자막이면서 생략된 연결 객체를 복원하는 연구를 할 것이다.

## 참 고 문 헌


[1] R. Lyu, J. Song, M. Cai, "A Comprehensive Method for Multilingual Video Text Detection, Localization and Extraction", *IEEE Transaction on Circuits and Systems for Video Technology*, Vol. 15, No. 2, pp. 243-255, 2005.

[2] 정종면, 차지훈, 김규훈, "디지털 비디오를 위한 획 기반 자막 추출 알고리즘", *퍼지 및 지능시스템학회 논문지*, Vol. 17, No. 3, pp. 297-303, 2007.

[3] K.C. Jung, K.I. Kim, A.K. Jain, "Text Information Extraction in Images and Video: A Survey", *Journal on Pattern Recognition*, Vol. 37, No. 5, pp. 977-997, 2004.

[4] E.K. Wong, M. Chen, "A Robust Algorithm for Text Extraction in Color Video", *IEEE Int'l Proc. Multimedia and Expo 2000(ICME 2000)*, Vol. 2, pp. 797-800, 2000.

[5] K.C. Jung, E.Y, Kim, "Automatic Text Extraction for Content-Based Image Indexing", *Lecture notes in Computer Science, Proc. 8th Pacific -Asia Conf. (PAKDD 2004)*, Vol. 3056, pp. 497-507, 2004.

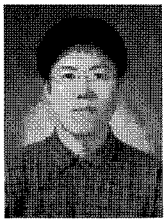
[6] H. Byun, I. Jang, Y. Choi, "Text Extraction in Digital News Video Using Morphology", *Lecture notes in Computer Science, Proc. 5th, Int'l Workshop on Document Analysis System*, Vol. 2423, pp. 341-352, 2002.

[7] Y.M.Y. Hasan, L. J. Karam "Morphological Text Extraction from Images", *IEEE Trans. Image Processing*, Vol. 9, No. 11, pp. 1978-1983, 2000.

[8] H.E. Jiaying, L.I. Shaofa, "Hybrid Chinese/English Text Identification in Web Images", *Proc. 3rd Int'l Conf. Image and Graphics(ICIG '04)*, pp. 361-364, 2004.

[9] R. C. Gonzalez, *Digital Image Processing, 2nd edition*, Prentice Hall, New Jzersey, 2001.


## 저 자 소 개

**정제희(Je-Hee Jung)**

2007년 : 목포 해양대학교 소프트웨어 과 학사

2007년~현재 : 성균관 대학교 전기전자 컴퓨터학과 석사과정

관심분야 : 인공지능, 자막 추출, 영상 처리

E-mail : gulingi@skku.edu

**윤태복(Tae Bok Yoon)**

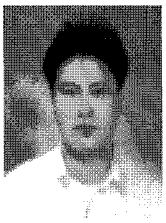
2001년 : 공주대학교 컴퓨터공학과 학사

2005년 : 성균관대학교 컴퓨터공학과 석사

2005년~현재 : 성균관대학교 컴퓨터공학과 박사과정

관심분야 : 게임AI, Data mining

E-mail : tbyoon@skku.edu



**김동문(Dong-Moon Kim)**

2006년 : 동국대 컴퓨터 공학 석사

2006년~현재 : 성균관대학교 전자 전기 컴퓨터공학과 석사 과정

관심분야 : 유비쿼터스 컴퓨팅, 데이터 마이닝

E-mail : skyscrape@skku.edu

**이지형(Jee-Hyoung Lee)**

1993년 : 한국과학기술원 전산학과 학사

1995년 : 한국과학기술원 전산학과 석사

1999년 : 한국과학기술원 전산학과 박사

2002년~현재 : 성균관대학교 정보통신공학부 조교수

관심분야 : 지능시스템, 기계학습, 온톨로지

E-mail : jhlee@ece.skku.ac.kr