k-평균 클러스터링을 이용한 필기 문서 영상의 단어 분리법

류제웅, 조남익
서울대학교 전기정보공학부
youjw@ispl.snu.ac.kr, nicho@snu.ac.kr

Word Segmentation Algorithm for Handwritten Documents based on k-means Clustering

Jewoong Ryu, Nam Ik Cho
Seoul National University

요 약

본 논문에서는 필기 문서 영상을 분석하여 단어 단위로 요소들을 분할하는 방법을 제안한다. 일반적으로 인쇄 문서에 비하여 필기 문서에서는 글자 간 간격이 일정하지 않을 뿐만 아니라 필기자 또는 작성된 언어에 따라 특성이 매우 다르게 나타나기 때문에 단어를 분리하는 것은 어려운 문제로 간주되었고 많은 연구가 진행되었다. 제안하는 방법은 이 문제를 해결하기 위하여 글자 획의 두께를 고려하여 정규화시킨 각 연결 요소간 간격과 간격 안에 존재하는 글자 픽셀의 수로 구성된 2 차원의 특징값을 추출하였다. 이 특징값을 바탕으로, 제안하는 방법은 k-평균 클러스터링을 이용하여 각 텍스트라인을 구성하는 연결 요소간 간격을 단어 사이의 간격과 단어 내부 글자간의 간격으로 분류하였다. ICDAR 2013 Handwriting Segmentation Contest 데이터베이스에 대한 실험 결과 제안하는 방법은 가장 우수한 성능을 나타내었다.

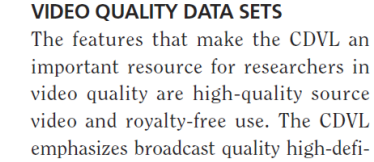
VIDEO QUALITY DATA SETS
The features that make the CDVL an important resource for researchers in video quality are high-quality source video and royalty-free use. The CDVL emphasizes broadcast quality high-defi-

그림 1 인쇄문서 영상의 예

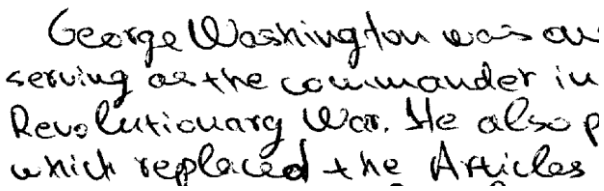

그림 2 필기 문서 영상의 예

1. 서론

스캐너 또는 카메라로 취득된 문서 영상을 텍스트 라인 및 단어 단위로 분리하는 과정은 문서의 구조 분석 및 오프라인(off-line)으로 수행하는 광학 문자 인식(Optical Character Recognition)등에 매우 필수적인 과정이다[1]. 그런데, 기계로 인쇄된 문서는 그림 1 과 같이 글자 간 간격 및 텍스트 라인간의 간격이 일정하여 각 요소 별로 분리가 비교적 쉬운 반면, 필기 문서 영상의 경우 그림 2 와 같이 텍스트 라인 및 문자 간 간격이 일정하게 나타나지 않을 뿐만 아니라, 필기자 및 작성된 언어에 따라서도 특성이 크게 다르게 나타나게 된다. 따라서, 이는 어려운 문제로 간주되었으며 관련된 연구가 진행되어 왔다 [2]-[5].

기존 연구에서 필기 문서 영상에서 단어 단위로 분리하는 문제는 주로 문서 영상을 텍스트 라인 별로 분리하고 이를 입력으로 이용하여 수행된다. 이렇게 분리된 각 텍스트 라인을 구성하는 글자들 사이의 간격을 단어 사이의 간격 (inter-word gap)과 단어 내부 글자의 간격 (intra-word gap)으로 분류하고, 그 결과를 바탕으로 단어 분리를 수행하였다. 일반적으로는 글자 단위의 픽셀 집합을 표현하기 위하여 연결 요소(Connected component)를 추출한 뒤, 각 요소간 간격을 분류하는데, 주로 가우스 혼합 모델(GMM)[2,5], SVM[3] 등을 이용하여 수행하였다. 이러한 기존 방법들은 사전 학습이 필요하므로 다른 종류의 데이터베이스마다 따로 학습이 필요하다는 단점이 있다.

본 논문에서는, 글자 획의 두께로 정규화한 연결요소간 간격과 세로방향 투영 영상을 이용한 두 가지 특징 값을 이용하여 단어 분리를 수행하는 알고리즘을 제안한다. 먼저, ICDAR 2013 Handwritten Contest[6]에서 가장 좋은 성능을 보인 텍스트 라인 분리 알고리즘[7]을 이용하여 주어진 필기 영상을 텍스트 라인 단위로 분리하고, 연결 요소를 추출한다. 그리고 글자 자획으로 정규화한 간격 및 간격 내부에 존재하는 글자 픽셀 수에 해당하는 특징 값을 추출한 뒤, 이를 k-평균 클러스터링 알고리즘을 이용하여 단어 내부 글자 간격과 단어 간 간격으로 분류하여 단어 분리를 수행한다. 제안하는 방법은 글자 획의 두께를 기반으로 정규화를 수행한 특징 값을 이용하므로, 스케일의 변화에 강인하고 각 텍스트 라인에서 k-

평균 클러스터링 방법을 이용하여 분리를 수행하므로 사전 학습이 필요하지 않다는 장점이 있다. ICDAR 2013 Handwriting Segmentation Contest 데이터베이스에 대한 실험 결과, 제안하는 방법은 FM 90.12%로 가장 좋은 성능을 나타내었다.

2. 제안하는 알고리즘

2.1 문제 설정

문서 영상에서 단어 분리는 주로 텍스트 라인 검출 알고리즘의 결과를 이용하므로, 단어 분리 알고리즘의 입력은 주로 한 줄의 텍스트라인 영상이다. 한 텍스트라인 입력 영상에서 추출한 연결 요소를 c_j ($j=1,2,3,...$) 이라고 하고, 이들이 왼쪽에서 오른쪽으로 정렬되어 있다고 가정하자. 그리고 이웃한 연결 요소인 c_j 와 c_{j+1} 에 사이의 간격(gap)을 g_j 라 정의하자. 그러면, 단어 분리 문제는 레이블,

$$L=\{l_j\}_{j=1}^{N} \tag{1}$$

을 연결 요소간 간격의 집합인

$$G=\{g_j\}_{j=1}^{N} \tag{2}$$

로 할당하는 문제로 볼 수 있다. 여기서 $l_j \in \{0,1\}$ 이고 (1: 단어 간 간격, 0: 단어 내부 글자간 간격) N 은 해당 텍스트 라인의 간격의 수 이다.

2.2 특징 값 추출 (Feature extraction)

이번 절에서는 연결 요소간 간격 g_j 에 해당하는 특징 값 추출법을 설명한다.

2.2.1 정규화된 글자간 거리

추출한 연결 요소 간의 간격 g_j 에서 단어 분리를 위해 쓸 수 있는 가장 큰 특징은 이웃한 연결요소 c_j 와 c_{j+1} 사이의 거리이다. 그런데, 이 거리는 스캔 해상도에 따라 달라질 수 있으므로, 본 논문에서는 문서 영상에서 글자 획의 평균 두께인 $\bar{W}$ 을 추정하여 이를 정규화 하여 이용한다. $\bar{W}$ 을 추정하기 위하여, 먼저 각 연결요소 c_j 에 해당하는 글자의 두께인 W_j 을 각각 추정하고, 이를 평균 내어 이용한다. 먼저, W_j 는 c_j 에서 뽑은 n 개의 점 p_k 를 중심으로 4 개 방향으로 측정한 길이 $W_d(p_k)$ 중에서 가장 짧은 길이가 평균 자획일 가능성이 높으므로,

$$W_j=\frac{1}{n}\sum_{k=1}^{n}\min_{d\in D}\left(W_d(p_k)\right) \tag{3}$$

와 같이 추정할 수 있다. 여기서 D 는 {동-서, 남-북, 동남-서북, 동북-서남} 의 4 방향이다.

이렇게 구한 $\bar{W}$ 를 바탕으로 g_j 에 해당하는 정규화된

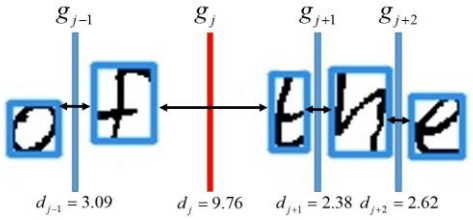

그림 3 d_j 의 몇 가지 예

거리 d_j 를 구할 수 있다. $L_x(\cdot)$, $R_x(\cdot)$ 을 각각 해당하는 연결요소의 왼쪽 끝 x 좌표와 오른쪽 끝 x 좌표라고 하자. 그러면 정규화된 거리 d_j 는,

$$d_j=\frac{L_x(c_{j+1})-R_x(c_{j+1})}{\bar{W}} \tag{4}$$

로 구할 수 있다. 그림 3 에 몇 가지 d_j 의 예제를 나타내었다. 그림에서 볼 수 있듯이, 단어 간 간격에 해당하는 g_j 의 정규화된 거리인 d_j 가 다른 것들에 비해 확연히 크게 나타나는 것을 확인할 수 있다.

2.2.1 간격 내부에 존재하는 글자 픽셀의 수

앞 절에서 이용한 d_j 는 해당 g_j 가 단어 간 간격에 해당하는지 판단하기에 좋은 특징이긴 하지만, 간격 사이에 잡음이 존재하거나 필기 방식으로 인하여 자획의 두께가 매우 얇게 나타나는 경우에 잘못된 정보를 제공하게 된다. 본 논문에서는 여기서 발생하는 오류를 방지하기 위하여, g_j 중심에 존재하는 글자 픽셀의 수를 나타내는 코스트 p_j 을 이용하여 두 번째 특징 값으로 이용한다. 먼저 입력된 텍스트 라인 영상에서 글자 픽셀의 분포를 얻기 위하여,

$$I_p(x)=\sum_{y=1}^{h}\frac{I(x,y)}{\bar{W}}, x=1,\ldots,w \tag{5}$$

와 같이 평균 자획의 두께로 정규화한 세로방향 투영 영상을 얻는다. 여기서 $I(x,y)$ 는 입력된 텍스트라인 영상의 해당하는 (x,y) 좌표이며, (h,w) 는 각각 영상의 높이와 너비를 나타낸다. d_j 와 마찬가지로 스캔 해상도에 영향을 받지 않게 하기 위하여, $\bar{W}$ 로 정규화 하였다. 이렇게 얻어진 투영영상 $I_p(x)$ 을 바탕으로 간격 내부에 존재하는 글자 픽셀의 수를 추정하게 되는데, g_j 의 중심인 $\bar{x}_{g_j}$ 에 가까운 픽셀일수록 더 고려를 해야 하므로, 가우시안 커널 $\kappa_j(i)\sim N(\bar{x}_{g_j},\sigma^2)$ 을 이용하여 중심에 가까운 글자 픽셀이 코스트 p_j 에 미치는 영향이 크도록 하였다. 위와 같은 관측을 바탕으로, 간격 내부의 존재하는 글자의 픽셀 수를 반영하는 코스트 p_j 는

$$p_j=\sum_{i=1}^{M}\kappa_j(i)I_p(\bar{x}_j-\lfloor M/2\rfloor+i-1), \tag{6}$$

로 계산할 수 있다. 여기서 M 은 가우시안 커널의 크기로 $4\sigma+1$로 설정하였다.

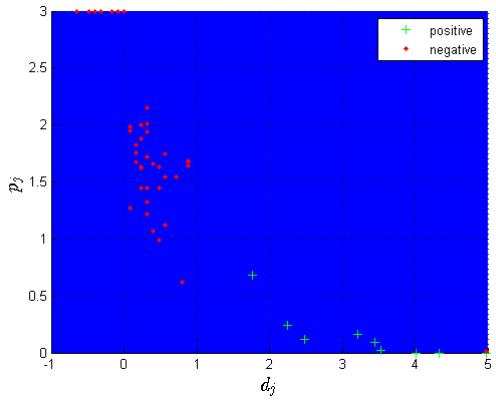

그림 4 텍스트 라인의 (d_j, p_j) 예. 녹색 점은 단어간 간격에 해당하는 g_j 이고 빨간 점은 단어 내부 글자간 간격에 해당한다.

2.3 k-평균 클러스터링

이번 절에서는 두 개의 특징 값 d_j 와 p_j 를 이용하여 g_j 의 레이블인 l_j 을 할당하는 방법을 서술한다. 일반적으로 한 장의 필기 문서 영상은 보통 15 줄 내외의 텍스트 라인으로 구성되어 있고, 한 줄의 텍스트 라인은 보통 수십 개의 g_j 로 구성되어 있다. 그런데, 필기 문서 영상은 필기자 및 사용한 언어에 따라 특성이 크게 다르게 나타나고, 심지어는 같은 문서 영상 내에서도 부분적으로 특성이 변화한다. 따라서, 연결 요소간의 간격 g_j 에 레이블을 할당하여 단어 분리를 수행하기 위해 문서 전체를 고려하기 보다는 부분적으로 판단하여 할당하는 것이 좀 더 성능을 높일 수 있다. 게다가, 단어 분리 알고리즘의 입력 영상이 텍스트 라인 분리 알고리즘으로 추출된 각 텍스트 라인이므로, 한 텍스트 라인에 대하여 특징값을 추출한 뒤, 해당하는 g_j 의 레이블인 l_j 을 설정하는 것이 효율적이다. 그림 4 에 한 텍스트 라인에 해당하는 g_j 의 2 차원 특징값 (d_j, p_j) 을 나타내었다. 그림에서 볼 수 있듯이, 하나의 텍스트 라인에서는 특성이 크게 바뀌지 않기 때문에 비교적 쉽게 분리될 수 있다. 이런 특성을 가지는 데이터들의 경우 비교사 분류 (unsupervised clustering) 방법을 이용하여 분류 할 수 있는데, 본 논문에서는 k-평균 클러스터링 알고리즘[8]을 이용하였고 g_j 에 해당하는 l_j 가 2 가지이므로 $k=2$ 로 설정하였다. k-평균 클러스터링을 그림 4 에 데이터에 적용한 결과를 그림 5 에 나타내었다. 일반적으로 단어간 간격일수록 d_j 는 큰 값을 가지고 p_j 는 작은 값을 가지므로, 위 그림상에서 우측 아래에 위치한 클러스터에 $l_j=1$ 을 할당하고 다른 클러스터에 $l_j=0$ 을 할당하였다.

3. 실험 결과

제안하는 단어 분리 알고리즘의 성능을 평가하기 위해, ICDAR 2013 Handwriting Segmentation Contest [6] 데이터

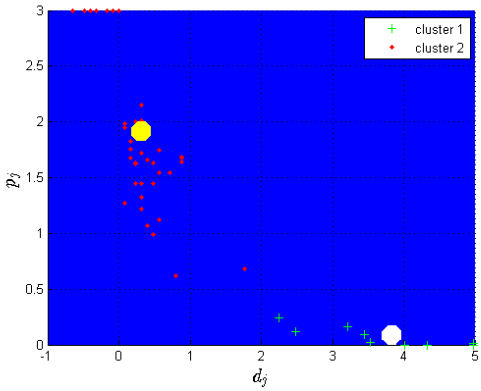

그림 5 k-평균 클러스터링 결과.

표 1 실험 결과. 다른 실험 결과들은 [6]에서 참조.

알고리즘	DR(%)	RA(%)	FM(%)
CUBS [5]	87.86	86.91	87.38
GOLESTAN-a	89.66	90.44	90.05
GOLESTAN-b	89.59	90.07	89.83
INMC	88.18	90.36	89.26
LRDE	86.75	86.94	86.85
MSHK	75.93	83.94	79.73
NUS	87.28	**91.07**	89.13
QATAR-a	88.19	83.10	85.57
QATAR-b	87.94	80.52	84.07
NCSR [2]	88.31	90.98	89.62
ILSP [3]	87.93	88.37	88.15
TEI [4]	87.15	88.15	87.65
제안하는 방법	**90.38**	89.85	**90.12**

베이스에 대하여 실험하였다. 이 데이터베이스는 다양한 필기자에 의해 영어, 그리스어 및 벵갈어로 작성된 총 150 장의 필기 문서 영상으로 구성되어 있다. 모든 영상은 스캔을 수행한 뒤 이진화(binarization)되어 있으며, 각 필기 영상에 대하여 픽셀 단위로 진리값(ground truth)이 설정되어있다. 평가를 위하여 [6]에서 이용한 방법인,

$$\text{MatchScore}(m,n)=\frac{|G_m \cap R_n|}{|G_m \cup R_n|} \qquad (7)$$

을 이용하였다. 여기서 G_m 은 실제 m 번째 단어로 설정된 픽

(a) (b)

그림 6 단어 분리 예제. (a) 입력 문서 영상, (b) 제안하는 알고리즘의 결과.

셀들의 집합이고, R_n 은 알고리즘 결과 n 번째 단어로 분리된 픽셀의 집합이다. MatchScore 가 0.9 이상일 때, 분리된 결과와 참값이 일대일 대응($o2o$)이라 간주하고,

$$DR = \frac{o2o}{N}, RA = \frac{o2o}{M}, FM = \frac{2 \times DR \times RA}{DR + RA} \quad (8)$$

와 같이 검출율(DR), 인식 정확도(RA) 및 F-measure (FM)을 정의하고 최종 성능 비교는 FM 으로 수행하였다. 표 1 에 제안하는 알고리즘과 다른 방법들의 결과를 나타내었다. 실험 결과, 제안하는 알고리즘은 ICDAR 2013 Handwriting Segmentation Contest 에 참가한 9 개의 알고리즘을 포함하여 기존 3 편의 논문[2-4]에 대해서도 가장 좋은 성능을 보이는 것을 확인하였다. 또한 학습 데이터베이스를 이용하여 기계학습을 수행한 [3]에 비해서도, 제안하는 방법은 별다른 학습 과정 없이 비교사 분류 방법을 이용하여 더 좋은 성능을 얻을 수 있었다. 알고리즘 수행시간은 1 장의 문서 영상당 텍스트 라인 검출 시간을 포함하여 평균 3~4 초 가량이 소요되었다. 그림 6 에 본 알고리즘의 결과 예제를 나타내었다.

4. 결론

본 논문에서는 필기 문서 영상을 위한 단어 분리 알고리즘을 제안하였다. 제안하는 방법은 먼저 텍스트 라인 입력 영상에서 연결 요소를 추출하고, 각 연결 요소간 간격에 대해 2 가지 특징값을 스캔 해상도에 관계없도록 정규화하여 추출하였다. 이렇게 추출된 특징 값을 이용하여 각 텍스트 라인에 대해 k-평균 클러스터링 알고리즘을 이용하여 단어 간 간격 및 단어 내부 글자의 간격에 해당하는 것을 분리하여 필기 영상에서 단어를 성공적으로 분리하였다. ICDAR 2013 Handwriting Segmentation Contest 데이터베이스에 대한 실험 결과, 제안하는 방법은 FM 90.12%로 가장 좋은 성능을 보였다.

참고문헌

[1] L. O'Gorman, "The document spectrum for page layout analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1162–1173, Nov. 1993.

[2] G. Louloudis, B. Gatos, I. Pratikakis and C. Halatisis, "Text line and word segmentation of handwritten documents", *Pattern Recognition*, vol. 42, no.12, pp.3169-3183, 2009.

[3] T. Stafylakis, V. Papavassiliou, V. Katsouros and G. Carayannis, "Handwritten document image segmentation into text lines and words", *Pattern Recognition*, vol. 43, no. 1, pp. 369-377, 2010.

[4] A. Nicolaou and B. Gatos, "Handwritten Text Line Segmentation by Shredding Text into its Lines", in *10th International Conference on Document Analysis and Recognition*, 2009, pp. 626-630.

[5] Z. Shi, S. Setlur, V. Govindaraju, "A Steerable Directional Local Profile Technique for Extraction of Handwritten Arabic Text Lines", in *International Conference on Document Analysis and Recognition (ICDAR'09)*, July 2009, pp.176-180

[6] N. Stamatopoulos, B. Gatos, G. "ICDAR 2013 handwriting segmentation contest," *in International Conference on Document Analysis and Recognition (ICDAR) 2013*, 2013, pp. 1402–1406.

[7] J. Ryu, H. I. Koo and N. I. Cho, "Language-Independent Text-Line Extraction Algorithm for Handwritten Documents", *IEEE Signal Processing Letters*, vol. 21, no. 9., pp. 1115-1119, Sep. 2014.

[8] J. A. Hartigan and M. A. Wong, "Algorithm as 136: A kmeans clustering algorithm," Applied statistics, pp. 00–108, 1979.