색상레이어를 이용한 스팸메일 영상에서의 텍스트 영역 추출

김 지 수[†] · 김 수 형[††] · 한 승 완[†††] · 남 택 용[††††] · 손 화 정[†††††] · 오 성 열[††††††]

요 약

본 논문에서는 스팸메일 영상에서 텍스트 영역의 추출을 위한 색상 레이어기반의 알고리즘을 제안한다. CLTE(color layer-based text extraction)는 색상 레이어를 사용하여 영상을 8개로 나눈다. 8개 각각의 영상에서 연결요소를 추출한 후, 연결요소의 크기에 의해서 텍스트 영역과 비텍스트 영역을 분류하고 텍스트 영역을 추출한다. 또한, 추출된 텍스트 영역으로부터 회손된 획 정보를 복구하는 알고리즘을 제안한다. 이진영상내의 한글 문자에는 두 가지 형태의 손상된 획이 존재한다. 첫째 중성 획에 해당하는 'ㅣ' 나 'ㅡ' 등의 획들이 지워지는 경우와, 둘째 초·종성 획에 해당하는 'ㅁ'이나 'ㅇ'이 흑화소로 채워지는 경우가 있다. 제안한 알고리즘은 이러한 두 가지 손상된 획들을 복구해준다. 200개의 스팸메일 영상을 사용한 실험 결과 제안한 알고리즘이 기존의 텍스트 추출 알고리즘보다 10% 이상 우수함을 관측하였다.

키워드 : 텍스트 추출, 색상 레이어, 스팸메일 필터링

Extraction of Text Regions from Spam-Mail Images Using Color Layers

Ji-Soo Kim[†] · Soo-Hyung Kim[††] · Seung-Wan Han[†††] · Taek-Yong Nam[††††]
Hwa-Jeong Son[†††††] · Sung-Ryul Oh[††††††]

ABSTRACT

In this paper, we propose an algorithm for extracting text regions from spam-mail images using color layer. The CLTE(color layer-based text extraction) divides the input image into eight planes as color layers. It extracts connected components on the eight images, and then classifies them into text regions and non-text regions based on the component sizes. We also propose an algorithm for recovering damaged text strokes from the extracted text image. In the binary image, there are two types of damaged strokes: (1) middle strokes such as 'ㅣ' or 'ㅡ' are deleted, and (2) the first and/or last strokes such as 'ㅇ' or 'ㅁ' are filled with black pixels. An experiment with 200 spam-mail images shows that the proposed approach is more accurate than conventional methods by over 10%.

Key Words : Text Extraction, Color Layer, Spam-mail Filtering

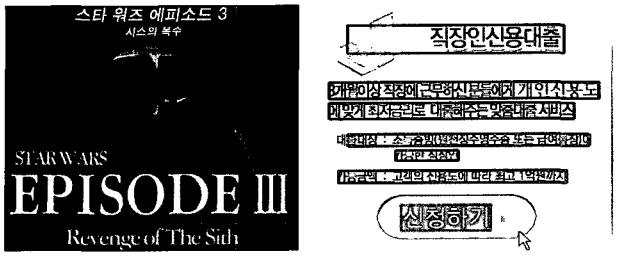
1. 서 론

인터넷과 월드와이드웹의 급속한 보급 그리고 컴퓨터 통신 기술의 발전으로 전자 메일은 많은 사람들이 사용하는 편리하고 효율적인 통신 수단으로 자리를 잡았다. 그러나 이러한 편리함에도 불구하고 전자메일 서버들이 늘어감에 따라 우리는 스팸메일(spam-mail)이라는 또 다른 전자 공해와 부딪히게 되었다. 스팸메일의 해결책으로 텍스트 필터링 기술을 통하여 텍스트 스팸메일을 필터링 할 수 있었지만, 텍스트메일 위주에서 영상메일로 변화된 이후부터 영상(정지영상, 동영상)과 함께 전달되어져 오는 불법적인 스팸메일은 정신적 스트레스, 집중도 저하, 정보에 대한 불신 등과 같은 정신적 손실뿐만 아니라 메일 장비의 과부하, 인력 낭비, 업무 시간 낭비 등 계량적 손실도 매우 커서 기업의 효율성에 매우 심각한 영향을 주고 있다. 이러한 스팸메일 및 유해정보들을 영상내의 텍스트 정보 인식 연구를 이용하여 텍스트 영역을 추출 및 인식하여 스팸메일인지 여부를 판단할 수 있다면 스팸메일이나 유해 정보가 사용자에게 전달되기 전에 미리 차단함으로써 경제적, 정신적 손실을 최소화하는데 크게 기여할 수 있을 것이다.

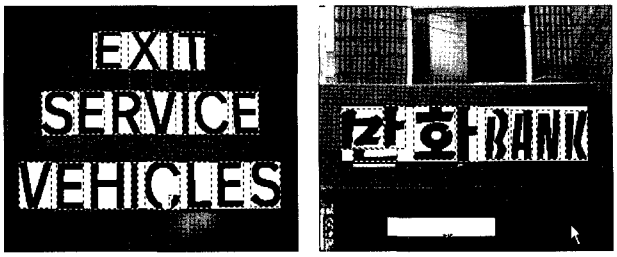
영상내의 텍스트 정보 인식을 위한 전처리 기술 연구는 크게 두 가지로 나누어 볼 수가 있다. 첫 번째 연구는 영상내에 포함된 텍스트 영역의 위치를 찾아서 추출해내는 텍스트 후보 영역 추출에 관한 연구이고[1-5, 11-14], 두 번째는 추출된 텍스트 영상을 이진화(binarization) 후 OCR(optical character recognition)을 사용하여 인식하는 연구이다[6-10].

본 논문에서는 스팸메일 영상에서 텍스트 정보를 효과적

† 준 회 원 : 전남대학교 전산학과 박사과정
†† 정 회 원 : 전남대학교 전자컴퓨터공학부 부교수
††† 정 회 원 : 한국전자통신연구원 선임연구원
†††† 정 회 원 : 한국전자통신연구원 정보보호 연구본부 능동보안기술연구팀 팀장
††††† 준 회 원 : 전남대학교 전산학과 박사과정
†††††† 준 회 원 : 전남대학교 전산학과 석사과정
논문접수 : 2006년 4월 27일, 심사완료 : 2006년 7월 24일



(그림 1) 텍스트 추출



(그림 2) 텍스트 이진화

으로 추출하기 위한 색상 레이어 기반 텍스트 추출 및 텍스트 획 복구 알고리즘을 제안한다. 제안한 시스템은 기존의 시스템과 비교하여 두 가지의 장점이 있다. 첫째 단순한 알고리즘으로 더 높은 텍스트 추출 성공률을 얻을 수 있으며, 둘째 텍스트 추출 결과가 곧바로 이진영상 형태로 제공된다는 점이다.

논문의 구성은 다음과 같다. 2장에서는 영상내의 텍스트 정보 인식을 위한 관련 연구를, 3장에서는 제안한 알고리즘을, 4장에서는 제안한 방법에 대한 실험 결과를 기술하며, 5장에서는 본 논문의 결론을 맺는다.

2. 관련 연구

영상내의 텍스트 정보 인식을 위한 전처리 기술로는 텍스트 영역 추출과 텍스트 영역의 이진화 및 인식 두 가지로 구성된다.

2.1 텍스트 영역 추출

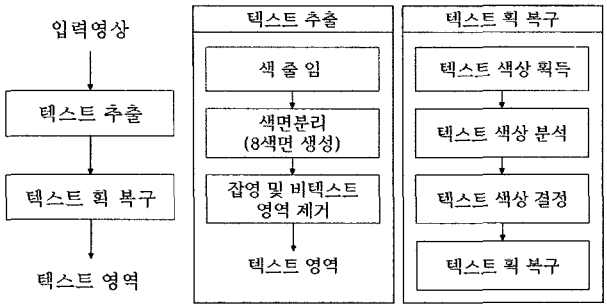
텍스트 추출(text extraction) 기술은 영상내에 인위적으로 삽입되거나 자연적으로 포함된 텍스트들의 영역을 찾아내주는 기술을 의미한다. (그림 1)은 텍스트 추출 사례를 보여준다.

Jain[1] 등은 입력 영상으로 자연영상, 웹영상, 동영상을 사용하였다. 전처리는 다중 값(multivalued) 영상 분해에 의한 9단계의 지역적 편차 영상을 구하게 되며, BAG 알고리즘과 연결 요소 분석을 통한 텍스트 영역을 추출하였다. Hoya[2] 등은 입력 영상으로 그레이 정지영상을 사용하였다. 전처리는 8×8 크기의 윈도우(단위 : 픽셀)을 이용한 지역적 이진화 방법을 사용하였으며, 인접 영역의 명도차를 고려하여 텍스트 영역을 추출하였으나, 영상 전체에 나타나는 불필요한 에지 값들 때문에 다양한 종류의 연산과 시간이 많이 걸린다는 단점이 존재하였다. Lienhart[3] 등은 입력 영상으로 동영상 프레임을 입력 영상으로 사용하였다. 전처리는 분할/병합 알고리즘을 적용하였으며, 블록 매칭과 콘트라스트(contrast) 분석을 이용한 텍스트 영역을 추출하였다. Messelodi[4] 등은 입력 영상으로 이진 정지영상, 그레이 정지영상을 사용하였다. 전처리는 지형적 특징, 지역적 대비 특징 및 필터들을 사용하였다. 추출되어진 텍스트 영역 각각의 중심점과 투영된 프로파일을 이용하여 텍스트의 기울기를 추정하는 방법 등을 제안하였다. Zhong[5] 등은 입력 영상으로 색상(자연색) 정지영상을 사용하였다. 전처리는 색상 히스토그램을 계산한 후에 비슷한 색상 영역별로 분할/합병 알고리즘을 적용하였다. 또한 전처리 부분에서 지워져 버린 텍스트 영역을 복원한 후에 텍스트 영역을 추출하는 방법을 제안하였다.

2.2 텍스트 영역 이진화 및 인식

텍스트 이진화(text binarization) 기술은 텍스트 추출 기술에 의해서 찾아진 텍스트 후보 영상에서 텍스트 영역과 배경 영역을 분류해주는 기술이다. (그림 2)는 텍스트 이진화 예제를 보여준다.

Hori[6] 등은 입력 영상으로 그레이 동영상을 사용하였다. 추출된 텍스트 영상의 이진화 방법으로는 픽셀의 명암 값에 의해서 히스토그램을 구한 후에 텍스트 객체와 배경 객체를 구분할 수 있는 임계값(threshold)을 찾아내어 이진화를 수행하였다. 이진화된 텍스트 영상을 인식하기 위해 상용 문서인식기를 사용하였다. Ohya[7] 등은 입력 영상으로 그레이 정지 영상(길거리 간판에서 취득한 영상, 자동차 번호판 영상)을 사용하였다. 추출된 텍스트 영상의 이진화 방법으로는 입력 영상을 8×8 크기의 작은 블록으로 나눈 후 각 영역의 로컬(local) 임계값을 구하고, 인접 영역의 로컬 임계값과 비교하여 텍스트 객체와 배경 객체를 이진화 하였다. 이진화된 텍스트 영상을 인식하기 위해 상용 문서인식기를 사용하였다. Wang[8] 등은 입력 영상으로 색상 정지 영상을 사용하였다. 추출된 텍스트 영상의 이진화 방법으로는 위상학적 특징(topological feature), 가중치가 부여된 유클라디안 거리(weight-euclidean distance) 및 변형된 코스-파인 퍼지 씨민스(modified coarse-fine fuzzy c-means) 알고리즘을 이용하여 텍스트 객체와 배경 객체를 이진화 하였다. 이진화된 텍스트 영상을 인식하기 위해 상용 문서인식기를 사용하였다. Wolf[9] 등은 입력 영상으로 그레이 동영상을 사용하였다. 추출된 텍스트 영상의 이진화 방법으로는 Niblack's 알고리즘을 이용하여 텍스트 객체와 배경 객체를 이진화 하였다. 이진화된 텍스트 영상을 인식하기 위해 상용 인식기 Finereader 5.0을 사용하였다. Wu[10] 등은 입력 영상으로 색상 정지영상과 그레이 영상(동영상, 사진, 신문, 잡지)을 사용하였다. 추출된 텍스트 영상의 이진화 방법으로는 그레이 값에 의해서 히스토그램을 구한 후에 텍스트 객체와 배경 객체를 구분할 수 있는 임계값을 찾아내어 이진화를 수행하였다. 이진화된 텍스트 영상을 인식하기 위해 상용 인식기 Caere's Omnipage Pro 8.0을 사용하였다.



(ㄱ) 시스템 구성도 (ㄴ) 텍스트 추출 (ㄷ) 텍스트 획 복구

(그림 3) 제안하는 시스템

〈표 1〉 스팸 메일 영상 분석 결과

영상 가로 크기(화소)	300 - 980
영상 세로 크기(화소)	250 - 970
색상 값	24비트 색상
영상 해상도(dpi)	72 - 96

(그림 4) 원본 영상의 예

3. 제안한 시스템

본 논문에서 제안하는 시스템은 (그림 3)과 같이 2단계로 구성되었다. 첫 번째 단계에서는 색줄임 연산, 색상 레이어 영상 생성, 잡영 및 비텍스트 영역 제거 및 연결요소 분석을 통해서 단어 단위 텍스트 영역을 추출하게 된다. 두 번째 단계에서는 텍스트 영역의 색상을 획득, 분석 및 결정하여 회손된 획을 복구하고 잡영 영역을 제거한 후 이진영상 형태의 텍스트 영역을 추출하게 된다.

(그림 4)는 실험에 사용한 원본 영상들을 나타낸다. 본 논문의 실험에서는 전남대학교 이메일 서버 및 웹메일 서버(다음, 네이트, 네이버)를 통해서 수집한 200개의 스팸메일 영상을 사용하였다. <표 1>은 실험에 사용한 원본 영상의 크기 및 해상도를 정리하였다.

3.1 텍스트 추출(color layer-based text extraction)

제안한 텍스트 추출 알고리즘은 기존의 다른 알고리즘보다 단순하지만 더 좋은 결과를 얻을 수 있도록 설계되었다.

3.1.1 색줄임 및 색상 레이어 영상

스팸메일 영상은 그 목적이 광고성을 가지고 있기 때문에 텍스트 영역과 배경 영역의 콘트라스트가 대조적인 특징을 가지고 있다. 이러한 특징으로 인해 색줄임 연산(color reduction process)을 수행해도 텍스트 영역과 배경 영역이 여전히 구별 가능한 특징을 갖는다. 영상처리에서 색줄임 연산을 통해서 얻을 수 있는 장점은 계산량을 줄일 수 있고, 잡영을 제거할 수 있다. 색 줄임 연산은 화소의 r, g, b 각 요소의

(ㄱ) 입력 영상 (ㄴ) 색줄임 영상

(그림 5) 입력 영상 및 색줄임 영상

〈표 2〉 레이어 영상 생성 알고리즘

For 색 줄임 영상에 있는 모든 화소
  if $C(R,G,B)_{y,x} = (0,0,0)$ *then* $L_1(y,x) = 1$(레이어 1번영상)
  else if $C(R,G,B)_{y,x} = (0,0,128)$ *then* $L_2(y,x) = 1$
    (레이어 2번영상)
    ...
  else if $C(R,G,B)_{y,x} = (128,128,0)$ *then* $L_7(y,x) = 1$
    (레이어 7번영상)
  else if $C(R,G,B)_{y,x} = (128,128,128)$ *then* $L_8(y,x) = 1$
    (레이어8번영상)
End for

하위 7비트를 제거하여 상위 1비트만을 남기는 비트 줄임(bit dropping) 방법을 이용한다. 색줄임 후 결과 영상은 최대 8개의 색(0=red,0=green,0=blue), (0,0,128), (0,128,0), (0,128,128), (128,0,0), (128,0,128), (128,128,0), (128,128,128)으로 표현된다. (그림 5)(ㄱ)는 입력 영상을 보여주며, (그림 5) (ㄴ)는 비트 제거에 의한 색줄임 결과 영상을 보여준다. (그림 5) (ㄴ)을 보면 색줄임 연산을 수행한 후에도 텍스트 영역과 배경 영역의 대조적인 콘트라스트 때문에 텍스트 정보가 여전히 강조되어 있다.

<표 2>의 알고리즘을 사용하여 8개의 색상 레이어 영상을 생성한다.

여기서 $C(R,G,B)_{y,x}$는 (그림 5) (ㄴ)과 같은 색줄임 영상, $L_i(y,x)$는 이진영상 형태의 레이어 영상을 나타낸다. (그림 6)은 <표 2>에 의해서 분류된 8개의 색상 레이어 영상의 예를 나타낸다.

3.1.2 잡영 및 비 텍스트 영역 제거

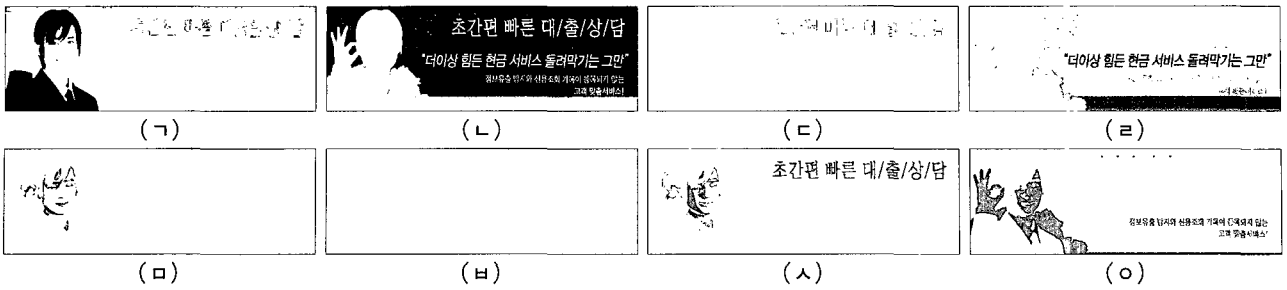
각 색상 레이어 영상에 있는 연결요소(connected components)를 blob coloring[15] 방법을 적용하여 분리한 후, 분리된 각 영역들의 사각상자(bounding box)의 크기를 구한 후에 식 (1), (2), (3)중 하나라도 만족하면 잡영 및 비텍스트 영역으로 판단하여 제거하게 된다. (그림 7) (ㄴ)은 잡영 및 비텍스트 영역을 제거하고 난 후의 영상이다.

$$CC\_H\_L_i < T_1 \text{ OR } CC\_W\_L_i < T_2 \cdots \quad \text{식 (1)}$$

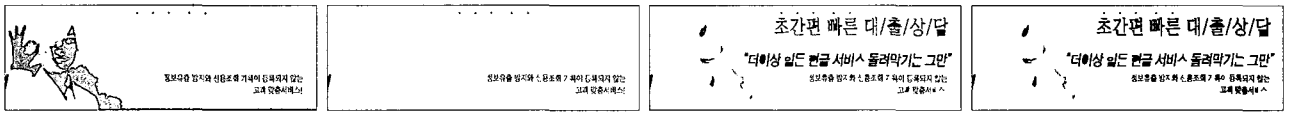
$$CC\_H\_L_i > T_3 \text{ OR } CC\_W\_L_i > T_4 \cdots \quad \text{식 (2)}$$

$$CC\_W\_H\_R_i > T_5 \cdots \quad \text{식 (3)}$$

여기서 $CC_H_L_i$는 i번째 연결요소의 세로 크기, $CC_W_L_i$는 i번째 연결요소의 가로 크기이며, $CC_W_H_R_i$는 i번째 연결요소의 가로와 세로의 비율을 나타내는데 이것은 연결요소의 가로와 세로 중 큰 값을 작은 값으로 나누어서 구하게



(그림 6) (그림 5) (ㄴ) 영상에 대한 8개의 색상 레이어 영상

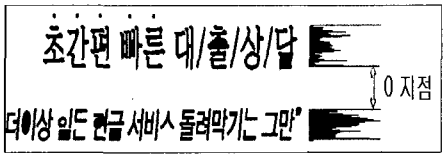


(ㄱ) 레이어 8번 영상 (ㄴ) 식 (1, 2, 3) 적용 후 영상

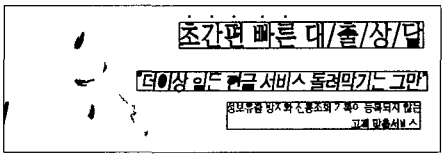
(그림 7) 잡영 및 비텍스트 영역 제거 영상

(ㄱ) 잡영 제거 된 영상 (ㄴ) 이진영상

(그림 8) 잡영 제거 및 이진영상



(그림 9) 수평 방향 투영 프로프일 영상



(그림 10) CLTE를 사용한 텍스트 추출 결과

된다. 본 논문의 실험에서는 T_1, T_2는 2픽셀을 설정하였으며, T_3, T_4는 각각 영상의 가로 및 세로 크기의 절반으로 설정하였으며, T_5는 5로 설정하였다.

(그림 8) (ㄱ)은 레이어 8번 영상을 나타내며, (그림 8) (ㄴ)은 식 (4)에 의해서 이진화된 영상이다.

$$L(y,x)=\begin{cases}1, \text{if } L_i(y,x)=1 \ \exists_i, i=1,2,\cdots,8 \ \cdots \\ 0 \ otherwise\end{cases} \quad \text{식(4)}$$

즉, $L(y,x)$는 8개의 색상 영상 $L_i(1 \le i \le 8)$을 하나로 논리합(or) 연산한 이진영상을 나타낸다.

3.1.3 텍스트 영역 추출

텍스트 영역 추출은 문자 및 단어 단위로 추출할 수 있다. 본 논문에서는 단어 단위 영역 추출을 수행한다. 영역 추출 과정은 다음과 같이 두 단계로 구성된다. 첫 번째 단계에서는 (그림 9)와 같이 수평 방향 투영 프로파일을 값을 구한 후에 수평 방향 문자열 분할 점을 탐색한다. 수평 방향 투영 프로파일 값이 0인 경우 해당 지점을 수평 방향 문자열 분할 점으로 선택하게 된다. 두 번째 단계에서는 blob coloring[15] 방법을 적용하여 연결요소를 분리한 후, 분리된 각 요소들의 사각상자의 크기를 구한다. 사각상자 사이의 거리가 수평 방향 투영 프로파일의 높이보다 작거나 외곽 상자의 좌표가 인접 상자들과 겹치면 한 개의 영역으로 결합하여 단어 단위 텍스트 영역을 구성하게 된다. (그림 10)은 (그림 8) (ㄴ)의 이진영상에 단어 단위로 추출된 텍스트 영역을 보여 주고 있다.

CLTE 시스템은 기존의 다른 텍스트 추출 시스템에 비해서 단순한 알고리즘을 사용하여 텍스트 영역을 추출하기 때문에 다음과 같은 두 가지의 문제가 발생한다. 첫째 식 (1, 2, 3)을 이용하여 잡영과 비텍스트 영역을 제거하기 때문에 'ㅣ'나 'ㅡ'획들이 지워지는 문제점이 발생 한다. 둘째 (그림 8)과 같이 분리된 8개의 색상 영상을 논리합 연산을 사용한 하나의 이진영상으로 만들기 때문에 'ㅇ', 'ㅎ' 및 'ㅁ' 같은 획 내부에 채움 현상이 발생한다. 이러한 두 가지 문제는 단순히 텍스트 영역의 위치만을 찾는 응용분야에서는 문제가 없지만 문자인식이나 키워드 검색(keyword spotting)시에는 인식 오류나 검색 오류의 주원인이 된다. (그림 10)은 이와 같은 문제점을 보여준다. 다음 절에서 제안하는 텍스트 획 복구 알고리즘으로 CLTE가 갖는 두 가지의 문제점을 해결할 수 있다.

3.2 텍스트 획 복구

CLTE를 통해 추출된 이진화 영상 형태의 텍스트 영상은 한글 중성 획 'ㅣ' 나 'ㅡ' 의 획이 지워지는 단점과 한글의 초성 및 종성 획 'ㅁ', 'ㅇ' 및 'ㅎ' 에 흑화소로 채워지는 문제점을 가지고 있다. (그림 5) (ㄱ)은 원본 입력 영상을 나타낸다. (그림 10)은 CLTE를 이용하여 텍스트 영역을 추출한 영상이다. 영상내의 사각상자는 추출된 텍스트 영역이다. 본 논문에서는 회손된 획 때문에 발생하는 인식 오류나 검색 오류 문제를 색상 정보를 이용한 회손된 획 정보를 복구하는 알고리즘을 제안하여 CLTE의 단점을 보완하였다.

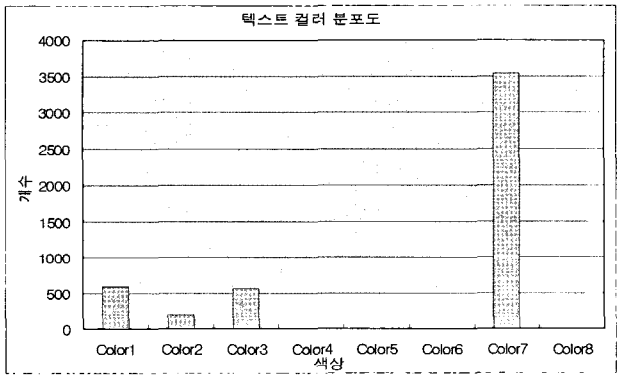
3.2.1 색상 정보 획득 및 텍스트 색상 결정

(그림 10) 영상에는 사각상자로 표시된 3개의 텍스트 영역이 있다. 박스내의 문자 중에는 중성 획이 지워지거나

(그림 11) 확대된 텍스트 영상

〈표 3〉 색상 정보 획득 및 텍스트 색상 결정 알고리즘

For $i=1\cdots n$: $CLTE$를사용하여추출된사각상자수
  $C\_1_i \cdots C\_8_i$ 값0으로 초기화
  For i번째 사각상자 좌표 내의 모든 픽셀 //색상 정보 획득
    if $L(y,x)=1$ AND $C(R,G,B)_{y,x}=(0,0,0)$ $then$ C_1_i += 1
    else if $L(y,x)=1$ AND $C(R,G,B)_{y,x}=(0,0,128)$ $then$
        C_2_i += 1
        ...
    else if $L(y,x)=1$ AND $C(R,G,B)_{y,x}=(128,128,0)$ $then$
        C_7_i += 1
    else if $L(y,x)=1$ $ADN C(R,G,B)_{y,x}=(128,128,128)$ $then$
        C_8_i += 1
  End for
  For $j=1\cdots 8$ //텍스트 색상 결정
    if(C_j_i 중 가장 큰 값) then TC_i = C_j_i(가장 큰 값)
  End j
End for i



(그림 12) 텍스트 색상 분포도

초·종성 획에 흑화소로 채워져 있음을 발견할 수 있다. (그림 11)은 이중 하나의 텍스트 영역을 확대한 영상이다.

텍스트 영역에서의 색상 정보를 획득하여 텍스트 색상을 결정하는 방법은 <표 3>의 알고리즘과 같다.

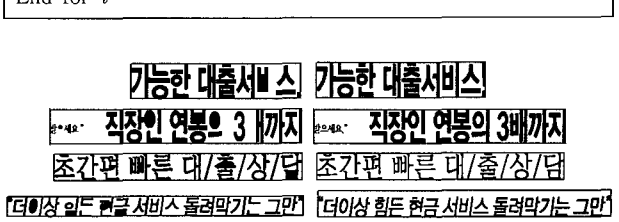
여기서 $L(y,x)$는 CLTE의 출력 영상이고, $C(R,G,B)_{y,x}$는 (그림 5)(ㄴ)에서와 같은 색줄임 영상을 나타낸다. $L(y,x)$의 화소 값이 1은 흑화소이고, $C_1_i \cdots C_8_i$는 i번째 사각 상자 내의 color1부터 color8까지의 누적 합을 나타내며, TC_i는 i번째 사각 상자의 텍스트 색상이다. <표 3> 알고리즘에 의해서 (그림 11)에 있는 color1부터 color8까지의 누적 합을 그려보면 (그림 12)와 같다. (그림 12)에서는 color7번이 텍스트 색상에 해당하는 것을 알 수 있다.

3.2.2 문자 획 복구

(그림 12)에서와 같이 색상의 분포도를 분석하면 텍스트 색상 및 잡영의 색상을 알 수 있게 되어 CLTE의 단점인 획이 제거되고 채움이 발생한 단점을 보완할 수 있다. <표 4>는 문자 획 복구를 위한 알고리즘이다.

〈표 4〉 획 복구 알고리즘

For $i=1\cdots n$: $CLTE$를사용하여추출된사각상자수
  For i번째 사각 상자 좌표 내의 모든 픽셀
    if $C(R,G,B)_{y,x}=TC_i$ $then$ $L(y,x)=1$ //텍스트 획 복구
    if $C(R,G,B)_{y,x}\neq TC_i$ $then$ $L(y,x)=0$ //잡영 제거
  End for
End for i



(ㄱ) 획 복구 전 이미지 (ㄴ) 획 복구 후 이미지

(그림 13) 획 복구 후 이미지

〈표 5〉 실험 데이터 분석

총 스팸메일 개수	1369개
영상 스팸메일 개수	1226개
텍스트 스팸메일 개수	143개

여기서 $C(R,G,B)_{y,x}$는 색줄임 영상이고, $L(y,x)$는 CLTE의 출력 영상을 나타내며, TC_i는 i번째 사각 상자의 텍스트 색상이다.

(그림 13)은 <표 4>에 의해서 회손된 획이 복구된 영상들을 보여준다. <표 4>에 의해서 CLTE가 가지는 단점인 지워진 텍스트 획이 복구가 되고 채움 현상이 발생한 획에서 불필요한 부분을 제거할 수 있다.

(그림 13)(ㄱ)의 영상은 획 복구 전 영상이고 (ㄴ)는 획 복구 후 영상이다. 본 논문에서 제안한 획 복구 알고리즘이 회손된 획 복구에 우수한 효과가 있음을 알 수가 있다. 제안한 시스템의 최종 결과는 이진영상 형태의 텍스트 추출 결과 영상이다.

4. 실험 및 평가

실험 및 평가를 위해 전남대학교 이메일 서버 및 웹메일 서버(다음, 네이트, 네이버)를 통해서 1369개의 스팸메일을 수집하였다. <표 5>는 메일에 영상의 포함 여부를 분류한 결과이다. 분류 결과를 보면 현재 스팸메일은 텍스트만을 사용하여 보내는 것 보다 영상을 포함하여 보내어지는 메일이 더 많다는 것을 알 수가 있다.

실험에서는 <표 5>의 스팸메일 영상 중 200개를 선택하고, 이들을 대상으로 2가지를 측정하였다. 즉, 텍스트 영역 추출 성공률과 텍스트 영역 추출에 걸린 시간을 측정하였다. 텍스트 영역 추출 및 시간 성능의 비교를 위해 제안한 시스템, Kim[13], Choi[14]을 사용하였다. Choi[14] 시스템은 입력 영상의 크기에 제한을 받는 관계로 입력 영상 자체는 그대로 두고 캔바스(canvas) 크기를 1024×768 크기로 재조정하여 실험에 사용하였다. 실험에 사용한 컴퓨터는 펜티엄4

〈표 6〉 텍스트 영역 추출 결과

Image	System	Total	True	Part	Error	False
spam images (200)	제안한 시스템	1241	984	133	124	89
		Precision = 81.6%		Recall = 79.3%		
	Kim[13]	1241	800	263	178	283
		Precision = 59.4%		Recall = 64.53%		
	Choi[14]	1241	769	12	460	130
		Precision = 84.4%		Recall = 62%		

〈표 7〉 추출에 걸린 평균 시간

System	Elapsed time
제안한 시스템	5.103 sec
Kim[13]	0.916 sec
Choi[14]	5.503 sec

1.7GHz CPU, 512M RAM을 사용한 PC이다.

<표 6>은 텍스트 추출 결과에 대한 실험 결과를 보여준다. Total은 영상내의 전체 텍스트 라인 개수, True는 시스템에서 정확히 추출한 텍스트 영역 개수, Part는 텍스트 영역의 1/2이상 찾은 개수, Error는 찾지 못한 텍스트 영역 개수 그리고 False는 비텍스트 영역을 찾은 개수를 말한다. 텍스트 영역은 영상내에 존재하는 모든 텍스트 영역을 대상으로 실험하였다. Precision은 식 (5)에 의해서 구했으며, Recall은 식 (6)에 의해서 구했다.

$$\text{Precision} = \frac{True}{True + Part + False} \quad \cdots \quad \text{식 (5)}$$

$$Recall = \frac{True}{True + Part + Error} \quad \cdots \quad \text{식 (6)}$$

제안한 시스템은 Presicion이 81.6%를 Recall은 79.3%를, Kim[13]은 Presicion이 59.4%를 Recall은 64.53%를, Choi[14]는 Presicion이 84.4%를 Recall은 62%를 보였다. 결과에서 알 수 있듯이 제안한 텍스트 추출 시스템이 다른 2개의 시스템보다 안정적으로 텍스트 영역을 추출하는 것을 알 수가 있다.

<표 7>은 텍스트 영역 추출에 걸린 평균 시간을 보여준다. 이 결과에서는 Kim[13]이 다른 2개의 시스템보다 5배 이상 빠른 속도를 보였다. 제안한 시스템은 레이어 8개를 생성하여 텍스트를 추출하기 때문에 속도가 느려지는 단점이 있다.

(그림 14)는 제안한 시스템을 사용하여 추출한 결과 영상에서 오류(Part, Error 및 False) 영상을 보여준다. 오류 영상들에서 나타나는 원인을 분석해 보면 크게 3가지로 분석할 수 있다. 첫째 비텍스트 영역을 추출하는 경우, 둘째 임계값보다 작은 텍스트 영역들을 놓치는 경우, 셋째 텍스트 영역보다 훨씬 크게 텍스트 영역을 추출하는 경우로 분석할 수 있다. 이 세 가지 문제들에 대한 향후 연구 방향으로는 텍스트 영역 검증기의 개발이다.

(그림 15)는 3가지 텍스트 추출 시스템의 결과를 보여준다. 제안한 시스템은 다른 시스템과는 달리 텍스트 추출 결과를 이진영상 형태로 추출해주는 장점을 가지고 있다.

5. 결 론

본 논문에서는 스팸메일 영상 내의 텍스트 영역을 추출하기 위해서 색상 정보를 이용하는 시스템을 제안하였고, 추출된 텍스트 영상내에서 획들이 지워지는 단점과 임의의 획이 흑화소로 채워지는 문제를 해결하기 위한 획 복구 알고리즘

(ㄱ) Part 영상 (ㄴ) False 영상 (ㄷ) Error 영상

(그림 14) 텍스트 추출 오류 영상

(ㄱ) 제안한 시스템 (ㄴ) Kim[13] (ㄷ) Choi[14]

(그림 15) 텍스트 추출 결과

을 제안하였다. 텍스트 추출 실험 결과를 보면 제안한 시스템이 기존의 다른 시스템들 보다 10% 이상 우수한 추출 성공률을 보였다. 즉, 제안한 알고리즘이 단순하지만 다른 시스템보다 안정적으로 텍스트 영역을 추출해줌을 알 수가 있다. 향후 연구 과제로는 텍스트 인식기를 개발하고 제안한 시스템과 결합한 후 스팸메일 필터 시스템을 개발하는 것이다.

참 고 문 헌

[1] A. K. Jain, B. Yu, "Automatic Text Location in Images and Video Frames," Pattern Recognition, Vol.31, No.12, pp.2055-2076, 1998.

[2] J. Hoya, A. Shio and S. Akamatsu, "Recognizing Characters in Scene Images," IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.16, No.2, pp.67-82, 1995.

[3] R. Lienhart, F. Stuber, "Automatic Text Recognition in Digital Videos," Image and Video Processing IV, SPIE, 1996.

[4] S. Messelodi and C. M. Modena, "Automatic Identification and Skew Estimation of Test Lines in Real Scene Images," Pattern Recognition, Vol.32, No.5, pp.701-810, 1999.

[5] Y. Zhong, K. Karu and A. K. Jain, "Locating Text in Complex Color Images," Pattern Recognition, Vol.28. No.10, pp.1523-1535, 1995.

[6] O. Hori, "A Video Text Extraction Method for Character Recognition," Proc. Fifth International Conference on Document Analysis and Recognition, pp.25-28, 1999.

[7] J. Ohya, A. Shio and S. Akamatsu, "Recognizing Characters in Scene Images," IEEE Trans. Pattern Analysis and Machine Intelligence, PAMI-16(2), pp. 214-220, 1994.

[8] X. Wang, X. Ding and C. Liu, "Character Extraction and Recognition in Natural Scene Images," Proc. Sixth International Conference on Document Analysis and Recognition, pp. 1084-1088, 2001.

[9] C. Wolf and J.M. Jolion, "Extraction and Recognition of Artificial Text in Multimedia Documents," Pattern Analysis and Applications, Vol.6, No.4, pp.306-326, 2003.

[10] V. Wu, R. Manmatha and E.M. Riseman, "An Automatic System to Detect and Recognize Text in Images," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.21, No.11, pp.1224-1229, 1999.

[11] J. Zhang, X. Chen, A. Hanneman, J. Yang and A. Waibel, "A Robust Approach for Recognition of Text Embedded in Natural Scenes," Proc. 16th International Conference on Pattern Recognition, Vol.3, pp.204-207, 2002.

[12] 김지수, 김수형, "명도 정보를 이용한 자연 이미지에서의 텍스트 영역 추출," 한국정보처리학회 호남·제주지부 학술발표논문집, Vol.3, No.1, pp.127-132, 2003.

[13] 김지수, 김수형, 최영우, "명도 정보와 Split/Merge 분할을 이용한 자연 이미지에서의 텍스트 영역 추출," 한국정보과학회 논문지 : 소프트웨어 및 응용, Vol.32, No.6, pp.502-511, 2005.

[14] Y.J. Song, K.C. Kim, Y.W. Choi, H.R. Byun, S.H. Kim, S.Y. Chi, D.K. Jang, Y.K. Chung, "Text Region Extraction and Text Segmentation on Camera-captured Document Style Images," Proc. of the 7th International Conference on Document Analysis and Recognition, Vol.1. pp.172-176, 2005.

[15] D.H. Ballard and C.M. Brown, Computer Vision, Prentice-Hall, 1982.

김 지 수

e-mail : kimjisoo@iip.chonnam.ac.kr
1996년 광주대학교 컴퓨터공학과(학사)
2003년 전남대학교 전산학과(이학석사)
2003년 3월~현재 전남대학교 전산학과 박사과정
관심분야 : 인공지능, 패턴인식, 문자인식

김 수 형

e-mail : shkim@chonnam.ac.kr
1986년 서울대학교 컴퓨터공학과(학사)
1988년 한국과학기술원 전산학과 (공학석사)
1993년 한국과학기술원 전산학과 (공학박사)
1993년~1996년 삼성전자 멀티미디어 연구소 선임연구원
1997년~현재 전남대학교 전자컴퓨터공학부 부교수
관심분야 : 인공지능, 패턴인식, 문서영상 정보검색, 유비쿼터스컴퓨팅

한 승 완

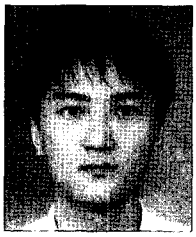
e-mail : hansw@etri.re.kr
1994년 전남대학교 전산학과(학사)
1996년 전남대학교 전산통계학과(석사)
2001년 전남대학교 전산통계학과(박사)
2001년~현재 한국전자통신연구원 선임연구원
관심분야 : 암호이론, 네트워크 보안, 계산이론, 알고리즘 등

남 택 용

e-mail : tynam@etri.re.kr
1987년 충남대학교 계산통계학과 이학사
1990년 충남대학교 계산통계학과 이학석사
2005년 한국외국어대학교 전자정보공학과 공학박사
1987년~현재 한국전자통신연구원 정보보호연구단 보안게이트웨이연구팀 팀장(책임연구원)
관심분야 : 개인정보보호, 콘텐츠 보안, 정보분류 등

손 화 정

e-mail : sonhj@iip.chonnam.ac.kr
2001년 전남대학교 통계학과(학사)
2004년 전남대학교 전산학과(이학석사)
2004년~현재 전남대학교 전산학과 박사과정
관심분야 : 패턴인식, 문자인식, 영상처리



오 성 열

e-mail : acecap@iip.chonnam.ac.kr
2005년 목포대학교 멀티미디어학과 (이학사)
2006년 3월~현재 전남대학교 전산학과 석사과정
관심분야 : 인공지능, 패턴인식, 문자인식 등