

부분 투영기법을 이용한 필기체 주소 영상에서의 문자열 분리

정선화, 남윤석
한국전자통신연구원, 우정기술연구센터

Text line separation in handwritten address image using partial projection technique

Seonhwa Jeong, Yunseok Nam
Postal Technology Research Center, ETRI
E-mail : {sh-jeong, ysnma}@etri.re.kr

Abstract

In this paper, we describe a method for separating text lines in handwritten Korean address images. The most remarkable feature of the proposed method is to use a modified projection technique, named a partial projection technique. A projection based text line separation method which projects the whole address image in horizontal direction to find split points for text line separation cannot avoid failing separation in case of images with a little skew or overlap between vertically neighboring text lines. To overcome this problem, we have introduced a partial projection technique which splits an address image into a few partial address images to be equal width and then project them each horizontally. The experiment done with 989 handwritten Korean address images extracted from live mails shows the superiority of the proposed method. The correct text-line separation rate for the testing images was about 91.5%.

I. 서론

본 논문에서는 다수의 문자열로 구성된 필기 주소 영상을 문자열 단위로 분리하는 방법에 관하여 기술하였다. 그림 1은 필기 주소 영상의 예이다. 그림에서 볼 수 있듯이 사람에게 따라서 필기 방식이 달라서 주소 문자열은 다양한 형태를 띄고 있다. 특히, 필기 문자의 경우 문자의 크기가 일정하지 않으면 문자열의 위치와 길이도 매우 다양하다. 또한, 문자열간의 겹침이나 접촉, 문자열의 기울어짐이 발생하기도 한다.

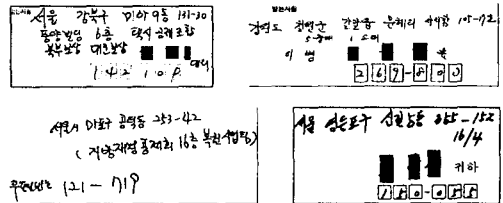


그림 1. 필기체 주소 영상의 예

기존의 문자열 분리 방법은 크게 연결요소기반기법 [1,2]과 투영기반기법 [3-4]으로 구분할 수 있다. 연결요소기반기법은 영상을 연결요소로 표현한 후 연결요소를 병합하면서 문자열을 생성하는 방법이다. 이 방법은 병합할 연결요소의 선택 및 연결요소간의 거리측정 방법에 따라 그 성능이 달라진다. 필기 영상의 경우 문자의 크기가 일정하지 않고 겹침이 많으며 그 유형 또한 다양하여 병합할 연결요소의 선택 및 거리 측정이 용이하지 않다. 투영기반기법은 영상을 수평방향으로 투영한 후 흰 런(white-run)을 찾아서 문자열을 분할하는 방법이다. 따라서, 문자열 간의 겹침이 있거나 기울어짐이 발생하는 영상에 대하여 문자열 분할이 어렵다.

본 논문에서는 투영기법기반의 필기체 문자열 분리 방법을 설명한다. 제안 방법에서는 투영기법기반의 문제점을 해결하기 위하여 주소 영상을 부분적으로 투영하여 문자열의 분할점을 찾고자 하였다.

II. 필기체 문자열 분리 방법

문자열 분리 방법의 순서는 다음과 같다. 먼저, 주소 영상을 N 개의 영역으로 수직방향 직선분할한다. 그 다음, 문자열 부분 영상 생성 단계에서는 각 등분된 영상을 수평방향 투영하여 문자열 부분 영상을 생성한다. 문자열 부분 영상은 문자열이 수평 또는 수직방향으로 분리된 영상을 의미한다. 좀 더 효율적인 결합을 위하여 작은 크기의 문자열 부분 영상을 주변 문자열 부분 영상에 결합하거나 큰 크기의 문자열 부분 영상을 분할한다. 마지막으로, 문자열 부분 영상 결합 단계에서는 문자열 부분 영상을 결합하여 문자열을 생성한다. 따라서, 문자열 부분 영상 생성 단계에서 최종 생성되는 영상은 하나의 문자열이 분할된 영상으로 이루어질 수 있지만 서로 다른 문자열에 속한 영상들로 이루어져서는 안된다. 참고로, 제안 방법에서 사용된 여러 임계치들은 200 DPI 해상도를 갖는 영상에서 최적화된 값이다.

2.1 수직방향 영상 N 등분

주소 영상을 동일한 폭을 가지도록 N 개의 영역으로 수직방향 직선 균등 분할한다. N 이 크면 수직방향 등분의 효과, 즉 문자열의 기울어짐이나 문자열간의 겹침이 발생하는 영상에서 문자열 분할점을 찾을 수 있는 확률이 증가하며 N 이 작으면 수직방향 등분의 효과는 감소하게 된다. 그러나, N 이 커지면 후에 생성되는 문자열 부분 영상의 개수가 증가하여 처리 속도가 감소한다는 단점을 갖는다. 따라서, 문자열 분할 성공률과 처리 속도를 모두 고려하여 적절한 N 의 선택이 이루어져야 한다. 제안 방법에서는 실험을 통하여 N 을 5 로 결정하였다. 그림 2 는 5 등분된 주소 영상을 보여준다.

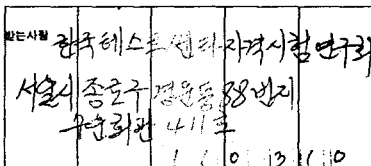


그림 2. 5 등분된 주소 영상의 예

2.2 문자열 부분 영상 생성

문자열 부분 영상 생성단계는 네 단계로 이루어져

있다. 첫번째, 각 등분된 영상을 수평방향으로 투영하여 흰 런(white run)을 기준으로 분할한다. 그림 3 은 이와 같은 방법으로 생성된 문자열 부분 영상을 보여준다.

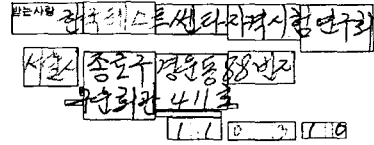


그림 3. 수평투영기반 문자열 부분 영상 생성의 예

두번째, 문자 인식에 영향을 미치는 않는 범위에서 장식으로 간주되는 수직획의 일부 및 문자열과 상관없는 잡영을 제거한다. 잡영이나 긴 수직획은 문자열 높이 값을 왜곡한다. 그런데, 제안 방법에서 문자열 높이에 따라 분할과 결합이 이루어지기 때문에 정확한 문자열 높이의 추정은 매우 중요하다. 제안 방법에서는 좀 더 정확한 문자열 높이의 추정을 위하여 이들을 제거하고자 하였다. 영상의 크기와 밀도 정보를 사용하여 제거가 이루어졌다. 그림 4 는 잡영 및 긴 수직획의 일부가 제거된 문자열 부분 영상을 보여주고 있다.

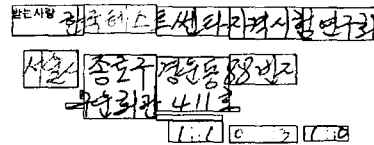


그림 4. 잡영 및 긴 수직획의 일부를 제거한 예

문자열 높이는 문자열 부분 영상의 높이 값을 사용하여 추정된다. 얻어진 높이 값을 올림차순 정렬한 후 이상치라고 간주되는 낮은 높이 값과 높은 높이 값을 제거한 후 나머지 높이 값을 평균하여 추정하였다. 문자열 높이는 문자열 부분 영상의 개수가 변할 때마다 다시 계산되어 사용된다.

세번째, 작은 크기의 문자열 부분 영상들을 수평 또는 수직 방향으로 이웃하는 문자열 부분 영상에 결합한다. 수평방향 결합 대상이 되는 영상은 폭과 높이가 임계치 이하로 작은 영상이다. 폭에 대한 임계치는 추정된 문자열 부분 영상의 폭의 50%이며 높이에 대한 임계치는 추정된 문자열 높이의 80%이다. 폭 추정 방법은 높이에 대한 추정 방법과 거의 동일하다. 결합 대상 영상은 겹침이 높이의 30% 이상 발생하면서 수평방향

거리가 30 화소 이내에 있는 문자열 부분 영상에 결합된다. 수직방향 결합 대상이 되는 영상은 높이가 추정된 문자열 높이의 60%보다 작은 영상이며, 수직방향으로 거리가 30 화소 이내에 있는 문자열 부분 영상과 결합된다. 그림 5 은 그림 4 의 영상에서 작은 크기의 문자열 부분 영상들이 결합된 예를 보여준다.

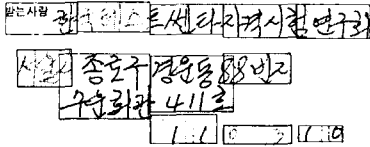


그림 5. 수평 및 수직 방향 결합의 예

네번째, 큰 크기의 문자열 부분 영상을 수평방향 직선 분할한다. 분할 대상 영상은 높이가 추정된 문자열 높이의 150%보다 큰 영상이다. 분할점은 영상으로부터 연결요소를 추출한 후 두 개의 군집으로 군집화한 후 찾아진다. 상위 문자열 군집에 속하는 연결요소가 갖는 y 좌표 중 가장 큰 y 좌표와 하위 문자열 군집에 속하는 연결요소가 갖는 y 좌표 중 가장 작은 y 좌표를 분할점 후보로 고려하였다. 두 분할점 후보 중 수평방향으로 투영하였을 때 낮은 밀도를 갖는 분할점 후보를 분할점으로 선택한다. 그림 6 은 그림 5 의 큰 크기의 문자열 부분 영상을 수평방향 분할한 예를 보여준다.

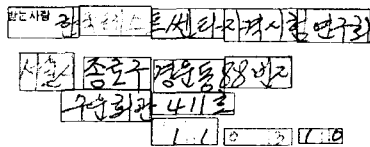


그림 6. 수평방향 분할의 예

2.3 문자열 부분 영상 결합

이와 같이 최종 생성된 문자열 부분 영상을 결합하여 문자열 영상을 형성하기 위하여 네 단계를 재귀적으로 거치게 된다. 주소 영상을 균등 분할하였기 때문에 문자 영상이 분할된 경우가 발생한다. 첫번째, 수평방향으로 이웃하면서 하나의 문자를 공유하고 있는 경우 두 문자열 부분 영상을 결합한다. 이때, 두 영역의 수평방향 겹침 정도가 높이의 50% 이상일 때 결합하였다. 그림 7 은 그림 6 의 문자를 공유하고 있는 문자열 부분 영상들을 수평방향으로 결합한 예를 보여주고 있다.

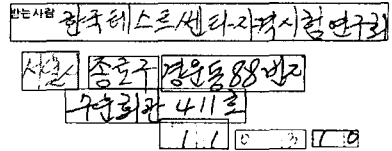


그림 7. 문자를 공유하고 있는 영상들의 결합의 예

두번째, 문자열 부분 영상을 최소인접사각형으로 표현하였을 때 임계치 크기 이상의 영역을 공유하고 있는 경우 두 문자열 부분 영상을 결합한다. 작은 크기의 문자열 부분 영상의 면적의 40% 이상을 공유하고 있는 경우 두 문자열 영상을 결합하였다.

세번째, 문자열을 생성한다. 문자열 부분 영상이 수평방향으로 150 화소 이내에 이웃하면서 두 영상의 수평방향 겹침 정도가 임계치 이상이면 결합하였다. 임계치는 대상 문자열 부분 영상의 높이의 50% 이다. 그림 8 은 그림 7 의 문자열 부분 영상으로부터 수평방향 결합을 통하여 문자열을 생성한 예를 보여주고 있다.

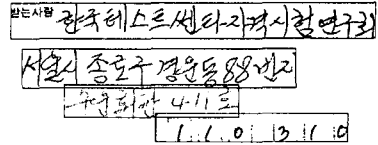


그림 8. 수평방향 문자열 부분 영상 결합 예

네번째, 중성이 상당히 떨어져서 작성됨으로써 앞 단계의 결합 알고리즘들에 의해서 결합되지 않은 경우를 처리하기 위하여 수직방향 결합을 시도한다. 작은 크기의 높이를 갖는 문자열 영상을 중성들의 모임으로 간주하고 수직방향으로 이웃하는 문자열에 결합하였다. 문자열의 폭이 50 화소 이하이거나 높이가 40 화소 이하이면 수직방향으로 30 화소 이내에 있는 문자열과 결합한다.

III. 실험결과

3.1 문자열 분리 성능

제안 방법의 성능 평가를 위하여 실제 우편영상으로부터 추출된 989 개의 주소 영상을 사용한 결과 문자열 분리 성공률 91.5%을 얻었다. 표 1 은 제안 방법의 성능을 제시하고 있다. 성능 측정 시 주소의 작성 순서를 유지하면 하나의 문자열이 여러 개의 문자열로 수평

방향 분리되는 경우에도 문자열 분리 성공으로 간주하였다. 그러나, 문자 영상이 분리되어서는 안된다.

표 1. 문자열 분리 성능

구분		영상 개수	합
문자열 분리 성공		904	989
문자열 분리 실패	과대분리	46	
	과소분리	19	
	기타	16	

그림 9 는 문자열 분리에 성공한 주소 영상의 예를 보여주고 있다. 제안 방법은 문자열의 기울어짐이 발생하거나 문자열간에 접촉이 발생하는 영상에 대해서도 성공적으로 문자열 분리를 수행할 수 있었으며 문자열의 길이나 문자열을 구성하는 문자의 크기에 편차가 많이 발생하는 경우에도 문자열 분리 성공을 보여주었다.

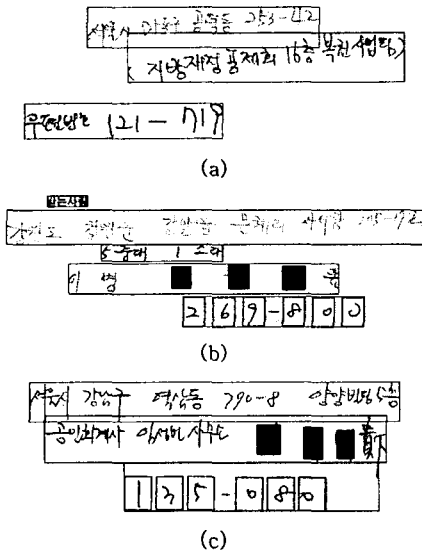


그림 9. 문자열 분리 성공의 예

3.2 문자열 분리 실패 분석

문자열 분리 실패는 하나의 문자열이 여러 개의 문자열로 분리될 때 문자가 분할되거나, 다수의 문자열이 하나의 문자열로 병합되는 경우이다. 전자를 과대분리 오류라 하며 후자를 과소분리 오류라 한다. 총 84 개의 문자열 분리 실패 영상 중 과소분리 오류로 분류되는 영상은 46 개이었으며 과대분리 오류로 분류되는 영상은 19 개이었다. 나머지 16 개의 오류는 기타로 분류되

었다. 다른 문자열의 일부가 포함된 형태를 기타의 오류라 하였다.

실패 영상을 제안 방법의 수행 순서에 따라 분석하여 보면 총 28 개의 영상이 문자열 부분 영상 생성에서 오류가 발생하였으며, 46 개의 영상이 문자열 부분 영상 결합에서 오류가 발생하였으며, 나머지 10 개의 영상이 무늬가 있는 주소 영상이거나 다른 문자열의 정보를 잡영의 형태로 포함하고 있었다. 문자열 부분 영상 생성 단계에서 발생하는 오류는 대부분 큰 크기의 문자열 부분 영상을 분할하지 못하거나 분할하더라도 잘못 분할하면서 발생하였으며, 문자열 부분 영상 결합 단계에서 발생하는 오류는 대부분 수평방향 결합 과정에서 발생하였다.

IV. 결론

본 논문에서는 필기 한글 주소 영상에서 문자열 분리 방법을 제안하였다. 제안 방법은 기존의 투영기반 문자열 분리 방법의 단점을 극복하기 위하여 주소 영상을 부분적으로 투영하여 문자열 분할점을 찾고자 하였다. 그 결과 문자열이 기울어지거나 문자열간에 접촉이 발생한 주소 영상에 대해서도 문자열을 성공적으로 분할할 수 있었다. 제안 방법의 문자열 분리 성공률은 총 989 개의 필기 한글 주소 영상에 대하여 91.5%이다.

참고문헌

- [1] S. Messlodi and C.M. Modena, "Automatic identification and skew estimation of text lines in real scene images," Pattern Recognition, Vol. 32, No. 5, pp. 789-808, 1999.
- [2] 박동열, 정선화, 김수형, "우편 봉투 영상에서 수신자 주소 영역의 텍스트 라인 추출", 전남대학교 정보통신논문지, 제 3 권 제 1 호, pp. 167-176, 1999.
- [3] A. Zahour, B. Taconet, P. Mercy, S. Ramdane, "Arabic handwritten text-line extraction," Proceedings of the ICDAR01, pp.281-285, Seattle, USA, 2001.
- [4] F. Venturelli, "A successful technique for unconstrained handwritten line segmentation," Progress in Handwriting Recognition, pp.563-568.