연결요소 특징을 이용한 복잡한 문서영상의 구조 분석

이상협, 이경무
홍익대학교 전파공학과

A new segmentation method for non-manhattan layout document images using connected component analysis

Sang Hyup Lee Kyoung Mu Lee
Dept. of Radio Science & Communication engineering , Hong - Ik University

요약

본 논문은 일반적으로 제약 없는 형식 문서 즉, 논-맨하탄 (non-manhattan) 형식의 이진문서영상을 분석하는 기법으로서, 연결요소기법에 기반한 특징추출과 이를 이용한 영역분리 및 분류에 관한 새로운 방법을 제안한다. 제안한 방식은 바텀-업 (bottom-up)방식으로서 먼저 처리속도의 고속화와 축소시 특징 영역보존을 위해 임계치 축소기법을 사용하고, 축소된 이진 문서영상내의 각 연결된 검은 화소의 집합을 개체화 하고 개체의 특성에 따라 텍스트, 선성분, 해프톤, 도형 그리고 표등으로 분류한다. 영역분류는 두단계로 이루어지는 데, 1 차분류에서는 우선, B/W 비, 면적, 외각 테두리의 높이와 너비 비, 테두리선유무 등의 특징을 이용하여 해프톤, 수평 수직선, 테두리(표 및 도형)영역을 분리한다. 이후 2 차분류에서는 문자성분의 수평결합을 통한 텍스트행 성분을 추출한다. 마지막 후처리 과정으로 표분석 알고리듬을 통하여 테두리 영역 중 표와 도형을 정확히 구분하고, 또한 도형에 관련한 문자성분을 해당 도형 개체에 연결하는 작업을 수행함으로써 완벽한 영역분류를 한다. 다양한 문서영상을 이용한 시뮬레이션을 통하여 제안한 알고리듬의 성능을 입증한다.

1. 서 론

현재 방대한 양의 문서정보를 자동적으로 처리 인식하는 시스템개발에 대한 요구는 멀티미디어 및 정보화에 부응하여, 디지탈 도서관, 복합 사무자동화시스템, 전자출판 등에의 응용과 함께 날로 증대되고 있는 실정이다. 일반적으로 이러한 다양한 형식과 내용으로 구성된 문서정보를 디지탈 형태로 변환하기 위해서는 먼저 텍스트 및 넌-텍스트 영역을 구분하여야하고, 나아가 넌-텍스트영역도 해프톤, 선성분, 표 및 도형등 특징에 따라 분리됨으로써 차후 문자인식 (OCR)에 필요한 적절한 입력을 제공한다.

기존의 개발된 문서 영역 분리 및 분류 알고리듬들은 대부분 탑-다운 방식과 바톰-업 방식[4, 5, 6] 으로 구별할 수 있는데, 탑-다운 방식은 문서전체 구성에 대한 가정을 기반으로, 조건에 맞는 문서에 대해서 매우 효과적인 영역화를 수행하나 논-맨하탄 문서에의 적용이 용이하지 않다. 반면 바톰-업 방식은 단위 특정 개체를 점진적으로 결합시킴으로써 개체의 크기를 확대하고 이의 특성에 따라 분류하는 방식이다. 최근 들어 모폴로지와 웨이브릿 등 다양한 시각에서 접근을 시도하고 있다. 그러나 실용적인 면에 있어서 아직 성능상 많은 문제점을 가지고 있는 실정이다.

본 논문에서는 바톰-업방식에 기초한 실용적이고 정확한 논-맨하탄 문서의 영역 분리 및 분류 알고리듬을 제안한다.
제안한 알고리듬은 먼저 수행속도 및 노이즈처리를 위해 입력된 이진문서영상을 임계치축소기법에 의해 문서를 축소시킨 후 연결요소기법을 기초로한 문서 영역의 개체화, 그리고 특징추출을 통한 제안한 새로운 다단계 분리 및 분류기법 과정을 통하여 효과적인 논-맨하탄 문서의 영역화를 수행한다.

2. 모폴로지를 이용한 영상 축소

스캐너를 통하여 입력되는 원본 문서영상의 크기는 일반적으로 매우 크다. 따라서 실시간 문서영상처리를 위해서는 문서영상을 축소함으로써 데이터의 양을 줄일 수 있는 데, 이때 축소시 문서영상내의 특징영역들의 특성이 훼손되지 않고 동시에 불필요한 노이즈도 제거되도록 하는 것이 바람직하다. 이를 위하여 본 논문에서는 모폴로지를 이용한 임계치축소방법을 사용하여 입력 이진문서 영상을 축소하였다.

임계치축소방법은 이진 영상에서 텍스트, 그림, 선등을 이루는 흑화소(BLACK)를 ON, 배경화소, 즉 백화소(WHITE)를 OFF 라고 정의할 경우, 이진 영상을 N x N 의 비율로 축소한다면 먼저 원본영상을 N x N 크기의 블럭들로 나누고 각 블럭들의 내부 ON , OFF 분포에 따라 하나의 ON 또는 OFF 의 값으로 매

핑시키는 것이다. 블럭 내부의 모든화소가 ON 일 경우만 ON 으로 매핑하는 경우 모폴로지 AND에 해당하는 반면 블럭내부의 화소중 최소 하나 이상 ON화소가 있으면 ON 화소로 매핑하는 것을 모폴로지 OR에 해당된다. 또한 임계치를 이용하여 매핑을 일반화 시킬 경우 임계치 이상 또는 이하의 여부에 따라 좀 더 확장된 개념으로의 임계치 AND와 임계치 OR 연산을 정의할 수 있다. 위의 제안된 방법에 따라 노이즈 제거 및 특징영역을 최대한 보존하는 축소영상을 구현 할 수 있다. 본 논문에서는 수평방향으로 OR 그리고 수직방향으로 AND 연산한 영상과 수직방향으로 AND 그리고 수평방향으로 OR 연산한 두 영상을 다시 OR 연산한 결과를 사용하였는 데, 각 수평, 수직 OR 연산 시 임계치는 1 로 설정하였다..

3. 연결요소기법을 이용한 개체 라벨링

개체 라벨링은 바톰-업 방식의 연결요소기법을 이용하여 상호 연결된 흑화소들의 집합을 독립된 개체화하는 것으로서 이러한 개체 라벨링을 통하여 문서영상내에서 각 개체들을 효율적으로 분리하고 또, 분류에 필요한 정보를 얻어낼 수 있다. 그림 1은 연결요소기법에 의해 각 개체들의 라벨값과 검은화소의 갯수, 그리고 개체를 둘러싼 좌표에 의한 테두리선을 나타내는 과정을 도시한다.

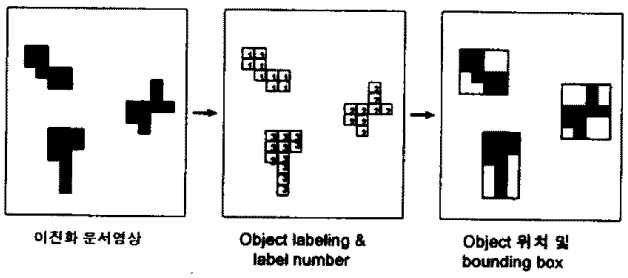


그림 1. 라벨링 알고리듬

본 논문에서 사용한 8-방향 연결요소 개체 라벨링기법 알고리듬의 전체 개략도는 그림 2 와 같다

4. 특징영역 분류 알고리듬

연결요소기법에 의해 구해진 문서내의 각 개체들을 그 특징데이타에 따라 제안한 다단계분류 방법을 통하여 분류함으로써 입력문서영상내의 각 영역들을 효율적으로 추출 한다. 전체 분리 및 분류 알고리듬의 개략도는 그림 3 에 도시된 바와 같다.

4.1 1 차 분류

먼저 각 개체의 특징값중 면적 그리고 개체의 외각 사각형 테두리 정보를 이용하여 입력영상내의 **해프톤, 수평, 수직선 성분, 표 관련 특징영역(개체)**를 분류해낸다. 구체적인 각 개체의 특징들은 다음과 같다.

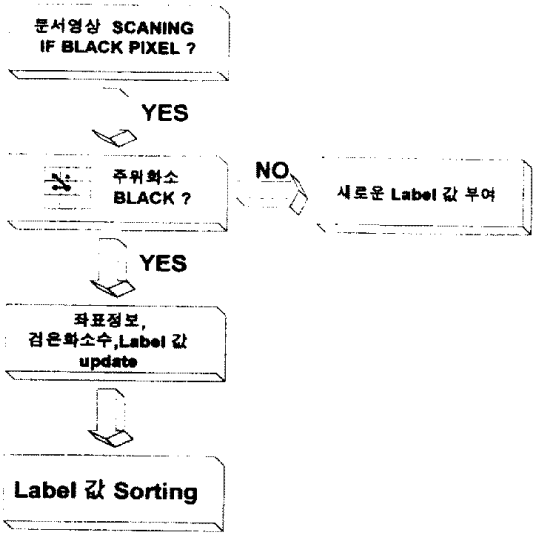


그림 2. 연결요소기법에 의한 라벨링 알고리듬 개략도

- 해프톤 이미지: 문서영상내의 해프톤 이미지의 크기는 비교적 크고 이를 둘러싼 사각테두리 안의 검은 화소의 갯수가 매우 많다. 또 해프톤 이미지의 크기가 크지않을 경우에는 B/W 비 (흑화소 대 백화소 수의 비)가 매우 높다.
- 수평, 수직선 성분 : 사각테두리의 높이 대 너비의 비 또는 너비 대 높이의 비가 임계치이상 크면 선 성분으로 추출한다
- 표 관련 영역: 각 개체 테두리의 검은 화소와 이 개체를 둘러싼 좌표가 임계치 이상 일치하면 표 관련 영역으로 인식한다. 표 관련 영역이란 표뿐만 아니라 사각 테두리 선을 가지고 있는 도형과 그림도 이에 해당한다. 따라서 순수한 표만을 추출하기 위해서는 차후 표 성분과 도형 분리를 위한 후처리 알고리듬이 필요하다.

4.2 2 차 분류

1 차 분류알고리듬을 통해 추출, 분류된 해프톤, 수평, 수직선 성분, 표 관련 영역을 제외한 문서영상내의 특징영역은 크게 문자개체들과 일반적인 도형개체들로나눌 수 있다. 따라서 2 차 분류 단계에서는 이러한 문자개체와 도형개체를 분류하고, 또한 문자개체들을 서로 결합하여 텍스트행을 추출한다. 문자와 도형개체는 다음과 같은 특징정보를 이용함으로써 효과적으로 분류할 수 있다.

- 문자 개체의 특징: 일반적으로 문자개체는 그 크기가 도형의 경우보다 매우 작으며 연속한 개체와 비슷한 크기와 특성을 가지고 있다.
- 일반적인 도형 개체의 특징: 문자개체 보다 크기가 비교적 크며 테두리 사각형내의 B/W 비가 문자개체의 경우보다 작다.

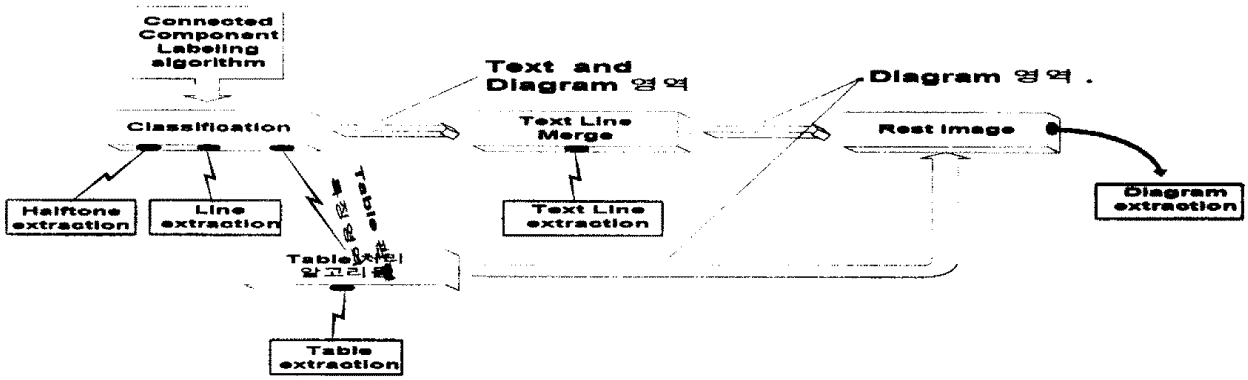


그림 3. 다단계 분류 알고리듬의 전체 개략도

문자개체들의 결합을 통한 텍스트행 추출은 서로 이웃한 개체끼리의 크기와 B/W 비 특징을 비교하여 서로 유사한 경우 하나의 개체로 결합하는 방식을 통하여 수행한다. 이때 개체들의 결합으로 인한 개체 라벨 및 관련 특징데이타 정보의 갱신도 함께 이루어 져야 한다. 텍스트 행이 추출된 후 남은 개체는 도형과 개별문자 또는 길이가 매우 짧은 문자열들 이다. 문자들이 도형에 연관된 경우 이들을 하나의 개체로 결합해주는 것이 필요한데, 지리적인 관계 즉, 문자열이 도형 내부 또는 근접한 거리에 위치하는 가를 파악하여 수행한다.

4.3 표처리 알고리듬

1 차 분류 과정에서 분류된 표 관련 영역에는 표 뿐만 아니라 사각 외곽 테두리선을 가진 일반적인 도형들도 포함이 되어 있기 때문에 다시 순수한 표와 그렇지 않은 도형을 분류하는 후처리 알고리듬이 필요하다. 본 논문에서는 아래와 같은 특징을 사용하여 표와 도형을 분류 한다.

- 표 성분의 특징 : 표를 구성하고있는 사각 외곽 테두리선과 2 개이상으로 구성된 셀이 있으며 각 셀안은 데이타 또는 텍스트로 채워져 있으므로 셀안의 검은 화소수가 비교적 많다.
- 사각 외곽 테두리선을 가진 도형 또는 그림의 특징: 표와 유사하게 사각 외곽 테두리선을 구성하고 있으나 표에 비해 셀의 갯수가 매우 적고, 테두리 선 안의 정보가 비교적 적기 때문에 검은화소 대 흰 화소의 비가 작은 수치를 갖는다.

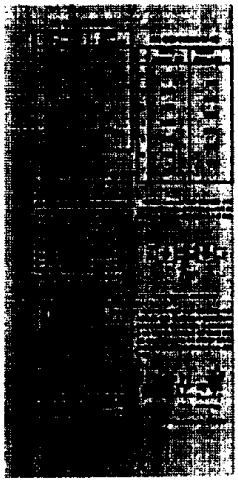
따라서 위의 특징을 고려하여 표와 도형의 구별 및 분류방법은 먼저 특징값을 강조하기 위하여 대상영역을 RLS (Run-Length Smoothing) 시킨 후 이 영역의 검은 화소의 평균 연속값(average black run length)과 B/W 비를 조사함으로써 표와 도형을 분류한다. 이 과정에서 분류된 도형은 2 차 분류과정에서 텍스트행 성분이 추출된 후 남은 도형과 함께 전체적인 도형영역을 구성한다.

5. 시뮬레이션 및 결과

스캐너로 입력된 다양한 형식의 문서에 대하여 가로 대 세로의 비가 4 : 8 의 비로 임계치 축소한 후 제안한 알고리듬으로 시뮬레이션을 수행하였고 그결과를 그림 4, 5 에 도시하였다. 먼저 1 차분류과정에서 해프톤, 수평, 수직선 성분, 표관련 영역을 추출하는데 해프톤 추출시의 임계값은 B/W 비가 0.6 이상, 개체를 둘러싼 사각테두리안의 넓이가 3000 화소이상, 검은 화소의 수가 2000 이상으로 설정했고, 개체를 둘러싼 사각테두리안의 넓이가 3000 이하일 경우에는 검은 B/W 비가 0.9 이상으로 임계치를 설정하여 분리, 분류하였다. 수평, 수직선은 사각테두리의 가로 세로의 비가 20 이상인경우로 설정 추출 하였고, 표 관련 영역은 개체가 그를 둘러싼 사각 테두리선의 3/4 이상 일치하면 표 관련 영역으로 추출, 분리하였다.
2 차 분류과정을 통하여 문자성분을 결합한 후 텍스트행을 추출, 분류하였으며, 마지막 과정으로 제안된 표처리 알고리듬에서는 평균 검은화소의 길이 20, 그리고 B/W 비 0.8 을 기준으로 표와 도형을 분류하였다. 그림 4 는 맨하탄형식의 문서로서 표와 텍스트 그리고 도형의 특징영역으로 구성되어있고, 그림 5 는 전형적인 논-맨하탄형식의 문서로서 텍스트, 수평,수직선과 해프톤 등의 특징영역으로 구성되어있다. 결과에서 볼 수 있듯이 텍스트성분, 표성분, 수평수직선 성분 그리고 도형성분들이 정확히 추출됨을 알 수 있다.

6. 결 론

본 논문에서는 자동문서인식 처리에 필요한 문서내의 해프톤, 선성분, 표와 도형 그리고 텍스트등의 특징영역을 효율적이고 정확하게 분리 및 분류하는 알고리듬을 제안 하였다. 제안한 알고리듬은 방법은 연결요소기법을 이용하여 문서영상내의 개체들을 분리하고 이후 각 개체의 특징값을 이용한 다단계 분류기법을 통하여 정확하게 영역분리를 수행한다. 실제 문서영상에 대한 실험을 통하여 제안한 알고리듬의 성능을 입증하였다. 향후 알고리듬의 최적화 및 고속화에 대한 연구를 통하여 보다 실용적인 문서영상 전처리 시스템의 개발이 기대된다.



원본 영상

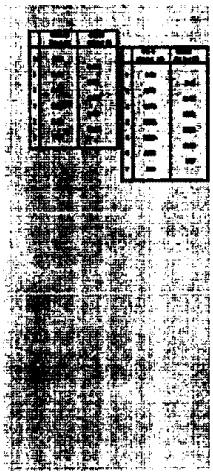
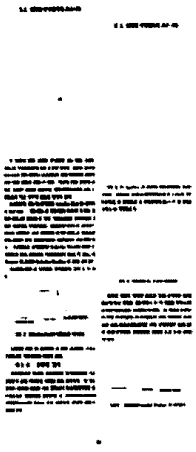
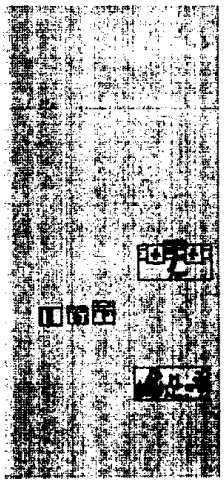


표 영역



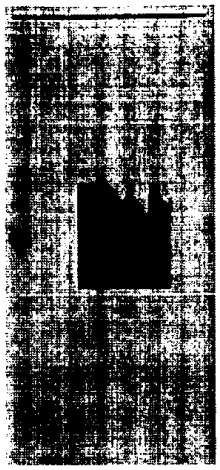
텍스트 영역



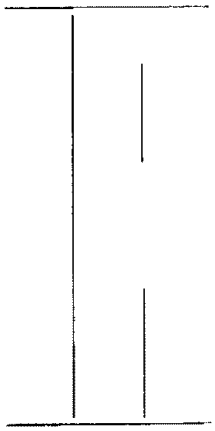
도형 영역

그림 4. 시뮬레이션 결과 1

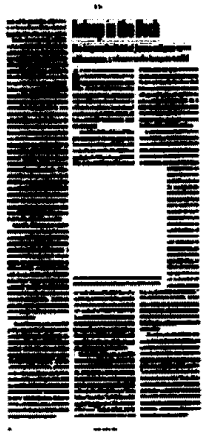
원본 영상



해프톤 영상



선성분 영역



텍스트 영역

그림 5. 시뮬레이션 결과 2

7. 참고 문헌

[1] 이상협, 이경무 ,"모폴로지를 이용한 문서영상내의 특징영역 추출" '96 년도 한국방송공학회 학술대회 논문집 pp. 67-75.

[2] J. L. Fisher, S. C. Hinds and D. P. D'Amato. "A Rule -Based system for document image segmentation", Proc. 10th ICPR, Los Alamitos, CA,1990, PP. 567-572.

[3] T. Pavlids and J. Zhou, " Page segmentation and classification", CVGIP: Graphical Models and Image Processing , Vol. 54, No. 6, Nov.1992. pp.484-496.

[4] J. Sauvola and M. Pietikainen, "Page segmentation and classificarion using fast feature extraction and connentivity anaysis".

[5] Dimitrios Drivas and Adnan Amin, "Page Segmentation and Classification Utilising Bottom- Up Approach".

[6] Pietro Parodi and Giulia Piccioli, " An Efficiect Pre-Processing of Mixed- Content Document Image for OCR Systems"