

論文97-34C-7-7

온라인 연속 필기 문자의 인식을 위한 문자간 구분 및 문자 종류의 결정에 관한 연구

(A Study on Character Segmentation and Determination of Linguistic Type for Recognition of On-line Cursive Characters)

朴康領*, 全炳煥**, 金暢秀*, 金雨成***, 金在熹*

(Kang Ryoung Park, Byung Hwan Jun, Chang Soo Kim, Woo Sung Kim, and Jaihie Kim)

요 약

온라인 문자 인식에 관한 연구가 활발해짐에 따라 최근에는 필기자의 보다 자연스러운 문자 입력을 허용하기 위해, 문자간의 구분 없이 연속적으로 필기하는 연속 필기 인식과 한 종류가 아닌 여러 종류의 문자들을 혼합 필기하는 혼합 문자 인식(예) '한글 Window 95'의 필요성이 증대되고 있다. 일반적으로 연속 필기된 혼합 문자의 인식시에는, 문자의 끝 및 문자의 종류에 대한 정보를 인식기가 알지 못하므로 모든 경우에 대해 인식 경로가 발생하여 인식 도중에 많은 단어 후보들이 발생한다. 본 논문에서는 이와 같이 인식 도중에 발생하는 많은 단어 후보들을 효과적으로 처리하기 위해 이들을 문자가 완성된 후보 그룹과 아직 문자 인식이 진행중인 후보 그룹으로 나누어 관리하였으며, 5개의 단어 선택 속성들을 사용하여 문자가 완성된 후보들중 가장 높은 확신도를 갖는 후보 단어만을 선택함으로써, 인식 진행중에 생성될 불필요한 많은 인식 경로를 제거하였다. 속성들을 결합하여 결정 판단을 내리는 과정에서는 본 논문에서 제안하는 가중치 부여 순위 방법(weighted ranking method)에 의해 각 속성들에 가중치를 반영함으로써 보다 우수한 결정 판단 능력을 나타내게 되었다. 이때 가중치는 최소 자승법(least mean square method)에 의해 실제 필기 데이터를 학습시킴으로써 구했다. 실험 결과, 가중치가 부여된 속성 집합을 사용하여 단어 후보를 선택하였을 경우 가중치를 사용하지 않은 경우보다 우수한 인식 성능을 나타냈다.

Abstract

With the vigorous researches in the character recognition, the need to recognize run-on multilingual handwritten characters is increasing to provide users with more comfortable PUI(Pen User Interface) environments. In general, many intermediate word candidates are generated in run-on multilingual recognition because there is no information of ending position and linguistic kind of character. To remove unnecessary word candidates which are generated in run-on multilingual recognition, we classify them into two groups and select the best candidate among the word candidates in the group where the final character is completed using 5 attributes. In this research, we propose a method in order to select the best one candidate. It is called WRM(Weighted Ranking Method). The weights are adaptively trained by LMS(Least Mean Square) learning rule. Results show that the abilities of decision making using weights are much better than those not using weights.

* 正會員, 延世大學校 電子工學科

(Dept. of Electronic Engineering, Yonsei University)

** 正會員, 國立公州大學校 電子計算學科

(Dept. of Computer Science, Kongju National University)

*** 正會員, 韓國通信 研究開發團

(Research Center, Korea Telecom)

※ 본 논문은 1997년도 한국통신의 정보통신 기초연구사업비의 일부 지원에 의하여 연구되었음.

接受日字:1996年11月7日, 수정완료일:1997年7月4日

I. 서 론

컴퓨터의 보급이 보편화와 함께 초기의 대용량의 컴퓨터는 점점 사라지고, 보다 작고 휴대하기 간편한 컴퓨터에 대한 요구가 증대되고 있다. 이에 따라 부피가 크고 이동중 사용이 어려운 기존의 키보드나 마우스 대신에, 음성이나 필기와 같이 보다 손쉬운 입력 방법을 도입하고자 하는 노력이 지속되어왔다.

온라인 필기 문자 인식은 전자 펜(stylus)과 전자 평판(tablet)을 이용하여 보다 손쉽게 컴퓨터에 입출력을 대신할 수 있는 기술이다. 이와 같은 필기 문자 인식 기술은 인식 성능과 인식 속도적인 측면에서 계속 발전되어 왔으며,^{[1] [2]} 최근에는 사용자에게 보다 자연스러운 문자 입력을 허용하기 위해 연속 필기 인식(run-on handwritten recognition)과 혼합 문자 인식(multilingual character recognition)에 대한 연구의 필요성이 대두되고 있다. 연속 필기 인식이란 연속하여 필기된 문자열을 한 문자씩 분할하여 인식하는 방법이며 크게 외부 분할(external segmentation)과 내부 분할(internal segmentation)의 두 가지 방법으로 구분된다. 외부 분할 방법은 문자 박스나 시간 및 문자 간격 등을 이용하여 필기자가 매번 한 문자의 끝 위치를 인식기에 알려주어야 하기 때문에, 사용자로부터 부담스러운 필기감을 느끼게 한다.^{[1] [2]} 반면 내부 분할 방법에서는 사용자가 문자의 끝 정보를 알려줄 필요 없이 인식기가 자체적으로 한 문자의 끝을 알아내므로 보다 자연스러운 필기를 할 수 있게 된다.^{[3] [4] [5]}

혼합 문자 인식이란 사용자가 여러 언어의 혼합 문자열(예: '한글 windows 95')을 입력하였을 경우에 인식기가 자체적으로 입력 문자의 종류를 결정하여 인식하는 방법이다. 혼합 문자의 인식 방법에는 가장 대표적인 것으로 주종 인식 방법(master-slave recognition)과 병렬 인식 방법(parallel recognition)이 있다. 주종 인식 방법은 입력 획에 대해 하나의 주 인식기만을 가동시키는 방법으로 상대적으로 인식 시간은 빠르지만 오인식과 미인식은 높아지게 된다.^[6] 반면 병렬 인식 방법은 입력 획에 대해 동시에 모든 인식기를 가동시켜 인식하는 방법으로 여러 인식기에서 많은 후보들이 발생하므로 인식 시간은 늘어날 수 있지만 오인식과 미인식은 크게 감소시킬 수 있다. 이 논문에서는 연속해서 필기된 정자체 및 흘림체 한글과

숫자의 혼합 문자열을 인식 대상으로 하며, 필기자에게는 자연스러운 필기 환경을 제공하고 또 보다 나은 인식 성능을 얻기 위해 내부 분할 방법과 병렬 인식 방법을 사용하였다. 이때 흘림체 문자 인식은 곡률 모델링 기법(curvature modelling approach)을 사용하였다.^[7]

연속 필기된 혼합 문자의 인식을 위해 내부 분할 방법과 병렬 인식 방법을 사용하게 되면, 일반적으로 한 문자의 끝이나 입력된 언어의 종류에 대한 사전 정보 없이 각 언어별 인식기가 동시에 가동되므로 많은 인식 후보들을 생성하게 된다. 따라서, 실시간 처리 및 과다하게 발생한 후보 단어로 인한 메모리 확보 문제의 해결을 위해서는 인식 중에 발생하는 많은 단어 후보들을 간추리는 후보 단어의 선택이 필수적으로 요구된다. 그러므로 이 논문에서는 인식 중간에 발생하는 후보 단어들을 문자가 완성된 그룹(G1)과 완성되지 않은 그룹(G2)으로 나누어 인식을 진행하고, 문자가 완성된 그룹내의 후보 단어들에 대해서는 가중치 부여 순위 방법(weighted ranking method)에 의한 단어 선택 속성 집합을 사용하여 최상의 후보 단어만을 선택한다. 이때 가중치는 각 속성의 구분력의 차이를 나타내며, 실제 연속 필기된 혼합 문자열로부터 최소 자승법(least mean square method)을 이용한 적응적인 학습 과정을 통해 구한다.

II. 연속 필기 혼합 문자의 분할 및 인식 방법

일반적으로 정자체 문자보다는 흘림체 문자가, 그리고 한 종류의 문자보다는 여러 종류의 문자가 연속 필기될수록 많은 단어 후보들이 발생한다. 이러한 후보 단어들은 독립적인 인식 경로를 형성하므로 인식 중간의 단어 선택 과정없이 모든 경로를 다 진행시키는 경우, 처리 시간과 메모리 사용량의 폭주(overflow)가 발생한다. 본 논문에서는 연속 필기된 혼합 문자의 인식중에서 생성된 많은 후보 단어들을 처리하기 위해 인식 중간의 후보 단어들을 두 그룹, 즉 현재 문자의 인식이 끝나고 다음 문자의 인식을 위해 대기중인 그룹(G1)과 현재 문자의 인식이 진행 중인 그룹(G2)으로 나누어 관리한다. 이때, G2에 속하는 후보들은 이후 진행되는 인식 결과에 따라 후보 단어의 최종 확신도(certainty)가 달라지기 때문에, 후보 단어의 선택

과정 없이 그대로 유지해야 한다. 반면, G1에 속하는 후보들은 모두 단어의 완성된 지점이 일치하고, 현재의 후보 단어 선택이 향후 진행되는 인식 결과 값에 영향을 미치지 않으므로, 가장 가능성이 높은 하나의 후보만을 남기고 나머지는 제거한다. 이렇게 함으로써, 인식 도중에 생기는 불필요한 후보 단어의 인식 경로를 줄일 수 있으며 그 결과 실시간 인식 및 효율적인 메모리 관리가 가능해진다. 다음 그림 2.1은 본 논문에서 사용한 혼합 문자의 분할 인식 방법에 관한 예를 나타낸 것이다.

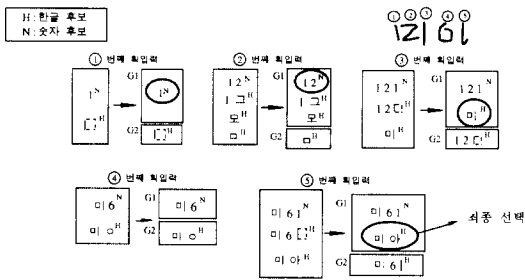


그림 1. 혼합 문자의 분할 인식 방법
Fig. 1. The segmentation and recognition of multilingual characters

그림 1에서 '미아'의 첫 번째 획이 입력되었을 때, 인식 결과 숫자 '1'과 한글의 자음 '미'의 일부분이 생성된다. 이때 이 후보들을 G1과 G2로 분류하고 G1의 후보들은 문자가 완성되었으므로 두 번째 획 입력시 한글과 숫자의 인식을 모두 가동시키고, G2의 후보들은 아직 문자 인식이 진행 중이므로 이전에 진행시키던 인식기만을 진행시킨다. 그 결과 두 번째 획까지의 인식으로 문자가 완성된 단어 후보인 '1 그'와 '1 2'와 '모'가 생성되고, 인식 진행중인 후보로 한글 자음 '미'가 생성되었다. 이때 앞에서와 마찬가지로 이 후보들을 G1과 G2로 분류한 후, G1에 문자가 완성된 단어 후보가 2개 이상 존재하므로 단어의 선택 과정을 통해 최상의 단어 후보인 '1 2'만을 선택하고 나머지 후보들은 제거한다. 이처럼 매 입력획에 대해 생성된 단어 후보들을 그룹별로 나누어 관리함으로써 인식중에 발생하는 많은 후보들을 효과적으로 처리할 수 있다. 다음 그림 2.2는 전체 시스템의 흐름도이다.

III. 단어 선택 속성 집합

연속 필기된 혼합 문자의 인식 중에 마지막 문자가

완성된 단어 후보들에 대해서는 여러 속성들을 적용하여 최상의 단어 후보를 선택하게 된다. 이 논문에서는 5개의 단어 선택 속성들을 동등하게 적용하여 최상의 후보 단어를 선택한다. 이때 사용하는 속성들은 다음과 같다.

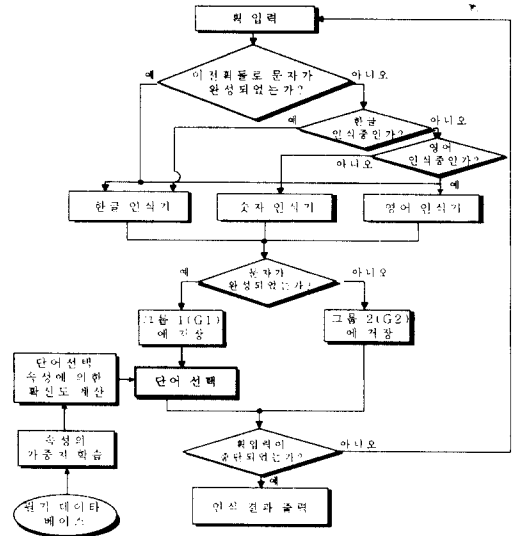


그림 2. 인식 시스템의 흐름도
Fig. 2. The control flow of recognition system.

- [속성1] 후보 단어를 구성하는 개별 문자 내와 문자간에서의 최대 X축 투영 거리차의 합
- ▶ 문자내의 최대 gap과 문자간의 최대 overlap의 합이 가장 작은 단어 후보를 선택

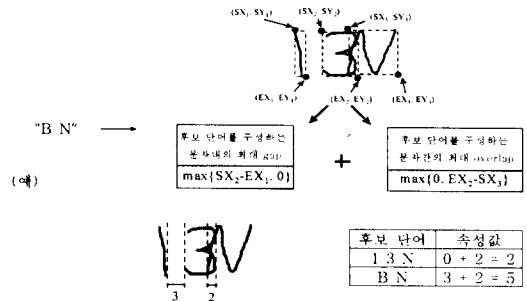


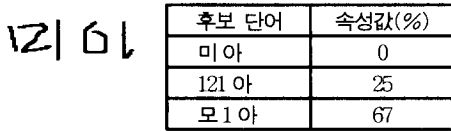
그림 3. 문자내와 문자간에서의 최대 X축 투영 거리차의 합
Fig. 3. The sum of maximum gap and overlap of a word candidate.

- [속성2] 후보 단어 내의 문자 종류에 따른 변환 비율

▶ 문자 종류의 변환 비율이 가장 적은 후보 단어를 선택

$$\text{속성값}(\%) = \frac{\text{후보 단어내의 문자 종류의 변환 횟수}}{\text{후보 단어내의 전체 문자수}} \times 100 \quad (1)$$

(예)



후보 단어	속성값(%)
미아	0
121아	25
모1아	67

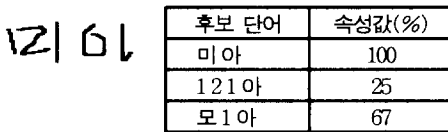
그림 4. 후보 단어내의 문자 종류에 따른 변환 비율
Fig. 4. The language conversion ratio of a word candidate.

[속성3] 후보 단어내의 구성 문자수

- ▶ 가장 적은 수의 문자로 구성된 단어 후보를 선택
- [속성4] 후보 단어내의 사용 빈도수가 높은 문자수
- ▶ 서로 다른 종류의 문자들을 혼합 필기하는 경우, 사용 빈도수가 높은 문자(한글>숫자)의 수가 가장 많은 후보 단어를 선택

$$\text{속성값}(\%) = \frac{\text{후보 단어내의 한글 문자수}}{\text{후보 단어내의 전체 문자수}} \times 100 \quad (2)$$

(예)



후보 단어	속성값(%)
미아	100
121아	25
모1아	67

그림 5. 후보 단어내의 사용 빈도수가 높은 문자의 수
Fig. 5. The number of high frequency language in a word candidate.

[속성5] 후보 단어를 구성하는 문자들의 평균 확산도

- ▶ 후보 단어를 구성하는 문자들에 대한 평균 인식값이 가장 높은 단어 후보를 선택

IV. 가중치 부여 순위 방법과 속성의 가중치 학습 방법

일반적으로 패턴 인식이나 정보 융합 분야 등에서 어떤 특정한 입력 대상에 대해 사용하는 다수의 속성들이 서로 다른 결정을 내릴 때, 이 결정들을 효과적으로 결합하여 최종적인 판단을 내릴 수 있는 방법이

필요하다.

기존의 연구들은 여러 속성값의 범위와 분포를 정규화(normalization)함으로써 다차원 특징 공간상에서 여러 속성들을 결합하는 접근 방법을 사용하였다^[8] [9]. 그런데 속성의 정규화를 위해서는 일반적인 비계수 추정 방법(nonparametric estimation)에서처럼 입력 데이터의 분포에 대한 사전 지식(a priori knowledge)을 갖고 있거나, 혹은 입력 데이터의 분포에 대한 가정((예) 가우시안 분포)을 두어 정규화 계수를 구하는 방법을 사용한다.^[10] 그러나 대부분의 경우 실제 입력 데이터의 분포 형태가 미리 설정된 가정과 일치하지 않는 경우가 많이 존재하며, 이러한 경우 계산된 정규화 계수는 부정확한 결과를 낳게 된다. 또한 각 속성값의 정규화 계수에 대한 실제 계산 과정에서는 충분히 많은 수의 학습 데이터가 필요한데, 본 연구와 같이 다양한 언어가 혼합되어 연속 필기되는 경우에는 발생할 수 있는 후보 단어의 수가 매우 방대해 지므로, 실제로 이 많은 수의 데이터들을 구하여 각 속성값들의 정확한 정규화 계수를 구하는 것은 매우 어렵다. 그러므로 이 논문에서는 속성값들을 그대로 이용하는 것이 아니라 가중치가 부여된 각 속성의 특성에 의해 우수하게 평가되는 순서대로 속성값들을 배열하는 가중치 부여 순위 방법(weighted ranking method)을 이용함으로써, 각 속성값들의 범위와 분포에 대한 정규화 과정 없이 손쉽게 여러 속성들을 결합할 수 있는 새로운 방법을 제안한다.^[11]

일반적으로 순위 방법(rank method)은 각 속성들을 특징 공간축상에 정규화할 필요없이 단지 크기 순으로 배열함으로써 사용할 수 있으므로 계산하기 쉽고 필기 데이터의 실제 분포에 대한 가정이 불필요하며, 사용하는 속성의 종류나 특성에 무관하게 사용할 수 있는 장점이 있다.^[12]

하나의 입력된 데이터에 대해 사용하는 n 개의 속성 A_1, A_2, \dots, A_n ($n \geq 2$)에 의한 순위 벡터 $R = \langle r_1, r_2, \dots, r_n \rangle$ 가 구해지는 경우, 이 순위 벡터를 종합하여 최종 확산도를 구하는 방법에는 다음과 같은 것들이 있다.

최종 확산도 : $C = f(R)$

- [방법1] $f_1(R) = \min(r_1, r_2, \dots, r_n)$ (3)
- [방법2] $f_2(R) = \max(r_1, r_2, \dots, r_n)$ (4)
- [방법3] $f_3(R) = r_1 + r_2 + \dots + r_n$ (5)
- [방법4] $f_4(R) = w_1 r_1 + w_2 r_2 + \dots + w_n r_n$ (6)

[방법1] 과 [방법2] 는 오직 하나의 속성값에 의해 최종 확신도가 결정되므로, 확신도 값이 같은 클래스의 수가 많아지게 되어 올바른 판단을 내리기 어려운 단점이 있다. [방법 3] 은 일종의 다수결의 방법으로 사용하는 각 속성의 구분력이 동일하다는 가정하에서 사용하는 방법이다. 그러나 실제로 사용하는 속성들 사이의 구분력에는 차이가 있으므로 본 논문에서는 [방법 4] 를 이용하여 입력 클래스에 대한 최종 확신도 값을 구한다. 이때 각 속성의 가중치는 실제 필기 데이터로부터 점진적 학습 과정을 통해 구하며, 이는 각 속성의 구분력의 차이를 나타낸다.

하나의 입력된 클래스에 대한 순위 벡터 $R=(r_1, r_2, \dots, r_n)$ 가 구해지는 경우, 각 속성에 의해 높은 순위로 평가된 것을 높은 값으로 할당하고 낮은 순위로 평가된 것은 낮은 값을 할당함으로써 순위 벡터 R 을 역순위 벡터 $IR=(Ir_1, Ir_2, \dots, Ir_n)$ 로 변환한다.

R : n 개의 속성으로부터 구해진 순위 벡터 (rank vector)

IR : n 개의 속성으로부터 구해진 역순위 벡터 (inverse rank vector)

$$R=(r_1, r_2, \dots, r_n) \longrightarrow IR=(Ir_1, Ir_2, \dots, Ir_n)$$

$Ir_i = m - r_i + 1 (1 \leq i \leq n)$ m : 후보 클래스의 수

(7)

입력 클래스에 대한 역순위 벡터 IR 가 구해진 경우 그 입력 클래스에 대한 두 개의 사후 확률(a posteriori probability)이 구해진다.

$p(x \in TrueClass|IR)$: 역순위 벡터 IR 이 주어진 경우 그 입력 클래스가 참일 확률 (8)

$p(x \notin TrueClass|IR)$: 역순위 벡터 IR 이 주어진 경우 그 입력 클래스가 거짓일 확률 (9)

식 (8)과 (9)의 $p(x \in TrueClass|IR)$ 과 $p(x \notin TrueClass|IR)$ 는 실제 입력 데이터로부터 구한 실측 확률이다. 이 경우 임의의 입력 클래스가 n 개의 속성에 의해 높은 순위로 평가될수록, 즉 역순위 벡터의 각 벡터 성분들이 큰 값을 가질수록 사후 확률 $p(x \in TrueClass|IR)$ 는 1에 가까워지게 된다. 또한 각 속성에 의해 낮은 순위로 위치될수록 $p(x \in TrueClass|IR)$ 는 0에 가까운 값을 갖는다. 이와 같이 역순위 벡터 $IR = (Ir_1, Ir_2, \dots, Ir_n)$ 과 $p(x \in TrueClass|IR)$ 사이에는 단조 증가 관계(monotonic increasing relation)이지만,

$p(x \in TrueClass|IR)$ 의 범위가 확률의 성질에 의해 0과 1사이에서 제한되기 때문에 비선형적 단조증가 관계를 이루게 된다. 그러므로 실측 사후 확률 $p(x \in TrueClass|IR)$ 는 0과 1사이의 비선형 값을 갖는 시그모이드 함수(sigmoid function)를 통해 가장 가깝게 표현할 수 있다. 이때 이 시그모이드 함수를 추정 사후 확률 $\hat{p}(x \in TrueClass|IR)$ 라고 나타낸다. 다음 그림 4.1은 1개의 속성을 사용하는 1차원 특징 공간상에서 4개의 입력 클래스에 대한 실측 사후 확률 $p(x \in TrueClass|IR)$ 와 추정 사후 확률 $\hat{p}(x \in TrueClass|IR)$ 사이의 관계를 나타낸 것이다.

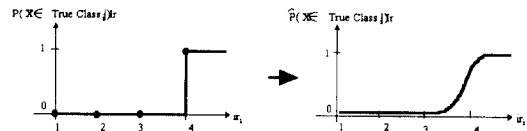


그림 6. 실측 사후 확률과 추정 사후 확률
Fig. 6. The real a posteriori probability and the estimated a posteriori probability.

이와 같은 추정 사후 확률 $\hat{p}(x \in TrueClass|IR)$ 은 시그모이드 함수의 형태를 갖기 때문에 다음과 같이 표현할 수 있다.

$$\hat{p}(x \in TrueClass|IR) = \frac{\exp(\hat{w}_0 + \hat{w}_1 Ir_1 + \dots + \hat{w}_n Ir_n)}{1 + \exp(\hat{w}_0 + \hat{w}_1 Ir_1 + \dots + \hat{w}_n Ir_n)} \quad (10)$$

$$= \frac{\exp(\hat{W}' \cdot IR)}{1 + \exp(\hat{W}' \cdot IR)} \quad (11)$$

$IR = (1, Ir_1, Ir_2, \dots, Ir_n)'$: n 개의 속성으로부터 구해진 역순위 벡터

$\hat{W} = (\hat{w}_0, \hat{w}_1, \dots, \hat{w}_n)'$, \hat{w}_0 : 바이어스 항(bias term)

추정 사후 확률 $\hat{p}(x \in TrueClass|IR)$ 의 모양은 식 (10)에서처럼 가중치와 역순위 벡터값과의 내적(inner product)인 $\hat{w}_0 + \hat{w}_1 Ir_1 + \dots + \hat{w}_n Ir_n$ 로 결정되며, 이때 가중치 벡터 $(\hat{w}_0, \hat{w}_1, \hat{w}_2, \dots, \hat{w}_n)$ 는 실제 필기 데이터에 의한 학습 과정을 통해 구한다. 그리고 이러한 학습 과정을 통해 $p(x \in TrueClass|IR)$ 에 가장 가깝도록 $\hat{p}(x \in TrueClass|IR)$ 가 결정된 순간에 구해진 가중치 벡터 $(\hat{w}_0, \hat{w}_1, \hat{w}_2, \dots, \hat{w}_n)'$ 는 n 개의 속성들의 후보 단어에 대한 구분 능력을 나타내게 된다.

이때 학습에 필요한 가중치와 역순위 벡터값과의 내적을 구하기 위해서는 실측 사후 확률과 추정 사후 확

를 각각 로그 변형(logistic transform)을 거쳐야 한다.

실측 사후 확률의 로그 변형

$$\ln \frac{p(x \in \text{TrueClass} | IR)}{1 - p(x \in \text{TrueClass} | IR)} \quad (11)$$

→ $RPLT(IR)$: Real Probability Logistic Transform

추정 사후 확률의 로그 변형

$$\ln \frac{\hat{p}(x \in \text{TrueClass} | IR)}{1 - \hat{p}(x \in \text{TrueClass} | IR)} = \hat{W}^t \cdot IR \quad (12)$$

$$= \hat{w}_0 + \hat{w}_1 Ir_1 \dots \hat{w}_n Ir_n \quad (13)$$

→ $EPLT(IR)$: Estimated Probability Logistic Transform

로그 변형을 거친 실측 사후 확률과 추정 사후 확률은 각각 n 차원상의 선형 함수(n -dimensional linear function)의 형태를 나타낸다. 이때 선형 함수의 매개변수 $\hat{w}_0, \hat{w}_1, \dots, \hat{w}_n$ 는 선형 회귀법(linear regression method)과 같은 직접적인 방법^{[13] [14]}이나 실제 학습 데이터를 통한 반복적인 학습 방법 등으로 구할 수 있다. 이 중 이 논문에서는 입력 데이터들의 반복적 학습을 통해 매개변수 $\hat{w}_0, \hat{w}_1, \dots, \hat{w}_n$ 를 구하기 위해 최소 자승법(least mean square method)을 이용한다.^[10] 이때 학습 결과 얻어진 가중치 벡터 $\hat{W} = (\hat{w}_0, \hat{w}_1, \hat{w}_2, \dots, \hat{w}_n)$ '는 각각 n 개의 속성 (A_1, A_2, \dots, A_n) ($n \geq 2$)들의 후보 단어에 대한 구분 능력의 차이를 나타내는 값이 된다. 다음은 최소 자승법을 이용한 학습 방법이다.

$$\hat{W}_{k+1} = \hat{W}_k + \rho(RPLT(IR) - EPLT(IR))IR \quad (14)$$

$$= \hat{W}_k + \rho(RPLT(IR) - \hat{W}_k^t \cdot IR)IR \quad (15)$$

$$= \hat{W}_k + \rho(RPLT(IR) - (\hat{w}_0 + \hat{w}_1 Ir_1 \dots \hat{w}_n Ir_n))IR \quad (16)$$

$k = 1, 2, \dots$: 학습의 반복 횟수

가중치 벡터 $\hat{W}_k = (\hat{w}_0, \hat{w}_1, \hat{w}_2, \dots, \hat{w}_n)$ '는 초기에 동일한 값으로부터 시작하여, $EPLT(IR)$ 값이 $RPLT(IR)$ 에 가장 유사해지도록 반복적인 학습 과정을 수행한다.

V. 실험 결과 및 결과 고찰

실험에 사용한 컴퓨터는 IBM PC 486 DX2

(66MHz)이며, WACOM사의 테블릿 디지털타이저를 입력 장치로 사용하였다. 본 논문에서는 가중치의 학습 및 계산된 학습 값에 의한 인식 실험을 한글과 숫자의 혼합 문자열에 대하여 수행하였다. 특히 한글과 숫자의 혼합 문자열은 3가지 방법으로 나누어서 실험했는데 첫째, 한글 문자열만 존재하는 경우(예) : '대한' 둘째, 숫자 문자열만 존재하는 경우(예) : '1996' 그리고 마지막으로 한글과 숫자의 혼합 문자열이 존재하는 경우 등이다(예) : '미아 1가'. 보다 객관성 있는 실험을 위해 한글 문자열은 '연세대학교 한국어 사전 편찬실'에서 선정한 어휘 빈도수가 높은 단어를 사용하였으며, 한글과 숫자의 혼합 문자열은 '1995년 대한전자공학회 추계학술대회 논문집'에서 발췌하였고, 숫자 문자열은 숫자간의 임의의 조합을 고려하여 선정하였다.

그리고 각 실험 데이터에 대해 문자 크기에 대한 자간 거리의 백분율 비를 -10, 10, 50%로 나누어 실험하였으며, 학습 데이터의 수는 한글의 경우 270단어(540문자), 숫자 데이터의 경우 270단어(540문자), 한글과 숫자의 혼합 데이터의 경우 270단어(810문자)를 대상으로 하였다. 그리고 인식 데이터는 각각 450단어(900문자), 450단어(900문자), 450단어(1350문자)를 대상으로 하였다. 다음 그림 5.1은 실험 데이터의 예이다.

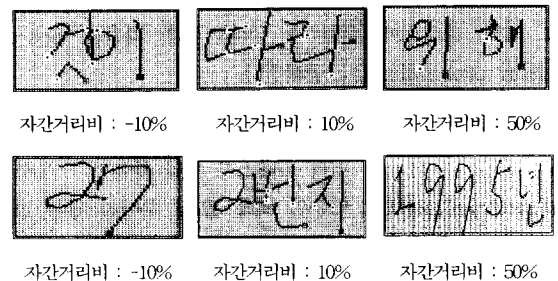


그림 7. 실험 데이터

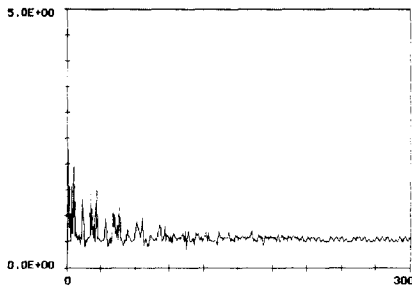
Fig. 7. The experimental data.

식 (4-15)에서 초기 가중치 벡터로 $\hat{W}_0 = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$ '을 사용하였으며, ρ 는 0.1과 0.01의 두 경우에 대해 각각 실험해 보았다. 최소 자승법을 통한 학습 과정의 수렴성을 보이기 위해, 식 (5-1)에서와 같이 실측 확률의 로그 변형($RPLT(IR)$)과 추정 확률의 로그 변형($EPLT(IR)$)의 차에서 계산된 평균 제곱 오차(MSE)를 통해 수렴 여부를 조사하였다.

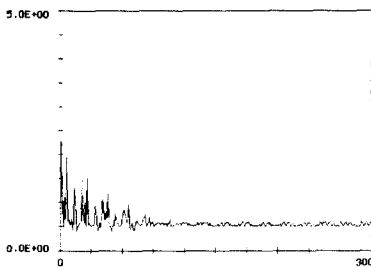
$$MSE(\text{Mean Square Error}) = \frac{1}{N} \sum_{i=1}^N (RPLT(IR) - EPLT(IR))^2 \quad (17)$$

N : 입력 클래스의 수

그림 8은 학습 횟수의 증가에 따른 MSE의 감소를 나타낸 것이다. 각 경우에 대하여 MSE 값은 수렴되는 결과를 나타냈지만, ρ 가 0.1인 경우가 0.01일 때 보다 좀더 빨리 수렴되는 결과를 나타냈다. 학습은 각 속성의 가중치 값이 이전 단계의 학습 결과 얻어진 가중치 값과 동일한 값이 될 때까지 수행하였다.



(a) $\rho = 0.01$, 반복 제한 횟수 = 1000번 최종 MSE값 ≈ 0.54



(b) $\rho = 0.1$, 반복 제한 횟수 = 1000번 최종 MSE값 ≈ 0.54

그림 8. 학습 과정에서의 MSE의 수렴 정도
Fig. 8. The convergence of MSE.

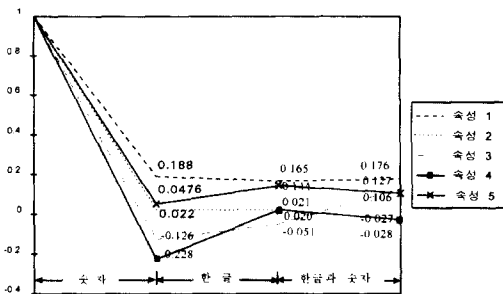


그림 9. 각 속성의 가중치
Fig. 9. The weights of attributes.

위의 실험 결과로부터 실제 필기 데이터의 n차원

특징 공간상의 분포가 본 논문에서 사용한 최소 자승 법과 같은 선형 함수(linear function)에 의해 분류(classification)가능함을 알 수 있었다.

다음 그림 9는 실험 결과 구해진 각 속성의 가중치 변화이다.

처음에 동일한 초기값으로부터 시작한 각 속성의 가중치는 숫자 데이터를 학습시킨 결과 속성 1>속성 5>속성 2> 속성 3>속성 4의 순이었으나, 최종 혼합 데이터의 학습까지 마쳤을 때에는 속성 1>속성 3>속성 5>속성 4>속성 2의 순위로 되었다. 이는 숫자나 한글 데이터의 학습시에는 문자 종류의 변화가 없으므로 속성 2가 중요한 구분력을 나타냈으나, 혼합 데이터의 경우 후보 단어 내에 한글과 숫자의 변환 상태가 존재하므로 상대적으로 속성 2의 구분력이 낮아졌기 때문이다. 또한 속성 3의 경우 적은 수의 문자로 구성된 단어를 선택하는 속성의 특징으로부터 숫자보다는 한글을 우선적으로 선택하기 때문에 숫자 데이터의 학습 결과보다는 혼합 데이터의 학습 결과 가중치가 보다 향상되었음을 알 수 있었다. 다음 표 1에서는 앞에서 구한 각 속성의 가중치를 사용하여 한글, 숫자 및 한글과 숫자의 혼합 데이터의 인식 성능을 나타낸 것이다. 이때 인식 성능의 조사는 인식기 자체에서 발생하는 오인식 문자를 제외한 필기 데이터를 대상으로 문자 인식률을 나타낸 것이다.

VI. 결론

기존의 연속 필기된 혼합 문자의 인식 방법에서는 인식 대상이 정자체로 국한되어 있기 때문에 인식 중간에 발생할 수 있는 문자의 종류가 한정되어 있으므로 모든 인식 경로를 진행시켜 최종 단계에서 단어 후보를 판단할 수 있었다. 그러나 이 논문에서와 같이 인식 대상이 흘림체인 경우 한 획으로도 문자를 완성할 수 있으므로 인식 중간에 발생할 수 있는 후보 문자의 수는 엄청나게 증가된다. 그러므로 메모리 관리 및 실시간 인식을 위해 인식 중간의 단어 선택 과정이 필수적으로 요구된다. 이 논문에서는 이와 같은 문제점을 해결하기 위해 인식 중간의 후보들을 그룹별로 나누어 관리하며 문자가 완성된 그룹내의 단어 후보들은 5개의 단어 선택 속성을 적용하는 가중치 부여 순위 방법에 의해 가장 높은 확신도를 나타내는 후보 단어만 인식을 진행시킨다. 이때 각 속성의 가중치는 최

표 1. 자간 거리에 따른 인식 성능

Table 1. The recognition performance according to intervals between characters.

속성 판정의 결합 방법	데이터 종류	숫자			한글			한글과 숫자		
	자간 거리비(%)	-10	10	50	-10	10	50	-10	10	50
가중치를 부여하지 않은 순위 방법 [방법3]	인식률(%)	67.5	72.7	73.8	68.3	69.7	70.7	66.9	68.8	69.2
가중치를 부여한 순위 방법 [방법4]	인식률(%)	98.5	99.7	99.7	98.3	99.7	99.7	96.9	98.9	98.9

소 자승법을 이용하여 적응적인 학습 과정을 통해 구하였다. 향후 보다 나은 인식 성능을 얻기 위해 현재의 단어 선택 속성 이외에 추가적인 속성의 보완이 필요하며 한글과 숫자의 실제 필기 비율에 대한 통계적인 정보 등을 이용할 수 있는 보다 다양한 학습 데이터를 통한 실험이 필요하다.

참 고 문 헌

- [1] Charles C. Tappert, Ching Y. Suen and Tour Wakahara, "The State of the Art in On-Line Handwriting Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 12, no. 8, pp. 787-808, August 1990.
- [2] M. Hosaka and K. Fumihiko, "An interactive geometrical design system with handwriting," in *Proc. Int. Fed. Information Processing Soc. Congr.*, pp.167-172, 1977
- [3] S. Hanaki, T. Temma and H. Yoshida, "An On-Line Character Recognition Aimed at a Substitution for a Billing Machine Keyboard," *Pattern Recognition*, vol. 8, pp. 63-71, 1976.
- [4] M. Hosaka and K. Fumihiko, "An interactive geometrical design system with handwriting," in *Proc. Int. Fed. Information Processing Soc. Congr.*, pp. 167-172, 1977.
- [5] 정 진영, 전 병환, 김 우성, 김 재희, "온라인 연속 필기 한글의 인식을 위한 내부 문자 분할에 관한 연구," 대한전자공학회논문지, 제32권 B편 제9호, pp. 95-102, 1995년 9월
- [6] 강 민석, 김 진환, 박 수근, 조 문중, 이 현주, "한글 영문 혼용 인식 구현을 위한 마스터 슬라이브 인식 시스템의 설계 기법," 제1회 문자 인식 워크샵, pp. 175-180, 1993년 5월
- [7] B. H. Jun, M. Y. Kim, C. S. Kim, W. S. Kim and Jaihie Kim, "On-Line Cursive Korean Character Recognition by Using Curvature Models," *International Conference on Document Analysis and Recognition*, pp. 1051-1054, Montreal, Canada, 1995.
- [8] L. Xu, A. Krzyzak and C.Y. Suen. "Methods of Combining Multiple Classifiers and Their Applications to Handwriting Recognition," *IEEE Transactions on Systems, Man and Cybernetics*, vol. 22, no. 3, pp. 418-435, May/June 1992.
- [9] Y.S. Huang and C.Y. Suen, "An Optimal Method of Combining Multiple Experts for Handwritten Numerical Recognition," *the Third International Workshop on Frontiers in Handwriting Recognition*, pp.11-20, Buffalo, New York, USA, 1993.
- [10] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, Wesley, New York, 1973.
- [11] K. R. Park, C. S. Kim, B. H. Jun, W. S. Kim and Jaihie Kim, "A Weighted Ranking Method in the On-Line Recognition of Run-On Multilingual Characters," *ITC-CSCC*, pp. 385-388, Seoul, Korea, 1996.
- [12] Tin Kam Ho, Jonathan J. Hull, Sargur N. Srihari, "Decision Combination in Multiple Classifier System," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 16, no. 1, pp 66-75, January 1994.
- [13] A. Agresti, *Categorical Data Analysis*, John Wiley & Sons, 1990.
- [14] D.R. Cox and E. J. Snall, *Analysis of Binary Data*, 2nd. ed., Chapman and Hall, 1989.

— 저 자 소 개 —



朴 康 領(正會員)

1994年 2月 연세대학교 전자공학과 공학사. 1996年 2月 연세대학교 대학원 전자공학과 공학석사. 1996年 3月 ~ 현재 연세대학교 대학원 전자공학과 박사과정. 주관심분야는 컴퓨터 비전, 영상처리, 3D 컴퓨터 그래픽

등임.

金 暢 秀(正會員)

1992年 2月 연세대학교 전자공학과 공학사. 1994年 2月 연세대학교 대학원 전자공학과 공학석사. 1994年 3月 ~ 현재 연세대학교 대학원 전자공학과 박사과정. 주관심분야는 인공지능, 패턴인식, 문자인식 등임.

金 在 燾(正會員) 第 32卷 B編 第 2號 參照

현재 연세대학교 전자공학과 교수

全 炳 煥(正會員)

1989年 2月 연세대학교 전자공학과 공학사. 1991年 8月 연세대학교 대학원 전자공학과 공학석사. 1996年 8月 연세대학교 대학원 전자공학과 공학박사. 1997年 3月 ~ 현재 국립공주대학교 전자계산학과 전임강사. 주관심분야는 패턴인식, 컴퓨터비전, 인공지능, 분산처리 등임.

金 雨 成(正會員)

1981年 연세대학교 전자공학과 학사. 1983年 KAIST 전기 및 전자공학과 석사. 1991年 KAIST 전기 및 전자공학과 박사. 1983年 ~ 현재 한국통신 연구개발원 선임 연구원