Communications in Information and Systems
Volume 20, Number 4, 443–459, 2020

3D facial landmark detection based on differential cylindrical projection and multi-task learning

Takuma Terada, Ryusuke Kimura, and Yen-Wei Chen

Facial landmark detection is a fundamental step for face and expression recognition, and identification of personal attributes, analysis of race, and personal authentication. Numerous methods have been proposed for 2D facial landmark detection. However, 3D landmark detection is still a challenging task. In this paper, we propose a 3D facial landmark detection method based on differential cylindrical projection and multi-task learning. We first transform the 3D image to a 2D gray-scale image using cylindrical projection. We further enhance edges and facial parts (i.e. eyes, nose, mouth), which are useful for landmark detection, by using differentiation of the transformed 2D gray-scale image. Then we applied a convolutional neural network to detect the landmarks in the transformed 2D gray-scale image (differential cylindrical projection). Finally, we transformed the detected landmarks back to the original 3D image. Furthermore, we propose to use multi-task learning based on multi-labels pertaining to gender and age to improve detection accuracy. The code is available at: https://github.com/RU-IIPL/landmark_detection.

1. Introduction

3D facial data is more useful and more accurate for measurement of facial features than 2D facial data in many applications such as facial recognition and facial analysis [1, 2, 3]. In our previous studies, we used 3D facial data to identify the relationship between facial shape and genetic factors, based on the assumption that facial shape features can help classify individuals based on their place of birth, as observed between individuals from Hondo and Ryukyu in Japan [4, 5, 6].

Facial landmark detection and facial alignment are fundamental preprocessing in such applications [7, 8, 9]. Many landmark detection methods have been proposed for 2D facial images [12, 13, 14, 15, 16, 17]. Recently, the high-level feature representation of deep convolutional neural networks (DCNNs) has achieved great successes in the 2D landmark detection [16, 17].

On the other hand, there are a few researches on 3D landmark detection [18, 19, 20, 21]. Segundo [10] proposed a process of detecting facial landmarks and Nair [11] proposed to align the 3D data using the Point Distribution Model. However, when using 3D facial images alone, it is difficult to detect facial landmarks in detail.

Inspired by current achievements of DCNN for 2D facial landmark detection, we propose a DCNN-based landmark detection method for 3D facial images. We first transform the 3D image to a 2D gray-scale image using cylindrical projection. We further enhance edges and facial parts (i.e. eyes, nose, mouth), which are useful for landmark detection, by using differentiation of the transformed 2D gray-scale image. Afterwards we applied a convolutional neural network (CNN)-based multi-task learning approach to detect the landmarks in the transformed 2D gray-scale image (differential cylindrical projection). Finally, we transformed the detected landmarks back to the original 3D image. The overview of our proposed method is shown in Fig. 1.

The rest of this paper is organized as follows. In Section 2, we present a review of the related work. The proposed method is presented in Section 3. Experiments are presented in Section 4; our conclusion is given in Section 5.

2. Related work

In this section, we present some of the most relevant methods related to our work on facial landmark detection in 2D and 3D images.

2.1. Landmark detection for 2D images

Active shape model [12] and active appearance model [13] proposed by Cootes are widely used for localization of facial landmarks in facial shape or facial appearance. Some extended models, such as the deformable part model [14], are also proposed for facial landmark detection. However, in such model-based approaches, fitting the model with original image is a challenging task.

Recently, deep learning techniques have achieved great success in numerous computer vision tasks including facial landmark detection. Sun [16] proposed a CNN-based cascade network for facial point detection. Zhan [17] demonstrated that deep multi-task learning can significantly improve landmark detection accuracy.

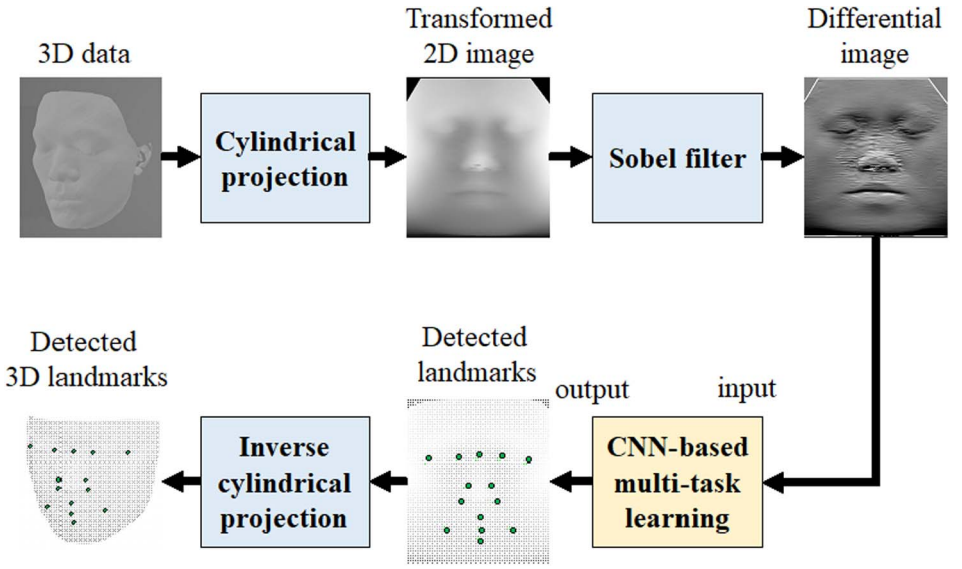

Figure 1: The overview of the proposed facial landmark detection approach for 3D facial image. The method consisting of four steps: transforming the 3D image to a 2D image using cylindrical projection; differentiating the transformed image using a sobel filter to enhance edges and facial parts (i.e. eyes, nose, mouth); extracting facial landmarks on the transformed 2D image using a DCNN; transforming the detected 2D landmarks back to the original 3D space.

2.2. Landmark detection for 3D images

Compared with the numerous 2D facial landmark detection methods, there are few 3D landmark detection methods. Böckeler [18] proposed an efficient 3D facial landmark detection algorithm with haar-like features [20] and anthropometric constraints, which generate a gradient image from depth data [19]. In order to avoid false-positive detection, the landmark detection is based on pre-defined sub-regions, such as the eye and mouth. However, it is not easy to define the accurate sub-regions automatically. Whitmarsh [21] proposed a 3D facial model for landmark detection on 3D face scans. The method is a 3D model-based approach, fitting the model with the original 3D scan is still a challenging task.

In our previous work, we proposed a deep learning-based 3D facial landmark detection method [22], in which the 3D data was first transformed to a 2D gray-scale image by cylindrical projection and then we used a deep

convolutional neural networks to detect the landmarks on the transformed 2D images. Zhang et al. [23] also proposed a similar method for 3D facial landmark detection, in which the transformed 2D image is called a position map.

In this paper, we present an improved version of our previous work [22]. We further enhance edges and facial parts (i.e., eyes, nose, mouth), which are useful for landmark detection, by using differentiation of the transformed 2D gray-scale image to improve the detection accuracy. We also show that the landmark detection accuracy can be improved by using a multi-task learning approach.

3. The proposed method

In this section, we describe details about the proposed method. As shown in Fig. 1, the proposed method consists of four steps: (1) we first transform the 3D image to a 2D gray-scale image using a cylindrical projection method; (2) we further enhance edges and facial parts (i.e. eyes, nose, mouth) by using differentiation of the transformed 2D gray-scale image; (3) we applied a convolutional neural network to detect the landmarks in the transformed 2D gray-scale image with a multi-task learning; (4) we transformed the detected landmarks back to the original 3D image. Compared with the previous method [22], our main improvements are steps 2 and 3. The comparison of the proposed method with the previous method is shown in Fig. 2.

3.1. Data acquisition and preprocessing

3.1.1. 3D facial dataset Our dataset consists of 750 3D facial data obtained from a 3D Mobile Scanner (ZScanner700CX), which are collected by the University of the Ryukyus, Japan [4, 5, 6]. The 750 subjects are Japanese including both males and females from 20s to 60s. The data distribution is summarized in Table 1. The scanned data is a group of 3D surface points (about 350K points) with three-dimensional coordinate (x, y, z) in each point, primarily focused on the facial region. A rough alignment was performed on each facial image manually so that the head was centered. Several 3D facial images are shown in Fig. 4(a).

3.1.2. Cylindrical projection (transformation of 3D data to a 2D gray-scale image) We first resample the 3D data and transform the resampled 3D image to a 2D gray-scale image using cylindrical projection method. And then perform the landmark detection on the transformed 2D

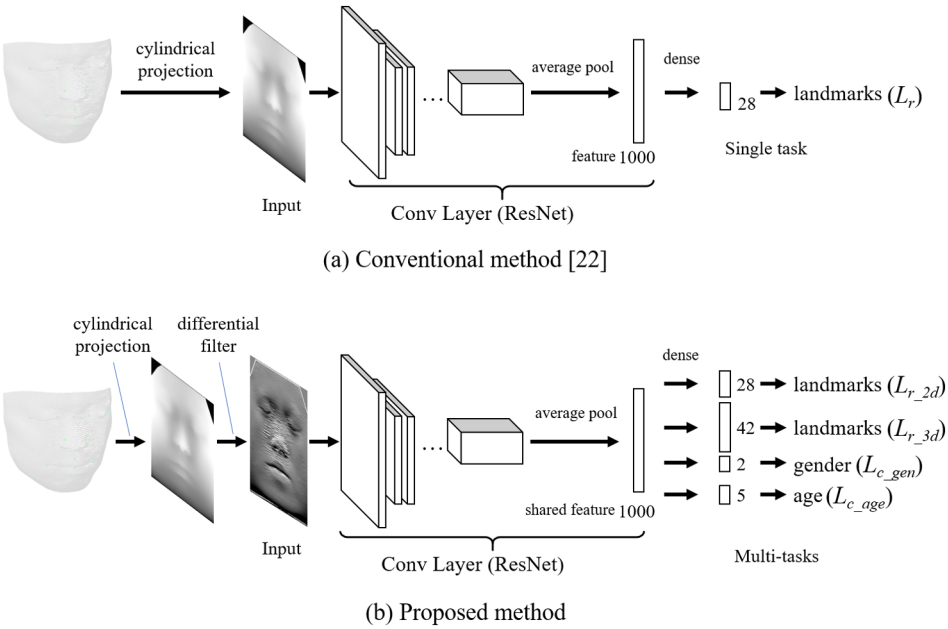

Figure 2: The architectures of the conventional method [22] (a) and the proposed method (b).

Table 1: Subject distribution of our dataset

	Twenties	Thirties	Forties	Fifties	Sixties
Male	297	93	23	6	1
Female	254	63	10	3	0

gray-scale image. We convert the Cartesian coordinate to the cylindrical coordinate as shown in Fig. 3. The cylindrical axis is the Cartesian z axis. The cylindrical coordinate (ρ, θ, z) is shown in:

$$(\rho, \theta, z) = \begin{cases} \rho = \sqrt{x^2 + y^2} \\ \theta = \tan^{-1} \dfrac{y}{x} \\ z = z \end{cases} \tag{1}$$

where ρ is a distance from the longitudinal axis, θ is the azimuthal angle and z is the longitudinal axis position.

After cylindrical coordinate conversion, we resample 144,000 surface points (400 points along z-axis and 360 points along θ). Regions such as

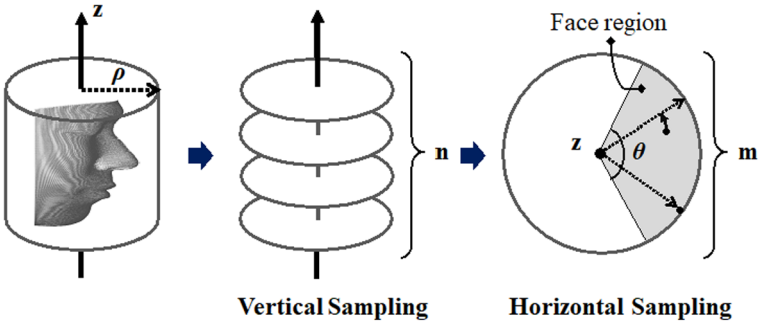

Figure 3: Cylindrical projection and resampling.

the ears and neck are excluded from facial parts. If the process is unable to obtain the corresponding points by sampling, we interpolate the same based on points of the nearest neighbor. Thus, the 3D image (surface image) can be represented as a 2D image $\rho(\theta, z)$ with a size of 360×400, and the distance ρ is used by the pixel value of the 2D image. Some image samples converted from the 3D to the 2D facial image are shown in Fig. 4. Figure 4(a) shows 3D facial images and their transformed 2D images are shown in Fig. 4(b). We attempt to extract the 2D landmarks from the transformed 2D image and then transform the landmarks back to the 3D image. The transformation from the 2D data to the 3D data is obtained using following equation:

$$(x, y, z) = \begin{cases} x = \rho \cos\theta \\ y = \rho \sin\theta \\ z = z \end{cases} \tag{2}$$

3.1.3. Differential image As shown in Fig. 4(b), the transformed 2D image is not clear enough to extract landmarks compared with normal facial images. To solve this problem, we propose to apply a differential filter to enhance the edges and facial features in the transformed 2D image. The differential image is obtained by the use of a Sobel filter for vertical derivative approximations Gy. The filter uses a 3×3 kernel which is convolved with the source image to calculate approximations of the derivatives. The filter can be written by following equation:

$$\mathbf{Gy} = \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ 1 & 2 & 1 \end{bmatrix} \tag{3}$$

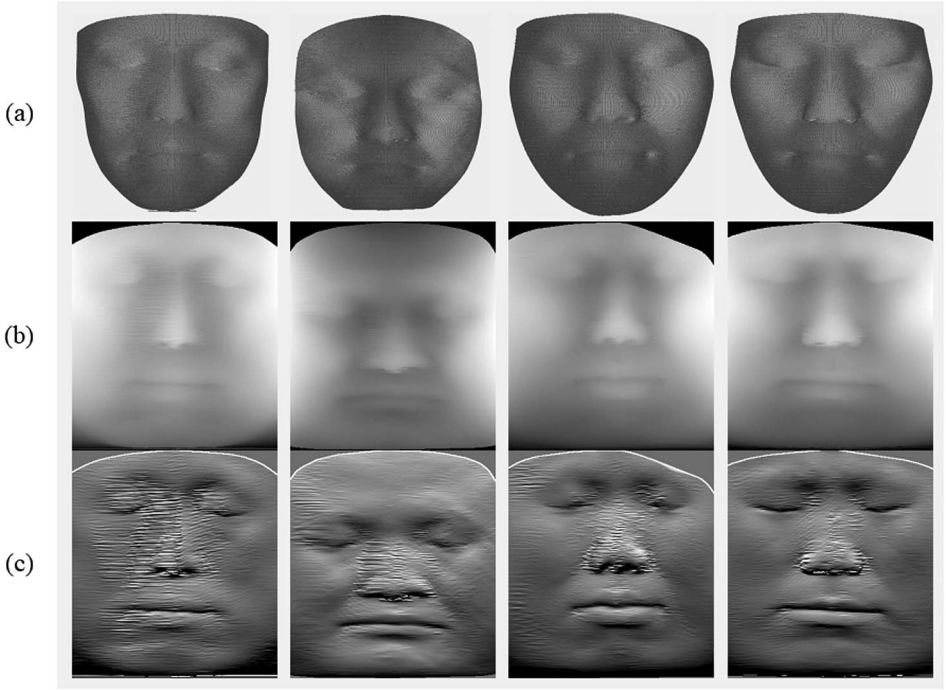

Figure 4: (a) 3D images, (b) transformed 2D images and (c) their differential images.

Differential images of the transformed 2D image (Fig. 3(b)) are shown in Fig. 3(c). We can see that facial parts such as eyes, nose, mouth become clearer and easier to recognize the landmarks.

Based on the 2D differential image, we manually annotated 14 landmarks as shown in Fig. 7 (red points), which are used as ground truth for training. Each landmark on the 2D transformed image has a pixel value ρ and coordinates (θ, z). By using equation (2), we can easily obtain 3D coordinates of the landmark on the 3D facial image.

3.1.4. Data augmentation In order to increase the number of samples, we adopted various data augmentation strategies. The augmentation is included a scaling, translation, flip of vertical axis and a combination of them. In scaling, the image is resized by a combination of three different scaling factors (0.9, 1.0, 1.1) in the horizontal and vertical axes. In translation, the image is moved either along the horizontal or vertical axis by an amount of width/10 or height/10. In flip, the image is flipped horizontally or verti-

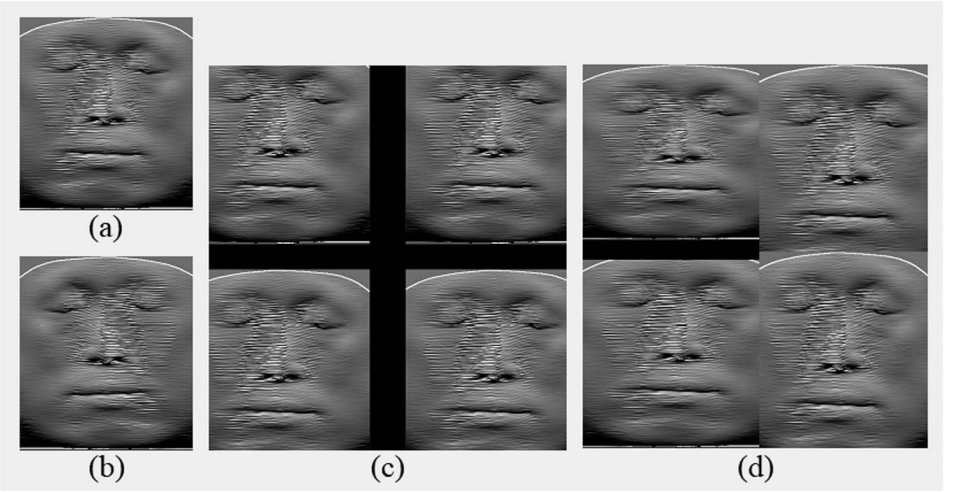

Figure 5: Examples of data augmentation. (a) original image, (b) its flipped image, (c) images with different translations, (d) images with different scales.

cally. The number of images is expanded eight (2^3) times by the scaling, four times by the translation and twice by the flip. In addition to above three augmentations, we also use a combination of the scaling and flip to augment the number of images for training. Example images of data augmentations are shown in Fig. 5.

3.2. Facial landmark detection

We used a regression method for facial landmark detection, and adopted a multi-task learning approach for improvement of detection accuracy.

3.2.1. The regression-based approach Assume the input image to the network is $I \in R^{W \times H}$, and the output estimated position $\mathbf{p} \in R^{N \times D}$, where $W \times H$ is the image size and N is the number of facial landmarks ($N = 14$). D is the dimensionality of the transformed image ($D = 2$). So, the dimension of the output $\mathbf{p}$ is 28. A regression function f, is given by the following equation:

$$\mathbf{p} = f(\mathbf{I}, \mathbf{W}) \tag{4}$$

where $\mathbf{W}$ are trainable parameters of function f.

The parameters $\mathbf{W}$ are optimized to minimize the error between estimated position $\hat{\mathbf{p}}$ and ground truth $\mathbf{p}$. The ground truth (correct coordinates of the landmarks) are annotated manually on the 2D differential image. The mean square error between $\hat{\mathbf{p}}$ and $\mathbf{p}$ is used as a loss function Lr as in:

$$Lr = \frac{1}{M}\sum_{i=1}^{M} \|\hat{\mathbf{p}}_\mathbf{i} - \mathbf{p}_\mathbf{i}\|^2 \tag{5}$$

where M is number of training samples.

3.2.2. Multi-task learning In our dataset, each sample contains gender, age and place of birth as meta-data in addition to 3D facial images. The objective is to converge learning effectively by using these multiple labels. As our second contribution, we proposed a multi-task learning approach to improve the detection accuracy. In the proposed multi-task learning method, there are following four tasks:

1. The first task is to estimate the 2D coordinates of the landmarks in the projection space, which is the same as the single-task learning described in Sec. 3.2.1. Since we have 14 landmarks, the output dimension is 28.
2. The second task is to estimate the 3D coodinates of the landmarks in the original 3D space. In this case, $D = 3$ and the output dimension is $3\times14 = 42$.
3. The third task is gender classification, which is a two-class classification. The output dimension is two (male or female).
4. The fourth task is age classification, which is a five-class (20s, 30s, 40s, 50s, and 60s) classification. The output dimension is five.

The cross-entropy loss Lc as shown in Equation (6) is used for classification tasks, where $\hat{p_k}^{(i)}$ is the predicted probability of sample i that belongs to class $k, y_k^{(i)}$ is equal to 1 if the target class of sample i is k. K is the number of classes. $K = 2$ for the task 2 (gender classification) and $K = 5$ for the task 3 (age classification).

$$Lc = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{K} y_k^{(i)} \log(\hat{p_k}^{(i)}) \tag{6}$$

In the multi-task learning, the loss functions Lm with the ratio of learning weight λ is given by the following equation:

$$Lm = \lambda_{r(2d)}L_{r(2d)} + \lambda_{r(3d)}L_{r(3d)} + \lambda_{c(gen)}L_{c(gen)} + \lambda_{c(age)}L_{c(age)} \tag{7}$$

Since the first two tasks are main tasks and the last two tasks are auxiliary tasks, we set both $\lambda_{r(2d)}$ and $\lambda_{r(3d)}$ as 1.0 and both $\lambda_{c(gen)}$ and $\lambda_{c(age)}$ as 0.001 based on our experiments. The proposed network architecture for multi-task learning is shown in Fig. 2(b). Note that the previous single task learning method without differential image [22] is shown in Fig. 2(a)

4. Experiments

4.1. Implementation details

In order to validate the effectiveness of the proposed method, we conducted computer experiments with our 3D facial dataset, described in Sec. 3.1. We divided 750 samples into three groups (each group contains 250 facial data). We used three-fold cross-validation to validate the detection accuracy. Two groups (500 data) were used as training dataset and one group (250 data) is used as test dataset. We repeat the experiment three times by changing the test dataset. The input is the 2D transformed gray-scale image or its differential image. The size of the input image is down-sampled to 90×100 from 360×400.

We used ResNet18 [24] as our baseline network. We also compared its performance with other network architecture, i.e. ResNet34 [24] and and SENet [25]. All methods were implemented using Tensorflow. We chose a mini-batch size of 32 and training epochs of 50, and the parameters were optimized using an Adagrad optimizer with an initial learning rate of 0.005 and the final learning rate of 0.003 (decay). We ran experiments on a GeForce GTX 1080 GPU with 8 GB of Video Memory.

We projected the estimated landmarks from the 2D transformed image to the original 3D image (Sec. 3.1.2), the distance (error) between the estimated 3D landmark and the ground truth landmark is used for quantitative evaluation of detection accuracy. The total mean error defined in Eqs. (8) and (9) is used as a measure of detection accuracy.

$$TMD = \frac{1}{MN} \sum_{m=1}^{M} \sum_{n=1}^{N} D_{mn} \tag{8}$$

$$D_{mn} = \sqrt{(x_{mn} - gx_{mn})^2 + (y_{mn} - gy_{mn})^2 + (z_{mn} - gz_{mn})^2} \tag{9}$$

where M and N are number of samples and nember of landmarks, respectively. (x, y, z) and (gx, gy, gz) are 3D coordinates of the estimated landmark and the ground truth landmark, respectively.

Table 2: Comparison of different network architechers

Backbone Network	ResNet18	ResNet34	SENet
Total Mean Error (pixel)	4.2 ± 2.5	4.2 ± 2.8	5.3 ± 3.4

Table 3: Comparison of the proposed methods with conventional method. "x" is indication of whether the method is applied. The total mean error is represented by $mean \pm SD$

Method	Differential Image	Multi-task Learning	Total mean Error (pixels)
CM [22]			5.5 ± 2.9
PM1	x		5.2 ± 2.6
PM2		x	4.4 ± 2.5
PM3	x	x	4.2 ± 2.5

We performed experiments with following four different projection-based methods and compared their performance: 1) CM, the conventional method (single task learning and without differential image) [22], which can be considered as a baseline of this paper. 2) PM1, the proposed method 1 (single task learning, but with differential image). 3) PM2, the proposed methods (multi-task learning, but without differential image). Note that PM3 is our final proposed method. PM1 and PM2 are used to show the effectiveness of the proposed differential image and multi-task learning, respectively.

4.2. Results and discussion

Firstly, we compared the performance of the proposed method (PM3) with different network architectures in Table 2. As shown in Table 2, both ResNet18 and ResNet34 achieved better performance than SENet, but no statistically significant difference between ResNet18 and ResNet34. Though ResNet34 has a deeper network architecture than ResNet18, its performance was restricted by the limited training samples. Since ResNet18 has the simplest architecture and the comparable performance as ResNet34, we use it as our backbone network in further experiments (Table 3, Figs. 6 and 7). Note that the proposed method can be combined with any state-of-the-art backbone networks.

In order to make a quantitative evaluation, we summarized the mean error (the average distance over testing samples) for each landmark between the detected result and the ground truth in Fig. 6. The total mean error over all 14 landmarks are summarized in Table 3.

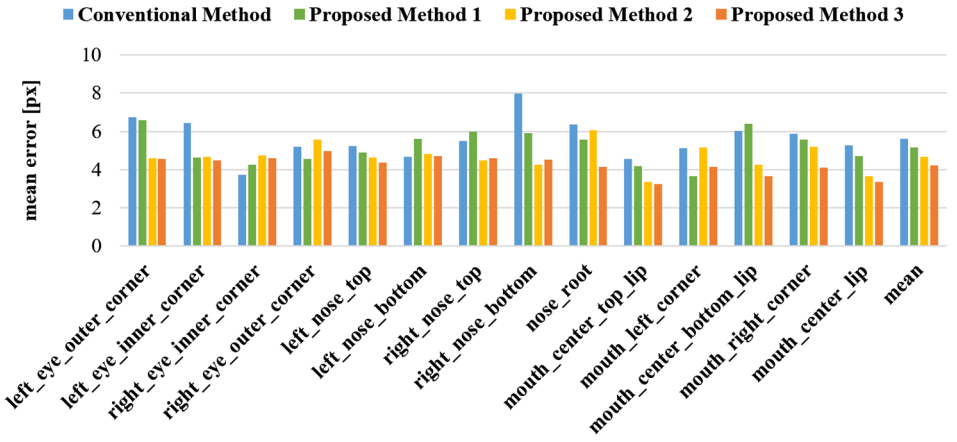

Figure 6: The result of facial landmarks detection related each facial part with the conventional method [22] and proposed methods.

We can see that the detection accuracy (total mean error) can be significantly improved from 5.5 ± 2.9 (pixels) to 5.2 ± 2.6 (pixels) and 4.4 ± 2.5 (pixels) by using the differential image and multi-task learning, respectively. If we combine both to detect the landmarks, the accuracy (total mean error) can be improved to 4.2 ± 2.5 (pixels). To test the statistical significance of the detection accuracy differences between the proposed methods and conventional method, we used the distance (error) for each landmark in each sample to employ t-test. The result of t-test confirmed the statistically significant (p-value < 0.05) superior performance of all proposed methods (PM1, PM2, PM3) against CM. We also confirmed the statistically significant (p-value < 0.05) superior performance of PM3 against PM1 and PM2, respectively. It means that using both differential image and multi-task learning is more effective than using only differential image or multi-task learning.

Qualitative landmark detection examples are shown in Fig. 7, which are displayed on the 2D transformed image, the differential image and their original 3D image. Green points are automatically detected landmarks by the conventional method and the proposed methods. Red points are ground truth (manually annotated landmarks on the 2D image). We can see that 3D landmarks are almost accurately detected by the proposed methods.

5. Conclusions

In this paper, we proposed a facial landmark detection method for 3D facial images based on a combination of differential cylindrical projection and

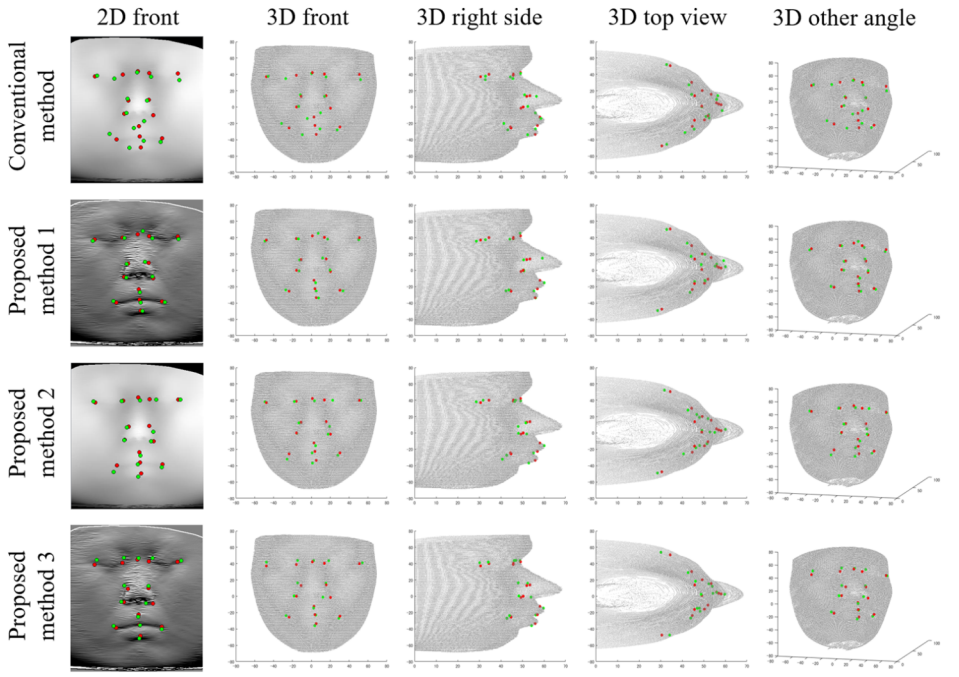

Figure 7: Qualitative landmark detection examples. Landmarks (green) in 2D and 3D images detected by the conventional method, the proposed method 1, the proposed method 2 and the proposed method 3 vs. ground truth (red).

multi-task learning. The effectiveness of the proposed method was validated by experiments. The proposed method achieved statistically significant superior perfomance compared with the conventional projection-based method [22]. The detection accuracy (total mean error) was improved from 5.5 (pixels) to 4.2 (pixels). As future works, we are going to modify the fixed sobel filter by learning an adaptive filter (adding a learnable layer as a learnable filter after the input layer) that is optimized for enhanced landmarks in the transformed 2D image. We are also going to add a self-attention module to automatically identify and enhance the useful features.

We should also note that the proposed method needs to transform the 3D point sets to a 2D image by cylindrical projection as a preprocessing, which is a time-consuming process. We are also going to combine the ideas of differentiation and multi-task learning with PointNet [26] or its improved version of PointNet++ [27], which can use 3D point sets directly for classification and segmentation, to improve the detection efficiency.

References

[1] X. Dong, Y. Yan, W. Ouyang, and Y. Yang, *Style aggregated network for facial landmark detection*. arXiv:1803.04108 Cs, Mar. 2018.

[2] R. B. Rusu, N. Blodow, and M. Beetz, *Fast Point Feature Histograms (FPFH) for 3D registration*. In: 2009 IEEE International Conference on Robotics and Automation, pp. 3212–3217, 2009.

[3] R. Zhao, Y. Wang, and A. M. Martinez, *A simple, fast and highly-accurate algorithm to recover 3D shape from 2D landmarks on a single image*. IEEE Trans. Pattern Anal. Mach. Intell., 2017.

[4] Y. Yamaguchi-Kabata, K. Nakazono, A. Takahashi, S. Saito, N. Hosono, M. Kubo, Y. Nakamura and N. Kamatani, *Japanese population structure, based on SNP genotypes from 7003 individuals compared to other ethnic groups: effects on population-based association studies*. Am. J. Hum. Genet., vol. 83, no. 4, pp. 445–456, Oct. 2008.

[5] D. Kitabayashi, G. Duan, T. Tateyama, R. Kimura, and Y. W. Chen, *Facial morphology analysis for a genetic association study—a scheme of 3D face shape alignment*. In: 2011 6th International Conference on Computer Sciences and Convergence Information Technology (ICCIT), pp. 870–873, 2011.

[6] M. Nakatsu, R. Kimura, X. H. Han, and Y. W. Chen, *Discriminant statistical analysis of local facial geometrical regions*. In: 2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR), 2015, pp. 351–355, 2011.

[7] Y. Sun, X. Wang, and X. Tang, *Hybrid deep learning for face verification*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 38, no. 10, pp. 1997–2009, 2016.

[8] T. Terada, T. Fukui, T. Igarashi, K. Nakao, A. Kashimoto, and Y. W. Chen, *Automatic facial image manipulation system and facial texture analysis*. In: 2009 Fifth International Conference on Natural Computation, vol. 6, pp. 8–12, 2009.

[9] K. Chen, S. Gong, T. Xiang, and C. C. Loy, *Cumulative attribute space for age and crowd density estimation*. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2467–2474, 2013.

[10] M. P. Segundo, C. Queirolo, O. R. P. Bellon, and L. Silva, *Automatic 3D facial segmentation and landmark detection*. In: 14th International

Conference on Image Analysis and Processing (ICIAP 2007), pp. 431–436, 2007.

[11] P. Nair and A. Cavallaro, *3-D face detection, landmark localization, and registration using a point distribution model*. IEEE Trans. Multimed., vol. 11, no. 4, pp. 611–623, 2009.

[12] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, *Active shape models-their training and application*. Comput. Vis. Image Underst., vol. 61, no. 1, pp. 38–59, Jan. 1995.

[13] T. F. Cootes, G. J. Edwards, and C. J. Taylor, *Active appearance models*. IEEE Trans. Pattern Anal. Mach. Intell., vol. 23, no. 6, pp. 681–685, Jun. 2001.

[14] M. Uricár, V. Franc, and V. Hlavác, *Facial landmark tracking by tree-based deformable part model based detector*. In: 2015 IEEE International Conference on Computer Vision Workshop (ICCVW), pp. 963–970, 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, *Imagenet classification with deep convolutional neural networks*. In: Advances in Neural Information Processing Systems, pp. 1097–1105, 2012.

[16] Y. Sun, X. Wang, and X. Tang, *Deep convolutional network cascade for facial point detection*. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3476–3483, 2013.

[17] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, *Facial landmark detection by deep multi-task learning*. In: European Conference on Computer Vision, 2014, pp. 94–108, 2013.

[18] M. Böckeler and X. Zhou, *An efficient 3D facial landmark detection algorithm with haar-like features and anthropometric constraints*. In: 2013 International Conference of the BIOSIG Special Interest Group (BIOSIG), pp. 1–8, 2013.

[19] F. C. Crow, *Summed-area tables for texture mapping*. In: Proceedings of the 11th Annual Conference on Computer Graphics and Interactive Techniques, New York, NY, USA, pp. 207–212, 1984.

[20] P. Viola and M. Jones, *Rapid object detection using a boosted cascade of simple features*. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), vol. 1, pp. 511–518, 2001.

[21] T. Whitmarsh, R. C. Veltkamp, M. Spagnuolo, S. Marini, and F. B. ter Haar, *Landmark detection on 3d face scans by facial model registration.* In: 1st International Symposium on Shapes and Semantics, pp. 71–75, 2006.

[22] T. Terada, Y.-W. Chen and R. Kimura, *3D facial landmark detection using deep convolutional neural networks.* In: Proceedings of 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD), pp. 390–393, 2018.

[23] J. Zhang, K. Gao, K. Fu and P. Cheng, *Deep 3D facial landmark localization on position maps.* Neurocomputing, vol. 406, pp. 89–98, 2020.

[24] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition.* In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778, 2016.

[25] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, *Squeeze-and-excitation networks.* arxiv preprint arXiv:1709.01507, 2018

[26] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, *Pointnet: deep learning on point sets for 3D classification and segmentation.* arXiv preprint arXiv:1612.00593, 2016.

[27] C. R. Qi, L.Yi, H. Su, and L. J. Guibas, *PointNet++: deep hierarchical feature learning on point sets in a metric space.* In: Proc. of International Conference on Neural Information Processing Systems (NIPS), 2017.

Takuma Terada
Graduate School of Information Science and Engineering
Ritsumeikan University
Shiga
Japan
E-mail address: t.terada.j@gmail.com

Ryusuke Kimura
Graduate School of Medicine
University of the Ryukyus
Okinawa
Japan
E-mail address: rkimura@med.u-ryukyu.ac.jp

Yen-Wei Chen
Graduate School of Information Science and Engineering
Ritsumeikan University
Shiga
Japan
E-mail address: chen@is.ritsumei.ac.jp

Received July 30, 2020