2000년 한국정보처리학회 추계 학술발표논문집 제7권 제2호

텍스트 영역에 대한 단어 단위 분할 시스템

정창부*, 곽희규*, 정선화*, 김수형*
*전남대학교 전산학과
e-mail:cbjeong@cs.chonnam.ac.kr

A System for the Decomposition of Text Block into Words

Chang-Boo Jeong*, Hee-Kue Kwag*, Seon-Hwa Jeong*, Soo-Hyung Kim*
*Dept. of Computer Science, Chonnam University

요약

본 논문에서는 주제어 인식에 기반한 문서영상의 검색 및 색인 시스템에 적용하기 위한 단어 단위 분할 시스템을 제안한다. 제안 시스템은 영상 전처리, 문서 구조 분석을 통해 추출된 텍스트 영역을 입력으로 단어 단위 분할을 수행하는데, 텍스트 영역에 대해 텍스트 라인을 분할하고 분할된 텍스트 라인을 단어 단위로 분할하는 계층적 접근 방법을 사용한다. 텍스트라인 분할은 수평 방향 투영 프로파일을 적용하여 분할 지점을 구한다. 그리고 단어 분할은 연결요소들을 추출한 후 연결요소간의 gap 정보를 구하고, gap 군집화 기법을 사용하여 단어 단위 분할 지점을 구한다. 이때 단어 단위 분할의 성능을 저하시키는 특수기호에 대해서는 휴리스틱 정보를 이용하여 검출한다. 제안 시스템의 성능 평가는 50개의 텍스트 영역에 적용하여 99.83%의 정확도를 얻을 수 있었다

1. 서론

날로 발생량이 증가하고 있는 인쇄문서에 대한 자동 정보 추출에 대한 요구는 당연한 것이라 하겠다. 대규모 문서의 컴퓨터 데이터베이스화에 따른 기존 고비용 수작업 입력을 대체할 수 있는 자동 정보 추출 방식은, 문서를 디지털 영상의 형태로 저장한 후, 문서의 기하학적인 구조를 분석하고 의미를 이해하는 문서 인식(Document Recognition)을 말한다[1]. 이러한 문서 인식의 중요한 첫 단계가 문서 영상 분할이며 인식률에 큰 영향을 미친다.

문서 인식에 대한 접근 방법은 크게 광학 문자 인식(OCR: Optical Character Recognition)을 기반으로 하는 방법과 주제어 검색(keyword spotting)을 기반으로 하는 방법을 들 수 있다[2]. OCR 기반 방법은 문서 영상의 내용 전체를 텍스트 코드 형태로 변환한 후 인식을 수행하는 것으로, 인식을 위한 문서 분할은 영상 전처리, 문서 구조 분석, 문자 분리 등의 처리 과정을 거쳐 이루어진다[3]. 그러나 OCR 기반 문서 인식 방법을 실제 응용 시스템에 적용하기에는 낮은 인식률, 비텍스트(non-text) 부분 처리의 어려움, 수작업에 의한 OCR 결과의 검증 등의 아직도 해결해야 할 많은 문제점을 안고 있다. 반면, 주제어 검색 방법은 사용자가 지정한 주제어를 단어 영상이 갖는 고유한 특징들을 이용하여 인식하는 영상-기반 인식 방법으로써 대용량 문서 영상을 다루기에 적합한 고속의 성질을 가지며, 사용자가 원하는 단어들이 있는 페이지 및 위치까지 바로 찾아갈 수 있다는 장점이 있다. 여기서에서는 단어 영상의 특징을 추출하기 위해서 문서 영상을 띄어쓰기 단위인 단어 영상의 집합으로 분할한다.

본 논문은 주제어 인식에 기반한 검색 및 색인 시스템에 적용하기 위한 영상 분할로서, 영상 전처리, 문서 구조 분석을 통해 추출된 텍스트 영역을 띄어쓰기 단위인 단어 영상의 집합으로 분할하는 방법에 대한 연구이다. 기존의 연구들이 상향식 또는 하향식 접근 방법으로 분류되는 반면, 제안 방법은 고속의 투영 프로파일 방법과 연결요소 분석 방법을 혼합하여 사용한다[4, 5, 6, 7]. 텍스트 영역에 대해 계층적인 방법으로 텍스트 라인을 분할하고, 각 텍스트 라인들에 대해 단어 단위로 분할한다. 텍스트 라인분할에는 수평 방향 투영 프로파일(HPP: Horizontal Projection Profile)을 계산하여 분할 지점을 구하고, 재귀적 투영 프로파일(RPPC : Recursive Projection Profile Cut) 분석 방법을 추가하여 정확도를 개선한다. 단어 단위 분할은 각 텍스트 라인에 연결 요소 분석을 수행하여, 연결요소간의 다양한 거리 값에 대해 클러스터링 기법을 사용하여 단어 단위의 분할 지점을 추출한다. 또한, 단어와 단어 사이에 위치함으로써 단어 분할에 어려움을 주는 특수 기호들을 인식 단계를 거치지 않고 휴리스틱 정보들을 이용하여 검출함으로써 분할에 따른 정확도를 개선한다. 성능 평가를 위한 실험은 50개의 텍스트 영역을 사용하여 정확도를 계산한다. 제안한 단어 분할 시스템은 2장에서

더 자세하게 기술한다. 또한 3장에서는 실험 결과 및 분석에 대해서, 4장에서는 결론 및 향후 연구에 대해 논한다.

2. 단어 단위 분할 시스템

문서 영상의 단어 분할 시스템은 영상 전처리와 문서 구조 분석을 통해 추출된 텍스트 영역을 단어단위 영상으로 분리·추출한다. 제안된 시스템의 흐름도는 그림 1과 같이 크게 두 단계로 구성된다. 먼저 텍스트 영역을 텍스트 라인 단위로 분리하기 위하여, 텍스트 영역에 대한 수평방향 투영 프로파일을 구하여 분할 지점을 찾는다. 다음으로 분리된 텍스트 라인 각각에 대해 연결 요소 분석 방법을 이용하여 간격(gap) 정보를 구하고, 평균-결합 클러스터링(Averager-Linkage Clustering) 방법으로 단어간의 간격(IWG : Inter-Word Gap)을 분류하여 단어 영상을 분리한다. 이때 단어와 단어 사이에 위치하여, 인식 전 단계에서는 단어영상 분할에 오류를 유발하는 특수기호들을 검출하여 단어영상 분리의 신뢰성을 높이도록 한다.

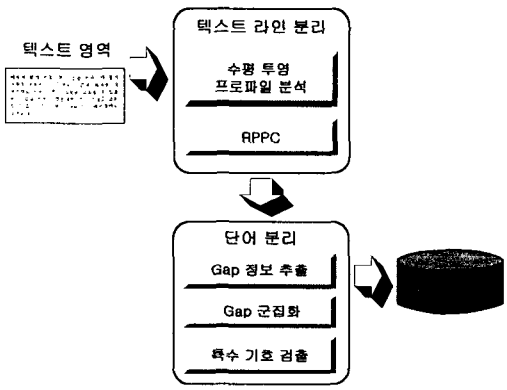

그림 1. 단어 영상 분할 시스템 구성도

2.1 텍스트 라인 분리

텍스트 라인 분리 단계는 텍스트 영역을 입력으로, 텍스트 라인의 집합을 추출하는 것이다. 제안 방법에서는 텍스트 영역에 대한 수평 방향 투영 프로파일을 분석한다. 이는 수평 투영 프로파일에서 텍스트 라인과 행간 영역의 peak-valley 특성을 이용하는 것으로써, valley 부분을 각 텍스트 라인의 분리점으로 간주한다. 그림 2와 같이 ①과 ②에서 프로파일 값이 0이 되는 valley 부분이 나타나며, 이 부분의 중간 지점을 텍스트 라인 분리점으로 결정한다.

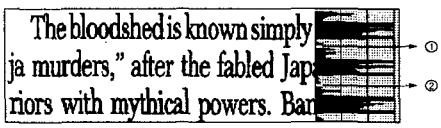

그림 2. 텍스트 영역의 수평 방향 투영 프로파일 결과

일반적으로 대부분의 텍스트 라인들은 수평 방향 투영 프로파일을 이용하면 쉽게 분리 가능하다. 그러나 문서 구조 분석을 통해 분리된 텍스트 영역이 다양한 형태로 구성되어 있을 경우, 단순히 수평 방향 투영 프로파일만으로는 분리되지 않을 경우가 있다. 그림 3의 상단 부분처럼 하나의 문자가 두 개의 텍스트 라인에 걸쳐 있어서 라인 분리의 조건이 되는 수평 방향 투영 프로파일 값이 0인 부분이 나타나지 않음을 보여준다. 이러한 문제를 해결하기 위한 제안 방법으로 RPPC 방법을 사용한다[5]. RPPC 처리의 대상이 되는 라인은 전체 라인들의 높이를 고려하여 임계치 (T_{RPPC} = 라인 평균 높이×2) 보다 큰 높이의 라인이다. RPPC 알고리즘은 다음과 같다.

> ⓐ 해당 텍스트 라인의 수직방향 투영 프로파일(VPP: Vertical Projection Profile)을 구한다.
> ⓑ 라인의 좌측으로부터 처음으로 VPP 값이 0인 곳을 찾고, 그곳을 기준으로 영역을 두 개로 분리한다.
> ⓒ 분리된 두 개의 영역 중, 우측 영역의 HPP을 구한다. 라인의 분리가 가능한, 즉 HPP값이 0인 곳을 찾을 수 있으면 라인 분리점을 결정하고 텍스트 라인 분리 작업을 종료한다.
> ⓓ ⓒ에서 HPP값이 0인 곳을 찾을 수 없을 경우에는 우측영역에 대하여 ⓐ~ⓒ를 다시 수행한다.

위의 ③번 과정에서 라인 분리가 가능하면 RPPC 과정을 종료하고, 라인분리가 가능한 우측 영역만이 다음 단계인 단어 분리의 입력 데이터로 설정된다. 그림 3는 RPPC의 적용 예로써 반복적으로 수행됨을 보여준다. 영상을 수직 투영하면 '9'와 '7' 사이에서 프로파일 값이 0인 곳을 찾아서 그 지점을 기준으로 영상을 수직 분할한다. 분할된 우측 영상에 대해 수평 투영하여 프로파일 값이 0인 곳을 찾지 못하므로 앞선 과정을 반복한다. 그래서 라인 분리가 가능한 영상을 얻는다.

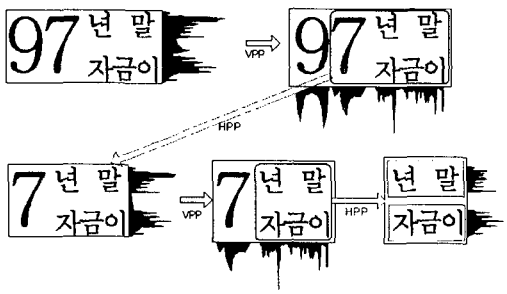

그림 3. RPPC 적용의 예

2.2 단어 분리

텍스트 영역에서 추출된 텍스트 라인으로부터 단어단위 영상을 분할한다. 분리된 텍스트 라인의 단어 분리를 위하여 단어들의 간격정보(IWG : Inter-Word Gap)이 필요하다. 그러나 텍스트 라인에는 IWG만 있는 것이 아니라 문자들의 간격정보(ICG : Inter-Character Gap), 심지어 단일 문자 속에서도 간격정보가 나타날 수 있다. 그러므로 텍스트 라인에서 단어 영상을 얻기 위해서는 모든 간격정보를 구하여 이들 gap에서 IWG를 분류하여야 한다. 또한 단어단위 영상 추출의 정확도를 높이기 위해서 단어 사이에 존재하는 특수기호를 검출한다.

2.2.1 Gap 정보 획득

이 단계에서는 텍스트라인에 존재하는 연결요소간의 gap 정보를 구한다. 임의의 텍스트라인에서 gap 정보를 추출하는 방법은 다음과 같다.

ⓐ 8방향 연결요소 분석방법을 통하여 CC와 CC의 외접 사각형(BB : Boundary Box) 정보를 구한다.

ⓑ ⓐ에서 구해진 사각형의 정보들을 이용하여 수직 방향으로 겹치는 사각형을 병합한다. 단, Gap 정보를 변형시킬 수 있는 ·, ,, ', " 등과 같은 특수기호는 병합과정 전에 (조건 1)과 같이 CC의 BB크기 분석을 통해 BB의 정보를 삭제함으로써 처리된다.

(조건1) $BB_{폭} < \frac{Line_{평균높이}}{4}$ 와 $BB_{높이} < \frac{Line_{평균높이}}{3}$

ⓒ BB간 수평거리를 계산하여 gap 정보를 추출한다.

2.2.2 gap 분류

단어 분리을 위해서 앞서 구해진 gap들을 IWG와 ICG으로 분류하고, 분류된 IWG 정보를 이용하여 단어간 분할 지점을 찾는 것이다. 그러나 텍스트 라인에는 다양한 크기의 ICG와 IWG가 존재하고, 언어의 종류에 따라 각각의 ICG와 IWG의 차이가 상당하다. gap 정보들이 문서 영상의 특성(언어, 해상도 등)에 따라 가변적이기 때문에 임계치를 이용한 gap 분류는 효과적이지 못하다[8]. 그래서 문서 영상의 특성에 관계없이 다양한 gap 정보를 ICG와 IWG로 분류하기 위해서 계층적 군집화 방법 중 평균-결합 클러스터링(average-linkage clustering) 방법을 사용한다[9]. 이 방법은 클러스터 구성원들간의 평균 거리를 정의함으로써 얻어진다. 그림 5의 수형도(dendrogram)는 6개의 gap이 두 개의 군집으로 묶여 가는 순서를 알파벳을 사용하여 보여주고 있다. 그림의 (e)와 같이 2개의 클러스터가 남았을 때 클러스터링 과정을 멈춘다. 위 예제의 gap들은 ICG = {2, 4, 4, 5}, IWG = {23, 25}로 분류된다.

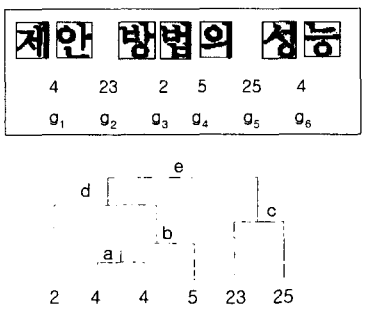

그림 5 수형도 - 평균 결합 클러스터링 적용 과정

2.2.3 특수 기호 검출

단어 영상 추출의 성능을 저하시키는 텍스트 요소 중에는 특수 기호(~, -, (,), {, }, [,] 등)가 있다. 일반적으로 특수 기호들은 단어와 단어 사이에 위치함으로써 단어 구분에 도움을 주지만 이는 인식기를 걸쳐 특수 기호로 판별되었을 때 가능한 일이다. 인식 전의 이들은 단어간의 gap 정보를 변질시키는 요소에 지나지 않는다. 다음은 특수 기호들에 의해 단어간의 gap 정보가 분명하지 못한 예제를 보여준다.

그림 6. 특수 기호로 인하여 단어 영상 추출 오류

그림 6의 ① ② ③은 단어 분할 지점인데도 특수 기호로 인하여 두 개의 단어가 하나의 단어가 되는 오류가 발생한다. 이런 오류에 대한 처리는 특수 기호를 찾아 그 지점을 IWG부분으로 설정을 하면 된다. 특수 기호 '~', '-' 등은 (조건 2)와 같이 연결요소의 높이와 높이에 대한 폭의 비율을 분석함으로써 특수 기호의 위치를 알아내어 단어 분리에 이용한다.

(조건 2) CC의 높이가 $Th \div 3.5$보다 크고 CC의 높이가 CC의 폭 2배보다 커야한다.

이때 임계치 Th는 BB 높이의 최빈값이다. 특수기호 '(', '{', '[' 등은 두 단계의 처리를 통해 검출한다. 단계 1에서는 연결요소의 크기를 분석하여 특수 기호와 유사한 CC를 선별적으로 검출한다. 이렇게 해서 검출되는 CC 후보군은 대략 '(', '{', '[' 등의 특수 기호들과 'I', 'l', '1', 'j', '|' 등의 문자들이 있다. 이들 중에서 특수 기호가 아닌 문자들은 단계 2에서 걸러진다.

단계 1. 후보군 검출 : CC의 폭과 폭에 대한 높이의 비율, CC의 블랙 픽셀의 개수와 CC의 넓이의 비율을 분석하여 특수 기호와 유사한 CC를 선정한다.

(조건 3) CC의 폭이 $Th \div 2.5$ 보다 크고 CC의 폭이 CC의 높이 2배보다 커야한다.

(조건 4) CC의 블랙 픽셀 합이 CC의 넓이 75%보다 적어야 한다.

단계 2에서는 특수기호 '(', '{', '[' 등이 수직으로 대칭을 이루고 있는 특성을 이용하여 특수 기호를 검출한다.

단계 2. 특수 기호 검출 : 특수 기호의 런(run)들이 수직으로 대칭적 형태를 이루는 특성을 이용(조건 5)하여 특수 기호를 검출한다. 그리고 특수기호는 아니지만 수직으로 대칭을 이루는 문자의 처리를 위하여 CC의 수직방향으로 중앙부분에 위치한 런이 좌우로 치우쳐 있는 정도를 이용(조건 6)한다.

(조건 5) CC 높이의 10% 길이에 해당하는 상위와 하위의 런들에 대한 BB들의 넓이를 각각 $High\_BB_{area}$, Low_BB_{area}라 한다면,

$$|HIGH\_BB_{area} - LOW\_BB_{area}| < Th \div 5$$

을 만족하지 않으면 특수 기호 후보에서 제외한다.

(조건 6) 수평방향으로의 CC 중앙을 CC_{Center}, 수직방향으로 CC 중앙에 위치한 런의 수평 방향 시작점과 끝점을 각각 x_1, x_2라 하였을 때,

$$MAX(CC_{Center} - x_1, x_2 - CC_{Center}) > Th \div 10$$

을 만족하는 CC는 특수 기호에서 제외시킨다.

3. 실험 결과 및 분석

본 논문에서 제안한 방법은 펜티엄-III 450 개인용 컴퓨터에서 Windows NT환경 하에 Visual C++ 6.0으로 구현되었다. 성능 평가를 위한 데이터는 '전남대학교 정보통신연구소 논문집 제3권'을 300dpi의 해상도로 스캔한 영상들 중, 임의로 추출한 50개의 텍스트 영역이다.

단 계	정확률 (분리 개수 / 총 개수)
특수 기호 제거 전	96.1% (4462/4643)
'-', '~' 제거 루틴 포함	97.29% (4517/4643)
'-', '~' 제거 루틴 + '{', '(', '[' 제거 루틴 포함	99.83% (4635/4643)

표 1. 단어 영상 추출 결과

텍스트 영역의 텍스트 라인 분리는 수평 방향 투영 프로파일 및 RPPC를 적용하여 100%의 정확도를 보였다. 그리고 단어 분리는 표 1과 같이, 특수 기호 검출에 대한 처리가 추가되면서 정확도가 개선되었다. 따라서 제안 시스템의 단어 단위 분할은 99.83%의 높은 정확도를 보였다. 그림 7은 입력 텍스트 영역을 단어 단위로 분할한 결과를 보여주는데, 특수 기호들이 성공적으로 처리되었음을 잘 나타내고 있다.

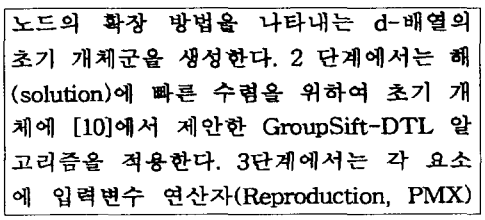

그림 7. 텍스트 영역

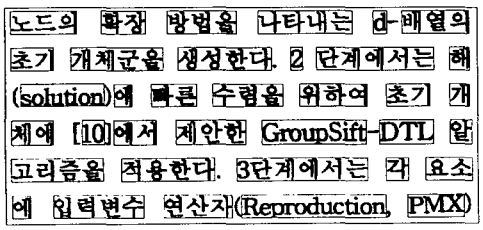

그림 8. 특수 기호를 처리한 단어 분리 결과 화면

4. 결론

본 논문에서는 주제어 검색(keyword spotting) 방법의 요소기술인 텍스트 영역의 단어 단위 분할 시스템을 제안하였다. 제안 시스템은 영상 전처리, 문서 구조 분석을 통해 추출된 텍스트 영역을 입력으로 단어 단위 분할을 수행하는데, 텍스트 영역에 대해 텍스트 라인을 분할하고 분할된 텍스트 라인을 단어 단위로 분할하는 계층적 접근 방법을 사용하였다. 텍스트 라인 분할은 수평 방향 투영 프로파일을 적용하여 분할 지점을 구하고, 단어 분할은 연결요소들을 추출한 후 연결요소간의 gap 정보를 구하고, gap 군집화 기법을 사용하여 단어 단위 분할 지점을 구하였다. 단어 단위 분할의 성능을 저하시키는 특수 기호에 대해서는 휴리스틱 정보를 이용하여 검출하였다. 제안 시스템의 성능 평가는 50개의 텍스트 영역에 적용하여 99.83%의 매우 높은 정확도를 얻을 수 있었다.

참고문헌

[1] AIIM'96 *Conference Handbooks*, Association for imaging and information methodologies, 1996.

[2] 정규식, 권희웅, "내용기반의 인쇄체 영문 문서 영상 검색을 위한 특징기반 단어 검색", 정보과학회논문지(B), 제26권, 제10호, pp.1204-1218, 1999. 10.

[3] F.R. Jenkins, T.A. Nartker and S.V. Rice, "Result of the fifth annual test of OCR technology by UNLV's Information Science Research Institute," Inform *Magazine*, pp.20-25, Sep. 1996.

[4] Yuan Y. Tang, Seong-whan Lee and Ching Y. Suen, "Automatic Document Processing: A Survey,", Pattern Recognition, Vol. 29, No. 12, pp. 1931-1952, 1996.

[5] 장명욱, 천대녕, 양현승, "연결화소를 이용한 문서 영상의 분할 및 인식," 한국정보과학회 논문지, Vol.20, No.12, pp. 1741-1750, 1993.

[6] 조현목, 이경무, 최영우, "Projection Profile을 이용한 새로운 자동 문서영상의 영역분리 및 분류 알고리즘". 영상처리 및 이해에 관한 워크샵, pp. 136-140, 1997.

[7] 김두식, 이성환, "한글과 영·숫자가 혼용된 문서를 위한 효과적인 문자 분할 방법," 제 8회 영상 처리 및 이해에 관한 워크샵 발표논문집, 제8권, pp. 19-26, 1996.

[8] Dr. Sargur N. Srihari, Stephen Lam, Dr. V. Govindaraju, Dr. Rohini Srihari and Dr. Jonathan Hull, "Document Understanding: Research Directions," CEDAR-TR-92-1. May 1992.

[9] E. Gose, R. Johnsonbaugh and S. Jost, Pattern recognition and image analysis, Prentice Hall, 1996.