論文 96-21-12-23

CombNET 신경망을 이용한 혼용 문서 인식 시스템의 구현

正會員 최 재 혁*, 손 영 우*, 남 궁 재 찬**

An Implementation of the Mixed Type Character Recognition System Using CombNET

Jae Hyuk Choi*, Young Woo Shon*, Jae Chan Namkung** *Regular Members*

요 약

문자인식에 대한 연구는 주로 한글인식에 대해서만 이루어져 왔는데, 대부분의 문서는 한글 뿐만 아니라 여러 종류의 문자가 포함되어 있다. 따라서, 본 논문에서는 다중 크기, 다중 활자체, 다자종 문자가 포함되어 있는 한글 문서를 인식할 수 있는 문자인식 시스템을 구현하였다. CombNET 구조를 갖는 신경회로망을 자종별로 구성하여, 문자인식시에 문자를 구별하지 않고 인식하는 방법을 제안하였다. CombNET 구조의 상단부를 차지하는 Kohonen의 SOFM 신경망을 이용하여 한글과 한자는 36개, 영숫자는 16개의 유형으로 분류하고 각 유형에 대해서 CombNET 구조의 하단부에 있는 BP 네트워크를 이용하여 문자인식을 수행하였다. 실험결과 학습 데이타에 대해서는 95.6%의 인식율을 나타내었고, 실제문서에 대해서도 92.6%의 인식율과 초당 10.3자의 인식속도를 보임으로써 제안된 인식 시스템의 유효성을 입증하였다.

ABSTRACT

The studies of document recognition have been focused mainly on Korean documents. But most of documents composed of Korean and other characters. So, in this paper, we propose the document recognition system that can recognize the multi-size, multi font and mixed type characters. We have utilized a large scale network model, "CombNET" which consists of a 4 layered network with a comb structure. And we propose recognition method that can recognize characters without discrimination of character type. The first layer constitutes a Kohonen's SOFM network which quantizes an input feature vector space into several sub-spaces and the following 2-4 layers

*광운대학교 대학원 컴퓨터공학과
**광운대학교 컴퓨터공학과 교수, 신기술연구소
論文番號:95419-1211
接受日字:1995年 12月 11日

constitutes BP network modules which classify input data in each sub-space into specified catagories. An experimental result demonstrated the usefulness of this approach with the recognition rates of 95.6% for the training data. For the mixed type character documents we obtained the recognition rates of 92.6% and recognition speed of 10.3 characters per second.

Ⅰ. 서 론

기술이 발전하고 정보화 시대에 접어들면서 인간은 정보의 홍수 속에 살게 되었다. 이러한 정보들은 다양한 매체를 이용하여 저장, 보관되고 있으므로 지금 같은 정보화 시대에 가장 중요한 것은 인간이 얼마나 빠르고 정확하게 정보를 획득, 추출 그리고 전송하느냐 등이 큰 과제로 등장하고 있다. 그러나 문서 자체는 보관과 검색상의 어려움 때문에, 정보의 데이타베이스화 작업이 필수적이다. 하지만 키보드를 이용한 기존의 입력 방법은 많은 시간과 인력을 필요로 하므로 보다 많은 정보를 자동 입력하기 위해서 문자인식과 음성인식에 관한 연구가 발전하였으며, 이 중에서 문자인식 기술은 우리 나라에서도 괄목할 만한 성과를 보여주고 있다.

현재까지 개발된 문서인식 시스템은 주로 다중 크기, 다중 활자체 한글 및 영어, 숫자 및 각종 특수 기호 등의 자동 인식에 관한 연구가 중심을 이루어 왔다. 그러나 우리나라가 아직 까지는 한자 문화권에 속하고 거의 모든 문서가 한글, 한자와 영문자가 혼용되어 사용되고 있으나 혼용 문서에 대한 연구는 미비한 실정이다. 이러한 이유는 그 동안의 연구가 대부분 한글의 구조적인 특성을 정보로 이용하여 한글을 자소 별로 인식하였기 때문에, 한글이 아닌 다른 문자에는 적용할 수가 없었기 때문이다. 따라서 한글 이외의 다른 문서가 혼용된 문서에서는 한글인지 아닌지를 구별하는 과정이 선행되어야 했고, 한글의 경우에는 6형식으로 문자를 분류한 후, 각 형식에 대한 구조적인 특징을 이용하여 한글 인식을 하였으며, 한자나 영숫자의 경우에는 그에 따른 형식분류 과정을 거친 후에 각 문자를 인식하는 방법을 사용하였다. 그러나 이러한 방법은 문자를 구별하는 단계나 형식분류 과정에서 오류가 발생하면 이를 보상할 수 없다는 단점이 있다. 그리고 대부분의 문서가 한글 및 한자, 영숫자가 혼용되어 사용되고 있으므로 고속의 문자 인식을 위해서는 문자 구별이라는 선행 단계를 거치지 않고 인식할 수 있는 방법이 필요하다.

김우성[1]은 한글 520자와 한자 900자를 문자 구별을 하지 않고 여러 유형으로 군집화한 후에 APC(Adaptive Pattern Classifier)를 이용하여 한글의 구조적인 정보를 고려하지 않고 한글과 한자를 동시에 인식하였다. 이 방법은 오류가 발생한 문자에 대해서 신경망의 노드를 증가시키므로써 추가 학습이 가능하지만, 많은 글자로 인한 오류 보상을 위해 새로운 뉴런이 추가되므로 인식기의 규모가 계속적으로 커져야하는 문제점을 갖고 있다. 김현종[2]은 이러한 단점을 지적하고, 인식 속도가 늦더라도 한글 전용 인식기와 한자 전용 인식기와의 접목을 가능 하게 하기 위해 한글과 한자의 분리를 위한 인식기를 제안하였다. 그러나 이 방법은 한글과 한자의 완전 분리를 위하여 부가적인 LVQ3 학습을 반드시 거쳐야 하고, 새로운 자종을 추가로 학습할 때에도 군집화 과정을 처음 부터 다시하여야 하는 문제점이 있다. 따라서 본 연구에서는 개선된 ComNET[3][4][5][6]을 자종별로 모듈화한 신경 회로망을 제안하여, 문서 인식시에 입력 문자가 한글, 한자, 영숫자인지를 구별하지 않고 인식하는 방법을 제안한다.

Ⅱ. CombNET의 구조

3층 이상의 계층 구조를 갖는 신경망을 이용한 R_m 공간에서 R_n 공간로의 사상(mapping)이 이루어져 왔지만, 대용량의 신경망을 구현하는 것이 아직까지는 어렵다. 이러한 대형의 신경망의 학습 과정은 종종 국소 최소점(local minimum state)에 빠지거나 학습 시간이 오래 걸리는 문제가 있다. 이러한 문제를 해결하기 위해서는 동작하기 쉽게 가능한 한 소형으로 구성한 몇 개의 모듈화된 신경망의 조합으로 구성하는 것이 효과적이다. 따라서, 그림 1과 같이 몇 개의 신경망으로 조합된 구조를 갖는 4층 신경망으로 구성된

CombNET이 제안되었다.

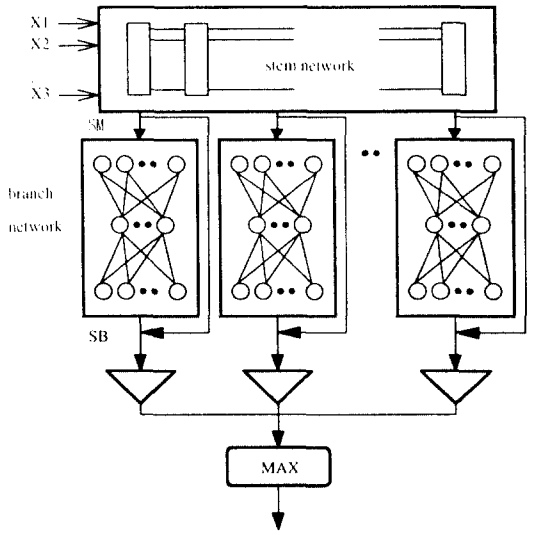


그림 1. CombNET의 구조
Fig. 1 Structure of CombNET

CombNET의 구조는 벡터 양자화 역할을 해주는 1층의 stem network과 그것에 연결된 3층의 branch network들로 구성되어 있다. CombNET에서는 stem network의 학습을 위해서 kohonen의 self organizing feature map(SOFM)을[7] 이용하고 branch network에서는 BP network[8][9]을 이용한다. branch network의 수는 stem network의 뉴런 수와 같으며, stem network으로 데이타가 입력되면 높은 유사도를 갖는 몇 개의 뉴런이 선택된다. 하나의 뉴런만을 선택했을 때 벡터 양자화를 잘못 분류하면 그 유형 내에는 정확한 패턴이 없기 때문에 오류가 발생되므로, 이를 줄이기 위해서 복수의 뉴런을 선택한다. 그 다음에, 입력 데이타는 stem network의 선택된 뉴런과 연결되어 있는 몇 개의 branch network에 입력된다. 최종적인 결정은 다음과 같으며, 입력 데이타는 maximum final score Z를 갖는 category로 결정된다.

$$Z=(SM)^{\alpha}\cdot(SB)^{\beta}$$

where, SM = similarity measure in stem network
SB = maximum output score in a branch network

CombNET은 2965자의 인쇄체 한자(JIS 1st level set)을 분류하기 위한 대형의 신경망을 NEURO-TURBO 상에서 구현하였다. NEURO-TURBO 상에서의 전체 learning process를 수행하는 데에는 약 4시간이 소요되며, 이것은 SUN4-260 workstation에서 보다 20배 가량 빠른 것이다. $\alpha/\beta=5$인 경우에 99.5%의 인식률을 얻었다. SM과 SB는 0에서 1의 값이기 때문에 "$\alpha/\beta=5$"라는 것은 stem network의 similarity 측정보다는 branch network의 출력 결과가 더 많은 영향을 갖는다는 것을 의미한다. 즉 입력 데이타가 복수개의 branch network 중의 하나로 분류될 때, correct category에 해당하는 뉴런은 높은 SB를 제공하기 때문에 stem network에서의 matching score SM이 비록 작더라도 final score Z는 커지므로 stem network의 오분류는 branch network에 의해서 복구된다.

Ⅲ. 제안된 문서인식 시스템

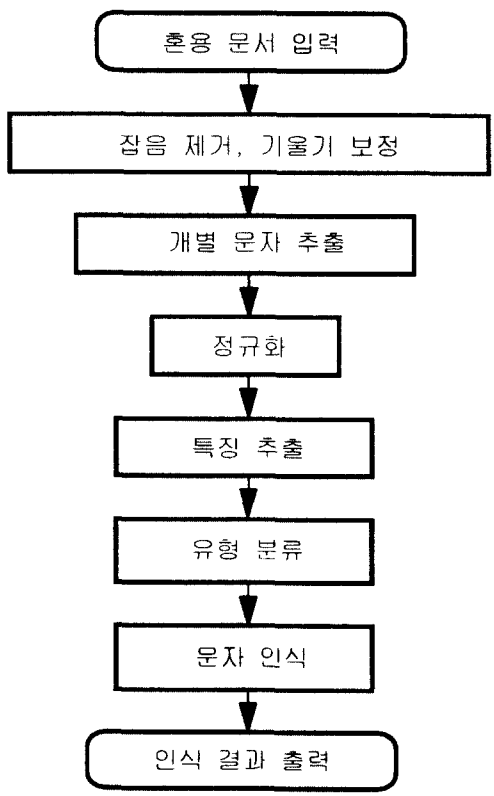


그림 2 제안된 문서 인식 시스템의 구성도
Fig. 2 Block diagram of document recognition system

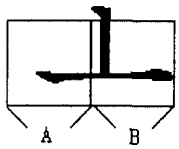
본 논문에서는 다양한 활자체와 다양한 크기의 한글, 한자, 영숫자 그리고 특수 문자가 혼합되어 있는 한글 문서를 인식할 수 있는 문자인식 시스템을 구성하였다. 그림 2는 문서인식 시스템을 나타내었고 인식 대상 문서는 다단(milticolumn)이 아닌 가로쓰기 문서로 제한하며, 입력된 문서 영상은 전처리 과정으로 잡음 제거, 기울기 보정[10], 개별 문자 추출[11], 정규화를 행한다.

3.1 유형분류에 사용되는 특징추출

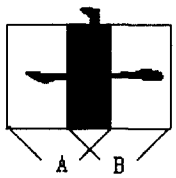
문자 영상 자체를 문자의 특징으로 사용하면 기억용량 및 계산량이 많아지므로 문자의 특성을 잘 표현할 수 있는 특징을 추출하여 입력 패턴으로 사용해야 한다. 어떠한 특징을 사용하느냐에 따라 인식률이 크게 좌우되므로 문자의 특징추출은 매우 중요하다[12]. 특징추출은 2단계로 구성되는데, 첫단계는 유형분류를, 두번째 단계는 문자 인식을 위한 것이다. 본 연구에서는 유형분류를 위한 특징으로 망 특징(Mesh Feature) 및 투영 특징(Projection Feature), 교차거리 특징(Cross Distance Feature) 등을 사용하였다.

3.1.1 망 특징

그림 3에 (a)와 같이 특징추출 영역을 고정된 위치에서 정적으로 분할하면, 문자의 위치 변동이나 다중 활자체를 사용하였을 경우에 특징값이 민감하게 변할 수 있다. 따라서 본 연구에서는 그림 3에 (b)와 같이 이웃하는 메쉬 사이에 빗금 친 영역을 특징추출 영역에 중복되게 포함시켜서 문자 변위와 약간의 변형에 대한 적응성을 갖게 하였다[13]. A의 특징 추출 영역은 빗금친 영역을 포함하여 14×14의 망을 사용하였고, B의 특징 추출도 빗금친 영역을 포함하여 14×14의 망을 사용하였다. 이러한 방법을 사용한 경우 문자의 위치 변동과 변형에 어느 정도의 적응성을 갖고 있다.



(a)고정분할된 영역에서의 특징추출



(b)중복된 위치에서의 특징추출

그림 3. 망 특징
Fig. 3 Mesh feature

3.1.2 투영 특징

투영 특징은 그림 4와 같이 수직과 수평축으로 주사했을 때, 만나는 흑화수의 수를 더해 값을 구하는 방법이다. 이 방법은 동일축 상에 존재하는 긴 획을 잘 표현할 수 있으나, 다양한 서체에서 발생할 수 있는 획의 두께 차이에 의해 민감한 반응을 보이고 있다. 두께의 변화에 따른 특징값의 차이를 해결하는 방법으로 퓨리에 스펙트럼을 얻어 유효한 주파수를 특징으로 사용하는 방법이 있지만, 본 연구에서는 망 특징 추출에서와 같이 특징추출 영역을 중복하여 투영 특징을 추출하였다.

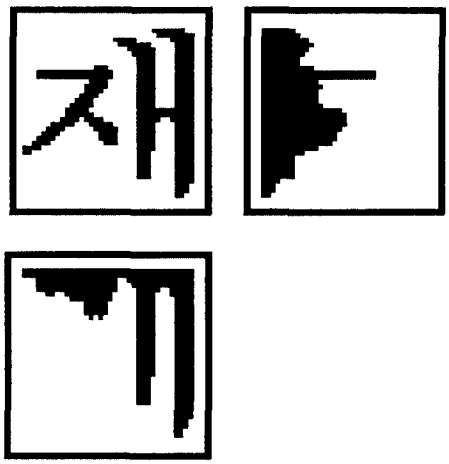
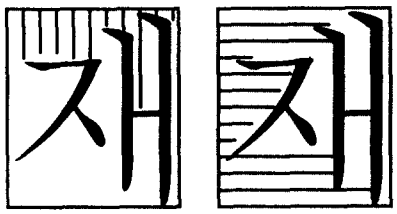


그림 4. 투영 특징
Fig. 4 Projection feature

3.1.3 교차 거리 특징

교차 거리 특징은 수직과 수평으로 주사를 했을 때, 최초의 흑화소를 만날 때까지의 거리를 구하는

방법이다. 이 방법은 문자의 전체적인 외형을 잘 표현하지만, 잡음 성분에 의해 민감한 반응을 보일 수도 있다. 그림 5와 같은 방법을 이용하여 문자 영상의 상, 하, 좌, 우의 4 방향에서 교차거리 특징을 구해 대략적인 외형을 파악할 수 있다.



(a)수직교차거리 (b)수평교차거리

그림 5. 교차 거리 특징
Fig. 5 Cross distance feature

3.2 유형 분류

combNET 구조를 갖는 본 연구의 인식 시스템은 유형 분류를 위해 비지도 학습을 하는 SOFM을 이용하였고, 분류된 유형 내에서는 BP network로 문자인식을 수행하였다. 입력패턴이 어느 유형에 속할지를 정확하게 알 수 없는 경우에는 분류될 유형의 갯수를 미리 정하는 것보다는, 입력 패턴들만의 자연스러운 경쟁을 바탕으로 한 비지도 학습이 바람직하다. 일반적으로 사용되는 K-means 알고리즘을 이용하여 분류될 군집의 갯수를 미리 정해 놓고 군집화를 행하게 되면, 군집내의 구성원들이 강제적인 군집화로 인해 자연스러운 군집을 형성하기가 어렵다. 반면에 SOFM은 출력층의 뉴런수를 임의로 지정하고 적당한 변수로써 학습을 행하면 자연스러운 군집화를 이룰 수 있다.

다자종 문자의 학습을 위해서 모든 문자를 유형 분류망인 SOFM에 한 번에 입력하는 것이 아니라, 자종별로 분리하여 각각을 개별적으로 학습시켰다. 즉 한글 및 한자, 영숫자에 대한 유형분류망이 그림 6과 같이 자종 단위로 구성된다. 이러한 모듈화된 신경망을 이용하면 자종별 단위의 학습이기 때문에 학습이 용이하고, 일어 등의 다른 문자의 추가적인 인식이 가능하다. 또 하나의 유형 내에는 하나의 자종만으로 군집화가 이루어지므로, 특정 자종만을 사용하는 문서의 인식에는 필요한 자종만을 선택해서 사용할 수가 있다. 이러한 모듈화된 신경망에 한글 패턴을 입력하면 한자만으로 구성된 유형이 선택되거나, 한자 패턴을 입력하면 한글로만 구성된 유형이 선택되는 등의 상이한 자종간에 발생할 수 있는 유형 결정의 오류를 예측할 수 있다. 그러나 영숫자는 한글에 비해 단순한 모양을 갖고 있으며, 한자는 한글에 비해 구조 자체가 복잡하고 특징값이 상대적으로 크다. 실험에서는 33차원의 망 특징, 24차원의 교차거리 특징, 12차원의 투영 특징을 추출해 모두 69차원의 특징을 유형분류를 위한 특징으로 사용했기 때문에, 특징 공간상에 위치하는 서로 다른 자종간의 분포는 큰 차이를 보이고 있다. 따라서 서로 다른 자종간에 발생할 수 있는 유형선택의 오류는 매우 적다.

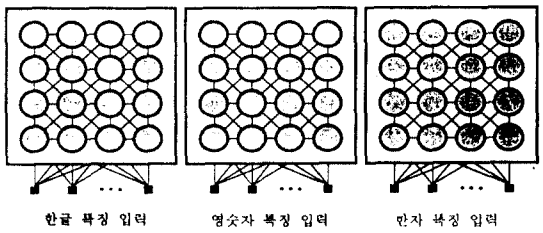


그림 6. 자종 단위의 SOFM
Fig. 6 Modular SOFM

본 논문에서는 유형 분류상의 오류를 줄이기 위해서 어떤 패턴이 입력되었을 때 하나의 유형만을 선택하는 것이 아니라 정합도가 높은 복수개의 유형을 선택하여 그 후보 유형과 연결된 모든 BP network를 활성화하고, 인식시에 정규화 이전의 문자 영상이 갖고 있는 크기 정보를 추가로 사용하기 때문에 오인식을 줄일 수 있다. 이때 정합도가 높은 복수개의 유형을 고정된 갯수로 선택하는 경우에, 입력 패턴에 대해서 올바른 유형이 그것에 포함되어 있지 않으면 유형 분류의 오류가 발생하게 되어서 오인식이 되는 문제를 갖고 있다. 따라서 본 논문에서는 처음에 소수의 유형만을 선택해서 얻은 인식결과가 임의의 임계치 보다 작은 경우에는, 유형의 갯수를 더 확장하여 다시 인식을 하는 방법을 사용했다.

서로 다른 서체의 같은 문자를 다른 패턴으로 간주하면 처리해야 할 패턴의 수가 너무 많기 때문에 다중 활자체의 동일 문자는 같은 문자로 취급하는 것이 바람

직하다. 따라서 실험에서 사용한 모든 서체의 각 문자에 대해 앞에서 설명한 69 차원의 특징을 추출하여, 이를 평균한 평균 특징 벡터를 한 문자의 특징으로 간주하고 이 값을 유형 분류망의 입력으로 사용한다.

3.3 문자인식에 사용되는 특징추출

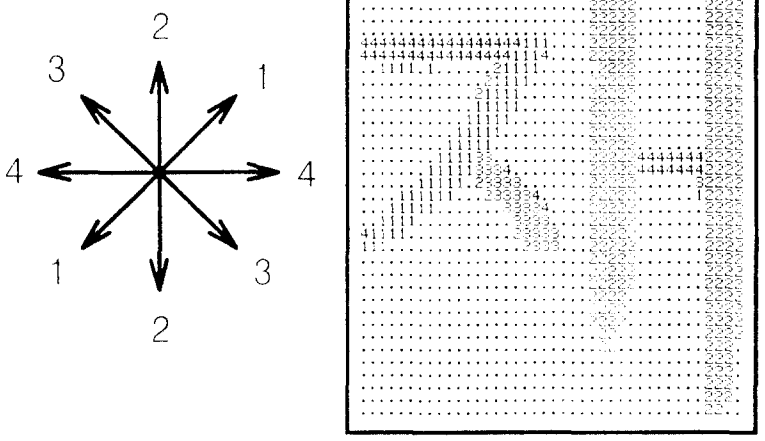
문자인식에 사용되는 특징은 유형분류에 사용했던 망 특징, 교차 거리 특징, 투영 특징 외에 문자영상의 방향밀도 벡터를 추출하여 추가한다. 한 유형내의 모든 패턴은 특징 공간상에서 매우 유사한 위치에 분포하므로, 유형분류에 사용했던 특징만을 문자인식에도 그대로 사용하면 인식 결과가 좋지 않다. 따라서 잡음에 비교적 영향을 적게 받고 문자의 구조를 잘 표현할 수 있는 방향밀도 벡터를 추출하여 문자인식의 특징으로 이용하였다. 여기서 유형 분류에 사용했던 특징들을 문자인식을 위한 특징에 포함시키는 이유는, 복수개의 유형을 선택하기 때문에 문자인식시에 다른 유형내의 문자와 구별할 수 있는 적응성을 갖게 하기 위해서이다.

3.3.1 방향 코드화

방향밀도 벡터를 구하기 위한 전 단계로써, 문자획의 방향성을 파악하기 위해서 방향 코드화를 행한다. 문자의 획 정보를 8 방향에 대해서 구하거나 문자의 외각선에 대해서만 방향성을 조사하는 방법도 있지만, 본 연구에서는 문자 영상의 모든 흑화소에 대해서 4방향으로 방향성을 구하였다. 8 방향으로 획 정보를 구하는 것은 다양한 서체에서 발생할 수 있는 획 길이의 변화에 민감하게 반응할 수 있으며, 문자의 외각선에 대해서만 방향성을 조사하면 문자 영상의 입력 상태가 좋지 않거나 문자 외각선에 잡음 성분이 많은 경우에 방향성을 제대로 표현할 수 없다는 단점이 있다. 따라서 본 연구에서의 방향 코드화는 문자 영상내의 모든 흑화소에 대해서 수평, 수직, 대각, 역대각의 방향성을 조사해서 방향성이 가장 큰 값을 주복 화소의 방향 코드로 결정하였다. 그림 7에는 방향 코드와 방향 코드화된 결과를 나타내었다.

3.3.2 방향 밀도 벡터

방향 코드화 방법에 의해 구해진 문자 영상을 임의 갯수의 영역으로 분할한 후, 영역 내에서 각 방향별로 방향 코드 갯수를 구해 이를 방향 밀도 벡터로 사



(a)방향코드화 (b)방향코드화된 결과

그림 7. 방향 코드화
Fig. 7 Directional Coding

용한다. 아울러 문자영상의 위치변화에 따른 특징값의 차이를 줄이기 위해, 중첩된 특징추출 영역에서 방향밀도 벡터를 구하였다. 그림 8에는 방향 코드화된 문자 영상과 방향 밀도 벡터를 나타내었다.

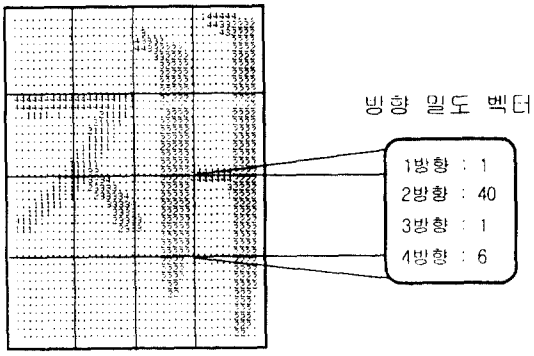


그림 8. 방향 밀도 벡터
Fig. 8 Directional density vector

3.4 문자 인식

유형 분류망에 의해 분류된 각 유형내의 문자는 BP network을 이용하여 인식을 행한다. 본 연구에서는 한글과 영숫자는 6개, 한자는 3개의 서체를 실험 대상으로 하였으며, 이들 서체의 학습뿐만 아니라 각 서체의 계열별로 평균한 특징값을 추가로 학습시켜서 비학습 문자에 대한 일반화 능력을 증가시켰다. 즉, 명조체 계열의 서체들과 고딕체 계열의 서체들을 각각 평균한 2개의 평균 특징 벡터를 추가로 학습시켰다. 실제로 평균 방향 밀도 벡터만을 사용하여 인식 실험을 한 결과에서도, 평균 특징값이 미지의 패턴에 대해서 어느 정도의 적응성이 있음을 확인하였다.

SOFM으로 유형을 분류할 때 이웃 범위의 모양은 정사각형으로 하였다. 학습 후 생성되는 전체 출력층의 분포를 관찰해 보면, 비슷한 패턴끼리는 거리적으로 가까운 범위 안에 위치하는 것을 알 수 있다. 따라서 그림 9와 같이 각 유형간의 경계 위치 상에 존재하는 패턴들은 정확한 유형을 선택하기가 매우 어렵다.

이러한 모호함으로 인해, 어떤 문자가 처음에는 유형 1로 학습되었으나 두번째 인식시에는 유형 2로 판명되었을 경우가 발생되어 유형 분류의 오류로 인한 오인식이 일어날 수 있다. 따라서 본 논문에서는 각 유형내의 문자를 학습시킬 때, 그림 10과 같이 이웃하는 8 방향에 있는 유형내의 문자들도 함께 학습시켰다. 가까운 거리상에 있는 유형들의 유형 분류율은 유사하다. 따라서 각 유형간의 경계 위치 상에 있는 문자들은 정확한 유형을 선택하기가 어렵다. 본 연구에서는 이러한 문제점을 인식단에서 해결하기 위해서, 이웃하는 유형내의 문자들을 함께 학습하는 방법을 사용하였다.

예를 들어, 그림 10과 같이 5번째 유형에는 "돔, 둡, 몸, 송, 웅, 읍"의 문자가 존재한다. 이문자들을 학습할 때, 이웃하는 유형내의 문자들도 함께 학습한다. 이때 이웃하는 유형내의 문자들에 대해서는, 5번째 유형의 인식단에서 반응되지 않기 위해서 모든 출력

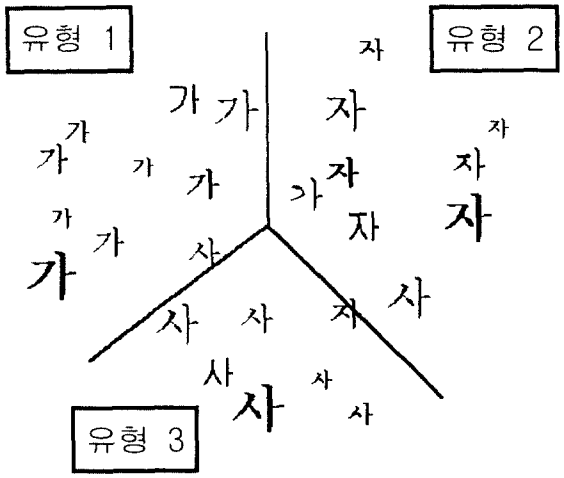


그림 9. 유형 경계상의 패턴 분포
Fig. 9 Patterns on cluster boundary

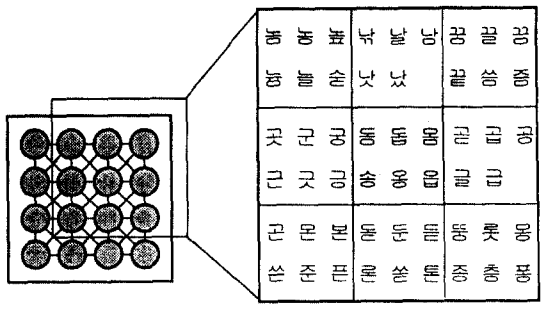


그림 10. 이웃하는 유형내 문자의 추가 학습
Fig. 10 Training with patterns in neighboring clusters

값을 0으로 한다. 이렇게 하면, 8번째 유형내의 '돋'이라는 문자가 유형 분류의 오류로 인하여 5번째 유형이 선택되더라도 인식단에서의 반응이 억제되어 있기 때문에 문자 인식율은 매우 낮게 나타난다.

Ⅳ. 실험결과 및 고찰

본 연구는 PC 상에서 Hewlett Packard의 ScanJet 스캐너를 이용하여 300 dpi로 입력받은 문서 데이타에 대해서 PC와 LAN으로 연결된 SUN Sparc-II workstation을 사용하여 X-window 상에서 C 언어로 구현하였으며 그림 11에는 실험 환경을 나타내었다.

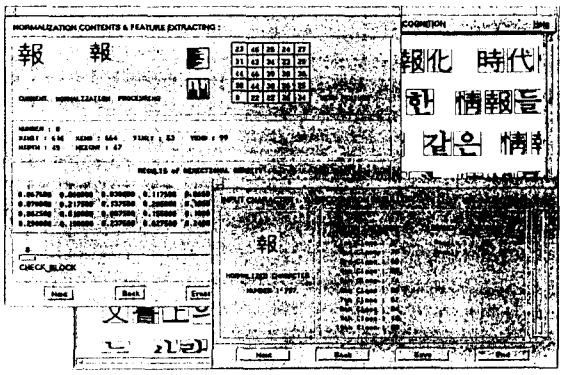


그림 11. 실험 환경
Fig. 11 Experimentation tool

표 1. 학습 서체
Table 1. Training font

	아래아 신명조	아래아 중고딕	서울 신명조	시울 세명조	시울 세고딕	보스-M 바탕체
한글	한	한	한	한	한	한
영숫자	A	A	A	A	A	A
한자	漢	漢	漢			

학습에 사용된 서체는 표 1과 같이 한글과 영숫자의 경우에는 아래아 한글의 신명조와 중고딕, MS-Windows의 바탕체, (주) 서울 시스템의 신명조, 세명조, 세고딕의 6개 서체를 사용하였으며, 한자의 경우에는 아래아 한글의 신명조와 중고딕, (주) 서울 시스템의 신명조의 3개 서체를 사용하였다. 실험에서는 각 서체의 한글 찾기순 상위 990자, 영숫자와 특수기호 79자, 교육부 지정 중학교 교육용 한자 900자를 사용하였으며, 전처리 과정으로서 잡음제거, 기울기 보정, 개별 문자 추출, 정규화를 행하였다.

4.1 유형 분류부

분류부에서 사용하는 SOFM은 이웃의 모양을 정사각형으로 하고 학습율과 이웃의 범위는 시간이 지남에 따라 감소시키면서 500회를 수행하였다. 먼저 분류수와 분류율의 관계를 조사하는 실험을 행하였다. 유형 분류율을 높이기 위해서는 분류수를 작게 하는 것이 좋고, 인식율을 높이기 위해서는 한 유형내의 패턴 수를 작게 하는 것이 좋다. 분류될 유형의 갯수가 너무 적으면 중심 벡터와 패턴들 간의 분산이 너무 커서, 패턴의 분포를 제대로 표현할 수 없으며 인식단에서 부담이 커지는 문제가 있다. 또 유형의 갯수가 너무 많으면, 유형의 분포가 패턴 공간상에서 유사한 위치에 존재하므로 유형 분류의 오류가 생기며 오인식의 가능성이 크고 인식 속도가 늦어지는 단점이 있다. 본 논문에서는 최적화된 유형의 갯수를 조사하기 위해서, 유형의 갯수를 달리하면서 유형 분류에 대한 실험을 하였다. 실험에서는 33차원의 망 특징, 24차원의 교차 거리 특징, 12차원의 투영 특징을 추출해서 평균을 취한 모두 69차원의 평균 특징 벡터를 유형 분류를 위한 특징으로 사용하여 대분류를 하였다. 문자 영상에서의 정보량 분포를 조사한 연구[14]에 의하면, 대부분의 정보량은 활자체에는 크게 영향을 받지 않으며 문자의 가장자리 부분에서 얻어진 특징들이 문자 인식에 의미 있게 사용됨을 알 수 있다. 따라서 본 논문에서는 69차원의 유형 분류에 사용되는 특징중에 48차원의 특징을 문자의 가장자리 부분에서 추출하였고, 자종간의 특징값 차이가 크게 나타나는 망 특징을 가장 많이 사용하였다. 표 2와 그림 12에는 유형의 수를 다양하게 변화하면서 행한 유형 분류에 대한 실험결과를 수치와 도표로 나타내었다.

표 2와 그림 12에 보인바와같이 선택한 유형의 갯

표 2. 다양한 유형수에 대한 유형 분류율
Table 2. The classification rate of number of clusters

순 위 / cluster 수	1순위	3순위	5순위	10순위	20순위
16 clusters	72.1	97.6	99.5	100	
25 clusters	72.0	95.0	98.4	99.8	100
36 clusters	71.0	92.5	96.2	99.5	100
49 clusters	68.7	91.7	96.2	99.5	100
100 clusters	51.4	76.5	85.3	92.3	96.5

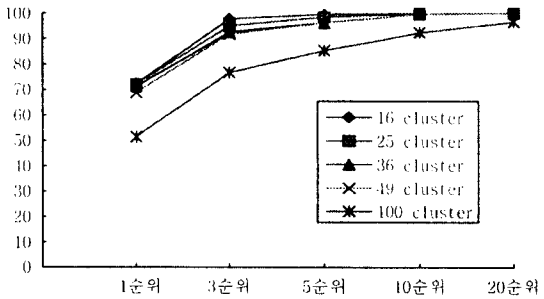


그림 12. 다양한 유형수에 대한 유형 분류율
Fig. 12 Number of clusters vs. fine classification rate

수가 적을수록 유형 분류율이 높게 나타났고, 16개에서 49개의 유형을 사용하였을 때는 대체로 비슷한 결과를 나타내었다. 본 논문에서는 유형 분류율도 높고 패턴간의 분포도 잘 나타난 것으로 판단된 36개의 유형을 한글 990자에 적용하였으며 영숫자는 16개, 한자는 36개의 유형을 선택하였다.

표 3. 다양한 활자체에 대한 유형 분류율
Table 3. The fine classification rate of various fonts

순위 / 서체명	1순위	3순위	5순위	10순위	20순위	30순위	40순위
아래아 중고딕	33.1	48.0	69.6	91.8	95.3	99.2	100
아래아 신명조	66.4	86.5	91.9	96.9	99.8	100	
서 울 세명조	38.9	69.5	84.1	95.0	96.6	98.7	100
서 울 신명조	68.1	89.5	93.9	97.9	99.5	100	
서 울 세고딕	54.8	78.9	88.4	97.0	97.0	99.6	100
MS-W 바탕체	76.0	95.0	98.9	100			

다음은 실험에 사용한 각 서체에 대한 유형 분류 실험을 하였다. 한글 유형 36개, 영숫자 유형 16개, 한자 유형 36개에 대해서 각 서체별 유형 분류율 측정하였다. 표 3과 그림 13에서 보듯이 아래아 한글의 중고딕체의 유형 분류율이 가장 낮게 나타났다. 이러한 원인은 실험에서는 명조체 계열의 서체를 4개를 사용하였고 고딕체 계열의 서체를 2개를 사용하였기 때문에, 명조체 계열의 특징값이 평균 특징 벡터에 많이 반영되는 것으로 판단된다. 즉, 아래아 한글 중고딕체의 특징값이 vector 공간상에서 평균 특징 벡터와 어느 정도의 차이를 보이기 때문에 유형 분류율이 낮게 나타났다.

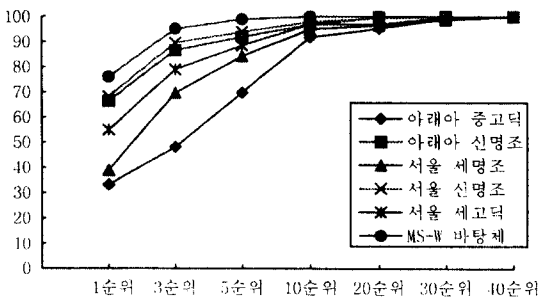


그림 13. 다양한 활자체에 대한 유형 분류율
Fig. 13 Various fonts vs. fine classification rate

4.2 인식부

분류부에서 구한 69 차원의 특징과 방향 밀도 벡터 160 차원의 전체 229 차원의 특징을 인식을 위한 특징 벡터로 사용한다. 일반화 능력을 갖기 위해서, 학습 서체 이외에 명조체 계열의 평균 벡터와 고딕체 계열의 평균 벡터를 구해 함께 학습을 행한다. 표 4에서와 같이, 유형 분류율이 가장 낮은 아래아 중고딕체의 인식율이 가장 낮게 나타났다. 여기서 측정한 인식율은 유형 분류의 오류를 포함한 수치이기 때문에, 유형 분류율이 낮은 서체는 인식율도 비교적 낮다.

표 4. 각 서체에 대한 인식율
Table 4. The recognition rate of various fonts

서체명	아래아 중고딕	아래아 신명조	서울 세명조	서울 신명조	서울 세고딕	MS-W 바탕체	평균
인식율	90.6	95.9	97.1	96.7	94.8	98.2	95.6

본 논문에서는 학습 서체 외에 명조체 계열의 평균 특징 벡터와 고딕체 계열의 평균 특징 벡터를 학습에 참여시켜 일반화 능력을 갖도록 하였다. 이번 실험에서는 상기 2개의 평균 벡터만을 학습시킨 후에, 학습에 참여한 6개의 서체 한글 990자를 대상으로 인식율을 구하였다. 실험 결과, 표 5에 나타낸 바와같이 고딕체 계열의 인식율이 낮게 나타났다. 이러한 원인은 아래아 중고딕체는 두꺼운 획을 갖고 있어서 복잡한 문자의 경우에는 인접한 획들이 서로 붙어 있기 때문에 문자의 방향성을 정확하게 추출할 수 없고, 평균 특징 벡터와의 거리값이 상대적으로 크기 때문이다. 또, 명조체 계열의 경우에는 MS-WINDOWS 바탕체의 인식율이 가장 좋게 나타났는데, 이것은 바탕체의 특징값과 평균 특징 벡터가 가장 유사하기 때문인 것으로 판단된다.

표 5. 평균 특징 벡터에 대한 인식율
Table 5. The recognition rate for average feature vector

서체명	아래아 중고딕	아래아 신명조	서울 세명조	서울 신명조	서울 세고딕	MS-W 바탕체	평균
인식율	22.3	37.5	35.3	41.4	28.5	44.2	34.9

다음은 자주 오인식 되는 문자에 대해서 조사를 하였다. 오인식되는 경우는 서로 다른 자종간에 발생되는 유형 선택의 오류로 인한 오인식과, 동일 자종내에서 유사 문자로 인한 오인식이 있다. 오인식의 주요 원인은 유형 선택의 오류로 인한것이 대부분이기 때문에, 유형 선택의 오류를 최소화할 수 있는 추가적인 학습 방법이 요구된다. 표 6에는 자주 오인식이 발생되는 문자의 예를 나타내었다.

다음으로 인식 시스템의 중요한 성능평가 요소인 인식속도를 측정하였다. 본 논문에서는 복수개의 후보유형을 선택하고 그것에 연결된 BP 네트워크를 기동시켜 인식을 행하기 때문에, 후보유형의 수와 각 유형에 연결된 BP 네트워크의 중간층 뉴런 수에 따라 인식속도는 차이를 보이고 있다. 실제 실험에서는 5순위 까지의 후보유형을 선택하여 인식을 행하고 인식 결과가 임계치 보다 낮은 경우에는 5개의 후보 유형을 더 설정하여 재인식하는 방법을 사용하였는데, 그 결과 초당 10.3자를 인식하였다. 이것은 정규화 과정, 특징추출 과정, 유형분류와 세분류를 모두 포함한 속도이다. 유형선택에 대한 부분을 보완하여 유형 분류율을 더 향상시키면 후보 유형의 수를 줄일 수 있기 때문에, 인식 속도는 더 향상될 수 있다고 판단된다.

4.3 미학습 데이타와 실제 문서의 적용

본 논문에서는 제안된 인식 시스템의 성능을 평가하기 위해서, 미학습 데이타인 아래아 한글의 신명중명조체와 실제 문서에 대해 적용해 보았다. 신명중명조체의 유형 분류율은 5순위 까지는 86.0%, 10순위 까지는 95.2%를 나타냈으며 인식율은 93.1%를 나타내었다. 이것은 학습 데이타의 평균 인식율인 95.6%에 비해 다소 떨어진 수치이다.

최종적으로 실제 문서에 대해 인식 실험을 하였다. 문서내의 총 문자수는 774자이고, 이 중 한글은 584자이고 영숫자는 25자, 한자는 165자이다. 개별문자 추출의 오류는 13자이고 이를 포함한 인식율은 91.1%이며, 개별문자 오류를 제외한 인식율은 92.6%이다.

4.4 고 찰

제안된 인식 시스템은 SOFM을 이용하여 유형 분

표 6. 오인식 문자의 예
Table 6. An example of character recognition error

구 분	오인식 문자의 예
한 글	구⇒T 괴⇒피 그⇒2 깨⇒열 끈⇒끝 몸⇒릎 뮬⇒期 배⇒베 뱅⇒탱 베⇒떼 북⇒묵 븕⇒昔 솟⇒大 숲⇒舍 연⇒면 온⇒본 윤⇒유 쟁⇒젬 즘⇒줌 쫓⇒美 키⇒外 커⇒걱 콘⇒론 투⇒早 패⇒폐 폭⇒목 회⇒희 후⇒혹
영숫자	I⇒l O⇒口 T⇒丁 Z⇒之 x⇒k l⇒l 2⇒그 .⇒,
한 자	貴⇒봅 갑⇒뿌 已⇒巳 刀⇒力 明⇒朋 間⇒聞 上⇒寸 事⇒봅 右⇒左 陽⇒楊 日⇒日 王⇒조 肉⇒丙 入⇒八 者⇒老 直⇒面 晴⇒떤 牛⇒生 貝⇒兒 合⇒습 向⇒同

류를 수행하고 각각의 유형내에서는 BP 네트워크를 이용하여 인식을 행하였다. 대부분의 오인식은 유형 선택의 오류로 인한 것이 대부분이다. 또 인식속도를 증가시키기 위해서는 후보유형의 수를 줄여야 하는데, 이를 위해서는 유형 분류율을 향상시켜야 한다. 본 연구에서는 유형 분류시에 평균 특징 벡터를 구해서 사용했는데, 평균 특징 벡터와 거리값의 차이가 큰 아래아 한글 중고딕체에 대해서는 유형 분류의 오류가 자주 발생되는 문제점이 있다. 따라서 평균 특징 벡터에 의한 유형 분류를 바탕으로 하여, 다중 활자체의 동일 문자가 같은 유형에 속하도록 추가적인 지도 학습이 필요하다고 생각된다. 또 획이 두꺼운 문자는 이웃하는 획과 접촉되어서 방향성을 제대로 추출하기 어렵기 때문에, 이를 보완할 수 있는 세부 특징의 추출이 요구된다.

V. 결 론

본 연구에서는 신경회로망을 이용하여 다중 크기, 다중 활자체, 다자종 문자가 포함되어 있는 한글 문서를 인식할 수 있는 문자 인식 시스템을 구현하였다. 기존의 문자 인식 시스템은 한글의 구조적인 특성에 의존하였기 때문에, 한글이 아닌 다른 문자가 혼용된 문서의 경우에는 추가적인 문자 구별 과정이 선행되었고, 각 문자의 형식 분류에서도 구조적인 특성에 따라 강제적으로 분류하였다. 따라서 본 논문에서는 기존의 문자 인식 시스템에서 발생되는 문자 구별 오류와 형식분류 오류를 줄이기 위해서, CombNet 구조를 갖는 신경 회로망을 이용하여 비지도 학습을 통해 각 문자의 자연스러운 유형분류를 행하였다. 대분류시에 자종별로 학습을 행함으로써, 특정한 자종만을 사용하는 문서의 인식에는 필요한 자종만을 선택해서 사용할 수가 있으며 새로운 자종의 추가 학습이 용이하다. 각 유형내의 문자를 학습시킬 때 이웃하는 8 방향에 있는 유형내의 문자들은 반응이 억제되도록 함께 학습시켜서 각 형식간의 오분류로 인한 오인식을 인식단에서 줄였다. 또, 유형분류에서 오류가 발생된 문자에 대해서도 잘못 할당된 유형에서는 반응이 억제되도록 추가 학습을 함으로써 인식율을 향상시킬 수 있다. 제안된 인식 시스템이 일반화 능력을 갖을수 있도록 학습 서체뿐만 아니라 각 서체별 평균 특징 벡터를 추가로 학습하였다. 특징 추출시에는 문자 변위에 대한 적응성을 갖기 위해서 특징 추출 영역을 중복되게 사용하였고, 유형 분류에 사용되는 특징은 문자 정보량이 많이 존재하는 문자 패턴의 가장자리 부분에서 많이 추출하였으며 각 자종간에 발생할 수 있는 유형 분류의 오류를 줄이기 위해 자종간의 특징 값 차이가 큰 망 특징을 많이 사용하였다.

실험에서는 한글과 영숫자는 6개의 서체, 한자는 3개의 서체로 학습한 결과, 미학습 서체에 대해서는 93.1%의 인식율을 보였고 실제 문서에 대해서도 92.6%의 인식율과 초당 10.3자의 인식 속도를 보임으로써 제안된 인식 시스템의 유용성을 입증하였다.

향후 진행 과제는 제안된 인식 시스템을 실용화하기 위해서 한자의 경우 1,800자 이상의 문자에 대한 추가적인 학습이 필요하다. 이 문제는 패턴의 분포를 고려하여 유형의 갯수를 변화시켜 학습시키면 충분히 가능하리라 본다. 본 연구에서는 유형 분류시에 다중 활자체의 동일 문자를 같은 패턴으로 간주하기 위해 평균 특징 벡터를 구해서 사용했는데, 평균 특징 벡터와 거리값의 차이가 큰 패턴에 대해서는 유형 분류의 오류가 발생되는 문제점이 있다. 또, 유형 분류율과 인식 속도를 향상시키기 위해서는 패턴의 분포를 제대로 표현할 수 있는 범위 안에서 유형의 수를 가능한 한 작게 하는 것이 필요하다. 따라서 평균 특징 벡터에 의한 유형 분류를 바탕으로 하여, 다중 활자체의 동일 문자가 같은 유형에 속하도록 LVQ 등의 신경망을 이용한 추가적인 지도 학습이 필요하다고 생각된다. 인식시에 주로 사용한 방향 밀도 벡터는 문자의 획이 두껍고 문자 구조가 복잡한 패턴에 대해서는 좋은 결과를 보이지 못하기 때문에, 이러한 문제들을 극복할 수 있는 새로운 특징의 선택이 요구된다. 영숫자가 혼합된 문서에서는 개별 문자 추출시에 오류가 발생되는 경우가 종종 있는데, 이에 대한 추가적인 보완이 필요하며, 문서 영상 전반에 대한 구조 분석과 이해가 필요하다.

참 고 문 헌

1. 김우성, 방승양, "신경회로망을 이용한 한글 한자 혼용 문서 인식에 관한 연구", 전자공학회 논문지, vol.29-B, no.2, pp.162-171, 1992년

2. 김헌종, "SOFM 신경망을 이용한 계층적 구조의 한글과 한자의 분리에 관한 연구", 경북대학교 석사 학위 논문, 1994년
3. IWATA, HOTTA, MATSUO, SUZUMURA, MATSUDA, "A large scale neural network CombNET on neural network accelator(NEURO-TURBO)," proceedings of the Internai Conference on Fussy Logic & Neural Networks, pp.329-333, 1990.
4. Iwata A, Ino Y, Hotta K. and Suzumura N, "Hand-written Japanese Kanji character recognition by selfgrowing neural network CombNET-II," Artificial Neural Network, 2 (Proc. of ICANN 92), pp.1189-1192, 1992.
5. Masayuki A, Kenzo O. and Jyuichi M, "Thousands of Hand-written Kanji Recognition by HoneycombNET," 電子情報通信學會論文誌 D-II, vol. J176-D-II, no.11, pp.2316-2323, 1993年 11月
6. Masayuki A, Kenzo O. and Jyuichi M, "Thousands of Hand-written Kanji Recognition by HoneycombNET-II," 電子情報通信學會論文誌 D-II vol. J177-D-II, no.9, pp.1708-1715, 1994年 9月
7. Kohonen T, "Self organization and associative memory," Spriger verlag, 1984.
8. R.P. Lippmann, "Neural Nets for Computing," ICASSP, vol.1, pp.1-5, 1988.
9. D.E. Rumelhart, G.E Hinton, R.J Wiliams, "Learning Internal Representations by Error Propagation," Parallel Distributed Processing, vol.1, MIT press, 1986.
10. R.O. Duda and P.E. Hart, "Use of Hough TRansformation to detect Lines and Curves in Pictures," Communications the ACM, vol.15, no.1, January 1992.
11. 홍성붕, "인쇄체 한글 문서에서의 동적 문자추출에 관한 연구", 광운대학교 석사학위논문, 1994년
12. 김우태, 윤병식, 박인규, 진성일, "인쇄체 한글 문자인식을 위한 특징 성능의 비교", 한국정보과학회지 논문지 vol.20, no.8, pp.1103-1110, 1993년
13. S. Ling, M. Shridhar and M. Ahmadi, "Efficient Algorithms for Segmentation and Recognition of printed characters in Document Processing," IEEE Pacific Rim Conf. on Communication, Computer and Signal processing, pp.240-243, Victoria, Canada, May 19-21, 1993.
14. 이성환, "한글 문자 영상에서의 정보량 및 엔트로피의 분포", 제2회 한글 및 한국어 정보처리 학술대회 발표 논문집, 서울 1990년 11월, pp.208-215.

최 재 혁(Jae Hyuk Choi) 정회원
1993년 2월 : 광운대학교 전자계산기공학과 졸업(공학사)
1995년 6월 : 광운대학교 대학원 컴퓨터공학과 졸업(공학석사)
1995년 7월~현재 : 현대정보기술(주) 의료사업부 연구원
※관심분야 : 영상처리, 패턴인식, 신경회로망, 문서인식

손 영 우(Young Woo Shon) 정회원
1981년 2월 : 광운대학교 전자공학과 졸업(공학사)
1983년 2월 : 광운대학교 대학원 전자공학과 졸업(공학석사)
1996년 2월 : 광운대학교 대학원 컴퓨터공학과 박사과정 수료
1991년 4월~현재 : 산업기술정보원(KINITI) 책임연구원
※관심분야 : 패턴인식, 신경회로망, Chaos 이론, 문서인식

남궁재찬(Jae Chan Namkung) 정회원
1970년 2월 : 인하대학교 전기공학과 졸업(공학사)
1976년 8월 : 인하대학교 대학원 전자공학과 졸업(공학석사)
1982년 2월 : 인하대학교 대학원 전자공학과 졸업(공학박사)
1982년~1984년 : 일본 동북대학 객원교수
1979년 3월~현재 : 광운대학교 컴퓨터공학과 교수
※관심분야 : 패턴인식, 신경회로망, Chaos 이론, 문서인식