Dataset variation for masked face detection using YOLO-v5 method

By Register 3249

WORD COUNT 2880

TIME SUBMITTED 30-JAN-2023 11:18AM

PAPER ID 96034494

Dataset variation for masked face detection using YOLO-v5 method

ABSTRACT

The Covid-19 pandemic necessitates that health standards take precedence in companies and public areas that employ a large number of people. Normally, officers supervise the use of masks in public places, however, this can be accomplished through the use of computer vision. Face detection using a mask is proposed in this work utilizing the YOLO-v5 algorithm on a variety of datasets and resolutions. Three datasets were used: a face dataset with masks (M dataset), a synthetic dataset (S dataset), and a combined dataset (G dataset), with picture resolutions of 320 pixels and 640 pixels. The training results indicate that the training time is longer at a higher resolution, but the prediction results are excellent. The system test results indicate that the detection of facial images utilizing masks using the YOLO-v5 method is very good at a resolution of 640 pixels, with a detection rate of 99.2 percent for the G dataset, 98.5 percent for the S dataset, and 98.9 percent for the M dataset. The proposed S dataset can be utilized to conduct research in computer vision.

1. Introduction

Due of the Covid-19 outbreak, everyone is compelled to wear a mask [1]. Health procedures are the primary concern for companies and public venues that employ a large number of people due to the wide variety of community interactions [2, 3], which necessitates supervision in public spaces. Manual supervision, such as placing officers in public places such as airports or shopping centers to watch people wearing masks or not, is inefficient and promotes greater virus transmission. As a result, research on face detection while wearing masks is essential. Computer vision can be used to perform surveillance technology, particularly object detection [4, 5].

The usage of computer vision has increased significantly in recent years, with applications and automation systems now being developed that have great capabilities for detecting, distinguishing, and locating objects in images and videos in real time [4, 6, 7]. When compared to classification, detection is a more complicated process. Classification can distinguish objects in an image but cannot determine where the object is located in the image [8]. More importantly [9], if the image contains more than one item, but the feature extraction method combined with a classifier for face detection is claimed to be less than optimal [10].

Researchers have investigated object detection, particularly face detection, in public locations in real time [11-13] using a variety of methods, including the Viola-Jones method [11], Holistic Matching Methods [14], Local Binary Pattern [13], CNN [1] and several others. There are several methods for detecting faces that are based on the eyes, nose, and mouth [15, 16]. In order to recognize faces, a variety of constraints must be overcome [3, 17, 18], including light, face position, noise, as well as other factors. Face detection becomes even more difficult when people wear accessories, such as masks [1, 7]. Deep learning-based methods for face detection while wearing masks have been used in several investigations, such as those conducted by researchers [5] utilizing the YOLO-v5 method with datasets from the AIZOOTe team's FaceMaskDetection. Researchers [19] carried out a face detection study with a mask, employing the YOLO-v3 approach with a quicker R-CNN to achieve better results. Researchers [19] used a combination of the YOLO-v3, DBSCAN, DFSD, and MobileNetv2 algorithms to detect people who were wearing masks on their faces. For face detection while wearing masks, the method of choice is one of the most essential variables [20], in addition to gathering datasets for training data [21], which is not commonly available because face datasets while wearing masks are not widely available.

This work presents a method of face detection utilizing a mask, which is based on the YOLO-v5 algorithm, to address this challenge. Because of the unified model created by this deep learning method, the procedure is more efficient. For example, the bounding box will represent a person's recognized face, and the relevant label will indicate whether the person is wearing a mask or not. The use of generated datasets is what distinguishes this study from others. When a synthetic dataset is generated from an image of a face that is not wearing a mask to an image of a face that is wearing a mask, the dataset can be recognized by the system as being worn by a mask. In order to determine the accuracy of the system, it will be tested on three separate datasets, including the face dataset while wearing a mask (dataset M), the synthetic dataset (dataset S), and the combined dataset (dataset G), and at three different resolutions. Computer vision researchers will find the proposed synthetic dataset in this study to be much more beneficial.

2. Material and Methods

Generally, the detection system employs a classifier or localizer to do detection by applying the model to an image at various locations and scales [22] with the classifier or localizer performing the detection. The YOLO method, on the other hand, takes a completely different approach, in that it applies a single neural network to the entire image [23]. Using a single neural network, the image is divided into areas, and then bounding boxes and probabilities are predicted. For each bounding region box, the chance of categorization is predicted, allowing for the determination of whether an object is classified as such or not. The highest value in the bounding box will be chosen to be used as a separator between objects, with the lowest value being ignored. With its basic architecture and convolutional layer, the YOLO algorithm detects objects in real-time. It is a clever neural network that detects objects in real-time [19]. The YOLO algorithm was invented by a researcher [24], who claims that it can be used to get high accuracy and good prediction results.

By providing a bounding box for the image, the YOLO algorithm predicts with high accuracy on the image [25]. Each grid in the image is labelled, and image classification and object localization algorithms are applied to each grid in the image using the methods described above. This algorithm examines each grid individually and marks the labels that include items as well as the bounding boxes that are within each grid. R-CNN and YOLO share similar characteristics. With the convolution feature [24], each lattice cell offers a hypothetical bounding box and scores the box based on the proposal. As a result, in his research, the bounding box approach (Bounding) is employed for object localization in order to address the shortcomings of the sliding window method [26]. The development of the sliding window method is based on the CNN method, which entails cutting all parts of the image to the size of the window so that a set of cut areas will have several objects, along with classes and object bounding boxes [19], and then assembling these objects into a single object. Because of its improved spatial feature extraction capabilities and cheaper processing costs, CNNs play a vital role in pattern recognition linked to computer vision [1]. With the help of feature learning and transformation invariant components, CNN has made tremendous advancements in facial recognition technology.

Several studies suggest novel approaches for detecting faces hidden behind masks, and the majority of them resolve this issue utilizing a simple CNN algorithm as a binary classification problem. As a result, in research on face detection while wearing masks, it must be handled by employing an object detection model to detect a large number of people and provide a bounding box. It then draws a bounding box around them in a specific color based on whether or not they are wearing a mask and presents an analysis of the ratio of persons wearing masks using the YOLO-v5 method. The YOLO-v5 ultralytics approach, a development of YOLO-v4.5, is employed as an object identification algorithm with a speed of 140 frames per second [5]. This YOLO-v5 is capable of overcoming the problem associated with identifying several facial scales. This technique performs admirably on things of standard size yet is incapable of detecting small objects. This alleviates the difficulty associated with identifying different face scales, hence enhancing face detection performance.

A concise description of the methods utilized in this article consists of three distinct steps. The first stage involves the creation of the dataset, the second stage involves the training of the model, and the last stage involves testing the produced model. NumPy, Pytorch, and OpenCV were used in this study.

2.1. Dataset preparation process

The dataset in this study used a face dataset wearing a mask, a synthetic dataset, and a combined dataset.

a. The M dataset is a face dataset wearing a mask, the dataset used in this study is a dataset from [27].
b. The S dataset is a synthetic dataset.
c. The G dataset is a combined dataset.

The synthetic dataset is a dataset created using a face image dataset without wearing a mask, using a machine learning algorithm to produce a face image wearing a mask. The process of making a synthetic dataset as presented in figure 1. Face images without wearing a mask are collected using a smartphone camera.

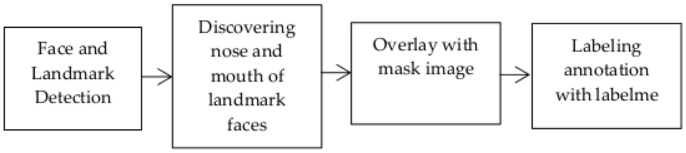
Face and Landmark Detection
Discovering nose and mouth of landmark faces
Overlay with mask image
Labeling annotation with labelme

Fig. 1. A sample Synthesizing Dataset Process

As illustrated in Figure 1, the process of creating a synthetic dataset begins with the preparation of a face image without the use of a mask, followed by face and landmark detection. The second objective is to establish the position of the nose, mouth, and chin in relation to facial landmarks. Thirdly, fill (overlay) the face with the mask's picture; the mask's image can only be filled if the face's image is facing the camera / frontal. The final step is to provide labeling/annotation using the labelme program; the outcome is data in the form of.xml that has a bounding box and its label. Annotation is the process of producing a label for an image by specifying a bounding box followed by the class name of the object in the image, as seen in Figure 2.

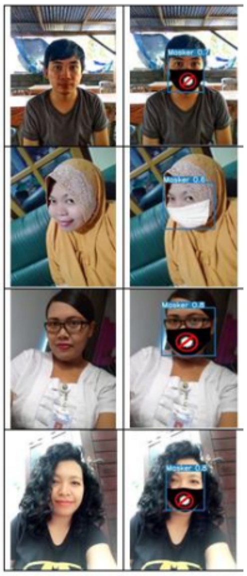

Fig. 2. Example of a mask to the face image

Figure 2 illustrates that the image of the face in the left column is the face without a mask, whereas the image of the face in the right column is the face wearing a mask as a consequence of the mask addition procedure. The resulting dataset is employed as a synthetic dataset as a result of the addition of this mask. The combined dataset is comprised of mask-wearing facial picture datasets and synthetic datasets.

2.2. Training

The training procedure is depicted in Figure 3. Three datasets were used in this investigation. The face image dataset includes up to 853 face images that were trained with masks. A synthetic dataset in this study comprised of 262 facial photos. 1115 photos comprise the merged dataset. Each dataset was supplemented with up to 192 face photos that are not wearing masks. Prior to training, the image dataset was separated into two folders, one for the training set and another for the validation set, each with an 80% and 20% weighting. Additionally, the image of the face wearing a mask was translated to Darknet format, as is the image of the face without a mask. Following that, the facial image was labeled with symbol 0 for faces wearing masks and symbol 1 for faces not wearing masks. The YOLO-v5 algorithm was used for training. As a result of this training, a database of faces wearing and not wearing masks had been created.

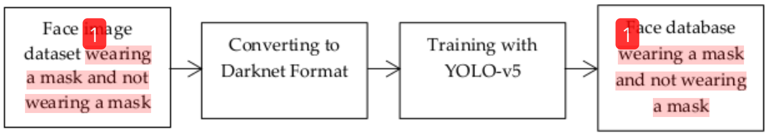
Face image dataset wearing a mask and not wearing a mask
Converting to Darknet Format
Training with YOLO-v5
Face database wearing a mask and not wearing a mask

Fig. 3. Training Process Using YOLO-v5

The training in this study was carried out using OpenCV (Open Source Computer Vision Library) and NumPy, with each dataset and pixel being trained separately. As a result, the resulting databases were M, S, and G.

2.3. Testing

The tests conducted in this study were to develop a system using the Pytorch tool and then to test the system on several databases. The method was evaluated on a variety of facial photos, both without and with masks.

To determine face detection using a mask, this study made suggestions for a face detection procedure that utilizes two models: the first model is used to detect faces in the image, and the second model is used to assess the presence or absence of masks within the bounding box discovered.

3. Results and Discussion

According to the training procedure outlined in Figure 3, computational training was performed on each dataset using a GPU running the YOLO-v5 algorithm on the Google Colab server. Epoch 30, batch size 16, and a threshold of 0.5 to 0.95 were used as pre-train weight values. The training was performed on images with a resolution of 320 pixels and 640 pixels; the results are presented in Table 1.

Table 1. Training Results with Three Dataset

No	Dataset Type	Number of images wearing mask (pieces)	Image resolution (pixel)	Training time (hour)	mAP@ 0.5	mAP@ 0.5:.95	Prediction
1	Dataset M	853	320	0.080	0.583	0.343	0.9
2	Dataset M	853	640	0.145	0.613	0.389	0.932
3	Dataset S	262	320	0.126	0.986	0.625	0.955
4	Dataset S	262	640	0.204	0.986	0.694	0.964
5	Dataset G	1115	320	0.104	0.604	0.392	0.929
6	Dataset G	1115	640	0.275	0.634	0.44	0.965

Table 1 summarizes the outcomes of training with three datasets. For images with a greater resolution, the training process takes longer. The prediction value indicates fairly good results, as the results of the prediction on various datasets and resolutions indicate that the prediction results or accuracy of the results are rather good. The ratio of true positive predictions to the overall positive projected outcome is called prediction. For example, mAP (mean Average Precision) is the average value of Average Precision (AP), which serves as an evaluation parameter for object detection performance. The results are quite good based on the mAP value with a threshold of 0.5 to 0.95. A higher threshold results in fewer false positives. However, if we set it too high, the model will lose a significant amount of detection, resulting in a low confidence value for the correct value. The assessment ratings acquired during the training process demonstrate that the training was conducted extremely well and very precisely.

System testing results

After receiving satisfactory training results, the built system is tested using the training result database. The findings of the system's implementation are utilized to detect face images. After the face is recognized, a new image is displayed with a bounding box that includes the class name of each object and a file providing the bounding box's coordinates and the estimated likelihood of detecting a face wearing a mask or not wearing one.

The test was carried out on one face or several faces in public places. Figure 4 illustrates the results of system testing.

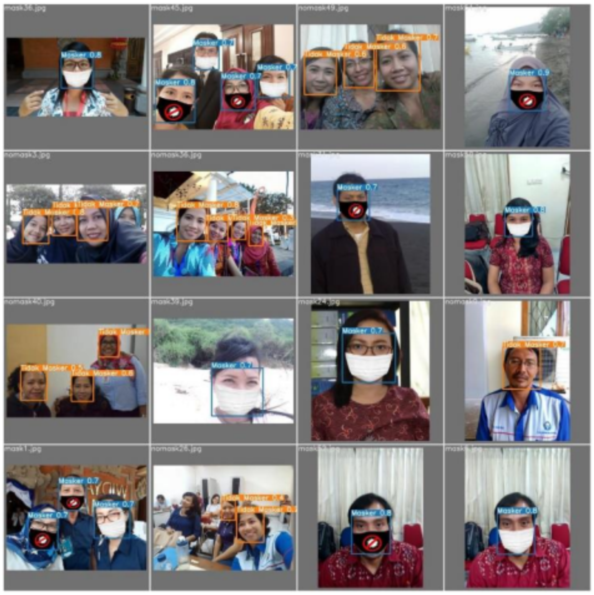

Fig. 4. Training System test results with predictive probabilities

As illustrated in Figure 4, the system constructed with YOLO-v5 is capable of accurately predicting the face that is wearing a mask and the face that is not wearing a mask. The YOLO-v5 approach produces excellent results for face detection with a mask because the object detector being trained is utilized to detect the bounding box and the relevant label. The bounding box displays the recognized faces, and the corresponding label indicates whether or not the individual is wearing a mask. The system is designed to conduct real-time tests on two-dimensional images and video. The test results demonstrate face detection that is constrained by a bounding box containing probability information. The system was evaluated utilizing multi-facial picture test data, both with and without masks, totaling 475 face images.

Table 2. The results of face detection using masks on different datasets and resolutions

Dataset Type	Image resolution (pixel)	Detection results
Dataset M	320	97,8%
Dataset M	640	98,9%
Dataset S	320	97,4%
Dataset S	640	98,5%
Dataset G	320	98,2%
Dataset G	640	**99,2%**

The test results described in Table 2 obtained an average of 98.3 percent accuracy across multiple datasets and resolutions. Additionally, the test results for this system indicate that the G dataset provides the best face detection when utilizing a mask.

With remarkably accurate results, it is believed that this method can be used to help prevent the transmission of viruses and germs by anticipating or monitoring the presence of masks or the absence of masks of people face.

4. Conclusions

This research prioritizes face detection while wearing a mask by analyzing a number of datasets, including a dataset of faces wearing masks (M dataset), a synthetic dataset (S dataset), and a combined dataset (G dataset). The system was developed using the YOLO-v5 algorithm, and tests were conducted on multi-face photos with a variety of datasets and resolutions. The test results obtained the best detection results at a resolution of 640 pixels, namely 98.9 percent for the M dataset, 98.5 percent for the S dataset, and 99.2 percent for the G dataset. While the average face detection rate is 98.3 percent when wearing a mask. The study's proposed usage of a synthetic dataset (dataset S) offers remarkably satisfactory findings and can be used for face detection studies while wearing masks. As a result of the consistent test findings, our recommendation for face detection while wearing a mask is applicable to all datasets.

5. References

Dataset variation for masked face detection using YOLO-v5 method

ORIGINALITY REPORT

5%
SIMILARITY INDEX

PRIMARY SOURCES

1 Shuang Li, Xiaoli Dong, Yuan Shi, Baoli Lu, Linjun Sun, Wenfa Li. "Multi - angle head pose classification with masks based on color texture analysis and stack generalization", Concurrency and Computation: Practice and Experience, 2021
Crossref
62 words — 2%

2 Zaigham A. Randhawa, Shivang Patel, Donald A. Adjeroh, Gianfranco Doretto. "Chapter 2 Learning Representations forMasked Facial Recovery", Springer Science and Business Media LLC, 2022
Crossref
17 words — 1%

3 Ameema Zainab, Dabeeruddin Syed. "Deployment of Deep Learning Models on Resource-Deficient Devices for Object Detection", 2020 IEEE International Conference on Informatics, IoT, and Enabling Technologies (ICIoT), 2020
Crossref
14 words — < 1%

4 medium.com
Internet
14 words — < 1%

5 Dongwei Yu, Liang Shan, Yifei Zhou, Wei Tang, Xun Luo. "Improved Face Detection Algorithm under Mask Occlusion", 2022 International Conference on Cyber-Physical Social Intelligence (ICCSI), 2022
13 words — < 1%

Crossref

6 www.ncbi.nlm.nih.gov
Internet
13 words — < 1%

7 docs.microsoft.com
Internet
11 words — < 1%

8 Ziwei Song, Kristie Nguyen, Tien Nguyen, Catherine Cho, Jerry Gao. "Camera-Based Security Check for Face Mask Detection Using Deep Learning", 2021 IEEE Seventh International Conference on Big Data Computing Service and Applications (BigDataService), 2021
Crossref
10 words — < 1%

9 Liang Niu, Cheng Qian, John-Ross Rizzo, Todd Hudson, Zichen Li, Shane Enright, Eliot Sperling, Kyle Conti, Edward Wong, Yi Fang. "A Wearable Assistive Technology for the Visually Impaired with Door Knob Detection and Real-Time Feedback for Hand-to-Handle Manipulation", 2017 IEEE International Conference on Computer Vision Workshops (ICCVW), 2017
Crossref
6 words — < 1%

EXCLUDE QUOTES	OFF	EXCLUDE SOURCES	OFF
EXCLUDE BIBLIOGRAPHY	ON	EXCLUDE MATCHES	OFF