JKSCI

한국컴퓨터정보학회논문지
Journal of The Korea Society of Computer and Information
Vol. 26 No. 5, pp. 1-8, May 2021
https://doi.org/10.9708/jksci.2021.26.05.001

Adaptive Face Mask Detection System based on Scene Complexity Analysis

Jaeyong Kang*, Jeonghwan Gwak*

*Post-Doc., Dept. of Software, Korea National University of Transportation, Chungju, Korea
*Professor, Dept. of Software, Korea National University of Transportation, Chungju, Korea

[Abstract]

Coronavirus disease 2019 (COVID-19) has affected the world seriously. Every person is required for wearing a mask properly in a public area to prevent spreading the virus. However, many people are not wearing a mask properly. In this paper, we propose an efficient mask detection system. In our proposed system, we first detect the faces of input images using YOLOv5 and classify them as the one of three scene complexity classes (Simple, Moderate, and Complex) based on the number of detected faces. After that, the image is fed into the Faster-RCNN with the one of three ResNet (ResNet-18, 50, and 101) as backbone network depending on the scene complexity for detecting the face area and identifying whether the person is wearing the mask properly or not. We evaluated our proposed system using public mask detection datasets. The results show that our proposed system outperforms other models.

▸ **Key words**: Artificial intelligence, Machine learning, Object detection, Deep learning, Mask detection, COVID-19

[요 약]

코로나바이러스-19(COVID-19)의 대유행에 따라 전 세계 수많은 확진자가 발생하고 있으며 국민을 불안에 떨게 하고 있다. 바이러스 감염 확산을 방지하기 위해서는 마스크를 제대로 착용하는 것이 필수적이지만 몇몇 사람들은 마스크를 쓰지 않거나 제대로 착용하지 않고 있다. 본 논문에서는 영상 이미지에서의 효율적인 마스크 감지 시스템을 제안한다. 제안 방법은 우선 입력 이미지의 모든 얼굴의 영역을 YOLOv5를 사용하여 감지하고 감지된 얼굴의 수에 따라 3가지의 장면 복잡도(Simple, Moderate, Complex) 중 하나로 분류한다. 그 후 장면 복잡도에 따라 3가지 ResNet(ResNet-18, 50, 101) 중 하나를 기반으로 한 Faster-RCNN을 사용하여 얼굴 부위를 감지하고 마스크를 제대로 착용하였는지 식별한다. 공개 마스크 감지 데이터셋을 활용하여 실험한 결과 제안한 장면 복잡도 기반 적응적인 모델이 다른 모델에 비해 가장 성능이 뛰어남을 확인하였다.

▸ **주제어**: 인공지능, 딥러닝, 기계학습, 객체 감지, 마스크 감지, 코로나바이러스-19

• First Author: Jaeyong Kang, Corresponding Author: Jeonghwan Gwak
*Jaeyong Kang (kjysmu@ut.ac.kr), Dept. of Software, Korea National University of Transportation
*Jeonghwan Gwak (jgwak@ut.ac.kr), Dept. of Software, Korea National University of Transportation
• Received: 2021. 02. 16, Revised: 2021. 04. 22, Accepted: 2021. 05. 03.

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

I. Introduction

코로나바이러스-19(COVID-19)가 전 세계로 확산함에 따라서 공공장소에서 올바르게 마스크를 착용하는 것이 어느 때보다 중요해졌다. 또한, 과학자들이 마스크를 착용하는 것이 코로나 전파를 차단하는 데 매우 효율적이라는 것을 증명하였다. 코로나바이러스가 발병한 지 6개월도 되기 전에 총 188개국에서 5백만 명이 넘는 사람들이 바이러스에 감염되었다고 보고되었다. 전 세계 정부 관계자들이 코로나바이러스 전파와 확산을 막는 것에 많은 어려움을 호소하고 있다. 확산을 저지하기 위해서 여러 나라에서 사람들이 공공장소에서 마스크를 제대로 착용하는 것이 법으로 제정되었음에도 여전히 수많은 사람이 마스크를 착용하지 않거나 제대로 착용하지 않고 있다. 또한, 마스크를 착용하지 않거나 제대로 착용하지 않는 수많은 사람을 일일이 다 감시하기는 쉽지 않다.

본 연구에서는 영상 이미지로부터 사람들이 얼굴에 마스크를 제대로 착용했는지 효율적으로 식별하는 장면 복잡도 기반 적응적 얼굴 마스크 감지 시스템을 제안한다. 여기서 장면 복잡도란 하나의 영상 이미지에서 보이는 사람들의 숫자로 정의된다. 여기서 핵심 아이디어는 딥러닝 기반 특징 추출 모델인 ResNet[5]의 경우 사용된 계층의 수에 따라 ResNet-18, 50, 101 등으로 나뉘는데 계층이 많은 모델의 경우는 장면 복잡도가 큰 이미지의 특징을 잘 추출하며 반대로 계층의 수가 적은 모델의 경우는 장면 복잡도가 작은 이미지의 특징을 잘 추출한다. 즉 우리가 제안 방법은 우선 이미지의 장면 복잡도를 YOLOv5 알고리즘[6]을 사용하여 3단계로 분류(Simple, Moderate, Complex)하고 장면 복잡도에 따라 각각 다른 계층을 가진 ResNet 네트워크를 기반으로 한 Faster-RCNN을 사용하여 사람의 얼굴 부위를 감지하여 마스크를 제대로 착용하였는지 식별한다. 웹상에 공개된 마스크 감지 데이터셋을 활용하여 실험한 결과 제안한 장면 복잡도 기반 적응적인 모델이 다른 모델에 비해 가장 성능이 뛰어남을 확인하였다.

본 논문의 구성으로 2장에서는 본 연구에 활용되고 있는 기존의 물체 감지 기술들을 소개한다. 3장에서는 제안한 마스크 감지 시스템에 대해서 자세히 기술한다. 4장에서는 웹상에 공개된 마스크 감지 데이터셋을 가지고 여러 가지 모델의 성능을 측정한 결과를 보여준다. 마지막으로 5장에서 결론을 통하여 활용방안을 제시하고 마무리한다.

II. Related Work

2.1 Convolutional Neural Networks (CNN)

합성곱 신경망(Convolutional Neural Networks, CNN)은 이미지 분류 및 객체 인식과 같은 컴퓨터 비전 관련 작업에서의 매우 중요한 기술로 여겨지며 그 이유로는 공간적 특징을 잘 잡아내고 또한 계산량이 적기 때문이다. CNN은 기존 영상 이미지에서 높은 수준의 특징을 추출하기 위해 합성곱 커널을 사용한다. 하지만 CNN 구조를 어떻게 더 잘 설계할 것인가에 대해서는 현재까지도 많은 연구가 진행되고 있다. CNN은 세 가지 빌딩 블록으로 구성되어 있다. 첫 번째는 특징을 학습하기 위한 합성곱 계층이고 두 번째는 입력 이미지의 차원을 줄이기 위한 맥스 풀링(max-pooling) 계층, 그리고 세 번째는 입력 이미지를 어떤 한 클래스로 분류하기 위한 완전 연결(Fully-connected, FC) 계층이다. Fig. 1은 CNN의 아키텍처를 보여준다.

이러한 CNN 기반의 객체 인식 기술은 이미지넷 대회라고 잘 알려진 ILSVRC(ImageNet Large Scale Visual Recognition Challenge)[2]를 통해 발전되었으며, 2012년 AlexNet[1], 2014년 VGGNet[3], GoogLeNet[4]과 2015년 ResNet[5] 등 시각 인식 문제에서 괄목할 만한 성장을 기록하게 되었다.

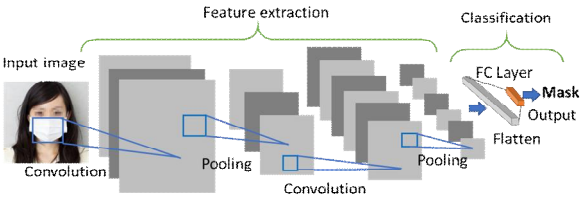

Fig. 1. CNN architecture

2.2 Object Detection

전통적인 물체 감지 시스템은 다중 단계 처리를 사용한다. 잘 알려진 감지 알고리즘 중 하나는 Viola-Jones이며 실시간 처리가 가능하다[12]. 이 알고리즘은 적분 영상(Integral Image)기법과 함께 Haar 특징 기술자(descriptor)를 사용하여 이미지의 특징을 추출하고 그중 유용한 특징을 선택해서 물체를 감지한다. 하지만 계산량이 복잡하여 느리다는 단점이 있다. 최근 몇 년간 직접 수동으로 정한 특징(handcrafted)을 사용하기보다는 딥러닝을 기반으로 자동으로 추출한 특징을 사용한 객체 감지 기법이 월등히 좋은 성능을 보여주었다. 딥러닝 기반 객체

검출 기법 중 CNN을 기반으로 한 객체 검출 알고리즘인 R-CNN(Region-based Convolutional Neural Network[7])이 초기에 제안되었다. 이 모델은 후보 영역 생성을 위해 선택적 탐색(Selective search)기법을 도입하였으며 각 후보 영역별로 CNN을 거쳐서 나온 특징을 기반으로 객체를 검출한다. 하지만 모든 후보 영역에 대해 CNN 연산을 하므로 학습하고 추론하는 데 많은 시간이 소요되어서 실시간 처리가 불가능하다는 단점이 존재한다. 이를 개선하기 위해 모든 후보 영역 대신 입력 이미지에 한 번만 CNN 연산을 한 뒤 나온 특징 맵(feature map)의 후보 영역에서 RoI Pooling기법으로 객체 검출을 위한 특징을 추출하는 모델인 Fast R-CNN(Fast Region-based Convolutional Neural Network[8])이 제안되었다. 더 나아가 선택적 탐색기법 대신 CNN을 사용하여 후보 영역을 생성하는 RPN(Region Proposal Network)의 도입을 통해 처리 속도 및 성능을 개선한 모델이 Faster R-CNN이다. Fig. 2는 Faster R-CNN의 아키텍처를 보여준다. 이 모델은 우선 CNN을 통해 추출한 입력 이미지의 특징을 RPN에 전달하여 객체가 존재할 만한 영역인 RoI를 계산한다. 그 이후 RoI pooling을 거쳐 나온 특징을 가지고 최종적으로 객체 검출 및 분류를 진행한다.

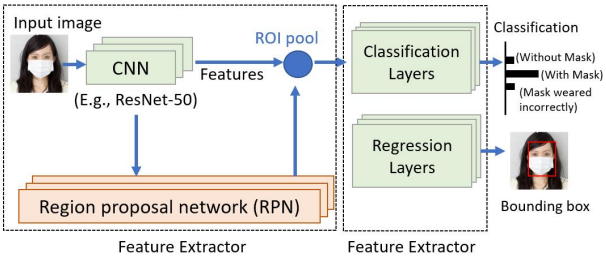

Fig. 2. Faster R-CNN Architecture

2.3 Transfer Learning

CNN은 일반적으로 대용량의 데이터셋을 사용하는 것이 그렇지 않은 경우보다 좋은 성능을 보여준다. 전이 학습 기법은 대용량의 데이터셋을 확보하기 어려울 때 사용할 수 있는 기법의 하나로 사용된다. 전이 학습의 원리는 Fig. 3에서 보이듯이 대용량의 데이터셋인 ImageNet[1]을 통해 미리 학습된 모델의 파라미터 가중치 값을 불러와서 얼굴 마스크 감지하는 작업에 활용될 수 있다. 최근 몇 년 사이 전이 학습은 화물 검색, 의료영상 분류 및 분할 등 다양한 분야에서 성공적으로 사용됐다[13-17]. 전이 학습을 통해 딥러닝 모델을 처음부터 학습하면서 긴 학습 시간을 단축할 수 있고 또한 많은 양의 데이터가 없이도 충분히 좋은 성능을 내는 모델을 생성할 수 있다[9,10].

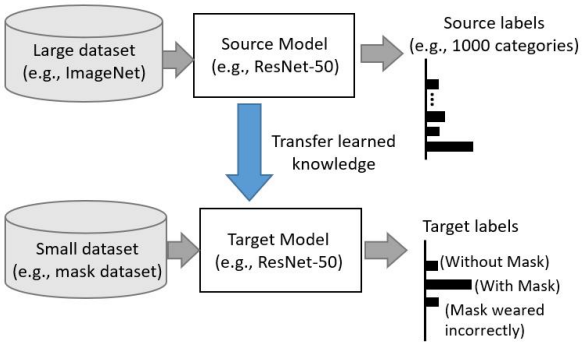

Fig. 3. Concept of transfer learning

III. The Proposed Method

제안한 시스템 아키텍처는 Fig. 4와 같다. 우선 입력 영상 이미지가 들어오면 장면 복잡도 분석 모듈을 통해 이미지를 먼저 분류한다. 얼굴이 하나만 있는 단순한 이미지로 분류가 되면 ResNet-18을 특징 추출 네트워크로 사용하는 Faster-RCNN 물체 감지 모델을 사용한다. 만약 얼굴이 2~5개로 감지된 이미지의 경우 ResNet-50을 사용한 Faster-RCNN 감지 모델을 사용하고, 얼굴이 6개 이상으로 감지된 이미지의 경우 계층의 수가 많은 ResNet-101을 사용한 Faster-RCNN 감지 모델을 사용한다. 또한, 전이 학습 기법을 사용해서 ImageNet 데이터셋으로 미리 학습한 모델이 가중치를 불러와 ResNet 파라미터를 초기화하는 데 사용되었다. ResNet 모델은 최종적으로는 Faster-RCNN을 통해 얼굴의 영역을 감지하고 마스크의 착용 여부 및 제대로 착용했는지를 감지하게 된다.

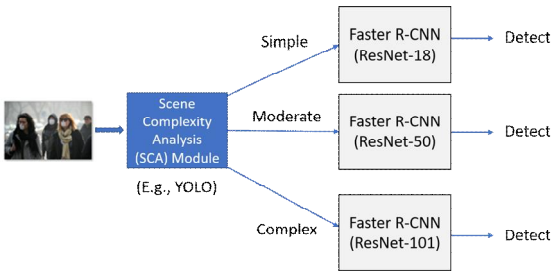

Fig. 4. Proposed method

3.1 Scene Complexity Analysis Module

장면 복잡도(Scene, Complexity Analysis, SCA) 모듈로서는 YOLOv5[6]를 사용하였다. YOLO는 한 단계만 사용해서 물체를 빠르게 감지하는 알고리즘으로 잘 알려져 있다. YOLO는 첫 출시 이후 지금까지 계속해서 버전이 상향되어왔고 현재는 객체 감지 속도 면에서 월등한 성능

을 보이는 다섯 번째 버전인 YOLOv5 가 출시되었다. 물론 YOLOv5만을 사용해서 마스크 착용 여부와 같은 분류 결과도 추출할 수 있지만, Faster R-CNN보다 정확도가 떨어진다는 단점이 존재하여 우리는 YOLO를 장면 복잡도를 계산하는 데에만 사용하였으며 각 장면 복잡도의 수준에 따라 다른 계층을 가진 특징 추출 네트워크를 사용한 Faster R-CNN을 감지 및 분류기로 사용하였다. 여기서 장면 복잡도는 사람들의 얼굴 숫자들로 구분이 되며, 사람의 얼굴이 하나만 있는 경우, 사람의 얼굴이 2~5개가 있는 경우, 사람의 얼굴이 6개 이상 있는 경우 총 3가지 경우를 분별하여서 복잡도를 Simple, Moderate, Complex로 분류하게 된다. Fig. 5는 YOLOv5를 사용해서 사람의 얼굴을 감지한 결과를 보여준다. 하나의 영상 이미지로부터 얼굴을 감지하는 데 총 0.008초가 걸렸고 결과에서 알 수 있듯이 YOLOv5가 사람의 얼굴 영역을 잘 감지함을 알 수 있다. 이 예시에서는 총 6개 이상의 객체가 감지되었기 때문에 장면 복잡도는 Complex로 분류가 된다.

Fig. 5. Face detection result using YOLOv5

3.2 Mask Detection System

장면 감지기를 통해 복잡도가 분류된 이미지를 입력값으로 받아서 Faster-RCNN을 사용하여 얼굴을 감지하고 마스크의 착용 여부 및 제대로 착용하였는지 판별하게 된다. 장면 복잡도가 Simple이면 ResNet-18을 특징 추출 네트워크로써 사용하였고, Moderate이면 ResNet-50을 사용하였고, Complex이면 ResNet-101을 사용하였다. 그 이유는 ResNet-18의 경우 계층의 수가 적고 따라서 비교적 단순한 이미지에 대해서 충분히 특징을 잘 추출하기 때문이고 반면에 계층의 수가 많은 ResNet-101의 경우 사람이 많은 이미지와 같은 복잡한 이미지에 대해서 유용한 특징을 잘 잡아내는 성질을 지니고 있기 때문이다. 우선 ResNet-18을 사용한 Faster-RCNN의 학습 과정에서는 복잡도가 Simple인 데이터셋, 즉 얼굴의 개수가 실제 하나인 이미지를 사용해서 학습하였고, ResNet-50을 사용한 Faster-RCNN의 경우는 복잡도가 Moderate인 이미지, 즉 얼굴의 개수가 실제 2개에서 5개 사이인 이미지를 사용해서 학습하였고 마지막으로 ResNet-101을 사용한 Faster-RCNN의 경우는 복잡도가 Complex인 데이터셋, 즉 얼굴의 개수가 6개 이상인 이미지를 가지고 학습을 하였다. 그 이후 실제 새로운 이미지가 들어왔을 때의 추론 과정에서는 장면 복잡도 모듈을 통해 복잡도가 구해진 이미지를 가지고 해당 복잡도에 맞게 학습이 된 ResNet 모델을 선택해서 최종적으로 새로운 이미지에 대한 얼굴 마스크 감지 및 분류가 이루어지게 된다.

Table 1. Face Mask Detection Dataset

Complexity	Train	Test	Total
Simple	264	65	329
Moderate	240	60	300
Complex	180	44	224
All	684	169	853

Fig. 6. Example images from our dataset

IV. Experiment

4.1 Dataset

데이터셋으로는 웹상에 공개된 얼굴 마스크 감지 데이터셋[11]을 사용하였다. 데이터셋은 총 3가지의 클래스(마스크 미착용, 마스크 제대로 착용, 마스크 제대로 미착용)로 구분되며 각 이미지에 포함된 사람의 개수에 따라 3가지 형태의 복잡도(Simple, Moderate, Complex)로 나누었다. Fig. 6에서 보이듯이 마스크를 착용했지만, 입과 코를 완전히 가리지 않는 경우 마스크 제대로 미착용으로 구분된다. Simple에 해당하는 데이터셋은 사람의 얼굴이 한 개만 있는 영상 이미지로 구성되며, Moderate에 해당하는 데이터셋은 사람의 얼굴이 2~5개 있는 영상 이미지, Complex에 해당하는 데이터셋은 사람의 얼굴이 6개 이상이 있는 영상 이미지로 데이터로 구성된다. 또한, 각 3가지 복잡도에 해당하는 데이터셋은 모델을 학습하기 위한 학습 데이터셋과 검증하기 위한 테스트 데이터셋으로서

8:2의 비율로 나뉜다. Table 1은 얼굴 마스크 감지 데이터셋에 대한 자세한 사항을 보여준다. 또한, Fig. 7은 데이터셋의 클래스별 분포를 보여준다.

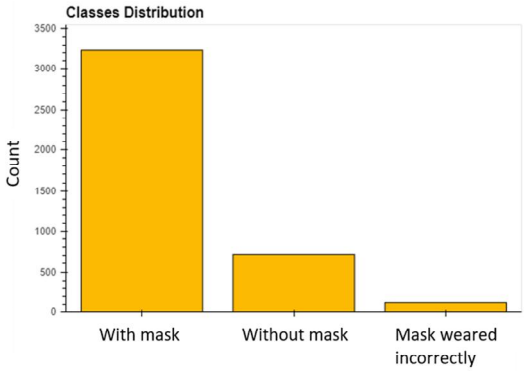

Fig. 7. Class distribution

4.2 Experimental setting

장면 복잡도 계산의 경우 YOLOv5를 사용하였고 마스크 감지 및 분류는 Faster-RCNN을 사용하였다. Faster-RCNN에서의 백본(Backbone) 네트워크로는 장면 복잡도에 따라 ResNet-18, ResNet-50, ResNet-101 총 3가지를 사용하였다. 백본 네트워크로 사용된 각각의 ResNet은 전이 학습의 한 과정으로서 방대한 이미지로 구성되는 이미지넷(ImageNet) 데이터셋을 통해 미리 학습되었다. 그런 후 ResNet 아키텍처의 합성곱 블록 총 5개 중 상단의 3개 블록(conv3~5)만 파인튜닝(Fine-tuning)을 할 수 있게끔 설정을 하였다. 예를 들어 Fig. 8은 파인튜닝한 ResNet-50 모델을 나타낸다. 학습을 위한 옵티마이저로는 SGD를 사용하였고 학습률은 0.005로 설정을 하였다. 총 100번의 Epoch를 돌며 ResNet-18, ResNet-50, ResNet-101을 기반으로 한 세 가지 Faster-RCNN 모델을 학습하였고 각 Epoch중 성능이 가장 잘 나온 모델을 저장하여 활용하였다. Table 2는 모델을 학습하는 데 사용된 시스템 환경을 나타낸다.

Table 2. System Environment

Item	Value
CPU	Intel i5-10400
Memory	32GB
GPU	NVIDIA Geforce RTX 3080
CUDA ver	10.2
Python ver	3.7
Pytorch ver	1.7.1

4.3 Evaluation metric

평가지표로는 클래스별로 AP(Average Precision), Precision, Recall, F1을 사용하였다. 또한, 클래스별로 계산한 각 지표의 점수를 평균을 낸 평균값인 mAP, mean Precision(mPrec), mean Recall(mRec), mean F1(mF1)도 계산하였다.

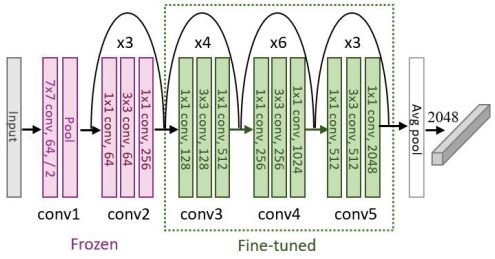

Fig. 8. Three fine-tuned blocks of ResNet-50 pretained on ImageNet

Table 3. Comparison between the model pretrained on ImageNet and the model without pre-training during 10 epoches

	without pre-trained model		with pre-trained model	
Epoch	Train loss	mAP	Moderate	mAP
1	0.1717	0.0617	0.3276	0.0000
2	0.1594	0.0072	0.1768	0.0445
3	0.1422	0.0819	0.1235	0.1810
4	0.1370	0.0818	0.1251	0.5344
5	0.1301	0.1825	0.1191	0.5912
6	0.1338	0.2463	0.1073	0.6398
7	0.1277	0.4438	0.0958	0.6512
8	0.1269	0.2429	0.0927	0.6537
9	0.1290	0.4576	0.0897	0.6463
10	0.1259	0.3793	0.0805	0.6294

Fig. 9. Mask detection result using our proposed system (green: with mask, orange: Mask weared incorrectly, red: without mask)

4.4 Results

Table 3은 전이 학습 과정으로서 이미지넷(ImageNet) 데이터셋을 통해 미리 학습한 ResNet을 기반으로 한 Faster-RCNN 모델과 전이 학습을 사용하지 않는 모델 2가지에 대해서 처음 10번의 Epoch 과정을 통한 모델 성

능 향상 정도를 비교한 결과를 나타낸다. Table 3의 결과에서 알 수 있듯이 전이 학습을 기반으로 한 모델이 그렇지 않은 모델에 비해 성능이 많이 좋아짐을 알 수 있었다. Table 4는 서로 다른 복잡도를 가진 데이터셋을 가지고 각각의 3가지 CNN 기반 특징 추출 네트워크(ResNet-18, ResNet-50, ResNet-101)를 사용한 Faseter-RCNN모델(M1,M2,M3) 및 제안한 장면 복잡도 모듈(SCA)을 사용한 모델의 각 클래스에 대한 AP, Precision, Recall, F1 값을 보여준다. 참고로 M1, M2, M3는 각각 Simple, Moderate, Complex 데이터셋을 통해 학습되었고 SCA가 잘 동작하여 해당 복잡도에서만 학습이 된 모델과 유사한 성능을 보이는지 관측하기 위해 설계되었다. 또한, Table 5~8은 각 모델에 대해 전체 클래스의 평균값을 나타내는 mAP, mPrecision, mRecall, mF1 값을 보여준다. Table 9는 전체 데이터셋을 가지고 YOLOv5만 사용해서 마스크를 감지하는 모델, 전체 데이터셋으로 학습된 M1, M2, M3 모델 및 제안한 장면 복잡도 모듈(SCA)을 사용해 감지하는 모델의 성능 비교를 보여준다. Fig. 9은 우리가 제안한 모델을 사용해서 마스크를 감지한 결과를 나타낸다. 결과 테이블에서도 알 수 있듯이 단순한 이미지만 사용한 데이터셋의 경우 ResNet-18을 사용한 모델이 다른 모델에 비해 성능이 좋았으며, 마찬가지로 SCA를 사용한 모듈도 ResNet-18과 유사한 성능을 보여주었다. 이와 마찬가지로 복잡도가 Moderate인 데이터셋을 사용하였을 때 ResNet-50, 복잡도가 Complex인 데이터셋을 사용하였을 때 ResNet-101가 다른 모델에 비해 좋은 성능을 나타내었고 SCA 모듈을 사용했을 때도 좋은 성능을 나타내는 각 모델과 유사한 성능을 보여주었다. 이는 SCA 모듈이 잘 작동하여서 이미지의 복잡도에 따라 최적의 ResNet을 잘 선택할 수 있다는 것을 의미한다. 참고로 복잡도가 Complex인 데이터셋에서는 마스크를 제대로 착용하지 않은 이미지가 매우 적을뿐더러 얼굴의 크기 또한 매우 작아서 Simple과 Moderate 데이터셋으로 각각 학습된 M1과 M2에선 얼굴을 전혀 감지하지 못하여 성능 지표 값이 0이 도출되었고, Complex 데이터셋으로 학습된 M3는 마스크를 제대로 착용하지 않은 얼굴에 대해 다 감지하여 Recall 및 AP의 지표 값이 1이 도출되었다. 또한, 전체 데이터셋에 대해서는 YOLOv5 모델만 사용하거나 단일 ResNet 모

Table 4. Results of each model using simple, moderate, complex, and all dataset (M1: Faster R-CNN with ResNet-18 as backbone, M2: Faster R-CNN with ResNet-50 as backbone, M3: Faster R-CNN with ResNet-101 as backbone, SCA: our proposed adaptive model using Scene Complexity Analysis (SCA) module. M1, M2, M3 are trained on simple, moderate, complex dataset, respectively).

Simple Dataset												
	Without Mask				With Mask				Mask weared incorrectly			
	AP	Prec	Rec	F1	AP	Prec	Rec	F1	AP	Prec	Rec	F1
M1	**1.0000**	0.9231	**1.0000**	0.9600	0.9692	0.8824	**0.9783**	0.9278	**0.6762**	**0.6667**	**0.8571**	**0.7500**
M2	**1.0000**	**1.0000**	**1.0000**	**1.0000**	0.9495	0.7333	0.9565	0.8302	0.0952	0.2222	0.2857	0.2500
M3	0.6830	0.6250	0.8333	0.7143	0.6409	0.4925	0.7174	0.5841	0.4286	0.6000	0.4286	0.5000
SCA	**1.0000**	0.9231	**1.0000**	0.9600	**0.9701**	**0.9000**	**0.9783**	**0.9375**	**0.6762**	**0.6667**	**0.8571**	**0.7500**
Moderate Dataset												
	Without Mask				With Mask				Mask weared incorrectly			
	AP	Prec	Rec	F1	AP	Prec	Rec	F1	AP	Prec	Rec	F1
M1	0.2074	0.6364	0.2188	0.3256	0.7387	0.8934	0.7517	0.8165	0.0476	0.0909	0.1429	0.1111
M2	**0.8898**	**0.7838**	**0.9062**	**0.8406**	0.9601	**0.9400**	0.9724	0.9559	**0.4286**	**0.7500**	**0.4286**	**0.5455**
M3	0.7740	0.7429	0.8125	0.7761	0.8754	0.8897	0.8897	0.8897	0.1837	0.1765	**0.4286**	0.2500
SCA	0.8342	0.7568	0.8750	0.8116	**0.9664**	0.9396	**0.9790**	**0.9589**	**0.4286**	0.6000	**0.4286**	0.5000
Complex Dataset												
	Without Mask				With Mask				Mask weared incorrectly			
	AP	Prec	Rec	F1	AP	Prec	Rec	F1	AP	Prec	Rec	F1
M1	0.0104	0.2500	0.0104	0.0200	0.2821	0.6745	0.3373	0.4497	0.0000	0.0000	0.0000	0.0000
M2	0.7276	**0.8902**	0.7604	0.8202	0.8545	**0.9246**	0.8679	**0.8954**	0.0000	0.0000	0.0000	0.0000
M3	0.8625	0.8318	**0.9271**	**0.8768**	**0.8826**	0.8863	**0.9009**	0.8936	**1.0000**	**0.6786**	**1.0000**	**0.8261**
SCA	**0.8638**	0.8302	0.9167	0.8713	0.8802	0.8881	0.8986	0.8933	**1.0000**	**0.6786**	**1.0000**	**0.8261**
All Dataset												
	Without Mask				With Mask				Mask weared incorrectly			
	AP	Prec	Rec	F1	AP	Prec	Rec	F1	AP	Prec	Rec	F1
M1	0.1408	0.7143	0.1429	0.2381	0.4476	0.7755	0.4829	0.5952	0.1466	0.2500	0.2188	0.2333
M2	0.7694	**0.8626**	0.8071	0.8339	0.8854	**0.9110**	0.8992	0.9051	0.1068	0.3333	0.1562	0.2128
M3	0.8262	0.7962	0.8929	0.8418	0.8602	0.8432	0.8829	0.8626	0.6823	0.5200	0.8125	0.6341
SCA	**0.8646**	0.8205	**0.9143**	**0.8649**	**0.9054**	0.9013	**0.9233**	**0.9122**	**0.8063**	**0.6667**	**0.8750**	**0.7568**

델만 사용한 Faster-RCNN보다 제안한 SCA 모듈을 사용한 모델이 가장 높은 성능을 보여줌을 확인하였다.

Table 5. mAP results

	Simple	Moderate	Complex	Total
M1	0.8818	0.3312	0.0975	0.2450
M2	0.6816	**0.7594**	0.5274	0.5872
M3	0.5842	0.6110	**0.9150**	0.7896
SCA	**0.8821**	0.7431	0.9147	**0.8587**

Table 6. mPrecision results

	Simple	Moderate	Complex	Total
M1	0.8240	0.5402	0.3082	0.5799
M2	0.6519	**0.8246**	0.6050	0.7023
M3	0.5725	0.6030	0.7989	0.7198
SCA	**0.8299**	0.7655	**0.7990**	**0.7962**

Table 7. mRecall results

	Simple	Moderate	Complex	Total
M1	**0.9451**	0.3711	0.1159	0.2815
M2	0.7474	**0.7691**	0.5428	0.6209
M3	0.6598	0.7102	**0.9427**	0.8628
SCA	**0.9451**	0.7609	0.9384	**0.9042**

Table 8. mF1 results

	Simple	Moderate	Complex	Total
M1	0.8793	0.4177	0.1566	0.3555
M2	0.6934	**0.7807**	0.5719	0.6506
M3	0.5995	0.6386	**0.8655**	0.7795
SCA	**0.8825**	0.7568	0.8636	**0.8446**

Table 9. Comparison between our proposed model (SCA) and M1, M2, M3 models trained on all dataset, YOLOv5 with 4 different models sizes, using all dataset

	mAP	mPrec	mRec	mF1
M1(all)	0.7864	0.7693	0.8143	0.7893
M2(all)	0.7970	0.7864	0.8219	0.8030
M3(all)	0.8284	0.7995	0.8600	0.8274
YOLOv5s	0.6992	0.8482	0.7252	0.7683
YOLOv5m	0.6627	0.7627	0.6838	0.7204
YOLOv5l	0.7491	0.8480	0.7710	0.8037
YOLOv5x	0.7189	**0.8737**	0.7346	0.7785
SCA	**0.8587**	0.7962	**0.9042**	**0.8446**

V. Conclusions

본 논문에서는 이미지 장면 복잡도에 따른 적응적 얼굴 감지 시스템을 제안하였다. 실험 결과 장면 복잡도를 사용한 제안한 모델이 장면 복잡도를 사용하지 않은 모델보다 높은 성능을 나타냄을 보여주었다. 제안한 시스템을 통해 여러 공공장소에서 사람들이 마스크를 제대로 잘 착용했는지 실시간으로 감지하여 마스크를 착용하지 않았거나 제대로 착용하지 않는 사람을 감지할 때 사진 및 위치 정보가 모니터링 요원에게 실시간으로 전송이 되어 벌금이나 관련 제재를 가함으로써 결과적으로 코로나의 확산을 저지하는 데 크게 이바지할 것으로 기대한다.

ACKNOWLEDGEMENT

This work was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (Grant No. NRF-2020R1I1A3074141), the Brain Research Program through the NRF funded by the Ministry of Science, ICT and Future Planning (Grant No. NRF-2019M3C7A1020406), and “Regional Innovation Strategy (RIS)” through the NRF funded by the Ministry of Education.

REFERENCES

[1] A. Krizhevsky, I. Sutskever, and G.E. Hinton, “Imagenet classification with deep convolutional neural networks,” Advances in Neural Information Processing Systems, Vol. 25, pp. 1097-1105, 2012.

[2] 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,” International Journal of Computer Vision, Vol. 115, pp. 211-252, 2015.

[3] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, and A. Rabinovich, “Going deeper with convolutions,” In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1-9, 2015.

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

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770-778, 2016.

[6] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, “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] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 580-587, 2014.
[8] R. Girshick, "Fast R-CNN," In Proceedings of the IEEE International Conference on Computer Vision, pp. 1440-1448, 2015.
[9] N. Tajbakhsh, J.Y. Shin, S.R. Gurudu, R.T. Hurst, C.B. Kendall, M.B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?," IEEE Transactions on Medical Imaging, Vol. 35, No. 5, pp. 1299-1312, 2016.
[10] S.J. Pan, and Q. Yang, "A survey on transfer learning," IEEE Transactions on Knowledge and Data Engineering, Vol. 22, No. 10, pp. 1345-1359, 2009.
[11] Mask dataset, https://makeml.app/datasets/mask.
[12] P. Viola, and M. Jones, "Rapid object detection using a boosted cascade of simple features," In Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, Vol. 1, pp. I-I, IEEE, 2001.
[13] S. Akçay, M.E. Kundegorski, M. Devereux, and T.P. Breckon, "Transfer learning using convolutional neural networks for object classification within x-ray baggage security imagery," In 2016 IEEE International Conference on Image Processing (ICIP), pp. 1057-1061, IEEE, 2016.
[14] H. Yang, S. Mei, K. Song, B. Tao, and Z. Yin, "Transfer-learning-based online Mura defect classification," IEEE Transactions on Semiconductor Manufacturing, Vol. 31, No. 1 pp. 116-123, 2018.
[15] I.M. Baltruschat, H. Nickisch, M. Grass, T. Knopp, and A. Saalbach, "Comparison of deep learning approaches for multi-label chest X-ray classification," Scientific Reports, Vol. 9, No. 1, pp. 1-10, 2019.
[16] S. Christodoulidis, M. Anthimopoulos, L. Ebner, A. Christe, and S. Mougiakakou, "Multisource transfer learning with convolutional neural networks for lung pattern analysis," IEEE Journal of Biomedical and Health Informatics, Vol. 21, No. 1, pp. 76-84, 2016.
[17] J. Kang, and J. Gwak, "Ensemble of instance segmentation models for polyp segmentation in colonoscopy images," IEEE Access, Vol. 7, pp. 26440-26447, 2019.

Authors

Jaeyong Kang received the Ph.D. degree in electrical engineering and computer science from the Gwangju Institute of Science and Technology (GIST), Gwangju, South Korea, in 2017. From 2018 to 2019, he was a research scientist at the Biomedical Research Institute, Seoul National University Hospital, Seoul, South Korea. He is currently a postdoctoral researcher at the Korea National University of Transportation (KNUT), Chungju, South Korea. His current research interests include deep learning, computer vision, natural language processing, agent-based information retrieval, semantic web, social media analysis, and recommender systems.

Jeonghwan Gwak received the Ph.D. degree in machine learning and artificial intelligence from Gwangju Institute of Science and Technology, Gwangju, Korea in 2014. From 2002 to 2007, he had worked for several companies and research institutes as a researcher and a chief technician. From 2014 to 2016, he worked as a postdoctoral researcher in GIST, and from 2016 to 2017 as a research professor. From 2017 to 2019, he was a research professor in Biomedical Research Institute & Department of Radiology at Seoul National University Hospital, Seoul, Korea. From 2019, he joined Korea National University of Transportation as an assistant professor, and he is the director of the Applied Machine Intelligence laboratory. His current research interests include deep learning, computer vision, signal and image processing, AIoT, evolutionary algorithms and optimization, fuzzy sets and systems, and relevant applications of biomedical and visual surveillance systems.