Facial Landmarks Detection using Scalable Deep Learning Model

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

***Abstract*—Facial landmark detection has achieved great performance by learning important features from 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. Facial landmarks detection for animal faces is more difficult than for human faces due to the wide variety of shapes. This paper proposes a scalable model for predicting facial landmarks on the detected animal faces from digital images or video.**

***Index Terms*—facial landmarks, animal faces, scalable model, inception module**

I. Introduction

Animals are an important part of our world and their needs are often expressed through faces, which can help us improve the well-being of animals in labs, farms and homes if understood properly. The study of animal faces is of major importance. Facial landmarks can help us better understand animals and foster their well-being via interpreting their facial expressions. Facial expressions reveal the internal emotions and psychological state of an animal being. Powerful technologies could be adopted or developed with useful indicators from facial expressions, e.g., cows widen their eyes and flatten their ears when they are scared, horses close their eyes in depression and sheep position their ears backwards when facing unpleasant situations. Though significant development has been made toward automatic understanding and interpretation of human faces, animal face analysis is largely unexplored in the computer vision community. Animal facial landmarks are detected using a scalable model and predict nine landmarks on unseen data.

II. Experiment

A. Implementation Details

The facial landmarks model is trained with the AnimalWeb dataset [1] with 21,921 faces from 334 diverse species and 21 animal orders across biological taxonomy. The faces are captured 'in the wild' conditions and are consistently annotated with nine landmarks on key facial features. The images in the dataset are resized to 112×112 resolution in grayscale. They are trained with a batch size of 32 and epochs of 150. The dataset is split into train, test and validation data. 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.

B. Landmarks Model

The model is adopted from [2] which consists of EfficientNet [3] as a baseline model which uses the compound scaling concept to scale the width, depth and resolution of the network. The model uses the MS-FC layer to extend the single-scale feature maps into multi-scale feature maps to enlarge the receptive field and catch better global features of animal faces. The model is further modified with a tunable inception module which consists of different filters such as 1×1, 3×3 and 5×5 to extract features. The different feature extraction from filters helps to focus on the different parts of face images to detect facial landmarks. Fig. 1 shows the architecture with EfficientNet, MS-FC layers and inception module.

C. Results

The model is 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 [4] is faster than Faster R-CNN since it does not need an initial object proposals generation step. The facial landmarks detection model achieves a detection accuracy of 54% with 0.8 M of total parameters. The inception module filters help to extract features at different scales, especially in different parts of the animal face region. The pig image is not used as one of the species of the dataset hence it was able to detect the animal face with appropriate landmarks detection. The model fails to localize well for different head poses of the animal faces and mostly works for frontal images.

III. Conclusion

This paper predicts facial landmarks using a scalable model such as EfficientNet. It is modified with MS-FC layers and an inception 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.

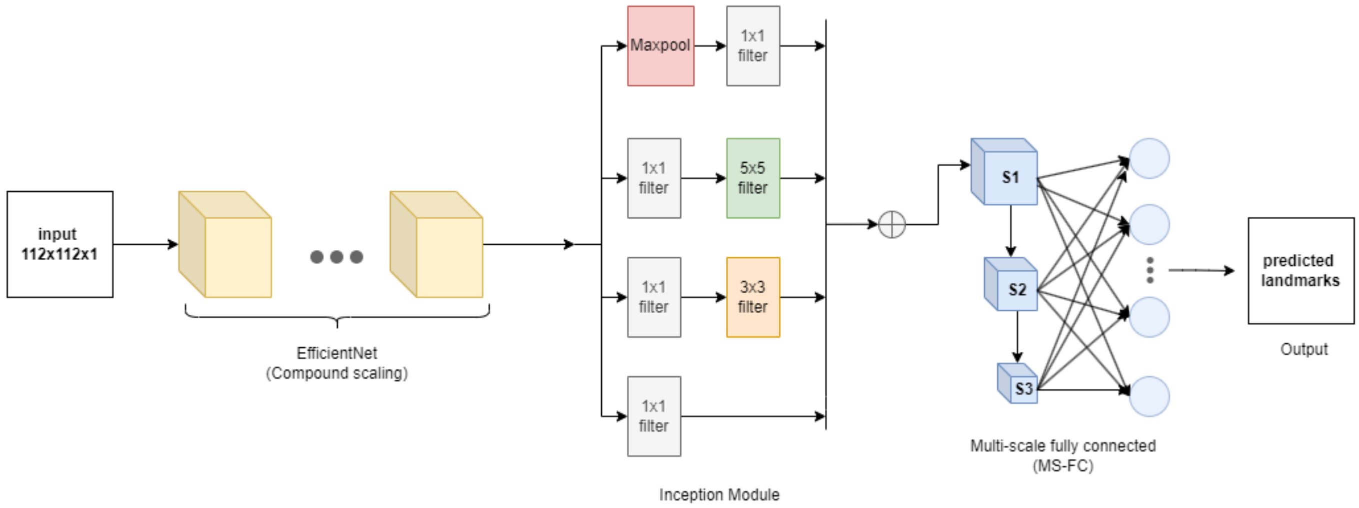

Fig. 1. Facial landmarks detection model design.

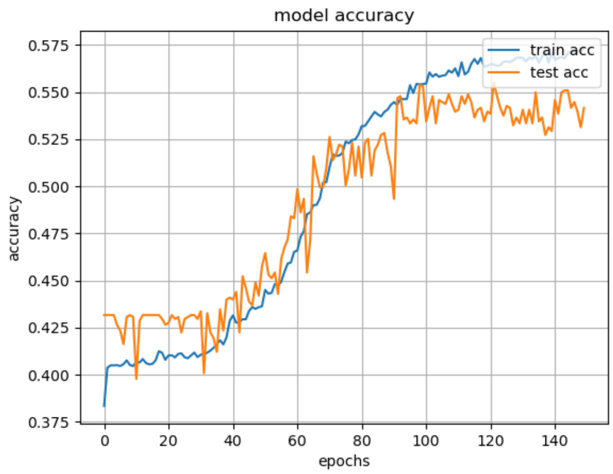

Fig. 2. Facial landmarks detection model accuracy.

Acknowledgment

This research was performed with the support of the 5G-based autonomous driving convergence technology demonstration platform project (P0013840) supported by the Ministry of Trade, Industry and Energy and the Korea Institute of Industrial Technology Promotion.

References

[1] M. H. Khan, J. McDonagh, S. H. Khan, M. Shahabuddin, A. Arora, F. S. Khan, et al., "AnimalWeb: A Large-Scale Hierarchical Dataset of Annotated Animal Faces," 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6937-6946, 2020.

[2] S. J. Colaco and D. S. Han, "Deep Learning-Based Facial Landmarks Localization Using Compound Scaling," IEEE Access, vol. 10, pp. 7653-7663, 2022.

[3] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in International Conference on Machine Learning, 2019, pp. 6105-6114.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C.-Y. Fu, et al., "SSD: Single Shot MultiBox Detector," in ECCV, 2016.