웹 영상에 포함된 문자 영역의 추출

김 상 현, 심 재 창 , 김 중 수
안동대학교 컴퓨터공학과
전화 : (0571) 850-5645 / 팩스 : (0571) 850-5480

Text Extraction In WWW Images

Sang Hyun Kim, Jae Chang Shim, Joong Soo Kim
Department of Computer Engineering Andong National University
E-mail: aerobee@comeng.andong.ac.kr jcshim,kimjs@andong.ac.kr

Abstract

In this paper, we propose a method for text extraction in the Web images. Our approach is based on contrast detecting and pixel component ratio analysis in mouse position. Extracted data with OCR can be used for real time dictionary call or language translation application in Web browser.

Ⅰ. 서론

컴퓨터와 네트워 관련 기술이 급속히 발달함에 따라 각종 멀티미디어 매체를 비롯하여 문자와 그래픽 영상은 인터넷에서 없어서는 안될 정보 전달의 주요 매개체로 자리잡고 있다. 예전 환경에 비하여 정보 전달의 역할을 맡고 있던 문서들이 모두 전자화되어 간다고 해도 과언이 아닌 것이다[1].

그 중에서 텍스트 정보는 번역이나 편집, 해석이 용이하지만 영상이나 그래픽은 그림 형태로 구성되어 있기 때문에 당장 직접 활용에는 많은 제약점들이 따르게 된다[2,3]. 기존의 많은 연구들에서 알 수 있듯이, 스캐닝 된 영상이나 입력된 비디오 영상으로부터 배경과 문자 영역을 분리하여 처리하는 방식을 이용하고 있다[3,4].

그러나 정보 전달의 방법에 있어서 사용자와의 상호작용으로 이루어지는 인터넷상에서는 전체 영상을 기준으로 모든 문자들을 추출해내는 기존의 방법들이 적합하지 않다. 즉, 대부분 기존의 방법들은 이미 획득된 전체 영상을 대상으로 하여 문서 처리 기법을 사용하고 있기 때문이다[1-9].

홈페이지 상에서는 필요한 블록 영상만 전송, 압축, 해독 등 다양한 처리가 요구된다. 만약 선전을 위한 웹 페이지에서 그래픽 영상에 포함된 문자는 사람이 쉽게 이해할 수 있다. 그러나 컴퓨터로 이를 처리하기 위해서는 먼저 문자 영역을 추출하는 것이 필요하다. 그렇게 되면 그래픽 영상에 포함된 선전문구나 설명 등도 문서 자동 인식 시스템으로 처리가 가능해질 것이다.

본 논문에서는 웹 영상에 포함된 문자 영역의 추출 기법을 제안한다. 제안한 방법에서는 마우스가 위치한 지점을 기준으로 명도 대비 탐색 알고리즘을 적용한다. 이는 문자 영역을 추출하는데 있어서 웹 영상의 특성상 의미 전달을 위하여 문자열을 배경과 구분되게 강조한다는 점에 착안한 것이다. 처리 과정은 먼저 웹 영상에서 사용자가 지정한 기준 위치로부터 많이 벗어난 픽셀을 대표 색상으로 지정하는 오류를 방지하기 위하여 9x5 픽셀 크기내의 모든 명도 대비 값들에 대해 수평 방향으로 가중치를 곱한 후, 가장 큰 명도 대비 값을 가지는 픽셀을 대표 색상으로 지정한다. 구성된 대표 색상의 R, G, B 비율을 조사하고 각각의 허용 오차 결정에 반영하여 문자 영역을 추출한다.

Ⅱ. 기존의 추출 방법

영상에서 문자 영역과 비문자 영역을 분리하는 기존의 연구로는 [1-4]의 외국 논문과 [5-9] 등의 국내 논문들이 있다. 박영석 [8]은 화상의 흑백 화소에 대한 통계적 특징량을 정의하여 5부류로 식별하는 고정도 영역 식별법을 이용하였으며, 이인동 [7]등은 문자와 비문자 영역의 경계 위치를 알아내어 동시에 분리 추출을 하였다. 최봉희 [9]등은 문서영상의 누적분포를 이용하여 문자열 영역을 추출하고, 문자열 영역의 누적분포를 이용하여 개별 문자 영역을 추출하였다. 웹 영상을 기반으로 한 연구로 Jiangying Zhou [2]등은

페이지에 삽입된 한 영상을 기준으로 EMST (Euclidean minimum spanning tree)[10] 알고리즘을 이용하여 컬러 공간(color space)을 양자화한 후, 연결 구성(connected component)을 분석한 color clustering 알고리즘을 사용하였다. 그러나 이 방법을 적용하였을 경우 wrapped text, outline font, stylized font와 같은 특수한 상황의 문자열들은 추출이 불가능함을 알 수 있다. 입력된 영상에서 모든 문자열들을 추출하는 기존의 방법들은 사전과 같은 응용 프로그램에 바로 적용할 경우 비효율적인 문제점이 있어 개량이 요구된다. 예를 들어 일본어 사이트나 기타 다른 언어로 구성된 웹 페이지에서 특정 단어의 의미를 알기 위해 추출한 후 사전에 적용한다고 가정하였을 경우, 전체 영상에 대하여 모든 문자 영역들을 추출할 필요는 없기 때문이다.

본 논문에서는 이러한 기존의 문제점을 보완하기 위하여 사용자의 마우스 위치 정보와 가중치를 곱한 명도 대비 탐색 알고리즘을 적용하여 사용자가 원하는 부분의 단어만 추출할 수 있는 방법을 제시하고자 한다.

Ⅲ. 자동 문자 영역 추출 알고리즘

웹 영상에서 배경 그림과 함께 섞여 있는 문자들은 대부분 공통적인 특징으로 사람들이 보기 쉽게 주변과 확연히 구분되어지는 색상들로 이루어져 있다.

본 논문에서는 사용자 마우스의 위치 정보를 적극 활용하여 원하는 곳의 문자열을 명도대비 탐색을 이용하여 추출한다. 마우스가 위치한 부근의 문자를 추출하기 위하여 다음과 같이 가정하였다.

1) 마우스의 위치는 문자가 존재하는 수평선상의 범위 내에 있으며, 문자 구간의 위치에 있다.
2) 추출하려는 문자들은 눈에 잘 띄는 명도 대비를 가진다.
3) 추출하기 위한 글자는 너무 크거나 작지 않다.
4) 문자는 단색으로 구성되어 있거나 서로 완전히 다른 색상들로 구성되어 있지 않다. 따라서 그림 1의 "MoSaJi"와 같이 서로 다른 색상일 경우는 지정하는 위치에 따라서 한 글자만 추출된다.

명도 대비는 임의의 픽셀 P_1(x, y)과 P_2(x, y)가 존재할 때, P_1의 R, G, B 중에서 가장 큰 값과 P_2의 R, G, B 중에서 가장 큰 값을 서로 빼서 나타낸 값으로 정의한다.

$$C_{max} = max\ in\ P_1 (R, G, B) - max\ in\ P_2 (R, G, B)$$

그림 1. 가정 4)의 예

제안된 문자 추출 방법의 흐름도는 다음과 같다. 먼저 사용자가 마우스로 추출할 위치를 지정한 다음 기준 위치를 중심으로 그림 2와 같이 9x5 픽셀 크기내의 모든 명도 대비 값들에 대하여 그림 3과 같이 수평 방향으로 가중치를 곱한다.

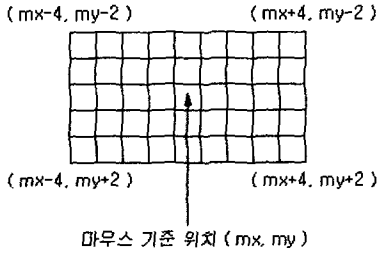


그림 2. 명도 대비 탐색을 할 픽셀 영역

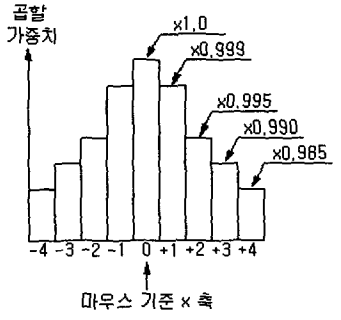


그림 3. 가중치 참조 테이블

이는 기준 위치에서 많이 벗어난 픽셀의 R, G, B를 대표 색상으로 지정하는 오류를 방지하게 된다. 구체

적으로 적용되는 알고리즘은 그림 4와 같다.

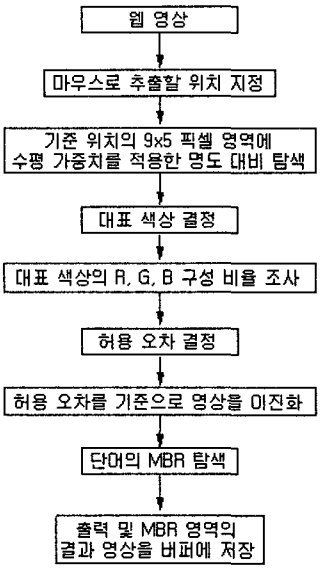


그림 4. 제안된 알고리즘의 흐름도

이때 가장 큰 명도 대비 값을 가지는 픽셀의 R, G, B를 대표 색상으로 지정한 후, 구성된 R, G, B의 비율을 조사하여 각각의 허용 오차를 결정한다. 기본적으로는 대표 색상의 R, G, B에서 ±32 만큼의 범위를 가지나 각 요소에 구성비를 기준으로 추가적인 32를 더한다. 이는 허용 오차의 범위를 대표 색상과 비슷한 색상들이 많이 포함될 수 있도록 더욱 강조를 하게 된다. 이렇게 허용 오차의 범위가 결정되면 전체 영상에서 참조를 하여 이진화를 수행한 후, 단어의 최소외접 사각형(MBR)을 탐색하여 최종 결과를 출력하고 결과 영상을 버퍼에 저장한다.

Ⅳ. 실험 및 고찰

본 연구는 Intel Coppermine 500Mhz, 128MB RAM, Windows 98 SE 환경의 PC에서 Visual C/C++ 언어를 사용하였다. 제안된 알고리즘을 적용하여 문자를 추출한 결과는 표1과 같다.

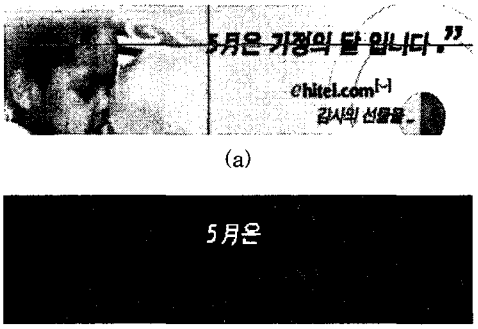
표 1. 실험 결과

실험영상개수	추출대상문자블록수	정추출개수	오추출개수
50	362	279	83

웹 영상중 문자가 포함된 영상 50개를 추출하였고, 총 추출대상 문자 블록의 수는 362개이다. 실험결과 77%가 정확히 추출되었다.

컬러 영상의 특성상, 구성하는 화소가 흑백 영상에 비하여 3배나 많기 때문에 명도 대비 탐색을 통한 문자 영역의 추출이 잘 이루어질 수 있었으며, 특히 웹 영상에서 광고와 같은 특정 문구들은 강조를 하기 위하여 주변 배경과 확연하게 다른 색상으로 구성되어져 있으므로 좋은 결과를 나타내었다. [그림 5,6]

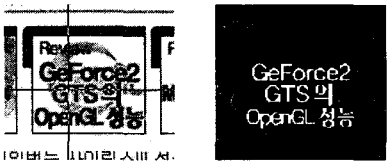
또한 본 연구의 알고리즘을 적용하였을 경우, 그림 7처럼 약간 기울어져 있거나, 그림 8과 같이 둥글게 회전된 문자들에 대해서도 추출이 가능하다.



(a)

(b)

그림 5. 단어 추출의 실험 결과 (a) 원 영상 (b) 결과 영상



(a) (b)

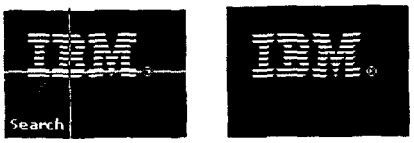
그림 6. 문단을 기준으로 한 실험 결과 (a) 원 영상 (b) 결과 영상

(a) (b)

그림 7. 기울어진 문자 영역의 추출 (a) 원 영상 (b) 결과 영상

(a) (b)

그림 8. 회전된 문자 영역의 추출 (a) 원 영상 (b) 결과 영상



(a) (b)

그림 9. Stylized font (a) 원 영상 (b) 결과 영상

(a) (b)

그림 10. Outline font (a) 원 영상 (b) 결과 영상

여기서 영상에 그어진 십자선은 마우스 위치를 나타낸다.

실험결과 그림 8-10은 Jiangying Zhou [2]등의 연구에서 처리가 불가능했던 문자열에 대해서도 잘 추출됨을 알 수 있다.

대표 색상에서 효과적인 허용 오차의 범위를 결정하는 문제 및 불규칙적인 띄어쓰기 간격의 인식, 배경과 글자가 뒤집혀서 반대로 이진화된 경우, 큰 문자에 대해서는 한 글자나 자소로 추출되는 경우, 다양한 색상으로 이루어진 문자열의 경우들과 같은 문제점들은 추후 연구를 통하여 지속적으로 개선해야 될 부분으로 판단된다.

V. 결론

본 논문에서는 웹 영상에 포함된 문자 영역을 추출하기 위해 마우스 기준 위치에서 국부영역의 픽셀 영역들에 대하여 가중치를 적용하고 명도 대비 탐색을 수행하여 사용자가 원하는 문자만을 추출할 수 있도록 알고리즘을 제안하였다. 이는 웹 영상내의 원하는 부분에 대해서만 추출하므로 사전이나 번역 등의 응용 프로그램들에 효과적으로 활용되어질 수 있게 된다. 또한 사용자와의 상호작용으로 필요한 정보만을 선별하여 습득할 수 있는 웹 페이지의 특성에 부합되어 본 논문에서 제시한 방법이 영상 내에서 모든 문자열들을 추출하는 기존의 방법들보다 효율적임을 알 수 있었다.

참고문헌

[1] Yu Zhong, Kalle Karu and Anil K. Jain, "Locating Text In Complex Color Images", Pattern Recognition, Vol. 28, No. 10, pp. 1523-1535, 1995

[2] Jiangying Zhou, Daniel Lopresti, "Extracting Text from WWW Images", Proceedings of the 4th International Conference Document Analysis and Recognition (ICDAR '97), pp. 248-252, 1997

[3] Anil K. Jain, Fellow, Bin Yu, "Document Representation and Its Application to Page Decomposition", IEEE Transactions On Pattern Analysis and Machine Intelligence, Vol. 20, No. 3, pp. 294-305, 1998

[4] Rainer Lienhart, "Automatic Text Recognition for Video Indexing", ACM Multimedia 96, Boston MA USA, pp. 11-20, 1996

[5] 함영국, 김인권, 정홍규, 박래홍, 이창범, 김상중, 윤병남, "텍스트와 그래픽으로 구성된 혼합문서 인식에 관한 연구", 전자공학회논문지-B, Journal of the KITE 1994, Vol. 31-B, No. 7, 1994, pp. 76-90

[6] 성연진, 어진우, "텍스트- 배경무늬 혼합문서로부터 수리형태학을 이용한 문자열 추출", 전자공학회논문지-S 1997, Vol. 34-S, No. 10, pp. 104-111

[7] 이인동, 권오석, 김태균, "문서 영상에서 문자와 비문자의 분리추출방법", 정보과학회논문지, Vol. 17, No. 3, pp. 247-258, 1990

[8] 박영석, "일반적인 문서화상의 영역식별법", 정보과학회논문지, Vol. 21, No. 5, pp. 757-767, 1994

[9] 최봉희, 이인동, 김태균, "문자영역 추출과정에서의 오분리의 교정", 정보과학회논문지, Vol. 21, No. 1, pp. 86-93, 1994

[10] R. Graham and P. Hell. On the history of minimum spanning tree problem. *Annals of History of Computing, 7, 1985.*