Lightweight-VGG based Facial Landmarks Detection with Inception-ResNet Module

Savina Jassica Colaco, MyungChul Park, DaeWoong Cha, ChaeHyun Lee and Dong Seog Han*
School of Electronics and Electrical Engineering
Kyungpook National University
Daegu, Republic of Korea
savinacolaco@knu.ac.kr, mcpark@knu.ac.kr, dwcha@knu.ac.kr, hyeu330@knu.ac.kr, dshan@knu.ac.kr*

***Abstract*—Facial landmark detection has achieved great performance by learning discriminative features from the rich deformation of face shapes and poses. The landmarks such as eye centres, nose centres, jawlines, etc are localized to give vital information to computer vision-related applications. The paper proposes a modified light-weight VGG based model for predicting facial landmarks on the detected faces from digital images or video.**

***Index Terms*—facial landmarks, convolutional neural networks, inception-ResNet**

I. Introduction

Facial landmarks detection locates a set of vital points in the digital images of faces taken in unconstraint conditions. This important information is used in different applications such as 3D facial structure estimation, attributes estimation or facial expression recognition [1]. Since convolutional neural networks (CNNs) outperform conventional methods in different applications, CNN also shows significant performance in the detection of facial landmarks. Recent facial landmark detection approaches mainly focus on features from face shapes, poses, different expressions, occlusions and other conditions. A simple facial landmark detection model uses CNN to detect and predict landmarks simultaneously. In this paper, simple CNN is trained to predict 68 facial landmarks and then mapped onto the detected faces in real-time.

II. Experiment

A. Facial landmarks detection model

The VGG16 [2] architecture acts as a baseline model for predicting 68 facial landmarks. The convolution layers are replaced with depthwise convolution layers which apply a single convolutional filter to each input channel. An inception-ResNet module is introduced between the depthwise layers to provide a different level of feature extraction at different scales. The Inception-ResNet , as shown in Figure 1, is highly adjustable giving several possibilities to change the number of filters in the layers. The different number of filters such as 1×1, 3×3 and dilated filters are used to extract features. This helps to focus on different regions of the face images to predict facial landmarks. A skip connection is presented to perform identity mapping where the original input features are added to the outputs of the stacked layers. Each layer is followed by batch normalization and Hswish activation. The Hswish activation function replaces the expensive sigmoid with its piece-wise linear in swish which could be a disadvantage for mobile devices. Figure 2 depicts the model design of facial landmarks detection.

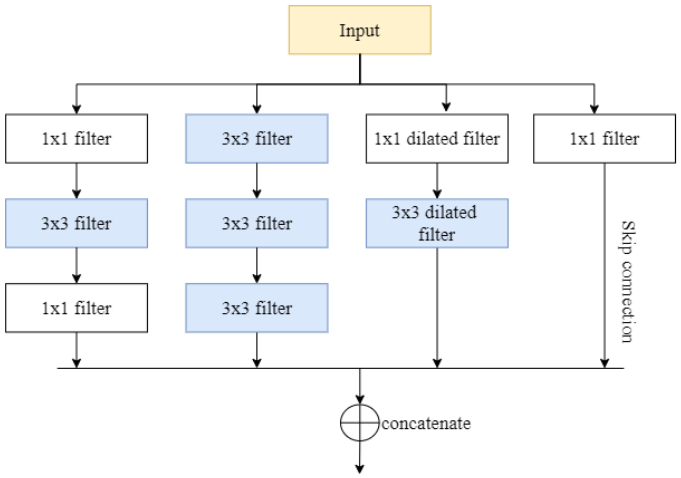

Fig. 1. Inception-ResNet module.

B. Implementation details

The facial landmarks model is trained with a combined dataset of 300W [3] and 300VW [4]–[6] with a total number of 112,111 images. The 300W dataset comprises AFW, HELEN, LFPW, XM2VTS, and IBUG datasets where images are annotated with 68 landmarks. The images in the dataset are resized to 112×112 resolution in grayscale. The Keras framework is used for model implementation and trained with a batch size of 32 and epochs of 50. The model is continuously optimized with the Adam optimization technique with a learning rate of 10^{-3}. For the model training, mean squared error (MSE) is used between the ground truth keypoint coordinate vector and the predicted one.

C. Results

The faces are detected with the ResNet-single-shot face detector (SSD) from images or video. The SSD [7] is faster than Faster R-CNN since it does not need an initial object proposals

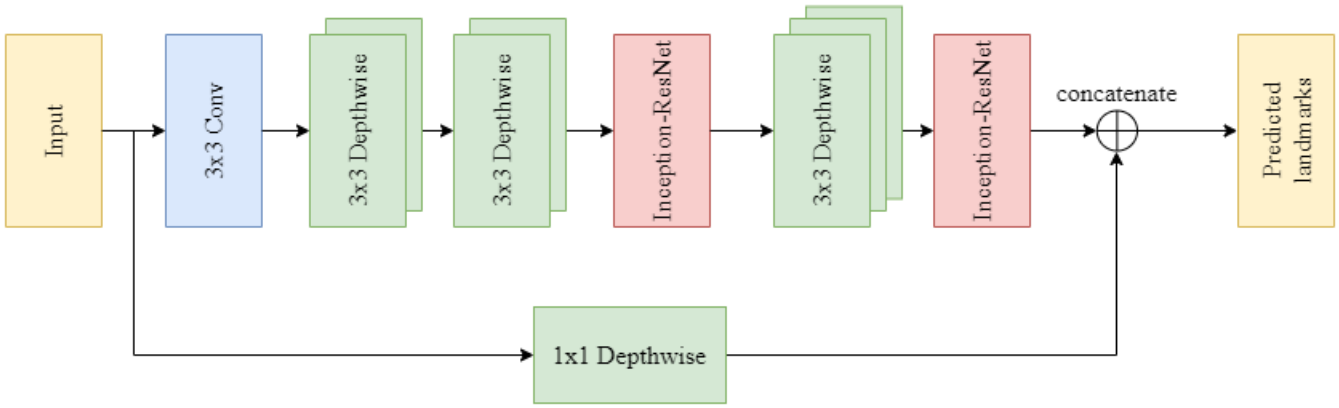

Fig. 2. Facial landmarks detection model design.

generation step. The model achieves a prediction accuracy of 71% with 16M parameters. The predicted landmarks are detected approximately well for frontal face and small head pose images. The facial landmarks are not detected well with extreme head poses such as yaw, pitch and roll conditions. The landmarks are detected well around the eye and nose area due to the inception-ResNet module that concentrates on important regions of the face. This information can be useful in the application such as driver monitoring and assistance systems that require driver's status information. The normalized mean error (NME) on the 300W dataset for the model achieves 6.5 on the full-set, 18.33 on the challenging subset and 7.2 on the common subset. The overall NME is greater than the challenging subset with contains extreme conditions of head poses, illumination and occlusion. Figure 3 shows the real-time facial landmarks detection with various head poses.

Fig. 3. Facial landmarks detection in real-time.

III. Conclusion

This paper predicts facial landmarks using VGG16 as a baseline model. It is replaced with the depthwise convolutions layers and an inception-resnet module to make the model light and extract a different level of features. The model still suffers from extreme head poses and illumination conditions. The future work is focused on improving the model to make it robust to different imaging conditions.

Acknowledgment

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2020-0-01808) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) and Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education(2019R1D1A3A0310384913).

References

[1] Shi, S., Facial Keypoints Detection. ArXiv 2017, abs/1710.05279.
[2] Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR 2015, abs/1409.1556.
[3] Sagonas, C.; Tzimiropoulos, G.; Zafeiriou, S.; Pantic, M. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In Proceedings of the IEEE International Conference on Computer Vision Workshops,Sydney, Australia, December 2 – December 8 2013 2013; pp. 397-403.
[4] Chrysos, G. G.; Antonakos, E.; Zafeiriou, S.; Snape, P. Offline deformable face tracking in arbitrary videos. In Proceedings of the IEEE international conference on computer vision workshops,Santiago, Chile, December 7 – December 13 2015; pp. 1-9.
[5] Shen, J.; Zafeiriou, S.; Chrysos, G. G.; Kossaifi, J.; Tzimiropoulos, G.; Pantic, M. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In Proceedings of the IEEE international conference on computer vision workshops,Santiago, Chile, December 7 – December 13 2015; pp. 50-58.
[6] Tzimiropoulos, G. Project-out cascaded regression with an application to face alignment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,Boston, MA, USA, June 7 – June 12 2015; pp. 3659-3667.
[7] Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.-Y.; Berg, A. C. Ssd: Single shot multibox detector. In Proceedings of the European conference on computer vision, Amsterdam, The Netherlands, October 8 – October 16 2016; pp. 21-37.