이미지 필터와 제한조건을 이용한 문서영상 구조분석

장 대 근† · 황 찬 식††

요 약

문서영상 구조분석은 문서영상을 세부 영역으로 분할하는 과정과 분할된 영역을 문자, 그림, 표 등으로 분류하는 과정을 포함한다. 이 중 영역분류 과정에서 영역의 크기, 흑화소 밀도, 화소 분포의 복잡도는 영역을 분류하는 기준이 된다. 그러나 그림의 경우 이러한 기준들의 범위가 넓어 경계를 정하기 어려우므로 다른 형태에 비해 상대적으로 오분류의 비율이 높다. 본 논문에서는 그림과 문자를 분류하는 과정에서 영역의 크기, 흑화소 밀도, 화소 분포의 복잡도에 의한 영향을 줄이기 위하여 메디안 필터를 이용하고, 영역확장 필터(region expanding filter)와 제한조건들을 이용하여 영역분류에서의 오분류를 수정함으로써 상용제품을 포함한 기존 방법에 비해 그림과 문자의 분류가 우수한 문서영상 구조분석 방법을 제안한다.

Document Image Layout Analysis Using Image Filters and Constrained Conditions

Dae Geun Jang† · Chan Sik Hwang††

ABSTRACT

Document image layout analysis contains the process to segment document image into detailed regions and the process to classify the segmented regions into text, picture, table or etc. In the region classification process, the size of a region, the density of black pixels, and the complexity of pixel distribution are the bases of region classification. But in case of picture, the ranges of these bases are so wide that it's difficult to decide the classification threshold between picture and others. As a result, the picture has a higher region classification error than others. In this paper, we propose document image layout analysis method which has a better performance for the picture and text region classification than that of previous methods including commercial softwares. In the picture and text region classification, median filter is used in order to reduce the influence of the size of a region, the density of black pixels, and the complexity of pixel distribution. Futhermore the classification error is corrected by the use of region expanding filter and constrained conditions.

키워드 : 문서영상(document image), 영역해석(region analysis), 영역분할(page⟨region⟩ segmentation), 영역분류(region classification)

1. 서 론

정보화와 더불어 전자문서의 사용이 증가함에 따라 인쇄문서의 사용은 감소할 것이라는 예상과는 달리 프린터와 같은 컴퓨터를 이용한 출력장치의 개발로 인해 예전보다 인쇄문서의 양은 더욱 늘어나고 있는 추세다. 따라서 인쇄문서를 직접 손으로 입력하지 않고 편집 가능한 전자문서로 자동전환의 필요성이 갈수록 증가하고 있다. 인쇄문서를 전자문서로 자동 전환하려면 문서영상 구조분석, 문자인식 등의 요소기술이 필요하며 이 중 문서영상을 문자, 그림, 표 등의 세부 영역으로 분할하는 문서영상 구조분석은 전표, 수표 등의 형식문서 인식, 문서영상 색인 및 검색, 다계층 문서영상 압축(multi-layer document image compression) 등의 전처리로 사용되는 기반기술이다. 문서영상 구조분석은 문서영상을 세부영역으로 분할하는 영역분할과정과 분할된 영역을 문자, 그림, 표 등의 영역으로 분류하는 영역분류과정을 포함하며 특히 전자문서로의 자동 전환에서 영역분류에서의 오류는 문서 본연의 형태와는 다른 전자문서를 만든다.

기존에 개발된 영역분류 방법으로는 문자를 구성하는 연결요소의 크기와 밀도는 그림의 경우와 다르다는 특성을 이용하는 방법 [1]과 문자열 부분의 흑화소 분포의 복잡도를 수치화하는 방법으로 cross correlation approach[6], Kolmogorov complexity measure[7], texture pattern analysis[8] 등이 있다. 그러나 기존의 [1, 6-8]의 방법으로는 그림의 경우 영역의 크기, 흑화소 밀도, 화소 분포의 복잡도의 범위가 넓어 문자와 구분할 수 있는 기준을 정하기 어렵다. 또한 문

† 정 회 원 : 경북대학교 대학원 전자 · 전기 · 컴퓨터학부
†† 정 회 원 : 경북대학교 전자 · 전기 · 컴퓨터학부 교수
논문접수 : 2002년 5월 3일, 심사완료 : 2002년 6월 3일

자의 크기도 다양하여 상대적으로 크고 밀도가 높은 문자들이 그림으로 분류되는 경향이 있다.

본 논문에서는 그림과 문자의 분류과정에서 영역의 크기, 흑화소 밀도, 화소 분포의 복잡도에 의한 영향을 줄이기 위하여 문자제거 효과가 우수한 메디안 필터를 사용함으로써 크기가 작고 밀도가 낮은 그림이 문자로 분류되거나 상대적으로 크기가 크고 밀도가 높은 문자들이 그림으로 오분류되는 비율을 감소시켰다. 또한 영역확장 필터와 제한조건들을 이용하여 메디안 필터링에 의해 오분류된 영역을 수정함으로써 그림과 문자의 분리에서 오분류의 확률을 더욱 감소시켜 문서영상 구조분석 성능을 향상시키는 방법을 제안한다.

2. 기존 문서영상 구조분석 방법 분석

문서영상 기하학적 구조분석은 1. 서론에서 언급한 바와 같이 문서영상을 세부영역으로 분할하는 방법과 분할된 영역을 문자, 그림, 표 등으로 분류하는 방법을 필요로하며 각각에 대한 설명은 다음과 같다.

2.1 영역분할

2.1.1 상향식 영역분할

기본이 되는 화소단위에서 시작하여 유사성을 갖는 부분을 점차적으로 크고 의미를 부여할 수 있는 단위로 단계적으로 병합하는 방법으로 기울어진 문서를 포함하여 여러 가지 다양한 형태의 문서를 처리할 수 있다는 장점이 있는 반면 많은 계산량과 버퍼를 필요로하는 단점이 있다. 상향식 방법으로는 연결요소를 이용하는 방법 [1, 2]와, 인접선분밀도를 이용하는 방법 [3]이 있다.

2.1.2 하향식 영역분할

문서의 전체적인 영역에서 시작하여 문서를 점점 작은 영역으로 분할하는 방법으로 알고리즘이 간단하고 빠르며 영역이 사각형 블록으로 구성된다는 장점이 있으나 복잡한 형태의 문서나 기울어진 문서에는 적용하기 어렵다는 단점이 있다. 하향식 방법으로는 투영 윤곽(projection profile) 이용법 [4]와 런 길이 평활화(run length smoothing)를 이용하는 방법 [5]가 있다.

2.2 영역분류

임의의 한 가지 방법으로 문자, 사진, 그래프, 차트, 표, 선과 같은 다양한 항목들을 효과적으로 분류하기는 어렵다. 따라서 다양한 항목을 구분하기 위한 효과적인 방법들을 복합적으로 적용해야 하며 이러한 방법들은 해당 속성을 구분하는 능력이 우수할 뿐 아니라 실시간 처리를 위해 알고리즘이 간단하고 계산량이 적어야 하는 어려움이 있다. 영역분류를 위한 기존의 방법은 다음과 같다.

2.2.1 연결요소의 크기와 밀도 이용

영역분류의 가장 일반적인 방법으로 문자를 구성하는 연결요소의 크기, 종횡비, 밀도가 그림과 다르다는 점을 이용하여 속성을 분류하는 방법이다. 이 방법은 알고리즘이 간단하다는 장점이 있는 반면 그림의 경우 연결요소의 크기와 밀도가 다양하고 흑화소의 분포가 광범위하여 문자와 구분할 수 있는 기준을 정하기 어렵다. 또한 문자의 크기도 다양하여 상대적으로 크고 밀도가 높은 문자들이 그림으로 분류되는 경향이 있다.

2.2.2 cross correlation approach

백화소가 0 흑화소가 1로 표현되는 이진수열에서 가로방향과 세로방향으로 이웃하는 화소와 xor 연산을 수행하여 0과 1이 바뀌는 횟수를 구하고 이 값을 이용하여 영역을 문자와 그림으로 구분한다. 이 방법은 그림을 구성하는 흑화소 분포가 다양하여 문자와 그림의 분류기준을 정하기 어렵고, 영역을 구성하는 문자수가 적은 경우 오분류의 확률이 높다는 단점이 있다. cross correlation은 다음 식 (1)을 이용하여 계산한다.

$$Cr = 1 - \frac{2}{MN} \sum_{i=0}^{N-1} \sum_{j=0}^{M-1} [p(i,j) \oplus p(i+1, j)] \quad (1)$$

C_r : cross correlation
M : 영역의 가로방향 화소 수,
N : 영역의 세로방향 화소 수
$\oplus$: exclusive or operation

2.2.3 Kolmogorov complexity measure

백화소가 0 흑화소가 1로 표현되는 이진수열을 1차원으로 배열한 후 Kolmogorov가 제안한 식 (2)와 식 (3)을 이용하여 Complexity(KC)를 계산하여 그림과 문자로 구분하는 방법이다. 이 방법 또한 Cross Correlation을 이용한 경우와 장단점 및 성능이 비슷하다.

$$KC = \frac{c(n)}{b(n)} = \frac{1}{n} c(n) \log_2 n \, n \quad (2)$$

$$\lim_{n \to \infty} c(n) = b(n) = \frac{n}{\log_2 n} \quad (3)$$

KC : Kolmogorov Complexity
$c(n)$: complexity for finite strings of length n
$0 \leq KC \leq 1$

2.2.4 Texture Pattern Analysis

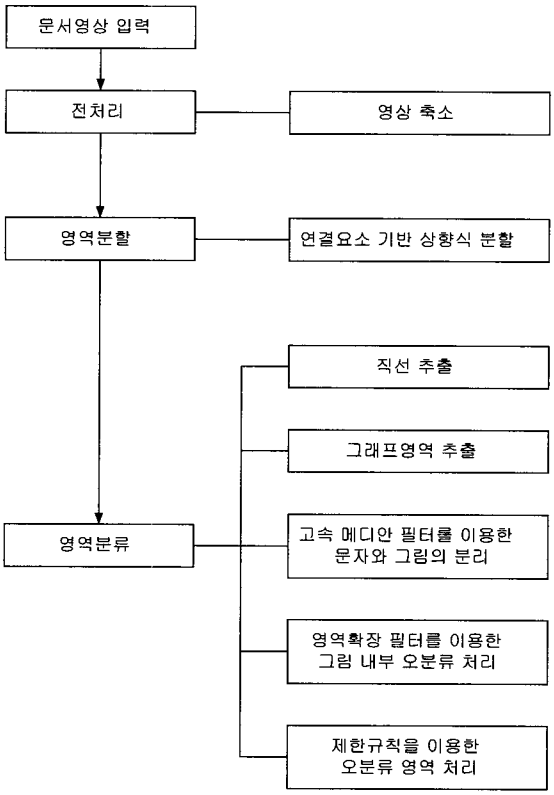
입력영상을 n-차원의 texture pattern 벡터들로 표현하고 많은 영상으로부터 미리 준비한 n-차원의 texture pattern 벡터 bank와의 비교를 통하여 문자, 그림, 배경으로 구분하는 방법이다. 이 방법은 다양하고 많은 입력영상으로부터 texture pattern vector bank를 생성해야하며, 그림은 흑화소

분포가 다양하여 문자와 패턴이 비슷한 경우가 있어 오분류가 발생한다. 또한 영역을 구성하는 문자수가 적은 경우 오분류의 확률이 높다는 단점이 있다.

3. 제안한 문서영상 구조분석 방법의 전체구성 및 처리과정

3.1 전체구성

제안한 문서영상 구조분석 방법의 전체구성은 (그림 1)과 같이 전처리, 영역분할, 영역분류의 3단계 과정으로 구성된다. 전처리에서는 실시간 처리를 위해 영상을 축소한다. 영역분할은 연결요소 기반 상향식 방법을 고속화한 Xingyuan. Li's 방법 [1]을 사용한다. 영역분류에서는 1차원 메디안 필터링을 이용한 직선추출, 흑화소 밀도를 이용한 그래프영역 추출, 고속 메디안 필터링을 이용한 그림과 문자의 분류, 영역확장 필터를 이용한 그림영역 내부의 오분류 처리, 제한조건을 이용한 오분류 처리의 순으로 과정이 수행한다.



(그림 1) 제안한 문서영상 구조분석 방법의 전체 구성

3.2 처리과정

3.2.1 전처리

PC를 이용하여 실시간으로 구조분석을 수행하기 위해 입력영상을 축소하여 처리한다. 축소비율(r)은 디지털 카메라를 이용한 영상입력을 고려하여 조정한다. 해상도가 너무 낮은 경우 표를 구성하는 선과 데이터가 붙어 훼손되며 너무 높은 경우 실시간 처리가 어려우므로 1000×1000pixels 보다 약간 작은 크기로 축소되도록 아래의 식 (4)에서 입력영상의 가로, 세로 길이 중 큰 값($MaxLen$)을 분모 값 1000으로 나누어 반올림한다.

$$r = Q[\frac{MaxLen}{1000}] \quad (4)$$

r : 축소비율

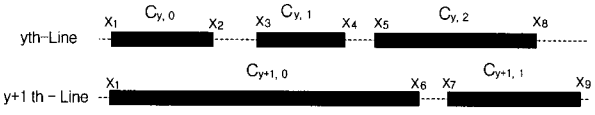
$Q[\alpha]$: α를 반올림하여 정수화

$MaxLen$: 입력영상의 가로, 세로 길이 중 큰 값 (pixel 단위)

영상축소는 입력영상을 $r \times r$ 크기의 영역으로 겹치지 않게 분할한 후 분할영역 내부에 흑화소가 한개 이상 존재하는 경우는 $r \times r$ 영역을 1개의 흑화소로 표현하고 나머지는 1개의 백화소로 표현하여 영상을 축소한다.

3.2.2 영역분할

영역분할은 연결요소를 기반으로 하는 상향식 분할방법을 고속화한 Xingyuan Li's 방법 [1]을 이용하여 복잡한 구조의 문서도 세밀하게 영역분할 하는 것이 가능하다. 분할과정을 (그림 2)를 예로 보면 $C_{y,i}$와 같은 연결요소를 가로방향 라인단위로 먼저 생성한다. 라인단위로 생성한 연결요소들을 결합하기 위하여 첫 번째 라인부터 차례로 기준라인(base line)과 그 다음라인을 비교라인(comparative line)으로 설정하고 식 (5)를 만족하는 즉 서로 연결 관계가 있는 두 라인 간 연결요소들을 결합함으로써 영역을 생성한다. (그림 2)의 경우 두개 라인의 모든 연결요소들은 하나로 결합되는 결과가 된다.



(그림 2) 인접한 라인에서 연결요소 예

$$\min[\max(C_{y,m}), \max(C_{y+1,n})] \geq \max[\min(C_{y,m}), \min(C_{y+1,n})]$$

$$\text{example} > X1 = \min(C_{y,0}),\ X_2 = \max(C_{y,0}) \quad (5)$$

3.2.3 영역분류

3.2.3.1 1차원 메디안 필터를 이용한 직선분류

문서영상에서 직선은 영상입력 장치의 오차로 인하여 두께가 일정하지 않으며 테두리 또한 요철로 이루어진다. 그리고 끊어지거나 노이즈가 포함된 경우도 있어 직선 분류에 1차원 직선의 방정식을 적용하기에는 어려움이 있다. 따라서 제안한 시스템에서는 언급한 직선분류의 문제점들을 해결하기 위하여 영역의 종횡비와 1차원 메디안 필터를 이용하여 직선을 추출한다. 직선은 문자, 그림에 비해 영역의

종횡비가 크다. 따라서 종횡비가 5이상인 영역을 직선이 될 수 있는 후보로 분류하며 임계값 5는 많은 문서영상에서 직선을 대상으로 실험하여 정한 값이다. 메디안 필터는 해당 화소값을 주변 화소값들을 포함한 값 가운데 중간값으로 바꾸는 필터링으로 impulse noise를 제거하는 효과가 있다. 따라서 적당한 크기의 탭을 갖는 1차원 메디안 필터를 종횡비를 이용하여 추출한 후보영역을 대상으로 종회비가 큰 방향으로 적용하면, 직선의 경우 내부의 흑화소는 필터링 후에도 대부분이 그대로 남게 된다. 또한 선의 끊어진 부분이 필터 탭 길이보다 짧은 경우 중간값을 택하는 필터의 특성으로 인해 다시 연결되는 장점도 있다. 직선분류는 메디안 필터링 전, 후의 영역내부 흑화소 수의 비(r_d)를 식 (6)를 이용하여 구하고 $r_d \geq 0.98$인 경우 직선으로 추출한다.

$$r_d = \frac{n_{bp}(R_i)}{n_{ba}(R_i)} \tag{6}$$

$n_{bp}(B)$: 메디안 필터링 후의 영역 R_i 의 흑화소 수
$n_{ba}(B)$: 메디안 필터링 전의 영역 R_i 의 흑화소 수

3.2.3.2 그래프영역 분류

표, 차트, 그래프를 구성하는 연결요소의 두께는 문자와 비슷한 경우가 많아 메디안 필터링에 의해 그림으로 분류되지 않는다. 그러나 내부의 데이터영역들을 제거한 경우 영역의 밀도는 25%이하가 대부분이므로 이 조건을 이용하여 표, 차트, 그래프를 모두 그래프영역으로 분류하며 분류과정에서 조건이 비슷한 문자가 그래프로 오분류된 경우는 3.2.3.5의 오분류 처리 과정에서 문자로 수정한다.

3.2.3.3 메디안 필터링을 이용한 그림, 문자 분류

3.2.3.1에서 설명한 메디안 필터를 문서영상에 적용하면 문자를 구성하는 흑화소는 제거되고 밀도가 높은 사진과 같은 그림부분의 흑화소는 남게 되어 문자와 그림의 분리가 가능하다. 따라서 메디안 필터링 결과 제거되지 않고 남아 있는 흑화소 부분을 포함하는 분할영역은 그림영역으로 나머지는 문자영역으로 분류한다. 필터링에서의 탭 크기는 문자는 제거되고 그림은 제거되지 않을 정도의 크기로 정해주어야 하므로 다양한 입력문서의 해상도와 문자 크기에 맞게 정하기 위하여 3.2.2에서 분할한 영역들의 평균길이 ($L_{ave}(R)$)의 2배 크기로 설정한다. $L_{ave}(R)$은 분할된 각 영역의 길이를 평균한 값이며 식 (7)과 같다.

$$L_{ave}(R) = \frac{\sum_{i=0}^{n(R)-1} L(R_i)}{n(R)} \tag{7}$$

$L_{ave}(R)$: 분할 영역들의 평균길이
$L(R_i)$: index i 영역의 세로길이
$n(R)$: 분할 영역의 총 수

3.2.3.3.1 문서영상 특성을 고려한 히스토그램 이용 메디안 필터링의 중간값 추출

일반적으로 메디안 필터링의 중간값 추출은 sorting을 이용하여 값들을 정렬한 다음 중간값을 찾아낸다. 그러나 1000×1000 pixels 크기의 문서영상에 sorting을 이용한 메디안 필터링을 적용할 경우 실시간 처리가 어려우므로 필터 탭 내의 화소값에 대한 히스토그램을 구하여 분포를 누적시킨 빈도가 필터탭 크기의 $\frac{1}{2}$ 이상이 될 때까지만 계산하여 중간값을 구한다. 히스토그램 분포를 누적시킬 때는 문서영상은 흑화소 보다 백화소 수가 많다는 특성을 이용하여 밝은 단계에서 어두운 단계로 누적시킴으로써 연산수를 감소시킨다. 이진영상의 경우 화소의 밝기 단계가 0과 1의 2단계뿐이므로 계산량은 더욱 감소한다.

3.2.3.3.2 separable 메디안 필터링

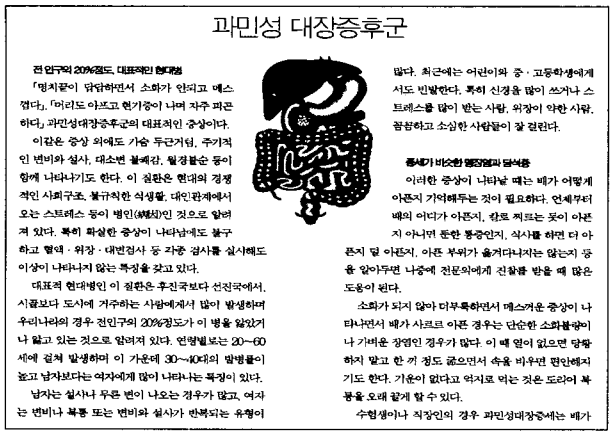
현재 PC의 연산능력으로는 1000×1000 pixels 크기의 256 그레이 영상이나 이진 영상을 대상으로 2차원 메디안 필터링을 실시간으로 수행하기 어려우므로 1차원 메디안 필터링을 수평방향으로 수행한 결과를 다시 수직방향으로 수행하는 separable 방식의 필터링을 수행함으로써 필터링을 실시간으로 수행한다.

3.2.3.3.3 그림, 문자 분류 예

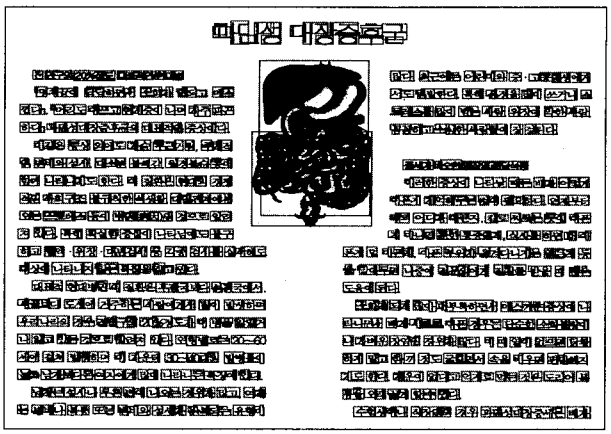
(그림 3)은 제안한 메디안 필터링을 이용하여 분할된 영역들을 문자와 그림으로 분류한 예이다. (a)는 입력영상이고 (b)는 Xingyuan Li's 방법 [1]을 이용하여 영역분할한 예이다. (c)는 제안한 메디안 필터링에 의해 제거되지 않고 남은 흑화소 부분이다. (d)는 (c)의 결과를 이용하여 문자(음영이 있는 사각형)와 그림(사각형 테두리 영역)영역을 분류한 예이다.

3.2.3.4 영역확장 필터를 이용한 그림 내부 오분류 처리

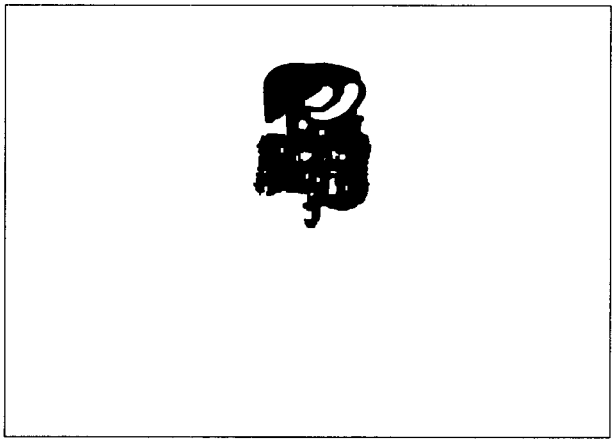
3.2.2의 연결요소 기반 상향식 영역분할은 문자의 자소 단위까지 상세한 영역분할이 가능하다. 그러나 그림영역 내부의 문자를 분류하는 경우 그림을 형성하는 연결요소 조각 중에 크기와 밀도가 문자와 비슷한 것들이 있어 메디안 필터링을 이용하여 문자와 그림을 분류할 경우 그림 조각이 문자로 오분류 되는 경우가 있다. 이 문제를 해결하기 위하여 그림영역 내부의 문자영역을 모두 제거한 후 영역확장 필터링을 수행한다. 영역확장 필터링은 메디안 필터링과 처리과정이 같고 해당 화소값을 중간값 대신 최소값(밝기가 가장 어두운 값)으로 바꾼다는 점이 다르다. 따라서 흑화소 부분이 확장되는 효과가 있어 문자로 오분류된 그림을 구성하는 연결요소 대부분은 영역 확장 필터링 결과 확장된 흑화소 영역에 존재하므로 제거가 가능하다. (그림 4)의 경우를 보면 (a)는 입력영상을 제안한 방법



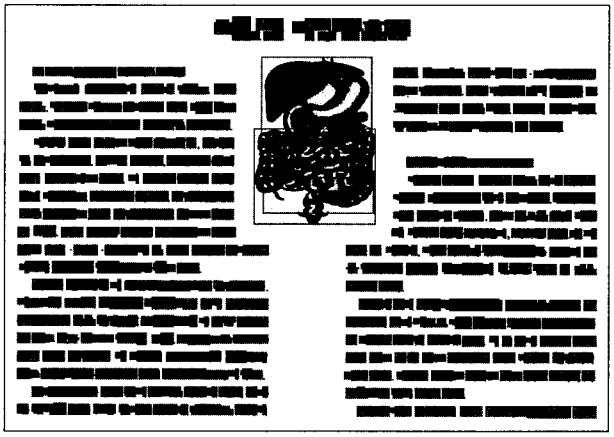
(a) 입력 영상(2037×1713 pixels 크기의 이진 영상)



(b) (a)영상을 3.2.2의 방법을 이용하여 영역분할 한 결과



(c) (a) 영상을 histogram 분포+separable 방법을 이용하여 메디안 필터링한 결과



(d) 메디안 필터링한 결과를 이용하여 문자와 그림영역을 분류한 결과

(그림 3) 제안한 메디안 필터링을 이용하여 문자와 그림영역을 분류한 예

으로 메디안 필터링 한 결과이고 (b)는 (a)의 결과를 이용하여 분할된 영역 중 문자영역(사각형 테두리 영역)만 분류한 결과이다. (b)의 결과를 보면 그림부분에 일부 그림조각 영역들이 문자영역으로 오판된 결과를 볼 수 있다. (c)는 (b) 영상에서 문자영역을 제거한 후 영역확장 필터링을 수행한 결과이다. (c)의 결과를 (a)의 메디안 필터링 결과와 비교해 보면 그림부분이 확장되어 있음을 확인 할 수 있다. (d)는 (b)에서 문자로 분류한 영역 중 (c)영상의 흑화소 부분에 포함되는 문자로 오분류된 그림조각들을 제거한 결과이다.

3.2.3.5 제한 조건을 이용한 오분류 영역 처리

제안한 영역분류에서 오분류의 유형을 보면 일부 'ㅣ', '—'와 같은 문자의 획이 직선으로 추출되거나 볼드체의 문자가 그림으로 오분류된 경우, 크기가 작고 밀도가 낮은 그래프나 그림이 문자로 오분류되는 경우가 발생한다. 따라서 문자의 최대크기와 문자들끼리 이웃하는 특성을 이용하여 오분류를 수정한다.

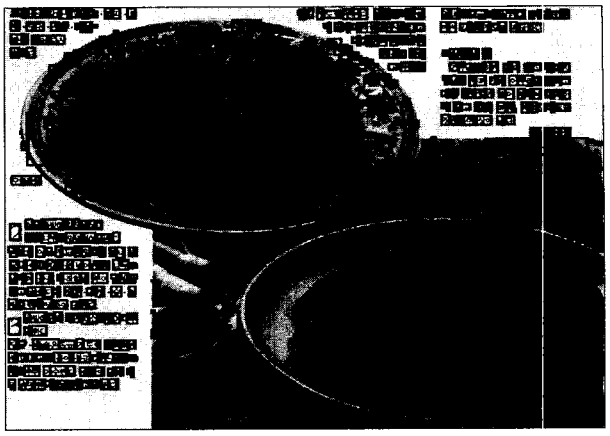
3.2.3.5.1 문서의 방향성 검사

문서에서 문자배열의 방향 결정은 메디안 필터링 결과 문자로 분류된 영역을 대상으로 한다. (그림 5)를 예로 설명하면, 먼저 각 기준영역마다 탐색구간을 설정한다. 탐색구간은 기준영역의 가로, 세로 길이 중 긴 쪽을 택하여 해당 길이만큼 기준영역의 상, 하, 좌, 우로 확장한 구간을 말한다. 탐색구간에서 기준영역과 수평으로 가장 가까이 인접한 비교영역과의 거리 (d_h)와 수직으로 가장 가까이 인접한 비교영역과의 거리 (d_v)를 비교하여 $d_h \le d_v$를 만족하는 개수가 반대인 경우 보다 더 많으면 문자배열을 가로방향으로 반대인 경우는 세로방향으로 설정한다.

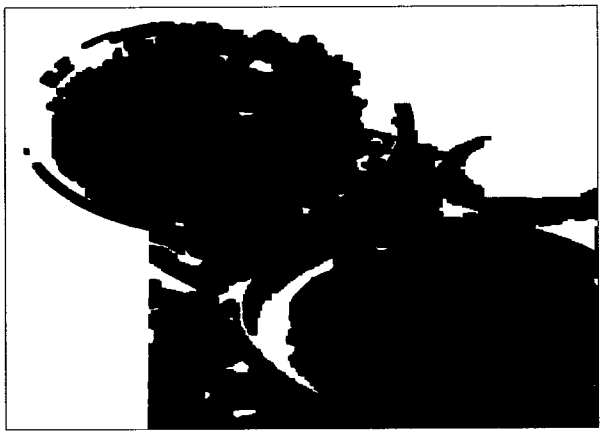
3.2.3.5.2 최대문자크기 (C_{max}) 결정

문자의 최대크기는 오분류된 문자, 직선, 그래프, 그림을 환원하는데 사용된다. 최대문자크기 (C_{max})는 문자배열 방향과 수직한 방향으로 문자영역의 길이 중 최대 길이를 C_{max}로 결정한다.

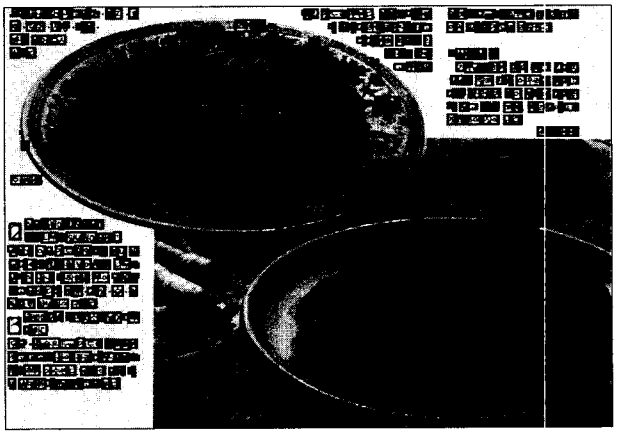
(a) 입력영상을 제안한 방법으로 메디안 필터링 한 결과



(b) (a)의 결과를 이용하여 분할된 영역 중 문자영역만 분류한 결과

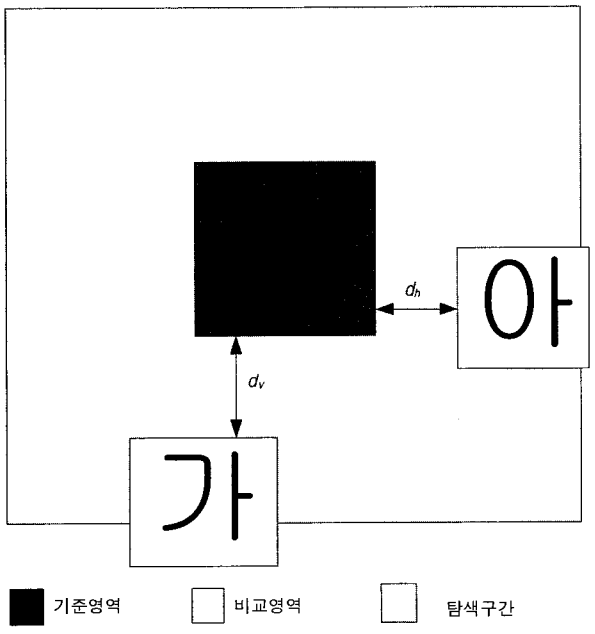


(c) (b)영상에서 문자영역을 제거한 후 영역확장 필터링을 수행한 결과



(d) 문자영역 중 (c)영상의 흑화소 부분에 포함되는 영역을 제거한 결과

(그림 4) 영역확장 필터를 이용한 그림영역 내부 오분류 문자영역 제거



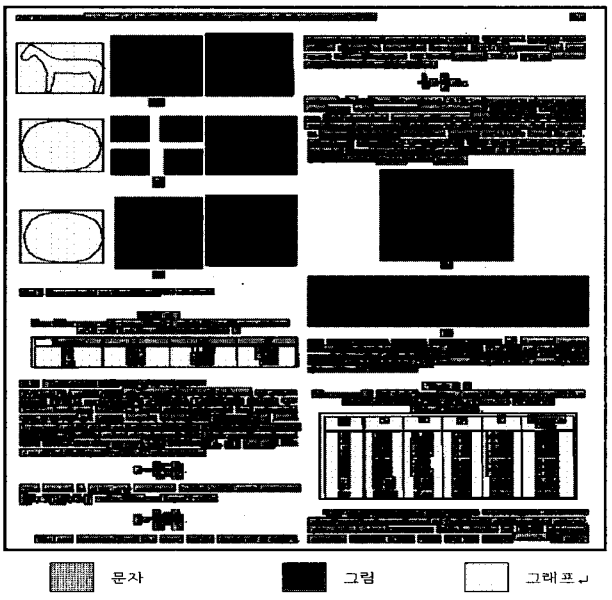
(그림 5) 인접 문자영역 탐색 예

3.2.3.5.3 오분류된 직선, 그래프 처리

3.2.3.1에서 추출한 직선 중에는 'ㅣ', 'ㅡ'와 같은 문자의 획도 포함되어 있다. 또한 3.2.3.2에서 내부 데이터를 제외한 영역의 흑화소 밀도가 25%이하라는 조건을 이용하여 그래프 영역을 추출하였으므로 이 조건을 만족하는 문자가 그래프영역으로 추출되는 경우가 있다. 따라서 문자배열 방향과 수직한 방향으로의 영역의 길이가 C_{max}이하인 직선과 그래프는 문자로 환원한다.

3.2.3.5.4 오분류된 볼드체 처리

메디안 필터링을 이용하여 그림으로 분류한 영역 중에는 다른 문자에 비해 상대적으로 크기가 크고 흑화소 밀도가 높은 볼드체의 문자열이 오분류된 경우가 있다. 이러한 오분류된 볼드체 문자열은 문자열 방향으로 탐색거리 $(2\times C_{max})$)내에 인접하는 문자가 있는 경우 해당영역을 문자로 수정한다. (그림 6)은 제안한 방법을 이용한 문서영상 구조분석을 수행하여 문자, 그림, 그래프의 세부영역으로 분류한 결과 예이다.



(그림 6) 제안한 방법을 이용한 문서영상 구조분석 실행 결과 예

4. 실험 및 고찰

제안한 문서영상 구조분석 방법은 Xingyuan Li's 방법 [1]과 현재 판매되고 있는 상용제품 3종류와 성능비교를 하였으며 성능평가에서 주관적 판단부분이 많은 영역분할보다는 영역분류의 정확성을 기준으로 실험하였다. 문서영상은 사용하는 언어와 문서구조가 다양하여 단일화된 시험영상을 생성하기 어렵다. 따라서 본 실험에서는 자체적으로 마련한 문서영상 40장을 대상으로 실험하였다. 대상 문서는 신문, 잡지, 논문, 서류, 영수증 등 다양한 종류에서 주로 영역분할이 어려운 형태로 되어 있는 문서들 위주로 선별하였으므로 기존의 방법이나 상용제품의 성능평가 결과가 해당 제품에서 발표한 것보다 낮다는 것을 밝혀둔다.

실험에서는 대상이 되는 문서들을 먼저 수(手)작업으로 영역분할 및 영역분류를 수행한 다음 상기된 방법들을 수행한 결과와 비교함으로써 각 방법의 성능을 평가하였다.

〈표 1〉 메디안 필터링 수행시간 비교(second)

문서영상 이름	크기	2차원 메디안 필터링		histogram 분포 이용 separable 방식
		quick sort 이용	histogram 분포 이용	
paper.bmp	676 * 985	21.89	2.06	0.61
XML.bmp	546 * 920	19.66	1.95	0.52
desktop.bmp	988 * 836	28.58	2.65	0.71

<표 1>은 quick sort, 히스토그램 분포, separable 방법을 이용한 메디안 필터링의 수행시간을 기록한 결과이다. 실험에 사용한 영상은 256 밝기단계의 그레이 영상이며 메디안 필터의 탭 크기는 19이고 Pentium-4 1.7GHz microprocessor, RAM 384MB의 PC를 사용하여 실험하였다.

<표 2>는 제안한 영역분류 방법을 상용제품인 P사의 A 6.0 Pro (국내제품), Scansoft사의 Omni page pro 11.0, ABBYY Software House사의 Fine Reader 5.0 Office와 영역분류에서의 인식영역 수를 비교한 결과이다.

〈표 2〉 영역분류에서 상용제품과의 인식 영역 수 비교(개)

제품 / 형태	수(手) 작업	A 6.0 Pro	Omni page pro 11	Fine Reader 5.0	제안한 방법
문 자	392	325	313	311	389
그 림	121	80	78	70	111
그래프	43	38	35	22	38
총 계	556	443	426	403	538

<표 3>은 <표 2>에서 수(手) 작업으로 분류한 영역 수를 분모로 하고 상용제품과 제안한 방법에서 인식영역 수를 분자로 하여 구한 인식률 비교 결과이다.

〈표 3〉 영역분류에서 상용제품과의 인식률 비교(%)

제품 / 형태	A 6.0 Pro	Omni page pro 11	Fine Reader 5.0	제안한 방법
문 자	83	80	79	99
그 림	66	64	58	92
그래프	88	81	51	88
총 계	80	77	72	97

<표 4>는 가장 성능이 우수한 영역분할 방법 중의 하나인 Xingyuan Li's 방법 [1]과의 비교결과이다. 인식률은 수(手) 작업으로 분류한 영역 수를 분모로 한 비율이다.

〈표 4〉 Xingyuan Li's 방법과의 영역분류 성능 비교

제품 / 형태	수(手) 작업	Xingyuan Li's method (Form Reader 0.9)		제안한 방법	
	인식 영역 수 (개)	인식 영역 수 (개)	인식률 (%)	인식 영역 수 (개)	인식률 (%)
문 자	392	376	**96**	389	**99**
그 림	164	141	**86**	149	**91**
총 계	556	517	**93**	538	**97**

<표 5>는 제안한 문서영상 구조분석 방법을 이용하여 문자, 그림, 그래프, 직선 4가지로 분류한 경우의 성능평가 결과이다.

〈표 5〉 제안한 문서영상 구조분석 방법을 이용 문자, 그림, 그래프, 직선 4가지로 분류한 경우의 성능평가

Li의 제품 / 형태	수(手) 작업	제안한 방법	
	인식 영역 수(개)	인식 영역 수(개)	인식률(%)
문 자	404	396	**98**
그 림	121	111	**92**
그래프	32	29	**91**
직 선	29	26	**90**
총 계	586	562	**96**

5. 결 론

본 논문에서는 문서영상 영역분류 과정에서 문서영상의 특성을 고려한 고속 메디안 필터를 사용하여 문자와 그림을 분리한다. 또한 문자, 그림, 그래프, 직선의 영역분류 과정에서 발생한 오분류를 영역확장 필터와 제한조건들을 이용하여 수정하는 문서영상 구조분석 방법을 제안하였다.

영역분류 성능시험에서는 상용제품을 포함한 Xingyuan Li's 방법 [1]과 비교를 통하여 성능을 평가한 결과 평균 인식률이 96~97%로 방법 [1]과 상용제품보다 우수함을 확인할 수 있었다. 또한 제안한 문서영상 구조 분석 방법에서는 표, 차트, 그래프 영역의 내부에 포함된 문자도 추출함으로써 보다 상세한 구조분석이 가능하였다.

제안한 문서영상 구조분석 방법이 인간이 수행하는 것과 같은 수준의 성능을 발휘하기 위하여 보완해야 할 점들을 보면, 영역분할에서 연결요소 기반 상향식 방법을 사용함으로써 많은 양의 버퍼가 필요하다는 문제점과, 그림영역 내부의 일부 그림조각 영역이 문자로 오판되는 것을 해결하기 위하여 더욱 정확한 문자와 그림의 판단 기준이 필요하며, 점선으로 된 직선을 추출하는 문제에 대한 대책이 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] K. Kise, M. Iwata and K. Matsumoto, "A Computational Geometric Approach to Text-line Extraction from Binary Document Images," *Proc. 3th Int. Work. Document Analysis System,* pp.346-355, 1998.

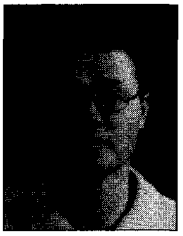
[4] H. Fujisawa and Y. Nakano, "A Top-Down Approach for the Analysis of Document Images," *Proc. Work. Syntatic and Structural Pattern Recognition, Murray Hill, USA,* pp.113-122, 1990.

[5] Y.Y. Tang, C.Y. Suen, C.D. Yan and M. Cheriet, "Document Analysis and Understang : A Brief Survey," *Proc. 1st Int. Conf. Document Analysis and Recognition, Saint-Malo, France,* pp.17-31, 1991.

[6] 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.

[7] 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.

[8] Mario I. Chacon Murguia, "Document Segmentation Using Texture Variance and Low Resolution Images," *IEEE Southwest. Symp. Image Analysis and Interpretation,* pp. 164-167, 1998.



장 대 근

e-mail : ssendol@palgong.knu.ac.kr

1997년 경북대학교 전자공학과(공학석사)

1998년~현재 경북대학교 전자·전기·컴퓨터학부 박사과정

1997년~현재 한국전자통신연구원 컴퓨터소프트웨어기술연구소 연구원

관심분야 : 문서영상처리, 영상부호화 및 압축, 인공지능

황 찬 식

e-mail : cshwang@ee.knu.ac.kr

1979년 한국과학기술원 전자공학과(공학석사)

1996년 한국과학기술원 전자공학과(공학박사)

1979~현재 경북대학교 전자·전기·컴퓨터학부 교수

관심분야 : 영상통신, 암호통신, 초고속망