필기 한자 고문서의 디지털 라이브러리화를 위한 입력 시스템 개발

장만대° 김민수 이택헌 김진형 *곽희규
한국과학기술원 전자전산학과, *㈜ 동방 SnC
{mdjang°, mskim, three, jkim, *hkkwag}@ai.kaist.ac.kr

Development of Input System for Digitalizing Handwritten Hanja Historical Documents

Man Dae Jang° Min Soo Kim Taik Heon Rhee Jin Hyung Kim *Hee Kue Kwag
Dept. of Electronic Engineering and Computer Science, KAIST, *Dongbang SnC

요 약

본 논문에서는 필기 한자로 쓰여진 고문서를 보다 효율적으로 디지털 라이브러리화하기 위한 입력 방법을 제안한다. 제안한 입력 방법은, 문자 인식 방법과 수작업을 병행하는 방법으로서, 인식 및 기각 방법을 사용하여 유사한 글자들을 자동 군집화한 후, 수작업으로 교정 및 검증을 거쳐 최종 입력하는 방식이다. 한국학 고문서인 승정원일기를 대상으로 한 실험에서, 제안한 방법이 높은 정확률과 낮은 기각 비율을 보임으로써, 기존의 수작업 입력 방법을 대체할 경우 상당한 시간 및 노동력의 절감을 가져올 것으로 기대한다.

1. 서론

국내에 존재하는 전통적인 고문서에는 한자로 쓰여진 문서가 많이 존재한다. 이와 같은 고문서들은 그 시대의 생활상이나 제도 및 경제적인 상황 등을 이해하는데 중요한 단서가 된다는 점에서 그 역사적 가치나 보존 가치가 높다. 전통적, 문화적 가치가 높은 자료를 학술적으로 영구히 보존 및 이용하기 위해서는 효율적인 자료 관리와 검색을 위한 디지털 라이브러리 구축이 필요하다. 그러나 기존 디지털 라이브러리는 사람이 직접 입력 및 교정하는 방식으로 이루어졌으며, 한자의 수가 워낙 많기 때문에 직접 손으로 입력하는 방법은, 상당한 노동력과 작업시간을 필요로 하는 작업이며, 일의 정확성도 많이 떨어진다.

패턴 인식 기술을 이용하여 OCR 시스템과 같이 문자의 입력 과정 전체를 자동화하면 이러한 문제점들의 상당 부분을 해결할 수 있을 것으로 기대되지만, 현재의 기술 수준으로는 이와 같이 완전히 자동화된 인식시스템의 개발은 어렵다. 필기체 한자 인식기의 성능이 아직 실용적인 단계에 미치지 못하고 있을 뿐 아니라[1], 원본의 노후화로 인해 저하된 화질의 개선, 독특한 한자들의 처리 등 인식 이외에도 해결해야 할 문제가 산적해 있기 때문이다 (그림1). 따라서, 현실적인 절충안으로서, 입력된 영상을

(a) 획이 훼손된 한자 (b) 이체자의 예

그림 1. 고문서 영상에 존재하는 한자

문자 인식 기법을 통해 각 문자별로 자동으로 군집화한 후, 해당 군집 별로 한자의 입력 작업을 수행하는 방법을 제안한다. 단, 군집 내에 오분류된 글자들은 사람이 수작업으로 제거하고, 정분류된 글자만을 일괄적으로 자동 입력한다. 이 경우, 기존의 낱자 단위로 이루어지던 입력 작업보다 훨씬 증가된 처리율과 입력의 정확도를 얻을 수 있으리라 기대된다.

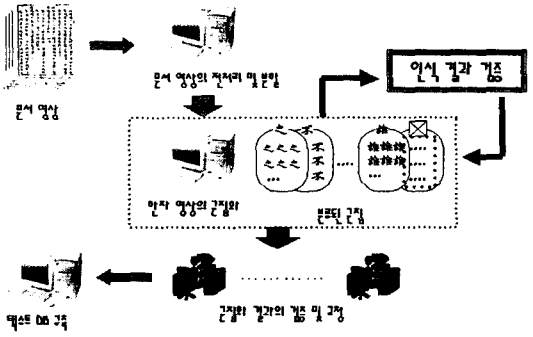

그림 2. 시스템 전체 구성도

또한, 신속하고 효율적인 디지털 라이브러리를 만들기 위하여, 인식된 결과를 신뢰하기 어려운 경우에는 인식 시점에서 자동으로 기각하여 기각 군집으로 보낸 후 수작업으로 일괄 입력할 수 있도록 기각 시스템을 추가하는 방법도 고려된다.(그림2)

본 논문에서는 자동 군집화와 자동 기각을 함으로써, 수작업에 의한 시간 및 비용을 절감시켜줄 수 있는 시스템

을 두 가지 제안하고 서로 비교해 본다.

2. 유클리디언 거리 기반 시스템의 개요

고문서 디지털 라이브러리화를 위한 시스템의 전체 프로세스는 다음과 같다.

먼저, 시스템의 입력으로 한자 문서 영상들이 들어오면 전처리와 분할을 수행하여 문서내의 각 글자들이 낱자 단위로 분할한다.(그림3)

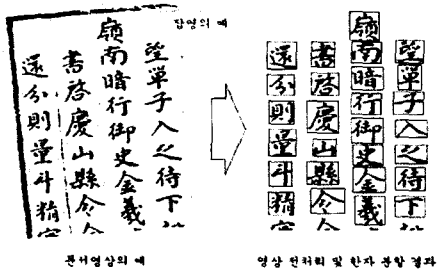

그림 3. 전처리 및 낱자 분할

분할된 각각의 낱자 이미지들은 어떤 클래스에 속하는지 인식되기 위하여 인식 모듈에 의해 호출된다(그림4). 인식 모듈은 분할된 낱자 영상을 입력으로 받아서 비선형 모양 정규화[2]와 그물눈 윤곽선 방향 특징 추출 방법[3]을 수행하고 각각의 글자 모델들과 유클리디언 거리를 비교하여 가장 거리가 짧은 것이 가장 유사하다고 판단하여 그 글자의 레이블을 출력으로 내게 된다.

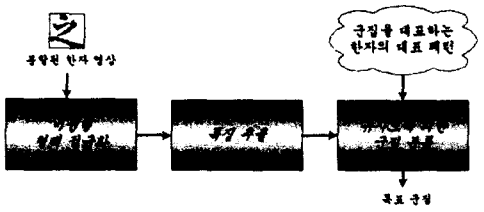

그림 4. 한자 영상의 군집화

이렇게 특정 레이블로 인식된 글자들은 그 거리와 해당 모델이 갖는 특정한 임계 값과의 크기를 비교하여 임계값보다 크기가 작을 경우 그 레이블을 가지는 미리 정의된 그룹으로 군집화 된다. 반면에, 임계 값보다 크기가 클 경우에는 따로 정의된 기각 분류 그룹에다 글자를 할당하고 나중에 이 그룹의 모든 글자들은 오퍼레이터에 의해 직접 손으로 입력되게 된다. 각각의 레이블을 가지는 그룹내의 이미지들은 그 군집의 정확성을 검증하기 위해 오퍼레이터에게 그래픽 유저 인터페이스 환경으로 보여지며, 그룹 내에 잘못된 글자가 존재하게 되면 지울 수 있다. 이와 같이, 그룹들이 정확한지 아닌지 몇 번의 교정 및 검증단계를 거치게 된 후에 정확하다고 판단되면 최종적으로 그 글자의 코드가 각 위치에 자동으로 입력된다.

본 연구에서 모든 글자들을 인식하지 않고 기각시스템을 이용하여 특정 글자만을 인식하는 이유는, 입력비용에 비해 교정비용이 훨씬 많이 들기 때문이다. 즉, 잘못 인식된 글자를 제거하고 교정하는 비용이 새로 입력하는 비용보다 훨씬 크다. 그러므로, 기각시스템을 이용하여 오류 가능성이 많은 글자들을 배제한다.

이 시스템은 직관적으로도 쉽게 이해할 수 있는 간단한 인식 방법을 사용하기 때문에 구현하기 쉬울 뿐 아니라, 입력 영상과 각 글자 모델의 평균값과의 거리만을 비교하기 때문에 처리 속도가 빠르다는 장점이 있다.

반면에, 이 방법은 인식을 수행할 때, 각 모델과의 유클리디언 거리만을 비교하기 때문에, 특징 간의 상관 관계 및 그 분포를 전혀 반영하지 못하며, 기각 시스템을 구축하기 위하여 훈련 데이터 외의 별도의 임계 값 계산을 위한 데이터가 필요하기 때문에 그 데이터에 따라 정확률과 기각 비율에 차이가 심할 수 있다는 단점이 있다.

3. 마할라노비스 거리 기반 시스템

이 시스템은 유클리디언 거리 기반 시스템의 인식 및 기각 방법에 대한 단점을 개선하기 위해 제안된다.

유클리디언 거리 기반 시스템에서는 유사도를 측정하기 위하여 유클리디언 거리 정합을 인식 방법으로 사용하였다. 이 시스템에서는 식(1)과 같이 유사도 비교를 위해 특징 간의 상관 관계 및 분포를 반영할 수 있는 마할라노비스 거리 정합 방법으로 거리 계산 방법을 변경하였다.

$$r_j = (\mathrm{x}-\mu_j)^T \Sigma^{-1} (\mathrm{x}-\mu_j) \qquad (1)$$

이 때, 마할라노비스 거리를 구하기 위한 각 모델의 공분산은 모두 같다고 가정하였다.

또한, 유클리디언 거리 기반 기각 방법은 각 모델의 특정 임계 값을 가지고 기각 여부를 판단하였다. 유클리디언 거리 기반 기각 시스템의 임계 값 설정 방법은 신뢰도가 95%이라 하고, 특정 그룹으로 모인 글자가 100개라고 할 때, 글자들을 거리가 작은 것부터 차례로 정렬하고, 5%에 해당하는 5번째까지 오분류된 글자들은 받아들인 후, 6번째 오분류된 글자에 대해 그 글자 바로 앞에 있는 글자가 가지는 거리를 그 모델의 임계 값으로 설정한다. 이 방법은 앞에서 언급한 것처럼 기각 시스템을 만들기 위해 쓰이는 데이터에 의해 정확률과 기각 비율에 차이가 심할 수 있다.

마할라노비스 거리 기반 시스템에서 사용한 기각 방법의 특징은 유클리디언 거리 기반 기각 시스템과 달리 기각 시스템을 만들기 위해 별도의 데이터를 사용하지 않으며, 사후 확률 기반의 기각 규칙이 적용된다.

사후 확률은 임의의 입력 특징(x)이 주어졌을 때 특정 클래스일 확률을 말하며, 베이스 정리 하에서 식(2)와 같다.

$$P(\omega_j \mid \mathrm{x}) = \frac{p(\mathrm{x} \mid \omega_j) P(\omega_j)}{p(\mathrm{x})} \qquad (2)$$

$$\text{, where } \quad p(\mathrm{x}) = \sum_{k=1}^{c} p(\mathrm{x} \mid \omega_k) P(\omega_k)$$

이 때, 식(2)에서 $P(\omega_j)=P(\omega)$, $\Sigma_j=\Sigma$ 라 하고, $x|\omega_j$ 가 각각 다변량 정규 분포를 따른다고 가정하면, 식(3)과 같이 나타낼 수 있다.

$$P(\omega_j|x)=\frac{\exp(-\frac{1}{2}\times r_j^2)}{\sum_{k=1}^{c}\exp(-\frac{1}{2}\times r_k^2)} \qquad (3)$$

여기서 r은 식(1)에 나오는 마할라노비스 거리를 나타낸다.

사후 확률을 이용한 기각 규칙은 사후 확률이 가장 큰 값을 가지는 클래스에 대해 그 크기가 특정 임계 값보다 작을 경우에는 기각하고 클 경우에는 그 클래스로 할당하는 방법이다(그림5). 여기서 인식된 결과는 앞에 나오는 마할라노비스 거리를 통한 인식 방법과 똑같은 결과를 가진다.

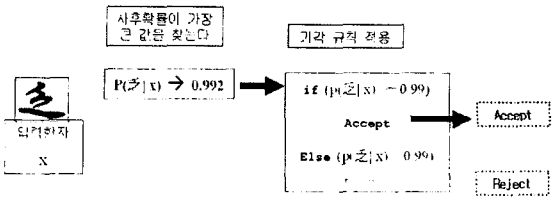

그림 5. 사후 확률 기반의 기각 시스템

4. 실험 결과 및 분석

본 연구를 위해 사용된 실험 데이터는 승정원일기 29책 3,897장(약1474000자)이며, 인식을 위한 훈련 데이 터는 1,066클래스에 대해 클래스 당 100자로 훈련 시킨 것과 2,568클래스를 클래스 당 최대 300자로 훈련시킨 것으로 나누어 실험하였다. 테스트를 위하여 사용된 데이터는 1,000장(약 400,000자)의 문서, 5,599클래스를 200장씩 나누어 5번에 걸쳐 실험하였다. 실험 결과는 [표1]과 같다.

[표 1] 인식기 성능 개선 실험 결과 (단위 %)

문서	set1	set2	set3	set4	set5	Total
(1)	68.4	80.6	83.0	83.6	88.3	80.8
(2)	70.8	83.7	86.0	86.0	89.6	83.2
(3)	75.0	89.1	90.0	90.5	94.6	87.8

여기서 (1)은 유클리디언 거리와 1,066개로 훈련시킨 모델을 이용한 것이고 (2)는 마할라노비스 거리와 1,066개로 훈련된 모델을 이용하였고, (3)은 마할라노비스 거리와 2,568개의 모델을 이용하였다. (1)과 (2)의 Total을 비교해 볼 때, 마할라노비스 거리를 이용한 것이 유클리디언 거리를 이용한 것보다 약 3% 성능 이 개선된 것을 알 수 있다. 또한, 훈련 클래스 수를 1,066개에서 2,568개로 늘렸을 때, 추가로 약 4% 증가 했음을 알 수 있다.

다음으로, 기각 방법 개선에 관한 실험 결과는 (그림6)과 같다. 기각시스템에서 사용한 훈련 데이터는 2,568클래스에 대해 클래스 당 최대 300자를 사용했으며, 테스트 데이터는 5,599클래스로 이뤄진 승정원 일기 문서 200장(77,597자)을 가지고 실험하였다.

마할라노비스 거리 기반의 시스템이 유클리디언 거리 기반의 시스템보다 처리 속도면에서 더욱 오래 걸린다는 단점이 있지만, 더 높은 정확률과 낮은 기각 개수를 보임을 알 수 있다. 현재 실제 시스템에서 사용하는 방법은 마할라노비스 거리 기반 시스템을 채택하여 운영중이다.

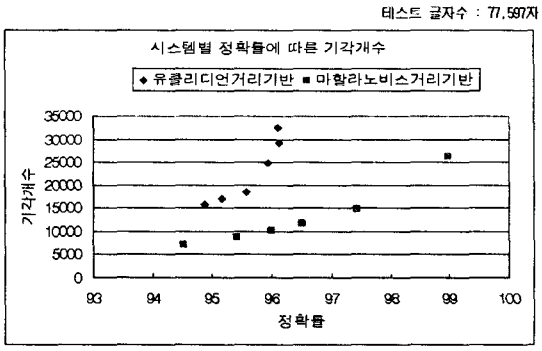

그림 6. 기각 시스템별 정확률에 따른 기각개수

5. 결론 및 향후 연구

본 논문에서는 기존의 수작업만을 통한 필기 한자 고문서의 입력 방법에 문자 인식 기법을 결합하여 좀더 효율적인 입력 방법을 제안하였다.

제안한 방법은 유클리디언 거리 기반의 인식 및 기각방법과 마할라노비스 거리 기반의 인식 및 기각 방법을 사용하였는데 전자는 처리 속도가 빠른 반면 정확률이 떨어지며, 후자는 전자보다 높은 정확률과 낮은 기각 개수를 가지지만 처리 속도가 느린 것을 알 수 있었다.

두 방법을 통한 장, 단점과 작업환경에서의 시간 및 비용 측면 등을 고려하여, 제안된 시스템을 적절히 적용하면 기존의 수작업으로만 입력하던 방법을 효율적으로 개선할 수 있을 것으로 기대한다.

향후 연구로는 첫째, 마할라노비스 거리의 임계 값을 적용하여 훈련 과정에서 배제된 한자에 대해 기각할 수 있는 방법을 고려하고자 한다. 둘째, 언어 모델을 이용하여 기각 시스템을 완전히 없앨 수 있는 가능성을 검토하고자 한다.

참고 문헌

[1] S. Hara, "OCR for CJK classical texts preliminary examination," *Proc. Pacific Neighborhood Consortium Annual Meeting*, Taipei, pp. 11-17, 2000

[2] S.W. Lee and J.S. Park, "Nonlinear shape normalization methods for the recognition of large-set handwritten characters," *Pattern Recognition*, Vol. 27, No. 7, pp. 895-902, 1994.

[3] 곽희규, 김승태, 류성호, 김진형, "필기 한자 인식을 위한 개선된 윤곽선 방향 특징," 2002 정보과학회 가을 학술발표논문집, 제29권 2호, pp. 463-465, 2002.