순위에 따른 가중 페널티를 이용한 처방전의 특정 정보 인식

*이병모° *강동구 **김성우 *차의영
*부 산 대 학 교 일반대학원 전자계산학과
**부 산 대 학 교 일반대학원 정보시스템공학과
{lbmo, dkkang1, swkim, eycha}@harmony.cs.pusan.ac.kr

Recognition of the Specific information in Medical Prescription Using Weighted Penalty by Order

*Byong-Mo Lee° *Dong-Gu Kang **Seong-Woo Kim *Eui-Young Cha
*Dept. of Computer Science, Pusan National University
**Dept. of Information System, Pusan National University

요 약

본 논문에서는 한글, 숫자, 영문자, 기호 등이 혼용된 컬러용 폼지 또는 A4지 처방전을 자동으로 인식하는 시스템을 설계하는 방법을 제안한다. 이를 구현하기 위해 먼저 처방전을 스캐너를 이용하여 스캔하고 컬러 정보를 이용하여 회전된 처방전을 보정한 다음 처방전의 종류를 결정한다. 그리고, 문자의 형태학적 특징에 따라 한글과 그 외의 문자(비한글)를 구분한다. 그리고, 구분된 비한글의 경우는 다양한 특징 벡터를 이용한 가중 페널티 방법을 이용하여 인식한다.

1. 서 론

지금까지 처방전 인식에 관한 기술은 새로운 분야였다. 2000년 7월 의약분업과 동시에 발생한 원외 처방전은 환자가 약국에 가지고 가면, 약사는 처방전의 필요한 정보를 다시 컴퓨터로 옮겨야 했으며, 컴퓨터에 익숙하지 않은 약사들은 자연히 데이터를 저장하는데 오랜 시간이 소모되었으며, 때로는 잘못 입력하기도 했다. 이에 2차원 바코드에 처방전의 내용을 함축한 후 이를 스캐너를 이용하여 읽어 들이는 방식이 소개되었으나, 병원의 협조 없이는 기술의 사용이 무의미했다. 따라서, 병원에 의존하지 않고 단지 처방전을 스캐닝함으로써 처방전의 정보를 인식하는 기술을 개발하기에 이르렀다.

처방전은 크게 2종류가 있는데, 폼지에 쓰여진 것과 A4지에 쓰여진 것이 있다. 폼지에 쓰여진 것은 컬러 서식 위에 인쇄체 문자를 기록한 것으로, 문자와 서식이 서로 겹치는 경우가 자주 발생한다. 따라서, 이와 같은 문제를 해결하기 위한 방법으로는 모폴로지를 이용한 방법[1], 인식기를 이용한 방법[2], 그리고, 색상 정보를 이용한 방법[3,4] 등이 있다. 모폴로지를 이용하는 방법은 문자의 외각선 정보가 소실될 위험이 있고, 인식기를 이용한 방법은 인식하는데 시간이 오래 걸리는 단점이 있다. 따라서, 본 실험에서는 컬러 정보를 이용한 방법을 통하여 문자를 분리하고 있다.

한글 여부는 문자의 구조적 특징을 이용한다. 처방전에는 한글과 비한글이 함께 쓰이고 있는데, 인식률을 높이는 차원에서 한글과 비한글을 따라 분리하여 각각의 인식기를 통하여 인식한다.

본 논문의 2장에서는 전체 시스템 구성을 보이고, 3장에서는 문자 분리에 관하여 설명하고, 4장에서는 한글 여부를 파악하고, 5장에서는 비한글 인식기에 대해서 설명한다. 6장에서는 실험 결과 및 분석을 하고, 끝으로 7장에서는 본 논문의 결론을 맺는다.

2. 전체 시스템 구성

본 논문에서 제안하는 설계도는 그림 1과 같다. 처방전을 스캔하는 부분, 타입을 결정하는 부분, 기울기 보정, 그리고, 문자 분류에 따른 문자 인식으로 이루어진다.

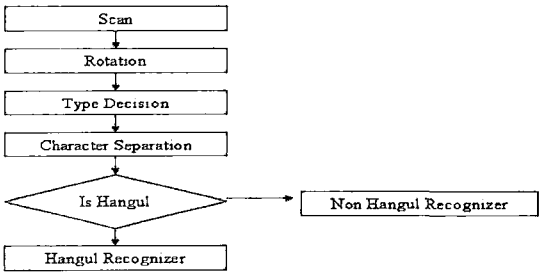
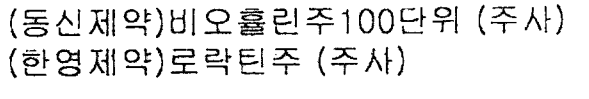


그림 1. 시스템 순서도

3. 문자 분리

처방전의 약품명을 기입하는 항목에는 다중 라인으로 약품명이 기입된다. 따라서, 가로 방향으로 문자를 분리한 다음, 세로 방향으로 문자를 분리한다. 가로 방향 분리는 가로 프로젝션을 통하여 이루어지고, 세로 방향 분리는 각 라인에 대해서 세로 프로젝션을 한다. 그리고, 분리된 각각의 문자에 대해서 최종 크기 보정을 함으로써 정확한 크기의 문자를 추출한다. 다음은 이러한 과정을 통하여 추출한 문자들이다[그림 2].



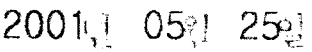
a. 원영상

b. 가로 방향 분리 후 세로 방향 분리

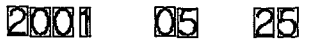
c. 최종 크기 보정

그림 2. 문자 분리

문자 분리 때의 이진화는 HSI 컬러 정보를 이용한다. 폼지는 미리 인쇄된 글자와 프린터에 의해 씌어진 글자가 겹쳐져 쓰지는 경우가 발생하기도 하는데, 이 경우 단지 그레이 영상으로 처리해서는 제대로 문자를 분리하기가 힘들다[그림 3]. 그래서, 본 논문에서는 HSI의 saturation과 intensity 정보를 이용하여 문자를 분리한다[3,4].



a. 원영상



b. 문자 분리

그림 3. 문자의 접촉과 분리의 예

4. 한글 여부

처방전에서 사용되고 있는 문자는 한글, 숫자, 영문자, 기호로 구성되어 있는데, 문자가 한글인지 아닌지를 구분하는 것은 매우 중요하다. 한글은 완성형의 경우 2350자로 구성되어 있어서, 한글을 비한글과 구분하여 각각 인식하는 방법을 사용한다. 본 논문에서는 한글이 다른 문자에 비해서 평균 문자 폭이 넓은 점과 앞 뒤 문자의 종류를 보고 한글의 여부를 판단한다.

5. 비한글 인식

처방전에서 사용되는 문자에는 한글이 차지하는 비율이 약 18%, 숫자가 차지하는 비율이 68%, 영문자의 경우 6%, 그리고, 기호는 약 8% 정도의 비율을 차지한다. 따라서, 비한글이 차지하는 비율이 82%로 비한글의 인식률이 전체 인식률에 크게 영향을 주고 있음을 알 수 있다.

비한글 문자 인식기는 문자를 11가지의 세부 특징 벡터와 그물망을 이용한 순위에 따른 가중 페널티 방법을 통하여 구현하였다[5,6].

5.1 순위에 따른 가중 페널티

각각의 특징 벡터와 그물망을 통한 일대일 비교 오차에 따른 오차 순위를 메기고, 순위에 따른 자연 log 함수를 이용한 가중 페널티를 적용하여 합이 최소 페널티를 가지는 것을 인식값으로 했다. 특징 벡터는 서로 비슷한 문자를 구별하는데 특히 효과를 발휘하며, 페널티로 자연 로그 함수를 적용하여 순위가 낮은 것에 대해서는 차별화를 크게 하고, 순위가 높은 것에 대해서는 차별화를 적게 하여, 비록 한 두 개의 특징 벡터에서 순위가 높다하더라도 대부분의 다른 특징 벡터의 순위가 낮으면 최종적으로 낮은 페널티를 적용 받도록 하였다.

순위에 따른 가중 페널티(P_W)는 다음과 같은 식 (1)에 의해 구해진다.

$$P_W = P_S / N * \log10(rank) \qquad (1)$$

여기서 P_S는 주어진 페널티로서 경험치로 정한 값이고, N은 인식 대상의 수, rank는 순위를 나타낸다.

다음 표 1은 숫자 '8'에 대한 순위에 따른 가중 페널티의 예를 보인 것으로 표에 나온 대문자는 각 특징 벡터에 대한 페널티를 영문자 오름차순으로 나타낸 것으로 각 특징 벡터에 대한 페널티를 모두 합한 값을 종합 벌점으로 표현하였으며, 그림 4는 숫자 '0'에 대한 순위에 따른 가중 페널티의 다른 표현 예를 그래프를 이용하여 보인 것이다.

표 2는 무작위로 선택한 처방전에서의 비한글에 대한 인식결과를 보인 것이다.

6. 실험 결과 및 분석

본 논문에서 사용한 실험 데이터는 4종류의 처방전을 대상으로 하여 글꼴 크기 10PT의 문자를 200DPI 해상도로 스캔하여 인식하였다. 시스템 개발 환경은 Pentium 800MHz, Memory 512Mbyte, Windows 98에서 Visual C++ 6.0을 사용하였다.

단, 여기서 알파벳이 A쪽에 가까울수록 순위가 낮다.

표 1. 숫자 '8'에 대한 순위에 따른 가중 페널티의 예

	4방향 특징	X방향 평균	Y방향 평균	X방향 분산	Y방향 분산	X방향 0/1	Y방향 0/1	가:세 비율	가로 Proj	세로 Proj	크기	가로 유사	세로 유사	시작점	레이 블링	HT	종합 벌점
3	A	M	V	I	I	I	G	K	I	O	A	K	A	F	A	W	1318.45
6	A	G	A	C	J	D	A	K	I	F	A	K	A	F	A	I	967.46
8	**A**	**A**	**D**	**A**	**H**	**C**	**B**	**G**	**A**	**J**	**A**	**A**	**A**	**A**	**A**	**W**	**690.71**
9	A	I	J	J	G	B	D	K	Q	F	A	H	G	F	A	G	1225.00
S	A	H	G	D	B	G	C	E	D	F	A	K	A	A	A	H	932.97
)	K	R	L	V	M	N	U	Y	I	O	A	T	I	U	A	V	1881.36

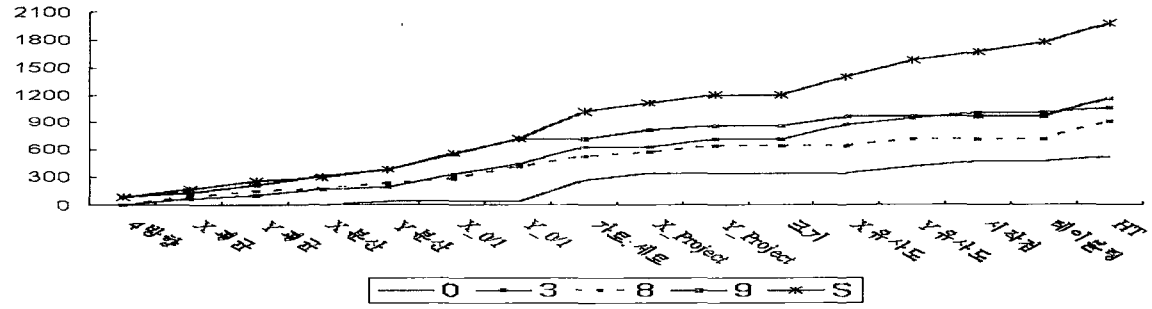


그림 4. 숫자 '0'에 대한 순위에 따른 가중 페널티 예

표 2. 비한글 인식 결과

폰트에 따른 인식률 / 문자 (테스트 데이터 수)	인식률(에러 개수)	
	폰트에 상관없는 Target Value	폰트에 따라 다른 Target Value
숫자(3147)	99.2%(24)	99.6%(12)
영문자(274)	98.5%(4)	98.9%(3)
특수기호(382)	98.9%(4)	99.2%(3)
합 계(3803)	99.1%(32)	99.5%(18)

7. 결 론

본 논문에서는 처방전을 스캔하는 단계에서 시작하여 회전한 처방전은 Hough transform을 이용하여 회전각을 구하고, 최근이웃 보간법을 이용하여 보간한다. 그리고, 처방전 타입을 결정되면, 문자를 분리한 다음, 한글 여부를 파악했다. 그리고, 비한글의 경우 다양한 특징 벡터, 즉, 가로 세로 비율, 평균, 분산값 등을 이용한 순위에 따른 가중 페널티 기법을 이용하여 인식하였다.

본 논문에서 보완해야 할 사항은 좀 더 다양한 특징 벡터를 이용해야 한다는 것과 지금 진행 중인 한글 인식기에 적용하는 것이다.

[참고 문헌]

[1] Su Liang, M.Ahmadi, M.Shridhar, "A Morphological Approach to Text String Extraction from Regular Periodic Overlapping Text/Background Images", Image Processing, 1994. Proceedings. ICIP-94., IEEE International Conference , Vol.1 , 144 -148, 1994

[2] O. Hori, D.S.Doermann, "Robust Table-form Structure Analysis Based on Box-reasoning", Proceedings of the Third International Conference on Document Analysis and Recognition, Vol. 2, pp. 218-221, 1995.

[3] Hiroyuki Hase, Toshiyuki Shinokawa, Masaaki Yoneda, Ching Y. Suen, "Character string extraction from color documents", International Journal of the Pattern Recognition Society, Vol.37,No.7,pp.53-69 2001.

[4] Yu Zhong, Kalle Karu, Anil K.Jain, "Locating Text in Complex Color Images", Proceedings of the Third International Conference on Document Analysis and Recognition, Vol. 1, pp. 146-149, 1995.

[5] 박창순, 김두영, "오프라인 필기체 숫자 인식을 위한 다양한 특징들의 성능 비교 및 인식률 개선 방안", 한국정보처리학회 논문지, Vol. 3, 제 4호, pp. 915-925, 1996.

[6] 이성환, 박희선, "고리 투영을 이용한 위치, 크기 회전 변형에 무관한 패턴 인식", 인지 과학회 논문지, Vol. 13, 제 1호, pp. 1103-1111, 1993