유사 문자쌍을 구분하기 위한 한글 인식의 후처리

장승익° 김진형
한국과학기술원 전자전산학과
{sijang, jkim}@ai.kaist.ac.kr

Post-processing of Hangul Recognition for Discriminating Pairs of Characters

Seung-Ick Jang° Jin-Hyung Kim
Dept. of EECS, KAIST

요 약

유사한 형태의 필기 한글 문자쌍은 한글 인식 시 발생하는 오류의 많은 부분을 차지한다. 이는 유사한 문자들의 작은 차이를 인식기가 충분히 반영하기 어렵기 때문이다. 본 논문에서는 최근 주목 받고 있는 Support Vector Machine을 이용해 유사한 문자쌍을 검증하는 한글 인식 후처리 방법을 제안한다. 제안하는 방법은, 대부분의 문자 유사쌍이 한 두개의 자모만이 상이한 점에 착안하여 자모 단위로 문자 유사쌍을 구분한다. 기존 랜덤그래프를 이용한 한글 인식기를 이용하여 자모 분할을 수행하고, Support Vector Machine을 이용하여 분할된 결과를 검증한다. 제안한 방법은 유사쌍 구분에 중요한 자모만을 선택적으로 고려하여, 기존 한글 인식기의 부족한 점을 보완한다. 실험 결과, 자주 혼동되는 문자쌍들의 인식 오류가 정정되는 것을 볼 수 있었으며 그에 따라 한글 인식의 전체 성능이 향상되었다.

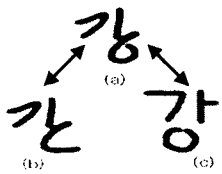
1. 서론

한글은 자모의 조합으로 이루어져 있으며, 이런 특성상 조합 가능한 글자의 수는 11,172자이다. 하지만 하나의 인식기로 한글 전체 클래스 11,172자를 구분해내기는 매우 어렵다. 특히 필기체 문자의 경우 필기자간의 변이가 상당히 커서, 인식은 더욱 어려워진다. 예를 들어 필기자가 바뀌어서 생기는 변이(intra-class variation)는 동일한 필기자가 다른 글씨를 쓸 때 생기는 변이(inter-class variation)보다 큰 경우가 많다. [그림 1]에서 (a)와 (b)는 동일 필기자가 쓴 것이고, (c)는 다른 필기자가 쓴 것이다. 하지만 이 경우 (a)와 (b)는 다른 글자이지만 변이가 비교적 작은 반면, (a)와 (c)는 같은 글자이지만 변이가 (a)와 (b)의 변이보다 더 크다.

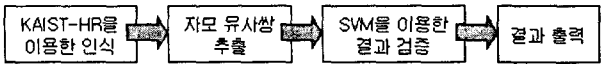
이런 경우는 하나의 인식기가 두 상반된 변이를 동시에 흡수하기는 매우 어려우며, 인식기의 성능도 떨어지게 된다. 따라서, 인식기의 부족한 점을 보완하고 전체 시스템의 성능을 향상시켜줄 수 있는 후처리기의 도입이 필요하다.

기존 인식기의 오류를 살펴보면, 글자 내의 일부 자모만이 틀리는 경우가 많다. 따라서 기존의 한글 인식기의 결과를 검증하기 위해서, 인식기에서 모델링 되지 못한 자모의 세부적인 특징들을 별도로 검증하는 방법을 적용하는 것이 바람직하다.

본 논문에서는 기존의 랜덤그래프를 이용한 한글 인식기의 결과를 바탕으로 자모 유사쌍을 선별하고, 자모 유사쌍에 대한 구분기를 이용해 인식 결과를 검증하



[그림 1] 유사한 형태의 글자



[그림 2] 전체 시스템의 구성

는 후처리 방법을 제안한다. 제안하는 시스템의 흐름은 [그림 2]와 같다. KAIST-HR[1]을 이용하여 인식을 수행하고, 그 결과를 바탕으로 자모들을 분할 및 추출하였다. 추출한 자모 유사쌍은 유사쌍 구분기를 이용해 검증을 수행하고, 그 결과를 출력한다. 본 논문에서 제안한 후처리기는 Support Vector Machine[9]을 이용해 구현하였다.

본 논문의 구성은 다음과 같다. 2장에서는 유사 문자쌍 구분에 대한 관련연구를, 3장에서는 KAIST-HR의 분할 결과를 이용한 자모 분할방법과 후처리에서 사용한 특징을 설명한다. 4장에서 실험을 수행한 결과를 보이며, 5장에서 본 논문의 결론을 맺고있다.

2. 관련연구

Takahashi, Chiang 등은 영상 기반 인식기에서 유사 문자쌍을 구분하는 방법을 제안하였다[2,3]. Takahashi의 경우 각각의 유사 문자쌍에 대한 이분 구분기(binary classifier)의 집합을 이용하여 인식 결과를 검증하였다[2]. 각각의 구분기는 인식기가 사용한 특징과 같은 특징들을 사용하고 있으나 각각의 구분기는 하나의 유사 문자쌍만을 구분하면 되기 때문에 전체 시스템의 성능은 더 향상되었다. Chiang의 경우는 새로운 특징을 추가해서 성능을 향상시키는 것이 아니라, 오히려 인식에 방해가 되는 특징을 제거함으로써 인식기의 성능을 높였다[3]. 그는 특징들이 인식에 얼마나 도움이 되는지를 정량화 시켜, 그것을 바탕으로 인식기에 대한 기여도가 낮거나 부정적인 영향을 주는 특징들은 제거하였다.

Chou, 권재욱, 김인중 등은 획 기반 인식기에서 유사 문자쌍을 구분하는 방법을 제안하였다[4,5,6]. Chou는 특정 획들의 상대적 위치나 형태 등의 특징을 사용해 한자의 유사 문자쌍을 구분하였다[4]. 그는 각각의 유사 문자쌍을 구분하는 규칙을 heuristic을 사용해 정의하였다. 그는 heuristic을 이용하는 방법이 성능에 있어 가장 효과적이라고 주장을 하고있다. 하지만 그의 방법은 많은 시간과 노력을 필요로 한다. 뿐만 아니라 전체 시스템의 성능은 시스템을 디자인하는 사람에 크게 의존적이며, 성능을 더 향상시키기도 상당히 어렵다. 권재욱은 한글 필기체 인식기에 사용하는 유사 문자쌍 구분기를 제안하였다[5]. 그 역시 구조적 정보를 사용하였으며, Chou가 사용한 방법과 동일한 방법인 heuristic을 이용해 접근하였다. 하지만 그는 학습 데이터로부터 구조적 정보들의 분포를 얻어, 그것을 이용해 유사 문자쌍 구분을 수행하였다. 김인중은 획들에 중요도를 부가하여 혼동되기 쉬운 한자의 유사 문자쌍을 구분하는 방법을 제안하였다[6]. 각 획의 중요도는 신경망의 역전파 학습을 통하여 자동으로 구하였다.

3. 자모 유사쌍 구분기

3.1 한글 인식기에 기반 한 유사쌍 선택

본 논문은 필기체 한글 인식기인 KAIST-HR을 기반으로 유사 문자쌍 구분을 수행한다. KAIST-HR은 인식에서 좋은 성능을 보이고 있으며, 획 기반 인식기로서 인식 결과와 함께 자모 분할 정보를 얻을 수 있는 장점을 가지고있다.

KAIST-HR의 인식 오류를 분석하고, 유사쌍 집합을 구하기 위해서 SERI database[7]를 이용해 인식 성능 실험을 수행하였다. 실험 결과, 1순위가 정답인 경우는 84.3%이었으나, 2순위 후보까지 고려할 경우 인식률은 91.7%까지 증가될 수 있었다. 바꾸어 말하면, 1순위와 2순위 후보에서 정답만을 선택한다면 최대 47.1%의 오류를 감소시킬 수 있음을 의미한다. KAIST-HR의 인식 결과에서 1순위와 2순위의 차이점을 살펴보면 60% 정도가 자모 하나만이 상이한 경우이며, 나머지 40% 정도만이 두개 이상의 자모가 상이하다. 이런 특성 때문에 낱자 유사쌍을 사용하는 것보다 자모 유사쌍을 사용하는 것이 더욱 효율적이다.

낱자 유사쌍은 유사한 형태를 가지는 두 글자로 구성된 문자쌍이다. 예를 들어 '마'와 '아'를 자주 혼동되는 유사한 형태를 가지는 글자들이라고 했을 때 {'마',

'아'}를 하나의 낱자 유사쌍이라 한다. 또, '막'과 '악'을 자주 혼동되는 글자들이라면, {'막', '악'}은 다른 하나의 낱자 유사쌍이 되는 것이다.

반면, 자모 유사쌍은 낱자 유사쌍의 두 글자를 각각 자모의 집합으로 표현했을 때, 두 집합간의 차집합 쌍이다. 예를 들어 '마'와 '아'가 낱자 유사쌍이라면, 집합 A = {'ㅁ', 'ㅏ'}, 집합 B = {'ㅇ', 'ㅏ'}로 표현하고 A – B와 B – A의 결과인 {'ㅁ'}과 {'ㅇ'}이 자모 유사쌍 {{'ㅁ'}, {'ㅇ'}}이 된다. 다른 예로 '쾨'와 '리'가 자주 혼동이 되는 문자쌍일 때, 'ㅣ'를 제외하고 서로 상이한 자모의 조합 {'ㅋ', 'ㅗ'}와 {'ㄹ'}이 자모 유사쌍 {{'ㅋ', 'ㅗ'}, {'ㄹ'}}이 된다.

[표 1]은 SERI database의 200개 set, 약 10만자를 이용해 얻은 낱자 유사쌍과 자모 유사쌍의 분포이다. 이를 보면 자모 유사쌍을 사용할 경우 낱자 유사쌍보다 훨씬 적은 수의 유사쌍만으로 동일한 수의 인식 결과를 검증할 수 있음을 알 수 있다.

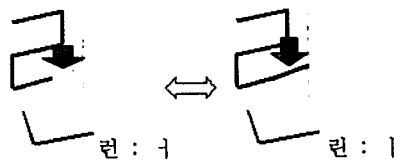
3.2 자모 영역의 선택

자모 유사쌍을 사용하기 위해서는 입력 영상을 각각의 자모들로 분할하고, 자모 영역을 추출하는 과정이 필요하다. 본 논문에서는 KAIST-HR에서 나오는 분할 정보를 바탕으로 자모 영역을 추출한다.

그런데, 하나의 입력 영상에 임의의 글자를 매칭시켰을 때, 글자의 label에 따라 자모의 영역이 다르게 분

[표 1] 유사쌍의 분포

유사쌍 누적 비율	낱자 유사쌍 수	자모 유사쌍 수
10%	36	1
20%	87	3
30%	160	6
40%	258	11
50%	397	23
60%	613	39
70%	936	69
80%	1469	137
90%	2222	344
100%	2975	1031



[그림 3] 서로 다르게 분할된 글자의 예

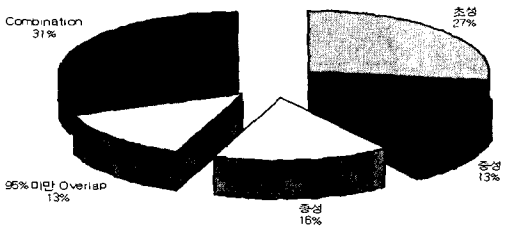
할될 수 있다. [그림 3]은 동일한 영상에서 자모의 영역이 다르게 분할된 경우이다. 화살표가 가리키고 있는 부분은 '런'에서는 중성으로, '린'에서는 초성으로 분할되었다.

본 논문에서는 인식기에서 나온 분할 결과가 서로 얼마나 상이한지 측정하기 위해 *overlapping rate*를 정의하였다. 이 overlapping rate는 유사한 두 글자에서 자모 유사쌍을 제외한 뒤, 각 글자의 나머지 자모 획들의 길이의 합에 대한 비율이다. KAIST-HR에서 나온 결과에서, 1순위 후보를 C_1이라 하고, 2순위 C_2라고 하자. 그리고 이들의 자모 유사쌍은 초성으로만 구성되었다고 하자. 다음으로 C_1과 C_2를 구성하는 획들의 집합을 S_1과 S_2라고 하고, 각각의 집합은 부분획인 s_i와 s_j로 이루어졌다고 했을 때 overlapping rate $O(C_1, C_2)$는 식 (1)과 같이 표현할 수 있다.

$$O(C_1, C_2) = \frac{\min(\sum_{i \neq 초성} length(s_i), \sum_{j \neq 초성} length(s_j))}{\max(\sum_{i \neq 초성} length(s_i), \sum_{j \neq 초성} length(s_j))} \quad (1)$$

입력 영상의 잡영과 길이가 짧은 획들에 의해 자모의 분할 영역이 완전히 일치하지 않는 경우가 발생한다. 이런 이유로, 본 논문에서는 입력의 overlapping rate가 95% 이상이면 자모 분할 영역이 동일한 것으로 간주하고, 이를 후처리 대상으로 선택하였다. [표 2]는 KAIST-HR의 1순위 후보와 2순위 후보에서 추출한 자모 유사쌍의 분포이다. Overlapping rate가 95%이상이면서 하나의 자모로 구성된 유사쌍은 전체의 56%를 차지하고있다. 인식오류 감소가 최대47.1%까지 가능한 점을 감안하면, 자모 하나로 구성된 유사쌍만을 고려하더라도 최대 26.4%까지 오류 감소가 가능하다.

[표 2] 1순위와 2순위로부터 추출한 유사쌍 분포



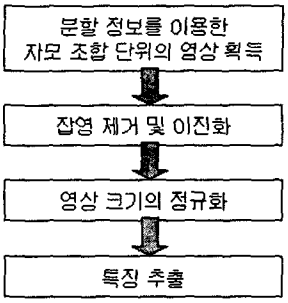
3.3 자모 구분기

KAIST-HR은 형태의 다양성으로부터 오는 변이를 흡수할 수 있는 장점을 가지고있다. 하지만 자모를 모델링을 하는 과정에서 다양한 필기 형태의 변이를 흡수하기 위해서 자모의 모델을 일반화(generalization)하게 된다. 이러한 과정에서 자모의 형태가 충분히 모델링 되지 못하는 경우가 발생한다.

일례로서 [그림 4]는 '디'가 '티'로 잘못 인식된 경우를 보여주고 있다. 왼쪽 부분은 추성인 'ㄷ'으로, 오른쪽 부분은 중성인 'ㅣ'로 인식이 되었다. 이는 자모의 형태가 모델의 일반화로 인해서 잘 표현하지 못하고 있음을 보여준다. 본 논문에서는 이런 문제점을 해결하기 위해, 부족한 자모의 모양에 대한 모델링을 보완하는 방법으로 mesh feature를 사용해 자모의 모양에 대한 표현을 하였다.

후처리기에 사용할 특징을 추출하는 과정은 [그림 5]와 같다. KAIST-HR의 인식 결과와 분할 정보를 바탕으로 입력 영상에서 비교 대상인 자모 단위의 영상을 얻는다. 예를 들어 자모 유사쌍이 초성으로만 구성되었다면, 초성의 영상만을 얻게 된다. 얻어진 영상에서 잡영을 제거하고 영상에 대해 이진화를 수행한다. 다음 단계에서 영상의 크기를 정규화 시킨다. 본 논문에서 사용한 정규화 방법은 Yamamoto의 비선형 정규화 방법이다

[그림 4] '디'가 잘못 인식된 경우



[그림 5] 특징 추출 단계

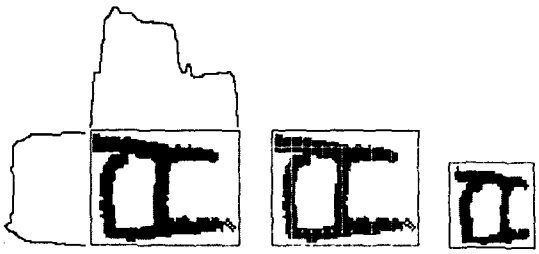
[8]. 이 방법은 선의 밀도를 선이 교차한 수와 내접원의 크기 등을 이용해 측정한 후, 선의 밀도를 평활화(equalization) 시켜주는 방법이다. [그림 6] 참조.

정규화는 30 x 30의 크기의 영상이 되도록 수행하였으며, 정규화 된 영상을 후처리기의 특징으로 사용하였다. 본 논문은 자모 구분기로 Support Vector Machine을 이용하였으며, 실제 후처리기 시스템은 Joachims가 구현한 SVM^{light} 라이브러리를 기반으로 구현하였다[10].

4. 실험 및 실험결과

학습과 인식률 실험에 사용한 데이터는 SERI database이다. SERI database는 1000개의 set으로 이루어져 있다. 하나의 set은 제약 없이 쓰여진 필기체 한글 520자의 낱자 영상으로 구성되어 있으며, 이 글자들은 한글에서 발생 빈도가 높은 순으로 뽑은 것이다.

학습 데이터는 SERI database의 80개 set으로 구성되어 있으며, 글자의 이름이 잘못 표기된 글자를 제외한



(a) 수직 수평에 대한 투영 (b) 표본화 (c) 결과

[그림 6] Yamamoto의 비선형 정규화 방법

총 40,580자이다. 각각의 낱자 영상을 KAIST-HR로부터 얻은 분할 정보를 이용해 자모 영역을 분할하고, 이것을 이용해 자모 유사쌍 학습 데이터를 생성했다. 각각의 자모 유사쌍의 학습 데이터 수는 유사쌍에 따라 약 1,000개에서 8,000개 정도이다.

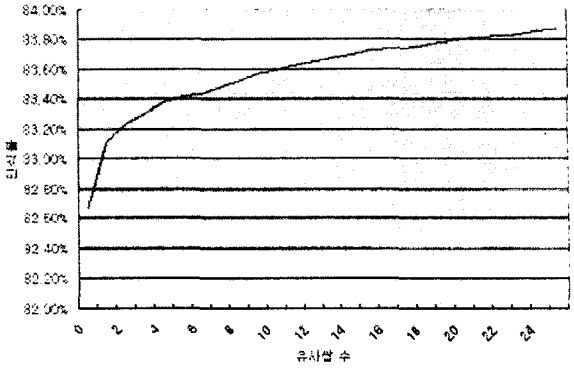
테스트 데이터는 SERI database 120개 set, 총 62,398자로 구성되어있다. 그리고, 실험에서 사용한 유사쌍 집합은 발생 빈도가 높은 초성 자모 유사쌍 25개로 이루어져 있으며, 초성 자모 유사쌍의 87%를 차지한다.

실험은 테스트 데이터를 KAIST-HR을 이용해 인식하고, 인식 결과와 분할 정보를 이용해 후처리 수행하는 순서로 진행하였다. 인식 결과가 자모 유사쌍 집합에 속할 경우 본 논문에서 제안한 후처리기를 수행하였다. 각

[표 3] 실험결과

유사쌍	전체수	정답수	오류정정	오류발생	오류감소
ㅁ↔ㅇ	5522	4946	346	64	49.00%
ㄸ↔ㅍ	473	375	77	6	72.40%
ㅈ↔ㅊ	1251	1162	63	19	49.40%
ㅇ↔ㅎ	392	314	55	8	60.30%
ㄷ↔ㅌ	1104	1048	27	4	41.10%
ㄹ↔ㅌ	366	305	27	14	21.30%
ㄱ↔ㅈ	428	353	34	7	36.00%
ㅁ↔ㅂ	742	674	34	8	38.20%
ㅇ↔ㄷ	233	184	33	2	63.30%
ㄷ↔ㄴ	586	557	20	3	58.60%
ㄹ↔ㄴ	498	445	27	7	37.70%
ㄹ↔ㅎ	169	135	20	6	41.20%
ㅇ↔ㅂ	156	130	16	0	61.50%
ㅁ↔ㅍ	242	213	12	3	31.00%
ㄲ↔ㅈ	207	178	20	0	69.00%
ㄷ↔ㄹ	338	300	11	3	21.10%
ㅍ↔ㅌ	61	43	13	8	27.80%
ㄴ↔ㅅ	398	371	13	3	37.00%
ㄷ↔ㅈ	122	99	16	0	69.60%
ㅎ↔ㅌ	84	68	14	3	68.80%
ㅁ↔ㄲ	280	263	11	4	41.20%
ㅇ↔ㄴ	150	135	11	5	40.00%
ㅅ↔ㅈ	435	399	12	4	22.20%
ㅎ↔ㅂ	126	104	17	2	68.20%
ㄹ↔ㅂ	76	63	9	1	61.50%
계	14439	12864	938	184	47.90%

[표 4] 유사쌍 수에 따른 인식률



유사쌍수	0	5	10	15	20	25
인식률	82.66%	83.41%	83.60%	83.72%	83.80%	83.87%
오류감소율	0.00%	4.33%	5.42%	6.11%	6.57%	6.98%

각의 유사쌍에 대한 실험 결과는 [표 3]과 같다. [표 3]에 나열된 유사쌍의 순서는 유사쌍 발생 빈도가 높은 순으로 나열한 것이다. 'ㅁ'과 'ㅇ'의 경우 5,522자 중에서 KAIST-HR이 정인식을 한 수는 4,946자였다. 후처리를 통해서 64자가 추가적으로 오인식 되었지만 346자의 오류 정정을 하여, 49.0%의 오류 감소를 보였다.

[표 4]는 후처리 후의 시스템 전체 인식률을 보여준다. 테스트 데이터 62,398자에 대한 KAIST-HR의 인식률은 82.66%이였다. 자모 유사쌍 집합에 유사쌍 발생 빈도 순서에 따라 5개씩 자모 유사쌍을 추가하였을 때 인식률은 그에 상응해서 올라갔다. 자모 유사쌍 집합의 크기가 25를 넘었을 때 인식률은 더 이상 크게 증가하지 않았다. 이는 초성 자모 유사쌍 중에서 상위 25개가 차지하는 비중이 매우 크기 때문이다.

[그림 7]은 오류 정정의 예를 보여주고 있다. [그림 7]의 (a)는 KAIST-HR이 입력 영상을 '일'로 오인식 한 것을 제안한 후처리 방법을 사용해 '밀'로 정정한 경우이다. (b)와 (c)도 각각 '먕'과 '퍄'가 '양'과 '따'로 오류가 정정되었음을 보여주고 있다.

(a) '일'에서 '밀'로 오류 정정

(b) '먕'에서 '양'으로 오류 정정

(c) '파'에서 '따'으로 오류 정정

[그림 7] 오류 정정의 예

5. 결론

본 논문에서는 Support Vector Machine을 이용해 유사한 형태의 문자쌍을 검증하는 한글 인식 후처리 방법을 제안하였다. 입력 영상을 KAIST-HR을 이용해 자모 단위로 분할하고, 자모 영역이 올바르게 분할되었는지 검사한 뒤 자모 구분기의 입력 대상으로 사용하였다. Support Vector Machine으로 구현된 자모 구분기는 자모의 모양을 잘 표현할 수 있는 mesh feature를 사용해 학습시켰다.

실험에서는 초성으로만 구성된 발생 빈도가 높은 상위 25개로 구성된 유사쌍 집합을 사용하였다. 후처리 대상에 속한 글자들의 인식률은 89.1%에서 94.3%로 인식오류를 47.9% 감소시켰다. 제안한 방법을 적용해 중성 유사쌍, 종성 유사쌍 또는 자모 조합의 유사쌍 등에도 적용해 인식률 향상이 가능하다.

6. 참고문헌

[1] H.Y. Kim, Representation and parameter estimation of hierarchical random graph and its application to handwritten Hangul recognition. Ph.D. Dissertation. KAIST (1999)

[2] H. Takahashi, T.D. Grffin, Recognition enhancement by linear tournament verification, Proc. Second International Conference on Document Analysis and Recognition, (1993) pp. 585-588

[3] C.C. Chiang, S.S. Yu, A method for improving the machine recognition of confusing Chinese characters. Proc. Thirteenth International Conference on Pattern Recognition, (1996) pp. C79-83

[4] K.S. Chou, K.C. Fan, C.K. Lin, A knowledge based approach to the recognition of on-line confusing Chinese characters, Proc. Fourth International Workshop of Frontiers of Handwriting Recognition, (1994) pp. 185-194

[5] J.O. Kwon, B.K. Sin, J. Kim, Recognition of on-line cursive Korean characters combining statistical and structural method, Pattern recognition 30 (8) (1997) pp. 1255-1263

[6] I.J.Kim, Handwritten Chinese Character Recognition using Statistical Structure Modeling and Stroke Weighting. Ph.D. Dissertation. KAIST (2001)

[7] D.-I. Kim, S.-W. Lee, An automatic evaluation of handwriting qualities for offline handwritten Hangul character database KU-1, Proc. of the 25th Korea Information Science Society Conference 25 (1) (1998) pp. 707-709

[8] H. Yamada, K. Yamamoto and T. Saito, A Nonlinear Normalization method for Handwritten Kanji Character Recognition-Line Density Equalization, Pattern Recognition, vol. 23, no. 9 (1990) pp.1023-1029

[9] Vladimir N. Vapnik, The Nature of Statistical Learning Theory. Springer, (1995)

[10] http://ais.gmd.de/~thorsten/svm_light/