칼라정보에 기반한 텍스트 영역 추출에서의 지워진 획 복구

김선형*, 김지수*, 김수형*
*전남대학교 전산학과
e-mail:skimh@chonnam.ac.kr

Recovery of Erased Character Strokes in the Extraction of Text Using Color Information

Seon-Hyung Kim*, Ji-Soo Kim*, Soo-Hyung Kim*
*Dept of Computer Science, Chonnam National University

요 약

자연영상이나 스팸메일 영상으로부터 텍스트 영역을 추출하고 추출한 텍스트 영역에 이진화를 수행하고 나면 가로 방향이나 세로획 방향으로 놓여 있는 "l" 그리고 "ㅡ" 에 해당하는 한글의 종성부분이 이미지 내의 잡영을 지울 때 종종 지워지는 결과를 볼 수 있다. 이렇게 지워진 획 부분을 되살리기 위한 방법으로 텍스트 Hinting 알고리즘을 제안한다. 텍스트 Hinting 알고리즘은 이진화된 이미지의 텍스트 픽셀 위치와 동일한 좌표에 해당하는 원본 이미지의 RGB 값을 추출하여 추출된 텍스트 후보 영역의 색상을 알아낸다. 추출된 텍스트 색상 레이어 이미지와 이진화된 이미지에 OR연산을 수행하게 되면 지워진 획 부분을 복원할 수 있다. 제안한 방법을 스팸 이미지에 적용한 결과 텍스트 추출결과를 획기적으로 개선할 수 있음을 보였다.

1. 서론

21세기 멀티미디어 시대를 맞아 인터넷과 최첨단 정보통신 기술의 발전은 개인 휴대 통신 및 영상 장비들의 비약적인 발전을 통해 대량의 영상 및 음성 정보가 교환되고 인간의 모든 생활영역에서 중요한 자리를 차지하고 있다. 멀티미디어 시대의 핵심은 영상정보로서 하루에도 엄청난 양의 정지 영상 및 동영상이 생성 및 저장되고 있다. 이러한 영상들내에 포함되어진 텍스트 정보들은 영상내의 내용을 함축적이고 구체적으로 표현해줄 수가 있다. 이러한 정보들을 실시간에 추출 및 인식 할 수만 있다면 로봇 자동 주행 보조 시스템 및 시각 장애인 주행 보조 시스템, 고용량 비디오 프레임의 자동 검색 및 색인 시스템, 텍스트 자동 번역 시스템, 스팸 메일 필터링 시스템등과 같은 다양한 분야에서 사용될 수 있다.

이미지에서 텍스트 영역을 찾고 찾아진 영역에 이진화를 수행하고 나면 한글 중성에 해당하는 "l" 나 "ㅡ"의 획들이 지워지게 되고 한글의 초성에 해당하는 "ㅁ"이나 "ㅇ"에 흑화소로 채워지는 문제가 발생한다. 이러한 문제를 해결하기 위하여 본 논문에서 제안한 텍스트 힌팅을 사용한 결과 지워진 획들이나 흑화소로 채워진 글자들을 복구할 수 있다.

논문의 구성은 다음과 같다. 2장에서는 장면 텍스트 추출에 대한 관련 연구를 3장에서는 텍스트 힌팅 시스템에 대해서 설명하고 4장에서는 제안한 방법에 대한 실험 결과를 보이며 5장에는 본 논문의 결론을 맺는다.

2. 장면 텍스트 추출에 관한 연구

장면 텍스트 추출에 관한 연구는 크게 명도 정보를 이용한 텍스트 추출 연구와 색 정보를 이용한 추출 연구로 분류할 수 있다.

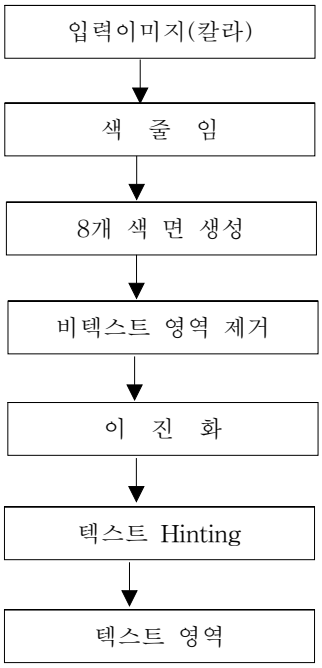
명도 정보를 이용한 텍스트 추출 연구로 Kim [1]

등은 전처리 과정으로 메디안 필터와 케니 에지 필터를 사용하였으며 최소자승법과 반복적인 RLS를 이용하여 텍스트 영역을 추출하여 외곽사각형 정보를 이용하여 검증하였다. Roh [2] 등은 긴 선 요소를 제거하고 반복적인 RLS를 적용한 후에 텍스트 후보 영역을 추출하였다. Kim [4] 등은 명도 정보 분석(GIA)과 분할/합병 분석(SMA)과 두 방법을 결합한 HAM을 이용하여 텍스트 영역을 추출하였고 RFFA 방법을 이용하여 검증하였다. Choi [5] 등은 텍스트의 색 연속성, 밝기 변화 및 색 변화와 같은 낮은 수준의 이미지 특징을 이용하여 텍스트 영역을 추출하였고 다해상도 웨이블릿 변환을 이용하여 검증하였다. Lin [7] 등은 에지 이미지 및 그레이 이미지 내에 존재하는 연결요소들의 히스토그램 분석 및 연결요소의 크기에 따라 문자단위로 그룹화하여 텍스트 영역을 추출하였다. Ewerth [8] 등은 전처리 이미지에 웨이블릿 변환을 적용하였고 텍스트 영역과 배경영역의 분류 클러스터로 최소 유클라디안 거리를 사용하여 텍스트 영역을 추출하였고 다해상도 웨이블릿 변환을 이용하여 검증하였다.

색 정보를 이용한 텍스트 추출 연구로 Roh [2] 등은 색 병합된 영상에 모폴로지 닫힘(Closing) 연산을 수행하고 연결 요소 분석 방법에 의하여 텍스트 영역을 추출하였다. Ezaki [3] 등은 수학적인 모폴로지 연산을 적용한 후의 이미지에 Otsu 이진화를 수행하였으며 수정된 Top-hats 연산처리를 사용하여 텍스트 후보 영역을 추출하였다. Ezaki [6] 등은 각 컬러 채널의 히스토그램을 구해 계산하여 임계치 값을 설정하여 세 개의 칼라 채널을 모두 이진화를 수행하여 8개의 이진 이미지를 가지고 텍스트 영역을 추출하였다. LIU [9] 등은 텍스트 후보 영역은 웨이블릿 변환을 통해서 특징들을 획득한 후 EM 알고리즘을 사용하여 텍스트 영역을 추출하였고 2차원 Harr 웨이블릿 변환을 사용하여 텍스트 영역에 대해 검증하였다. Jung [10] 등은 MLP 알고리즘과 MultipleCAMShift 알고리즘을 사용하여 텍스트 후보 영역을 추출하였다.

3. 제안한 시스템

본 논문에서 제안한 시스템은 그림 1과 같다.
입력된 이미지는 다양한 크기와 다양한 컬러를 가지고 있어서 색상 레이어(Layer)를 분리하는데 많은 계산 시간이 필요하기 때문에 색 줄임을 먼저 수행한다. 색 줄임 연산은 화소의 R, G, B 각 요소의 하위 7비트를 제거하여 상위 1비트만을 남기는 Bit Dropping 방법을 이용한다. 색 줄임은 이미지 처리에서 계산 시간량을 줄일 수 있고 잡영을 제거할 수 있다는 장점이 있다. 색 줄임 연산 후의 결과 이미지는 최대 8개의 색상 레이어 이미지로 ((red=0,green=0,blue=0), (0,0,128), (0,128,0), (0,128,128), (128,0,0), (128,0,128), (128,128,0), (128,128,128) 표현된다.



(그림 1) 텍스트 힌팅 시스템

8개 색상 이미지 내의 비텍스트 영역을 제거하기 위해 스택을 이용한 영역 라벨링 방법을 적용하여 연결요소들을 분리한 후에 분리된 각 영역들의 외곽지 상자(Bounding Box)의 크기를 구하게 된다. 구해진 값이 식 (1)을 만족하면 잡영으로 판단하여 제거하고, 식 (2)를 만족하면 비텍스트 영역으로 판단하여 제거한다.

$$CC\_H\_L_i < T_1 \ AND \ CC\_W\_L_i < T_2 \quad \text{.....식(1)}$$
$$CC\_H\_L_i > T_3 \ AND \ CC\_W\_L_i > T_4 \quad \text{.....식(2)}$$

(단, $CC\_H\_L_i$: i번째 CC의 세로크기
$CC\_W\_L_i$: i번째 CC의 가로크기
T_1, T_2, T_3, T_4 : 임계값)

8개 레이어 이미지에 식 (1)과 식 (2)을 각각 적용하고 다시 8개의 레이어 이미지를 OR 연산을 수행하여 하나의 이미지로 만든 후에 식 (3)을 적용하여 이진화를 수행한다.

$$\text{if } F(y,x)_{RGB} < 255 \text{ then } F(y,x) = 0 \quad \text{.... 식 (3)}$$

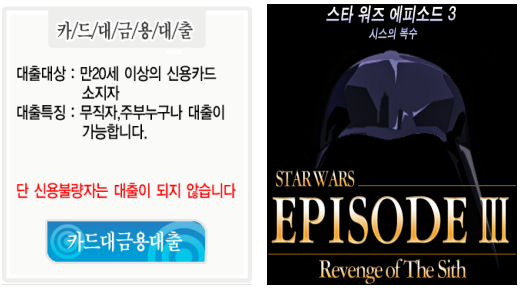
$$else \ F(y,x) = 255$$

텍스트 영역을 추출하기 위해 이진화를 수행하고 나면 가로 방향이나 세로획 방향으로 놓여 있는 " ㅣ " 나 " ㅡ "에 해당하는 중성 획들의 지워진 텍스트들을 추출하거나 "ㅁ"이나 "ㅇ"에 해당하는 초성들에 흑화소가 채워진 텍스트들을 추출하게 된다. 이렇게 지워진 획이나 흑화소로 채워진 영역을 복구하기 위해서 텍스트 힌팅 알고리즘을 적용한다.

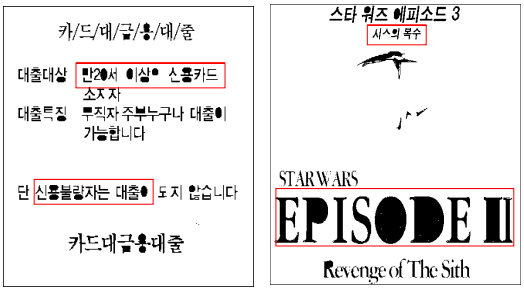
텍스트 힌팅 알고리즘은 이진화로 인해 손상된 텍스트 영역의 색상을 찾아내어 원래의 텍스트를 복원하는 방법으로 텍스트 추출 시 발생하는 문제점을 보완할 수 있다. 알고리즘은 다음과 같다.

이진화된 이미지의 텍스트 영역에 해당하는 픽셀(검정색 픽셀) 위치 좌표($II_{x,y}$)를 읽어 색 줄임 이미지에서 동일한 좌표($II_{x,y}$)에 해당하는 RGB 값을 추출하게 된다. 텍스트 영역에 해당하는 모든 픽셀의 RGB 값을 조사하고 나면 이미지 내의 텍스트를 이루는 컬러 값을 알 수 있게 된다. 이렇게 구해진 텍스트 색에 해당하는 색상 레이어 이미지와 획이 손상된 이진화 이미지를 OR연산을 수행하게 되면 지워진 획 부분을 복원할 수 있게 된다.

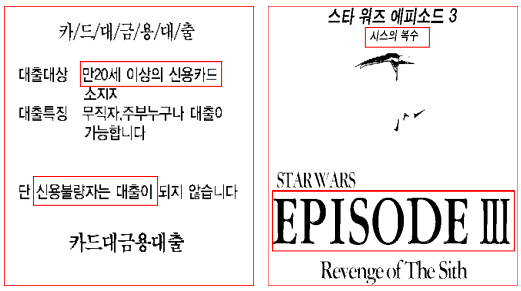
그림 2는 실험에 사용한 원본 컬러 이미지이며, 그림 3은 색상레이어를 이용하여 텍스트 영역을 추출한 후 이진화까지 수행하고 난 후의 이미지이다. 이진화를 수행하고 나면 그림 3내 빨간 박스 부분과 같이 문자 획 부분이 지워지거나 채움 현상이 발생하게 된다. 이는 이미지내 잡영 제거를 하다보면 문자를 이루고 있는 획들이 지워지게 된다. 그림 4는 그림 3 이미지에 텍스트 힌팅 알고리즘을 적용한 후 문자 획 안에 불필요한 흑화소 들이 제거되고 지워진 획들이 복구가 되어 있는 것들을 알 수 가 있다.



(그림 2) 원본 이미지



(그림 3) 텍스트 힌팅 적용 전 이미지



(그림 4) 텍스트 힌팅 적용 후 이미지

4. 실험 및 평가

실험 및 평가에 사용한 이미지는 이메일에서 추출한 스팸 이미지 156개를 대상으로 Visual C++ 6.0 환경에서 실험을 수행하였다. 입력된 이미지의 크기는 제한을 두지 않았으며, 취득한 이미지들은 단순한 이미지와 복잡한 배경을 갖는 다양한 이미지들을 실험에 사용하였다. 제안한 알고리즘의 성능 평가를 위해서 상용문자인식기로 많이 사용 중인 아르미 6.0을 사용하여 이진화 영상 내의 문자 인식 실험을 줄 단위로 수행하였다. 표 1은 텍스트 힌팅을 적용하기 전의 문자 인식 결과와 텍스트 힌팅을 적용하고 난 후의 문자 인식 결과를 보여 주고 있다. 실험 결과에서도 알 수 있듯이 텍스트 힌팅을 적용한 이미지에서 인식률이 약 10% 정도 좋아진 것을 알 수가 있다.

<표 1> 힌팅 적용 전후의 실험 결과 비교

		인식개수	평균
힌팅 적용 전	단어	42(239)	17.6%
	문자	1103(2268)	48.6%
힌팅 적용 후	단어	58(239)	24.2%
	문자	1300(2268)	57.3%

5. 결론

본 논문에서는 문자 영상에서 지워진 획이나 흑화소로 채워진 부분을 복원하는 알고리즘을 제안한다. 제안된 힌팅 시스템은 지워진 부분의 원래 색상 정

보를 찾아내어 이진화된 이미지와의 OR 연산을 통해 복원하는 방법으로, 실험을 통해 향상된 성능을 보였다.

본 논문에서 제안한 텍스트 힌팅 시스템은 획이 지워진 이미지를 가지고 OCR를 사용하여 인식을 한다거나 word spotting을 구현하고자 할 때 지워진 획 때문에 인식오류를 일으키거나 원하는 값을 찾아낼 수가 없게 된다. 이러한 부분에 텍스트 힌팅 시스템을 적용하여 지워진 획 부분을 되살린다면 좀 더 정확한 텍스트 검출이 이루어지리라 본다.

참고문헌

[1] 김길천, 최영우, 변혜란, “명도이미지에서의 장면 텍스트 추출,” 한국정보과학회, 컴퓨터비젼 및 패턴인식 연구회 추계워크샵 논문집, pp. 159-160, 2001. 11.

[2] 노명철, 최영우, 이성환, “색 및 명도 정보를 이용한 장면(Scene) 텍스트 추출,” 제 14회 영상처리 및 이해에 관한 워크샵 논문집, pp. 515-520, 2002, 1.

[3] N. Ezaki, M. Bulacu and L. Schomaker, "Text Detection from Natural Scene Images Towards a System for Visually Impaired Persons," *proc. 17th Int. Conf. on Pattern Recognition*, pp.683-686, 2004.

[4] 김지수, 김수형, 최영우 “명도 정보와 분할/합병 방법을 이용한 자연 영상에서의 텍스트 영역 추출,” 정보과학회논문집, 소프트웨어 및 응용 제 32권 제 6호, pp. 502-511, 2005.

[5] 최영우, 김길천, 송영자, 배경숙, 조연희, 노명철, 이성환, 변혜란, “계층적 특징 결합 및 검증을 이용한 자연이미지에서의 장면 텍스트 추출,” 정보과학회논문집, 소프트웨어 및 응용 제 31권 제4호, pp. 420-435, 2004.

[6] N. Ezaki, K. Kiyota, B. T. Minh, M. Bulacu and L. Schomaker, “Improved Text-Detection Methods for a Camera-based Text Reading System for Blind Persons," *proc. 17th Int, Conf Pattern Recognition*, pp. 977-980, 2005.

[7] L. Lin, C. L. Tan, "Text Extraction from Name Cards with complex Design," *proc, 8th Int, Conf Document Analysis and Recognition*, 2005.

[8] G. J. Ewerth, and R. Freisleben, "Text detection in images based on unsupervised classification of high-frequency wavelet coefficient," *proc. 17th Int, conf Pattern Recognition*, pp. 425-428, 2004.

[9] Y. LIU, S. Goto, T. Ikenaga, "A Robust Algorithm for Text Dection in Color Images," *proc. 8th Int, conf ICDAR* pp. 399-401 2005.

[10] K. Jung, K. I. Kim, T. Kurata, M. Kourogi, and J. H. Han, "Text Scanner with Text Detection Technology on Image Sequences," *Proc. 16th Int, conf Pattern Recognition*, pp. 473-476, 2002.

[11] 김지수, 김수형, "색상 레이어를 이용한 스팸메일 영상에서의 텍스트 영역 추출," 2005년 한국정보처리학회 추계학술발표대회 발표논문집, 제12권 제3호, pp. 124-128, 조선대학교, 2005년 10월.