히스토그램 분석 기반의 인쇄체 문자열 분할 방법

장승익° 임길택 남윤석
한국전자통신연구원 우정기술연구센터
{sijang°, ktlim, ysnam}@etri.re.kr

A Method of Character String Segmentation using Histogram Analysis

Seunglck Jang°, Kil-Taek Lim, Yun-Seok Nam
Postal Technology Research Center, ETRI

요 약

본 논문에서는 인쇄체 우편주소 영상에서 smearing과 히스토그램 분석을 이용한 고속의 문자열 기울기 보정 및 분할 방법을 제안하였다. 제안한 방법에서는 입력 영상을 가분할 하고, 각각의 가분할 영상에 대한 수평 히스토그램을 분석하여 기울기 측정 및 보정을 수행하였다. 문자열 분할 단계에서는, 기울기가 보정된 영상에 smearing을 수행하고, 영상에 존재하는 잡영 및 각종 바코드를 제거하고, 수평 히스토그램 분석을 통해 최종 문자열 분할 결과를 도출하였다. 제안한 방법을 사용한 실험에서 2,000 장의 테스트 영상 중 1,989 장의 영상에서 정확한 문자분할 결과를 얻을 수 있었으며, 제안한 방법이 유효함을 보였다.

1. 서론

문서 자동화 처리분야의 필수요소인 문자인식 기술과 관련한 많은 연구가 있어왔으며, 현재 인쇄체 한글 문자의 인식률은 99%에 달하는 성능을 보여주고 있다[1-2]. 하지만, 문서 자동화 처리분야의 하나인 다량 우편물 자동화 처리 시스템에 기존 문자인식 기술을 적용할 경우 문자인식률에 비해 현저히 낮은 처리율밖에 얻지 못한다. 이는 처리하지 못한 우편영상의 대부분이 문자인식 이외의 부분에서 오류가 발생하기 때문이다. 오류의 원인은 문자간의 접촉, 우편물 자체에 존재하는 잡영, 바코드, 기울어짐 등에 의해서 나타난다. 즉, 높은 우편물 처리율을 얻기 위해서는 잡영 제거, ROI 추출, 기울기 보정, 문자열 및 문자 분할 등 전처리 단계의 성능을 높여야 한다. 특히 문자열 분할 성능은 문자 분리 및 문자 인식의 성능을 결정짓는 중요한 변수이며, 전체 우편물 처리율에도 큰 영향을 미치게 된다.

본 논문에서는 영상 smearing과 수평 히스토그램 투영을 이용한 문자열 기울기 보정 및 분할 방법을 제안한다. 제안한 방법에서는 수평 히스토그램 투영을 통해 입력 영상의 기울기를 측정하고, 이를 보정한다. 다음으로, 우편물의 수취인 영상을 수평 방향으로 smearing을 수행한 뒤, 연결요소를 추출하고, 잡영과 각종 바코드를 제거한다. 마지막으로, 각각의 연결요소에대한 수평 히스토그램을 이용하여 미세 문자열 분할을 수행하고, 최종 문자열 분할 결과를 도출한다. 제안한 방법을 사용하여 우편영상에서 추출한 200dpi 해상도의 2,000개의 영상에 대해 실험한 결과, 99.45%인 1,989개의 영상에서 정확한 문자열 분할 결과를 얻을 수 있었다.

2. 인쇄체 문자열 분할 시스템

2.1 시스템 흐름

인쇄체 문자열 분할 시스템의 흐름은 그림 1과 같다. 입력된 우편주소 영상에서 수평 히스토그램을 추출하고, 이를 분석하여 입력영상의 기울기를 측정하고, 기울기가 있을 경우 입력영상의 기울기를 보정한다. 다음으로, 기울기가 보정된 영상에 대해서 수평 방향으로 smearing을 수행하고, smearing이 된 영상을 연결요소 단위로 분할한다. 다음으로, 각 연결요소 중에서 잡영이나 바코드로 사료되는 연결요소를 제거한다. 잡영과 바코드 제거가 끝난 각각의 연결요소들에 대해서 히스토그램 분석을 이용해, 문자열 분할이 필요한 경우 연결요소를 분할한다. 마지막으로, 각 연결요소들을 문자열 단위의 영상으로 병합하여 최종 문자열 분할 결과를 도출하게 된다.

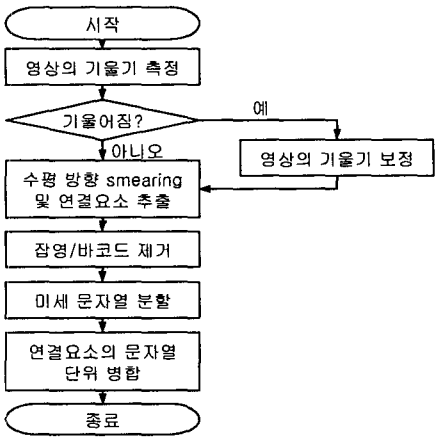
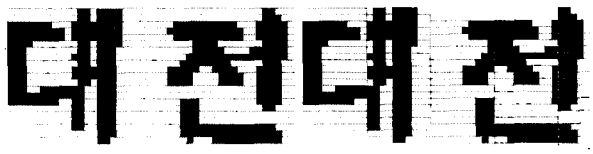


그림1. 문자열 분할 시스템 흐름도



(a) 0도 (b) -5도

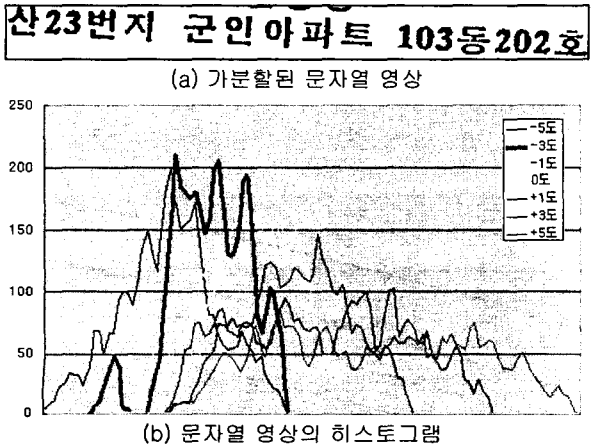
그림2. 수평 히스토그램 투영의 예

2.2 기울기 측정 및 보정

영상의 기울기를 측정하고 이를 보정하는 방법과 관련한 연구들이 있었다[3]. 하지만, 대부분의 방법은 많은 실수 연산으로 인해 그 속도가 느려, 고속으로 기울기를 측정하고 보정하기에는 적합하지 못하다.

본 논문에서는 실수 연산을 최소화하여 기울기 측정 및 보정을 수행한다. 기울기 측정의 과정은 다음과 같다. 먼저, 입력된 영상에서 연결요소들을 추출하여 수평으로 겹치는 연결요소들을 병합해 나감으로써 문자열 가분할을 수행한다. 다음으로, 가분할된 각각의 문자열 영상에서 +5도에서 -5도까지 1도 간격으로 11개의 히스토그램을 추출한다. 히스토그램의 추출 시 영상을 회전하여 히스토그램을 추출하지 않고, 히스토그램을 추출하는 과정에서 기울기를 감안하여 추출하였다. 각각의 기울기 θ에 해당하는 $1/\arctan(\theta)$ 를 미리 계산하여, 영상의 x축으로 매 $1/\arctan(\theta)$ 마다 히스토그램의 y축의 인덱스를 가감하여 수평히스토그램을 추출하였다. 그림 2의 (a)와 (b)는 각각 θ가 0과 -5인 경우의 히스토그램 투영 방법이다.

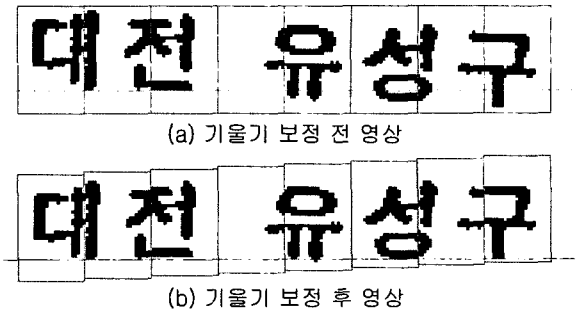
이렇게 추출된 문자열의 히스토그램을 이용하여 기울기를 측정하게 된다. 하나의 문자열에 대한 기울기는 11개의 히스토그램 중에서 산술평균이 가장 큰 히스토그램의 θ를 기울기로 설정한다. 이는 동일한 수의 픽셀로 이루어진 문자열 영상의 높이가 가장 낮을 때, 수평 히스토그램의 평균값이 가장 크다는 점을 이용한 것이다.



(a) 가분할된 문자열 영상

(b) 문자열 영상의 히스토그램

그림 3. 수평 히스토그램 및 분석



(a) 기울기 보정 전 영상

(b) 기울기 보정 후 영상

그림 4. 기울기를 보정한 예

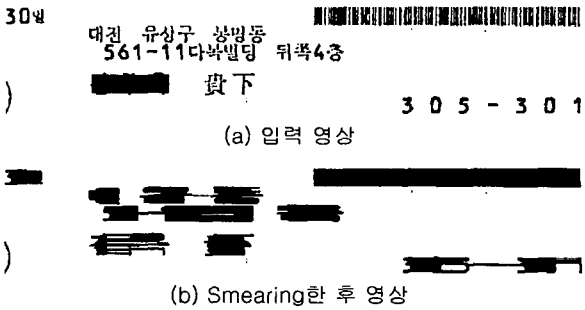
그림 3의 (b)는 (a)영상에 대해 각각의 θ에 대해 히스토그램을 구한 것이다. 그림에서 볼 수 있듯이 θ의 값이 -3도일 때 히스토그램의 폭이 가장 좁고, 높이가 가장 높다. 이와 같은 방법으로 각각의 가분할된 문자열의 기울기를 구한다음, 이들의 산술평균을 계산하여 입력 영상의 기울기로 사용하였다.

기울기 보정은 수평 히스토그램 추출과 유사한 방법을 사용한다. 앞에서 측정된 기울기 θ에 대한 $1/\arctan(\theta)$ 값을 이용해 기울기를 보정한다. 그림 4의 (a)와 같이 문자열이 기울어졌을 경우, 영상의 x축으로 매 $1/\arctan(\theta)$ 마다 y축 방향으로 영상을 밀거나 당김으로써 영상의 기울기를 보정한다. 그림 4의 (b)는 (a)의 영상에서 기울기를 보정한 영상이다.

2.3 문자열 분리

문자열 분리는 크게 smearing 단계와 분할 단계로 나누어 진다. 먼저 smearing 단계에서는 영상을 수평 방향으로 smearing한다. Smearing은 각 흑화소를 기준으로 우측의 30픽셀 이내(200dpi의 영상 기준)에 다른 흑화소가 있을 경우, 이 두 화소 사이의 백화소를 흑화소로 전환하는 방식으로 수행한다. 그림 5는 smearing의 예이다.

입력 영상에서 기울기를 보정한 뒤, smearing을 수행할 경우 대부분의 입력영상은 문자열 또는 단어 단위로 분할이 된다. 하지만, 그림 6의 (b)에서 보는것과 같이 문자열



(a) 입력 영상

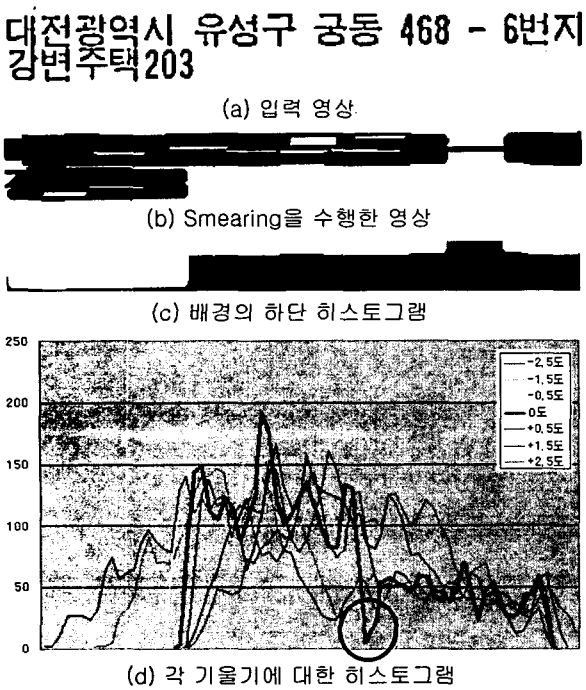
(b) Smearing한 후 영상

그림 5. Smearing 수행의 예

의 간격이 좁은 경우 두개 이상의 문자열이 하나의 연결요소로 나타나는 경우가 발생한다.

분할 단계에서는 이러한 연결요소를 분할하기 위해서 다음과 같은 일련의 작업을 수행한다. 먼저, 영상의 배경 히스토그램을 사용해 각 연결요소가 2개 이상의 문자열이 병합되었는지 검사를 한다. 상단과 하단에서 추출된 히스토그램에서 문턱치를 기준으로 문턱치 이상인 부분과 미만인 부분으로 나눈다. 문턱치 이상인 부분의 폭이 15를 초과하면 해당 연결요소는 2개 이상의 문자열로 구성되었다고 판별한다. 그림 6의 (c)는 (b)의 하단에 대한 배경 히스토그램이다. 본 논문에서는 문턱치를 20으로 설정하였으며, 위의 상수값들은 200dpi로 스캔된 영상에서 실험적으로 좋은 결과를 보여주는 값들을 선택한 것이다.

제안하는 방법에서는 2개 이상의 문자열로 구성된 연결요소를 각각의 문자열로 분할하기 위해서 수평 히스토그램을 사용한다. 히스토그램을 계산하는 방법은 기울기 측정에서 사용한 방법과 동일하다. 단, 0.25도 간격으로 히스토그램을 계산함으로써, 더욱 미세한 문자열 분할이 가능하도록 하였다. 그림 6의 (d)는 (b)의 각 기울기에 대한 히스토그램이다. 그림에서 볼 수 있듯이, 원으로 표신한 부분에서 히스토그램의 값이 가장 낮게 나타남을 알 수 있다. 그리고, 히스토그램의 가장자리의 값이 더 작은 값을 가지고 있지만, 이는 문자열의 분할 예상 범위에서 벗어나기 때문에 후보에서 제외시켰다. 분할 예상 범위는 배경 히스토그램에서 문턱치를 넘는 값들의 산술평균에서 문턱치를 넘지 못하는 값들의 산술 평균을 감하여 분할 예상점을 찾고, 분할 예상점의 상하 10픽셀의 범위를 분할 예상 범위로 설정한다. 문자열 분할점은 이 범위내에서만 선택하게 된다. 이렇게 찾은 분할점을 기준으로 영상을 이분하게 되고, 더 이상 분할이 되지 않을때까지 모든 연결요소에 대해서 문자열 분할을 수행한다.



(a) 입력 영상

(b) Smearing을 수행한 영상

(c) 배경의 하단 히스토그램

(d) 각 기울기에 대한 히스토그램

그림 6. 문자분할의 예

3. 실험 및 결과

본 논문에서 제안한 방법을 사용하여 문자열 분할 실험을 수행하였다. 문자열 분할 실험에 사용한 우편영상 DB는 실제 우편물을 200dpi 해상도, 256단계의 회색조로 스캔한 50,000개의 영상 중에서 무작위로 2,000개의 영상을 선택하여 실험하였다. 실험에 사용한 영상은 Otsu의 이진화 방법을 사용하여 이진화하였으며, 수취인 영역은 수작업을 통해서 추출하였다.

2,000개의 수취인 영상 중 1,989장의 영상에서 문자열을 성공적으로 추출하였으며, 11장에서 문자열 분할 오류가 발생하였다. 이중 9개의 영상은 잡영에 의해 문자열 분할 오류가 발생하였으며, 2개의 영상은 이진화 오류로 인해 문자열의 일부가 소실되어 나타난 오류이다.

4. 결론

본 논문에서는 인쇄체 우편영상의 문자열 기울기 보정 및 분할 방법을 제안하였다. 영상의 기울기 측정 및 보정을 수행함으로써, 문자열 분할 성능을 향상시킬 수 있었으며, smearing과 히스토그램 분석을 통하여 정확한 문자열 분할을 수행할 수 있었다. 또한, 히스토그램을 추출하는 과정에서 실수 연산을 최대한 배제함으로써, 고속으로 문자열 기울기 보정 및 분할을 수행할 수 있었다. 제안한 방법을 사용한 실험에서는 99.45%라는 높은 문자열 분할 성공률을 얻을 수 있었으며, 제안한 방법이 유효함을 보였다.

참고문헌

[1] 임길택, 김호연, 이상호, 송재관, 남윤석, " 우편물 자동구분을 위한 문자인식 시스템" , 대한전자공학회 컴퓨터/반도체 소사이어티 추계학술대회, 제 25권 제 2호, pp. 103-106, 2002.

[2] 장승익, 임길택, 김호연, 정선화, 남윤석, " 낱자 인식기와 자소 조합 인식기를 혼용한 인쇄체 한글 인식방법" , 한국정보과학회 봄 학술발표논문집, 제 30권 제 1호(B), pp. 244-246, 2003.

[3] C. Sun and D. Si, " Skew and Slant Correction for Document Images Using Gradient Direction" , DAR, *Proc. of 4th International Conf.*, Vol. 1, pp. 142-146, 1997.

[4] R. G. Casey and E. Lecolinet, " A Survey of Methods and Strategies in Character Segmentation" , IEEE Trans. on PAMI, Vol. 18, No. 7, pp. 690-706, 1996.