고문서 전산화를 위한 문서 인식 기법

조규태° 김진식 이성훈 김자환 김민수 김진형
한국과학기술원 전자전산학과
{ktcho°, jskim, leesh, jahwan, mskim, jkim}@ai.kaist.ac.kr

Recognition for Digitizing Historical Document Pages

Kyutae Cho°, Jinsik Kim, Seonghun Lee, Jahwan Kim, Minsoo Kim, JinHyung Kim
Dept. of Electrical Engineering & Computer Science, KAIST

요 약

역사적 가치가 높은 고문서의 훼손을 방지하고 접근을 용이하게 하기 위해서 고문서 전산화가 필요하다. 이를 위한 작업에서는 고문서의 방대한 양을 빠르고 정확하게 처리하는 기술이 필수적이다. 본 논문은 고문서 전산화를 위한 분할 방법과 인식 방법을 제안한다. 인식을 이용한 분할 방법을 통해 신속하면서도 정확하게 문서내의 문자영역을 찾아낸다. 또한 인식기로부터 생성된 점수를 확률화 하여 신뢰도를 높이고 이를 문자의 모양 및 문맥정보와 결합을 통해 분할과 후처리를 수행한다. 제안하는 방법은 고문서 전산화 과정에서 사람의 수작업을 최소화하기 위해 유용하다.

1. 서 론

국내에 존재하는 고문서는 그 시대의 생활상이나 제도 및 경제적인 상황 등을 이해하는데 중요한 단서가 된다는 점에서 그 역사적 가치와 보존 가치가 높지만 종이의 노화로 인한 훼손 문제와 파손 위험으로 인해 일반인이 접근하기 어려운 문제가 존재한다(그림1). 따라서 고문서의 파손 위험을 줄이고 접근을 용이하게 하기 위하여 고문서의 전산화가 필수적이다[1].

그림 1. 원본의 노후화와 관리의 어려움

고문서 전산화를 효율적으로 수행하기 위해서는 수작업을 최소화 한 자동 분할 및 인식 시스템이 필요하다. 그러나 고문서가 두꺼운 붓을 사용한 필기체 문자로 이루어졌기 때문에 문자들이 겹치거나 접촉 된 경우가 많아서 기존의 문자 분할 알고리즘을 그대로 적용할 수 없다. 또한 인식클래스가 5000개 이상으로 방대하고 새로운 문자가 추가되고 있는 상황이기 때문에 기존의 신경망 기반의 인식 구조는 높은 재훈련 비용으로 인해 비효율적인 접근 방법이라 할 수 있다. 따라서 본 논문에서는 고문서 분할의 문제점을 해결하기 위해서 인식기 점수, 문자의 모양 및 문맥정보를 이용한다. 또한 인식기 훈련의 비용이 저렴한 거리 기반의 인식기법을 사용하여 다양한 문자 정규화 방법과 특징에 대한 실험을 수행하였다. 인식기로부터 생산된 점수는 확률 값으로 변환되어 분할 및 후처리에 활용된다.

본 논문의 구성은 다음과 같다. 2장은 제안하는 시스템의 구조를 설명한다. 3장은 문자 분할 및 인식에 대해 설명한다. 4장에서는 인식을 위한 정규화 및 특징에 대해 설명하고 5장에서는 실험 결과를 보인다.

2. 시스템 개요

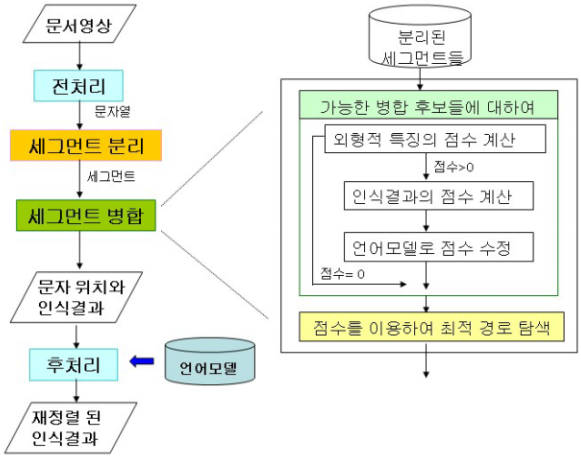

그림 2. 시스템의 흐름도

본 시스템의 대략적인 흐름은 그림 2와 같다. 우선 문서영상이 입력으로 들어오면 투영과 스무딩 기법을 사용하여 문자열을 찾아내고, 글자를 분리할 수 있는 최소단위인 세그먼트들로 분리해낸다. 다음으로 세그먼트들을 합쳐가면서 인식을 수행하는데 이때 문자의 외형적 특징을 사용하여 병합후보를 줄이게 된다. 또 문맥정보를 사용하여 분리된 세그먼트 각각이 글자가 될 수 있는 가능성을 줄인다. 이를 통해 문자의 위치, 인식결과 그리고 인식결과에 대한 확률 값이 생성되게 된다. 언어모델과 인식결과 값을 사용하여 인식 후보를 재정렬하여 후처리

를 수행한다.

3. 분할 및 인식

고문서 분할에서 가장 큰 문제점은 글자의 겹침과 접촉문제이다[2]. 따라서 겹치거나 접촉된 부분을 잘라내고 글자들을 최소단위로 나누기 위해 비선형 분할 경로를 이용하였다[2][3](그림3).

그림 3. 비선형 분할 경로

다음으로 비선형 분할 경로로 나누어진 세그먼트들을 글자의 가능성을 고려하여 병합하여야 한다. 글자의 가능성을 계산하기 위한 병합 점수는 입력영상에 대하여 결정되는 문자와 외형적 특징의 사후확률을 사용하였다. 문자열 내에 결정되는 문자 개수를 k라고 하면 분리된 세그먼트를 병합한 입력 영상은 $X = x_1 x_2 .....x_k$ 로 정의되고, 결정되는 전체 문자열은 $S = s_1 s_2 .....s_k$, 문자의 외형적 특징은 $G = g_1 g_2 .....g_k$ 로 정의한다. 이때 사후 확률은 $P(G,S \mid X)$ 로 표현 할 수 있다. 이를 최대화 하는 식은 다음과 같이 전개된다.

$$S = \arg\max_s P(G,S \mid X)$$

$$= \arg\max_s P(G,S,X) \quad (1)$$

$$\approx \arg\max_s P(G)^{\lambda} P(S,X) \quad (2)$$

$$= \arg\max_s P(G)^{\lambda} P(S)P(X \mid S) \quad (3)$$

$$= \arg\max_s \lambda \log P(G) + \log P(S) + \log P(X \mid S) \quad (4)$$

식(1)에서 (2)으로 전개될 때 외형적 특징 G가 문자열 S와 입력영상 X에 독립이라고 볼 수 없기 때문에 이를 보정하기 위하여 λ 를 사용하였다. λ 는 실험을 통해 결정하였다. 식(4)에서 $P(G)$ 는 외형적 특징 모델, $P(S)$ 는 언어모델, $P(X \mid S)$ 는 인식기 모델이라고 볼 수 있다. 따라서 이 세가지 모델로부터 얻어낸 점수의 합의 형태로 병합점수를 계산한다.

외형적특징

문자의 외형적인 모양을 이용하여 글자의 가능성을 계산한다. 이를 위해 파즌(Parzen) 윈도우를 사용하여 비모수적 분포를 추정하였다. 사용한 특징은 문자높이, 바운딩 박스 모양, 바운딩박스 내의 공간 정도를 이용하였다.

인식기 점수 및 확률화

외형적 특징에 의해 기각되지 않은 병합 후보 영상에 대하여 인식을 수행한다. 이때 인식은 클래스 간의 상관 관계를 고려한 마할라노비스(Mahalanobis) 거리 기반의 인식을 수행한다[1]. 거리 기반의 인식기를 사용하여 생성되는 값은 입력 영상의 해당 군집으로부터의 거리이고 이를 다른 모델과 결합하기 위해서는 확률 값으로의 변화가 필요하다. 인식기 점수를 확률 값으로 추정하는 여러 가지 방법 중 본 논문에서는 isotonic regression을 사용하였다[4]. 이는 비모수적 추정방법이기 때문에 모수를 추정하기 위한 비용이 없고 구현이 간편한 장점을 가진다. 또한 이 방법은 결정 이론(decision theory)적인 관점에서 최적의 해를 구함을 증명할 수 있다(submitted). 인식에 사용한 정규화 방법 및 특징은 다음 장에 기술하였다.

문맥정보

훈련데이터로부터 통계적으로 생산한 바이그램(bi-gram) 언어모델을 사용한다. 이는 Katz의 back-off 모델을 사용하여 구현하였다[5].

위의 세가지 기준을 이용하여 계산된 병합 점수를 사용하여 문자열 내의 최적 경로를 탐색하고 그에 대한 문자 영역, 인식 결과 및 인식 확률을 생성해 낸다(그림4).

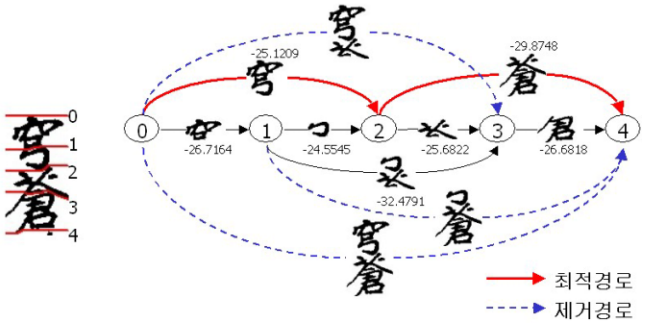

그림 4. 최적 경로 탐색

4. 인식을 위한 특징추출

4.1. 정규화 및 격자 설정

4.1.1 선간격 정규화

선간격 정규화 방법[8]은 획의 간격을 가급적 동일한 간격으로 조정해 주기 위한 정규화 방법이다. 한문 획의 간격은 동일한 글자라고 해도 조금씩 차이가 날 수 있기 때문에 이를 균일하게 해주면 특징에서 나타나는 차이를 줄여줄 수 있다.

4.1.2 격자 설정

격자 설정에는 일정 길이로 나누는 선형격자와 흑화소 밀도 등을 이용한 비선형 격자가 존재 한다. 일반적으로 비선형 격자가 성능이 나은 것으로 알려져 있지만 선 간격 정규화에서 이미 비선형 정규화가 되었으므로 선형 격자를 이용하였다.

4.2. 특징 추출

4.2.1 윤곽선 방향 특징 (CDF)

윤곽선 방향 특징[6]은 하나의 격자에서 흑화소의 경계 화소에서 3x3 수직 수평 소벨 연산자(Sobel operator)의 결과로 방향을 결정한다. 해당 방향은 비슷한 방향으로 그룹화하여 각 방향 그룹별로 누적한 값을 특징값으로 사용하며 전체를 수직, 수

평, 두 가지 대각방향의 4방향이나 더 세분화한 8방향 값 등을 사용한다(그림6). 이러한 방향 특징은 해당 격자의 가로세로 길이 합으로 나누어 정규화한다. 수평방향 소벨 연산자 결과를 G_x, 수직방향 소벨 연산자 결과를 G_y 라 하면 해당 방향 $\alpha = \tan^{-1}\frac{G_x}{G_y}$ 에 의해 정의된다.

4.2.2 방향각 특징 (ADF)

방향각 특징[7]은 CDF가 흑화소의 경계방향을 따라 방향을 계산하는 것에 반하여 모든 점에 대하여 방향을 계산한다. 이때 방향이 없는 방향, 즉 방향을 계산하는 해당 흑화소 주변의 3x3 윈도우가 대칭인 경우 해당 포인트가 흑화소일 때와 해당 포인트가 흑화소가 아닐 때를 각각 2가지 다른 방향으로 계산한 후 이 방향이 없는 특징에 대해서는 전체 문자 이미지의 흑화소, 백화소 수로 나누어 정규화 한다. 방향각 특징에 대해서도 CDF와 마찬가지로 4, 8방향으로 각각 나누며 8방향에 방향이 없는 방향 2방향이 추가되면 10방향으로 간주한다.

[α 값에 대한 4방향 분할] [α 값에 대한 8방향 분할]

그림 6. 방향 설정 예제

5. 실 험

실험은 대표적인 고문서인 승정원 일기 29책의 2927장의 문서 영상을 사용하였다. 문서 내의 5600개 이상의 클래스 중 빈도수로 상위 2556클래스를 대상으로 하였고 이는 전체의 99%을 차지한다. 학습 데이터는 각 클래스당 최대 300자를 이용하였고 테스트는 학습에 참여하지 않은 문서 200장의 78,767개의 데이터를 사용하였다.

기존의 실험[1]에서 사용된 방법은 선 간격 정규화와 8x8 선형격자에 4방향 CDF이고 이를 적용하면 약 92.7%의 정확도를 보인다. 새로운 특징을 사용하여 실험한 결과는 표1에 정리 되어 있다.

표 1. ADF 적용 결과(정확도)

	ADF4방향	ADF8방향
6x6 선형 격자	92.11%	93.09%
8x8 선형 격자	92.93%	93.25%

표 2. 언어모델을 이용한 후처리 적용 결과(정확도)

후처리 적용 전	후처리 적용 후
93.25%	96.4%

8x8 선형 격자와 ADF 8방향을 사용한 경우가 93.25%로 가장 좋은 것으로 나타났다. 또한 표2와 같이 후처리를 적용하면 96.4%로 인식률 향상을 보인다. 이는 인식결과 값을 확률로 추정함으로써 언어모델과의 결합이 보다 정확한 방향으로 이루어졌음을 의미한다.

6. 결 론

본 논문에서는 고문서의 전산화를 위해 문자영역을 발견하고 인식하는 방법을 제안하였다. 제안하는 방법은 수작업을 최소화할 수 있고 분할과 인식 결과가 동시에 생성됨으로써 서로간의 정확도를 높일 수 잇다. 또한 인식 결과를 확률로 변환함으로써 후처리의 정확도를 높였다. 추후에 인식 확률을 이용하여 인식기의 결합을 수행한다면 더욱 높은 인식 성능을 생성할 수 있을 것으로 기대된다.

참고 문헌

[1] 장만대, 김진형, "한국학 고문서 전산화 작업을 위한 한자 인식 및 그 결과의 기각 방법", 제 16회 영상처리 및 이해에 관한 워크샵 발표 논문집, pp.245-250. 2004

[2] 조규태, 김진형, "문자의 외형적 특징과 문맥정보를 이용한 고문서 분할-인식 통합기법", 제 16회 영상처리 및 이해에 관한 워크샵 발표 논문집, pp.239-244. 2004

[3] Y.H. Tseng, H.J. Lee, "Recognition-based handwritten Chinese character segmentation using a probabilistic Viterbi algorithm", Pattern Recognition Letters, Vol.20, pp.791-806, 1999.

[4] Bianca Zadrozny, Charles Elkan, "Transforming Classifer Scores into Accurate Multiclass Probability Estimates", SigKDD, 2002

[5] D. Jurafsky, J.H. Martin, "Speech and Language Processing: An Introduction to Natural Language Processing", Computational Linguistics and Speech Recognition, Prentice-Hall, 2000

[6] Yi-bong Tseng et al, "Speeding up Chinese character recognition in an automatic document reading system", Pattern Recognition, Vol. 31, pp. 1601-1612, 1998

[7] 이상호, 김호연, 임길택, 남윤석, "방향각 특징기반의 필기 숫자 인식", 한국정보처리학회, 추계학술발표대회 논문집 제9권 제2호, 2002

[8] Jun Tsukumo, "Classification of Handprinted Chinese Characters Using Non-linear Normalization and Correlation methods", 1988,