論文2003-40SP-3-2

1차원 메디안 필터 기반 문서영상 영역해석

(The Region Analysis of Document Images Based on One Dimensional Median Filter)

朴 承 浩 *, 張 大 根 *, 黃 燦 植 **

(Seung Ho Park, Dae Geun Jang, and Chan Sik Hwang)

요 약

인쇄문서를 전자문서로 자동변환하기 위해서는 문서영상 영역해석과 문자인식 기술이 필요하다. 이들 중 영역해석은 문서영상을 세부 영역으로 분할하고, 분할한 영역을 문자, 그림, 표 등의 형태로 분류한다. 그러나 문자와 그림의 일부는 크기, 밀도, 화소분포의 복잡도가 비슷하여 정확한 분류가 어렵다. 따라서 영역해석에서의 오 분류는 자동변환을 어렵게 만드는 주된 원인이 된다. 본 논문에서는 문서영상을 문자와 그림영역으로 분할하는 영역해석 방법을 제안한다. 문자와 그림의 분류는 1차원 메디안 필터링을 기반으로 한 방법을 이용하여 언급한 문제점을 해결한다. 또한 메디안 필터링에 의해 발생하는 볼드체 문자와 그래프나 표와 같은 그림영역의 오 분류 문제를 표피 제거 필터와 문자의 최대크기를 이용하여 해결한다. 따라서 상용제품을 포함한 기존의 영역해석 방법보다 그 성능이 우수하다.

Abstract

To convert printed images into electronic ones automatically, it requires region analysis of document images and character recognition. In these, region analysis segments document image into detailed regions and classifies these regions into the types of text, picture, table and so on. But it is difficult to classify the text and the picture exactly, because the size, density and complexity of pixel distribution of some of these are similar. Thus, misclassification in region analysis is the main reason that makes automatic conversion difficult. In this paper, we propose region analysis method that segments document image into text and picture regions. The proposed method solves the referred problems using one dimensional median filter based method in text and picture classification. And the misclassification problems of boldface texts and picture regions like graphs or tables, caused by using median filtering, are solved by using of skin peeling filter and maximal text length. The performance, therefore, is better than previous methods containing commercial softwares.

Keywords : document image, region analysis, region segmentation, region classification

* 正會員, 경북대학교 전자·전기·컴퓨터학부
(Department of Electronic Engineering Kyungpook National University)
** 正會員, 한국전자통신연구원
(Electronics and Telecommunications Research Institute)
接受日字:2000年8月11日, 수정완료일:2003年3月31日

1. 서 론

정보화와 더불어 전자문서의 사용이 증가함에 따라 인쇄문서의 사용은 감소할 것이라는 예상과는 달리 프린터와 워드프로세서를 이용함에 따라 예전보다 인쇄문서의 양은 더욱 늘어나고 있는 추세다. 따라서 인쇄

문서를 직접 손으로 입력하지 않고 편집 가능한 전자문서로 자동변환의 필요성이 갈수록 증가하고 있다. 인쇄문서를 전자문서로 자동변환 하려면 문서영상 영역해석, 문자인식 등의 기술이 필요하며 이 중 영역해석은 문서영상을 세부 영역으로 분할하고 분할한 영역을 문자, 그림, 표 등의 형태로 분류한다. 그러나 문자와 그림의 일부는 크기, 밀도, 화소분포의 복잡도가 비슷하여 정확한 분류가 어렵다. 그리고 인쇄문서를 전자문서로 자동변환 하는 나머지 과정들은 영역해석의 결과를 이용하므로 영역해석에서의 오 분류는 인쇄문서의 전자문서로의 자동변환을 어렵게 만드는 주된 원인이 된다.

영역해석에서 분할한 영역을 문자와 그림으로 분류할 수 있는 방법으로 문자와 그림의 크기와 밀도 차이 그리고 문자간의 인접성을 이용하는 방법[1,2]과 문자와 그림을 구성하는 화소 분포의 복잡성(complexity)을 계산하는 방법[5,6] 그리고 문자배열의 반복성을 이용하는 방법[7]이 있다. 그러나 이러한 방법들은 크기, 밀도, 화소분포의 복잡도가 비슷하거나 문자배열의 반복성이 없는 경우 문자와 그림을 분류하기 어려운 문제점이 있다.

따라서 본 논문에서는 메디안 필터를 이용하여 문자를 구성하는 흑화소를 제거함으로써 문서영상을 문자와 그림으로 분류하고 메디안 필터링에 의해 발생하는 볼드체 문자 제거와 저 밀도 그림분류 문제를 표피 제거 필터와 문자의 최대크기를 이용하여 해결함으로써 기존의 방법들보다 영역해석 성능이 우수한 방법을 제안한다.

제안한 방법은 영어의 대, 소문자 52자와 한글 완성형 2350자의 normal fonts가 그 외각을 둘러싸는 직사각형에 대해 흑화소의 비율이 평균이 30% 내외라는 점을 이용하여 메디안 필터링을 수행함으로써 문자를 구성하는 흑화소를 제거하여 문자와 고 밀도 그림을 분류한다. 또한 문자를 구성하는 획 가운데 가로와 세로 방향 획의 비율이 영어의 경우 50%, 완성형 한글의 경우 평균 84%라는 점을 이용하여 1차원 윈도의 메디안 필터를 문자 획의 수직 방향인 가로와 세로방향으로 수행하여 필터 윈도 내부의 백화소 비율을 극대화함으로써 문자를 구성하는 흑화소 제거율을 높인다. 그리고 흑화소 비율이 높아 메디안 필터에 의해 제거되지 않은 볼드체 문자는 표피 제거 필터(skin peeling filter)를 이용하여 문자 획을 엷게 한 후 메디안 필터를 적용하고 메디안 필터에 의해 분류되지 않은 그래프, 표와 같은 저 밀도 그림은 문자의 최대크기를 이용하여 문자와 구분함으로써 문자와 그림 분류의 정확성을 높인다.

Ⅱ. 기존 문서영상 영역해석 방법 분석

문서영상 영역해석이란 문서영상을 세부 영역으로 분할하고, 분할한 영역을 문자, 그림, 표 등의 형태로 분류하는 것을 말한다. 따라서 문서영상을 작은 영역으로 분할하는 방법과 분할한 영역을 분류하는 방법의 종류와 특징에 관한 고찰이 필요하며 그 내용은 다음과 같다.

1. 영역분할(region segmentation)

문서영상을 세부 영역으로 분할하는 방법에는 크게 기본이 되는 화소단위에서 시작하여 유사성을 갖는 부분을 점차적으로 크고 의미를 부여할 수 있는 단위로 단계적으로 병합하는 상향식(bottom up)[1,2]과 문서의 전체적인 영역에서 시작하여 문서를 점점 작은 영역으로 분할하는 하향식(top down)[3,4]이 있다. 상향식은 기울어진 문서를 포함하여 복잡한 형태의 문서를 분할할 수 있다는 장점이 있는 반면 많은 계산량과 버퍼를 필요로 하는 단점이 있다. 하향식은 알고리즘이 간단하고 빠르며 영역이 사각형 블록으로 구성된다는 장점이 있으나 복잡한 형태의 문서나 기울어진 문서에는 적용하기 어려운 단점이 있다. 상향식의 예로는 연결요소(connected components) 이용법[1,2]이 있으며 하향식의 예로는 투영윤곽(projection profile) 분석법[3]과 런 길이 평활화(run length smoothing) 방법[4]이 있다.

1.1. 연결요소 이용법

연결요소는 임의의 흑화소 또는 백화소에 대하여 4방향 또는 8방향의 인접화소가 같은 경우 이 화소를 모두 연결하여 얻은 화소의 집합으로 분할하고자 하는 영역에 대한 연결요소의 특성을 이용하여 필요한 영역을 분할한다. 즉 서로 특징이 유사하고 거리가 가까운 연결요소들은 병합하여 영역을 분할한다.

1.2. 투영윤곽 분석법

문서가 주로 사각형의 블록으로 구성된다는 점에 근거하여 투영 축 상의 흑화소의 수를 계산함으로써 얻어지는 윤곽을 이용하여 순환적으로 문서를 사각형의 영역으로 분할하는 방법이다. 즉 블록과 블록 사이의

여백을 기준으로 수평과 수직방향 분할을 번갈아 반복함으로써 전체 영상을 분할한다.

1.3. 런 길이 평활화 방법

백화소가 0 흑화소가 1로 표현되는 이진수열에서 연속된 0의 개수가 임계값보다 적을 경우 0을 1로 바꿈으로써 즉 흑화소들 사이의 백화소 공백을 흑화소로 치환하여 가까이 있는 흑화소들을 병합함으로써 영역을 분할한다.

2. 영역분류(region classification)

영역분류는 분할한 영역을 문자, 그림, 표와 같은 형태로 구분하는 과정으로 문자와 그림의 크기와 밀도 차이 그리고 문자간의 인접성을 이용하는 방법[1]과 문자와 그림을 구성하는 화소 분포의 복잡도를 계산하는 방법[5] 그리고 문자배열의 반복성을 이용하는 방법[7]이 있다.

2.1. 문자와 그림의 크기와 밀도 차이 그리고 문자들 간의 인접성 이용법

영역분류의 가장 일반적인 방법으로 문자의 크기와 밀도가 그림과 다르며 또한 문자는 서로 인접해 있는 경우가 많다는 점을 이용하는 방법이다. 이 방법은 알고리즘이 간단하지만 크기와 밀도가 비슷한 일부 문자와 그림이 오 분류되는 경우가 발생한다.

2.2. 화소 분포의 복잡도 계산법

백화소가 0 흑화소가 1로 표현되는 이진수열에서 화소가 분포하는 정도를 수치화 함으로써 문자와 그림을 구분하는 방법으로 상호 상관도(Cross Correlation)를 이용하는 방법 []이 있다. 이 방법은 식 (1)을 이용하여 상호 상관도(C_c)를 구하고 이 값이 0.88 이하이면 문자 나머지는 그림으로 분류한다.

$$C_c = 1 - \frac{2}{MN} \sum_{y=0}^{N-1} \sum_{x=0}^{M-1} [I(x,y) \oplus I(x+1,y)] \quad (1)$$

이 방법 또한 화소 분포의 복잡도가 비슷한 문자와 그림이 오 분류되는 경우가 발생한다. 수식에서 M, N은 각각 해당 영역의 가로, 세로방향 화소 수이고 $I(x,y)$는 좌표 (x,y)에서의 화소값(0 or 1)이며 $\oplus$는 exclusive OR 연산을 나타낸다.

2.3. 문자배열 반복성 이용법

문서에서 문자배열은 줄간을 경계로 서로 이웃하게 배열되어 있어 문자배열 방향으로 화소 히스토그램을 구하면 문자배열 사이에 있는 줄간을 확인할 수 있고 이 줄간의 반복을 검사하여 문자배열의 존재를 확인함으로써 그림과 구분이 가능하다. 그러나 문자배열의 반복성이 없는 1줄로 된 문자나 문자배열의 경우 반복성을 찾기 어려운 문제점이 있다. 또한 문자배열 방향으로 화소 히스토그램을 구하는 것은 투영윤곽 분석법 즉 하향식 영역분할과 같은 방법이므로 기울어지거나 복잡한 구조의 문서에 적용하기 어렵다.

Ⅲ. 메디안 필터 기반 문서영상 영역해석

제안한 영역해석 방법은 연결요소 생성을 고속화한 X. Li의 방법[1]을 이용하여 문서영상을 세부 영역으로 분할하고 분할한 영역은 1차원 메디안 필터를 이용하여 문자와 고 밀도 그림으로 분류한다. 그리고 메디안 필터에 의해 제거가 어려운 볼드체 문자는 표피 제거 필터를 이용하여 문자 획을 엷게 한 후 메디안 필터를 적용하고 메디안 필터에 의해 분류되지 않은 그래프, 표와 같은 저 밀도 그림은 문자의 최대크기를 이용하여 문자와 구분함으로써 상용제품을 포함한 기존의 방법들보다 문자와 그림 분류의 정확성을 높인다.

1. X. Li의 방법[1]을 이용한 영역분할

방법[1]은 연결요소를 생성하고 인접한 같은 형태의 연결요소들을 결합함으로써 영역을 생성하는 상향식 방법이다. 이 방법은 가로방향 라인단위로만 연결요소를 생성하고 인접한 라인간 연결요소들을 결합하여 완성된 형태의 연결요소를 만드는 방식이므로 4 또는 8 방향으로 탐색하여 연결요소를 생성하는 방식보다 계산량이 적어 수행시간이 짧다.

2. 메디안 필터를 이용한 문자와 그림분류

2.1 이진 문서영상에서의 메디안 필터링

메디안 필터는 해당 화소 값을 주변의 화소 값을 포함한 중간 값으로 바꾸는 필터이다. 따라서 메디안 필터를 0과 1의 값만 존재하는 이진(binary) 영상에 적용할 경우 중간 값은 필터 윈도 내의 0과 1 중 다수를 차지하는 값이 된다. 그리고 영어 대, 소문자 52자와 KS 완성형 한글 2350자를 대상으로 사용빈도가 높은 영어 폰트 3종류와 한글 폰트 3종류를 실험한 결과 각 문자를 둘러싸는 직사각형에 대해 문자를 구성하는 흑화소 비는 영어가 평균 27%, 한글이 평균 31%로 그 결과를 <표 1>에 기록하였다.

표 1. 문자를 둘러싸는 사각형에 대한 내부 흑화소 비율

Table 1. The ratio of inner black pixels v. rectangle surrounding the text.

항목 / 폰트종류(크기 10points)		영역크기(R) (pixels)	흑화소 수(B) (pixels)	$\frac{B}{R}$ (%)
영어	Times New Roman	3871	1069	28
	신명조	6206	1448	23
	고딕	3826	1100	29
한글	바탕체	302352	105937	35
	신명조	471210	131579	28
	고딕	421620	128629	31

<표 1>의 결과와 같이 이진 문서영상을 메디안 필터링 하면 문자부분은 필터 윈도 내부에 백화소 수가 더 많게 되어 다수 값을 택하는 필터의 특성에 의해 구성 흑화소는 백화소로 치환되어 제거된다. 그러나 사진과 같은 밀도가 높은 그림의 경우 흑화소의 비율이 높아 제거되지 않고 남게 되어 문자와 그림의 구분이 가능하다.

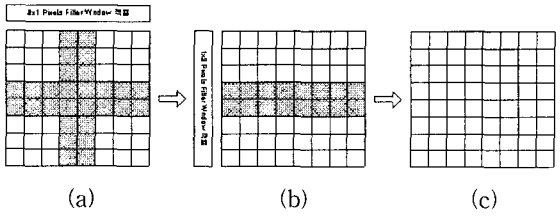
2.2 필터 윈도 모양 및 필터링 방향 결정

2.1에서 이진 문서영상에 메디안 필터를 적용하여 문자와 그림의 분류가 가능함을 설명하였다. 그러나 필터 윈도 내부의 백화소 비율을 극대화하여 문자를 구성하는 흑화소 제거 성능을 높이려면 필터 윈도의 모양을 1차원, 2차원 중 어떤 모양으로 적용하는 것이 좋은가를 결정해야한다. 따라서 문자를 둘러싸는 영역의 크기 R와 영역 내부의 흑화소 가운데 메디안 필터링 결과 제거된 수(N_B)와 제거되지 않고 남은 수($N_{\overline{B}}$) 를 이용하여 필터 윈도 내부의 백화소 비(R_{W_FW})가 최대가 되도록 필터 윈도 모양을 결정한다. 그리고 메디안 필터링에서 해당 화소는 자신을 둘러싸는 필터 윈도의 가운데에 위치하도록 하며, 필터 윈도 내부의 백화소 수가 흑화소 수보다 많은 경우만 해당 위치의 흑화소를 백화소로 치환하여 R, $N_{\overline{B}}$, N_B를 계산한다.

$$R_{W\_FW} = \frac{(R-N_{\overline{B}}) - N_B}{R-N_{\overline{B}}} \qquad (2)$$

계산 과정을 <그림 1(a)>를 예로 보면 "+" 모양으로 흑화소가 분포하는 R(8×8 = 64 pixels) 영역의 각 화소값을 2 차원 필터 윈도(8×8 pixels)의 메디안 필터링을 수행하면 <그림 1(c)>와 같이 영역내의 흑화소는 모두 제거되어 $N_{\overline{B}} = 0, N_B = 28$이 되므로 $R_{W\_FW} = \frac{36}{64}$ 이다. 그러나 1차원 필터 윈도(8×1 pixels)의 메디안 필터링을 영역의 각 점에 수직방향으로 적용하면 <그림 1(b)>의 흑화소 부분은 제거되지 않아 $N_{\overline{B}}$ = 16, N_B = 12가 된다. 그리고 <그림 1(b)>의 결과를 다시 1×8 pixels 크기의 필터 윈도를 갖는 메디안 필터링을 수평방향으로 각 점에 적용하면 흑화소는 모두 제거되어 $N_{\overline{B}}$ = 0, N_B = 16 이다. 따라서 1차원 필터 윈도를 수평방향으로 적용 후 수직방향으로 2번 적용 할 경우 R = 2×64 = 128 pixels 이고 $N_{\overline{B}}$와 N_B는 수평, 수직방향의 값을 각각 더한 형태가 되므로 $R_{W\_FW} = \frac{84}{112}$ 이다. 즉 메디안 필터링을 적용할 때 1차원 필터 윈도를 사용하는 경우가 75%로 2차원의 경우 56%보다 필터 윈도 내부의 백화소 비율이 높아 다수 값 선택에 의한 문자를 구성하는 흑화소 제거 성능이 더 우수함을 확인할 수 있다.

또한 메디안 필터링에서 필터 윈도 내부의 백화소 수를 극대화하기 위해서는 문자 획의 수직 방향으로 1차원 필터 윈도를 적용하여야 한다. 따라서 영어 대, 소문자 52자와 KS 완성형 한글 2350자를 대상으로 문자를 구성하는 획의 방향을 조사한 결과, <표 2>와 같이



(a) (b) (c)

그림 1. 필터 윈도 내부의 백화소 비율 계산 예

Fig. 1. Example of the calculating white pixel ratio in the filter window.

표 2. 영어와 한글에 대한 문자 획 방향 통계

Table 2. Statistics of the direction of character strokes about English and Korean language.

방향 / 언어	가로		세로		기타	
	수(개)	비율(%)	수(개)	비율(%)	수(개)	비율(%)
영어	26	**20**	39	**30**	69	**50**
한글	9519	**46**	7759	**38**	3266	**16**

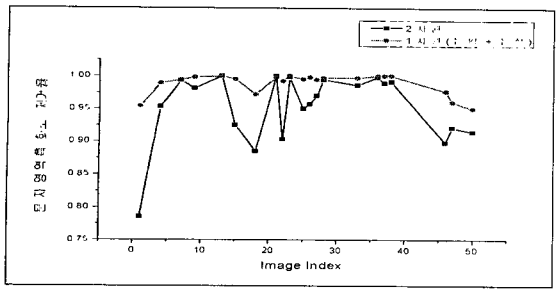


그림 2. 1차원과 2차원 필터 윈도를 이용한 문자영역 흑화소 제거 성능 비교

Fig. 2. Performance comparison of deleting the black pixels in text regions using 1D and 2D filter window.

영어는 가로와 세로방향의 획을 합쳐서 50%, 한글은 84%이므로 1차원 필터 윈도를 수평과 수직 방향으로 적용하는 것이 효과적임을 알 수 있다.

<그림 2>는 21장의 문서영상을 대상으로 메디안 필터링에서 $n \times n$ pixels 크기의 2차원 필터 윈도와 n pixels 크기의 1차원 필터 윈도를 수평과 수직방향으로 적용하여 문자영역 흑화소를 제거한 예로 2차원 필터 윈도를 사용한 경우 평균 95%이고 1차원 필터 윈도를 사용한 경우는 평균 99%로 2차원으로 적용하는 것보다 우수함을 확인하였다.

2.3 필터 윈도 크기 결정

필터 윈도의 크기 (l_w)는 내부의 백화소 수를 극대화할 수 있도록 결정해야 하지만 문서의 경우 문자의 크기가 다양하여 l_w를 정하기 어렵다. 그러나 문서에서 문자의 분포가 가장 많으므로 분할 영역들의 평균 크기 (l_{ave_seg})는 문자의 평균 크기 (l_{ave_char})와 비슷하다는 점을 이용하여 l_w를 결정한다. 또한 필터링 했을 때 모든 문자를 다 분류하지는 못해도 모든 크기의 문자를 분류할 수 있으면 후 처리에서 오류 수정이 가능하므로 l_w 크기를

$$l_w = 4 \times l_{ave\_char} \tag{3}$$

로 하여 l_{ave_char}보다 큰 모든 크기의 글자가 분류되도록 한다. 식 (3)에서 l_w는 시험 문서영상 50장을 대상으로 실험하여 결정한 값이며 그 결과의 한 예를 <그림 3>에 나타내었다. 그리고 필터링 결과 문자로 분류된 영역을 음영이 있는 사각형으로 표시하였다.

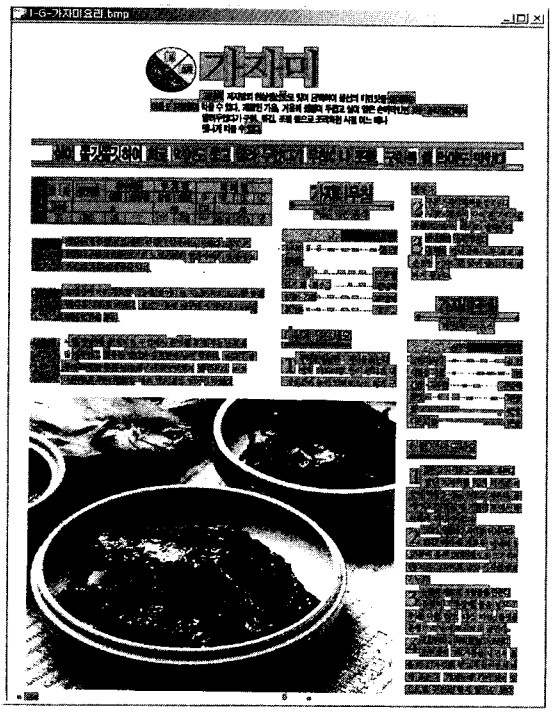


그림 3. 1차원 메디안 필터를 이용한 문자분류 예

Fig. 3. Example of text classification using one dimensional median filter.

3. 메디안 필터링에 의한 오 분류 처리

식 (3) 크기의 필터 윈도를 갖는 메디안 필터를 이용하여 이진 문서영상을 문자와 그림으로 분류하면 <그림 3>의 예와 같이 흑화소 밀도가 높은 일부 볼드체 문자는 그림으로 분류되고 그래프나 표와 같은 밀도가 낮은 그림은 문자로 분류되는 문제가 있다. 따라서 문자와 그림분류의 정확성을 높이기 위해서는 이런 부분에 대한 처리가 필요하다.

3.1 표피 제거 필터를 이용한 볼드체 문자 처리

일반적으로 문서에서는 문자의 분포가 가장 많기 때문에 문서영상 분할 영역들의 평균 크기 (l_{ave_seg})는 문자의 평균 크기 (l_{ave_char})와 비슷하다. 그리고 필터 윈도 크기는 $4 \times l_{ave_seg}$이므로 l_{ave_char} 보다 크기가 작고 <표 1>과 같이 흑화소 밀도가 30% 이내인 문자의 경우 메디안 필터링을 이용하여 문자로 분류할 수 있다. 따라서 볼드체 문자의 흑화소 밀도를 30% 정도로 낮출 필요가 있으며 표피 제거 필터(skin peeling filter)를 이용하여 이러한 기능을 수행한다. 표피 제거 필터는 메디안 필터와 원리가 같고 필터 윈도 내부의 값

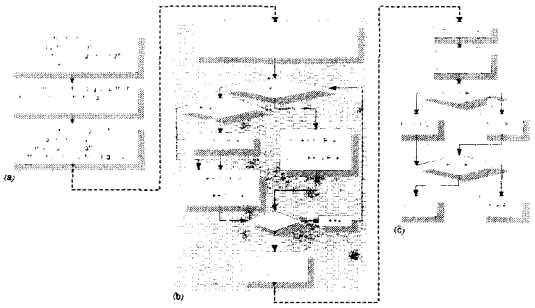
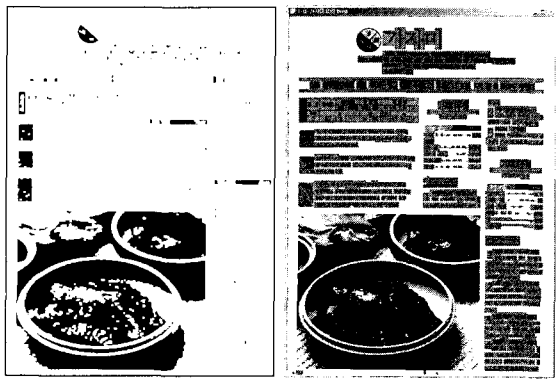


그림 4. 표피 제거 필터 수행 횟수 결정을 위한 흐름도
Fig. 4. Flowchart to decide the iteration number of skin peeling filter.

중 최대(가장 밝은) 값으로 해당 화소 값을 치환하는 것만 다르다. 따라서 문자 획의 외각을 구성하는 흑화소는 백화소로 바뀌게 되어 획이 얇아지게 된다. 제안한 방법에서는 3×3 pixels 크기의 필터 윈도를 적용한 표피 제거 필터를 사용하므로 필터링 횟수에 따라 문자 획이 얇아지는 정도가 다르다. 또한 필터링을 수행하면 그림부분의 흑화소도 같이 제거되므로 적절한 필터링 횟수를 결정 할 필요가 있다. 따라서 <그림 4>의 과정과 같이 문자의 크기와 밀도를 고려하여 적절한 횟수를 정한다.

<그림 4>는 기능에 따라 ⓐ, ⓑ, ⓒ 3개의 부분으로 구성된다. ⓐ 과정에서는 2절에서 수행한 메디안 필터링 결과 문자로 분류된 분할 영역을 대상으로 ±10% 크기 범위 내에서 가장 많은 수를 차지하는 문자의 크기(L_m)와 흑화소 밀도(D_m)를 구한다. ⓑ 과정에서는 크기가 L_m보다 작은 문자들은 평균 밀도(D_l)만 구하고 L_m보다 큰 문자들은 평균 밀도(D_h)와 최대크기(L_{h_max})를 함께 구한다. 이것은 문자의 밀도가 크기가 L_m보다 크면 밀도가 낮더라도 문자의 크기에 의해 획이 굵어져서 윈도 크기가 $4\times l_{ave_seg}$인 메디안 필터링에 의해 제거되지 않는 경우가 발생하므로 크기와 밀도를 모두 고려하고 나머지는 밀도만 고려하여 필터링 횟수를 결정하기 위해서이다. ⓒ 과정에서는 L_{h_max}/L_m, D_h/D_m, D_l/D_m 가운데 가장 큰 값을 표피 제거 필터의 수행 횟수(n)로 결정한다. 즉 표피제거 필터링을 크기와 밀도비 수만큼 반복함으로써 문자획의 흑화소 밀도를 감소시킨 후 필터 윈도 크기가 l_{ave_seg}인 1차원 메디안 필터를 다시 적용하여 문자와 그림을 분류함으로써 언급한 문제점을 해결한다. <그림 5(a)>는 표피 제거 필터링을 수행한 결과이며 <그림 5(b)>는 <그림 5(a)>를 메디안 필터링 하여 문자를 분류한 결과로 <그림 3>에서 제외되었던 볼드체 부분도 문자로 분류됨을 확인 할 수 있다.



(a) (b)

그림 5. 표피 제거 필터를 이용한 문자 분류 예 (a) 표피 제거 필터링 결과, (b) (a) 영상을 메디안 필터링 하여 문자를 분류한 결과
Fig. 5. Example of text classification using skin peeling filter. (a) The result of skin peeling filtering, (b) The result of text classification from image (a) by median filtering.

3.2 저 밀도 그림 분류

메디안 필터를 이용하여 문자와 그림을 분류하면 <그림 5(b)>의 예와 같이 표, 그래프와 같은 밀도가 낮은 그림과 직선은 문자로 분류된다. 그러나 저 밀도 그림을 직접 분류할 수 있는 특징을 찾기는 어렵다. 따라서 3.1절에서 분류한 문자들을 대상으로 문자의 최대크기(L_{max_char})를 기준으로 문자를 먼저 분류함으로써 나머지 저 밀도 그림과 직선을 분류한다. 그리고 문자의 최대크기를 구하기 전에 다음 5가지에 해당하는 불확실한 문자들을 제외함으로써 최대크기에 대한 정확성을 높인다.

① 내부에 분할 영역을 포함하는 문자
② 흑화소 밀도가 0.2 이하 또는 1인 문자
③ 문자배열 방향으로의 분할 영역 종횡비가 5보다 큰 문자
④ 문자배열 방향의 수직 방향으로 분할 영역 종횡비가 2.8보다 큰 문자
⑤ 상호 상관성 방법에 의해 그림으로 분류된 문자

표는 내부에 문자를 포함하므로 ①의 경우는 제외한

다. <표 1>을 보면 문자의 평균 흑화소 밀도가 0.3 내외이므로 0.2 이하는 문자가 아닐 확률이 높고 밀도가 1인 경우는 직선일 가능성이 있어 ②의 경우는 제외한다. ③과 ④는 직선의 가능성이 있어 제외한다. ③에서 종횡비의 기준 5는 입력 문서영상의 축소 처리로 인하여 문자배열 방향으로 문자들이 붙어 하나의 분할 영역을 형성하는 경우를 고려하여 정한 실험 값이다. ④에서 종횡비 기준 2.8은 영어 대, 소문자 52자와 KS 완

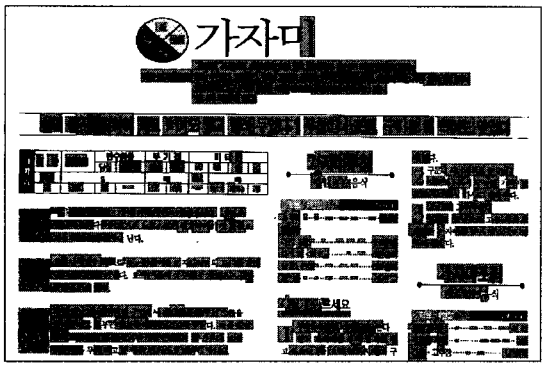


그림 6. 문자영역 추출 예
Fig. 6. Example of text region classification.

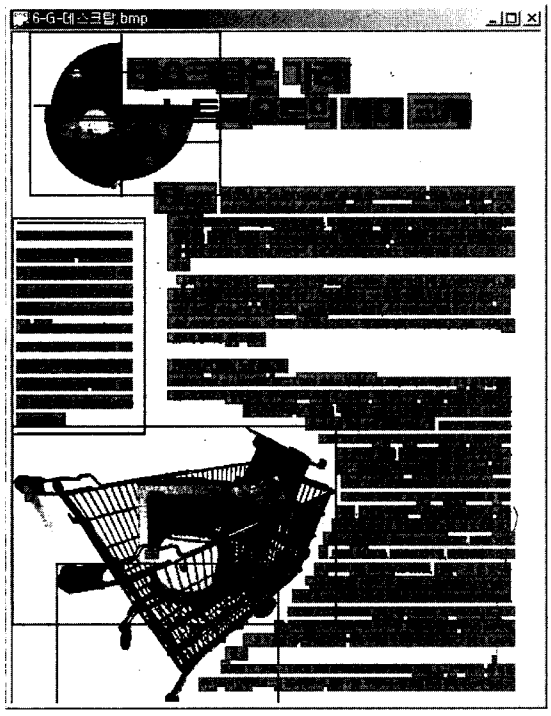


그림 7. 제안한 방법을 이용한 문자와 그림분류 예
Fig. 7. Example of text and picture classification using proposed method.

성형 한글 가운데 직선으로 오 분류될 가능성이 있는 '一', 'ㅣ'를 제외한 2399자를 대상으로 각 문자의 종횡비를 계산한 결과 최대 값이 2.8이므로 이 값보다 종횡비가 큰 문자는 직선일 가능성이 있어 제외한다. ⑤는 Ⅱ장 2.2절에서 소개한 방법을 이용한 문자와 그림분류 방법이다.

<그림 6>은 3.1절에서 분류한 문자 가운데 ①~⑤에 해당하는 문자를 제외한 나머지 문자를 음영이 있는 사각형으로 표시한 예로 이들을 대상으로 $L_{\max_char}$를 결정한다.

$L_{\max_char}$가 결정되면 3.1절에서 분류한 문자들 가운데 크기가 이 값보다 작은 영역만 문자로 분류하고 나머지는 그림으로 분류하며, <그림 7>은 제안한 방법을 이용한 문자와 그림분류의 예로 문자는 음영이 있는 사각형으로 그림은 음영이 없는 사각형 테두리선으로 나타내었다.

Ⅳ. 실험 및 고찰

문서영상 영역해석에서 영역분할은 문단을 나누거나 합치는 등 주관적 차이가 있을 수 있다. 그러나 영역분류에서의 오류는 다른 형태의 전자문서로 전환하게 되므로 실험에서는 영역분류의 정확성 (C_r)

$$C_r = \frac{(N_r - N_{mcr})}{N_r} \times 100 \tag{4}$$

을 측정하여 그 성능을 평가하였다. C_r은 문서들을 수(手)작업으로 영역분류 한 결과와 상기된 방법들을 이용한 결과의 비율로 각 방법의 성능이 인간이 수행한 결과에 근접한 정도를 나타낸다. 그리고 식 (4)에서 N_r은 수작업으로 분할한 영역의 수이고 N_{mcr}은 수작업으로 분할한 영역 중 해당 방법을 이용한 오 분류 영역 수이다.

성능평가는 제안한 영역해석 방법을 기존 방법 3종류, 상용제품 3종류와 영역분류 성능을 비교하여 평가하였다. 그리고 입력 문서영상은 300 dpi 해상도의 스캐너 영상과 디지털 카메라를 이용한 1024×768 pixels 크기의 영상을 사용하였으며 대상 문서는 사용하는 언어와 문서구조가 다양하여 단일화된 시험영상을 생성하기 어려우므로 본 실험에서는 자체적으로 마련한 문서영상 40장을 대상으로 실험하였다. 그리고 대상 문서

표 3. 영역분류 성능 비교
Table 3. The performance comparison of region classification.

형태 / 방법	문자		그림		계	
	개수(개)	C_r(%)	개수(개)	C_r(%)	개수(개)	C_r(%)
수(手) 작업	392	**100**	164	**100**	556	**100**
방법[1]	376	**96**	141	**86**	517	**93**
방법[5]	369	**95**	129	**75**	498	**90**
방법[7]	358	**92**	123	**79**	481	**87**
A 6.0	325	**83**	118	**72**	443	**80**
Omnipage Pro 12	313	**80**	113	**69**	426	**77**
Fine Reader 5.0	311	**79**	92	**56**	403	**73**
제안한 방법	389	**99**	158	**96**	547	**98**

는 신문, 잡지, 논문, 서류, 영수증 등 다양한 종류에서 주로 영역분할이 어려운 형태의 문서들 위주로 선별하였으므로 기존의 방법이나 상용제품의 성능평가 결과가 해당 방법이나 제품에서 발표한 것보다 낮다는 것을 밝혀둔다.

<표 3>은 기존의 영역해석 방법 가운데 연결요소를 이용하여 영역을 분할하고 문자와 그림의 크기와 밀도 차이 그리고 문자들간의 인접성을 이용하여 분할한 영역의 형태를 분류하는 방법[1]과 연결요소를 이용하여 영역을 분할하고 상호 상관성을 이용하여 분할한 영역의 형태를 분류하는 방법[5]과 투영윤곽을 이용하여 영역을 분할하고 문자배열의 반복성을 이용하여 분할한 영역의 형태를 분류하는 방법[7]과 상용으로 국내 제품인 P사의 A 6.0과 Scansoft사의 Omnipage Pro 12와 ABBYY사의 Fine Reader 5.0과 제안한 영역해석 방법의 C_r을 비교한 결과로 제안한 방법의 문자와 그림분류 성능이 우수함을 확인할 수 있다.

V. 결 론

본 논문에서는 연결요소를 이용하여 문서영상을 영역분할하고 1차원 메디안 필터를 기반으로 하여 분할한 영역의 형태를 문자와 그림으로 분류하는 문서영상 영역해석 방법을 제안하였다.

기존의 영역분류 방법들[1,5]은 영역의 크기, 흑화소 밀도와 분포의 복잡도가 비슷한 문자와 그림의 경우 오분류가 자주 발생한다. 또한 문자배열의 반복성을 이용하여 문자를 분류하는 방법[7]의 경우 글자 수가 적은 영역은 분류하기 어려운 문제점이 있다. 그러나 제안한 방법의 경우 1차원 메디안 필터를 이용하여 문자와 그림을 분류하고 메디안 필터링에 의해 발생하는 문제점들을 해결함으로써 특징이 비슷한 문자와 그림간 오분류가 적고 연결요소 단위로 분할한 작은 영역까지 상세한 분류가 가능하여 구조가 복잡한 문서도 영역해석이 가능하다.

영역분류 성능평가에서는 상용제품을 포함한 기존 방법들[1, 5, 7]과의 비교를 통하여 제안한 방법의 성능을 평가하였다. 성능평가 결과 기존 방법들의 경우 87~93%의 영역분류도를 나타냈으며 상용제품의 경우 방법 73~80%의 영역분류도를 나타낸 반면 제안한 방법의 경우 98%로 신문, 잡지, 논문, 서류, 영수증 등 다양하고 복잡한 구조를 갖는 문서영상에 대해 상용제품을 포함한 기존 방법들보다 영역분류 성능이 우수함을 확인 할 수 있었다.

제안한 문서영상 영역해석 방법은 연결요소를 이용하여 영역을 분할하므로 하향식에 비해 버퍼 사용량과 연산량이 많은 문제점이 있다. 그리고 카메라를 이용하여 문서영상을 생성 할 경우 조명의 변화로 인해 이진화가 잘 되지 않는 문제와 카메라 각도의 변화로 인한 기울어짐이 발생하는 문제를 해결하기 위한 효과적인 전처리 방법의 개발이 필요하다. 그리고 제안한 방법이 인간이 수행하는 것과 같은 수준의 영역해석 성능을 발휘하기 위해서는 그림 가운데 표를 분류하고 분석하기 위한 처리부분의 추가가 필요하며 칼라 문서영상에 대한 영역해석 방법의 개발이 필요하다.

참 고 문 헌

[1] X. Li, W. Gao, S. Y. Chi, K. A. Moon and H. J. Kim, "An Efficient Method for Page Segmentation," Proc. ICICS, vol.2, pp. 957~961, 1997.

[2] D. Drivas and A. Amin, "Page Segmentation and Classification Utilizing Bottom-up Approach," Proc. ICDAR, pp. 610~614, 1995.

[3] L. O'Gorman, "The Document Spectrum for Page Layout Analysis," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.15, No.11, pp. 1162~1173, Nov. 1993.

[4] N. Papamarkos, J. Tzorttzakis and B. Gatos,

"Determination of Run-Length smoothing values for document segmentation," IEEE 3th Int. Conf. on Electronics, Circuits and Systems, ICECS, pp. 684~687, 1996.

[5] S. K. Yip and Z. Chi, "Page Segmentation and Content Classification for Automatic Document Image Processing," Proc. Int. Symp. Intelligent Multimedia, Video and Speech Processing, pp. 279~282, 2001.

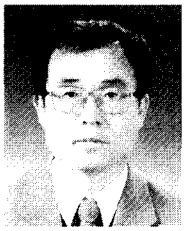
[6] J. Kong and Z. Chi, "Image Classification Using Kolmogorov Complexity Measure with Extracted Blocks," IEICE Trans. Inf. & Syst., Vol.1, E81-D, pp. 1239~1246, 1998.

[7] S. W. Lee and D. S. Ryu, "Parameter-Free Geometric Document Layout Analysis," IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.23, No.11, Nov. 2001.

저 자 소 개

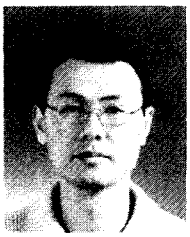
朴 承 浩(學生會員)

2002년 2월 : 경북대학교 전자·전기공학부 학사. 2002년 3월~현재 : 경북대학교 전자 공학과 대학원 석사과정. <주관심분야 : 문서영상처리, 문서자동분류, 영상처리>



黃 燦 植(正會員)

1979년 : 한국과학기술원 전자공학과(공학석사). 1996년 : 한국과학기술원 전자공학과(공학박사). 1979~현재 : 경북대학교 전자·전기·컴퓨터 학부 교수. <주관심분야 : 영상통신, 암호통신, 초고속망>



張 大 根(正會員)

1997년 : 경북대학교 전자공학과(공학석사). 2003년 : 경북대학교 전자·전기·컴퓨터공학부(공학박사). 1996년~현재 : 한국전자통신연구원·컴퓨터소프트웨어기술연구소 공간정보기술센터·영상인식연구팀 선임연구원. <주관심분야 : 문서영상처리, 정지영상 및 동영상 부호화, 위성영상분석, 인공지능>