Multi-task learning in facial analysis applications

Maryam Darei - The George Washington University

April and May 2022

Abstract

Facial landmark localization is one of the important intermediary steps for many facial recognition and analysis applications and it can be a challenging task in computer vision fields as its performance heavily depends on the resolution and size of the image as well as face positions. The subsequent steps such as facial expression recognition and head position estimation benefit from the information provided by facial feature points, the performance of a facial landmark localization is influenced by the face expression, head position and the image resolution therefore multi-task frameworks are used in variety of recent studies to allow sharing information between multiple tasks during learning process. In this review three papers [1,2,3] are chosen each of which uses a multi-task framework to achieve different goals in the facial analysis area for more than one task simultaneously. Facial landmark localization is a common intermediary step in all three.

1 Introduction

Typically, machine learning approaches are mainly focused on solving a single problem and while it achieves acceptable results; it ignores information from learning other related tasks that are beneficial for the task in hand. Finding a robust solution for face detection, face expression detection, head pose recognition and facial landmarking localization as a single task is not easy and the result of each of these can be fully or partially beneficial in the other's learning process. Hence Multi-task learning is introduced to allow multiple tasks that are related to each other to share information while training simultaneously.

In this review we describe three different multi-task learning frameworks in facial analysis applications and compare their application as well as their strengths and weaknesses.A brief description of each paper is discussed in section 2, in 3 the results of each paper is discussed and compared. Future work and conclusion can be found in section 4.

2 A brief description of the three papers

2.1 Residual multi-task learning for facial landmark localization and expression recognition

In this paper a new Residual multi-task learning is proposed for facial landmark localization and expression recognition simultaneously. The proposed architecture, named RMT-Net, consists of four different parts. The first layer, which is a front-end convolutional encoder that outputs a deep feature. The second layer outputs the enhanced features for the both tasks in hand and the third later consists of a regressor and a classifier, that generates facial landmarks and predicts the expression. The final layer is a learning module to enhance the overall results.

The Evaluation metrics used in this framework for facial landmark localization is Normal Mean Error (NME) , accuracy is used for evaluating facial expression recognition.

2.2 Super-FAN: Integrated facial landmark localization and super-resolution of real-world low-resolution faces in arbitrary poses with GANs

In this paper, an architecture called Super-FAN is proposed, which aims to increase the resolutions and alignment of the face at the same time by integrating a sub-network for facial landmark localization through thermal map regression in a network with GAN-based ultra-high resolution and losing a new thermal map. Also, they show the benefit of this approach. In addition, they also defined a residual-based architecture for super-resolution. In the last part they showed large

improvement over the state-of-the-art on both super resolution and face alignment and also the good visual results on real-world low resolution facial images taken from the WiderFace dataset.

2.3 MOS: A Low Latency and Lightweight Framework for Face Detection, Landmark Localization, and Head Pose Estimation

This article proposes a light-weight architecture that detects, locates, and landmarks a face and estimates the head pose simultaneously. It consists of a Feature Pyramid Network (FPN), Single Stage Headless (SSH) face detector module and a multi-task head (MTH). This article also proposes two more steps to the framework to improve the performance; an uncertainty Multi-task Loss (UML) and an online data-feedback augmentation. Multi-task head (MTH) used in this article is of a multi-task architecture (hard parameter sharing) [4]. In this architecture each head shares information through a cross stitch unit and combines features linearly that are used for all the tasks.

3 Experiments and result

3.1 RMT-Net Architecture

The RMT-Net method achieves the highest average test accuracy on AffectNet and RAF datasets in comparison with other methods. Also this method improves the performance of NME by nearly 10% and 17% compared to the state-of-the-art method on 300-W common and challenging sets, respectively. It achieves the very competitive NME results, which are better than the previous state-of-the-art by more than 11% on AFLW-Full. On the AFLW-Front testing set, the network's result is also better than state-of-the-art by more than 22%. For qualitative evaluation, it can be found that this method is able to handle the occlusion and gives more accurate landmarks. Also, based on the multi-task learning feature, it achieves the result of facial landmarks and the expression simultaneously.

3.2 Super-FAN Architecture:

There are 4 methods for super-resolution. The best results based on PSNR are achieved by Ours-pixel-feature-heatmap and the best result based on SSIM comes from Ours-pixel. But high quality and more detailed facial images are produced by Ours-pixel-feature-heatmap and Ours-Super-FAN. It should be mentioned that while the Ours-pixel achieved the good result based on SSIM but the produced image by that is blurry and unrealistic (need better metrics). For Facial landmark localization results, there are nine methods. The best method is named Super-FAN while FAN-Ours-pixel-feature-heatmap-GAN. For real word data result, 200 low resolution blurry images taken from the WiderFace as input data for Super-FAN. The result was not good enough and for improving the performance apply additionally random Gaussian blur (kernel size between 3 and 7 px) to the input images, and also perform some other preprocess on the input images. A few bad results happened because of extreme poses, large occlusions and heavy blurring.

3.3 MOS Architecture:

This paper compares the proposed approach and previous studies for face detection, landmark localizations and head pose estimation. Since our goal in this review is to compare landmark localizations accuracy of all three papers; landmark localization is only described here.

4 Conclusion and Future Work:

In the three reviewed papers, multi-task learning is used for landmark localization as well as other facial analysis tasks. The RMT-net approach tries to find the facial landmark while enhancing the results and predicting the facial expression [2] and the Super-FAN approach tries to enhance resolution of the input images while localizing the facial landmark using a multi-task learning framework [3] and performs well on low quality images. While these two approaches focus only on static face recognition (SFR), MOS architecture claims to be effective for dynamic face recognition (DFR) and real time facial analysis. Unfortunately, we can’t compare the performance of these three approaches because the Super-FAN paper uses accuracy as the metric to evaluate the facial landmarking while the other two papers use NME. Moreover, RMT-net and MOS framework architecture use different datasets for training.

RMT-Net can learn multitasking by learning to associate data sets with single-task labels. Extensive experiments on two labels databases and two expressions have shown its effectiveness. In this regard, new results have also been presented on face-landing positioning tasks at 300-W and AFLW, and superior performance in AffectNet and RAF face-mode detection. Second architecture is Super-FAN, claims to improve the state-of-the-art for both face super-resolution and alignment for

different facial poses and also it is shown good results on real-world low resolution facial images for the first time. In MOS architecture the results of face landmarking shows a better NME compared to RetinaFace and MTCNN however RetinaFace is proven to be faster. In Realtime applications, accuracy can be sacrificed to achieve acceptable results in less amount of time hence RetinaFace seems still to be the one that works better in real time applications. The speed of all these three approaches still need to be improved. Moreover, there is no evidence on how RMT-Net and the MOS architecture will perform on low quality, low resolution datasets.

References:

[1] Liu, Yepeng, et al. "MOS: A Low Latency and Lightweight Framework for Face Detection, Landmark Localization, and Head Pose Estimation." arXiv preprint arXiv:2110.10953 (2021).

[2] Chen, Boyu, et al. "Residual multi-task learning for facial landmark localization and expression recognition." Pattern Recognition 115 (2021): 107893.

[3] Bulat, Adrian, and Georgios Tzimiropoulos. "Super-fan: Integrated facial landmark localization and super-resolution of real-world low resolution faces in arbitrary poses with gans." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.

[4] Crawshaw, Michael. "Multi-task learning with deep neural networks: A survey." arXiv preprint arXiv:2009.09796 (2020).