연결요소 방법과 메디안 필터를 이용한 문서영상 기하학적 구조분석

정회원 장 대 근*, 황 찬 식*

The Geometric Layout Analysis of the Document Image Using Connected Components Method and Median Filter

Dae-geun Jang*, Chan-sik Hwang* ***Regular Members***

요 약

인쇄문서를 전자문서로 자동전환하기 위해서는 먼저 문서영상의 기하학적 구조를 분석하여 문자, 그림, 표 등의 세부 영역으로 분류해야한다. 그러나 문서구조의 복잡성과 그림의 크기와 밀도의 다양함은 기하학적 구조분석을 어렵게 만드는 원인이 되고 있다.

본 논문에서는 연결요소 기반의 방법을 이용하여 복잡한 구조의 문서도 세부적 영역분할이 가능하며, separable 메디안 필터를 이용하여 크기와 밀도가 다양한 문자와 그림을 분류하고, 1차원 메디안 필터를 수평, 수직방향으로 각각 적용하여 표를 구성하는 직선이 훼손되거나 직선에 문자가 붙어있는 경우에도 추출을 가능하게 함으로써, 상용제품이나 기존의 방법에 비해 영역분할 및 분류 그리고 표를 구성하는 직선추출이 우수한 방법을 제안한다.

ABSTRACT

Document image should be classified into detailed regions as text, picture, table and etc through the geometric layout analysis if paper documents can be converted automatically into electronic documents. However, complexity of the document layout and variety of the size and density of a picture are the reason to make it difficult to analyze the geometric layout of the document images.

In this paper, we propose the method which have a better performance of the region segmentation and classification, and the line extraction in the table region than the commercial softwares and previous methods. The proposed method can segment the document into detailed regions by using connected components method even if its layout is complex. This method also classifies texts and pictures by using separable median filter even though their size and density are diverse, In addition, this method extracts the lines from the table adapting one dimensional median filter to the each horizontal and vertical direction, even though lines are deformed or texts attached to them.

Ⅰ. 서 론

PC의 보급과 워드프로세스의 개발로 인해 전자문서의 작성과 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 있으며 이와 같은 기능을 수행하는 소프트웨어 제품과 방법들이 국내외적으로 개발되어 왔다. 그러나 이러한 제품이나 방법들이 워드프로세

* 경북대학교 전자·전기·컴퓨터학부 데이터 통신 시스템 연구실(ssendol@palgong.knu.ac.kr)
논문번호 : 020274-0615, 접수일자 : 2002년 6월 15일

스 소프트웨어에 비해 현재 널리 사용되지 못하는 이유는 문서영상을 세부영역으로 분할하고 분할된 영역을 문자, 그림, 표 등의 영역으로 분류하는 문서영상 기하학적 구조분석의 성능이 인간이 수행하는 수준에 미치지 못하기 때문이다. 특히 문자, 그림, 표와 같은 영역분류과정에서의 오분류는 문서 본연의 형태와는 다른 형태의 전자문서를 만들어 낸다.

영역분류 방법으로는 문자를 구성하는 연결요소의 크기와 밀도는 그림의 경우와 다르다는 특성을 이용하는 방법 [2]와 문자열 부분의 흑화소 분포를 수치화하는 방법으로 cross correlation approach [3], Kolmogorov complexity measure [4], texture pattern analysis [5] 등이 있다. 그러나 기존의 [2]-[5]의 방법으로는 그림의 경우 연결요소의 크기와 밀도가 다양하고 흑화소 분포의 범위가 넓어 문자와 구분할 수 있는 기준을 정하기 어렵다. 또한 문자의 크기도 다양하여 상대적으로 크고 밀도가 높은 문자들이 그림으로 분류되는 경향이 있다.

본 논문에서는 문자제거 효과가 우수한 메디안 필터와 영역 내부 흑화소 밀도, 영역간 포함관계를 이용하여 크기와 밀도가 다양한 그림과 문자을 분류함으로써 크기가 작고 밀도가 낮은 그림이 문자로 오분류되거나 상대적으로 크기가 크고 밀도가 높은 문자들이 그림으로 오분류되는 것을 방지한다. 그리고 표영역분류에서 1차원 메디안 필터를 수평과 수직방향으로 각각 적용하여 필터방향의 직선을 추출함으로써 표를 구성하는 직선이 끊어지거나 직선에 노이즈나 문자가 붙어 있는 경우에도 추출이 가능하다. 따라서 제안한 방법은 문서영상 기하학적 구조분석을 어렵게 만드는 문제점들을 극복함으로써 기존 방법을 포함한 국내외 상용제품보다 성능이 우수하다.

Ⅱ. 기존 문서영상 기하학적 구조분석 방법 분석

문서영상 기하학적 구조분석은 문서영상을 세부영역으로 분할하는 방법과 분할된 영역을 문자, 그림, 표 등으로 분류하는 방법을 필요로 하며 각각에 대한 설명은 다음과 같다.

1. 영역분할

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 연결요소와, 인접선분밀도를 이용하는 방법이 있다.

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)과 런 길이 평활화(run length smoothing)를 이용하는 방법이 있다.

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.

2.1. 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. 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 correlation(Cr)은 다음 수

식 (1)을 이용하여 계산하며 M과 N은 각각 가로방향과 세로방향 화소수이고 ⊕은 exclusive OR 연산을 의미한다.

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

2.3. Kolmogorov complexity measure

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

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

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

$c(n)$: complexity for finite strings of length

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.

Ⅲ. 제안한 문서영상 기하학적 구조분석 방법의 구성 및 처리과정

1. 전체 구성

제안한 문서영상 기하학적 구조분석 방법은 그림 1과 같이 전처리, 영역분할, 영역분류의 3단계 과정으로 구성된다. 전처리로는 노이즈 제거, 문서 기울기 보정, 영상이진화, 영상축소 등이 있으나 제안한 방법에서는 전처리로 256단계의 그레이영상을 이진영상으로 변환하는 과정과 실시간 처리를 위하여 영상을 축소하는 과정만 사용한다. 영역분할은 연결요소기반 상향식 방법을 고속화한 Xingyuan Li's 방법 [2]를 사용한다. 영역분류에서는문자와 그림을 분류한 후 표와 표를 구성하는 직선과 데이터를 분류함으로써 문서영상을 문자, 그림, 표의 3가지 세 부영역으로 분류한다.

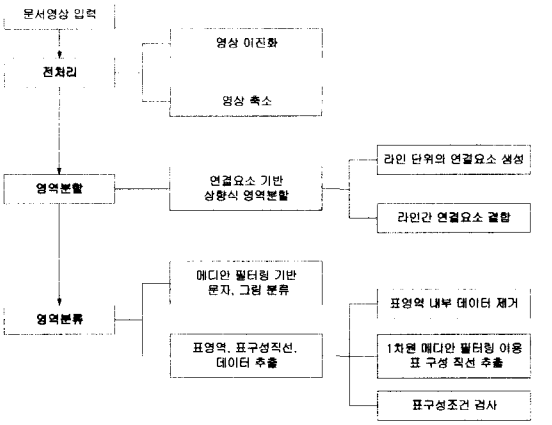


그림 1. 제안한 문서영상 기하학적 구조분석 방법의 전체 구성도

2. 처리과정

2.1. 전처리

제안한 방법에서는 스캐너로 입력한 300dpi 해상도 또는 디지털 카메라로 찍은 1000×1000 pixels 이상의 해상도를 갖는 256 그레이 문서영상 또는 이진 문서영상을 대상으로 기하학적 구조분석을 수행한다. 그러나 실제적으로 이진 문서영상의 경우만 처리가 가능하므로 256 그레이 문서영상의 경우 전처리과정에서 N. Otsu의 방법 [1]을 이용하여 이진영상으로 변환한 다음 구조분석을 수행한다.

또한 PC를 이용하여 실시간으로 기학적 구조분석을 수행하기 위해 입력영상을 축소하여 처리한다. 축소비율()은 디지털 카메라를 이용한 영상입력을 고려하여 조정한다. 해상도가 너무 낮은 경우 표를 구성하는 선과 데이터가 붙어 회손되며 너무 높은 경우 실시간 처리가 어려우므로 1000×1000 pixels 보다 약간 작은 크기로 축소되도록 아래의 수식 (4)에서 입력영상의 가로, 세로 길이 중 큰 값 ($MaxLen$)을 분모 값 1000으로 나누어 반올림한다. $Q[\alpha]$는 α 를 반올림하여 정수화하는 연산을 의미한다.

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

영상축소는 입력영상을 크기의 영역으로 겹치지 않게 분할한 후 분할영역 내부에 흑화소가 한개 이상 존재하는 경우는 영역을 1개의 흑화소로 표현

하고 나머지는 1개의 백화소로 표현하여 영상을 축소한다.]

2.2. 영역분할

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

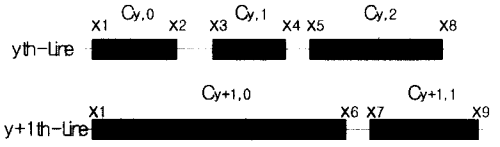


그림 2. 인접한 라인에서 연결요소 예

$$\min[\max(C_{y,m}),\max(C_{y+1,n})] \geq \max[\min(C_{y,m}),\min(C_{y+1,n})] \quad (5)$$

example> $X_1=\min(C_{y,0})$, $X_2=\max(C_{y,0})$

2.3. 영역분류

2.3.1. 그림, 문자분류

2.3.1.1. 메디안 필터링의 적용

제안한 방법에서는 메디안 필터링을 수행하여 사진과 같은 흑화소 밀도가 높은 그림을 분류한다. 메디안 필터는 자신의 화소와 주변의 화소값을 포함하는 필터탭 내부의 화소값 중 중간값을 택하는 필터링으로 impulse noise 제거에 탁월한 효과가 있다. 그림 3은 1차원 메디안 필터링의 예로 자기 자신을 포함하여 좌우로 2개씩의 pixels을 포함하는 탭 크기가 5인 경우이다. 5개의 pixel 값은 화소의 밝기이며 가운데 있는 자신의 화소값 255를 주변의 화소값을 포함한 5개 화소값의 중간값 128로 바꿈으로써 주변값과 차이가 많은 impulse noise를 제거한다.

그림 3. 1차원 메디안 필터링 예

이 원리를 문서영상에 적용하면 문자를 구성하는 흑화소는 제거되고 밀도가 높은 사진과 같은 그림 부분의 흑화소는 남게 되어 문자와 그림의 분리가 가능하다. 따라서 메디안 필터링 결과 제거되지 않고 남아 있는 흑화소 부분을 포함하는 분할 영역은 그림영역으로 나머지는 문자영역으로 분류한다. 필터링에서의 탭 크기는 문자는 제거되고 그림은 제거되지 않을 정도의 크기로 정해주어야 하므로 다양한 입력문서의 해상도와 문자 크기에 맞게 정하기 위하여 2.2.에서 분할한 영역들의 평균길이 ($L_{avg}(R)$)의 2배 크기로 설정한다.

$L_{avg}(R)$은 분할된 각 영역의 길이를 평균한 값이며 수식 (6)과 같다.

$$L_{avg}(R)=\frac{\sum_{i=0}^{n_R-1} L(R_i)}{n_R} \quad (6)$$

$L_{avg}(R)$: 분할 영역들의 평균길이

$L(R_i)$: index i 영역의 길이

n_R : 분할 영역의 총 수

index i 영역의 길이($L(R_i)$)는 문자의 배열이 가로방향인 문서는 영역의 세로길이를, 문자의 배열이 세로방향인 문서는 영역의 가로길이를 사용한다. 문서에서 문자배열의 방향 결정은 메디안 필터링 결과 문자로 분류된 영역을 대상으로 한다. 그림 4를 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) ≤ (d_v) 를 만족하는 개수가 반대인 경우 보다 더 많으면 문자배열을 가로방향으로 반대인 경우는 세로방향으로 설정한다.

2.3.1.2. 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

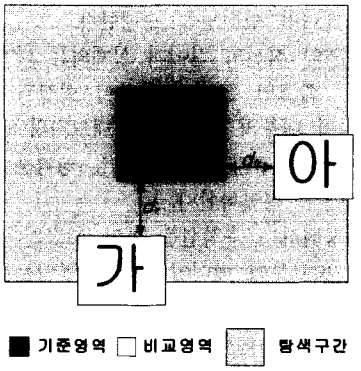


그림 4. 인접 문자영역 탐색 예

방식의 필터링을 수행함으로써 필터링을 실시간으로 수행한다.

2.3.1.3. 초기 문자영역 분류

메디안 필터링에 의해 그림으로 분류된 영역은 사진과 같은 밀도가 높은 영역이므로 상대적으로 밀고가 낮은 차트, 그래프와 같은 그림과 표는 메디안 필터링에 의해 문자로 분류된다. 따라서 메디안 필터링에 의해 문자로 분류된 영역 중 다음의 2가지 조건을 이용하여 문자로서 확률이 낮은 영역들을 제외함으로써 확실한 문자영역들만 초기 영역으로 분리한다. 첫째, 많은 문서영상들을 대상으로 실험한 결과 그래프, 차트, 표 영역은 내부 데이터를 제외한 분할영역의 흑화소 밀도가 대부분 25% 이하이며 이 조건을 만족하는 분할영역들을 원소로 하는 집합 S_d를 수식 (7)에 표현하였다.

$$S_d = \{ R_i \in R \mid D_e(R_i) \leq 0.25 \} \quad (7)$$

R : 분할영역

R_i : R 에 속하는 index i 분할영역

$D_e(R_i)$: 내부에 포함된 영역들을 제거한 영역 R_i의 흑화소 밀도

둘째, 영역분할 결과 문자는 대부분 내부에 다시 영역을 포함하지 않으며 이 조건을 만족하는 영역들을 원소로 하는 집합을 S_s 로, 여집합을 S_s^C로 표현한다.

초기 문자영역은 메디안 필터링에 의해 문자로 분류된 영역 중 집합 ($S_d \cup S_s^C$)에 속하는 영역들을 제외한 영역이며, 수식 (8)에서 초기 문자영역을 원소로 하는 집합 S_t를 표현하였다.

$$S_t = \{ R_i \mid R_i \in (S_m - (S_d \cup S_s^C)) \} \quad (8)$$

2.3.1.4. 문자영역 확장 및 병합

일반적으로 영역분할은 상세한 부분까지 분할하기 위하여 Ⅱ의 1.1.에서 언급한 상향식 영역분할을 많이 사용한다. 그러나 작은 요소들을 결합하여 영역을 형성하는 과정에서 요소들을 문자와 비문자로 분류하기 어렵다. 따라서 구조가 복잡한 문서의 경우 이런 요소들을 결합한 영역 중에는 문자와 그림이 섞여있거나 오분류된 경우가 발생한다. 그러나 제안한 방법에서는 확실한 문자들을 먼저 분리한 후 이 문자들을 seeds로 하여 주변에 조건이 비슷한 영역들을 병합함으로써 Xingyuan Li's 방법 [2]를 포함한 기존의 상향식 방법에 비해 문자영역 분류에서의 오류를 감소시켰다.

영역 확장에 필요한 요소들을 보면, 수식 (8)에서 분류한 문자영역의 가로, 세로방향 길이 중 2.3.1.1.에서 구한 문자배열 방향과 수직한 방향으로의 길이를 해당 영역의 크기로 정의하고 이 중 가장 큰 크기를 최대문자크기($C_{\max}$)로 설정한다. 또한 $C_{\max}$에 해당하는 영역의 상, 하, 좌, 우로 인접하며 크기가 1.2배 이상 차이나지 않는 영역과의 거리 중 가장 긴 거리를 $d_{\max}$로 설정하며 이 거리는 영역확장 폭의 최대값이 된다.

영역확장은 먼저 (8)에서 분류한 문자영역을 대상으로 각 문자영역의 상, 하, 좌, 우로 영역 크기의 1/2 거리내에 있으며, 해당 문자영역과의 크기 비가 1.2 이하인 영역들을 선별하여 포함시킨다. 선별된 영역들을 대상으로 이 과정을 다시 적용하여 주변의 영역들을 포함시키며 더 이상 포함할 영역이 없을때 까지 반복 한다. 이 과정에서 선별된 영역과 초기 문자영역을 수식 (9)의 크기로 인접한 영역이 있는 방향으로만 확장하여 겹쳐지는 부분이 있는 영역들을 병합함으로써 문자영역을 완성한다. 수식에서 $L_e(R_i)$는 R_i 영역의 확장크기이고 $L_e(R_i)$는 영역 R_i의 가로, 세로방향 길이 중 문자배열 방향과 수직한 방향으로의 길이 즉 영역 R_i 의 크기를 의미한다.

$$L_e(R_i) = d_{\max} \cdot \frac{L(R_i)}{C_{\max}} \quad (9)$$

그림 5는 제안한 방법을 이용하여 영역분할 후 분할영역을 문자와 그림영역으로 분류한 예이다. 문자영역의 경우 초기 문자영역(음영이 있는 영역) 및 확장 후 병합한 영역을 직사각형 영역으로 표시하였다.

그림 5. 문자, 그림영역 분류 예

2.3.2. 표분류

2.2.에서 영역분할 결과 표를 구성하는 수평선과 수직선은 하나의 연결요소로 추출되며 이들을 분리하기위한 방법으로 Xingyuan Li's 방법 [6]와 같은 morphological filters를 이용한 방법이 사용된다. 그러나 이 방법은 끊어진 직선을 어느 정도 연결하면서 추출이 가능하지만 직선에 문자가 붙어있는 경우 직선과 문자의 분리가 어렵다. 따라서 이들을 분리하기 위하여 Jain-Shiue Chen 's 방법 [7]과 같은 방법의 추가가 필요하며 이러한 방법의 추가는 계산량의 증가를 가져오는 문제점이 있다. 따라서 제안한 방법에서는 1차원 메디안 필터를 수평, 수직방향으로 각각 적용하여 해당방향의 직선을 추출한다. 메디안 필터는 중간값을 택하는 특성으로 인해 별도의 추가과정이 필요 없이 이진영상의 경우 끊어진 직선을 이어주고 직선에 붙어있는 문자를 분리하면서 필터링 방향의 직선을 추출하는 장점이 있다.

2.3.2.1. 표영역 추출을 위한 후보영역

2.3.1.3.과 2.3.1.4.에서 언급한 $S_d, S_s^C, L(R_i), C_{\max}$ 를 이용하여 수식 (10)에 표 후보영역을 원소로 하는 집합 S_c를 구하는 것을 표현하였다. 즉 후보 영역은 내부에 데이터 영역을 포함하며 데이터를 제외한 영역의 흑화소 밀도가 25%를 넘지 않는다. 또한 영역의 크기는 최대문자크기의 2배보다 커야한다.

$$S_c = \{ R_i | R_i \in (S_d \cap S_s^C) \cap L(R_i) \geq 2 \times C_{\max} \} \quad (10)$$

2.3.2.2. 1차원 메디안 필터를 이용한 표 구성 직선 추출

후보영역 내부의 문자들을 백화소로 치환하여 제거한 후 메디안 필터링을 수행한다. 그리고 필터탭이 너무 크면 직선의 길이가 실제와는 다르게 추출되며 너무 짧으면 끊어진 부분이 연결되지 않으므로 해당영역 내부의 문자 중 최대크기의 1/2 로 설정한다. 필터링 후에는 해당영역을 영역분할하여 추출한 직선들을 영역화한다.

추출과정에서는 수직선을 먼저 추출한다. 그림 6의 (c)는 (b)로부터 메디안 필터와 영역분할을 이용하여 수직선을 추출한 결과로 끊어진 부분은 연결되고 직선에 붙어있던 문자가 분리되어 추출됨을 확인할 수 있다. 추출한 직선들은 수식 (11)의 검사를 이용하여 수직방향으로 교차관계에 있는 영역 내부의 모든 문자들보다 길이가 긴 경우만 표를 구성하는 직선으로 설정하고 나머지는 제거함으로써 직선에 붙은 문자를 분리한다. 즉 수식 (11)을 만족하는 직선 l_k 는 직선에 붙어있는 문자부분이므로 제거한다. (d)는 수직선만 추출한 결과이다.

$$\exists\, r_b, l_k \subset R_i \rightarrow$$
$$(\min[ey(l_k), ey(r_i)] \geq \max[sy(l_k), sy(r_i)]) \cap$$
$$(\ L(l\_k\ ) \leq L(r\_b\ )) \quad (11)$$

r_i : 영역 내부에 존재하는 index 의 문자영역
l_k : 영역 에서 추출한 index 의 직선영역
$\max[a, b]$: a,b 중 최대값을 선택
$\min[a, b]$: a,b 중 최소값을 선택
$sy(a)$: 영역 의 y방향 시작 좌표
$ey(a)$: 영역 의 y방향 끝 좌표
$L(a)$: 영역 의 세로방향 길이

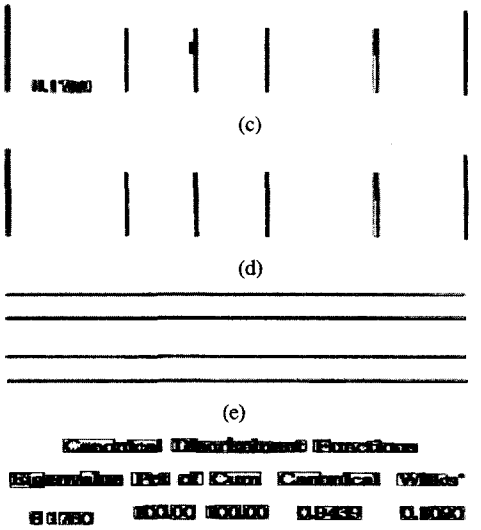
수평선 추출은 먼저 제거된 수직선 부분을 문자영역으로 설정후 백화소로 치환한다음 메디안 필터와 영역분할을 이용하여 직선영역을 추출한다. 그림 6의 (e)는 그 결과이며 (f)는 데이터(문자)를 추출한 결과로 직선에 붙어있던 문자들이 추출됨을 확인할 수 있다.

Canonical Discriminant Functions				
Eigenvalue	Pct of	Cum	Canonical	Wilks'
8.1760	100.00	100.00	0.9439	0.1090

(a)

	1			
8.1760				

(b)



(a) 입력 영상
(b) 영역 내부 데이터를 제거 후 표 구성 직선들을 하나의 연결요소로 추출한 결과
(c) 1차원 메디안 필터와 영역분할을 이용한 수직선 추출 결과
(d) 선에 붙어있던 문자들이 수직선으로 추출된 경우를 제거한 결과
(e) 1차원 메디안 필터와 영역분할을 이용한 수평선 추출 결과
(f) 표 구성 데이터(문자)추출 결과

그림 6. 1차원 메디안 필터를 이용한 표 구성 수평선, 수직선, 데이터 추출 결과 예

2.3.2.3. 표 구성조건 검사

표의 형태는 다양하므로 인간의 시각으로 표를 판별하는 경우에도 구분이 어려운 경우가 있으며 그래프, 차트와 같은 그림 중 일부는 표와 구조가 비슷하여 표로 오판되는 경우가 있다. 따라서 표를 구분하기 위해서는 먼저 표의 범주를 명확히 정의하여야 할 필요가 있으며 제안한 방법에서는 표 전체를 관통하는 수평선과 수직선이 각각 3개 이상, 최외각을 구성하는 4개의 직선 중 3개 이상 존재하는 경우만 표로 정의하고 나머지는 차트나 그래프 형태의 그림으로 분류한다.

Ⅳ. 실험 및 고찰

제안한 문서영상 기하학적 구조분석 방법은 Xingyuan Li's 방법 [2], [6]과 현재 판매되고 있는 상용제품 3종류와 성능비교를 하였으며 성능평가에서 주관적 판단부분이 많은 영역분할보다는 영역분류와 표를 구성하는 직선추출의 정확성을 기준으로 실험하였다.

문서영상의 경우 사용하는 언어와 문서구조가 다양하여 단일화된 시험영상을 생성하기 어렵다. 따라서 본 실험에서는 자체적으로 마련한 문서영상 40장을 대상으로 실험 하였다. 대상 문서는 신문, 잡지, 논문, 서류, 영수증 등 다양한 종류에서 주로 기하학적 구조분석이 어려운 형태로 되어 있는 문서들 위주로 선별하였으므로 Xingyuan Li's 방법 [2], [6]과 상용제품의 성능 평가결과가 해당방법이나 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. 그리고 실험 결과는 문자, 그림, 표분류의 성능과 표 구성 직선 추출 성능을 기록하였다.

표 1은 제안한 영역분류 방법을 상용제품인 P사의 A 6.0 (국내제품), Scansoft사의 Omni page pro 11.0, ABBYY Software House사의 Fine Reader 5.0 Office, Xingyuan Li's 방법 [2], [6]과 영역분류에서의 문자, 그림, 표의 인식 영역 수를 비교한 결과이다.

표 1. 문자, 그림, 표분류에서의 인식 영역 수(개)

제품 / 형태	수(手) 작업	A 6.0	omni page pro 11	Fine Reader 5.0	X. Li's method	제안한 방법
문자	392	325	313	311	376	389
그림	121	80	78	70	98	111
표	43	38	35	22	40	40
총계	556	443	426	403	514	540

표 2는 표 1에서 수(手) 작업으로 분류한 영역수를 분모로 하고 인식한 영역 수를 분자로 하여 구한 인식률 비교 결과이다.

표 2. 문자, 그림, 표분류에서의 인식률(%)

제품 / 형태	A 6.0	omni page pro 11	Fine Reader 5.0	X. Li's method	제안한 방법
문자	83	80	79	96	99
그림	66	64	58	81	92
표	88	81	51	93	93
총계	80	77	72	92	97

표 3은 총 43개의 표를 구성하는 563개의 직선(수평선+수직선)추출에서의 성능평가 결과이다.

표 3. 표를 구성하는 직선(수평선+수직선)추출에서의 성능평가

제품 / 구분	A 6.0	omni page pro 11	Fine Reader 5.0	X. Li's method	제안한 방법
추출 직선 수(개)(수평선+수직선)	526	495	512	535	557
인식률(%)	93	88	91	95	99

V. 결 론

본 논문에서는 separable 메디안 필터링 결과를 기반으로 분할된 영역 내부의 흑화소 밀도, 영역 사이의 포함관계를 이용하여 크기와 밀도가 다양한 문자와 그림을 분류하고, 1차원 메디안 필터를 수평과 수직방향으로 각각 적용하여 필터방향의 직선을 추출함으로써 표를 구성하는 직선이 끊어지거나 직선에 노이즈나 문자가 붙어 있는 경우에도 추출이 가능하여 기존의 방법 Xingyuan Li's 방법 [1], [6]과 상용제품에 비해 문서영상 기하학적 구조분석 성능이 우수한 방법을 제안하였다.

영역분류 성능시험에서는 문자, 그림, 표 3가지로 분류시험을 수행한 결과 평균 인식률이 97%로 Xingyuan Li's 방법 [1], [6]과 상용제품들보다 우수함을 확인 할 수 있었다. 특히 문자와 그림의 분류에서 인식률이 각각 99%, 92%로 이전의 방법들보다 우수하여 전체 인식률의 상승을 주도함을 확인할 수 있다.

제안한 문서영상 기하학적 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 필요하다. 또한 표 구조에서 사선을 처리하는 문제와 점선으로 된 직선을 추출하는 문제 등에 대한 개선책이 필요하다.

참 고 문 헌

[1] N. Otsu, "A Threshold Selection Method From Gray-level Histograms," IEEE Trans. Systems, Man, and Cybernetics, vol. SMC-9, No.1, pp. 62-66, 1979.

[2] 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.

[3] 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.

[4] 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.

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

[6] X. Li, J. Hong, Z. Zhang and B. Chen, "A Statistical Form Reading System," Proc. IEEE Region 10 Conf. Computer, Communication, Control and Power Engineering, vol.2 pp.1062-1065, 1993.

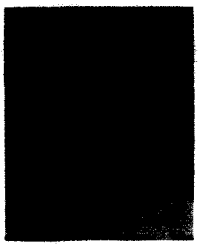
[7] Jain-Shiue Chen and Din-Chang Tseng, "Overlapped Charter Separation and Reconstruction for Table form Documents," Proc. Int. Conf. Image Processing, vol.1, pp.233-236, 1996.

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

[9] 장 명욱, 천 대녕, 양 현승, "연결화소를 이용한 문서영상의 분할 및 인식," 한국정보과학회 논문지, 제 20권, 제 12호, pp.1741-1751, 1993.

[10] 이 인동, 권 오석, 김 태균, "문서 영상에서 문자와 비문자의 분리추출 방법," 한국정보과학회 논문지, 제17권 제 3호, pp.247-258, 1990.

장 대 근(Dae-geun Jang) 정회원



e-mail : ssendol@plgong.knu.ac.kr
1998년 : 경북대학교 전자 · 전기 · 컴퓨터 학부 박사 과정
1997년 ~ 현재 : 한국전자통신 연구원 연구원

<주관심 분야> 문서영상처리, 영상압축, 인공지능

황 찬 식(Chan-sik Hwang) 정회원

e-mail : cshwang@ee.knu.ac.kr
1979년 : 한국과학기술원 전자공학과(공학석사)
1996년 : 한국과학기술원 전자공학과(공학박사)
1979 ~ 현재 : 경북대학교 전자 · 전기 · 컴퓨터 학부 교수

<주관심 분야> 영상통신, 암호통신, 초고속망