JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 25 No. 7, pp. 47-55, July 2020
https://doi.org/10.9708/jksci.2020.25.07.047

Multi-scale face detector using anchor free method

Dong-Ryeol Lee*, Yoon Kim*

*Student, School of Computer Science and Engineering, Kangwon National University, Chuncheon, Korea
*Professor, Dept. of Computer Science and Engineering, Kangwon National University, Chuncheon, Korea

[Abstract]

In this paper, we propose one stage multi-scale face detector based Fully Convolution Network using anchor free method. Recently almost all state-of-the-art face detectors which predict location of faces using anchor-based methods rely on pre-defined anchor boxes. However this face detectors need to hyper-parameters and additional computation in training. The key idea of the proposed method is to eliminate hyper-parameters and additional computation using anchor free method. To do this, we apply two ideas. First, by eliminating the pre-defined set of anchor boxes, we avoid the additional computation and hyper-parameters related to anchor boxes. Second, our detector predicts location of faces using multi-feature maps to reduce foreground/background imbalance issue. Through Quantitative evaluation, the performance of the proposed method is evaluated and analyzed. Experimental results on the FDDB dataset demonstrate the effective of our proposed method.

▸ **Key words**: Face detection, Anchor free method, Multi-scale detection, Deep learning, Feature pyramid learning

[요 약]

본 논문에서는 앵커 프리 방법을 이용한 FCN(Fully Convolutional Network)기반의 1단계 다중 크기 얼굴 검출기를 제안한다. 최근 대부분의 연구들은 사전 정의된 앵커를 사용하여 얼굴이 있을 만한 위치를 예측한다. 그러나 사전 정의 앵커를 이용함으로써 학습 시 하이퍼 파라미터의 설정과 추가적인 계산이 필요하다. 제안하는 방법의 핵심 아이디어는 앵커 프리 방법을 사용하여 하이퍼 파라미터를 없애고 여러 개의 특징 맵을 사용함으로써 클래스 내 불균형 문제를 완화하는 것이다. 이 방법들은 다음과 같은 효과가 있다. 첫째로 사전정의 앵커를 없앰으로써 앵커와 관련된 하이퍼 파라미터와 추가적인 계산을 피한다. 둘째로 클래스 내 불균형을 완화하기 위해 여러 개의 특징 맵으로부터 얼굴을 예측한다. 정량적 평가를 통해 제안하는 방법에 따른 검출 성능을 평가 및 분석한다. FDDB(Face Detection Dataset & Benchmark) 데이터 셋의 실험 결과에서 제안하는 방법이 효과가 있음을 증명했다.

▸ **주제어**: 얼굴 검출, 앵커 프리 방법, 다중 크기 검출, 딥 러닝, 특징 피라미드 학습

• First Author: Dong-Ryeol Lee, Corresponding Author: Yoon Kim
*Dong-Ryeol Lee (ryol8888@kangwon.ac.kr), School of Computer Science and Engineering, Kangwon National University
*Yoon Kim (yooni@kangwon.ac.kr), Dept. of Computer Science and Engineering, Kangwon National University
• Received: 2020. 06. 11, Revised: 2020. 07. 02, Accepted: 2020. 07. 03.

http://www.ksci.re.kr pISSN:1598-849X | eISSN:2383-9945

I. Introduction

얼굴 분석 작업은 생체 보안, 정보 보안, 접근 관리, 신분 확인 등 다양한 보안 응용 분야에서 사용되고 있다. 얼굴은 개인의 특징이 뚜렷하게 나타내는 특징으로 이 특성은 다른 생체 인식과는 달리 신체 접촉이나 행동을 사용자에게 요구하지 않고 쉽게 얻을 수 있다.

일반적으로 얼굴 분석 작업을 위해서 얼굴 검출이 선행되어야 한다. 얼굴 검출(Face Detection)은 얼굴 인식, 추적, 애니메이션 및 표현 분석 등과 같은 많은 얼굴 분석 작업에 가장 중요한 전 처리 단계이다. 얼굴 검출의 정확성은 얼굴 분석 작업에 직접적인 영향을 미치므로 매우 중요하다. 얼굴 검출의 정확성을 높이기 위하여 오랜 기간 컴퓨터 비전과 패턴 인식 분야에서 다양한 방법들이 연구되어왔다. 얼굴 검출은 복잡한 환경에서 다양한 크기와 비율을 가진 얼굴의 특징을 표현하는 특징 표현(Feature Representation)문제와 얼굴이 있을 만한 위치를 제안하는 지역 제안(Region Proposal)문제가 있다. 이를 해결하는 방식으로는 두 문제를 동시에 진행하는 1단계(one-stage) 검출기와 순서대로 진행하는 2단계(two-stage) 검출기로 나뉜다. 1단계 검출기는 물체의 특징 표현문제와 물체의 지역을 제안하는 문제를 한 네트워크에서 동시에 해결한다. 2단계 검출기는 물체가 있을만한 지역을 제안하는 네트워크와 제안한 지역에서 물체의 특징을 추출하는 네트워크로 두 문제를 나눠서 해결한다.

얼굴 검출에 대한 초기 작업[1,2,3,4,26]의 특징 표현 문제는 Haar[3], Control point set[26]과 같은 수작업 특징에 의존했다. 지역 제안 문제를 해결하기 위해 Arbelaez은 색상 대비, 특징의 밀도 및 돌출과 같은 시각적인 정보를 이용하였다[39]. 또한 선택적 검색(Selective Search)은 슈퍼 픽셀 병합을 기반으로 지역 제안을 하였다[40]. 이후 물체 검출(Object Detection)분야에서 Convolution Neural Network(CNN) 모델[5,6,7,8,9,10,41]을 이용하여 객체가 있을만한 위치를 제안하는 방법이 높은 성능을 얻음에 따라 연구가 활발해졌다. 최근 Fully Convolutional Network(FCN)[35]가 crowd counting [32,33,34], key-point detection[29,30], Object detection [9,21]과 같은 분야에서 의미 있는 성능을 달성했다. 얼굴 검출 분야도 FCN을 활용하여 여러 연구[22,23,24,25]들이 의미 있는 성능을 보여준다. 이 연구들은 얼굴 검출에 영감을 주었고 얼굴 검출 분야의 많은 연구에서 얼굴의 특징을 추출하기 위해 CNN을 적용하였다. 복잡한 환경에서 다양한 크기와 종횡비를 가진 얼굴의 특징을 잘 표현하는 CNN을 만들기 위해 다중 크기 특징 학습 방식, 지역 특징 인코딩, 상황 정보 학습 등 많은 방식을 사용한다.

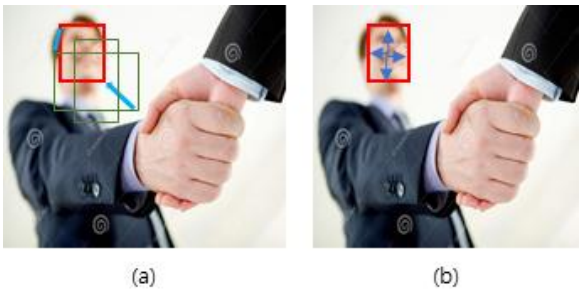

Fig. 1. (a) Anchor based method (b) Anchor free method. red is ground truth, green is pre-defined anchor, arrow is offset.

다중 크기 특징 학습 방식은 서로 다른 크기 정보를 가진 특징 맵 또는 이미지를 학습한다. 예를 들면 깊이가 얕은 CNN 망에서 나온 특징 맵은 더 높은 해상도와 더 작은 수용 필드(Reception field)를 갖는다. 얕은CNN의 특징 맵은 작은 물체를 검출하는데 더 적합하고 깊은 망에서 나온 특징 맵은 더 큰 물체를 검출하는데 더 적합하다.

지역 특징 인코딩은 2단계 검출기[5]에서 특징 맵을 고정 길이 벡터로 추출하는 중요한 단계이다. R-CNN은 전체 이미지에서 잘라낸 영역과 잘라낸 영역을 이중선 보간법(bilinear interpolation)을 통해 고정 크기 패치로 크기를 변환한 다음 CNN을 사용한다. 이 방법은 고해상도 영역 기능을 인코딩할 수 있다는 장점이 있지만 계산량이 많다.

상황 정보 학습은 상황 정보(Contextual information)를 이용하여 물체를 감지한다. 물체는 특정한 환경에서 나타나는 경향이 있으며 특히, 얼굴의 경우 사람의 몸, 머리카락 등이 그 예다. P. Hu은 상황 정보가 얼굴 검출 성능 향상에 도움이 된다는 것을 증명했고[27]. Li, J.는 Feature Enhance Module을 이용하여 더 높은 얼굴 검출 성능을 얻었다[18].

CNN을 적용하여 얼굴의 특징을 추출한 대부분의 연구에서 지역 제안은 일반적으로 앵커 기반 방법(Anchor-base method)와 앵커 프리 방법(Anchor free method)같은 방법을 사용한다.

앵커 기반 방법은 Fig 1 (a)과 같이 CNN에서 얻은 특징 맵에 사전 정의된 앵커 박스와 Ground Truth(GT)를 비교하여 얼굴의 위치와 크기를 얻는 방법이다. 사전 정의된 앵커 박스는 크기, 종횡비에 따라 앵커의 개수가 정해진다. 얼굴의 위치는 GT와 가장 비슷하게 위치한 사전 정의된 앵커 박스를 매칭해서 추정한다. 얼굴의 크기는 사전 정의

된 앵커 박스의 크기를 GT의 얼굴 크기로 회귀하여 추정한다. CNN을 적용한 대부분의 연구[12,13,14,15,16,17,18,19,20,42]는 앵커 기반 방법을 사용하여 높은 검출율과 정확도를 얻는다. 하지만 다음과 같은 단점이 존재한다. 1) [5]에서 볼 수 있듯이 얼굴 검출의 성능은 앵커 박스의 크기와 종횡비에 민감하다. 예를 들어 RetinaNet[9]에서 앵커 박스의 크기와 종횡비를 수정하면 COCO 벤치마크 데이터에서 AP(Average Precision) 성능이 최대 4%까지 영향을 받는다. 2) 높은 검출율을 위하여 앵커 박스는 이미지에 최대한 조밀하게 위치하여야 한다. 그러나 너무 촘촘하면 훈련 중 음성 샘플의 수가 많아지므로, 클래스 간 불균형 (Class imbalance) 문제에 직면한다. 3) 사전 정의된 앵커 박스의 개수만큼 추가적인 계산이 요구된다.

앵커 프리 방법은 Fig 1 (b)와 같이 특징 맵의 각 픽셀 좌표에서 직접적으로 얼굴 크기를 얻는다. 얼굴의 크기는 주어진 특징 맵의 좌표로부터 GT의 얼굴 크기에 맞게 예측한다[22,23,24,25]. 이 방법은 앵커 박스를 사용하지 않음으로써 사람이 조정해야하는 앵커 박스의 크기와 종횡비 같은 하이퍼 파라미터가 필요 없다. 이로 인하여 학습과 테스트 시간이 단축되고 필요 메모리가 감소하며 음성, 양성 샘플의 수로 인한 클래스 간 불균형 문제가 완화된다. 하지만 이 방법은 한 개의 특징 맵에서 얼굴 검출을 수행하기 때문에 클래스 내 불균형 문제를 일으킨다. 예를 들면, 크기가 100 × 100인 큰 얼굴과 25 × 25인 작은 얼굴은 같은 얼굴로 분류되지만 픽셀 수는 16배이다.

본 논문에서는 위의 문제에 중점을 두어 학습에 민감한 하이퍼 파라미터를 없애고 클래스 내 불균형 문제를 완화시키기 위하여 앵커 프리 방법을 이용한 FCN기반 네트워크를 제안한다. 제안하는 네트워크는 특징 피라미드 방식으로 이루어진 FCN(Fully Convolutional Network)이며, 병합 모듈을 통하여 네 가지 특징 맵을 얻고 검출 모듈을 이용하여 얼굴을 검출한다.

제안하는 방법을 사용함으로써 아래와 같은 장점을 얻을 수 있다. 1) 앵커 프리 방식을 사용함으로써 앵커 박스와 관련된 하이퍼 파라미터와 앵커로 인한 추가적인 계산을 없앤다. 이 방식은 학습의 안정성과 속도를 높인다. 2) 제안된 알고리즘은 1단계 방식뿐만 아니라 2단계 방식에서 지역 제안 알고리즘으로도 사용이 가능하다. 3) 여러 개의 특징 맵을 사용함으로써 클래스 내 불균형 문제를 완화했다. 4) 최근 연구에서 사용하고 있는 FDDB 데이터 셋에서 0.944, 0.723을 달성한다.

II. Preliminaries

1. Related work

1.1 Multi-scale feature learning

현재 다중 크기 특징 학습에는 이미지 피라미드, 예측 피라미드, 통합된 특징, 특징 피라미드와 같은 4가지 분류가 있다. Fig. 2는 다중 크기 특징 학습의 4가지 종류를 설명한다.

이미지 피라미드는 입력 이미지의 크기를 다양하게 조정하여 특정 범위의 크기를 담당하는 검출기를 학습시키는 방법이다. 테스트 시, 각 검출기에서 나온 결과는 합쳐진다. Zhang은 다양한 크기의 얼굴을 검출하기 위해 이미지 피라미드 방법을 사용한 2단계 얼굴 검출기를 제안하였다[36]. 그리고 얼굴 검출과 얼굴 정렬을 동시에 진행하는 다중 학습 방법을 제시했다. 이 방법은 여러 개의 이미지를 사용함으로써 다른 방법에 비해 계산 량이 높기에 학습과 테스트 속도가 떨어진다.

예측 피라미드는 망을 통과하며 나오는 여러 특징 맵에서 예측을 수행한다. 이 방법은 다양한 크기를 가진 얼굴을 검출하기 수월한 형태이지만 추정할 때 하나의 특징 맵에서만 예측해야하기 때문에 앞, 뒤 특징 맵에서 얻은 충분한 정보를 활용할 수 없다.

통합된 특징은 여러 레이어에서 나온 특징 맵을 결합하고 새로 구성된 단일 특징 맵을 구성하여 최종 예측을 수행한다. 하지만 이 경우 얼굴 검출 분야[22,25]에서 클래스 내 불균형 문제가 발생한다.

특징 피라미드는 통합된 특징과 예측 피라미드의 장점을 결합한 방법이다. 여러 레이어에서 나온 특징 맵을 하향식으로 통합한다. 일반적으로 특징 맵들을 요소별 곱셈 또는 덧셈으로 결합한다. 결합으로 얻은 특징 맵 셋은 다양한 크기의 얼굴을 검출하기 용이하다.

1.2 Anchor-based method

앵커 기반 방법은 얼굴의 크기와 위치를 검출하기 위해 각 좌표의 오프셋 회귀와 GT와 가장 가까운 박스를 판단하는 지역제안 방식이다. 이 방법은 전통적인 슬라이딩 윈도우 방식으로부터 아이디어를 얻었다. 슬라이딩 윈도우의 제안으로 영역을 얻은 Faster R-CNN[5]과는 달리 앵커 기반 방법은 CNN의 특징 맵을 사용하여 반복되는 특징 맵 계산을 줄이고 학습 및 테스트의 시간을 단축한다. 해당 방법은 YOLO[6], SSH[19], S3FD[20]과 같은 연구에 의해 널리 쓰이며 높은 검출율과 성능을 보인다. YOLO는 그리드(Grid)라 불리는 특정 크기로 특징 맵을 나눠서 그

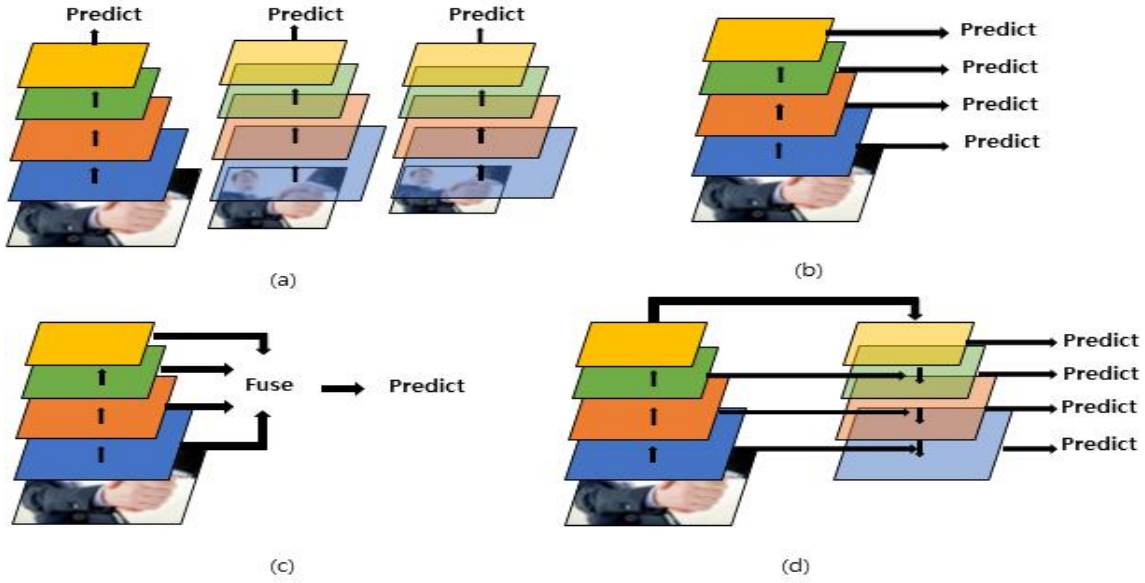

Fig. 2. Four paradigms for multi-scale feature learning. (a) Image pyramid (b) Prediction pyramid (c)Integrated features (d)Feature pyramid

리드 당 한 개의 앵커 박스를 추정한다. 이는 빠른 속도와 높은 검출율을 보이지만 한 그리드에 여러 객체가 존재할 경우 한 개의 객체만 검출하는 한계를 보인다. SSH는 앵커의 종횡비를 1로 고정하여 앵커 수를 줄여 학습과 테스트의 속도를 높인다[19]. 하지만 앵커 기반 방법은 크기와 종횡비같이 앵커의 모양을 설명하는 파라미터 외에도 각 앵커를 양성, 음성 샘플로 라벨링하기 위한 하이퍼 파라미터를 요구 한다.

1.3 Anchor free method

앵커 프리 방법은 위에 서술한 바와 같이 각각의 특징 맵 픽셀에서 얼굴과 얼굴의 크기를 추정한다. 앵커 기반 방법처럼 크기와 종횡비에 관한 파라미터가 존재하지 않기에 복잡한 계산, 후처리 및 클래스 간 불균형 문제를 피한다. 앵커 기반 방법과 달리 일반적으로 얼굴의 네 방향 좌표를 추정하는 방법을 사용한다. L. Huang는 전처리 시 물체 주위로 이미지를 240 x 240 크기로 자르고 다양한 크기로 변환한다. 변환한 이미지들을 이용하여 이미지 피라미드 방식으로 특징 맵을 추출한다[23]. 추출한 특징 맵으로부터 물체 크기에 대한 상하좌우 좌표를 L2 손실 함수를 이용하여 회귀시킨다. 하지만 이미지 피라미드 방식 특성상 추가적인 계산이 필요하고 L2 손실 함수의 경우 각각의 좌표를 독립적으로 회귀시키므로 좌표간의 연결성이 떨어진다. J. Yu은 L2 손실 함수 대신 4차원의 좌표를 공동으로 훈련시키는 IOU(Intersection Over Union) 손실 함수를 제안한다[24]. 또한 [23]과는 달리 학습 시 이미지 파리미드 방식을 사용하지 않는다. 하지만 단일 특징 맵에서 지역 제안을 함으로써 충분한 상황정보를 갖지 못한다.

1.4 Fully convolutional network

기존의 이미지 분류 모델들은 기본적으로 내부 구조와 관계없이 출력층이 Fully-connected(FC) 레이어로 구성되었다. 기존의 방법은 다음과 같은 단점이 있다. 1) FC 레이어로 인하여 이미지의 위치 정보가 사라진다. 2) FC 레이어의 가중치 개수가 고정되어있기 때문에 입력 이미지의 크기 역시 고정되어야 한다. 반면 Fully Convolutional Network(FCN)는 출력층을 FC 레이어와 다르게 컨볼루션 레이어를 사용하여 이미지의 위치 정보를 계속 가지면서 입력 이미지의 크기를 고정시킬 필요가 없어진다. FCN[35]은 의미론적 영상분할 분야에서 제안된 모델 네트워크 구조이며 물체 검출 분야에서도 의미 있는 성능을 보여준다. SSD[7]은 VGG16기반의 FCN 모델 네트워크로 빠른 속도와 최첨단 성능을 얻었다. 최근 Tian, Z. et al은 앵커 프리 방법을 이용한 FCN기반의 최첨단 성능을 얻었다[21].

III. The Proposed Scheme

1. Preprocessing

대부분의 앵커 기반 방법은 $anno_i = (x_0^i, y_0^i, x_1^i, y_1^i, c^i)$ 와 같은 GT들을 가지고

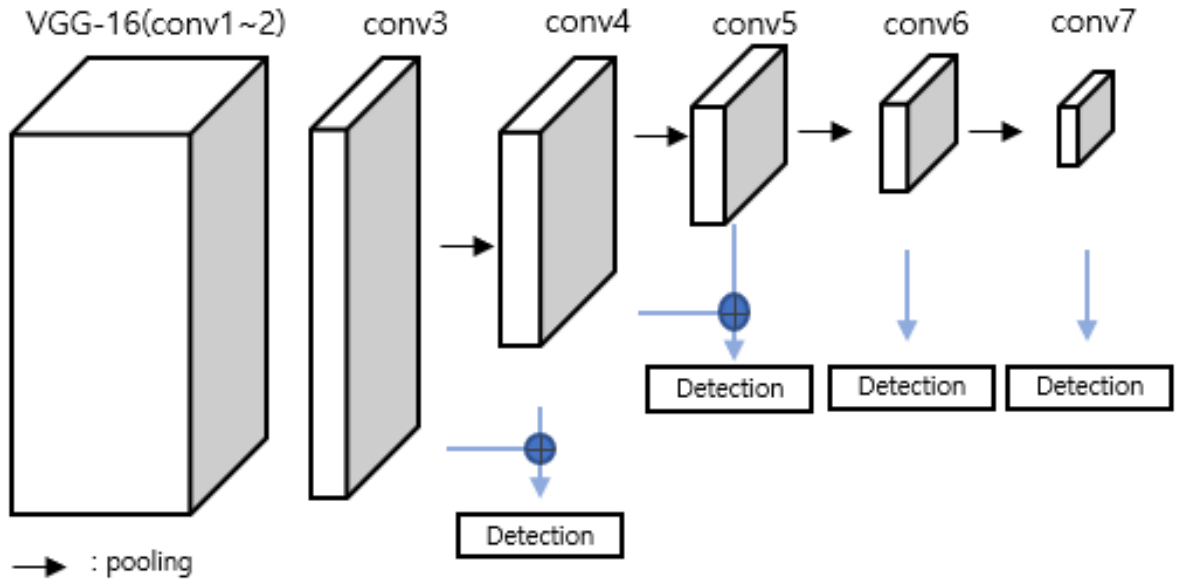

Fig. 3. Proposed network architecture

학습을 진행한다. i는 이미지의 번호, x, y는 얼굴의 너비, 높이 좌표이고 (x_0,y_0)와 (x_1,y_1)은 얼굴의 좌상좌표와 우하좌표이다. C는 객체의 종류를 의미하고 이 데이터셋에서 객체는 얼굴이다. 얼굴 검출의 특성상 얼굴의 개수가 배경에 비해 현저히 적기 때문에 클래스 간 불균형 문제가 발생한다. 본 논문에서는 얼굴 안의 모든 픽셀을 얼굴로 분류하여 클래스 불균형 문제를 완화한다. 얼굴 B의 크기는 각 특징 맵의 픽셀 위치 (x,y)에서 $B_{x,y}=(l,t,r,b)$로 표현하고 각 요소는 각각 얼굴 중앙으로부터 왼쪽, 위쪽, 오른쪽, 아래쪽의 길이를 나타낸다. 구하는 방법은 아래 수식 (1)과 같다.

$$\begin{aligned} l &= loc_{x,y} - x_0, \quad r = x_1 - loc_{x,y} \\ t &= loc_{x,y} - y_0, \quad b = y_1 - loc_{x,y} \end{aligned} \tag{1}$$

2. Network architecture

본 논문에서 제안하는 알고리즘의 네트워크 구조는 Fig. 3과 같다. 제안 모델은 PyramidBox[15], S3FD[20], Smallhardface[39]와 같은 VGG16[31]기반의 백본(Back-bone) 네트워크를 사용한다. 이 백본 네트워크는 VGG16의 conv1부터 conv5까지 탐지 레이어로 사용한다. 통합한 특징 맵을 사용하는 이전 연구[23]와 달리 클래스 내 불균형 문제를 해결하기 위해 제안하는 구조는 이전 연구[23]에서 겪었던 네 개의 특징 맵으로 얼굴을 검출한다.

FCOS[21]에 영감을 받아 첫 번째와 두 번째 특징 맵은 결합 모듈을 통해 각각 conv3, conv4와 conv4, conv5를 통합한 새로운 특징 맵을 만든다. 이 특징 맵들은 Fig. 4와 같이 결합 모듈을 통해서 결합된다. 제안하는 결합 모듈은 Hu [27]에 영감을 받아 상대적으로 저수준과 고수준의 특징과 서로 다른 수용 필드 그리고 풍부한 정보를 가진다. 세 번째와 네 번째 특징 맵은 두 번째 특징 맵을 입력으로 사용하여 추가한 conv6, conv7 레이어에 통과시켜 얻는다.

추가한 conv6, conv7 레이어는 각각 { 1x1 컨볼루션 레이어(convolution layer), 3x3 컨볼루션 레이어, pooling 레이어 }, { 3x3 컨볼루션 레이어, 3x3 컨볼루션 레이어, pooling 레이어 }로 이루어진다. 이렇게 얻은 4개의 특징 맵들은 검출 모듈을 통하여 얼굴의 위치와 크기를 추정한다. 이 블록은 가중치 공유를 통하여 파라미터를 줄이고 검출 성능을 향상시킨다.

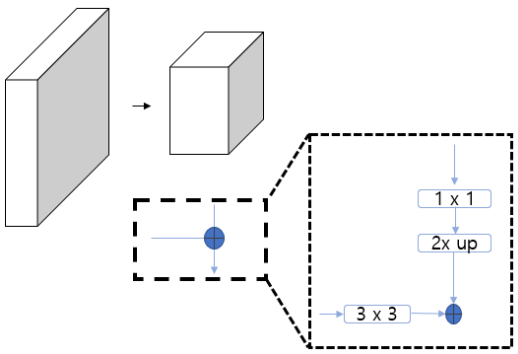

Fig. 4. Concat module

3. Network output

본 논문에서 제안한 네트워크로부터 얻은 특징 맵을 얻는다. 이 특징 맵들을 Fig 5의 검출 모듈을 통과하여 수식 (2), (3)과 같이 얼굴 크기에 대한 네 방향 좌표 회귀 특징 맵과 전경, 배경을 추정하는 분류 특징 맵과 네 방향 좌표 회귀 특징 맵을 얻는다.

$$regression_{x,y} = (l^*, t^*, r^*, b^*) \quad (2)$$

$$classification_{x,y} = y^*_{x,y} \in R^1 \quad (3)$$

수식 (2)에서 각 좌표는 양의 정수 값을 가지므로 각 좌표에 자연상수(exponential)를 적용하여 값의 크기가 $(0,\infty)$의 값을 갖도록 한다. 분류 문제는 $y^*_{x,y}$ 가 1개의 클래스만 가지므로 시그모이드(Sigmoid)함수를 사용하여 분류 값들을 (0,1)사이로 만들어준다.

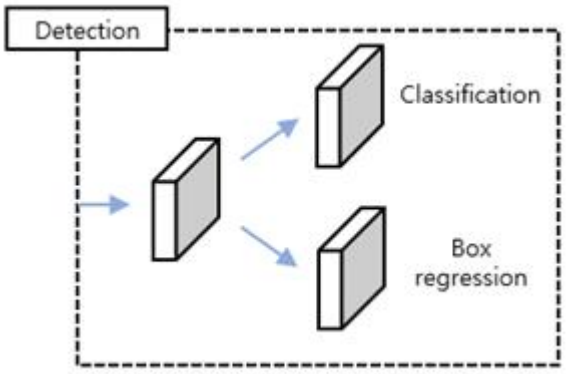

Fig. 5. Detection module

4. Loss function

전체 손실 함수(loss function)는 다음과 같이 정의한다.

$$\begin{aligned} L(F_{x,y}, T_{x,y}) = \lambda_1 L(F^1_{x,y}, T_{x,y}) \\ + \lambda_2 L(F^2_{x,y}, T_{x,y}) \\ + L(F^3_{x,y}, T_{x,y}) \\ + L(F^4_{x,y}, T_{x,y}) \end{aligned} \quad (4)$$

$F^i_{x,y}$는 i번째 특징 맵의 좌표 (x,y)에서 추정한 얼굴과 얼굴 크기이고 $T_{x,y}$는 GT의 얼굴과 얼굴 크기이다. λ_1, λ_2는 손실 함수에 대한 가중치로 사용한다. 본 논문에서는 각각 2,2로 사용하였다. i번째 특징 맵에 대한 손실 함수는 다음과 같다.

$$\begin{aligned} L(F^i_{x,y}, T_{x,y}) = \frac{1}{N}\sum_{x,y}^{N}(L_{cls}(C_{x,y}, C^*_{x,y}) \\ + \frac{\lambda_{reg}}{N}\sum_{x,y}^{N}(L_{reg}(B_{x,y}, B^*_{x,y})) \end{aligned} \quad (5)$$

N은 샘플의 개수이고 λ_{reg}는 해당 손실함수에 대한 가중치이다. 본 논문에서는 λ_{reg}는 0.1로 설정했다. 회귀 문제는 IOU 손실 함수[24]를 사용하고 분류 문제는 클래스 간 불균형 문제를 완화하기 위해 RetinaNet[9]에서 제안한 focal 손실 함수를 사용한다.

IV. Experimental Results

1. Training detail

학습 시 학습에 이용한 이미지의 입력 크기는 800이며 각각의 특징 맵들은 input 이미지 크기의 $\frac{1}{8}, \frac{1}{16}, \frac{1}{32}, \frac{1}{64}$를 가지고 (8 x 8), (16 x 16), (32 x 32), (64 x 64) 그 이상의 얼굴 크기를 학습한다. 배치사이즈는 4를 사용한다. 최적화 알고리즘은 아담 최적화 방식(Adam optimizer)를 사용하고, learning rate는 5e-4로 120 epoch 동안 학습을 진행한 후 10 epoch에 걸쳐 $\frac{1}{10}$ 씩 줄인다. 가중치 감소(Weight-decay)와 모멘텀은 각각 0.0001과 0.9를 사용한다. 백본 네트워크는 이미지넷(ImageNet)으로 미리 학습한 모델을 사용하고 직접 추가한 모델의 레이어는 He 초기화 방법을 이용하여 초기화한다. 테스트 시 이미지에서 추출된 얼굴 후보들은 NMS (Non-maximum suppression)를 통해 최종 후보들을 추린다. NMS의 threshold는 0.3으로 설정한다. 실험 환경은 아래의 Table 1.과 같다. 하드웨어는 CPU Ryzen 7 2700X, GPU는 RTX 2080 Ti 11GB, RAM은 32GB를 사용하였다. 사용한 모델은 Ubuntu 18.04 운영체제에서 파이토치(Pytorch) 1.4 버전 딥 러닝 프레임워크를 사용하였다.

Table 1. System Environment

OS	Ubuntu 18.04
H/W	CPU : Ryzen 7 2700X GPU: RTX 2080 Ti 11GB RAM : 32GB
Deep learning framework	Pytorch 1.4

2. Experiment

얼굴 검출분야에서 대표적으로 많이 사용되는 FDDB 데이터 셋[38]으로 실험을 진행했다. FDDB 데이터 셋은 2,845개의 이미지와 5,171개의 라벨링된 얼굴들이 있다.

본 논문에서 제안하는 방법을 [22],[24],[25],[41]과 비교하였다. Table 2, 3은 FDDB 데이터 셋에서 제공하는 평가 방법으로 평가한 성능을 보여준다. DiscROC와 ContROC Curves는 동일한 방법으로 평가하지만 ContROC Curves의 경우 DiscROC 보다 훨씬 낮은 IOU를 가지고 있어 상대적으로 오검출이 더 많다. 제안하는 방법은 [22],[24],[25],[41]과 비교하여 DiscROC Cureves의 경우 비슷한 수준의 성능을 가졌다. ContROC Curves에서는 다른 논문들의 방법과 큰 차이를 보이지 않지만 한 개의 특징 맵으로 검출하는 [24]와 비교하였을 때 Disc ROC Curves의 값이 상대적으로 오검출이 적은 것을 확인할 수 있다. 또한 가중치 공유를 한 모델과 그렇지 않을 모델을 비교하였을 때 가중치 공유를 한 모델의 오검출이 더 적은 것을 확인할 수 있다.

Table 2. DiscROC Curves score on FDDB for 1000 False Positives

Method	Disc ROC curves score
[25]	0.980
[22]	0.966
[24]	0.950
[41]	0.973
Ours (not share weight)	0.928
Ours	0.944

Table 3. ContROC Curves score on FDDB for 1000 False Positives

Method	Cont ROC curves score
[25]	0.732
[24]	0.721
[41]	0.724
Ours (not share weight)	0.717
Ours	0.723

V. Conclusions

본 논문에서는 앵커 프리 방법을 이용한 다중 크기 얼굴 검출기를 제안하였다. 기존의 최첨단(state-of-the-art) 모델들은 앵커 기반 방법을 이용하여 얼굴을 검출하지만 정밀 튜닝을 해야 하는 하이퍼 파라미터와 사전정의앵커에 민감한 성능을 가졌고 기존의 앵커 프리 방법을 이용한 모델들은 통합된 하나의 특징 맵으로 얼굴을 검출함으로써 클래스 내 불균형 문제가 존재했다. 이를 개선하기 위해 본 논문에서는 두 가지 방법을 제안하였다. 첫째로 앵커 프리 방법을 사용하여 하이퍼 파라미터의 수를 줄임으로써 학습 시 안정성이 증가한다. 둘째로 얼굴 검출 시 다중 크기 특징 학습 중 하나인 특징 피라미드방식으로 학습한 여러 개의 특징 맵을 사용함으로써 기존의 논문에 존재했던 클래스 내 불균형 문제를 완화했다. 또한 가중치 공유 모듈을 사용함으로써 검출 성능을 높였다. 제안하는 방법의 성능을 평가하기 위해 정량적 평가를 진행하였고 제안하는 방법의 효율성을 입증하였다. 하지만 제안하는 방법은 손실 함수와 후처리 알고리즘인 NMS 과정에서 가중치의 정밀 튜닝이 필요로 한다. 제안하는 구조 중 세 번째, 네 번째 특징 맵은 백본 네트워크의 깊은 망으로 인해 학습이 잘 이루어지지 않았다. 향후 과제로 사용자가 적절한 하이퍼 파라미터를 선택할 수 있는 방법을 연구하고 입력 크기에 강건한 특징 맵을 추출할 수 있는 방법을 연구하고 할 필요성이 있다.

ACKNOWLEDGEMENT

This study has been worked with the support of a research grant of Kangwon National University in 2017

REFERENCES

[1] J. Li, T. Wang, and Y. Zhang, “Face detection using surf cascade,” in Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on, pp. 2183-2190, IEEE, 2011.

[2] K.-K. Sung and T. Poggio. “Example-based learning for viewbased human face detection.” IEEE Transactions on pattern analysis and machine intelligence, 20(1):39-51, 1998.

[3] P. Viola and M. J. Jones. “Robust real-time face detection.”

International journal of computer vision, 57(2):137-154,2004

[4] X. Zhu and D. Ramanan. “Face detection, pose estimation,and landmark localization in the wild.” In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on,pages 2879-2886. IEEE, 2012

[5] Ren, S., He, K., Girshick, R., & Sun, J. “Faster r-cnn: Towards real-time object detection with region proposal networks.” In Advances in neural information processing systems (pp. 91-99). 20115

[6] Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. “You only look once: Unified, real-time object detection.” In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788). 2016

[7] Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C. Y., & Berg, A. C. (2016, October). “Ssd: Single shot multibox detector.” In European conference on computer vision (pp. 21-37). Springer, Cham.

[8] Lin, T. Y., Dollár, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. “Feature pyramid networks for object detection.” In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 2117-2125). 2017

[9] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollar. “Focal loss for dense object detection.” In ´ Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pages 2980-2988,2017.

[10] He, K., Gkioxari, G., Dollár, P., & Girshick, R. “Mask r-cnn.” In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969). 2017

[11] S. Yang, P. Luo, C.-C. Loy, and X. Tang. “Wider face: A face detection benchmark.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5525-5533, 2016.

[12] H. Jiang and E. Learned-Miller. “Face detection with the faster r-cnn.” In Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, pages 650-657. IEEE, 2017.

[13] Wang, H., Li, Z., Ji, X., & Wang, Y. “Face r-cnn.” arXiv preprint arXiv:1706.01061. 2017

[14] Yang, W., & Jiachun, Z. “Real-time face detection based on YOLO.” In 2018 1st IEEE International Conference on Knowledge Innovation and Invention (ICKII) (pp. 221-224). IEEE. September 2018

[15] Tang, X., Du, D. K., He, Z., & Liu, J. “Pyramidbox: A context-assisted single shot face detector.” In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 797-813). 2018

[16] Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., & Zafeiriou, S. “Retinaface: Single-stage dense face localisation in the wild.” arXiv preprint arXiv:1905.00641. 2019

[17] Cakiroglu, O., Ozer, C., & Gunsel, B. “Design of a Deep Face Detector by Mask R-CNN.” In 2019 27th Signal Processing and Communications Applications Conference (SIU) (pp. 1-4). IEEE. April 2019

[18] Li, J., Wang, Y., Wang, C., Tai, Y., Qian, J., Yang, J., ... & Huang, F. “DSFD: dual shot face detector.” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 5060-5069) 2019

[19] Najibi, M., Samangouei, P., Chellappa, R., & Davis, L. S. “Ssh: Single stage headless face detector.” In Proceedings of the IEEE International Conference on Computer Vision (pp. 4875-4884). 2017

[20] Zhang, S., Zhu, X., Lei, Z., Shi, H., Wang, X., & Li, S. Z. “S3fd: Single shot scale-invariant face detector.” In Proceedings of the IEEE International Conference on Computer Vision (pp. 192-201). 2017

[21] Tian, Z., Shen, C., Chen, H., & He, T. “Fcos: Fully convolutional one-stage object detection.” In Proceedings of the IEEE International Conference on Computer Vision (pp. 9627-9636). 2019

[22] Wang, C., Luo, Z., Lian, S., & Li, S. “Anchor Free Network for Multi-Scale Face Detection.” In 2018 24th International Conference on Pattern Recognition (ICPR) (pp. 1554-1559). IEEE. August 2018

[23] L. Huang, Y. Yang, Y. Deng, and Y. Yu, “Densebox: Unifying landmark localization with end to end object detection,” arXiv preprint arXiv:1509.04874, 2015.

[24] J. Yu, Y. Jiang, Z. Wang, Z. Cao, and T. Huang, “Unitbox: An advanced object detection network,” in Proc. ACM Multimedia, 2016, pp. 516-520.

[25] Xu, Y., Yan, W., Sun, H., Yang, G., & Luo, J. “ CenterFace: Joint Face Detection and Alignment Using Face as Point.” arXiv preprint arXiv:1911.03599. 2019

[26] Yotam Abramson, Bruno Steux, and Hicham Ghorayeb. “Yet even faster (yef) real-time object detection.” International Journal of Intelligent Systems Technologies and Applications, 2(2-3):102-112, 2007. 2

[27] P. Hu and D. Ramanan, “Finding tiny faces,” in Proc. CVPR, 2017, pp.1522-1530.

[28] Zhang, Z., Shen, W., Qiao, S., Wang, Y., Wang, B., & Yuille, A. “Robust face detection via learning small faces on hard images”. In The IEEE Winter Conference on Applications of Computer Vision (pp. 1361-1370). 2020

[29] Yu Chen, Chunhua Shen, Xiu-Shen Wei, Lingqiao Liu, and Jian Yang. Adversarial “PoseNet: A structure-aware convolutional network for human pose estimation.” In Proc. IEEE Int. Conf. Comp. Vis., 2017.

[30] Sun, K., Xiao, B., Liu, D., & Wang, J. “Deep high-resolution representation learning for human pose estimation." In Proceedings of the IEEE Conference on Computer Vision and

Pattern Recognition (pp. 5693-5703). 2019

[31] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv:1409.1556, 2014.

[32] Lokesh Boominathan, Srinivas SS Kruthiventi, and R Venkatesh Babu. “Crowdnet: A deep convolutional network for dense crowd counting.” In Proc. ACM Int. Conf. Multimedia, pages 640–644. ACM, 2016.

[33] Sam, D. B., Peri, S. V., Kamath, A., & Babu, R. V. “Locate, Size and Count: Accurately Resolving People in Dense Crowds via Detection.” arXiv preprint arXiv:1906.07538. 2019

[34] Sindagi, V. A., & Patel, V. M. “Inverse attention guided deep crowd counting network.” In 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS) (pp. 1-8). IEEE. September 2019.

[35] Jonathan Long, Evan Shelhamer, and Trevor Darrell. “Fully convolutional networks for semantic segmentation.” In Proc. IEEE Conf. Comp. Vis. Patt. Recogn., pages 3431–3440, 2015

[36] Zhang, Kaipeng, et al. "Joint face detection and alignment using multitask cascaded convolutional networks." IEEE Signal Processing Letters 23.10 (2016): 1499-1503.

[37] Huang, Gary B., et al. "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments." 2008.

[38] Jain, Vidit, and Erik Learned-Miller. “Fddb: A benchmark for face detection” in unconstrained settings. Vol. 2. No. 6. UMass Amherst technical report, 2010.

[39] B. Alexe, T. Deselaers, V. Ferrari, Measuring the objectness of image windows, in: Proceedings of the TPAMI, 2012.

[40] J.R. Uijlings, K.E. Van De Sande, T. Gevers, A.W. Smeulders, Selective search for object recognition, in: Proceedings of the IJCV, 2013.

[41] Liu, Li, et al. "Deep learning for generic object detection: A survey." International journal of computer vision 128.2 (2020): 261-318.

[42] Zhang, Shifeng, et al. "Refineface: Refinement neural network for high performance face detection." IEEE Transactions on Pattern Analysis and Machine Intelligence (2020).

Authors

Dong-Ryeol Lee received the B.S. degrees in Computer Science and Engineering from Kangwon National University, Korea, in 2018. He is currently a M.S. Student in the department of Computer Science and Engineering at Kangwon National University. His research interests are in the areas of machine learning and computer vision.

Yoon Kim received a B.S. degree in 1993, an M.S. degree in 1995, and a Ph.D. degree in 2003, in electronic engineering with the Department of Electronic Engineering from Korea University In 2004, he joined the Department of Computer Science and Engineering, Kangwon National University, where he is currently a professor. From 1995 to 1999, he was with the LG-Philips LCD Co., where he was involved in research and development on digital image equipment. His research interests are in the areas of machine learning, multimedia communications, and computer vision.