Multimedia Tools and Applications (2019) 78:3221–3238
https://doi.org/10.1007/s11042-018-6482-7

Robust facial landmark extraction scheme using multiple convolutional neural networks

Hyungjoon Kim, et al. *[full author details at the end of the article]*

Received: 20 March 2018 / Revised: 1 June 2018 / Accepted: 29 July 2018 /
Published online: 23 August 2018

Abstract

Facial landmarks are a set of features that can be distinguished on the human face with the naked eye. Typical facial landmarks include eyes, eyebrows, nose, and mouth. Landmarks play an important role in human-related image analysis. For example, they can be used to determine whether there is a human being in the image, identify who the person is, or recognize the orientation of a face when taking a photograph. General techniques for detecting facial landmarks can be classified into two groups: One is based on traditional image processing techniques, such as Haar cascade classifiers and edge detection. The other is based on machine learning techniques in which landmarks can be detected by training neural network using facial features. However, such techniques have shown low accuracy, especially in some special conditions such as low luminance and overlapped faces. To overcome these problems, we proposed in our previous work a facial landmark extraction scheme using deep learning and semantic segmentation, and demonstrated that with even a small dataset, our scheme could achieve reasonable facial landmark extraction performance under such conditions. Nevertheless, for more extensive dataset, we found several exceptional cases where the scheme could not detect face landmarks precisely. Hence, in this paper, we revise our facial landmark extraction scheme using a deep learning model called Faster R-CNN and show how our scheme can improve the overall performance by handling such exceptional cases appropriately. Also, we show how to expand the training dataset by using image filters and image operations such as rotation for more robust landmark detection.

Keywords Convolutional neural networks · Facial landmark · Semantic segmentation · Object detection · Faster R-CNN

1 Introduction

Facial landmarks extraction plays an important role in diverse areas such as human-related image analysis, personal color analysis, makeup recommendation and security. For instance, in the human image analysis, it can be used to determine whether there is a human being in the image, identify who the person is, or recognize the orientation of a face when taking a photograph.

Many techniques based on image processing have been proposed for face landmarks extraction with a limited accuracy. Recently, convolutional neural networks (CNN) [12] have shown overwhelming performance in the field of image classification compared to other existing threshold or edge-based image processing techniques, such as Haar classifiers or linear spatial pyramid matching [9, 22]. In addition, CNNs have shown quite good performance in object recognition that requires locating a specific object within an image. More recently, semantic segmentation, which extracts the exact position and shape of a specific object in the image, has drawn much attention. Hence, semantic segmentation using CNNs has been actively researched, and high-quality object recognition techniques have been proposed in various fields, including autonomous vehicles, medical image analysis, artificial intelligence robots, defective product detection, camera software, and image indexing.

In our previous work, we proposed a scheme for diagnosing personal color and applying virtual makeup based on facial landmarks [16]. To do that, we first extracted facial landmarks using an open library called dlib [11]. We then evaluated personal color, and determined makeup colors that matched well with the personal color. Finally, we put virtual makeup on the face, considering the facial landmark coordinates and their relative position. Hence, the quality of the virtual makeup application depends on the accurate extraction of facial landmarks. However, landmark extraction using an open library gave an unsatisfactory performance in the detection of landmark shape and location, especially when the landmark had various shapes and colors, such as hair.

To overcome this problem, we proposed a method for extracting facial landmarks using deep learning and semantic segmentation [10]. In the work, we constructed a deep learning model using pre-trained CNN layers, trained a small amount of face image data as ground truth to tune the layer weights, and then demonstrated the performance of our scheme by measuring how well pixels of the final output matched with ground truth. Even though its performance was acceptable in most cases, the scheme showed very poor performance in some exceptional cases.

In this paper, we solve the problem of our previous scheme by constructing another deep learning model and enlarging its training set to consider abundant cases similar to those that had show a poor outcome in our previous work. In addition, we use a new face detection model to make landmark detection more efficient.

The paper is organized as follows: Section 2 introduces several studies related to facial landmark detection. Section 3 describes our scheme in detail and in Section 4, the performance of our scheme is evaluated through experiments. Finally, Section 5 concludes this paper.

2 Related work

The first step in the extraction of facial landmarks in an image is to determine whether the image contains a face or not. This is important in terms of the efficiency and accuracy of facial landmark extraction. Many studies have been done to perform object or multi-object recognition in an image. After detecting the candidates in the face region, semantic segmentation is performed to extract facial landmarks. The semantic segmentation classifies an object not by rectangular unit, but by pixel unit. In other words, it determines to which class each pixel in the image belongs. Because deep learning techniques work well in image classification, there are many studies where semantic segmentation is performed using deep learning.

2.1 Object detection using deep learning

Ross Girshick et al. proposed Region-based Convolutional Neural Networks (R-CNNs) for object detection [7]. They first used a selective search algorithm to find all the regions that might contain an object. Those candidate regions are called region proposals and are represented using a bounding box. Each candidate is then evaluated using a CNN. However, the processing speed of an R-CNN is quite slow because every candidate must undergo feature extraction by CNN. To alleviate this, Shaoqing Ren et al. proposed Fast R-CNN [6]. This method first creates a feature map using a CNN, and then detects objects using RoI pooling and the softmax algorithm. Although Fast R-CNN improved the processing speed noticeably, creating region proposals still required a considerable amount of time. Due to this, Shaoqing Ren et al. further improved the processing speed using a revised model called region proposal network (RPN) [18]. They replaced the selective search algorithm with RPN. Because this model reduces the processing time considerably, it is called Faster R-CNN. On the other hand, Joseph Redmon et al. proposed You Only Look Once (YOLO) for real time object detection [17]. YOLO divides each image into N × N grids, and calculates the reliability of the grid. The reliability represents the accuracy of recognizing an object in each grid. Based on the reliability, the bounding box is moved around until the box with the highest object recognition accuracy is found.

2.2 Semantic segmentation

Jonathan Long et al. first proposed a deep learning model for semantic segmentation in 2015 [14]. They named their networks fully convolutional networks (FCNs), substituting the fully-connected layer that was used for image classification in the general CNN model with a convolution layer for pixel-level classification. They also proposed to skip the layer method to improve accuracy during the up-sampling process. Bilinear interpolation in FCNs is an up-sampling method that has shown a low-resolution problem. To solve this problem, Noh. H et al. proposed DeconvNet, and tried to add a deconvolutional network symmetrical to the convolutional network [15]. At the same time, they used the switch variable concept to remember the location of the max value during the max pooling calculation, and locate the position. V. Badrinarayanan et al. proposed SegNet [1, 2], which is a network that combines the advantages of DeconvNet and U-net. It reduces parameterization by removing the fully-connected layer used in DeconvNet. In addition, it improves memory cost by using pooling indices as opposed to copying and cropping the entire feature in U-net [19].

2.3 Facial landmark extraction

Erjin Zhou et al. localized facial landmark with Coarse-to-fine Convolutional Network Cascade [4]. They designed four-level convolutional network cascade. The first level estimates two bounding boxes: inner points and contour points. The second level predicts an initial estimation of the position. The third level refines each component. The fourth level is used for improving quality of detecting mouth and eyes by considering rotated image patches. Umut et al. performed face segmentation using conditional random fields and deep learning [8]. They formulated a conditional random field (CRF) over a four-connected graph as end-to-end trainable, convolutional, and recurrent networks, and then estimated them via an adversarial process. They showed that their model could achieve state-of-the-art performance through evaluation against two standard benchmark datasets for semantic face segmentation.

3 Facial landmark extraction

In this section, we briefly describe how facial landmarks are extracted from an image. As a first step for facial landmark extraction, we crop the facial region from the image using the Faster R-CNN for efficiency and accuracy. This is because Faster R-CNN is known to give superb detection performance despite a relatively slow processing speed. To extract facial landmarks, we construct a network architecture for semantic segmentation, and then build a training dataset with the ground truth that was created manually on the constructed network. The network architecture used in this paper is SegNet, which is known to supplement the disadvantages of existing semantic segmentation methods. SegNet is also suitable for classifying images into pixel units.

The overall architecture of our scheme is shown in Fig. 1. Firstly, the facial region in the input image is detected using Faster R-CNN. Then, the facial region is cropped from the image and given to SegNet for semantic segmentation. Lastly, facial landmarks are extracted and represented in the input image.

3.1 Faster R-CNN

Usually, portraits images contain a face with different sizes and locations. For more effective landmark extraction, we resize every input image to 224 ×224 pixels. Hence, if the face occupies a significant portion of the image, face landmark detection can be done easily. However, if the image contains the whole body or the upper body, then its face size is relatively small in the image, and face landmark detection could be difficult. The situation becomes even worse after resizing because the face size becomes smaller. Figure 2 shows an example of landmark extraction for small faces. For the original image in Fig. 2 (a), Fig. 2(b) shows inaccurate facial landmark extraction results.

To solve this problem, we determine candidate facial regions before facial landmark detection by using Faster R-CNN. Faster R-CNN is one of popular high-performance object detection methods based on deep learning. The network architecture of Faster R-CNN is

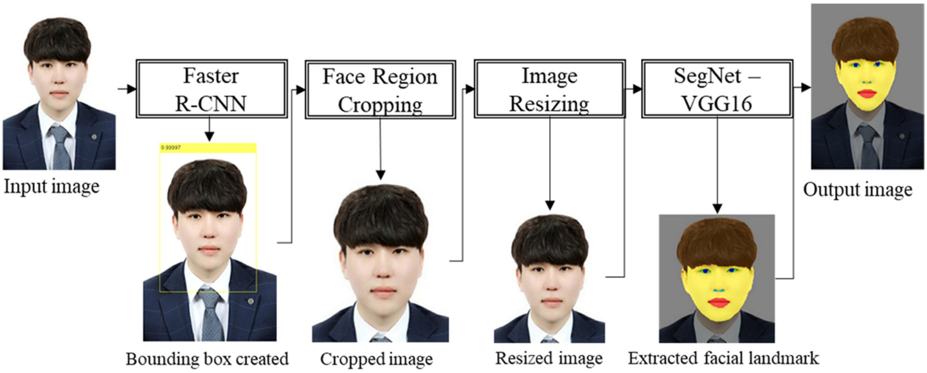
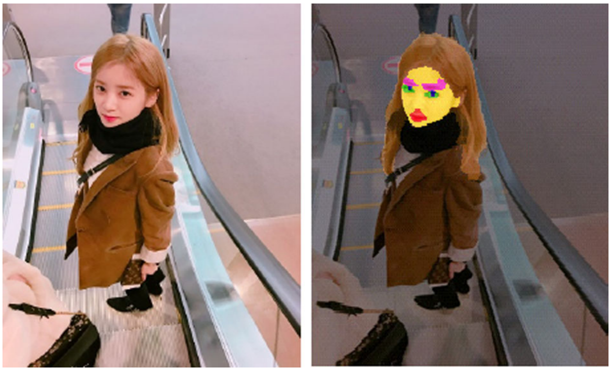


Fig. 1 Steps for facial landmark extraction



(a) Original image (b) Extracted landmarks

Fig. 2 Inaccurate landmark extraction result

shown in Fig. 3. Because we only need an approximate facial region, we construct a relatively lightweight network model.

We collected 304 face images for the training set. For more robustness and better performance, we created more images by adding noise such as Gaussian noise and Poisson noise to the original face images. As a result, we constructed a training dataset of 1216 facial images. A few examples of the training set are shown in Fig. 4. Figure 4(a) shows three original face images, and Fig. 4(b) shows the corresponding modified images using Gaussian noise. The face in the image is marked using a yellow bounding box. Hence, all the faces in the bounding box become the ground truth of our model. Later, we explain in detail why we added such noise to populate the dataset.

3.2 SegNet

As mentioned above, we used SegNet to extract facial landmarks. The basic network structure and method are almost the same as those used in [1]. SegNet consists of an encoder network and a decoder network with 13 convolutional layers based on DeconvNet and U-net. The

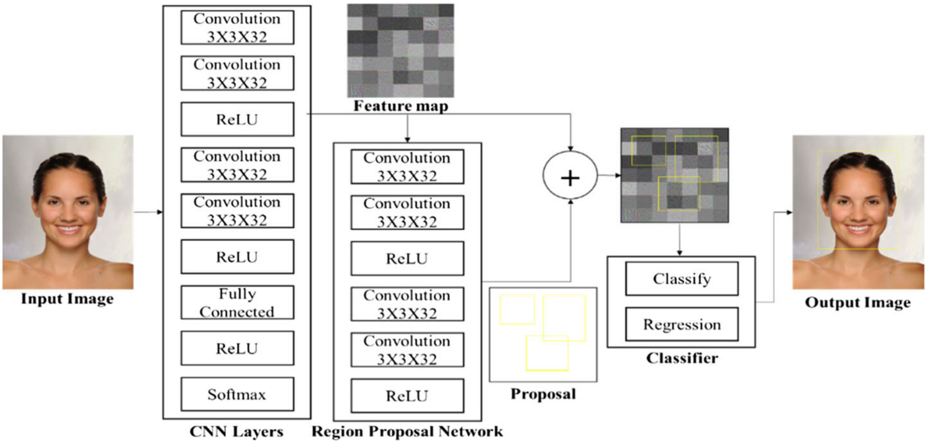
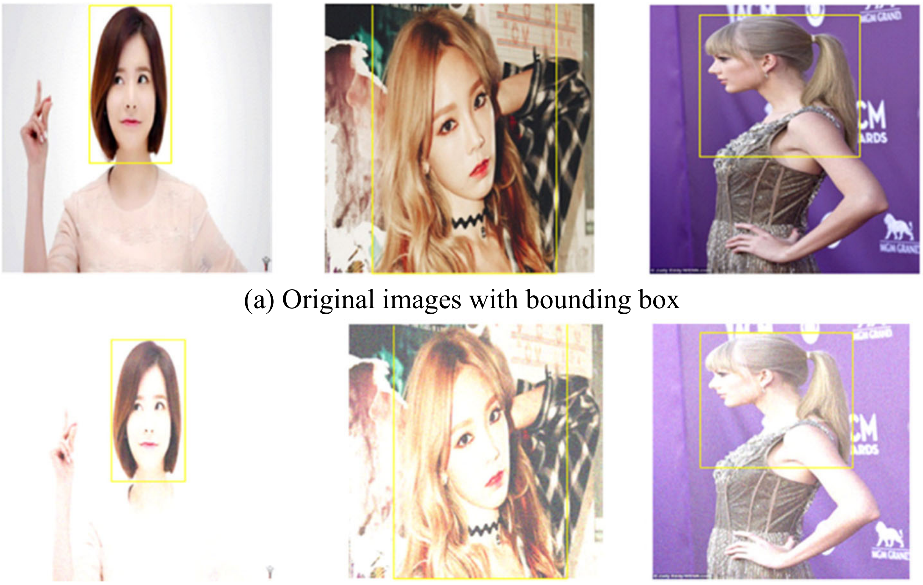


Fig. 3 Faster R-CNN architecture


(a) Original images with bounding box

(b) Modified images with bounding box

Fig. 4 Examples of Faster R-CNN training dataset

encoder network is constructed in the same way as the first 13 convolutional layers of the VGG16 [21] network. The VGG16 is a deep learning technique designed to classify an ILSVRC dataset [20] with 1000 classes. At the back of the encoder network, however, three fully-connected layers at the end of the VGG16 are removed to reduce the number of parameters in SegNet. The network has also 13 decoder layers, which correspond to the encoder layers. The encoder network adjusts the weight with a large dataset and finally provides a set of features, which are batch-normalized at each layer. During this process, the output is sub-sampled through max pooling, and at the same time, resolution loss of the feature map occurs because the max pooling calculation removes all the values except the max value within the size of the window. Therefore, it is necessary to store the information of the feature maps in the encoder network before sub-sampling. Thus, our proposed scheme stores pooling indices, which indicate the location of the maximum feature value in the pooling window. This

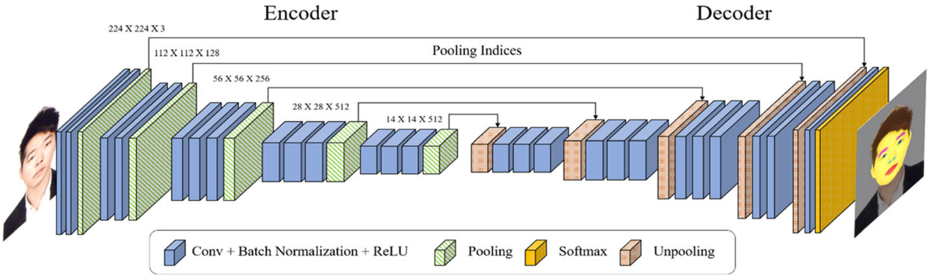
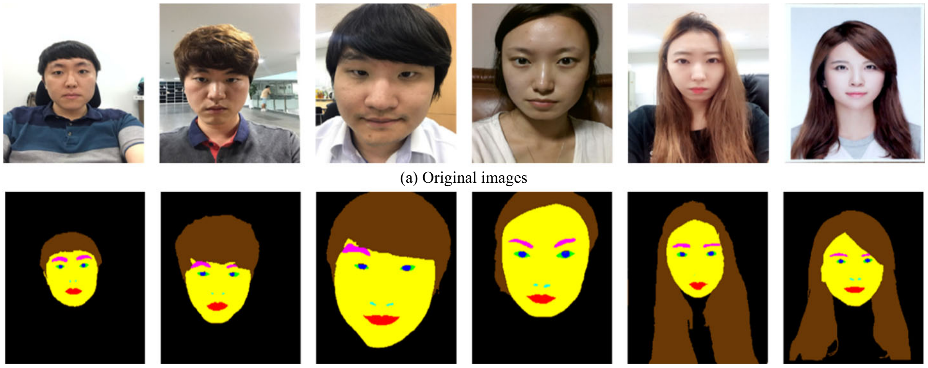


Fig. 5 SegNet architecture



(a) Original images

(b) Ground truth images

Fig. 6 Samples in the dataset

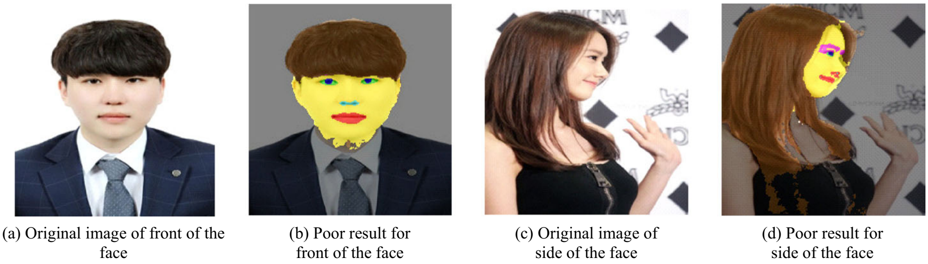
approach is more efficient than storing the feature map as U-nets, as it requires less memory. These pooling indices are used for up-sampling in the decoder layer, and finally convolution by the decoder filter. Lastly, the final decoder output is fed to the multi-class soft-max classifier, which provides a classification for each pixel. The detailed architecture of SegNet used in our work is described in Fig. 5.

3.3 Dataset construction

We defined 9 classes to describe facial landmark features: face skin, sclera, pupil, eyebrow, hair, lip, between-the-mouth, nostril, and background. "Face skin" is the area of skin not identified as other landmarks on the human face. "Sclera" is the white part of the eye, and "pupil" is the circular area of the center of the eye. "Between mouth" indicates the region where a shadow is formed when a person's mouth is open, or an area where the teeth are visible, and which does not belong to the skin or lips. All areas except face and hair are defined as "background."

Table 1 Facial landmark and rgb values

	R	G	B	Color
Hair	106	57	6	🟫
Face skin	255	255	0	🟨
Sclera	0	255	0	🟩
Pupil	0	0	255	🟦
Eyebrow	255	0	255	🟪
Nostril	0	255	255	🟦
Lip	255	0	0	🟥
Between mouth	255	255	255	⬜
Background	0	0	0	⬛


(a) Original image of front of the face

(b) Poor result for front of the face

(c) Original image of side of the face

(d) Poor result for side of the face

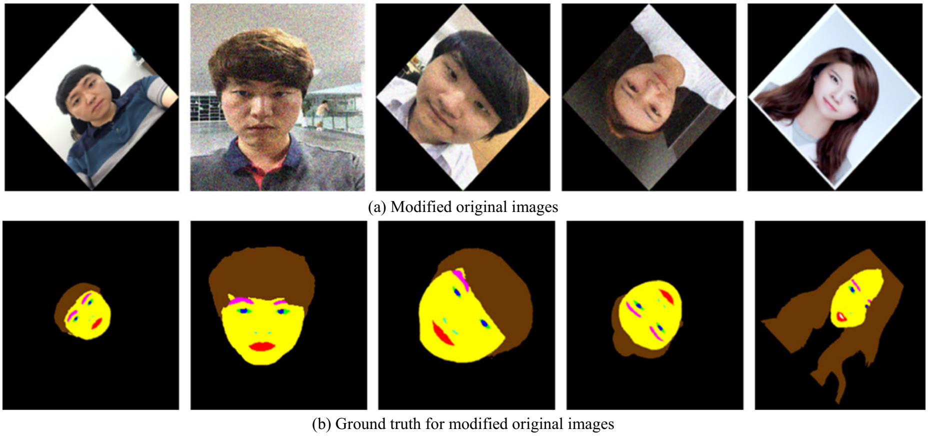
Fig. 7 Imprecise landmark extraction

In our previous work, we used a total of 150 front view face images as a dataset. Among those 150 images, 140 are Korean and 10 are Westerners. As for gender, 80% are women and 20% are men. In general, there is almost no difference between male and female in regarding to the position of facial landmarks, and their appearance is very similar. Women's hair, on the other hand, has various lengths, colors, and shapes compared to men. Therefore, it is necessary to train them using diverse hair shapes and colors. As for age, 80% are in their teens to 30s and 20% are in their 40s or older. Each image is a colored image with a resolution of 224 × 224 pixels. We also created a ground truth for evaluating the accuracy of facial landmark extraction. The ground truth represents nine classes with different RGB values. The values are described in Table 1 in detail.

Figure 6 shows a few samples in the created dataset. For the six original color images in Fig. 6(a), (b) represents the ground truth images extracted from the original image, which consist of the landmarks expressed using the different RGB colors defined in Table 1.

In some cases, this method resulted in imprecise landmark extraction. Figure 7 shows such cases, where the borderline between the jaw and the neck or the side view of the face were not clear, as shown in Fig. 7(b) and (d) for the input image in Fig. 7(a) and (c), respectively.

We solved this problem by constructing a more extensive training set. More specifically, in addition to the original 150 front view face images, we collected 30 images of faces in profile view. Usually, facial landmarks can look different for the same person in terms of color and shape, depending on diverse factors such as camera hardware, light, or


(a) Modified original images

(b) Ground truth for modified original images

Fig. 8 Ground truth for modified images

the external environment. Hence, for a total of 180 original images, we performed diverse operations such as image rotation and noise filtering to produce images of different features. This approach is known to be effective for improving the robustness of the model by scaling up the dataset during machine learning. We considered two types of operations for this purpose, which were noise filtering and image rotation. The noise filters that we considered in this paper included Gaussian noise and Poisson noise. With the combination of noise filtering and image rotation, we enlarged the dataset considerably. Figure 8 shows the result of face landmark extraction for the noisy and rotated images. Figure 8(a) shows the modified images, and Fig. 8(b) shows their ground truth images. The ground truth of the images with Gaussian noise and Poisson noise is the same as the ground truth of the non-noised images.

To summarize, we used a total of 180 original front view and side view facial images. By introducing noise and performing rotation, we obtained a total of 6840 images for the experiment. Details for the dataset are shown in Table 2.

3.4 Training weight

Depending on the facial landmark, the total number of pixels in the class could be different. For example, because the background includes all the areas except the face and the hair, the background could occupy more than 60% of the total pixels, which is the largest portion of the total pixels. Because the difference in the number of pixels of each class could degrade extraction accuracy, weights for the nine classes are calculated using median frequency balancing [3], which can be defined by Eq. (1).

$$weight_c = median(freq)/freq(c_n) \tag{1}$$

Here, c is a set of classes, $median(freq)$ is the median value of 9 $freq(c_n)$ values, and $freq(c_n)$ is defined by Eq. (2).

$$freq(c_n) = pixels_{c_n}/pixels_{ic} \tag{2}$$

$pixels_{c_n}$ is the number of pixels in the class c_n among all images, and $pixels_{ic}$ is the total number of pixels of all the images in which class c_n is found.

Table 2 Construction of dataset for training SegNet

Operations	No. of images	Remarks
Original	180	Frontal and side view facial images
Rotation	1440	45 to 315 degrees in 45-degree increments
Gaussian noise	720	Sigma = 0.05, 0.1
Poisson noise	2880	45 to 315 degrees in 45-degree increments
Rotation + Gaussian noise	180	45 to 315 degrees in 45-degree increments, Sigma = 0.05, 0.1
Rotation + Poisson noise	1440	45 to 315 degrees in 45-degree increments
TOTAL	6840	

Table 3 Evaluation result of facial region detection

Image index	No. of pixels	No. of inbound pixels	Accuracy
1	29,136	29,136	1
2	18,310	18,310	1
3	18,168	17,441	0.96
4	86,369	84,641	0.98
5	13,628	12,537	0.92
6	102,741	102,741	1
7	178,789	178,789	1
8	131,839	130,520	0.99
9	12,241	12,118	0.99
10	34,019	33,338	0.98
…	…	…	…
498	13,652	13,652	1
499	15,548	15,081	0.97
500	113,347	103,145	0.91
Average			0.97

4 Experiment

We evaluated the performance of our proposed scheme by comparing extracted facial landmarks with the constructed ground truth through diverse experiments. Experiments were conducted on an Intel® Core™ i5-7700 k CPU with 32 GB DDR4 memory, and an NVIDIA GeForce GTX 1080 Ti graphics card in a MATLAB 2017b environment. The mini-batch size was 12, and we performed a total of 17,000 iterations for training Faster R-CNN, and 50,000 iterations for training SegNet. In the first experiment, we evaluated the performance of Faster R-CNN. We then evaluated the accuracy of facial landmark extraction. In the last experiment, we applied our model to other face datasets, the results of which are reported in this paper.

4.1 Bounding box detection

In this experiment, we apply our Faster R-CNN model on various face images and check whether the bounding box contains a facial region or not. We evaluate our method by calculating how many pixels of face skin region are included in the bounding box. If the bounding box captures the facial region correctly, the accuracy should be 1. We used a total of 500 images for evaluation. The images were collected from our dataset for SegNet, Google image search, and other image datasets. Table 3 shows a part of entire evaluation, with an average accuracy of 0.97.

Figure 9 shows a few examples of facial region detection using our model, where facial regions were perfectly captured in the bounding box.

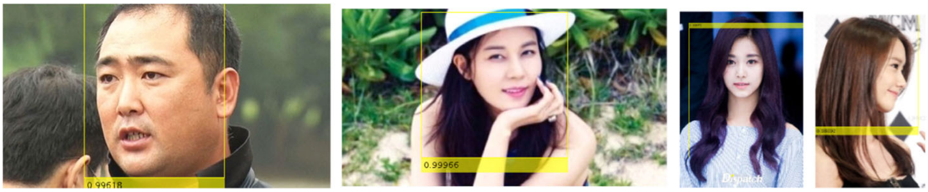
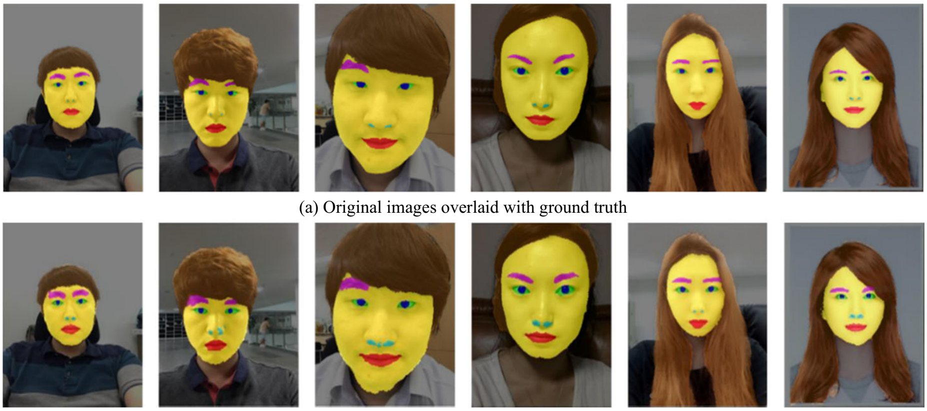


Fig. 9 Facial region detection results


(a) Original images overlaid with ground truth

(b) Original images overlaid with output by our method

Fig. 10 Facial landmark extraction result (Overlaid)

4.2 Facial landmark extraction

In this experiment, we extracted facial landmarks from 30 images and compared them with their ground truth by pixel units. Figure 10 shows the result. First, Fig. 10(a) shows the original images overlaid with the ground truth, and Fig. 10(b) shows the original images overlaid with the outcome using our scheme.

By comparing these figures, we can see that our proposed scheme detects facial landmarks quite well. To show the performance of our scheme quantitatively, we measure the detection accuracy of major landmarks, including hair, skin, and iris, which play an important role in personal color diagnosis and identification. Informally, the accuracy indicates how many pixels in the actual landmark are detected by our scheme. For hair and iris, the accuracy was higher than 95%, and for skin, the accuracy was higher than 90%, as shown in Table 4.

Detailed pixel-matching accuracies for all images and all classes are shown in Table 5. Each table entry indicates the probability that one landmark in the row is recognized as the landmark in the column. Hence, the accuracy on the diagonal indicates the correct recognition, and others indicate false recognition. For instance, an eyebrow was detected as eyebrow with an accuracy of 0.9437, and was not detected as hair or nose, as their entries are zero. From the table, we can see that all the landmarks, including major landmarks, are well detected with an accuracy higher than 90%. Interestingly, when a detection error occurs, the mistaken landmark is usually adjacent to the right landmark.

Table 4 Extraction accuracy of major landmarks

Person Landmark	#1	#2	#3	#4	#5	#6	AVG
Hair	0.964	0.9714	0.9829	0.964	0.9803	0.9862	0.964
Pupil	0.9787	0.9552	1	0.9706	0.9867	0.9273	0.9787
Face Skin	0.9091	0.9273	0.9535	0.9633	0.949	0.9379	0.9091

Table 5 Accuracy of pixel-matching for facial landmarks

Truth Recognized									
Background	0.9942	0.0034	0	0	0	0	0	0	0.0019
Hair	0.0122	0.9429	0	0.0002	0.0011	0	0	0	0.0436
Eyebrow	0	0	0.9437	0	0	0	0	0	0.0563
Sclera	0	0	0	0.9692	0.0308	0	0	0	0
Pupil	0	0	0	0.0290	0.9710	0	0	0	0
Nostril	0	0	0	0	0	0.9891	0	0	0.0109
Mouth	0	0	0	0	0	0	0.9286	0.0621	0.0093
Inner Mouth	0	0	0	0	0	0	0.0032	0.9852	0.0116
Face Skin	0.0016	0.0130	0.0130	0.0083	0.0016	0.0107	0.0094	0.0003	0.9942

For example, most of the detection errors for sclera are the iris, and the iris is adjacent to the sclera.

Overall, the experimental results show that although we use relatively small datasets for machine learning, we were able to achieve excellent facial landmark detection accuracy using pre-trained VGG16.

4.3 Additional experiment

In this experiment, we show that our scheme can achieve satisfactory levels of facial landmark extraction for other human face databases, such as the Helen dataset [13] and another dataset [5]. The extraction results for images in other face databases are shown in Fig. 11.

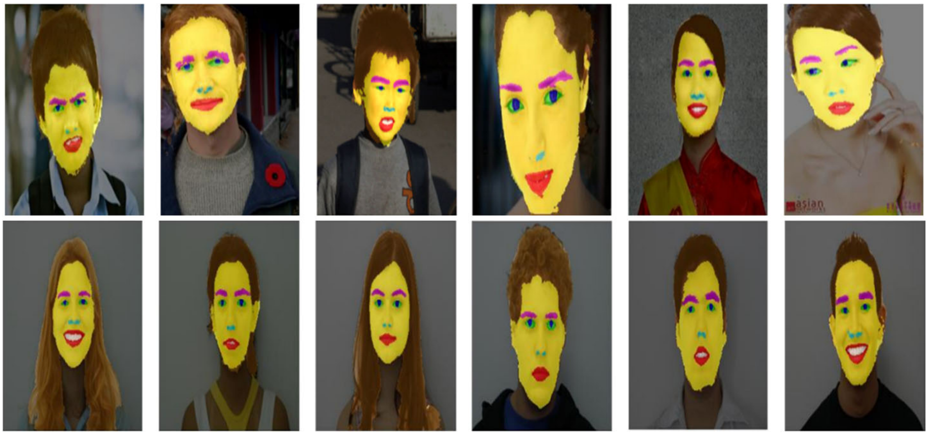


Fig. 11 Facial landmark extraction using other face dataset

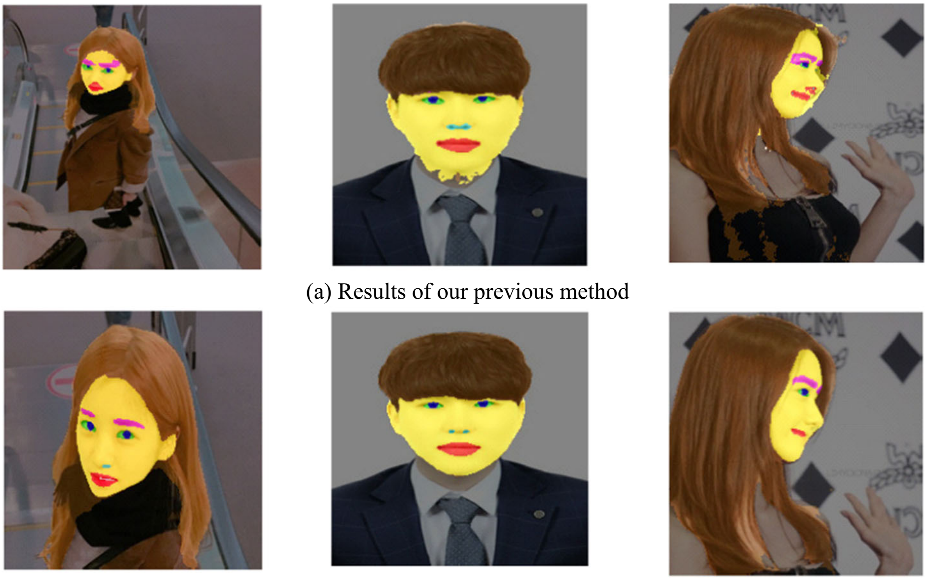
Table 6 Evaluation result of facial landmark detection on other face images

Facial landmarks	Perfect detection	Reasonable detection	Wrong detection	Missed detection	Total	$\frac{\text{Perfect+Reasonable}}{\text{Total}}$
Hair	367	133	0	0	500	1
Face skin	331	169	0	0	500	1
Sclera	838	114	2	0	954	0.99
Pupil	810	142	2	0	954	0.99
Eyebrow	667	238	3	27	935	0.96
Nostril	233	231	1	22	487	0.95
Lip	416	84	0	0	500	1
Between mouth	226	59	12	0	297	0.96

From the figure, you can see that the shape and position of most landmarks are well detected, even though there are some landmarks with partial distortion. We roughly evaluated how well our method could detect facial landmarks using the 500 face images. For evaluation, we classified the detection results into 4 types and calculated the portion of perfect and acceptable detections as shown in Table 6.

Even though most landmarks are detected well, landmarks such as skin, hair, and nostril showed a minor problem. This is due to the different colors and shapes of the skin and hair and can be improved by considering a more extensive training dataset.

So far, we have shown that our model works superbly for typical face images. As mentioned earlier, our goal in this paper was to improve the accuracy of facial landmark extraction for the exceptional cases where the detection quality was poor. In fact, we constructed our model to overcome these cases. The result of applying our model to the exceptional cases we mentioned earlier is shown in Fig. 12. In the figure, we can see that the facial landmark is well extracted when using our proposed method.


(a) Results of our previous method

(b) Results of our proposed method

Fig. 12 Improvement for some exceptional cases

5 Conclusion

In this paper, we proposed a way to improve our previously proposed facial landmark extraction scheme by using SegNet based on the convolution layer of VGG16, and to make it more robust even for the exceptional cases by using Faster R-CNN. We used Faster R-CNN for extracting facial regions in the image, and then cropped the region for training. SegNet detected the facial landmarks from a cropped face region. Also, we showed how to scale up the dataset using diverse operations such as noise filtering and image rotation. We performed several experiments to evaluate the performance of our model. First, we evaluated the facial region detection of our Faster R-CNN. Although the model is not deep, it detects facial regions precisely. Second, we showed that our proposed model could achieve good performance in terms of extraction accuracy, even for a relatively small dataset. If extra classes such as accessories and glasses were added to the current nine classes, or if images with various skin colors are used as a training set, facial landmark extraction could be more accurate and robust. Finally, we showed that our current model can process exceptional cases that our previous model could not process effectively. Because we can obtain the exact location and shape of the facial landmark, our method can be used effectively for diverse applications, such as personal color diagnosis and virtual makeup.

Acknowledgements This work was supported by Korea Environment Industry & Technology Institute (KEITI) through Public Technology Program based on Environmental Policy, funded by Korea Ministry of Environment (MOE)(2017000210001).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

References

1. Badrinarayanan V, Kendall A, Cipolla R (2015) Segnet: a deep convolutional encoder-decoder architecture for image segmentation. arXiv preprint arXiv:1511.00561
2. Badrinarayanan V, Handa A, Cipolla R (2015) Segnet: a deep convolutional encoder-decoder architecture for robust semantic pixel-wise labelling. arXiv preprint arXiv:1505.07293
3. Eigen D, Fergus R (2015) Predicting depth, surface normals, and semantic labels with a common multi-scale convolutional architecture. in ICCV, pp 2650–2658
4. Erjin Z et al (2013) Extensive facial landmark localization with coarse-to-fine convolutional network cascade. Comput Vis Workshops (ICCVW) 2013 IEEE Int Conf IEEE
5. Face datasets – http://ac.aua.am/Skhachat/Web/CS322/Face/FEI/. Accessed: 2017-11-03
6. Girshick R (2015) Fast r-cnn. arXiv preprint arXiv:1504.08083
7. Girshick R et al (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. Proc IEEE Conf Comput Vis Pattern Recognit
8. Güçlü U et al (2017) End-to-end semantic face segmentation with conditional random fields as convolutional, recurrent, and adversarial networks. arXiv preprint arXiv:1703.03305
9. Kasinski A, Schmidt A (2010) The architecture and performance of the face and eyes detection system based on the Haar cascade classifiers. Pattern Anal Applic 13(2):197–211
10. Kim H, Park J, Kim H, Hwang E (2018) Facial landmark extraction scheme based on semantic segmentation. 2018 International Conference on Platform Technology and Service (PlatCon-18), Jeju, Korea.01
11. King DE (2009) Dlib-ml: a machine learning toolkit. J Mach Learn Res pp 1755–1758
12. Krizhevsky et al (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Proces Syst
13. Le V, Brandt J, Lin Z, Bourdev LD, Huang TS (2012) Interactive facial feature localization. Interactive facial feature localization. Eur Conf Comput Vis pp 679–692

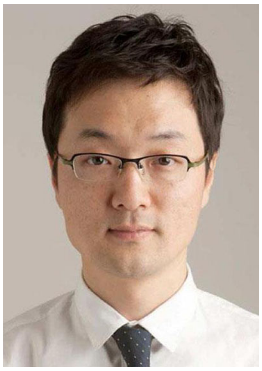
14. Long J, Shelhamer E, Darrell T (2015) Fully convolutional networks for semantic segmentation. Proc IEEE Conf Comput Vis Pattern Recognit pp 3431–3440
15. Noh H, Hong S, Han B (2015) Learning deconvolution networks for semantic segmentation. Proc IEEE Int Conf Comput Vis pp 1520–1528
16. Park J et al (2018) An automatic virtual makeup scheme based on personal color analysis. International Conference on Ubiquitous Information Management and Communication (IMCOM 2018), Langkawi, Malaysia. 01
17. Redmon J et al (2016) You only look once: unified, real-time object detection. Proc IEEE Conf Comput Vis Pattern Recognit
18. Ren et al (2015) Faster r-cnn: towards real-time object detection with region proposal networks. Adv Neural Inf Proces Syst
19. Ronneberger O, Fischer P, Brox T (2015) U-net: convolutional networks for biomedical image segmentation. Int Conf Med Image Comput Comput Assist Interv pp 234–241
20. Russakovsky O et al. (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis (IJCV) pp 1–42
21. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556
22. Yang J et al (2009) Linear spatial pyramid matching using sparse coding for image classification. Comput Vis Pattern Recognit CVPR 2009. IEEE Conference on. IEEE 2009

Hyungjoon Kim received the M.S. degree in School of Engineering from Korea University, Korea, in 2017. He is a Ph.D. student at the School of Electronic Engineering from Korea University, Seoul, South Korea. His current research interests image processing, deep learning, bio-informatics analysis

Jisoo Park received the B.S. degrees in multimedia engineering from Sungkyul University, Korea. She received M.S. degree from Korea University, Seoul, South Korea. Her current research interests image processing, database, text mining, beauty applications

Hyeonwoo Kim received the B.S. degrees in information and communication engineering from Hansung University, Korea, in 2017. Currently, he is a M.S. student at the School of Electronic Engineering, Korea University, Seoul, South Korea. His current research interests image processing, machine learning.

Eenjun Hwang received his B.S. and M.S. degrees in Computer Engineering from Seoul National University, Seoul, Korea, in 1988 and 1990, respectively; and his Ph.D. degree in Computer Science from the University of Maryland, College Park, in 1998. From September 1999 to August 2004, he was with the Graduate School of Information and Communication, Ajou University, Suwon, Korea. Currently, he is a member of the faculty at the School of Electrical Engineering, Korea University, Seoul, South Korea. His current research interests include database, multimedia systems, information retrieval, big data processing, and healthcare applications.



Seungmin Rho received his M.S. and Ph.D. degrees in Computer Science from Ajou University, South Korea, in 2003 and 2008, respectively. From 2008 to 2009, he was a postdoctoral research fellow at the Computer Music Laboratory of the School of Computer Science in Carnegie Mellon University. Currently, he is a member of the faculty at the Department of Media Software, Sungkyul University. His research interests include database, music retrieval, multimedia systems, machine learning, knowledge management

Affiliations

Hyungjoon Kim[1] • Jisoo Park[1] • HyeonWoo Kim[1] • Eenjun Hwang[1] • Seungmin Rho[2]

✉ Eenjun Hwang
ehwang04@korea.ac.kr

Hyungjoon Kim
hyungjun89@korea.ac.kr

Jisoo Park
jisoo_park@korea.ac.kr

HyeonWoo Kim
guihon12@korea.ac.kr

Seungmin Rho
smrho@sungkyul.edu

[1] School of Electrical Engineering, Korea University, Seoul, Republic of Korea

[2] Department of Media Software, Sungkyul University, Anyang, Republic of Korea