웹상의 영상 내의 문자 인식과 음성 전환 시스템

안 희 임, 정 기 철
성균관대학교 전기, 전자 및 컴퓨터 공학부

Text to Speech System from Web Images

HeeIm An, Keechul Jung
Dept. of Electrical and Computer Engineering, Sungkyunkwan University
E-mail : heeim@graphics.skku.ac.kr

Abstract

The computer programs based upon graphic user interface(GUI) became commonplace with the advance of computer technology. Nevertheless, programs for the visually-handicapped have still remained at the level of TTS(text to speech) programs and this prevents many visually-handicapped from enjoying the pleasure and convenience of the information age. This paper is, paying attention to the importance of character recognition in images, about the configuration of the system that converts text in the image selected by a user to the speech by extracting the character part, and carrying out character recognition.

Ⅰ. 서론

전통적으로 디지털 비디오는 주로 수 작업에 의해 인덱싱되고 있는데, 이는 상당한 시간과 노동력을 요구하는 작업이다. 최근 영상 처리, 문자 인식 등의 기술 발전과 더불어 영상 내의 문자 자동 추출에 대한 연구가 진행되고 있다 [1-4, 6, 7]. 영상 내의 문자 추출에 관한 연구는 멀티미디어 시스템(multimedia system), 전자 도서관(digital library), 비디오 인덱싱(video indexing) 등의 다양한 응용 분야로 인하여 많은 연구가 진행되고 있으며, 문서 구조 분석, 우편 영상 내의 주소 영역 추출, 자동차 번호판 추출 등 다양한 관련 분야로 인하여 많은 연구가 필요하다.

또한 인터넷 기술의 발전과 함께 그래픽 사용자 인터페이스를 기반으로 하는 프로그램들이 보편화됨에 따라, GUI 형식의 웹브라우저의 보급으로 시각 장애인들은 정보생활영위에 많은 지장을 받게 되었다. 이에 대한 해결책으로 텍스트 기반의 웹 브라우저의 이용과, 윈도우용 화면 읽기 프로그램과 시각장애인 전용 웹 브라우저의 이용법이 있을 수 있다 [5].

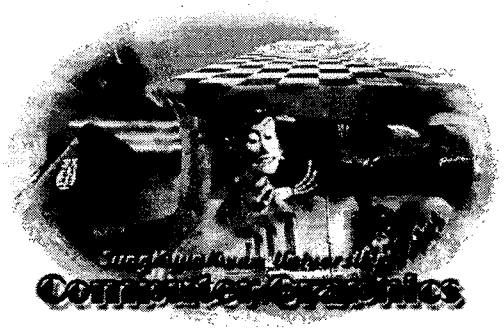


그림 1. 이미지 문자를 포함한 웹 페이지

그림 1은 이미지 문자를 포함한 웹 페이지의 한 예이다. 현재 많은 웹사이트들이 전달하고자 하는 문자들을 이미지로 보여주는 경우가 많음에도 불구하고 많은 시

각장애인용 프로그램들(시각장애인용 인터넷 프로그램인 EYES 2000과 음성합성 TTS제품인 리얼스피크 등)은 아직까지 문자 중심의 읽기 프로그램의 수준에 머무르고 있는 실정이다. 이제는 웹 브라우저 안에 포함된 영상 내의 문자인식이 한층 중요해 졌다고 하겠다.

기존의 문자 추출 방법은 크게 연결 성분 (connected component) 방법과 텍스춰 (texture) 방법으로 나눌 수 있다 [1]. 문자 영역의 텍스춰 성질을 이용한 방법들 중에는 gabor filter, wavelet, spatial variance 등을 이용한 방법들이 있다 [2]. 이러한 텍스춰 기반 방법들은 일반적으로 최적 필터 디자인의 어려움, 문자 추출 필터의 전역탐색 (exhaustive convolution)에 의한 속도 저하, 텍스춰 분석 후 문자 영역 획득을 위한 별도의 후처리 과정 수행 등의 단점이 있다 [2, 4]. 연결 성분 방법은 상향식 방법으로써 작은 영역에서 점차 큰 영역으로 합쳐가면서 문자 영역과 비-문자 영역으로 나누는 방법이다. 연결 성분 방법은 구현이 쉬운 반면, 문자 크기와 문자간의 거리등에 대한 사전 지식이 필요하다. 또한 하나의 문자를 배경과 구별되는 몇 개의 연결 성분들로 분할할 수 있는 알고리즘이 필요하다. 이 방법은 비디오 영상 등의 잡음이 많은 저해상도 영상에는 적합하지 않은 단점이 있으나, 웹 브라우저내의 이미지 문자들은 잡음이 적고, 문자의 크기 변화가 크지 않으며, 사용되는 색상 개수가 많지 않아 간단한 방법으로도 쉽게 분할되기 때문에 연결성분 방법이 유리하다.

본 논문은 영상 내의 문자인식의 중요성에 착안하여 웹 서핑시 읽고자하는 이미지를 가지고, 연결성분 방법을 이용하여 문자영역을 추출하고, 망 피쳐(mesh feature)값을 사용한 nearest neighbor classifier(NNC) 모델에 의해 문자인식을 수행한 후, 음성 전환 API를 사용하여 소리로 읽어주는 시스템의 구성에 관한 논문이다.

본 논문의 구성은 다음과 같다. 2절에서는 본 시스템이 사용한 방법으로 2.1절에서는 문자를 추출하는 방법을, 2.2절에서는 문자를 인식하는 방법을, 2.3절에서는 음성으로 전환하는 방법에 대해, 3절에서는 실험 및 결과에 대해, 4절에서는 결론 및 향후 과제에 대해 설명한다.

II. 본론

본 시스템은 이미지내의 이미지 문자의 음성으로의 전환을 통해 이전의 많은 시각 장애인용 프로그램의 한계를 극복하고자 한다.

그림 2는 전체적인 시스템의 흐름도 이다. 시스템은 크게 문자 추출 단계, 문자 인식 단계, 음성으로의 전환 단계를 거치는 세 단계로 구성된다.

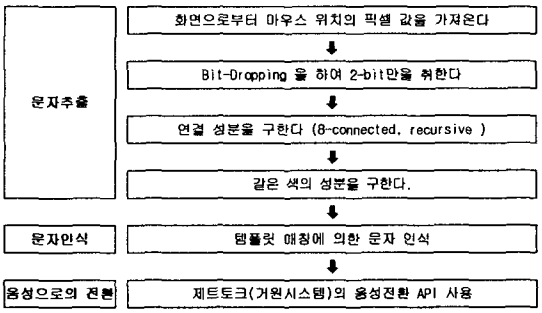


그림 2. 시스템 흐름도

첫째, 문자 추출 단계에서는 마우스 위치의 픽셀 값을 가져서 연결 성분을 구하고, 문자 영역의 특성을 이용하여 필터링을 수행하여 영상 내의 문자 부분을 추출한다.

둘째, 문자 인식 단계에서는 NNC 알고리즘을 이용한 템플릿 기반으로 문자를 인식한다.

셋째, 음성으로의 전환 단계에서는 상용화된 제트토크의 API를 사용하였다[8].

2.1 문자추출

본 시스템에서 문자 추출은 다음과 같은 과정을 거친다. 웹으로부터 이미지를 언제 가져올 것인지에 대한 판단은 마우스의 움직임으로부터 얻는다. 마우스 커서가 일정 시간동안 한곳에 있다면 그곳의 문자를 읽기를 원한다고 간주하고, 스크린으로부터 읽고자하는 이미지를 가져와 배열에 저장한다. Bit dropping을 하여 이미지는 보다 적은 수의 색을 가지고 표현된다. 이는 웹 상의 영상들이 가지는 특징으로 최상위 비트 몇 비트만을 취해도 이미지의 피쳐(feature)들이 많이 변하지 않음으로부터 기인한다. 다음 과정은 연결 성분을 구하고 필터링을 수행, 같은 색의 성분을 구하는 과정이다. 연결 성분을 구하고, bounding box의 가로 세로 픽셀의 크기가 일정 크기 이하 또는 이상인 것들은 버린다. 영상내의 글자라고 하더라도 하나의 문자열은 단일한 색을 가지는 성질을 이용하여 같은 색깔의 성분을 구해낸다. 마지막으로 이미지 내에서 읽고자 하는 문자를 추출하는 과정을 거친다. 읽고자 하는 문자는 마우스위치와 관계가 있으므로 이미지의 가운데 스캔라인을 지나는 성분을 텍스트 파일로 저장한다.

그림 3은 문자추출 단계의 한 예이다. 그림 3의 (a)는 실제 웹 상의 이미지에서 마우스가 위치한 부분이고, (b)는 (a)이미지를 bit dropping한 모습이다. (c)는 본 시스템을 통해 문자라고 판단되어지는 부분이 텍스트 파일로 저장이 된 모습이다. 그림에서 (b)의 이미지는 (a)의 이미지의 bit dropping 모습이기 때문에 색의 변화가

있고, 비슷한 색은 하나로 통일되는 결과를 낳는데, 흑백 모드로 출력이 되는 경우 같아 보일 수가 있다.

(a)

(b)

(c)

그림 3. 문자 추출 예제
(a)웹상에 마우스가 위치한 곳의 이미지
(b)Bit dropping한 이미지
(c)텍스트 검출 결과

2.2 문자인식

문자 인식은 NNC 에 의해 이루어진다. 이는 입력 문자 패턴과 가장 유사한 기준 패턴의 레이블을 제공해준다. 본 연구에서 인식 대상으로 하는 문자들은 크기나 모양들에서 심한 변형을 보인다. 스캐너로 입력받은 양질의 문서에 대해서 현재의 실용화된 오프라인 한글 문자 인식기의 평균 인식률을 90%정도라고 했을 때, 영상 내의 문자 인식은 결코 쉬운 문제가 아니다. 이러한 문제를 해결하기 위해서 다양한 연구가 진행중이며, 다양한 글자 형태와 크기에 적응할 수 있는 특징 값들의 추출, 컨텍스트에 기반한 문자 인식 등에 관한 연구가 있다. 본 연구에서는 교차 피쳐 (cross feature), 망 피쳐 (mesh feature)를 사용, 두 특징 값에 대해서 Euclidean 거리를 사용하였다.

NNC 모델은 테스트 패턴과 모든 기준 패턴과의 Euclidean 거리를 계산하여야 하므로 많은 시간이 소요된다. 그러나 본 실험에서는 실험 대상인 웹 이미지내의 글자 수가 극히 적어서, 많은 수의 기준 패턴을 사용할 수 있었다. 학습 샘플을 위해서 한글 잣기순 500자를 대상으로 다양한 크기와 폰트 모양의 데이터를 생성하여 사용하였다. 본 실험에서는 글자 크기 10, 13, 16, 20 화소 크기의 문자를 각각 명조체와 고딕체에 대해서 생성하였다. 각 샘플 문자들을 10 화소 크기의 문자로 정규화 한 후에, 특징 피쳐들을 추출하였다.

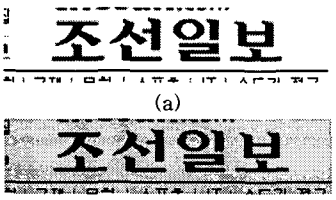
2.3 음성으로의 전환

본 시스템에서 문자를 음성으로 전환해주는 작업은 상용화된 제트토크의 음성합성(TTS) 기능 API를 사용하였다[8].

Ⅲ. 실험 및 고찰

이 절에서는 본 시스템을 사용하여 웹의 실제 이미지 속의 이미지 문자를 가지고 실험한 결과를 고찰한다. 본 시스템은 마우스 커서가 0.5초 동안 한곳에 있다면 그곳의 문자를 읽기를 원한다고 간주하고, 마우스의 커서를 중심으로 높이 55 픽셀, 너비 250 픽셀을 화면으로부터 마우스 위치의 픽셀 값을 가져와 배열에 저장한다. Bit dropping을 하여 최상위 비트 2비트를 취하면 이미지는 총 64가지의 색을 가지게 된다. Bit dropping한 이미지를 가지고 연결 성분을 구한다. 각 연결 성분의 Bounding Box의 가로세로 픽셀의 크기가 3×3 미만인 것들과 60 픽셀이상인 것들은 버린다. 같은 색깔의 성분을 구해낸다. 마지막으로 이미지의 가운데 스캔라인을 지나는 성분을 텍스트 파일로 저장한다. 저장된 텍스트 파일로 NNC 기반으로 문자 인식을 하고, 결과를 TTS API를 이용하여 음성으로 전환해준다.

그림 4, 5, 6은 본 시스템으로 실험 해 본 결과들이다. 각 그림의 (a)는 마우스의 위치로부터 이미지를 가져온 모습이고, (b)는 MSB 2 bit로 bit dropping 한 모습이다. (c)는 본 시스템으로 문자 인식한 결과이다.



(a)

(b)

조선일보

(c)

그림 4. 실험 결과

(a)

(b)

확술전쟁체던

(c))

그림 5. 실험 결과

(a)

(b)

창천만원의선용대출

(c)

그림 6. 실험 결과

현 실험 단계에서는 전체적으로 30% 내외의 문자 인식 결과를 보인다. 이는 기술한 바와 같이 문자의 모양, 크기의 다양성, 정규화와 이진화에 따른 잡음 발생 등으로 나타나는 결과이다. 현재 이러한 문제를 보완할 수 있는 인식 기법들에 대한 연구가 활발히 진행중이며, 이러한 연구 결과를 사용한다면 보다 실용적인 시스템을 구성할 수 있으리라 생각된다.

이미지로부터 문자를 추출하는 과정에서 100개의 이미지를 가지고 실험한 결과는 표 1과 같다.

검출 률	샘플의 개수
0%~10%	48
11%~40%	1
41%~60%	5
61%~90%	7
91%~100%	39
평균 검출 률 = 43 %	

표 1. 텍스트 검출 성능

실험결과 43%의 문자 검출 률을 보였다. 이는 영상 내에서의 문자를 추출하는 과정에서 일정 크기 이하와 이상의 것들은 제거함을 통해 추출할 수 있는 문자크기의 한계와 비-문자 영역이면서도 문자 추출 단계에서 언급되었던 필터링 단계에서 제거되지 않은 비-문자들이 추출되는 이유 때문이다. 크기와 관계없이 문자는 보존하면서 노이즈만을 제거하는 필터링을 현재 연구중이다.

Ⅳ. 결론

본 시스템은 스크린으로부터 마우스의 절대적 좌표 값으로부터 픽셀정보들을 가져오기 때문에 그 이용 가능성이 무한하다고 기대된다. 하지만 영상 내에서의 문자를 추출하는 과정에서의 필터링의 한계로 추출된 문자열들은 크기나 모양 면에서 변형이 심하고, 문자 추출과 문자 이진화 단계에서 발생하는 잡음으로 인하여, 기존의 문자인식 방법은 적당하지 않다. 아직 한글에 대한 템플릿만을 사용하여 숫자나 영문 등에 대한 인식을 하지 못하는 한계가 있다. 전체적으로는 30% 정도의 문자 인식률을 보였다. 아직 시험 단계이므로 문자추출 단계에서 보다 정확한 노이즈 필터링을 수행하고, 문자 인식 단계에 보다 다양한 글자들을 인식하도록 하면 인식률은 높일 수 있을 것으로 기대된다.

참고 논문

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

[2] K. Y. Jeong, K. Jung, E. Y. Kim, and H. J. Kim, "Neural Network-based Text Location for News Video Indexing," International Conference on Image Processing99, 1999.

[3] Rainer Lienhart and Frank Stuberm, "Automatic Text Recognition In Digital Videos," SPIE-The International Society for Optical Engineering, pp. 180-188, 1996.

[4] Huiping Li, David Doerman, and Omid Kia, "Automatic Text Detection and Tracking in Digital Video," IEEE Transactions on Image Processing, Vol. 9, No. 1, pp.147-156, January 2000.

[5] 이승수, 김석일, "웹 브라우져에서의 시각장애인 사용자 인터페이스 설계," "http://john.chungbuk.ac.kr".

[6] Jiangying Zhou and Daniel Lopresti, "Extracting Text from WWW Images," Proceedings of the Fourth International Conference on Document Analysis and Recognition, Vol. 1, pp.248 -252, 1997.

[7] K. Jung and H.J. Kim, "Texture-based Text Locations for Video Indexing," The 4th Character Recognition Workshop, pp. 49-58, 2000.

[8] 거원 제트토크 API, "http://www.cowon.com/SynthSW.htm#Talk".