Journal of Korean
Data & Information Science Society
2005, Vol. 16, No. 2, pp. 313~325

Digitalization System of Historical Hanja Documents using Mahalanobis Distance-based Rejection

Min Soo Kim[1] · Jin Hyung Kim[2]

Abstract

In Korea, there exists a large corpus of handwritten historical documents that serve as a valuable resource. Most of them are hand-written by the King's chroniclers and secretaries. Recently, the historical archives of Lee dynasty have been digitalized. Since it is extremely difficult to utilize conventional OCR system, most of the processes have been performed manually. In this paper, we propose OCR-based digitalization system using Mahalanobis distance-based rejection and interface for eye inspection about historical Hanja documents. Compared with our previous work, experimental results show that the proposed system can help enhancing the overall efficiency of the process.

Keywords : classification, digital library, pattern recognition

1. 서론

역사연구에 있어서 고문서의 중요성은 아무리 강조해도 지나치지 않다. 즉, 고문서의 수집과 정리, 고문서의 고증과 성격에 대한 분석, 해석, 적절한 활용이 곧 역사연구의 출발이자 맺음이라 해도 과언은 아닐 것이다. 또한, 고문서는 귀중한 문화유산이고, 역사연구 뿐만 아니라, 당시의 사회적, 경제적 등의 다양한 면을 살펴볼 수 있는 중요한 자원이기도 하다.

우리나라에는 한자에 의해 기록된 다양하고, 상당한 양의 고문서들이 있다. 우리나라의 도서관과 정부기관들은 이러한 고문서를 보관해오고 있지만, 고문서들이 종이라는 특수성 때문에 보존와 접근에 상당한 한계를 가지고 있다. 즉, 고문서를 안전하게 보존하기 위해 정부기관은 엄청난 양의 예산을 사용하고 있으며, 또한, 보존을 위해 고문서를 연구하거나 읽고자 하는 많은 사람들의 접근에 제한이 가해지고 있다.

---

1) 제1저자 : 대전광역시 유성구 구성동 373-1 한국과학기술원 전자전산학과 인공지능연구실, 박사후연구원
E-mail : mskim@ai.kaist.ac.kr
2) 대전광역시 유성구 구성동 373-1 한국과학기술원 전자전산학과 교수

최근 고문서의 영구적 보존, 연구, 검색 등의 이유로 정부기관의 주도하에 디지털화 작업이 진행되고 있다. 디지털화된 고문서는 유지비용을 줄여줄 뿐 아니라, 일반인의 접근도 쉽게 할 수 있다. 고문서 디지털화가 고문서의 효율적 접근과 연구에 큰 도움이 될 것이라는 기대감과 함께, 수년전부터 고문서를 디지털화 하기 시작했다. 하지만, 대부분의 작업들은 거의 수작업에 의해 이루어져왔다. 예를 들면, 약 160,000면과 약 56,000,000자로 이루어진 목판 고려대장경<그림1>의 경우 2000년에 디지털화 작업이 완성되었다. 하지만, 이 작업은 완전한 수작업 입력에 의해 7년 동안 약 80억 원의 예산으로 100여명의 작업에 의한 것이었다.

<그림1> 목판 고려대장경의 예

최근 국사편찬위원회와 한국전산원 등에서는 2000년부터 지식정보자원관리사업의 한국역사분야 정보화전략사업(The Project of Information Strategy Plan in the Korean History Part)을 수행하고 있다. 이 사업에서 역사분야 문서의 자료는, 1년간 약 68억 원씩, 5년간 약 340억 원의 예산으로 디지털화를 수행한다. 하지만, 이 예산으로 디지털화할 수 있는 고문서의 양은 10억자 (1년에 2억자씩)이다. 현재, 우리나라 28개 기관에서 보유하고 있는 고문서의 양이 약 800억 자 이상으로 추정됨을 고려할 때 현재의 고문서 디지털화 속도로 우리나라 모든 고문서를 디지털화 하기 위해서는 엄청난 비용과 함께, 약 400년의 시간이 걸릴 것으로 예상된다<표1>.

<표1> 우리나라 28개 기관의 역사자료 소장 현황 총계

구분	전적류(책)	문서류(건)	기타자료(건)
총계	2,034,871	2,365,561	5,868,689

추정되는 고문서의 총 글자수 : 800억자≈약 2백만 자 * 200 * 200

* 산출근거 : 전적류만을 한정하여 책당(200면), 면당(200자)로 산출하여 계산한 수치임.

2. 고문서 디지털화 방법

2. 1 수작업에 의한 방법

한자로 쓰여진 고문서의 디지털화 방법으로 수작업 입력 방법과 광학문자인식(OCR) 방법이 고려될 수 있다. 먼저, 수작업 입력을 고려할 때, 일반적인 한자의 각 낱자는 몇 개의 획으로 이루어져 있다. 그리고, 모든 획들은 서로 다른 키코드(key code)를 가지며 획들의 조합은 하나의 글자를 이룬다. 그러므로 한 개의 한자코드 입력을 위해 적어도 몇 번의 타이핑이 필요하다(<그림2>참조). 또한 한자의 모형은 너무도 많다. 이것은 한자를 입력하기 위해서는 오퍼레이터(operator)가 입력방법을 익히는데 상당한 시간이 필요하다는 것을 의미한다. 결과적으로 전체 디지털화 작업은 매우 힘들고 시간 소모적인 작업이 될 것이며 상당한 노동력이 필요하고 시간과 비용이 대단히 많이 들게 된다.

<그림 2> 수작업 입력 인터페이스의 예

2.2 광학문자인식 방법

요즘, 광학문자인식의 사용이 큰 주목을 받아오고 있다. 최근, 인쇄체 문서에 대한 문자인식기술이 좋은 효과를 거두고 있고 필기체 문자에 대한 광학문자인식 기법에 대한 많은 연구가 있었다(Hara(2000), Tung et al(1994)). 하지만, 필기체 한자 고문서의 디지털화에 이러한 기술을 완전히 이용하는 것은 쉽지 않다. 첫 번째 어려움은 필기자마다 필기습관이 달라서 오는 글자모양의 변형(<그림3(a)>)이고, 두 번째는 붓으로 쓰여짐에 의한 글자의 번짐 현상이 문서에 자주 등장하는 것이다(<그림3(b)>). 결국, 고문서에 대한 광학문자인식의 수행력의 저하에 따라 완벽한 결과를 기대하는 것은 불가능하고, 따라서 고문서에 대한 광학문자인식의 사용이 제한적일 수밖에 없다(Tseng et al(1998)).

따라서, 광학문자인식에 의한 방법을 사용하지 못하고 수작업에 의한 디지털화가

수행되고 있는 가장 큰 이유는 광학문자인식의 수행력이 높지 않다는데 있다. 현재 대부분의 고문서 디지털화는 거의 수작업에 의해 이루어져 왔다. 광학문자인식 방법의 성능이 높지 않을 뿐만 아니라 오 분류된 글자를 고치는 비용이 순수하게 한자를 입력하는 비용보다 훨씬 크기 때문이다. 실제로, 오 분류된 글자를 교정하는 비용이 순수하게 입력하는 비용보다 3-4배 더 크기 때문에 약 80% 정도의 정 분류율 성능을 가지는 광학문자인식기를 사용하고 교정하는 경우보다 차라리 순수하게 수작업 입력을 하는 것이 비용적인 면이나 시간적인 면에서 유리하다고 여겨진다.

(a) 각 글자에 대한 두가지 표현들 (b) 번짐 현상이 있는 글자들

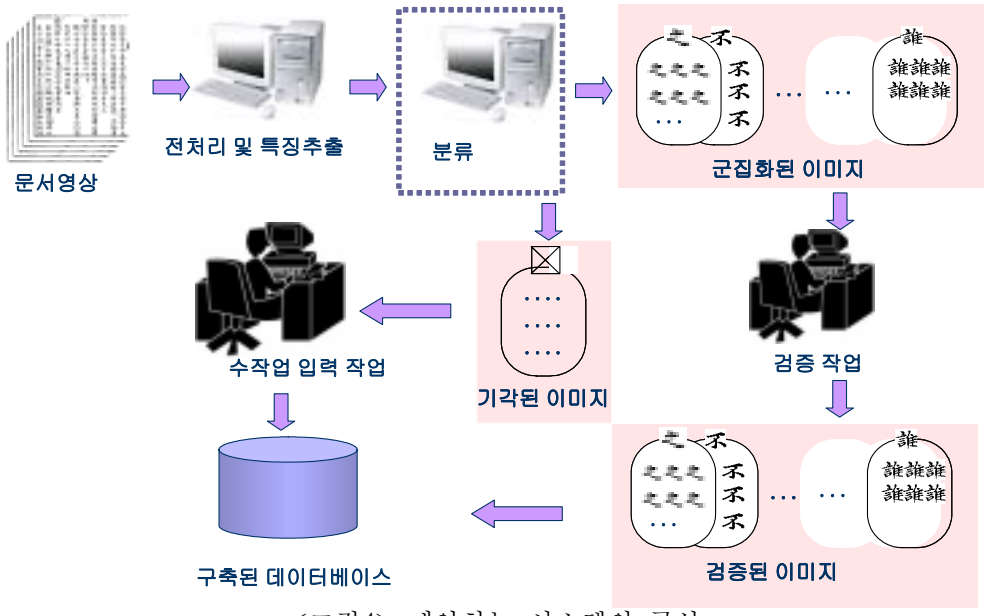
<그림3> 고문서 한자 영상의 특징

3. 제안하는 고문서 디지털화 방법

본 논문에서는 고문서의 디지털화에 시간과 비용을 절약할 수 있는 광학문자인식과 수작업 입력방법의 장단점을 서로 보완한 새로운 방법을 제안한다. 현실적으로, 필기로 쓰여 지고, 특히 붓에 의해 쓰여진 고문서의 광학문자인식의 성능은 제한적이다. 제안하는 방법의 핵심은 분류에 대해서, 기각방법을 이용하는 것이다. 즉, 분류기가 오 분류의 가능성 높은 낱자영상에 대해서는 분류를 보류하고 오퍼레이터에게 넘기는데 있다. 실제로, 오퍼레이터에 의해 순수하게 입력되는 비용이 오 분류된 글자를 수정하거나 교정하는데 드는 비용보다 훨씬 저렴하다. 제안하는 방법은 마할라노비스 거리(Mahalanobis distance) 기반의 분류 및 기각방법을 이용한다.

즉, 제안된 시스템은 분류의 정확도를 판단하여 오 분류의 가능성이 큰 글자들에 대해서는 기각하는 과정을 두는데, 분류기는 실제로 선형판별분석(linear discriminant analysis)과 같은 결과를 주지만, 선형판별함수를 사용하지 않고 사후확률(posterior probability)을 계산하여 분류와 기각에 이용한다.

제안된 시스템의 구성은 <그림4>와 같이 1) 전처리 및 특징추출단계, 2) 마할라노비스 거리를 이용한 분류 및 군집화 단계, 3) 수작업에 의한 검증 및 교정단계의 3단계로 크게 이루어져 있다. 전처리 및 특징추출단계에서는 수백 장의 문서영상이 빠르게 각각의 낱자영상으로 분할되고 특징추출된다. 이 단계에서는 비선형 모양 정규화(Nonlinear Shape Normalization)방법, 윤곽 방향 특징(Contour Direction Feature)추출 방법, 다음으로 고문서한자 인식을 위한 마할라노비스 거리기반 고문서 한자전용 광학문자인식 단계가 적용된다. 분류기는 같은 글자로 할당된 낱자영상들끼리 군집화가 수행된다. 마지막으로 군집화된 낱자영상들은 검증 인터페이스를 통해 오퍼레이터의 수작업 검증에 의해 텍스트 데이터베이스(database)로 구축된다.



<그림4> 제안하는 시스템의 구성

위의 설명에서 군집화란, <그림5(a)>와 같이 같은 글자로 분류된 낱자영상들을 모아서 오퍼레이터가 검증하기 쉽게 하는 것을 의미한다. 이것은 광학문자인식과 수작업 교정의 효율적 결합에 큰 도움을 준 부분으로 기존의 광학문자인식을 통한 디지털화에서는 오 분류된 글자를 하나하나 찾아가면서 검증했던 것과는 달리 본 시스템에서 오퍼레이터가 쉽게 눈으로 검증할 수 있는 인터페이스를 사용함으로써 시간적, 경제적으로 매우 효율적인 결과를 얻을 수 있었다. <그림5(a)>는 '之'라고 여겨지는 낱자영상들을 군집화한 예이다. 그리고, <그림5(b)>는 오퍼레이터에 의해 오 분류된 낱자영상들이 제거되는 예이다. 검증작업이 끝나면 오퍼레이터는 하나의 레이블을 갖는 낱자영상들은 한꺼번에 입력할 수 있다.

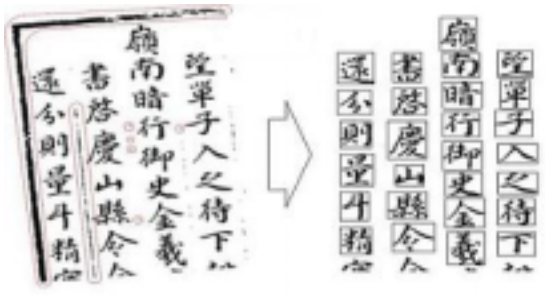
(a) (b)

<그림 5> (a) '之'로 여기지는 영상들을 군집화한 예 (b) 오 분류된 문자영상들이 제거되는 예

3.1 전처리

3.1.1 문서 기울기 교정

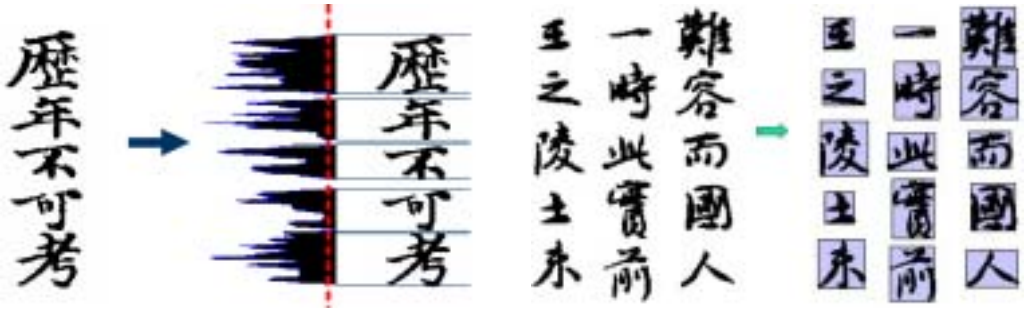
입력 영상은 수작업에 의해서 스캔닝된 문서이므로 기울기가 존재한다. 문서의 기울기를 교정하면 문자열을 보다 정확하게 추출할 수 있다. 이 때 문서 기울기가 심하지 않고 문자열이 세로로 되어있다고 가정한다. 수직선을 0^o라고 할 때 -3^o에서 $+3^o$ 사이에 투영을 수행하여 이웃하는 투영값간의 차이의 합이 가장 큰 각도로 기울기를 교정한다. <그림 6>는 기울기 교정을 거쳐 완전히 낱자로 분할된 결과를 보여주고 있다.



<그림 6> 기울기 교정 후 낱자 분할된 예

3.1.2 문자열 추출 및 문자 분할

고문서는 문서 내 문자열이 세로로 되어 있다고 가정하여 수직으로 투영한다. 그리고, 투영된 값들에 대해 가우시안 커널 함수(Gaussian kernel function)를 이용하여 비모수적 커널 평활(kernel smoothing)을 수행한다(Parzen(1962), Nadaraya(1964)). 수정된 투영 값을 좌에서 우로 스캔하면서 기울기의 변화를 이용하여 투영값의 계곡(valley)을 찾는다. 각 계곡의 하얀색 봉우리(peak)가 존재하는 부분이 문서 내 문자열의 위치이다. 추출된 문자열로부터 글자를 구성할 수 있는 낱자 단위로 분리하기 위해서 문자분할 단계가 필요하다. 본 연구에서는 투영법(projection method)을 사용한다. 투영법은 세로로 구성된 문자열에 대해서 수평방향으로 투영을 수행하여 누적된 픽셀 값을 계산하는 방법이다. 여기서 임계 값을 넘는 연속된 누적 값을 가지는 부분을 하나의 낱자로 결정한다(<그림7>참조).

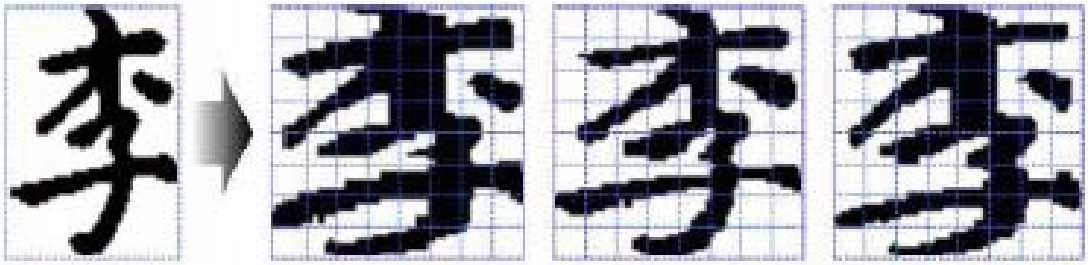


(a) (b)

<그림7> (a) 투영법에 의해 분리되는 문자영상과 (b) 문자 분할된 영상의 결과

3.2 비선형 형태 정규화 및 윤곽선 방향 특징 추출방법

필기로 쓰여진 글자들의 변형 또는 왜곡을 개선하기 위해, 정규화 작업을 거친다. 비선형 형태 정규화 방법에는 다양한 방법들이 제안되어 있는데, 점밀도에 의한 방법, 교차횟수에 의한 방법, 획 간격에 의한 방법 등이 있으나, 비선형 형태 정규화 방법들의 정량적 평가를 통해, 윤곽선 방향 특징을 사용하는 경우 획 간격을 이용하는 방법(Tsukumo 와 Tanaka(1988))이 우수한 수행 성능을 보여서 본 연구에서는 획 간격에 의한 방법을 사용하였다<그림8 (d)>.

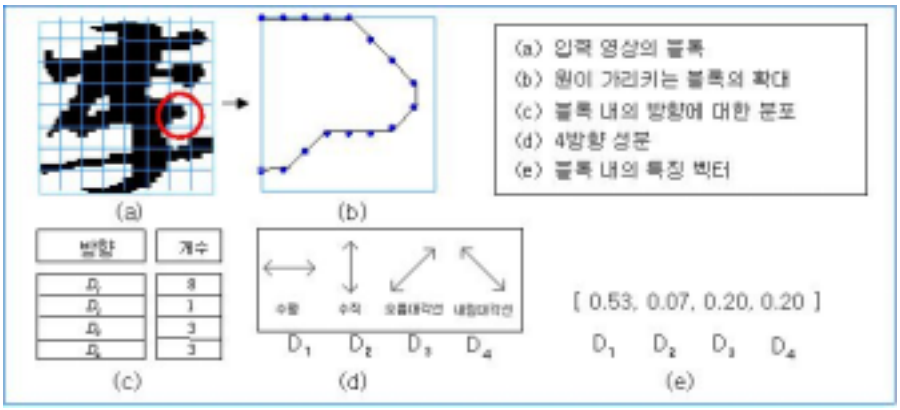


(a) (b) (c) (d)

<그림8> 8×8 블럭으로 정규화된 영상

[(a) 원래의 영상과 각각 (b) 점밀도 (c) 교차횟수 (d)획 간격 방법을 이용한 영상들]

윤곽선 방향 특징은 입력영상을 여러 개의 격자 단위로 나누어서, 각 격자 내의 화소의 방향성분 분포를 결정하는 것이다(Kimura 와 Shridhar(1991)). 먼저, 낱자 분할되고 정규화과정을 거친 낱자 영상들로부터 4가지의 방향성분을 표현하는 윤곽선 방향 특징을 추출한다. 방향성분은 영상의 외곽선화소에서 정의되고, 각 픽셀 방향은 경사도 값에 의해 결정된다. 그 후 결정된 방향을 4방향으로 양자화해서 방향성분 분포를 개수하고 격자의 크기에 의해 정규화 한다. <그림9>의 8×8로 격자된 낱자영상으로부터 임의의 한 블록에서 4가지의 방향 특징을 추출하는 방법을 보여주고 있다. (a)와 같이 낱자영상의 한 블록으로부터 방향성분(b)들을 만들고 4방향의 각 개수를 계산한다. 결과적으로, 이러한 과정에서 256차원(8×8×4)의 특징을 얻을 수 있다.



<그림 9> 윤곽선 방향 특징추출 방법

3.3 마할라노비스 거리 기반 사후확률에 따른 분류 및 기각방법

본 논문에서 고문서 낱자인식을 위해 사용된 분류와 기각은 마할라노비스 거리에 기반한 방법이다. 상당수의 패턴인식 시스템은 이론적으로 베이즈(Bayes) 판별규칙에 기반한다. 베이즈 판별규칙은 각 모형의 확률밀도함수와 주어진 손실함수를 사용하여 패턴들을 소속모형에 분류하는 방법이다. 베이즈 판별규칙을 이용하는 통계적인 방법

은 i번째 모형에 속하는 패턴(특징벡터) $\boldsymbol{x} = (x_1, \cdots, x_p)$가 평균 μ_i와 분산 Σ_i를 갖는 다변량 분포를 따른다는 가정 하에서 만들어진 판별함수를 주로 이용한다. 여기서 조건부 확률밀도함수 $p(x|\pi_i)$, $i = 1, \cdots, c$ 를 추정하고자 할 때, 즉 모형 π_i에서 패턴 $x \in R^p$의 다변량 정규확률밀도함수는 다음과 같다.

$$p(\boldsymbol{x}|\pi_i) = (2\pi)^{-p/2}|\Sigma_i|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\mu_i)^T\Sigma_i^{-1}(\boldsymbol{x}-\mu_i)\right\}$$

이 때, π_i는 i번째 모형, c는 모형의 개수, p는 특징벡터의 수(혹은 변수의 수)를 의미한다. 본 연구에서는 3.2절에서 설명한 윤곽선 방향 특징벡터를 사용하여 고문서 한자를 분류 한다. 그런데 분류될 모형의 수가 2500개 이상이고 특징벡터의 수가 256개 ($p = 256$)이므로, 만약 모형들 간의 공분산이 다르다고 가정하는 위의 확률밀도함수를 사용하여 문제를 해결하려면, 8천만 개 이상의 모수(parameter)를 추정해야 한다. 그러므로 본 연구에서는 각 모형별 공분산 행렬이 모두 같다고 가정하였다(선형판별분석과 같은 조건이다). 모든 모형에 대하여 공분산 행렬이 Σ로 동일하게 주어진다면, i번째 모형에 속하는 패턴 x의 확률밀도함수는

$$p(\boldsymbol{x}|\pi_i) = (2\pi)^{-p/2}|\Sigma|^{-1/2}\exp\left\{-\frac{1}{2}(\boldsymbol{x}-\mu_i)^T\Sigma^{-1}(\boldsymbol{x}-\mu_i)\right\}$$

이 된다. 일반적인 통계적 판별분석에서는 이 경우 선형판별함수(linear discriminant function)를 사용하여 문제를 해결하지만, 본 시스템은 기각방법으로 사후확률 $p(\pi_i|x)$의 임계치를 사용하고자 하므로 선형판별함수를 사용하지 않고 직접 사후확률을 계산한다. 이 경우에 사전확률들이 모두 같다고 가정하고 $p(\pi_i|x)$을 계산하면 다음 식과 같음을 알 수 있다. 그래서 사후확률이 가장 큰 모형에 표본 x를 할당하는 분류를 수행할 수 있다.

$$p(\pi_i|\boldsymbol{x}) = \frac{\exp(-\frac{1}{2}\times r_i^2)}{\sum_{k=1}^{c}\exp(-\frac{1}{2}\times r_k^2)},\ r_k = ((\boldsymbol{x}-\mu_k)^T\Sigma^{-1}(\boldsymbol{x}-\mu_k)^{\frac{1}{2}}$$

즉, 본 시스템은 모든 분류결과를 채택하는 것이 아니라 사후확률에 대해서 미리 임계치를 두어, 계산된 가장 큰 사후확률 값이 임계치 보다 적으면 분류결과를 채택하지 않고 기각한다.$\max\ p(\pi_j|\boldsymbol{x}) < \theta$ 인 낱자 영상들은 광학문자인식기에 의해 자동분류되지 않고 모여져서 오퍼레이터에게 넘겨지고 수작업으로 입력하게 된다. <그림 10>은 승정원 일기(필사본)의 영상으로부터 기각방법을 이용하여 분류한 예이다. <그림10>의 오른쪽 결과에서 ? 로 표시된 낱자 영상은 시스템이 분류를 보류하여 기각한 결과이다. 이 낱자 영상들에 대해서는 분류가 이루어지지 않고 바로 오퍼레이터

에게 넘어가서 수작업 입력이 수행된다.

<그림 10> 고문서(승정원 일기)에 대한 기각방법을 이용한 분류결과의 예
(? 는 기각한 낱자 영상을 의미한다)

4. 실험

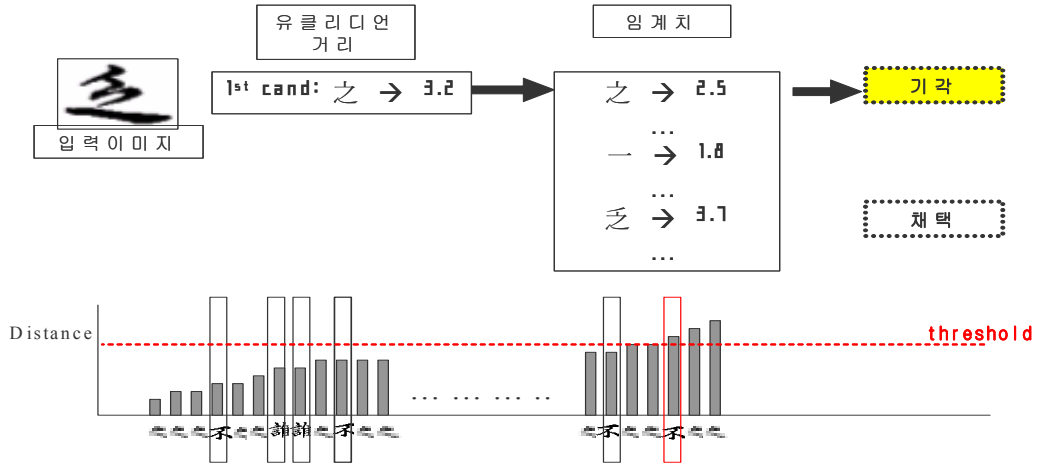
아직까지 다른 연구기관에서 발표된 고문서 한자 전용 광학문자인식 기반 디지털화 시스템이 없는 것으로 조사되었기 때문에, 실험에서는 본 연구팀에 의해 기 개발된 고문서 한자 디지털화 시스템의 성능과 비교한다. 4.1에서는 기존의 시스템에 대한 간략한 설명을, 4.2에서는 실험결과를 제시한다.

4.1 기 개발된 고문서 한자 디지털화 시스템

현재, 국내 고문서 디지털화 작업은 거의 수작업에 의해 이루어져 왔었고, 2003년 초에 처음, KAIST에서 유클리디언(Euclidean) 거리 기반 패턴 매칭 방법을 이용한 고문서 디지털화 시스템을 개발하여 운영하기 시작했다. 이 시스템은 유클리디언 거리를 이용한 분류 및 군집화 방법을 사용함과 동시에 기각방법을 이용하는데, <그림 11>를 참조하자.

이 방법은 유클리디언 거리를 이용하여 거리가 가장 짧은 모형에 입력 낱자영상을

분류하고 그 결과를 기각 또는 채택한다. 즉, 모형별 거리의 임계치를 설정하고 실제 할당된 모형과 입력 영상의 특징벡터 사이의 거리를 해당 모형의 임계치와 비교하여 임계치보다 작으면 분류결과를 채택하고, 크면 기각하는 방법을 사용한다. 이때, 모형별 임계치 설정 방법은 임계치 설정을 위한 별도의 데이터 집합을 두어 그 데이터를 이용하여 분류를 수행한 후, 모형 별로 거리가 작은 순서로 분류 결과를 정렬한다. 그리고 오 분류된 결과를 각 모형에 대해 어느 정도 포함시킬 것인가를 결정한 뒤 그에 따른 모형별 임계치를 설정한다. <그림11>의 예는 '之' 글자라고 인식되어 모여진 표본이 100개인 경우, 6번째의 오 분류가 발생한 거리를 95% 거리 임계치로 설정한다. 그래서 만약, 설정된 '之 'ㄱ글자모형의 거리 임계치가 2.5인 경우이고 입력 영상과 모형과의 거리가 3.2라면 이 경우는 분류가 '之'로 되었다고 할지라도 기각이 되는 것이다(이것을 여기서는 거리 임계치 방법이라고 한다).

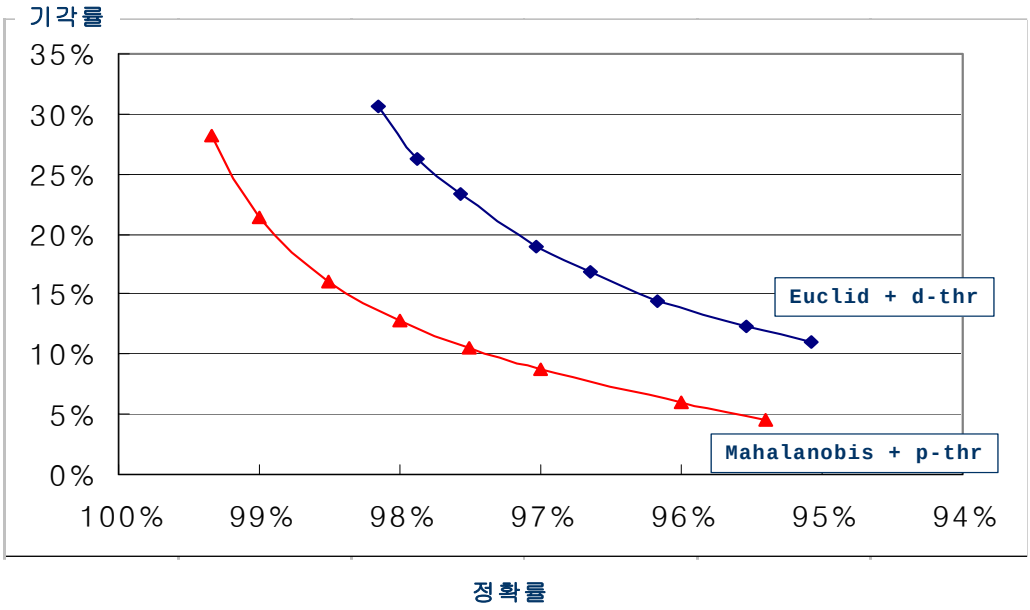


<그림11> 유클리디언 거리 분류방법과 거리임계치 기각방법

4.2 실험 결과

제안된 방법의 수행력을 알아보기 위해, 한국학 고문서인 승정원 일기 29책의 3,927장의 문서 영상을 사용하였다. 문서 내에 존재하는 한자의 수는 약 150만 자에 달하며 5,600개 이상의 글자모형을 가지고 있다. 학습 데이터는 문서 내에서 빈번히 사용되는 글자 2,556개 모형을 정의하여 사용했으며, 이것은 문서 내 전체 한자 빈도수의 약 99%을 차지하고 있다. 학습 샘플 수는 모형 당 최대 300자, 최소 20자를 이용하여 학습시켰다. 글자모형별 표본수가 다른 것은 글자별로 빈도수가 매우 큰 차이를 보이기 때문에 어떤 글자 모형에 대해서는 20개 밖에 얻을 수가 없는 경우도 있었다. 테스트 데이터로는 학습에서 사용하지 않은 문서 중에서 200장을 임의로 선택하여 사용하였고, 이 문서는 학습 시 정의하지 않은 모형도 모두 포함하고 있으며, 테스트 글자수는 78,756개이다. 실험 결과는 <그림12>에 제시되었다. 그림에서 윗쪽 그래프는 유클리디언 거리 기반 인식 방법과 거리 임계치 설정을 통한 기각 규칙을 수행했을 때의 결과를 나타낸다. 아래쪽 그래프는 본 논문에서 제안한 마할라노비스 거리를 이용한 인식과 사후확률에 기반한 기각 임계치를 적용한 결과를 나타낸다. 제안된 방법이

같은 정확률(채택된 낱자 영상들의 인식률)에서 보다 더 낮은 기각률을 가짐으로써 좀더 효과적인 수행력을 보임을 알 수 있다. 예를 들어, 같은 정확률 97%에서 기존의 시스템은 약 20% 정도의 한자를 기각하는 반면, 제안된 시스템은 10% 미만을 한자를 기각하는 것을 의미한다. 또한 같은 기각률에서 비교해보면, 약 10%의 기각률에서 기존의 시스템은 약 95%의 정확률을, 제안된 시스템은 97.5%의 정확률을 보이고 있다.



<그림12> 기존시스템(유클리디언 거리와 거리임계치 기반)과 제안된 시스템(마할라노비스 거리와 사후확률 기반)에 대한 수행력 비교

5. 결론

최근 향상되고 있는 우리나라의 인터넷 인프라구조(infrastructure)와 함께, 정부기관들은 역사 연구자들이나 일반 대중들이 인터넷 등을 통해 고문서에 대한 정보를 이용하고 검색할 수 있도록 고문서 디지털화 작업을 해오고 있다. 하지만, 지금까지의 디지털화는 거의 수작업에 의해 이루어져왔다. 그 이유는 필기로 쓰여진 고문서, 특히 붓으로 쓰여진 고문서에 대해 컴퓨터를 이용한 자동분류를 이용하기가 상당히 어렵기 때문이다. 또한, 수작업을 이용했을 때의 비용이 오히려 자동분류방법을 사용했을 때보다 저렴했기 때문이다. 그래서, 본 연구에서는 고문서 디지털화 작업에 광학문자인식과 수작업 입력의 효과적인 상호보완 및 결합을 이용하여 고문서 디지털 작업을 보다 효율적으로 수행할 수 있도록 전체 시스템을 개발하고 제안하였다. 특히, 본 논문은 마할라노비스 거리 기반 분류 및 기각 방법을 제안하여, 실험 결과에서 알 수 있듯이 기존 시스템의 성능을 월등히 향상 시켰다.

참고문헌

1. Hara, S.(2000). OCR for CJK classical texts preliminary examination. *Proc. Pacific Neighborhood Consortium(PNC) Annual Meeting*, Taipei, Taiwan, 11-17.
2. Kimura, F. and Shridhar, M.(1991). Handwritten Numerical Recognition based on Multiple Algorithm, *Pattern Recogntion*, Vol. 24, No. 10, 969-983.
3. Nadaraya, E. A.(1964). On estimating regression. *Theory of Probability and Its Applications*, Vol. 9, No. 1, 141-142.
4. Parzen, E.(1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, Vol 33, No. 3,105-1076.
5. Tseng, Y. H., Kuo, C. C. and Lee, H. J.(1998). Speeding up Chinese character recognition in an automatic document reading system, *Pattern Recognition*, vol 31, no. 11, 1601-1612.
6. Tsukumo, J. and Tanaka, H.(1988). Classification of Handprinted Chinese Characters using Nonlinear Normalization Methods. *In proceedings of 9th International Conference on Pattern Recognition*, 168-171.
7. Tung, C. H., Lee, H. J. and Tsai, J. Y.(1994). Multi-stage pre-candidate selection in handwritten Chinese character recognition systems, *Pattern Recognition*, vol. 27, no. 8, 1093-1102.

[2005년 1월 접수, 2005년 5월 채택]