Facial Landmarks Detection with MobileNet Blocks

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

***Abstract*—Facial keypoint detection is a challenging problem in the field of computer vision. The keypoint detection is done by predicting the coordinates of certain facial features. In this paper, facial keypoint detection is predicted using MobileNet techniques. The model is trained to predict facial key points using the webcam input data. The facial keypoints includes eyebrow corners, nose tip, eye corners and center, and lip points. The predicted keypoints are mapped onto the webcam input data and tested with various head poses. The adaptive wing loss is used to estimate the loss of the model.**

***Index Terms*—Facial key point detection, Facial landmarks, MobileNet**

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 is closely related to many domains such as computer and pattern recognition, multimedia processing, security, biometrics, neuroscience, and psychology. Face recognition can be considered as a problem of identifying an individual person from face images. It is an important field of research with the interaction between psychologists and computer scientists. Face recognition in videos requires face tracking possibly with three dimensions for head pose estimation. This helps us to analyze a person's attention and estimate gaze which is useful in an application such as a driver monitoring system. This task can be completed by detecting keypoint positions on a person's face. 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, nose center, etc. This information from human faces can be widely used in face applications such as face recognition, verification, emotion recognition, etc. 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 landmark detection can help to prevent such problems. Facial landmark detection can be used to estimate head variations, emotion, or gaze movement. In recent years, deep learning strategies have shown immense performances. The convolution neural network with deep structure processes the raw pixels of image input into multiple levels of feature representations which have the semantic abstracting property required for a variety of applications. In the paper, deep learning methods such as MobileNet techniques are used to predict the facial landmarks on human faces. The predicted facial points are tested with various human poses.

II. Experiment

A. Model

The backbone network architecture is adopted from the paper [3] for predicting the facial landmarks on the user's face. Figure 1 shows the model architecture. The neural network is built based on MobileNet version 2 [4] with architecture changes to predict facial keypoint detection.

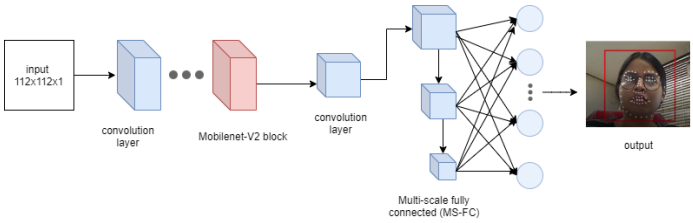

Figure 1: Face keypoint detection with MobileNet version 2

The MobileNet techniques have depthwise separable convolutions, linear bottlenecks, and inverted residuals. The performance of MobileNet techniques is satisfactory from the traditional convolution operations. The MobileNet can be compressed by adjusting the width multiplier hence making it a lightweight model with accelerated speed. The computational load of the network is significantly decreased. Table 1 gives the network configuration. The architecture input size is $112 \times 112 \times 1$. Each line represents the number of identical layers with symbol n. All layers in the same sequence have the same number of output channels c. Stride s is applied to each first layer of sequence. The expansion factor t is applied to the input size. Multi-scale fully connected layer is applied, which extends the single-scale feature maps into multi-scale feature maps. The S1, S2, and S3 denote the multi-scale fully connected layers which are later concatenated.

B. Implementation Details

The images are resized to 112×112. The dataset used to predict facial landmarks is the 300W dataset [5]. The dataset consists of XM2VTS, AFW, HELEN, LFPW, and IBUG with

Table I: Architecture configuration

Input	Operator	t	c	n	s
$1 \times 112 \times 112$	Conv3 $\times$ 3	-	64	1	2
$64 \times 56 \times 56$	Depthwise Conv3 $\times$ 3	-	64	1	1
$64 \times 56 \times 56$	Bottleneck	2	64	5	2
$64 \times 28 \times 28$	Bottleneck	2	128	1	2
$128 \times 14 \times 14$	Bottleneck	4	128	6	1
$128 \times 14 \times 14$	Bottleneck	2	16	1	1
(S1)16 $\times$ 14 $\times$ 14	Conv3 $\times$ 3	-	32	1	2
(S2)32 $\times$ 7 $\times$ 7	Conv7 $\times$ 7	-	128	1	1
(S3)128 $\times$ 1 $\times$ 1	-	-	128	1	-
S1,S2,S3	Full Connection	-	136	1	-

68 landmarks. The model is implemented with Keras framework using batch size 100 and 300 epochs. We employ the Adam optimization technique [6] with the learning rate fixed to 10^{-3} throughout the training. The network is trained on Nvidia GeForce GTX 980Ti GPU. The total number of parameters is 1,878,968 where, 1,859,544 is trainable parameters and 19,424 is non-trainable parameters.

C. Results and discussion

The difference between wing loss [7] and other loss functions such as smooth L1 [8] is the curvature which is controlled so for the difference in loss value, the error is more focused on harder images since the gradients are much larger. It is harder to get a zero value for wing loss and it never converges. The loss function used for the experiment is the adaptive wing (Awing) loss [9] which was mainly used for heatmap regression, defined

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

where y and $\hat{y}$ are ground truth and predicted values respectively. Unlike wing loss, ω is used as a threshold, θ is the new variable threshold to switch between the linear and non-linear part. The ω,θ,ε, and α are positive values. A = $\omega(1/(1+(\theta/\varepsilon)^{(\alpha-y)}))(\alpha-y)((\theta/\varepsilon)^{(\alpha-y-1)})(1/\varepsilon)$ and C = $(\theta A - \omega \ln(1+(\theta/\varepsilon)^{(\alpha-y)}))$ are used to make the loss function smooth and continuous at $|y-\hat{y}| = \theta$. The similar settings from the paper [9] is used such as θ = 14, θ = 0.5, ε = 1, and α = 2.1.

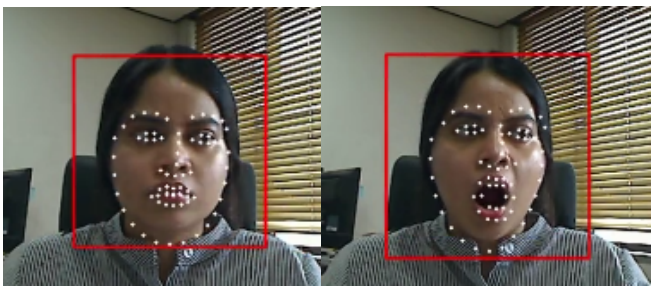

Figure 2: Real-time keypoints detection with MobileNet version 2

The face is detected using the ResNet and SSD model. The predicted facial landmark model is mapped onto real-time webcam input. Figure 2 shows the facial keypoints predicted on real-time webcam input. The accuracy of face landmarks prediction is approximately 90% with adaptive wing loss and it can be increased by increasing the number of epochs and change other parameters. Figure 3 shows the model accuracy plotted with adaptive wing loss for 300 epochs.

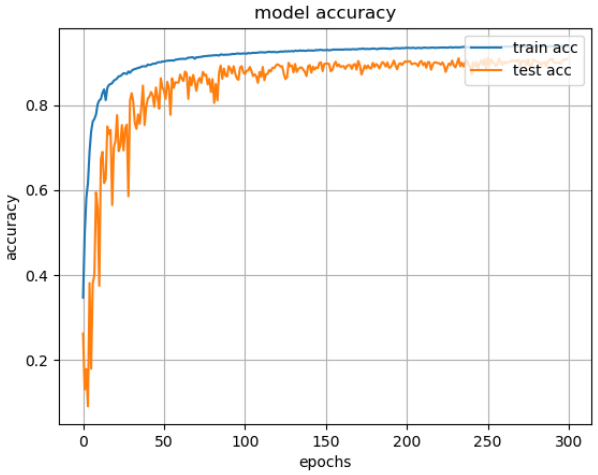

Figure 3: Model accuracy of Face keypoint detection with adaptive wing loss

The face keypoint detection experiments with various head pose to check the robustness. The detection with extreme head poses gives approximate key point localization. The keypoint information can be used to estimate the head pose which can be used in applications such as driver monitoring systems. Figure 4 shows the facial key points detection with various head poses.

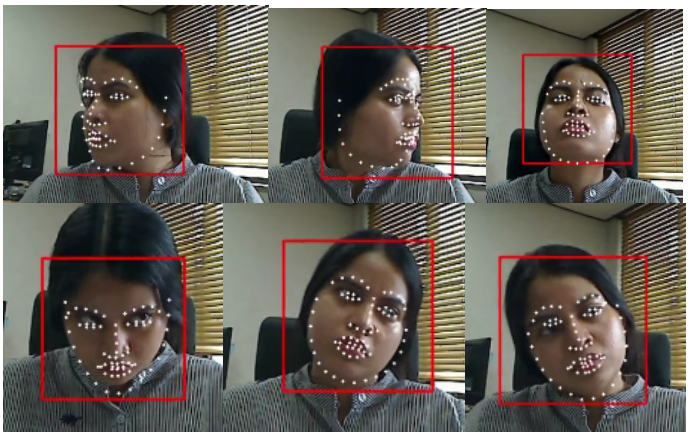

Figure 4: Keypoints detection with various head poses with adaptive wing loss

III. Conclusion

In this paper, the model predicts the key points which are mapped onto the webcam input data using the MobileNet version 2 network. The model was able to predict approximate keypoints on the input face in real-time. The future work is concentrated on building a better model considering different cases. Using a better model, we can analyze the head position of the user. Data augmentation can be applied to improve the performance of the models. Custom loss function can be designed for better prediction of key points. Boundary aware methods can be considered for the alignment of keypoints 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 long term driving drivers (20003519) funded By the Ministry of Trade, Industry & Energy (MOTIE, Korea).

References

[1] S. Shi, "Facial Keypoints Detection," arXiv preprint arXiv:1710.05279, 2017.

[2] U. S. D. o. Transportation, "Traffic Safety Facts: A Compilation of Motor Vehicle Crash Data," 2017.

[3] X. Guo, S. Li, J. Zhang, J. Ma, L. Ma, W. Liu, and H. Ling, "PFLD: A practical facial landmark detector," CoRR, vol. abs/1902.10859, pp. 1–11, Feb. 2019.

[4] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, 2018, pp. 4510-4520.

[5] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-Wild challenge: The first facial landmark localization challenge," in Proc. IEEE Int. Conf. Comput. Vis. Workshops, Dec. 2013, pp. 397–403.

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

[7] Z.-H. Feng, J. Kittler, M. Awais, P. Huber, and X.-J. Wu, "Wing loss for robust facial landmark localisation with convolutional neural networks," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Salt Lake City, UT, USA, 2018, pp. 2235–2245.

[8] P. J. Huber, "Robust Estimation of a Location Parameter," The Annals of Mathematical Statistics, vol. 35, pp. 73-101, 1964.

[9] X. Wang, L. Bo, and L. Fuxin, "Adaptive Wing Loss for Robust Face Alignment via Heatmap Regression," in Proc. IEEE/CVF Int. Conf. Comput. Vis, Seoul, Korea, 2019, pp. 6971-6981.