

유사 문자쌍 구분을 위한 지역적 공헌도 기반 비선형 정규화

류상진*, 김인중**

한동대학교 전산전자공학부

e-mail : *pardon6@naver.com, **ijkim@handong.edu

Non-linear Normalization for Pair-wise Discrimination Based On Local Contribution Measure

Sang-Jun Ryu, In-Jung Kim

School of Computer Science and Electronic Engineering, Handong Global University

요 약

지금까지 필기 변이를 완화하기 위한 다양한 비선형 정규화 방법들이 제안되었으며 실제 인식 시스템에서 상당한 인식률 개선 효과를 나타내었다. 그러나, 필기 한글 인식에 있어서는 필기 변이 외에도 문자간의 높은 유사도로 인해 높은 인식률을 얻는데 어려움을 겪고 있다. 한글과 같이 문자간 유사도가 높은 언어를 효과적으로 인식하기 위해서는 필기 변이를 흡수하는 것뿐 아니라, 유사 문자간의 차이를 정확히 찾아내어 그 차이점을 부각시키는 것이 요구된다. 본 논문에서는 유사 문자간의 차이점을 부각시킬 수 있는 비선형 정규화 방법을 제안한다. 기존의 비선형 정규화 방법들이 영상의 지역적 복잡도를 균일화 함으로써 정규화를 수행했던 것에 반해, 제안하는 방법에서는 유사 문자쌍의 구분에 있어 지역적 공헌도에 기반하여 영상을 정규화한다. 즉, 유사 문자쌍 구분에 공헌도가 높은 지역은 확대하고 그렇지 않은 지역은 축소한다. 그 결과, 문자간에 서로 상이한 지역을 강조 함으로써 유사 문자쌍에 대한 구분력을 높인다. 실험 결과, 제안하는 방법으로 정규화된 영상에서는 유사 문자쌍의 차이점이 확대되었으며, 문자쌍의 구분 성능 또한 향상되었다.

1 서론

지난 수 십 년 동안 필기 문자 인식에 관한 연구는 많은 발전을 이루었다. 그 결과, 영어와 중국어의 경우 인식 기술이 이미 상당한 수준에 이르렀다. 그러나, 필기 한글의 경우 높은 문자간 유사도와 필기 변이로 인해 높은 성능을 얻지 못하고 있다.

문자 인식 시스템은 일반적으로 전처리, 특징 추출, 인식, 등 세 단계로 구성된다. 필기 문자 인식의 전처리 단계에서는 필기 변이를 효과적으로 흡수할 수 있는 비선형 정규화가 많이 사용된다. 현재까지의 비선형 정규화 알고리즘들은 주로 필기변이를 완화시키는 것에 중점을 두었다. 그러나, 한글과 같이 유사 문자가 많은 언어를 인식하기 위해서는 유사 문자쌍들을 잘 구분하는 것이 매우 중요하다. 유사 문자쌍이란 ‘홍’-‘홍’, ‘데’-‘테’, ‘품’-‘푹’ 등과 같이 형태가 매우 유사하여 약간의 지역적 차이에 의해서만 구분되는 문자들의 쌍이다. 이들을 잘 구분하기 위해서는 유사 문자간 차이점에 집중해야 한다.

본 논문에서는 유사 문자쌍의 구분을 위한 지역적 공헌도를 이용하여 두 문자의 차이점을 부각시키는 비선형 정규화 방법을 제안한다. 제안하는 방법은 유사 문자간에 상이한 지역은 확대하고 유사한 지역을

축소함으로써 두 문자간 차이점을 강조한다. 그 결과 유사 문자간의 차이를 정확히 검출하게 함으로써 구분 성능을 향상시킨다.

본 연구의 2 장에서는 기존에 제안되었던 비선형 정규화 방법과 유사 문자쌍 구분에 관한 연구를 소개한다. 3 장에서는 유사 문자쌍 구분을 위한 지역적 공헌도 기반 비선형 정규화 방법을 제안하고, 4 장에서는 실험 결과를 제시한다. 마지막으로 5 장에서는 결론을 맺는다.

2 관련 연구

2.1 비선형 정규화

비선형 정규화 알고리즘은 주로 필기 문자의 변이를 완화하기 위한 목적으로 개발되었다. Yamashita 는 문자 영상의 화소밀도 히스토그램이 균일하게 분포되도록 영상을 리샘플링함으로써, 화소밀도가 높은 부분은 확대하고 그렇지 않은 부분은 축소하는 비선형 정규화 방법을 개발하였다[1]. 그 후, Yamada 는 화소밀도 대신 선밀도를 균일화 하는 정규화 방법을 제안하였다[2]. Yamashita 의 방법에서는 긴 획들이 굵게 확대되는 문제점이 있었으나, Yamada 의 방법에서는 이러한 단점이 개선되었다. Tsukumo 도 역시 선밀도를

이용한 정규화 방법을 개발하였다[3]. 그러나, Yamada 가 선 교차점(line crossing)의 수를 이용하여 선밀도를 측정 한 것과는 달리 Tsukumo 는 선들 사이의 간격(line interval)을 이용하여 선밀도를 측정하였다. Lee 는 다양한 정규화 방법들을 비교 평가하였다[1]. 그 결과, Tsukumo 의 방법은 Yamada 의 방법에 비해 속도는 느리나, 성능은 우수한 것으로 나타났다. Liu 는 기존의 방법들이 히스토그램 균일화를 통한 1 차원적 알고리즘을 사용했던 것에 비해 2 차원적인 균일화 알고리즘을 이용하여 정규화를 수행하였다[5]. 특히, Liu 는 선밀도 투영보간(line density projection interpolation: LDPI) 방식을 이용하여 속도를 많이 저하시키지 않으면서도 기존 1 차원 정규화 방법보다 필기변이를 더욱 효과적으로 흡수할 수 있는 방법을 제안하였다. Liu 의 방법은 실제 인식 성능 향상에 우수한 성능을 나타내어 필기 인식에 널리 사용된다.

2.2 유사 문자쌍 구분

유사 문자가 많이 포함된 언어를 인식하는 것은 쉽지 않은 일이다. 특히, 일반적인 인식기만을 이용하여 형태가 매우 유사한 문자들을 구분할 경우 충분한 성능을 얻기 어렵다. 이러한 한계를 극복하기 위하여 유사 문자쌍들을 전문적으로 구분해 주는 유사 문자쌍 구분 방법들이 제안되었다. 유사 문자쌍 구분기는 주로 인식 시스템의 후처리 단계에서 사용되어 인식기의 결과를 검증함으로써 전체 시스템의 인식 성능을 향상시킨다. 이들은 주로 문자간 차이를 중심으로 분석을 수행함으로써 유사 문자들을 효과적으로 구분한다.

Kim 은 구조적 인식 시스템을 위한 유사 문자쌍 구분 방법을 제안하였다[6]. 입력 문자들의 획들과 문자 모델들의 획들을 매칭시킨 후 획별 매칭 스코어를 평균하여 각 문자 클래스의 스코어를 계산한다. 이 때, 각 획들의 공헌도에 따라 가중치를 지정함으로써 중요한 획들에 집중할 수 있도록 하였다. Jang 도 문자간 차이점에 중점을 두어 유사 문자쌍들을 구분하였다[7]. Jang 의 방법은 문자 영상에서 중요한 문자획에 해당하는 영역으로부터 특징을 추출하여 통계적인 방법을 통해 유사 문자쌍을 구분한다. 두 방법 모두 유사 문자쌍의 구분에 상당한 성능 향상을 얻었다.

3 유사 문자쌍 구분을 위한 지역적 공헌도 기반 비선형 정규화

본 논문에서 제안하는 정규화 방법은 그림 1 과 같이 학습 단계와 정규화 단계로 이루어진다. 학습 단계에서는 각 유사 문자쌍의 구분에 있어서 문자 영상의 각 부분의 공헌도를 정량화하여 저장한다. 먼저 기존의 방법을 이용해 학습 영상을 사전 정규화(pre-normalization)함으로써 인식에 불필요한 필기 변이를 최소화한다. 그 후, 유사 문자쌍 부분에서 각 지역의 공헌도를 정량화하여 저장한다. 정규화 과정에서는 역시 사전 정규화를 통해 입력 영상으로부터 필기 변이를 최소화 한 후, 지역적 공헌도를 반영하여 다시

리샘플링한다.

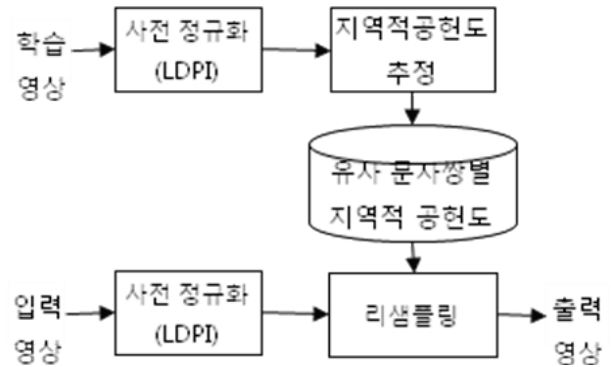


그림 1. 지역적 공헌도 기반 비선형 정규화

유사 문자쌍 (c_1, c_2)가 주어졌을 때 학습 단계에서는 먼저 학습 영상들에 대하여 사전 정규화를 수행함으로써 필기 변이를 최소화한다. 본 연구에서는 성능이 우수하다고 알려진 LDPI 알고리즘을 사용하였다[5].

그 후, 사전 정규화 된 학습 데이터들로부터 특징 벡터를 추출한다. 특징으로는 문자 인식에서 널리 사용되는 그물망 기반 외각선 기울기 특징을 사용하였다[8]. 문자 영상으로부터 외각선 화소들을 기울기 별로 D 방향(D 는 4 또는 8)으로 분류한 후, $M \times N$ 그물망을 이용하여 각 영역에서 추출한 특징들을 누적함으로써 특징 벡터를 구성한다. 이와 같이 추출한 특징 벡터는 DMN 차원의 벡터가 된다.

인식에 있어서 각 특징 차원의 공헌도는 두 클래스의 평균간 거리에 비례하고, 표준편차에 반비례한다[9]. 따라서, i 번째 특징 차원의 공헌도는 식 (1)과 같이 정의된다.

$$r_i^2 = \left(\frac{\mu_{1i} - \mu_{2i}}{\sigma_i} \right)^2 \quad (1)$$

여기에서, μ_{1i} 와 μ_{2i} 는 각각 클래스 c_1 과 c_2 에 속한 샘플들로부터 추출한 i 번째 차원의 평균들이며, σ_i 는 모든 샘플들에 대한 i 번째 차원의 표준편차이다.

그런데, 위에서 기술한 방식으로 추출한 특징 벡터에서는 각 그물망 영역마다 D 가지 방향의 특징 차원이 대응하게 된다. 본 연구에서는 각 그물망 영역의 공헌도를 그에 대응하는 D 개 차원의 공헌도를 누적한 값으로 정의하였다. 특징 벡터 중 j 번째 그물망 영역에 대응하는 차원들은 $(j+dMN)$ 번째 특징 차원들이다($0 \leq d < D$). 따라서, j 번째 ($0 \leq j < MN$) 그물망 영역의 공헌도는 식 (2)와 정의할 수 있다.

$$R_j^2 = \sum_{d=0}^{D-1} \left(\frac{\mu_{1(j+dMN)} - \mu_{2(j+dMN)}}{\sigma_{(j+dMN)}} \right)^2 \quad (2)$$

D 는 각 지역에 대응하는 특징 차원의 개수를 나타내며 $\mu_{1(j+dMN)}$ 과 $\mu_{2(j+dMN)}$ 는 각각 클래스 c_1 과 c_2 의 학습

영상에 대한 $(j+dMN)$ 번째 차원의 평균을 의미한다. $\sigma_{(j+dMN)}$ 는 모든 샘플들에 대한 $(j+dMN)$ 번째 특징 차원의 표준편차이다.

그런데, 이와 같은 각 그물망 영역의 공헌도를 그대로 사용할 경우 그물망 영역간의 경계 부분에서 지나치게 급격한 변화가 나타난다. 이를 완화하기 위하여 각 그물망 별 공헌도에 가우시안 보간법을 이용해 각 영역에 대한 지역적 공헌도를 계산한다. 이와 같이 계산된 지역적 공헌도의 예는 표 1의 가운데 열과 같다. 이와 같이 추출된 지역적 공헌도는 정규화 과정에 사용된다.

정규화 과정은 크게 두 단계로 구성된다. 먼저, 기존 비선형 정규화 방법을 이용하여 필기 변이를 최소화 한다. 그 후, 학습단계에서 추출한 지역적 공헌도가 균일화되도록 입력 영상을 리샘플링하여 출력 영상을 생성한다. 지역적 공헌도를 균일화 하기 때문에 공헌도가 높은 지역은 확대되고, 그렇지 않은 지역은 축소된다. 리샘플링을 위해서는 Yamashita 가 사용한 히스토그램 평활화 방법을 사용하였다[1]. 상세한 내용을 [1]을 참조하기 바란다.

표 1은 제안하는 방법으로 정규화를 수행한 몇가지 예를 보여준다. 좌측 두 열은 사전 정규화에 의해 필기 변이를 최소화 한 유사 문자쌍의 영상의 예이다. 가운데 열은 학습단계에서 추출한 영역별 공헌도이다. 우측 두 열은 각각 좌측 두 영상들에 대하여 지역적 공헌도를 균일화 함으로써 수행한 비선형 정규화 결과이다. 공헌도가 높은 영역이 확대되고 공헌도가 낮은 영역은 축소됨으로써 두 문자의 차이가 강조되었음을 알 수 있다.

표 1. 유사 문자쌍 별 지역적 공헌도 및 이를 균일화 한 비선형 정규화 예

사전 정규화 결과		지역적 공헌도	지역적 공헌도 기반 비선형 정규화	
항	향		항	향
거	저		거	저
홍	흥		홍	흥

4 실험

실험을 통하여 제안하는 유사 문자쌍 구분을 위한 지역적 공헌도 기반 비선형 정규화 방법의 성능을 측정하였다. 실험에는 SERI95a 문자 영상 데이터베이스를 사용하였다[10]. SERI95a 데이터베이스는 자주 사용되는 한글 520 개 클래스로 구성된다.

실험에 사용한 문자 인식기는 최소 거리 구분기

(minimum distance classifier)를 이용하여 구현하였다[8]. 정규화된 영상으로부터 외곽선 기울기 특징을 추출한 후 입력 영상의 특징 벡터로부터 유클리드 거리가 가장 가까운 문자 모델의 클래스를 인식 결과로 출력한다. 성능 향상을 위해 LVQ(Linear Vector Quantization)를 이용하여 구분력 강화 학습을 수행하였다[8].

유사 문자쌍 구분기도 동일한 방법으로 구성하였다. 다만, 인식기가 520 개 클래스를 인식 대상으로 하는 반면 유사 문자쌍 구분기는 구분 대상이 되는 두 개의 클래스만을 인식 대상으로 한다. 유사 문자쌍을 적용할 클래스는 인식기의 혼동표(confusion table)로부터 추출하였다. 각 클래스별 혼동 빈도를 측정한 후 혼동이 가장 많이 발생하는 클래스 쌍 10 개를 선택하여 이들에 대하여 유사 문자쌍 구분기를 적용하였다. 학습 데이터와 테스트 데이터로는 SERI95a 데이터를 각각 2:1의 비율로 나누어 사용하였으며, 실험은 Q6600 2.4GHz CPU와 4GB 메모리를 가진 PC에서 수행하였다.

4.1 유사 문자쌍의 평균간 거리

먼저, 제안하는 방법으로 정규화 하였을 때 특징 공간에서의 유사 문자 클래스간 거리가 어떻게 변하는지를 측정하였다. 필기 변이 흡수에 우수하다고 알려진 LDPI 제안하는 방법으로 유사 문자쌍의 데이터를 정규화 한 후 특징 벡터를 추출하여 두 클래스의 평균 간 유클리드 거리를 측정하였다. 그 결과는 표 2와 같다. 제안하는 방법을 적용하였을 때 유사 문자 클래스의 평균간 거리가 증가했음을 알 수 있다. 이는 제안하는 방법이 문자간 구분력 향상에 효과가 있음을 보여준다.

표 2. 클래스 평균간 거리 비교 결과

유사문자쌍	LDPI[5]	제안하는 방법	변화량
레 레	47094	57279	10185
술 술	82652	87381	4729
허 허	48149	59063	10914
홍 홍	48292	49145	853
항 향	79282	115105	35823
념 념	64147	93488	29341
런 런	33136	40653	7517
품 품	60268	68505	8237
고 교	104518	119936	15418
너 녀	89409	106358	16949
평균	89409	106358	16949

4.2 유사 문자쌍 구분 성능

또한, LDPI와 제안하는 방법 유사문자구분 방법을 유사 문자쌍 구분기에 실제로 적용하여 그 성능을 측정하였다. 그 결과는 표 3과 같다.

표 3. 유사 문자쌍 구분 성능

유사 문자쌍	LDPI[5]	제안하는 방법	변화량
레 레	77.00	84.00	7.00
술 술	88.00	87.50	-0.50
허 허	83.00	86.00	3.00
흥 흥	77.78	76.77	-1.01
향 향	80.90	84.42	3.52
넘 넘	82.00	80.00	-2.00
런 런	74.37	74.87	0.50
품 품	85.00	85.50	0.50
고 교	87.00	87.50	0.50
녀 녀	85.00	89.50	4.50
평균	82.01	83.61	1.60

520 개 클래스에서 혼동 확률이 가장 높은 10 개의 클래스를 선택했기 때문에 이들간의 평균 구분을 역시 높지 않게 나타났다. 그런데, 제안하는 정규화 방법을 유사 문자쌍 구분기에 적용하였을 때 LDPI 에 비하여 구분률이 더 높게 나타났다. 이는 제안하는 방법이 두 문자간 차이점을 강조함으로써 유사 문자간 구분력을 개선하는데 효과가 있음을 보여준다.

5 결론

필기한글에서 유사 문자들을 구분하는 것은 매우 중요하다. 유사 문자쌍 구분기를 사용하는 것이 바람직하다. 본 논문에서는 유사 문자쌍의 구분을 위한 비선형 정규화 방법을 제안하였다. 기존의 비선형 정규화 방법이 문자 영상의 복잡도를 균일화 함으로써 영상을 정규화 한 것에 반해 제안하는 방법에서는 문자쌍의 구분에 있어서 지역적 공헌도를 반영하여 영상을 정규화 한다. 즉, 두 문자간에 상이한 부분은 확대하고, 유사한 부분은 축소함으로써 문자간 차이점을 강조한다. 실험 결과 제안하는 방법은 기존 방법에 비해 문자간 거리를 부각시키는 효과가 있었으며, 실제로 유사 문자쌍 구분기에 적용하였을 때 구분 성능도 향상되었다.

사사 (Acknowledgement)

본 연구는 교육과학기술부와 한국산업기술재단의 지역혁신인력양성사업으로 수행된 연구결과임.

참고문헌

- [1] Y. Yamashita, K. Higuchi, Y. Yamada and Y. Haga, "Classification of handprinted Kanji characters by the structured segment matching method," Pattern Recognition Lett. 1, 475-479 (1983)
- [2] H. Yamada, T. Saito and K. Yamamoto, "Line density equalization--a nonlinear normalization for correlation method," Trans. IECE Japan J67-D(11), 1379-1383 (in Japanese) (1984).
- [3] J. Tsukumo and H. Tanaka, "Classification of

handprinted Chinese characters using nonlinear normalization methods," Proc. 9th Int. Conf. on Pattern Recognition, Rome, Italy, pp. 168-171, November (1988).

- [4] 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, 1994, pp. 895-902.
- [5] Cheng-Lin Liu, Katsumi Marukawa, "Pseudo two-dimensional shape normalization methods for handwritten Chinese character recognition," Pattern recognition, vol. 38, no. 12, 2005.
- [6] I. J. Kim and J. Kim, "Pairwise Discrimination Based on a Stroke Importance Measure," Pattern Recognition, vol.35, no.10, pp 2259-2266, October 2002.
- [7] Seung-Ick Jang, "Postprocessing of Handwritten Hangul Recognition using Pair-wise Grapeme Discrimination," MS Thesis, KAIST, 2002.
- [8] C. L. Liu, I. J. Kim, and J. H. Kim, "High Accuracy Handwritten Chinese Character Recognition by Improved Feature Matching Method," Proc. 4th ICDAR, Ulm, Germany, pp.1033-1037, 1997.
- [9] R.O. Duda, P.E. Hart and D.G. Stork, "Pattern Classification 2nd ed.," pp.110, Wiley-Interscience, 2001.
- [10] KAIST 인공지능 연구실 홈페이지 (<http://ai.kaist.ac.kr/Resource/dbase/Image%20Database.htm#HangulCharacter>)
- [11] R.O. Duda, P.E. Hart and D.G. Stork, "Pattern Classification 2nd ed.," pp.36-45, Wiley-Interscience, 2001.