회전된 문서영상에서의 구성요소 분석 및 분류

모문정*, 김욱현*
영남대학교 컴퓨터공학과
E-mail : mmj@star.tfc.ac.kr

Component Analysis and Classification for Rotated Document Image

Moon-jung Mo*, Wook-Hyun Kim*
Dept. of Computer Engineering, Yeungnam University
241-1, Dae-Dong, Kyungsan, Kyungpook, 712-749, KOREA

요 약

본 논문에서는 회전된 문서에서의 회전각 검출과 문서에 포함된 그림, 글자, 표, 직선과 같은 구성요소를 자동으로 분석하고 분류하는 방법을 제안한다. 본 연구는 입력영상을 획득하는 과정에서 발생되는 회전각에 의해 발생되는 오류를 최소화하기 위한 회전각 검출단계, 각 구성요소 검출에 불필요한 배경제거 단계, 각 구성요소의 특성을 통한 구성요소 분류단계로 이루어진다. 제안한 문서 인식 시스템의 성능 평가를 위해서 다양한 문서에 제안한 방법을 적용하고, 성공적인 결과를 보인다.

1. 서 론

현대사회에서 컴퓨터는 이제 더 이상 숫자 · 문자정보의 입 · 출력만을 위한 수단이 아니며 문자, 도형, 음성, 음향, 정지영상, 동영상 등 다양한 표현형식을 통해서 원하는 정보를 얻는다. 영상 처리란 영상을 카메라나 스캐너 등을 통하여 전자적으로 얻은 후, 여러 가지 목적에 따라 다양한 알고리즘을 적용하여 처리하는 것이다. 1960년대부터 시작된 이러한 영상처리 작업은 최근 섬유산업, 공장자동화, 문서 처리, 의료 진단 영상 시스템 등의 여러 응용분야에서 실용화되고 있으며, 특히 전자문서의 사용이 증대됨에 따라 입력된 문서영상을 편집상태로 전환하는 작업이 많이 연구되어지고 있다. 문서 처리에서 글자나 그림, 도표 등의 자동적인 획득과 처리는 은행이나, 보험회사, 관공소에서 업무를 자동화하는데 큰 도움을 주고 있다. 그러나 문서에 내재된 각 구성요소들을 자동으로 분석하고 분류하는 작업은 미흡한 실정이기 때문에 본 논문에서는 문서영상에서 각 구성요소들을 자동으로 분석하고 분류하는 방법을 제안한다.

입력영상을 획득하는 과정에서 문서가 기울어지는 경우가 많이 발생되며, 문서내에 내재된 각 구성요소를 검출하는 과정에서 이와 같은 오류는 잘못된 결과영상을 나타내기 때문에 문서영상을 처리하는 과정에서 회전각을 제외하는 제약조건을 갖는다. 본 연구에서는 이러한 제약조건을 없애고, 회전각 검출에 소요되는 시간을 최소화하는 방법을 제안한다. 본 연구는 문서영상을 처리하는 과정에서 발생되는 오류를 최소화하기 위한 회전각 검출단계, 불필요한 연산의 수행을 최소화하기 위해 배경영역을 제거하는 구성요소 영역 검출단계, 구성요소 분류단계를 수행하며, 마지막으로 고주파 필터링을 통한 구성요소 영역 검출 단계로 구성된다.

2. 문서 인식 시스템

본 논문에서 제안하는 문서인식시스템은 그림 1에 나타낸다.

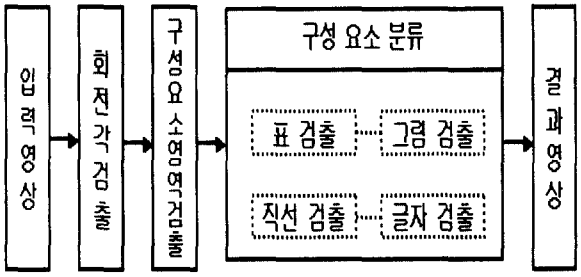


그림 1. 문서인식 시스템

2.1 회전각 검출

입력영상을 획득하는 과정에서 기울어짐과 같은 오류가 발생될 가능성이 높으며, 구성요소를 검출하는 과정에서 이러한 오류는 잘못된 결과를 제시할 가능성이 매우 높다. 본 실험에서는 문서에 내재된 각 구성요소를 검출하는 과정에서 발생될 수 있는 오류를 최소화하기 위해서 먼저 회전각의 검출과 회전되기 이전의 상태로 되돌리는 전처리 작업을 수행한다.

영상의 회전각을 검출하기 위해서 일반적으로 허프변환이 사용된다. 허프변환은 입력된 영상에 포함된 직선, 곡선성분의 검출을 통해서 회전각에 대한 정보를 얻는다. 허프변환은 노이즈를 포함하는 이진 영상에서 선이나 원 등의 도형을 추출하는 데에는 유효한 방법이나 거대한 파라미터 공간을 설정하지 않으면 안되므로, 기억용량, 처리속도 등에서 불리하다. 그러므로, 입력 영상 중의 픽셀의 좌표뿐만 아니라, 그 픽셀에 있어 선의 방향 등을 이용하여 파라미터 공간에서 차원 수를 줄이는 방법, 변환 대상 픽셀수를 줄이는 방법 등 추가 작업이 필요하다.

본 논문에서는 회전각을 검출하는데 소요되는 시간과 추가작업을 최소화하기 위해서 먼저 입력된 영상의 최상위 수평성분에 해당되는 픽셀들과 좌측 수직성분에 해당하는 픽셀들을 검출한다. 행과 열을 기준으로 가장 먼저 나타나는 픽셀들을 검출함으로써 수평성분과 수직성분에 해당되는 픽셀을 검출한다. 회전각은 수평선과 수직선이 동일한 값을 나타내므로 검출된 수평성분과 수직성분에 대한 회전각을 검출한다.

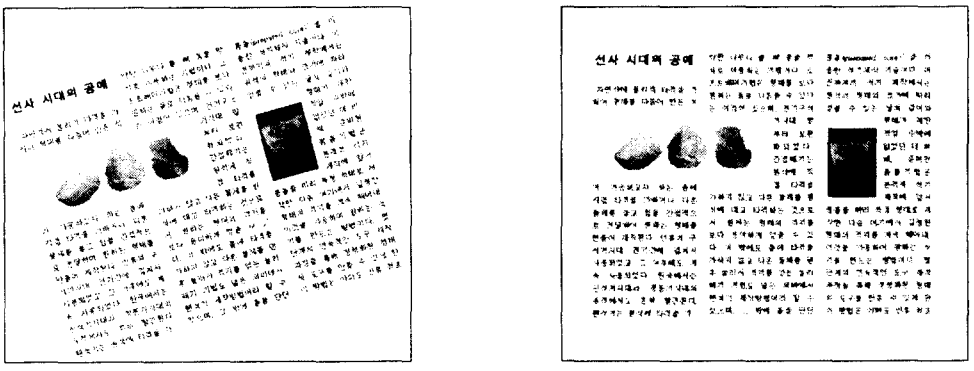
제목이나 들여 쓰기와 같은 편집형식에 의해서 잘못된 회전각 검출을 최소화하고, 입력영상의 크기에 제한을 받지 않도록 입력영상을 수직, 수평에 대해서 10개의 영역으로 분할하고 각 영역에서 검출된 첫 번째 픽셀을 기준으로 나머지 픽셀들과의 회전각을 검출함으로써 보다 정확하고 빠른 회전각 검출이 이루어진다. 검출된 회전각을 기준으로 입력영상을 회전되기 이전의 상태로 복원한다. 그림 2는 시계 반대 방향으로 15 ° 회전된 영상에 대한 실험 결과를 나타낸다.

2.2. 구성요소 영역 검출

회전에 의한 오류를 제거하고 각 구성요소의 특성을 분석하기 위해서 각 구성요소를 포함하는 영역을 검출하는 과정을 수행한다. 배경부분은 각 구성요소를 검출하는 과정에서 불필요한 연산을 수행하게 되므로 본 단계에서는 배경을 제거하는 과정을 수행한다. 문서는 최소한의 테두리 영역을 포함하기 때문에 입력영상의 수평, 수직길이와 동일한 크기를 갖는 흰색 픽셀들의 그룹은 배경으로 인식하고 오브젝트 영역에서 제외하는 1차 단계를 수행하며, 검출된 오브젝트 영역 중에서 다단이나 줄 바꿈에 의한 배경부분도 오브젝트 영역에서 제외하는 2차 단계를 수행한다. 그림 3은 배경과 오브젝트 분리결과를 나타내며, 각 영역의 무늬는 분리단계를 구분하기 위해서 임의로 나타내었다.

2.3 구성요소 분류

구성요소 분류단계는 문서에 내재된 그림, 글자, 표,



(a) 입력영상 (b) 입력영상에 대한 복원

그림 2. 회전된 영상의 복원

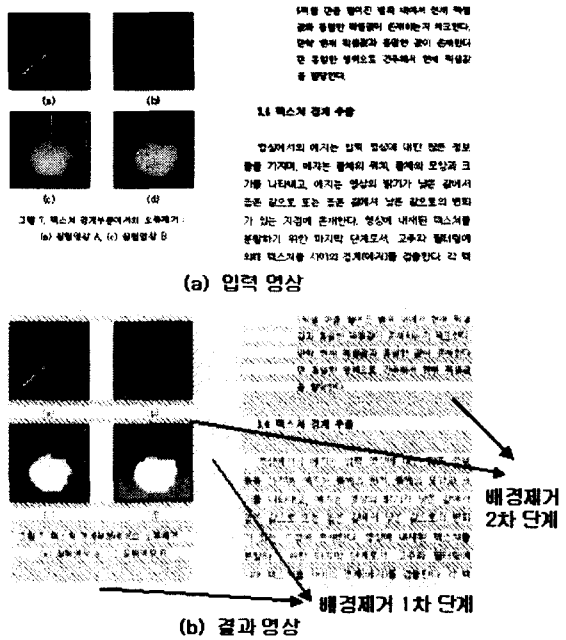


그림 3. 오브젝트와 배경 분리

직선 등의 각 구성요소의 특징을 파악하고 각 구성요소를 분류를 작업을 수행한다. 각 구성요소는 표, 그림, 직선, 글자의 순으로 검출한다.

2.3.1 표 검출

표와 그림은 일반적으로 글자보다는 큰 영역으로 나타나며, 장평으로 글자가 확대되었을 경우에도 수평, 수직의 어느 한 부분에 대해서는 보다 작은 영역을 갖는다. 먼저 입력 영상의 수평, 수직크기에 대해서 모두 5%이상의 크기를 갖는 영역들을 검출하고, 표는 내부에 일정한 크기를 갖는 흰색 픽셀들의 그룹을 포함하므로 각 영역의 내부에 수평, 수직크기의 90%에 해당되는 흰색 픽셀들의 그룹이 규칙적으로 존재할 경우 이 영역을 표영역으로 정의한다.

2.3.2 그림 검출

표 검출과정에서 제외된 블록들에 대해서 그림영역을 검출하며, 큰 글자가 그림영역으로 정의되는 오류를 최소화하기 위해서 각 영역에서 나타나는 값들이 균일하지 않을 경우는 글자나 직선영역에서 제외되므로 그림영역으로 정의한다.

2.3.3 직선 검출

표와 그림 영역에서 제외된 부분들은 글자나 직선영역으로 정의되며, 나머지 영역들을 수평선이나 수직선은 일정한 값들로 표시되기 때문에 글자 부분과 구분하여 정의할 수 있다.

2.3.4 글자 검출

검출되지 않은 영역들 중에서 글자 영역을 정의하며, 글자는 일반적으로 흰색에 가까운 값들과 검은색에 가까운 값들의 비율이 3:7에서 7:3의 비율로 나타나기 때문에 검출되지 않은 영역들 중에서 이와 같은 비율의 조건을 만족하는 영역을 글자 영역으로 정의한다.

2.3.5. 구성요소 영역 검출

문서영상에 내재된 각 구성요소 검출을 위한 단계로서, 고주파 필터링에 의해 각 구성요소들 사이의 경계(에지)를 검출한다. 각 구성요소들의 경계를 검출함으로써 입력된 문서영상을 편집한 가능한 형태로 변환하는데 용이하게 사용될 수 있다.

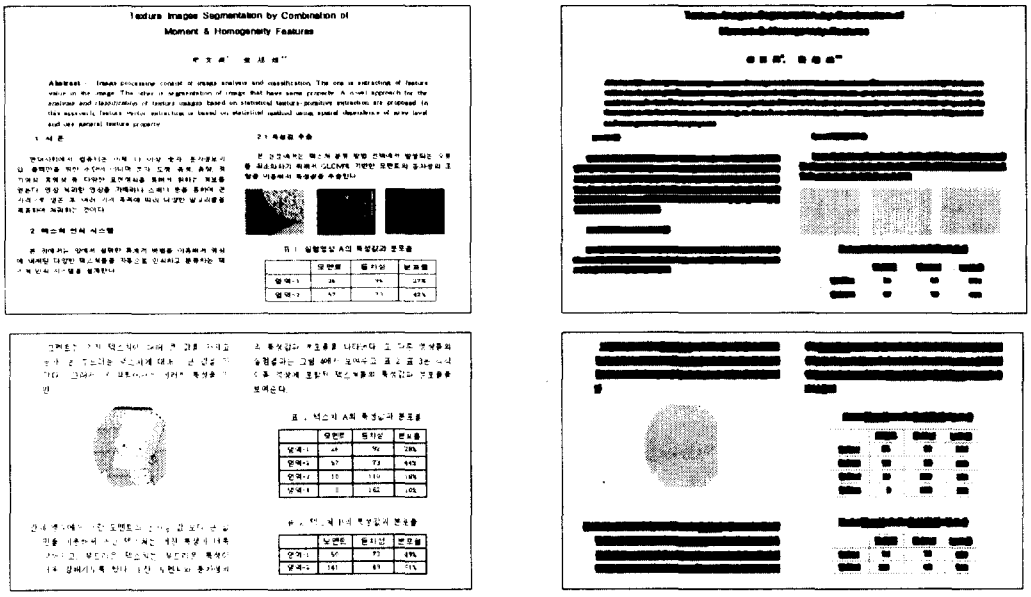
3. 실험결과 및 검토

실세계는 다양한 크기와 종류의 문서가 존재하기 때문에 본 연구에 사용된 실험 문서영상은 A4크기의 문서뿐만 아니라 다양한 크기를 갖는 256 그레이 레벨 영상을 선택하였으며, 각 단계별 처리과정은 SUN SPARC 워크스테이션의 X-WINDOWS하에서 C언어로 구현되었다.

본 논문에서 제안한 방법은 회전각 검출 단계, 배경제거에 의한 구성요소 영역 검출 단계, 각 구성요소 특성정의에 의한 구성요소 분류 단계, 그리고 고주파 필터링을 통한 구성요소 영역 검출 단계로 구성되며 본 논문에서 제안한 문서 인식 시스템을 실험영상에 적용해서 얻어진 결과 영상을 그림 4에 나타낸다.

4. 결 론

본 논문에서는 회전각 검출에 소요되는 계산량을 최소화하고 문서의 크기나 회전각의 범위에 영향을 받지 않으며 간단하고 빠르게 회전각을 검출하는 방법과 문서영상에서 내재된 각 구성요소를 자동으로 인식하고 분류하기 위한 방법을 제안하였다. 글자뿐만아니라 그림영역이나 표영역을 동시에 검출함으로써 문서영상을 전자문서 형태로 쉽게 전환할 수 있다. 문서에 포함된 각 구성요소를 검출하는 과정에서 불필요한 배경영역으로 먼저 제거함으로써 불필요한 계산량을 줄임으로써 빠른속도로 검출이 가능하다. 제안된 방법은 은행이나, 보험회사, 관공소에서 업무를 자동화에 적용이 기대된다. 향후 과제로는 도표와 같은 부분에 대한 추가적인 연구가 필요하며,



(a) 원영상 (b) 구성요소 검출결과 영상

그림 4. 영역 분류 결과

회전된 문서를 회전되기 이전의 상태로 복원하는 전처리 작업없이 회전된상태에서 회전각의 검출만으로 각 구성요소를 검출하는 연구가 필요하다.

참고문헌

[1] Q. Luo and N. Sugie, "Layout recognition of multi-kinds of table-form documents", IEEE Computer Society, vol. 17, no. 4, 1995.

[2] Y. S. Chen and Y. T. Yu, "Thinning approach for noisy digital patterns", Pattern Recognition 29(11), 1847-1862 (1996).

[3] J. L. Chen and H. J. Lee, "An efficient algorithm for form structure extraction using strip projection", Pattern Recognition 31(9), 1353-1368 (1997).

[4] Z. lu, "Detection of text regions from digital engineering drawings", IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 4, 1998.

[5] L. Y. Tseng and R. C. Chen, "Recognition and data extraction of form documents based on three types of line segments", Pattern Recognition 31(10), 1525-1540 (1998).

[6] D. Dori and Y. Velkovitch, "Segmentation and Recognition of Dimensioning Text from Engineering Drawings", Computer Vision and Image Understanding", vol. 69, no. 2, 1998.

[6] 金 熙 昇, "영상인식", 生能, 1993.