필기 한자 인식을 위한 개선된 윤곽선 방향 특징

곽희규° 김승태 류성호 김진형
한국과학기술원 전산학과
{hkkwag°, stkim, shryu, jkim}@ai.kaist.ac.kr

Improved Contour Directional Feature for Handwritten Hanja Recognition

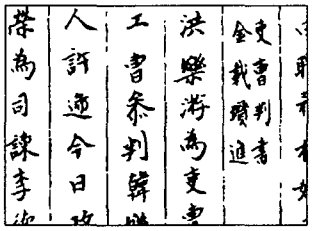
Hee Kue Kwag° Seung Tai Kim Sung Ho Ryu Jin Hyung Kim
Dept. of Computer Science, KAIST

요 약

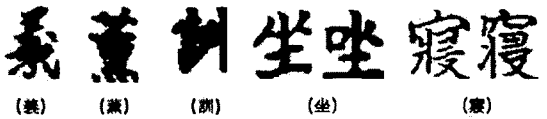
본 논문은 필기한자 인식의 성능 향상을 위한 개선된 윤곽선 방향 특징 추출에 대한 연구이다. 제안한 특징 추출은 기존의 방법에서 나타나는 계단현상을 완화함으로써 한자의 기본 요소인 획의 방향을 표현하는 통계적 성질을 두드러지게 하였다. 한국학 고문헌 상에 나타나는 필기 한자들을 대상으로 한 실험에서, 제안한 특징의 변별력이 뛰어나고, 오인식률이 감소하였음을 보였다.

1. 서 론

필기 한자 인식에 관한 연구는 중국 및 일본 등에서 주도적으로 연구되고 있지만 아직도 매우 어려운 패턴인식(pattern recognition) 문제로 여겨지고 있다[1-7]. 이것은 한자가 매우 방대한 양의 어휘를 포함하고 있을 뿐만 아니라, 매우 복잡한 획(stroke) 구조, 다양한 필기 변이, 그리고 다른 클래스간의 상호 유사성 등에서 기인한다. 국내에서는 지금까지 수작업에 의존하던 한국학 고문헌들의 전산화에 필기 한자 인식 기술을 접목하려는 시도가 일고 있다. 그러나 이런 고문헌들의 고전적인 한자 인식 문제는 기존의 어려움 외에도, 노후한 원본에서 나타나는 한자의 훼손과 이체자(異體字) 문제 등을 안고 있다. (그림 1)



(a) 품질 향상된 고문헌 영상



(b) 획이 훼손된 한자 (c) 이체자의 예

그림 1. 고문헌 영상에 존재하는 한자

필기 한자 인식의 우수한 성능 향상을 위해서는 다양한 필기상의 형태 변이를 흡수할 수 있는 적절한 특징의 설계가 필수적이다. 기존 연구들에서는 획 개수(stroke count), 윤곽화소 개수(contour-pixel count), 윤곽선 방향(contour direction), 배경 면적(background area), 교화 횟수(crossing count), 투영(projection) 등의 통계적인 특징들을 사용하였다[2]. 특히, 윤곽선 방향 특징(CDF: Contour Directional Feature)은 획을 기본 요소로 하는 한자의 경우에 성능이 매우 뛰어나기 때문에 많은 문헌들에서 언급되어지고 있다[2-7]. 그러나 0과 1로 표현되는 디지털 평면에서는 윤곽선 상에 나타나는 계단현상(staircases) 때문에 CDF의 통계적 성질(statistical property)이 두드러지게 부각되지 못한 단점을 가지고 있다. 몇몇 연구들에서는 이러한 문제를 해결하기 위해 평활화(smoothing) 연산을 사용하여 윤곽선을 부드럽게 변환하거나[3], 한 화소에 두 개의 방향을 할당하는[4] 등의 조치를 취하고 있다.

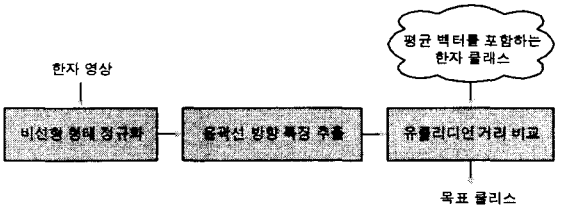
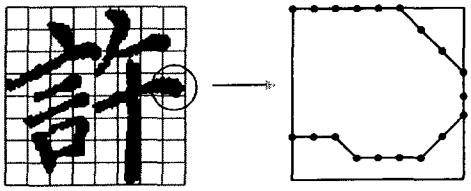


그림 2. 필기 한자 인식시스템의 구성도

본 논문에서는 필기 한자 인식의 성능 향상을 위한 개선된 CDF 추출에 대해 기술한다. 제안하는 CDF 추출에서는 디지털 평면상의 계단현상을 제거하는 앤티에일리어싱(anti-aliasing) 기법을 적용하여 아날로그 평면을 구성한 후, 특징을 추출함으로써 CDF의 통계적 성질을 유지할 수 있다. 이것은 윤곽선 상의 각 화소의 방향이 실제 획의 방향과 거의 일치한다는 것을 의미한다. 한국학 고문헌의 필기 한자들을 대상으로 기존 CDF의 문제점을 기술하고 개선된 특징의 성능을 비교하였다. 또한, 개선된 CDF의 효용성을 입증하기 위해 다양한 형태 정규화(shape normalization) 방법들을 적용하여 실험하였다. 본 연구의 필기 한자 인식의 구성은 그림 2와 같다.

2. 윤곽선 방향 특징의 설계

한자는 주로 수평, 수직, 사선, 역사선 방향의 획들로 구성된다. 따라서 윤곽선 상의 화소는 네 가지 방향성을 갖는다고 가정할 수 있다[2-7]. 따라서 CDF의 추출은 먼저 이웃하는 윤곽선 화소들로 구성되는 선 성분의 방향을 네 방향 중 하나로 분류하고, 미리 분할된 블록(block)에서 각 방향에 대한 화소의 개수를 계산하여 할당한다. 본 논문에서는 입력 한자를 64×64 크기의 영상으로 정규화 한 후, 8×8 화소 영역을 하나의 블록으로 하여 64(8×8) 블록을 형성한다. 각 블록에서 네 방향의 특징값이 존재하므로, CDF는 256(8×8×4)차원 공간을 갖게 된다(그림 3).



(a) (b)

(a) 정규화 영상의 64블록
(b) 블록내의 윤곽 화소
(c) 네 방향에 대한 윤곽화소의 개수

수평	10
사선	2
수직	2
역사선	4

(c)

그림 3. 정규화 영상의 윤곽선과 방향성

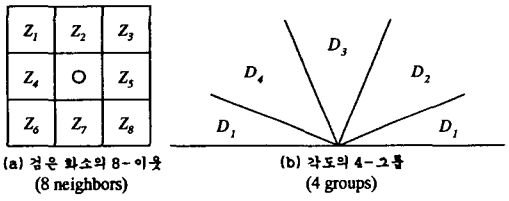
윤곽선 화소의 방향은 이웃하는 8개 화소의 정보를 이용하여 계산하는데, 그림 4(a)와 같은 3×3 윈도우를 정의한다. $\alpha(x,y)$가 화소 (x,y)에서 획의 방향을 나타내는 각도를 표현한다고 하면, 다음과 같이 정의할 수 있다.

$$\alpha(x,y) = \tan^{-1}(G_x/G_y)$$

이 때, G_x와 G_y는 각각 x축과 y축에서 정의되는 도함수(derivative)라고 한다. 이 도함수들은 소벨 연산자(Sobel operator)에 기초하여 다음과 같이 계산된다.

$$G_x = (z_6 + 2z_7 + z_8) - (z_1 + 2z_2 + z_3)$$
$$G_y = (z_3 + 2z_5 + z_8) - (z_1 + 2z_4 + z_6)$$

한자는 주로 네 방향의 획으로 구성되기 때문에 각도의 범위를 그림 4(b)에서와 같이 네 그룹으로 나눌 수 있고, 화소의 각도는 네 그룹 중 하나로 분류된다[2-5].



(a) 중심 화소의 8-이웃 (8 neighbors) (b) 각도의 4-그룹 (4 groups)

그림 4. 3×3 윈도우와 네 각도 그룹

그런데, 0과 1로 표현되는 디지털 평면에서 계산된 화소의 방향들은 실제 획의 방향과 일치하지 않는 경우가 다수 발생하였다. 이것은 지역적(local) 화소 정보의 계단현상에서 기인한 것으로, 획의 방향에 대한 통계적 성질을 표현해야 하는 CDF의 역할에 적합하지 못하다. 그림5와 같은 한자에서 볼 수 있듯이, 실제 획의 방향과 일치하지 않는 수평과 수직 방향 성분이 다수 검출되었다. 직관적으로 획의 사선과 역사선 방향 성분이 두드러지게 보이지만, 각 방향에 대한 전체 성분 비율은 각각 수평 22%, 수직 28%, 사선 22%, 역사선 28%로 나타났다.

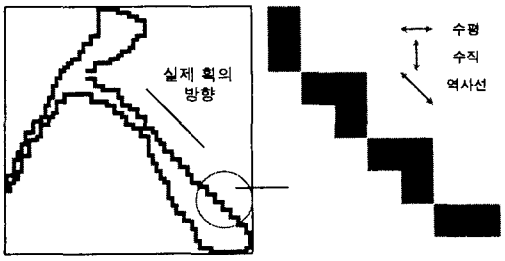


그림 5. 디지털 평면에서 윤곽선 방향

본 논문에서는 디지털 평면의 계단현상을 완화할 수 있는 앤티에일리어싱 기법을 적용하였다. 그림 6과 같이 화소의 값을 이웃하는 화소들의 값으로부터 가중치 평균(weighted average)을 취하여 아날로그 평면을 구성한다.

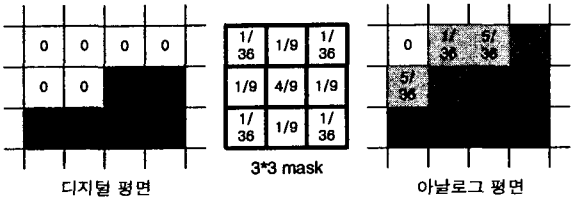


그림 6. 3×3 마스크에 의한 아날로그 평면

따라서 윤곽선 화소의 방향을 구하기 위한 소벨 연산자는 새롭게 생성된 아날로그 평면에 적용된다. 그림 7의 결과와 같이, 아날로그 평면의 계단현상이 완화되면서 윤곽선 화소의 방향이 실제 획의 방향과 거의 일치하였다. 또한, 각 방향에 대한 화소의 성분 비율도 수평 11%, 수직 15%, 사선 31%, 역사선 43%로 나타남으로써, 제안한 CDF가 통계적인 성질을 잘 표현하고 있다고 할 수 있다.

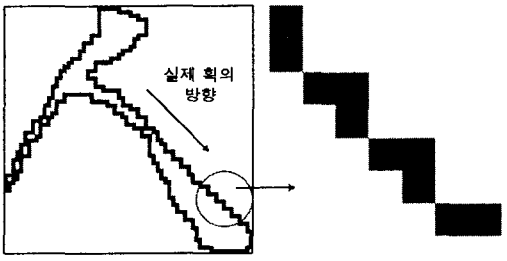


그림 7. 아날로그 평면에서 윤곽선 방향

3. 실험 결과 및 분석

제안한 CDF의 효용성을 입증하기 위한 실험은, 두 가지 관점에서 진행하였다. 먼저, 입력 한자의 특징벡터와 한자 클래스 대표벡터와의 유클리디언(Euclidean) 거리 비교에서 유사도 1순위와 2순위간의 거리 차이의 변화를 살펴보았다. 다음으로, 고문헌 상에서 자주 출현하는 한자 1,189클래스(훈련: 100, 테스트: 100)를 대상으로 인

식 성능을 측정하여 오인식률의 변화가 있는지를 살펴보았다. 이때, 자주 사용되는 세 가지 비선형 형태 정규화 방법들[8] 상에서 비교하였다(그림 8).

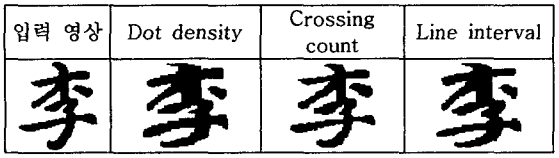


그림 8. 형태 정규화 방법들의 결과 영상

필기 한자 인식의 어려움은 유사한 형태를 가진 한자가 많다는 것에서도 기인한다. 따라서 적합한 특징은 다른 클래스간의 작은 변이도 두드러지게 표현할 수 있어야 한다. 이것은 입력 한자의 특징벡터 비교시 1순위와 2순위간의 유사도 거리의 차이가 크면, 그만큼 변별력이 높다고 할 수 있다. 다음 그림 9의 두 한자는 유사도 순위가 자주 바뀔 정도로 매우 유사하게 나타났다. 그런데, 기존 CDF에 비해 개선된 CDF의 1순위와 2순위간 거리 차이가 더 크게 나타났다. 또한, 실제 유사도의 순서가 바뀌어서 바람직한 결과를 낳았다.

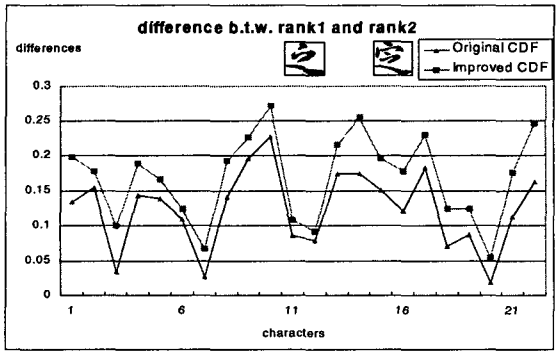


그림 9. 1순위와 2순위간 유사도거리 차이 비교

한편, 1,189클래스에 대한 테스트 한자들을 대상으로 실험한 결과, 형태 정규화 방법에 따라 90~92% 정도의 인식 성능을 나타내었다. 실제 오인식의 비율을 조사하였는데, 형태 정규화 방법들간 차이는 다소 있지만 기존 CDF에 대해 제안한 특징의 오인식률이 약 10% 정도 감소하였음을 알 수 있었다(그림 10).

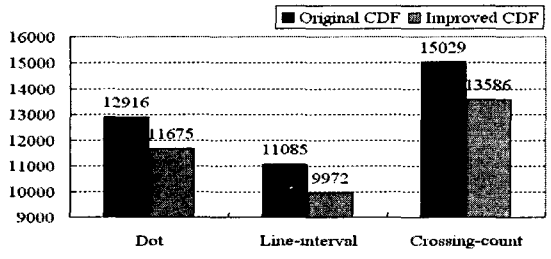


그림 10. 기존 CDF와 개선된 CDF의 오인식률 비교

4. 결 론

본 논문에서는 필기 한자 인식의 성능 향상을 위한 개선된 CDF의 추출에 대한 기술하고, 고문헌 한자 데이터를 대상으로 성능을 평가함으로써 효용성을 입증하였다. 제안한 CDF는 디지털 평면의 계단현상을 완화하는 앤티에일리어싱 기법을 적용하여 아날로그 평면으로 구성한 후, 윤곽선 화소의 방향을 구하였다. 이것은 윤곽선 화소의 방향 성분이 실제 획과 거의 일치하는 효과를 가짐으로써 통계적인 성질을 두드러지게 표현할 수 있는 효과를 낼 수 있었다.

참고문헌

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

[2] Y.H. Tseng, C.C. Kuo and H.J. Lee, "Speeding up Chinese character recognition in an automatic document reading system," *Pattern Recognition*, Vol. 31, No. 11, pp. 1601-1612, 1998.

[3] C.L. Liu, I.J. Kim and J.H. Kim, "High accuracy handwritten Chinese character recognition by improved feature matching method," *Int. Conf. on Document Analysis and Recognition*, Ulm, pp. 1033-1037, 1997.

[4] N. Kato, S. Omachi, H. Aso and Y. Nemoto, "A handwritten character recognition system using directional element feature and asymmetric Mahalanobis distance," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 21, No. 3, pp. 258-262, 1999.

[5] C.H. Tung, H.J. Lee and J.Y. Tsai, "Multi-stage pre-candidate selection in handwritten Chinese character recognition systems," *Pattern Recognition*, Vol. 27, No. 8, pp. 1093-1102, 1994.

[6] F. Kimura, T. Wakabayashi, S. Tsuruoka and Y. Miyake, "Improvement of handwritten Japanese character recognition using weighted direction code histogram," *Pattern Recognition*, Vol. 30, No. 8, pp. 1329-1337, 1997.

[7] Z. Lixin and D. Ruwei, "Off-line handwritten Chinese character recognition with nonlinear pre-classification," *Int. Conf. on Multimodal Interfaces*, pp. 473-479, 2000.

[8] 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.