은행 전표에서 한글 금액 인식기의 구현

지태창, 김은진, 이일병
연세대학교 컴퓨터과학과

Korean Amount Recognizer in Bank Slips

Tae-Chang Jee, Eun-Jin Kim, Yillbyung Lee
Dept. of Computer Science, Yonsei University

요 약

본 논문에서는 은행 전표의 한글 금액열을 인식하는 시스템을 제안한다. 한글의 낱자를 인식하는 연구가 활발히 진행되고 있는데 반하여, 본 연구에서는 한글의 낱자 인식 결과를 가지고 후처리를 한 후, 그 결과를 금액 숫자의 인식결과와 통합하는 시스템을 구성하였다. 한글 낱자 인식기는 MDC(Minimum Distance Classifier) 기법을 응용한 방법을 사용하였고, 그 후처리는 금액의 구조적인 특징을 사용하였다. 마지막으로 숫자 인식기의 결과와 상호 참조하여 인식기를 완성하였다. 인식 결과를 보면 한글 금액 문자열의 낱자에 대해서는 후처리를 하기 전에는 96.29%, 후처리를 한 후에는 97.72%의 인식률을 보였고, 한글 금액 문자열에 대해서는 후처리를 하기 전에는 79.96%, 후처리를 한 후에는 98.24%의 신뢰도를 보였다.

1. 서론

[illegible] 문서의 사용이 꾸준히 증가하고 있는 상황에서도, 종이에 작성된 대부분의 기존 문서조차 적절히 처리하여 저장하지 못하고 있기 때문에 대용량의 자료 입력이 필수적인 데이터베이스 구축에서도 대부분 타자에 의한 입력이 사용되고 있는 실정이다. 이와 같은 상황에서 문서, 심볼, 테스트, 그래픽, 영상 등 문서의 기본 구성 요소를 컴퓨터로 인식하고, 더 나아가 문서 자체의 구조를 분석하여 그 의미를 이해하고자 하는 연구가 활발히 진행되고 있다. 특히, 인쇄체의 문자 단위 인식에 대한 연구는 상당한 수준에까지 진척되어 이미 실용화되고 있는 시스템이 소개된 바 있으며, 요즈음은 필기문자를 인식하고자 하는데 까지 그 영역을 넓히고 있다.

이에 반하여 실제 문서 인식에 필수불가결한 문서의 구조를 분석하고 이해하는 연구, 인식을 효과적으로 수행하기 위해 영상의 질을 강화시키는 전처리에 대한 연구나 인식된 문자열의 오류교정에 대한 연구 등 실제 문서 인식 시스템 개발을 위해 필요한 연구는 상대적으로 미흡한 형편이다. 하지만, 그간의 기반 연구에 힘입어 최근에 그 필요성이 크게 고조되고 있다. 1991년부터 문서분석 및 인식에 관한 국제 학술회의 (ICDAR), 첨단 필기인식에 관한 국제 워크샵 (IWFHR)이 격년으로 개최되고 있으며, 최근에는 우리 나라와 불란서 간의 연구교류를 위한 한불 문자인식 워크샵이 개최되는 등 국내외적으로 활발한 연구개발이 진행되고 있다.

본 논문에서는 실제 문서 인식 시스템 개발을 위해 필요한 연구 중 하나인 한글 금액열 인식기를 개발하고, 은행 전표 내의 한글 금액열을 대상으로 실험하였다.

* 본 논문은 정보통신부 '산학연 공동기술개발사업'의 지원을 받음

한글 금액열의 특징을 살펴보면 다음과 같다.

첫째, 한글 낱자의 개수가 제한적이다. 금액에 쓰이는 한글의 개수는 '일...구, 십...억, 원, 정'의 16자로 상당히 적다. 그래서, 낱자 인식기의 성능을 일반적인 인식기보다는 비교적 쉽게 올릴 수 있다.

둘째, 금액 자체에 구조적인 정보가 내재되어 있다. 3장에서 언급하겠지만, 금액열이 이루어지기 위해서는 일정한 규칙을 만족해야만 한다. 이 규칙은 금액열이 자체적으로 가지고 있는 성질이고, 이 성질을 이용하여 구조적인 후처리가 가능하다.

셋째, 숫자 금액일과의 상호 참조가 가능하다. 일반적으로 은행 전표에는 한글 금액과 동시에 숫자 금액을 기입하게 되어있기 때문에 이 둘을 상호 참조 함으로써 2차 후처리를 해서 최종 결과를 생성한다.

본 논문의 구성은 다음과 같다. 2장에서는 한글 금액 인식기의 전체 구조에 대해 간단히 살펴보고, 3장에서는 한글 금액 인식기의 구성 요소에 관해 설명한다. 4장에서는 실험 환경에 관해 언급하고, 5장에서 실험 결과를 보인 후, 6장에 결론과 앞으로의 연구 과제에 대하여 살펴본다.

2. 한글 금액 인식기의 구조

한글 금액 인식기의 전체 구조는 그림 1과 같다. 우선, 은행 전표에서 한글 금액 영상에 해당되는 부분만이 추출되어 낱자 단위로 분

리된 후 한글 금액 인식기에 돌이와서, 한글 낱자 인식기의 입력으로 사용된다. 낱자 인식기의 인식 결과는 금액의 구조적 후처리기의 입력으로 사용되어 1차적인 문자열의 출력을 보인다. 은행 전표에 한글 금액열만 있는 경우에는 이 결과가 금액으로 결정된다. 하지만, 숫자 금액열이 같이 있는 경우에는 이 인식 결과를 받아들여서, 한글과 숫자 금액열의 통합 후처리기의 수행을 거쳐서 최종적인 인식 결과를 생성한다.

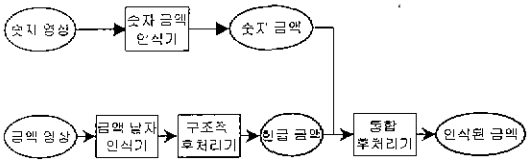

그림 1 한글 금액 인식기의 전체 구조

3. 한글 금액 인식기의 구성 요소

3.1 낱자 인식기

낱자 인식기의 구성은 그림 2와 같다.

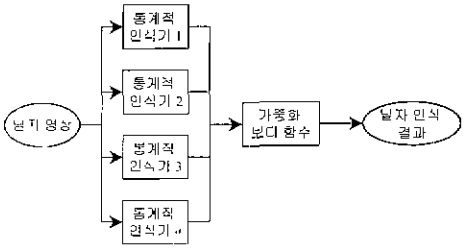

그림 2 낱자 인식기의 구성도

통계적 인식기는 모두 네 개가 사용된다. 각각의 통계적 인식기는 서로 다른 네 가지의 특징을 인식 자료로 사용하다. 이 특징은 기로, 세로 방향성분 추출의 방법[1]과 Gradient 방향성분을 이용한 특징추출 방법[2]과 오목성 특징(Up Down Left Right Hole UDLRH)[3]과 윤곽선의 위상 변화 (Gradient)를 이용한 특징 추출[3] 방법이 사용된다. 통계적 인식기의 인식방법은 MDC 방법을 사용했다. MDC 방법은 각각의 특징을 대표할 표준벡터를 구한 후, 인식하고자 하는 낱자에서 추출한 특징벡터와의 차를 구하여 그 차의 절대값이 가장 작게 나온 표준벡터의 클래스를 그 낱자의 인식 결과로 결정하는 방식이다. 표준벡터를 만들기 위해서 학습 실험에서 각각 네 개의 특징을 뽑아낸 후 각각의 특징에 대해 평균벡터를 만들고, 이 평균벡터를 LVQ1 알고리즘으로 학습시켜 표준벡터를 구했다.[4]

전체 패턴이 M개의 클래스 ($A=\{1,\ldots,M\}$) 로 이루어져 있을 때, 보다 함수는 클래스 i에 대하여 문자 인식기 e_k가 출력한 순위 r_k^i에 따라 "$M-r_k^i$"를 할당하고, 이 값을 모든 인식기에 대하여 합산한 것을 그 클래스의 보다 점수로 정하고, 이 점수가 큰 순서대로 순위를 결정하는 방식이다.[5] 가중화 보다 함수는 기본 보다 함수에 문자 인식기의 중요도 W_k에 따라서 별도의 가중치를 부여하는 방식이다. 본 논문에서는 훈련 데이터에 대한 각 개별 인식기의 인식률을 (1, 10) 사이의 값으로 변환하여 사용하였다. 가중화 보다 함수는 아래와 같이 정의할 수 있다.

$$F(x)=\max_{i\in A}(B_i(x)) \qquad (1)$$

$$\text{단, } B_i(x)=\sum_{k=1}^{K} W_k\times(M-r_k^i(x))$$

가중화 보다 함수를 거쳐서 나오는 결과 중 상위 세 개를 낱자 인식 후보로 선정한다. 이 상위 세 개 후보는 구조적 후처리기의 입력으로 사용된다. 상위 세 후보까지의 인식률이 99.5%까지 나오고, 상위 네 후보 이상의 후보를 선택해도 인식률의 증가가 눈에 띄지 않기 때문에 상위 세 후보 만 문자열 인식에 사용했다.

3.2 구조적 후처리기

구조적 후처리기는 한글 금액열에 내재되어 있는 구조적인 정보를 이용한다. 금액열의 구조적인 정보를 간단히 보면 표1과 같다. 한글 낱자 인식기의 인식 결과 중 1후보를 기준으로 해서 표1의 정보와 비교하여 틀리는 경우에 그 다음 후보로 넘어가는 방법을 사용하다.

숫자	연속 숫자 불가.
십	각 단계에서 한 번만 나옴.
백	각 단계에서 한 번만 나오고, 십의 앞에 나옴.
천	각 단계에서 한 번만 나오고, 십, 백의 앞에 나옴
만	금액열에 한 번만 나옴
억	금액열에 한 번만 나오고, 만의 앞에 나옴
원	금액열의 마지막이나 두 번째에만 나타남
정	금액열의 마지막에만 나타남

표 1 금액열의 구조적인 정보

3.3 통합 후처리기

한글 금액열의 구조적인 정보를 이용하는 후처리기는 한글 금액의 단위에 대한 후처리를 할 수 있지만, 한글 금액내의 숫자 낱자에 대한 인식 결과가 틀린 경우에는 보정이 불가능하다. 따라서, 숫자 인식기의 결과와 상호 참조하는 단계가 필요하다. 본 연구에서는 이미 개발되어 있는 숫자 인식기를 사용하였다.[5] 다중 인식기의 다단계 결합을 통한 숫자 인식기인데, 숫자 낱자의 신뢰도가 98.1% 정도 되고, 실험에 사용된 숫자열의 평균길이가 7자 정도일 때, 숫자열에 대해서 94.12% 정도의 신뢰도를 보였다.

이 숫자 인식기의 인식 결과를 받아들여서 한글 금액 인식기와 비교하는데, 중요한 것은 한글 금액 인식기의 숫자 부분만 비교 대상이 되는 것이다. 한글 인식 결과와 숫자 인식 결과를 비교하여 더 신뢰도가 높은 인식 결과를 최종 인식 결과로 선정한다.

4. 실험 환경 및 데이터

실험에 사용한 시스템은 중앙처리장치는 Pentium 150 MHz, 주메모리는 32MB였다. 운영 체제는 Linux OS 4.0을 사용하였다.

실험을 위해서 포항공대에서 제작한 PE92 한글 데이터[6]를 이용

하였다. 낱자 인식기는 PE92 100 set에서 16자씩 추출하고, 각각 50set은 학습 집합으로 50 set은 실험 집합으로 정하여 실험하였다.

금액 열은 금액의 특성상 특별히 정해진 것이 없기 때문에 임의로 문자열을 만들어서 실험하였다. 전체 문자열은 50개이고, PE92 데이터를 이용하여, 50set은 학습 집합으로 50 set은 실험 집합으로 결정하였다. 전체 사용한 문자의 개수는 320자이고, 한 문자열당 평균 6.4자가 사용되었다.

5. 실험 결과

5.1 낱자 인식률

현금 금액 낱자 16자에 대해 실험에 사용된 네 개의 통계적 인식기의 성능을 비교해 보면 다음과 같다.(표 2)

통계적 인식기	인식률
사로-세로 방향 성분	87.375%
Gradient(1) 특징	84.75%
UDLRH 특징	86.00%
Gradient(2) 특징	92.5%

표 2 사용된 통계적 인식기의 성능 비교

위의 인식기를 보면 Gradient(2)의 인식률이 가장 좋다는 것을 알 수 있는데, 이는 특징 벡터의 크기가 다른 세 개의 특징은 100개 정도인데 반하여 이 특징은 200개로 벡터 공간이 커지기 때문이라고 할 수 있다. 이 네가지 통계적 인식기의 결과를 통합하는 방법으로 본 논문에서 제안한 방법은 가중화 보다 함수이다. 이 함수의 성능을 보기 위하여 위의 네가지 특징의 모든 벡터를 하나의 벡터로 통합해서 표준 벡터를 만들어 MDC에 적용하는 방법과 MPCNN을 이용하여 인식하는 방법[7]과 비교한 결과가 표 3에 있다.

인식 방법	실험 데이터	인식률	2후보	3후보
가중화 보다	PE92	95.375%	98.5%	99.5%
MDC	PE92	92.64%		
MPCNN	360명 데이터	95.3%		

표 3 낱자 인식기의 결과 비교

5.2 금액열 인식률

실험에서 얻어진 금액열 데이터를 이용해서 후처리를 한 결과는 표 4와 같다. 이 금액열에 사용된 금액 낱자들은 후처리를 하기 전에는 80.28%의 인식률을 보였고, 통합 후처리를 한 후에는 97.72%의 인식률을 보였다. 3절에서 얻어진 숫자 인식기의 금액열에 대한 신뢰도 결과와 비교하면 최종 신뢰도가 98.24%로 약 4% 정도 성능이 향상된 것을 확인할 수 있다.

후처리 기	정인식률	기각률	오인식률	신뢰도
후처리 전	74.72%		25.28%	74.72%
구조적 후처리기	88.60%	3.88%	7.52%	92.48%
통합 후처리기	93.76%	4.48%	1.76%	98.24%

표 4 후처리 후 금액열 인식 결과

6. 결론 및 향후 연구 과제

지금까지 한글을 인식하려는 시도는 여러 방법으로 진행되어 왔다고 하지만, 대부분의 방법이 한글의 낱자 인식기에만 초점을 맞추고, 후처리나 상호 참조에는 그리 신경을 쓰지 않은 것이 사실이다.

그러나 본 논문에서는 한글 낱자 인식기 뿐만 아니라, 구조적 후처리기와 숫자 인식기의 인식결과와 상호 참조할수 있는 후처리기를 설계하였다. 통계적 인식기와 가중화 보다 함수를 사용하여 제한된 수의 한글 낱자 인식기를 만들고, 금액의 구조적인 정보를 이용하여 1차 후처리를 한 후, 숫자 인식기와 상호 참조하는 2차 후처리를 수행하는 시스템을 구성하였다.

본 시스템의 의의는 낱자 인식기의 결과가 95% 내외로 나오긴 하지만, 금액열로 들어가면 인식률이 극도로 저하되는 상황에서 구조적 후처리기와 통합 후처리기를 사용하여 오인식률을 최소화 했다는데 있다.

앞으로의 과제는 첫째, 숫자 인식기와의 상호 참조를 더 강화하는 데 있다. 현재는 한글 금액열에서 숫자로 인식된 것하고만 비교를 하고 있는데, 한글 금액열 전체와 숫자 금액열 전체를 비교할 수 있는 시스템을 개발하면 더 신뢰도가 좋아질수 있을 것이다.

둘째, 지금의 낱자 인식기는 상위 세 후보까지만 사용하고, 그 인식률은 99.5%이다. 그런데, 낱자 인식기의 성능을 개선해서 상위 네 후보 이내에서 인식률이 99.9% 이상이 되는 시스템을 만든다면, 전체 성능은 향상될 것이라고 생각된다.

참고문헌

[1] 중앙대학교, 기아정보시스템, "필기체 문자인식 기술개발", 정보통신부 제조업 경쟁력 강화사업 연구 보고서, Sep, 1995

[2] R. G. Gonzalez, R. E. Woods, "Digital Image Processing", Addison Wesley, 1992.

[3] 고태석, 김종렬, 정규식, "오프라인 필기체 한글 자소 인식에 있어서 특징성능의 비교", 인지과학, 제 7권, 제 1호, pp. 57-71, 1996

[4] T. Kohonen, et al., LVQ_PAK - the learning vector quantization program package Version 3.1, April 1995

[5] 백중현, 조성배, 이관용, 이일병, "이중 결합 구조를 갖는 다중 인식기 시스템", 한국정보과학회 '96 봄 학술발표논문집, 제 23권, 1호, pp. 281-284, 1996.

[6] Dae-Hwan Kim, Young-Sup Hwang, Sang-Tae Park, Eun-Jung Kim, Sang-Hoon Paek and Sung-Yang Bang, "Handwritten Korean Character Image Database PE92", IEICE Transaction on Information and Systems, Vol. E79-D, No. 7, pp. 943-950, 1996.

[7] Yung-Mok Baek, Kil-Taek Lim, Sung-Il Chien and Jee-Sung Park, "Off-Line Handwritten Hangul Recognition Based on Mutiple Features and Modular Partially Connected Multi-Layer Perceptron", IWFHR-IV, Aug. 12-14, pp. 269-278.