UIRNet: Facial Landmarks Detection Model with Symmetric Encoder-Decoder

Savina Colaco, Young Jin Yoon and Dong Seog Han*
School of Electronic and Electrical Engineering
Kyungpook National University
Daegu, Republic of Korea
savinacolaco@knu.ac.kr, skag2603@knu.ac.kr, dshan@knu.ac.kr*

***Abstract*—One of the challenging problems for facial landmarks detection is learning important features from faces that contain different deformation of face shapes and pose. These important features include eye centres, jawline points, nose points, mouth corners etc that are helpful in various computer vision-related applications. The detection of facial landmarks is difficult when faces have a lot of variation in different conditions. These conditions could be various imaging conditions such as illumination, occlusion, or head poses. In this paper, we propose a deep learning-based facial landmarks detection model called Unet-Inception-ResNet (UIRNet) to predict distinct feature points. The model predicts 68-point landmarks from the detected faces from digital images or video.**

***Index Terms*—facial keypoint detection, convolutional neural network, encoder-decoder**

I. Introduction

Due to various implications of face recognition, the performance gap between machines and the human visual system domain becomes a huge obstacle. Since recognising faces for humans can be done effortlessly but it is a challenging problem for the machines in the computer vision area over many years [1]. In particular, the identification methods for fingerprint or iris scans are more accurate than face recognition. Extensive research has been carried out for face recognition since it is an important method for the identification of the person. Face recognition is related to many domains such as computer and pattern recognition, security, biometrics, neuroscience, and multimedia processing. One of the difficult fields in face recognition is face alignment or facial landmark detection. The facial landmark detection goal is to detect the location of predefined facial landmarks, such as the corners of the eyes, eyebrows, the tip of the nose. It has been widely applied to a large variety of computer vision applications. For example, head pose estimation, facial re-enactment, 3D face reconstruction, etc. Recent advances in facial landmark detection focus on learning vital features from different deformation of face shapes and poses, different expressions, partial occlusions and so on. A simple framework is to construct features to depict the facial appearance and shape information by the convolutional neural networks (CNNs), and then learn a model, to map the features to the landmark locations. A CNN captures the complex semantic relationship between the features for a variety of applications. In this paper, we propose a symmetric encoder-decoder network with an Inception-ResNet module to better capture the landmarks for the detected faces in real-time.

II. Experiment

A. Implementation Details

The facial landmarks model is trained with a combined dataset of 300W [2] and 300VW [3]–[5] 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 100. We also apply early stopping once the model performance stops improving on the validation data. The model is continuously optimized with the Adam optimization technique [6] with a learning rate of 10^{-4}. The whole dataset is split with a ratio of 60 to 20% for training and testing subsets. The testing subsets are further split with 20% of validation subsets from testing subsets. For the model training, mean squared error (MSE), which is defined as the average of the square of all of the errors, is used between ground truth and predicted points.

B. Models

The proposed model adopts symmetric architecture called Unet [7] as a baseline model as shown in Fig. 1. The original Unet architecture is also called a contraction-expansive path or encoder-decoder. The Unet model is effective where the output is of similar size as the input and the output needs that amount of spatial resolution. As the name, the architecture resembles a U shape, which has a downsampling or encoder path on the left part and upsampling or decoder path on the right part. The downsampling path consists of recurring layers of two 3×3 convolutions followed by the Hswish activation function and batch normalization. The spatial dimensions are reduced with the application of a 2×2 max-pooling operation. This downsampling or encoder path helps to capture the contextual information in the image. Moreover, the number of feature channels is doubled with spatial dimensions being halved. In the upsampling step, several transposed convolutions are used and the corresponding feature map from the encoder is concatenated with 3×3 convolution. The final layer has a

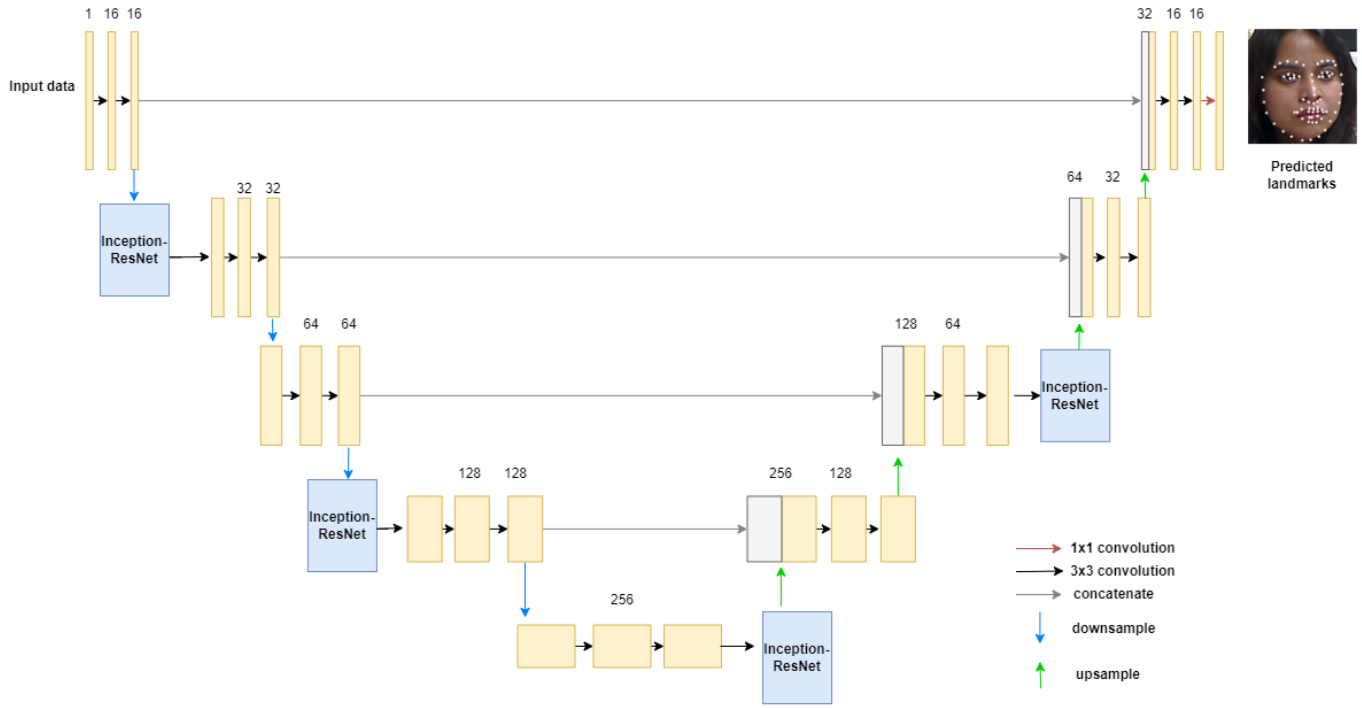

Fig. 1. Proposed symmetric encoder-decoder with Inception-ResNet module.

1×1 convolution to map the channels to the desired number of classes. The upsampling step helps to get the precise localization which is important for facial landmarks detection. The number of landmarks to be predicted can depend on the different target tasks. The Unet architecture combines the high-level features which are semantically low from the encoder and reused with upsampled output in the decoder. The proposed model called Unet-Inception-ResNet (UIRNet) is further extended with an inception-resNet module to get a better level of abstraction.

The Inception-ResNet module as shown in Fig. 2 is a tunable structure that gives several possibilities to change the number of filters in the layers. For the model, a different number of filters such as 1×1, 3×3, dilated filters are used to extract features. The different filters help to concentrate on the different parts of face images to detect facial landmarks. A skip connection performs an identity mapping by adding the original input features to the output of the stacked layers. The Inception-ResNet modules are placed after the 1^{st}, 3^{rd}, 5^{th} and 7^{th} stacked convolutions layers of UNet. Each layer in the module 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.

III. Discussion

The facial landmarks detection is being experimented with three models such as simple encoder-decoder, Unet and UIR-Net. The simple encoder-decoder model used in the experiment has a similar structure with Unet without the concatenation of a higher-level feature map from the encoder to upsampled output from the decoder. All the models are evaluated with the MSE loss function to measure the average of the squares of the errors. The faces are detected with the ResNet- single-shot detector(ResNet-SSD) face detector from images or video. The SSD [8] is faster than Faster R-CNN since it does not need an initial object proposals generation step.

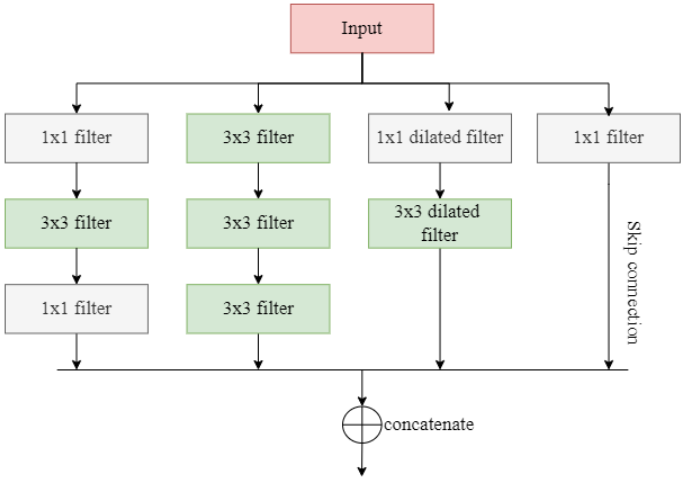

Fig. 2. Inception-ResNet module structure.

The simple encoder-decoder has a similar structure with Unet but without concatenation. As shown in Table 1, it

Table I: Comparison with different CNN models with proposed model

Model	Accuracy	Parameters (in Millions)
Simple encoder-decoder	64%	3.8M
UNet	39%	3.6M
UIRNet	73%	4.3M

achieves 64% of prediction accuracy with 3.8 M total parameters on the combined dataset. The encoder reduces the spatial dimensions in every layer and increases the channels. But decoder increases the spatial dimensions while the reduction in channels. Hence the spatial dimensions are restored to predict each pixel in the input image. Real-time detection of facial landmarks with simple encoder-decoder in Fig. 3, shows approximate localization of facial landmarks. It does not align well around the nose, mouth and jaw area.

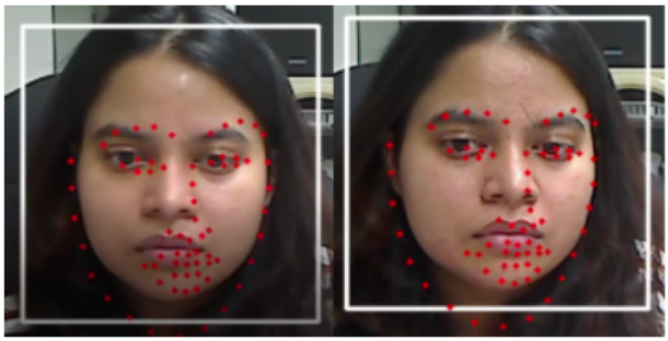

Fig. 3. Facial landmarks detection with simple encoder-decoder.

The Unet which is a symmetric encoder-decoder with the concatenation of high-level features with upsampled output gives 39% of prediction accuracy on the combined dataset with 3.6 M parameters. One of the reasons it shows lower accuracy on the combined dataset is less variation in data needed to tackle different conditions such as head pose, occlusion and illumination. The Unet is limited in extracting complex features from images. In Fig. 4, facial landmarks detected suffers completely with extreme variations in head poses.

The proposed model, UIRNet, is extended with the Inception-ResNet module and achieves 73% of prediction accuracy with 4.3 M parameters on the combined dataset. The Inception-ResNet allows changes to the number of filters in various layers without affecting the quality of the fully trained network. The different filters help to extract features at different scales especially in a different part of the face region. The skip connection added with the Inception makes the architecture deeper to prevent degradation problems. Fig. 5 shows the real-time detection of facial landmarks with the proposed model UIRNet. The UIRNet has better localization of facial landmarks compared to the other two models with most of the facial landmarks. It also shows approximate alignment with different head poses but suffers distortion with roll rotation around the z-axis.

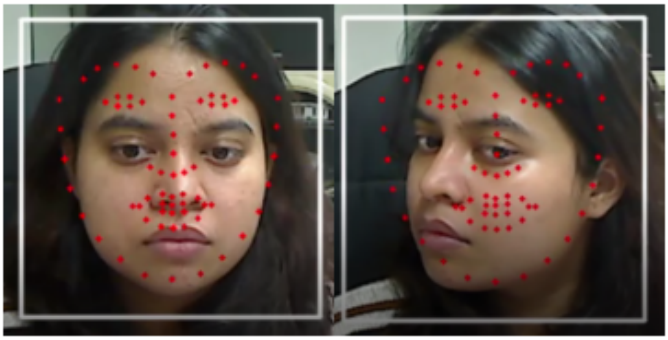

Fig. 4. Facial landmarks detection with Unet.

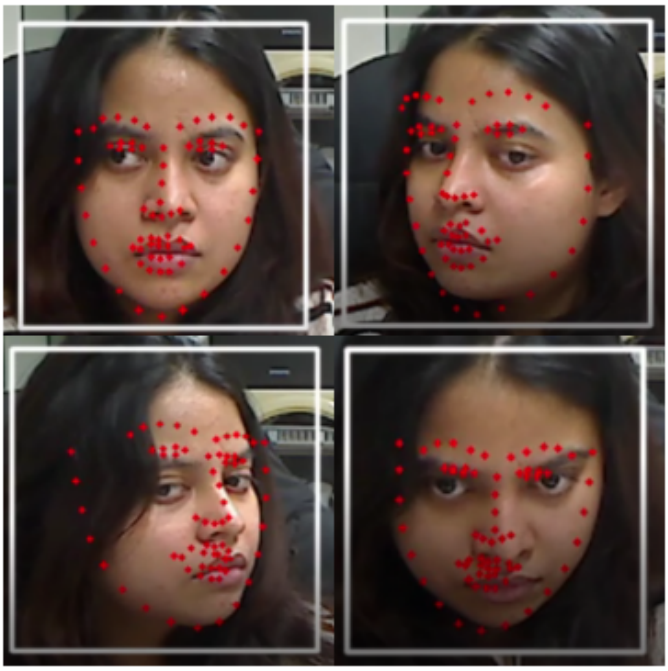

Fig. 5. Facial landmarks detection with UIRNet.

IV. Conclusion

In this paper, we proposed a model called UIRNet for predicting facial landmarks in real-time. The UIRNet uses a Unet as a baseline network with the Inception-ResNet module to improve prediction accuracy. We trained and compared our proposed model with other CNN models such as simple encoder-decoder and Unet with the combined dataset of 300W and 300VW. The proposed model showed better prediction accuracy than the other two models. Though prediction accuracy is improved, the model still suffers from large localization errors for extreme variations. For future work, we aim to improve our model for different unconstrained conditions for robust detection.

ACKNOWLEDGMENT

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (2021R1A6A1A03043144).

REFERENCES

[1] Shi, S., Facial Keypoints Detection. ArXiv 2017, abs/1710.05279.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv preprint arXiv:1412.6980, 2014.

[7] Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional Networks for Biomedical Image Segmentation. MICCAI.

[8] 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.