문자 가분할과 Support Vector Machine을 이용한 필기 한글 단어 고속 검증기

이 충 식, 김 인 중, *신 종 탁, 김 진 형
KAIST 전자전산학과, *(주)레코그램
전화 : 042-869-5557 / 핸드폰 : 016-680-2279

Hangul Segmentation and Word Verification System for Automatic Address Processing

Chung-Shik Lee, In-Jung Kim, Jong-Tak Shin, Jin-Hyung Kim
Dept. of EECS, KAIST
E-mail : cslee@ai.kaist.ac.kr

Abstract

A fast method of Hangul address word verification is presented in this paper. Pre-segmentatin and recognition by DP matching is adopted in this paper. An address line image is over-segmented by analyzing the topology of connected components and the projection profile. A fast individual Hangul character verifier was developed by applying SVM (Support Vector Machine). The segmentation hypothesis was represented by lattce structure, and a best path search by dynamic programming generates the most probable segmentation path and the final verification score. The word verifier was tested on 310 address image DB, and it show the possibility of improvements of this method.

I. 서론

우편물의 자동 분류 및 처리를 위한 기술은 외국의 경우 60년대 후반부터 꾸준히 연구되어 왔으나 국내에서는 아직도 실질적인 기술의 개발이 미미한 실정이다. 국내에서는 주로 한글 낱자 인식 연구에 치중하고 있으며 이를 응용한 단어인식이나 주소인식 기술 개발이 시급하다. 국내의 우정자동화 현황을 보면, 우편영상으로부터 우편번호 필드를 추출, 인식하는 시스템이 사용되고 있다. 그러나 우편물을 배달경로순으로 정렬하는 순로자동화 시스템을 구축하기 위해서는 궁극적으로 주소필드의 인식이 필요하다.

우편물에는 우편번호와 주소에 동일한 정보가 중복되어 나타나며, 문자보다 상대적으로 인식이 쉽고 정확도가 높은 숫자인식 결과를 최대한 이용하는 것이 중요하다. 따라서 우편번호를 인식한 후, 이를 주소영역에서 검증하는 방식이 효과적이라고 판단되며, 본 연구에서는 이를 위한 기반 연구로서 주소영상의 필기 단어 검증에 관한 연구를 수행하였다.

단어인식에 관한 연구는 크게 세가지로 분류된다. [1] 첫째, 분할 후 인식법은 단어영상을 문자로 분할한 후, 각각의 분할된 문자를 인식하는 방법이다. 비교적 적은 정보만으로도 문자 분할이 가능한 인쇄체 문자인식에서 주로 사용되고 있으나, 문자의 분할이 상대적으로 어려운 필기체 인식에서는 분할에서 비롯되는 오류가 전체 시스템 성능에 큰 영향을 미친다. 둘째, 가분할 후 인식기를 이용한 탐색 방법은 분할의 오류를 인식기가 모델링하고 있는 정보를 이용하여 극복하도록 하는 방법이다. 오류의 가능성이 큰 문자 분할 과정에서 실제 분할점보다 많은 분할점을 지정하고, 인식기의 인식신뢰도를 이용하여 최적의 분할점을 뒤늦게 결정하여 오류를 최소화한다. 셋째, 전체단어인식은 단어 전체를 하나의 단위로 인식하는 방법으로, 사용되는 단어의 수가 많지 않을 경우에 주로 사용된다. 전체 단어영상으로부터 개략적 특징을 추출하여 HMM

등의 모델링 틀을 사용하는 방법 등이 소개되고 있다. 그러나 인식하고자 하는 단어의 수가 많은 경우에는 사용하기가 어렵고, 또한 훈련에 필요한 단어영상을 얻기 어려운 환경에서의 사용이 제약된다. 본 연구에서는 두 번째 방법인 가분할 후 인식방법을 사용하여 단어검증에 사용하였다.

문자 가분할에 관한 연구는 다수 존재하나 수학적인 모델링 방법을 사용한 연구는 거의 없다. 문자분할에 관한 연구의 대부분이 필기의 특성에 기초한 휴리스틱을 이용하여 문자를 분할한다. 접촉된 형태의 문자를 정확히 분리하기 위해서는 분리된 문자의 인식에 필요한 정보보다 많은 정보량을 요구하나 실제로 분할시에는 더 적은 정보만으로 문자의 분할이 이루어지기 때문에 오류의 발생 가능성이 높다. 또한, 숫자, 영어, 한글, 한자 등 문자간의 특성에 따라 접촉형태가 다양하게 변하며 필요로하는 정보도 달라지게 된다.

본 연구에서는 고속처리된 우편번호 인식결과를 우편주소영상으로부터 검증방법을 제시한다. 고속처리가 필수적이기 때문에 문자가분할시에 다양한 정보를 조사하기 어렵다는 제약이 존재하며, 동적탐색 적용시에 사용하는 문자인식기의 속도가 빠른 것이 요구된다. 한글의 경우 낱자 인식 성능이 90% 정도인 인식기가 존재하나 속도가 느리며, 고속 인식이 가능한 인식기는 성능이 50% 전후의 낮은 정확도를 가지고 있다. 이러한 상황에서 주어진 인식기를 이용하여 가장 높은 성능을 나타낼 수 있는 단어 검증기를 구성하는 것이 필요하다.

본 논문에서 구성한 시스템은 주소영상을 가분할 후 문자후보를 래티스로 표현하고, SVM을 이용한 낱자검증기와 동적탐색기법을 적용하여 최적경로를 탐색함으로써 단어검증을 수행한다. II절에서는 주소영상을 문자후보로 가분할하는 방법을 설명하고, III절에서는 SVM을 이용한 한글 낱자 검증기에 대하여, IV절에서는 가분할 결과와 낱자검증기를 이용한 최적경로 탐색 단어검증방법에 대해 설명한다. V절에서는 실제 우편영상에 적용한 실험결과를, 마지막으로 VI절에서는 결론을 맺는다.

II. 주소영상 가분할

문자가분할이란 실제의 문자분할점보다 많은 수의 분할점을 찾음으로써 실제 분할점을 놓침으로써 비롯되는 오류를 최소화하는 방법을 말한다.

본 논문에서 제시하는 문자가분할 방법은 다음과 같이 구성된다.

1) 연결화소 분리
2) 수직인접 연결요소 병합
3) 화소, 획 투영정보를 이용한 분할함수 계산
4) 분할함수의 국소최대점 선택
5) 연결요소와의 관계 고려에 의한 분할점 선택
6) 분할후보 래티스 구성

문자 필기의 많은 부분은 연결화소를 분리한 후 그룹을 묶는것만으로도 문자분할이 쉽게 이루어진다. 이렇게 쉽게 판별이 가능한 문자분할점을 찾기 위하여 1) 연결요소 분석을 수행한다.

다음으로는 분리된 연결요소들 중에서 하나의 문자를 구성하는 그룹으로 판단되는 연결요소들을 병합한다. 한글의 특성상 하나의 문자를 구성하는 연결요소들은 수직투영시에 서로 겹침이 심하게 존재한다. 이러한 특징을 이용하여, 2) 수직투영시에 겹침의 비율이 임계치를 넘는 연결요소들을 병합함으로써 정보를 단순화한다.

병합이 완료된 연결요소들 중에는 인접한 문자간의 접촉을 포함하고 있는 연결요소가 존재한다. 이들을 분리하기 위한 최소한의 정보로써 일반적으로 문자분할에 널리 사용되고 있는 화소수직투영정보와 획수직투영정보를 사용하였다.

문자화소를 수직으로 투영하였을 때, 문자와 문자간에는 일반적으로 계곡이 존재할 가능성이 높은 것으로 알려져 있다. 또한, 한글은 수직모음이 문자의 우측에 존재한다는 사실과, 대부분의 접촉지점에서는 획수직투영시 투영값이 크지 않은 것으로 관측되고 있다.

이러한 특성을 이용하여 다음과 같은 분할점 예측함수를 정의하였다.

$$f(x)=\frac{\alpha p(x-t)+(1-\alpha)p(x+t)-2p(x)}{p(x)s(x)}$$

분할 후보점 x에서의 분할함수값 $f(x)$는 현재 위치 x에서 좌우측 t의 위치의 수직투영값 $p(x-t)$, $p(x+t)$을 참조하여 기울기의 변화를 측정한다. 상수 α는 한글의 특성상 문자의 우측에 모음이 존재할 확률이 높다는 점을 반영하기 위하여 $\alpha>0.5$를 선택하였다. 획수직투영값 $s(x)$는 분할점에서의 획의 복잡도가 일반적으로 낮다는 점을 반영하고 있다.

이 함수값은 이용하여 분할점을 예측하고, 함수값의 국소최대점을 예비분할점으로 판단하며, 예비분할점과 연결요소와의 위치관계를 고려하여 최종적인 분할점을 결정한다.

이렇게 결정된 분할점을 이용하여 주위의 영상요소들을 결합시에 넓이가 문자의 추정넓이와 비슷한 요소들을 문자후보로 지정하여 문자후보 래티스를 구성한

다. 그림1은 실제 영상을 분할함수를 적용하여 가분할한 영상을 보여주고 있다.

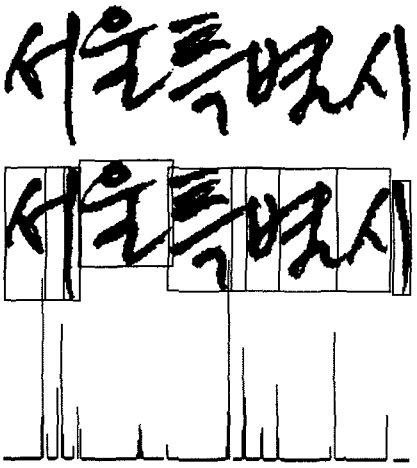

그림 1 문자분할함수를 적용한 영상 가분할 결과

III. SVM을 이용한 문자 검증기

SVM은 1995년 Vapnic에 의해 개발된 패턴인식 방법이다. 신경망의 한계점을 수학적으로 모델링하여 개선하기 위한 방법을 제시하고 있으며, 신경망에 비하여 일반화(generalization)에 있어서 강점을 보인다는 것을 실험적으로 보이고 있다. 그러나, 두가지 클래스에 대한 구분에 대해서는 좋은 성능을 보이지만, 여러 클래스를 구분하기 위해서는 앞으로도 많은 연구가 필요하다. 본 연구에서는 [3]에 기술된 방법을 이용하여 훈련 및 검증기를 구성하였다.

우편주소에는 2350개의 한글 완성형 문자 중 467개가 실제로 사용되고 있는 것으로 조사되었다. 본 과제에서는 467개의 각 문자에 대해서 이진 인식기(binary classifier)를 구성함으로써 검증기를 구성하였다. 사용한 특징벡터로는 동양문자의 통계적 인식방법에 많이 사용되는 방향성 특징을 사용하였으며, 비선형 정규화를 적용하여 한글 필기의 변이를 흡수하였다[2]. 문자 X의 검증기를 구성하기 위한 훈련 알고리즘은 아래와 같다[3].

클래스 X의 모든 훈련 데이터 영상에 대하여

1) X에 속한 영상을 클래스 A, 그렇지 않은 영상을 클래스 B로 분류
2) 클래스 A와 B에 속하는 모든 영상으로부터 특징벡터 추출
3) 추출된 특징벡터를 커널함수를 사용하여 변환
4) 클래스 A와 클래스 B의 최적경계면 계산

이렇게 구성된 낱자 검증기는 입력특징벡터와 최적경계면까지의 거리를 출력값으로 가지며, 클래스 A에 속하면 0보다 큰 값을, 클래스 B에 속하면 0보다 작은 값을 갖는다. 이를 [0-1]의 스코어로 변환하기 위하여 다음과 같은 sigmoid 함수를 사용하였다.

$$f(x)=\frac{2}{1+e^{-\lambda(x+a)}}$$

IV. 동적정합에 의한 단어 검증

단어 검증의 목적은 주소영상의 시작부분에 주어진 주소단어가 존재하는지의 여부를 판단하는 것이다. II절의 문자가분할 과정을 거쳐 주소영상은 문자분할후보 래티스의 형태로 주어지고, 검증하고자 하는 주소단어와의 최적 정합을 찾는다. 최적 정합 스코어가 임계치보다 클 때에 주소단어가 영상내에 존재하는 것으로 판단하며, 최적 정합 결과로부터 영상내 단어의 존재 범위를 알 수 있다.

단어인식을 위한 동적정합은 다양한 방법론이 연구되어 있으며, 단어사전의 이용방식이나, 스코어링 기준 등에 따라서 정합 방법이 다양해질 수 있다. 본 논문에서는 단어사전이 주어졌을 때의 동적정합 방법을 제시한 Kim[4]의 연구를 응용하여 최적 정합 검색에 사용하였다.

Kim의 연구와는 달리 주소영상의 시작부분에 주어진 단어가 존재하는지의 여부를 판단하는 것이므로, 동정정합에서의 끝점의 판별이 모호하다. 이러한 점을 고려하여 다음과 같은 방식으로 최적 정합을 탐색하였다.

찾고자 하는 주소단어를,

$$L=(L_1, L_2, \ldots, L_k), where\ k=\text{단어의길이}$$

라고 하고, 주어진 문자후보 래티스는, 노드 $n_i\ (i=1,\ldots,N)$와 에지 $e(n_i, n_j)$로 구성된다. 이때 각 노드는 가능한 분할점을 나타내며, 두 노드를 연결하는 에지는 두 분할점 사이의 영상이 하나의 문자를 형성할 수 있는 후보임을 의미한다. 에지의 문자후보를 낱자 검증기로 검증한 결과값은 $D(n_i, n_j)$로 표현한다.

이와 같을 때, 영상의 첫부분부터 노드로 표현되는 분할후보점 n_i까지 단어의 일부인 $L_1,\ldots,L_m$을 사

용하여 인식한 결과는 $D_{n_i}(L_1,\ldots,L_m)$으로 표현된다. 이와 같을 때, 최적 정합을 찾는 알고리즘은,

1) 초기화 :

$D_{n_i}(\varnothing)=0, \text{for } all\ i$

$D(n_i,n_j)=$ 낱자검증기의 검증스코어

2) 반복 :

$D_{n_i}(L_1,\ldots,L_m)=$
$Max_j\,[D_{n_j}(L_1,\ldots,L_{m-1})+D(n_i,n_j)],$
$where\ there\ exist\ an\ edge\ e(n_i,n_j)$

3) 최적정합판별 :

$D^*=Max_i\,[D_{n_i}(L_1,\ldots,L_k)]$

4) 검증 결과 :

$\text{if } D^*\geq\theta\ then\ TRUE\ else\ FALSE$

이상과 같은 방법으로 주소영상의 시작부분에 주어진 주소단어가 존재하는지의 검증을 수행한다.

V. 실험 및 분석

낱자 검증기의 성능 실험에는 PE92 한글낱자 데이터베이스를 사용하였다. PE92 데이터는 완성형 코드 2350개의 문자 클래스에 대하여 각각 100개의 영상을 가지고 있다. 본 실험에서는 주소에 나타나는 문자 467셋의 짝수번 영상을 훈련에, 홀수번 영상을 테스트에 사용, 클래스당 각각 50개의 데이터를 실험에 사용하였다. 검증에 사용되는 출력값은 0.5를 임계값으로 사용하였다.

실험은 크게 두가지로 구성되었다. 실험1에서는 바른 문자가 주어졌을 때 승인 여부를 조사하였으며 실험2에서는 틀린 문자가 주어졌을 때의 부인(rejection)의 정확도를 측정하였다. 실험2에서의 PE92 DB로부터 해당 문자와 다른 클래스로부터 임의로 50개의 데이터를 추출하여 실험에 사용하였다. 실험에 사용된 변수값은 $\lambda=5, a=1.2$를 사용하였으며, 각 실험의 검증 성공률은 다음과 같다.

오류 유형	낱자 검증 성능
Type I 오류 (실험1)	93.38% (21368 / 22883)
Type II 오류 (실험2)	97.75% (22367 / 22883)

실험은 펜티엄III 600MHz에서 수행하였으며, 검증속도는 평균 0.000176초 (8.06초 / 45776문자)를 얻었다.

단어검증기의 성능은 실제 우편봉투로부터 추출한 주소영상을 사용하여 실험하였다. 총99개의 주소영상으로부터 각각 도,시,군,동 등 3~4개의 질의를 구성하였다. 각각의 질의는 정답질의와 오답질의의 2개로 구성되어 각각 310개의 질의DB를 구성하였다.

	검증률	평균검증속도
정답질의	81.94%	0.31초
오답질의	92.92%	0.30초

오류의 원인으로는 문자가분할 오류가 전체 오류의 많은 부분을 차지하고 있었으며, 낱자검증기의 검증값이 실제 데이터에서 큰 오차를 보이는 경우가 다수 존재하였다. 문자 가분할이 여전히 큰 영향을 미치고 있다는 것을 알 수 있었으며, 검증기의 성능 향상이 필요하고, 문자가분할과 낱자 검증기의 오류를 단어검증 단계에서 보완하기 위한 방법론의 개발이 필요한 것으로 판단된다.

VI. 결론

우편영상의 고속인식에 필요한 문자검증 방법을 제시하였다. 고속 한글문자 가분할 방법을 제시하였으며, SVM을 이용한 고속 문자검증기를 개발, 이를 이용한 주소단어 검증 방법을 제시하였다. 실험 결과 문자 가분할 과정에서의 오류가 많은 것을 확인하였으며, 가분할의 성능 향상이 필요하다. 또한, 우편번호 인식 결과를 고속, 저성능의 문자검증기를 이용하여 주소영상으로부터 효과적으로 검증하기 위해서는 유용한 정보를 효과적으로 판별하는 방법론 개발이 필요하다.

참고문헌

[1] Casey, R.G., Lecolinet, E., "A survey of methods and strategies in character segmentation", IEEE Trans. on PAMI, V.18, No.7, 1996, pp.690-706

[2] C-L. Liu, I.-J.Kim and J.H.Kim, Hi-accuracy handwritten Chinese character recognition by improved feature matching method. Proc. ICDAR-4, pp.1033-1037, 1997

[3] T. Joachims, Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning, B. Scholkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999

[4] Gyeonghwan Kim, Venu Govindaraju, "A Lexicon Driven Approach to Handwritten Word Recognition for Real-Time Applications", IEEE Trans. on PAMI, Vol.19, No.4, 1997