

단위벡터를 이용한 한글 인식에 관한 연구

전 수 언

조 동 섭

이 화 여 자 대 학 교

A study on Korean character recognition using the sum of unit vector

요 약

본 논문은 한글 자모의 인식에 관한 새로운 방법을 제시한다.

본 연구는 한글패턴을 독립된 자소의 부분패턴으로 나누어서 특징점들(골점, ㄱ골국점, T골국점)을 추출하여 각 골국점에서 연결되는 점과의 벡터를 골자의 크기에 관계없이 하기 위해 크기가 1인 단위벡터를 구한 후 이들의 합성벡터를 생성한다. 생성된 합성벡터들의 수, 벡터들이 기준축과 이루는 각, 그리고 특징점들의 수로부터 한글의 기본 자모를 분류하는 기법을 연구하였다. 일차원 한글패턴은 이미 세션화가 된 자오패턴으로 하였다.

1. 서 론

인공지능의 분야 중에서 인간의 시각 능력과 관련된 패턴인식 분야 중 문자인식에 관한 연구가 많이 진행되고 있다. 우리나라 문자인 한글에 대하여도 1970년대를 전후해서 시작하여(1, 2) 인쇄체에 대해서는 성숙단계에 이르렀다고 볼 수 있다. 1980년대 초반부터는 필기체에 관한 연구가 시도되었으며, 여러가지 방법이 제시되어 왔다.(3, 4)

한글은 기본 자모에 의해 조합하여 사용하기 때문에 문자인식이 어렵고, 또한 자모가 극히 단순하고, 문자에 동일한 자모가 패턴이 다소 다른 문자로 반복조합되기 때문에 인식하는데 어려움이 있다.(5)

본 연구는 입력된 Binary image의 한글 패턴에서 특징점들을 추출하여 각 특징점들을 골점, ㄱ골국점, T골국점 등으로 분류하였다. 이 때, 각 골국점에서 단위벡터들을 합성하여 합성벡터를 구하여 구해진 합성벡터들의 수, 그 벡터들이 기준축과 이루는 각, 그리고 특징점들의 수(골점의 수, ㄱ골국점의 수, T골국점의 수)를 갖고 부분패턴을 기본자모로 분류한다.

1.1. 한글 자모의 인식방법

특징점들을 이용하여 부분패턴을 인식하는 과정에서 사용하는 정의들은 다음과 같다.

정의 1. P_c 는 각 특징점에서 연결되는 점이다. N_c 는 연결점(P_c)의 갯수이다.

정의 2. 골점(P_g)은 $N_c = 1$ 인 점이다. 골점들의 갯수는 N_g 로 표시한다.

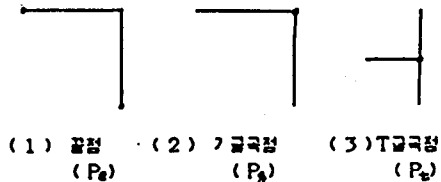
정의 3. ㄱ골국점(P_g)은 $N_c = 2$ 인 점이다. ㄱ골국점의 갯수는 N_g 로 표시한다.

정의 4. T골국점(P_t)은 $N_c = 3$ 인 점이다. T골국점의 갯수는 N_t 로 표시한다.

정의 5. 특징점(P_f)은 골점, ㄱ골국점, T골국점을 말한다.

정의 6. 특징점들의 수(N_f)는 (골점들의 수(N_g), ㄱ골국점들의 수(N_g), T골국점들의 수(N_t))로 정한다.

부분패턴의 각 골국점(P_g 혹은 P_c)에 대해 그 점에 연결되는 점(P_c)들을 찾아서 골국점으로부터 연결점으로의 단위벡터를 구한 다음, 각 골국점에서 단위벡터들을 합성하여 합성벡터(V)들을 얻는다. 이때 단위벡터는 임의의 주어진 벡터를 그것의 크기로 나누어서 구하며, 따라서 단위벡터의 크기는 1이다. 또한 단위 벡터의 방향은 원래의 벡터가 갖는 방향과 같다. 기준축은 임의로 정하여 줄 수 있다. (그림1)과 (그림2)에 각 특징점들의 예인 골국점에서 단위벡터를 구하는 예를 보여준다.



(그림1) 골점, ㄱ골국점, T골국점의 예



(1) 단위벡터를 구하기 전의 패턴

(2) ㄱ골국점 b에서 연결점 a, c로의 벡터에서 단위벡터를 구한 패턴

표1. 종류에 따라 분류한 기본자모

종류	기본 자모	id	r _i	(N _e , N _g , N _h)
종류1	ㄱ	1	7	(2,1,0)
	ㄴ	2	3,4	(2,1,0)
	ㄷ	7	1,2	(3,0,1)
	ㄹ	15	2,3	(3,0,1)
	ㅁ	17	6,7	(3,0,1)
	ㅂ	19	4,5	(3,0,1)
	ㅅ	21	8	(3,0,1)
종류2	ㅇ	3	1-4	(2,2,0)
	ㅈ	9	7	(3,1,1)
	ㅋ	11	7,8	(3,1,1)
	ㅊ	16	2,3	(4,0,2)
	ㅋ	18	6,7	(4,0,2)
	ㅌ	20	4,5	(4,0,2)
	ㅍ	22	1,8	(4,0,2)
	ㅊ	12	1-4	(4,2,0)
	ㅋ	10	7	(5,1,1)
	ㅌ			1,2
종류3	ㅌ	10	4,5	(4,1,2)
	ㅌ		7	
종류4	ㅇ	12	1-4	(3,2,1)
	ㅌ		2,3	
	ㅌ		1-4	
	ㅌ	4	7,8	(2,4,0)
	ㅌ		5,6	
종류5	ㅇ	5	1,2	(0,4,0)
	ㅌ		7,8	
	ㅌ		3,4	
	ㅌ		5,6	
	ㅌ	6	2,3	(2,2,2)
종류6	ㅌ		2,3	(2,2,2)
	ㅌ		3,4	
	ㅌ		5,6	
	ㅌ	13	1,8	(4,0,4)
종류5	ㅇ	8	loop	(0,0,0)
	ㅌ	14	loop	(3,0,1)
	ㅌ		4,5	
종류6	ㅌ	14	loop	(4,0,1)
	ㅌ	23	2,3	(2,0,0)
종류6	ㅌ		6,7	
	ㅌ	24	1,4	(2,0,0)
종류6	ㅌ		5,8	
	ㅌ			

(그림2) 글꼴점에서 단위벡터를 구하는 예
 이 과정 중에서 두 벡터의 인상을 구하는 과정은 다음과 같다.
 글꼴점을 기준으로하여 연결점(P₀)도와의 벡터의 크기를 각각 x, y라 하고, 벡터들이 이루는 각도를 각각 α, β라고 하자. 이 때, 두 벡터의 인상된 벡터의 크기(z)는 (식1)과 같다.

$$z = \sqrt{x^2 + y^2 + 2xy \cos(\beta - \alpha)}$$

---(식1)

구해진 z로부터 인상된 벡터의 각(d)을 (식2)와 같이 구한다.

$$rd = ((x^2 + y^2 - z^2) / (2xy))$$

$$Y = 1.5708 - \text{ATAN}(rd / \sqrt{1 - rd^2})$$

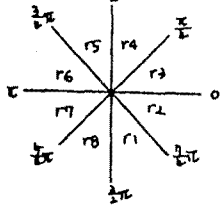
상향벡터의 각 $dv = \alpha + Y$

---(식2)

이러한 벡터의 인상을 구하기 전에 몇몇의 자모는 따로 분류하여 인식하여야 한다. 이러한 자모의 예로는 '0', '8', '1', 그리고 '1'가 있다. 복장점을 추출할 때 loop가 있는지 조사하여 loop가 있으면 '0' 나 '8'으로 분류하여, 또한 loop가 없으면서 글꼴점도 없으면 '-' 나 '1'로 분류한다.

앞에서 분류한 자모를 제외한 자소들은 인상벡터들을 구하여, N_v와 dv, 그리고 N_φ를 비교하여 기본 자모로 분류한다. 이때 dv는 자소의 형태가 여러가지의 변형이 있을 수 있으므로, 이에 따라 생성될 수 있는 각의 범위를 주었다. 각의 범위는 (그림3)과 같이 정하고 각 범위에 1부터 8까지의 번호를 부여한다.

(r_i, 1 ≤ i ≤ 8)는



(그림3) 각의 범위 설정

N_v를 기준으로 하여 각 종류로 분류하고, dv_i (1 ≤ i ≤ N_v)와 N_φ에 의한 전후반부 인상을 (표1)에 나타내었다.
 (여기에서 id는 각 기본자모를 나타내는 번호이다.)
 (r에서 '0', '1' 나 '-'는 앞뒤의 범위에 속한다는 것)을 의미한다.

III. 인식의 예

본 장에서는 2장에서 제안한 방법을 사용하여 한글의 기본 자모를 인식해 내는 과정을 설명한 예를 들어 설명한다.

예로는 P_e, P_g, 그리고 P_h가 모두 있는 '3' (id = 11)을 예제 문자로 선택하였다. 10X12 matrix에 '3'을 나타내어 설명하였다. 설명에 사용된 '3'의 matrix는 (그림4)에 나타내었다.

(그림4) 10X12 matrix의 '3'

(그림4)의 matrix에서 특징점들을 찾아내어 곱셈 (F₂)은 e, 7 글극점 (P₂)은 e, 그리고 T글극점 (P₃)은 7로 표시하여 (그림5)에 나타내었다.



Ne = 7
N_g = 1
N_t = 1

(그림5) P₂ 등이 추출된 matrix와 N

수미진 matrix에서 양을 y축, 굵을 x축으로 가정하여, 각 특징점들의 좌표를 나타내면 다음과 같다.

- (2, 3)
- (7, 3)
- (9, 5)
- (3, 6)
- (7, 10)

(표2)에 각 글극점 (P₂ 혹은 P₃) 등에서 인식벡터 (V)의 dv와 r_k를 표시했다.

P ₂ OR P ₃	P ₂	d _v	r _k
(2, 3)	(2, 3) (9, 6)	0.627	1
(7, 5)	(7, 5) (3, 6) (7, 10)	0.047	6

표2. 각 글극점에서의 P₂와 r_k

(그림6)은 '가'의 10x12 matrix를 갖고 인식한 결과를 보여준다.

THE NUMBER OF VECTORS : 2
 1. VECTORS SINGLE RANGE : 0
 2. VECTORS DOUBLE RANGE : 6
 THE FEATURE POINTS : 1, 1, 1
 THE CHARACTER'S ID. NO. : 11

(그림6) '가'를 인식한 결과

IV. 결 론

본 논문은 각 글극점에서 단위벡터들의 합성에 의해 생성된 인식벡터들의 수와 그 벡터들이 기준축과 이루는 각도, 그리고 특징점들의 수(곱셈의 수, 7글극점의 수, T글극점의 수)에 의해 한글의 기본 차이를 인식할 수 있는 새로운 방법을 제안하였다.

본 방법은 자음과 모음을 따로 분류하지 않고 인식할 수 있으며 인식되는 문자의 패턴에 따라 적용되지 않는 부분이 있으나 정상적인 필기체의 대부분은 연구된 방법에 의해 처리 가능하다.

참 고 문 헌

- (1) 이 주근, "한글 문자인식에 관한 연구(II)," 전자공학회지, 제7권, 제3호, pp5-11, 1970년
- (2) 이 주근, "한글 문자인식에 관한 연구(IV)," 전자공학회지, 제9권, 제4호, pp25-32, 1972년
- (3) 이 은주, 김 태근, "한글의 on-line 인식에 관한 연구," 한국정보과학회 '85 가을 학술 발표 논문집, 제12권, 제2호, pp278-284, 1985년
- (4) 왕 박영, "구조적 특성을 이용한 필기체 한글의 인식," 서강대학교 석사 학위 논문, 1985년
- (5) 이 주근, 인공 지능, pp258-259, 정문각, 1986년
- (6) Laurent Miclet, "structural methods in pattern recognition," Springer-Verlag New York Inc., pp130-134, 1986년