기울기 보정과 블록 분할 합병을 통한 문자 추출

[0]김도현* 강민경** 차의영*
*부산대학교 전자계산학과
**부산대학교 멀티미디어 학과
{uliminer, dragon, eycha}@harmony.cs.pusan.ac.kr

Text Extraction by Skew Normalization and Block Split & Merge

[0]Do-Hyeon Kim* Min-Kyeong Kang** Eui-Young Cha*
*Dept. of Computer Science, Pusan National University
**Dept. of Multimedia, Pusan National University

요 약

신문, 잡지, 공문서, 영수증 등의 문서로부터 필요한 정보를 자동화하여 처리할 수 있는 문서영상 이해 시스템의 구현에 있어서 문서영상에 존재하는 문자를 추출하는 연구는 문자 인식의 전처리 단계로서 매우 중요한 의미를 지니고 있다. 하지만 현 시점에서 문서 자체가 가지는 다양한 형태 및 배경 등에 의하여 범용화되고 일반화된 방법을 찾기란 매우 어려운 실정이다. 본 논문에서는 특히 배경이 선이나 도표 등으로 이루어진 문서 영상에서 Hough Transform을 사용하여 기울어짐을 보정하고 문자들이 선에 겹친 부분을 효과적으로 보정하며 추출된 영역에 대한 분할 및 합병 과정을 거쳐 최종적으로 완전한 문자 영역을 추출하는 방법에 대하여 다룬다.

1. 서 론

컴퓨터가 산업의 전반에 다양하게 활용됨에 따라 컴퓨터를 통한 정보 처리의 필요성이 날로 두각되고 있다. 특히 신문, 잡지, 공문서, 영수증 등 다양한 양식의 문서로부터 필요한 정보를 추출하여 이를 자동화하고 해석하하는 문제는 벌써 오래 전부터 많은 연구와 개발이 진행되고 있다[1-5]. 이와 같이 다양한 형태의 공문서나 영수증으로부터 필요한 정보를 추출하기 위한 일련의 문서 영상 이해 시스템의 한 부분으로써 선행되어야 할 문자 추출 방법에 대해 본 논문에서는 간단하고 직관적인 방법을 제시하고자 한다.

일반적으로 문서영상 이해 시스템은 크게 영역 분할(Block Segmentation) 과정과 영역 식별(Block Classification) 과정, 그리고 문자 인식 (Character Recognition) 과정으로 이루어진다. 영역 분할 과정은 문서상에서 필요한 정보가 존재하는 영역을 구분 짓는 과정이라 할 수 있으며, 영역식별 과정은 그 구분된 영역에 대한 특성을 결정하여 분류하고 검증하는 과정이다. 이와 같은 과정을 거친 후 최종적으로 문자 인식과정을 거쳐 필요한 정보를 얻게 된다. 그러나 현실적으로 요구되는 문서영상 이해시스템에서는 해결되어야 할 많은 문제들이 제기되고 있으며 특히 문자 인식을 위한 전처리 단계로써 올바른 문자 영역을 추출하는 것은 올바른 문자 인식을 위하여 반드시 선행되어야 하는 중요한 과정이다. 본 논문에서는 특히 배경이 선이나 도표 등으로 이루어진 문서 영상에서 그림 1에서 보는 바와 같이 Hough Transform을 통하여 기울어진 문서에 대한 보정을 수행하며, 문자들이 선에 겹친 부분에 대하여 배경 선을 제거하면서 소실된 문자 획에 대하여 보정하는 과정을 거치며 추출된 영역에 대한 분할 및 합병 과정을 거쳐 최종적으로 완전한 문자 영역을 추출하는 방법에 대하여 다룬다.

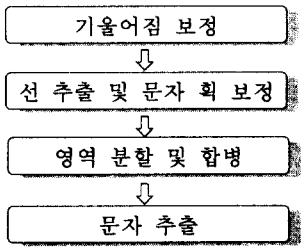

그림 1. 문자추출 과정

2장에서는 Hough Transform을 이용한 기울어짐 보정 과정을 설명하고, 3장에서는 배경선을 제거하면서 소실된 문자 획에 대한 보정 방법을 설명한다. 그리고 4장에서는 추출된 문자 영역에 대하여 완전한 문자 영역을 확정짓는 단계로서 문자 블록 분할 및 합병 과정을 설명하고 이어 예를 통한 실험과 향후 연구 과제에 이어 결론을 맺는다.

2. 기울어짐 보정 (Skew Normalization)

문서영상에서 기울어짐을 보정하는 것은 정확한 문자 영역을 결정하기 위해 선행되어야 하는 작업이다. 즉, 비뚤어진 문서 화상의 기울어짐을 보정함으로써 문자에 대한 외접사각형 영역을 보다 정확하게 추출해낼 수 있게 된다. 기존의 방법으로는 이진화된 문서 영상에 대해 수평 투영(Projection) 과정과 작은 각도의 회전을 반복하면서 문자열과 문자열이 분리되는 간격이 최대가 되는 지점을 찾아 기울어짐을 보정하는 방법을 사용하였으나 본 논문에서는 영상공간의 좌표를 기울기(Θ)와 원점으로부터의 거리(ρ) 공간으로 변환하는 Hough Transform (2)을 사용[6]하여 문서 화상의 최장 직선 성분을 검출

함으로써 이에 해당하는 각도로 회전하여 정확한 문서 기울기 보정을 수행하게 된다.

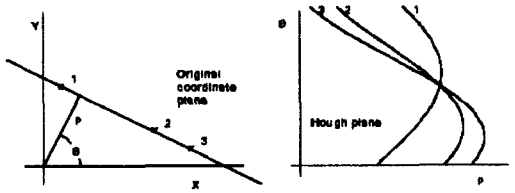

그림 2. Hough Transform

$$y = ax + b \quad (1)$$

$$\rho = x \cdot \cos\theta + y \cdot \sin\theta \quad (2)$$

그림 2에서 보는 바와 같이 카테시안 좌표계에서 식 (1)로 표현된 직선상의 x, y 좌표들(1,2,3)은 식 (2)에 의해 변환된 ρ와 Θ로 정해지는 좌표계에서 하나의 점을 공유한다. 이와 같은 ρΘ 파라메타 공간을 나타내는 2차원 배열을 이용하여 각 점의 누적도를 조사하면 최대가 되는 지점의 각도(Θ)와 거리(ρ)를 구할 수 있으며 이에 따라 문서를 정확한 각도로 회전보정 할 수 있다.

3. 선 추출 및 문자 획 보정

배경이 도표나 선으로 이루어진 문서에서 선에 해당하는 배경 부분을 제거해 버리면 문자 획을 유실하게 될 가능성이 많다. 특히 문자 획이 선에 겹쳐지는 경우 획이 끊겨버리는 현상이 발생하게 된다.

(a) (b) (c)

그림 3. 문자 획 유실 유형

그림 3은 수평선에 의하여 끊어진 현상을 대표적인 3가지 유형을 보여주고 있으며 이를 보정하기 위하여 다음과 같은 과정을 수행한다.

3.1 배경 선 추출

문서 영상에서의 배경 선을 추출하기 위하여 2장에서 사용한 것과 같은 Hough Transform을 이용하여 도표나 선 등의 직선군을 검출한다. 이때 ρΘ 파라메타 공간상에서 최대가 되는 $g_{\rho'\theta'}$ 에 대하여 G''(ρ,Θ)에 대한 역변환을 수행하면 최대가 되는 각도상에서의 직선군들을 검출할 수 있다.

$$g_{\rho'\theta'} = \mathrm{MAX}(G(\rho,\Theta)) \quad (3)$$

$$G(\rho,\Theta) = \{\ \rho,\Theta \mid \rho=1,2,\ldots,M,\ \Theta=0°,1°,\ldots,180°\ \} \quad (4)$$

$$G'(\rho,\Theta) = \{\ \rho,\Theta \mid \rho=1,2,\ldots,M,\ \Theta=\Theta'\ \} \quad (5)$$

$$G''(\rho,\Theta) = \{\ \rho,\Theta \mid G'(\rho,\Theta) > g_{\rho'\theta'} * \mathrm{match\_ratio}\ \} \quad (6)$$

여기서 G(ρ,Θ)는 ρΘ 파라메타 공간을 나타내며, M은 최대 거리를 나타낸다. match_ratio는 참조할 파라메타 공간배열값의 범위를 제한하는 비율값으로 최대값 $g_{\rho'\theta'}$ 대한 비율을 나타낸다.

3.2 모폴로지 닫힘연산을 이용한 획 보정

추출된 직선군 영상을 원영상에 대하여 차연산을 수행하면 원영상에서 직선군과 교차되는 부분의 문자 획이 유실되는 현상이 발생하게 된다. 이때, 이 유실 획 영역을 보정하기 위하여 모폴로지 연산(Morphological Operation)[6]을 수행한다. 먼저 그림 10에서 검출된 직선군의 에지 영역을 차례대로 검색하면서 해당하는 위치가 원영상과 교차하는 부분을 표시해둔다. 모든 좌표에 대한 표시 과정이 끝나면 교차되는 지점만 표시된 하나의 새로운 영상이 생성되며 이 영상에 대해서 그림 4와 같은 모폴로지 mask로 Closing 연산을 수행하면 그림 3에서와 같이 상하, 대각선 방향으로 유실된 문자 획을 주변 획과 참조하여 보정하게 된다.

1	0	1	0	1
0	1	1	1	0
1	1	1	1	1
0	1	1	1	0
1	0	1	0	1

(a) 5x5 mask

1	1	1
1	1	1
1	1	1

(b) 3x3 mask

그림 4. 모폴로지 mask

4. 문자 블록 분할 및 합병(Block Split&Merge)

기울어짐 보정과 문자 획 보정 과정을 거쳐 추출된 단순한 텍스트 영역에서 실제적으로 정확한 문자 영역을 분할할 필요가 있다. 일반적으로 문자열에서 문자영역을 추출하는 방법으로는 수직 투영(Vertical Projection)과 수평 투영(Horizontal Projection)을 이용[6]하고 있지만 실제적으로 투영만으로는 그림 5에서 보는 바와 같이 정확한 분할 영역을 찾기가 곤란하다. 따라서 본 논문에서는 그림 6에서 보는 바와 같이 투영을 바탕으로 문자의 외접 사각 블록을 최소한의 길이 폭으로 먼저 분리(split)한 다음 이웃 블록을 통합(merge)하면서 각 블록들의 폭에 대한 분산을 구하여 이를 최소로 하는 영역으로 외접 사각 블록을 결정하였다.

그림 5. 투영만을 이용한 문자 분할

그림 6. 글자폭의 분산값을 이용한 문자 분할

5. 실험 결과

본 연구의 실험으로는 Windows 운영체제 하에서 Microsoft Visual C++을 사용하여 구현하였으며 실제적으로 도표와 선이 있는 영상 및 문자 획이 겹치는 영상인 입출금 영수증을 HP ScanJet 스캐너로 scan하여 사용하였다. 원영상은 그림 7 에서 보는 바와 같이 회전보정이 필요하며 이진화하였을 경우 문자 획이 선에 겹

치는 현상이 발생하였다.

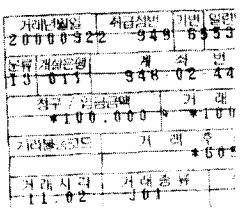

그림 7. 원영상

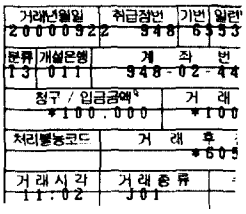

그림 8. 회전 이진영상

그림 9. Hough Transform

그림 10. 수평 직선군

그림 11. 직선제거 영상

그림 12. 획 보정 영상

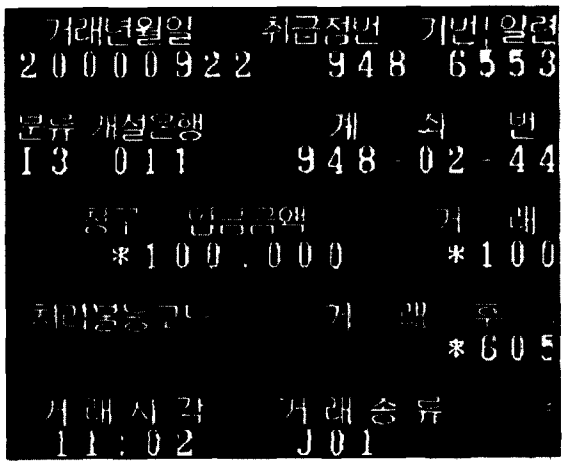

그림 13. 최종 문자 추출 영상

이와 같은 원영상(그림 7)에 대하여 회전 보정된 이진 영상은 그림 8과 같고, 이 이진 영상에 대하여 반전(Invert)을 시킨 후 Hough Transform(그림 9)으로 직선군을 검출한 영상은 그림 10과 같다. 그림 9에서 보이는 점들은 ρΘ공간상에서 일정한 각도(Θ:0°)에 대하여 일정한 거리에 직선군이 존재함을 보여주고 있다.

그림 11과 12는 그림 10을 이용하여 직선군을 제거한 영상과 획에 대한 보정 결과를 보여준다. 그림 12에서 보는 바와 같이 가로선에 의하여 겹쳐진 문자들에 대하여 효과적으로 문자 획이 보정되었음을 알 수 있다. 이와 같은 과정을 수평 직선군과 수직 직선군에 반복 적용하면 직선 제거 영상을 추출할 수 있으며 이 영상에서 문자 블록 분할 및 합병을 통하여 최종적으로 그림 13과 같은 글자 추출 결과를 얻음을 알 수 있다.

6. 향후 연구 과제 및 결론

본 연구는 도표나 선이 있는 배경으로부터 문자를 효율적으로 추출하는 방법에 대하여 Hough Transform을 통한 직선군 제거 및 모폴로지 연산을 사용한 보정 과정을 거치면서 문자 획을 추출한 다음 추출된 문자 영역에 대한 외접사각영역을 투영을 통해 구하고, 영역의 폭을 기준으로 최소 분산값을 가지는 영역으로 올바른 문자 영역을 확정지을 수 있었다. 하지만 문자 획이 선에 정확히 평행하게 일치하는 경우에는 이를 정확히 보정할 수 없었던 문제점과 Hough Transform 시 각 각도에 대하여 sin값과 cos값에 대한 Lookup Table을 미리 작성하고 제한된 각도 범위만으로 연산하였음에도 불구하고 변환 자체가 가지는 많은 연산량으로 인하여 속도 저하를 가져오는 원인이 되어 이를 개선할 수 있는 방법에 대한 연구도 필요할 것으로 생각된다. 또한 문자 폭의 분산을 이용한 영역설정에 있어서 문자 폭이 다른 전각문자나 반각문자가 섞여 있거나 여러 가지 기호 등이 포함되어 있는 경우에는 영역 분할 상의 오류가 발생하였으며 이런 문제는 추출된 영역의 문자 인식을 통한 feedback으로 문자 인식의 오류를 줄이는 방안을 모색해야 할 것으로 생각된다.

참 고 문 헌

[1] 고견, 이일병, "한국 문서 인식 시스템 개발 연구", 한국 정보과학회 추계 학술 발표회 논문집, 제 14권 2호, pp. 167~169, 1987년 10월

[2] 이인동, 권오석, 김태균, "문서 영상에서 문자와 비문자의 분리 추출 방법", 한국 정보과학회 논문지, 제 17권 3호, pp.247~258, 1990년 5월

[3] 박영석, "일반적인 문서화상의 영역 식별법", 한국정보과학회 논문지, 제 21권 5호, pp.757~767, 1994년 5월

[4] Seong-Whan Lee, "Direct Extraction of Topographic Features for Gray Scale Character Recognition", IEEE Transaction on Pattern Analysis and Machine Intelligence, VOL17, NO. 7, JULY 1995

[5] Dong-June Lee, "A New Methodology for Gray-Scale Character Segmentation and Recognition", IEEE Transaction on Pattern Analysis and Machine Intelligence, VOL18, NO. 10, OCTOBER 1996

[6] Ramesh Jain, Rangachar Kasturi, Brian G. Schunck, *Machine Vision*, McGRAW-HILL INTERNATIONAL EDITION, pp. 61~69, 218~223