Journal of the Korean Data & Information Science Society
2010, **21**(4), 689–697

한국데이터정보과학회지

서포트벡터머신과 정칙화판별함수를 이용한 비디오 문자인식의 분류 성능 개선[†]

임수열[1] · 백장선[2] · 김민수[3]

[123]전남대학교 통계학과

접수 2010년 5월 11일, 수정 2010년 7월 5일, 게재확정 2010년 7월 8일

요 약

본 연구에서는 비디오이미지로부터 추출된 텍스트영역으로부터 문자인식을 수행하였다. 비디오영상으로부터 추출된 문자열은 한글, 영어, 숫자, 특수문자 등으로 혼합되어 있거나, 또는 다양한 폰트와 크기, 그래픽 형태의 글자 존재, 영상의 기울어짐, 끊김, 잡영, 접촉, 저해상도의 글자 등으로 인하여 일반적인 문자인식에 비해 많은 어려움이 존재한다. 이와 같은 어려움을 극복하기위해 본 연구에서는 모든 글자에 대해서 인식하지 않고 가장 빈번하게 등장하는 글자만을 인식하고 나머지는 버리는 방법을 사용하였으며 지지도벡터기계와 정칙화판별분석의 2단계 문자인식 방법을 이용하여 인식률을 개선하였다. 또한 인식률이 좋지 못한 4형식과 5형식 글자에 대해 모음별로 중분류를 실시하였다. 실험결과 지지도벡터기계와 정칙화판별분석을 동시에 사용하는 방법이 다른 문자인식의 방법들보다 인식률이 우수하였으며, 부분적인 중분류의 방법을 이용한 경우 향상된 인식 성능을 나타냈다.

주요용어: 문자인식, 분류, 정칙화판별함수, 지지도벡터기계.

1. 서론

문자인식은 입력 문자로부터 특징을 추출하고, 그 특징들의 정보를 이용하여 미리 정해진 글자집단 중의 한 집단으로 분류하는 과정이라고 할 수 있다 (Kim과 Baek, 1999). 오프라인의 문자인식은 필기체 문자인식과 활자체 문자인식으로 나뉘는데, 필기문자의 변형이 많기 때문에 필기체 문자인식이 훨씬 어렵다고 알려져 있다.

그림 1.1은 비디오 영상으로부터 문자인식을 하는 단계를 보여주고 있으며, 그림 1.2와 같이 비디오 영상내의 자막을 인식하고자 하는 비디오 문자인식의 경우 영상속의 글자에 다양한 활자체와 필기체가 섞여있을 뿐만 아니라 빛의 방향과 명암까지 고려해야하므로 인식이 더욱 어렵게 된다 (김영화와 남지호, 2009). 또한 그림 1.3을 보면 모두 한글 '다'의 의미를 갖는 글꼴이지만, 글자의 크기와 치우침, 그리고 해상도 등이 모두 다르다. 즉 같은 의미를 갖는 글자이지만 표기되는 방법들이 매우 다양하기 때문에 문자인식에 어려움이 있음을 알 수 있다.

기존의 한글 영상을 인식하는 방법으로는 인식 단위에 따라 크게 3가지로 구분 할 수 있다. 먼저 낱자 영상을 입력으로 받아서 낱자 대상 클래스 중에서 해당 클래스를 찾는 낱자단위 인식방법 (이성환,

[†] 이 논문은 2007년도 전남대학교 학술연구비의 지원을 받아 연구되었음.
[1] (500-757) 광주광역시 북구 용봉동 300번지, 전남대학교 통계학과, 박사과정.
[2] (500-757) 광주광역시 북구 용봉동 300번지, 전남대학교 통계학과, 교수.
[3] 교신저자: (500-757) 광주광역시 북구 용봉동 300번지, 전남대학교 통계학과, 조교수.
E-mail: kimms@chonnam.ac.kr

그림 1.1 비디오영상의 문자인식그림

그림 1.2 다양한 비디오영상속의 자막그림

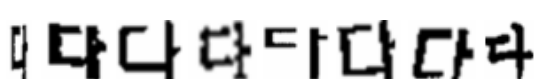

그림 1.3 비디오 영상속의 다양한 글꼴

1993), 낱자 영상을 자소 영상으로 분리하여 각 자소별로 인식하고 최종적으로 각 자소의 결과를 조합해서 낱자로 출력하는 자소단위 인식방법 (권재욱 등, 1992; 이진수 등, 1996; 이판호 등, 1994), 그리고 두 가지 방법의 중간 형태인 자소조합 인식방법 (장승익과 남윤석, 2004)으로 분류할 수 있다. 본 연구에서는 낱자 영상을 자소 영상으로 나누어 자소별로 인식하는 방법인 자소단위 인식방법을 이용하여 한글 영상의 인식에 사용하였으며, 이 방법은 한글의 조합적인 성격을 잘 반영함과 동시에 자소의 결합에 의해 모든 한글 문자를 인식할 수 있다는 큰 장점을 가지고 있다.

본 논문의 구성은 다음과 같다. 제2장에서는 문자영상의 인식에 사용된 방법들에 대하여 살펴보고, 제3장에서는 실제 자료들을 이용한 문자 인식의 방법별 성능에 대하여 비교 분석한다. 마지막으로 제4장에서는 결론 및 토의를 통해 본 연구를 정리한다.

2. 문자영상 인식 방법의 고찰

실생활에서 사용되는 한글의 글자 수는 2,350자로서 매우 많다. 하지만 그림 2.1과 같이 609개의 글자만으로도 비디오 영상에서 사용되는 글자의 99%를 차지할 만큼 실제 사용빈도가 매우 적은 글자가 많

음을 알 수 있다.

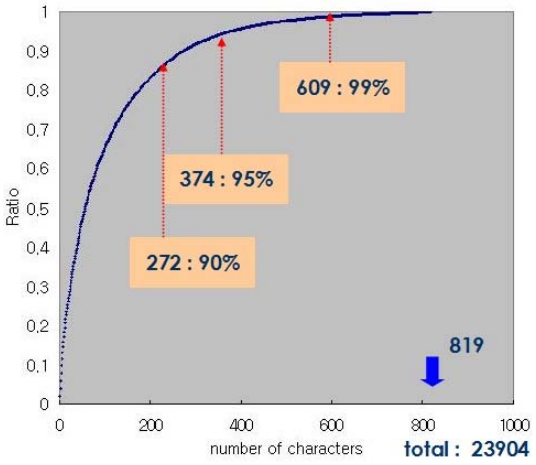

그림 2.1 비디오 영상에 사용되는 글자의 개수와 비율

전처리란 입력된 문자 데이터를 사용하여 문자를 구성하는 선의 굵기를 고르게 하거나, 또는 작은 오염이나 탈락을 제거하는 잡음 제거 성격의 처리와 다음 단계의 인식처리에 적합한 형태로 문자 데이터를 변환하는 작업 등을 의미하며 방송 등 동영상 자료를 토대로 문자의 정보만을 인식하여 저장하는 것을 비디오 색인화라고 한다 (조완현, 1999). 비디오 문자를 인식하기 위해서는 다음의 네 가지 과정을 거치게 된다.

① 문자 영역 추출: 비디오 영상을 입력으로 받아 문자가 존재하는 영역의 위치를 탐색하여 문자 영역을 추출
② 문자 영역 분리: 추출된 문자 영역의 영상으로부터 배경을 제거하고 이진화하여 글자의 화소만 남김
③ 문자 영역 분할: 일반적으로 자막은 문자열 형태로 문자들이 붙어 있는 것이 대부분으로 인식을 위하여 이를 낱자 단위로 분할
④ 문자 인식: 분할된 각각의 낱자 영상을 인식하여 이에 상응하는 문자 코드로 변환

그림 1.1은 위의 네 단계를 거쳐 낱자 결과를 결합한 문자열을 출력하게 되는 과정으로, 마지막 네 번째 단계인 문자인식 과정에서는 문자 영상의 종류와 크기 등을 정규화하고 잡영을 제거하는 전처리 과정을 거치게 된다. 그 결과, 각각의 문자를 판별하기 위해서 필요한 특징을 추출하여 여러 모형 중 가장 유사한 모형을 선택하여 해당 모형에 대한 문자 코드를 출력하는 과정을 거치게 되는 것이다. 하지만 실제로 비디오 영역에서의 추출된 문자영역 인식을 실행하는 경우 그림 2.2와 같이 문자들이 한글, 영어, 숫자, 특수문자 등으로 혼합되어 있거나, 또는 다양한 폰트와 크기, 그래픽 형태의 글자 존재, 영상의 기울어짐, 끊김, 잡영, 접촉, 저해상도의 글자 등으로 인하여 많은 어려움에 직면해 있다.

따라서 문자인식의 문제는 쉽게 해결할 수 있는 문제가 아니며 하나의 일반화된 알고리즘으로 해결하기보다는 각 형태에 맞는 규칙기반접근 (rule-based approach)을 적극적으로 활용한 끈질긴 노력이 뒷받침되어야 한다.

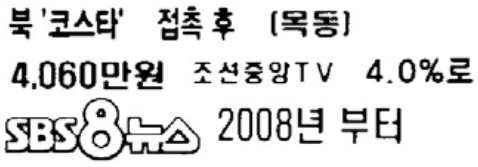

그림 2.2 어려운 문자인식의 예

2.1. 문자의 형식 분류

실제로 사용되고 있는 한글의 인식 대상 수가 매우 많은 만큼 실제로 정확하게 문자를 인식하는 과정에는 많은 어려움이 존재하고 있다. 하지만 한글의 특성을 고려해보면 종성의 유무, 또는 수직·수평으로의 자음과 모음의 분할 여부 등에 따라 그림 2.3과 같이 한글을 6가지 형식으로 분류할 수 있다. 이와 같

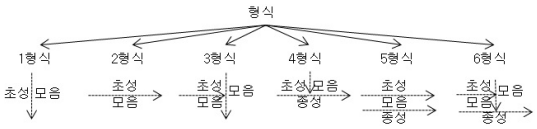

그림 2.3 한글의 조합 방식에 따른 형식 분류

이 형식 분류에 의해 분류된 낱자 영상들은 그림 2.4와 같이 낱자 영상으로부터 추출된 방향각 특징 벡터를 입력으로 받아서 각 자소를 개별적으로 인식한 후 그 결과를 결합함으로써 낱자 레이블을 출력하는 비분할 자소인식 방법으로 인식하게 된다.

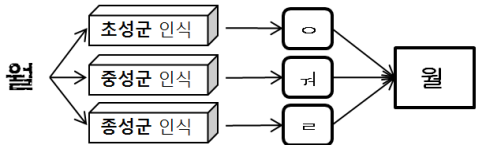

그림 2.4 비분할 자소인식의 예

2.2. 형식분류기 및 관련 연구

자소군이란 공통의 자소를 가지는 낱자 영상 집합을 의미하며, 각 자소군에는 각 형식별로 해당 자소의 크기나 위치가 유사한 공통점이 반영되어 있다. 예를 들어 '가'와 '궁'의 경우 초성군이 'ㄱ'군으로 같지만 '가'는 1형식, '궁'은 5형식으로 'ㄱ'의 크기나 위치가 서로 다른 특징을 가지게 된다. 이와 같은 특징을 기반으로 형식분류기를 훈련하게 되고, 훈련이 끝난 후 입력된 문자들의 형식을 구분할 수 있게 된다.

2.2.1. 지지도벡터기계 (Support Vector Machines: SVM)

SVM은 Vapnik (1995)에 의해 개발된 분류 기법으로 입력공간과 관련된 비선형 문제를 고차원의 특징 공간에서의 선형문제로 대응시켜 나타내기 때문에 수학적으로 분석하는 것이 쉽다고 알려져 있는 이

진분류기법이다. 또한 조정해야 할 모수의 수가 많지 않기 때문에 학습에 영향을 미치는 요인들을 비교적 간단하게 규명할 수 있으며, 전역 최적해를 구할 수 있다는 점에서 인공신경망보다 우월한 기계학습적 기법으로 주목을 받고 있다. 최근 몇 년간 SVM을 이용한 다양한 연구가 진행되고 있으며 문서분류, 영상인식, 문자인식 등에서 뛰어난 성능을 보여주고 있다 (Joachims, 1998). 또한 채권신용의 등급예측, 기업부도의 예측 및 주가지수의 예측 등에서도 우수한 성능을 보여주고 있다 (박정민 등, 2005; Huang 등, 2004). 한글의 경우와 같이 다중 클래스를 분류하기 위해서는 다중 SVM을 이용할 수 있다 (Hwang 등, 2006). 하지만 클래스의 수가 많은 다중 분류의 경우 SVM의 수가 많아져 계산량이 증가하게 되고, 한 클래스에 속하는 자료의 수가 많지 않을 경우 다른 클래스와의 자료 수의 불균형으로 인하여 학습에 나쁜 영향을 미치게 되기도 한다.

2.2.2. 신경망모형 (Neural Network: N-N)

신경망모형은 생물학적인 프로세스를 컴퓨터를 이용하여 모형화하기 위한 노력에서 비롯된 것으로, 복잡한 구조를 가진 자료의 예측문제를 해결하기 위해 사용되는 유연한 비선형 모형의 하나로 분류된다. 즉 신경망은 비선형적인 현상을 분석하는데 유용하고, 학습을 통한 분석으로 분석시간이 짧고 비용이 적으며 패턴인식, 예측, 분류 등에 효과적인 장점을 가지고 있다 (오광식 등, 1997). 이와 같은 장점을 갖는 신경망모형을 이용하여 Kim과 Lee (2003)는 신용평가모형, Cho와 Park (2008)은 보험회사의 이탈고객에 관한 분석을 하였다. 하지만, 은닉층과 은닉마디가 많아질수록 모형은 더욱 복잡해지고 추정해야할 계수의 수가 급격히 증가하기 때문에 최적화가 어려우며 어떤 입력변수가 중요한지 또한 그것들이 어떻게 상호작용을 하는지에 대한 결정을 하기가 어려워 결과에 대한 간편한 해석의 어려움을 겪게 된다.

2.2.3. 선형 및 이차판별분석 (LDA와 QDA)

판별분석이란 이미 알려진 상호배반적인 몇 개 집단에 속하는 다변량 관측치로부터 각 집단의 차이를 분류할 수 있는 함수를 추정하거나, 소속집단이 알려지지 않은 새로운 관측치를 추정된 함수를 이용하여 어떤 집단으로 분류할 것인가를 결정하는 다변량분석기법이다. 따라서 임의의 관측치 $\boldsymbol{x}$가 어느 특정한 집단 G_i의 표본평균 $\bar{x}_i$로부터 떨어져 있는 거리를 측정하면 그 관측치가 어느 집단에 가장 가깝게 위치해 있는가를 판단할 수 있다. 또한 각 집단의 공분산행렬이 동일한 경우와 동일하지 않은 경우에 따라 다음과 같이 관측치 $\boldsymbol{x}$에서 집단 G_i까지의 일반화 거리 자승, $D_i^2(\boldsymbol{x})$을 정의할 수 있다 (Johnson과 Wichern, 1992).

①집단의 공분산 행렬이 같은 경우 (선형판별분석: Linear Discriminant Analys, LDA)

$$\hat{L}_i(\boldsymbol{x}) = \bar{\boldsymbol{x}}_i' \boldsymbol{S}_p^{-1} \boldsymbol{x} - \frac{1}{2} \bar{\boldsymbol{x}}_i' \boldsymbol{S}_p^{-1} \boldsymbol{x} + \ln p_i \tag{2.1}$$

이때, 새로운 관측치 $\boldsymbol{x}$는 $\hat{L}_j(\boldsymbol{x}) = \max(\hat{L}_1(\boldsymbol{x}),\ \hat{L}_2(\boldsymbol{x}),\ \cdots,\ \hat{L}_k(\boldsymbol{x}))$인 집단 G_j로 분류한다.
②집단의 공분산 행렬이 다른 경우 (이차판별분석: Quadratic Discriminant Analys, QDA)

$$\hat{Q}_j(\boldsymbol{x}) = \frac{1}{2} \ln |\boldsymbol{S}_i| - \frac{1}{2} (\boldsymbol{x} - \bar{\boldsymbol{x}}_i)' \boldsymbol{S}_i' (\boldsymbol{x} - \bar{\boldsymbol{x}}_i) + \ln p_i \tag{2.2}$$

이때, 새로운 관측치 $\boldsymbol{x}$는 $\hat{Q}_j(\boldsymbol{x}) = \max(\hat{Q}_1(\boldsymbol{x}),\ \hat{Q}_2(\boldsymbol{x}),\ \cdots,\ \hat{Q}_k(\boldsymbol{x}))$인 집단 G_j로 분류한다.

2.2.4. 정칙화판별분석 (Regularized Discriminant Analysis: RDA)

일반적으로 표본의 크기가 작은 때, 추정해야 하는 모수의 수가 많을 경우 모수의 추정량은 매우 불안정하고 분산이 커지게 된다. 이 경우 정칙화 방법은 표본에 근거한 불편추정량을 사용하는 대신 편의추정량을 사용함으로써 추정량의 정밀도를 향상시켜 잠재적으로는 편의를 증가시키지만 추정량의 분산은 줄일 수 있게 된다. Friedman (1989)은 표본공분산행렬의 추정에 있어 식 (2.3)의 정칙화 방법을 제안하였다.

$$\widehat{\Sigma_i}(\lambda,\gamma) = (1-\gamma)\widehat{\Sigma_i}(\lambda) + \frac{\gamma}{p} tr[\widehat{\Sigma_i}(\lambda,\gamma)]\mathrm{I} \tag{2.3}$$

여기서

$$\widehat{\Sigma_i}(\lambda) = \frac{(1-\lambda)n_i\widehat{\Sigma_i} + \lambda n\widehat{\Sigma}}{(1-\lambda)n_i + \lambda n}$$

이다. i번째 그룹에 대한 판별점수는 식 (2.4)이고 이를 사용하여 판별분석을 한다.

$$\hat{d}_i(\boldsymbol{x}) = (\boldsymbol{x}-\bar{\boldsymbol{x}}_i)^T\widehat{\Sigma_i}^{-1}(\lambda,\gamma)(\boldsymbol{x}-\bar{\boldsymbol{x}}_i) + \ln|\widehat{\Sigma_i}(\lambda,\gamma)| - 2\ln\pi_i \tag{2.4}$$

이러한 접근을 정칙화판별분석 (RDA)이라 한다.

3. 인식 방법의 성능 비교

본 연구에서 사용되는 문자 인식기에 대한 성능 비교 실험을 위해 뉴스에 있는 비디오 자막 영상들에 대해서 인식을 실행하였다. 비디오 뉴스 자막은 각 방송사별로 여러 시간대에서 추출된 자료이며, 분류기의 훈련과 성능의 검증에 사용된 자료의 수는 각각 2,000개씩 사용하였다. 각 글자로부터의 특징벡터는 6×6 비선형 메쉬 (nonlinear mesh)에 의한 8방향 특징추출법으로 추출하였다. 또한 자료의 1-6형식 구분은 SVM을 이용하였으며, 각 형식에 대한 세부 분류에는 N-N, LDA, RDA의 방법으로 비교 분석하였다.

표 3.1은 SVM을 통한 형식 분류 결과로서 검증집단에 대한 인식률이 99.59%이며, 이는 거의 정확한 인식 성능을 나타내고 있음을 알 수 있다. 또한 N-N, LDA, RDA의 방법으로 세부 분류한 결과 1-6형식에 대한 각각의 검증집단에 대한 정분류율과 전체 정분류율이 제시되고 있다. 각 형식별 자료들의 인식률을 살펴보면 RDA가 N-N이나 LDA보다 모든 형식에서 높은 인식률을 나타내고 있으며, 검증자료에 대한 전체 정분류율은 N-N과 LDA가 비슷한 반면 RDA는 95.08%로 다른 방법들보다 인식성능이 높음을 알 수 있다. 즉 SVM을 이용하여 문자의 형식을 분류한 후, RDA를 이용하여 문자들을 인식할 경우 가장 높은 인식률을 갖는다고 할 수 있다.

표 3.1 각 형식별 자료의 인식률에 대한 정분류율 (%)

		형식						
		SVM을 이용한 형식분류율: 99.59%						
		1형식	2형식	3형식	4형식	5형식	6형식	정분류율
방법	N-N	91.50	97.00	91.45	90.71	89.30	93.97	91.86
	LDA	93.21	97.40	94.30	88.40	89.00	97.68	91.85
	RDA	95.90	98.58	95.64	93.15	92.22	99.30	95.08

하지만 RDA가 다른 방법들보다는 높은 인식률을 갖는 방법임에는 틀림이 없으나, 4형식과 5형식 자료의 경우 다른 형식의 자료들과 비교하여 인식률이 상대적으로 낮음을 알 수 있다. 이것은 상대적으로

이 두 형식의 구조가 복잡하며, 특히 5형식 문자의 경우 영상이 자소간에 접촉되어 있는 경우가 많기 때문으로 보인다. 따라서 이미 분리된 4, 5형식의 자료들에 대하여 모음별로 SVM을 이용한 중분류를 수행하고, RDA로 자료의 인식률을 다시 구하였다.

표 3.2는 4형식 자료에 대한 인식률의 결과이다. 위 결과 4형식 글자 자료의 정분류율은 94.38%로 표 3.1의 4형식에 대한 정분류율 93.15%보다 높은 정분류율을 나타낸다.

표 3.2 4형식 자료의 인식률에 대한 정분류율 (%)

	형식							
	ㅏ	ㅐ	ㅑ	ㅓ	ㅔ	ㅕ	ㅣ	정분류율
RDA	94.75	90.48	89.23	95.15	84.09	94.41	94.48	94.38

표 3.3은 5형식 자료에 대한 인식률의 결과로써, 5형식 글자 자료의 정분류율은 92.84%로 표 3.1의 92.22%보다 약간 높은 정분류율을 나타냈다.

표 3.3 5형식 자료의 인식률에 대한 정분류율 (%)

	형식					
	ㅗ	ㅛ	ㅜ	ㅠ	ㅡ	정분류율
RDA	90.75	96.00	93.59	91.67	91.84	92.84

마지막으로 표 3.4는 4, 5형식 자료에 대한 중분류까지 시행한 후의 전체 형식에 대한 정분류율로서 4, 5형식에 대하여 중분류를 하기 전의 SVM과 RDA를 이용한 전체 정분류율은 95.08%였으나, 4, 5형식을 중분류한 후의 전체 정분류율은 95.50%로서 중분류를 한 후의 인식률이 향상되었음을 알 수 있다.

표 3.4 SVM+RDA 방법과 중분류까지 이용했을 때의 정분류율 비교

SVM (형식분류)과 RDA를 이용한 전체 정분류율	95.08%
중분류 (4, 5형식)까지 시행한 후 RDA를 이용한 전체 정분류율	95.50%

4. 결론 및 토의

비디오로부터 추출된 글자영상들은 해상도가 다양하고 글자의 변이가 많을 뿐만 아니라 빛의 영향 등으로 인식하기가 쉽지가 않다. 본 연구에서는 한글의 모든 클래스 (2,350자)를 목표로 하지 않고 실제로 가장 빈번하게 등장하는 609자의 글자클래스를 대상으로 수행함으로써 오히려 전체적인 인식성능을 향상시켰다. 문자인식기법은 대분류와 소분류를 기반으로 하였는데, 대분류는 지지도벡터기계 (SVM)를 통한 한글의 6형식 분류를 기본으로 하였고, 각 형식글자들에 대한 소분류는 정칙화판별분석 (RDA)을 기반으로 수행하여 인식 성능의 향상이 있었다. 또한, 인식률이 상대적으로 낮은 4형식과 5형식에 대해서는 각 형식에서 모음에 따라 다시 중분류를 수행하여 인식률이 향상됨을 알 수 있었다.

참고문헌

권재욱, 조성배, 김진형 (1992). 계층적 신경망을 이용한 다중 크기의 다중활자체 한글 문서인식. <한국정보과학회 논문지>, **19**, 69-79.

김영화, 남지호 (2009). 영상 잡음의 분산 추정에 관한 통계적 알고리즘 및 응용. <한국데이터정보과학회지>, **20**, 869-878.

박정민, 김경재, 한인구 (2005). 지지도벡터기계를 이용한 기업부도예측. <경영정보학연구>, **15**, 51-63.
오광식, 김상민, 이동로 (1997). 문자인식을 위한 로버스트 역전파 알고리즘. <한국데이터정보과학회지>, **8**, 163-171.
이성환 (1993). 다양한 활자체 및 크기를 갖는 대용량 한글의 고속 인식을 위한 최적 트리 분류기. <한국정보과학회 논문지>, **20**, 1083-1092.
이진수, 권오준, 방승양 (1996). 개선된 자소 인식 방법을 이용한 인쇄체 한글 인식. <한국정보과학회 논문지>, **23**, 841-851.
이판호, 장희돈, 남궁재찬 (1994). 동적 자소 분할과 신경망을 이용한 인쇄체 한글 문자 인식에 관한 연구. <한국통신학회 논문지>, **19**, 2133-2146.
장승익, 남윤석 (2004). 낱자 특징 기반 자소 인식을 이용한 한글 인식방법. <한국정보처리학회 춘계학술발표대회 논문집>, **11**, 351-354.
조완현 (1999). 문자인식의 기본 원리와 여러 가지 인식방법의 성능비교. <한국자료분석학회 논문지>, **1**, 1-14.
Cho, M. and Park, E. (2008). Analyzing customer management data by data mining: Case study on Churn prediction models for Insurance company in Korea. *Journal of the Korean Data & Information Science Society*, **19**, 1007-1018.
Friedman, J. H. (1988). Regularized discriminant analysis. *Journal of the American Statistical Association*, **84**, 165-175.
Huang, Z., Chen, H., Hsu, J. and Chen, H. (2004). Credit rating analysis with support vector machine and neural network. *Decision Support Systems*, **37**, 543-558.
Hwang, J. S., Lee, J. Y. and Kim, J. Y. (2006). A comparison study of multiclass SVM methods in microarray data. *Journal of the Korean Data & Information Science Society*, **17**, 311-324.
Joachims, T. (1998). Text categorization with support vector machines. *Proceedings of the European Conference on Machine Learning (ECML)*, **10**, 137-142.
Johnson, R. A. and Wichern, D. W. (1992). *Applied multivariate statistical analysis*, 3rd Ed., Springer, New Jersey.
Kim, K. and Lee, C. (2003). A study of data mining optimization model for credit evaluation. *Journal of the Korean Data & Information Science Society*, **14**, 825-836.
Kim, M. S. and Baek, J. S. (1999). Feature extraction and statistical pattern recognition for image data using wavelet decomposition. *The Korean Communications in Statistics*, **6**, 831-841.
Vapnik, V. N. (1995). *The nature of statistical learning theory*, Springer, New York.

Journal of the Korean Data & Information Science Society
2010, **21**(4), 689–697

한국데이터정보과학회지

Video character recognition improvement by support vector machines and regularized discriminant analysis[†]

Suyeol Lim[1] · Jangsun Baek[2] · Min Soo Kim[3]

[123]Chonnam National Univisity

Received 11 May 2010, revised 5 July 2010, accepted 8 July 2010

Abstract

In this study, we propose a new procedure for improving the character recognition of text area extracted from video images. The recognition of strings extracted from video, which are mixed with Hangul, English, numbers and special characters, etc., is more difficult than general character recognition because of various fonts and size, graphic forms of letters tilted image, disconnection, miscellaneous videos, tangency, characters of low definition, etc. We improved the recognition rate by taking commonly used letters and leaving out the barely used ones instead of recognizing all of the letters, and then using SVM and RDA character recognition methods. Our numerical results indicate that combining SVM and RDA performs better than other methods.

Keywords: Character recognition, classification, regularized discriminant analysis, support vector machines.

[†] This research was supported by Chonnam National University Research funds, 2007.
[1] Doctor of philosophy candidate, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea.
[2] Professor, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea.
[3] Corresponding author: Professor, Department of Statistics, Chonnam National University, Gwangju 500-757, Korea. E-mail: kimms@chonnam.ac.kr