한국통신학회 1988년도 추계학술발표회 논문집(1988. 11. 5) 88691

# OCR 한글폰트의 설계에 관한 연구

○김 재 성 김 홍 윤 김 홍 일 이 균 하
인하대학교 , 전자계산학과

# Research on the Design of OCR Hangul Font

**Jae Sung Kim, Hong Yon Kim, Hong Il Kim and Kyoon Ha Lee**
**Dept . of Computer Science , INHA University.**

**Abstract**

The size and shape of a phonetic symbol in a Hangul character frame changes depending on the types of the other phonetic symbols in the same character frame. It makes difficulty in machine recognition. And thus, we designed **OCR Hangul Font** which led easy machine recognition in consideration of the structural characteristics of the Hangul character frame.

## 1. 서 론

컴퓨터의 기억 용량이 대폭적으로 증가되어 취급하는 정보의 양이 기하 급수적으로 증가하고 있으나, 정보의 입력 과정은 대부분 키보드 조작에 의한 극심한 수작업으로 진행되고 있다. 이러한 이유로 정보의 입력 과정은 자료 처리의 병목 현상을 초래하여, 정보 입력에 대한 자동화 요구가 날로 증가 되었다. 또한, 기존의 정보들이 대부분 문서로 저장되어 있기 때문에 이를 컴퓨터화하는데 방대한 양의 입력을 요구하여 보다 효율적인 문서 입력 방법이 요구된다. 이러한 요구에 발 맞추어 기계 판독이 용이한 정보들을 신속하게 컴퓨터로 자동입력 시킬 수 있는 OCR(Optical Character Recognition)이 각광을 받기 시작하여 이와 관련된 연구 발표들이 있었다 [1,3,4,13].

OCR에 있어서 필수 요건인 문자 인식 기술과 이미지 센서의 기술이 점차 발전하고, 반도체 기술의 눈부신 발전으로 중앙 처리 장치, 기억 장치 등의 속도 및 기억 용량이 개선되고 있으나, 현재의 기술로써 사람의 시각으로 판독이 가능한 여러 종류의 모든 문서를 OCR로 자동 입력하는데는 아직 거리감이 있다. 따라서 현재의 OCR이 문서를 인식하기 위해서는 OCR 전용의 폰트가 필요하다. 현재 영어권에서 널리 사용되고 있는 OCR 폰트로는 ANSI 에서 인식이 용이한 활자체 폰트로 개발한 OCR - A 폰트와 ECMA(European Manufactures Association)의 OCR - B 폰트등이 있다 [1]. 그러나 이러한 폰트는 영 문자 폰트이며, 한글 문서의 인식을 용이하게 하기 위해서는 **OCR 한글 폰트의 개발**이 요구된다. 이와 관련하여 한글 수서문자의 표준화를 위한 연구 발표가 있었으나 [13], 필기 작성과 관련된 연구로서 OCR 전용의 인쇄체 한글 폰트로는 거리감이 있다.

문자 인식에 있어서 글자체(typeface), 경사(slant), 곡률(curvature), 획의굵기, 획의 돌기(serif), 글자크기 등의 폰트 민감도(font sensitivity)에 따라 OCR 의 인식율과 속도를 좌우한다 [2]. 특히 한글의경우, 영문자와는 달리 조합 문자라는 구조적 특성으로 인하여 문자의 기본 성분인

획과 획간의 접촉 현상 및 기본 문자들의 종류와 수효에 따라 조합되는 문자의 형태가 많고 다양한 속성을 지니고 있어서 한글 인식 문제를 어렵게 만들고 있다[10,11,12]. 따라서, 본 논문에서는 OCR 한글 폰트의 인식 과정에서 수반되는 이러한 문제점들을 극소화하여 인식을 용이하게 할 수 있는 OCR 한글 폰트를 설계하였다. 이때 한글 자형의 심미성 과 폰트의 인식율을 동시에 만족시키기 위하여, 벌수의 개념을 절충한 공간 개념 [14,15] 방식을 채택하였다. 또한 본 논문에서 설계한 OCR 한글 폰트로 시험용 한글문서를 작성하여 이미 개발되어 실효를 거둔 한글 문서 인식 소프트웨어 [10,11]로 인식 실험을 수행하여 OCR 한글 폰트로서 만족할만한 결과를 얻었다.

## 2. 한글 자획의 특성 분석

문자 구성은 자획으로 되어있으며 자획이 갖는 특성은 문자 인식에 중요한 요소가 된다. 따라서 자획의 특성을 다음과 같이 획의 굵기 및 획의 수효에 따라 고찰해 보기로 한다. 첫째, 그림 1 과 같이 서체의 종류에 따라 자획의 굵기가 다르거나, 같은 종류의 서체에 있어서도 가로획과 세로획간에 굵기비율이 심한 불균형을 이루고 있다. 둘째, 한글의 경우 2~7 개의 기본 자모가 하나의 문자를을 형성하기 때문에 그림 2 (a)에서와 같이 가로획의 수효가 2:7의 비율이거나, 그림 2 (b) 의 경우와 같이 세로획의 수효가 2:6의 비율로 문자가 구성될 수도 있다. 이에 따라 극심한 획의 밀도 변화를 보여주고 있으며, 기본 자모의 조형이 좌우되고 있음을 알 수 있다. 특히 자획의 밀도가 높은 문자의 경우 획이 가늘어져서 연결성이 손실되거나,다른 획과의 접촉 가능성이 높아져 문자 인식을 어렵게 하기 쉽다.

앞서가고 앞서가고 앞서가고

(a) 고 딕 체 (b) 그라픽 체 (c) 환타일 체

그림 1. 자획의 굵기에 따른 불균형 .

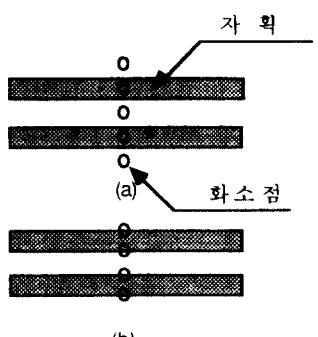
그 를 니 빼

(a) (b)

그림 2. 자획의 수효에 따른 획의 밀집 현상.

## 3. 문자 영상의 입력 특성

문자 인식을 위하여 문자영상을 입력 시킬때에 스캐너의 해상도에 따라 문자 영상의 입력 상태가 변화될 수 있다. 그림 3은 자획의 굵기와 획의 간격이 스캐너의 해상도와 같은 경우, 문자 영상에 끼치는 영향에 대하여 설명한 것이다. 특히, 획의 굵기및 획의 간격이 스캐너의 해상도와 비슷한 경우 그림 3 (a)와 같이 스캐닝 위치가 획의 위치와 일치할 때에는 자획이 분명하게 나타난다. 그러나 그림 3 (b) 와 같이 스캐닝 위치가 획의 위치와 일치하지 않을때에는 화소점 값의 흑, 백이 분명하지 않게 된다. 화소점의 값이 흑으로 취급되면 두개의 획이 접촉되는 것으로 나타나며, 화소점의 값이 백으로 취급되면 두개의 획이 손실되어 나타나지 않게 된다. 즉, 문자 영상의 입력 과정에서 자획의 손실과 자획간의 접촉 현상을




(b)

그림 3. 스캐닝 과정에서의 획의 손실 및 접촉 현상.

발생시켜 문자의 오인식을 초래하는 요인이 될 수 있다. 따라서 획의 굵기 및 간격의 크기는 해상도보다 충분히 커야 된다. 임의의 획이나 획간의 공간이 최소한 2 화소점 이상으로 작성 된다면 입력 과정에서 획의 손실 및 접촉을 피할 수 있다. 이때 필요한 2화소점을 1 스롯(slot)이라고 정의하면, 수평 밀도가 가장 높은 /불/자의 경우를 위하여 13스롯이 필요하며, 수직 밀도가 가장 높은 /빼/자의 경우에서는11 스롯이 필요하게 된다.

## 4. OCR 한글 폰트의 설계

일반적으로 컴퓨터를 이용하여 폰트를 설계하는 방법으로는 크게 2가지로 분류할 수 있다. 첫째, 화소들의 행렬로서 문자 패턴을 기술하는 비트맵 폰트(bitmap font)가 있다. 이러한 폰트는 특정 크기를 갖고 있어 정교한 패턴 묘사를 할 수 있다는 잇점은 있으나 출력 장치가 고해상도의 레이저 프린터이거나 보다 큰 크기의 폰트를 요하는 경우, 비트맵 폰트의 표현를 위하여 막대한 기억장소를 요하게 된다. 또한 확대,축소등의 폰트 변환시에 문자 패턴의 품질 저하를 초래한다. 둘째, 직선과 Bezier curve [7,9]와 같은 그래픽 프리미티브 (graphic primitive)들로 각 문자패턴을 윤각선만으로 정의하는 윤각선 폰트(outline font)가 있다. 이 폰트는 비트맵 폰트보다 압축된 형태로 문자를 기술 할 수 있다. 또한, 레이저 프린터의 기술 언어 (PDL, Page Description Language) 인 포스트 스크립(postscript) 언어로 폰트의 변환을 용이하게 조정할 수 있으며, 이때 포스트스크립 프로그램에서 벡터 정보를 주어 문자를 그리는 알고리즘적 폰트를 얻게 된다 [5,6,7,8]. 본 논문에서는 두번째 방식을 이용하여 윤각선에 따라 한글 자소에 대한 원시 폰트를 생성하고, 그림 4와 같은 과정으로 OCR 한글 폰트를 설계하였다. OCR 한글 폰트를 설계하는 과정에서 한글 자획의 특성과 스캐너 특성등을 고려하여 임의의 획 또는 획과 획사이의 간격을 최소한 1 스롯 이상으로 문자 도안을하였다.

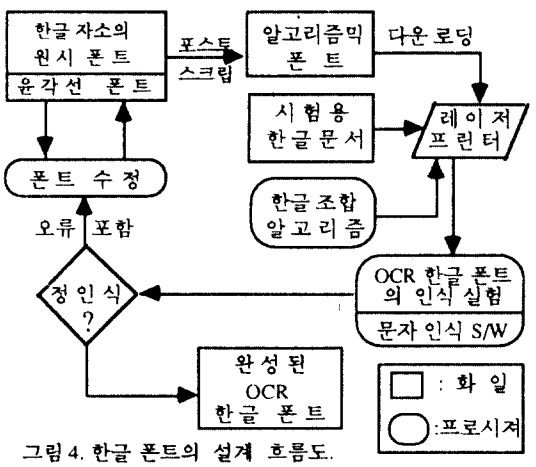


그림 4. 한글 폰트의 설계 흐름도.

이에 따라서 그림 5 (a)에서는 /째/자와 같이 자획간의 접촉 부분을 줄이고자 새리프와 같은 자획의 돌기 부분을

그림 5. 한글 자획의 속성에 따른 문자 설계.

깎아내렸고, 획의 밀도가 높은 문자에서의 스롯 수를 최소화 하기 위하여 그림 5 (b)처럼 /빼/자에 필요한 11 스롯을 9 스롯으로 줄여 인식을 용이하게 설계하였다.

특히, 한글 폰트로서 적합하게 설계하기 위해서는 한글자형의 심미성과 폰트의 인식율을 동시에 만족시켜야한다. 이 목적을 위해서는 공간개념[14,15]의 도입이 적합하나 폰트의 기억 공간이 너무 방대하게 된다. 따라서 본 논문에서는 한글 폰트의 문자 설계 과정에서 모든 문자에 대하여 충분한 벌수의 기본 자소들을 마련한 벌수개념을 절충하여 공간 개념 방식을 채택하였다. 즉, 획의 작성 과정에서는 초성, 중성, 종성이 차지하고 있는

그림 6. 공간 개념에 입각한 한글 자형.

개별 공간, 획간의 공간, 획의 밀집 상태 등을 고려해서 문자를 설계하였다. 그림 6(a)는 초성 획의 밀도에 따라서 중성 " ㅐ"의 위치를 달리 준 예이며, 그림 6(b)는 중성의 돌출부가 초성의 개별 공간을 파고드는지의 여부에 따라 문자를 설계한 예이다. 이에 따라서, 본 논문에서 설계한 OCR 한글 폰트에 대하여 총 480 가지의 기본 자소 원시폰트를 설계 하였으며, 설계된 기본 자소로 조합 가능한 한글 자형은 총 11,172자가 된다. 이렇게 작성된 원시 폰트중 한국 표준 연구소에서 선정한 한글 사용 문자[16]를 그림 4의 시험용 한글 문서로 작성하여 설계한 OCR 한글 폰트에 대한 인식 실험을 하되, 오류가 발생되었을시에는 원인 분석을 하여 설계된 원시 폰트를 수정하고 문자 인식 소프트웨어[10,11]를 통해 재인식 실험을 거쳐 만족할만한 수준의 OCR 한글 폰트를 얻게 되었다.

## 5. 결 론 및 기대효과

본 논문에서는 스캐너의 특성과 한글 자획이 갖고있는 다양한 속성등을 분석하여 획의 손실 및 획과 획간의 접촉 현상을 극소화 시키고, 한글 자형의 미적 요소와 폰트의 인식이 용이하도록 총 11,172자가 조합 가능한 OCR한글 폰트를 설계 하였으며 한국 표준화 연구소에서 선정한 2,350자로 OCR 한글 문서를 작성, 인식 실험을 하여 양호한 결과를 얻게 되었다. 이러한 결과는 한글 문서 입력의 기계화를 용이하게 하며, 인쇄 매체를 사람과 컴퓨터가 공유 할수 있는 길을 마련하였다. 또한 정보화 사회에서 폭주하는 문자 정보들의 입력 과정을 자동화하는데 기여 하리라 기대된다.

## [참 고 문 헌]


[1] C. H. Suen, "Character Recognition by Computer and Application," Handbook of Pattern Recognition and Image Processing, pp 569 - 578, 1986.

[2] S. Kahan, T. Pavidis, H. S. Baird, "On the Recogniton of Printed Characters of Any font and size, " IEEE Conf. on PR&IP, pp 274-287, 1987.

[3] K. Sato, I. Isshiki, A. Ohoka, K. Yoshida, "Hand-Scan OCR with a one-dimentional image sensor ," PR, VOL.16, NO. 5, pp 459 -467, 1983 .

[4] F. W. M. Stentiford, "Automatic Feature Design for Optical Character Recognition Using an Evolutionary Search Procedure," PAMI, VOL.7, NO.3, pp 349 - 355 , 1985.

[5] Adobe System,PostScript Language Tutorial and Cookbook, Addison-Wesley, 1985.

[6] Adobe System, PostSript Language Reference Manual, Addison-Wesley, 1985.

[7] Altsys Corp., Fontographer User's Guide,1988.

[8] Apple Computer, Inside Macintosh VOL-I, Addison-Wesley, 1985.

[9] William M. Newman, R. F.Spoul, Principles of Interactive Computer Graphics , Mcgraw-Hill, 1979.

[10] K. H. Lee, K. B. Eom, R. L. Kashyap,"Character Recognition Using Attribute Grammar," IEEE Proc., Computer Vision & Pattern Recognition, Ann Arbor, Michican, pp 418 -423, June 1988.

[11] K. H. Lee, K. B. Eom,R. L. Kashyap,"Recognition of Korean Characters," Proceedings of the 30th Midwest Symposium on Circuit and System, NY, 1987.

[12] 이 귄 하, 속성에 구속을 받는 문법을 이용한 문자 패턴 인식, 인하대학교 공학 박사 학위 논문 ,1981.

[13] 황 종 선, 원 유 헌, 박 도 순 외 8 명, 광학식 문자 인식을 위한 한글 수서 문자의 표준을 위한 연구, 공업 진흥청, 1985.12.

[14] 송 현, 한글 자형학 ,디자인 출판부,1985 . 9.

[15] 김 홍 련, 문자 디자인 , 한글 레터링 디자인 시리즈 NO. 1, 미진사, 1976 - 1987.

[16] 한국 표준 연구소, 한글.한자 코드 표준화에 관한 연구 , 과학 기술처, 1987'.7.