특징분포를 고려한 AdaBoost 약분류기의 성능 개선방법

이경주°, 최형일**, 김계영*
°*숭실대학교 컴퓨터학과, **숭실대학교 미디어학과
e-mail: {lkj0917, gykim11@ssu.ac.kr, hic@ssu.ac.kr

A Method to Improve the Performance of Weak Classifier in AdaBoost by Considering Features Distribution

Gyung-ju Lee°, Hyung-il Choi**, Gye-young Kim*
°*School. of Computer Science and Engineering, Soongsil University
**The Global School. of Media, Soongsil University

● 요 약 ●

본 논문에서는 AdaBoost알고리즘에서 약분류기(Weak Classifier)의 성능을 개선하기 위한 임계값 설정 방법을 제안한다. 일반적으로 약분류기에 사용되는 임계값은 특징들의 평균값을 많이 사용하지만 이는 특징들의 분포가 고려되지 않았기 때문에 분별력이 많이 떨어진다. 그러므로 각 특징들의 분포를 고려한 약분류기의 임계값 설정방법을 제안한다. 이는 얼굴에 대한 간단한 학습 및 테스트를 통하여 기존 방법에 비하여 더 나은 성능을 보임을 입증한다.

키워드: 아다부스트(AdaBoost), 약분류기(Weak Classifier), 임계값(Threshold)

I. 서 론

얼굴인식, 음성인식 등 객체를 인식하기위하여 많은 연구가 진행되고 있다. 그중에서 Freund와 Schapire가 AdaBoost알고리즘은 여러 개의 약분류기(Weak Classifier)를 선형결합하여 성능이 좋은 강분류기(Strong Classifier)를 생성하는 부스팅 기법 중 하나이다[1].

AdaBoost알고리즘은 실시간에서 적용 가능하도록 Cascade 구조를 도입하여 높은 검출 성능과 속도를 보이도록 설계되었다[2]. 약분류기는 positive와 negative데이터의 특징값의 분포를 분석하여 두 분포를 가장 잘 분류할 수 있는 특징을 선택하여 약분류기를 만든다. 약분류기가 하나 만들어지면 분류하지 못한 데이터는 분류된 데이터에 비해 데이터에 대한 가중치를 높이고 반대로 분류가 된 데이터에 대해서는 가중치를 낮게주어 다음 약분류기를 선택할 때에 오분류된 데이터에 대하여 집중적으로 학습할 수 있도록 한다. 이 약분류기는 단일 임계값을 이용한 분류방법으로 positive와 negative의 특징 분포에서 최적의 임계값을 찾는 것으로 이것은 학습 성능에 영향을 미치게 된다. 이 두 분포를 분류하는 임계값을 찾는 방법으로 일반적으로 평균값을 많이 이용한다. 그러나 이 경우는 특징 공간을 너무 단순하게 해석한 경우이고, 실제 복잡한 데이터를 분류하기 위한 임계값으로 충분하지 않다. 이런 경우 캐스케이드의 뒷단계로 갈수록 분류 난이도가 높아지는데 비해 분별력은 훨씬 더 약해져서 부스팅의 효과가 없을 정도가 된다. 그렇기 때문에 특징공간을 분석하여 분별력이 좋은 임계값을 설정하는 것은 중요한 문제이다.

본 논문에서는 각 특징의 분포를 고려하여 특징공간의 평균에 분포에 대한 분산을 가중치로 줘서 더 정확한 임계값을 설정하고자 한다. 이에 대한 성능을 평가하기 위하여 정면 얼굴 2000개의 샘플과 얼굴이 포함되지 않은 비얼굴 4000개의 샘플을 훈련데이터로 사용하고, 얼굴 500개와 비얼굴 1000개의 샘플을 테스트데이터로 사용하였다. 제안한 방법은 훈련에 참가하지 않은 테스트 데이터를 분류함에 있어서 더 좋은 분류결과를 보였다.

II. 관련 연구

1. AdaBoost

AdaBoost알고리즘은 여러 약분류기들을 결합하여 성능이 좋은 강분류기를 만드는 대표적인 지도학습방법중의 하나이다. 학습 과정은 반복적인 계산으로 학습이 이루어지는데 특징선택과 관측값에 해당하는 에러율 산출 그리고 학습데이터에 대한 가중치 수정 단계로 이루어져 있다. 학습을 반복할수록 분류 능력이 높아지는데 이는 잘못 분류된 데이터에대한 가중치는 증가시키고 바르게 분류된 데이터의 가중치는 감소시키는 과정을 반복하면서 좋은 분류율을 보이는 약분류기가 결합하여 강분류기로 만들어지기 때문이다.

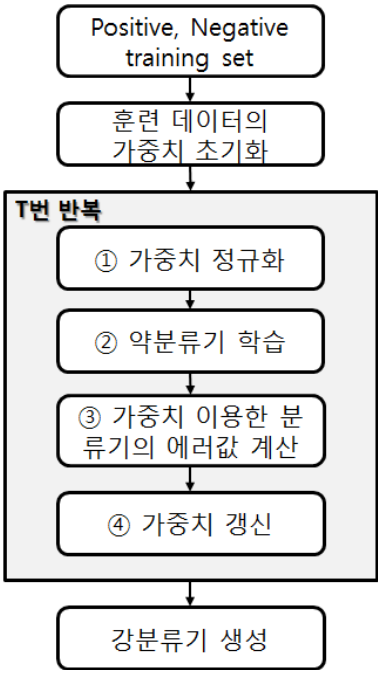

그림 1. AdaBoost 순서도
Fig. 1. AdaBoost Flow Chart

약분류기는 다음 식1과 같이 정의된다.

$$h_j = \begin{cases} 1 & p_j f_f(x) < p_j \theta_j \\ 0 & otherwise \end{cases} \quad (1)$$

여기서 f는 훈련데이터에 대한 j번째 특징값을 나타내는 함수이고, θ는 j번재 특징의 임계값이며, p는 부호를 결정하는 패러티 값이다. 이때에 임계값을 각 positive와 negative 집합의 평균의 평균으로 많이 사용한다.

다음 그림 1은 Viola와 jones가 제안한 AdaBoost의 훈련과정을 나타낸다. 약분류기는 데이터에대한 가중치를 이용하여 학습한다. 입력받은 훈련 데이터에 대하여 잘못 분류하면 가중치를 증가시키고 옳게 분류하면 가중치를 감소시킨다. 이때 가장 낮은 에러율을 나타내는 특징을 약분류기로 선택한다. 하나의 특징이 선택되면 다시 훈련 데이터를 입력으로 받아들이고 이와같은 과정을 T번 반복하여 가중치를 변환하고 에러율이 가장 낮은 약분류기를 반복 횟수만큼 선택한다. 이후 최종적으로 선택된 약분류기들을 선형적인 결합으로 강분류기를 생성하는 것이다.

2. Cascade

객체 검출 시에 수행 속도를 향상시키기 위해 그림 1과 같이 AdaBoost 분류를 여러 단계로 만들어 적용한다.

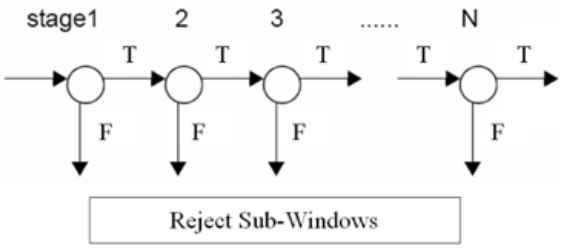

그림 2. Cascade 구조
Fig. 2. Cascade Structure

훈련에서의 cascade방법의 원리는 positive데이터는 그대로 유지하면서 대부분의 negative데이터를 초기 단계에서 걸러냄으로써 나머지 단계에서는 좀 더 복잡한 nagative데이터에 집중할 수 있도록 한 것이고, 검출에서의 cascade는 서브윈도우에서 초기단계에 negative로 분류를 하면 다음단계로 진행하지 않고 다음 서브윈도우를 수행하기 때문에 검출 수행속도도 빨라진다.

III. 개선된 Weak Classifier

일반적으로 약분류기에서의 임계값은 각 특징의 분산을 고려하지 않은 각 특징의 평균의 평균값으로 취한다. 따라서 보다 더 정확한 임계값 설정을 위하여 각 특징의 분산을 고려한 임계값 설정 방법을 제안한다. 방법은 식 1과 같다.

$$threshold = \frac{\mu_p \sigma_n + \mu_n \sigma_p}{\sigma_p + \sigma_n} \quad (2)$$

여기서 μ와 σ는 각각 평균과 표준편차를 의미하고, p와 n은 positive와 negative를 의미한다. σ는 식 2에 의해 계산된다. 그러나 이 수식대로 계산을 하면 먼저 특징 벡터의 평균 μ를 계산하고 또다시 분산을 계산하기 위해 다시 특징을 읽어야 하므로 식 3을 적용하여 분산을 빠르게 계산한다.

$$\sigma^2 = \frac{1}{N}\sum_{i=0}^{N-1}(x_i - \mu)^2 \quad (3)$$

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma^2 = \left(\frac{1}{N}\sum_{i=0}^{N-1} x_i^2\right) - \left(\frac{1}{N}\sum_{i=0}^{N-1} x_i\right)^2 \quad (4)$$

결과적으로 서로 다른 집합의 표준편차를 가중치로 사용하여 더 정확한 분별력을 가질 수 있도록 결정경계를 만들어 주는 것이다.

Ⅳ. 실험 결과

본 논문에서는 분류기의 성능을 실험하기 위하여 CBCL Frontal face database[3]에서 24x24 크기의 정면 얼굴 1000장, 비얼굴 2000장을 훈련 데이터로 사용하였다[그림 3].

그림 3. 학습 데이터. (위: 얼굴, 아래: 비얼굴)
Fig. 3. Training Data (top: face, bottom: non-face)

Cascade 학습을 위한 각 설정값은 false positive rate = 0.5, Detection Rate = 0.995, 종료조건은 전체 false positive rate = 10-7으로 설정하였다. 또한 학습을 위한 특징은 Haar-like Feature를 사용하였다[그림 4].

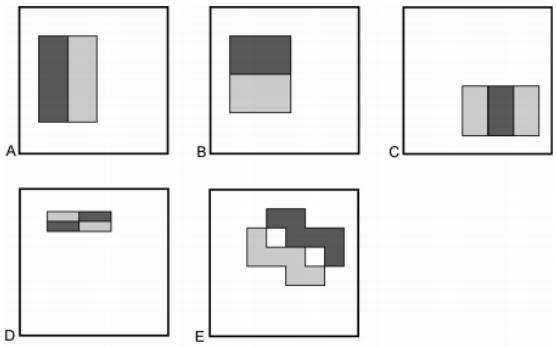

그림 4. 훈련에 사용된 Haar-like feature.
Fig. 4. Haar-like features which used in training.

훈련 결과를 확인하기 위하여 훈련에 참여하지 않은 얼굴 500장, 비얼굴 500장으로 분류한 결과를 그림 4와 같이 ROC 커브로 나타내었다. 그림 4에서 보는바와 같이 제안한 방법이 더 높은 검출율과 낮은 오검출율을 보이는 것을 확인할 수 있다.

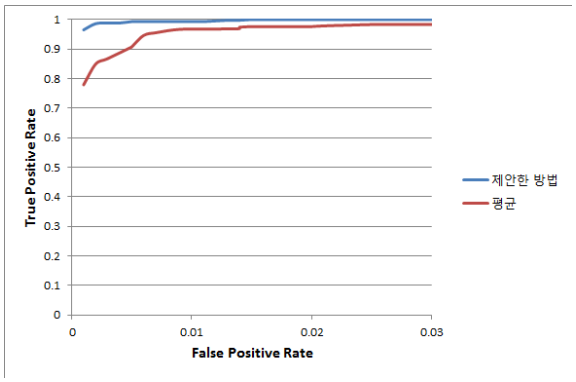

그림 5. 테스트 결과 ROC 커브
Fig. 5. ROC Curve of Test Result

Ⅴ. 결 론

학습기의 분류율을 간단히 측정하기 위하여 적은 데이터를 가지고 측정하였지만 테스트 결과 일반적으로 각 특징의 평균을 이용한 약분류기의 임계값 보다는 특징들의 분산을 고려한 제안된 방법이 기존 방법에 비해 높은 검출율과 낮은 오검출율을 보여 더 효율적인 것을 확인할 수 있었다.

참고문헌

[1] Y. Freund and R. Schapire. “A short introduction to boosting,” Journal of Japanese Society for Artificial Intelligence, 14(5): 771-780, 1999.

[2] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of simple features,” In Proceedings, IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[3] http://cbcl.mit.edu/software-datasets/FaceData2.html