Reflective Regression of 2D-3D Face Shape Across Large Pose

Xuhui Jia[1]
xhjia@cs.hku.hk
Heng Yang[2]
yanghengnudt@gmail.com
Xiaolong Zhu[3]
lucienzhu@gmail.com
Zhanghui Kuang[4]
kuangzhanghui@sensetime.com
Yifeng Niu[2]
niuyifeng@nudt.edu.cn
Kwok-Ping Chan[1]
kpchan@cs.hku.hk

[1] The University of Hong Kong
[2] National University of Defense Technology
[3] Tencent Inc.
[4] Sensetime Inc.

Abstract

In this paper we present a novel reflective method to estimate 2D-3D face shape across large pose. We include the knowledge that a face is a 3D object into the learning pipeline, and formulate face alignment as a 3DMM fitting problem, where the camera projection matrix and 3D shape parameters are learned by an extended cascaded pose regression framework. In order to improve algorithm robustness in difficult poses, we introduce a reflective invariant metric for failure alert. We investigate the relation between reflective variance and face misalignment error, and find there is strong correlation between them. Consequently this finding is exploited to provide feedback to our algorithm. For the samples predicted as failure, we restart the algorithm with *better* initialisations based on explicit head pose estimation, which enhances the possibility of convergence. Extensive experiments on the challenging AFLW and AFW datasets demonstrate that our approach achieves superior performance over the state-of-the-art methods.

1 Introduction

Over the past decades, face alignment, a process of localising semantic facial landmarks such as eyebrows and nose tip, has been intensively studied as it is an essential prerequisite for many face analysis tasks, e.g., face animation [21], 3D face modelling [17] and face recognition [6]. Some previous works [23, 25, 40] have reported close-to-human performance on academic datasets such as 300W [31]. However, face alignment across large poses in real scenarios remains challenging and very limited attention and progress has been made.

There are mainly two inherent challenges associated with this problem: First, most existing face alignment algorithms attempt to learn a mapping from image appearance to land-

Pages 135.1-135.14
DOI: https://dx.doi.org/10.5244/C.30.135

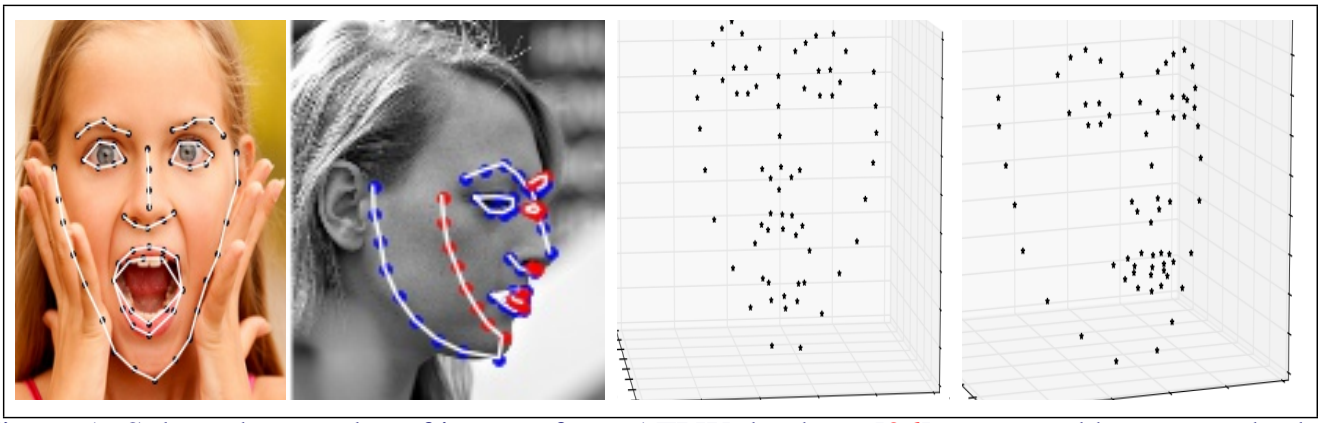

Figure 1: Selected examples of images from AFLW database [26] processed by our method. Left: The projection of landmarks to the 2D image plane. Right: estimated sparse 3D face shapes in world coordinate system. Red point denote self occlusion which can be directly computed from the fitted dense model by [18]

mark spatial distribution. Most of them are heavily initialisation dependent, from gradient descent based methods (e.g. Active Appearance Models [12] and Supervised Descent Method [40]) to more recent popular non-parametric Cascaded Pose Regression [16]. Hence for large-pose problem, feeding these algorithms with different initializations for the same image might lead to quite different output, because dramatic appearance variation (e.g., self occlusion) may cause the algorithm to be trapped in local optimum.

Second, spatial configurations of landmarks are highly different when faces deviate from frontal to profile, and the existing 2D shape models have difficulty in modeling 3D out-of-plane rotation. Although multi-view shape models [48] partially solve the pose variation problem, they cannot cover unlimited possibilities of view changes. Whereas introducing an additional 3D model is able to handle continuous pose views, traditional optimization approach achieved by fitting a 3D morphable model [6] is inefficient and it also assumes the 2D landmarks are provided manually or by a separate face alignment method.

In this paper, we aim to address the above discussed challenges and make Cascaded Pose Regression framework (CPR) perform better across large pose variations. Based on the fact that 2D face image is a projection of 3D face model, we parameterize the configuration of landmarks into 3D Morphable Model and the projection matrix, and regress them in a unified framework. First, two regressors are learned for each cascaded stage, one for predicting the update of camera projection matrix, and the other for predicting the update of 3D shape parameters. They work collaboratively to refine the predicted shape towards true shape; Second, to tackle failures which always occur in large-pose problem, we propose a novel reflective invariant metric to quantitatively estimate the alignments, subsequently the estimation will guide the model whether there is a need to restart the algorithm with different initialization. This is motivated by the fact that CPR are more sensitive to the horizontal reflection of an image, and the reflective variance are highly correlated to the misalignment error as we will demonstrate later; Third, instead of using mean shape or random shapes for initialisation [4, 24], we propose a head pose based initilisation scheme, which will relax failure alignments by explicitly incorporate a ConveNet head pose estimator. New initilisations can then be found by searching samples with similar head pose in the training set.

We summarize the main contribution of our works as: 1) Large pose face alignment by fitting a dense 3DMM; 2) Introduce a novel reflective invariant metric, by investigating

the relation between reflective variance and misalignment error; 3) A Reflective Cascaded Collaborative-Regressor algorithm that reduces large pose face alignment failures greatly.

2 Related work

There is a wide range of literature on 2D facial landmark localization. Based on the types of underlying models, we categorize them into parametric and non-parametric methods. The seminal work of both AAM [2, 12, 34] and CLM [3, 14] are classic parametric algorithms, which inspired many follow-on works. A fast AAM [36] was proposed for real-time face alignment, an ensemble AAM [11] was presented to align landmarks on a image sequence and a tree-structured model [50] was introduced to efficiently reformulate the CLM problem. Recently, non-parametric regression methods receive high interests due to its high accuracy evaluated on unconstrained face dataset and its fast performance. This include boosting regressions [13, 28, 37], regression trees [15, 20, 22, 25], linear regressions [4, 39, 40, 44], etc. Among them SDM [40] and CPR [39] are very popular, which work in a similar cascaded way but differ in several aspects such as how to select initialisations.

Despite the continuous improvement, large-pose face alignment remains very challenging and there are very few work reported in literature. Existing methods either exploit a multi-view framework that uses different landmark configurations for different views, or fitting a 3D model that attempt to achieve pose-invariant. TSPM [50] and CDM [46], for instance, combine face detection, pose estimation and face alignment in a DPM-like framework. However, since every view has to be evaluated, the computational cost is very high. On the other hand, fitting a parametric 3D face model can theoretically cover arbitrary face poses, and many works have done done in the context of pose-invariant face recognition. For 3D face alignment, Cao [10] jointly estimated a 3D face shape together with 2D landmarks for face animation, and Sergey et al.[35] directly regressed a 3D face shape from a single image with CPR. Their target was to advance performance on constrained images. Only recently, Liu et al. [24] aligned arbitrary face with the assistance of a 3D sparse point distribution model via a cascaded framework. However the performance is still far from mature when compared with frontal-view face alignment. Additional efforts are still in need for practical application, i.e., a failure-alarm mechanism, and a robust initialization scheme.

In this paper, we propose a reflective-regression of 2D-3D face shape model, which is related to works like failure alert [47] for failure prediction, face recognition score analysis [38] and meta recognition[32]. Our work differs from these works in two prominent aspects: 1) we focus on fine-grained semantic points localisation while they focused on instance level recognition or detection. 2) we do not train any additional models for prediction while all those methods rely on meta systems. In human perception, objects in horizontally reflected images will be perceived as the same object; computer vision methods should do the same in applications such as object recognition and scene classification. Inspired by this, [43] utilizes observed appearance symmetry to assess object part localisation. However, in large poses where up to half of the face can be occluded and there are will inevitably be no data in some parts of the image. To overcome these issues, we introduce the reflective invariant metric to estimate the prediction of 2D-3D face alignment. This valuable feedback greatly improves the robustness of our method to large pose variations as shown in experiments.

3 Methdology

3.1 3D Dense Face Modelling

Various 3D face models have been used in computer graphics and computer vision. In our proposed method, we exploit the 3D Morphable Model (3DMM) [6] to describe the dense 3D shape of an individual face:

$$\boldsymbol{S} = \boldsymbol{S}_0 + \sum_{i=1}^{N_{id}} w_{id}^i \boldsymbol{S}_{id}^i + \sum_{i=1}^{N_{exp}} w_{exp}^i \boldsymbol{S}_{exp}^i, \tag{1}$$

where $\boldsymbol{S} = [\boldsymbol{x}_1, ..., \boldsymbol{x}_k, ..., \boldsymbol{x}_K]$ is the 3D shape, $\boldsymbol{x}_k = [x_k, y_k, z_k]^T$ is a point in the world coordinate system, $\boldsymbol{S}_0$ is the average shape, $\boldsymbol{S}_{id}^i$ and $\boldsymbol{S}_{exp}^i$ are the ith identity and ith expression basis of the 3DMM, and w_{id}^i and w_{exp}^i represent the ith identity and ith expression parameter respectively. We built the identity basis $\boldsymbol{S}_{id}^i$ based on BFM [30], and the expression basis based on FaceWarehouse [8]. Therefore, any 3D face shape can be estimated by a collection of both identity and expression parameters $\boldsymbol{w} = [\boldsymbol{w}_{id}, \boldsymbol{w}_{exp}]^T$. With the assistance of 3DMM, we are able to have a rich representation of rigid and non-rigid face shape transformation. Subsequently, a 3D point $\boldsymbol{x}_k$ on the mesh can be transformed to image space once a projection parameter is given: $\mathbf{m} = [s, R_\alpha, R_\beta, R_\gamma, Q, t]^T$:

$$[u_k, v_k]^T = s * Q * R_\alpha * R_\beta * R_\gamma * \boldsymbol{x}_k + t, \tag{2}$$

the angles β and γ control the in-depth rotations around horizontal and vertical axis, α defines a rotation around the camera axis, s is the scaling factor, Q is the orthographic projection matrix $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$, $t = [t_x, t_y]$ is a spatial shit, and $\boldsymbol{u}_k = [u_k, v_k]^T$ defines the corresponding image-plane projection position. Similar to prior work[24], we model 2D face shape by a sparse set of facial landmark position (i.e., features on eyes, brows, nose, mouth and chew) $\mathbf{U} = [u_1, v_1; u_2, v_2; ..., u_N, v_N]^T$, which is a projection of 3D shape $\mathbf{S}^{(\mathbf{vt})}$. Note that, each landmark position in the image corresponds to a semantic vertex in the 3D facial shape, such that projection of the 3D vertex on the image should match the 2D landmark. We denote the correspondence by $\mathbf{vt} = [vt_1, vt_2, ..., vt_N]$. For convenience, we define $\prod_{\mathbf{m}}$ as the mapping function that transform a 3D shape in world coordinate to the 2D image plane, by taking projection parameter $\mathbf{m}$, 3D shape parameter $\mathbf{w}$, and vertex correspondence $\mathbf{vt}$:

$$\mathbf{U} = \prod_{\mathbf{m}}(\mathbf{S}_0, \mathbf{S}_{id}, \mathbf{S}_{exp}; \mathbf{w}, \mathbf{vt}). \tag{3}$$

Data Augmentation. To learn a 3D face shape predictor is difficult due to lacking of data of 2D image and 3D model pairs in unconstrained environment. While 2D annotation $\mathbf{U}$ is available for most face datasets, it is still hardly possible even for a human annotator to estimate the z-coordinate by observing just a single 2D image. In order to leverage the existing dataset, we first generate reflective images by taking horizontal flipping, and then propose a data augmentation method for 2D face image, based on the assumption that a 3D model can be accurately fitted given sufficient 2D landmarks [9, 51].

$$\mathbf{E(m,w)} = \sum_{i=1}^{N} ||\prod_{\mathbf{m}}(\mathbf{S}_0, \mathbf{S}_{id}, \mathbf{S}_{exp}; \mathbf{w}, vt_i) - \mathbf{u}_i||^2, \tag{4}$$

which minimise the difference between the projection of 3D landmarks in fitted model and 2D ground truth $\mathbf{U}$. This objective function is minimised using the coordinate-descent method, by alternately optimizing one parameter $\mathbf{m}$ or $\mathbf{w}$ while fixing the other in each iteration until convergence. Therefore 3D model can be constructed by estimating $\mathbf{m}$ and $\mathbf{w}$.

Cheek Landmark Marching. For each 2D landmark, there is a corresponding vertex index vt_i on the 3D model. While the indices for internal landmark are fixed, the contour landmarks will move along face silhouette when face deviate from the frontal view. Different from a sparse 3D landmark shape model in [24], our dense 3D model will adaptively adjust the 3D landmarks during the fitting process assisted by landmark marching method [51], so that the cheek landmarks consistently contribute to the prediction.

3.2 Reflective Invariant Metric

Intuitively, it is reasonable to expect that a salient location that is *interesting* in an image should also be *interesting* in the same image that is horizontally reflected. As human perception is invariant to horizontal reflection: they are equally able to locate and recognize object parts regardless whether they are looking at the original images, or the horizontally reflected image, as if looking at a mirror. However, algorithms in various computer vision tasks struggle to exhibit such consistency when an image is reflected. In a study [19], feeding a pair of reflected face images to the popular Microsoft's How-Old.net [1] surprisingly returned quite different age results though a consistent output is expected.

In our observation, such inconsistency also exist in large-pose 2D-3D face alignment. Furthermore, this inconsistency very likely indicates the misalignment of the original image. Inspired by this, we introduce a reflective invariant metric: to be reflective invariant, an algorithm must show that a set of keypoints found in its image are equivalent to those found in a horizontally reflected image. To enable quantitative analysis, a direct way is to project the predicted dense 3D model to 2D image plane, and for efficiency, which can further be measured by a sparse set of landmarks. Specifically, let $\mathbf{U}^p = \{\mathbf{u}_i^p\}_{i=1}^N$ denote the projection from predicted 3D of one image, and $\mathbf{U}^q = \{\mathbf{u}_i^q\}_{i=1}^N$ denote the projection from predicted 3D of its reflected image, such that sample-wise reflective error can be obtained:

$$\mathcal{E}_r = \frac{1}{N * f(I)} \sum_{i=1}^{N} ||\mathbf{u}_i^p - ref_{q\to p}(\mathbf{u}_i^q)||, \tag{5}$$

where $f(I)$ is the image-dependent normalisation factor, $ref_{q\to p}(\mathbf{u}_i^q)$ represents the function which transfers ith 2D point from one to another for mutually reflected images. Generally this process contains two part: 1) find a corresponding bilateral symmetric index i' with assistance of a look up table (e.g., a left eye becomes right eye in mutual reflected image); 2) apply horizontal coordinate flip, that is, $x_i' = w_I - x_i$, where w_I is the width of image. Given a pair of mutual reflected face images, should be no difference whether the image or its reflected version is considered to be original. Thus, it is safe to say there is misalignment alarm when the reflective error is significant. To be complete, given the ground truth $\mathbf{U}^{gt}$, the alignment error can be expressed as:

$$\mathcal{E}_a = \frac{1}{N * f(I)} \sum_{i=1}^{N} ||\mathbf{u}_i^p - \mathbf{u}_i^{gt}||, \tag{6}$$

Studying relationship between reflective error and misalignment error would help us to

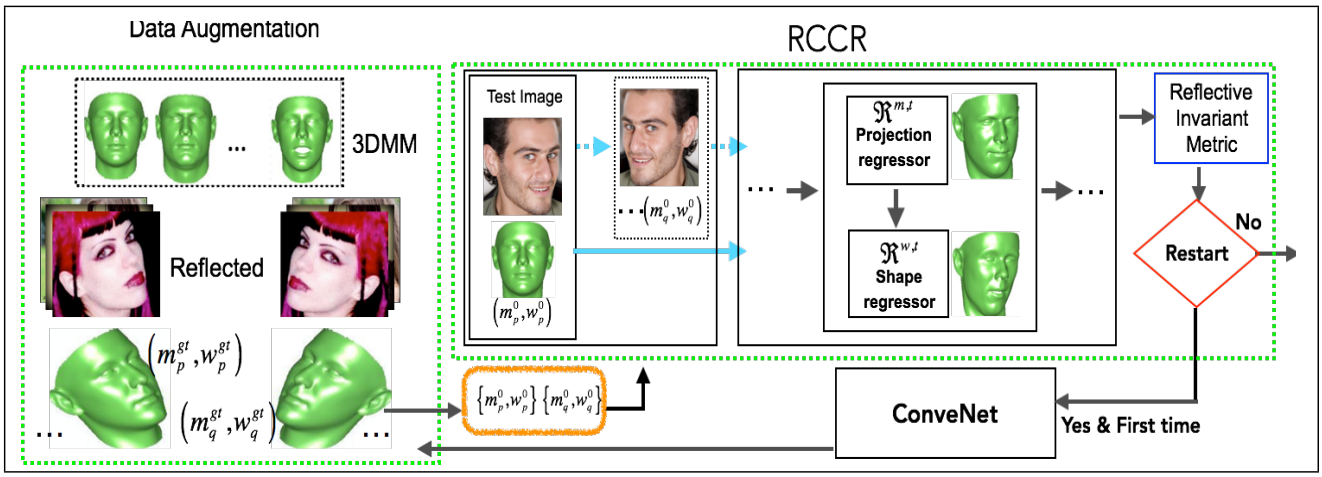

Figure 2: The proposed framework of our work

design a robust algorithm, especially for our large pose face alignment problem. In the next section, we show how to incorporate reflective feedback to our prediction process.

3.3 Reflective Cascaded Collaborative-Regressor

With a set of M training data $\{(I_i, \mathbf{m}_i, \mathbf{w}_i)\}$ augmented with estimated ground truth, we are interested in learning an efficient regression function that predict $\mathbf{m}$ and $\mathbf{w}$ from a single image. Therefore, the problem of 2D-3D face alignment is transformed to learning a projection parameter $\mathbf{m}$ and shape parameter $\mathbf{w}$ based on our representation. However, the challenges lie in: a) The appearance of unconstrained face image vary dramatically across large pose due to self occlusion (ranging from frontal view to profile view). Hence, complex input appearance demand a more advanced non-linear mapping function. b) Projection and shape parameters typically exhibit different amount of variation over large pose. Thus, regressing two parts together is very difficult and cause slow convergence and degraded accuracy.

Motivated by the success of cascaded pose regression framework (CPR)[33, 39], we propose a cascaded collaborative-regressor algorithm, which starts from an initialisation $\{\mathbf{m}^0, \mathbf{w}^0\}$ and progressively refines the parameters in an additive manner via a sequence of T regressors, $\mathcal{R}^{1...T}$, such that the estimated parameters gradually approximate the ground truth, where, the input feature built on current estimation for each regressor $\mathcal{R}^t$ provide geometric invariance. Specificially, there are two regressors to be learned at the cascaded tth layer for our algorithm, namely $\mathcal{R}^{m,t}(*)$ and $\mathcal{R}^{w,t}(*)$, each of them is learned by fitting the residual parameter between the ground truth and the current estimation as in [39], which enable them to work collaboratively during runtime:

$$\mathbf{m}^t = \mathbf{m}^{t-1} + \mathcal{R}^{m,t}(h^{m,t}(I, \mathbf{U}^{m,t-1})), \tag{7}$$

$$\mathbf{w}^t = \mathbf{w}^{t-1} + \mathcal{R}^{w,t}(h^{w,t}(I, \mathbf{U}^{w,t})), \tag{8}$$

where $\mathbf{U}^{m,t-1}$ comes from Eq. (3) by finding $\{\mathbf{m}^{t-1}, \mathbf{w}^{t-1}, \mathbf{vt}^{t-1}\}$. After applying $\mathcal{R}^{m,t}$, we collaboratively update $\mathbf{U}^{w,t}$ with $\{\mathbf{m}^t, \mathbf{w}^{t-1}, \mathbf{vt}^t\}$, and $h^{*,t}$ is the feature extractor indexed based on current 2D estimation. The cascaded collaborative-regressor algorithm is independent of particular feature extractor and regressor. In our paper, we use nonlinear HOG-based regressor as $\mathcal{R}^{m,t}(*)$ [42] and the famous fern regressor as $\mathcal{R}^{w,t}(*)$ [7]. Consequently they are adopted to alternate between the estimation of $\mathbf{m}$ and $\mathbf{w}$, similar to face construction[29].

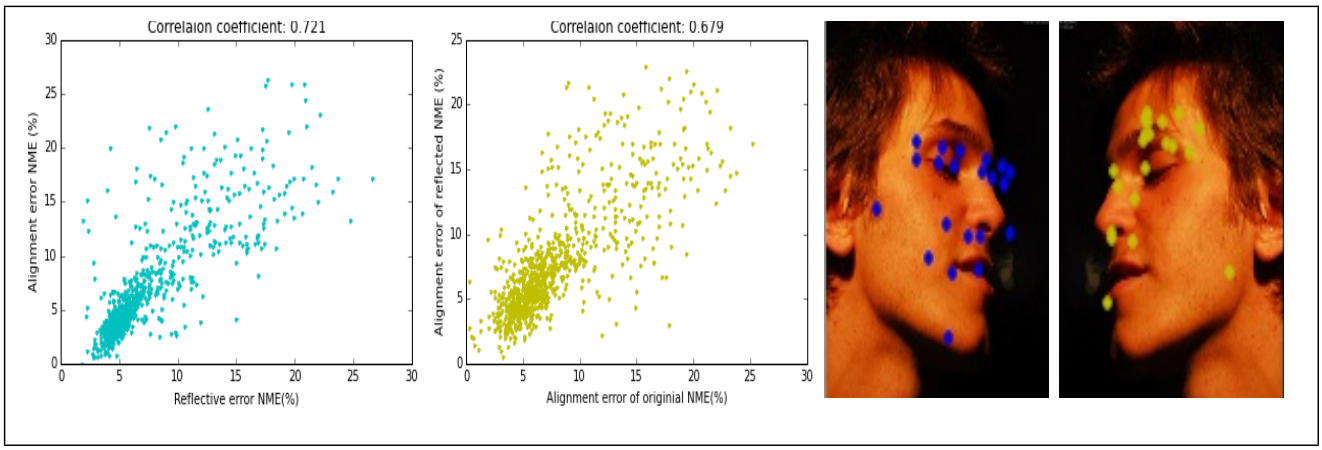

Figure 3: From left to right are: (a) correlation between alignment error and reflective error; (b) correlation of alignment errors between mutually reflected images; (c,d) misalignment examples of mutually reflected images.

Motivation of Reflective Feedback. CPR is initialization dependent. Previous works [7, 39] suggested to run the algorithm with several random initialisations and take the median as the final output. However, as stated in [41] only initialisations that are in the range of the optimal shape can converge to the correct solutions. Having several randomly generated initialisations does not guarantee that correct solutions are reached. In our case, firstly, 3D face alignment is inherently challenging in literature; Secondly, the spatial distribution of landmark is highly pose dependent; The representation is composed of 3D parameters, and the mapping is likely more complicated than CPR in 2D face alignment, which result in more failures. To overcome these issues, a reliable feedback mechanism is in needed. With this goal in mind, we decide to investigate how to marry our proposed reflective invariant metric with cascaded collaborative regressor. We augment the training set with reflected images to avoid bias. The results on testing set in Fig. 3 show that: a) the alignment error is strongly correlated with reflected error, when significant misalignment presents on one image, it is very likely to fail on the reflected image; b) The failure on mutual reflected images are not symmetrically consistent. With these nice properties, we use proposed reflected invariant metric as a *mirror* to reflect the *beauty* of prediction during runtime. Specifically, given a test image, we generate its reflected one, then they are processed by our Collaborative-Regressor and we compute its reflected error. If the reflected error is above a threshold (by cross-validation the threshold can be set to 0.1, metric described in experiment part) we restart with different initialisations, otherwise the original one is kept. Although more sophisticated approaches are possible, this straightforward scheme is already quite effective.

Smart Restart Based on our observation, the alignment error highly depends on parameters estimation, e.g., the error of a profile face is mainly caused by a yaw angle, the error of an open-mouth is always caused by a close-mouth parameter. We therefore relax the problem by incorporating a fast ConvNet head pose estimator[45]. For test images that require restart, we first get a reasonable head pose estimation, then we search samples in the training set that are with similar head pose as the test image. Subsequently, we calculate the similarity transformation between the two face bounding boxes to adjust scale and translation of selected samples. Once we get a set of better initializations, we feed one back to our RCCR in 3.3, this process can be continued until reflective error below a threshold, or the entire set is explored (in this case, we choose the output with smallest reflective error). The general framework of our proposed method can be found in Fig. 2.

4 Experiment

4.1 Experiment Preparation

In this paper, we propose a robust face alignment method for unconstrained large-pose setting. We therefore conduct the experiments on the following publicly available datasets:

300W and 300W-LP: The popular benchmark 300W [31], which include 3837 unconstrained face images from several datasets with 68 landmarks annotation. 300W-LP extended the original dataset across large poses by face profiling, which include 61,225 samples (1,786 from IBUG [31], 5207 from AFW [50], 16,556 from LFPW [5] and 37,676 from HELEN [27]), which is further expanded to 122,450 samples with horizontal flipping.

AFLW: AFLW [26] contains about $25,000$ real-world face images, each image is annotated with 21 landmarks upon visibility, and a bounding box. Given the dataset with large yaw pose variation (from -90° to $+90^\circ$), it is very suitable for evaluating our method across large poses.

AFW: AFW [50] was built using Fickr images. it has 205 images with 473 labeled faces. The annotations for each face include a bounding box, 6 landmarks and head pose angles ($\pm 90^\circ$) .

Experiment setup To build our proposed model, we select a subset of 2407 images from 300W-LP (excluding AFW) and 2754 images from AFLW for training, where the numbers of images whose absolute yaw angle within [0°, 30°], [30°, 60°], [60°, 90°] are evenly distributed, roughly $\frac{1}{3}$ in each category. Test set come from AFLW, samples fall in pose [0°, 30°], [30°, 60°] and [60°, 90°] are 428, 365 and 372 respectively. We first get the tightest bounding box of each image from the ground truth landmarks, and then expand its size by 10%, and add 10% noise to the top-left corner, width and height size of the bounding box, which mimic the imprecise face detection. For each image, we also generate its reflected image by flipping horizontally, note that, we will use the mutually reflected ones for both training and testing. We construct augmented information for each image using method described in Sec. 3.1, which contains estimated ground truth 3D face and corresponding 68 landmark for each sample. For training, We augment each training sample with 15 sets of initialisations $\{\mathbf{m}^0, \mathbf{w}^0\}$, one set is mean parameter (mean projection parameter in training samples, mean 3D shape by setting all shape parameters to zero), others from randomly selected sets of parameters in training data. The stage for our RCCR is $T = 10$, $R^{m,t}$ are learned to predict 6-dimensional projection parameters. $R^{w,t}$ regressor is composed of 300 primitive ferns that learned to predict 199-dimensional identity parameters w_{id} and 29-dimensional expression parameters w_{exp}. For evaluation, we use two conventional metrics for measuring the error of landmarks: 1) Normalised Mean Error (NME) [46] which is the average of landmark error normalised by the bounding box size rather than the common eye-to-eye distance. Since it is not well defined in large pose such as profile face, NME is obtained by setting the $f(I)$ as the square root of the face size in Eq.(5, 6). 2) Mean Average Pixel Error (MAPE) which is simply obtained by setting $f(I) = 1$.

First of all, we evaluate the effectiveness of our proposed method in component-wise manner on AFLW test set. We compare to 1) RCCR without reflective feedback (CCR). 2) RCCR with reflective feedback and 5 random restart initialisations (RCCR). 3) RCCR with reflective feedback and 5 smart restart initialization (RCCR + SR). The comparison is shown in Fig. 4. As can be seen on the left figure, by using the reflective feedback, we achieve big improvement over CCR, which suggests us a failure-alarm mechanism is indeed very useful. Moreover, by using the head pose based initialisations, we achieve even better performance,

though the improvement is relatively minor.

4.2 Comparison with Baselines

Baselines: Though remarkable progress have been reported on face alignment, there are very few works claimed as pose-free. We therefore choose recent popular methods with released codes: CDM [46], RCPR [7], SDM [40], and TCDCN [49]. Among them CDM [46] is the first one claimed to perform pose-free face alignment, which is a CLM-type method that can handle all the poses and outperforms multiple-view tree based methods. RCPR [7], which was proposed to handle heavy occlusion, and potentially able to deal with self occlusion caused by large pose. SDM[40] has shown superior performance in various tasks e.g. 3D pose estimation and face alignment. TCDCN [49] is a multi-task deep learning algorithm that represents recent development in face alignment. These methods not only well represent the major categories of the state of the arts in face alignment, but are also consistent with our goal of face alignment across large poses.

Comparison on AFLW We train RCPR, and SDM methods by using the same partition as mentioned in Sec. 4.1. To have a fully understanding of the benefit of 3D modelling, we train each method on the '300W' dataset (without 3D), and on the 'Full' set (contains 300W-LP and AFLW, with 3D) respectively. We then divide the test set of AFLW images into three subsets[0°, 30°], [30°, 60°], [60°, 90°] and evaluate each method on each subset respectively. The NME results for different methods are shown in Tab. 1. As expected, our proposed RCCR achieve state of the art performance when compared with baselines under different poses, and more impressively, it demonstrates robustness to pose-variation according to standard deviation. The improvements on 'Full' set over '300W' set for each method are significant, 51.20% for RCPR and 46.99% for SDM in [60°, 90°], due to its rich representation of 3D modelling, which can be used either for face profiling or shape parameter variations.

Comparison on AFW In addition to AFLW, we also report on AFW with two baselines, CDM and TCDCN. The reasons are: they were both evaluated on AFW though different metrics were exploited in the original papers, and they only provide the executable file of the trained models, which predicted on fewer landmarks. CDM does not rely on bounding box input since it integrates face detection and face alignment. TCDCN reports on 5 landmarks on the subset of images whose absolute yaw angle are up to 60°. To make a fair comparison, we use the metric MAPE, and all methods are evaluated on 5 landmarks. We choose a subset of 242 faces with successful face detection by CDM which share similar pose range as TCDCN. The results are showed in Fig. 4. Again, we see the consistent improvement of our proposed method over the baseline methods.

Methods	[0, 30]	[30, 60]	[60, 90]	Mean	Standard deviation
RCPR[7]	6.28	17.53	33.62	18.53	11.32
RCPR (on 300W)	5.51	9.94	22.37	12.28	7.15
RCPR (on Full)	5.54	6.37	10.91	7.51	2.35
SDM (on 300W)	5.33	7.26	18.15	10.03	5.61
SDM (on Full)	**5.32**	6.85	9.62	7.17	2.34
RCCR (on Full)	5.38	**5.87**	**6.86**	**6.01**	**1.32**
RCCR (Full + SR)	**5.27**	**5.34**	**6.53**	**5.70**	**1.24**

Table 1: The NME(%) results on AFLW.

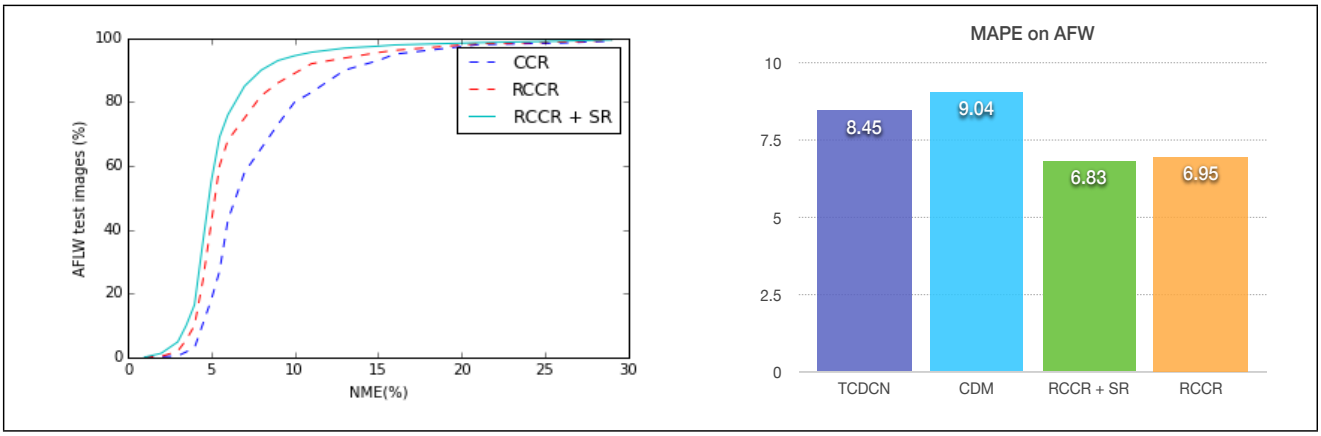

Figure 4: Left: the analysis results on AFLW with different versions of our proposed RCCR. Right: MAPE results of face alignment methods on AFW.

5 Conclusions

With the assistance of 3D Morphable Model, we in this paper propose a novel Reflective Cascaded Collaborative-Regressor algorithm, which achieves the state-of-the-art performance for the challenging problem of face alignment across large pose. Different from traditional CPR framework, we propose an "reflective invariant metric" and successfully marry it with CPR. We have shown that reflective variance is highly correlated with alignment error in our problem, which in turn provide us an failure-alarm signal and helps us to improve the algorithm robustness due to failure. We therefore suggest researchers to incorporate "reflective invariant" as a measure of success of the algorithm in the future.

Acknowledgement

The work of Xuhui Jia and Kwok-Ping Chan is supported by Hong Kong Research Grant Council, Project Code HKU 710412E. Niu is partially sponsored by National Natural Science Foundation of China (61403410).

References

[1] https://how-old.net/.

[2] Brian Amberg, Andrew Blake, and Thomas Vetter. On compositional image alignment, with an application to active appearance models. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1714–1721. IEEE, 2009.

[3] 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.

[4] 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.

[5] 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.

[6] Volker Blanz and Thomas Vetter. Face recognition based on fitting a 3d morphable model. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 25(9):1063–1074, 2003.

[7] 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.

[8] Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. Facewarehouse: a 3d facial expression database for visual computing.

[9] Chen Cao, Yanlin Weng, Stephen Lin, and Kun Zhou. 3d shape regression for real-time facial animation. *ACM Transactions on Graphics (TOG)*, 32(4):41, 2013.

[10] Chen Cao, Qiming Hou, and Kun Zhou. Displaced dynamic expression regression for real-time facial tracking and animation. *ACM Transactions on Graphics (TOG)*, 33(4): 43, 2014.

[11] Xin Cheng, Sridha Sridharan, Jason Saragih, and Simon Lucey. Rank minimization across appearance and shape for aam ensemble fitting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 577–584, 2013.

[12] T.F. Cootes, G.J. Edwards, and C.J. Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, 2001.

[13] D Cristinacce and T Cootes. Boosted regression active shape models. In *Proc. Brit. Mach. Vis. Conf.*, volume 2, pages 880–889, 2007.

[14] D Cristinacce and T Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 41(10):3054–3067, 2008.

[15] 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.

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

[17] Gabriele Fanelli, Matthias Dantone, Juergen Gall, Andrea Fossati, and Luc Van Gool. Random forests for real time 3d face analysis. *Int. J. Comput. Vis.*, 101(3):437–458, 2013.

[18] Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. Effective face frontalization in unconstrained images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4295–4304, 2015.

[19] Craig Henderson and Ebroul Izquierdo. Reflection invariance: an important consideration of image orientation. *arXiv preprint arXiv:1506.02432*, 2015.

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

[21] Pei-Lun Hsieh, Chongyang Ma, Jihun Yu, and Hao Li. Unconstrained realtime facial performance capture. In *Computer Vision and Pattern Recognition (CVPR)*, 2015.

[22] Xuhui Jia, Xiaolong Zhu, Angran Lin, and Kwok-Ping Chan. Face alignment using structured random regressors combined with statistical shape model fitting. In *2013 28th International Conference on Image and Vision Computing New Zealand (IVCNZ 2013)*, pages 424–429. IEEE, 2013.

[23] Xuhui Jia, Heng Yang, Angran Lin, Kwok-Ping Chan, and Ioannis Patras. Structured semi-supervised forest for facial landmarks localization with face mask reasoning. In *Proc. Brit. Mach. Vis. Conf.*, 2014.

[24] Amin Jourabloo and Xiaoming Liu. Pose-invariant 3d face alignment. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3694–3702, 2015.

[25] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1867–1874. IEEE, 2014.

[26] M. Kostinger, P. Wohlhart, P.M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, pages 2144–2151. IEEE, 2011.

[27] Vuong Le, Jonathan Brandt, Zhe Lin, Lubomir Bourdev, and Thomas S. Huang. Enhanced pictorial structures for precise eye localization under incontrolled conditions. In *Proc. Eur. Conf. Comput. Vis.*, pages 1621–1628. Springer, 2012.

[28] 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.

[29] Narendra Patel and Mukesh Zaveri. 3d facial model construction and animation from a single frontal face image. In *Communications and Signal Processing (ICCSP), 2011 International Conference on*, pages 203–207. IEEE, 2011.

[30] Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. A 3d face model for pose and illumination invariant face recognition. In *Advanced video and signal based surveillance, 2009. AVSS'09. Sixth IEEE International Conference on*, pages 296–301. IEEE, 2009.

[31] 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.*, pages 397–403, 2013.

[32] Walter J Scheirer, Anderson Rocha, Ross J Micheals, and Terrance E Boult. Meta-recognition: The theory and practice of recognition score analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(8):1689–1695, 2011.

[33] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. Cascaded hand pose regression. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 824–832, 2015.

[34] Philip A Tresadern, Patrick Sauer, and Timothy F Cootes. Additive update predictors in active appearance models. In *Proc. Brit. Mach. Vis. Conf.*, page 4, 2010.

[35] Sergey Tulyakov and Nicu Sebe. Regressing a 3d face shape from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3748–3755, 2015.

[36] Georgios Tzimiropoulos and Maja Pantic. Optimization problems for fast aam fitting in-the-wild. In *Proc. IEEE Int. Conf. Comput. Vis.*, pages 593–600, 2013.

[37] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 2729–2736, 2010.

[38] Peng Wang, Qiang Ji, and James L Wayman. Modeling and predicting face recognition system performance based on analysis of similarity scores. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(4):665–670, 2007.

[39] 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.

[40] 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.

[41] Xuehan Xiong and Fernando De la Torre. Supervised descent method for solving nonlinear least squares problems in computer vision. *arXiv preprint arXiv:1405.0601*, 2014.

[42] 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.

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

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

[45] Heng Yang, Wenxuan Mou, Yichi Zhang, Ioannis Patras, Hatice Gunes, and Peter Robinson. Face alignment assisted by head pose estimation. *arXiv preprint arXiv:1507.03148*, 2015.

[46] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, pages 1944–1951, 2013.

[47] Peng Zhang, Jiuling Wang, Ali Farhadi, Martial Hebert, and Devi Parikh. Predicting failures of vision systems. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 3566–3573, 2014.

[48] Shaoting Zhang, Yiqiang Zhan, Maneesh Dewan, Junzhou Huang, Dimitris N Metaxas, and Xiang Sean Zhou. Sparse shape composition: A new framework for shape prior modeling. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1025–1032. IEEE, 2011.

[49] 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.

[50] 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. http://www.ics.uci.edu/~xzhu/face/.

[51] Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. High-fidelity pose and expression normalization for face recognition in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 787–796, 2015.