다중 언어에서 다중 활자체 및 다중 크기의 문자 인식을 위한 2 계층 분류기

지수영*, 문경애*, 오원근*, 김태윤**
*전자통신연구소 부설 시스템공학연구소
**고려대학교 전산과학과

A Two-Layer Classifier for Recognition of Multi-font and Muiti-size Characters in Multi-lingual Documents

Su-Young Chi*, Kyung-Ae Moon*, Weon-Geun Oh*, Tai-Yun Kim**
*Systems Engineering Research Institute / ETRI
**Department of Computer Science & Engineering, Korea University

요 약

본 논문에서는 2 계층 분류기를 이용하여 일반적인 문서(보고서, 책, 잡지, 워드프로세서에서 출력된 양식) 내의 다중 크기 및 다중 활자체의 인식을 위한 효과적인 방법을 제안하고 구현하였다. 다중 언어 문자를 효과적으로 인식하기 위한 2 계층 분류기를 제안 하였는데 이는 폰트 독립적 분류기와 폰트 의존적 분류기로 구성되어 있다.

제안된 방법의 성능 평가를 위하여 사무실에서 많이 사용하는 59 종류의 폰트와 각 폰트 당 3 가지 크기의 글꼴과, 스캐너에서 지원되는 3 가지 농도의 총 489 개의 서로 다른 부류를 갖는 3,593,172 자를 대상으로 학습 시킨 뒤에 일반 문서를 가지고 펜티엄 PC 상에서 인식 실험을 수행 하였다. 실험 결과, 2 계층 분류기를 갖는 시스템에서 96-98% 의 인식률과 초당 40 자 이상의 인식 속도를 보여줌으로써 일반적인 문서에서 다중 크기 및 다중 활자체의 문자 인식에 매우 실용적인 가치가 있음을 확인 했다.

1. 서론

일반적으로, 우리가 취급하고 있는 문서는 한글 뿐 아니라 영어, 숫자, 특수 기호, 한문 등 다양한 문자들이 다양한 크기와 활자체로 혼용되어 있기 때문에, 이러한 형태의 문서에서 사용되고 있는 활자체를 알 수 없을 때 인식은 그리 쉽지 않다. 또한, 각종 활자체의 통계적인 특성들이 조금씩 다르기 때문에 활자체의 종류에 무관한 모든 경우에 적용할 수 있는 공통적이고 구체적인 특징을 발견 하기란 쉽지 않다. 과거 몇 년 동안 이러한 문제점을 극복 하고자 많은 연구가 진행되어 몇몇 인식 시스템들이 좋은 결과를 보여 주고 있다[1,2,3,4]. 그러나, 이러한 연구들은 단일 폰트만을 위한 인식 방법을 제공 했으며, 최근 들어 다중 언어, 다중 활자체 및 크기의 문자들을 효과적으로 인식하기 위한 연구가 있어 왔다[5,6,8].

본 논문에서는 2 계층 분류기를 이용하여 다중 언어에서 다중 크기 및 다중 활자체의 인

식을 위한 효과적인 방법을 제안 한다. 다중 크기의 문자를 인식 하기 위하여 입력된 문자 형태를 일정한 크기로 정규화 하여 인식한다.

여기서 사용된 4 개의 활자체는 HWP, WORD, MAC, BITMAP 이며, 다시 각 활자체는 명조, 고딕, 궁서로 전체 표준 문자는 3107 자 이며 인식한 문자는 13,500 자를 인식하였고 인식률은 96-98%이고 인식 속도는 초당 35-40 자 이다.

본 논문의 구성은, 2 장 에서는 문자 인식에 사용되는 표준 특징을 추출하는 방법에 대하여 자세하게 기술 하였으며 3 장 에서는 추출된 특징을 바탕으로 하여 일반적인 문서내의 다중 크기 및 다중 활자체의 효과적인 인식을 위한 2 계층 분류기에 대하여 기술 하였으며, 4 장 에서는 실험 및 결과 분석으로써 실험 환경과 인식된 결과를 분석해 보았으며, 마지막은 결론으로 이 논문이 보완해야 할 부분에 대해서 기술을 하였다.

2. 문자 특징 추출

2.1 그물눈 특징 추출 (Mesh Feature)

그물눈 특징 추출은 한 영상을 지역 영역의 크기가 N x M 영상으로 균일하게 분할 한 후 각 부분 영역 내에 포함된 문자 정보의 화소 수를 계산하여 이 값을 특징 벡터로 사용하는 방법이다[7]. 일반적으로 이 방법은 구현이 간단하여 문자 인식에 흔히 사용되는 특징이다.

본 논문에서는 48 x 48 크기의 문자 영상을 16 x 16 크기의 이차원 부분 영상으로 분할하고 각 부분 영역에 포함된 문자 정보의 화소 수 256 개를 그물눈 특징으로 사용 하였다.

2.2 평균 문자 특징 추출 (Average Feature)

이 방법은 앞 절에서 추출한 256 개의 그물눈 특징을 가지고 격자 모양으로 128 개의 값을 취하여 특징 벡터를 추출하는 방법이다.

이 특징은 본 논문에서 사용한 문자 영상 데이타 베이스를 가지고 동일한 문자들에 대하여 128 개의 특징을 모두 누적하여 평균한 벡터 값이다. 그림 1 은 256 그물눈 특징에서 격자 모양으로 128 개를 추출한 특징을 나타내고 있다.

[가 110]

```
. 3 . 51 . 51 . 51 .  0 .  0 . 50 .  0
6 . 102 . 102 . 104 . 96 .  0 . 101 .  0 .
. 0 .  0 .  0 .  3 . 128 .  0 . 101 .  0
0 .  0 .  0 .  3 . 128 .  0 . 101 .  0 .
. 0 .  0 .  0 .  3 . 128 .  0 . 101 .  0
0 .  0 .  0 .  3 . 128 .  0 . 101 .  0 .
. 0 .  0 .  0 .  3 . 128 .  0 . 101 .  0
0 .  0 .  0 .  5 . 80 .  0 . 102 . 50 .
. 0 .  0 .  0 .  8 . 48 .  0 . 104 . 116
0 .  0 .  0 . 56 .  0 .  0 . 101 .  0 .
. 0 .  0 .  1 . 132 .  0 .  0 . 101 .  0
0 .  0 .  6 . 97 .  0 .  0 . 101 .  0 .
. 0 .  2 . 133 . 16 .  0 .  0 . 101 .  0
2 . 121 . 64 .  0 .  0 .  0 . 101 .  0 .
. 71 . 66 .  0 .  0 .  0 .  0 . 101 .  0
0 .  0 .  0 .  0 .  0 .  0 . 83 .  0 .
```

그림 1: 평균 문자 특징

2.3 배타적 OR 특징 추출 (Exclusive OR Feature)

이 방법은 문자간의 유사성 및 상이성을 특징으로 집적하여, 문자를 구별하는 새로운 특징 추출 방법이다. 즉, 각 문자 자체의 고유한 특성(문자의 획 부분과 여백 부분)을 상이한 문자들과 비교하여 자기 자신 만을 나타내는 부분을 특징으로 집적하여 특징을 추출하는 방법이다.

문자 A 를 기준 문자라 하고 Z 를 A 를 제외한 모든 대상 문자의 집합이라 하면, Z={B,C,D, ...}이다. A 의 글자 부분과 대상 문자들의 글자 부분을 배타적 OR 하면 A 자신의 글자 부분만 남는다.

If (A == 1 && (A XOR Z))

then Cnt++ ..... (1)

(여기서 Cnt는 자기 자신의 누적좌표 변수)

동일하게 A의 글자가 아닌 여백 부분과 대상 문자들의 여백 부분을 배타적 OR 하면 A 자신의 고유한 여백 부분만 남는다.

If (A == 0 && (A XOR Z))

then Cnt– ..... (2)

이렇게 하여 글자 부분 (+) 과 여백 부분 (-)에서 큰 순서로 12개씩 정렬하여 총 24개의 특징을 추출하는 것이다. 그림 2는 문자 '가'의 배타적 OR 특징을 나타낸 것이다.

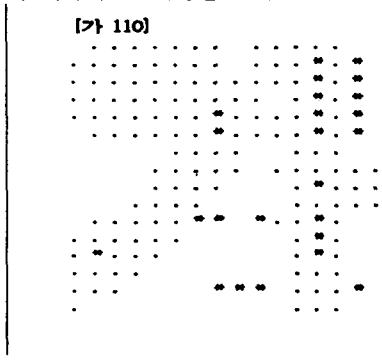

그림 2: 배타적 OR 특징

3. 2 계층 분류기

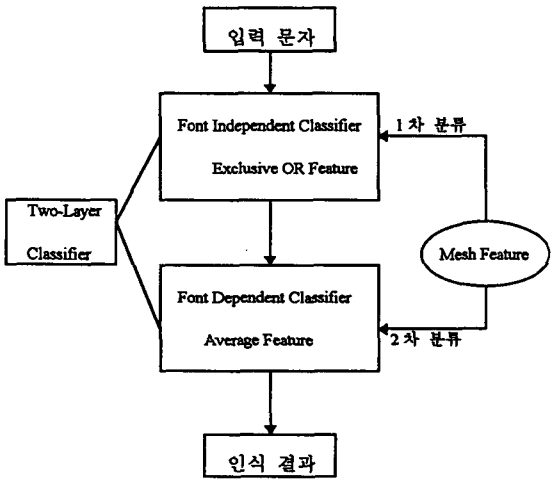

그림 3: 2 계층 분류기 구조

이 장에서는 앞 절에서 추출된 특징을 바탕으로 하여 일반적인 문서(보고서, 책, 잡지, 워드프로세서에서 출력된 양식) 내의 다중 크기 및 다중 활자체의 인식을 위한 효과적인 2 계층 분류기를 제안한다.

본 논문에서 사용한 2 계층 분류기의 구조는 그림 3 과 같다.

3.1 폰트 독립적 분류기 (Font Independent Classifier)

다양한 종류의 크기 및 활자체가 포함된 문서를 인식함에 있어서 인식 속도와 인식률의 향상을 위하여 여러 단계의 분류기를 두고 인식 할 수 있다. 그러나 인식 단계들이 특정한 폰트의 특징에 의존적이다 보면 인식 시스템 자체가 범용성을 가질 수 없고 매우 제한된 폰트 영역에서만 인식이 가능한 문제점이 있다. 본 논문에서는 이러한 문제점을 해결하기 위하여 모든 폰트 집합에서 각 문자 자체의 고유한 특징 만을 집적하여 학습 시킴으로써 폰트에 독립적이고 일반적인 문서에 적용 시킬 수 있는 폰트 독립적 분류기를 제안 한다.

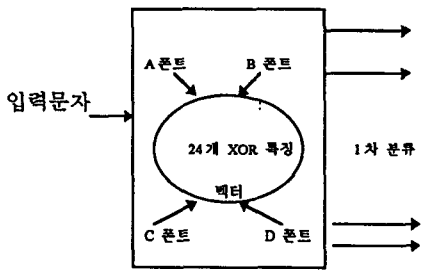

그림 4: 폰트 독립적 분류기

이 분류기는 학습된 24개의 배타적 OR 특징 벡터를 사용하여 1차 대분류를 수행한다. 그림 4는 폰트 독립적 분류기를 나타낸다

3.2 폰트 의존적 분류기 (Font Dependent Classifier)

이 절에서는 앞 절의 폰트 독립적 분류기에서 1 차적으로 분류된 문자를 대상으로, 입력 문자가 속해 있는 폰트를 식별하여 인식 하는 폰트 의존적 분류기를 설명한다. 이 분류기에 사용되는 특징으로 각 폰트가 갖고 있는 다양한 크기 및 활자체를 모두 누적하여 평균 한 후 128 개의 특징 벡터를 갖는 평균 문자 특징 피처를 사용한다. 먼저, 입력된 문자의 폰트를 식별하고 그 폰트에 포함된 활자체를 식별하여 정확하게 분류를 하는 것이다.

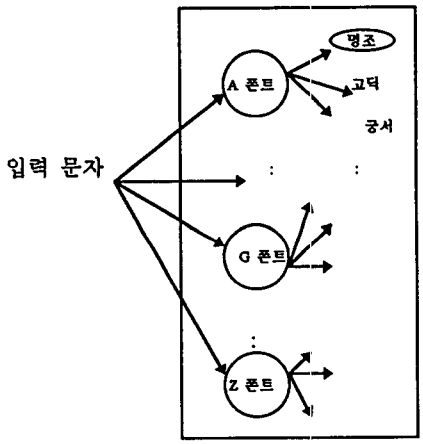

그림 5: 폰트 의존적 분류기

4. 실험 결과 및 분석

4.1 다중 언어 문자 영상 데이타 베이스

본 논문에서 사용한 데이타는 여러 가지 크기와 활자체의 문자들의 특성을 분석하기 위해서 KS 코드 테이블에 있는 영문, 숫자, 한글, 한자 7348 자로 구성된 문자 코드표를 사용하였다. 각기 다른 워드 프로세서에서 폰트별로 10, 12, 14 포인트의 크기를 레이저 프린터로 A4 크기의 종이에 문자를 각각 출력하고 3 종류의 스캐너(HP Flatbed 스캐너, Microtek 스캐너, PhotoScan10 스캐너)를 사용하여 각 스캐너의 3 가지 농도로 300 dpi 의 해상도로 입력 받은 영상을 각각 사용하였다.

4.2 실험 결과

본 논문에서는 제안된 2 계층 분류기에 사용된 문자 특징 추출 및 학습을 위하여 삼보 워크스테이션에서 사용하였으며, 제안된 방법을 펜티엄 PC Windows 95 상에서 Visual C4.0++ 로 구현하여 실험을 수행하였다.

실험 방법으로는 4 종류의 워드 프로세서에서 글자 크기가 중간 크기인 10 point 로 각 활자체 당 2 혹은 3 가지 체 모두 1,500 자씩으로 모두 13,500 자로 된 문서를 레이저 프린터로 출력한 후 광학 스캐너로 영상을 입력 받아 인식 하였다. 표 1 에 2 계층 분류기의 인식률 을 비교 하였다.

표 1 각 활자체 별 인식률

워드프로세서	활자체	2 계층 분류기 인식률	
		1 차분류	2 차분류
Sample 1	명조체	97.4%	96.8%
	고딕체	97 %	96.4%
Sample 2	명조체	98.2%	97.2%
	고딕체	97.3%	96.3%
Sample 3	명조체	98.8%	98.1%
	고딕체	98.4%	97.8%
Sample4	바탕체	97.4%	96.8%
	돋움체	98.2%	97.4%
	궁서체	93.5%	92.8%

Sample 1 은 아래아 한글로 작성된 문서
Sample 2 는 한글 워드로 작성된 문서
Sample 3 은 메킨토시로 작성된 문서
Sample 4 는 윈도우 에디터로 작성된 문서

이 비교에서 보면 궁서체가 여러 활자체 중에서 가장 낮은 인식률을 보이는데 이것은 다른 활자체에 비해서 굵기가 굵고 형태도 매우 달라 각 문자 전체의 뼈대를 특징으로 사용하는 폰트 독립적 분류기에서 다른 문자로 인식을 하게 되어 인식률이 낮게 나타났다 또한, 충분하지 못한 학습 데이타를 사용한 결과로 학습 데이타 수를 증가 하면 더 좋은 결과를 기대 할 수 있다.

5. 결론

본 논문에서는 2 계층 분류기를 이용하여 일반적인 문서(보고서, 책, 잡지, 워드프로세서에서 출력된 양식) 내의 다중 크기 및 다중 활자체의 인식을 위한 효과적인 방법을제안하고 구현 하였다. 2 계층 분류기는 폰트 독립적 분류기와 폰트 의존적 분류기로 구성된다. 먼저, 입력된 문자에 대하여 문자의 공통적인 뼈대를 특징으로 하는 폰트 독립적 분류기에서 1 차 분류를 하고 폰트 마다 각기 고유한 특징을 갖는 폰트 의존적 분류기에서 2 차 분류를 한 후, 인식 결과를 얻는다. 제안된 방법의 성능 평가를 위하여 다양한 크기와 활자체로 구성된 문서를 가지고 실험을 하였다. 그 결과 2 계층 분류기를 갖는 시스템이 일반적인 문서에서 다중 크기 및 다중 활자체의 문자 인식에 매우 실용적인 가치가 있음을 확인했다.

향후 과제는 본 논문에서와 같이 인식 해야 하는 문자의 수가 많고 문자들 사이의 유사성과 다양한 활자체를 인식하기 위해서는 다양한 활자체의 문자들의 특징을 잘 표현하는 피처(Feature)들의 연구와 또한, 다수의 서로 다른 인식기를 동시에 사용하고, 그들의 인식 결과를 결합 함으로써 인식 성능을 향상 시키려는 다수 인식기 시스템에 대한 연구가 필요하다 하겠다.

참고문헌

1.김우태, 윤병식, 박인규, 진성일, "인쇄체 한글 문자 인식을 위한 특정 성능의 비교," 한국정보과학회 논문지, 제 20 권 제 8 호, pp. 1103-1111, 1993.

2.이성환, "다양한 활자체 및 크기를 갖는 대용량 한글의 고속 인식을 위한 최적 트리 분류기," 한국정보과학회 논문지, 제 18 권, 제 3 호, pp.1083-1092, 1993.

3.권재욱, 조성배, 김진형, "신경망 기법을 이용한 다중 크기 및 다중 활자체 한글 문서의 인식," 제 3 회 영상 처리 및 이해에 관한 워크샵 발표 논문집, pp. 129-136, 1991.

4.김상우, 전윤호, 최종호, "신경 회로망을 이용한 인쇄체 한글 문자의 인식," 전자 공학 회지, 제 27 권, 제 2 호, pp. 65-71, 1990.

5.송희헌, 김종수, 이성환, "계층적 신경망을 이용한 다중 언어, 다중 활자체 및 다중 크기의 대용량 문자인식," 한국정보과학회 추계 학술 발표 논문집, Vol. 22, No. 2, pp. , 1995.

6.문경애, 지수영, 오원근, "혼용 문서에서의 유사 문자 분류," 제 5 회 한글 및 한국어 정보처리 학술 발표 논문집, pp. 485-492, 1993.

7.S. Mori, K. Yamamoto, and M. Yasuda, "Research on Machine Recognition of Handprinted Characters," IEEE Trans, on PAMI-6 No. 4, pp. 386-405, July, 1984.

8.S. Kahan, T. Pavlidis and H. S. Baird, "On the Recognition Printed Characters of any Font and Size", IEEE Trans. On Pattern Analysis and Machine Intelligence, Vol. 2, No. 5, pp. 274-288, Sep. 1987.