Joint Multi-view Face Alignment in the Wild

Jiankang Deng, Student Member, IEEE, George Trigeorgis, Yuxiang Zhou, and Stefanos Zafeiriou, Member, IEEE

Abstract—The de facto algorithm for facial landmark estimation involves running a face detector with a subsequent deformable model fitting on the bounding box. This encompasses two basic problems: i) the detection and deformable fitting steps are performed independently, while the detector might not provide best-suited initialisation for the fitting step, ii) the face appearance varies hugely across different poses, which makes the deformable face fitting very challenging and thus distinct models have to be used (e.g., one for profile and one for frontal faces). In this work, we propose the first, to the best of our knowledge, joint multi-view convolutional network to handle large pose variations across faces in-the-wild, and elegantly bridge face detection and facial landmark localisation tasks. Existing joint face detection and landmark localisation methods focus only on a very small set of landmarks. By contrast, our method can detect and align a large number of landmarks for semi-frontal (68 landmarks) and profile (39 landmarks) faces. We evaluate our model on a plethora of datasets including standard static image datasets such as IBUG, 300W, COFW, and the latest Menpo Benchmark for both semi-frontal and profile faces. Significant improvement over state-of-the-art methods on deformable face tracking is witnessed on 300VW benchmark. We also demonstrate state-of-the-art results for face detection on FDDB and MALF datasets.

Index Terms—Joint multi-view face alignment, Cascade face detection

I. INTRODUCTION

Object detection in computer vision has seen a huge amount of attention in recent years [1], [2], [3]. The advances in deep learning and the use of more elaborate models, such as Inception [4] and ResNet [5], have allowed for reliable and fine-scale non-rigid object detection even in challenging scenarios. Out of all the objects probably the most studied one is the human face. Face detection, although having embedded in our everyday lives through the use of digital cameras and social media, is still an extremely challenging problem as shown by the recent survey [6].

Human face in images captured in unconstrained conditions (also referred to as “in-the-wild”) is a challenging object, since facial appearance can change dramatically due to extreme pose, defocus, low resolution and occlusion. Face detection “in-the-wild” is still regarded as a challenging task. That is, considerable effort was needed in order to appropriately customise a generic object methodology, e.g. Deformable Part-Based Models [7] and Deep Convolutional Neural Networks (DCNNs) [1], in order to devise pipelines that achieve very good performance in face detection [8], [7], [9].

This work was partially funded by EPSRC project EP/N007743/1 (FACER2VM), as well as by the European Community Horizon 2020 [H2020/2014-2020] under grant agreement no. 689520 (TeSLA).

Fig. 1: Facial landmark response maps generated by Multi-view Hourglass Model (MHM). The profile and frontal faces are trained jointly, and the model is robust under large pose variations.

Specifically, when dealing with human face we are also interested in detailed face alignment, that is, localising a collection of facial landmarks on face images. This step plays an important role in many face analysis task, such as face recognition [10], [11], [12], expression recognition [13], [14], and face animation [15]. Due to the importance of the problem, a large number of facial landmark localisation methods have been proposed in the past two decades [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], and the previous works can be categorised as parametric fitting based [16], [17], [18], [27] and non-parametric regression based [19], [20], [21], [22], [23], [24], [26]. The former aims at minimising the discrepancy between the model appearance and the input image. The latter extracts features from the image and directly regresses to the ground truth landmarks. With the increasing number of training data [28], the performance of regression-based methods is generally better than that of parametric fitting based methods.

Recently, it was shown that it is advantageous to perform jointly face detection and facial landmark localisation [29], [9]. Nevertheless, due to the high cost of facial landmark localisation step, only few landmarks were detected [9]. Furthermore, in [9] the method made use of extra 400K facial images from the web which are not publicly available. To avoid this, we propose a coarse-to-fine joint multi-view landmark localisation architecture. In the coarse step, few landmarks are localised, while in the fine stage, we detect a large number of landmarks (e.g., 68/39). In our methodology, for reproducibility, we made use of publicly available data only.

Face alignment and tracking across medium poses, where all the landmarks are visible, has been well addressed [21], [22], [23]. However, face alignment across large poses is still a challenging problem with limited attention. There are two
main challenges: Firstly, there is a controversy on landmark definition, from 2D view or 3D view? As is shown in Figure 2, facial landmarks are always located at the visible face boundary in the 2D annotation. Faces which exhibit large facial poses are extremely challenging to annotate, because the landmarks on the invisible face side stack together. Since the invisible face contour needs to be always guessed to be consistent with 3D face models, labelling the self-occluded 3D landmarks is also ambiguous for annotators. Secondly, since occlusions can occur on both frontal and profile face images, designing a single shape constraint is hard for large pose face alignment. As view variation is continuous, view-specific modelling inevitably brings the problem of view classification and increases the computation cost.

In this work we present the first, to the best of our knowledge, method for deformable face modelling which jointly detects the face and localises a large amount of landmarks.

1) We employ a coarse-to-fine strategy where a face detector is first applied to find a coarse estimate of the facial shape using a small subset of landmarks. After removing the similarity transformation, a refining step is performed to estimate the dense facial shape of each person.

2) We formulate a novel Multi-view Hourglass Model (MHM) which tries to jointly estimate both semi-frontal and profile facial landmarks. Different from the other methods which employ distinct models, we try to capitalise on the correspondences between the profile and frontal facial shapes.

3) We demonstrate huge improvement over the state-of-the-art results in the latest benchmarks for deformable face fitting such as IBUG, 300W, COFW and the latest Menpo Benchmark. We demonstrate state-of-the-art results for the deformable face tracking on the 300VW benchmark and face detection on FDDB and MALF.

II. RELATED WORK

To better understand the problem of deformable face fitting, we review three of the major related elements.

Besides traditional models (such as AAMs [17], CLMs [18] and regression models [21], [32], [33], [34], [35], [36]), recently DCNNs has been employed in face alignment [24], [37], [38]. The resolution loss within the pooling step in DCNN was compensated by the image enlargement in a global to local way. Zhang et al. [39] adopted the similar coarse-to-fine framework with auto-encoder networks. Ranjan et al. [40] combined outputs of multi-resolution convolutional layers to predict the landmark locations. After the presentation of the fully-convolutional network (FCN) [38], which takes input of arbitrary size, produces a correspondingly-sized dense label map and shows convincing results for semantic image segmentation, direct landmark coordinated prediction changed to the landmark response map prediction. Lai et al. [41], Xiao et al. [42] and Bulat et al. [43] employed the convolutional and de-convolutional network to generate the response map for each facial landmark, and added a refinement step by utilising a network that performs regression. In the area of articulated human pose estimation, Alejandro et al. [44] proposed a novel stacked hourglass model, which repeated bottom-up and top-down processing in conjunction with intermediate supervision and obtained state-of-the-art result. Bulat et al. [45] further explored binarized Hourglass-like convolutional network for face alignment with limited resources.

Despite the large volume of work on semi-frontal face alignment, literature on the large-pose scenario is rather limited. This is attributed to the fact that large-pose face alignment is a very challenging task, until now there are not enough annotated facial images in arbitrary poses (especially with a large number of landmarks). A step towards this direction is the data presented in the new facial landmark competition [46]. The most common method in large-pose image alignment is the multi-view AAMs framework [30], which uses different landmark configurations for different views. However, since each view has to be tested, the computation cost of multi-view method is always high. In [47] the methods utilised the DPM framework to combine face detection and alignment, and the best view fitting was selected by the highest possibility. Since non-frontal faces are one type of occlusions, Wu et al. [48] proposed a unified robust cascade regression framework that can handle both images with severe occlusion and images with large head poses by iteratively predicting the landmark visible status and the landmark locations.

To solve the problem of large pose face alignment, 3D face fitting methodologies have been considered [49], [50], [25], which aims to fit a 3D morphable model (3DMM) [51] to a 2D image. [49] aligned faces of arbitrary poses with the assist of a sparse 3D point distribution model. The model parameter and projection matrix are estimated by the cascaded linear or nonlinear regressors. [50] extended [49] by fitting a dense 3D morphable model, employing the CNN regressor with 3D-enabled features, and estimating contour landmarks. [25] fitted a dense 3D face model to the image via CNN and synthesised large-scale training samples in profile views to solve the problem of data labelling. 3D face alignment methods model the 3D face shape with a linear subspace and achieve fitting by minimising the difference between image and model appearance. Although 3D alignment methods can cover arbitrary poses, the accuracy of alignment is bounded by the linear parametric 3D model, and the invisible landmarks are predicted after the visible appearance are fitted. In this paper, we focus on non-parametric visible landmark localisation.

Finally, we assess our methodology for facial landmark
tracking in 300VW [52]. The current state-of-the-art around face deformable tracking boils down to a pipeline which combines a generic face detection algorithm with a facial landmark localisation method [53]. Variants of this pipeline with different detectors or deformable models appear in the related paper [53]. The pipeline is quite robust since the probability of drifting is reduced due to the application of the face detector at each frame. We demonstrate that by applying the proposed methodology, large improvements over the state-of-the-art can be achieved.

III. OUR METHOD

In Figure 3, we shown the pipeline of the proposed coarse-to-fine joint multi-view deformable face fitting method. First, face proposals are generated by a small fully convolutional network on the image pyramid. Then, these face boxes are classified and regressed to predict the five facial landmarks. Afterwards, the similarity transformation between faces are removed using the five facial landmarks, and the response map for each landmark estimate is calculated by the joint multi-view hourglass model. Lastly, we make the final prediction of each landmark based on the corresponding response map.

A. Face Region Normalisation

The training of our face detection module follows the exact design of three cascade network and sampling strategies in [54]. In that, we minimise an objective function with the multi-task loss. For each face box i, its loss function is defined as:

\[ L = L_1(p_i, p^*_i) + \lambda_1 p^*_i L_2(t_i, t^*_i) + \lambda_2 p^*_i L_3(l_i, l^*_i), \]

where \( p_i \) is the probability of box i being a face; \( p^*_i \) is a binary indicator (1 for positive and 0 for negative examples); the classification loss \( L_1 \) is the softmax loss of two classes (face / non-face); \( t_i = \{t_{x_i}, t_{y_i}, t_{w_i}, t_{h_i}\} \) and \( t^*_i = \{t^*_{x_i}, t^*_{y_i}, t^*_{w_i}, t^*_{h_i}\} \) represent the coordinates of the predicted box and ground truth box correspondingly. \( l_i = \{l_{x_i}, l_{y_i}, \ldots, l_{x_n}, l_{y_n}\} \) and \( l^*_i = \{l^*_{x_i}, l^*_{y_i}, \ldots, l^*_{x_n}, l^*_{y_n}\} \) represent the predicted and ground truth five facial landmarks. The box and the landmark regression targets are normalised by the face size of the ground truth. We use \( L_2(t_i, t^*_i) = R(t_i - t^*_i) \) and \( L_3(l_i, l^*_i) = Rv^*(l_i - l^*_i) \) for the box and landmark regression loss, respectively, where \( R \) is the robust loss function (smooth-L1) defined in [2]. In Figure 4, we give the network structure of the third cascade network with multi-task loss.

One core idea of our method is to incorporate a spatial transformation [55] which is responsible for warping the original image into a canonical representation such that the later alignment task is simplified. Recent work (e.g., [56]) has explored this idea on face recognition and witnessed an improvement on the performance. In Figure 5, the five facial landmark localisation network (Figure 4) as the spatial transform layer is trained to map the original image to the parameters of a warping function (e.g., a similarity transform), such that the subsequent alignment network is evaluated on a translation, rotation and scale invariant face image, therefore, potentially reducing the trainable parameters as well as the difficulty in learning large pose variations. Since different training data are used in face region normalisation (CelebA [57] and AFLW [58]) and multi-view alignment (300W [28] and Menpo Benchmark [46]), end-to-end training of these two networks with intermediate supervision on the face region normalisation step is equal to step-wise training. In this paper, we employ step-wise cascade structure, and the face region normalisation step benefits from larger training data as annotation of the five facial landmarks is much easier than dense annotation.

B. Multi-view Hourglass Model

Hourglass [44] is designed based on Residual blocks [5], [59], which can be represented as follows:

\[ x_{n+1} = H(x_n) + F(x_n, W_n), \]

where \( x_n \) and \( x_{n+1} \) are the input and output of the n-th unit, and \( F \) is the stacked convolution, batch normalisation, and ReLU non-linearity. Hourglass is a symmetric top-down and bottom-up full convolutional network. The original signals are branched out before each down-sampling step and combined together before each up-sampling step to keep the resolution information. \( n \) scale Hourglass is able to extract features from the original scale to \( 1/2^n \) scale and there is no resolution loss in the whole network. The increasing depth of network design helps to increase contextual region, which incorporates global shape inference and increases robustness when local observation is blurred.

Based on the Hourglass model [44], we formulate the Multi-view Hourglass Model (MHH) which tries to jointly estimate both semi-frontal (68 landmarks) and profile (39 landmarks) face shapes. Unlike other methods which employ distinct models, we try to capitalise on the correspondences between the profile and frontal facial shapes. As shown in Figure 6, for each landmark on the profile face, the nearest landmark on the frontal face is regarded as its corresponding landmark in the union set, thus we can form the union landmark set with 68 landmarks (U-68). Considering that the landmark definition varies in frontal and profile data, we also enlarge the union set to 86 landmarks (U-86) by dissimilating two landmarks from eyebrow and seven landmarks from the lower part of face contour for profile annotation. During the training, we use the view status to select the corresponding response maps for the loss computation.

\[ L = \frac{1}{N} \sum_{n=1}^{N} (v^*_{n,i} \sum_{i,j} \| m_n(i,j) - m^*_n(i,j) \|^2_2), \]

where \( m_n(i,j) \) and \( m^*_n(i,j) \) represent the estimated and the ground truth response maps at pixel location (i, j) for the n-th landmark correspondingly, and \( v_{ni} \in \{0,1\} \) is the indicator to select the corresponding response map to calculate the final loss. We can see from Figure 6 that the semi-frontal response maps (second and fourth examples in third row) benefit from the joint multi-view training, and the proposed method is robust and stable in a range of poses.

Based on the multi-view response maps, we extract shape-indexed patch (24 x 24) around each predicted landmark from...
Fig. 3: Proposed coarse-to-fine joint multi-view face alignment. Face regions are generated by the multi-scale proposal, then classified and regressed by the following network. Five facial landmarks are predicted to remove the similarity transformation of each face region. Multi-view Hourglass Model is trained to predict the response map for each landmark. The second and third rows show the normalised face regions and the corresponding response maps, respectively.

Fig. 4: The architecture of 3rd cascade network [54]. “Conv” means convolution, “MP” means max pooling, and N is the number of landmarks. The step size in convolution and pooling is 1 and 2 respectively.

Fig. 5: Face Region Normalisation. The five facial landmark localisation network acts as the spatial transform layer and the subsequent alignment network is evaluated on a translation, rotation and scale invariant face image, therefore, potentially reducing the trainable parameters as well as the difficulty in learning large pose variations.

the down-sampled face image (128 × 128). As shown in Figure 7, a small classification network is trained to classify face / non-face. This classifier is not only used to remove high score false positives for face detection, but also can be employed as a failure checker for deformable face tracking.

IV. EXPERIMENTS

A. Experiment Setting

1) Training Data: Face Detection Model: The face detection module before the multi-view face alignment step follows the cascaded network design and sampling strategies as in [54]. We crop positive faces (IoU > 0.6), negative faces (IoU < 0.3) and part faces (IoU ∈ (0.4, 0.65)) from Wider Face [60] training set. To guarantee a high accuracy in predicting five facial landmarks, we employ additional labelled faces from the AFLW [58] dataset besides labelled faces from CelebA [57]. For the additional classifier after the multi-view alignment step, the positive (IoU > 0.5) and negative samples (IoU < 0.3) are generated from the previous cascaded face detector. Multi-view Hourglass Model: We train the face alignment module MHM on the 300W database [28], and the Menpo Benchmark database [46], where faces are manually annotated with either 68 (semi-frontal face) or 39 (profile face) landmarks. The training set of the 300W database (we denote as 300W-68) consists of the LFPW trainset [61], the Helen trainset [62] and the AFW dataset [8], hence, a total of 3148 images are available. The Menpo Benchmark database [46] (denoted as Menpo-39-68) consists of 5658 semi-frontal face images and 1906 profile face images. In this paper, we defined two training sets (300W-68-Menpo-39 and 300W-68-Menpo-39-68) for different evaluation purposes. 300W-68-Menpo-39 includes the 300W-68 data and the profile faces of Menpo-39, while 300W-68-Menpo-39-68 groups all the available training images in 300W-68 and Menpo-39-68.

2) Testing data: Face detection: We evaluate the performance of our face detection module in two challenging
datasets, FDDB and MALF. FDDB consists of 5171 faces in 2845 images from the unconstrained environment. MALF is a fine-grained evaluation dataset, in total, there are 5250 images with 11931 annotated faces. The “hard” subset contains faces (larger than 60 × 60) with huge variations in pose, expression, or occlusion. In particular, we give detailed pose-specific evaluations on MALF. **Face alignment in images & videos:** Evaluations of single face alignment and face tracking are performed in several *in-the-wild* databases. For alignment in static image, we test on *IBUG* dataset, 300W testset [28], *COFW* [63], [64], and *Menpo-test* [46]. All these databases are collected under fully unconstrained conditions and exhibit large variations in pose, expression, illumination, etc. In particular, *Menpo-test* [46] collects faces of all different poses, which are categorised into 5535 semi-frontal faces and 1946 profile faces based on [46]. For face tracking experiment, 300VVW is the only publicly available *in-the-wild* benchmark. It consists of 114 videos (about 218k frames in total), captured in the wild with large pose variations, severe occlusions and extreme illuminations.

3) **Evaluation Metric:** Given the ground truth, the landmark localisation performance can be evaluated by Normalised Mean Error (NME), and the normalisation is typically carried out with respect to face size.

\[
err = \frac{1}{M} \sum_{i=1}^{M} \frac{1}{N} \sum_{j=1}^{N} \frac{|p_{i,j} - g_{i,j}|^2}{d_i},
\]

where \(M\) is the number of images in the test set, \(N\) is the number of landmarks, \(p\) is the prediction, \(g\) is the ground truth, and \(d\) is the normalise distance. According to the protocol of difference facial alignment benchmarks, various normalise distances are used in this paper, such as eye centre distance [22], outer eye corner distance [28] and diagonal distance of ground truth bounding box [53]. The permissible error (localisation threshold) is taken as a percentage of the normalise distance.

4) **Training of Multi-view Hourglass Model:** The training of the proposed method follows a similar design as in the Hourglass Model [44]. Before the training, several pre-processing steps are undertaken. We firstly remove scale, rotation and translation differences by five facial landmarks among the training face images (referred as the spatial transformer step), then crop and resize the face regions to \(256 \times 256\). We augment the data with rotation (+/- 30 degrees), scaling (0.75-1.25), and translation (+/- 20 pixels) that would help simulate the variations from face detector and five landmark localisation. The full network starts with a \(7 \times 7\) convolutional layer with stride 2, followed by a residual module and a round of max pooling to bring the resolution down from 256 to 64, as it could save GPU memory while preserving alignment accuracy. The network is trained using Tensorflow [65] with an initial learning rate of 1e-4, batch size of 12, and learning steps of 100k. The Mean Squared Error (MSE) loss is applied to compare the predicted heatmaps to the ground-truth heatmaps. Each training step takes 1.2s on one NVIDIA GTX Titan X (Pascal) GPU card. During testing, face regions are cropped and resized to \(256 \times 256\), and it takes 12.21ms to generate the response maps.

### B. Ablation Study

We consider different training strategies and validate these setting on the challenging IBUG dataset in Table 1. (1) Hourglass Model (HM) trained on 300W-68. (2) HM trained on 300W-68, with spatial transformer step based on five facial landmarks. (3) HM trained on 300W-68 with simulated response maps from the output five landmarks. The input channel increases from 3 to 8, and this Hourglass model is trained with the spatial facial clue from face detector. The result of Method 3 is worse than that of Method 2, which indicates that the spatial transformer step for each face region is better than the spatial indication. (4) Multi-view Hourglass Model (MMH) trained on 300W-68-Menpo-39 with 68 union landmarks. (5) MMH trained on 300W-68-Menpo-39 with 86 union landmarks. (6) MMH trained on Menpo-39-68 with 68 union landmarks. (7) MMH trained on 300W-68-Menpo-39-68 with 68 union landmarks. (8) Two-stage Multi-view Hourglass with intermediate supervision. This model barely improves the performance but doubling the computation cost.

| Method | AUC   | FR (%) | NME (%) |
|--------|-------|--------|---------|
| 1      | 0.4470| 8.14   | 6.09    |
| 2      | 0.4737| 1.48   | 5.30    |
| 3      | 0.4629| 2.96   | 5.49    |
| 4      | 0.5076| 0.74   | 4.92    |
| 5      | 0.5141| 0.74   | 4.86    |
| 6      | 0.5226| 0.74   | 4.78    |
| 7      | 0.5324| 0.74   | 4.68    |
| 8      | 0.5409| 0.74   | 4.59    |

**Table 1:** Landmark localisation results on the IBUG dataset using 68 landmarks. Accuracy is reported as the Area Under the Curve (AUC) of the Cumulative Error Distribution curve, the Failure Rate (FR) at threshold 0.1, and out eye corner distance Normalised Mean Error (NME).

From the ablation experiments, we could conclude that by integrating the spatial transformer step, joint multi-view training and feeding more quality training data, the robustness and accuracy of proposed method improve hugely. As shown in Figure 8, although responses are more evident on facial organs than those on face contour, owing to more available profile training data, the proposed joint Multi-view Hourglass Model is able to deal with large pose variation.
C. Face Alignment on Images

We present experimental results on three face image databases, 300W database [28], COFW [63], [64] dataset and Menpo Benchmark [46]. The alignment method we evaluate here is the proposed Multi-view Hourglass Model (MHM), where the -Norm means the spatial transformer, and the - U-86 means the union 86 landmarks. Experiment results on 300W database are shown in Figure E where we compared the proposed methods with the best results in the 300W competition [28], such as Deng et al. [51] and Fan et al. [66]. Besides, we also compare with the state-of-the-art face alignment method “DenseReg + MDM” [67]. It is obvious that our model (Menpo-39-68-300W-68-U-68-Norm) outperforms those methods by a large margin. Table II reports the area under the curve (AUC) of the CED curves, as well as the failure rate for a maximum error of 0.1. Apart from the accuracy improvement shown by the AUC, we believe that the reported failure rate of 0.33% is remarkable and highlights the robustness of our MHM. Additionally, we found that the union landmark definition only has little influence on semifrontal face alignment accuracy. Thus we stick to the union 68 landmarks definition to avoid any confusion.

| Method                  | AUC     | FR (%) |
|-------------------------|---------|--------|
| Fan et al.              | 0.4802  | 14.83  |
| Deng et al.             | 0.4752  | 5.5    |
| DesenReg + MDM          | 0.5219  | 3.67   |
| Menpo-68                | 0.5485  | 1.00   |
| Menpo-68-Norm           | 0.5656  | 1.17   |
| Menpo-39-68-U-68-Norm   | 0.5973  | 0.17   |
| Menpo-39-68-U-86-Norm   | 0.5987  | 0.17   |
| Menpo-39-68-300W-U-68-Norm | 0.6071 | 0.33   |

TABLE II: Landmark localisation results on the 300W (indoor and outdoor) testing dataset using 68 landmarks. Accuracy is reported as the Area Under the Curve (AUC) and the Failure Rate of the Cumulative Error Distribution of the RMS point-to-point error normalised with out eye corner distance. “Norm” stands for the spatial transformer step from five facial landmarks, “U” stands for the union set number of profile and frontal data. We also present the performance of the MHM on the COFW [63], [64] dataset. Robust face alignment under occlusion and occluded landmark prediction are coupled problem that could be resolved simultaneously. Given the landmark occlusion status, local observation noise can be removed and the occluded landmark location can be predicted by shape context or constraint. Given a good fitting result, exploiting the fact that appearance of occluded region is quite different from the normal face appearance, even the simplest binary classifier could achieve excellent performance on occlusion classification. In Figure 10 we show the result of the proposed method comparing with state-of-the-art methods on COFW [63], such as HPM [68], SAPM [69], CFSS [23], TCDCN [70], and RCPR [63]. It can be clearly seen that even the baseline Hourglass model obtains a much better result because the bottom-up and top-down processing steps model the scale variations that would benefit the context inference. Moreover, by adding the spatial transformer, joint multi-view training and combined training data step-by-step, we gradually improved the alignment result, with the final success rate approaching 97.44%. Based on our best result, we employ the adaptive exemplar dictionary method [71] to predict occlusion status and refine the occluded landmarks. The normalised mean error decreases from 5.09% to 5.58%, and the occlusion prediction obtains a recall rate of 70.36% at the precision rate of 85.97%. In Figure 11 we give some fitting examples on COFW under heavy occlusions. To our surprise, responses of the occluded parts are still very clear and evident, which would prevent weird fitting results. This suggests that the proposed method captures and consolidates information across whole face images under different conditions, and incorporates local observation and global shape context in an implicit data-driven way, and thus improves the model’s robustness under occlusions.

![Fig. 9: Landmark localisation results on the 300W dataset. Accuracy is reported as Cumulative Error Distribution of RMS point-to-point error normalised with the eye corner distance.](image)

![Fig. 10: Landmark localisation results on the COFW dataset.](image)

In Figure 12 we also report the test results of our model on the Menpo Benchmark by comparing with the best three
D. Face Alignment on Videos

We employ the 300VW challenge testset for the challenging task of deformable face tracking on videos. Using our joint MHM method, We perform a frame-by-frame tracking on the video, and we initialise the next frame by the previous facial bounding box. The classifier based on the multi-view response maps is used as the failure checker during tracking. The face detector will be called if the fitting fails. The MHM takes 12.21 ms per face, and the classifier takes 2.32ms per face. The proposed multi-view face alignment and tracking method can run at about 50 FPS on the 300VW testset. We compare our method against the winners of the 300VW challenge: Yang et al. [75] and Xiao et al. [76]. Figure 15 reports the CED curves for all three video scenarios, and Table III reports the AUC and Failure Rate measures. The proposed MHM achieves the best performance, by a large margin compared to the winner of the 300VW competition (≥15% at RMSE = 0.02 in Scenario1&2, ≈10% at RMSE = 0.02 in Scenario3) as well as the best setting for CFSS method [23], [53] (≈15% at RMSE = 0.02 in Scenario1&2, ≈10% at RMSE = 0.02 in Scenario3), despite the fact that our approach is not fine-tuned on the training set of 300VW, while the rest of the methods were trained on video sequences and sometimes even with temporal modelling. Besides, our frame-by-frame tracking result is good enough that additional smoothing step (Kalman Filter) might be unnecessary.

In Figure 14 we select some frames from most challenging videos in Scenario3 and show their corresponding response maps for visualisation purpose. The response maps of proposed method is robust under large pose variation (yaw + pitch angles) and occlusion. In addition, response maps of invisible face parts are also reasonable, which indicates an implicit facial shape constraint within our method.

E. Face Detection

We evaluate the effectiveness of the multi-view response maps to remove high score false positives and obtain a state-of-the-art result on the FDB dataset. As in [9], we review the annotation of FDB, and add 67 unlabelled faces in FDB dataset to make sure all the false alarms are correct. We enlarge FDB images by 1.6, and the average resolution is about 639 × 604. We test the model on a single NVIDIA GTX Titan X (Pascal) GPU setting minimum face as 20. As shown in Table IV and Figure [16(a)], we observe the precision rate of recall within the high precision section (150 false positives, precision rate 97.1%). The baseline method refers to our re-implementation of MTCNN [53], due to adopting additional labelled faces from AFLW, our implementation is slightly better than the original MTCNN. Our method th1 sets a higher thresholds (0.6, 0.7, 0.7, 0.7) for cascaded classifiers, while our method th2 employs a lower thresholds (0.5, 0.5, 0.3, 0.7). As can be seen from Table IV and Table V the setting of th2 is slightly better than th1, but increases the running time from 49.8 ms to 62.9ms per image. The proposed joint multi-view response maps contribute to removing high score false positives from previous cascade classifiers. At the precision rate of 99.9%, the proposed method improves

Fig. 11: Example results by MHM on COFW. Response maps on the occluded parts are still very clear and evident.

Fig. 12: Landmark localisation results on the Menpo Benchmark. Accuracy is reported as Cumulative Error Distribution of RMS point-to-point error normalised with the diagonal of the ground truth bounding box.
Fig. 13: Example landmark localisation results on the test set of the Menpo Benchmark.

(a) Menpo Semi-frontal set

(b) Menpo profile set
the recall from 65.1% to 84.5%. At the precision rate of 99%, the proposed method improves the recall from 89.9% to 90.5%. The result is obviously higher than HR-ER [78] and Conv3D [79], and comparable with the best academic face detectors, e.g. STN [9], Xiaomi [80], and DeepIR [81]. After investigating our false positives, we surprisingly find some tiny regions (shown in Figure 16(b)) that can hardly be removed by our method, since they have very similar appearance and structure of the face, and may only be resolved by context-based model.

We also submitted our face detection results to http://www.cbsr.ia.ac.cn/faceevaluation/ and obtained the true positive vs. false positive curve on MALF. In Figure 17 our submission is named “sub_v1” and the threshold setting is (0.5,0.5,0.3,0.7). We compared with the off-the-shelf face detectors including HeadHunter [7], ACF [82], DPM [7], JDA [29], and DenseBox [83]. The proposed method obtains the best performance on MALF compared to the best academic algorithms including cascade models (HeadHunter [7], ACF [82], JDA [29]), structure models (DPM, JDA) and the structure-constrained deep model (Densebox). We also outperform the big data driven commercial models such as the FacePP-v2 and Picasa algorithms. Compared to the state-
of-the-art method DenseBox, our joint multi-view response maps achieve a significantly better detection result in large pose data (yaw angle > 40 degrees). A similar improvement could also be observed on the “hard” subsets.

V. CONCLUSION

In this paper, we proposed a coarse-to-fine multi-view face alignment method where a face detector is used to estimate a coarse estimate of the facial shape using a small subset of landmarks and then after removing similarity transformations a refining subsequent step is performed that estimates the high-resolution facial shape of each person. We formulate a novel multi-view hourglass model which tries to jointly estimate both semi-frontal and profile facial landmarks, and the joint training model is stable and robust under continuous view variations. We demonstrate huge improvement over the state-of-the-art results in the latest benchmarks for face alignment such as 300W, COFW and the latest Menpo Benchmark. We also demonstrate state-of-the-art results for the deformable face tracking on the 300VW benchmark and face detection on Fddb and MALF datasets.

REFERENCES

[1] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in Computer Vision and Pattern Recognition, 2014, pp. 580–587.

[2] Ross Girshick, “Fast r-cnn,” in International Conference on Computer Vision, 2015, pp. 1440–1448.

[3] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun, “Faster r-cnn: Towards real-time object detection with region proposal networks,” in Advances in neural information processing systems, 2015, pp. 91–99.

[4] Sergey Ioffe and Christian Szegedy, “Batch normalization: Accelerating deep network training by reducing internal covariate shift,” in International Conference on Machine Learning, 2015, pp. 448–456.

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, “Deep residual learning for image recognition,” in Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[6] Stefanos Zaferiou, Cha Zhang, and Zhengyou Zhang, “A survey on face recognition with pose variations and misalignment via orthogonal procrustes regression,” in European Conference on Computer Vision, 2014, pp. 1701–1708.

[7] Don Li, Huiling Zhou, and Kin-Man Lam, “High-resolution face verification using pose-scale facial features,” IEEE transactions on image processing, vol. 24, no. 8, pp. 2317–2327, 2015.

[8] Ying Tai, Jian Yang, Yigong Zhang, Lei Luo, Jianjun Qian, and Yu Chen, “Face recognition with pose variations and misalignment via orthogonal procrustes regression,” IEEE Transactions on Image Processing, vol. 25, no. 6, pp. 2673–2683, 2016.
Fig. 17: Fine-grained evaluation on MALF dataset.
IEEE Transactions on Image Processing, vol. 24, no. 8, pp. 2393–2403, 2015.

[36] Qingshan Liu, Jiankang Deng, Jing Yang, Guanccan Liu, and Dacheng Tao, “Adaptive cascade regression model for robust face alignment,” IEEE Transactions on Image Processing, vol. 26, no. 2, pp. 797–807, 2017.

[37] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin, “Extensive facial landmark localization with coarse-to-fine convolutional network cascade,” in International Conference on Computer Vision Workshops, 2013, pp. 386–391.

[38] Zhujin Liang, Shengyong Ding, and Liang Lin, “Unconstrained facial landmark localization with backbone-branches fully-convolutional networks,” arXiv preprint arXiv:1507.03409, 2015.

[39] 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. Springer, 2014, pp. 1–16.

[40] Rajeev Ranjan, Swami Sankaranarayanan, Carlos D Castillo, and Rama Chellappa, “An all-in-one convolutional neural network for face analysis,” arXiv preprint arXiv:1611.00851, 2016.

[41] Hanjiang Lai, Shengtao Xiao, Yan Pan, Zhen Cui, Jiasheng Feng, Chunyan Xu, Jian Yin, and Shuicheng Yan, “Deep recurrent regression for facial landmark detection,” IEEE Transactions on Circuits and Systems for Video Technology, 2016.

[42] Shengtao Xiao, Jiasheng Feng, Junliang Xing, Hanjiang Lai, Shuicheng Yan, and Ashraf Alimi, “Robust facial landmark detection via recurrent attentive-refinement networks,” in European Conference on Computer Vision. Springer, 2016, pp. 57–72.

[43] Adrian Bulat and Georgios Tzimiropoulos, “Convolutional aggregation of local evidence for large pose face alignment,” 2016.

[44] Alejandro Newell, Kaiyu Yang, and Jia Deng, “Stacked hourglass networks for human pose estimation,” in European Conference on Computer Vision. Springer, 2016, pp. 483–499.

[45] Adrian Bulat and Georgios Tzimiropoulos, “Binaryized convolutional landmark localizers for human pose estimation and face alignment with limited resources,” arXiv preprint arXiv:1703.00862, 2017.

[46] Stefanos Zafeiriou, George Tzimiropoulos, Jiankang Deng, and Jie Shen, “The menpo facial landmark localisation challenge: A step towards the solution,” in Computer Vision and Pattern Recognition Workshop, 2017.

[47] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N Metaxas, “Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model,” in International Conference on Computer Vision, 2013, pp. 1944–1951.

[48] Yue Wu and Qiangu Jia, “Robust facial landmark detection under significant head poses and occlusion,” in International Conference on Computer Vision, 2015, pp. 3658–3666.

[49] Amin Jourabloo and Xiaoming Liu, “Pose-invariant 3d face alignment,” in International Conference on Computer Vision, 2015, pp. 3694–3702.

[50] Amin Jourabloo and Xiaoming Liu, “Large-pose face alignment via cnns: A dense model and model fitting,” in Computer Vision and Pattern Recognition, 2016, pp. 4188–4196.

[51] Volker Blanz and Thomas Vetter, “Face recognition based on fitting a 3d morphable model,” IEEE Transactions on pattern analysis and machine intelligence, vol. 25, no. 9, pp. 1063–1074, 2003.

[52] Jie Shen, Stefanos Zafeiriou, Grigoris G Chrysos, Jean Kossaifi, Georgios Tzimiropoulos, and Majid Pantic, “The first facial landmark tracking in-the-wild challenge: Benchmark and results,” in Computer Vision Workshop (ICCVW), 2015 IEEE International Conference on, IEEE, 2015, pp. 1003–1011.

[53] Grigoris G Chrysos, Epameinondas Antonakos, Patrick Snape, Ackshay Asthana, and Stefanos Zafeiriou, “A comprehensive performance evaluation of deformable face tracking in-the-wild,” arXiv preprint arXiv:1603.06015, 2016.

[54] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao, “Joint face detection and alignment using multitask cascaded convolutional networks,” IEEE Signal Processing Letters, vol. 23, no. 10, pp. 1499–1503, 2016.

[55] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., “Spatial transformer networks,” in Advances in Neural Information Processing Systems, 2015, pp. 2017–2025.

[56] Oren Tadmor, Tal Rosenwain, Shai Shalev-Shwartz, Yonatan Weksler, and Amnon Shashua, “Learning a metric embedding for face recognition using the multibatch method,” in Advances In Neural Information Processing Systems, 2016, pp. 1388–1399.

[57] Ziyue Liu, Ping Lao, Xiaogang Wang, and Xiaoou Tang, “Deep learning face attributes in the wild,” in International Conference on Computer Vision, December 2015.
[81] Xudong Sun, Pengcheng Wu, and Steven CH Hoi, “Face detection using deep learning: An improved faster r-cnn approach,” arXiv preprint arXiv:1701.08289, 2017.
[82] B. Yang, J. Yan, Z. Lei, and S. Z. Li, “Aggregate channel features for multi-view face detection,” in International Joint Conference on Biometrics. IEEE, 2014, pp. 1–8.
[83] L. Huang, Y. Yang, Y. Deng, and Y. Yu, “Densebox: Unifying landmark localization with end to end object detection,” arXiv:1509.04874, 2015.