3D Face Alignment in the wild

Shriya Kak[1] and Jilliam Diaz Barros[2]

Abstract. Face Alignment is the process in computer vision for identifying the geometric structure of human faces in digital images. It is an underlying task for many applications like face recognition, 3D face modeling and face de-identification. Recently, there have been considerable developments in 3D face alignment methods but the task of detecting 3D facial landmarks still remains a challenge in the wild(unconstrained conditions). In this paper, we have focused on extracting 68-3D facial landmarks in uncontrolled conditions using a convolutional-neural-network-based heatmap regression method. We are using the state-of-the-art method on 3D benchmark dataset and projection of 3D coordinates to train the existing model to make it more robust and increase its pixel-wise accuracy. Overall this study achieves comparable results with FAN, however, with regards to large pose face images, our model performs better.

Keywords: 3D face alignment, stacked hourglass network, Normalized mean error.

1 Introduction

The procedure of localizing a set of facial landmarks on a face image is called face alignment. Given the location and size of a face, it automatically determines the shape of the face components such as eyes, nose, lips. Localizing the facial landmarks in an unconstrained environment has been an active topic of research because of its potential applications for numerous vision tasks. It serves as a prerequisite for tasks such as facial behaviour analysis, face recognition, lip reading, 3D face construction, face de-identification to name a few. Moreover, localizing facial landmarks has been a daunting task due to challenges like illumination, pose and dramatic occlusion.

Ohlyan et al.[14] have broadly categorized the cause of the variation in facial appearance in two groups: Intrinsic factors and Extrinsic factors. Intrinsic Factors: are due to the physical nature of the face and it is independent of the observer e.g. facial expression, ageing etc. Extrinsic factors: are due to different conditions in which images are taken e.g. illumination, large pose, resolution, scale, noise etc. There have been several attempts to overcome and reduce the effect of these factors and develop a model which is robust to an unconstrained environment.

With the advent of deep learning there has been an increasing development in methods for localizing the facial landmarks. These methods can be divided into two categories 2D solutions and 3D solutions. According to a prior study[7], the 2D solutions treats the face as a 2D object and loses correspondence in any rotation in plane and for large poses. 3D face alignment has been proposed to overcome this problem of 2D face alignment. Convolution Neural Network(CNN) methods have been proved to be successful in different domains like human pose estimation, hand pose estimation etc. Recently, a CNN-based on heatmap regression has revolutionized human pose estimation producing significant results on very challenging datasets. We extend the proposed model [13] in the current work for the problem of face alignment.

In this paper, we have used a very strong baseline implemented in [4] by combining state-of-the-art architecture for landmark localization with the state-of-the-art(SOTA) residual block named as stacked hourglass network. The objective of this work is to localize the 3D points which are highlighted in green in fig 1 in unconstrained settings. Our work is broken down in sections as follows : 1) The Dataset 2) Prior study 3) Network architecture 4)Evaluation Metric 5) Experiment 6) Results. Our contribution is to improve the accuracy of the existing model and track the facial landmarks in the wild and increase in the pixel-wise accuracy.

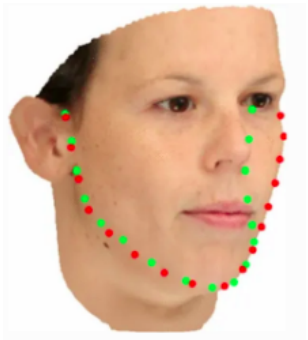

Fig. 1. Landmark annotation on face contour between 2D and 3D views. Red annotation is from 2D view, and green annotation is from 3D view (Color figure online). Source: [7]

2 Datasets

Datasets play a significant role in creating a robust model which would work in an unconstrained environment. Previously, the datasets available were collected under controlled conditions with neutral expression, frontal pose and normal lighting. The models which are trained on these dataset were not robust against rotation, pose and illumination changes. This inconsistency has been improved with the introduction of dataset created in uncontrolled environment. Below are some of the datasets which marked the beginning under unconstrained conditions and are used in this paper and also by the [4].

300-W: is currently the most widely-used in-the-wild dataset for 2D face alignment. The dataset itself is a concatenation of a series of smaller datasets: LFPW [2], HELEN [12], AFW [21] and iBUG [15], where each image was re-annotated in a consistent manner using the 68 2D landmark configuration of Multi-PIE [8]. The dataset contains in total 4,000 near frontal facial images. For training and validation of existing FAN network dataset 300W-LP-2D and 300W-LP-3D was used which is a synthetically generated dataset obtained by rendering the faces of 300-W into larger poses, ranging from -90°to +90°, using the profiling method of [22]. The dataset contains 61,225 images providing both 2D (300W-LP-2D) and 3D landmark annotations (300W-LP-3D). Since there are not enough 3D publicly available datasets we have trained our network using two types of dataset with 3D coordinates:

AFLW2000-3D(Annotated Facial Landmarks in the Wild): This dataset is prepared by [22], and contains 2, 000 images with yaw angles between ±90. This dataset is constructed by re-annotating the first 2000 images from AFLW [22] using 68 3D landmarks in a consistent manner.

Menpo: Menpo 3D benchmark[7], is composed of new datasets for multi-pose 2D and 3D facial landmark localisation and tracking. In contrast to the previous benchmarks such as 300W and 300VW[16], it contains facial images in both semi-frontal and profile pose. In Menpo, the landmark configuration is designed for both semi-frontal and profile faces based on the correspondence with a 3D face model, thus making face alignment not only full-pose but also corresponding to the real-world 3D space. Menpo Benchmark dataset [18] consists of 5658 semi-frontal and 1906 profile face images, which are selected from FDDB [9] and ALFW [11]. These annotated face images are collected from completely unconstrained conditions, which exhibit large variations in pose, expression, illumination etc. Fig 2 shows the ground truth in 3D Menpo dataset. In this dataset, 3DA-2D landmarks are defined as the 2D projections of the 3D landmarks on the image plane.

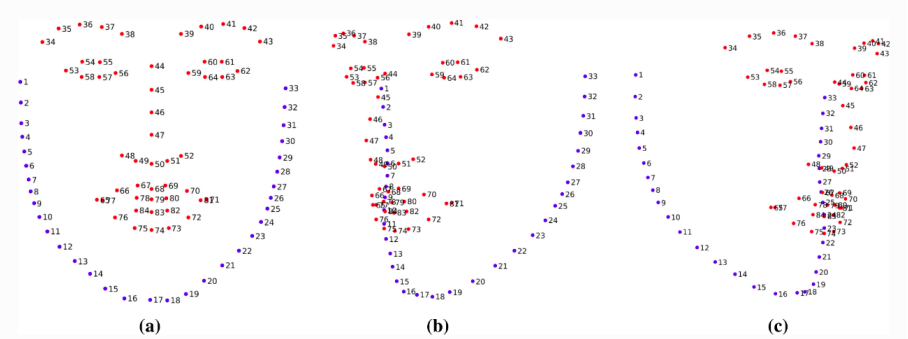

Fig. 2. Configuration for 3D and 3DA-2D landmarks, used in the Menpo 3D benchmark. The configuration includes 84 landmarks, is independent from the facial pose and corresponds directly to the 3D structure of the human face. (a) Frontal pose, (b) left yaw pose, (c) right yaw pose Source:[7]

3 Related Work

The 2D face alignment was performed using classical approaches such as Active Appearance Models (AAM) [6] and Constrained Local Model (CLM) [6] under controlled condition. Later, different regression based methods [5] have been proposed to directly estimate landmark locations from the discriminative features around landmarks. Most regression based algorithms do not consider the visibility of facial landmarks under different view angles. As a result, their performance can degrade substantially for input face images with large poses.

With the introduction of deep learning, many methods like CFAN [19], Casaded regression [17] were proposed to regress facial landmarks. Zhang et al. [20] combines face detection, face alignment, and other tasks into the training of CNN but alignment of faces with large pose variation is still remains a challenging problem, because each face might have a different set of visible landmarks. To overcome this problem and to increase pixel-wise accuracy 3D facial landmarks were extracted. The 3D face alignment has a strong advantage over 2D with respect to representational power and robustness to illumination and pose. So far there has been limited research on localizing 3D Face alignment in comparison to 2D Face alignment. Some of the recent method such as combination of 3D Morphable Model (3DMM) with CNN [10] which includes a set of well-controlled 2D face images associated 3D face scans were propose to handle self-occluded landmarks and large-pose landmark detection but were computationally complex. Due to the type and amount of training data, as well as the linear bases, the representation power of 3DMM can be limited. In this work, we use Face alignment network by [4] which serve as the basis for many SOTA networks.

4 Network Architecture

In the current work, we employ a 3D Face alignment network to extract 3D facial landmarks based on [4]. The architecture is made by stacking a four hourglass network which generates the heatmaps. In this hourglass network all the bottleneck blocks are replaced with hierarchical, parallel and multi-scale blocks of models which is very well mentioned in [13]. The concept of hourglass network is taken from [13] and fig. 4 shows the basic working hourglass network. The hourglass network pools down the image to a very low resolution, then up-samples and combines features across multiple resolutions.

Heatmaps have the advantage of providing higher output resolution, which helps in accurately localizing the keypoints. The Architecture is divided in 2 stages: one is localizing 3DA-2D(x,y) and

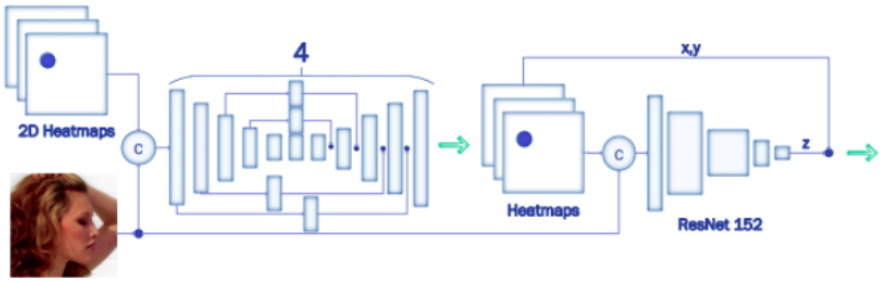

Fig. 3. Basic architecture of 3D FAN. Source:[4]

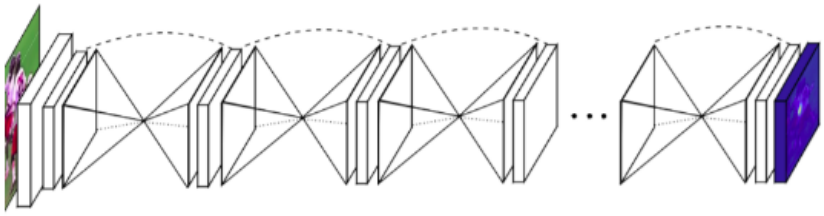

Fig. 4. Stacking of Hourglass Network. Source:[13]

then passing that output for calculating depth coordinates (z). The 2D points are extracted by 2D FAN and depth By ResNet [3] depth model. In this paper we'll call it 3D FAN. The hourglass model takes image as an input of size of 3*256*256 and generates heatmaps of size 68*64*64. Then the output of the first model is concatenated with the input image and given as an input to the ResNet model which is 152 layer deep. The input channel to ResNet model is 71 which includes 68 heatmaps for each landmark corresponding to the input image. The output shape of the ResNet model is 1*68 i.e. it predicts the depth(z) coordinate which is later concatenated with (x, y) to visualize 3D view.

5 Evaluation Metric

The traditional metric for face alignment is the point-to-point Euclidean distance normalized by the interocular distance(the distance between the two pupils when the visual axes are parallel). However, as mentioned in [21], this error metric is biased for profile faces for which the interocular distance can be very small. Similarly to the FAN [4], we have used Normalized Mean Error(NME) as shown in (1). The evaluation is done by calculating point-to-point euclidean distance normalized by bounding box size.

$$\text{NME} = \frac{1}{N}\sum_{k=1}^{N}\frac{\|\mathbf{x}_k - \mathbf{y}_k\|_2}{d} \tag{1}$$

In the equation, $\boldsymbol{x}$ and $\boldsymbol{y}$ denotes the ground truth landmarks and predicted landmarks for a given face respectively and $\boldsymbol{d}$ is the square root of the ground truth bounding box, computed as $\sqrt{w_{bbox} * h_{bbox}}$, were $\boldsymbol{w}$ and $\boldsymbol{h}$ are the width and height of the bounding box.

6 Experiments

6.1 Data pre-processing

The first step in data pre-processing in to find Region Of Interest(ROI). Extraction of ROI can be done either by using ground truth i.e. the given landmarks or by using a face detector. In the original paper [4], the data pre-processing and training method are not detailed. In our work , we have used Dlib face detector [1] to extract the bounding box. For the Menpo dataset we perform additional pre-processing step since all the images are of different sizes with corresponding landmarks in projected space and model space. We select 68 landmarks out of 84 landmarks for our experiments.

6.2 Training

We started training the model on the AFLW2000-3D dataset by extracting the bounding box from the face detector. If there was more than one face per image, instead of we selected the face which was closest to the given landmarks by calculating the center of the bounding box and the given landmarks. The model takes an input image of 3*256*256 as mentioned earlier. We used the 2D coordinates to generate the heatmaps which are used to calculate the loss. Mean Squared loss function was used with optimizer Adam. The model was trained with a learning rate of 10^{-6}.
Furthermore, we have also trained our model on the Menpo 3D dataset. Due to scarcity of the 3D dataset, we divided the dataset into 3 sets: train, validation and test. We have trained the model on around 7000 images and evaluated on 1500 images. For predicting 3D-2D points, we took the (x,y) coordinates from projected space and for depth prediction, we took the z coordinate from the model space in menpo dataset.

7 Results

We compared our trained model with the ground truth and the existing FAN model using NME metric mentioned in section 5. We have used the existing trained model of FAN [4] for comparison. The steps performed in evaluation are : 1) Firstly, detecting the face in the image; 2)cropping it to 256*256; 3)Calculating the NME for both models(our model and FAN's model). Quantitative and qualitative results are provided below.

7.1 NME Graph for 3D-2D

As we can see from the fig. 5, Considering the average NME of our model is similar to that of 2D FAN. However, in case of large poses our model gives better qualitative result as can be seen in section 7.2.

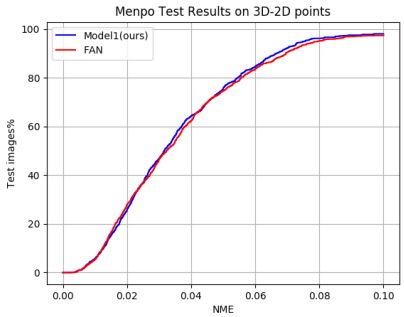

Fig. 5. NME on Menpo dataset(all 68 points used)

7.2 FAN vs Ours

We have evaluated 2D FAN and our model on Menpo dataset using NME metric. The NME for 2D FAN is 3.13 and for our model it is 2.97. There are some cases where we noticed a performance drop in FAN's model for large poses. Below are some of the result of our testing.

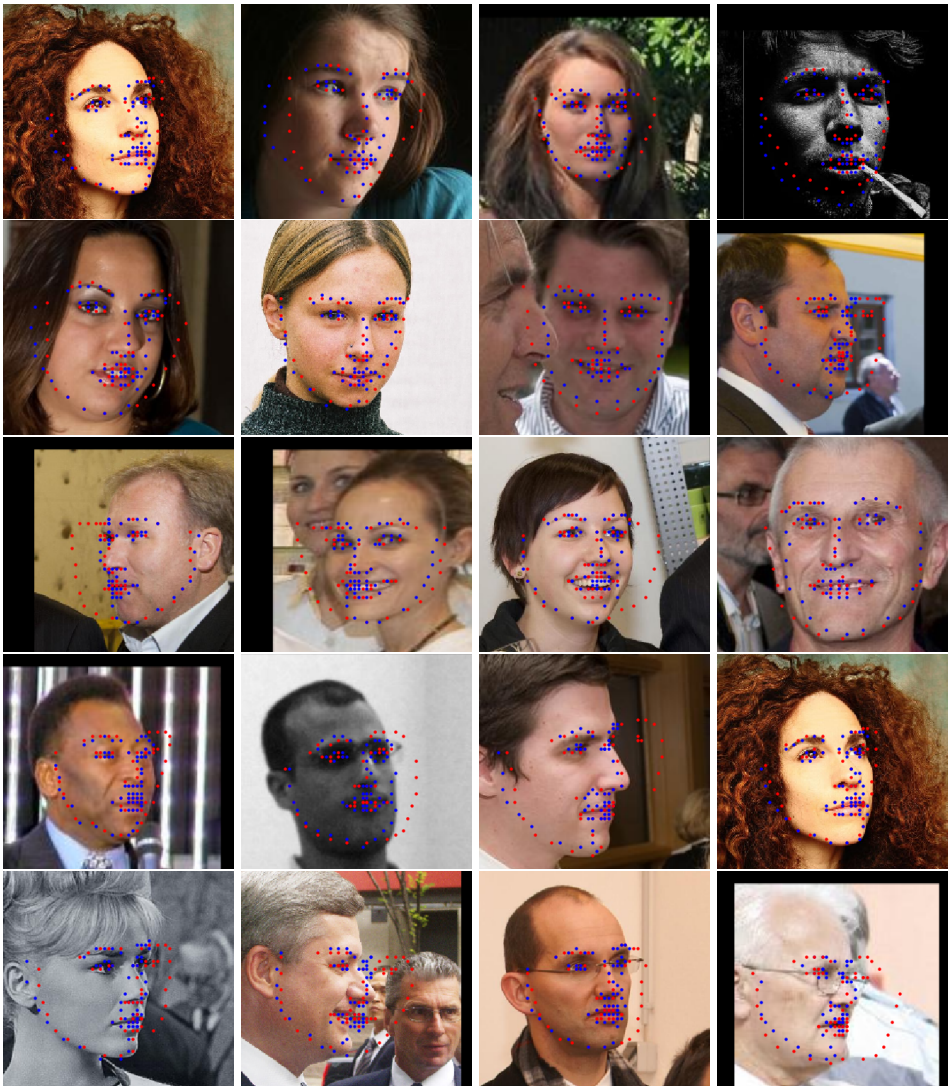

Fig. 6. Comparison between our trained model and existing FAN model on menpo dataset. Landmarks in blue are our prediction and in Red are FAN's prediction.

7.3 Ground Truth and Ours

We also compared our 2D model with groudtruth landmarks. According to fig. 7 we can observe that our prediction in blue is overlapping the ground truth in magenta in some images. Although it fails in some cases (see Section 7.6).

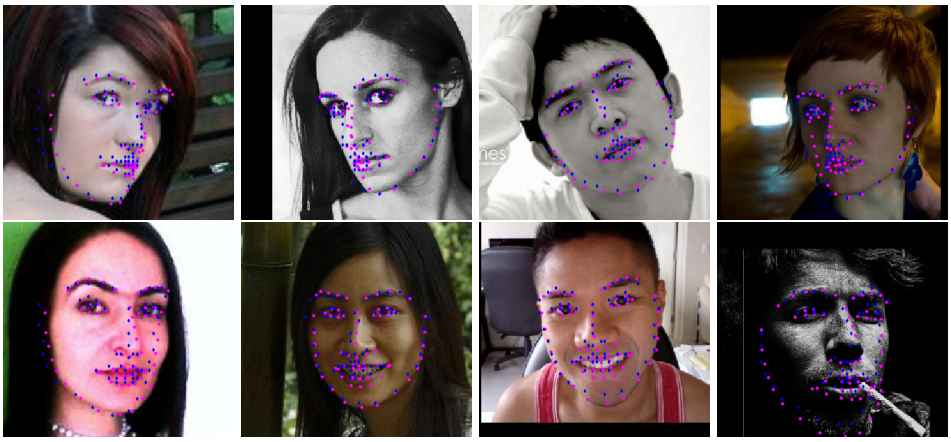

Fig. 7. Comparison shows how far we are from the ground truth on Menpo dataset. Landmarks in magenta are the ground truth and blue landmarks are our prediction

7.4 NME Graph for depth prediction

Figure 8 reports the numerical error of our trained 3D FAN on Menpo Dataset.

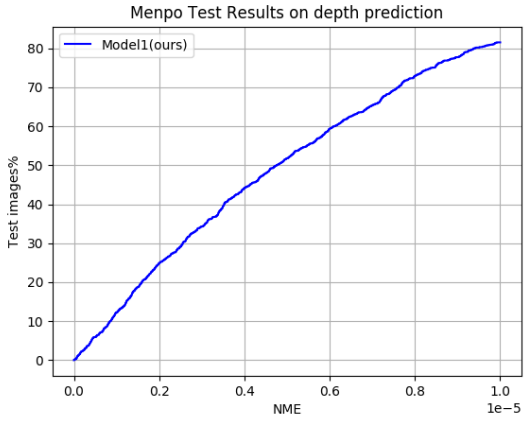

Fig. 8. NME on Menpo Dataset for depth prediction.

7.5 Depth prediction

Figure 9, shows the result of our ResNet Model. According to the results, we can say that we have successfully achieved depth prediction.

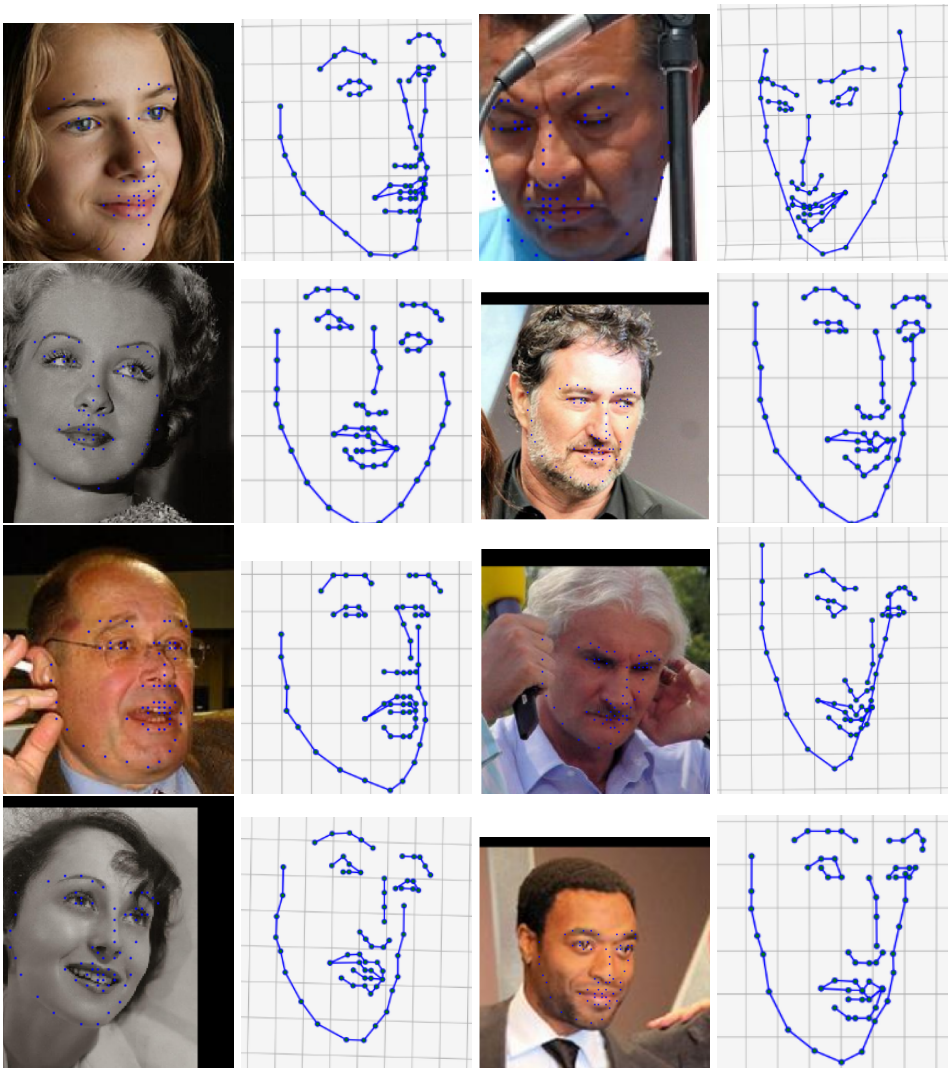

Fig. 9. Full 3D fitting examples on Menpo dataset.

7.6 Failure cases

While testing, we encountered some of the issues. Fig. 10 shows some of the failure cases which lead to the wrong prediction of landmarks. In first two examples faces were detected correctly but still predictions were inaccurate(Blue landmarks: our model prediction and red landmarks: 2D FAN model). Also, we noticed that face detector fails for some of the large pose face image e.g last 2 example in fig 10(Mangenta : ground-truth). To address this issue, we used groundtruth(landmarks) to create the bounding box.

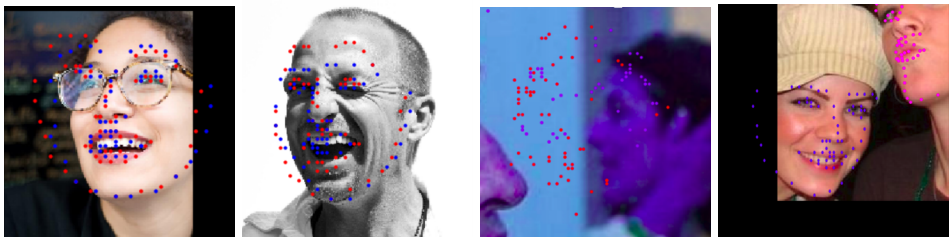

Fig. 10. Results using face detector on Menpo dataset.

7.7 Tracking by detection

We have also tested our model for tracking faces in the videos. For tracking, we used face detector on each frame and then performed 3D face alignment. Fig. 11 shows some of the results of the tracking by detection.

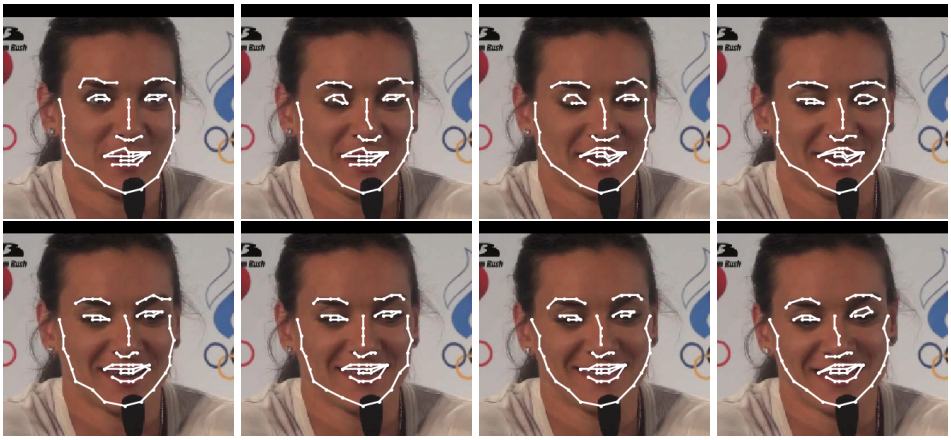

Fig. 11. Result of tracking on 3D Menpo Dataset.

8 Conclusion

We provide a improved model to estimate 2D and 3D landmarks than existing 2D FAN Network[4]. We discuss the shortcoming of existing FAN in section 7 such as large pose, self occlusion, appear to be solved by training the network on 3D Menpo benchmark dataset which exhibit large variations in pose, expression and illumination. This would allow the network to be more robust in realistic scenarios. We also tested our model for tracking with detection by applying face detector on each frame. For further research, we intend to use a temporal model to perform tracking by using developed 3D face alignment model.

References

1. https://dlib.net/face-landmark-detection.html. N.D.
2. Peter N Belhumeur, David W Jacobs, David J Kriegman, and Neeraj Kumar. Localizing parts of faces using a consensus of exemplars. *IEEE transactions on pattern analysis and machine intelligence*, 35(12):2930–2940, 2013.
3. Adrian Bulat and Georgios Tzimiropoulos. Binarized convolutional landmark localizers for human pose estimation and face alignment with limited resources. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3706–3714, 2017.
4. Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1021–1030, 2017.
5. Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proceedings of the IEEE international conference on computer vision*, pages 1513–1520, 2013.
6. David Cristinacce and Timothy F Cootes. Feature detection and tracking with constrained local models. In *Bmvc*, volume 1, page 3. Citeseer, 2006.
7. Jiankang Deng, Anastasios Roussos, Grigorios Chrysos, Evangelos Ververas, Irene Kotsia, Jie Shen, and Stefanos Zafeiriou. The menpo benchmark for multi-pose 2d and 3d facial landmark localisation and tracking. *International Journal of Computer Vision*, 127(6-7):599–624, 2019.
8. Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. *Image and Vision Computing*, 28(5):807–813, 2010.
9. Vidit Jain and Erik Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, UMass Amherst technical report, 2010.
10. Amin Jourabloo and Xiaoming Liu. Large-pose face alignment via cnn-based dense 3d model fitting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4188–4196, 2016.
11. Martin Koestinger, Paul Wohlhart, Peter M Roth, and Horst Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *2011 IEEE international conference on computer vision workshops (ICCV workshops)*, pages 2144–2151. IEEE, 2011.
12. Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S Huang. Interactive facial feature localization. In *European conference on computer vision*, pages 679–692. Springer, 2012.
13. Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016.
14. Sonia Ohlyan, Sunita Sangwan, and T Ahuja. A survey on various problems & challenges in face recognition. In *International Journal of Engineering Research & Technology (IJERT)*, volume 2, 2013.
15. Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. A semi-automatic methodology for facial landmark annotation. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 896–903, 2013.
16. Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Maja Pantic. The first facial landmark tracking in-the-wild challenge: Benchmark and results. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 50–58, 2015.
17. Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.
18. Stefanos Zafeiriou, George Trigeorgis, Grigorios Chrysos, Jiankang Deng, and Jie Shen. The menpo facial landmark localisation challenge: A step towards the solution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 170–179, 2017.
19. Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European conference on computer vision*, pages 1–16. Springer, 2014.
20. Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European conference on computer vision*, pages 94–108. Springer, 2014.
21. Xiangxin Zhu and Deva Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2879–2886. IEEE, 2012.
22. Xiangyu Zhu, Zhen Lei, Xiaoming Liu, Hailin Shi, and Stan Z Li. Face alignment across large poses: A 3d solution. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 146–155, 2016.