Detection of Facial Landmarks using Encoder-Decoder Network

Savina Jessica 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 landmark detection has achieved great performance by learning distinct features from the rich deformation of face shapes and poses. The facial landmarks such as eye centres, nose centres, jawlines, etc are localized to provide crucial information to various computer vision-related applications. The paper proposes an encoder-decoder network for detecting facial landmarks on the detected faces from digital images or video.

Index Terms—facial landmarks, convolutional neural networks, encoder-decoder

I. INTRODUCTION

Computer vision is a field that deals with how computers can be made to achieve high-level understanding from digital images or videos [1]. Particularly, the detection of fiducial points from digital images of faces under unconstrained conditions such as illumination, occlusion, head poses etc. These points can be used in various applications such as emotion recognition, 3D face reconstruction, identity verification and so on. In recent years, convolutional neural networks (CNNs) show high performance in solving tasks, which we, humans, are good at, such as natural language processing or image classification. In this paper, CNN is used to detect facial landmarks, which process face images to provide 68 fiducial points such as nose centre, eye centre, eye corners, jaw points and so on.

II. EXPERIMENT

A. Facial landmarks detection model

The detection of facial landmarks is done with a symmetric architecture called Unet [2]. The architecture can be divided into a contraction-expansive path or encoder-decoder path equivalently. The contraction path consists of repeated layers of two $3 \times 3$ convolutions where each convolution is followed by ReLU and batch normalization. A $2 \times 2$ max-pooling operation is applied to reduce the spatial dimensions. Furthermore, at each downsampling step, the number of feature channels is doubled while the spatial dimensions are halved. The contraction path is used to capture the context in the image. The expansive path consists of an upsampling of the feature map by $2 \times 2$ transpose convolutions, where the number of feature channels is halved. The corresponding feature map from the encoder is concatenated with $3 \times 3$ convolutions. At the final layer, a $1 \times 1$ convolutional is used to map the channels to the desired number of classes. The expansive path is used to enable precise localization using transposed convolutions. In the Unet architecture, the high-level architecture from the contraction path is combined and reused with the upsampled output from the expansive path. The number of landmarks to be predicted can depend on the different target task. Fig. 1 shows the facial landmarks detection model using Unet.

B. Implementation details

The facial landmarks model is trained with a 300W dataset [3]. The 300W dataset comprises AFW, HELEN, LFPW, XM2VTS, and IBUG datasets where images are annotated with 68 landmarks. The 300W dataset is divided into 3,148 training images and 689 testing images. The images in the dataset are resized to $112 \times 112$ resolution in grayscale. The Keras framework is used for model implementation and trained with a batch size of 32 and epochs of 100. The model is continuously optimized with the Adam optimization technique [4] with a learning rate of $10^{-3}$ and reduces the learning after every 20 epochs. For the model training, mean squared error (MSE), which is defined as the average of the square of all of the errors, is used between ground truth and predicted points.

C. Results

The Resnet-SSD face detector is used to detect faces from images or videos. The single-shot detector (SSD) [5] is faster than Faster R-CNN since it does not need an initial object proposals generation step. The model achieves a prediction accuracy of 55% with 3.6M parameters. The model localizes 68 facial landmarks as shown in fig. 2, with better alignment around the nose, mouth and eye region of the face. Since Unet enables precise localization of the landmarks, the points are detected with yaw, pitch and roll angles. The detected landmarks are not completely misaligned with extreme head poses. Fig. 3 shows the real-time results when the face is occluded at least 50%, though it is not the best alignment, the points are not completely lost, hence the points are aligned with the help of neighbouring landmark coordinates. The normalized mean error (NME) on the 300W dataset for the model achieves 11.87 on the full-set, 12.09 on the challenging subset and 11.24 on the common subset. The NME error gives a slight difference with all subsets of the 300W dataset. The challenging subset has extreme unconstrained conditions and hence can achieve acceptable value with Unet. The Unet,
Fig. 1. Facial landmarks detection model architecture with Unet [2].

which has an encoder-decoder structure captures the image context and specific localization.

Fig. 2. Facial landmarks detection in real-time.

III. CONCLUSION

This paper predicts facial landmarks with an encoder-decoder network called Unet. Unet architecture provides a contraction-expansive path which gives more context and accurate localization of the facial landmarks. Since the model suffers from occlusion and extreme head pose conditions, future work is focused on improving the model to make it robust to different imaging conditions.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2021-2020-0-01808) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation) and Technology Innovation Program (20003519, Development of Vehicle ICT convergence advanced driving assistance system and service for safe driving of longterm driving drivers) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea)
REFERENCES