Face Alignment Assisted by Head Pose Estimation

Heng Yang[1]
heng.yang@cl.cam.ac.uk
Wenxuan Mou[2]
w.mou@qmul.ac.uk
Yichi Zhang[3]
yichizhang@fas.harvard.edu
Ioannis Patras[2]
i.patras@qmul.ac.uk
Hatice Gunes[2]
h.gunes@qmul.ac.uk
Peter Robinson[1]
peter.robinson@cl.cam.ac.uk

[1] Computer Laboratory
University of Cambridge
Cambridge, UK

[2] School of EECS
Queen Mary University of London
London, UK

[3] Faculty of Arts & Sciences
Harvard University
Cambridge, MA, US

Abstract

In this paper we present a supervised initialisation scheme for cascaded face alignment based on explicit head pose estimation. We first investigate the failure cases of most state of the art face alignment approaches and observe that these failures often share one common global property, i.e. the head pose variation is usually large. Inspired by this, we propose a deep convolutional network model for reliable and accurate head pose estimation. Instead of using a mean face shape, or randomly selected shapes for cascaded face alignment initialisation, we propose two schemes for generating initialisation: the first one relies on projecting a mean 3D face shape (represented by 3D facial landmarks) onto 2D image under the estimated head pose; the second one searches nearest neighbour shapes from a training set according to head pose distance. By doing so, the initialisation gets closer to the actual shape, which enhances the possibility of convergence and in turn improves the face alignment performance. We demonstrate the proposed method on the benchmark 300W dataset and show very competitive performance in both head pose estimation and face alignment.

1 Introduction

Both head pose estimation and face alignment have been well studied in recent years given their wide application in human computer interaction, avatar animation, and face recognition/verification. These two problems are interlaced and combining enables mutual benefits. Head pose estimation from 2D images remains a challenging problem because of the high diversity of face images [13, 18]. Recent methods [10] attempt to estimate the head pose by using depth data. , face alignment has made significant progress and several methods

Pages 130.1-130.13
DOI: https://dx.doi.org/10.5244/C.29.130

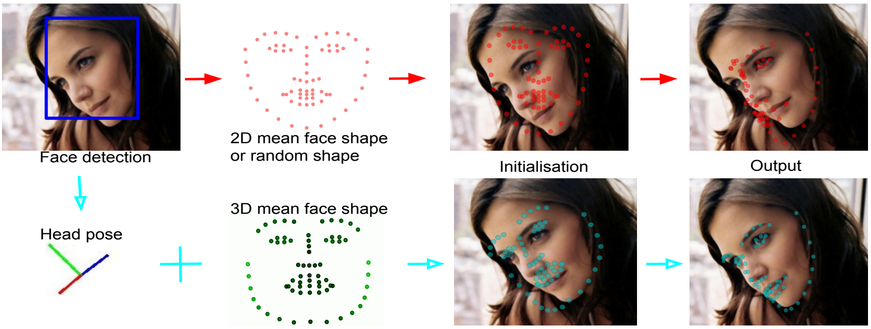

Figure 1: Our proposed head pose based cascaded face alignment procedure (path in cyan color) vs. conventional cascaded face alignment procedure (path in red color).

[2, 20, 30, 36] have reported good performance on images *in the wild*. However, they also show some failures. When we look into their failures cases, we find that those samples share one significant property, i.e., the head (face) in such images is usually rotated from frontal pose by big angles.

The best performing face alignment methods proposed in recent years ([30], [2] and [36]) also share a similar cascaded pose regression framework, i.e., face alignment starts from a raw shape (a vector representation of the landmark locations), and updates the shape in a coarse to fine manner. The methods in this framework are usually initialisation dependent. Therefore, the final output of one cascaded face alignment system might change if a different initialisation is provided to the same input image. Moreover, each model has a convergence radius, i.e., if the initialisation lies within the range of the actual shape, the model will be able to output a reasonable alignment result, otherwise it might lead the shape to a wrong location, as shown in Fig. 1. The methods like [2, 30] perform initialisation using a mean shape within the face bounding box or from a randomly selected shape from training set. There is no guarantee the initialisation lies within the convergence radius, especially when head pose variation is large.

In this paper, we aim to address the these problems and make cascaded face alignment perform better under large head pose variations. The difference between our proposed method and the conventional cascaded method procedure is illustrated in Fig. 1. By contrast to using mean shape or random shapes for initialisation by other methods, our proposed method aims to produce better initialisation schemes for cascaded face alignment based on explicit head pose estimation. This is motivated by two facts: 1) most current methods fail on face images with large head pose variation-as we will demonstrate later; 2) most recent face alignment methods work in a cascaded fashion and perform initialisation with mean shape. More specifically, we first estimate the head pose using a deep Convolutional Network (ConvNet) directly from face image. Given the estimated head pose, we propose two schemes of producing the initialisations. The first scheme projects a canonical 3D face shape under the estimated head pose to the detected face bounding box. The second scheme searches shape(s) for initialisation from the training set by nearest neighbour method in the head pose space. We build on our proposed scheme on the Robust Cascaded Pose Regression (RCPR) to demonstrate the effectiveness of supervised initialisation. We note that the proposed initialisation scheme can be naturally applied to any other cascaded face alignment. In summary, we make the following contributions:

- We investigate the failure cases of several state of the art face alignment approaches and find that the head pose variation is a common issue across those methods.

- Based on the above observation, we present a ConvNet framework for explicit head pose estimation. It is able to achieve an accuracy of 4° absolute mean error of head pose estimation for face images acquired in unconstrained environment.

- We present two initialisation schemes based on reliable head pose estimation. They enable face alignment method (RCPR) perform better and reduce large head pose failures by 50% when using only one initialisation.

To summarise, we propose better initialisation schemes based on explicit head pose estimation for cascaded face alignment, to improve the performance, especially in the case of large head pose variation.

2 Related Work

Face alignment has made considerable progress in the past years and a large number of methods have been proposed. There are two different sources of information typically used for face alignment: face appearance (i.e., texture of the face image) and the shape information. Based on how the spatial shape information is used, the methods are usually categorized into local-based methods and holistic-based methods. The methods in the former category usually rely on discriminative local detection and use explicit deformable shape models to regularize the local outputs while the methods in the latter category directly regress the shape (the representation of the facial landmarks) in a holistic way, i.e. the shape and appearance are modelled together.

2.1 Local-based methods

Local based methods usually consist of two parts. One is for local facial feature detection, which is also called local experts and the other is for spatial shape models. The former describes how image around each facial landmark looks like in terms of local intensity or color patterns while the latter describes how face shape, that is the relative location of the face parts, varies. This captures variations such as wide forehead, narrow eyes, long nose etc.

There are three types of local feature detection. (1) Classification methods include Support Vector Machine (SVM) classifier [4, 19] based on various image features such as Gabor [28], SIFT [15, 30], HOG [31] and multichannel correlation filter responses [11]. (2) Regression-based approaches are also widely used. For instance, Support Vector Regressors (SVRs) are used in [16] with a probabilistic MRF-based shape model and Continuous Conditional Neural Fields (CCNF) are used in [3]. (3) Voting-based approaches are also introduced in recent years, including regression forests based voting methods [6, 8, 32] and exemplar based voting methods [22, 23].

One typical shape model is the Constrained Local Model (CLM) [7]. The CLM steps can be summarised as follows: first, sample a region from the image around the current estimate and project it into a reference frame; second, for each point, generate a “response image" giving a cost for having the point at each pixel; third, searching for a combination of points which optimises the total cost, by manipulating the statistical shape model parameters. The methods built on CLM mainly differ from each other in terms of local experts, for instance CCNF in [3] and the Discriminative Response Map Fitting (DRMF) in [1]. There are many other local based methods either using CLM or other models such as RANSAC in [4], graph-matching in [38], Gaussian Newton Deformable Part Model (GNDPM) [26] and mixture of trees [39].

2.2 Holistic-based methods

Table 1: Holistic methods and their properties.

Methods	SDM [30]	RCPR [5]	IFA [2]	LBF [20]	CFAN [36]	TCDCN [37]
initialisation	mean pose	random	mean pose	mean pose	supervised	supervised
features	SIFT	pixel	HOG	pixel	auto-encoder	ConvNet feature
regressor	linear regression	random ferns	linear regression	random forests	linear regression	ConvNet

Holistic methods have gained high popularity in recent years and most of them work in a cascaded way like SDM [30] and RCPR [5]. We list very recent holistic methods as well as their properties in Table 1. The methods following the cascaded framework differ from each other mainly in three aspects. First, how to set up the initial shape; Second, how to calculate the shape-indexed features; Third, what type of regressor is applied at each iteration. For initialisation, there are mainly three strategies are proposed in literature: random, mean pose, and supervised. In order to make it less sensitive to initialisation, previous approaches such as [5, 29] propose to run multiple different initialisations and pick the median of all the predictions as the final output. Each initialisation is treated independently way until the output is calculated. However, such a strategy has several issues, first the theoretical support for selecting the median value is not well understood; second, there is no guidance on how to choose the multiple initialisations; third, using multiple initialisations is computationally expensive. A similar supervised initialisation scheme was proposed in [35] where the initialisation shapes were selected by using an additional regression forest model for sparse facial landmarks estimation. More recent work [33] has proposed a re-initialisation scheme based on mirrorability to improve the face alignment performance.

3 Data preparation

In this section we describe how the data is prepared in order to support our further discussion. More specifically, we discuss how we provide ground truth head pose and face bounding boxes from different face detectors for the benchmark dataset.

We use face image data from the benchmark face alignment in the wild dataset, 300W [21]. Since their testing samples are not publicly available, we follow the partition of recent methods [20] to set up the experiments. More specifically, we use face images from AFW [39], HELEN [25], LFPW [4] and iBug [21], which include 3148 training images and 689 test images in total. 3148 training images are from AFW (337 images), HELEN training set (2000 images) and LFPW training set (811 images), and 689 test images are from HELEN test set (330 images), LFPW test set (224 images) and iBug (135 images).

It is intractable to get the ground truth 3D head pose for face images collected in unconstrained conditions. In order to generate reasonable head pose (Pitch, Yaw and Roll) values, we use the pose estimator provided by Supervised Descent Method (SDM) [30]. Note that, when calculating the head pose, we feed the ground truth facial landmark locations instead of

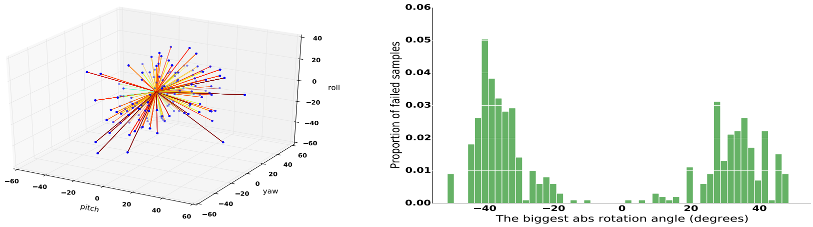

Figure 2: Distribution of the most erroneous samples.

using the detected landmarks. Technically, head pose is estimated by solving the projection function from an average 3D face model (49 3D points) to the input image, given the 3D to 2D correspondences. We also use the 3D head pose estimator provided by [1] for head pose calculation for evaluating the results. It produces very similar results to [30]. We calculate the head pose for all images in 300W.

The benchmark dataset only provides two types of face bounding boxes: one is the ground truth bounding box calculated as the tight box of the annotated facial landmarks; the other is the detection results from model of [39], which is quite similar to the ground truth face bounding box. However, several models like SDM [30] and RCPR [5] are trained with different face bounding boxes, thus their performance deteriorates significantly when using the provided face bounding boxes. We therefore provide different face bounding boxes to the test images by employing Viola-Jones detector [27] and HeadHunter detector [17] for fair comparison. For the input images on which the face detector fails we manually set reasonable bounding boxes.

4 Method

4.1 Motivation

We first run several state of the art methods, including 6 holistic based methods (SDM [30], IFA [2], LBF [20], CFAN [36], TCDCN [37], RCPR [5]) and 3 local based methods (GNDPM [26], DRMF [1], CCNF [3]) given their good performance and availability of source code. For each method, we provide the *best* type of face bounding boxes in order to get the best performance. For each method, we select 50 difficult samples out of the 689 test samples that provide the biggest sample-wise alignment error. Then we plot their head poses in Fig. 2 (left). As can be seen, most of the points are far away from the original point, i.e. they have big rotation angle(s). We further plot the histogram of the biggest absolute rotation angles of those samples in Fig. 2 (right). The biggest absolute rotation angle is calculated as the one of the three directions with the biggest absolute value. As can be seen, those samples are distributed at big absolute angles. There are very few samples that have small rotation angles. Based on this observation, we can conclude that, large head pose rotation is one of the main factors that make most of the current face alignments fail. Based on this fact, we develop a head pose based initialisation scheme for improving the performance of face alignment under large head pose variations.

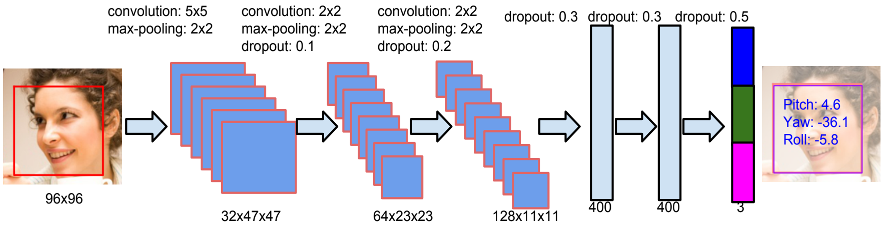

Figure 3: ConvNet model for head pose estimation.

4.2 Head Pose Estimation

Giving the training data from 300W with augmented head pose annotation, we train a convolutional network (ConvNet) [14] model for head pose estimation on the training set of 300W with 3148 images. The samples are augmented by 3 times with small permutations on the face bounding box. The ConvNet structure is shown is shown in Fig. 3. The input of the network is 96x96 gray-scale face image , normalised to the range between 0 and 1. The feature extraction stage contains three convolutional layers, three pooling layers, two fully connected layers and three drop-out layers. As we pose it as a regression problem, the output layer is 3x1 representing the head pose pitch, yaw and roll angle respectively. The angles are normalised between -1 and 1. We use Nesterov's Accelerated Gradient Descent (NAG) method [24] for parameter optimisation and we set the momentum to 0.9 and learning rate to 0.01. The training finishes in two hours on Tesla K40c GPU after around 1300 epochs, controlled by early-stop strategy. The learning curve is shown in Fig. 4 (left). The forward propagation of this network on GPU only takes 0.3ms per image on average.

4.3 Pose based Cascaded Face Alignment

4.3.1 General Cascaded Face Alignment

In order to make this work stand alone, we first summarise the general framework of cascaded face alignment. Face shape is often represented as a vector of landmark locations, i.e., $S = (\mathrm{x}_1, ..., \mathrm{x}_k, ..., \mathrm{x}_K) \in \mathbf{R}^{2K}$, where K is the number of landmarks. $\mathrm{x}_k \in \mathbf{R}^2$ is the 2D coordinates of the k-th landmark. Most of the current holistic-based method works in a coarse-to-fine fashion, i.e., shape estimation starts from an initial shape S^0 and progressively refines the shape by a cascade of T regressors, $R^{1...T}$. Each regressor refines the shape by producing an update, ΔS, which is added on the current shape estimate, that is,

$$S^t = S^{t-1} + \Delta S. \tag{1}$$

The update ΔS returned from the regressor that takes the previous pose estimation and the image feature I as inputs:

$$\Delta S = R^t(S^{t-1}, I) \tag{2}$$

An important aspect that differentiates this framework from the classic boosted approaches is the feature re-sampling process. More specifically, instead of using the fixed features, the input feature for regressor R^t is calculated relative to the current pose estimation. This is often called pose-indexed feature as in [9]. This introduces weak geometric invariance into the cascade process and shows good performance in practice. The CPR is summarized in Algorithm 1 [9].

Algorithm 1 Cascaded Pose Regression

```
Require: Image $I$, initial pose $S^0$
Ensure: Estimated pose $S^T$
for $t$=1 to $T$ do
    $f^t = h^t(I, S^{t-1})$                ▷ Shape-indexed features
    $\Delta S = R^t(f^t)$                  ▷ Apply regressor $R^t$
    $S^t = S^{t-1} + \Delta S$             ▷ update pose
end for
```

4.3.2 Head Pose based Cascaded Face Alignment

In section 4.2 we have presented how a ConvNet model can be used for head pose estimation. We propose two head pose based initialisation schemes for face alignment. One is based on an average 3D face shape projection and the other is based on nearest neighbour searching.

Scheme 1: 3D face shape based initialisation Given a 3D mean face shape, represented by 68 3D facial landmark locations, as shown in Fig. 1, we first project this shape under the estimated head pose to a set of canonical 2D locations. More specifically we use constant translation and focus length in order to get a reasonable projection for all images. Then we re-scale the canonical 2D projection by the face bounding box scale of the test image to get the initialisation. We can represent the initialisation process by function $\mathcal{F}$ as follows.

$$S_0 = \mathcal{F}(\theta, bb, \bar{S}^{3D}) \tag{3}$$

with bb the face bounding box, $\bar{S}^{3D}$, the 3D mean face shape, θ, the estimated head pose, which can be represented by:

$$\theta = \mathcal{G}(I, bb) \tag{4}$$

where $\mathcal{G}$ is the deep convolutional model described in section 4.2.

Scheme 2: Nearest Neighbour based initialisation We propose a second scheme for head pose based initialisation by nearest neighbour search. Since we have provided the training samples with head pose information as well, we can easily search samples that are with similar head pose of a test sample. Then we calculate similarity transformation between two face bounding boxes in order to calculate the initialisation shape for the test sample. In this way, we can also provide K initialisations by searching k-Nearest Neighbors from the training set.

Once we get a reliable initialisation (or several ones), we feed it to Algorithm 1 and apply the cascade of regressors in the same way to the baseline approach. In the case of the multiple initialisations, we calculate the output in a similar fashion to [5, 29], i.e., to pick up the median value of their estimations. We build our proposed head pose based initialisation schemes on top of the popular Cascaded Pose Regression (CPR) method due to its simplicity and popularity. We train its recent variant Robust Cascaded Pose Regression (RCPR) [5] model by using its new interpolated feature extraction, which is re-implemented by the author of [34]. We do not use its full version as occlusion status annotation is not available. We trained the baseline RCPR model on our 300W training set using Viola-Jones [27] face detection. 20 random initialisations are used for data augmentation at the training time.

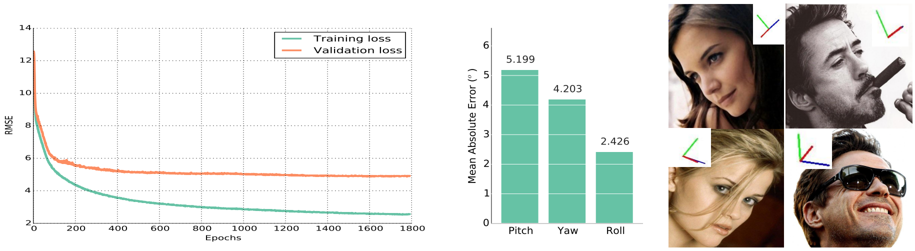

Figure 4: Head pose estimation result. Left, learning curve of head pose network, with y axis the Root Mean Square Error (RMSE) and x axis the number of epochs; middle, absolute mean error on test set; right, example results of head pose estimation.

5 Evaluation

5.1 Head Pose Estimation

We first evaluate the performance of head pose estimation. As we discussed before, it is very difficult to get the ground truth head pose for face images acquired in uncontrolled conditions. We calculate the pose based on the annotated facial landmark locations. We apply the trained deep ConvNet model on the test images of 300W and measure the performance. The result is shown in Fig. 4. The absolute mean errors of the head pose pitch, yaw, roll angles are 5.1°, 4.2° and 2.4°, respectively. Some example results are shown on the right. Despite the work by Zhu & Ramanan [39] is conceptually similar to our work in terms of simutaneous head pose and facial landmarks estimation, we do not compare to it here because their work can only estimate very sparse head pose yaw angles (e.g. -15°, 0° , 15 °).

5.2 Face Alignment

We first show the effectiveness of head pose based initialisation by comparing with the baseline strategy of the CPR framework [5, 29], i.e., generating random initialisations from training samples. The comparison is shown in Fig. 5. As can be seen on the left figure, by using one initialisation projected from 3D face shape, we obtain similar performance to the baseline approach with 5 initialisation shapes, and much better performance than that uses only one random initialisation shape. Similar superior performance is obtained by using nearest neighbour initialisation scheme, as shown on the right. By using more head pose based initialisations, we gain even better results, though the improvement is minor. It is worthy noting that by using our proposed initialisation scheme, we are able to decrease the number of failure cases (sample-wise average alignment error $>$ 0.1) from 130 to 69 (scheme 1) and from 130 to 72 (scheme 2), nearly 50%. Those samples are usually with large head pose variations and difficult for conventional face alignment methods. Moreover, by using one set of initialisation, the whole test procedure on one typical image takes 3.8 ms (0.3 ms for head pose estimation and 3.5 ms for cascaded face alignment).

We further compare the proposed method with recent state of the art methods including 5 holistic based methods (SDM [30], IFA [2], LBF [20], CFAN [36], TCDCN [37]) and 3 local based methods (GNDPM [26], DRMF [1], CCNF [3]). SDM and DRMF are trained using the Multi-PIE [12] dataset and detect 49 and 66 facial landmarks respectively. The rest of them are with models trained on 300W datasets. When we run their model on the test

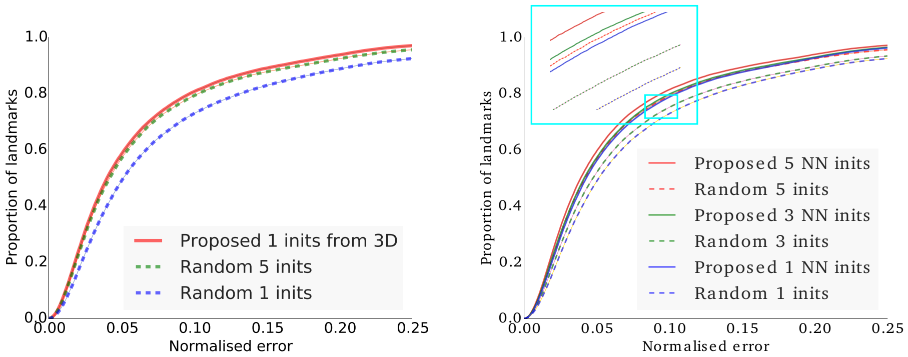

Figure 5: Our proposed head pose based initialisation scheme vs. random initialisation scheme. Left, our 3D face shape based scheme; right, our Nearest Neighbour (NN) based scheme.

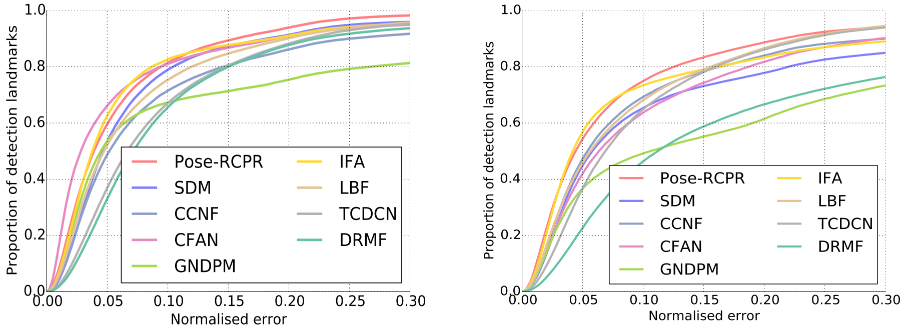

Figure 6: Comparison with recent methods. Left, results from the *best* face detection of each method; right, results from the common HeadHunter face detection. Pose-RCPR is our proposed method using only 1 initialisation from 3D.

images, we use the *best* bounding boxes for a fair comparison. Best bounding box refers to Viola-Jones detection for SDM and RCPR and tight face detection provided by 300w dataset for the rest of them. The comparison is shown in Fig. 6. As can be seen, our proposed method shows competitive performance. We also compare the performance on another type of common face detection, HeadHunter, given its best performance in face detection. The result is shown on the right of Fig. 6. We observe that the performance of most methods deteriorate significantly when testing on HeadHunter face bounding boxes. Our method provides most stable result, despite the fact that the HeadHunter face bounding box is more overlapped with the face detection from 300W (both are tight boxes of facial landmarks) than with Viola-Jones face detection. We believe this robustness to face bounding box changes is partially due to our head pose based initialisation strategy.

6 Conclusion and Future Work

In this paper we have observed that most recent face alignment methods show failure cases when large head pose variation is present. Based on the fact that cascaded face alignment is initialisation dependent, we proposed supervised initialisation schemes based on explicit head pose estimation. We use deep convolutional networks for head pose estimation and produce initialisation shape by either projecting a 3D face shape to the test image or searching nearest neighbour shapes from the training set. We demonstrated that using a more reliable initialisation is able to improve the face alignment performance with around 50% failure decreasing. It also shows comparable or better performance when comparing to other recent face alignment approaches.

Although we have managed to decrease the failure cases to a certain degree, we have not fully solved this problem. There are several interesting directions for future research. First, using head pose based initialisation shapes in the training stage may further boost the performance. Second, we only test our method on RCPR, we believe the proposed scheme can be naturally applied to other cascaded face alignment methods. It also raises several interesting questions. Do we need to make the cascaded learning model better for face alignment or to make the initialisation more reliable? Do we need more uniformly distributed data or a better model in order to make face alignment work better in wider range of head pose variations? We are going to investigate on these problem in our future research.

Acknowledgement

We gratefully acknowledge NVIDIA for the donation of the Tesla GPU used for this research.

References

[1] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3444–3451, 2013.

[2] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1859–1866, 2014.

[3] Tadas Baltrušaitis, Peter Robinson, and Louis-Philippe Morency. Continuous conditional neural fields for structured regression. In *Proc. Eur. Conf. Comput. Vis.*, pages 593–608. Springer, 2014.

[4] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 545–552, 2011.

[5] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1513–1520, 2013.

[6] T.F. Cootes, M. C.Lindner Ionita, and Sauer P. Robust and accurate shape model fitting using random forest regression voting. In *Proc. Eur. Conf. Comput. Vis.*, pages 278–291. Springer, 2012.

[7] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *Proc. Brit. Mach. Vis. Conf.*, volume 2, page 6, 2006.

[8] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2578–2585, 2012.

[9] P Dollár, P Welinder, and P Perona. Cascaded pose regression. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1078–1085, 2010.

[10] G Fanelli, J Gall, and L Van Gool. Real time head pose estimation with random regression forests. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 617–624, 2011.

[11] Hamed Kiani Galoogahi, Terence Sim, and Simon Lucey. Multi-channel correlation filters. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 3072–3079, 2013.

[12] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. Multi-pie. volume 28, pages 807–813. Elsevier, 2010.

[13] Murad Al Haj, Jordi Gonzalez, and Larry S Davis. On partial least squares in head pose estimation: How to simultaneously deal with misalignment. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pages 2602–2609. IEEE, 2012.

[14] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[15] David G Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004.

[16] B Martinez, M Valstar, X Binefa, and M Pantic. Local Evidence Aggregation for Regression Based Facial Point Detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, pages 1149–1163, 2012.

[17] Markus Mathias, Rodrigo Benenson, Marco Pedersoli, and Luc Van Gool. Face detection without bells and whistles. In *Proc. Eur. Conf. Comput. Vis.*, pages 720–735. Springer, 2014.

[18] Erik Murphy-Chutorian and Mohan M Trivedi. Head pose estimation in computer vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31 (4):607–626, 2009.

[19] V Rapp, T Senechal, K Bailly, and L Prevost. Multiple kernel learning svm and statistical validation for facial landmark detection. In *Proc. IEEE Int'l Conf. on Autom. Face Gesture Recognit.*, pages 265–271, 2011.

[20] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1685–1692, 2014.

[21] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, pages 397–403, 2013.

[22] Xiaohui Shen, Zhe Lin, Jonathan Brandt, and Ying Wu. Detecting and aligning faces by image retrieval. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3460–3467, 2013.

[23] Brandon M Smith, Jonathan Brandt, Zhe Lin, and Li Zhang. Nonparametric context modeling of local appearance for pose-and expression-robust facial landmark localization. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1741–1748, 2014.

[24] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *ICML*, pages 1139–1147, 2013.

[25] X Tan, F Song, Z H Zhou, and S Chen. Enhanced pictorial structures for precise eye localization under incontrolled conditions. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1621–1628, 2009.

[26] Georgios Tzimiropoulos and Maja Pantic. Gauss-newton deformable part models for face alignment in-the-wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1851–1858, 2014.

[27] Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages I–511, 2001.

[28] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using gabor feature based boosted classifiers. In *Proc. IEEE Int'l Conf. Systems, Man, and Cybernetics*, pages 1692–1698, 2005.

[29] Cao X., Y. Wei, F. Wen, and Jian Sun. Face alignment by explicit shape regression. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 177–190. Springer, 2012.

[30] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 532–539, 2013.

[31] Junjie Yan, Zhen Lei, Dong Yi, and Stan Z Li. Learn to combine multiple hypotheses for accurate face alignment. In *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, pages 392–396, 2013.

[32] Heng Yang and Ioannis Patras. Sieving regression forests votes for facial feature detection in the wild. In *Proc. Int'l Conf. Computer Vision*. IEEE, 2013.

[33] Heng Yang and Ioannis Patras. Mirror, mirror on the wall, tell me, is the error small? In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 4685–4693, 2015.

[34] Heng Yang, Changqing Zou, and Ioannis Patras. Face sketch landmarks localization in the wild. *IEEE Signal Processing Letters*, pages 1321 – 1325, 2014.

[35] Heng Yang, Xuming He, Xuhui Jia, and Ioannis Patras. Robust face alignment under occlusion via regional predictive power estimation. *IEEE Trans. Image Processing*, 2015.

[36] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *Proc. Eur. Conf. Comput. Vis.*, pages 1–16. Springer, 2014.

[37] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *Proc. Eur. Conf. Comput. Vis.*, pages 94–108. Springer, 2014.

[38] Feng Zhou, Jonathan Brandt, and Zhe Lin. Exemplar-based graph matching for robust facial landmark localization. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1025–1032, 2013.

[39] D. Zhu, X. Ramanan. Face detection, pose estimation and landmark localization in the wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2879–2886, 2012.