문서 이미지에서 문자 추출과 3차원 면적-가중치 그래프를 이용한 단어 그룹핑

옥세영[O] 박환철 조환규
부산대학교 전자계산학과 그래픽스 응용 연구실

Text Extraction and Word Grouping using 3D Area-Weighted Graph in Document

Se-Young OK[O] Hwan-Chul Park Hwan-Gue Cho
Graphics Application Lab., Dept. of Computer Science, Pusan National University
{seok,hcpark,hgcho}@pearl.cs.pusan.ac.kr

요 약

이미지 분석이나 데이타 베이스 인덱싱 또는 종이 문서를 전자 문서화 하는 문제는 컴퓨터 비전 응용분야에서 중요 관심사가 되어 왔다. 이러한 문제들을 처리하기 위해서는 제일 먼저 이미지와 문자가 혼합되어 있는 문서에서 자동으로 문자와 이미지들을 분리해내는 과정이 필수적이다. 본 논문에서는 신문이나 광고등에서 볼 수 있는 이미지, 음각 문자와 양각 문자가 섞여 있는 문서에서 문자만을 추출하는 알고리즘을 제안한다. 이 알고리즘은 Run-length code를 이용하여 문자나 이미지의 경계선(bound) 모양의 특징을 추출하여 음각 문자와 이미지, 양각 문자를 구분한다. 그리고 추출된 글자들을 3차원 공간상에 매핑한 후 3차원 면적 가중치 그래프를 이용하여 관련된 단어들로 묶어주는 3차원 그룹핑 알고리즘을 제시한다. 실험결과로는 추출된 문자와 그룹핑된 결과를 보여준다

1. 서론

래스터 문서 이미지에서 문자 정보를 추출하는 방법에 대해서는 오래전부터 많은 연구가 이루어 졌다. 최근에 인터넷과 같은 네트웍의 발달로 기존에 종이로 구성된 자료들에 대해 전자 문서화의 요구가 늘고 있다 이때 일간지나 광고지등 기타문서들을 전자 문서화할때 사람이 일일이 수작업으로 재구성하는 것은 비용이 많이 요구될 뿐만 아니라 비능률적이다. 그래서 본 논문은 잡지나 신문 등의 문서에서 문자 인식(OCR)없이 자동적으로 문자와 문자 이외의 요소들을 추출하고자 한다. 문자의 위치를 추적해내는 방법에는 Gabor filtering과 같은 texture 분석 방법이 있는데, 이 방법은 문자의 크기와 종류에 의존적이며 처리 시간이 많이 소요되어 비효율적이다[1] 연결요소(connected component) 추출에 기반을 두고 피라미드(Pyramid) 탐색방법을 사용하여 빠른 시간내에 효율적으로 문자들을 추출해서 단어로 연결시켜주는 방법이 제안되었다 그러나 이 방법은 관련 단어를 연결할때 정확성이 떨어진다 [3]. 비슷한 방법으로 연결요소를 추출한 후 이 요소들을 병합(merge) 시킴으로써 얻어지는 정보를 통해 문자를 추출할 수 있다[4]. 이렇게 지금까지 다양한 문자 추출 방법들이 제시되고 있으나 그림과 비슷한 크기의 문자 추출이나, 신문의 헤드 라인같은 음각 문자의 추출에 관한 방법은 많이 제시되어지지 않았다. 따라서 양각 문자와 음각 문자가 동시에 존재할 경우 각각을 동시에 문자로 추출할 수 있다면 더 유용하게 사용될 것이다.

따라서 본 논문은 신문이나 광고지등과 같은 문서 이미지에서 연결 요소 분석 방법을 사용하여 문자 위치를 찾아내고 개개의 문자들을 연결된 단어로 인식할 수 있는 알고리즘을 소개한다.

전반적인 과정을 간략히 소개하면 다음과 같다.
1 단계 : 래스터 이미지에서 8-방향 연결 요소를 생성 하여 정보를 저장한다.
2 단계 . 각 연결 요소의 Run-length를 이용한 경계선(boundary shape)의 규칙성(regularity)을 검사하여 양각 문자, 음각 문자, 그림, 잡음등으로 분류한다.
3 단계 : 앞서 구해진 문자들에 대해서 3차원 면적 가중치 그래프를 이용해 단어로 연결시킨다.

2. 문자 추출

2.1 연결 요소(connected component) 추출

흑백 이미지에서 문자를 추출하기 위해서 먼저 8-방향 연결 요소 생성 알고리즘을 사용하여 연결 요소(C)를 구한 후 각 연결요소에 대한 다음과 같은 정보들을 얻어낸다.

- 연결 요소를 포함하는 사각형의 좌표(P_{max}, P_{min})
- 검정 픽셀의 갯수(C_{num}), 밀집도(C_{num}/C_{area})
- 가로 대 세로의 비율 (w/h)

입력 이미지가 일반 문서를 스캐닝한 데이타이므로 여러가지 예상하지 못한 잡음들이 생기는데, 문자 추출 알고리즘을 적용하기 전에 구해진 C_{area}가 주어진 한계값보다 작을 경우 잡음(noise)으로 보고 제거한다

2.2 연결 요소의 분류

이 장에서는 각 연결 요소에 대해서 양각 문자, 음각문자, 그림 등으로 분류해서 문자만을 추출하는 알고리즘을 설명하기로 한다 문서에 나타날 수 있는 긴 라인은 연결 요소의 가

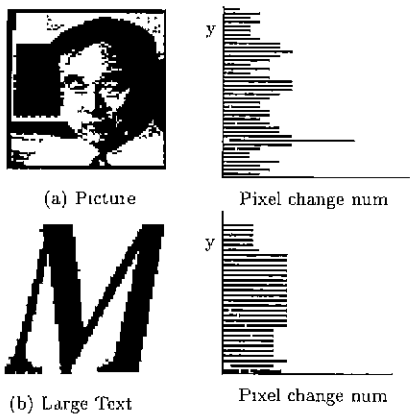


그림 1: 그림과 큰글자의 B/W 픽셀 변화 수

로 대 세로 비율과 픽셀 밀집도를 이용하여 제거할 수 있다. 그림과 음각 문자, 양각 문자를 구별함에 있어서 가장 어려운 점은 두 가지로 분류된다. 하나는 그림과 같은 크기의 양각 문자를 어떻게 문자로 인식할 수 있는냐와 또 다른 하나는 그림과 음각 문자와의 구분이다. 이 두가지의 문제를 해결하기 위해서 다음과 같은 알고리즘을 제시한다.

각 연결 요소의 경계선모양(boundary shape)의 정규성(regularity)을 검사함으로써 양각 문자와 그림,음각 문자를 구분할 수 있다. 일반적으로 문자의 경우에는 그림1(b)와 같이 바깥 경계선의 모양의 변화가 일정하다. 그러나 그림1(a)같은 그림의 경우 경계선을 따라 탐색을 해보면 매우 불규칙적으로 경계선의 모양이 변한다는 것을 알 수 있다. 이렇게 경계선의 정규성을 알아내기 위해서 x축으로 프로파일링을 한후 Run-length를 구하여 문자와 그림을 정확히 구분할 수 있다. 이 방법을 이용한 문자의 추출 알고리즘은 아래와 같다.

Algorithm : 프로파일링 알고리즘
Input 연결 이미지
Output. 문자인지 그림인지의 여부
1 각 연결 요소에 대해 x축으로 프로파일링하면서 Run-length 정보를 구한다
(a) $flag_i = C_i$(연결 요소)에 대한 x축으로 픽셀의 색의 변화 횟수
(b) μ_{flag}, σ_{flag}를 구한다
2. 연결요소에 대해 아래와 같은 조건을 검사한다
If $((C_{width} \quad C_{height}) \geq l)$ **then**
If $((\sigma_{flag}/\mu_{flag}) \leq 1)$ **then**
C_i = 문자
C_i에 해당하는 영역을 반전시키고,그 부분에 대해 다시 연결요소를 구한다
Else C_i = 그림
delete C_i
End
Else then
If $((\sigma_{flag}/\mu_{flag}) \leq 1)$ **then**
C_i = 문자
Else C_i = 그림
delete C_i
End

위 알고리즘에서 연결 요소의 x축 프로파일링할 때 검정에서 흰색으로 또는 반대로 픽셀 값이 변하는 수($flag_i$)의 평균(μ_{flag})과 분산(σ_{flag})은 다음과 같은 방법으로 얻을 수 있다.

$$\mu_{flag} = \frac{\sum_{i=P_{min\ y}}^{P_{max\ y}} flag_i}{(P_{max\ y} - P_{min\ y})}$$

$$\sigma_{flag} = \frac{\sum_{i=P_{min\ y}}^{P_{max\ y}} (flag_i - \mu_{flag})^2}{(P_{max\ y} - P_{min\ y})}$$

알고리즘의 이해를 돕기 위하여 그림1을 예로 들어 설명하기로 한다. 문자의 경우는 문자를 둘러싸고 있는 바깥 경계선을 탐색해보면 픽셀들의 변화량이 매우 규칙적이라는 것을 알 수 있다. 즉 경계선이 정규성을 띄기 때문에 x축의 각 라인에서 Run-length가 바뀌는 횟수 즉 $flag_i$가 규칙적이다. 그림 1의 (b)의 픽셀의 색의 변화 플래그에 대한 그래프를 보면 플래그의 값들이 균일하다는 것을 알 수 있다 반면에 그림의 경우에는 그림의 경계선을 따라가보면 불규칙한 특성을 지니고 있다. 바꾸어 말하면 그림은 다양한 Run-length의 집합으로 되어 있다는 뜻이다. 그러므로 그림의 경우에는 픽셀 값의 전환 횟수, $flag_i$가 커지게 된다. 그림 1의 (a)의 그래프를 보면 $flag_i$의 값들이 다양하다는 것을 알 수 있다. 따라서 문자는 일정한 변화값을 수용하므로 플래그값들의 흩어져 있는 정도(분산, σ_{flag}) 가 작고, 그림의 경우는 반대로 Run-length 차가 크기 때문에 (하나의 픽셀에서 연결 요소의 폭에 이르기까지 다양한 Run-length) μ_{flag}이 커진다 여러번의 실험을 통해 분산과 플래그의 평균을 이용하여 그림과 문자로 구분이 가능하다. 특히 음각 문자는 연결 요소의 밀집도가 크고 가로, 세로의 비가 다르므로 이점을 활용하였다 음각 문자는 단어 연결을 위해 반전시켜 양각 문자로 바꾸었다

3. 3차원 면적 가중치 그래프를 이용한 단어의 연결

문서에서 추출한 문자들의 크기와 위치정보를 이용하여 3차원 면적 가중 우선 그래프를 만들게되면 각 문자들을 연결하는데 필요한 많은 기하학적인 정보들을 얻을 수 있다.

3.1 단어 연결의 전처리 단계

추출된 문자들을 살펴보면 아주 작은 크기의 문자들이 많이 있다. 이렇게 작은 문자들이 인접하게 존재하는 데이타를 스캐닝하면 이들 문자사이의 가까운 거리로 인해서 하나의 연결 요소로 나타난다. 그러므로, 그룹핑을 하기전에 정확한 단어 연결을 위하여 붙어 있는 문자들을(하나의 연결 요소를) 데이타에서 여러 개의 문자로(여러 개의 연결 요소로) 끊어 주는 과정이 필요하다

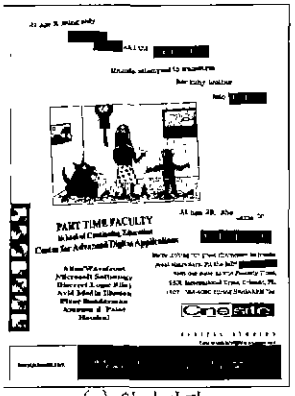
3.2 3차원 면적 가중치 그래프의 생성

이미지 데이타의 문자들은 크기와 문자들이 놓여 있는 방향이 매우 다양하다. 따라서, 연관성 있는 단어를 정확하게 연결하는것은 매우 어렵다. 3차원 면적 가중치 그래프 $G(V,E)$의 주 아이디어는 각 문자v_i를 3차원 공간상의 좌표(x_i, y_i, z_i)로 위치시키는 것이다. 여기서 x와 y좌표는 2차원 공간상에있는 문자들의 중심 좌표이고, z좌표는 전체 문자의 면적에 대한 분산과 각 문자의 면적의 비로써 표현된다. 이렇게 하면 문자들이 서로 교차하는 경우, 문자의 면적에 가중치를 주어 엉뚱한 단어로 오인되는 결과를 방지할 수 있다. z좌표는 아래의 식으로 표현된다. Υ은 실험에 의하여 나온 값이다.

$$V = \{v_i \parallel position = (x_i, y_i, z_i)\}$$

$$z_i = \frac{Area(c_i) \quad \Gamma}{\Upsilon}, \ where \Upsilon = \frac{1}{n}\sum_{k=1}^{n} Area(c_k),$$

$$\Gamma = 1/\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Area(v_i) - \frac{1}{n}\sum_{k=1}^{n} Area(c_k))^2}\right)$$


(a) 원이미지

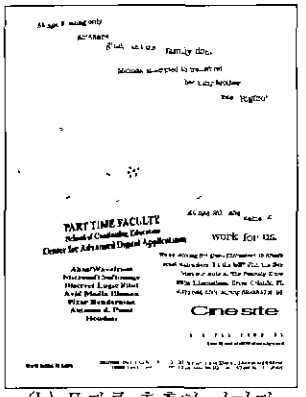
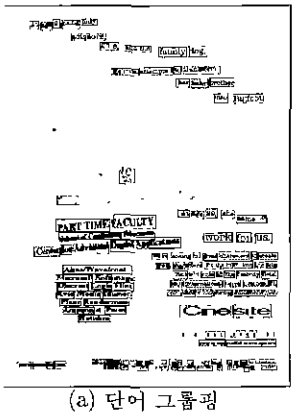
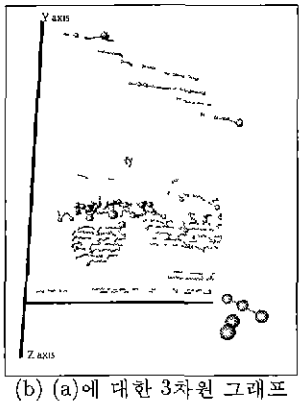

(b) 문자를 추출한 이미지

그림 2: 문자 추출


(a) 단어 그룹핑



(b) (a)에 대한 3차원 그래프

그림 3: 3차원 가중치 그래프를 이용한 단어 그룹핑

$G(V,E)$의 E_i는 위에서 만들어진 노드(V)들의 3차원 공간상에서 가장 가까운 거리에 있는 노드들에 우선순위를 주어 두번째 노드까지 연결한다.

3.3 3차원 그래프의 깊이 우선 탐색

$G(V,E)$에서 연관된 문자들의 집합 즉, 단어가 될 수 있는 후보를 결정하기 위하여 서브 그래프 G_{dfs}를 찾는다. $G(V,E)$를 깊이 우선 탐색을 하면서 연결된 노드들간의 상대적인 거리와 기울기를 이용해서 G_{dfs}를 찾는다. 두 에지 사이의 각을 $\theta_i = \angle v_{i-1}v_iv_{i+1}$라고 할 때 단어로 연결될 수 있는 기울기의 허용 범위는 아래와 같이 구할 수 있다. δ_0는 실제 문자들이 직선이 아닌 곡선으로 연결되어 있는 경우의 에러 허용값이다.

$$\theta_{avg} = \sum_{j=1}^{i-1} \theta_j / n, \frac{1}{n} \sum_{i=1}^{k} (\theta_{avg} - \theta_i)^2 \leq \delta_0$$

3.4 클러스터내의 단어 연결

마자막으로 위에서 만들어진 G_{dfs}에서 정확한 단어끼리 연결시켜 주는 작업을 한다. 전체 G를 각각 클러스트링한 결과가 G_{dfs}인데, 각 클러스트, 즉 G_{dfs}내에서 각 노드들간의 거리에 대한 정규성(regularity)을 검사해서 규칙에서 벗어나는 노드는 제거해서 완전한 단어로 연결시킨다.

4. 실험결과

본 논문에서 제시한 문자 추출 알고리즘과 단어 그룹핑 알고리즘의 성능 평가를 위하여 Visual C++을 이용한 Pentium 133에서 실험을 하였다. 음각 문자와 양각 문자가 동시에 존재하고 문자의 크기와 방향이 다양한 신문이나 잡지등을 데이타로 사용하였다. 그림2(a)는 원래의 이미지 데이타이고 그림2(b)는 추출된 문자들의 결과를 보여준다. 음각 문자부분은 문자로 인식해서 해당 부분을 반전시켜 양각 문자의 형태로 나타내었다 그리고 그림3(a)는 그림2(b)에 대해 3차원 그룹핑을 한 결과로 연결된 단어들을 보여준다. 그림3(b)는 그림3(a)의 각각의 연결된 문자들을 3차원 공간상에 위치시킨 3차원 면적 가중치 그래프를 보여준다. 각 문자들의 면적에 따라서 z축상에서 다른 높이에 위치하게 됨을 볼 수 있다.

사용한 데이타들에 대한 실험 결과는 아래의 표와 같다.

Name	Size	ESR	FRR
$Data_1$	480 × 890	94	2.6
$Data_2$	885 × 614	96	0
$Data_3$	747 × 992	96	1.4
$Data_4$	847 × 800	97	1.7
$Data_5$	842 × 976	99	2.1

위 표는 논문에서 제시한 알고리즘으로 여러 데이타들을 검사한 결과이다. ESR(Extract Success Rate)은 추출한 문자에 대한 그룹핑의 정확성을 나타내고, FRR (False Recognition Rate) 은 문자로 잘못 인식된 정도를 나타낸다.

5. 결론

본 논문은 그림과 문자가 섞여 있는 신문이나 잡지의 광고 등의 문서에서 문자들을 추출해 단어로 연결하는 시스템을 제안하였다. 이 알고리즘은 Run-length code를 이용하여 그림과 문자들의 바깥 경계선(bound shape)의 규칙성(regularity)을 검사하는 알고리즘으로 양각 문자와 음각 문자, 문자 이외의 요소를 분리해 내었다. 이렇게 추출된 문자들을 의미있는 단어들로 연결시키기 위하여 각 문자에 대해 3차원 공간상에 위치시키는 3차원 면적 가중치 그래프를 이용하여 그룹핑하는 알고리즘을 제시하였다. 문자 추출이나 그룹핑은 문자의 종류나 크기, 문자가 기울어져 있는 방향에 상관없이 성공적으로 결과를 보여주었다. 그러나, 데이타에서 그림이 기호같이 매우 단순한 모양일땐 문자로 오인되는 오류가 있다 본 논문은 신문, 잡지등의 문서와 공학 설계도면들을 저장하거나 분석할 때 유용하게 사용될 수 있다. 앞으로 영어뿐만 아니라 한글로 된 문서에 적용될 수 있도록 연구할 것이다.

참고문헌

[1] A.Jam and S Bhattacharjee. Text segmentation using gabor filters for automatic document processing In *Mach. Vision Applic*, volume 5, pages 169–184, 1992

[2] N AMAMOTO, S TORIGOE, and Y HIROGAKI. Black segmentation and text area extraction of vertically/horizontally written document In *ICDAR '93 Proc*, pages 739–742, Oct 1993

[3] C L TAN and P.O.NG Text extraction using pyramid. In *Pattern Recognition '97 Proc*, volume 31, pages 63–72, Feb 1997

[4] T Saito, M. Tachikawa, and T Yamaai Document iamge segmentation and text area ordering. In *ICDAR '93 Proc*, pages 323–329, Oct 1993