패턴 인식에 대하여

——한글 글씨의 기계적 인식을 중심으로——

강 인 구*

1. 서 론

글자를 읽어서 알고 말을 들어서 그뜻 뿐아니라 상대가 누구인 가를 아는 것은 물론 그 상대의 기분까지 아는 능력은 인간 고유의 인식 능력으로 간주되어 왔다.

패턴은 이러한 음성이나 글자뿐 아니라 도형, 병의 증상 등을 들 수 있다. 그러니까 패턴인식이면 글자의 인식, 음성의 인식, 지문의 감정, 병의 자동 진단, 천기 예보, 사진·음향·전파에 의한 정찰에서 얻은 정보의 분석, 게임의 운영, 암호의 해독등이 포함되며 관상도 일종의 패턴 인식이라고 할 수 있다. 이러한 일은 20년전만해도 인간만이 하는 일로 의심하지 아니했다.

그러나 전자 계산기의 발달이 인간의 계산능력을 대치했듯이 이것을 발달시켜 패턴 인식의 능력 마저도 전자 계산기가 맡을 수 없겠는지 모색하기에 이르렀다.

이 중에서도 문자와 음성을 기계로 인식하는 문제는 비교적 개발이 빨리 이루어지고 있는데, 이는 전자 계산기 자체의 약점을 보완하기 위해서이다. 즉 전자 계산기의 보급에 따라 보다 손쉽게 기계와 인간사이에 의사를 소통시키는 수단을 요구하게 되는데, 인간의 가장 손쉬운 방법이라면 말로 하든지 글을 써서 전하는 것인 반면에 기계는 전기적 신호가 있어야 한다. 그래서 현재로서는 펀치 카—드라든지 테프와 같은 매개체를 일단 만들어 가지고 그것을 전자 계산기가 읽도록 하고 있으나 이로 인한 시간과 인력의 소모가 클뿐 아니라 인간에게 편리한 정보의 형태도 못된다. 또 하나의 이유는 전자계산기가 특수한 용도로 쓰임에 따라 입력장치로서 필요하다. 즉 수표나 인쇄물을 분류한다든지 화물과 화차를 자동 분류하고 선별한다든지 요금의 증수서를 자동처리한다든지 차표를 자동으로 개찰하는 등에 쓰이는 입력 장치는 글씨나 기호를 식별하는 능력이 있어야 한다. 이러한 패턴을 인식하는 기계는 그 인식의 대상을 제한시켜 표준화된 패턴만을 읽게하는 방법만이 현재 주로 실용화 단계에 있으며 여러가지 유형의 패턴을 전부 인식할 수 있는 보다 고급의 인식장치는 아직도 연구 단계에 있다.

이 글은 주로 글씨를 기계적으로 인식하는 문제에 대하여 고찰하고 특히 한글 글씨를 인식하는 문제에 대한 문제점과 연구 동향을 소개하고저 한다.

2. 글씨의 인식 기계

패턴을 인식하는 기계의 일반적 구성은 그림 1과 같다.

그림에 주사(走査)기 구란 패턴을 기계가 처리할 수 있는 형태로 바꾸는 장치로서 주로 광

*기술사(전기부문)
울산공과대학 전기과과장

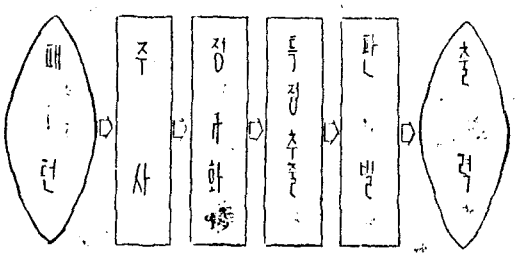


그림 1 패턴 인식의 구성

학적으로나 자기(磁氣)적인 방법으로 전기(電氣) 신호를 발생시킨다.

정규화(正規化) 기구란 필요 없는 정보로 제거하고 약간의 변형등을 바로 잡으며 주위의 잡음을 제거해서 보통 입력신호를 양자화(量子化)하는 역할을 맡는다.

특징 추출 기구란 판별에 필요한 정보를 주사기구와 정규화기구를 거쳐 양자화된 신호에서 능율적으로 집약하는 기구로서 어떤 방식의 특징을 추출할 것인가에 대해서는 그동안 많은 연구가 거듭되어 왔으며 여러 방식에 대해서는 뒤에 소개하겠다.

판별기구란 입력에서 추출된 특징과 표준 글자의 각 특징을 비교해서 입력의 패턴을 결정하는 장치로서 이것은 보통 표준 글자의 패턴에 가장 근사한 것으로 결정하는 확율적인 수법을 흔히 사용한다.

고급의 패턴 인식 기구에서는 특징 추출과 판별 기구의 기능을 환경이나 실적(實績)을 참고해서 개조해나가는 학습(學習)기구가 있는 수가 있다.

글자를 인식하는 기계중에서도 필기한 글자를 읽는다든지 여러가지 자형(字形)의 인쇄체를 전부 읽는 기계에는 이런 학습기구가 필요하지만 현재 실용되고 있는 기계는 주로 특수하고 일정한 자형만을 판별하게 되어있다.

이글은 광학적인 방법에 대해서 설명하겠는데 주로 글자의 어떤 특징을 추출하느냐에 따라서 여러 종류로 분류된다. 이 종류에 대해서 간단히 아래에서 설명하겠으나 실제로 실용되고 있는 방식은 매트릭스 맞추기 법과 스트록 분석법 정도이다.

(1) 매트릭스 맞추기법

가장 단순하고 안전한 방법으로 글자의 형태와 크기가 일정하면 이 글자를 일정 간격의 격자(mesh)속에 넣고 각 바둑눈안이 글씨로 덮혔는가 아닌가를 1과 0의 신호로 양자화해서 표준 패턴과 대조하는 것으로 초기의 인식기계가 표준 패턴의 필림과 비교하는 것과 상통한 점이 있다.

근래에는 1, 0으로만 양자화하지 않고 그자리에 있어서의 존재 확율이나 다른 문자와의 차이 정도를 감안해서 일정한 값으로 곱하는 가중(加重)의 방식도 많이 쓰이고 있다.

(2) 정점(定點)샘프링

역시 일정 간격의 격자위에 글자를 놓았을 때 특정한 점에 있어서의 흑점 또는 백점(白點)의 유무 조합으로 판정하는 방법인데 기억하고 처리할 정보의 량은 작으나 오염등으로 잘못 식별할 경향이 높다. 인쇄체 한자의 인식에서도 이 방법이 지도된 바 있다. [9]

(3) 손데(Sonde)법

평면상의 몇개의 Sonde라는 선을 놓고 이 선이 글자와 교차하는가 않는가에 따라 스토록의 위치를 검출하는 방식으로 글자의 크기, 형태, 위치, 경사등의 변화에도 비교적 예민한 장점이 있다.

(4) 스릿트(Slit) 이용법

글자에 수직한 스릿트를 수평 방향으로 주사하거나 수평한 스릿트를 수직 방향으로 주사(走査)시켜 도형 점유비(占有比)로 나타나는 신호 파형의 변화에 의해서 판별하는 방식이다.

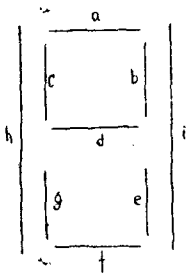
(5) 글자의 기하학적 특징 이용

가. 문자의 각 부분의 접속 관계

만체스터(manchester) 대학에서 처음 거론된 방법으로 수직방향의 상태가 변화하는 것을 경계(境界)로 해서 분할한 다음 각 부록의 접속 장소 및 방향의 조합으로 특징짓고 판별하는 방법이다. 일본에서 가다가나를 이와 유사한 방법으로 식별하는 기도가 있었다. [8]

나. 스트록 분석

그림 2에서와 같이 숫자를 식별하려면 정해진 2개의 스트록으로 분할해서 그 스트록의 어느 것이 있고 어느 것이 없는 가에 따라 식별되는 방식으로 예를 들면 "7"은 a, b, g 라는 기본 스트록으로 구성됨과 동시에 d 나 f와 같은 스트록이 있어서는 안된다는 조건을 만족시켜야만 된다.



(가) 기본스트록

0	1	2	3	4	5	6	7	8	9
a		a	a	c	a	a	a	a	a
i		b	d	d	c	d	b	b	c
h	h	d	i	i	d	e	g	c	d
f	f	g	f		e	f	(d)	d	(g)
(d)	(a)	f	(c)	(f)	f	h	(f)	e	(f).
	(d)		(g)		(b)	(a)		f g	

(나) 분류 *( ) 안은 없는 스트록

그림 2 스트록 분석 방식

다. 문자 주변 각변의 방향성 분석

글자의 주변을 그림 3에서와 같이 그 방향이 수평〔T(위), B(밑)〕 수직〔L(좌), R(우)〕 혹은 경사를 파라미터로 추출해서 글자의 주변에 따라 이 순서를 가지고 판정하는 방식인데 그림 3에서는 그 특징은 L-B-L-B-R-T가 된다.

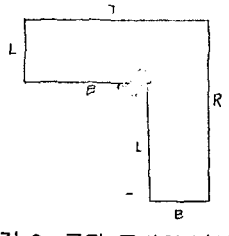


그림 3 글자 주변의 방향성

라. 단점(端點) 굴절점(屈折點) 분기점(分技點) 루—프등을 파라미터로 하는 방법

단점, 굴절점, 분기점, 루—프 혹은 고립점등을 파라미터로 생각하고 그것이 있는가 없는가 또는 방향성등을 검출해서 그러한 파라미터의 배열순서로 글자의 특징을 나타내서 판정하는 것이다.

마. 글자 주변의 각을 파라미터로 해서 그 수(數)로 판정하는 방법

2차원적으로 글자를 읽어서 그 주변을 정형(整形)하고 그 정형된 각의 점을 특징 파라미터로 해서 그 갯수(個數)로 판정하는 방식이다.

바. 여러 방향에서 본 도형의 형상을 특징 파라미터로 하는 방식

어떤 글자를 어느 특정한 각도로 회전시켰을 때 나타나는 어떤 특징을 파라미터로 하는 방식인데 회전 각도와 회전후의 어떤 형태상 특징이 파라미터가 된다. 형태상 특징의 예로는 도형의 경계를 미분한 출력 파형도 있고 흑면(글자가 차지하는 면적)의 좌우 비교등이 있다.

3. 한글 글자의 특색

위에서 주로 살핀 방법은 모두 외국문자 특히 알파벳과 숫자를 대상으로 한 것이다.

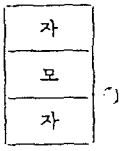
이런 방법을 응용하거나 혹은 다른 방법을 모색하기에 앞서 우리 한글이 인식면으로 봐서 어떤 특징이 있는가를 알 필요가 있다.

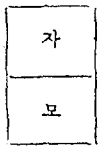
한글은 첫째로 그 글씨가 자모의 복합체라는 점을 특색으로 들 수가 있다. 즉 자모가 2에서 최고 7개가 합쳐서 한 글자를 이루고 있으므로 이를 자모로 분리해야만 그 식별이 간단하지 그렇지 않으면 매우 많은 패턴을 식별해야 하는 문제가 생긴다. ㅐㅒㅔㅖ의 네 복모음을 자모로 간주하고 복합체의 유형을 나누면 18가지의 패턴으로 분류된다. 주어진 글자를 어떻게 효과적으로 분리해서 어느 패턴에 속하는지를 결정할 것인가에 대해서 이미 발표된 바 있다.〔5〕

18가지나 되는 패턴이 있으나 그중에서 다음 빈도순으로 번호를 붙인 4가지 패턴의 총 생기빈도가 90%이상이다.

1. 

2. 

3. 

4. 

둘째로 한글의 자모는 비교적 간단한 구조를 갖고 있다. 즉 그 구성 요소가 선(線)과 점(點)과 그리고 원(圓)으로 되어 있을 뿐아니라 그 연결이 또한 단순해서 각 구성요소는 서로 한점에서만 접촉한다.

그래서 이주근 교수는 [6] ㄷ속에 ㄴ의 요소 포함되고 ㅁ속에 ㄷ, ㄱ, ㄴ등의 요소가 포 된다고 지적한 바 있다.

셋째로 한글 자모중 |모음자모는 그 형태보다 그 위치 즉 다른 자모와의 상대적 위치가 매 중요하다. 실제로 모음의 자모는 10개이지만 ㅏ" "ㅑ" "ㅣ"의 세가지 모양을 다른 자모와 상 적으로 어떻게 놓았느냐에 달린 것임은 곧 알 가 있다.

이러한 특징의 고려 없이 외국에서 다른 나라 자를 위해서 개발된 방식을 그대로 적용할 수 음은 분명하다.

4. 한글 자모의 식별 방식

한국에서 전자 계산기가 도입되고 그 이용에 한 인식이 높아진 것은 극히 최근의 일이므로 글 자모에 대한 인식 문제에 대해서도 거의 미 척이라고 할 수 있다. 이제까지 한글 자모에 한 인식 방식으로 시도된 방식을 소개하면

(가) 개량매트릭스 맞추기 형식[6]

비교적 굵은 매트릭스(5×3)로 그 눈금안에 흑 (黑白)의 존재와 동시에 그 순서를 판정기준 하나로 사용하여 비교적 간단한 특징 함수를 게 된다.

이 방식은 간단한 반면 대상 글자의 형태가 일 적이어야 한다.

(나) 가지수와 절점수 및 좌우 흑점차의 비교 [7]

직선과 직선이 맞나는 점을 절점이라하고 절 사이의 선분을 가지라고 하는데 이 가지수와 점수로 자음 자모를 분류하고 모음 자모는 자 의 위치와 가지 수로 분류를 시도 했으나 그 도 분리 안되는 자모가 있어서 다시 좌우 흑 상하로 흑색면 즉 자면(字面)의 차지하는 정 를 가지고 분류하는 방법인데 그러한 비교가 데면에서는 매우 힘들 것으로 예상되므로 기 가 별로 안되는 방식이다.

(다) 결합행렬 법[5]

이 방법은 위의 방식에서 출발한 것으로 절점 절점사이의 가지의 연결 관계와 그 연결 방 을 가지고 특징을 추출한 것이다.

그림 4와 같은 방향 지수를 행렬 소자로 한 각

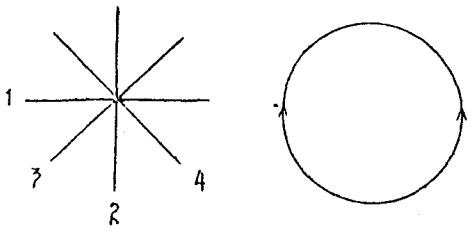


그림 4 방향지수

자모의 특징 행렬은 전부가 3각행렬이며 그 크기는 최소 2×2부터 최대 8×8이다. 이 방식은 좀 복잡하지만 특정한 인쇄체 이어야 한다는 제한도 없고 글자의 크기에도 별 영향을 받지 않는다.

5. 앞으로의 연구 방향

이제까지 외국에서 개발되고 있는 방식과 한글에 대해서 시도된 방법을 소개하였는데 이러한 방식외에도 아직 발견못한 방법이 있으리라고 믿으며 이의 발견을 위해서 힘써야 할 것은 물론이다. 앞으로 이 방면에 관심을 갖는 분이 더욱 늘어날 것을 기대한다.

보다 현실적인 문제는 현재 외국에서 실용화되고 있는 기계의 경우처럼 우리 한글도 속히 OCR나 MICR용의 자체(字體)가 표준화되어야 할 줄 안다. 그림 5와 6에서 보는 바와 같이 이미 영어 불어 독일어의 알파벳에 대해서는 국제 표준화 기구(ISO)에서 두개의 표준 자형을 권고한 바 있거니와 우리나라에서도 여러 외국 제작 회사에서 잡다하게 개발한 후에 그것을 통일하

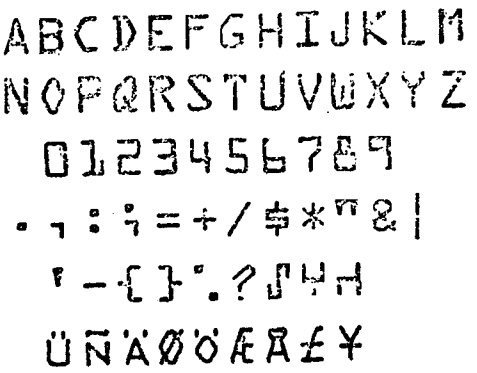


그림 5 OCR-A

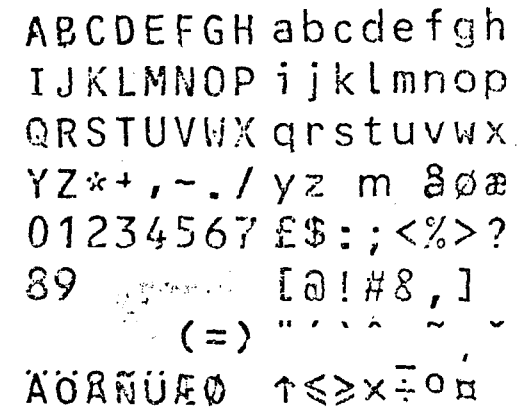


그림 6 OCR-B

기 위해서 현재 한글타자기가 겪고 있는 것과 같은 홍역을 치루지 않으려며는 하루속히 표준화가 이루어 지어야 하] 이 표준 글자형은 사람이 읽기도 쉽고 동시에 기계가 읽어 식별하기도 쉽도록 설계되어야 할 것이다.

현재 외국에서 실용화된 OCR를 한글도 읽어내도록 개조해 보려는 연구가 일부 진행되고 있다고 듣고 있는데 이는 현재 계산기 시장이 한정된 한국에서는 현실적으로 매우 실질적인 방향이라고 생각된다.

원래 글자의 인식은 필기체의 인식을 지향하고 연구되는 실정이나 실용적인 면에서 많은 반론을 이르키고 있다 한글의 필기체에 대해서도 그 기본적인 연구는 시작되어야 할 것이다.

또한 음성의 인식에 대한 문제도 외국에서는 여러 분야의 학자가 팀이 되어서 수행하고 있거니와 한국말에 대한 연구는 그 기초적인 자료조차도 미비하니 만큼 이 방면의 연구도 많이 해야 될 줄 생각한다.

이러한 각각의 인식 대상에 대한 방식을 개척하는 것과 아울러 특징 추출에 대한 일반적인 이론도 아직 세계적으로 미개척인 만큼 한국의 학자가 공헌할 길이 있다고 본다.

6. 결론

한글 글자를 어떻게 기계적으로 인식시킬 수 있겠는가 하는 문제를 중심으로 "패턴 인식"이란 전자 공학의 한 분야가 어떤 것이며 그 방법론을 설명하고저 쓴 글이나 미비한 점이 많은것 같다.

이제 한국에서는 차차 전자 계산기가 실용화되고 있는 만큼 필연적으로 한글 글자를 전자 계산기의 입력으로 써야만 일이 쉽게 처리될 응용분야가 많이 생길 줄 믿으며 적어도 우리 글의 인식만큼은 우리의 지혜로 이룩되기를 소망하면서 끝맺고저 한다.

참고 문헌

1. 情報處理學會編, 電子計算機 핸드북 제9편, 제4장, 옴社, 日本東京(1968)
3. 山崎一生, 최근의 패턴 인식 기계를 말한다. 電氣計算 1969. 10월호, p. 128~134
5. 강인구, 한글 자체 인식 방식—자모 분할방식을 주로한—1970. 1. 31 대한 전자공학회 학술발표회
6. 이주근, 한글 문자의 인식을 위한 특징 패턴의 추출 및 문자의 2가부호화 (I), 1969. 12. 15 대한 전기 학회 학술연구발표회
7. 강인구, 이행세, 한글 자체의 특징 추출의 한 방식, 전자학회지 6권2호 P. 1~5, 1969. 9
8. 富田등, Recognition of Hend written Kata kana Characters, 電通學誌 50권 p.650~663, 1967. 4
9. P. R. Casey & G. Nagy; Recognition of Printed Chinese Characters, *IEEE Trans. on EC*-Vol 15, p.91~101, Feb. 1966