Facial Landmarks Localization with Compound Model Scaling

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

Abstract—Facial landmarks localization has widespread use in various facial analysis applications which can solve major problems in the computer vision field. Localizing the face key points gives the vital information needed for face analysis. The paper addresses a deep learning-based facial landmarks localization approach. EfficientNet is used to predict the facial landmarks on human faces which are mapped on the detected face in real-time. The model is evaluated with the adaptive wing loss function. The detector is examined with various head poses and occlusion conditions.

Index Terms—Facial landmarks localization, Facial landmarks, Adaptive wing loss, EfficientNet

I. INTRODUCTION

People can recognize faces effortlessly without giving much thought to it, this has been a challenging problem in computer vision area over many years [1]. Even though identification methods for fingerprint or iris scans are more precise, the research for face recognition has been the main focus since it is an important method for the identification of the person. Face recognition can be considered as a problem of identifying a person from face images. It is an important field of research with the interaction between psychologists and computer scientists. Facial landmarks detection, also known as face alignment, is one of the most researched areas in the computer vision. The facial landmark detection goal is to localize a group of facial points on human faces. These points could be eye corners, mouth corners, eye center, etc. This information from human faces can be widely used in face analysis. The driver inattention was one of the factors for fatal accidents recorded by the U.S Department of Transportation [2]. Understanding and monitoring the driver’s status could prevent fatal accidents. Face alignment or facial landmarks detection can help to detect keypoint positions on a person’s face which can be useful for understanding information such as driver face status in a driver monitoring system. In recent years, the deep learning approaches have shown outstanding performance in the computer vision. Facial landmarks localization along with CNN as made significant advancements over the past few years [3]. In this paper, we use the EfficientNet model to predict the facial landmarks on human faces. The predicted landmarks are mapped on to detected faces in real-time. The facial landmarks localization is tested with different head poses and occlusion as well to check its robustness.

II. EXPERIMENT

A. Model

The baseline EfficientNet network is developed by performing a neural architecture search using the AutoML MNAS framework, which optimizes both accuracy and efficiency. The architecture from the neural search architecture uses mobile inverted bottleneck convolution (MBCov), which similar to [4] and MnasNet. The regular convolutional layers mostly focus on finding the best architecture. The efficient Net model uses the concept of compound scaling to scale network length, width, and image resolution without changing the architecture’s predefined baseline network to improve the overall performance. The compound scaling method is performed with a grid search to find the relationship between different scaling dimensions of the baseline network with fixed resource constraints. The scaling of dimensions are done in a principled way.

Figure 1 Model scaling a) baseline network b) compound scaling.
\[ d = \alpha \phi, \quad w = \beta \phi, \quad r = \gamma \phi, \quad \text{s.t} \quad \alpha \cdot \beta^2 \cdot \gamma^2 \approx 2, \quad \alpha \geq 1, \beta \geq 1, \gamma \geq 1 \] (1)

The dimensions \( \alpha, \beta, \) and \( \gamma \) in equation 1 are constants that can be determined by a small grid search. \( \phi \) is a user-specified coefficient that controls how many more resources are available for model scaling, while \( \alpha, \beta, \) and \( \gamma \) specify how to assign these extra resources to network width, depth, and resolution respectively. Grid search allows us to get coefficients to get the desired target model size or computational budget. The design space can be reduced by restricting all layers to be scaled uniformly with the constraint ratio.

**B. Implementation Details**

The dataset used to predict facial landmarks is the 300W dataset and other images added as well. The 300W dataset \[5\] consists of XM2VTS, AFW, HELEN, LFPW, and IBUG with 68 landmarks. The images are resized to \( 112 \times 112 \) in grayscale. The network is trained on Nvidia GeForce GTX 980Ti GPU. The model is implemented with the Keras framework using a batch size of 100 and 300 epochs. We employ the Adam optimization \[6\] technique with the learning rate fixed to \( 10^{-3} \) throughout the training. The training of the dataset is done with 80-20% train and test split to avoid overfitting.

**C. Loss Function**

The loss function used for the experiment is the adaptive wing (Awing) loss \[7\] which was mainly used for heatmap regression

\[ \text{Awing}(y, \hat{y}) = \begin{cases} \omega \ln(1 + \frac{|y - \hat{y}|}{\varepsilon}) & \text{if } |y - \hat{y}| < \theta \\ A|y - \hat{y}| - C & \text{otherwise} \end{cases} \]

where \( y \) and \( \hat{y} \) are ground truth and predicted values respectively. Unlike wing loss, \( \omega \) is used as a threshold, \( \theta \) is the new variable threshold to switch between linear and non-linear part. The \( \omega, \theta, \varepsilon, \) and \( \alpha \) are positive values. \( A = \omega[1/(1 + \theta/\varepsilon)^{(\alpha - y)/(\alpha - y)}][(\alpha - y)](\theta/\varepsilon)^{(\alpha - y - 1)}](1/\varepsilon) \) and \( C = [\theta A - \omega \ln\{1 + (\theta/\varepsilon)^{(\alpha - y)}\}] \) are used to make loss function smooth and continuous at \( |y - \hat{y}| = \theta \). Similar settings from the paper \[7\] is used such as \( \theta = 14, \theta = 0.5, \varepsilon = 1, \) and \( \alpha = 2.1 \).

**D. Results and Discussion**

ResNet and single-shot detector (SSD) \[8\] models are used to detect the user’s face in the input image. The facial landmarks are predicted with baseline EfficientNet model. The predicted facial landmarks are mapped onto real-time webcam input. The red bounding box is the detected face. The baseline EfficientNet model with 4.2 parameters. The model is experimented with adaptive wing loss function. Figure 2 shows the facial landmarks detected with the EfficientNet model.

The efficientNet model with a dropout rate of 0.2 gives 86% of approximate accuracy. The localization of landmarks is close to ground truth values. The area around the eyes is better detected than the mouth. With different head poses such as roll, yaw, and pitch, the detection does not lose the detection drastically. The figure shows the landmarks detected with a 0.2 dropout rate.

The model with a 0.5 dropout rate gives better converges than a 0.2 dropout rate. The detection around eyes and mouth is better approximated. The landmarks are better localized with extreme head poses such as roll and yaw. The accuracy
achieved with a 0.5 dropout rate is 87%. Figure 6 shows the detection of landmarks for a 0.5 dropout rate.

![Figure 6 Landmarks detection with dropout rate = 0.5.](image)

The face landmarks detection is experimented with various head poses and occlusion to check its robustness. The detection with extreme head poses gives approximate key points localization. The landmarks localize roughly around the eye and nose area with glasses and mask occluding the face area. The eye and nose area are quite important features in the application such as driver monitoring system to get information such as eye movement, drowsiness, alertness, etc. Figure 7 shows the detection with occlusion.

![Figure 7 Landmarks detection with occlusion.](image)

III. Conclusion

In this paper, the model predicts the facial landmarks using baseline EfficientNet which are mapped onto the face shape of the human face. The model was able to predict approximate facial key points on the input face in real-time. The adaptive wing loss is examined as the loss function for the model. The future work is concentrated on building a better model considering different imaging conditions. Using a better model, we can analyze the head orientation of the user which can be useful aspects in the application such as the driver monitoring system. Data augmentation for various head poses, occlusions, and illumination scenarios can be applied to improve the performance of the models. The custom loss function can be designed for better prediction of key points. Boundary aware methods can be considered for the alignment of key points on the user’s face.

ACKNOWLEDGMENT

This work was supported by the Industrial Strategic Technology Development Program-Development of Vehicle ICT convergence advanced driving assistance system and service for safe driving of longterm driving drivers (20003519) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea).