Journal of KIIT. Vol. 22, No. 3, pp. 1-9, Mar. 31, 2024. pISSN 1598-8619, eISSN 2093-7571
http://dx.doi.org/10.14801/jkiit.2024.22.3.1

인파 밀집 분석을 위한 보행자 검출 방법에 대한 연구

임혜연*, 안명수**, 강대성***

A Study on Pedestrian Detection Method for Crowd Density Analysis

Hye-Youn Lim*, MingShou An**, and Dae-Seong Kang***

이 논문은 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No.RS-2023-00247045)

요 약

영상에서 사람을 검출하는 기술에 대한 연구는 꾸준히 진행되어 왔으며, 최근에는 밀집 인파 분석을 위한 사람의 추적, 행동 인식 등의 다양한 연구가 진행되고 있다. 밀집된 인파에서의 보행자 검출은 보행자간의 가려짐 또는 겹침으로 인해 정확도 저하 및 오검출 등 문제를 발생시킬 수 있다. 따라서 본 논문에서는 이러한 문제를 해결하고자 YOLOv5 모델 기반의 개선 방법을 제안한다. 먼저 네트워크 특징 융합 단계에서 어텐션 메커니즘 기반 신경망을 구성하여 특징 추출 성능 향상과 동시에 NMS(Non-Maximum Suppression)의 계산 부담을 줄였다. 다음으로 업샘플링을 통한 각 신경망 층의 피라미드 특징의 해상도를 향상시켜 중첩 대상에 대한 차별적인 어텐션 특징 맵을 생성 및 결합하는 구조를 제안하였다. 제안하는 방법을 CrowdHuman 데이터 셋에서의 실험을 통해, 실시간 검출 테스트 결과 mAP 2.3%, loss 0.013이 개선되었음을 보여주었다.

Abstract

Research on technology for detecting people in video has been steadily conducted, and various studies such as human tracking and behavior recognition for dense population analysis have recently been conducted. Pedestrian detection in dense crowds can cause problems such as poor accuracy and false detection due to cover or overlap between pedestrians. Therefore, this paper proposes an improvement method based on the YOLOv5 model to solve this problem. First, an attachment mechanism-based neural network was constructed in the network feature fusion stage to improve feature extraction performance and reduce the computational burden of non-maximum suppression(NMS). Next, we propose a structure to create and combine differential attention feature maps for overlapping targets by improving the resolution of pyramid features of each neural network layer through upsampling. Experiments on CrowdHuman datasets showed that the proposed method improved mAP 2.3% and loss 0.013 as a result of real-time detection tests.

Keywords
crowd density, attention mechanism, saliency mechanism, object detection

* 동아대학교 전자공학과 조교수
- ORCID: https://orcid.org/0000-0002-5189-461X
** 중국 서안공업대학교 컴퓨터공학과 강사
- ORCID: https://orcid.org/0000-0002-1180-8916
*** 동아대학교 전자공학과 교수(교신저자)
- ORCID: https://orcid.org/0000-0003-0186-2430

· Received: Oct. 30, 2023, Revised: Mar. 13, 2024, Accepted: Mar. 16, 2024
· Corresponding Author: Dae-Seong Kang
Dept. of Dong-A University,37 NaKdong-Daero 550, beon-gil saha-gu, Busan, Korea
Tel.: +82-51-200-7710, Email: dskang@dau.ac.kr

Ⅰ. 서 론

디지털 서비스의 확산은 우리 사회에 많은 변화를 가져왔다. 특히 스마트 시티와 스마트 교통이 빠르게 발전함에 따라 보행자 검출 및 추적 기술에 대한 수요가 크게 증가하고 있다. 보행자 검출이란 영상에서 보행자의 유무와 위치를 확인하는 기법을 말한다. 지능형 영상 감시 시스템, 자율주행 자동차, 인간-컴퓨터 상호 작용 등 다양한 응용 분야에서 보행자 검출은 다년간 광범위하게 연구되어 온 중요한 문제 중 하나이다. 또한 최근 몇 년 동안 우리 사회에는 이태원 압사, 지하철 인파사고 등 인파로 인한 다중 밀집 사고가 빈번하게 발생한다. 따라서 이와 같은 인파 대형 사고 예방 및 긴급상황 대비를 위한 영상 기반 분석 기술이 필요하다. 이러한 영상 분석 기술은 인파의 규모 및 밀집도, 이동방향 등 다양한 데이터를 수집하고 분석할 수 있어야 한다. 밀집된 인파, 즉 사람을 검출하는 기술에 대한 연구는 컴퓨터 비전 분야에서의 이슈 중 하나이다. 영상에서 보행자를 검출함에 있어서 보행자 특징의 다양성, 복잡한 배경, 다른 물체에 의한 가려짐 및 유사특성을 가진 보행자간의 겹침현상 등은 아직도 어려운 과제다.

보행자 검출은 주로 두 개 단계를 거쳐 발전하였다, 첫 번째는 2014년 이전의 전통적인 머신러닝 방법을 기반으로 하는 단계이고, 두 번째는 2014년 이후 딥러닝을 기반으로 하는 단계이다. 전통적인 머신러닝에서는 사람이 선택한 특징을 기반으로 비교적 간단한 모델을 학습시키고 예측을 수행한다. 반면에 딥러닝에서는 데이터 자체에 포함된 고차원적인 특징을 자동으로 추출하고, 이를 기반으로 모델을 학습하고 예측을 수행한다. 따라서 딥러닝에서는 학습을 위해 사람이 직접 특징을 선택하는 작업이 필요 없고, 기존의 비교적 단순한 모델에 비해 딥러닝 모델은 계층수를 증가시킨 심층신경망을 통해 학습하므로 더욱 복잡한 특징에 대한 학습을 통해 데이터에 대한 표현 능력을 크게 증가시킬 수 있다. 또한 특징 추출과 분류기가 하나의 모델에 통합되어 데이터에 대한 인식 및 분류 성능을 극대화 할 수 있다[1].

인공지능 및 딥러닝 기술의 발전이 심화함에 따라 딥러닝 방법은 보행자 검출 분야에서 뛰어난 성능을 보였으며, 주요 기술로 거듭났다. 딥러닝 기반 보행자 검출 방법은 크게 두 가지 부류로 나눌 수 있다. 하나는 2-stage 검출 방법으로 바운딩 박스를 그려 객체의 위치 정보를 설정하여 심층 아키텍처를 통해 검출과 분류를 각각 처리하는 방법이다. 대표적인 모델로는 R-CNN(Region Convolutional Neural Network)[1], Faster R-CNN[2], Mask R-CNN[3] 등이 있다. 다른 하나는 1-stage 검출 방법으로 모든 공간 영역에 대해 직접적으로 검출 및 분류를 동시에 수행하므로 비교적 간단한 아키텍쳐를 통해 보행자를 검출할 수 있다는 장점이 있다. 대표적인 모델로 YOLO(You Look Only Once) 시리즈[4], SSD(Single Shot MultiBox Detector)[5] 등이 있다. 즉, 2-stage 검출 방법에서는 먼저 객체를 인식하고 다시 인식된 객체의 영역을 분류하는 방식으로 처리하였다면, YOLO와 같은 1-stage 검출 방법에서는 단 한번의 과정을 통해서 객체의 인식과 분류를 동시에 처리하도록 설계되었다.

본 논문에서는 인파 다중 밀집 분석을 위한 YOLOv5 기반의 새로운 보행자 검출 방법을 제안한다. 제안하는 방법은 보행자간의 겹침 현상을 고려할 뿐만 아니라, 업샘플링을 통한 피라미드 특징을 기반으로 하는 상하관계를 통해 정확도를 개선하도록 한다.

Ⅱ. 관련 이론

2.1 YOLO 모델

YOLO는 대표적인 객체 탐지 모듈로서 CVPR2016에서 처음 공개된 후 현재 YOLO v8 버전까지 출시되었다[6]. 그중 가장 많이 응용되는 버전은 v5와 v8 버전으로, v5는 사용성이 편리하다는 점에서 많은 응용에 쉽게 접근할 수 있다는 장점이 있고, v8 버전은 더 빠르고 정확하다는 장점이 있어, 결국에는 애플리케이션의 요구 사항에 따라 사용되는 버전도 달라지지만, 대체적으로 실시간으로 처리할 것인지 아니면 접근성을 확보할 것인지에 따라서 선택할 수가 있다[7]. YOLO v5는 2020년에 Ultralytics 연구팀이 제안한 1-stage 객체 검출 알고리즘으로 기존의 모델을 간소화함과 동시에 강인성

및 처리 속도와 정확도를 개선하였다. 다른 모델에 비해 YOLO v5는 처리 속도와 정확도를 확보함과 동시에 상대적으로 낮은 하드웨어 사양에서도 구현 가능하다는 장점이 있다.

일반적으로 YOLO v5 기반의 보행자 검출 또는 객체 검출 방법은 크게 세 개 부분으로 구성할 수 있다. (1) 입력 영상 데이터의 특징을 추출하는 핵심 네트워크 Backbone에서는 ResNet, DenseNet 그리고 VGGNet 등과 같은 CNN 구조를 통해 fine-tune 과정을 수행하게 된다. 이렇게 higher semantics와 같은 심층 구조에 의한 다양한 라벨의 특징을 생성하는 구조는 보행자 검출 신경망의 후반부에서 유용한 정보로 제공된다. (2) 다양한 스케일의 피라미드 특징을 융합하여 핵심 네트워크 Backbone과 연결해 주는 Neck에서는 FPN(Feature Pyramid Network), PANet, Bi-FPN 등이 사용 될 수 있다. Neck은 이름 그래도 backbone과 뒤에 나오는 head 사이를 연결해 주는 연결부이다. 여기서는 다양한 층, 즉 피라미드 특징을기반으로 하기 때문에 다양한 스케일 변화를 주어 객체를 효과적으로 검출할 수 있다는 것이다. (3) Head에서는 객체 분류를 위한 바운딩 박스의 분류 및 회귀와 같은 검출 기능이 이루어 지는 실질적인 부분이다. 이렇게 세 모듈을 하나의 과정으로 통합해서 보면 그림 1과 같다.

2.2 YOLO v5의 원리

YOLO 모델의 공통점은 입력 이미지 또는 특징맵(Feature map)을 특정 격자망(Grid)으로 나누고, 각 셀(Cell)마다 객체 검출을 수행하는 것이며, 이는 SSD 모델에서의 특징맵의 각 포인트마다 객체 검출을 수행하는 원리와 같다고 볼 수 있다.

YOLO v5의 경우 backbone으로 SCP-Darknet53을 사용하며, 이는 주로 Focus, CSP(Cross Stage Partial), SPP(Spatial Pyramid Pooling) 세 개 부분으로 구성된다. 그중 Focus 모듈은 빠른 다운샘플링 작업을 수행할 수 있다. CSP 모듈은 CSPNet 구조를 따르며, YOLO v4의 Scaled버전에서 제안된 방법이다. CSP 구조는 이전 계층으로부터 입력받은 데이터를 각각 컨볼루션 연산을 수행하는 두 개의 경로로 분리한다. 그중 하나의 경로는 CBL 블록을 이용하여 컨볼루션, 정규화, 활성화 함수 등 처리를 거친후 Multiple Residual 구조로 전달된다. 다른 경로에서는 컨볼루션 정보만 전달한다. 상기 두 경도로에 전달되는 데이터를 결합하여 다음 계층으로 전달하여 신경망 모델의 학습력을 향상함과 동시에 정확도를 확보할 수 있다. SPP는 CNN 기반 모델[8]들이 고정적인 입력 이미지 크기만 사용한다는 단점을 해결하고자 제안된 방법으로써, CNN 수행 후 SPP를 수행하게 되면 고정된 스케일의 텐서가 출력되기 때문에 이미지의 화질을 수정하거나 잘라낼 필요가 없이 바로 검출이나 분류 학습을 수행할 수 있다는 장점이 있다. SPP 작동 원리는 SPM(Spatial Pyramid Matching)을 기본으로 이미지의 관심 영역별로 특징을 추출한 뒤 각 특징별로 이미지에서 나타나는 빈도수를 카운팅하는 기법이다. SPP 블록은 Conv, max-pooling 및 concat 세 부분으로 구성된다.

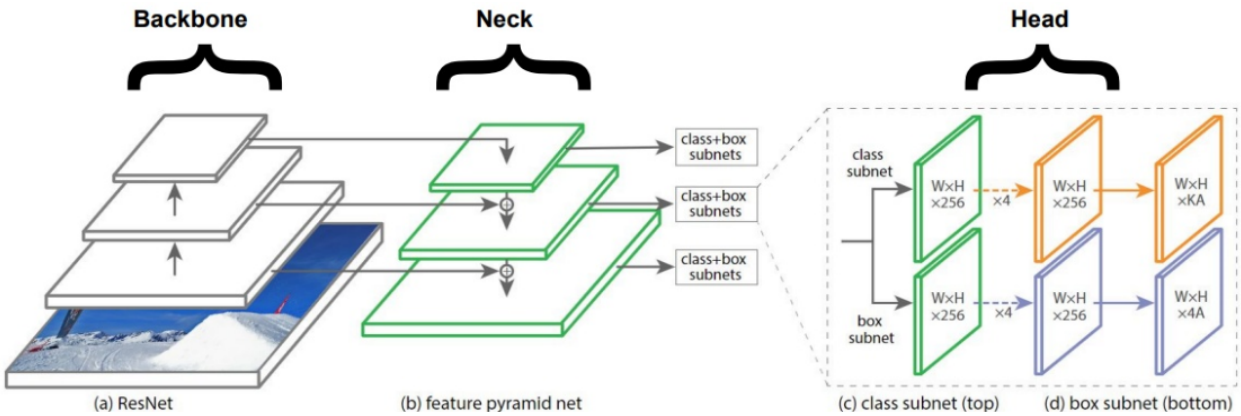


그림 1. YOLO 아키텍쳐
Fig. 1. YOLO architecture

III. 제안하는 방법

본 논문에서는 YOLO v5 모델을 기반으로 프레임워크, 측정 스켈링 및 손실함수에 대한 개선을 통해 밀집인파 환경에서의 보행자 검출을 구현하였다.

3.1 베이스 네트워크 개선 방안

객체 검출이나 분할 등 영역에서 관심 객체에 더욱 집중하여 학습할 수 있는 어텐션 메커니즘(Attention mechanism)[9] 개념을 도입하면 영상에서의 관심 영역 즉 관심 특징에 초점을 맞춰, 불필요한 특징을 억제하는 작용을 한다.

본 논문에서는 밀집 인파 영상에서 복잡한 배경 또는 목표 객체와 배경이 비슷한 경우 등 요소로 인해 목표 객체의 특징이 불명확하거나 뚜렷하지 못한 상황을 대비하여 관심 영역 검출 기법인 어텐션 네트워크를 추가적으로 구성하여 특징 추출 네트워크로 사용한다. 밀집 인파 즉 보행자 영역을 제외한 배경 영역에 대한 특징은 보행자의 특징 추출 시 잘못된 특징으로 학습될 수 있기 때문에 보행자 내부 영역을 관심 영역으로 구분하여 핵심적인 특징으로 활용하고자 한다. 최종적으로 기존의 YOLO v5s 모델에서 보행자와 배경을 구분하는 가중치 특징 맵을 추출하여 사전 학습된 네트워크의 특징에 컨볼루션 연산을 수행하여 보행자의 위치 정보를 추가적으로 반영하여 바운딩 박스를 찾기 위한 특징의 공간적 중요도를 강조할 수 있게 모델링할 수 있다. 그림 2는 YOLO v5s 모델에 어텐션 네트워크를 조합을 통해 개선된 구조를 보여주고 있다.

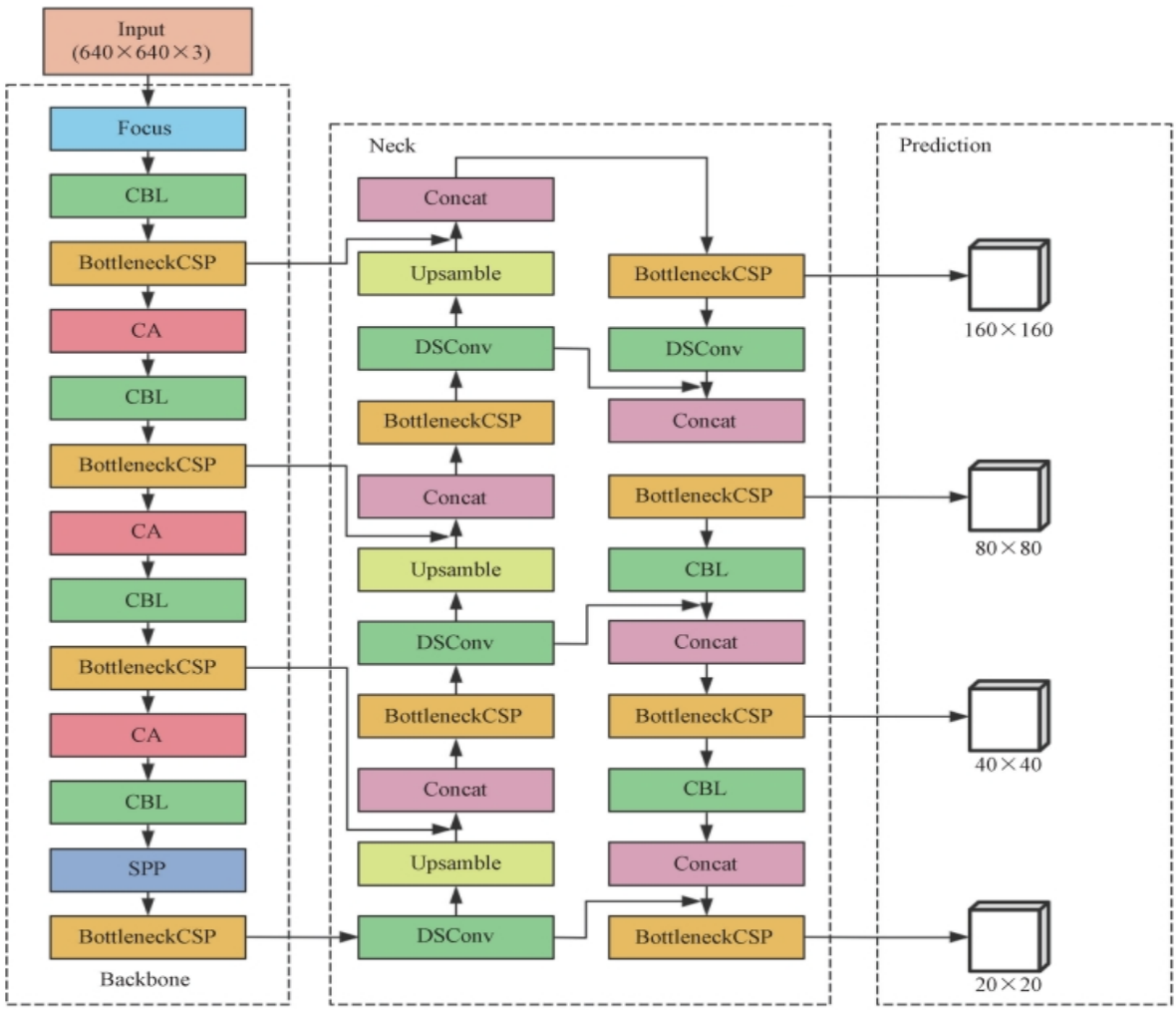


그림 2. YOLO v5 네트워크 구조도
Fig. 2. Structure of modified YOLO v5 network

개선된 베이스 네트워크는 두 단계로 구성된다. 하나는 어텐션 융합 특징 추출이고, 다른 하나는 인파 밀도 맵의 생성이다. 첫 번째 단계에서는 기존 YOLO v5s의 각 컨볼루션 층에 어텐션 특징을 조합형 융합 특징을 추출한다. 입력된 밀집 인파 영상 S에 각각 사이즈가 $(h,1)$과 $(1,w)$인 커널을 이용하여 인코딩 작업을 수행하여 height가 H, weight가 W인 C번째 채널의 독립된 수용 특징 맵 $Z(H)$와 $Z(W)$를 생성하며, 사이즈는 각각 $c\times1\times h$, $c\times1\times w$이다. 특징맵 계산 수식은 다음과 같다.

$$Z(H)_C=\frac{1}{w}\sum_i S(H,i) \tag{1}$$

$$Z(W)_C=\frac{1}{h}\sum_i S(W,i) \tag{2}$$

각 컨볼루션 층에서 풀링(Pooling)을 수행했기 때문에 네트워크 층이 깊어질수록 특징 맵의 사이즈는 축소되면 최종적으로 원 영상의 1/16로 되며, 인코더의 출력에 더 많은 특징 정보를 부여하게 된다. 여러 번의 풀링작업을 통해 얻은 특징 맵은 해상도가 현저히 저하되며 일부의 공간적 정보를 손실하게 된다. 따라서 디코더의 역할은 영상의 사이즈와 공간적 정보를 복원함과 동시에 스케일 어텐션 맵을 생성하도록 한다. 그 원리는 디코더시 핵심적인 4개의 층(사이즈가 원 영상 사이즈의 1/16, 1/8, 1/4, 1/2인 특징 맵)의 깊이 특징을 유지하게 한다. 또한 high layer의 특징들은 유용한 특징들을 포함하게 되고, low layer의 특징들은 보행자 및 기타 배경 특징 간의 경계 에지 정보를 포함하게 된다. 따라서 본 논문에서는 디코더 단계에서의 4 장의 특징 맵을 스케일 어텐션 네트워크를 위한 특징 맵으로 사용하게 된다.

두 번째 단계에서는 Concat융합을 통해 상기 생성된 $Z(H)$과 $Z(W)$을 컴볼루션 변환 함수 $F()$ 을 사용하여 변환시킨다. 즉, 수평과 수직방향에서 공간적 특징에 대해 인코딩된 어텐션 특징 맵 A를 생성한다.

$$A=\delta[F(Z(H)+Z(W))] \tag{3}$$

식 (3)에서 δ는 비선형 활성화 함수를 뜻한다. 식에서 사용된 커널 함수 $F()$는 사이즈가 1인 커널을 사용한다. 어텐션 네트워크는 주로 컨볼루션 모델 를 학습하여 밀집 인파 영상 S에서 어텐션 특징 맵을 추출하고 loss함수를 통해 예측 값과 실제 값의 손실 정도를 분석하여 학습을 수행한다. 각 층에서의 특징 맵 융합과정은 다음 수식을 통해 구현한다.

$$C_i^n=F_i^n * A_i^n,\ n=1,2,3,4 \tag{4}$$

위 식에서 C_i^n는 i번째 입력 영상이 특징 융합 네트워크에서의 n번째 특징 융합과정을 의미하며, F_i^n는 n번째 융합의 입력값을 의미한다. A_i^n는 입력 영상에 대응되는 어텐션 맵을 의미한다.

3.2 NMS 구조 개선

NMS(Non-Maximum Suppression)의 기본 원리는 다음과 같다. 대다수의 객체 검출 알고리즘은 객체 영역에 여러 개의 $C_A:S\rightarrow A^{4\times i}$ score가 비교적 높은 바운딩 박스를 생성하며, 결론적으로는 이중 하나의 바운딩 박스만을 선택하여 최종 검출된 객체의 위치로 확정되는데, 이때 적용되는 기법이 NMS 방법이다. 즉, NMS는 객체 검출기가 예측하는 바운딩 박스 중에서 가장 정확한 바운딩 박스만을 선택하도록 하는 기법이다. 전통적인 NMS는 구조가 비교적 간단하고 효율적이지만, 일부 특수 환경에서는 다음과 같은 문제점들이 발생할 수 있다. 첫 번째는 score가 비교적 낮아 겸침이나 많은 부분이 가려진 특징을 가지는 박스에 대해서는 미검출 문제나 mAP(mean Average Precision)를 저하시킬 수 있다. 두 번째는 NMS의 임계값을 설정할 수 없다. 임계값을 낮게 설정하면 두 개의 독립된 객체에 대한 예측 바운딩 박스 중에 적은 양의 특징을 가지는 박스에 대해서 억제하여 하나로 인식할 수 있어 미검출을 초래할 수 있다는 것이다. 반대로 임계값을 너무 크게 설정하면 두 개의 겹치는 객체 사이에 잘못된 예측 바운딩 박스가 추가될 수 있어 오검출을 초래할 수도 있다.

오직 score만을 표준 척도로 사용하게 되면 score가 높은 바운딩 박스가 특정 상황에서는 정확한 위치가 아닐 수도 있다. 세 번째는 처리속도 이다. NMS는 반복적인 연산을 수행하게 된다. 따라서 GPU의 병렬처리 효율을 떨어뜨릴 수 있다.

이에 본 논문에서는 일정 비율 이상의 겹치는 바운딩 박스들의 신뢰도(Confidence)를 0으로 만들어 억제하는 방법을 사용하지 않고 줄여서 최종 mAP를 향상시키는 Soft NMS를 베이스 네트워크에 사용하여 구현한다.

3.3 활성화 함수

활성화 함수는 크게 두 가지 종류로 나눌 수 있다. 하나는 선형 함수이고, 다른 하나는 비선형 함수이다. 일반적으로 활성화 함수의 미분과정을 통해 역전파 학습을 수행하는 과정에서 손실 값을 줄이는 연산을 하게 된다. 신경망 모델에서 예측값과 가중치 간에는 비선형 관계를 가지게 되므로 비선형 함수를 사용하게 된다. 현재 신경망 모델에서 가장 많이 사용하는 활성화 함수는 ReLU(Rectified Linear Unit) 함수이다. 하지만 ReLU의 경우 음의 영역에서는 Dying 현상이 발생하여 학습 성능을 저하시킬 수 있다. 이에 Swish의 특성을 기반으로 하는 활성화 ReLU의 Dying 문제를 해결하여 좋은 성능을 보였다[10]. 본 논문에서는 그림 3과 같은 형태의 활성화 함수인 CoS(Consist of Sigmoid) 함수를 사용한다.

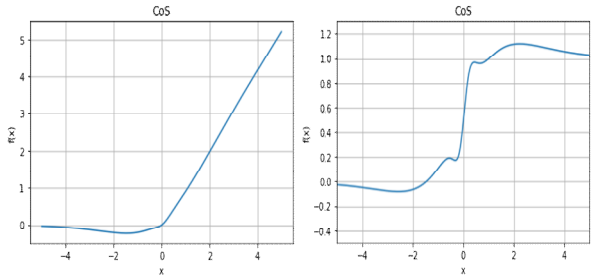


그림 3. CoS 활성화 함수 및 미분 그래프
Fig. 3. CoS activation function and differential graph

CoS 함수는 음의 영역 값을 더 적게 줄여 학습 과정에서의 음수 값의 영향을 줄이고, 양의 값은 ReLU의 특성을 그대로 사용하고, Swish의 비단조성 특징을 강조하여 표현력을 증가시킬 수 있다. 활성화 함수는 식 (5)로 표현된다.

$$f(x) = ((\sigma(5x) - 0.5)^2 + x) \cdot \sigma(x) \tag{5}$$

식에서 σ는 표준편차를 나타내며, 추가된 시그모이드 함수의 제곱식을 x의 계수 5와 상수값 0.5로 위치를 조절하여 Swish의 특성을 유지하면서 부드러운 곡선을 구사하여 초기값 및 학습률에 강인하고 기울기 변화에 따른 표현력 향상을 통해 네트워크의 정확도를 개선할 수 있다.

Ⅳ. 실험 결과

4.1 실험 데이터셋

본 논문에서 제안한 모델의 성능을 검증하기 위하여 공개 데이터 셋인 CrowdHuman Dataset을 이용하여 실험을 진행한다. 상기 데이터 셋은 객체의 밀집도가 높고 더욱 광범위한 장면 정보를 포함하고 있어 데이터의 규모가 아주 크며 풍부한 라벨링 정보와 다양성 데이터 정보를 포함하고 있다. 이 데이터 셋은 총 24000장의 영상을 포함하고 있으며, 그중 15000장의 학습용 영상, 4370장의 검증용 영상과 5000장의 테스트 영상을 포함하고 있다. 학습 및 검증용 데이터에 포함된 라벨링 보행자수는 470k에 달하며, 각 영상마다 다양한 겹침현상이 존재한다. 본 논문에서는 데이터 셋 중에서 임의로 10000장의 영상을 사용하여 제안하는 방법에 의한 모델을 학습시켰으며, 3000장의 테스트 영상을 사용하여 모델을 검증하였다. 또한 실제 환경에서의 테스트를 위하여 실내에 설치된 CCTV 카메라 영상을 통해 모델의 성능을 검증하였다.

4.2 성능지표

본 논문에서는 모델의 성능을 평가하기 위하여 AP(Average Precision), mAP(mean Average Precision), FPS(Frames Per Second), R(Recall) 등 성능지표를 사용한다.

precisiondms 객체를 검출한 결과가 실제 객체와 일치하는지를 나타내는 성능지표이고, recall은 실제 객체를 기준으로 얼마나 검출했는지에 대한 검출율을 나타내는 성능지표이다.

$$P = \frac{T_P}{T_P + F_P} \tag{6}$$

$$R = \frac{T_P}{T_P + F_N} \tag{7}$$

$$AP = \int_0^1 P(R)\,dR \tag{8}$$

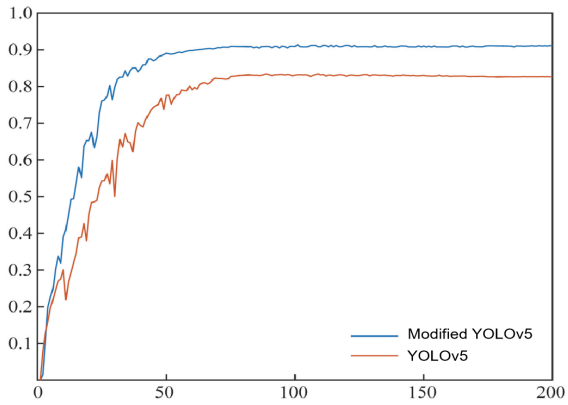
$$mAP = \frac{1}{n}\sum_{i=0}^{n} AP_i \tag{9}$$

상기 수식에서 P는 precision, R은 racall, T_P는 정확히 검출된 보행자 수(True positive), F_P는 오 검출된 보행자 수(False positive), F_N는 미 검출된 보행자 수(False negative)를 나타낸다.

4.3 실험 결과 분석

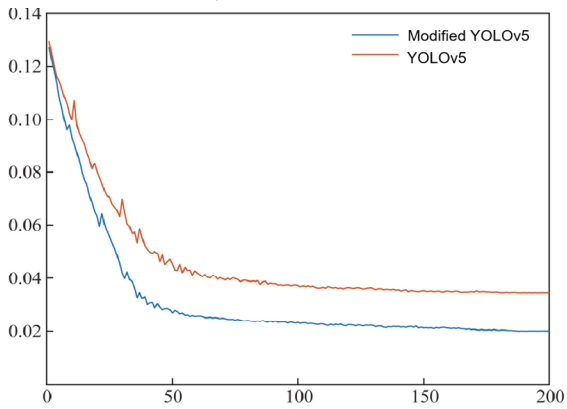
본 논문에서 제안한 모델의 성능을 검증하기 위하여 기존 YOLO v5s 모델을 기반으로 개선되기 전후의 mAP 및 loss 손실함수 수렴 그래프를 비교하였으며, 결과는 그림 4에서 나타내고 있다.

그림 4(a)에서 보여주는 바와 같이 개선된 모델의 mAP 곡선이 더 빠르게 증가하며, 최종적으로 0.916에서 안정되며, 기존 모델보다 8%정도 개선됨을 확인할 수 있다. 또한 실험을 통해 개선된 YOLO v5의 더욱 안정적인 성능과 개선된 검출 정확도를 확인할 수 있었다. 그림 4(b)에서는 손실함수의 수렴도를 보여준다. epoch는 총 200 수행하였으며, 50 정도에서 안정적으로 변화하는 것을 확인할 수 있다. 최종적으로 loss 값이 0.02에서 수렴되며, 기존 모델의 0.033보다 0.013정도 개선된 것을 확인할 수 있다.



(a) mAP 결과 비교

(a) Comparison of mAP results



(b) 손실함수 결과 비교

(b) Comparison of mAP results

그림 4. YOLO v5 학습 결과 비교

Fig. 4. Comparison of results YOLO v5 training

개선된 모델의 성능을 검증하기 위하여 CrowdHuman Dataset[11]에서 실험을 통해 기존 YOLO v5와 본 논문에서 개선한 방법을 비교 분석하였다. 표 1에서 보여주는 바와 같이 어텐션 네트워크와 soft NMS를 추가했을 때의 검출율을 기존 모델과 비교하였을 때 표 1에서와 같이 두 가지를 동시에 적용한 결과가 68.6%로 가장 높은 것으로 기존 모델보다 2.3% 개선된 것을 확인할 수 있다.

표 1. mAP 결과 비교

Table 1. Comparison results of mAP

Methods	mAP(0.5)	mAP(0.9)
YOLO v5s	66.3%	39%
YOLO v5s + soft NMS	67.7%	39.1%
proposal method	68.6%	39%

또한 본 논문에서 제안하는 모델의 성능을 검증하기 위하여 실제 CCTV 카메라를 연결하여 실시간 영상에서의 검출율을 분석하였다. 그림 5는 실제 영상에서의 검출 결과를 보여준다. 실시간 영상에서의 가장 중요한 성능지표인 FPS는 77로 실시간 처리가 충분한 것으로 확인되었다.

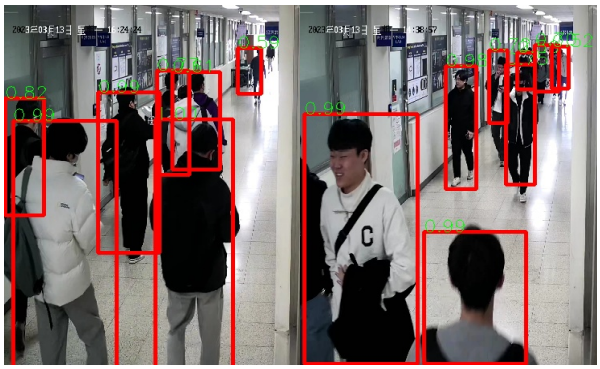


그림 5. 제안 방법의 검출 결과
Fig. 5. Results of detection based on proposed method

V. 결 론

본 논문에서 인파 밀집환경에서 보행자 검출은 기타 장애물에 의한 보행자 가려짐뿐만 아니라 보행자 간의 겹침으로 인해 정확도가 저하되거나 오검출, 미검출 등 문제를 해결하기 위하여 기존 YOLO v5를 기반으로 어텐션 네트워크 기반 특징 융합을 통해 정확도 개선 작업을 수행하였다. 또한 기존의 NMS는 실제로 존재하는 객체를 제거하여 mAP 가 낮아지는 문제점을 해결하고자 soft NMS를 적용하여 mAP를 향상시킴과 동시에 계산 부담을 줄여 실시간 영상 분석에 적용할 수 있는 모델로 구성하였다. 제안하는 방법을 CrowdHuman 데이터셋에서의 실험을 통해, 실시간 검출 테스트 결과 mAP 2.3%, loss 0.013 개선하였음을 보여주었다. 또한 실시간 영상에서 FPS 77를 확보할 수 있었다. 추후 연구에서 시스템 상용화를 위한 정확도 개선과 연산량 등에 대한 개선방안을 연구하고, 추적 알고리즘을 결합하여 인파의 동향 변화도 분석 가능한 방법을 연구할 계획이다.

References

[1] A. Brunetti, D. Buongiorno, G. F. Trotta, and V. Bevilacqua, "Computer vision and deep learning techniques for pedestrian detection and tracking: A survey", Neurocomputing, Vol. 300, pp. 17-33, Jul. 2018. https://doi.org/10.1016/j.neucom.2018.01.092.

[2] R. Girshick, F. Iandola, T. Darrell, and J. Malik, "Deformable part models are convolutional neural networks", In Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 437-446, Jun. 2015. https://doi.org/10.48550/arXiv.1409.5403.

[2] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks", Advances in neural information processing systems, Vol. 39, No. 6, pp. 1137-1149, Jun. 2017. https://doi.org/10.1109/TPAMI.2016.2577031.

[3] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn", In Proceedings of the IEEE international conference on computer vision, Venice, Italy, pp. 2961-2969, Oct. 2017. https://doi.org/10.1109/ICCV.2017.322.

[4] P. Jiang, D. Ergu, F. Liu, Y. Cai, and B. Ma, "A Review of Yolo algorithm developments", Procedia Computer Science, Vol. 199, pp. 1066-1073, Feb. 2022. https://doi.org/10.1016/j.procs.2022.01.135.

[5] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg. "Ssd: Single shot multibox detector", In European conference on computer vision, Vol. 9905, pp. 21-37, Sep. 2016. https://doi.org/10.1007/978-3-319-46448-0_2.

[6] M. Hussain, "YOLO-v1 to YOLO-v8, the Rise of YOLO and Its Complementary Nature toward Digital Manufacturing and Industrial Defect Detection", Machines, Vol. 11, No. 7, pp. 677, Jun. 2023. https://doi.org/10.3390/machines11070677.

[7] M. K. Cho, M. J. Kim, J. H. Kim, J. W. Kim, B. S. Hwang, S. W. Lee, J. H. Seon, and J. Y. Kim, "A Deep Learning-Based Image Recogntion Model

for Illegal Parking Enforcement", Journal of IIBC, Vol. 24, No. 1, pp. 59-64, Feb. 2024. https://doi.org/10.7236/JIIBC.2024.24.1.59.

[8] Y. S. Kwon, S. Y. Hwang, D. J. Shin, and J. J. Kim, "A Study on Application Method of Contour Image Learning to improve the Accuracy of CNN by Data", Journal of IIBC, Vol. 22, No. 4, pp. 171-176, Aug. 2022. https://doi.org/10.7236/JIIBC.2022.22.4.171.

[9] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design", In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, pp. 13713-13722, Jun. 2021. https://doi.org/10.48550/arXiv.2103.02907.

[10] J. Han and D. S. Kang, "Cos: An Emphasized Smooth Non-Monotonic Activation Function Consisting of Sigmoid for Deep Learning", Journal of KIIT, Vol. 19, No. 1, pp. 1-9, Jan. 2021. https://doi.org/10.14801/jkiit.2021.19.1.1.

[11] S. Shao, Z. Zhao, B. Li, T. Xiao, G. Yu, X. Zhang, and J. Sun, "Crowdhuman: A benchmark for detecting human in a crowd", arXiv:1805.00123, Apr. 2018. https://doi.org/10.48550/arXiv.1805.00123.

저자소개

임 혜 연 (Hye-Youn Lim)

2013년 : 동아대학교 전자공학과(공학박사)
2013년 ~ 현재 : 동아대학교 전자공학과 조교수
관심분야 : 영상처리, 패턴인식, 인공지능

안 명 수 (MingShou An)

2016년 : 동아대학교 전자공학과(공학박사)
2017년 ~ 현재 : 중국 서안공업대학교 컴퓨터공학과 강사
관심분야 : 영상처리, 패턴인식, 인공지능

강 대 성 (Dae-Seong Kang)

1994년 : Texas A&M 대학교 전자공학과(공학박사)
1995년 ~ 현재 : 동아대학교 전자공학과 교수
관심분야 : 영상처리, 패턴인식, 인공지능