저품질 이진 우편 영상에서의 고속 문자 분할

김 두 식, 남 윤 석
한국전자통신연구원 정보화기술연구본부 우정자동화팀
전화 : 042-860-5809 / 핸드폰 : 017-391-1606

High-Speed Character Segmentation from Low-Quality Binary Letter Image

Doo-Sik Kim, Yun-Seok Nam
Postal Automation Team, Electronics and Telecomunications Research Institute
E-mail : doosik@etri.re.kr

Abstract

This paper proposes a character segmentation method for Korean letter address image. The poor quality of image binarization results in broken character strokes. To overcome this problem, two steps of processing are introduced. The first one is to merge broken characters to generate character candidates, and the other one is to reduce the complexity of segmentation graph path. These two steps do not use recognition information to keep in high-speed.

I. 서론

순로구분자동화 시스템은 컴퓨터에 의한 문자인식 기술을 우편물 처리 환경에 도입하여 수작업에 의존하는 순로구분 작업의 속도를 개선하고, 정확도를 높이고자 한다. 우편물의 수신자 주소를 인식함에 있어서 문자의 추출 과정을 고려할 때, 다양한 우편영상에 대하여 고속의 처리 성능을 요구하고 있다. 문자 추출 과정에서는 우편영상의 이진화 과정을 거치면서 문자의 획이 끊기거나 획 자체가 소실되는 문제와 우편물 인쇄 과정에서 다양한 폭을 갖는 문자체가 사용되는 등의 어려움이 있다. 문자 추출의 실패가 우편 영상의 처리 성능에 직접적인 영향을 주기 때문에 빠른 속도와 함께 정확성이 요구된다.

본 논문은 우편주소 영상으로부터 문자 추출 과정을 고속으로 수행하는 방법을 연구하는 것으로 문자의 외형적 특성을 이용하여 대략적인 문자 추출을 수행하고, 다양한 문자 분할 후보를 제시하되 문자 인식 횟수의 최소화를 위하여 문자 분할 후보 수를 줄이고자 노력하였다. 문자 패턴의 분할 및 병합 여부를 판단하는 부분에서는 결정 트리 방법을 변형하여 성능 개선이 용이하도록 하였고, 최적의 문자 분할 후보 위치를 선택하기 위하여 동적 프로그래밍 기법을 이용하였다.

II. 우편주소 인식

2.1 순로구분 자동화와 우편주소 인식

순로구분자동화는 우편물의 배달순로 구분을 자동화하는 것이다. 수작업에 의존하고 있는 배달순로의 구분 과정을 자동화하기 위해서는 우편물의 수신자 주소를 판독하는 기술이 필요하다. 카메라를 이용하여 우편영상이 획득되면 하드웨어에 의한 고속의 이진화를 수행하고, 이진영상으로부터 주소 영역을 추출한다. 추출된 주소 영역으로부터 문자열 추출이 이루어지며, 각 문자열에서 문자 분할을 수행하고, 이 결과가 문자 인식기에 전달된다. 문자 인식 결과와 주소 DB 정보를 이용하여 주소가 해석되며, 최종적인 배달순로 코드를

결정한다. 본 논문은 이러한 과정 중에 문자 분할에 관한 연구 결과를 다루고자 한다.

2.2 우편 영상에서의 문자 분할 문제

우편주소의 인식 문제에서는 정확성과 신속성이 동시에 요구된다. 우편물의 재질과 인쇄매체의 특성, 우편물에 인쇄되는 광고 등으로 인한 명도 히스토그램 분포의 왜곡으로 인하여 우편영상의 이진화를 위한 올바른 임계값을 추정하기가 곤란하며, 고속의 처리가 요구되기 때문에 장시간 처리를 요구하는 적응적(또는 지역적) 이진화를 수행하기가 곤란하다. 이러한 이진화 과정에서의 어려움은 문자 분할 문제에 직접적인 영향을 준다. 우편영상의 이진화가 너무 어둡게 수행되면 작은 잡영들이 나타나거나 문자 획들의 빈번한 접촉이 발생하고, 너무 밝게 수행되면 문자 획의 끊김 현상이 발생하여 문자 분할 단계에 영향을 주게 된다. 한편, 한글과 영숫자가 다양하게 나타나는 문자열 패턴에서 문자 인식 정보의 도움 없이 문자 분할을 수행하기는 어렵다. 따라서, 문자 분할 모듈은 다수의 문자 분할 후보 위치들을 제시하고, 최종 단계에서 문자 인식 정보를 활용하여 최적의 문자 분할 경로를 선택하게 된다. 문자 분할 후보가 많을수록 문자 인식을 수행해야 하는 횟수가 증가하므로, 실시간 처리가 요구되는 우편주소 인식의 문제에서는 문자 분할 후보 수를 최소한으로 줄이는 노력이 필요하다.

2.3 기존의 문자 분할 연구

기존의 대표적인 문자 분할에 관한 연구를 정리하면 다음과 같다. 숫자의 경우에는 접촉 위치의 각 유형별 특성을 분석하고 각 유형에 대한 분할 방법을 적용하였다[1]. 영어의 경우에는 문자열의 수직 윤곽 히스토그램을 분석하여 접촉된 문자의 분할 위치를 추출하였다[2]. 문자의 외형적 특성(종횡비 등)이 한글과 유사한 일본 문자 및 한자의 경우에는 문자 성분의 폭 정보를 이용하여 분리된 문자 성분의 결합을 시도하는데 문자 성분의 결합 관계를 그래프로 표현하고 각 경로의 비용을 인식결과의 신뢰값으로 계산하여 최고의 인식 신뢰값을 갖는 결합 조합을 찾는다[3]. 한글의 경우에는 사람의 조작에 의하여 입력된 문자의 평균 폭 정보를 이용하여 접촉/분리된 문자의 분리/결합을 시도하는 연구[4]와 문자 성분의 폭에 대한 최빈값을 기준으로 접촉 문자를 탐지 및 분리하는 연구[5]가 있다.

기존의 문자 분할 방법들은 인식 정보에 대한 의존 여부에 따라 세 가지 방법으로 구분할 수 있는데[6], 문자 인식 정보의 도움 없이 문자 분할을 수행하면 좋은 성능을 기대하기 어려우며, 문자 인식 정보를 전적으로 의존한다면 속도의 저하를 초래하게 된다. 따라서, 문자 인식 정보를 이용하되 문자 인식기의 호출 횟수를 줄이기 위하여 문자 분할 후보를 최소화하는 노력이 필요하다.

III. 우편영상에서의 문자 분할

3.1 문자 성분의 추출

그림 1과 같이 각 문자열에 대하여 수직 투영을 수행하여 문자 성분의 좌우 좌표를 구하고(수직선), 각 문자 성분에 대하여 위에서 아래로 또는 아래에서 위로 흑화소를 추적하여 문자 성분의 상하 좌표를 구한다(수평선). 이러한 방법은 연결 요소 분석을 수행하는 것보다 문자 성분을 빠르게 추출하는 이점이 있다.

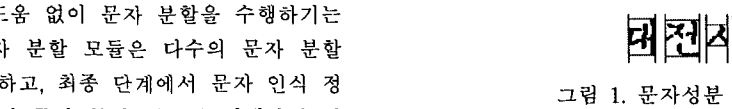

그림 1. 문자성분 추출 예

3.2 한글 문자 폭의 추정

(1) 방법 1

문자열의 높이(LH)를 기준으로 폭이 너무 작은 문자 성분($W<T_1*LH$)과 큰 문자 성분($W>T_2*LH$)을 제외한 나머지 문자 성분들의 폭에 대한 평균값(avg)과 중앙값(med)을 구한다. 평균값이 문자열의 높이와 근사하면($T_3*LH<avg<T_4*LH$), 평균값을 한글의 문자 폭(CW)으로 간주하며, 그렇지 않으면, 평균값과 중앙값 중 문자열의 높이에 가까운 값을 한글의 문자 폭으로 간주한다.

(2) 방법 2

문자열의 높이(LH)를 기준으로 폭이 너무 작은 문자 성분($W<T_1*LH$)과 큰 문자 성분($W>T_2*LH$)을 제외한 나머지 문자 성분들의 폭에 대한 평균값(avg_1)을 구하고, 이 평균값을 기준으로 폭이 너무 작은 문자 성분($W<T_5*avg_1$)과 큰 문자 성분($W>T_6*avg_1$)을 제외한 나머지 문자 성분들의 폭에 대한 새로운 평균값(avg_2)을 한글의 문자 폭으로 간주한다.

3.3 접촉 문자의 분리

한글의 문자 폭을 기준으로 접촉 문자($W>T_7*CW$)

를 탐지하고, 이에 해당하는 문자 성분을 이등분한다. 실험 분석에 의하면, 우편주소 영상에 나타나는 문자의 접촉 사례는 그다지 복잡하지 않았기 때문에 접촉 문자에 관한 처리 과정이 단순할 수 있었다.

3.4 문자 병합 후보의 생성

N개의 문자 성분에 관한 위치와 크기 정보는 다음과 같은 방법으로 2차원의 데이터 구조체 $L_{N,N}$에 저장한다. 우선 각 문자 성분의 정보는 문자열의 가장 왼쪽의 문자 성분(C_1)부터 마지막의 문자 성분(C_N)까지 데이터 구조체의 $L_{1,1}$부터 $L_{N,N}$까지 저장된다.

문자 성분 $L_{i,i}(1<=i<=N)$와 이웃하는 문자 성분 $L_{i+1,i+1}$을 병합한 문자 성분의 폭(MW)이 한글의 문자 폭보다 너무 크지 않다면($MW<T_7*CW$), 병합한 문자 성분의 정보를 $L_{i,i+1}$에 저장하며, 조건을 만족하는 동안 $L_{i,i+j}(1<=j<=N-i)$에 대하여 반복한다.

이처럼 그림 2의 좌측 그래프와 같은 문자 병합 후보의 정보는 우측의 2차원 데이터 구조체에 저장된다.

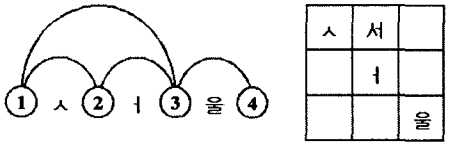

그림 2. 문자 병합 정보의 저장 예

이러한 정보의 표현 방법은 문자 성분을 병합하기 전과 병합한 후의 정보를 모두 포함하고 있을 뿐만 아니라 그래프 경로 비용의 계산에도 쉽게 사용된다[7].

3.5 문자 병합 판단을 위한 다양한 정보의 활용

3.4절에서는 문자의 병합 조건에 대하여 병합한 문자 성분의 폭(MW)과 한글의 문자 폭(CW)만을 고려하였다. 문자 병합 후보의 수를 줄이기 위해서는 보다 다양한 정보를 이용한 판단이 필요하다. 이 때에는 문자 성분의 폭 뿐만 아니라 각 문자 성분의 위치 및 크기, 병합 문자의 위치 및 크기, 두 문자 성분의 간격 등의 정보를 이용한다. 이처럼 다양한 정보들을 고려하는 여러 번의 판단들은 "*if ... then ...*"의 복합적인 형태를 가지므로, 결정 트리와 같은 구조로 표현될 수도 있으나, 결정 트리는 구조의 변경 및 실험을 통한 효과 분석이 용이하지 않기 때문에, 비트 표현 방법으로 구현하였다. 즉, 각 조건들의 참과 거짓을 비트 값으로 표현하면, 앞에서 나열한 다양한 정보들에 의한 판단 결과는 비트열로 나타낼 수 있다. 판단 조건이 n개인 복합 판단문의 결과는 2^n개의 표현을 갖는 길이 n의 비트열 중의 하나와 대응되며, 그 값을 1차원의 표에 저장하여 문자 성분의 병합 여부 판단시에 참조한다. 이때에 판단 조건의 수가 증가할수록 표의 크기가 기하급수적으로 증가하기 때문에, 판단 결과 값이 편중되어 있는 표의 경우에는 <비트열, 판단결과>의 한 쌍을 함께 저장하거나, 판단 과정을 여러 단계로 적절히 구분하여 표의 크기를 $a2^b(a*b\cong n)$보다 작게 줄일 수 있다. 계산의 효율성을 위하여 표의 검색에 의한 판단 결과를 다음과 같이 세분한다.

(1) 문자 성분을 병합 및 결과를 저장($L_{i,i+j}$)하고, 검사를 진행(j++)한다.

(2) 문자 성분을 병합 및 저장하고, 병합 전의 문자 성분들($L_{i,i+j-1}$,$L_{i+j,i+j}$)은 삭제하며, 검사를 진행(j++)한다.

(3) 검사를 진행(j++)한다.

(4) 검사를 중단(i++, j=1)한다.

앞의 각 판단 결과에 대한 대표적인 예는 (1) "이"와 "01"과 같이 문자 성분의 병합 여부가 혼동되는 경우, (2) "ㅐ"와 같이 문자 성분의 병합이 확실한 경우, (3) 획의 끊김이 많고 각 문자 성분의 크기가 작아서 계속 문자 성분의 병합을 시도할 필요가 있는 경우, (4) 병합된 문자 성분의 폭이 너무 커서 더 이상의 검사가 불필요한 경우를 들 수 있다.

3.6 문자 분할 경로의 생성

기존의 최적 경로 탐색 알고리즘을 활용한 문자 분할 방법은 문자 인식 정보에 전적으로 의존하기 때문에 인식기의 실행 횟수가 많아서 속도의 저하를 가져온다. 본 연구에서는 문자 인식 정보가 아닌 문자 성분의 위치 및 크기 정보를 이용하여 최적 경로 탐색 알고리즘의 경로 비용을 계산한다. 이 경우에는 최적의 경로를 확정하기가 곤란하므로 다양한 후보 경로를 제시하여야 하므로 각 분할 위치에서 1순위 경로뿐만 아니라 2순위 경로까지 고려한다.

(1) 경로 비용의 계산

문자 성분의 폭과 넓이에 따라 전폭 문자, 반폭 문자, '-'(hyphen)으로 구분한다. 종횡비가 1:1에 가까운 한글과 'W' 등은 전폭 문자의 예이며, 종횡비가 2:1에 가까운 대부분의 영숫자와 특수문자는 반폭 문자로 간주한다. 이들 중 어느 유형에도 속하지 않은 작은 크기의 문자 성분은 획의 일부이거나 잡영에 해당하므로 경로 비용을 높게 설정하고자 한다.

유형 c에 해당하는 문자 성분의 폭(W)과 높에(H)에 대한 형태(shape)적 경로 비용은,

$C_s = (|W - CW_c| + |H - CH_c|)$

$(CH_{선폭} = CH_{반폭} = CH,\ CH_{하이픈} = T_8*CH,$

$CW_{선폭} = CW_{하이픈} = CW,\ CW_{반폭} = T_9*CW)$

이고, 이웃하는 문자 성분과의 간격(G_L, G_R)에 대한 위치(position)적 경로 비용은,

$C_p = -(G_L + G_R)$

이며, 문자 분할 경로 상에서 각 문자 후보에 대한 경로 비용은,

$C = T_{10} * (T_{11} * C_s + T_{12} * C_p) / CH$

로 계산된다.

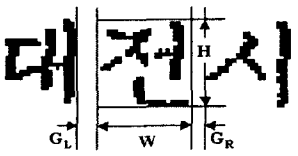

그림 3. 경로 비용의 산출을 위한 정보

(2) 불필요한 경로의 제거

- 각 분할 위치에서 2순위까지만 고려: 동적 프로그래밍 기법을 이용하여 최적의 경로를 찾는 방법[7]과 유사한 과정을 수행한다. 단, 각 분할 위치에서 1순위 뿐만 아니라 2순위의 경로도 고려한다. 3순위 이하에 해당하는 경로는 제거한다(그림 4a).

- 잡영의 제거: 경로 상에서 크기가 작고 이웃 문자 성분과 멀리 떨어져 있으면, 잡영으로 간주하여 제거한다(그림 4b).

- 겹친 경로의 제거: 경로 i,j가 서로 겹치고, 두 경로에 대한 경로 비용의 차이가 크면($|C_i - C_j| > T_{13}$), 그 중에서 비용이 큰 경로를 제거한다(그림 4c).

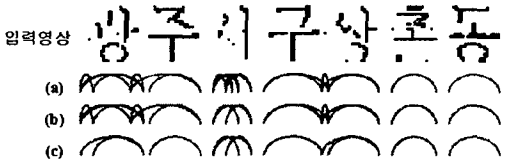

그림 4. 분할 경로의 추출 과정

Ⅳ. 실험 결과 및 분석

실제 우편물로부터 수집된 50,000통의 우편 영상 중에서 500통을 무작위로 추출하여 실험을 하였다. 처리 과정의 특성상 문자 분할만의 실행 시간을 측정하기가 곤란하였다. 우편영상의 이진화와 주소영역의 추출이 미리 실행되어진 주소영역의 영상에 대하여 주소 문자열의 추출 및 문자 추출에 소요되는 평균 시간은 Pentium 440MHz 상에서 0.02초이며, 문자 분할 경로 상의 문자 후보 중에서 올바른 문자가 결여되는 오류율은 3.2%, 문자 후보 수/실제 문자수는 156%이다.

Ⅴ. 결론

본 논문은 자획이 끊어진 저화질의 문자열 영상에서 고속으로 문자를 추출하는 방법을 제시하였다. 이 방법은 문자 인식 결과의 도움 없이 문자 분할 후보를 최소화하여 문자 분할 및 인식 속도를 향상시켰다. 앞으로도 문자 후보 수를 낮추기 위한 연구가 계속 진행될 예정이다.

참고문헌

[1] H. Fujisawa, Y. Nakano, and K. Kurino, "Segmentation methods for character recognition: from segmentation to document structure analysis," Proc. of the IEEE, Vol. 80, No. 7, pp. 1079-1092, July. 1992.

[2] S. Tsujimoto and H. Asada, "Major components of a complete text reading system," Proc. of the IEEE, Vol. 80, No. 7, pp. 1133-1149, July 1992.

[3] S. Ariyoshi, "A character segmentation method for Japanese printed documents coping with touching character problems," The 11th IAPR Int. Conf. on Pattern Recognition, Vol. 2, pp. 313-316, 1992.

[4] 최봉희, 이인동, 김태균, "문자영역 추출과정에서의 오분리의 교정," 한국정보과학회논문지, 제 21권 1호, pp. 86-93, 1994년 1월.

[5] 김의정, 김태균, "오프라인 문서에서 개별 문자 추출과 한자 인식에 관한 연구," 한국정보처리학회 논문지, 제 4권 5호, pp. 1277-1288, 1997년 5월.

[6] R. G. Casey and E. Lecolinet, "Strategies in character segmentation: a survey," Proc. of the 3rd Int. Conf. on Document Analysis and Recognition, Vol. 2, pp. 1028-1033, Aug. 1995.

[7] L. Y. Tseng and R.-C. Chen, "A new method for segmenting handwritten Chinese characters," Proc. of the 4th Int. Conf. on Document Analysis and Recognition, Vol. 2, pp. 568-571, 1997.