혼용 문자 코드 집합을 위한 계층적 다중 문자 인식기

김도현* · 박재현* · 김철기* · 차의영*

Hierarchical Multi-Classifier for the Mixed Character Code Set

Do-Hyeon Kim* · Jae-Hyeon Park* · Cheol-Ki Kim* · Eui-Young Cha*

이 논문은 2004년도 한국학술진흥재단의 지원에 의하여 연구되었음 (KRF-2004-036-D00385)

요 약

문자 인식은 인공지능의 한 분야로써 자동화 시스템, 로봇, HCI 분야에서 그 응용성이 증대되고 있는 첨단 기술이다. 본 논문에서는 숫자, 기호, 영어, 한글이 여러 가지 형태로 조합되어 사용될 수 있는 영역에서의 문자 인식을 위해 인식 문자 집합과 대표 문자를 도입하였다. 여러 가지 조합의 언어 집합에 따른 소규모 인식기를 계층적으로 조합하여 인식 결과의 정확성을 높이고 시간 비용을 줄일 수 있는 효율적인 인식기 구조를 제안하였다. 그리고 학습 성능이 우수한 Delta-bar-delta 알고리즘을 이용하여 개별 소규모 인식기를 학습한 다음 다양한 개별 문자를 대상으로 그 인식 성능을 살펴본 결과 99%의 인식률을 획득함으로써 혼용 언어 문자 인식의 효율성과 신뢰성을 증명하였다.

ABSTRACT

The character recognition technique is one of the artificial intelligence and has been widely applied in the automated system, robot, HCI(Human Computer Interaction), etc. This paper introduces the character set and the representative character that can be used in the recognition of the image ROI. The character codes in this ROI include the digit, symbol, English and Korean, etc. We proposed the efficient multi-classifier structure by combining the small-size classifiers hierarchically. Moreover, we generated each small-size classifiers by delta-bar-delta learning algorithm. We tested the performance with various kinds of images and achieved the accuracy of 99%. The proposed multi-classifier showed the efficiency and the reliability for the mixed character code set.

키워드
혼용 문자 인식, 계층적 인식기, Delta-bar-delta

I. 서 론

사람들이 다루는 대부분의 정보는 사진, 그림, 도형 등의 영상 정보와 한글, 한자, 숫자, 부호 등의 문자 정보로 구성되어 있으며, 이러한 정보를 자동으로 컴퓨터에 입력하기 위하여 OCR(Optical Character Recognition)에 관한 연구가 꾸준히 진행되고 있다. 문자 인식 분야는 실용적으로 컴퓨터와 인간의 보다 원활한 인터페이스(interface)를 추구하려는 목적에서부터 출발하였으며 자동화 시스템, 로봇, HCI(HMI) 분야에서 그 활용성은 매우 크다. 학문적으로 문자 인식 기술은 인간의 우수한 능력 중의 하나인 패턴 인식의 능력을 컴퓨터에게 부여

* 부산대학교

접수일자 : 2007. 5. 14

하여 인간처럼 사고하고 판단할 수 있는 인공지능 컴퓨터의 실현에 목적이 있다고 할 수 있다. 특히, 근래에는 휴대가 가능한 초소형 모바일 컴퓨터의 등장과 함께 작고 간편한 명령 입력 도구의 개발을 위한 온라인 문자 인식 기술에 대한 연구가 크게 주목받게 되었으며, 대용량 멀티미디어 시대의 도래와 함께 기존의 방대한 문서 정보를 일괄적으로 전산화화 할 수 있는 오프라인 문자 인식에 관한 연구도 지속적으로 이루어지고 있다.

문자 인식 분야에서 인식의 대상이 되는 문자는 한글을 비롯하여 숫자, 기호, 영어 등 다양한 언어를 포함한다. 숫자나 기호, 영어 등은 인식해야할 문자 클래스의 수가 작기 때문에 간단한 분류기(classifier)로도 만족할 만할 성능을 보이지만 한글은 완성형의 경우 2,350자, 유니코드의 경우 11,172자(21*19*28)의 방대한 클래스를 가지므로 단일 인식기로 인식하는 것은 거의 불가능에 가깝다. 뿐만 아니라 이들 언어들이 혼용될 경우 인식 클래스의 수는 더욱 많아진다.

대부분의 기존 방법들에서는 한글만을 대상으로 하거나 다른 언어들과 혼용될 경우 확정된 문자 집합만을 인식하기 위한 방법들이 연구되었다. 이렇게 고정된 문자 집합을 대상으로 설계된 인식기를 사용해서 문자를 인식할 수도 있지만 그 영역이 숫자만으로 되어 있거나 영어+숫자 조합 등 특정한 언어 조합으로 구성되어 있는 경우가 많다. 이와 같은 특정한 부분 영역에 대한 사전 정보가 있을 때에는 해당 영역별로 인식 문자 집합을 변경해서 해당되는 조합의 언어로만 학습된 인식기를 사용하는 것이 인식 대상이 되는 모든 언어가 조합된 인식기를 사용하는 것보다 효율적이며 합리적이다. 예를 들어, 이름 영역을 한글 전용 인식기를 사용하여 인식하게 되면 한글이 아닌 숫자나 영어 등으로 특정 문자가 인식되는 결과를 미연에 방지할 수 있다. 이와 같은 점을 고려하여 본 논문에서는 숫자, 기호, 영어, 한글이 여러 가지 형태로 조합되어 사용될 수 있는 영역에서의 문자 인식을 위해 인식 결과의 정확성을 높이고 시간 비용을 줄일 수 있는 효율적인 인식 방법과 인식기 구조를 제안하고자 한다.

II. 관련 연구

문자는 입력된 문자의 생성 형태에 따라 인쇄체 문자(printed character)와 필기체 문자(handwritten character)로 구분된다. 인쇄체 문자 인식 분야에서는 다중 글꼴(font)에 대한 인식의 문제가 비교적 오랫동안 연구되어 왔으며, 글꼴에 의한 변형을 최소화하는 특징 추출기의 설계와 글꼴의 특성에 구애받지 않는 인식기의 구현을 연구하고 있다. 이에 반해 필기체 문자는 인쇄체와는 달리 필기자의 필기 특성에 따른 문자의 변형이 심하고 동일 필자에 대해서도 필기시마다 문자의 형태가 다양하기 때문에 전처리 및 특징 추출 단계에서 문자의 변형을 최소화하고, 획의 변형에 대한 적응적 정합 기능을 갖는 인식기의 개발을 연구하고 있는 실정이다.

오프라인 문자 인식 방법은 크게 원형 정합 방법(template matching), 확률 통계적 방법(statistical approach), 구조적 방법(structural approach), 신경회로망을 이용한 방법(neural network based approach), SVM(support vector machine)을 이용한 방법 등으로 분류할 수 있다[1].

원형 정합 방법은 입력 문자 영상을 인식 대상이 되는 모든 문자 모델과의 정합을 통하여 유사도나 거리를 구하여 인식하는 단순한 방법으로 단일 글꼴과 같이 변형이 심하지 않은 문자 영상에는 적합하나, 필기체와 같이 변형이 심한 경우 적용하기 어렵다.

확률 통계적 방법은 수학적인 해석에 바탕을 두고 문자 영상의 통계적인 특징을 분석하는 방법으로 표현하고자 하는 대상 패턴을 확률 값으로 표현하고 입력 패턴에 대해 이를 주어진 모델로 생성해 낼 수 있는 확률 값을 계산하여 가장 높은 확률 값을 갖는 모델로 분류하는 방법이다. 하지만 이와 같은 방법은 문자 패턴이 다양하고 유사한 글자가 많은 문자 집합에 대해서는 문자의 본질적인 구조적 특징을 이용하기가 쉽지 않은 단점이 있다.

구조적 방법은 문자의 구성 원리에 입각하여 획 등과 같이 문자를 구성하는 기본 요소와 그들과의 연관성을 추출하여 문자를 인식하는 방법으로 특히 복잡한 구조를 가지는 문자 집합의 인식이나 글자 모양의 변형이 심한 필기체 문자의 인식에는 적합하지만 인식 알고리즘의 기본이 되는 기본 요소의 추출이 어렵고, 문자의 구조를 표현할 수 있는 규칙의 생성과 규칙에 기반한 문법적 추론 기법에 관한 연구가 아직 미흡한 실정이다.

신경회로망을 이용한 방법은 인간의 두뇌를 모델화한 방법으로 패턴의 국부적인 변형 및 잡영에 민감하지 않다는 장점을 가지고 있으나, 패턴의 수가 많아질수록

학습에 걸리는 시간이 길어지며, 인식 후보 대상의 개수(클래스)가 많은 경우 성능이 저하된다는 단점을 가지고 있다.

SVM을 이용한 방법은 통계적 학습 이론을 바탕으로 최근 기계 학습 분야에서 주목 받고 있는 패턴 분류 방법으로 일반화 성능이 우수하다. SVM이 기본적인 2분류 인식기이므로 다 클래스 패턴 인식에 적용할 경우 이를 확장해야 하며 인식 클래스의 수가 많은 경우에는 학습을 위해 필요한 비용이 매우 많이 든다.

최근 들어 진행되는 대부분의 문자 인식에 관련된 연구는 특정한 응용 목적에 따른 문자만을 대상으로 한 연구[2]와 범용 문자 인식기로 숫자, 기호, 영어, 한글, 한자 등 다국어 인식을 위한 연구[3-5]로 구분된다. 특히, 범용 문자 인식기 구현은 인식해야 할 문자 클래스가 매우 많기 때문에 기존의 간단한 분류기(classifier)로는 해결할 수 없는 경우가 많다. 그러므로 인식의 대상이 되는 문자들을 몇 개의 유형으로 대분류한 다음 분류된 유형 내에서 다시 개별 코드를 인식하는 방법이 주를 이루고 있다. 이와 관련된 연구로는 템플릿 매칭에 의해 지로 서식 문서에서의 인쇄체 숫자열을 인식하는 방법에 대한 연구[2], 한글과 한자가 혼용된 문서를 빠르게 인식하기 위해 SOFM을 수정한 신경회로망을 사용하여 한글 및 한자를 유형 분류한 다음 분류된 각 유형 내에서 APC(Adaptive Pattern Classifier)를 이용하여 인식하는 방법에 대한 연구[3], 한글 유형 분류에 의해 탐색 공간을 줄이고 이후 분류된 유형에서 음소를 인식한 후 피드백에 의해 오류를 수정하는 혼합형 제어 전략 방법[4] 등이 있다.

III. 계층적 다중 문자 인식기

3.1. 인식 문자 집합 및 대표 문자 정의

표 1 . 인식 문자 집합
Table. 1 Character set for recognition

Set	인식 언어 조합	Set	인식 언어 조합
0	숫자 전용	7	숫자+한글
1	영어 전용	8	영어+한글
2	숫자+영어	9	숫자+영어+한글
3	기호+숫자	10	기호+숫자+한글
4	기호+영어	11	숫자+영어+한글
5	기호+숫자+영어	12	기호+숫자+영어+한글
6	한글 전용		

표 2. 대표 문자(Set5)
Table. 2 Representative Characters(Set5)

No	대표 문자	유사 중복 문자				중복 구분
		숫자	기호	영어		
				소	대	
1	c			c	C	분할 문자의 상대 높이
2	l	1		l	I*	주변 인식문자 특성 단어단위 의미분석(*)
3	o	0		o	O	분할 문자의 상대 높이 주변 인식문자 특성
4	p			p	P	분할 문자의 상대 높이
5	s			s	S	
6	u			u	U	
7	v			v	V	
8	w			w	W	
9	x			x	X	
10	z			z	Z	
11	-		-/_			문자의 수직위치

표 3. Set5의 문자 예
Table. 3 Characters of Set5

구분	개수	문자 집합
숫자	10개	0 1 2 3 4 5 6 7 8 9
기호	31개	‘ ~ ! @ # $ % ^ & * ( ) - _ = + \ [{] } ; : ’ ” , . / < > ?
영어	52개	A B C D E F G H I J K L M N O P Q R S T U V W X Y Z a b c d e f g h i j k l m n o p q r s t u v w x y z
대표	11개	c l o p s u v w x z -
제외	13개	C I I 0 O P S U V W X Z _
Set5	80개 (93-13)	2 3 4 5 6 7 8 9 ‘ ~ ! @ # $ % ^ & * ( ) - = + \ [{] } ; : ’ ” , . / < > ? a b d e f g h i j k l m n o p q r s t u v w x y z A B D E F G H J K L M N Q R T Y

문자 인식기를 설계하기 위해서는 먼저 인식해야 할 문자 코드를 클래스로 정의하는 과정이 선행되어야 한다. 예컨대, 숫자 영상의 경우 0~9의 10개 클래스로 각각의 영상을 구분해야 한다. 이 클래스 집합을 인식 문자 집합이라 명명한다.

문자 인식에 있어서 고정된 문자 집합을 대상으로 설계된 인식기를 사용해서 문자를 인식할 수도 있지만 그 영역이 숫자만으로 되어 있거나 영어+숫자 조합 등 특정한 언어 조합으로 구성되어 있는 경우가 많다. 예컨대, ‘날짜’, ‘금액’에 관련된 영역 인식에서는 숫자 단일 문

자 집합이거나 숫자+기호 조합 문자 집합, '이름'에 해당하는 부분 영역은 한글 단일 문자 집합, '모델명'에 해당하는 인식 영역은 대부분 영어+숫자 조합 문자 집합이 인식의 대상이 되는 문자 집합이 된다. 이와 같은 특성을 가지는 인식 영역의 문자 인식을 위해서 각 영역별로 인식 문자 집합을 지정하여 해당되는 조합의 언어로만 학습된 인식기를 사용하게 되면 모든 언어가 조합된 완전한 인식기를 사용함으로써 발생할 수 있는 오류를 줄일 수 있게 된다. 따라서 언어별 조합으로 구성된 문자 집합을 필요한 수만큼 인식 문자 집합으로 정의한다. 본 논문에서는 표 1과 같이 인식 문자 집합을 13가지로 설정하였다.

여러 가지 언어로 조합된 인식 문자 집합의 경우 서로 다른 언어가 조합되어 있으므로 유사한 형태를 가지는 문자들이 공존하게 된다. 영어의 대소문자의 경우에도 형태가 유사한 문자들이 존재한다. 예컨대, 숫자와 영어가 조합되는 문자 집합에서는 숫자 '0'과 영어 'O'가 거의 유사한 형태를 가지게 된다. 이런 문자 영상들은 인식기 자체만으로는 구분할 수 없으며 형태는 유사하지만 클래스가 서로 다른 문자 영상들이 인식기의 학습에 사용되면 클래스를 구분하는 분류 평면(hyperplane)의 생성이 잘못될 가능성이 높다. 따라서 이런 유사 문자를 하나의 대표되는 문자로 정의함으로써 인식할 클래스의 수를 줄이고 인식기의 학습을 보다 용이하게 할 수 있다. 표 2는 인식의 대상이 되는 숫자와 기호, 영어 문자가 조합된 인식 문자 집합(Set5)에서의 대표 문자를 나타내며 표 3은 Set 5에 해당하는 문자의 예이다.

3.2. 인식기 구조 설계

본 논문에서는 다 클래스 다 언어 문자 인식을 위한 인식기 구조를 그림 1과 같이 소규모의 인식기를 다단계로 조합한 계층적 형태로 설계하였다.

구분할 클래스 수가 50개 미만인 경우에는 단일 인식기로서 그 구조를 단순화할 수 있고 신뢰성 있는 인식 성능을 보장할 수 있기 때문에 이런 소규모 단일 인식기를 조합하게 되면 다 클래스 다 언어 문자 집합을 효율적으로 인식할 수 있다. 이러한 다단계 인식 구조에서는 2개의 유형 분류 인식기로 한글 및 비한글을 구분할 수 있는 한글/비한글 구분 인식기(2-type classifier)와 한글인 경우 그 형태에 따른 6가지 형식으로 분류해주는 한글 유

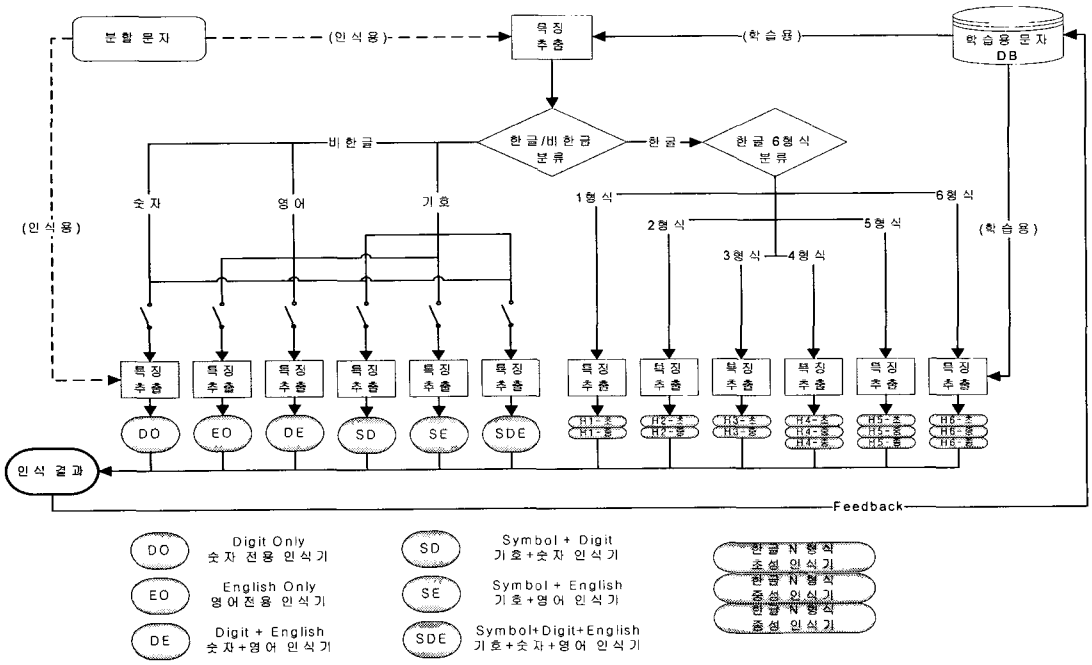

그림 1. 조합형 인식기 구조
Fig. 1 Structure of the combined classifier

형 분류 인식기(6-type classifier)가 있다. 유형 분류 인식기는 영상에 대한 직접적인 인식 결과를 산출하는 것이 아니라 유형만을 구분해주는 인식기이다.

한글/비한글 구분 인식기의 학습 방법은 전체 대상 문자를 비한글(class-0)과 한글(class-1)로 구분하여 목표 클래스(target class)로 설정하여 대상 문자들을 학습시킨다. 같은 방법으로 한글 유형 분류 인식기도 중성 모음의 조합 형태에 따라 한글을 6개의 목표 클래스로 설정하여 학습시킨다.

비한글인 경우 실제 문자 코드를 인식하는 인식기는 언어 조합에 따라 6개의 숫자 전용 인식기(DO), 영어 전용 인식기(EO), 숫자+영어 인식기(DE), 기호+숫자 인식기(SD), 기호+영어 인식기(SE), 기호+숫자+영어 인식기(SDE)가 있다. 한글의 경우에는 6가지 형식에 따라 각 형식별 초성 인식기, 중성 인식기, 종성 인식기로 구분하여 15개의 자소 인식기가 존재한다. 개별 인식기를 학습시키는 과정은 동일하다. 결과적으로 총 23개의 소규모 인식기를 적절히 선택하여 숫자, 기호, 영어, 한글이 조합된 다양한 경우의 인식 문자 집합을 인식할 수 있게 된다.

3.3. Delta-bar-delta 알고리즘

개별 소규모 인식기는 다양한 응용 분야에서 널리 사용되는 오류 역전파 알고리즘(BP)을 개선한 Delta-bar-delta 알고리즘을 사용하여 학습함으로써 생성한다.

오류 역전파 알고리즘(BP)은 순방향 다층 신경망의 학습에 널리 사용되는 대표적인 알고리즘이다. BP 알고리즘은 출력층 오차 신호를 이용하여 은닉층과 출력층간의 연결 강도를 변경하고, 이 오차 신호를 다시 은닉층에 역전파해서 은닉층과 입력층 사이의 연결 강도를 변경시키는 학습 방법을 사용한다. BP 알고리즘을 이용한 다층 신경망의 학습 절차는 다음과 같다[7].

Step 1. 입력층 패턴벡터 x, 은닉층 출력벡터 z, 출력층 출력벡터 y를 다음과 같이 표현한다.

$$x = [x_1, x_2, \dots, x_n] \quad (1)$$
$$z = [z_1, z_2, \dots, z_p]$$
$$y = [y_1, y_2, \dots, y_m]$$

입력층과 은닉층간의 연결강도 $v(p \times n)$, 은닉층과 출력층간의 연결강도 $w(m \times p)$를 임의의 작은 값으로 초기화하고 임의의 학습률 α를 설정한다.

Step 2. 학습 패턴 쌍을 차례로 입력하여 다음과 같이 은닉층의 가중합 NET_z 및 출력 z, 출력층의 가중합 NET_y 및 최종 출력 y를 구한다.

$$NET_z = xv^T \quad (2)$$

$$z = f(NET_z) = \frac{1}{1+\exp(-\neq T_z)} \quad (3)$$

$$NET_y = zw^T \quad (4)$$

$$y = f(NET_y) = \frac{1}{1+\exp(-NET_y)} \quad (5)$$

Step 3. 목표치 d와 최종 출력 y를 비교하여 오차를 구하고 출력층의 오차 신호 δ_y와 은닉층에 전파되는 오차 신호 δ_z를 구한다.

$$E = \frac{1}{2}(d-y)^2 \quad (6)$$

$$\delta_y = (d-y)y(1-y) \quad (7)$$

$$\delta_z = z(1-z)\sum_{i=1}^{m}\delta_y w \quad (8)$$

Step 4. 식 (9),(10)에 의해 k 학습 단계에서의 은닉층과 출력층간의 연결강도 변화량 Δw^k 및 입력층과 은닉층간의 연결강도 변화량 Δv^k 를 구하고 각각의 연결강도를 수정한다.

$$w^{k+1} = w^k + \Delta w^k = w^k + \alpha\delta_y z \quad (9)$$

$$v^{k+1} = v^k + \Delta v^k = v^k + \alpha\delta_z x \quad (10)$$

Step 5. 학습 패턴 쌍을 반복 입력하여 연결강도를 변경하며, 오차 E가 특정 범위 $E_{\max}$보다 적어지거나 지정된 반복회수를 수행하면 학습을 종료한다.

학습률 α와 오차신호 δ_z, δ_y에 의해서만 결정되는 연결강도 변화량 Δv, Δw에 이전 학습 단계에서의 연결강도 변화량을 보조적으로 활용하는 모멘텀(momentum) 알고리즘을 사용함으로써 학습 속도를 개

선시킬 수 있다. 모멘텀 알고리즘을 사용한 연결강도의 변화량은 다음과 같다. 여기서 η는 모멘텀 상수이다.

$$w^{k+1} = w^k + \alpha\delta_y z + \eta\Delta w^{k-1} \quad (11)$$

$$v^{k+1} = v^k + \alpha\delta_z x + \eta\Delta v^{k-1} \quad (12)$$

Delta-bar-delta($\Delta - \overline{\Delta}$) 학습 알고리즘은 각 학습 단계마다 연결강도의 변화에 따라서 학습률을 적응적으로 변경함으로써 학습 단계를 단축하고 궁극적으로 학습 시간을 효과적으로 감소시키는 방법이다. BP 알고리즘은 학습률 α에 의존하여 연결강도가 변하는 데[8], 일반적으로 여러 학습 단계에 걸쳐 연결강도가 계속 증가하거나 혹은 증가와 감소를 반복하는 경우가 발생한다. 만약, 연결강도가 계속 증가한다면 이것은 학습률 α가 너무 적기 때문에 연결강도를 적절히 변화시키는 데 상당히 많은 학습 단계가 요구될 것이라고 판단할 수 있으므로 학습률을 보다 큰 값으로 변경하여 학습 속도를 개선할 수 있다. 반면, 학습이 진행되면서 연결강도가 변화가 증가하거나 감소한다면 학습률이 너무 크기 때문에 연결강도가 적절히 변화되지 못한 것이라고 판단할 수 있으므로, 이 경우에는 반대로 학습률을 감소시킴으로써 학습 속도를 개선할 수 있다.

Delta-bar-Delta 알고리즘에서 Δ와 $\overline{\Delta}$는 다음과 같이 정의된다.

$$\Delta_w^k \equiv -\delta_y z^k \quad (13)$$

$$\overline{\Delta_w^k} \equiv (1-\beta)\Delta_w^k + \beta\overline{\Delta_w^{k-1}} \quad (14)$$

$$\Delta_v^k \equiv -\delta_z x^k \quad (15)$$

$$\overline{\Delta_v^k} \equiv (1-\beta)\Delta_v^k + \beta\overline{\Delta_v^{k-1}} \quad (16)$$

위와 같이 정의한 Δ와 $\overline{\Delta}$을 이용하여 $k+1$ 단계에서의 새로운 학습률 α^{k+1}은 다음과 같이 구한다.

$$\alpha^{k+1} = \begin{cases} \alpha^k + \kappa & ; \overline{\Delta^{k-1}} \bullet \Delta^k > 0 \\ (1-\gamma)\alpha^k & ; \overline{\Delta^{k-1}} \bullet \Delta^k < 0 \\ \alpha^k & ; \overline{\Delta^{k-1}} \bullet \Delta^k = 0 \end{cases} \quad (17)$$

여기서, β, κ, γ 는 임의의 상수이며 학습률은 Δ와 $\overline{\Delta}$가 같은 부호를 가지면 κ만큼 증가시키고 다른 부호를 가지면 $(1-\gamma)$만큼 감소시키는 방법으로 변경된다.

제안된 신경망에서의 연결강도의 변화는 다음과 같다.

$$w^{k+1} = w^k + \alpha^k\delta_y z + \eta\Delta w^{k-1} \quad (18)$$

$$v^{k+1} = v^k + \alpha^k\delta_z x + \eta\Delta v^{k-1} \quad (19)$$

Ⅳ. 실험 및 결과

문자 인식기에 대한 성능 평가는 인식기를 생성하기 위해 구축한 데이터베이스의 문자 153,792개를 대상으로 한글/비한글 구분 성능, 한글에 대한 6형식 분류 성능, 그리고 다양한 언어 조합에 따른 13가지 인식 문자 집합별 인식률을 조사하였다. 사용된 문자 영상은 잡영이 없는 깨끗한 문자 이미지이다.

4.1. 한글/비한글 분류 (2-type classification)

한글인지 비한글인지를 구분하는 유형 분류 인식기에서 대상 문자 집합은 총 36292개(학습24195개+테스트12097개)의 비한글 문자와 117500개(학습78334개+테스트39166개)의 한글 문자이다. 테스트 셋에 대한 분류 성능은 99.96%의 만족할 만한 분류 성능을 나타낸다. 한글은 영어나 숫자 등에 비해 대부분 그 형태가 복잡한 편이며 오류가 나는 문자들은 대부분 한글의 모양이 단순하여 비한글로 인식되는 경우이거나 그 반대로 영어나 숫자, 기호 등이 글꼴에 따라 한글과 유사한 경우가 대표적이다. 전체 분류 결과 비한글에 대한 신뢰도(99.85%)가 한글(99.99%)에 대한 신뢰도보다 약간 낮게 나타났다. 즉, 한글이 아님에도 불구하고 한글이라고 인식하는 경우가 그 반대인 한글을 비한글로 인식하는 경우보다 높다는 의미이다.

4.2. 한글 형식 유형 분류 (6-type classification)

한글에 대한 6가지 유형의 형식 분류기의 대상 문자 집합은 총 117500개의 한글 문자만을 대상으로 하며 형식별로 각각 학습셋과 테스트셋으로 나누었으며 그 성능은 99.90%로 매우 우수하게 산출되었다. 형식별로는 4형식에서 오류가 가장 많이 나타났다. 실제 4형식 문자

표 4. 한글/비한글 구분 성능 비교
Table. 4 Performance comparison of Korean/Non-korean classification

		비한글	한글	정인식	오인식	계	정확도
학습셋	비한글	24195	0	24195	0	24195	**100.00%**
	한글	23	78311	78311	23	78334	**99.97%**
	정분류	24195	78311	-	-	102506	-
	오분류	23	0	-	-	23	-
	계	24218	78311	102506	23	102529	**99.98%**
	신뢰도	**99.9%**	**100.0%**	-	-	**99.98%**	-
테스트셋	비한글	12094	3	12094	3	12097	**99.98%**
	한글	18	39148	39148	18	39166	**99.95%**
	정분류	12094	39148	-	-	51242	-
	오분류	18	3	-	-	21	-
	계	12112	39151	51242	21	51263	**99.96%**
	신뢰도	**99.85%**	**99.99%**	-	-	**99.96%**	-

가 다른 형식에 문자에 비해 월등히 많기 때문이기도 하지만 4형식인 문자의 형태가 5형식 또는 6형식인 문자와 비슷한 것들이 많이 때문에 발생하는 오류라고 판단된다.

4.3. 인식 문자 집합별 인식률

표 4에서 가로 항목은 해당하는 언어를 나타내며 세로 항목은 인식하기 위해 설정한 언어 집합을 나타낸다. 예를 들어, 영어인 문자들을 SET2 언어셋(숫자+영어)로 설정하여 인식한 경우 전체 24874개 중 24781개가 정인식, 나머지 93개가 오인식되어 99.63%의 인식률이 산출되었음을 나타낸다. 한글인 경우 SET6(한글 전용)으로 언어 집합을 설정하여 인식하는 경우가 그렇지 않은 경우보다 인식률이 좋게 나타남을 알 수 있다. 이것은 한글임에도 불구하고 한글이 아닌 다른 언어로 사전에 분류되는 경우가 있기 때문이다. 모든 언어의 문자들을 대상으로 모든 언어를 인식할 수 있는 SET12 언어 집합으로

표 5. 한글 형식 분류 성능 비교
Table. 5 Performance comparison of Korean character type classification

		1형식	2형식	3형식	4형식	5형식	6형식	정인식	오인식	계	정확도
학습셋	1형식	4967	0	0	0	0	0	4967	0	4967	**100.00%**
	2형식	0	3034	0	0	0	0	3034	0	3034	**100.00%**
	3형식	0	0	3634	0	0	0	3634	0	3634	**100.00%**
	4형식	0	0	0	35634	0	0	35634	0	35634	**100.00%**
	5형식	0	0	0	0	19500	0	19500	0	19500	**100.00%**
	6형식	0	0	0	0		11567	11567	0	11567	**100.00%**
	정분류	4967	3034	3634	35634	19500	11567	-	-	78336	-
	오분류	0	0	0	0	0	0	-	-	0	-
	계	4967	3034	3634	35634	19500	11567	78336	0	78336	**100.00%**
	신뢰도	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	**100.0%**	-	-	**100.0%**	-
테스트셋	1형식	2480	0	3	0	0	0	2480	3	2483	**99.88%**
	2형식	0	1516	0	0	0	0	1516	0	1516	**100.00%**
	3형식	5	0	1811	0	0	0	1811	5	1816	**99.72%**
	4형식	0	0	0	17810	0	6	17810	6	17816	**99.97%**
	5형식	0	1	0	1	9748	0	9748	2	9750	**99.98%**
	6형식	0	0	0	22		5761	5761	22	5783	**99.62%**
	정분류	2480	1516	1811	17810	9748	5761	-	-	39126	-
	오분류	5	1	3	23	0	6	-	-	38	-
	계	2485	1517	1814	17833	9748	5767	39126	38	39164	**99.90%**
	신뢰도	**99.80%**	**99.93%**	**99.83%**	**99.87%**	**100.00%**	**99.90%**	-	-	**99.90%**	-

표 6. 인식 문지 집합별 인식 성능 비교
Table. 6 Performance comparison according to the set of recognition language

클래스 \ 데이터		2Type 분류기	6Type 분류기	기호+숫자+영어				한글			전체
				기호	숫자	영어	전체	Gray	Bin	전체	
기본문자집합	SET0 숫자	×	×	-	99.94% 6389 /4	-	-	-	-	-	-
	SET1 영어	×	×	-	-	99.63% 24781 /93	-	-	-	-	-
	SET2 숫자+영어	×	×	-	99.55% 6364/ 29	99.59% 24773 /101	-	-	-	-	-
	SET3 기호+숫자	×	×	100% 5025 /0	99.84% 6383 /10	-	-	-	-	-	-
	SET4 기호+영어	×	×	99.94% 5022 /3	-	99.34% 24711 /163	-	-	-	-	-
	SET5 기호+숫자+영어	×	×	99.52% 5001 /24	99.58% 6366 /27	99.61% 24778 /96	99.59% 36145 /147	-	-	-	-
	SET6 한글전용	×	○	-	-	-	-	99.33% 70029 /471	99.46% 46745 /255	99.38% 116774 /726	-
확장문자집합	SET7 숫자+한글	○	○	-	99.92% 6388 /5	-	-	99.29% 70001 /499	99.44% 46735 /265	99.35% 116736 /764	99.41% 152878 /914
	SET8 영어+한글	○	○	-	-	99.62% 24779/ 95	-				
	SET9 숫자+영어+한글	○	○	-	99.53% 6363 /30	99.59% 24771 /103	-				
	SET10 기호+숫자+한글	○	○	100% 5025 /0	99.83% 6382 /11	-	-				
	SET11 기호+영어+한글	○	○	99.94% 5022 /3	-	99.34% 24709 /165	-				
	SET12 기호+숫자+영어+한글	○	○	99.52% 5001 /24	99.56% 6365 /28	99.61% 24776 /98	99.59% 36142 /150				
총 문자 개수				5025	6393	24874	36292	70500	47000	117500	153792

설정하여 인식한 경우 총 문자 153792개 중 152878개를 정인식하여 99.41%의 높은 인식 성능을 나타냄을 알 수 있다. 하지만 이러한 결과는 잡영이 비교적 적은 문자 영상을 대상으로 테스트 한 것이므로 실제적으로 다양하게 발생할 수 있는 획 사이의 잡영, 문자의 획 간 접합으로 인한 분할 오류, 다양한 글꼴에 따른 형태 왜곡 등이 발생할 경우 인식률이 더 낮아질 소지가 있다.

V. 결론 및 향후 연구 계획

본 논문에서는 효과적인 형식 문서 자동 인식 시스템 구현을 위한 전반적인 처리 과정과 다중 클래스, 다중 언어 집합의 문자 영상을 효율적으로 인식하기 위한 다단계 인식기 구조를 제안하였으며, 약 15만개의 개별 문자를 대상으로 그 인식 성능을 살펴본 결과 99%의 신뢰성

있는 결과를 획득함으로써 실제적으로 형식 문서 자동 인식 시스템에 응용하기 위한 유용성과 신뢰성을 증명하였다.

추후 보완되어야 할 사항으로는 잡영이 포함되어 있는 영상 처리 및 고해상도 대용량 이미지 처리에 따른 전처리 속도 개선에 대한 연구, 테두리와 겹쳐진 문자의 분리 및 복원에 관한 연구, 혼합 언어 집합인 경우 한글 '이'와 영어 'OI'와 같은 자모음 통합에 대한 후처리에 관한 연구가 필요하다. 그리고 다양한 유형의 문자 영상을 확보하여 개별 인식기의 성능을 지속적으로 갱신할 수 있도록 하는 것이 무엇보다 중요하다.

참고 문헌

[1] 안창, 이상범, "한글처리-문자 중심 인식 기술 고찰", 한국정보처리학회 논문지, 제5권, 제5호, pp.48-53, 1998년 9월

[2] 김진숙, 변영철, 김경환, 최영우, 이일병, "지로 서식 문서의 인쇄체 숫자 인식", 한국정보과학회, 한국정보과학회 학술발표논문집 한국정보과학회 1999년도 가을 학술발표논문집 제26권 제2호(II), 1999. 10, pp.446~448

[3] 김우성, 방승양, "신경회로망을 이용한 한글 한자 혼용 문서 인식기의 개발 ", 한국정보과학회, 한국정보과학회 학술발표논문집 한국정보과학회 1992년도 봄 학술발표논문집 제19권 제1호, 1992. 4, pp.677~680

[4] 심원태, 김진형, "혼합형 제어 전략을 사용한 인쇄체 한글 문자의 인식", 한국정보과학회, 한국정보과학회 학술발표논문집 한국정보과학회 1987년도 가을 학술발표논문집 제14권 제2호, 1987. 10, pp. 159~162

[5] 임길택, 김호연, "문자형식 분류 기반의 인쇄체 문자 인식에 관한 연구", 대한전자공학회, 전자공학회논문지-CI 전자공학회논문지 제40권 CI편 제5호, 2003. 9, pp.26~39

[6] 배규찬, 박형민, 오상훈, 최용선, 이수영, "SVM기반의 선택적 주의집중을 이용한 중첩 패턴 인식", 대한전자공학회, 전자공학회논문지-SP 전자공학회논문지 제42권 SP편 제5호, 2005. 9, pp.123~ 136

[7] 오창석, '뉴로컴퓨터', 내하출판사, 2000.

[8] 김광백, 박충식, "퍼지 제어 시스템을 이용한 학습률 자동 조정 방법에 의한 개선된 역전파 알고리즘", 한국해양정보통신학회논문지, 제8권, 2호, pp.464-470, 2004.

저자소개

김 도 현(Do-Hyeon Kim)

2001년 부산대학교 전자계산학과 졸업
2003년 부산대학교 전자계산학과 석사 졸업
2006년 부산대학교 컴퓨터공학과 박사 수료
2007년 부산대학교 컴퓨터공학과 박사 재학 중
※관심분야: 패턴인식, 영상처리 및 컴퓨터비젼, 퍼지 및 신경망, 제어자동화시스템.

박 재 현(Jae-Hyeon Park)

2001년 부산대학교 영상정보공학 석사 졸업
2003년 부산대학교 전자계산학과 박사 수료
※ 관심분야: 객체 추적, 영상처리 및 컴퓨터비젼

김 철 기(Cheol-Ki Kim)

1999년 부산대학교 전자계산학과 졸업
2001년 부산대학교 전자계산학과 석사 졸업
2003년 부산대학교 전자계산학과 박사 졸업
2003년 3월~2006년2월 밀양대학교 컴퓨터공학과 조교수
2006년3월~현재 부산대학교 디자인학과 조교수
※ 관심분야: 영상처리, 멀티미디어, 패턴분석, 색채분석

차 의 영(Eui-Young Cha)

1979년 경북대학교 전자공학과 졸업.
1982년 서울대학교 전자계산학과 석사 졸업.
1998년 서울대학교 컴퓨터공학과 박사졸업.
1981년~1985년 한국전자기술연구소 연구원.
1995년~1996년 University of London 방문교수.
1985년~현재 부산대학교 컴퓨터공학과 교수
※ 관심분야: 컴퓨터비젼, 신경망, 웨이블릿