

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/336952877>

Facial Mask Detection using Semantic Segmentation

Conference Paper · October 2019

DOI: 10.1109/CCCS.2019.8888092

CITATIONS

0

READS

467

3 authors:



Toshan Meenpal

National Institute of Technology Raipur

22 PUBLICATIONS 24 CITATIONS

[SEE PROFILE](#)



Ashutosh Balakrishnan

Indian Institute of Technology Delhi

2 PUBLICATIONS 0 CITATIONS

[SEE PROFILE](#)



Amit Verma

National Institute of Technology Raipur

8 PUBLICATIONS 20 CITATIONS

[SEE PROFILE](#)

Some of the authors of this publication are also working on these related projects:



m tech project [View project](#)



Designed a Frequency Counter to measure input of any incoming signal in Virtex-4 FPGA using VHDL [View project](#)

Facial Mask Detection using Semantic Segmentation

Toshanalal Meenpal

Dept. of Electronics and Telecomm.
National Institute of Technology
Raipur, India
tmeenpal.etc@nitrr.ac.in

Ashutosh Balakrishnan

Dept. of Electronics and Telecomm.
National Institute of Technology
Raipur, India
abalakrishnan1909@gmail.com

Amit Verma

Dept. of Electronics and Telecomm.
National Institute of Technology
Raipur, India
averma.phd2016.etc@nitrr.ac.in

Abstract—Face Detection has evolved as a very popular problem in Image processing and Computer Vision. Many new algorithms are being devised using convolutional architectures to make the algorithm as accurate as possible. These convolutional architectures have made it possible to extract even the pixel details. We aim to design a binary face classifier which can detect any face present in the frame irrespective of its alignment. We present a method to generate accurate face segmentation masks from any arbitrary size input image. Beginning from the RGB image of any size, the method uses Predefined Training Weights of VGG – 16 Architecture for feature extraction. Training is performed through Fully Convolutional Networks to semantically segment out the faces present in that image. Gradient Descent is used for training while Binomial Cross Entropy is used as a loss function. Further the output image from the FCN is processed to remove the unwanted noise and avoid the false predictions if any and make bounding box around the faces. Furthermore, proposed model has also shown great results in recognizing non-frontal faces. Along with this it is also able to detect multiple facial masks in a single frame. Experiments were performed on Multi Parsing Human Dataset obtaining mean pixel level accuracy of 93.884 % for the segmented face masks.

Index Terms—Fully Convolutional Network, Semantic Segmentation, Face Segmentation and Detection

I. INTRODUCTION

Face detection has emerged as a very interesting problem in image processing and computer vision. It has a range of applications from facial motion capture to face recognition which at the start needs the face to be detected with a very good accuracy. Face detection is more relevant today because it not only used on images but also in video applications like real time surveillance and face detection in videos. High accuracy image classification is possible now with the advancements of Convolutional networks. Pixel level information is often required after face detection which most face detection methods fail to provide. Obtaining pixel level details has been a challenging part in semantic segmentation. Semantic segmentation is the process of assigning a label to each pixel of the image. In our case the labels are either face or non-face. Semantic segmentation is thus used to separate out the face by classifying each pixel of the image as face or background. Also most of the widely used face detection algorithms tend to focus on the detection of frontal faces.

This paper proposes a model for face detection using semantic segmentation in an image by classifying each pixel as face and non-face i.e. effectively creating a binary classifier

and then detecting that segmented area. The model works very well not only for images having frontal faces but also for non-frontal faces. The paper also focuses on removing the erroneous predictions which are bound to occur. Semantic segmentation of human face is performed with the help of a fully convolutional network.

The next section discusses the related work done in the domain of face detection. In section III we describe the method followed for face segmentation and detection using semantic segmentation on any arbitrary RGB image. Finally, the generated facial masks are demonstrated in experimental results in section IV. Post processing on the predicted images has also been discussed at length which also entails the removal of erroneous predictions.

II. RELATED WORKS

Initially researchers focused on edge and gray value of face image. [1] was based on pattern recognition model, having a prior information of the face model. Adaboost [2] was a good training classifier. The face detection technology got a breakthrough with the famous Viola Jones Detector [3], which greatly improved real time face detection. Viola Jones detector optimized the features of Haar [4], but failed to tackle the real world problems and was influenced by various factors like face brightness and face orientation. Viola Jones could only detect frontal well lit faces. It failed to work well in dark conditions and with non-frontal images. These issues have made the independent researchers work on developing new face detection models based on deep learning, to have better results for the different facial conditions. We have developed our face detection model using Multi Human Parsing Dataset [5], based on fully convolutional networks, such that it can detect the face in any geometric condition frontal or non-frontal for that matter. Convolutional Networks have always been used for image classification tasks. Typical architectures like AlexNet [6] and VGGNet [7] comprise of stacked convolutional layers. AlexNet with 5 convolutional layers and 3 fully connected layers has been the winner of ImageNet LSVRC-2012 competition while VGGNet is an improvement over AlexNet as it replaces large kernels with 3x3 multiple kernels consecutively. The ILSVRC-2014 winning architecture GoogleNet [8] uses parallel convolution kernels and concatenating the feature maps together. In it 1x1, 3x3 and 5x5 convolutions and 3x3 max-pooling have been used. Smaller convolutions extract the

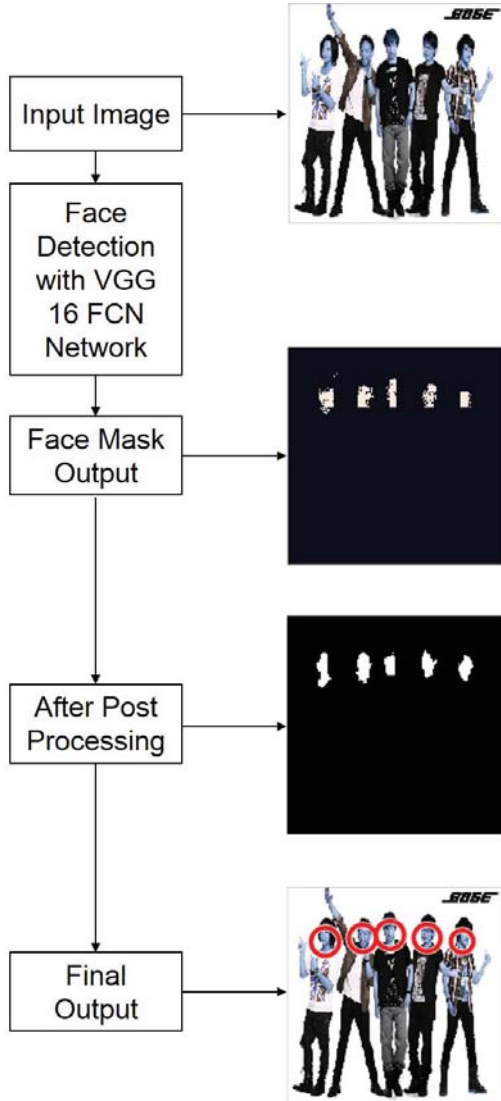


Fig. 1. Flowchart of the proposed method.

local features whereas larger convolutions extract high level features. More recent architectures such as ResNet [9] have introduced skip connections which allows deeper networks to avoid saturation in training accuracy. These architectures are often used for initial feature extraction in face detection networks. In our method, we are using VGG 16 architecture as the base network for face detection and Fully Convolutional Network for segmentation. VGG 16 network is sufficiently deep to extract features and computationally less expensive for our case. Though majority of segmentation architectures rely on downsampling and consecutive upsampling of input image, Fully Convolutional Networks [10], [11], [12] still are modest and have significantly accurate approach for segmentation.

III. METHODOLOGY

We propose this paper with twin objective of creating a Binary face classifier which can detect faces in any orientation

irrespective of alignment and train it in an appropriate neural network to get accurate results. The model requires inputting an RGB image of any arbitrary size to the model. The model's basic function is feature extraction and class prediction. The output of the model is a feature vector which is optimized using Gradient descent and the loss function used is Binomial Cross Entropy. Figure 1 represents the end to end pipeline of our method along with sample demonstration of obtained output at each step.

A. Proposed Work Flow

We propose a method of obtaining segmentation masks directly from the images containing one or more faces in different orientation. The input image of any arbitrary size is resized to $224 \times 224 \times 3$ and fed to the FCN network for feature extraction and prediction. The output of the network is then subjected to post processing. Initially the pixel values of the face and background are subjected to global thresholding. After that its passed through median filter to remove the high frequency noise and then subjected to Closing operation to fill the gaps in the segmented area. After this bounding box is drawn around the segmented area.

B. Architecture

The feature extraction and prediction is performed using pre-defined training weights of VGG 16 architecture. The basic VGG-16 architecture is depicted in Figure 2. Our proposed model consists of a total of 17 convolutional layers and 5 Max pooling layers. The initial image size which is fed to the model is $224 \times 224 \times 3$. As the image is processed through the layers for feature extraction its passed through convolutional layers and max pooling layers.

Convolutional layer convolutes the input image with another window while the max pooling operation ensures that the size of the feature vector being produced in every layer is halved so as to reduce the number of parameters. This is a very crucial step in feature extraction, if the number of parameters are not reduced then it will become very difficult to predict the classes of each pixel in a fully convolutional network. The initial layers extract the lower level features while as the subsequent layers extract the mid-level and higher level features. The segmentation task requires that the spatial information be stored in a pixel wise classification, this we have achieved by converting the VGG layers to convolutional layers. After the final max pooling layer – the image size is reduced to $28 \times 28 \times 2$. This is further upsampled to bring the image to standard size i.e. $224 \times 224 \times 2$ since it's a binary classifier – hence creates two channels for both the classes, face and background.

C. Face Detection and Avoiding Erroneous Prediction

Post processing on the predicted mask obtained is performed so that the irregularities in the region can be filled and to remove the unwanted errors (which may have crept during the processing). This we perform by first passing the mask through Median filter and then performing the Closing Operation.

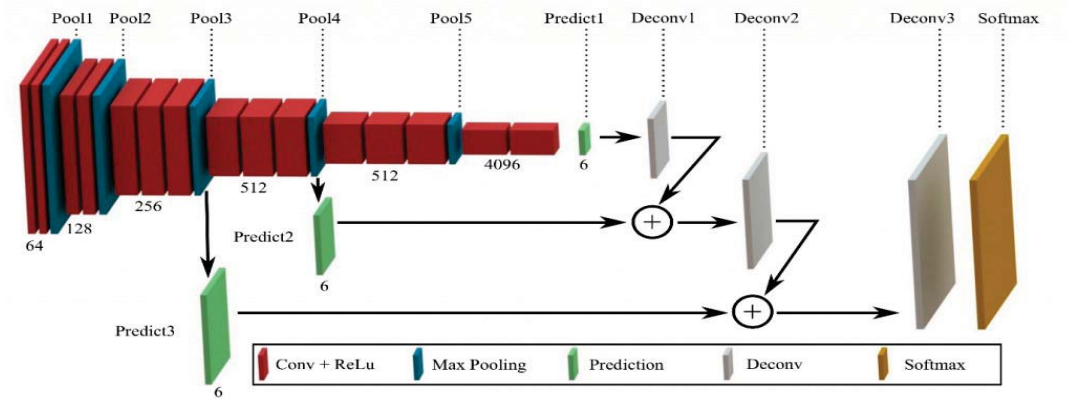


Fig. 2. The complete architecture of Fully Convolutional Network used generating segmentation masks.

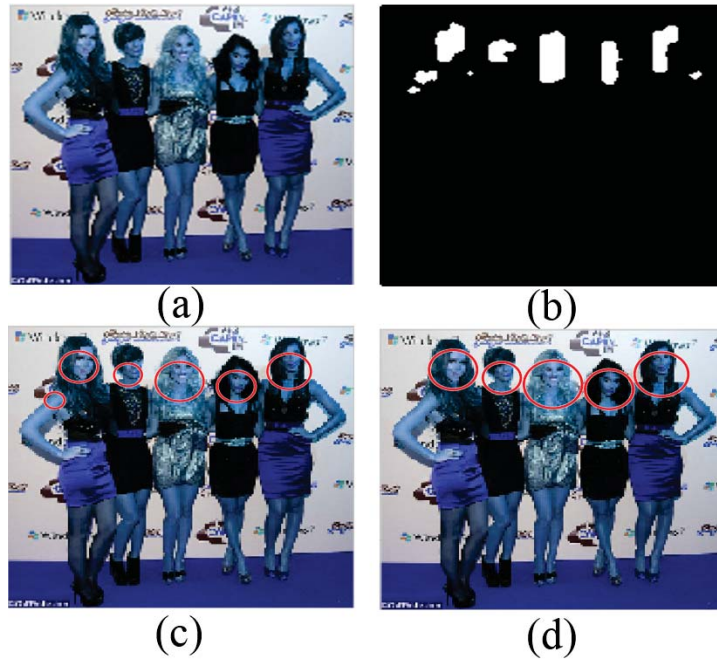


Fig. 3. (a) Actual Image (b) Erroneous Prediction (c) False Face Detection (d) Correct Face Detection.

This ensures that the gaps in the segmented region are filled and most of the unwanted false erroneous prediction removed. In spite of this there is a possibility that some large error may not have been removed. We have designed the model such that all those erroneous predictions are not considered while showing the final detected faces. We find out the following parameters in each region – Centroid, Major Axis Length and Minor Axis Length. These values for Figure 3 are depicted in Table 1 for all the facial (segmented) regions detected (including false predictions).

In the Figure 3(b), even after post processing through median filter and dilation, the unwanted erroneous predictions have not completely gone. This results in false face detection – Figure 3(c)

TABLE I
SEGMENTED REGION PARAMETER VALUES

S. No.	Centroid	Major Axis Length	Minor Axis Length
1.	9.414	11.62	7.2
2.	18.00	22.65	13.36
3.	13.18	14.84	11.51
4.	22.81	32.09	13.52
5.	18.07	27.35	8.8
6.	20.67	30.55	10.7

Using the centroid c_x , Major Axis ma_x and minor axis values mi_x for each of the segmented region, we calculate the diameter d_x of each region. We compute mean \bar{D} and standard deviation η_D for diameter vector D . Finally, we keep the most








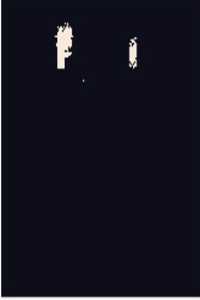
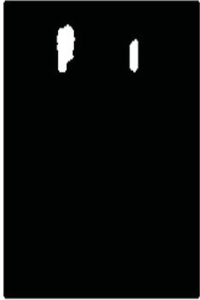



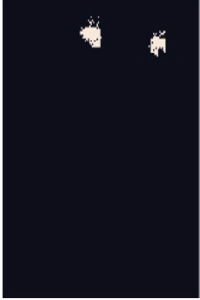




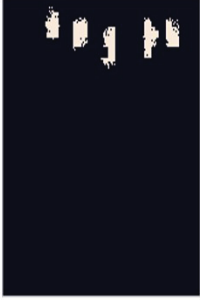
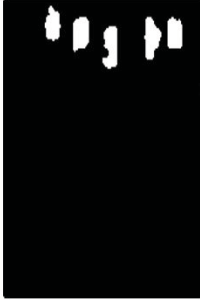

Original Image	True Class Mask	Predicted Class Mask	After Post Processing	Final Output	Pixel Level Accuracy
					93.227%
					94.682%
					93.761%
					89.221%

Fig. 4. Pixel level accuracy for predicted facial masks.

probable diameters lying within the first standard deviation. The detail procedure is shown in Algorithm 1.

Algorithm 1 Detail Distance Computing Procedure

```

 $X \leftarrow$  Number of regions
 $D \leftarrow \mathbb{R}^X$ 
for  $x \leftarrow 1, X$  do
   $d_x \leftarrow \frac{ma_x + mi_x}{2}$ 
   $D[x] \leftarrow d_x$ 
end for
 $\bar{D} \leftarrow \frac{1}{X} \sum_{x=1}^X D[x]$ 
 $\eta_D \leftarrow \sqrt{\frac{1}{X} \sum_{x=1}^X (D[x] - \bar{D})^2}$ 
 $D_{true} \leftarrow [ ]$ 
 $c \leftarrow 0$ 
for  $x \leftarrow 1, X$  do
  if  $\bar{D} - \eta_D < D[x] < \bar{D} + \eta_D$  then
     $c \leftarrow c + 1$ 
     $D_{true}[c] \leftarrow D[x]$ 
  end if
end for

```

IV. EXPERIMENTAL RESULTS

All the experiments have been performed on Multi Human Parsing Dataset containing about 5000 images, each with at least two persons. Out of these, 2500 images were used for training and validation while the remaining were used for testing the model. Figure 4 shows true and predicted class to a given input image of any arbitrary size. It also represents detected faces inside a bounding circle with respective pixel level accuracy. We have also shown the refined predicted mask after its subjected to post processing. The designed FCN semantically segments out the facial spatial location with a specific label. Furthermore, proposed model has also shown great results in recognizing non-frontal faces. Along with this it is also able to detect multiple facial masks in a single frame. The post processing provides a large boost to pixel level accuracy. The mean pixel level accuracy for facial masks : 93.884%.

V. CONCLUSION

We were able to generate accurate face masks for human objects from RGB channel images containing localized objects. We demonstrated our results on Multi Human Parsing Dataset with mean pixel level accuracy. Also the problem of erroneous predictions has been solved and a proper bounding box has been drawn around the segmented region. Proposed network can detect non frontal faces and multiple faces from single image. The method can find applications in advanced tasks such as facial part detection.

REFERENCES

- [1] T. Ojala, M. Pietikainen, and T. Maenpää, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, July 2002.
- [2] T.-H. Kim, D.-C. Park, D.-M. Woo, T. Jeong, and S.-Y. Min, "Multi-class classifier-based adaboost algorithm," in *Proceedings of the Second Sino-foreign-interchange Conference on Intelligent Science and Intelligent Data Engineering*, ser. IScIDE'11. Berlin, Heidelberg: Springer-Verlag, 2012, pp. 122–127.
- [3] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vision*, vol. 57, no. 2, pp. 137–154, May 2004.
- [4] 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, Dec 2001, pp. I–I.
- [5] J. Li, J. Zhao, Y. Wei, C. Lang, Y. Li, and J. Feng, "Towards real world human parsing: Multiple-human parsing in the wild," *CoRR*, vol. abs/1705.07206.
- [6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1097–1105.
- [7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.
- [8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," 2015.
- [9] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.
- [10] K. Li, G. Ding, and H. Wang, "L-fcn: A lightweight fully convolutional network for biomedical semantic segmentation," in *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec 2018, pp. 2363–2367.
- [11] X. Fu and H. Qu, "Research on semantic segmentation of high-resolution remote sensing image based on full convolutional neural network," in *2018 12th International Symposium on Antennas, Propagation and EM Theory (ISAPE)*, Dec 2018, pp. 1–4.
- [12] S. Kumar, A. Negi, J. N. Singh, and H. Verma, "A deep learning for brain tumor mri images semantic segmentation using fcn," in *2018 4th International Conference on Computing Communication and Automation (ICCCA)*, Dec 2018, pp. 1–4.