YOLOv3-based Mask and Face Recognition Algorithm for Individual Protection Applications

Roberta Avanzato[a], Francesco Beritelli[a], Michele Russo[b], Samuele Russo[c] and Mario Vaccaro[b]

[a]*Department of Electrical, Electronic and Computer Engineering, University of Catania, Catania, CT, Italy*
[b]*VICOSYSTEMS S.r.l V.le Odorico da Pordenone, 33, Catania, CT, Italy*
[c]*Sapienza University of Rome, Piazzale Aldo Moro 5, Roma, Italy*

Abstract

To combat the spread of the COVID-19 pandemic, it is essential to strictly obey social distancing measures, as well as have the possibility to possess and wear personal protective equipment. This paper proposes a mask and face recognition algorithm based on YOLOv3 for individual protection applications. The proposed method processes images directly in raw data format input to a neural network trained with deep learning techniques. System training was performed on a set of images appropriately obtained from the MAFA dataset by selecting those with surgical masks for a total of about 6,000 cases. The performances obtained indicate 84% accuracy in recognizing a mask and 96% in the case of a face.

Keywords

Image processing, Face recognition, Mask recognition, Computer vision, Deep learning

1. Introduction

TIn an emergency phase, the fight against the spread of COVID-19 contamination is regulated by procedures of medical-scientific rigor and official protocols adopted as regulations until the epidemic is definitively defeated on a global scale. For the return to normality, which is expected to be gradual and of medium-long duration, it is essential to strictly obey social distancing measures, as well as have the possibility to possess and wear personal protective equipment for those who continue to work in potentially contagious environments. Thus, it becomes strategic to focus on solutions that can remotely and non-intrusively monitor people's behaviour and health, while ensuring respect for privacy. One solution is represented by innovative video intelligence technologies for the automatic detection of body temperature and the proximity distance between individuals in order to guarantee, and possibly certify, in outdoor or indoor environments, compliance with the regulations on the constraints of the distance between individuals (and/or the maximum capacity in a given environment), access to indoor environments for individuals without critical health conditions, and, possibly, where necessary, compliance with the restrictions on individual protection (masks, gloves, overalls etc.). There are several important advantages: the safeguard of people's health, the mitigation of the risk of contamination return, the possibility of timely interventions by the law enforcement engaged in preserving public health orders, as well as safe and fast return to work.

The key issues forming the basis for the proposal described in this paper arises are the following:

- need for social distancing outdoors (streets, squares, parks, etc.) and indoors (offices, schools, shopping centers, theaters, restaurants, pubs, shops, etc.);
- need to manage quotas for access and use of public areas and public carriers;
- need for timely notification of gatherings to the managers of the frequented areas and, in the most serious cases, to the law enforcement, possibly via the certification of critical events;
- need to monitor the state of health (by checking the temperature) of people who access an indoor environment;
- need to monitor compliance with the use of protective equipment (masks, gloves, overalls), especially in the most at-risk work contexts.

The last point is the one the present study focuses on by proposing a mask/face recognition algorithm.

ICYRIME 2020: International Conference for Young Researchers in Informatics, Mathematics, and Engineering, Online, July 09 2020

roberta.avanzato@phd.unict.it (R. Avanzato); francesco.beritelli@dieei.unict.it (F. Beritelli); m.russo@vicosystems.it (M. Russo); samuelerussoct@gmail.com (S. Russo); m.vaccaro@vicosystems.it (M. Vaccaro)

In the state of the art there are many studies dealing with face recognition and, in particular, recognition of masked faces.

In [1] the authors propose a masked face detection technique useful for monitoring and identifying criminals or terrorists. They propose a CNN-based cascade framework, which consists of three carefully designed convolutional neural networks to detect masked faces. The accuracy in recognizing masked faces is 87.8%.

In [2] the authors propose a further method of identifying masked faces based on the LLE-CNN network and MAFA database [3]. In this approach, the authors achieved a performance of 76.4%. The authors in [4] address the issue of the importance of greater accuracy in face recognition during the period of COVID-19. The study proposes a face-eye based multi-granular recognition model. With this approach, the accuracy of masked face recognition goes from the initial 50% to 95%.

In the present study, a mask/face recognition technique is proposed using a very performing type of convolutional neural network called YOLOv3. This method allows to derive the detection and classification performance of the "faces" and "masks" within the video and/or images.

The paper is structured as follows: Section 2 describes the proposed method; Section 3 illustrates the neural network used; Section 4 describes the database used; section 5 shows the performances obtained by the proposed technique; the last section is dedicated to conclusions.

2. Proposed Method

This section describes the process of detecting masks and faces.

Figure 1 shows the block diagram of the proposed technique.

The first block represents the acquisition of the video signal by means of cameras, which can be installed in indoor or outdoor environments.

Once the video signal is acquired, a pre-processing phase is performed (Video processing block) which is responsible for extracting the frames with a frame-rate equal to 30 fps. Subsequently, the frames are fed into the previously trained YOLOv3 neural network. The output neural network provides a percentage of detection and classification accuracy of the face and masks present in the input data frames.

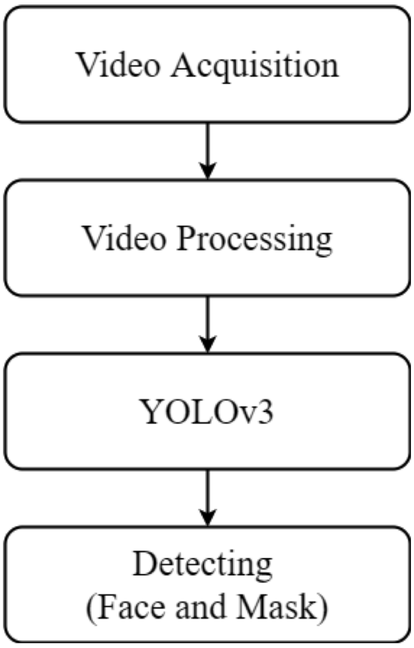

Figure 1: Block diagram of the proposed method.

3. Adopted Neural Network

The application of artificial intelligence and machine learning algorithms turns out to be a very complex approach if the problems requiring a solution are not highlighted [5, 6, 7, 8, 9, 10, 11, 12, 13]. In this study, we are interested in recognizing the face and any mask worn by the various people present in the video recordings.

The theme of face recognition and masks falls within the subject of object detection. Object detection is the basis of computer vision, and specifically for applications such as instance segmentation, image captioning and object detection/tracking. From an application point of view, it is possible to group object detection into two macro-categories:

- General object detection: the goal is to investigate methods for identifying different types of objects using a single framework, in order to simulate human vision and cognition;
- Detection applications: refers to the recognition of objects of a certain class in specific application scenarios. For example, there may be various applications for pedestrian detection, face detection or for text detection.

There are several models that implement object recognition, present in the state of the art.

One of them is the Faster R-CNN [14] which represents the current state of the art for models that divide the task of identifying objects into several phases. This network allows you to simultaneously train a recognizer and a bounding box designer within a single model. The procedure carried out by this network is of the "proposal detection and verification" type.

A second model is YOLO (You Only Look Once). In [15, 16] the authors have completely abandoned the pre-existing paradigm of "proposal detection and verification". Instead, YOLO follows a completely different philosophy: applying a single model to the entire image. YOLO, in fact, divides the image into regions, predicts the bounding boxes and for each of them, determines the probabilities of belonging to a certain class, all using a single network.

In [17] the authors define the SSD (Single Shot Detector) model. This method has greatly contributed to the change of perspective towards the generation of bounding boxes: unlike the previous models that were concerned with accurately predicting the location of an object within the image, SSD starts from a set of bounding boxes by default. Starting from this set a deviation and its classification are predicted for each of these boxes. Thanks to a set of operations and SSD filters, it also obtains excellent accuracy in the prediction of object classes.

In order to make an exhaustive comparison of the various convolutional models presented, to maintain a certain consistency in the results, it was decided to use the work done in [18] as a framework to compare the performances. In this study, the authors indicate that YOLOv3 is clearly superior, compared to the other CNNs, both in terms of computational time and accuracy. However, it should be noted that Fast R-CNN, despite the huge gap in terms of computational time, allows, among others, to identify very accurate segmentations (polylines) when compared with the "simple" bound boxes provided by YOLOv3 or SSD. Therefore, based on the specific application context there may be some cases in which Fast R-CNN is the optimal solution.

4. Database

Once the neural network model was defined, we moved to the search for a database containing faces and masks to train the model.

At first, MAFA [3] database designed to recognize faces partially occluded by objects was used as a reference, containing 25,000 images for training and 10,000 images for testing.

Table 1
Objects in training e testing dataset for each classs.

Dataset	Training	Testing
Mask	5555	1855
Face	4173	1299

Table 2
Confusion matrix.

	Face	Mask
Face	0.94	0.06
Mask	0.14	0.86

In this dataset, a great presence of images was noted in which the masks were not suitable for individual protection, such as: scarves, sweaters and hands covering the face, full masks used for masquerades, etc. For this reason, image filtering was performed; in particular, selecting those that contained surgical masks.

Subsequently, a re-labeling of the dataset was performed, in order to obtain an automatic recognition system of the presence of a protective mask on a face.

Via the new labeling, a dataset of 5,800 images was extracted, where 3,800 images were used for training the neural network and 2,000 images were used for testing.

Each image can contain one or more “Mask” and “Face” objects. In this regard, Table 1 shows the number of the two types of objects for the training and testing dataset.

5. Performance Evaluation

Once the neural network model and the dataset in use are defined, it is possible to analyse the performances obtained when the dataset described above is fed to the network.

After a training phase of the neural network model, the testing dataset was applied, containing images completely unknown to the network.

The performances on the testing dataset obtained from the network are shown in Tables 2 and 3. Table 2 shows the confusion matrix produced by the neural network. The performances obtained are quite high, implying that the network is able to perform good detection of the two classes on images that it has never seen before.

Table 3 shows the performances, in percentage, obtained using the statistical classification parameters: accuracy, recall or sensitivity, precision and F1 score

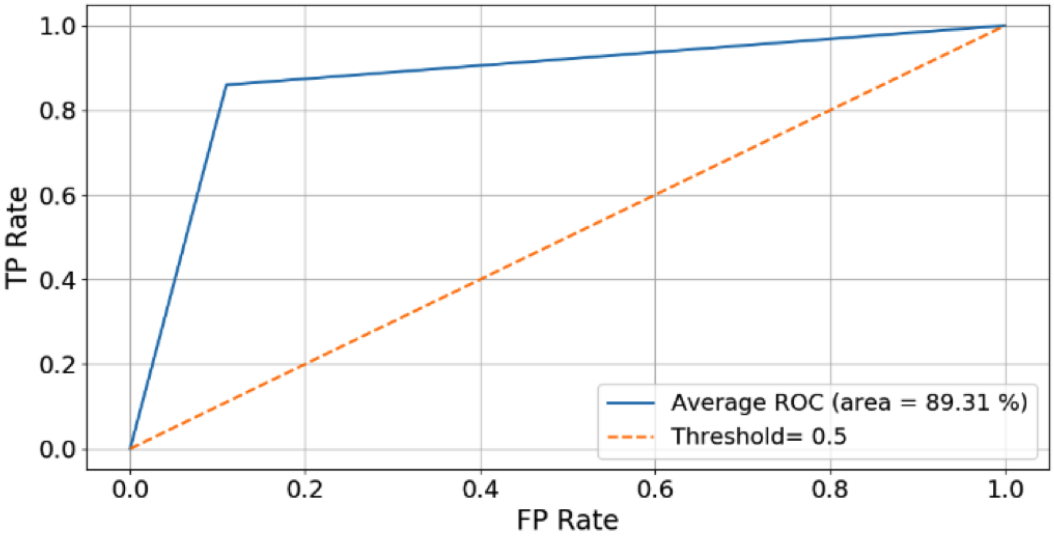

Figure 2: ROC curve.

Table 3
Global performance of the proposed method.

	A [%]	Recall [%]	PRE [%]	F1 [%]
Face	94	94	87	
Mask	86	86	93.5	
Mean	90	90	90.3	89.98

[19].

The table shows that the obtained results in terms of accuracy, recall, precision and F1 score are quite high.

For further validation of the network model and the performance obtained from the statistical classification parameters the ROC (Receiver Operator Characteristic) and ROC AUC (Area Under the Curve) graphs have been produced.

Applying this concept to our classification method, in Figure 2 we observe the resulting ROC curve which indicates that a certain degree of variance between the various parts and the average ROC AUC [20] lies between the perfect score (1.0) and the diagonal (0.5).

The graph shows that the area under the ROC curve is very large. This means that our model has excellent performance. In fact, the accuracy (average for the two classes) is equal to 89.31%.

The results show that the proposed neural network and the new re-labeled database perform better than the state of the art methods. In particular, comparing our study with the one presented in [2] it is clear that by conducting the research almost with the same dataset the obtained performances in the present study are 13.6

6. Conclusion

From the point of view of the recognition of the individual protective garment (mask), which is the subject of our study, we focused on the simple detection of faces and masks within a frame (image or video), obtaining an average accuracy of 90%. A future development includes the extension of the neural network by adding one or more decision-making layers (classification) in order to be able to identify not only the presence of the mask in the photo but also its position with respect to the person's face, so as to define whether it is worn properly or not.

References

[1] W. Bu, J. Xiao, C. Zhou, M. Yang, C. Peng, A cascade framework for masked face detection, in: 2017 IEEE International Conference on Cybernetics and Intelligent Systems (CIS) and IEEE Conference on Robotics, Automation and Mechatronics (RAM), IEEE, 2017, pp. 458–462.

[2] S. Ge, J. Li, Q. Ye, Z. Luo, Detecting masked faces in the wild with lle-cnns, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 2682–2690.

[3] MAFA open dataset, 2019. URL: http://221.228.208.41/gl/dataset/0b33a2ece1f549b18c7ff725fb50c561.

[4] Z. Wang, G. Wang, B. Huang, Z. Xiong, Q. Hong, H. Wu, P. Yi, K. Jiang, N. Wang, Y. Pei, et al., Masked face recognition dataset and application, arXiv preprint arXiv:2003.09093 (2020).

[5] R. Avanzato, F. Beritelli, F. Di Franco, V. F. Puglisi, A convolutional neural networks approach to audio classification for rainfall estimation, in: 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), volume 1, IEEE, 2019, pp. 285–289.

[6] S. Spanò, G. C. Cardarilli, L. Di Nunzio, R. Fazzolari, D. Giardino, M. Matta, A. Nannarelli, M. Re, An efficient hardware implementation of reinforcement learning: The q-learning algorithm, Ieee Access 7 (2019) 186340–186351.

[7] R. Avanzato, F. Beritelli, A cnn-based differential image processing approach for rainfall classification, Advances in Science, Technology and Engineering Systems Journal 5 (2020) 438–444.

[8] S. I. Illari, S. Russo, R. Avanzato, C. Napoli, A cloud-oriented architecture for the remote assessment and follow-up of hospitalized patients, in: Symposium for Young Scientists in Technology, Engineering and Mathematics, volume 2694, CEUR-WS, 2020.

[9] R. Avanzato, F. Beritelli, A. Raspanti, M. Russo, Assessment of multimodal rainfall classification systems based on an audio/video dataset, International Journal on Advanced Science, Engineering and Information Technology 10 (2020) 1163–1168.

[10] R. Avanzato, F. Beritelli, Automatic ecg diagnosis using convolutional neural network, Electronics 9 (2020) 951.

[11] C. Napoli, F. Bonanno, G. Capizzi, Exploiting solar wind time series correlation with magnetospheric response by using an hybrid neuro-wavelet approach, Proceedings of the International Astronomical Union 6 (2010) 156–158.

[12] C. Napoli, F. Bonanno, G. Capizzi, An hybrid neuro-wavelet approach for long-term prediction of solar wind, Proceedings of the International Astronomical Union 6 (2010) 153–155.

[13] G. Capizzi, C. Napoli, L. Paternò, An innovative hybrid neuro-wavelet method for reconstruction of missing data in astronomical photometric surveys, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) 7267 LNAI (2012) 21–29.

[14] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, IEEE Transactions on Pattern Analysis and Machine Intelligence 39 (2017) 1137–1149.

[15] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 779–788.

[16] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, arXiv preprint arXiv:1804.02767 (2018).

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A. C. Berg, Ssd: Single shot multibox detector, in: European conference on computer vision, Springer, 2016, pp. 21–37.

[18] X. Zhang, W. Yang, X. Tang, J. Liu, A fast learning method for accurate and robust lane detection using two-stage feature extraction with yolo v3, Sensors 18 (2018) 4308.

[19] C. Beleites, R. Salzer, V. Sergo, Validation of soft classification models using partial class memberships: An extended concept of sensitivity & co. applied to grading of astrocytoma tissues, Chemometrics and Intelligent Laboratory Systems 122 (2013) 12–22.

[20] A. P. Bradley, The use of the area under the roc curve in the evaluation of machine learning algorithms, Pattern recognition 30 (1997) 1145–1159.