최적선형변환에 의한 유사문자의 상세분류인식

김형원° 김성원 양윤모
고려대학교 전자정보공학과
{hwkim, swkim, ymyang}@hard.korea.ac.kr

Detailed Recognition of Similar Characters Based on Optimum Linear Transform

Hyung-Won Kim° Sung-Won Kim Yun-Mo Yang
Dept. of Electronics & Information Engineering, Korea University

요 약

본 논문에서는 문자 인식에서 두 단계의 식별과정을 통하여 인식률을 향상시키는 방법에 대하여 연구하였다. 한글 문자인식에서의 어려움은 인식대상 클래스가 많고 유사문자가 많은 반면, 여러 폰트의 글자를 하나의 클래스로 할 경우는 그 문자의 분산이 더욱 커지게 되는 점이다. 따라서 본 연구에서는 문자의 분포를 고려하여 거리를 계산하는 Bayes에 의한 식별함수를 1단계 인식과정에서 사용하여 1위 후보문자를 인식하였다. 2단계에서는 미리 준비된 1위 후보문자의 유사문자세트의 최적선형변환 공간에서 상세분류를 행하였다. 결과적으로 1단계의 Bayes거리만에 의한 인식률(91.1%)보다도, 또한 처음부터 모든 클래스에 대하여 최적선형변환에 의한 인식률(87.9%)보다 좋은 결과(92.9%)를 얻게되었다. 이로서 1단계의 대규모 문자세트에 대한 대분류에서는 문자의 분포를 고려하는 Bayes에 의한 인식이 유효하고, 2단계의 최적선형변환에 의한 인식은 소수의 유사문자들에 대한 변별력을 높이는데 유효함을 입증하였다.

1. 서론

한글 문자인식에서의 어려움은 인식대상 클래스가 많고 유사문자가 많은 반면, 여러 폰트의 글자를 하나의 클래스로 할 경우는 그 문자의 분산이 더욱 커지게 되는 점이다. 위와 같은 문제점을 피하기 위하여 제 1단계로서 문자들의 분포를 고려하는 Bayes 분류기에 의한 식별함수를 적용하여 인식대상문자와 가까운 유사한 문자들을 후보로 선택하고, 2단계로서 최적선형변환(Optimum Linear Transform: OLT)을 적용하여 유사한 문자들 중에서 식별을 최적화하는 새로운 특징량을 추출하여 상세분류를 행한다. 이와 같이 Bayes거리를 이용하여 후보문자들을 대분류하는 1단계(대분류)과정과, 최적선형변환을 적용하여 유사문자를 식별하는 2단계(상세분류)과정으로 나누어 적용함으로써 한글과 같은 대규모 문자세트에 대한 인식률을 향상시켰다. 전체 인식과정을 그림1에 나타내었다.

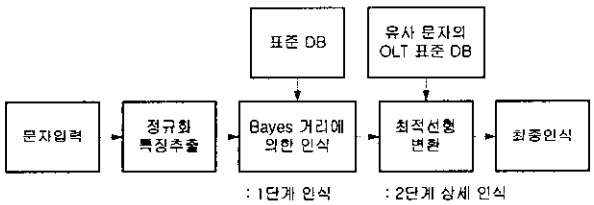

그림 1. 인식 과정

2. 전처리 및 특징추출

2.1 히스토그램을 이용한 선형 정규화

일반적인 정규화 알고리즘은 입력영상을 일정한 크기의 형상으로 선형 변형시키는 선형 정규화와 영상의 특징을 고려하는 비선형 정규화가 있다. 본 연구에서는 선형 정규화 방법을 수정, 보완하여 이용하였다. 이 때, 입력 영상이 정규화 영상의 크기보다 큰 경우, 중요한 정보를 갖는 단선성분이 제거되는 경우가 있어서 오인식의 원인이 되고있다. 이런 선형 정규화의 단점을 보완하기 위하여 문자영상의 히스토그램의 변화량을 이용하였다[4].

2.2 9진 트리 및 특징량 구성[1]

문자를 정규화한 후에 문자의 윤곽선에서 4가지 방향성분(0°, 45°, 90°, 135°)의 특징소를 필터를 사용하여 추출한다. 추출된 특징소를 이용하여 보다 안정성 있는 특징량을 구성하기 위해 계층적 9진 트리를 사용한다.

9진 트리는 그림 2와 같이 문자영상의 가로와 세로 각각 1/2 크기의 소영역 9개로 중첩하면서 분할하고, 그 소영역을 다시 9개로 반복 분할하여 영역을 계층적으로 9진 트리에 대응시킨다. 각 노드마다 4방향의 4차원 특징량을 할당하며 하위 노드의 특징량의 합으로 구한다. 입력문자는 히스토그램을 이용한 정규화 과정을 거쳐 32×32의 크기를 갖는다. 이러한 영상을 16×16크기의 작은 소영역 9개로 구분하고, 나뉘어진 각각의 소영역은 8×8크기의 9개의 소영역으로 나뉘어진다. 이렇게 반복하여 영역을 구분하면 결국 2×2크기의 소영역으로 나뉘어질 수 있다.

본 논문에서는 1세대 36차원 특징량을 사용하였다. 거리계산 과정에 각 문자의 공분산행렬을 이용하므로 324차원을 사용하면 계산량, 메모리, 학습데이터의 확보 등의 한계가 있기 때문에 36차원의 특징량을 사용한다.

3. Bayes Classifier에 의한 대분류

3.1 Bayes Classifier

1단계 인식과정에서 입력패턴과 표준패턴과의 거리계산방법은 Bayes 분류기에 의한 식별함수를 사용하였다[5]. k개 클래스의 m차원 특징벡터가 표준정규분포형태로 존재한다고 가정하고 입력패턴을 x라고 하면 x가 k클래스에 속할 확률은 다음과 같다.

$$g_k(x) = \Pr[w_k \mid x], \qquad k = 1, 2, \ldots, N_c \quad (1)$$

이것은 Bayes's rule에 의하여

$$g_k(x) = \Pr[x \mid w_k] \Pr[w_k], \qquad k = 1, 2, \ldots, N_c \quad (2)$$

이 되고, μ_k는 k클래스의 평균벡터, Σ_k를 k클래스의 공분산 행렬이라고 하면

$$g_k(x) = \frac{1}{(2\pi)^{m/2} |\Sigma_k|^{1/2}} \exp\left[-\frac{1}{2}(x-\mu_k)' \Sigma_k^{-1} (x-\mu_k)\right], \quad k = 1, 2, \ldots, N_c \quad (3)$$

여기에 -2 ln을 취하고 -m ln2π를 더하고 각 패턴의 사전확률 값을 모두 같게 하면

$$g_k = (x-\mu)' \Sigma_k^{-1} (x-\mu) + \ln|\Sigma_k| \quad (4)$$

와 같이 정리된 식을 식별함수로 사용한다.

3.2 Bayes Classifier에 의한 인식

대분류 단계에서는 넓은 분포를 갖는 다중폰트의 문자를 인식하기 위해서 Bayes 거리를 사용한다. 이 과정을 통하여 다양한 폰트나 잡음으로 인해 넓어진 문자의 분포에 적합한 거리계산을 할 수 있다. 이 거리를 이용하기 위해서는 식(4)에 나타난 것처럼 클래스의 평균과 공분산행렬을 필요로 한다. 입력되는 문자의 특징벡터, 각 표준특징벡터와 해당 공분산행렬을 이용하여 식(4)에 의하여 거리를 계산하게 된다. 이 값이 작을수록 해당 클래스일 확률이 크다고 할 수 있다. 제1단계에서는 가장 거리가 작은 1위 후보문자를 선택하고, 준비된 이 문자의 유사문자집합을 찾음으로써 대분류 과정이 이루어진다.

4. 최적선형변환에 의한 상세분류

4.1 최적선형변환(Optimum Linear Transform: OLT)

판별분석에 따르면, 클래스내의 분산, 클래스간의 분산행렬은 클래스의 식별능력의 척도로 사용된다[3]. 클래스내의 분산행렬(within-class scatter matrix)은

$$S_w = \sum_{i=1}^{L} P_i E[(X-M_i)(X-M_i)^T \mid \omega_i] = \sum_{i=1}^{L} P_i \Sigma_i \quad (5)$$

로 정의된다. 모든 클래스의 샘플들에 대한 전체 평균벡터를 M_0라 하면, 클래스간의 분산행렬(between-class scatter matrix)은

$$S_b = \sum_{i=1}^{L} P_i (M_i - M_0)(M_i - M_0)^T \quad (6)$$

$$M_0 = E[X] = \sum_{i=1}^{L} P_i M_i \quad (7)$$

로 정의된다.

클래스의 식별력을 높이기 위해서는 클래스 내의 분산이 작아야 하고, 클래스간의 분산이 커야한다. 이와 같은 의미의 평가함수는 클래스간의 분산이 커질수록 또는 클래스내의 분산이 작을수록 커져야 한다. 따라서 $J_1 = \mathrm{tr}(S_w^{-1} S_b)$을 최대화하는 변환행렬 A를 결정하면 각 클래스의 분산은 최소화하면서 클래스간의 분산은 최대화하는 특징성분을 구할 수 있다[2][3].

A를 클래스의 식별을 최대화하는 영역으로 입력 벡터를 변환하는 행렬이라고 하면, L개의 클래스를 가진 샘플로부터의 새로운 특징벡터는

$$Y = A^T X \quad (8)$$

로 표시할 수 있다. n 차원의 벡터 X를 m차원의 Y벡터로 변환시키는 선형변환을 나타낸다..

A는 n×m행렬이고 열벡터는 선형독립(linearly independent)이다. 그러나 신호를 표현할 때처럼 정규직교일 필요는 없다.

$J_1 = \mathrm{tr}(S_w^{-1} S_b)$이 최대값을 갖기위해서 $J_1 = \mathrm{tr}(S_w^{-1} S_b)$의 미분값을 0으로 놓으면 최적의 A는 다음 식을 만족한다.

$$(S_w^{-1} S_b)(AB) = (AB)\mu_m \quad (10)$$

이것은 μm 과 AB의 열벡터가 $S_w^{-1} S_b$의 m 고유값과 고유벡터임을 나타내다. 그러나, $S_w^{-1}S_b$ 는 항상 대칭적이지는 않기 때문에 고유벡터, 고유값계산이 불안정하다. $S_w^{-1}S_b$ 의 고유벡터와 고유값은 대칭인 각각의 행렬을 동시대각화(simultaneous diagonalization)하여 안정적으로 얻을 수 있다[3][4]. 여기서 S_b의 rank는 클래스의수(L)-1이므로 $S_w^{-1}S_b$ 의 rank또한 L-1이다. 이것은 $S_w^{-1}S_b$ 의 고유값은 L-1개만 0이 아닌 값을 가진다는 것을 의미한다. 따라서 입력벡터를 0이 아닌 고유값에 대응하는 L-1개의 고유벡터에 의해 span되는 L-1차원의 subspace로 사영할 수 있다. 이와 같이 OLT에 의해 변환된 특징의 최대차원은 원래의 특징량의 수가 L보다 큰 경우에는 식별해야 할 클래스의 수에 의해 정하여진다.

4.2 유사문자세트 구성 및 최적선형변환에 의한 인식

유사문자세트는 학습단계에서 작성한다. 제1단계의 인식결과, 제1후보에 의해서 선택하게 된다. 제1단계 인식의 학습에 사용된 60개의 문자세트를 인식하여 각 문자마다 1위부터 10위까지의 후보에 많이 들어온 빈도수대로 나열하여 N개의 문자를 갖는 유사문자세트를 준비한다. 표1에서 유사문자세트의 예를 보였다. 그리고 각 유사문자세트마다 N개의 클래스에 대하여 최적선형변환하여 각 유사문자세트내의 문자의 식별에 최적의 특징을 생성하여 제2단계 인식을 대비한다. 미지의 문자가 입력되면 1단계 인식후 제1후보를 OLT변환하고 제1후보의 유사문자세트에 대한 OLT 특징 공간에서 2단계 인식을 수행한다.

표 1. 유사문자 세트의 구성예

	유사문자										
감	감	강	캄	갈	잡	갈	칼	김	참	길	...
칸	칸	간	칼	캔	강	잡	같	길	김	탑	...
훌	훌	홀	돌	물	쿨	불	울	똘	들	흡	...
O	O	0	o	D	Q	.	C	G	c	n	...
:											

5. 실험 결과 및 분석

5.1 데이터 구성

사용한 문자데이터는 12종의 글자체(명조, 바탕, 견명조, 신명조, 돋움, 중고딕, 고딕, 견고딕, 굴림, 새굴림, 궁서, 궁서B)마다 6

개 세트, 총 72세트를 사용한다. 표준패턴을 만들기 위해서 1글자체마다 5개 세트(총 60세트)을 사용하였고, 테스트 데이터는 1글자체에 1개 세트 (총 12세트)를 사용하였다. 글자의 크기는 9.5, 10, 10.5의 3종류를 사용하였고 프린터는 잉크젯, 레이저 프린터 2종류를 사용하여 1글자체에 6개의 세트를 가지고 있다. 데이터세트 하나에는 한글 628문자, 숫자(10), 영어 대, 소문자(52), 특수문자(25)를 포함한 715문자로 구성되어 있다. 입력은 Epson GT9500스캐너를 사용하여 300dpi해상도로 읽어들였다.

5.2 실험결과

5.2.1 유사문자세트의 크기에 따른 인식률 비교

1단계 인식 후 1위 후보문자의 유사세트를 선택할 때 세트가 포함하는 문자수(N)에 따라 인식률을 비교하였다. 그림 2는 유사문자 후보들의 빈도순으로 5개에서 30개까지 유사문자의 세트크기에 따른 인식률을 나타내고 있다. 유사세트의 문자개수가 20개 이상에서는 안정된 인식률을 보이고 있고, 15개 이하의 유사문자세트는 클래스의 감소에 따른 특징량 수의 감소로 낮은 인식률을 나타내고있다.(OLT의 최대 특징량 수는 [클래스개수-1]이다). 최적선형변환을 하여 20개에서 30개의 차원을 사용하여 인식하더라도 원래의 36차원을 모두 사용하는 1단계의 인식률보다 높은 인식률을 나타내는 것을 확인할 수 있다. N=26일 때 최고 인식률(92.9%)을 얻었으며 이것은 Bayes 거리만에 의한 인식률(91.06%)보다 1.8% 개선되었다.

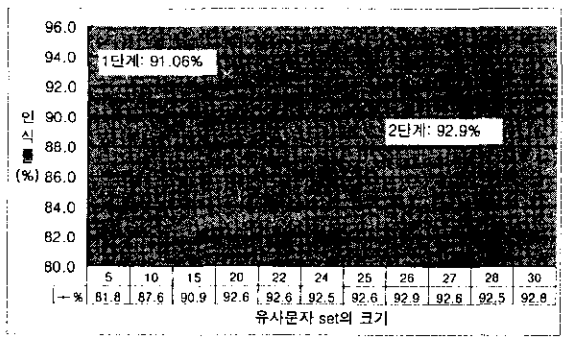

그림 2. 유사문자세트의 크기에 따른 인식률

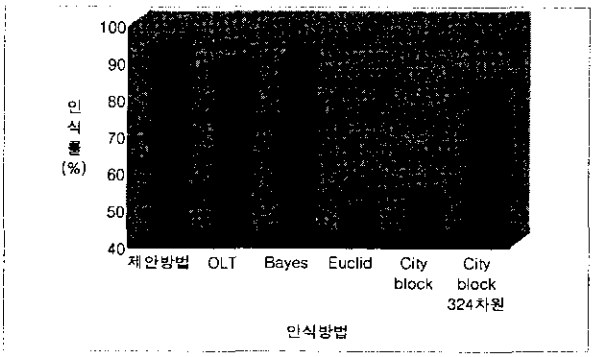

그림 3. 기존방법과 제안된 OLT인식방법의 인식률비교

그림 3에서는 제1단계의 Bayes와 제2단계의 OLT에 의한 제안방법과 기존의 인식방법들을 비교하였다. 제안방법에 대하여 전체문자에 대한 OLT변환에 의한 인식방법(OLT), Bayes 거리에 의한 인식방법(Bayes), Euclid 거리에 의한 방법, City Block 거리(절대거리)에 의한 방법, 324차원의 특징량을 이용한 City Block거리에 의한 방법을 비교하였다. 전체문자에 대한 OLT방법은 제안방법보다 낮은 인식률을 보이는데, 이것은 클래스가 715개나 되는 전체문자에 대한 OLT공간은 식별력이 좋아지지 않는다는 것을 보여준다.

City Block 거리법은 36차원(52.9%) 보다 324차원(80%)을 사용하였을 때 인식률이 27.1%나 높았으나 1/9의 특징량 수(즉 36차원)로써 문자의 분포를 고려한 Bayes 거리법이 91.1%를 나타내었다. 따라서 표준문자와의 절대거리만으로 거리를 계산하는 것은 많은 특징 차원수를 사용하더라도 적은 차원수로 분포를 고려한 방법보다는 인식률이 낮음을 보여주고 있다. 전체문자에 대한 OLT방법은 제안방법보다 약 5%낮은 인식률을 보인다. 따라서, 전체문자에 대한 최적선형변환결과는 식별해야 할 클래스가 많아 변환된 공간에서도 식별력이 좋지 않음을 알 수 있다. 제2단계의 OLT인식과정에서 오인식되는 문자(평균 51문자)중 유사문자세트에 입력문자가 포함되어있지 않아서 오인식된 경우는 715문자 중 평균4문자(0.48%)를 차지하였다.

6. 결론 및 발전방향

한글 문자인식과 같이 인식대상 클래스가 많고 유사문자가 많은 경우, 분포만을 고려하는 인식방법이나 최적 선형변환만에 의한 방법보다는, 제1단계에서는 각 클래스의 분포를 고려하는 Bayes 거리에 의한 인식방법이 대분류로서 유효하였고, 제2단계로서 소규모의 유사문자세트 내에서는 유사문자들의 변별력을 증강시키는 최적선형변환이 유효함을 입증하였다. 이와 같이 단계별로 적절한 인식방법을 적용함으로써 대규모 문자세트의 효율적인 인식이 가능하였다.

발전방향으로서 유사문자세트의 선정과정에 있어서 고정적인 유사문자수의 크기를 문자들의 특성에 따라서 가변적으로 정하는 것이 바람직하다. 또한 1단계 인식 후 5위 후보 이내의 인식률이 99%이상인 것을 감안하여 유사문자세트를 사전에 준비하지 않고 1단계인식 후의 후보 문자들로 구성하면 유사문자에 입력문자가 존재하지 않아서 오인식되는 것을 피할 수 있으리라 생각된다. 또한 두 단계의 인식과정에서 계산량이 많아지기 때문에 인식속도를 향상할 수 있도록 알고리즘의 고속화 연구가 필요하다.

참고문헌

[1] 강선미, 이기용, 황승욱, 양윤모, 김덕진, "고속문자인식을 위한 특징량 추출에 관한 연구", 전자공학회 논문지 29B, Vol. 11, pp.1047-1056, 1992.

[2] Daniel L. Swets and John Weng, "Using Discriminant Eigenfeatures for image Retrieval", *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, pp. 831-836.

[3] Keinosuke Fukunaga, Introduction to Statistical Pattern Recognition, second edition. Academic Press, 1990.

[4] 김형원, "최적선형변환에 의한 인쇄체 문자 인식", 석사학위논문, 고려대학교, 2001.

[5] Charles W. Therrien, "Decision, Estimation and Classification", John Wiley and Sons, 1989