Journal of Industrial Information Technology and Application *ISSN(Online):2586-0852*
Vol.7 No.1

3D facial Landmarks Detection and Head Pose Estimation using Multi-task Learning and Vision Transformer

Hyunduk Kim [1,*], Sang-Heon Lee [1], Myoung-Kyu Sohn [1]
[1] Division of Automotive Technology, DGIST, Daegu, Republic of Korea

Abstract. In this paper, we present 3D facial landmarks detection and head pose estimation algorithms. To solve these two tasks simultaneously, we apply the multi-task learning technique. Our architecture consists of three components: a multi-head to deal with different tasks, a backbone to represent common features, and linear layers to output results. For the real-time process, we apply MobileViT as a backbone network. Moreover, we employ the PCGrad algorithm for stable convergence during training. To evaluate the performance of the proposed algorithm, we trained and tested on AFLW200-3D datasets, respectively. In the experiments, we demonstrate the experimental results for comparing the accuracy between MobileNetV3 and MobileViT.

Keywords; 3d facial landmarks detection; head pose estimation, multi-task learning; vision transformer

1. Introduction

3D face alignment, including 3d facial landmarks detection, head pose estimation, and 3d face modeling, is essential for many applications on face recognition, tracking, image restoration, and anti-spoofing [1,2,3,4]. Facial landmark detection studies are mainly divided into two approaches: the 2D approaches usually regress direct facial landmark coordinates or heatmaps based on visible facial parts, and the 3D approaches predict aligned 3D faces with images. The advantage of the 3D approach is it can detect occluded facial landmarks better than the 2D approach. Head pose estimation predicts the Euler angle representing the face orientation, such as yaw, pitch, and roll. Tradition approaches focus on face orientation as a standalone task. Recently, 3D Morphable Model (3DMM) based approaches have been introduced to gain insight into full facial

[*] Corresponding author: hyunduk00@dgist.ac.kr

Received: Sep 8, 2022; Accepted: Dec 20, 2022; Published: Mar 31, 2023

geometry. 3D alignment approach based on 3DMM estimates the 3D face parameters, such as pose, shape, and expression parameters.

Guo et al. [5] adopted a dense 3DMM fitting to reconstruct face mesh from a single image via a cascaded convolutional neural network. Moreover, they released a large training dataset using the face synthesizing method to generate 68K samples across large poses. Valle et al. [6] used an encoder-decoder CNN with residual blocks and lateral skip connections to estimate head pose. They also applied a multi-task learning technique to improve the performance of the head pose estimation task. Wu et al. [7] proposed a multi-task, multi-model, and multi-representation facial landmark refinement network. Recently, Vision Transformer [8] (ViT) achieved excellent image recognition results. After that, many variations have been proposed to enhance the original ViT algorithm. Especially, MobileViT V1 [9] combines CNN and ViT to reduce the number of parameters, and MobileViT V2 [10] applies separable self-attention to reduce complexity.

Inspired by these approaches, in this paper, we develop 3D facial landmark detection and head pose estimation algorithms using multi-task learning. Moreover, we apply the MobileViT network as the backbone for real-time performance. Fig. shows the overview of the proposed 3D facial landmark detection and head pose estimation approach.

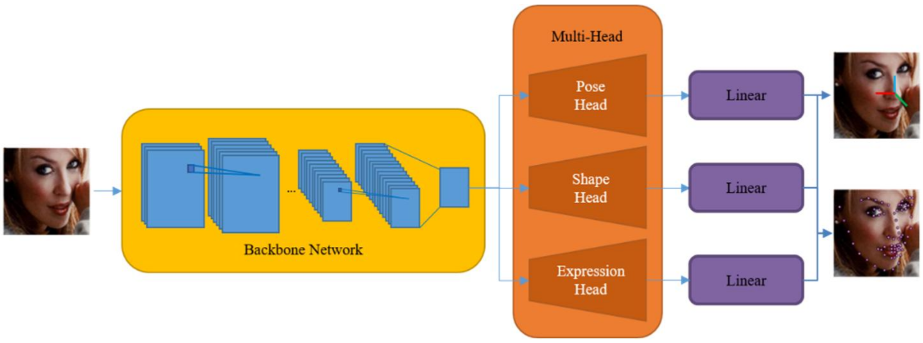

Fig 1. An overview of the proposed 3D facial landmark detection and head pose estimation approach.

2. Methodology

The proposed network, illustrated in Fig 1, consists of three parts: a backbone represents a common feature, a multi-head represents features to deal with each task separately, and linear layers extract output results. We use MobileViT V1-S and MobileViT V2-1.0 as a backbone network and extract features from the last 1×1 Conv. layer. A Multi-head includes two MBConv blocks with Depthwise Separable Conv. and Squeeze-and-Excitation layer and one Conv. layer with 1×1 kernel. The linear layers output pose, shape, and expression parameters. The Final outputs are calculated by the following equation.

$$S = R\left(\bar{S} + A_{shp}\alpha_{shp} + A_{exp}\alpha_{exp}\right) + t_{3d}, \tag{1}$$

where $R \in \mathbb{R}^{3\times3}$, $t_{3d} \in \mathbb{R}^{3\times1}$, $\alpha_{shp} \in \mathbb{R}^{40\times1}$ and $\alpha_{exp} \in \mathbb{R}^{10\times1}$ represent predicted the rotation matrix, the translation vector, the shape parameter, and the expression parameter. S and $\bar{S}$ are the output 3D face mesh and the given mean 3D face mesh, respectively. The rotation matrix and translation vector can be computed from the pose parameter.

In order to train the proposed network using a multi-task learning approach, we define the facial landmark loss, head pose loss, and parameter loss follows.

$$L_{lmk} = \sum_n L_{smL1}(l_n, l_n^*), n \in [1, N_l], \tag{2}$$

$$L_{pose} = \sum_{i=1}^{3} |p_i - p_i^*|, \tag{3}$$

$$L_{param} = \sum_{m\in\{r,s,e\}} ||\alpha_m - \alpha_m^*||^2, \tag{4}$$

where,N_t is the number of facial landmarks, * denotes ground-truth, $smL1$ is smooth L1 loss, m contains pose, shape, and expression. Finally, the multi-task loss is computed by combining them using a weighted sum of these losses as follows.

$$L_{total} = \lambda_1 L_{lmk} + \lambda_2 L_{pose} + \lambda_3 L_{param}, \tag{5}$$

3. Experiments

To evaluate the performance of the proposed 3D facial landmark detection and head pose estimation network, we use the 3DDFA dataset for training and the AFLW2000-3D dataset for testing [5]. All experiments are done on the PyTorch framework in NVIDIA RTX A5000 GPU. During the training, we use a SGD with momentum at 0.9 and weight decay at 0.0005 as an optimizer. We train the proposed network with 80 epochs, and the learning rate starts from 0.00001, rising to 0.0005 until 5 epochs, then decaying to 1/5 and 1/25 at 40 and 50 epochs. Moreover, we apply PCGrad [11] algorithm for stable convergence. During the testing, we use the Normalized Mean Error (NME) metric to quantify the 3D facial landmark detection error and the Mean Absolute Error (MAE) metric to measure the head pose estimation error.

$$\mathrm{NME} = \frac{1}{N}\sum_{i=1}^{N} \frac{\|l_i - l_i^*\|_2}{B}, \tag{6}$$

$$\text{MAE} = \frac{1}{N}\sum_i^N |p_i - p_i^*|, \qquad (7)$$

where N is the samples, * denotes the ground-truth, and B is the bounding box size.

We compare the accuracy between MobileNetV3 [12] and MobileViTs. As shown in Table I, while Mobile ViT V2-1.0 have fewer parameters than MobileNet V3-Small, it achieves similar accuracy. Moverover, MobileViT V1-S is more accurate than MobileNetV3-Small. However, MobileViT V1-S has most parameters than other networks. Even if we apply MobileViTs as the backbone, we may need more training data sets to achieve a more accurate result.

TABLE I. PERFORMANCE COMPARISON ON VARIOUS BACKBONE

		MobilNet V3-Small	**MobileViT V1-S**	**MobileViT V2-1.0**
Total Parameters		36,036,302	47,530,814	31,680,423
Facial Landmark (NME)	***0 to 30***	3.085	2.937	2.877
	30 to 60	3.964	3.564	3.770
	60 to 90	4.850	4.760	5.088
	All	3.967	**3.754**	3.911
Head Pose (MAE)	***Yaw***	3.651	3.659	3.696
	Pitch	4.777	4.531	4.891
	Roll	2.931	2.682	2.958
	Mean	3.786	**3.624**	3.848

4. Conclusions

In this paper, we proposed 3D facial landmark detection and head pose estimation algorithm using a multi-task learning approach and MobileViTs as the backbone network. The experimental results showed that the proposed methods are more accurate and efficient than MobileNet V3. In future works, we will develop remote PPG and Heart Rate estimation using multi-task learning and Vision Transformer. Finally, we will apply these face analysis algorithms to driver drowsiness detection.

Acknowledgment

This work was supported by the DGIST R&D Program of the Ministry of Science and ICT (23-IT-02). It was also supported by the Technology Development Program of MSS (S3237206).

References

[1] J. Shi, A. Samal, and D. Marx, “How effective are landmarks and their geometry for face recognition?,” Computer vision and image understanding, vol. 102, no. 2, pp. 117-133, 2006

[2] Q. Liu, J. Yang, J. Deng, and K. Zhang, “Robust facial landmark tracking via cascade regression,” Pattern Recognition, vol. 66, pp. 53-62, 2017.

[3] J. Cao, Y. Hu, H. Zhang, R. He, and Z. Sun, “Learning a high fidelity pose invariant model for high-resolution face frontalization,” Advances in neural information processing systems, vol. 31, 2018.

[4] Z. Yu, C. Zhao, Z. Wang, Y. Qin, Z. Su, X. Li, F. Zhou, and G. Zhao, “Searching central difference convolutional networks for face anti-spoofing,” In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5295-5305, 2020.

[5] X. Zhu, X. Liu, Z. Lei, and S. Z. Li, Face alignment in full pose range: A 3d total solution. IEEE transactions on pattern analysis and machine intelligence, vol. 41, no. 1, pp. 78-92, 2017.

[6] R. Valle, J. M. Buenaposada, and L. Baumela, “Multi-task head pose estimation in-the-wild,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.43, no. 8, pp. 2874-2881, 2020.

[7] C. Y. Wu, Q. Xu, and U. Neumann, “Synergy between 3dmm and 3d landmarks for accurate 3d facial geometry,” In 2021 International Conference on 3D Vision (3DV) pp. 453-463, 2021.

[8] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, T., M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv preprint arXiv:2010.11929, 2020.

[9] S. Mehta, and M. Rastegari, Mobilevit: light-weight, general-purpose, and mobile-friendly vision transformer. arXiv preprint arXiv:2110.02178, 2021.

[10] S. Mehta, and M. Rastegari, Separable Self-attention for Mobile Vision Transformers. arXiv preprint arXiv:2206.02680, 2022.

[11] T. Yu, S. Kumar, A. Gupta, S. Levine, K. Hausman, and C. Finn, “Gradient surgery for multi-task learning,” Advances in Neural Information Processing Systems, vol. 33, pp. 5824-5836, 2020.

[12] A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M. Tan, W. Wang, Y. Zhu, R. Pang, V. Vasudevan, Q. V. Le, and H. Adam, “Searching for mobilenetv3,” In Proceedings of the IEEE/CVF international conference on computer vision pp. 1314-1324, 2019.