확률적 VQ 네트워크와 계층적 구조를 이용한 인쇄체 한자 인식

이 장 훈[†] · 손 영 우[†] · 남 궁 재 찬[††]

요 약

본 논문에서는 확률적 VQ 네트워크와 계층적 구조를 가지는 다단계 인식기를 이용한 인쇄체 한자 인식 방법을 제안한다. 대용량 신경망은 구현하기가 매우 어렵기 때문에 모듈화된 신경망을 이용하였으며, 이 과정에서 발생되는 문제점을 확률적 신경망 모델을 이용으로 제거하였다. 또한 엔트로피 이론을 적용하여 오인식률이 높은 혼동 문자쌍에 대하여 재분류를 수행하였다.

실험대상은 KSC5601 코드의 한자 4,888자 중, 동자이음문자를 제외한 4,619자로 하였으며, 학습 데이타와 실험 데이타에 대하여 실험결과, 각각 평균 99.33%, 92.83%의 인식률과 초당 4-5자의 인식속도를 얻음으로써 본 방법의 유효성을 보였다.

The Recognition of Printed Chinese Characters using Probabilistic VQ Networks and Hierarchical Structure

Jang Hoon Lee[†] · Young Woo Shon[†] · Jae Chan Namkung[††]

ABSTRACT

This paper proposes the method for recognition of printed chinese characters by probabilistic VQ networks and multi-stage recognizer has hierarchical structure. We use modular neural networks, because it is difficult to construct a large-scale neural network. Problems in this procedure are replaced by probabilistic neural network model. And, Confused Characters which have significant ratio of miss-classification are reclassified using the entropy theory.

The experimental object consists of 4,619 chinese characters within the KSC5601 code except the same shape but different code. We have 99.33% recognition rate to the training data, and 92.83% to the test data. And, the recognition speed of system is 4-5 characters per second. Then, these results demonstrate the usefulness of our work.

1. 서 론

국내의 상용 문자인식기는 아직 실용화 초기단계라 할 수 있지만, 인쇄체 한글에 대해서 많은 연구가 진행되어 왔다. 반면, 한자인식에 관한 연구는 다른 한자문화권 국가에 비하여 미흡한 실정이다. 또한 한자의 입력과정은 한글에 비하여 복잡하며, 입력자가 한자에 대한 지식을 가지고 있어야 한다는 점을 감안

† 정 회 원:광운대학교 컴퓨터공학과
†† 종신회원:광운대학교 컴퓨터공학과 교수, 신기술 연구소
논문접수:1996년 11월 20일, 심사완료:1997년 5월 29일

한다면 그 필요성을 쉽게 짐작할 수 있다. 우리 나라의 문서와 책의 내용에 따라 다르지만, 많은 양의 문서와 책에서 한자가 쓰이고 있고, 인명, 지명 등의 고유명사가 한자로 표기된다는 점을 고려해 볼때 활발한 연구가 요구되고 있다.

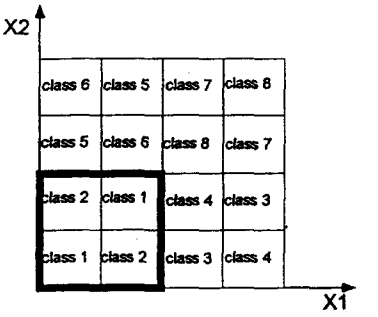
한자인식의 문제점으로 많은 글자수, 문자간의 유사성과 복잡한 문자의 구조를 들 수 있다. 국외의 연구 사례를 살펴보면, 문자인식 초창기라고 할 수 있는 80년대 초에는 원형정합 방법이나 확율통계적 방법, 구조적 방법 등이 사용되어 왔으며, 특히 C. Y. Suen은 엔트로피 이론을 Decision Tree에 접목하였다[1][2]. 최근에는 신경망을 이용한 방법이 주로 사용되고 있는데 대표적인 연구로는 A. Iwata의 CombNET을 이용한 JIS 제1수준의 인쇄체 한자 2,965자에 대한 인식을 들 수 있다[3]. 국내의 경우, 고등학교 교육용 한자 1,800자에 대상으로 부수의 위치에 따라 14형식으로 분류한 후, MLP를 이용 형식분류단과 인식단을 구성한 연구가 발표되었으나, 이는 문자의 유사도와 상관없이 강제적인 형식분류를 행함으로서 형식분류에서 발생하는 오류에는 적절히 대처하지 못하였다[4]. 최근에는 Self-Organizing Neural Network을 이용 한글 2,850자, 한자 4,888자를 대상으로 한 연구가 발표되었다[5]. KSC5601 부호체계에는 검색 및 정렬을 쉽게 하기 위해서 동자이음자(예:更 경/갱),두음법칙관련자(예:女 여/녀)를 포함하고 있다[6]. 이런 문자들은 문자의 모양으로는 구별해 낼 수 없으며 후처리 과정에서 교정하여야 한다.

본 논문에서는 KSC5601로 표현가능한 한자 4,888자중에서 동자이음자와 두음법칙관련자를 제외한 4,619자를 실험대상으로 하였다. 아울러 기존의 한자인식에 관한 연구에서 나타난 문제점을 해결하기 위해 계층적 구조 인식기 및 확률적 신경망 모델을 구성하였고, 엔트로피 이론을 적용하여 오인식률이 높은 혼동 문자쌍에 대하여 재분류를 수행하였다. 제2장에서는 기존 문자인식기의 문제점과 그 개선에 대하여, 제3장에서는 제안된 계층적 구조를 가지는 다단계 한자 인식기의 구조 및 그 구성방법을, 제4장에서는 실험 및 고찰, 제5장에서는 결론을 기술하였다.

2. 기존 문자인식기의 문제점과 그 개선

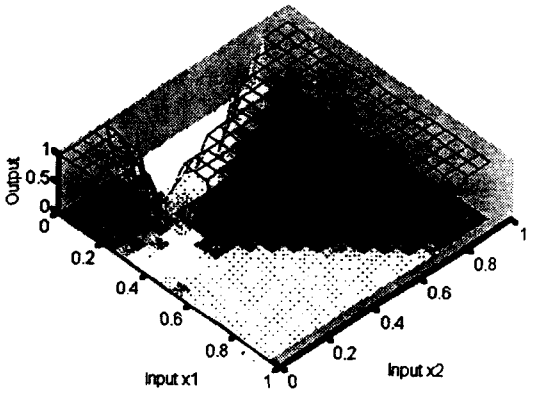
2.1 기존 다단계 문자인식기의 장점과 단점

A. Iwata의 연구에서 MLP(Multi-Layer Perceptron)를 오류역전파(Back-Propagation)알고리즘을 이용하여 많은 클래스의 패턴을 학습 시킬때, 학습의 수렴이 거의 불가능하다는 점을 지적 하고, 모듈화된 네트워크로 인식기를 구성하는 방법을 제안하여 CombNET을 구성하였다[3]. HoneyCombNET에 관한 연구에서는 입력패턴의 변동에 의한 VQ 오류를 줄이기 위하여, VQ 네트워크의 참조벡터를 복수로 할당하여 CombNET의 단점을 보완하였다[7]. HoneyCombNET Ⅱ의 연구에서는 대분류층 추가와 군집당 참조벡터의 조정으로 인식시간의 감소를 얻을 수 있었다[8]. 이러한 다단계 인식기는 학습 및 인식시간의 단축이라는 공통된 장점을 가지지만, 다음과 같은 문제점을 가지고 있다. MLP는 학습 데이타가 충분하고 실세계의 입력 데이타를 충분히 표현할 수 있는 경우 입력 데이타에 잡음이 첨가되어도 일반화된 결과를 낼 수 있는 능력을 가진다. 하지만, 모듈화된 MLP로 구성된 인식기는 이러한 조건을 만족 시키지 못한다. 하나의 MLP는 전체의 학습 데이타에 대하여 학습을 하지 않고, 하나의 군집에 속하는 학습 데이타만을 학습한다. 하나의 MLP가 학습되지 않은 군집에 속하는 입력을 받았을 경우, 이 MLP의 결과는 예측하기 힘들다. 따라서, VQ 네트워크의 출력결과가 산술적 또는 논리적으로 최종결과에 반영되어야 하며, 이로써 VQ network의 오류 역시 반영된다. 이러한 문제점의 예로써 4개의 XOR 문제를 (그림 2.1)에 나타내었다.



(그림 2.1) 4개의 XOR 문제
(Fig. 2.1) 4-XOR problem.

4개의 XOR 문제를 4개의 모듈화된 MLP에 의해 분류하고자 할때, 하나의 MLP는 두개의 클래스에 대하여 학습한다. 클래스 1과 2에 대해 학습한 MLP의 출력값은 (그림 2.2)와 같다. MLP의 은닉노드는 6개로 구성 되었으므로 6개의 결정선으로 조합된 영역으로 볼 수 있으며, 시그모이드 함수에 의해 결정선과의 거리에 상관없이 1의 높은 값을 보인다.

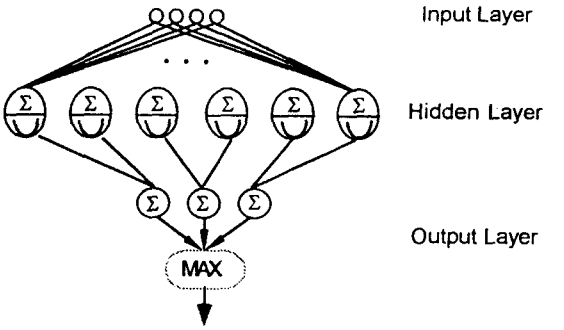


(그림 2.2) 모듈화된 MLP의 출력
(Fig. 2.2) Output of modular MLP.

2.2 확률적 VQ 네트워크 모델의 적용

D.F. Spectch는 수학적으로 정립이 잘 되어있는 Bayes 결정 전략(Bayes decision strategy)과 확률밀도 함수의 비모수적 측정(nonparametric estimator)방법인 Parzein-Window을 이용하여 Probabilistic Neural Network(PNN)을 제안하였다[9][10][11]. PNN은 입력층, 패턴층, 합산층과 출력층으로 이루어진 4층 구조이며, 가장 큰 장점은 별도의 학습과정이 필요 없다는 점이다. 하지만, PNN의 패턴층은 전체 학습 데이타로 이루어지므로, 특별한 하드웨어 없이는 구현시 문제가 많다. P. Burrascano의 연구에서는 이러한 단점을 지적하고, 패턴층(pattern layer)과 입력층(input layer)간의 가중치(weight)를 LVQ(Learning Vector Quantization) 학습 알고리즘을 이용하여 얻어진 최적의 참조벡터로 구성하였으며, 평활화 계수(smoothing parameter)에 따라 PNN과 동일한 성능을 보였다[12]. 균일한 분포를 가지는 데이타에 대해서는 LVQ보다 낮은 성능을 보였으며, 혼합밀도(mixture density)를 가지는 데이타에 대해서는 LVQ보다 높은 성능을 보였다.

이러한 연구들을 바탕으로 기존 다단계 인식기의 MLP를 대체하기 위하여 PNN을 간단화 시킨 확률적 VQ(Probabilistic Vector Quantization)네트워크를 고안하여 이용하였으며, 그 구조는 (그림 2.3)과 같다. 3층을 이루는 각층을 입력층, 은닉층, 출력층이라 부른다.



(그림 2.3) 확률적 VQ 네트워크의 구성
(Fig. 2.3) Structure of probabilistic VQ network.

i번째 은닉노드의 출력값은 입력벡터와 참조벡터의 내적, 즉 거리척도 $\cos\theta$에 지수 활성함수를 적용한 값이며, PNN과 마찬가지로 Parzen-Window방법을 나타낸다. σ는 평활화 계수이다.

$$y^h_i = \exp\left[\frac{(\sum_{k=1}^{N_i} x_k w^h_{ik}) - 1}{\sigma^2}\right] = \exp\left(\frac{(X \cdot W^h_i) - 1}{\sigma^2}\right) \quad (2.1)$$

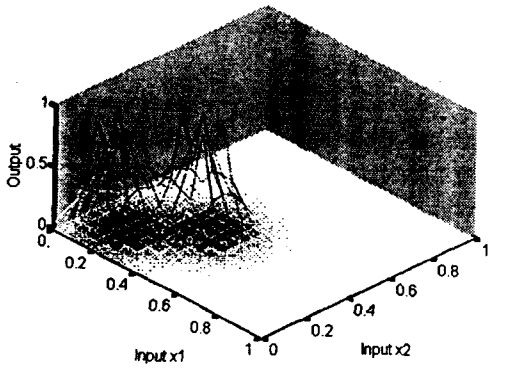
출력층의 노드는 분류 클래스를 나타내며, j번째 출력노드의 출력값은 노드 j와 연결을 가지는 은닉노드의 출력값과 가중치와의 곱의 합이다. N_{cj}는 참조벡터 수이고, h_j는 클래스 j에 대한 사전 확률, 즉 클래스 j에 해당하는 학습데이타의 수이다.

$$y^o_j = \sum_{k=1}^{N_{cj}} y^h_k w^o_k$$
$$w^o_k = \frac{h_j}{N_{cj}} \quad (2.2)$$

은닉층과 입력층간의 가중치는 k-means 알고리즘으로 각 클래스에 속하는 전체 학습데이타를 군집화

하여 얻은 k개의 평균벡터로 구성한다. k-means 수행시 네트워크의 구조에 맞도록, 거리의 척도로써 길이 정규화된 벡터간의 내적, $\cos\theta$를 이용한다. 은닉층과 입력층간의 가중치는 LVQ 학습알고리즘으로 학습이 가능하다. 출력층과 은닉층간은 부분연결을 가지며, 가중치는 각각의 출력층에 연결된 은닉노드의 수와 각각의 클래스에 속하는 학습 데이타수에 의해 얻어진다. 이 가중치는 PNN의 합산층과 출력층간의 가중치와 동일한 역할을 한다. 한편, PNN의 손실함수의 값은 주관적인 선택이 있어야 하며[9], 본 실험에서는 제외하였다. 출력값은 Parzen-Window방법에 의하여 측정된 확률값에 비례하며, 입력패턴은 최고 출력값을 가지는 출력노드의 클래스로 할당되며, PNN에서 이용된 하드 리미터(hard limiter)는 적용되지 않는다.

(그림 2.4)는 클래스 1과 2에 대하여 $\sigma^2 = 0.05$, 8개의 은닉노드로 구성하였을때 4개의 XOR문제에 대한 확률적 VQ 네트워크(PVQ)의 출력을 보인다. 결과에서 볼 수 있듯이 다른 입력에 대하여 0의 출력을 보인다.



(그림 2.4) 모듈화된 PVQ의 출력
(Fig. 2.4) Output of modular PVQ.

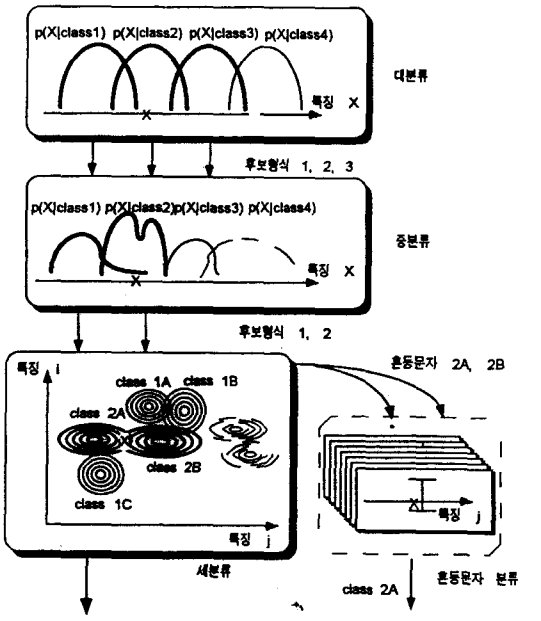
최대 출력값과 각 클래스 출력값의 비를 이용하여 확률분포의 겹침정도를 알 수 있으며, 이 정보를 이용하여 복수개의 결과를 선택할 수 있다. 한편, LVQ 학습 알고리즘은 1순위 정답률만을 높이기 위한 것이며 N순위까지의 정답률은 동일하기 때문에 복수의 후보결과를 선택하는 경우, 은닉층 참조벡터에 대하여 적용하지 않는다.

3. 제안된 계층적 인쇄체 한자 인식기

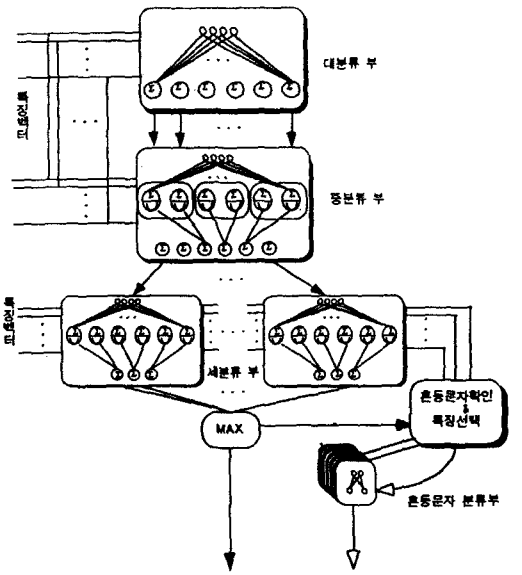
3.1 계층적 인식기의 필요성 및 전체적인 구성

한글과 한자 인식을 위한 인식기 구성시 인식 대상 수가 많다는 점과 유사문자가 많다는 점을 고려하여야 한다. 유사문자간의 혼동으로 발생하는 오분류를 줄이기 위하여 특징벡터의 차원수를 증가시킨다면 인식기의 일반화 능력이 감소될 것이며, 인식기의 복잡도는 증가할 것이다.

이러한 문제점들을 고려하여 전체 4단의 인식기를 구성하였으며, 상위계층에서 하위계층으로 진행됨에 따라 세밀한 분류가 이루어진다. 각단은 신경망 모델을 이용하였으며, 전반부의 3단은 형식분류와 문자인식, 후반부의 1단은 유사문자에 의한 혼동문자쌍의 재분류를 행한다. 각단을 대분류부, 중분류부, 세분류부, 혼동문자 분류부라 한다. 대분류부는 SOFM(Self-Organization Feature Map) 네트워크, 중분류부와 세분류부는 PVQ 네트워크로 구성하였으며, 혼동문자 분류부는 유클리디안 거리에 의해 분류를 행한다. 인식에 이용된 특징벡터는 계층적 구조에 적합하도록 추출하였다. 대분류와 중분류에 이용되는 특징으로 획 구조를 표현하는 64차원의 투영특징을 이용하였으며,



(그림 3.1) 전체 인식기의 개념도
(Fig. 3.1) Conceptual diagram of the overall recognizer.



(그림 3.2) 전체 인식기의 구성도
(Fig. 3.2) Structural diagram of the overall recognizer.

세분류에서는 128차원의 국소적 투영특징을 이용 하였다. 혼동문자 분류부에서는 정보이론을 바탕으로 선택된 16차원의 특징으로 혼동문자를 분류 하였다.

3.2 각단의 구성

3.2.1 대분류부

한자는 한글과 같이 곡선으로 구성되는 획이 없으며, 문자 구조의 대부분은 수직·수평 획에 의해 표현된다. KSC5601로 표현되는 한자중에는 매우 복잡한 구조를 가지는 한자가 많이 포함되어 있고, 이런 글자들은 인쇄상태와 입력상태에 따라 다르지만, 획간의 접촉이 많이 발생한다. 따라서, 흑화소의 방향성을 바탕으로 하는 특징 벡터의 경우, 획간의 접촉에 의해 특징값이 심하게 변화할 것을 예측할 수 있다.

본 실험에서는 인쇄체 한자인식을 위한 많은 특징중에서 특징추출 시간, 특징의 차원수, 획간의 접촉 등을 고려하여 투영특징을 이용하였다. 문자 영상 정규화 과정없이 추출된 특징의 차원수와 특징벡터의 길이를 정규화 하였으며, 차원수 정규화시 획의 이동에 의한 오류를 줄이기 위하여 문자의 크기에 비례하는 창을 이용하였다.

대분류부에서 이용하는 특징은 가로 32개, 세로 32개로 이루어진 64차원의 투영특징벡터이다. 식 3.1~3.4는 투영특징추출 및 창함수에 의한 차원수에 대한 정규화 과정을 보이며, p(x, y)는 x, y 좌표에서의 화소값을 나타낸다. (그림 3.3)은 창함수에 의해 추출된 특징을 보인다.

$$\text{수직축의 투영값}: f_y(x) = \sum_{i=1}^{M} p(x, i), \quad x = 1, \ldots, N \quad (3.1)$$

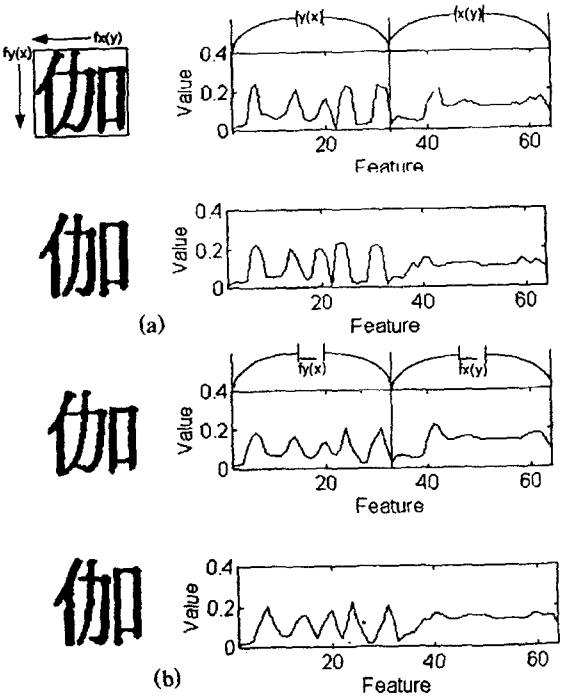
$$\overline{f}_y(x) = \sum_{i=-W_x}^{W_x} f_y(k-i), \quad x = 1, \ldots, 32,$$

$$k = x * \frac{N}{32}, \quad W_x = \lfloor \frac{N}{32} \rfloor \quad (3.2)$$

$$\text{수평축의 투영값}: f_x(y) = \sum_{i=1}^{N} p(i, y), \quad y = 1, \ldots, M \quad (3.3)$$

$$\overline{f}_x(y) = \sum_{i=-W_y}^{W_y} f_x(k-i), \quad x = 1, \ldots, 32,$$

$$k = y * \frac{M}{32}, \quad W_y = \lfloor \frac{M}{32} \rfloor \quad (3.4)$$



(그림 3.3) 추출된 특징벡터의 예 (a)창함수 적용전, (b)창함수 적용후
(Fig. 3.3) Examples of extracted feature vector. (a) before windowing, (b) after windowing

추출된 특징은 WHT에 의하여 직교변환 되며[13][14], FWHT(Fast Walsh-Hadamard Transform)에 적합하도록 특징의 수는 2^n이여야 한다. 가로, 세로방향 특징수 32는 획간의 접촉현상을 막을 수 있는 최소한의 갯수로써 뿐만 아니라, 효율적인 FWHT을 위하여 결정 되었다.

대분류부의 역할은 각 문자의 형식할당과 중분류부의 계산량 감소를 위한 1차 형식분류이며, 거리척도 $\cos\theta$를 이용한 SOFM 알고리즘을 이용 하였다. 하나의 문자가 복수의 형식에 속하는 것을 막기 위하여, 학습데이타의 각 문자별 특징벡터를 평균낸 평균 특징벡터를 이용하여 학습하였다. 출력노드수 변화에 따른 각 순위별 누적 분류율은 〈표 3.1〉과 같다. 순위를 백분율로 나타낸 이유는 중분류부 및 세분류부의 계산량을 비교하기 위해서이다. 본 실험에서는 256개의 형식을 이용 하였다.

〈표 3.1〉 군집수에 따른 분류율(학습데이타)
〈Table 3.1〉 Classification rate by the number of cluster (for training data).

순위 / 군집수	10% 순위	20% 순위	1순위	10순위
36	96.89% (4순위)	98.88% (7순위)	75.97%	99.42%
64	97.75% (6순위)	99.34% (13순위)	74.14%	94.02%
100	98.57% (10순위)	99.42% (20순위)	74.02%	98.57%
256	99.14% (26순위)	99.56% (51순위)	75.76%	97.71%
400	99.22% (40순위)	99.57% (80순위)	77.03%	97.39%

3.2.2 중분류부

중분류부는 대분류부에서 선택된 형식에 대하여 재평가하여 후보형식의 수를 감소시키는 2차 형식분류를 수행하며, 대분류부에서의 특징을 이용한다. 대분류부는 미리 정해진 수만큼의 후보군집을 선택하는 반면, 중분류부는 네트워크의 출력값에 따라, 즉 군집분포의 겹침정도에 따라 가변적으로 결정한다.

〈표 3.2〉는 $\sigma=0.06$일때 형식분류의 수와 형식당 참조벡터수 N_c에 따른 분류율을 보인다. 총 참조벡터수가 비슷한 경우, 군집수가 클수록 더 나은 결과를 보였다. 총 참조벡터수는 중분류기의 계산량과 비례하며, 문자수/99.5%는 학습데이타에 대한 분류율 5%를 유지할 경우 세분류기의 계산량과 비례한다.

〈표 3.2〉 군집수와 참조벡터수에 따른 분류율(학습데이타)
〈Table 3.2〉 Classification rate by the number of cluster and reference vetor (for training vector).

군집수 / N_c / 분류율	64		100			256		
	12	18	9	12	18	3	6	9
99.5%	14.1% 9순위	14.1% 9순위	13% 13순위	10% 10순위	8% 8순위	10.2% 26순위	5.1% 13순위	3.9% 10순위
100%	71.9% 46순위	78.1% 50순위	58% 58순위	73% 73순위	63% 63순위	96.1% 246순위	92.6% 237순위	63.7% 163순위
총 참조벡터수	768	1152	900	1200	1800	768	1536	2304
문자수/군집	72.6		46.5			18.2		
문자수/99.5%	649.6	649.6	600.5	461.9	369.5	469.1	234.6	162.4

최대 출력값의 β배 이상의 출력을 내는 군집을 선택함으로써, 입력패턴에 대해 분포의 겹침을 가지는 군집을 선택할 수 있다. 〈표 3.3〉은 군집수가 256개일 때, σ와 β에 따른 형식 분류율과 후보군집수를 나타낸다.

〈표 3.3〉 σ와 β에 따른 분류율(학습데이타)
〈Table 3.3〉 Classification by the value of σ and β (for training data).

σ / β	0.001	0.01
0.06 (평균/최대후보수)	99.64% (7.38/71)	99.21% (3.17/38)
0.07 (평균/최대후보수)	99.79% (15.86/103)	99.58% (6.03/60)
0.08 (평균/최대후보수)	99.91% (30.78/143)	99.78% (12.55/84)

3.2.3 세분류부

대분류부에서 구성된 군집은 유사도에 의해 결정되었기 때문에 상위단과 동일한 특징으로 군집안의 문자들을 분류하기 곤란하다. 세분류부에서는 세부적인 형태를 나타낼 수 있고, 대·중분류부의 특징과 계층적인 관계를 가지는 특징을 이용한다.

특징추출은 문자영상을 투영하는 영역을 2개의 영역으로 분할하여 흑화소의 갯수를 누적하는 국소적 투영과정을 통하여 이루어지며, 역시 정규화 과정과 WHT변환을 행한다. 식 3.5~3.8는 특징추출에 관한 식이며, (그림 3.4)는 추출된 특징을 보인다. 식 3.9는 한 번의 추출과정으로 얻을 수 있는 대분류부와 세분

류부 특징간의 관계를 나타낸다.

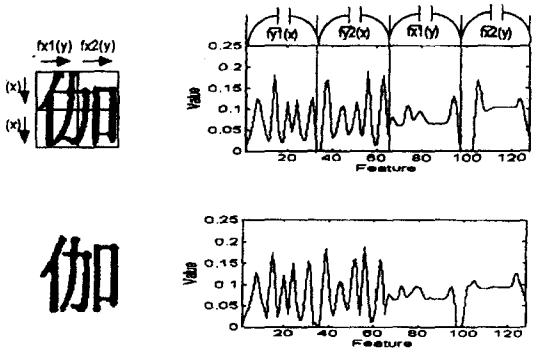
$$\text{수직축의 투영값 } 1: f_{y1}(x)=\sum_{i=1}^{M/2} p(x,i),\quad x=1,\ldots,N \tag{3.5}$$

$$\text{수직축의 투영값 } 2: f_{y2}(x)=\sum_{i=M/2+1}^{M} p(x,i),\quad x=1,\ldots,N \tag{3.6}$$

$$\text{수평축의 투영값 } 1: f_{x1}(y)=\sum_{i=1}^{N/2} p(i,y),\quad y=1,\ldots,M \tag{3.7}$$

$$\text{수평축의 투영값 } 2: f_{x2}(y)=\sum_{i=N/2+1}^{N} p(i,y),\quad y=1,\ldots,M \tag{3.8}$$

$$f_y(x)=f_{y1}(x)+f_{y2}(x)$$
$$f_x(y)=f_{x1}(y)+f_{x2}(y) \tag{3.9}$$



(그림 3.4) 창함수가 적용된 국소적 투영특징 $\{f_{y1}, f_{y2}, f_{x1}, f_{x2}\}$
(Fig. 3.4) Local projection feature after windowing $\{f_{y1}, f_{y2}, f_{x1}, f_{x2}\}$

〈표 3.4〉는 전역적 투영특징과 국소적 투영특징을 이용한 세분류 결과를 보인다. 각 문자에 대한 평균 벡터와 입력벡터의 거리값을 1-NN방법으로 인식한 결과이다.

〈표 3.4〉 특징에 따른 인식결과 (학습데이타)
〈Table 3.4〉 Recognition rate by the type of feature vector (for training data).

데이터 / 특징	HM20	HM21	HM30	HM31	SM20	SM21	SM30	SM31	평균
GPR	96.1%	96.4%	96.3%	96.3%	94.4%	94.7%	94.8%	94.4%	95.43%
LPR	97.5%	97.9%	97.8%	97.7%	98.0%	98.2%	98.2%	98.3%	97.95%

세분류는 중분류와 마찬가지로 PVQ를 이용하여 분류하며 하나의 출력노드는 하나의 문자를 나타낸다. 은닉층의 구성은 k-means 알고리즘을 이용하지 않고, 각 서체별로 평균벡터를 구하여 구성하였다. 하나의 클래스에 속하는 학습 데이타의 수가 적고, 학습서체 추가시 인식기의 확장이 쉽기 때문이다. 본 실험에서는 2종류의 명조체로 실험하였으며, 따라서, 세분류기의 총 은닉노드수는 4619*2=9238개 이며, 군집당 평균 약 36개 이다. 클래스당 은닉노드수에 따른 실험결과($\sigma=0.09$일때)는 〈표 3.5〉와 같다.

〈표 3.5〉 참조벡터수에 따른 인식결과 (학습데이타)
〈Table 3.5〉 Recognition rate by the number of reference vector (for training data).

데이터 / 참조벡터수	HM20	HM21	HM30	HM31	SM20	SM21	SM30	SM31	평균
1	97.5%	97.9%	97.8%	97.7%	98.0%	98.2%	98.2%	98.3%	97.95%
2	99.7%	99.8%	99.7%	99.6%	99.3%	99.4%	99.4%	99.5%	99.55%

3.2.4 혼동문자 분류부

한자는 몇개의 획 또는 점으로 구별되는 유사문자의 쌍이 많다. 오인식의 대부분 역시 이러한 문자들 간의 혼동에 의해 발생된다. 따라서, 분별력이 강하고 안정적인 특징을 선택하여 재분류하는 과정이 필요하다. 이런 특징선택의 기준으로써 엔트로피(entropy) 이론을 이용하였다[13].

어떤 클래스의 기본적인 구조을 나타내는 특징은 거의 일정하므로 조건부 엔트로피가 감소된다.

또한, 두 클래스를 분리할 수 있는 특징은 큰 상호정보량을 가진다.

특징의 선택은 먼저, 상호정보량 I(x;y)이 최대값 H(y)를 가지는 특징을 선택한다. 두 클래스는 똑같은 확률로 나타나므로 $H(y)=\log_2 2=1$이다. 이들 특징들을 조건부 엔트로피가 작은 순으로 정렬한다. 본 실험에서는 입력신호를 128등분 하였으며, 128개의 특징 요소중 평균 약 50개정도의 특징요소가 상호정보량의 최대값인 1을 보였다. WHT변환으로 특징간의 상관도를 없앰으로써 특징선택과정에 잇점이 있으나, 직류성분에 의해 분산이 매우 적기 때문에 WHT변환 이전의 특징을 이용하였다.

혼동문자 리스트의 작성은 학습 데이타와 실험 데이타에 대한 세분류부의 독립적인 인식결과로 이루

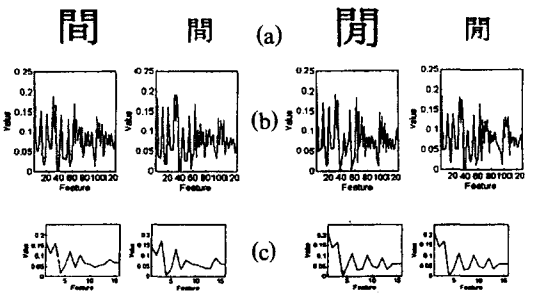
어진다. 오류율이 높은 문자들을 선택한 뒤, 세분류기의 기준벡터와 입력 데이타간의 유사도를 측정한다. 클래스 i가 j로 오인식 되었을때, 클래스 i의 기준벡터와 클래스 j의 입력벡터간의 유사도를 구하여 상위 50위에 속하는 문자의 쌍을 혼동문자 리스트로 작성하였다. 50쌍의 문자는 전체문자의 2.2%에 해당된다.

혼동문자 분류기의 기준벡터 생성은 상위 인식부와 마찬가지로 학습 데이타에 의해 구성되었다. 혼동문자 분류기는 특징벡터의 길이정규화가 곤란하기 때문에 내적에 의한 유사도가 아닌 유클리디안 거리에 의한 유사도에 의해 두 문자를 분류한다.

〈표 3.6〉은 추출 특징수를 8, 16, 128로 하였을 경우의 결과이고, (그림 3.5)는 혼동문자에 대하여 추출된 특징을 보인다. 본 실험에서는 일반화 능력과 특징선택에 관한 충분조건을 고려하여 16개의 특징을 선택하였다.

〈표 3.6〉 혼동문자에 대한 분류결과
〈Table 3.6〉 Classification rate for confused characters.

특징수 / 데이터	128	16	8
학습데이터	99.13%	99.13%	92%
실험데이터	53%	77.5%	73%

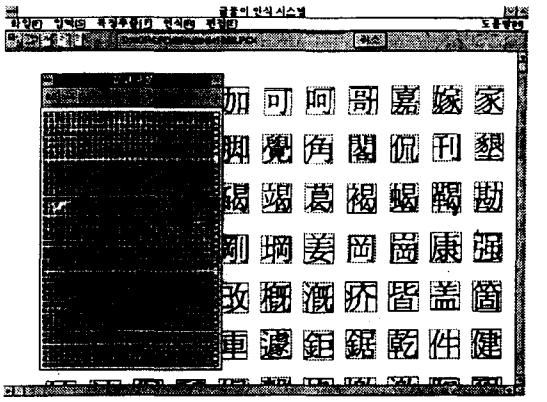


(그림 3.5) 혼동문자에 대하여 선택된 특징 (a)문자영상, (b)국소 투영특징, (c)선택된 특징
(Fig. 3.5) Selected feature about the confused characters. (a)images, (b)local projection features, (c) selected features

4. 실험 및 고찰

4.1 실험 환경

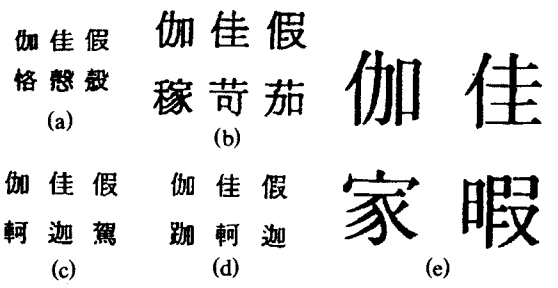
본 실험은 IBM 486PC-66MHz와 MS-WINDOWS 3.1 환경에서 구현되었다. (그림 4.1)에는 실험환경을 보였다. 문서영상의 출력은 HP LaserJet Ⅱp 프린터(300dpi), 입력은 HP ScanJet Plus 스캐너(300dpi)를 이용 하였다. 개발도구는 Borland C 3.1, gcc 2.5.3 컴파일러와 Pixel Translation®의 라이브러리를 이용하였다.



(그림 4.1) 실험환경
(Fig. 4.1) Experimentation tool.

학습 데이타와 실험 데이타의 구성은 〈표 4.1〉과 같다. 실험 데이타는 크기, 흑화소 잡영, 미지의 서체에 대하여 실험하기 위한 것이며, 모두 7 종류이다. 잡영의 첨가방법은 식 4.1과 같으며, M은 문자영상의 높이, N은 문자영상의 폭을 의미하며, α는 난수의 반영정도, 즉 0.1과 0.2를 나타낸다. 문서영상이 아닌 추출된 투영특징에 잡영을 첨가한 이유는 문자추출시 발생할 수 있는 오류를 배제하기 위해서이다.

(그림 4.2) 학습 데이타 영상의 예 : (a)와 (b)20포인트, (c)와 (d)30포인트
(Fig. 4.2) Examples of image of training data. (a) and (b) 20 point, (c) and (d) 30 point



(그림 4.3) 실험 데이타 영상의 예 : (a), (c)와 (d) 12포인트, (b) 20포인트, (e) 40포인트

(Fig. 4.3) Examples of image of test data. (a), (c) and (d) 12 point, (b) 20 point, (e) 40 point

$$f_y^{noise}(x) = \min(M, f_y(x) + \alpha * random(0, M)), \quad x = 1, ..., N$$
$$f_x^{noise}(y) = \min(M, f_x(y) + \alpha * random(0, N)), \quad y = 1, ..., M \tag{4.1}$$

본 실험에서 이용된 학습 데이타와 실험 데이타의 예는 (그림 4.2), (그림 4.3)와 같다.

4.2 실 험

4.2.1 전체 인식기의 조정

인식률과 인식속도간의 균형을 맞추어진 전체 인식기를 구성하기 위해서는 형식수와 중분류부의 참조벡터수, 대분류부의 후보결과수, 중분류부의 후보결과수를 결정하는 β에 관하여 산술·회귀적 분석이 추가되야 하지만, 본 논문에서는 독립적인 실험만을 통하여 분석 하였다.

대분류단은 〈표 3.1〉과 같이 51순위까지 99.56%의 정답률을 가진다. 대분류단의 후보형식수를 60과 100개로 하여 학습 데이타에 대하여 실험 하였을 경우, 각각의 평균 대분류 정답율은 99.61%, 99.77%을 가진다. 분류율은 99.77:99.61, 약 0.16% 증가하지만, 중분류단의 계산량은 100:60, 약 66.67% 증가한다.

중분류단의 분류율과 후보형식수는 〈표 3.3〉과 같이 σ와 β에 의해 결정되며, 세분류단의 계산량을 고려하여 $\beta = 0.01$로 고정 하였다. 〈표 4.2〉은 실험 데이타 중 미지의 서체인 매킨토시 서체에 대하여 실험한 결과이다.

〈표 3.3〉에 나타낸 학습 데이타에 대한 결과와 비교해보면, σ 증가에 따른 분류율의 증가율이 매우 크다.

표4.1 학습 데이타와 실험 데이터의 구성

표 4.1

	HM20	HM21	HM30	HM31	SM20	SM21	SM30	SM31
학습 서체	한글과 컴퓨터㈜				서울시스템㈜			
	신명조체				중명조체			
	20pt		30pt		20pt		30pt	
	원본	사본	원본	사본	원본	사본	원본	사본
실험 서체	HM12	HM22	HM201	HM202	SM40	MM112	MM212	
	한글과 컴퓨터㈜				서울시스템	매킨토시 컴퓨터		
	신명조체				중명조체	신명조체	신산명조체	
	12pt	20pt	20pt	20pt	40pt	12pt	12pt	
	원본	HM21 의 사본	10%잡음	20%잡음	원본	원본	원본	

〈표 4.2〉 미지 서체에 대한 중분류율

〈Table 4.2〉 Second classification rate for the unknown fonts.

실험데이터 / σ	MM112	MM212
0.06 (평균/최대후보수)	91.72% (6.11/44)	91.70% (6.08/46)
0.07 (평균/최대후보수)	95.18% (10.78/70)	95.07% (10.68/70)
0.08 (평균/최대후보수)	96.82% (18.62/104)	96.68% (18.45/112)

따라서, 일반화 능력을 고려하여 $\sigma = 0.08$로 결정한다. 세분류단은 σ의 변화에 민감하지 않으며, $\sigma = 0.09$로 결정하였다.

인위적인 잡영을 첨가한 2개의 데이타에 의한 오인식은 일반적인 오류가 아니기 때문에 혼동 리스트의 작성시 제외 시켰다. 혼동 리스트를 구성하는 50쌍의 문자는 〈표 4.3〉와 같다.

〈표 4.3〉 혼동문자 리스트

〈Table 4.3〉 List of the pair of confused character.

植↔殖	倣↔倣	棄↔奘	間↔間	瑾↔瑾	琦↔埼	郷↔郷	潰↔潰	種↔植	捞↔探
雨↔雨	賢↔賢	艘↔艘	杷↔把	細↔細	櫻↔櫻	枞↔枞	槁↔稿	瑯↔郷	椎↔推
徹↔徹	妹↔姝	梁↔梁	岾↔岾	迷↔迷	攪↔攪	餃↔餃	須↔項	琦↔熔	機↔機
徹↔徹	僧↔僧	凌↔漢	飯↔級	闇↔闇	檜↔檜	戊↔戊	作↔佯	勳↔勳	傅↔搏
橋↔僑	鈞↔釣	妹↔錶	英↔英	爆↔燦	菅↔管	瑯↔熔	睡↔唾	痴↔痴	斤↔斥

4.2.2 실험 결과

〈표 4.4〉은 대분류단의 후보수를 60으로 하였을 때, 학습 데이타에 대한 인식 결과를 보인다.

〈표 4.5〉는 대분류단의 후보수를 60으로 하였을 때, 실험 데이타에 대한 인식 결과를 보인다.

〈표 4.6〉는 대분류부 후보수에 따른 결과이며, 후보수 256은 대분류와 중분류가 없을 때를 의미한다.

〈표 4.4〉 학습 데이타에 대한 인식결과 (후보수 =60)
〈Table 4.4〉 Result of recognition for training data (number of candidate =60).

	HM20	HM21	HM30	HM31	SM20	SM21	SM30	SM31
대분류율	99.68%	99.68%	99.70%	99.72%	99.48%	99.51%	99.57%	99.53%
중분류율	99.55%	99.59%	99.59%	99.59%	99.29%	99.38%	99.36%	99.42%
세분류율	99.44%	99.53%	99.55%	99.48%	99.08%	99.14%	99.16%	99.29%
혼동문자/분류문자	98/96	97/96	97/95	98/96	97/97	97/97	97/96	97/96
평균후보(중분류)	11.2	12.0	10.0	10.4	12.1	17.1	12.1	14.7
최대후보(중분류)	58	59	59	59	60	60	58	60
인식시간(초)	913	947	862	881	946	1165	960	1096

〈표 4.5〉 실험 데이타에 대한 인식결과 (후보수 =60)
〈Table 4.5〉 Result of recognition for test data (number of candidate =60).

	HM12	HM22	SM40	HM201	HM202	MM112	MM212
대분류율	99.66%	99.74%	99.59%	99.63%	99.48%	98.52%	98.36%
중분류율	99.55%	99.69%	99.37%	99.56%	99.31%	96.45%	96.23%
세분류율	99.22%	99.27%	99.08%	99.29%	84.42%	84.33%	84.21%
혼동문자/분류문자	98/97	97/95	98/96	98/96	100/68	100/72	99/71
평균후보(중분류)	15.4	17.4	17.5	17.7	27.4	17.6	17.0
최대후보(중분류)	60	60	60	59	60	60	60
인식시간(초)	1089	1182	1187	1190	1586	1169	1139

〈표 4.6〉 실험 결과의 비교 (학습데이타)
〈Table 4.6〉 Comparsion of each result (for training data).

후보수(대분류부) / 결과	256	100	60
인식률(%/비율)	99.55 / 100%	99.38 / 99.83%	99.33 / 99.78%
인식시간(초/비율)	10560 / 100%	1184 / 11.21%	996 / 9.43%

4.3 고 찰

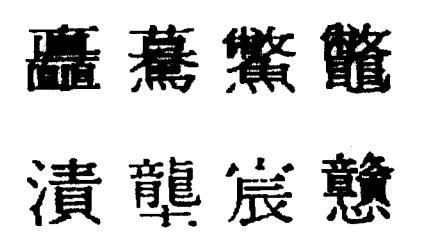
〈표 4.4〉에 나타난 바와 같이, 학습 데이타에 대한 실험에서 분류오류는 모든 단에서 고르게 발생하였다. 대부분의 형식분류오류는 인식속도를 감안하여 후보수를 결정함으로써 발생한다. 후보형식수를 증가시킴으로써 전체 인식률은 증가하지만, 전체 인식률의 증가율에 비하면 전체 인식시간의 증가율은 매우 크다.

계층적 구조를 가지는 다단 인식기에 의한 인식시간 단축효과를 입력벡터와 참조벡터간의 곱연산 횟수, 즉 계산량으로써 나타내면 다음과 같다. 전체 세분류부는 4619개의 클래스와 2개의 클래스당 참조벡터를 가지며, 참조벡터와 입력벡터는 128차원으로 구성된다. 따라서 대분류와 중분류 과정이 없을시, 하나의 입력벡터를 분류하기 위하여 요구되는 계산량은 $4619*2*128=1{,}182{,}464$이다. 혼동문자 분류부를 제외한 전체 인식기는 64개의 대분류부 참조벡터, 각 형식당 9개(총 576개)의 중분류부 참조벡터, 각 분류클래스당 2개(총 9238개)의 참조벡터로 구성된다. 총 계산량은 대분류부, 중분류부, 세분류부 계산량의 합이며, 대분류부의 후보수 60, 중분류부의 후보수 12일 경우, $256*64+256*9*64*(60/256)+4619*2*128*(12/256)\fallingdotseq 106{,}465$이다. 이것은 세분류부의 독립적 계산량의 약 9%이다. 실제 실험에서도 역시 세분류단을 독립적으로 실행하였을 경우의 인식시간, 2시간 56분,의 9.4%에 해당하는 평균 996.4초/4619자, 4~5자/초의 인식시간을 보였다.

명조계열의 서체라도 서체 종류에 따라 문자 구성요소의 위치 및 획의 모양에 약간의 차이가 있다. 실험 결과에서 알수 있듯이 잡영의 발생, 종이질과 입력상태의 변화 등으로 인한 오분류보다도 이러한 서체 변화에 가장 민감하다. 학습 데이타에 대한 전체 인식률은 혼동문자 분류부에 의해 평균 0.016%, 최대 0.04% 감소하였으나, 미지의 서체에 대하여 평균 1% 증가하였다.

잡영과 크기변화에 대한 실험에서 양호한 결과를 얻었으며, 특히 20%의 잡영을 추가한 데이타에 대한 실험에서 99.31%의 중분류율을 얻었다. 평균 27.4개의 후보군집을 선택 하였지만, 안정적인 성능을 나타내는 결과이다. 문자 크기가 12포인트이고 300dpi 해상도로 입력을 받았을 경우, 문자 획간의 접촉 빈도가 높으나, 양호한 인식률을 얻었다. (그림 4.4)에는 12포인트의 실험 데이타중 획간의 접촉, 획의 손실 등에 의하여 변형된 문자영상의 예를 보인다.

동일한 서체로 구성된 데이타, 즉 8개의 학습 데이



(그림 4.4) 변형된 문자영상의 예
(Fig. 4.4) Examples of distorted image.

타와 3개의 실험 데이타에 대하여 99.30%의 인식률을 얻었다.

5. 결 론

본 논문에서는 인쇄체 한자인식을 위해 확률적 VQ 네트워크와 계층적 구조를 갖는 다단계 인식기를 제안하였다. 인식대상은 KSC5601체계로 표현 가능한 명조체 계열의 4,619자로, 많은 글자수와 혼동문자에 대처하기 위하여 4단으로 구성된 계층적 구조의 인식기를 구성하였고, 확률적 VQ 네트워크를 이용함으로써 기존의 다층 퍼셉트론을 이용한 다단계 인식기에서 나타난 문제점을 해결할 수 있었다. 또한 출력값의 상태에 따라 후보결과를 선택함으로써 안정적인 형식분류율을 얻었다. 64차원의 전역투영특징과 128차원의 국소투영특징을 이용함으로써 한자 획의 위치와 길이를 분류 단계에 알맞게 나타낼 수 있었으며, 문자 획간의 접촉에 의한 특징값의 변화를 줄일수 있었다. 또한 엔트로피 이론을 바탕으로 선택된 안정적이고 분별력 있는 16개의 특징을 선택함으로써 혼동문자에 대하여 재분류 하였다. 실험결과 학습 데이타에 대하여 평균 99.33%, 실험 데이타에 대하여도 비교적 높은 평균 92.83%의 인식률을 얻을 수 있었으며, 초당 4-5자의 인식속도를 나타내었다.

확률적 VQ 네트워크는 학습 데이타에 의한 확률분포를 나타내기 위한 모델로써 다중서체문자가 나타내는 혼합밀도 분포를 근사화할 수 있는 능력이 있으며, 반복 학습시 신경망 모델보다 구성시간이 매우 짧다는 장점을 가진다. 하지만 최고의 성능을 나타낼 수 있는 은닉노드의 수와 평활화 계수와 같은 인자를 얻기 위해서는 비교적 많은 실험이 있어야 한다는 단점을 가진다. 또한 보다 복잡한 분류문제에 대하여 확률적 VQ 네트워크가 우수한 성능을 보이기 위해서는 가중치들을 최적화할 수 있는 추가 학습과정 및 알고리즘이 요구된다.

참 고 문 헌

[1] Y. X. Gu, Q. R. Wang, C. Y. Suen, "Application of a Multilayer Decision Tree in Computer Recognition of Chinese Characters," IEEE Trans. on PAMI, Vol. 5, No. 1, 1983.

[2] Q. R. Wang, C. Y. Suen, "Analysis and Design of a Decision Tree Based On Entropy Reduction and Its Application to Large Character Set Recognition," IEEE Trans. on PAMI, Vol. 6, No. 4, 1984.

[3] A. Iwata, T. Tohma, H. Matsuo, N. Suzumura, "A Large Scale Neural Network-CombNET," 日本 電子情報通信學會論文誌, Vol. J73-D-II, No. 8, 1990.

[4] 오종옥, "신경망을 이용한 인쇄체 한자의 인식에 관한 연구", 광운대학교 석사학위 논문, 1992.

[5] S. W. Lee, J. S. Kim, "Multi-lingual, Multi-font and Multi-size Large-set Character Recognition using Self-Organizing Neural Network," ICDAR, Vol. 1, 1995.

[6] 한국표준연구소, "한자부호 표준시안 작성을 위한 연구", KSRI-87-37-IR, 1987.

[7] M. Arai, J. Wang, K. Okuda, J. Miyamichi, "Thousand of Hand-Written Kanji Recognition by HoneyCombNET," 日本 電子情報通信學會論文誌, Vol. J76-D-II, No. 11, 1993.

[8] M. Arai, K. Okuda, J. Miyamichi, "Thousand of Hand-Written Kanji Recognition by HoneyCombNETII," 日本 電子情報通信學會論文誌, Vol. J77-D-II, No. 9, 1994.

[9] D. F. Spetch, "Probabilistic Neural Networks," Neural Networks, Vol. 3, 1990.

[10] T. Masters, "Advanced Algorithm for Neural Networks," Wiley, 1995.

[11] P. D. Wasserman, "Advanced Methods in Neural Computing," Van Nostrand Reinhold, 1993.

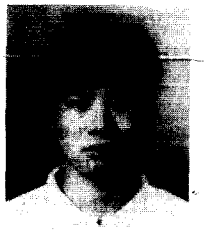
[12] P. Burrascano, "Learning Vector Quantization for the Probabilistic Neural Network," IEEE Trans. on Neural Network, Vol. 2, No. 4, 1991.

[13] K. G. Beauchamp, "Walsh Function and Their Applications," Academic Press, 1975.

[14] 김우태, 윤병식, 박인규, 진성일, "인쇄체 한글 문자인식을 위한 특징성능의 비교", 한국 정보과학회논문지, Vol. 20, No. 8, 1993.

[15] 송현경, 이영직, "계층적 패턴인식과정의 단계별

효율성 분석", 한국 정보과학회논문지, Vol. 20, No. 12, 1993.



이 장 훈

1994년 광운대학교 컴퓨터공학과 졸업(공학사)
1996년 광운대학교 대학원 컴퓨터공학과 졸업(공학석사)
1997년~현재 국방정보체계연구소 연구원
관심분야:패턴인식, 신경회로망, 화상처리

손 영 우

1981년 광운대학교 전자공학과 졸업(공학사)
1983년 광운대학교 대학원 전자공학과 졸업(공학석사)
1996년 광운대학교 대학원 전자계산기공학과 박사과정 수료
1991년~현재 산업기술정보원(KINITI) 책임연구원
관심분야:패턴인식, 신경회로망, Chaos이론, 문서인식

남 궁 재 찬

1970년 인하대학교 전기공학과 졸업(공학사)
1976년 인하대학교 대학원 전자공학과 졸업(공학석사)
1982년 인하대학교 대학원 전자공학과 졸업(공학박사)
1982년~1984년 일본 동북대학 객원교수
1979년~현재 광운대학교 컴퓨터공학과 교수
관심분야:패턴인식, 신경회로망, Chaos이론, 문서인식