인쇄체 문자 인식기의 성능 평가에 관한 연구

김 민 수[†] · 강 은 영[††] · 김 우 성[†††] · 한 선 화[††††] · 김 진 형[†††††]

요 약

본 논문에서는 국내의 대표적인 상용인식기들의 성능을 평가하기 위한 평가 방법과 평가 기준을 제안한다. 제안한 평가 기준으로 상용화된 오프라인 문자인식기들과 실험실 인식기를 비교해본 후 각각의 특성을 분석해 보았다. 인식에 필요한 대상 문서는 400 DPI로 스캔한 1000여개의 문서영상과 수작업으로 직성한 원문이 존재하는 KT 테스트 컬렉션을 사용하였다. 본 논문에서 인식기의 성능을 평가하기 위해 문자단위 인식률 측정 방법을 제안하였다. 비교를 위한 분석의 유형을 제안하여 단일 특성을 가지는 문서, 복합 특성을 가지는 문서 등으로 구분하여 비교·분석하였다.

A Study on Implementation of Printed Character Recognition System And Performance Evaluation

Min-Soo Kim[†] · Eun-Young Kang[††] · Woo-Sung Kim[†††] · Sun-Hwa Han[††††] · Jin-Hyung Kim[†††††]

ABSTRACT

In this paper we propose measure for performance evaluation of character recognition. We used three commercial character recognizers and one laboratory character recognizer for test. The characteristics of each recognizer is compared by proposed evaluation standard, and analyzed characteristics. For the input test data, KT test collection are used. KT test collection is composed of 1000 document images about and complete source text. In this paper we propose method for measuring recognition rate in character unit for evaluation of character recognition. The recognition rates are compared and analyzed by single feature characteristic or mixed feature characteristic.

1. 서 론

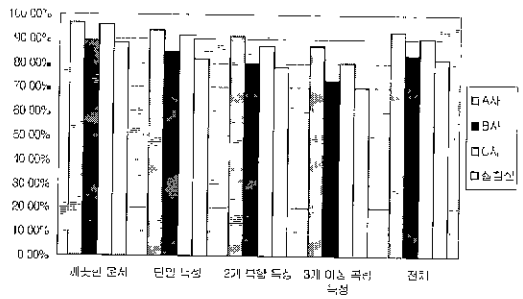
현대 사회는 정보화 사회로, 컴퓨터에 의한 정보는 모든 분야에 걸쳐 그 영향력이 점차 확대되고 있다. 컴퓨터는 인간에 비해 빠른 연산 처리와 거대한 정보의 저장, 검색 그리고 정보를 관리하는 측면에 있어서는 인간의 능력을 훨씬 능가하지만, 정보의 인식과 분석, 입출력 측면에서는 그 능력이 현저히 떨어지고 있다. 따라서 컴퓨터로 빠르고 정확한 처리를 하기 위해서는 사용자가 입력 자료를 구조적인 디지털 형식으로 변환해야 할 필요가 있다. 이러한 과정은 사용자들에게 많은 불편을 준다. 이 때문에 기존의 자료를 컴퓨터에 입력하기 위한 효율적인 자동 입력 시스템이 요구되고 있으며, 자료 입력 시 요구되는 시각 능력을 컴퓨터로 구현하고자 하는 노력이 꾸준히 진행되어 왔다. 문서인식은 자료의 자동 입력의 한 방식으로 오래 전부터 많은 연구가 진행되고 있는 분야이다.

이러한 문서의 자동 입력과 인식을 위해서는 무엇보다도 먼저 문자 인식이 기본적으로 이루어져야만 한다. 한글 문서인식에 관한 연구는 1960년대 이후 현재

† 준 회 원 : 호서대학교 벤처전문대학원 컴퓨터응용기술분야
†† 준 회 원 : (주)리드텍코리아 정보기술연구소 연구원
††† 정 회 원 : 호서대학교 컴퓨터공학과 교수
†††† 정 회 원 : 연구개발정보센터 한민족과학기술자네트워크지원단
††††† 종신회원 : 한국과학기술원 전산학과 교수

논문접수 : 2000년 1월 25일, 심사완료 : 2000년 10월 17일

을 나타내었으며, 문서 유형의 변화에 가장 둔감하므로 실험한 인식기 중에서 문자 인식 성능이 가장 좋다고 판단된다



(그림 7) 유형별, 인식기별 문서의 인식률 그래프

(그림 7)의 결과는 인식기의 성능을 평가 할 수 있는 전체적인 그래프를 나타내고 있다. (그림 7)에서 한눈에 알아 볼 수 있듯이 모든 면에서 A사 인식기가 가장 좋은 인식률을 나타낸다. 하지만 문서가 3개 이상의 복합 특성을 가질 때의 인식률이 90% 이하로 떨어지는 현상을 발견할 수 있다. 앞에서 가정한데로 사용자가 수정작업을 하더라도 쓸만한 인식결과가 91% 이상이라고 가정하고 인식 대상이 KT 테스트 컬렉션처럼 논문이 아닌 범용적인 문서라고 했을 때 A 인식기 또한 실용적이지 못하다.

실험 결과에서 보여준 비와 같이 평가 대상이 되는 4가지 종류의 인식기 중에서 모든 종류의 문서유형에 대하여 A사의 인식기가 대체로 높은 인식률을 갖는 것을 알 수 있다. C사의 경우는 인식하기 어려운 글자를 '♠', '♥'등의 특수기호에 대체되어 나타나고, B사의 인식기의 경우는 대부분의 글자를 인식하는 과정에서 공백 문자가 과도하게 삽입된다 A사 인식기의 경우도 공백문자가 삽입되는 경우가 발생하지만 B사 인식기의 경우는 공백문자의 삽입이 오류의 0.25%를 차지한다. 또한 다른 문자로 오인식 하는 오류도 많이 발생하는데 이 때문에 문자 해독이 어려워진다. 실험에 사용된 모든 문자 인식기에 있어서 인식된 문서에 특수 문자가 포함되거나 공백문자가 삽입되어 있는 등 상태가 좋지 않아 데이터베이스 생성과 같은 후처리에 적용하기에는 적합하지 않다

5. 결 론

상용 문자 인식기들을 제안한 인식률 평가 기준에 따라 비교한 결과를 종합해보면 한글 문자에 대한 인식률이 저조하다는 것을 알 수 있다. 인식률이 가장 좋게 평가된 A사의 인식률은 95.08%이다 영어 문화권에서 개발한 인식기의 경우 영문자-알파벳-의 인식률이 99.9%에 비하면 오인식률이 크다는 것을 알 수 있다 그 이유는 한글이라는 문자의 특수성 때문이다 한글은 알파벳과는 달리 초성, 중성, 종성으로 되어 있어 인식에 대한 처리 과정이 알파벳에 비해 복잡하다. 그렇기 때문에 그만큼 오인식률이 큰 것이다.

상용 문자 인식기들을 평가해봄으로써 한글 인식기술의 현재의 수준과 인식기들의 특성을 파악할 수 있었다 또한 인식기의 성능평가 기준을 제시함으로써 인식률의 객관적인 측정기준으로 사용할 수 있을 것이다

이번 실험에서 조금의 아쉬운 점이 있다면 인식대상 영상의 특성이 모두 테스트 영역으로만 구성되어 있어서 영역분할에 대한 실험을 거의 못했다는 점이다. 문자 인식률 뿐 아니라 기울기 보정이나 영역 분할 등의 전처리에 해당하는 과정도 인식에 많은 영향을 끼치므로 앞으로 그에 대한 성능 평가도 이루어져야 할 것이다.

문서 영상을 문자 인식기로 인식한 후 나온 결과 텍스트에서 정보 검색에 필요한 색인어를 찾아낸다면 문자인식에 기반한 전자도서관 구축에 효율적으로 활용될 수 있을 것이다.

참 고 문 헌

[1] J Liang, R M. Haralick, I. T. Phillips. "Performance Evaluation of Algorithms in ISL Document Layout Analysis Toolbox," ISL Technical Report, University of Washington, 1996

[2] G. Friensen. Suddenly. "OCR is a "Must BUY." Imageing Magazine, pp 22-25, 1992.

[3] 김동근, 황치정, "투영과 텍스쳐를 이용한 문서 영상의 블록 분류", 한국 정보처리학회 추계 학술 발표논문집 제4권 2호, 1997.

[4] 김우성, 심진보, 박용범, 문경애, 지수영, "문서 영상내의 테이블 백터화 연구", 정보처리논문지, 제3권 5호, pp 1147-1159. 1996. 9.

[5] Yasuo Korosu, Hidefumi Masuzaki, "A Method of

식(2)에서 n은 비교한 데이터의 개수, x_i는 비교한 각 데이터의 값, m은 비교한 전체 데이터의 평균을 의미한다. 인식에 사용된 알고리즘과 대부분인 문서 영상의 처리 방법에 따라 차이가 있을 수는 있지만 표준 편차가 크다는 것은 그만큼 인식기가 안정적이지 못하다는 것을 말한다. 표준 편차의 비교에서 보아도 A사의 인식기가 가장 우수하게 나왔다. 그러나 평균 인식률이 높게 나온 C사 인식기의 경우는 표준 편차가 가장 크게 나옴으로써 안정적이지 못한 것으로 평가되었다.

줄 간격이 좁은 문서(유형 6)와 글자의 굵기가 다른 문서(유형 11)들은 대체로 4종류의 인식기기 모두 인식을 잘 해냄으로써 줄 간격이 좁아도 글자들이 붙지 않으면 인식을 잘 수행함을 알 수 있었고, 글자의 굵기가 다른 것은 한글 인식에 별다른 영향을 주지 않음을 알 수 있었다. C사 인식기는 글자 사이의 간격이 좁은 문서(유형 8)에 대해 제일 낮은 인식률을 보인다. C사 인식기의 문자 분할 알고리즘이 다른 인식기의 그것에 비해 성능이 떨어지는 것을 알 수 있다. 또한 B사 인식기의 경우는 줄 간격이 좁은 문서는 잘 인식하지만 줄 간격이 넓은 문서는 인식률이 저조한 모습을 보여 문단 분할 알고리즘에 오류가 있는 것으로 판단된다.

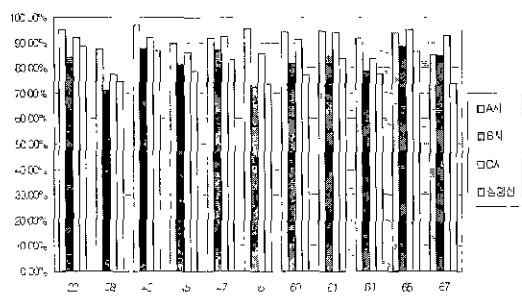
본 논문에서는 단일 특성이 2개씩 들어있는 문서들을 15번부터 68번까지 54개 유형으로 분류하고 비교 분석해보았다. 두 가지 복합 특성을 갖는 문서들은 그 경우의 수가 많아 데이터의 개수가 가장 많은 유형 10개만을 <표 2>에 나타내 비교해보았다 (그림 6)은 <표 2>를 비교하기 쉽게 나타낸 것이다. <표 2>의 제일 밑에 있는 수치는 54개의 모든 유형에 대한 개수와 평균인식률이다.

〈표 2〉 두 가지 특성을 가지는 문서의 각 인식기별 인식률

유형 번호	문서의 특성	개수	A사	B사	C사	실험실
22	진함 & 외곽선지저분	9	95.33%	84.22%	91.80%	88.58%
28	붕게짐 & 진함	42	87.21%	70.92%	77.37%	74.51%
42	줄간좁음 & 외곽선지저분	12	97.01%	87.60%	92.03%	86.92%
46	줄간좁음 & 붕게짐	21	89.68%	81.15%	86.13%	78.58%
47	줄간좁음 & 문자끊김	7	91.86%	87.17%	92.52%	83.22%
56	고딕 & 붕게짐	11	95.21%	72.56%	85.63%	73.55%
60	기울어짐&외곽선지저분	22	94.10%	81.86%	91.07%	77.11%
61	기울어짐 & 굵기다양	9	94.56%	81.83%	94.08%	83.67%
64	기울어짐 & 붕게짐	15	91.84%	78.56%	83.39%	77.33%
65	기울어짐 & 설영포함	8	94.02%	88.10%	95.14%	86.30%
67	기울어짐 & 줄간좁음	7	85.03%	84.41%	92.60%	73.77%
2개 특성을 가지는 문서		253	91.11%	79.69%	87.08%	78.44%

위의 <표 2>에서 살펴보면 단일 특성과 같이 평균적으로 A사 인식기의 인식률이 높은 것을 발견할 수 있다. 하지만 모든 인식기의 인식률이 단일 특성을 가지는 문서들의 평균 인식률보다 2%이상 떨어졌다. 이것은 문서가 포함하는 특성이 많아질수록 인식기의 성능이 떨어진다는 것을 나타낸다. B사, C사, 실험실 인식기는 모두 4%에 가까거나 그 이상의 인식률 저하를 보임으로서 문서 내 특성이 많아질수록 인식기 성능이 현저히 떨어지는 것을 볼 수 있다.

두 가지 특성을 가지는 문서에 대해서도 A사 인식기가 타 인식기에 비해 성능이 좋게 평가되었다. 두 가지 특성을 가지는 문서 유형에 대해서 표준편차를 구해본 결과 B사 인식기가 12.93, 실험실 인식기가 15.94로 나와 복합 특성 문서의 인식률 저하현상을 확실하게 보여준다. A사와 C사 인식기의 표준편차는 각각 9.71, 9.12로 나왔다.



(그림 6) 두 가지 특성을 가지는 문서의 각 인식기별 인식률 그래프

KT 테스트 컬렉션과 같이 논문의 경우에는 여러 가지 특성을 복합적으로 가지는 문서가 드물다. 하지만 인식기를 범용적으로 사용한다면 인식해야 할 문서는 1개 또는 2개의 특성만을 가지는 문서보다는 아래와 같이 3가지 이상 복합 특성을 가지는 문서들이 더 많다. 본 논문에서 KT 테스트 컬렉션을 대상으로 실험한 결과 3개 이상의 복합 특성을 가지는 문서에 대한 인식률은 A사 인식기가 87.26%, B사 인식기가 72.86%, C사 인식기가 80.16%, 실험실 인식기가 69.98%로 A사 인식기가 역시 가장 높은 인식률이 나왔다. 따라서 이 부분에서 가장 높은 인식률이 나온 A사 인식기가 실험한 인식기 중에서 가장 효율적이라 말할 수 있다. 물론 모든 부분에서 A사 인식기가 가장 높은 인식률

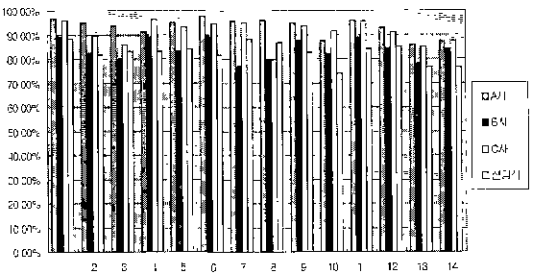
인식률이 0%에서 20까지인 데이터 수가 B사 인식기에서는 10개, 실험실 인식기에서는 9개나 나오지만 C사 인식기와 A사 인식기에서는 나오지 않는다. 인식률이 90% 이상인 데이터의 개수는 A사 인식기가 793개로 가장 높은 인식률을 보이고 있다. 위의 표에서 볼 수 있듯이 대부분의 인식기가 80% 이상의 인식률을 보여주지만 데이터의 상태에 따라서 인식률이 거의 70%에도 미치지 못하는 데이터도 존재한다.

본 논문에서는 단순히 전체 인식률만으로 인식기의 성능을 평가하지 않고 문서의 인쇄 특성에 따른 문서 유형을 구분하여 유형별 인식률을 분석하여 평가하였다. 3장에서도 언급한 것처럼 인식하고자 하는 KT 데스트 컬렉션의 문서 중에서 실험 대상 문자인식기에서 인식되는 960개의 문서를 선정하였다 960개의 문서들을 각각의 특성을 찾아서 분류하였는데, 각 문서들은 하나의 특성(글자가 진함, 영상이 기울어짐 등)만을 가지고 있는 경우도 있지만 복합 특성(글자가 진하면서 기울어진 문서 등)을 가지는 경우도 적지 않았다 그래서 실험문서들을 인식에 적합한 깨끗한 문서, 한가지의 특성만을 가지고 있는 문서, 두 가지의 특성을 가지고 있는 문서, 그리고 3가지 이상의 특성을 가지는 문서들로 구분하여 분석했다.

<표 1>은 선정된 960개의 문서 중 깨끗한 문서와 각 특성을 찾아서 분류한 것 중 비교적 뚜렷하게 하나의 특성만을 갖는 문서에 대한 실험 결과이다. (그림 5)는 <표 1>을 비교하기 쉽게 그래프로 나타낸 것이다. 깨끗한 문서(유형1)나 단일 특성을 가지는 문서들(유형 2~14)에 대해서 A사와 C사 인식기가 95%가 넘는 높은 인식률을 보인다. B사 인식기와 실험실 인식기도 깨끗한 문서에 대해서는 인식률이 90%에 가깝다는 것을 볼 수 있다. 단일 특성을 가지는 문서 유형에서 인식률이 가장 높은 것과 가장 낮은 것의 쌍을 보면 A사 인식기 - (줄간 좁음 · 한문 및 기호 포함), B사 인식기 - (기울어짐 : 줄간 넓음), C사 인식기 - (잡영 포함 : 자간 좁음), 실험실 인식기 - (줄간 넓음 . 흐림)이다. 이 결과에서 알 수 있듯이 각각의 인식기는 인식이 잘되는 유형과 안 되는 유형이 다르다. 즉 데이터베이스를 구성하기 위해 인식기를 선택할 경우에는 인식기 제작사에서 제공되어지는 평가용 버전 등으로 데스트를 해보고, 데이터베이스화하려는 문서의 유형을 고려해서 선택하는 것이 바람직하다. 물론 다른 목적으로 사용할 경우에도 좋은 방법이다. 또한 낮은 인식률을 보이는 문서와 경향이 같은 문서의 경우는 데이터베이스 역할을 제대로 수행할 수 없으므로 인식 과정에서 따로 분류하여 수작업에 의해 입력하는 것이 좋다

〈표 1〉 깨끗한 문서 및 단일 특성을 가지는 문서의 각 인식기별 인식률

유형 번호	문서의 특성	개수	A사	B사	C사	실험실
1	깨끗한 문서	207	96.45%	89.21%	95.79%	88.42%
2	진 함	28	95.03%	82.10%	89.36%	81.71%
3	몽 개 짐	63	93.69%	80.11%	85.89%	83.22%
4	잡영포함	9	91.26%	88.75%	96.20%	83.22%
5	문자 끊김	25	94.78%	83.03%	93.07%	83.06%
6	줄간 좁음	25	97.58%	89.00%	94.47%	81.82%
7	줄간 넓음	8	95.62%	76.92%	94.90%	88.22%
8	자간 좁음	5	95.79%	79.78%	79.78%	86.72%
9	기울어짐	62	94.75%	87.59%	93.62%	82.83%
10	흐 림	71	87.58%	81.91%	91.61%	73.73%
11	굵기 다양	41	96.03%	88.90%	95.62%	84.04%
12	의간선지서분	75	92.89%	64.30%	90.63%	85.18%
13	한문 및 기호	21	85.88%	77.99%	85.12%	76.64%
14	찌핑 넓음	2	87.01%	84.03%	87.30%	76.71%
1개 특성을 가지는 문서		435	93.17%	83.83%	90.96%	82.61%



(그림 5) 깨끗한 문서와 단일 특성을 가지는 문서의 각 인식기별 인식률 그래프

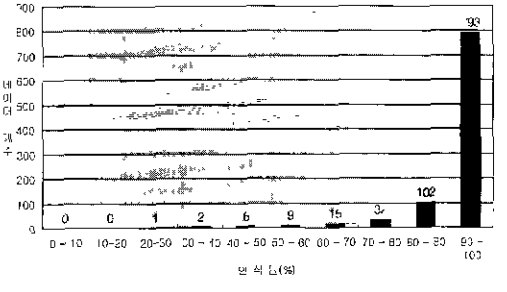
깨끗한 문서의 경우나 단일 특성을 가지는 문서의 경우에서 모두 A사 인식기가 우수한 인식결과를 보여주었다 평균 인식률로 따져보면 A사, C사, B사, 실험실 인식기 순으로 인식이 잘 된 것으로 나타난다. 하지만 단일 특성을 가지는 문서들에 대하여 식(2)를 이용하여 표준편차를 구하면 A사($\sigma = 3.75$), B사($\sigma = 4.04$), 실험실($\sigma = 4.25$), C사($\sigma = 4.81$) 순으로 점점 커짐을 볼 수 있다.

$$\text{표준편차}(\sigma) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - m)^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - m^2} \quad (2)$$

제끗한 문서 유형과 단일 특성에 대한 평가 기준인 13가지 문서 유형을 나타냈고, <표 2>에는 2가지 복합 특성을 가지는 54개의 문서 유형 중 빈도수가 높은 10개를 나타내었다 1997년 Hello-PC지에서 한 벤치마킹[10]에서는 7개의 유형에 대해 각 1개의 문서를 실험함으로써 성능 평가에 대한 판단이 사실상 불가능했던 것에 비해 본 논문에서 사용한 데이터 셋은 그 수와 유형이 다양함으로 국내의 대표적인 상용 인식기에 대한 성능평가가 가능하다.

KT 테스트 컬렉션에 대한 전체 평균 인식률은 A사 인식기가 92.96%으로 가장 높게 나왔고, 다음은 C사 인식기로 평균 인식률은 90.41%이고 다음으로 B사 인식기의 인식률은 83.27%이며 실험실 인식기는 81.51%이 나왔다.

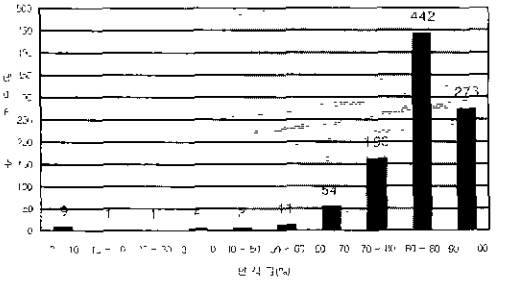
위의 결과에서도 알 수 있듯이 아직까지 한글 인식은 어려운 분야로 남아있다. 구체적인 성능평가를 위해 각 인식기의 자세한 인식률 분포를 살펴보면 다음과 같다. (그림1), (그림 2), (그림 3), (그림 4)에서 각 인식기 별 인식률의 분포를 나타내었다



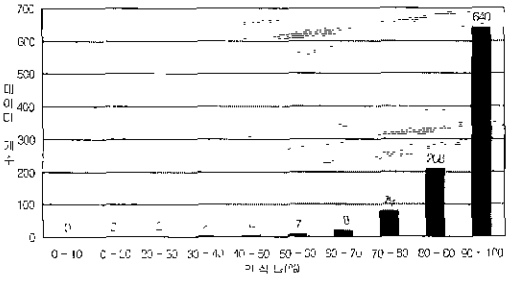
(그림 1) A사 인식기의 인식률 별 분포

본 논문에서는 인식기의 인식률이 0%에 가까운 문서가 적고, 100%에 가까워질수록 많아지는 형태로 나타날 때 인식률이 안정적이라고 정의한다. 또한 인식률이 100%에 가까운 문서가 많을수록 효율적이라고 정의한다. 실험한 인식기 중에서 A사의 인식기가 가장 안정적이고 효율적인 것을 (그림 2)를 통해 알 수 있다 91%에서 100%사이의 인식률을 나타내는 문서가 793개로 가장 많으며 인식률이 20% 미만인 문서는 존재하지 않는다.

B사 인식기는 81% 이상의 인식률을 나타내는 문서가 715개나 되지만 A사 인식기가 C사 인식기와 비교해 볼 때 낮은 수치이며 A사의 인식기에서 91% 이상의 인식률을 나타내는 문서가 800개 가까이 되는 것을



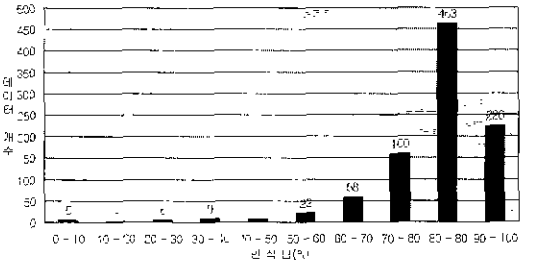
(그림 2) B사 인식기의 인식률 별 분포



(그림 3) C사 인식기의 인식률 별 분포

볼 때 실용적이지 못한 인식기로 판단된다.

C사 인식기는 실험에 사용된 인식기 중에서 두 번째로 높은 인식률을 보인다 그러나 사람이 적당히 오류를 수정한다라도 사용할 수 있는 인식률 기준을 91% 이상이라고 보았을 때, C사 인식기는 81%에서 90%사이의 인식률을 나타내는 문서가 전체 문서의 약 5분의 1가량을 차지하고 91% 이상의 인식률을 나타내는 문서의 수가 상대적으로 빈약하므로 B사 인식기와 마찬가지로 실용적이지 못한 인식기로 판단된다



(그림 4) 실험실 인식기의 인식률 별 분포

실험에 사용된 인식기 가운데 가장 낮은 인식률을 보인 실험실 인식기는 81%에서 90%사이의 인식률을 나타내는 문서의 수가 B사의 인식기보다 높다 하지만 인식률이 91%이상인 문서의 수도 적고 인식률이 70% 이하인 문서도 B사 인식기보다 많다.

식 실험을 수행할 수 있다.

② 수 작업된 데이터베이스가 이미 존재하고 있으므로 문자인식 문서와 원래의 문서를 비교하여 문자 인식률을 쉽게 측정할 수 있다

③ KT 테스트 컬렉션은 30여 개의 질의어와 검색된 문헌 정보에 대한 자료를 가지고 있다. 따라서 본 실험을 통해 구축된 데이터베이스는 추후 문자 인식된 데이터베이스에 대한 검색 방법 연구에 좋은 자료로 사용될 수 있다

본 논문에서는 스캔된 972개의 KT 테스트 컬렉션의 문서 중에서 실험 대상 문자인식기에서 인식되는 960개의 문서를 가지고 데이터베이스를 구축하였다. KT 테스트 컬렉션은 모두 400 DPI의 해상도로 스캔된 이진 영상이다

3.2 인식률 측정방법 및 타당성

기존의 인식률 측정방법은 개발자가 자신의 인식기의 성능을 알아보기 위한 내부적 인식률 측정방법이었다. 내부적 방법은 인식 알고리즘의 특성에 따라 여러 가지 방법으로 수행할 수 있고, 자소단위의 인식률 측정도 가능하다 그러나 이러한 내부적 측정방법에 의한 인식률이 높다고 하여 일반 사용자가 사용하기 좋은 인식기는 아니다 따라서 본 논문에서는 좀더 일반적이고 효율적인 인식기의 성능평가를 위해, 개발자 관점이 아닌 사용자 관점에서의 인식률 측정방법을 제안한다.

사용자 관점의 인식률 측정 방법에는 문자 단위 인식률과 어절 단위 인식률의 두 가지가 있다. 문자 단위 인식률이란 대상문서가 포함하는 전체 문자 중 몇 %의 문자가 옳게 인식되었는가를 측정하는 것이고, 어절 단위 인식률이란 띄어쓰기에 의해 분리된 문자 내의 각 어절을 대상으로 전체 어절 중 몇 %의 어절이 옳게 되었는기를 가지고 인식률을 측정하는 방법이다 본 논문에서는 문자 단위 인식률을 측정하였다

데이터베이스의 대상 문서를 KT 테스트 컬렉션으로 하였기 때문에 원문 텍스트에 대한 데이터베이스 파일이 존재한다. 따라서 문자 단위의 인식률 측정은 UNIX에서 사용하는 diff 명령어에 사용된 알고리즘을 사용하여 편리하게 구했다[8-9] 본 논문에서 사용한 인식률 측정 알고리즘은 다음과 같다.

① 공백 문자가 두 개 이상 연속적으로 발생하는 경우 하나의 공백 문자로 치환한다. 개행 문자는 공백 문자로 치환한다

② 한 줄에 하나의 문자가 들어가도록 개행 문자를 모두 삽입한다.

③ UNIX의 diff 명령에 사용된 알고리즘을 사용하여 원래의 텍스트와의 best-match를 구한 후 잘못 인식된 문자수를 계산한다.

1. 원문에 없는 내용이 인식기에 의해 추가된 경우 : 무시한다.
2. 원문에는 존재하니 인식기가 인식하지 못한 경우 : 없어진 문자의 수만큼 잘못 인식된 문자로 본다.
3. 원본의 문자가 다른 문자로 인식된 경우 : 잘못 인식된 원문의 문자수를 계산한다.

④ 공백을 포함한 원문 텍스트의 문자의 개수를 n이라고 하고, best-matching에서 계산 된 잘못 인식된 문자의 수를 c라 하면 인식률은 식(1)과 같이 계산된다

$$\text{인식률} = \frac{(n-c)}{n} \times 100 \qquad (1)$$

위와 같은 방식으로 인식률을 측정하는 것은 사람이 원문을 보고 타이핑 한 문서를 검토하는 것과 흡사하다. ③-1의 경우와 같이 원문에 없는 내용을 사람이 타이핑하는 경우는 드물지만, ③-2나 ③-3의 경우는 흔히 발생할 수 있다. ③-2의 경우는 원문에서 한 줄 내지는 두 줄을 건너뛰어 타이핑하는 경우에 속하고, ③-3의 경우는 쉽게 말하는 오타의 경우이다. 오타의 경우는 문자의 수가 늘어나서 ③-1의 경우도 발생할 수 있다.

따라서 위와 같은 방식으로 인식률을 측정하는 것이 사람이 판단하는 것과 가장 흡사한 인식률 측정 방법이라 할 수 있을 것이다.

4. 인식률 평가 기준 및 분석

본 논문에서는 2.1절의 인식률에 영향을 미치는 인쇄 특성을 기준으로 하여 특성 없이 깨끗한 문서 유형, 단일 특성 13가지 문서 유형, 13가지 단일 특성을 2개씩 가지는 54개의 문서 유형, 그리고 3개 이상씩 가지는 문서 유형으로 나누어 분석하였다. <표 1>에는

까지 꾸준히 진행되어 왔고, 최근에는 필기체 인식 등의 연구분야가 활발히 진행되고 있다.

본 논문에서는 자동 문서인식시스템을 현장에서 사용하기 위해 문자인식기의 성능을 평가하는 기준을 제안하고, 기존에 발표되어 일반인이 사용하고 있는 3개의 상용 인식기와 실험실에서 개발된 1개의 실험실 인식기를 대상으로 제안한 기준에 의한 인식률 성능평가[1, 2]를 수행하였다. 본 논문의 2장에서는 인식 대상이 되는 문서의 유형에 대해 살펴보고 3장에서는 실험 환경과 인식률을 조사하는 방법에 대해 알아본다. 4장에서는 문서의 특성에 따른 인식률 평가 기준을 제안하고, 이에 따라 인식기별 인식률을 조사하여 비교 및 분석하였다. 마지막으로 5장에서 결론을 내렸다.

2. 인식대상 문서의 특징

인식기의 성능을 테스트하기 위한 방법은 많은 수의 다양한 형태의 문서에 대해 인식 실험을 해야할 필요성이 있다. 특징이 있는 각 문서들의 인식률을 구하고, 분석해보면 인식기의 장단점을 알 수 있다. 인식하기 위한 문서의 특징은 여러 가지가 있는데 크게 구분해 보면 구조적 요소와 각 구조적 요소의 내부 속성이 해당된다. 각 구조적 요소와 포함하는 내부 속성에 대해서 알아본다.

2.1 텍스트

텍스트만으로 서술되는 형식으로 가장 보편적인 형태의 문서를 말한다. 그림이나 표 등을 포함하는, 즉 요소가 복합된 문서는 인식을 하기 전에 특징들을 제거하여 텍스트 영역만을 추출하여 문자 인식을 수행한다[3]. 그러므로 텍스트 영역에 대한 인식률이 저조하다는 것은 여러 요소를 포함하는 문서들의 인식률도 저조하다는 것을 의미한다. 즉 텍스트 영역의 인식률이 전체적인 인식기의 성능을 좌우한다고 볼 수 있다.

텍스트 영역에 대한 내부 속성으로는 폰트의 크기, 종류, 스타일 등이 있다. 이것은 인식기의 성능 측정에 중요한 척도가 되며 폰트의 변화에 덜 민감한 인식기가 더 좋은 성능의 인식기라 할 수 있다.

2.2 그림 · 표

그림이 포함된 문서에서는 그림 영역을 추출하여 그 부분을 제외한 영역만을 인식 대상범위로 설정하는 처리 과정이 필요하다. 또한 그림 영역 중 문자가 존재하더라도 그 문자는 그림 영역에 있으므로 인식 대상에서 제외된다.

표가 포함된 문서는 보고서나 은행 전표 등이 대표적인 예이다. 표의 구조와 크기를 분석하여 표를 구성하는 선들을 구분해내고[4], 표 안의 문자 영역을 찾아 문자들을 인식해야한다.

2.3 기울어짐

기울어진 문서는 스캐너와 같은 장비로 스캔할 때 원본이 기울어진 상태에서 스캔된 것이다. 기울어짐이 미약한 문서는 기울어짐에 대한 보정처리 없이 인식된다. 그러나 기울어짐이 있는 대부분의 문서는 기울어짐을 보정시킨 후 인식을 해야한다[5]. 기울어짐 보정이 제대로 되지 않으면 비록 사람이 보고 판독할 수 있다 하더라도 컴퓨터는 인식 불가능할 수 있다.

2.4 다단

신문, 논문 등 다양한 문서들이 다단 문서 구조로 되어있다. 다단 문서가 일정한 간격으로 단이 구분되면 처리 과정이 단순하겠지만 신문의 경우는 단의 개수가 불규칙적이며 단의 간격도 불규칙적이기 때문에 처리 과정이 복잡하다.

3. 실험환경 및 인식률 측정방법

3.1 실험 환경

인식 과정은 PC(Pentium 200 MMX)에서 OS는 Windows95를, 컴파일러는 Visual C++ 5.0을 사용하여 수행하였으며, 인식률 조사는 Axil 245에서 OS는 SPARC Solaris 2.5.1을, 컴파일러는 GCC 2.8.1을 사용하여 수행하였다.

본 실험의 대상 문서로서는 KT 테스트 컬렉션[6]을 구성하고 있는 논문을 사용했다. 1000여 개의 데이터를 가지고 있는 KT 테스트 컬렉션은 국내 정보 검색 분야의 연구자들이 실험 대상으로 가장 널리 이용하고 있는 자료이다. KT 테스트 컬렉션의 논문을 대상 문서로 선택함으로써 얻을 수 있는 장점은 다음과 같다[7].

(1) KT 테스트 컬렉션의 대상 논문은 수년에 걸쳐 수집된 두 종류의 논문지 및 학술 발표 집으로 구성되어 있으므로 다양한 인쇄 품질에 대한 인

Skew Detection and Correction in Document Images for Personal Computers," Transactions of Information Processing Society of Japan, Vol.39, No.8, pp.2466-2475, 1998 8

[6] 김성혁, 서은경, 이원규, 김빙철, 김영환, 김제균, "자동세인기 성능시험을 위한 Test Set 개발", 정보관리학회지, 제11권 제1호, pp.81-102, 1996. 6

[7] 이준호, 이충식, 한선화, 김진형, "문자인식에 의해 구축된 한글 문서 데이터베이스에 대한 정보검색", 정보처리논문지, 제6권 4호, pp.833-840, 1999 4.

[8] 한선화, 김진형, "문자 인식 기술을 이용한 데이터베이스 구축에 대한 기초 연구", KORDIC 연구보고서, 1997. 9.

[9] 한선화, 이충식, 한선화, 김진형, "문자 인식 기술을 이용한 데이터베이스 구축", 정보처리논문지, 제6권 7호, pp.1713-1723, 1999.7

[10] 김명진, 장원식, "OCR 소프트웨어 4종 한글, 영어 인식 능력 완벽 테스트", Hello-PC 1997년 5월호, pp.368-383, 1997 5.

김 민 수

e-mail holybest@hitel.net

1998년 호서대학교 컴퓨터공학과 (학사)

2000년 호서대학교 컴퓨터공학과 (석사)

2000년~현재 호서대학교 벤처 전문대학원 컴퓨터응용기술 분야 박사과정 재학

관심분야 : 영상처리, 문자인식, 뉴로-퍼지

강 은 영

e-mail : magic94@shinbiro.com

1998년 호서대학교 컴퓨터공학과 (학사)

2000년 호서대학교 컴퓨터공 학과 (석사)

2000년~현재 (주)리드텍코리아정보기술연구소 연구원

관심분야 : 영상처리, 문자인식, 데이터베이스

김 우 성

e-mail : wskim@office.hoseo.ac.kr

1980년 서강대학교(학사)

1983년 미국 Texas A&M 대학교 (석사)

1993년 서강대학교(박사)

1984년~1987년 한국전자통신연구소 연구원

1996년~1992년 호서대 지역협력연구센터 연구부장

1999년~2000년 미국 Univ. of Washington 방문 교수

1987년~현재 호서대학교 컴퓨터학부 교수

관심분야 : 영상처리, 문서인식, 지식관리, 정보검색

한 선 화

e-mail : shhahn@kordic.re.kr

1987년 성균관대학교 정보공학과 (학사)

1989년 한국과학기술원 전산학과 (석사)

1997년 한국과학기술원 전산학과 (박사)

1997년~현재 연구개발정보센터 한민족과학기술자네트워크지원단 단장

관심분야 : 데이터베이스/마이닝, 지식관리, Intelligent Tutoring, HCI

김 진 형

e-mail : jkim@cs.kaist.ac.kr

1971년 서울대학교 공과대학 (학사)

1973년~1976년 과학기술연구소 (KIST) 전산실 연구원

1976년~1977년 미 California State, 도로국, 프로그래머

1979년 UCLA 전산학과(석사)

1981년~1985년 Hughes Research Center, Malibu, Senior Computer Scientist

1983년 UCLA 전산학과(박사)

1990년~1991년 미 IBM Watson Research Center 초빙 연구원

1985년~현재 과학기술원 전산학과 교수

1991년~현재 과학재단 지정 과학기술원 인공지능 연구센터 부소장

1995년~현재 출연(연) 연구개발정보센터 소장

1997년~현재 공학한림원 회원

관심분야 : 문자인식, 지능형 인터페이스, 인공지능