TV 제어 메뉴의 다국적 언어 인식을 위한 특징 선정 기법

강근석 박현정 김호준
한동대학교 대학원 정보통신공학과
jackson80@lycos.co.kr hjpark79@empal.com hjkim@handong.edu

A Feature Selection Technique for Multi-lingual Character Recognition

Kang, Keun-Seok Park, Hyun-Jung Kim, Ho-Joon
Dept. of Information Technology, Handong Global University

요약

TV OSD(On Screen Display) 메뉴 자동검증 시스템에서 다국적 언어의 문자 인식은 표준패턴의 구조적 분석이 쉽지 않을 뿐만 아니라 학습패턴 집합의 규모와 특징의 수가 증가함으로 인하여 특징추출 및 인식 과정에서 방대한 계산량이 요구된다. 이에 본 연구에서는 학습 데이터에 포함되는 다량의 특징 집합으로부터 인식에 필요한 효과적인 특징을 선별함으로써 패턴 분류기의 효율성을 개선하기 위한 방법론을 고찰한다. 이를 위하여 수정된 형태의 Adaboost 기법을 제안하고 이를 적용한 실험 결과로부터 그 유용성을 고찰한다. 제안된 알고리즘은 초기의 특징 집합을 취약한 성능을 갖는 다수의 분류기(classifier)로서 고려하며, 이로부터 반복학습을 통하여 개선된 분류기를 점진적으로 선별해 나가게 된다. 학습의 원리는 주어진 학습패턴 집합에 기초하여 일종의 교사학습(supervised learning) 방식으로 이루어진다. 각 패턴에 할당된 가중치 값은 각 단계에서 산출되는 분류결과에 따라 적응적으로 수정되어 반복학습이 진행됨에 따라 점차 보완적 성능을 갖는 분류기를 선택할 수 있게 한다. 즉, 주어진 각 학습패턴에 대하여 초기에 균등한 가중치가 부여되며, 반복학습의 각 단계에서 적용되는 분류기의 출력을 분석하여 오분류된 패턴의 가중치 분포를 증가시켜 나간다. 본 연구에서는 실제 응용으로서 OSD 메뉴 검증 시스템을 대상으로 제안된 이론을 적용하고 그 타당성을 평가한다.

1. 서론

TV 등의 디스플레이 장치에는 사용자가 직접 환경을 설정할 수 있도록 하는 OSD(On Screen Display) 메뉴가 있다. TV 제품을 개발하는 과정에서 반드시 거쳐야 할 테스트 중 하나는 OSD 메뉴에 대한 정확성 검증 작업이다. 이것은 많은 시간이 소요되는 반복 작업이며 작업자의 피로나 집중력 저하에 따라 정확성 및 효율성이 떨어질 수 있는 작업이다. 이에 따라 효율적인 OSD 메뉴 테스트를 위한 자동화의 필요성이 절실히 요구된다. 요구되는 자동화 시스템은 개발자가 작성한 OSD 메뉴 스펙 시트가 입력되고 문자 인식 프로그램이 카메라를 통해 찍은 OSD 메뉴의 사진을 순차적으로 확인하면서 OSD 메뉴의 오류를 판정하여 결과를 파일로 출력하는 것이다. 이러한 시스템을 통해 OSD 메뉴의 결함을 비롯한 여러 통계 자료를 자동으로 출력함으로써 작업의 효율성을 극대화 할 수 있다.

TV OSD 메뉴 자동검증 시스템에서 다국적 언어의 문자 인식은 표준 패턴의 구조적 분석이 쉽지 않을 뿐만 아니라 학습 패턴 집합의 규모와 특징의 수가 증가함으로 인하여 특징 추출 및 인식 과정에서 방대한 계산량이 요구된다. 만약 (24x24) 크기의 표준 패턴에 대해 가능한 하아 특징(Haar-like feature)을 추출한다면 약 14만 개가 존재하게 될 것이고 이를 학습하는 과정에서 많은 계산량이 요구될 것이다[1][2]. 이에 본 연구에서는 학습 데이터에 포함되는 다량의 특징 집합으로부터 인식에 필요한 효과적인 특징을 선별함으로써 패턴 분류기의 효율성을 개선하기 위한 방법론을 고찰한다. 이를 위하여 수정된 형태의 Adaboost 기법을 제안하고 이를 적용한 실험 결과로부터 그 유용성을 고찰한다. 본 논문에서는 실제 응용으로서 OSD 메뉴 검증 시스템을 대상으로, 제안된 이론을 적용하고 그 타당성을 평가한다.

본 논문의 구성은 다음과 같다. 2장에서는 TV OSD 메뉴 자동검증 시스템의 구성에 대해서 살펴보고 특징에 의한 분류 방법 및 다국적 언어 패턴 인식의 한계에 대해서 고찰하게 된다. 3장에서는 본 논문에서 제안한 Adaboost에 의한 특징 선정 기법을 설명하며 4장에서 실험 결과를 통하여 제안된 이론의 타당성을 고찰한다.

2. TV OSD 메뉴 자동검증 시스템

가. 시스템 개요

TV OSD 메뉴 자동검증 시스템의 동작은 다음과 같다. 먼저 개발자가 작성한 OSD 메뉴의 스펙 시트가 입력되면 작업 스케줄링에 의해 메뉴의 구조적 분석이 이루어진다. 이후 TV에 나타난 OSD 메뉴를 카메라로 촬영하여 검증 시스템의 입력으로 들어가게 된다. 검증 시스템 내부의 문자인식 프로그램이, 입력으로 들어온 OSD 메뉴와 스펙 시트와의 일치 여부를 확인하여 OSD 메뉴의 오류 여부를 판정하고 그 결과를 화면 및 파일로 출력해낸다. OSD 메뉴 자동검증 시스템은 단순 작업에 의한 OSD 검증 업무의 부담을 줄이고 효율성을 극대화 할 수 있다는 점에서 실용적으로 큰 가치가 있다. 그림. ?은 TV OSD 메뉴 자동검증 시스템의 구조를 나타내고 있다.

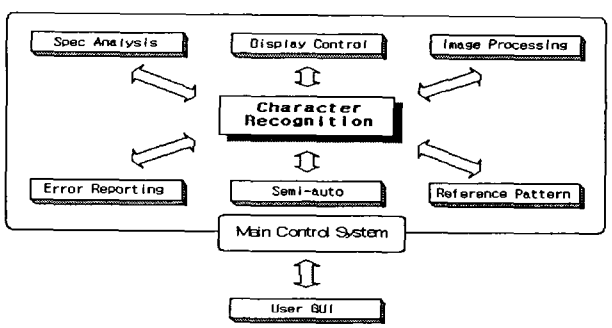


그림. 1 TV OSD 메뉴 자동검증 시스템의 구조

나. 특징(feature) 추출 및 패턴 분류(classification)

제안된 메뉴 자동검증 시스템에서 문자인식의 기본 틀은 특징 추출에 의한 방법이다. 특징(feature)이란 어떤 객체가 가지고 있는 객체 고유의 분별 가능한 측면(aspect), 양(quality) 혹은 특성(characteristic)이라고 정의될 수 있으며 인식 대상이 되는 객체들은 특징 공간상에서 특징 벡터가 형성하는 점들로 표현된다. 특징이 하나 이상의 수치 값을 가질 경우, 특징 벡터(feature vector)라고 하는 d-차원의 열벡터로 표현된다. 그리고 이러한 특징 벡터가 정의되는 d-차원의 공간을 특징 공간(feature space)라고 한다[3]. 이렇게 특징 벡터로 이루어진 글자의 여러가지 특징들, 예를 들어 구멍(hole), 개방 구간과 닫힘 구간, 뻗침(stretch) 등의 특징적인 정보들을 가지고 있어서 글자의 이미지에서 이들을 비교할 수 있다면 효과적인 인식이 가능할 것이다.

이러한 특징들에 의해 패턴 집합을 특정 클래스에 할당하는 것을 분류(classification)라고 하며 이렇게 패턴을 이루는 개별 특징의 선택은 패턴인식 알고리즘의 결정과 더불어 인식률에 결정적인 영향을 미친다. 그러므로 특징은 서로 다른 클래스 표본 간에 분별 가능한 성질을 되도록 많이 가지도록 선택되어야 한다[1][3].

다. OSD 메뉴의 다국적 언어 패턴 인식의 한계

다국적 언어로 된 OSD 메뉴의 문자인식은 표준패턴의 구조적 분석이 쉽지 않을 뿐만 아니라 학습패턴 집합의 규모와 특징의 수가 증가함으로 인하여 인식 과정에서 방대한 계산량이 요구된다. 다국적 언어의 패턴 인식에서 어떠한 특징을 선택할 것인가와 더불어 특징 벡터의 차원을 형성하는 특징의 개수 역시 인식률에 결정적인 영향을 미친다. 패턴이 가진 정보가 특징 벡터의 형태로 표현되므로 특징의 수가 적으면 적을수록 많은 경우와 비교하여 패턴 분류에 좋지 않은 영향을 미칠 것은 자명한 사실이다. 그러나 대부분의 패턴인식 시스템에서 패턴인식 성능은 차원을 올리면 어느 정도 인식률이 증가하다가 어느 지점에서부터 오히려 인식률이 감소된다. 패턴인식에서는 이런 현상을 "차원의 저주(curse of dimensionality)"라고 한다[2]. 특징 차원이 올라감에 따라 생기는 문제점들은 다음의 세 가지로 나타날 수 있는데, 1)잡음 특징들까지 포함된다. 2)패턴 분류기에 의한 학습과 인식 속도가 느려진다. 3)모델링에 필요한 학습 집합의 크기가 커진다[3][4]. 라는 점이다. 따라서 이를 최적화하기 위한 특징 선정의 방법이 필요하다.

특징을 최적하기 위한 방법은 일반적으로 주성분 분석법(PCA: principal components analysis), 선형 판별 분석법(LAD: linear discriminant analysis), 독립 요소 해석법(ICA: independent component analysis), 요인 분석법(FA: factor analysis) 등 고차원 특징 벡터를 저차원 특징 벡터로 축소하는 방법[4]이 주류를 이루고 있는데 본 논문에서는 효과적인 특징 선정 방법 중 하나인 Adaboost 알고리즘을 이용하여 다국적 언어의 특징을 선별하는 데에 적용하였다.

3. Adaboost를 이용한 학습 기법

Adaboost는 Freund와 Schapire가 1996년에 제안한 학습 알고리즘으로 취약한 성능을 갖는 다수의 분류기(weak classifier)들로부터 강력한 성능의 분류기(strong classifier) 구성하는 부스팅(boosting) 방법 중 하나이다.[2] 이는 역시 Freund와 Schapire에 의해 일반화된 버전으로 제안되었고 그 후 기계학습 분야(machine learning society)에서 활발한 연구가 이루어져 문자 인식(Schwenk, Bengio, 1997)[5] 및 얼굴 인식(Viola, Jones, 2001)[1] 등에 적용하여 우수한 성능을 나타내고 있다[6][7]. Adaboost 알고리즘은 분류의 성능이 뛰어나 여러 분야에 일반화시킬 수 있는 좋은 특성 때문에 효과적인 특징선정 방법으로 각광받고 있다. 제안된 알고리즘은 초기의 특징 집합을 빈약한 성능을 갖는 다수의 분류기(classifier)로서 고려하며, 이로부터 반복학습을 통하여 개선된 분류기를 점진적으로 선별해 나가게 된다[7].

학습의 원리는 주어진 학습패턴 집합에 기초하여 일종의 교사학습(supervised learning) 방식으로 이루어진다. 각 패턴에 할당된 가중치 값은 각 단계에서 산출되는 분류결과에 따라 적응적으로 수정되어 반복학습이 진행됨에 따라 점차 보완적 성능을 갖는 분류기를 선택할 수 있게 한다. 즉 주어진 각 학습

패턴에 대하여 초기에 균등한 가중치가 부여되며, 반복학습의 각 단계에서 적용되는 분류기의 출력을 분석하여 오분류된 패턴의 가중치 분포를 증가시켜 나간다. 본 연구에서 사용한 특징 선정을 위한 학습 방법은 다음과 같다.

- 주어진 학습패턴 집합을 $S = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_m, y_m)\}$ 라 한다. 여기서 $(\mathbf{x}_i, y_i) \in \mathcal{X} \times \{-1, +1\}$ 이며 m은 학습패턴의 수를 의미한다. 즉, 학습패턴 집합은 입력 특징과 기대 출력값의 쌍으로 이루어지며, 기대 출력값은 +1과 -1로 이루어지는 이진분류(binary classification)를 대상으로 한다. 주어진 학습패턴의 특징집합을 초기의 취약한 분류기 풀(pool) $\mathcal{H}$로 고려한다.
- 특정선정을 위한 반복 작업의 최초 단계에서 식 (1)과 같이 각 패턴에 대한 가중치값을 균일한 분포의 상수값이 되도록 초기화 한다.

$$D_1(i) = 1/m \tag{1}$$

- 학습패턴에서 T개의 취약한 분류기(weak classifier)를 선택하여 각각에 대하여 다음 과정을 반복하게 된다.

1. $\mathcal{H}$에 속한 각 분류기 h에 대하여 다음을 수행한다.
   a. 분류기 h에 의하여 패턴 공간 $\mathcal{X}$를 n개의 중첩이 없는(disjoint) 영역으로 나눈다.
   b. 가중치 분포 D_t에 의하여 다음 식을 산출한다.

$$W_l^j = P(\mathbf{x}_i \in X_j, y_i = l) = \sum_{i:\mathbf{x}_i \in X_j \wedge y_i = l} D_t(i) \tag{2}$$

   여기서 $l = \pm 1$의 값을 갖는다. 즉, W는 특정 영역에 속한 어떤 패턴이 주어진 기대 출력값을 가질 확률을 의미하며 식 (2)에 보인 바와 같이 해당 영역에서 동일 기대 출력값을 갖는 패턴에 대한 가중치 합으로 산출된다.
   c. 각 영역 X_j에 대한 출력값 h를 다음 식으로 산출한다.

$$\forall \mathbf{x} \in X_j, h(\mathbf{x}) = \frac{1}{2} \ln\left(\frac{W_{+1}^j + \varepsilon}{W_{-1}^j + \varepsilon}\right) \tag{3}$$

   여기서 ε는 작은 양의 상수값을 갖는다.
   d. 계산된 결과를 사용하여 정규화 인수 (normalizing factor)를 다음과 같이 계산한다.

$$Z = 2\sum_j \sqrt{W_{+1}^j W_{-1}^j} \tag{4}$$

2. 위에서 계산된 결과로부터 Z값을 최소화하는 분류기 h_t를 선택한다. 즉, 다음을 만족하는 h_t를 찾아내게 된다.

$$h_t = \arg\min_{h \in \mathcal{H}} Z$$
$$Z_t = \min_{h \in \mathcal{H}} Z \tag{5}$$

3. 새로운 가중치 분포를 다음 식을 적용하여 갱신한다. 즉 식 (6)에 의해 계산된 각 패턴별 가중치를 확률밀도함수를 적용하여 정규화하게 된다.

$$D_{t+1}(i) = D_t(i) \exp[-y_i h_t(\mathbf{x}_i)] \tag{6}$$

- 이상의 과정을 적용하여 최종적으로 강력한 성능을 갖는 분류기 H가 생성되는데, 그 출력값은 식 (7)로 결정된다. 식에서 b는 임계치값으로 기본값은 0을 갖는다.

$$H(\mathbf{x}) = sign\left[\sum_{t=1}^{T} h_t(\mathbf{x}) - b\right] \tag{7}$$

다시 말해서 총 T개의 선별된 분류기에 대한 출력값의 합에 임계치를 적용함으로써 최종 분류기의 결과가 생성된다. 또한 최종분류기의 신뢰도를 다음과 같이 정의할 수 있다.

$$Conf_H(\mathbf{x}) = \left|\sum_t h_t(\mathbf{x}) - b\right| \tag{8}$$

즉 각 분류 결과의 합에 대한 절대값의 크기는 분류결과를 확연하게 구분 짓는 척도가 되므로 일종의 신뢰도로서 고려될 수 있다.

본 연구에서 대상으로 하는 시스템을 위하여 위 알고리즘에서 식 (4)를 변형하여 특징값의 중요도 요소를 다음 식 (9)와 같이 정의 하였다.

$$R(i) = \frac{1}{N}\sqrt{\left(\sum_{i:\mathbf{x}_i \in X_j \wedge y_i = +1} D_t(i)\right) \bullet \left(\sum_{i:\mathbf{x}_i \in X_j \wedge y_i = -1} D_t(i)\right)} \tag{9}$$

즉 특징 i의 중요도 요소는 해당 특징에 의한 분류기가 주어진 m개의 패턴을 얼마나 정확한 영역으로 분류하는가에 대한 척도로서 평가하였다.

4. 실험 결과

본 논문에서 제안한 Adaboost 알고리즘을 실제 시스템에 적용하여 실제로 특징 선정에 효과적인지 여부를 알아보았다. 그림. 2는 OSD 메뉴 검증의 예이다. 그림에서 보인 바와 같이 시스템은 주어진 영상에서 나타나는 OSD 메뉴의 각 문자에 대하여 주어진 스펙시트의 내용과 일치여부를 판정하고 오류 여

부를 판단하여 결과를 출력하게 된다. 본 연구의 실험에서는 문자영역 분할 과정 이후의 데이터에 대한 인식문제를 대상으로 하였으며, 제안된 알고리즘에 따라 특징을 선별하고 특징별 중요도를 측정하여 최적화된 특징을 추출하도록 하였다.

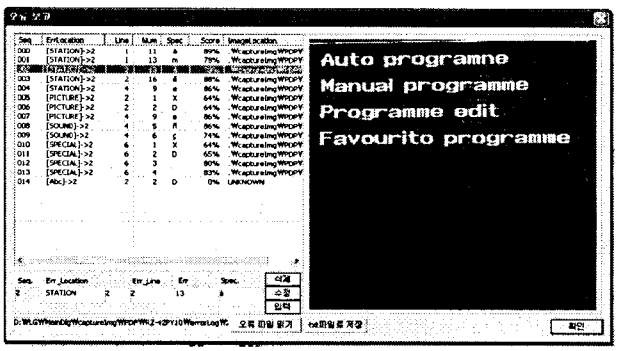
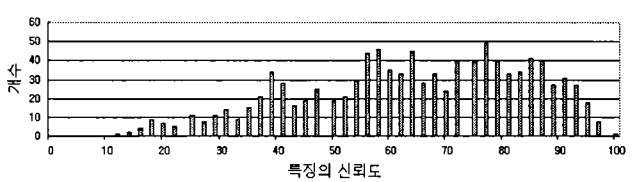
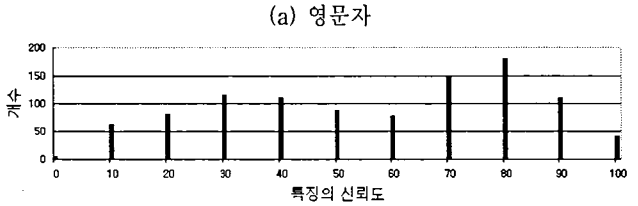


그림. 2 OSD 메뉴 검증 결과의 예

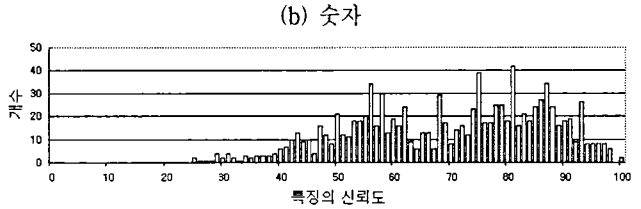
Adaboost 학습에 사용된 트레이닝 데이터 셋은 실제 OSD 자동 검증 시스템에서 쓰이는 각 언어의 폰트 셋으로부터 얻어졌다. 나라별로 영어, 프랑스어, 스페인어, 독일어, 이태리어 및 특수문자 등으로 이루어져 있으며 각각 (32x32)의 크기로 정규화하여 학습에 사용되었다. 그림. 3은 (32x32) 크기의 학습 데이터 셋에서 추출한 특징별 신뢰도를 나타낸 것이다. 각각은 신뢰도가 가장 높은 것을 100으로 가장 낮은 것을 0으로 하여 정규화 한 값이다.



(a) 영문자



(b) 숫자



(c) 5개 국어 통합

그림. 3 다국적 언어 학습 데이터의 특징별 신뢰도

그림. 4는 제안된 알고리즘으로 특징별 중요도에 가중치 요소를 부여하여 각각의 특징 지도를 추출한 결과를 나타낸 것이다. 그림에서 흰색 부분은 특징점의 상대적 중요도를 반영한다.

(a) (b) (c)

그림. 4 가중치 요소를 부여한 특징지도의 추출 결과: (a) 영문자 (b) 숫자 (c) 5개 국어 통합

각 특징별 신뢰도에 임계치를 적용하여 성능저하를 최소화하는 특징선별이 이루어 질 수 있다. 예를 들어, 위의 결과에서 임계치 40을 적용하면 영문자의 경우 18.8%, 숫자의 경우 29.7%, 그리고 총 8.3%의 특징수 감축 효과를 얻을 수 있다.

5. 결론

본 논문에서는 TV 제어 메뉴의 다국적 언어 인식을 위한 특징 선정 기법으로 Adaboost 알고리즘을 사용하였다. 이는 특징별 중요도 요소에 가중치를 부여하여 분류기의 성능저하 없이 특징수를 감축시킴으로써 시스템의 규모를 줄이고 수행시간을 개선할 수 있게 한다. 본 연구에서는 라틴 계열 5개 국어 문자 집합을 적용한 실제 OSD 메뉴를 대상으로 자동 검증 시스템을 구현하고 특징 분석 및 선별 알고리즘을 적용하였다. 향후 연구에서 성능 저하 요소와 특징수 선정 간의 이론적 분석 및 추가 실험이 요구되며 이러한 연구는 유사한 응용에서 시스템의 실시간 응답특성과 시스템 최적 설계를 위한 기반 연구가 될 것이다.

참고 문헌

[1] Paul Viloa, Michael Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", Computer Vision and Pattern Recognition, Vol.1, pp.511-518, 2001.

[2] 김형준, 정병희, 박성춘, 김회율, "AdaBoost의 학습 속도 향상을 위한 분산처리 시스템", 제 17회 신호처리합동학술대회 논문집, Vol.17, No.1, pp.245-248, 2004.

[3] 한학용 저, "패턴인식 개론: MATLAB 실습을 통한 입체적 학습", 한빛미디어.

[4] Zehang Sun, George Bebis, Ronal Miller, "Object detection using feature subset selection", Pattern Recognition, Vol.37, pp.2165-2176, 2004.

[5] Holger Schwenk, Yoshua Bengio, "AdaBoosting Neural Networks: Application to on-line Character Recognition", Int. Conference on Artificial Neural Networks, pp.967-972, 1997.

[6] R.E.Schapire and Y.Singer, "Improved Boosting Algorithm Using Confidence-rated Predictions", Machine Learning, Vol.37, pp.297-336, 1999.

[7] J.Amores, N.Sebe, P.Radeva, "Boosting the distance estimation Application to the K-Nearest Neighbor Classifier", Pattern Recognition Letters, Article in Press, Corrected Proof, 2005.