論 文

인쇄체 한글 및 한자의 인식에 관한 연구

正會員 金 晶 雨* 正會員 李 幸 世*

A Study on the Printed Korean and Chinese Character Recognition

Jeong Woo Kim* Haing Sei Lee* *Regular Members*

要 約

본 논문에서는 한자를 포함하는 한글 문서 인식을 위한 인쇄체 한글, 한자의 구분과 인식 방법에 대하여 연구하였다. 제안된 한글, 한자 구분 방법은 한글의 수직모음과 수평모음의 구조적 특징을 이용하였다. 한글은 6가지 형태로 분류하고 분류된 각 형태에 대하여 세선화 과정을 거치지않고 모음 우선추출에 의한 자모분리를 행하고 분리된 자음에 대하여 변형된 교차거리 특징을 이용하여 인식하였다. 한자에 대해서는 획교차수의 평균치를 이용하여 전체 한자 대상문자에 대해 분류를 하였으며, 문자의 획교차수와 흑점비율 특징을 이용하여 인식하였다. 한글과 한자의 구분에서는 90.5%의 분류율을 얻었다. 한글인식에 있어서는 대상문자 명조체 2512자에 대하여 90.0%의 형태 분류율을 얻었다. 인식 결과 실험 데이타 1278자에 대하여 92.2%의 인식율을 얻었다. 한자인식에 있어서는 대상문자 4585자에 대하여 분류한 결과 최대밀집 구간은 124자로서 약 1/40 정도로 분류되었음을 알 수 있었고, 인식실험 결과 89.2%의 인식율을 얻었다.

ABSTRACT

A new classification method and recognition algorithms for printed Korean and Chinese character is studied for Korean text which contains both Korean and Chinese characters. The proposed method utilizes structural features of the vertical and horizontal vowel in Korean character. Korean characters are classified into 6 groups. Vowel and consonant are separated by means of different vowel extraction methods applied to each group. Time consumming thinning process is excluded. A modified crossing distance feature is measured to recognize extracted consonant. For Chinese character, an average of stroke crossing number is calculated on every characters, which allows the characters to be classified into several groups. A recognition process is then followed in terms of the stroke crossing number and the black dot rate of character. Classification between Korean and Chinese character was at the rate of 90.5%, and classification rate of Ming-style 2512 Korean characters was 90.0%. The recognition algorithm was applied on 1278 characters. The recognition rate was 92.2%. The densest class after classification of 4585 Chinese characters was found to contain only 124 characters, only 1/40 of total numbers. The recognition rate was 89.2%.

* 건제전문대학
論文番號 : 92-117 (接受1992. 2. 25)

Ⅰ. 서 론

최근 컴퓨터의 사용이 확대됨에 따른 정보처리(information processing)나 컴퓨터와 인간의 맨-머신 인터페이스(man-machine interface)가 사회적으로 관심을 끌고 있으며 사무 자동화를 위한 정보량의 증가에 따른 방대한 양의 문서 데이타를 작성, 편집, 처리하는데 있어서 새로운 입력수단으로 인간의 수작업이 아닌 컴퓨터 스스로 문서의 자동 입력과 인식에 대한 문제가 절실히 요청되고 있다. 지금까지 한글에 대해서는 많은 연구가 되어져 왔으나, 일반 문서나 신문에서 한글과 함께 사용하고있는 한자에 대한 연구는 활발하지 못하였다. 이러한 일반 문서나 신문과 같은 문서 데이타를 처리하기 위해서는 한글과 함께 한자를 인식할 수 있는 연구가 필요하게 되었다.

한글 문자인식에 관한 연구는 1960년대 이후 현재까지 꾸준하게 연구되어져 오고 있으며[1,2,3] 최근에는 상당한 수준에까지 도달하고있다. 입력패턴으로부터 문자의 기본요소(primitive)들을 찾아내어 이들의 구조분석에 의해 인식하는 구조적 방법은 세선화 과정을 거쳐 인식하는 방법[2,4,5]과 세선화 과정을 거치지않고 문자패턴의 형상이나 윤곽선에 의하여 인식하는 방법[6,7]으로 나누어져 연구되었다.

한자인식은 1960년대에 처음 시작되었으나 1970년대 이후 한자를 많이 사용하는 일본에서 활발하게 진행되어 상당한 연구가 이루어졌다. 인쇄체 한자인식에 관한 첫 논문은 1966년 템플리트 매칭법을 이용하여 R.Casey 와 G.Nagy에 의해 발표되었으며[8] P. P. Wang[9]은 한자의 63개 부수에 대한 세 가지 변환(Fourier, Hadamard, Rapid)의 영향을 연구하였다. 계층적 매칭법은 Yamamoto[10]에 의해 개발되었다. Umeda[11]는 다중 문자체 한자에 대한 mesh와 주변패턴의 조합을 이용한 일차분류를 발표하였고 P. Chen[12] 등은 다중 문자체 한자에 대하여 복잡도를 이용한 일종의 획관계 코드를 정의하여 인식을하였다.

본 논문에서는 인쇄체 문자로 제한된 한글과 한자가 함께 존재하는 문서의 인식을 위해서 세선화 과정을 거치지않고 한글과 한자의 구조적 특징과 한글의 모음형태를 사용하여 한글과 한자를 분류하고, 분류된 한글은 구조적 방법으로 모음 우선 추출을 통한 자모분리 및 인식을 행하고, 한자는 획교차수의 통계적 성질을 이용하여 계층적인 분류를 하여 결정론적 방법에 의하여 인식하고자 한다.

Ⅱ. 한글과 한자 분류

한글과 한자가 포함되어 있는 문서로는 대표적인 것이 신문이다. 이러한 문서들을 인식하기 위해서는 한글과 한자를 분류하여야 한다. 한글과 한자를 분류하는데는 주로 한글과 한자의 구조적 특징의 차이를 이용한다.[13] 한글 모음의 위치와 크기를 기준으로 한 한글과 한자의 구조적 차이에 대하여 살펴보면 다음과 같으며 그림 1에 이러한 예를 나타내었다.

(특징1) : 한글은 긴 수평모음이 존재하는 경우 긴 수평모음은 1개만이 존재하고 그 수평모음의 위치는 문자 상단에는 나타나지 않는다.

(특징2) : 긴 수직모음이 존재하는 경우 한글의 긴 수직모음은 문자의 가로크기의 1/2 왼쪽 부분에는 나타나지 않는다.

(특징3) : 한글은 긴 수평모음이 존재하지 않는 경우 반드시 수직모음이 존재한다. 이때 수직모음이 존재하는 경우 가로 방향의 첫 번째 화소의 위치가 한글은 문자의 가로크기의 1/2 오른쪽 부분에서 수직모음부의 첫 화소가 존재한다.

이와같은 차이를 이용하여 본 논문에서는 먼저 한글의 긴 수평모음과 긴 수직모음의 성질에 의해 1차 분류를 행하고, 만약 1차분류에서 한자로 분류되지 않는것 중에서 수직성분이 존재하는것은 가장 오른쪽에 위치한 수직모음에 대하여 한글의 수직모음의

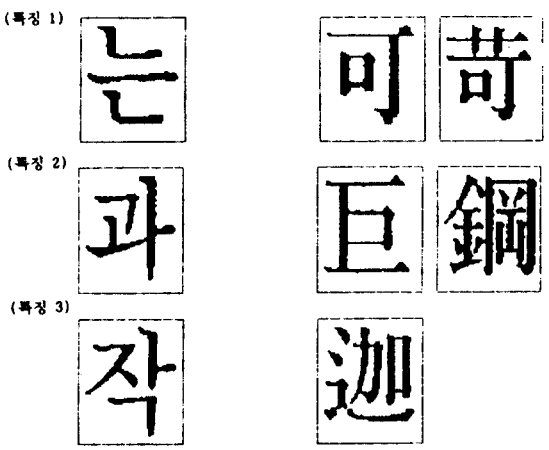


그림 1. 한글과 한자의 구조적 특징.

Fig. 1. Structural feature in Korean and Chinese character.

위치와 수직모음을 중심으로한 연결성분의 추출을 통한 획 성분의 존재(ㅏ, ㅑ, ㅓ, ㅕ, ㅣ의 형태)를 검사하여 다시 한글과 한자를 분류하였다. 그림 2는 한글과 한자를 분류하는 흐름도를 나타낸 것이다.

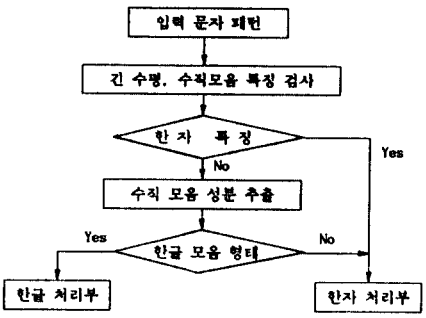


그림 2. 한글과 한자의 분류 흐름도.
Fig. 2. Flowchart for classification of Korean and Chinese characters.

Ⅲ. 한글 인식

입력된 한글문서로부터 한글로 분류된 문자패턴에 대하여 6가지 형태[1]로 분류하고 각 형태에 따라 모음의 위치와 형태를 이용하여 모음성분을 먼저 추출하여 모음인식 및 자모분리를 행하고, 자음의 구조적 특징을 이용하여 자음을 인식하는 과정을 통하여 한글을 인식하도록 하였다. 그림 3은 한글인식 과정에 대한 전체 흐름도를 나타낸 것이다.

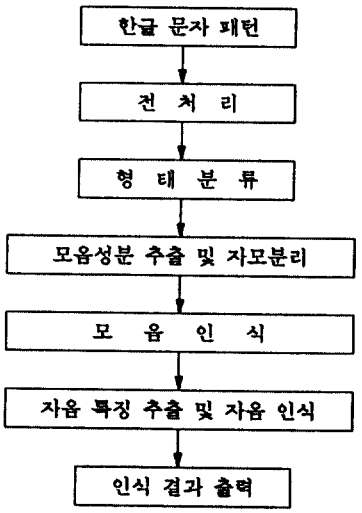


그림 3. 한글 인식 흐름도.
Fig. 3. Flowchart of Korean character recognition.

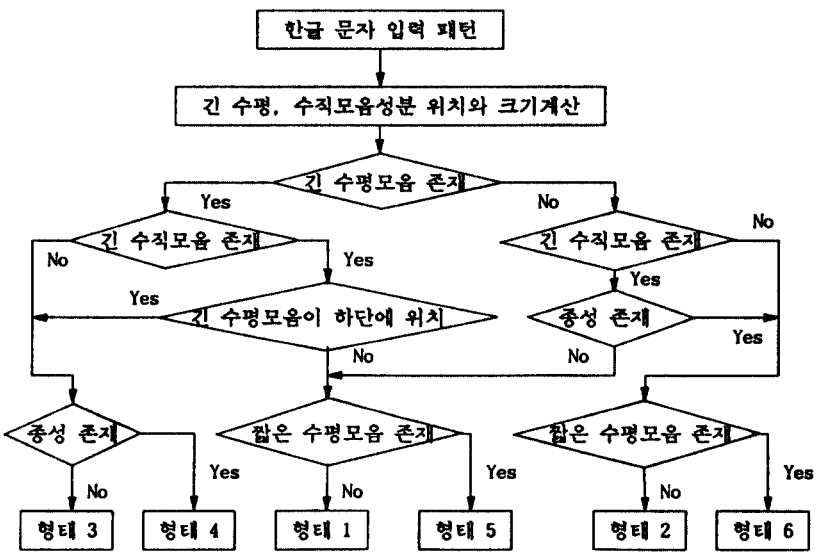


그림 4. 한글 형태 분류 흐름도.
Fig. 4. Flowchart for Korean character type classification.

1. 형태 분류

한글의 모음은 수평모음과 수직모음으로 나눌 수 있는데, 수평모음 중에서 긴 수평 모음은 크기의 변화는 심하지 않으나 위치의 변화가 있으며, 수직모음은 반대로 위치의 변화는 심하지 않으나 크기의 변화가 종성의 존재 여부에 따라 달라지는 성질이 있다. 이러한 모음의 성질을 이용하여 우선 긴 수직모음과 긴 수평모음의 존재 유무와 위치를 조사하여 3가지 경우로 먼저 대분류를 행한다. 이때 긴 수직모음과 긴 수평모음의 존재는 수직, 수평 방향의 연속된 화소들의 크기를 검사하여 문자의 세로크기의 90% 이상이 되는 경우에 긴 수직모음, 가로크기의 95% 이상이 되는 경우 긴 수평모음으로 간주하였다. 대분류가 행해지면 형태 1과 5의 제 1 그룹, 형태 3과 4의 제 2그룹, 형태 2와 6의 제 3 그룹으로 나누어지게 된다. 제 1 그룹과 제 3 그룹의 경우는 긴 수직모음과 종성 존재여부에 의해 형태 1과 형태 5 또는 형태 2와 형태 6을 나누는데 종성 존재여부에 따른 수직모음의 성분을 먼저 추출하여 원 패턴에서 제거한 뒤 남은 패턴으로부터 가장 왼쪽끝에 있는 화소의 위치를 중심으로 우측방향으로 연속된 화소들에 대한 크기와 상하 방향의 연속화소의 추출에 위해 짧은 수평모음 성분의 존재여부를 판단하게된다. 제 2 그룹의 경우는 수평모음을 추출, 수평모음(ㅗ, ㅛ, ㅜ, ㅠ, ㅡ)의 특성을 검사하여 종성존재 여부에 따라 형태 3과 형태 4로 나누게 된다. 그림 4는 한글형태 분류의 흐름도를 나타낸것이다.

2. 자모분리 및 모음인식

각 형태에 따른 수직 또는 수평모음의 위치를 중심으로 연속된 패턴을 추출하고 모음 성분이 아닌 부분을 제거하여 추출된 패턴에 대해 모음인식을 하고, 자모분리를 하였다. 자모분리와 모음 추출 및 인식의 과정은 다음과 같다. 또한 그림 5에 분류 형태중 가장 복잡한 형태에 대하여 예를 나타내었다.

(1)수직모음이 존재하면 수직모음은 대개 문자 가로크기의 1/2 우반부에 존재하므로 이 영역에서 수직모음 성분을 검사하여 단모음과 복모음을 구별한다.

①형태 1과 5의 경우는 문자의 세로크기에 대하여 그리고 형태 2와 6의 경우는 종성 받침이 존재하므로 실제 수직모음부까지의 문자 세로크기를 찾아 수직모음 성분을 조사한다.

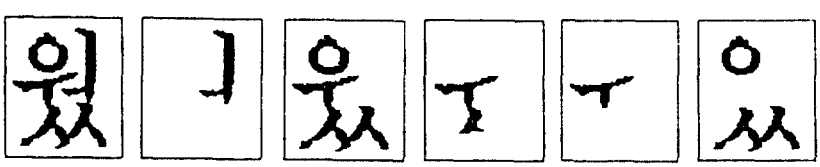
②조사된 성분들중 연속된 성분들은 하나로 간주하여 수직모음 갯수를 조사한다.

③조사된 수직모음 갯수가 1개 이면 단모음, 2개 이상이면 복모음으로 판정한다.

단모음인 경우는 가장 오른쪽 수직모음성분의 최대크기를 갖는 위치를 중심으로 좌우 방향으로 연속된 화소들을 추출하고 복모음의 경우는 가장 오른쪽 수직모음을 제거한 부분에서 다시 가장 오른쪽 수직모음 위치를 찾아내어 그 수직모음의 위치를 중심으로 마찬가지로 좌우 방향으로 연속된 화소들을 추출한다 (그림 5(b)).

추출된 형태로부터 수직모음 형태를 추정하여 추정된 형태에 따라 기준 위치로부터 왼쪽 방향으로의 연속 화소들의 크기를 계산하여 작은 곁줄기 성분의 선폭을 계산한다. 곁줄기 성분에 대하여 추가로 상하 방향의 연속된 화소들을 추출하여 선폭의 2배 이상이 되는 부분을 자음성분으로 간주하여 제거하게 된다. 추출된 형태에 의해 수직모음을 인식하고 원 패턴에서 제거한다(그림 5(c)).

(2)수평모음이 존재하면 최대크기를 갖는 위치를 중심으로 상,하 방향으로 연속된 화소들을 추출(그림 5 (d))하여 곁줄기 획성분에대한 선폭을 계산하고 각 수평모음의 형태에 따라 다시 좌우 방향의 연속화소를 추출하여 과정 (1)에서와 마찬가지로 초성 또는 종성 자음의 접촉부를 제거한 다음 수평모음을 인식



(a) (b) (c) (d) (e) (f)

그림 5. 모음 추출 예.
Fig. 5. Example of vowel extraction.

하고 원 패턴에서 제거한다.(그림 5 (e))

(3)위의 과정을 거치고나면 남는 패턴은 초성자음 또는 초성자음과 종성자음이 남게되어 자연스럽게 자모분리가 이루어지며, 남은 패턴은 자음인식과정으로 입력되게 된다 (그림 5 (f)).

3. 자음 인식

원 문자패턴으로부터 모음부로 판정된 부분을 제거하고나면 자음부분만 남게된다. 형태 1, 3, 5인 경우는 초성자음 부분만 남게되고, 형태 2, 4, 6인 경우는 초성과 종성 자음부분이 남게된다. 받침이 존재하는 경우 형태 2를 제외하고는 횡방향으로의 검색만으로도 자연스럽게 분리되어지나 형태 2의 경우는 초성부와 종성부가 서로 어긋나게 위치하는 경우에는 분리되지 않는다. 그러므로 형태 2의 경우 횡방향 검색으로 분리되지않는 경우에는 연속 화소 블럭의 추출과 초성, 종성의 분포위치를 이용하여 초성부분과 종성부분을 분리한다.

분리된 각 자음 패턴에 대하여 패턴을 둘러싸는 최소사각형에서 자음특징을 추출하게된다. 첫번째는 수직과 수평방향으로 연속된 흑점의 크기를 검사하여 설정한 문턱치를 넘는 행과 열의 갯수를 구한다. 이때 연속된 행과 열은 하나의 획 성분으로서 간주한다. 두번째는 문자를 둘러싸는 최소 외각틀에서부터 문자의 화소값이 처음 1이 될때까지의 거리값인 제 1 교차거리 특징을 구한다.

임의의 M×N 행렬 형태의 문자패턴에서 왼쪽에서 오른쪽, 아래에서 윗쪽, 오른쪽에서 왼쪽, 위에서 아래쪽으로의 4 방향에 대하여 구한 교차거리 특징을 각각 $L_m(i)$, $B_n(j)$, $R_m(i)$, $U_n(j)(0\le i<M-1,\ 0\le j<N-1)$이라 하자.

이와같이 구하여진 교차거리 특징에 대하여 3개의 문턱값을 정하여 문자패턴에 대한 주변특징으로 대치한다.

왼쪽에서 오른쪽 방향으로의 특징을 구하면

$$L_m(i)=\begin{cases}0 & L_m(i)<M/4\\ 1 & M/4\le L_m(i)<M/2\\ 2 & M/2\le L_m(i)\end{cases} \qquad (1)$$

으로된다. 마찬가지로 나머지 방향에 대해서도 같은 방법에 의해 특징을 추출할 수 있다. 그림 6에는 'ㄱ' 패턴에 대한 4 방향 특징값(예:02,20,0,0)을 나타내었다.

인식 과정은 수직과 수평 방향에 대한 획 성분의 갯수와 위치에 의해 후보자음이 하나로 결정되면 인식을 마치고, 후보자음이 여러개인 경우에는 4 방향에 대한 제 1교차거리 특징의 코드값을 비교하여 자음을 결정하게된다. 예를들어 'ㄱ'과 'ㅈ'의 경우와 같은 초성자음 패턴은 수평획의 갯수는 1, 수직획 갯수는 0이고 획성분의 위치는 최소사각형의 상단에 위치하는 점에서 유사하나 오른쪽에서 왼쪽방향으로의 교차거리 특징 코드값을 비교하면 구분할 수 있을것이다.

이와같이 자음이 인식되면 먼저 인식된 모음과 조합하여 하나의 한글문자를 인식하는 과정을 마치게 된다.

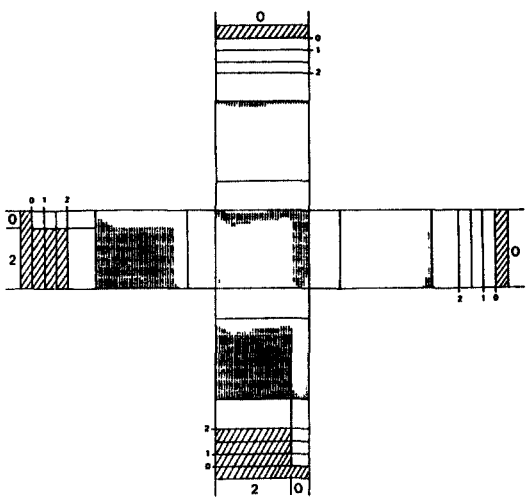


그림 6. 자음 특징 추출 예.
Fig. 6. Example of consonant feature extraction.

Ⅳ. 한자 인식

획교차수는 문자를 포함하는 최소 사각형에 수직으로 가로와 세로 방향으로 문자를 통과하는 직선을 각각 그은 경우 그 직선이 문자의 획과 교차하는 횟수로 정의한다. 이것은 문자의 복잡도에 관한 척도가 되며 그래프의 위상 기하학적 성질을 가진다. 따라서 이론적으로는 문자의 이동, 회전, 스케일링에 대하여 무관한 처리를 할 수 있다[7,14,15]. 또한 획교차수를 이용하면 문자 데이타의 전처리 과정인 세선화 과정이 필요없어 처리과정이 단순해지는 장점을 갖는다.

1. 분 류

스캐너를 통해 0과 1로 이루어진 입력 한자패턴이 M×N 행렬 형태라 한다면 각 행과 열에 대하여 0에서 1로 변화하는 횟수를 계산하여 얻은 획교차수를 fr_row, fr_col이라하면 가로방향과 세로방향의 획교차수 평균값은 다음과 같이 구하여진다.

$$mf\_row = \sum_{i=0}^{M-1} fr\_row(i) / M \quad (2)$$

$$mf\_col = \sum_{i=0}^{N-1} fr\_col(i) / N \quad (3)$$

다음으로 두 방향에서의 획교차수 평균값중 최대값과 최소값을 찾아 분류하고자 하는 최대구간을 정하고 다음과 같은 관계식에 의하여 입력된 문자의 해당구간을 구한다.

$$l_1 = [\{(\text{최대구간}-1) / (r\_max - r\_min)\} \times (mf\_row - r\_min)] \quad (4)$$

$$l_2 = [\{(\text{최대구간}-1) / (c\_max - c\_min)\} \times (mf\_col - c\_min)] \quad (5)$$

여기서 $0 \le l_1, l_2 <$ 최대구간, l_1, l_2 : 정수

l_1 : 해당하는 행의 위치

l_2 : 해당하는 열의 위치

r_max : 가로방향 획교차수 평균치중 최대값

r_min : 가로방향 획교차수 평균치중 최소값

c_max : 세로방향 획교차수 평균치중 최대값

c_min : 세로방향 획교차수 평균치중 최소값

이와같은 분류과정을 기준으로 정한 한자패턴에 대해 처리하여 같은 구간에 속한 문자끼리 그룹을 만들어 데이타 베이스를 구축한다. 또한 인식을 위해 입력되는 임의의 한자 문자패턴에 대해서도 같은 과정을 거쳐 해당구간을 결정하게된다. 그림7에 분류과정의 블럭도를 나타내었다

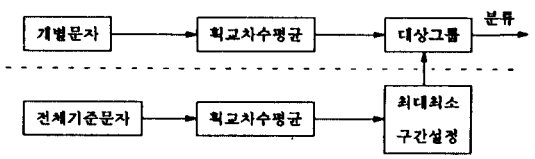


그림 7. 분류 과정 블럭도.
Fig. 7. Block diagram of the classification process.

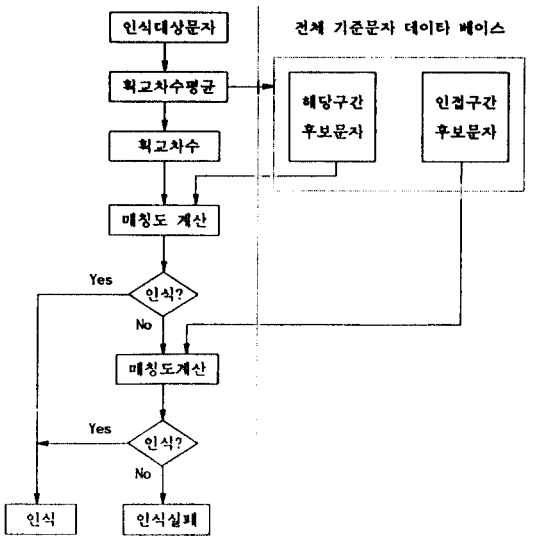


그림 8. 인식 과정 흐름도.
Fig. 8. Flowchart of the recognition process.

2. 인 식

인식 과정은 두 단계로 이루어진다. 첫단계는 입력된 임의의 문자에 대한 획교차수 평균치를 구한후 데이타베이스에서 인식대상그룹을 선정하고, 그 그룹내의 기준문자와 인식대상 문자 사이에 문자를 포함하는 최소사각형의 크기, 흑점비율, 최대빈도 획교차수 등의 특징들을 비교해 이들이 설정한 문턱값을 넘는 것들만 택하여 인식대상의 후보문자수를 줄인다. 두번째 단계에서는 후보문자들에 대해 모든 획교차수를 비교해서 그 매칭도가 문턱치를 넘으면 그 문자를 인식된 문자로 간주하고 다음 인식대상문자를 처리한다. 인식이 될때까지 그 그룹의 후보문자들에 대해 매칭을 하고 그 그룹내의 후보문자들 가운데 인식되는 문자가 없으면 매트릭스에서의 주변 8 그룹에 대해 똑같은 과정을 순차적으로 반복한다. 이 그룹들에서도 문턱치를 넘는 문자가 없으면 인식을 실패한 것으로 간주하고 문턱치를 넘는 문자가 있으면 그 문자를 인식한 것으로 간주한다. 그림 8에 인식과정에 대한 흐름도를 보였다.

V. 실험결과 및 검토

본 논문에서 사용한 문자패턴은 레이저 프린터로 출력한 명조체 한글 2512자와 명조체 한자 4585자를

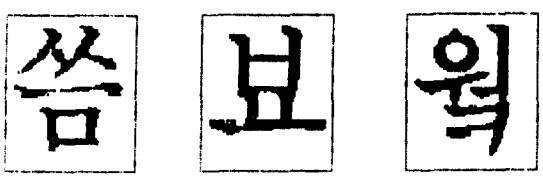
대상으로하여 MICROTEK사의 이미지 스캐너 MSF-300C를 이용하여 300dpi의 해상도로 얻은 이미지 데이타를 IBM-PC/AT 시스템에서 C언어를 사용하여 처리하였다.

1. 한글과 한자 분류

한글과 한자 분류 실험은 명조체 한글 2512자와 명조체 한자 4585자를 대상으로하여 한글의 수직, 수평 모음의 성질에 따른 한글과 한자의 구조적 특징의 차이를 이용하여 실험을 하였다. 분류실험 결과 한글 2512자에 대해서는 98.3%, 한자 4585자와 상용한자 범위의 1766자에 대해서는 각각 82.7%와 80.9%의 분류율을 얻었다. 표 1에 분류 실험 결과 데이타를 나타내었다. 한글 2512자 중에서 한자로 오분류된 데이타는 그림 9(a)와 같이 긴 수평모음 성분이 데이타 입력시 끊어짐에 의하여 수평성분의 검사에서 실패한 경우이고, 그림 9(b)와 같이 자음과 모음의 과다한 접촉에 의하여 수직모음 성분이 검출되는 경우이고, 그림 9(c)와 같이 수직모음과 종성 받침의 접촉으로 인하여 수직모음 성분의 추출 비교에 실패한 경우이다. 그러나 그림 9(a)와같이 입력시 발생한 문제점은 입력 조건을 조정 한다면 해결할 수 있으며 수직모음 성분 추출 알고리즘을 보완 한다면 한글의 오분류를 좀 더 줄일 수 있을것이다.

표 1. 한글과 한자 분류 실험결과 데이타.
Table. 1. Experimental results from classification between Korean and Chinese characters.

문 자	대상 문자수	오분류 문자수	분 류 율
한 글	2512 자	44 자	98.3%
한 자	상용 1766자	337 자	80.9%
	4585 자	795 자	82.7%



(a) (b) (c)

그림 9. 오분류 된 한글 데이타의 예.
Fig. 9. Example of mis-classified Korean character.

2. 한글 인식

1. 형태 분류 실험

한글 형태 분류 실험에서는 대상문자 명조체 2512자에 대하여 6가지 형태의 분류를 실험하였다. 분류 실험 결과 90.0%의 분류율을 얻었다. 형태분류시 먼저 형태 1과 형태 5, 형태3과 형태 4, 형태 2와 형태 6의 3그룹으로 우선 대분류를 하였다. 대분류 결과 오분류 문자를 제외하고 3그룹에 대해 다시 각 형태를 분류하였다. 표 2는 대상문자 2512자의 형태별 분포를 나타낸것이고 표 3은 형태 분류 결과를 나타낸 것이다.

표 2. 실험 대상문자의 형태별 분포.
Table. 2. Type distribution for characters used in experiment.

형 태	1	2	3	4	5	6	전 체
갯 수	155	1178	94	627	115	343	2512

표 3. 형태별 분류 결과
Table. 3. Result for type classification.

형 태	1	2	3	4	5	6	분류율
오분류 갯수	0	174	0	0	1	76	90.0%

형태 분류시 오분류는 문자 데이타의 입력시에 잡음에 의해 모음성분이 끊어지거나 짧은 수평모음성분의 잘못된 위치추적 그리고 초성자음과 종성자음간의 접촉에 의해 발생되었다. 입력시 잡음에 의하여 획성분이 끊어지는 경우는 주로 형태 3과 형태 4에서 발생하는데 입력시 조정을 잘 한다면 오분류를 막을 수 있다. 그림 10에 한글 형태 분류 결과 오분류된 데이타의 예를 나타내었다.

그림 10(a)와 (b)는 3그룹으로 대분류시 오분류된 데이타의 예를 나타낸 것으로 (a)의 예들은 수직 모음과 종성 받침의 접촉에 의하여 긴 수직모음 성분이 존재하는것으로 판단되어 형태 1, 5 그룹으로 오분류되었다. 그림 10(b)의 예들은 초성자음과 수직모음 또는 짧은 수평모음과 수직모음의 접촉에 의하여 긴 수평모음이 존재하는 것으로 판단되어 형태 3, 4 그룹으로 오분류되었다. 그림 10(c)에는 3개의 그룹내에서 오분류된 데이타의 예들을 나타낸것으로 형태 2가 형태 6으로 오분류된 원인은 초성 자음의 아래 부분이 형태 6의 짧은 수평모음 성분으로 간주되어 오분류되었으며 형태 6이 형태 2로 오분류된 원인은 초성 자음의 가로획 성분이 짧은 수평모음 성분으로 판단

되어 수평모음 성분의 위치를 잘못 계산하여 발생하였다.

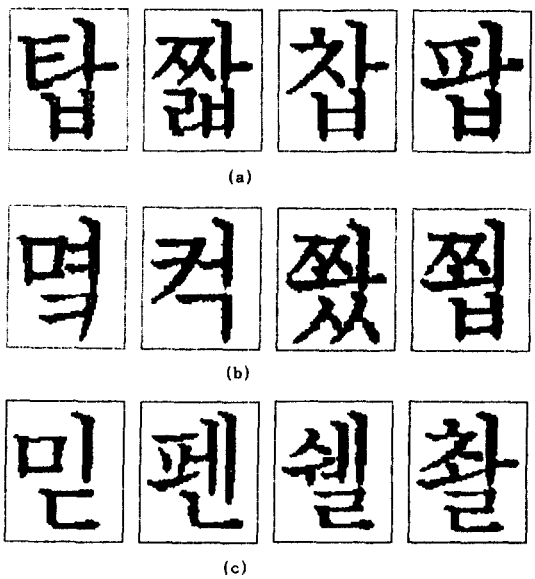


그림 10. 형태 분류시 오분류된 한글 데이타의 예.
Fig. 10. Example of mis-classified Korean character in type classification.

2. 인식 실험

분류된 한글 문자에 대하여 각 형태에 따른 모음 특성을 이용하여 모음성분을 먼저 추출 인식하고, 나머지 잔류 데이타인 자음성분에 대하여 초성 또는 종성을 자음특징에 의한 인식 트리를 가지고 인식한 후 인식된 자음과 모음에 대한 각 코드값을 조합하여 인식결과를 출력하였다. 인식 실험은 형태별로 분류된 데이타들에 대해 각각 형태별로 실험하였으며, 전체 문자수 2512자 중에서 형태 2와 형태 4, 그리고 형태 6에 대해서는 빈도가 높은 문자 1278자를 택하여 인식실험을 하여 92.2%의 인식율을 얻었다. 문헌 [16]에 의하면 빈도가 1번 이상 나오는 한글 문자의 갯수는 약 1100자 내외로 조사되어있다. 표 4에 인식에 사용한 문자의 형태별 분포와 인식율을 나타내었다. 오인식된 데이타는 추출된 모음의 오인식도 있었으나 주로 자모의 접촉이 과다하게 이루어져 모음성분의 추출후 남은 자음성분이 많이 훼손된 경우 자음 특징값의 변화에 의해 발생하였으며, 형태 2의 경우에는 모음 제거시 종성 부분의 훼손과 초성과 종성이 자연스럽게 분리되지 않는 경우 초성 자음과 종성 자음을 분리하는데에서 자음 형태의 훼손에 의해 발생하였다. 자소 접촉에 의한 문제점은 입력 데이타 상태의 조정이 필요하며, 모음과 자음의 접촉부의 정확한 검색을 위한 알고리즘이 개발되어야 할것으로 생각된다.

표 4. 형태별 인식 결과.
Table. 4. Result for recognition in each type.

형 태	1	2	3	4	5	6	총 합
실험대상갯수	155	557	94	304	115	53	1278
오인식 갯수	0	85	0	17	0	0	102
전체 인식율	92.2%						

추가로 명조체가 아닌 신문명조체와 고딕체 형태 1 (117자) 과 형태 3 (75자)에 대하여 각각 위의 인식실험을 반복하였다. 인식실험 결과 명조체와 거의 형태가 비슷한 신문 명조체에 대해서는 실험대상문자 192자에 대하여 155자를 인식하였고, 고딕체에 대해서는 실험대상문자 192자에 대하여 123자를 인식하여 평균 72.4%의 인식율을 얻었다. 오인식 원인의 대부분은 자모분리후 자음의 특징추출시 글자체의 변화에 대해 특징값이 변화하였기 때문에 발생한 것이다. 앞으로 글자체의 변화에 영향을 덜 받는 특징을 추가로 고려한다면 다중 문자체에서도 사용이 가능하리라 생각된다.

3. 한자 인식

입력 데이타에 대하여 64×64의 매트릭스로 문자부분은 좌상부에 위치하도록 정규화하여 획교차수와 최소사각형의 크기, 흑점수, 최대빈도 흑점수 등의 특징을 구한다. 구하여진 획교차수로부터 분류 과정을 거친다. 전체 대상문자 4585자에 대하여 20×20구간으로 분류한 결과를 표 5에 나타내었으며, 표 5에서 보는바와 같이 최대밀집구간의 문자수가 총문자수의 1/40 정도로 분류가 됨을 알 수 있다.

인식실험에 사용한 문자는 이미지 스캐너로 재입력한 한자 500자에 대하여 먼저 총 4585자를 이용하여 만든 데이타베이스에 적용하여 실험하였다. 먼저 입력된 문자에 대해 두방향에 대한 획교차수 평균치를 계산하여 기준 데이타베이스의 어느 구간에 속하는가를 결정한다. 그리고 해당구간의 흑점비율과 최빈획교차수 등을 이용하여 이차분류를 하여, 해당구

표 5. 획교차수 평균치를 이용한 분류

Table. 5. Classification using the average of stroke crossing number.

:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20:	
1:	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0:	2
2:	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0:	3
3:	0	0	0	2	3	3	1	5	3	3	2	1	0	0	0	0	0	0	0	0:	23
4:	0	0	0	2	4	5	9	8	12	10	7	5	2	4	1	1	0	0	0	0:	70
5:	1	0	0	3	6	11	13	16	18	26	22	26	15	8	6	5	4	0	0	0:	181
6:	0	0	0	1	6	13	20	15	27	54	50	56	43	28	17	4	4	2	0	0:	342
7:	0	1	0	0	4	16	27	35	40	62	87	78	65	49	35	15	8	4	1	0:	530
8:	0	0	0	0	4	22	25	31	34	62	92	120	97	70	42	22	11	5	2	1:	640
9:	0	0	0	0	1	11	21	32	53	80	86	124	114	90	57	34	15	9	4	0:	739
10:	0	0	0	0	1	5	22	37	55	71	85	100	116	94	51	24	13	7	3	0:	684
11:	0	0	0	0	0	7	13	34	45	57	74	76	82	74	42	12	15	9	2	0:	542
12:	0	0	0	0	1	2	10	33	54	46	45	50	37	41	23	18	7	0	0	0:	367
13:	0	0	0	0	0	1	10	20	30	31	44	27	31	12	13	6	3	2	1	0:	251
14:	0	0	0	0	0	0	0	9	22	20	18	15	10	11	0	1	0	0	0	0:	121
15:	0	0	0	0	0	0	1	14	9	10	8	11	6	3	2	1	0	0	0	0:	65
16:	0	0	0	0	0	0	0	2	2	8	4	7	3	1	1	1	0	0	0	0:	29
17:	0	0	0	0	0	0	0	0	1	3	2	3	1	0	0	0	0	0	0	0:	10
18:	0	0	0	0	0	0	0	3	1	2	1	0	0	0	0	0	0	0	0	0:	7
19:	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0:	3
20:	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0:	2
:	2	1	0	8	31	98	179	294	407	547	631	701	632	487	290	144	81	38	13	1:	4585

간과 인접구간의 데이타와 매칭을 실행하여 일정치 이상의 매칭도를 가지는 경우 인식된 것으로 한다. 여기서 매칭도의 문턱값은 70%로 정하였다.[15] 표6에 인식 실험에 대한 결과를 나타내었다.

표 6에서 인식소요시간은 하나의 문자데이타를 읽어서 특징을 조사하고 분류구간을 찾아 매칭하는데 걸리는 시간을 말한다. 인식 실패는 기준문자나 인식대상문자가 잡음에 의하여 왜곡됨으로서 후보그룹내에 원하는 문자가 존재하지만 매칭시 문턱값 이하로 매칭되어 실패로 간주되는 경우와 후보문자 그룹내에 원하는 문자가 존재하지 않으므로서 발생하였다. 또한 오인식은 후보문자 그룹내에서 문턱값이 넘는 유사문자가 먼저 매칭되어 인식되었기 때문이다. 인식 소요시간은 해당구간과 인접구간을 순차적으로 검색하기 때문에 해당후보 문자수가 많아지면 소요시간은 길어지게된다. 따라서 인식소요시간을 줄이기 위해서는 현재의 데이타베이스를 다차원 배열 형태로 구성하여 해당후보문자수를 줄여야만 할것이다.

표 6. 인식 실험 결과.

Table. 6. Result for recognition experiment.

인식대상문자수	500 자
오인식된 문자수	7 자
인식실패 문자수	47 자
인 식 율	89.2%

Ⅵ. 결 론

본 논문에서는 인쇄체 한글과 한자에 대하여 한글과 한자의 구분방법과 구분된 한글과 한자에 대하여 각각 인식 실험을 하였다.

한글과 한자의 구분은 한글의 긴 수직모음과 긴 수평모음의 위치와 성질에 의한 구조적인 차이를 이용하여 실험한 결과 한글 2512자에 대하여 98.3%, 한자 4585자에 대하여 80.9%의 한글 한자 구분율을 얻었다.

한글은 6가지의 형태로 분류하여 각 형태에 따른 모음의 특징을 이용하여 모음 우선 추출에 의한 자모분리와 모음 부분을 제거시 남은 자음 성분들에 대하여 획성분의 위치와 갯수 그리고 교차거리 특징을 변형한 4방향에서의 외형 특징 코드값들을 이용하여 인식 실험을 하였다. 빈도가 높은 1278자에 대하여 92.2%의 인식율을 얻었으며, 인식에 소모된 시간은 IBM-PC/AT(12MHz) 상에서 글자당 평균 1.5초 정도 이었다. 세선화 방법을 사용하지 않았기 때문에 세선화 과정에 소모되는 시간은 줄일 수 있었지만 문자 패턴의 과다한 접촉에 의한 오인식의 원인이 되기도 하였다.

또한 신문명조체와 고딕체 각 192자에 대하여 인식실험을 하여 평균 72.4%의 인식율을 얻었다. 잡음의 영향을 덜 받는 특징을 추가로 사용한다면 글자체가 다른 명조체나 고딕체와 같은 다중문자체의 인식에 사용될 수 있을 것이다.

한자는 제안한 획교차수의 평균값을 이용하여 분류과정을 거쳐 인식대상의 후보문자수를 줄이고 문자의 가로, 세로의 크기와 흑점비율, 그리고 획교차수의 특징을 매칭에 사용하여 인식실험을 하였다. 한자 4585자에 대한 분류를 해놓고 임의의 500자에 대하여 인식실험을 한 결과 89.2%의 인식율을 얻었으며 인식에 소모된 시간은 오인식과 인식실패의 경우를 포함하여 글자당 평균 47초 정도 걸렸다. 한자의 인식에 있어서 기준 문자에 대한 다차원적인 데이타베이스의 구축을 꾀함으로서 후보문자수를 줄여 인식을 위해 검색하는 시간을 줄이고 인식율을 좀 더 높일 수 있으리라 생각된다.

참 고 문 헌

1. 이주근, "한글 문자의 인식에 관한 연구." 대한전자공학회 논문집, 제9권 4호, 1972. 9.
2. T. Agui et al., "A method of recognition and representation of Korean character by tree grammar." IEEE Trans. Pattern Anal. Mach.

Intell. vol. PAMI-1, pp245~251, July 1979.
3. 이주근, 남궁재찬, 김영건, "한글 Pattern에서 Subpattern분리와 인식에 관한 연구," 대한전자공학회 논문집, vol. 18, No. 3, 1981.
4. 최병욱, 市川忠男, 藤田廣一, "한글 認識에 있어서의 子素抽出,"대한전자공학회 논문집, vol. 18, No. 3, 1981.
5. 이균하, "속성에 구속을 받는 문법을 이용한 문자 패턴 인식," 안하대학교 박사학위 논문, 1981.2.
6. 전종익, 조용주, 남궁재찬, "한글 Shape 문자 Pattern에서의 구조적 정보를 이용한 형식 분류와 인식에 관한 연구," 한국통신학회 논문지, 제16권, 제2호, pp180~195, 1991. 2.
7. 이행세, 최태영, 김영길, 김정우, "인공지능 기법을 이용한 텍스트 인식에 관한 연구," 대한전자공학회 논문집, 제26권 제11호 pp153~164, 1989.11.
8. R. Casey and G. Nagy, "Recognition of printed Chinese characters," IEEE Trans. on Electric Computer. Vol.EC-15. No.1,pp91~101,1966.
9. P. Wang and R.Shian, "Machine recognition of printed Chinese characters via transformation algorithms," Pattern Recognition, Vol.5,pp. 303~321,1973.
10. S. Yamamoto, A. Nakajima, K. Nakata, "Chinese Character Recognition by Hierarchical Pattern Matching -Study of Chinese Character Recognition-,"일본전자통신 학회지, Vol.56-D, No.12, pp714~721, 1973.12.
11. M. Umeda, "Pre-Classification for Recognition of Multi-Font Chinese Characters," 일본전자통신 학회지, Vol.J62-D, No.11, pp758~765, 1979.11.
12. P. Chen, Y. Chen, W. Hsu, "Stroke relation coding-a new approach to the recognition of multi-font printed Chinese characters," Int. Journal of Pattern Recognition and A.I., Vol. 2, No.1, pp.149~160, 1988.
13. 이승현, 조용주, 이성범, 남궁재찬, "문서 인식을 위한 한글과 한자의 구별과 한글의 형식 분류에 관한 연구,"대한전자공학회 하계종합학술대회 논문집, 제13권 제1호 pp232~235, 1990.7.
14. 이행세, 김정우, 김석동, 김태호, "한글 텍스트에 나타난 한자의 분류에 관하여," 제2회 신호처리 합동 학술대회 논문집, pp209~212, 1989.9.
15. 김정우, 김태호, 송도선, 이행세, "획교차수를 이용한 인쇄체 상용한자의 인식에 관한 연구," 대한전자공학회 추계종합학술대회 논문집, 제12권 제2호, pp476~479, 1989.
16. 한글·한자찾기 조사일람표, 한국과학 정보센터.

金 晶 雨(Jeong Woo Kim) 正會員
1960年 10月 5日生
1983年 2月 : 광운대학교 전자공학과 공학사
1985年 2月 : 아주대학교 대학원 전자공학과 공학석사
1992年 2月 : 아주대학교 대학원 전자공학과 공학박사
1992年 9月~현재 : 거제전문대학 전자과 조교수
※주관심분야 : 패턴인식, Computer Vision, 디지탈 신호처리, 인공지능, 신경회로망, Fuzzy 등임.

李 幸 世(Haing Sei Lee) 正會員
1943年 8月 29日生
1966年 : 전북대학교 전기공학과 공학사
1972年 : 서울대학교 전자공학과 공학석사
1984年 : 고려대학교 전자공학과 공학박사
1968年~1970年 : 해군사관학교 전자공학 교관
1973年~현재 : 아주대학교 교수
1982年~1983年 : 미국 Columbia Univ.(N.Y.) 객원교수
1987年~1988年 : 프랑스 INRIA(Paris) 객원교수
1992年~현재 : 거제 전문대학 학장
1973年~현재 : IEEE 회원
※주관심분야 : 문자 및 음성인식, 인간기계 인터페이스, 인공지능, 신경회로망, VLSI 회로, 디지탈 신호처리 등임