

대용량 필기 문자인식을 위한 최소거리 분류법의 성능 개선 전략

김 수 형[†]

요 약

본 논문은 한글이나 한자처럼 문자 부류의 개수가 많은 경우에 효과적인 오프라인 필기 문자인식 알고리즘을 제안한다. 이 알고리즘은 간단하며 구현하기 쉬운 최소거리 분류법에 기반을 두고 있는데, 최소거리 분류법의 인식 성능을 향상시키기 위해 다단계 선인식(multi-stage pre-classification) 및 신경망을 이용한 후보문자 재정렬(candidate reordering)의 두 가지 전략이 첨가되었다. 제안된 알고리즘의 성능은 PE92 데이터베이스 상의 574 종의 한글 문자들에 대한 실험을 통해 입증하였는데, 인식률은 86%, 처리 속도는 초당 15자로서 기존의 연구 결과보다 우수함을 관측하였다.

Performance Improvement Strategies on Minimum Distance Classification for Large-Set Handwritten Character Recognition

Soo-Hyung Kim[†]

ABSTRACT

This paper proposes an algorithm for off-line recognition of handwritten characters, especially effective for large-set characters such as Korean and Chinese characters. The algorithm is based on a minimum distance classification method which is simple and easy to implement but suffers from low recognition performance. Two strategies have been developed to improve its performance: one is multi-stage pre-classification and the other is candidate reordering. Effectiveness of the algorithm has been proven by an experiment with the samples of 574 classes in a handwritten Korean character database named PE92, where 86.0% of recognition accuracy and 15 characters per second of processing speed have been obtained.

1. 서 론

광학 문자인식(OCR: Optical Character Recognition)이란 스캐너를 통해 입력된 문서 영상에서 문자에 해당하는 부분의 내용을 인식하는 기술을 말한다. 통계에 의하면 컴퓨터를 사용하는 시간 중 65% 이상의

시간을 기존 문서의 내용을 입력하거나, 신규로 발생되는 정보를 입력하는데 사용하고 있다고 한다[3, 8]. 따라서, 이러한 정보입력 작업을 OCR을 이용하여 자동화할 경우 막대한 시간적·경제적 이익을 창출할 수 있다. 이러한 OCR 기술은 인쇄된 문서를 인식하는 경우와 필기 문자를 인식하는 경우로 크게 구분되어 연구·개발되고 있다[2, 10, 14].

한글이나 한자처럼 문자 부류의 개수가 많은 동양권 문자에 대한 오프라인 필기 문자인식은 매우 어려운 문제이다. 즉, 문자 부류가 많다는 사실 외에도, 동

* 이 논문은 1997년도 전남대학교 학술연구비 지원에 의하여 연구되었음.

[†] 경 회 원 : 전남대학교 전산학과 교수

논문접수: 1998년 3월 20일, 심사완료: 1998년 7월 16일

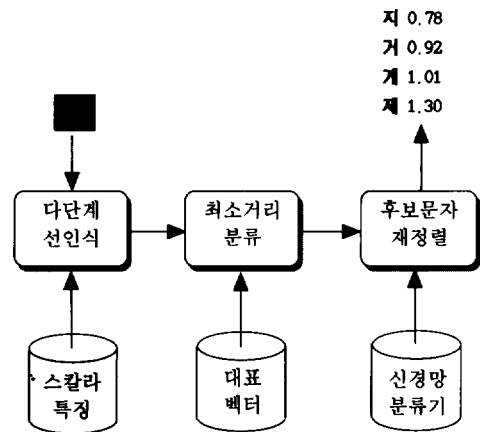
일한 문자에 대한 다양한 필체들 간의 변화를 흡수해야 하고, 또한 유사한 모양을 갖는 다른 문자들 간의 변별력을 높여야 하는 두 가지의 상충된 문제를 동시에 해결해야 한다. 특히, 한글의 경우 24 종류의 자모를 조합하여 최대 11,172 종류의 문자를 형성하기 때문에 유사한 모양을 갖는 문자의 종류가 매우 많아 필기 한글 문자의 인식은 매우 어렵다고 알려져 있다[4, 6, 17-21].

본 논문은 한글이나 한자처럼 문자 부류의 개수가 많은 경우에 효과적인 오프라인 필기 문자인식 알고리즘을 제안한다. 제안된 알고리즘은 최소거리 분류법(minimum distance classification)을 기반으로 구성하였는데, 이 기법은 단순하여 쉽게 구현할 수 있으며 인식 대상 문자 집합이 다양한 응용 분야에 보다 높은 유연성을 지닌다는 장점 때문에 대용량의 필기 문자인식을 위한 대분류 등에 많이 응용되고 있다[12, 13, 15, 17, 20]. 그러나 이 기법은 다른 기법들에 비해 상대적으로 낮은 인식 성능을 보인다는 단점이 있다. 본 논문에서는 최소거리 분류법을 기반으로 대용량의 필기 문자인식 시스템을 구축함에 있어 최소거리 분류법의 단점을 극복하기 위해 다단계 선인식(multi-stage pre-classification) 및 후보문자 재정렬(candidate reordering)의 두 가지 전략을 도입하였다. 다단계 선인식은 인식의 대상이 되는 클래스(class)의 개수를 줄여 최소거리 분류 시에 대표벡터와의 유사도(similarity) 비교에 소요되는 시간을 감소시키기 위한 것이며, 후보문자 재정렬은 최소거리 분류기가 출력한 인식 후보문자의 순위를 재조정하여 인식의 정확도를 향상시키기 위한 것이다.

다단계 선인식 모듈은 필기 한자의 인식을 위해 Tung 등이 제안한 대분류 기법[13]을 한글에 맞도록 수정하여 구현하였으며, 후보문자 재정렬 모듈은 다층 퍼셉트론 신경망(multi-layer perceptron neural network)[9]을 기반으로 유사한 문자의 변별력을 높일 수 있도록 구성하였다. 다단계 선인식, 최소거리 분류, 후보문자 재정렬의 세 가지 모듈로 구성되는 제안된 알고리즘의 성능은 주소 및 성명 등 전표인식 응용 분야에 자주 사용되는 574 종의 한글 문자 클래스에 대한 실험을 통해 입증하였는데, PE92라는 필기 한글 문자 데이터베이스[16]를 사용한 경우의 인식률은 86%, 처리 속도는 초당 15자로서 기존의 연구 결과보다 우수함을 알 수 있었다.

2. 제안된 알고리즘

제안된 알고리즘은 (그림 1)에서와 같이 독립된 세 가지 모듈로 구성된다. 다단계 선인식 모듈은 각각의 문자 클래스 별로 훈련된 몇 개의 스칼라 특징을 기반으로 최소거리 분류 시에 고려할 문자 클래스의 수를 감소시킨다. 최소거리 분류 모듈은 입력문자 패턴과 유사도가 높은 문자 클래스를 추출하기 위해 각 클래스의 대표 벡터와의 거리를 비교하여 유사도가 높은 몇 개의 클래스를 인식 후보로 출력한다. 후보문자 재정렬 모듈은 최소거리 분류기가 출력한 후보 클래스들의 순서를 재정렬하는데, 유사 문자간의 변별력을 높이기 위한 문자쌍별 분류기로서 신경망을 사용한다. 본 절에서는 이들 세 가지 모듈 중 가장 핵심적인 최소거리 분류 방법에 대해 먼저 기술한 후, 이를 개선하기 위한 두 가지 모듈 각각에 대해 설명한다.



(그림 1) 알고리즘 구성도
(Fig. 1) Algorithm structure

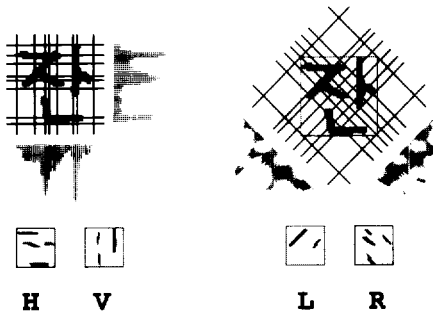
2.1 최소거리 분류

최소거리 분류(MDC: Minimum Distance Classification)는 각각의 인식 대상 클래스를 다차원 특징공간(feature space) 상의 대표벡터로 표현한 후, 입력 문자 영상을 동일한 특징 공간상의 특징벡터로 변환하여 이들 대표벡터와의 거리(distance)를 기준으로 문자간 유사도를 계산하여 입력 문자의 부류(class)를 결정하는 통계적인 인식 방법론이다[1]. 본 연구에서는 각각

의 문자 클래스를 대표하는 대표벡터를 학습벡터 양자화(LVQ: Learning Vector Quantization) 방법[5]으로 훈련하였으며, 임의의 특징벡터가 주어졌을 때 이것과 가까운 몇 개의 대표벡터들을 찾아내어 각 대표벡터의 소속 클래스 및 입력벡터와의 거리를 출력하도록 하였다.

2.1.1 대표벡터 훈련

본 논문에서는 문자 영상에 대한 특징으로서 Yamashita 등이 필기체 한자 인식을 위해 제안한 방향 성분 특징(directional segment feature)[6]을 채택하되, 필기 문자의 세로 대 가로 비율(aspect ratio)의 통계치를 고려하여 9×7×4차원의 벡터를 채택한다. 즉, 입력 문자에 대한 9×7차원의 비선형 메쉬(nonlinear mesh)와 수평(H), 수직(V), 사선(L), 역사선(R) 방향의 성분을 구한 후, 9×7차원 메쉬 상의 63개 셀(cell) 각각으로부터 이들 네 가지 방향 성분에 대한 특징값을 계산하여 9×7×4(=252)차원의 벡터를 형성한다. (그림 2)는 임의의 입력 문자 영상으로부터 7×7×4차원의 방향 성분 특징벡터를 추출하는 예를 보여준다.



(그림 2) 방향 성분 특징 추출의 예 (7×7×4)
(Fig. 2) Example of extracting directional segment feature (7×7×4)

각각의 인식 대상 클래스에 대한 대표벡터는 해당 클래스에 속하는 훈련 문자 영상들의 9×7×4차원의 특징 벡터를 사용하여 LVQ 방식으로 훈련한다. 본 연구에서 채택한 LVQ 알고리즘은 버전 2.1로서[5] 알고리즘의 동작 원리는 다음과 같다. 클래스 i와 클래스 j에 대한 참조 모델 m_i 와 m_j 사이에 클래스 j에 속하는 임의의 훈련 패턴 x 가 m_i 와 m_j 의 이등분 평면을 중심으로 하는 좁은 공간(window) 내에 존재할 때

즉,

$$\min \left(\frac{d(x, m_i)}{d(x, m_i)}, \frac{d(x, m_j)}{d(x, m_j)} \right) > s,$$

$$\text{where } s = \frac{1-w}{1+w}$$

일 때(여기서 $d(x, m_i)$ 는 패턴 x 와 참조 모델 m_i 와의 유클리드 거리이며, w 는 window 크기임), 이들 두 개의 참조 모델을 다음과 같이 갱신하는 과정을 반복함으로써 모든 인식 대상 문자 클래스에 대한 최적의 참조 모델인 대표벡터를 얻게 된다(α 는 학습률에 해당하는 상수임).

$$m_i' = m_i + \alpha [x - m_i],$$

$$m_j' = m_j - \alpha [x - m_j]$$

LVQ 알고리즘의 경우, 각 클래스에 대한 참조 모델의 초기값이 어떤 것인가에 따라 훈련 속도 및 훈련 결과가 달라지게 되는데, 본 연구에서는 각 클래스에 대한 초기의 참조 모델로서 해당 클래스에 속하는 모든 훈련 패턴들의 특징벡터를 평균한 평균벡터를 사용하였다.

2.1.2 최소거리 분류 알고리즘

최소거리 분류 알고리즘으로 임의의 입력 문자를 분류하는 방법은 다음과 같다. 먼저 주어진 입력 문자로부터 9×7×4차원의 방향 성분 특징벡터 x 를 계산한 후, 앞서 훈련된 대표벡터 m_1, m_2, \dots, m_N (N 은 인식 대상 클래스의 개수)와의 거리 $d(x, m_1), \dots, d(x, m_N)$ 를 계산한다. 이들 거리를 두 문자간 유사도의 기준으로 사용하되, 거리가 가까운 것과는 높은 유사도를, 거리가 먼 것과는 낮은 유사도를 갖는 것으로 간주하여 유사도가 높은 5개 내외의 인식 후보 클래스를 결정한다.

최소거리 분류법은 알고리즘이 단순하고 구현이 용이하며, 처리하고자 하는 문자영상 별로 인식 대상 클래스의 개수가 다른 경우에 융통성 있게 대처할 수 있다는 장점을 지닌다. 예를 들어, 한글 성명에 해당하는 문자를 인식할 때에는 인식 대상 문자 클래스가 250개 내외로 감소될 수 있으며, 본 알고리즘이 2,350개의 한글 문자 클래스에 대해 훈련되었다 하더라도 이들 250여개 클래스에 해당되는 대표벡터들만 참조하여 인식

함으로써 인식 대상 문자의 감소 효과를 얻을 수 있다. 그러나 최소거리 분류법은 필기체 문자의 다양성 흡수 및 유사 문자간의 변별력 측면에서 기존의 다른 기법에 비해 성능이 낮다고 알려져 있다. 따라서 이 기법의 장점은 그대로 유지하면서 단점을 보완하기 위해 다음의 두 가지 전략을 고안하였다.

2.2 다단계 선인식

다단계 선인식(MP: Multi-stage Pre-classification)이란 분별력이 뛰어난 몇 개의 스칼라 특징(scalar feature)을 순차적으로 적용하여 전체 시스템의 인식률을 가급적 저하시키지 않으면서 인식 대상 클래스의 수를 줄일 수 있는 전략이다. 스칼라 특징이란 단일값 특징(single-valued feature)으로서 예를 들면 이진영상으로 표현되는 패턴 내의 검은 화소의 수 등과 같은 것이다. 본 연구에서는 Tung 등이 필기 한자의 인식을 위해 제안한 대분류 기법[13]을 기반으로 다단계 선인식 전략을 구현하되, 한글의 특성에 맞는 6개의 스칼라 특징을 채택한 6단계 선인식(six-stage pre-classification)으로 구성하였다.

2.2.1 스칼라 특징의 훈련

선인식에 사용되는 이상적인 스칼라 특징 f 는 특정 문자 클래스 C_i 에 속하는 문자 패턴들에 대하여 동일한 값을 출력하는 함수이다. 즉, 이상적인 선인식은

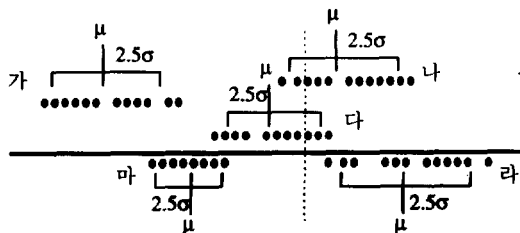
$$f(p) = f(q), \forall p, q \in C_i$$

을 만족하는 함수 f 의 값을 특징으로 사용하여 입력 패턴을 인식하려는 방법이다. 그러나 필기 문자의 경우 패턴의 다양성 때문에 위의 조건을 만족하는 특징을 찾기는 매우 어렵다. 따라서 본 논문에서는 함수 f 의 출력값이 특정 문자 클래스에 대하여 일정한 값을 갖는다는 가정 대신 임의의 값 근처에서 정규분포를 이룬다고 가정하여 입력 패턴을 선인식한다. 즉,

$$f(p) \sim N(\mu_i, \sigma_i^2), p \in C_i, (1 \leq i \leq N)$$

을 가정하면, 인식 대상 문자 클래스 C_1, C_2, \dots, C_N 에 대한 스칼라 특징값 f 의 분포는 개별 문자 클래스에 속하는 훈련 문자 패턴들로부터 추정된 $\langle \mu_1, \sigma_1^2 \rangle, \langle \mu_2, \sigma_2^2 \rangle, \dots, \langle \mu_N, \sigma_N^2 \rangle$ 들로 확률적으로 설명할 수 있다.

임의의 스칼라 특징 f 를 사용한 선인식 방법은 다음과 같다. 먼저 각 문자 클래스의 훈련 패턴들로부터 추정된 평균 및 표준편차 $\mu_i, \sigma_i, (1 \leq i \leq N, N$ 은 문자 클래스의 개수)를 기반으로 개별 문자 클래스에 대한 특징값 f 의 신뢰구간을 $[\mu_i - \alpha\sigma_i, \mu_i + \alpha\sigma_i]$ 과 같이 구한 후(α 는 신뢰구간의 신뢰도를 결정하는 상수로서 예를 들어 $\alpha = 3.0$ 일 때의 신뢰도는 99.73%임), 임의의 문자 패턴 p 를 인식할 때에는 앞서 정의된 N 개의 신뢰구간 중 입력 문자의 스칼라 특징 $f(p)$ 값과 겹치는(overlap) 것들만을 선택하여 선인식의 결과로 출력한다. (그림 3)에는 다섯 가지 클래스 중 입력 문자 패턴으로부터 계산된 스칼라 특징값(점선으로 표시된 위치)과 겹침 관계를 갖는 2개의 클래스만이 선인식에 의해 선택되는 상황을 예시하고 있다. 본 논문에서의 6단계 선인식은 이와 같이 하나의 스칼라 특징을 기반으로 인식 대상 문자 클래스의 개수를 줄이는 선인식 과정을 6번 순차적으로 적용하되, 매번 사용되는 스칼라 특징을 달리하여 감소되는 문자 클래스의 개수를 최소화하고 전체 시스템의 인식을 저하는 최소화하려는 방법이다.



(그림 3) 스칼라 특징에 의한 선인식의 예
(Fig. 3) Example pre-classification by a scalar feature

6단계 선인식을 위해서는 6개의 스칼라 특징을 선정한 후, 이들 각각에 대한 개별 문자 클래스의 신뢰구간을 다음과 같이 훈련해 두어야 한다:

$$[\mu_{ij} - \alpha_{ij}\sigma_{ij}, \mu_{ij} + \alpha_{ij}\sigma_{ij}], (1 \leq i \leq N, 1 \leq j \leq 6).$$

여기서 μ_{ij} 와 σ_{ij} 는 문자 클래스 C_i 에 대한 j 번째 스칼라 특징의 평균과 표준편차를 의미하며, α_{ij} 는 선인식 성능을 극대화하기 위해 개별적으로 선택된 신뢰도 상수이다. 이들 6개의 특징은 최소거리 분류 시에 사용되는 252개의 방향 성분 특징(2.1.1절 참조)을 기반

으로 생성한 64개의 스칼라 특징들 중 선인식 성능이 우수한 6개를 선정한 것이다. 즉, 방향 성분 특징 계산 시에 구성한 9×7 메쉬를 9×1 메쉬와 1×7 메쉬로 간략화한 후, 이들 16개의 메쉬 셀(cell)로부터 4가지 방향 성분에 대한 특징값을 계산한다. 이렇게 얻어진 64개의 특징 f_1, f_2, \dots, f_{64} 로부터 선인식 성능이 우수한 6개를 선정하기 위해 다음과 같은 hill-climbing 기법[1]을 적용한다.

훈련 문자 패턴으로부터 모든 인식 대상 클래스에 대해 64개 스칼라 특징의 신뢰구간을 계산한 후, 64개의 스칼라 특징 각각을 사용하여 모든 훈련 문자 패턴들을 대상으로 선인식을 수행한다. 이들 중 성능이 가장 우수한 특징을 선정하여 첫 번째 적용할 스칼라 특징 $f_{(1)}$ 으로 결정한다. 이때 성능 평가의 기준으로 감소율(reduction rate)을 사용한다. 즉, 각 스칼라 특징의 신뢰구간에 대한 신뢰도를 일정하게 고정시킨 후, 선인식 시에 인식 대상 문자 클래스를 가장 많이 제거시키는 특징을 가장 우수한 특징으로 간주한다. 다음으로 $f_{(1)}$ 을 제외한 나머지 63개의 스칼라 특징 각각을 사용하여 $f_{(1)}$ 을 적용한 후에 남게되는 모든 문자 클래스들을 대상으로 선인식을 수행하여 이들 중 감소율이 가장 우수한 특징 $f_{(2)}$ 를 결정한다. 이와 같은 과정을 반복하여 여섯 개의 스칼라 특징 $f_{(1)}, \dots, f_{(6)}$ 을 선정한다.

2.2.2 다단계 선인식 알고리즘

여섯 개의 스칼라 특징 $f_{(1)}, \dots, f_{(6)}$ 을 이용한 6단계 선인식은 다음과 같이 수행된다. 먼저, 입력 패턴 p 로부터 여섯 개의 스칼라 특징값 $f_{(1)}(p), \dots, f_{(6)}(p)$ 을 구한다. 앞 절에서 언급한 바와 같이 이들 여섯 개의 스칼라 특징은 최소거리 분류 시 입력 문자 영상으로부터 계산하는 252차원의 방향 성분 특징으로부터 간단하게 계산할 수 있다. 다음에는, 첫 번째 스칼라 특징값 $f_{(1)}(p)$ 과 $f_{(1)}$ 의 문자 클래스별 신뢰구간 $[\mu_{i1} - \alpha_{i1}\sigma_{i1}, \mu_{i1} + \alpha_{i1}\sigma_{i1}]$ ($1 \leq i \leq N$)을 기반으로 선인식을 수행하게 되는데, 이들 N 개의 신뢰구간 중 $f_{(1)}(p)$ 를 포함하는 것들을 선택한 후, 선택된 신뢰구간에 해당하는 문자 클래스들을 선인식의 결과로 출력한다. 두 번째 스칼라 특징값 $f_{(2)}(p)$ 를 사용한

선인식에서는 $f_{(1)}(p)$ 를 사용한 선인식에 의해 출력된 문자 클래스의 $f_{(2)}$ 신뢰구간들만을 대상으로 하여 인식 대상 문자 클래스를 줄이게 되며, $f_{(3)}(p)$ 를 사용한 선인식에서는 $f_{(1)}(p)$ 및 $f_{(2)}(p)$ 를 이용한 선인식을 연속으로 적용하여 생성한 클래스 집합을 대상으로 유사한 방식으로 선인식을 적용한다. 이와 같은 과정을 $f_{(6)}(p)$ 를 사용한 선인식까지 반복하면 6개의 스칼라 특징에 의한 6단계 선인식의 결과를 얻을 수 있다.

지금까지 설명한 6단계 선인식에 의해 출력되는 문자 클래스의 집합은 다음의 식을 만족하는 집합 C 와 같다:

$$C = \{C_i | \mu_{ij} - \alpha_{ij}\sigma_{ij} \leq f_{(j)}(p) \leq \mu_{ij} + \alpha_{ij}\sigma_{ij}, 1 \leq i \leq N, 1 \leq j \leq 6\}.$$

여기서 α_{ij} 는 스칼라 특징 j 에 대한 선인식 성능을 최적화하기 위해 문자 클래스 별로 설정된 상수로서 보통 3.2~4.4 사이의 값을 갖는다.

본 알고리즘에서 사용한 6단계 선인식기는 약 58%의 감소율(reduction rate)과 98.3%의 성공률을 갖는다. 감소율 58%는 최소거리 분류 시 인식 대상 클래스가 42%로 줄어들음을 의미하는데, 실제 6단계 선인식에 의해 추가되는 처리 시간을 고려하면 약 두 배 정도의 속도 개선 효과를 기대할 수 있다 - 6개의 특징값을 사용하는 다단계 선인식의 계산량은 252개의 특징값을 사용하는 최소거리 분류 모듈의 약 2.4%(= 6/252) 정도이다. 성공률 98.3%는 다단계 선인식에 의해 감축된 인식 대상 클래스 집합 내에 입력 문자 영상에 해당하는 클래스가 포함되어 있을 확률로서, 수치적으로만 보면 전체적으로 최대 1.7%의 인식을 감소를 초래할 수 있으나, 실험에 의해 최소거리 분류의 1등 인식을 이 오히려 증가함을 알 수 있었다(3절 참조).

2.3 후보문자 재정렬

후보문자 재정렬(CR: Candidate Reordering) 단계는 최소거리 분류기가 유사도 순으로 출력한 5개 내외의 후보 클래스(또는 후보문자)의 순서를 재정렬하는 과정이다. 본 논문에서는 후보문자의 재정렬을 위해 유사하게 생긴 한글 문자쌍 간의 변별력을 높일 수 있는 분류기를 사용한다. 예를 들어 최소거리 분류기가 1위

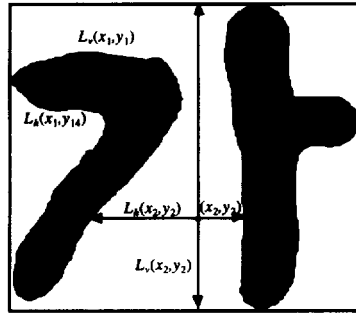
후보문자부터 5위 후보문자까지를 유사도와 함께 출력하였다면, 후보문자 재정렬 단계에서는 5위 후보문자와 4위 후보문자 쌍에 대한 변별력을 높일 수 있는 인식기를 적용하여 이들의 순서를 재조정 한 후, 재조정 에 의해 결정된 4위 후보문자와 기존의 3위 후보문자 쌍에 대한 변별력을 높일 수 있는 인식기를 적용한다. 이와 같은 과정을 반복하여 최종적으로 재조정된 2위 후보문자와 기존의 1위 후보문자 쌍에 대한 변별력을 높일 수 있는 인식기를 적용하면 최소거리 분류기가 출력한 후보문자들이 재정렬 된다. 본 연구에서는 문자쌍 별로 변별력을 높일 수 있는 분류기로서 다층 퍼셉트론 신경망(multi-layer perceptron neural network)을 채택하였다.

2.3.1 신경망 훈련

후보문자 재정렬을 위해 채택한 신경망은 3개의 층으로 구성된 전향(feed-forward) 신경망으로서 입력층에는 입력 문자 영상으로부터 추출한 9×7×4(252)차원의 방향 기여도 특징(directional contributivity feature)이 입력되며, 은닉층(hidden layer) 노드의 개수는 5개를 사용하고, 출력층 노드의 개수는 2개이다. 또한, 입력층과 은닉층에는 편기향(bias) 노드가 추가되었으며, 은닉층 및 출력층 노드에서의 활성화 함수로는 다음과 같은 시그모이드(sigmoid) 함수를 사용하였다.

$$f(x) = \frac{\exp(x)}{1 + \exp(x)}$$

입력 특징으로 사용되는 9×7×4차원의 방향 기여도 특징은 방향 성분 특징(directional segment feature - 2.1.1절 참조) 계산 시에 구성한 9×7 메쉬를 그대로 사용하여 계산하되, 각 셀(cell)에서의 특징으로는 (그림 4)에서와 같이 흑화소 (x_1, y_1) 에 대한 수평 및 수직 방향 기여도 $L_h(x_1, y_1)$, $L_v(x_1, y_1)$ 와 바탕화소 (x_2, y_2) 에 대한 수평 및 수직 방향 기여도 $L_h(x_2, y_2)$, $L_v(x_2, y_2)$ 각각을 셀 내의 모든 화소에 대해 평균하여 계산한다. 각 노드 사이의 연결강도(weight)를 훈련하기 위해 목표값과 출력값의 차이에 대한 평균 제곱합(MSE: mean square error)을 최소화하도록 조정하는데, 이때 MSE에 대한 최소값을 찾기 위해 오류 역전파(error backpropagation) 학습 알고리즘[9]을 채택하였다.



(그림 4) 흑화소 및 바탕화소의 방향 기여도 (Fig. 4) Directional contributivity for black and white pixels

두 개의 문자 클래스간의 변별력을 높이기 위한 신경망 분류기는 문자 클래스의 수가 N 개일 때 가능한 모든 문자 쌍의 개수 만큼에 해당하는 $N C_2$ 개가 필요하다. 따라서 본 재정렬 방법으로는 인식 대상 문자 클래스의 수가 많아짐에 따라 쌍별 분류에 필요한 신경망의 개수가 기하 급수적으로 증가하여 많은 기억 공간을 필요로 하며, 이들 신경망을 사전에 훈련하는데 소요되는 시간도 커다란 문제로 대두된다. 이러한 문제를 해결하기 위해 자주 발생하는 유사 문자쌍에 대해서만 신경망을 사용하고 이들 문자쌍 외에는 최소거리 분류기의 결과를 그대로 사용하는 절충적인 방법을 사용하여 시스템을 구현한다.

후보문자 재정렬시 자주 발생하는 유사 문자쌍을 선정하기 위해 훈련 문자 패턴에 대해 최소거리 분류기가 출력하는 후보문자 클래스들로부터 가능한 모든 문자쌍을 조사하고, 이들을 빈도수 기준으로 정렬한 후, 상위 K 개의 문자쌍만을 유사 문자쌍으로 선정하여 이들에 대한 신경망 분류기를 구축한다. 실험에 의하면 가능한 모든 문자쌍 중 빈도수 기준 상위 약 1.4% 정도의 문자쌍이 재정렬 시 발생하는 전체 유사 문자쌍의 50% 정도를 차지함을 알 수 있었다. 신경망의 개수를 증가시킬수록 전체 알고리즘의 인식률이 향상됨을 예측할 수 있으나, 신경망 개수의 증가는 곧 기억 공간 및 수행 시간의 증가를 의미하기 때문에 필요에 따라 신경망의 개수를 조정할 수 있다. 이들 신경망을 저장하기 위한 기억 용량 또한 많이 필요하지만, 이들을 하나의 화일로 병합(merge)하여 보조기억장치 상에 저장해 둬서 약간의 속도 저하를 감수하는 대신 이들 신경망을 위한 주기억장치 요구량을 최소화하였다.

2.3.2 후보문자 재정렬 알고리즘

후보문자 재정렬 단계에서는 두개의 문자쌍을 대상으로 인식하는 신경망 인식기를 최소거리 분류기가 출력한 후보 문자에 대해 순차적으로 적용하여 후보 문자들의 순서를 재조정한다. 최소거리 분류기가 다음과 같이 1위 후보문자 C_1 부터 k 번째 후보문자 C_k 까지를 유사도 S_i ($1 \leq i \leq k$)와 함께 출력하였다고 가정한다:

$$\langle C_1, S_1 \rangle, \langle C_2, S_2 \rangle, \dots, \langle C_k, S_k \rangle.$$

후보문자 재정렬 단계에서는 k 번째 후보문자 C_k 와 $k-1$ 번째 후보문자 C_{k-1} 의 두 문자쌍에 대한 신경망 인식기 $NN_{C_k, C_{k-1}}$ 를 적용하여 이 신경망의 출력 결과를 $k-1$ 번째 후보문자로 간주한다.

$$C_{k-1}' = NN_{C_k, C_{k-1}}$$

이렇게 얻은 $k-1$ 번째 후보문자 C_{k-1}' 와 $k-2$ 번째 후보문자 C_{k-2} 쌍에 대한 신경망 인식기 $NN_{C_{k-1}', C_{k-2}}$ 를 적용하여 이 신경망의 출력 결과를 $k-2$ 번째 후보문자로 간주한다.

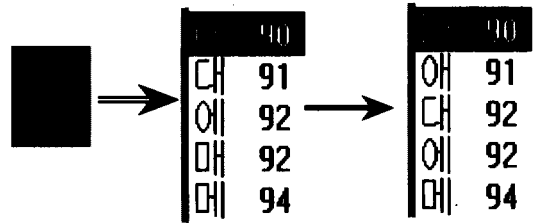
$$C_{k-2}' = NN_{C_{k-1}', C_{k-2}}$$

이와 같은 과정을 반복하여 최종적으로 2위 후보문자 C_2' 와 1위 후보문자 C_1 쌍에 대한 신경망 인식기 NN_{C_2', C_1} 를 적용하여 그 결과를 새로운 1위 후보로 간주한다면 최소거리 분류기가 출력한 후보문자들이 재정렬 된다.

$$C_1' = NN_{C_2', C_1}$$

즉, $\langle C_1, S_1 \rangle, \langle C_2, S_2 \rangle, \dots, \langle C_k, S_k \rangle$ 이 $\langle C_1', S_1 \rangle, \langle C_2', S_2 \rangle, \dots, \langle C_k', S_k \rangle$ 로 재정렬 된다.

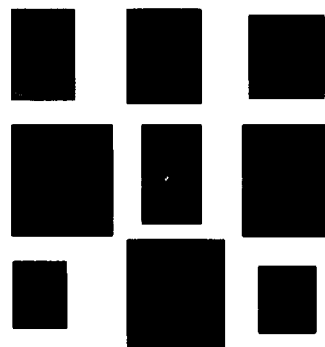
본 연구에서는 주기억 용량 및 훈련시간의 절약을 위해 자주 발생하는 유사 문자쌍에 대해서만 신경망을 구성한 관계로, 토너먼트 방식으로 진행되는 후보문자 재정렬 과정에서 특정 문자쌍에 대한 신경망 인식기가 존재하지 않을 수도 있다. 이 경우, 최소거리 분류기의 출력 결과를 그대로 사용한다. (그림 5)는 최소거리 분류기가 출력한 5개의 후보문자들을 재정렬하는 예를 보여준다.



(그림 5) 후보문자 재정렬의 예
(Fig. 5) An example of candidate reordering

3. 실험 결과

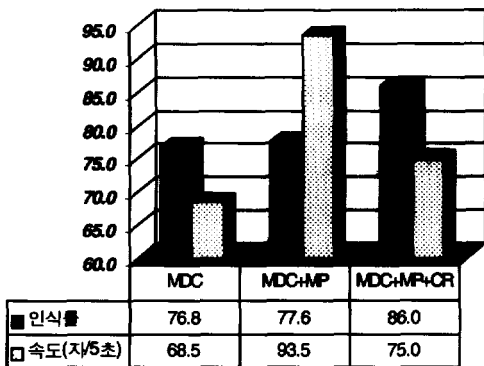
본 논문에서 제안한 대용량 필기 문자인식 알고리즘은 펜티엄 120MHz PC에서 C언어로 구현되었다. 구현된 알고리즘의 성능 분석을 위해 전표인식 응용 분야에 자주 사용되는 한글 574종[17]을 대상으로 하는 인식기를 구축하였는데, 시스템을 구성하는 각종 모델의 훈련 및 인식을 측정할 위해 PE92라는 필기 한글 문자 데이터베이스[16]를 사용하였다. PE92는 완성형 한글 2,350종에 대한 100세트의 문자 패턴으로 구성되어 있는데, 이 중 67세트는 각 모듈의 훈련을 위해 사용하고, 훈련에 사용되지 않은 나머지 33세트를 알고리즘의 성능 분석을 위해 사용하였다. (그림 6)에는 PE92의 상의 문자 클래스 '한'에 대한 몇 가지 패턴이 예시되어 있는데, 여기서 알 수 있듯이 잘못 분류된 문자 뿐 아니라 영상 자체가 매우 희미하여 사람도 판별하기 어려운 패턴이 다수 존재한다. 본 연구에서는 PE92 내의 전체 문자 패턴 중 약 4%에 해당하는 이들 잘못된 패턴을 훈련 및 테스트 과정에서 제외하였다.



(그림 6) PE92의 문자 패턴의 예
(Fig. 6) Sample patterns in PE92

기존의 필기 한글 인식 기법들로는 각 문자 클래스 모델을 자소 단위의 그래프 형태로 표현하여 이들과의 매칭에 의해 입력 문자를 인식하는 방법[4], 계층적 신경망을 이용한 방법[6], 군집화와 신경망을 결합한 통계적 인식 방법[17], 은닉 마르코프 메쉬 랜덤 필드에 기반한 방법[19] 등이 있다. 이들 기존 연구들과 본 연구와는 알고리즘의 동작 환경, 인식 대상 문자 등의 제약 조건 및 사용된 데이터베이스가 일치하지 않아 절대적 성능을 비교할 수는 없지만, 기존 연구의 경우 인식률이 약 70~84% 정도라고 발표되어 있다.

세 가지 모듈로 구성된 제안된 알고리즘의 인식률은 86.0%이며 처리 속도는 초당 약 15자이다. (그림 7)에는 세 가지 모듈의 선택적 구성에 따른 인식 성능이 비교되어 있는데, 최소거리 분류(MDC) 모듈만을 사용하는 경우의 인식률은 76.8%, 처리 속도는 13.7자/초인 반면, 최소거리 분류에 다단계 선인식(MP) 모듈을 결합할 경우의 인식률과 속도는 각각 77.6%, 18.7자/초로서 인식률과 처리속도 측면의 성능이 모두 향상되었음을 알 수 있다. 특이할 만한 사항으로는 다단계 선인식에 의해 인식률이 감소할 것으로 예측할 수 있으나 오히려 증가한 사실이다. 이는 다단계 선인식 모듈이 추후 최소거리 분류기에 의해 잘못 인식될 문자 클래스를 사전에 제거해 주는 효과가 있음을 입증하고 있다. 후보문자 재정렬(CR)을 포함한 세 가지 모듈 모두를 결합한 경우에는 인식률과 속도가 각각 86.0%, 15.0자/초로서 최소거리 분류 모듈만을 사용하는 경우에 비해 상당한 만큼의 성능 개선이 있음을 알 수 있다.



(그림 7) PE92에 대한 인식 성능
(Fig. 7) Recognition performance on PE92

4. 결 론

본 논문은 한글이나 한자처럼 문자 부류의 개수가 많은 경우에 효과적인 오프라인 필기 문자인식 알고리즘을 제안하였다. 제안된 알고리즘은 최소거리 분류법을 기반으로 다단계 선인식 및 후보문자 재정렬 기법을 결합하여 구성함으로써, 최소거리 분류법의 장점은 그대로 유지하면서 인식 속도가 느리고 인식률이 낮은 단점을 개선하여 대용량의 필기체 문자인식에 적합하도록 하였다. 제안된 알고리즘을 필기 한글 문자 인식 문제에 적용하여 그 성능이 우수함을 증명하였다.

대용량의 필기 문자인식 분야에 대한 기존의 연구뿐 아니라, 본 연구에서 제안된 알고리즘을 실제 응용 분야에 그대로 적용하기에는 아직도 절대 성능이 낮다. 따라서 현실적 응용을 위해서는 응용 분야의 환경적·문맥적 지식 등을 최대한 활용하여 해당 분야에 적합하도록 알고리즘을 적용시켜 성능을 극대화해야 하고, 아울러 문자분리, 문자인식 및 후처리 모듈 등 관련 기술의 강결합(tight-coupling)을 통해 각 모듈의 단점을 상호 보완할 수 있어야 한다. 본 연구의 향후 계획은 개발된 문자인식 모듈을 주변 요소 기술과 강결합하는 것이다.

참 고 문 헌

- [1] R.O. Duda and P.E. Hart, 'Pattern Classification and Scene Analysis', Wiley-Interscience Pub., 1973.
- [2] F.R. Jenkins, T.A. Nartker and S.V. Rice, "Result of the Fifth Annual Test of OCR Technology by UNLV's Information Science Research Institute," Inform, pp.20-25, Sep.1996.
- [3] M. Katz, "Recognition Technology," Technology Forecast: 1995, Price Waterhouse, pp.423-434, 1994.
- [4] H.Y. Kim, J.W. Suh, J.O. Kwon and J.H. Kim, "Current Hangul Recognition Activities in Korea," in Proc. French-Korean Workshop on Man-Machine Handwritten Communication, France, pp.1-14, 1996.
- [5] T. Kohonen, et al., "LVQ_PAK-the Learning Vector Quantization Program Package Version 3.1," April 1995.

[6] S.W. Lee and H.S. Park, "Multi-Lingual Large-Set Oriental Character Recognition Using a Hierarchical Neural Network Classifier," Computer Processing of Oriental Languages, Vol.10, No.2, pp.129-145, 1996.

[7] G.J. McLachlan, 'Discriminant Analysis and Statistical Pattern Recognition,' Wiley Series in Probability and Mathematical Statistics, Wiley, New York, 1992.

[8] M.S. Meley, "Competitive Outlook: OCR Solutions," Document Imaging Systems, Datapro Information Services Group, Oct. 1994.

[9] D.E. Rumelhart, G.E. Hinton and R.J. Williams, "Learning Representations by Back-propagating Error," Nature, Vol.332, pp.533-536, 1986.

[10] C.Y. Suen, M. Berthod and S. Mori, "Automatic Recognition of Handprinted Characters-The State of the Art," IEEE Proceedings, Vol.68, No.4, pp.469-485, 1980.

[11] H. Takahashi and T.D. Griffin, "Recognition Enhancement by Linear Tournament Verification," in Proc. 2nd Int. Conf. Document Analysis and Recognition, Japan, pp.585-588, 1993.

[12] J. Tsukumo, "Handprinted Kanji Character Recognition Based on Flexible Template Matching," in Proc. 11th Int. Conf. Pattern Recognition, pp.483-486, 1992.

[13] C.H. Tung, H.J. Lee and J.Y. Tsai, "Multi-Stage Pre-Candidate Selection in Handwritten Chinese Character Recognition Systems," Pattern Recognition, Vol.27, No.8, pp.1093-1102, 1994.

[14] R.A. Wilkinson, J. Geist, et al., "The First Optical Character Recognition Systems Conference," Technical Report NISTIR 4912, National Institute of Standards and Technology, 1994.

[15] Y. Yamashita, K. Higuchi, Y. Yamada and Y. Haga, "Classification of Hand-Printed Kanji Characters by the Structured Segment Matching Method," Pattern Recognition Letters, Vol.1, No.5, pp.475-479, 1983.

[16] 김대환, 방승양, "한글 필기체 영상 데이터베이스 PE92의 소개," 제4회 한글 및 한국어 정보처리 학술발표 논문집, pp.567-575, 1992.

[17] 김수형, 도정인, 김준호, 이상규, "통계적 군집화 및 신경망 기법을 이용한 필기체 문자의 오프라인 인식," 제2회 문자인식워크샵 발표논문집, 웨라튼워커 힐 호텔, pp.133-142, 1994.

[18] 김은진, "인간의 한글 지각 처리에 근거한 한글 인식 방법에 관한 연구," 박사학위논문, 연세대학교, 1997.

[19] 박희선, "은닉 마르코프 메쉬 랜덤 필드: 이론 및 필기체 문자 인식에의 응용," 박사학위논문, 충북대학교, 1996.

[20] 이성환, "문자인식-이론과 실제," 제1권, 홍릉과학출판사, pp.157-178, 1994.

[21] 황영섭, "제약조건 만족 그래프를 이용한 필기한글 인식," 박사학위논문, 포항공과대학교, 1997.



김수형

1982년 서울대학교 컴퓨터공학과 졸업(학사)

1986년 한국과학기술원 전산학과 졸업(공학석사)

1993년 한국과학기술원 전산학과 졸업(공학박사)

1990년~1996년 삼성전자 멀티미디어연구소 선임연구원

1997년~현재 전남대학교 전산학과 조교수

관심분야 : 패턴인식, 컴퓨터 비전, 신경망 학습