실시간 영상 기반 신경망을 이용한 마스크 착용 감지 시스템

고건혁, 최성진, 송도훈, 박종일I

한양대학교 컴퓨터소프트웨어학과
gogn93@hanyang.ac.kr, csj7480@hanyang.ac.kr, dohoons@hanyang.ac.kr,
jipark@hanyang.ac.kr

Face Mask Detection using Neural Network in Real Time Video Surveillance

요 약

본 논문에서는 합성곱 신경망을 활용하여 영상에서 마스크 착용 및 미착용 상태를 탐지하는 방법을 제안한다. 코로나바이러스감염증-19(COVID-19)의 유행에 따라 감염 및 확산방지를 위해 마스크 정상적 착용이 요구되는데 몇몇 사람들은 이를 지키지 않고 있으며 현재의 감시 시스템은 입구에서 마스크 착용 여부를 검사하는 방식으로 작동될 뿐 공간에 입장한 다음 착용 여부를 알 수 없다. 제안하는 방법은 합성곱 신경망을 통해 영상에서 얼굴을 탐지하여 얻은 데이터를 이용하여 다수사람들의 마스크 착용 및 미착용 상태를 판별하는 방법으로 설계하였다.

1. 서론

일부 국가 및 지역을 제외한 전 세계가 코로나바이러스감염증-19(COVID-19)의 확산 및 지속으로 고통받고 있다. 코로나바이러스 장기화에 따라 세계보건기구는 국제적 공중보건 비상사태를 선포하였고 대다수 국가에서도 해당 바이러스 확산을 대응 및 예방하기 위해 공공장소에서 마스크 착용이 의무화되었다. 하지만 몇몇 사람들이 마스크를 착용하지 않거나 비정상적으로 착용하고 있으며 현재의 감시 시스템으로 관리하기엔 사람의 인력으로는 어려움이 많다.

본 논문에서는 실시간 영상 데이터를 CNN을 이용하여 마스크를 정상적으로 착용하였는지 탐지하는 알고리즘을 제안한다.

2. 관련 연구

2.1 CNN(Convolution Neural Network)

합성곱 신경망(CNN, Convolutional Neural Networks)은 이미지 인식 및 분류 등의 컴퓨터비전 분야에서 주로 사용되는 기술이다. 최초로 LeNet[1]을 통해 제안되었고 AlexNet[2]이 2012년 ILSVRC[3]를 통해 발전되었으며, 2013년

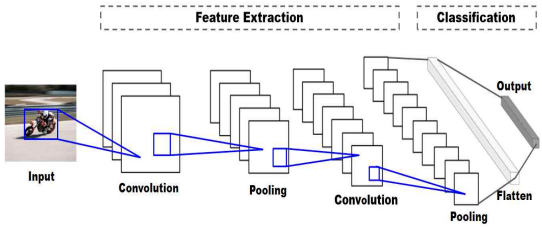

그림 1. CNN 구조

ZFNet[4], 2014년 VGG[5], 2015년 지금도 널리 사용되는 ResNet[6] 등 크게 발전하였다. CNN은 이미지 공간 정보를 유지한 상태로 학습이 가능한 기술로 이미지 특징 추출을 위해 합성곱 커널을 사용한다. CNN은 세 가지 계층으로 구성되어 있다. 들어온 이미지의 특징을 학습할 합성곱(Convolution) 계층과 해당 이미지의 차원을 줄일 풀링(Pooling) 계층, 마지막으로 이미지 분류를 위한 완전 연결(FC, Fully-Connected) 계층으로 그림 1.이 이를 나타낸 CNN의 구조이다.

2.2 YOLO(You Only Look Once)

YOLO(You Only Look Once)는 CNN을 사용하여 객체 탐지(Object Detection)하는 방법의 하나로 이미지 내부의 바운딩 박스(Bounding

I : 교신저자

Box)와 클래스 확률(Class Probability)을 단일 회귀 문제(Single Regression Problem)로 간주하여 이미지를 한 번만 보고 객체를 검출할 수 있다. 덕분에 YOLO는 다른 객체 탐지 모델과 비교해서 빠른 속도를 보여주며 이미지 전체를 보는 방식으로 객체에 대한 일반적인 특징을 학습하므로 속도 대비 높은 성능을 보여준다. YOLO에 대한 자세한 설명 및 구성은 원 논문 참조[7].

3. 제안하는 방법 및 실험 설계

실시간 영상 입력 데이터를 이용하여 마스크를 감지해야 하므로 우선 들어온 영상을 프레임 단위의 이미지 데이터로 받고 이를 YOLO를 통해 얼굴을 감지하여 마스크 착용 여부 및 정상 착용인지 판별하는 방법을 사용한다.

학습에 사용될 데이터는 Kaggle에서 제공하는 데이터셋[8]과 Google 검색 및 Youtube와 같은 비디오 스트리밍 사이트를 통해 직접 수집하여 사용하였다.

본 논문에서 제안하고자 하는 시스템에서 필요로 하는 객체는 사람 얼굴의 마스크 착용, 미착용 여부이므로 각 객체를 클래스로 바운딩 박스를 통해 라벨링 했다. 또한, 마스크를 코를 가리지 않는 등의 부적절하게 착용한 경우를 고려하여 3개의 클래스로 분류하여 라벨링 했는데 마스크 착용, 마스크 미착용, 마스크 비정상 착용으로 구분하였으며 콧구멍과 입을 기준으로 마스크 착용은 둘 다 가린 경우이고 마스크 미착용은 둘 다 가리지 않거나 아예 착용하지 않은 경우이며 마지막으로 마스크 비정상 착용은 콧구멍만 가리지 않은 경우를 칭한다. 입만 가리지 않는 경우는 마스크 미착용 외에는 찾아볼 수 있는 상황 및 데이터가 없어서 포함하지 않았으며 분류 예시는 그림 2.와 같다. 그리고 기존의 데이터셋[8]에서는 마스크를 올바르지 않게 착용한 얼굴이 없으므로 클래스를 추가하여 검수하였다.

데이터셋 정제과정에서 성능을 고려하여 다소 거리가 있거나 초점이 흐릿한 객체의 경우는 블랙 박스로 마스킹하여 학습에 방해되지 않도록

그림 2. 클래스 분류 예시

그림 2. 라벨링 된 이미지 예시

표 1. 학습 데이터셋

번호	종류	클래스	개수
0	마스크 착용	mask	7909
1	마스크 미착용	nomask	1290
2	마스크 비정상 착용	jaw_mask	394

하였다. 한 이미지의 객체 수는 최소 1명에서 최대 9명까지의 얼굴을 처리할 수 있도록 구성하였다. 학습 데이터셋 및 라벨링 된 이미지 예시는 각각 표 1.과 그림 3.와 같다.

본 논문에서는 실시간 영상을 고려하여 객체 탐지 모델로 YOLOv4[7]를 사용했다. 이미지 입력을 608x608해상도로 다운샘플링을 하고 출력은 바운딩 박스의 중앙 (x,y)좌표, 너비와 높이(w,h) 클래스(c) 정보이다. 학습 하이퍼파라미터로 학습률은 0.0013, 배치사이즈와 미니 배치사이즈는 각각 64, 1, 이터레이션은 9200을 사용했다. 데이터 증강 방법은 랜덤 리사이즈, 플립, 로테이션, 모자이크 기법을 사용하였다. 학습 데이터 클래스간의 불균형을 고려하여 loss마다 0.037, 0.225, 0.738 비율로 곱해서 최종적인 loss를 계산하였다.

4. 실험 결과 및 분석

YOLO[7] 성능을 확인하기 위해 71장의 이미지를 평가 데이터셋으로 이용한다. 표 2.는 평가 데이터셋에 대하여 클래스마다 AP, TP, FP와 전체 클래스의 mAP, Precision, Recall, F1-Score 값을 보여준다. 표 2.에서 mask는 마스크 착용, nomask는 마스크 미착용, jaw_mask는 비정상적인 마스크 착용을 나타낸다.

표 2.에서 알 수 있듯 각 클래스에서 전체적으로

표 2. 평가 데이터셋에 대한 모델 성능.

각 클래스			
	AP	TP	FP
mask	0.9641	412	131
nomask	0.8962	56	35
jaw_mask	0.8565	17	16
전체 클래스			
mAP	Precision	Recall	F1-Score
0.9056	0.7271	0.7650	0.7456

표 3. 시스템 환경

기기	사양
CPU	AMD Ryzen 9 5900HS
Memory	16GB
GPU	NVIDIA GeForce RTX 3080 Laptop
GPU Memory	16GB

그림 4. 학습된 모델로 이미지 탐지 결과

높은 AP 성능이 나왔다. 그림 4.은 YOLO[7]를 사용하여 얼굴 데이터를 탐지한 결과를 보여준다. 표 3.의 시스템 환경으로 구동 시 해당 이미지로부터 마스크 착용 및 미착용을 탐지까지 23.1밀리-초가 걸렸으며 높은 정확도를 보여줌을 알 수 있다. 해당 그림에서 객체마다 씌워진 바운딩 박스에 판별한 클래스 명칭과 정확도 수치가 적혀있다.

학습된 모델의 성능은 SCV[9] 모델과 비교를 이용하여 평가된다. 제안하는 모델이 mAP 성능에서 SCV모델에 비해 높은 성능을 보인다. 학습된 모델의 자세한 성능은 표 4.와 같다. 표 4.에

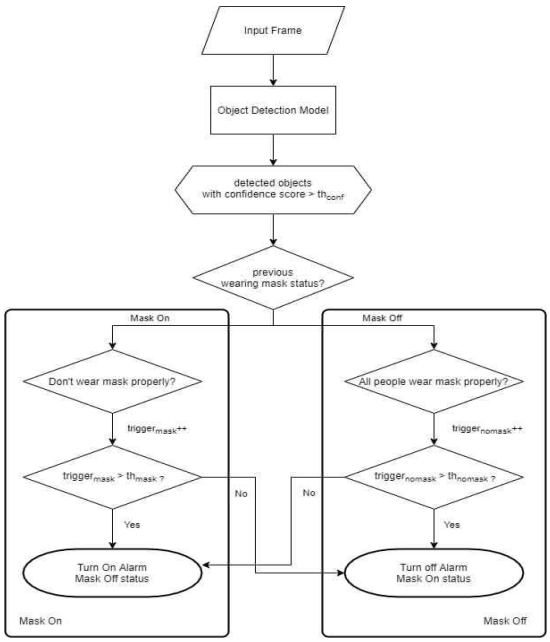

그림 5. 시스템 알고리즘

표 4. 학습된 모델(Ours)과 SCA[9] 모델의 성능 비교

	mAP	Precision	Recall	F1 Score
M1	0.7864	0.7693	0.8143	0.7893
M2	0.7970	0.7864	0.8219	0.8030
M3	0.8284	0.7995	0.8600	0.8274
SCA	0.8587	0.7962	0.9062	0.7785
Ours	**0.9056**	0.7271	0.7650	0.7456

서 Our는 본 논문에서 제안한 모델, M1, M2, M3는 각각 ResNet-18, ResNet-50, Resnet-101을 사용한 Faster R-CNN 모델로 진행하였다.

본 논문은 그림 5.와 같이 입력 이미지에 대해 객체 탐지 모델이 객체를 찾으면 해당 객체가 마스크를 썼는지 판별한다. 이후 시스템의 상태에 따라 나뉘어 동작한다. 기본적으로 마스크 착용(Mask On) 상태에서 시작한다. 마스크 착용 상태면 찾은 객체가 마스크를 미착용하거나 비정상 착용했는지 확인하고 해당 시 카운트하여 일정 카운트 이상일 때 마스크 미착용(Mask Off) 상태로 전환하면서 카운트를 초기화 및 경고 알람을 작동시킨다. 마스크 미착용 상태면 찾은 객체가 마스크를 착용했는지 확인하고 해당 시 카운트하여 일정 카운트 이상일 때 마스크 착용 상태로 전환하면서 카운트 초기화 및 경고 알람을 끈다. 그림 6.은 마스크 착용 상태에서 마스크 미착용 상태로 전환되는 시점을 보여준다.

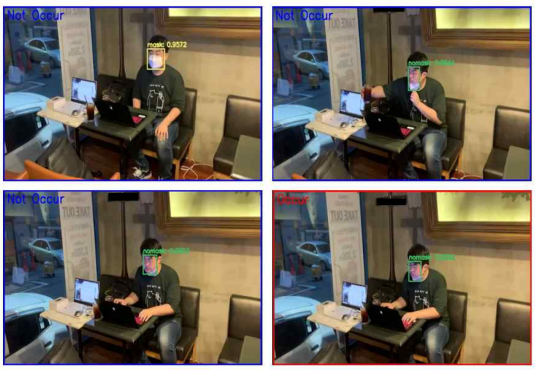

그림 6. 마스크 착용 상태(Not Occur) → 마스크 미착용 상태(Occur) 전환 시점 예시

5. 결론

코로나바이러스가 지속함에 따라 거리두기 단계도 개편되어 대한민국 등 여러 나라에서 위드코로나를 시행하기 시작하며 더욱 마스크 착용이 중요시되고 있으며 사람들이 마스크를 정상적으로 착용하였는지 관리하기엔 사람의 인력으로 어려움이 많다.

그래서 본 논문에서는 실시간 영상 기반 합성

곱 신경망을 이용한 마스크 탐지 시스템을 제안하였다. 실험 결과 실시간 영상에서 마스크 탐지에 높은 성능을 보였으며 이를 이용한 알고리즘을 통해 감지 시스템이 잘 동작할 수 있음을 확인하였다. 제안한 시스템은 공공장소에서 CCTV와 같은 실시간 영상 촬영 가능한 환경에서 마스크 착용 여부를 감지할 때 사용할 수 있으며 이를 통해 착용하지 않았거나 적절하게 착용하지 않는 사람들을 관리자가 파악하고 대처 및 제재하여 코로나의 확산방지에 많은 도움이 될 것으로 기대한다.

감사의 글

본 연구는 과학기술정보통신부 및 정보통신기획평가원의 SW중심대학지원사업의 연구결과로 수행되었습니다. (2016-0-00023)

5. 참고문헌

[1] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86(11), 2278-2324.

[2] Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems, 25, 1097-1105.

[3] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, and L. Fei-Fei “ImageNet Large Scale Visual Recognition Challenge.” IJCV, 2015

[4] Zeiler, M. D., & Fergus, R. (2014, September). Visualizing and understanding convolutional networks. In European conference on computer vision (pp. 818-833). Springer, Cham.

[5] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1-9).

[6] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 770-778).

[7] Bochkovskiy, A., Wang, C. Y., & Liao, H. Y. M. (2020). Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934.

[8] https://www.kaggle.com/datasets

[9] Kang, J., & Gwak, J. (2021). Adaptive Face Mask Detection System based on Scene Complexity Analysis. Journal of the Korea Society of Computer and Information, 26(5), 1-8.