Towards Robust Cascaded Regression for Face Alignment in the Wild

Chengchao Qu[1,2] Hua Gao[3] Eduardo Monari[2] Jürgen Beyerer[2,1] Jean-Philippe Thiran[3]
[1]Vision and Fusion Laboratory (IES), Karlsruhe Institute of Technology (KIT)
[2]Fraunhofer Institute of Optronics, System Technologies and Image Exploitation (Fraunhofer IOSB)
[3]Signal Processing Laboratory (LTS5), École Polytechnique Fédérale de Lausanne (EPFL)
firstname.lastname@iosb.fraunhofer.de firstname.lastname@epfl.ch

Abstract

Most state-of-the-art solutions for localizing facial feature landmarks build on the recent success of the cascaded regression framework [7, 15, 34], which progressively predicts the shape update based on the previous shape estimate and its feature calculation.

We revisit several core aspects of this framework and show that proper selection of regression method, local image feature and fine-tuning of further fitting strategies can achieve top performance for face alignment using the generic cascaded regression algorithm. In particular, our strongest model features Iteratively Reweighted Least Squares (IRLS) [18] for training robust regressors in the presence of outliers in the training data, RootSIFT [2] as the image patch descriptor that replaces the original Euclidean distance in SIFT [24] with the Hellinger distance, as well as coarse-to-fine fitting and in-plane pose normalization during shape update.

We show the benefit of each proposed improvement by extensive individual experiments compared to the baseline approach [34] on the LFPW dataset [4]. On the currently most challenging 300-W dataset [28] and COFW dataset [4], we report state-of-the-art results that are superior to or on par with recently published algorithms.

1. Introduction

Facial image analysis is an important research topic in the computer vision community. Localization of facial feature landmarks, a.k.a. face alignment, is an early but crucial step in this context for the latter processing stages, *e.g.*, face recognition [16], pose estimation [25], facial expression classification [8] and face hallucination [32]. In the wake of the explosive growth of Internet image and video data from social media, despite the broad interest and research effort since the seminal work Active Shape Model (ASM) [11] and Active Appearance Model (AAM) [9], there still remain challenges under uncontrolled conditions, *e.g.*, occlusion, extreme pose and shape variations.

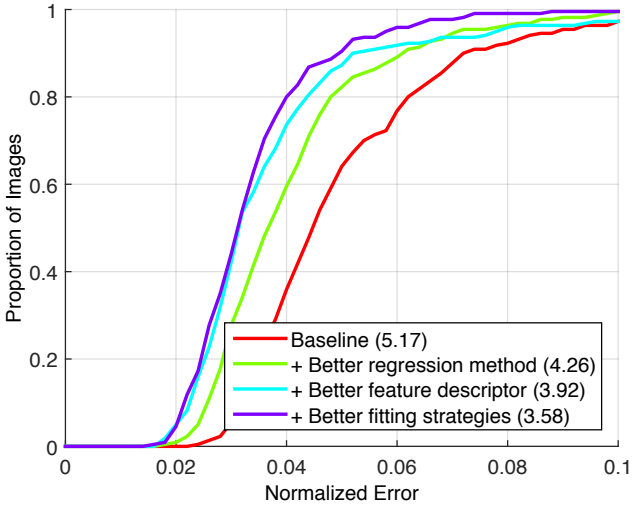


Figure 1: Overview of the performance gain with mean error through each proposed improvement on top of the baseline evaluated on the LFPW dataset [4]

Classic ASM and AAM approaches jointly optimize the shape parameters with local or global texture. In the last few years, a new family of face alignment algorithm emerges, which directly learns regressors from image feature descriptors to the target shape update. These regression-based methods are gaining increasing popularity due to their leading performance and high efficiency in the face alignment task. Although recent studies [7, 27] suggest that performance may have saturated on simple uncontrolled indoor (*e.g.*, BioID [20]) or outdoor (*e.g.*, LFPW [4]) datasets, reliable detection of facial feature points is still a distant promise on new challenging in-the-wild datasets (*e.g.*, 300-W [28] and COFW [6]). Unlike previous approaches that try to mitigate the impact of occlusion [6], feature selection [27] and initialization [35] with specific solutions, we instead revisit some of the low-level aspects of cascaded re-

gression. By reconsidering the essential assumptions and design choices, we argue that it is possible to achieve significant improvements and state-of-the-art performance on those datasets without bells and whistles.

In the spirit of the baseline cascaded regression-based approach [34], we further investigate the fundamentals and seek for enhancement in quest of successful in-the-wild landmark localization based on a series of experiments. Highlights of our approach include:

- **Robust regression** As a core component of the underlying framework, the quality of regression has a huge influence on the trained model. Iteratively Reweighted Least Squares (IRLS) alleviates the impact of outliers and noises which are inevitable in real-world data, especially in the presence of extreme pose, occlusion and illumination condition in unconstrained face datasets.

- **RootSIFT** The Hellinger distance proves to be preferable in histogram-based matching problems [2]. By applying square root during the feature map space conversion, small histogram bin values get more emphasized. In this way, face alignment accuracy is boosted dramatically.

- **Fitting strategies** Pose, novel expression and occlusion can all cause the initialized landmarks to drift far away from the true location. Thus, a larger local image patch size and compensation for in-plane face rotation account for fast convergence in early cascade stages, whereas a smaller patch size ensures high precision in the final stages.

The remainder of this paper is organized as follows. A brief introduction to the previous work in face alignment is given in §2. §3 recalls the baseline framework. The individual proposed improvements are discussed and analyzed in detail in §4, §5 and §6 respectively. Quantitative results of our final strong landmark detector are demonstrated in comparison with state-of-the-art methods in §7. Finally, we conclude our work in §8.

2. Related work

The fiducial facial landmarks, which face alignment algorithms aim to detect, are usually located at facial features that have semantic meaning, such as eyes, nose, mouth and chin. The sharp edges and corners in the facial texture near the feature points are exploited to approach the true landmark locations.

Following the pioneering work of ASM [11] and AAM [9], explicitly constraining shape variations by a linear shape subspace spanned by Principal Component Analysis (PCA) has become a standard methodology, jointly optimized with holistic appearance in a generative [1, 9] or discriminative [17, 23] fashion, as well as with part-based local classifiers [3, 12] or regressors [10, 14] using the Constrained Local Model (CLM) framework.

The latest trend in cascaded regression-based approaches [6, 7, 21, 27, 34, 35] has seen a great success. Inspired by the novel cascaded pose regression by Dollár *et al*. [15] that avails of cascaded simple regressors to approximate the mapping of rigid pose estimation, a progressive fitting algorithm with two-layered boosted ferns is introduced by Cao *et al*. [7] to regress the shape increment. Local shape-indexed feature instead of global ones [15] are adopted for the complex geometry of human faces. Together with correlation-based feature selection and random projection, state-of-the-art performance in both fitting accuracy and efficiency is reached. Later, Xiong and De la Torre [34] formulate the problem as a sequence of supervised gradient descent steps. The handcrafted SIFT feature extracted around the facial landmarks are fed to the linear least squares problem and the descent direction is learned to guide the current shape estimate towards the desired location. Though without explicitly modeling the shape mode, the implicit shape constraint still holds since each shape increment lies on the manifold of the training data, providing better generalization to novel shapes.

By virtue of the aforementioned advances, some standard indoor and outdoor benchmark datasets, *e.g*., BioID [20] and LFPW [4], can be safely regarded as resolved [7, 27]. However, with the introduction of newly published challenging datasets, *e.g*., 300-W [28] and COFW [6], which go one step further towards in-the-wild settings, face alignment performance is still far from satisfactory. Burgos-Artizzu *et al*. [6] extend [7] with occlusion handling by incorporating the occlusion information into regression and determine the location of the shape-indexed features based on two landmarks instead of one. Regarding adaptive feature selection, the greedy global scheme in [7] is redesigned by Ren *et al*. [27] to enable learning of representative local features by random forests [5]. Yan *et al*. [35] observe unstable shape initializations caused by challenges in unconstrained face detection. Multiple initial shape hypotheses of different facial parts are combined according to the structural SVM [29] outputs. Nevertheless, we approach the in-the-wild issue from an internal perspective of the cascaded regression framework. The reader is referred to [31] for a comprehensive review of up-to-date face alignment techniques.

3. Cascade of linear regressors

Within the framework introduced in [7, 15, 34], face alignment is naturally interpreted as a regression problem for the target output shape $\mathbf{x}$ given an input image $\mathbf{I}$ and an initial shape $\mathbf{x}^{(0)}$, which is commonly chosen as the mean shape of the training data scaled and translated with regard

to the bounding box of a face detector [30]. Here the vectorized shape $\mathbf{x} = [x_1, \dots, x_P, y_1, \dots, y_P] \in \mathbb{R}^{1\times 2P}$ is parametrized by the image coordinates of the P facial landmarks. The core idea is then to learn a regression function $\mathbf{r}(\cdot,\cdot)$ that returns an updated shape by minimizing

$$\sum_{i=1}^{N} \left\| \mathbf{r}(\mathbf{I}_i, \mathbf{x}_i^{(0)}) - \mathbf{x}_i^{\star} \right\|_2^2, \tag{1}$$

where i denotes the index of the totally N training samples and $\mathbf{x}^{\star}$ is the ground truth shape. While a one-pass regression is incapable of understanding the high complexity of the problem [34], composition of multiple regressors

$$\mathbf{r} = \mathbf{r}^{(T)} \circ \mathbf{r}^{(T-1)} \circ \cdots \circ \mathbf{r}^{(1)}, \tag{2}$$

a.k.a. a cascade of regression, proves to be effective [7, 15, 34], where the output shape of the previous regressor $\mathbf{r}^{(t-1)}$ is fed to the following one $\mathbf{r}^{(t)}$ as the input shape and T denotes the total number of stages.

As long as the initial shape $\mathbf{x}^{(0)}$ is valid, the subsequent shapes $\{\mathbf{x}^{(t)}\}$ are guaranteed to lie in the linear subspace of the training shapes by regression. This implicit shape constraint not only makes the algorithm exempt from an explicit shape model as in CLM, but also encourages to fit to novel shapes that share little similarity with the mean shape, which is favorable towards in-the-wild settings.

Next, the regression function $\mathbf{r}^{(t)}$ is specified as

$$\mathbf{r}^{(t)}(\mathbf{I}_i, \mathbf{x}_i^{(t)}) = \mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} + \mathbf{\Phi}(\mathbf{I}_i, \mathbf{x}_i^{(t)})\mathbf{R}^{(t)} + \mathbf{b}^{(t)}, \tag{3}$$

where $\mathbf{\Phi}(\mathbf{I}, \mathbf{x}) \in \mathbb{R}^{1\times PD}$ extracts the appearance feature mapping, such as raw intensity, shape-indexed features [6, 7, 27] or SIFT [34], in the proximity of $\mathbf{x}$ on the image $\mathbf{I}$, where D is the dimensionality of the feature. The descent direction $\mathbf{R}^{(t)} \in \mathbb{R}^{PD\times 2P}$ and bias term $\mathbf{b}^{(t)} \in \mathbb{R}^{1\times 2P}$ characterize the stage regressor $\mathbf{r}^{(t)}$ and are learned by incorporating Eq. (3) into Eq. (1)

$$\underset{\mathbf{R}^{(t)}, \mathbf{b}^{(t)}}{\arg\min} \sum_{i=1}^{N} \left\| \Delta\mathbf{x}_i^{(t)} - \mathbf{\Phi}(\mathbf{I}_i, \mathbf{x}_i^{(t)})\mathbf{R}^{(t)} - \mathbf{b}^{(t)} \right\|_2^2, \tag{4}$$

where $\Delta\mathbf{x}_i^{(t)} = \mathbf{x}_i^{\star} - \mathbf{x}_i^{(t)}$ is the desired optimal increment with reference to the current shape $\mathbf{x}_i^{(t)}$.

During the training phase, starting from the (perturbed) initial landmark locations $\{\mathbf{x}_i^{(0)}\}$, after $\mathbf{R}^{(0)}$ and $\mathbf{b}^{(0)}$ are learned by minimizing Eq. (4) using least squares fitting, a new set of training shapes $\{\mathbf{x}_i^{(1)}\}$ is generated by applying Eq. (3) to the regression output. A small number of iterations then suffice to successively converge $\{\mathbf{x}_i^{(t)}\}$ to $\{\mathbf{x}_i^{\star}\}$.

4. Which regression method?

Minimizing Eq. (4) is widely known as the linear least squares problem. Stacking all N training samples yields $\Delta\mathbf{X}^{(t)} = \left[\Delta\mathbf{x}_i^{(t)\top}, \dots, \Delta\mathbf{x}_N^{(t)\top}\right]^{\top} \in \mathbb{R}^{N\times 2P}$, $\widetilde{\mathbf{\Phi}}^{(t)} = \left[\left[\mathbf{\Phi}(\mathbf{I}_1, \mathbf{x}_i^{(t)}), 1\right]^{\top}, \dots, \left[\mathbf{\Phi}(\mathbf{I}_N, \mathbf{x}_N^{(t)}), 1\right]^{\top}\right]^{\top} \in \mathbb{R}^{N\times(PD+1)}$ and $\widetilde{\mathbf{R}}^{(t)} = \left[\mathbf{R}^{(t)\top}, \mathbf{b}^{(t)\top}\right]^{\top} \in \mathbb{R}^{(PD+1)\times 2P}$. To avoid the singular problem, a common practice is to append a regularization term to Eq. (4) as ridge regression [19]

$$\underset{\widetilde{\mathbf{R}}^{(t)}}{\arg\min} \left\| \Delta\mathbf{X}^{(t)} - \widetilde{\mathbf{\Phi}}^{(t)}\widetilde{\mathbf{R}}^{(t)} \right\|_F^2 + \gamma \left\| \widetilde{\mathbf{R}}^{(t)} \right\|_F^2, \tag{5}$$

which can be solved in closed form

$$\widetilde{\mathbf{R}}^{(t)} = \left(\widetilde{\mathbf{\Phi}}^{(t)\top}\widetilde{\mathbf{\Phi}}^{(t)} + \gamma\mathbf{I}_{\mathrm{id}} \right)^{-1} \widetilde{\mathbf{\Phi}}^{(t)\top}\Delta\mathbf{X}^{(t)}, \tag{6}$$

where $\mathbf{I}_{\mathrm{id}}$ stands for the identity matrix.

Due to the inevitable existence of noise in the training data, including annotation error, extremely difficult samples and local minima, ordinary linear regression is suboptimal, which assumes that the error is normally distributed. However, it is well known that even a small number of gross outliers can hugely bias the regressed model[1].

Iteratively Reweighted Least Squares (IRLS) IRLS offers an iterative solution to diminish the negative influence of noisy data samples [18]. Each iteration at stage s solves the weighted least squares problem

$$\underset{\mathbf{R}^{(s)}, \mathbf{b}^{(s)}}{\arg\min} \sum_{i=1}^{N} w_i^{(s)} \left\| \Delta\mathbf{x}_i - \mathbf{\Phi}(\mathbf{I}_i, \mathbf{x}_i)\mathbf{R}^{(s)} - \mathbf{b}^{(s)} \right\|_2^2, \tag{7}$$

where $w_i^{(s)}$ are the entries of the diagonal weighting matrix $\mathbf{W}^{(s)}$ with initial values set to $w_i^{(0)} = 1$ and for the purpose of clarity, the superscript (t) denoting regression stage is omitted. Similar to Eq. (6),

$$\widetilde{\mathbf{R}}^{(s+1)} = \left(\widetilde{\mathbf{\Phi}}^{\top}\mathbf{W}^{(s)}\widetilde{\mathbf{\Phi}} \right)^{-1} \widetilde{\mathbf{\Phi}}^{\top}\mathbf{W}^{(s)}\Delta\mathbf{X}. \tag{8}$$

The weighting matrix $\mathbf{W}^{(s)}$ is updated in contrast to the regression residuals. Specifically, in case of ℓ_1-norm,

$$w_i^{(s)} = \frac{K}{\left\| \Delta\mathbf{x}_i - \mathbf{\Phi}(\mathbf{I}_i, \mathbf{x}_i)\mathbf{R}^{(s)} - \mathbf{b}^{(s)} \right\|_1}, \tag{9}$$

where K as well as the regularization parameter γ are experimentally determined in §7.

[1] http://www.mathworks.com/help/stats/robustdemo.html

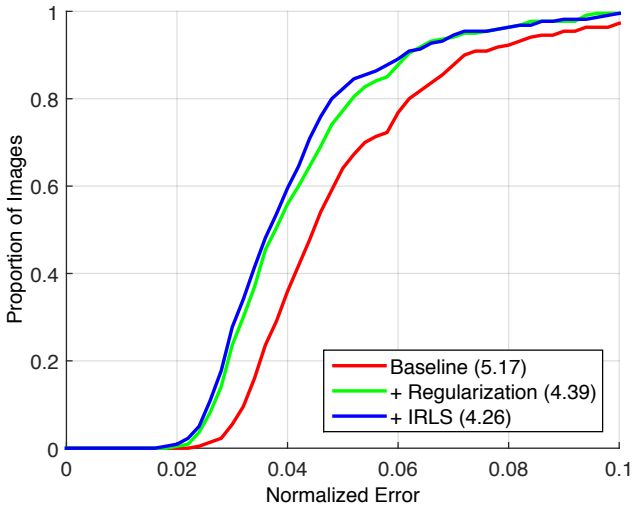


Figure 2: Performance on LFPW [4] by combining better regression methods

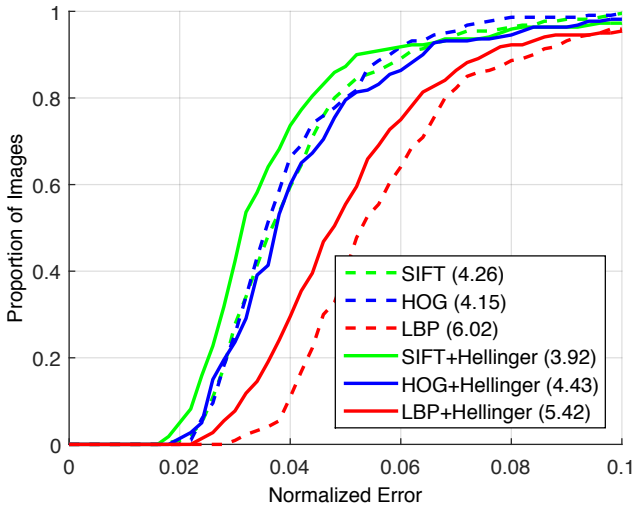


Figure 3: Performance on LFPW [4] by comparing various feature descriptors and the Hellinger feature map

The algorithm stops when $\mathbf{W}^{(s)}$ converges, which usually needs only a few iterations in our experiments. The mathematical formulation of IRLS reduces the contribution of outliers to a small extent, which keeps our regression model as little affected as possible and robust against unconstrained conditions in the training data.

Experiments and discussion We conduct intermediate experiments to validate the necessity of each proposed improvements for building our final landmark detector and the progress incrementally. To keep the compactness of the experiments here, more details are discussed in §7. The widely used Labeled Face Parts in the Wild (LFPW) dataset [4] is chosen as the benchmark. As some volatile URLs in LFPW are no longer valid, we only collect 810 and 220 images for training and testing respectively. Our baseline cascaded regression implementation resembles [34] with ordinary least squares and SIFT feature. It's worth a mention that due to different size of data (*c.f.* [4]), multiple initializations (*c.f.* [7]) or manual correction of erroneous landmarks in [34][2], we are unable to reproduce the same results as reported in respective papers on LFPW.

As illustrated in Fig. 2, simple ridge regression performs surprisingly well with considerable improvement in both precision and convergence of the curve approaching 100% in y-axis. With the adoption of IRLS, localization accuracy further increases by a small amount, indicating a more robust model against outliers during the learning. However, convergence remains almost unchanged, possibly because nearly all of the images already have a mean normalized error less than 10% of the interocular distance (IOD). In §7, the benefit is more evident as expected.

[2]By direct correspondence with the author

5. Which feature descriptor?

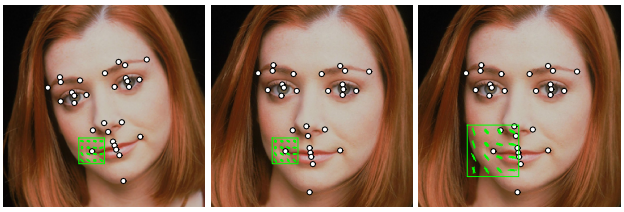
Since the choice of features is a key design factor, it is reasonable to experiment with other feature descriptors and mappings other than SIFT in the baseline method [34]. Given the fact that the popular descriptors like SIFT [24], HOG [13] and LBP [26] are all histogram-based ones, question naturally arises if the Euclidean distance employed in regression also yields inferior results compared to the Hellinger distance, observed in many other computer vision tasks like image retrieval [2] and face recognition [33].

The Hellinger kernel maps the original histogram to the element-wise square root of the ℓ_1-normalized

$$\mathbf{H}(\mathbf{x}, \mathbf{y}) = \sum_i \sqrt{x_i y_i}, \tag{10}$$

such that comparing transformed histograms in Euclidean space is equivalent to comparing the original descriptors in Hellinger space. It is obvious that the Hellinger distance augments small histogram bin counts, which are overwhelmingly suppressed by large bin values in Euclidean space. Therefore, to obtain higher localization precision, we apply square root to the ℓ_1-norm prior to the distance calculations to all evaluated feature descriptors.

Experiments and discussion With the IRLS algorithm fixed as our regression method according to the outcome of the previous section, we evaluate SIFT, HOG and LBP as feature descriptor with optional Hellinger distance mapping. Standard settings of HOG and LBP, namely 8×8 cell size, 2-by-2 blocks and 50% overlapping for HOG [13], as



(a) (b) (c)

Figure 4: Local image feature descriptor extracted from (a) the original image, pose-normalized image in (b) initial and (c) final cascade stages

well as $\text{LBP}^{u2}_{8,2}$ [26] are deployed. The Hellinger mapping is computed on the fly.

Fig. 3 presents the contribution of the respective feature descriptors. At first sight, only HOG and SIFT+Hellinger (RootSIFT) successfully bring smaller mean normalized error than the baseline SIFT after all. Both LBP variants in Euclidean and Hellinger space cannot compete with the rest. Interestingly, HOG+Hellinger performs a bit worse than the original HOG, which is the only one of the three histogram-based features that fails to improve under Hellinger feature map. SIFT+Hellinger (RootSIFT) reveals the best result in spite of the degradation in convergence. This trend is visible in HOG and LBP as well, though less obvious. The reason might be self-explanatory and referred to the definition. Whilst emphasis on small bin values improves fine fitting precision, suppression of larger bins leads to less sensibility to large shape variations. In the next section, we address this problem by looking for better fitting strategies to boost the convergence property on LFPW.

6. Which fitting strategies?

In face alignment, the initial shapes are usually determined as the mean shape of the training data scaled and translated with regard to the bounding box of a face detector [30]. In addition, random perturbation is imposed to the initial shapes in the training stage to take into account more harsh conditions and rough initializations in the testing phase. In the course of training the cascaded regressors, the variance of the face shapes reduces gradually, approaching the ground truth in the final stages. Hence, a rational strategy is to use large local patches for feature extraction at early stages to allow for large uncertainty, whereas at later stages, fine-scale local patches facilitate accurate landmark localization. A similar approach also appears in [35, 36].

Last but not least, modern face detectors, even trained for frontal upright face detection, can tolerate a certain degree of in-plane rotation. On the other hand, most widely used feature descriptors for face alignment, *e.g.*, standard SIFT, HOG and LBP, are *not* rotational invariant. The regressor

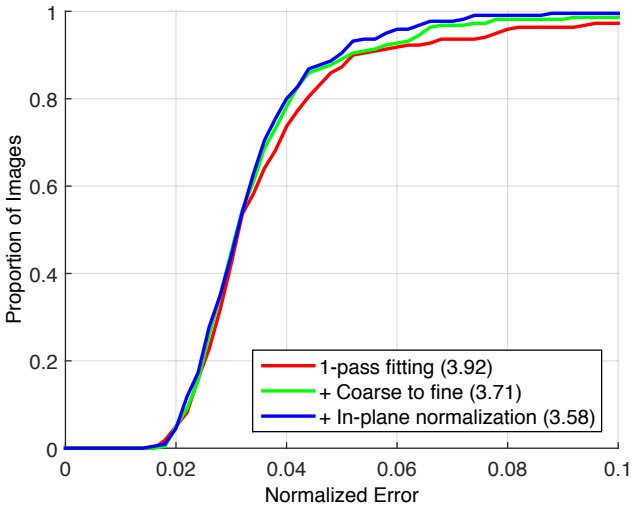


Figure 5: Performance on LFPW [4] by combining better fitting strategies

must then extra model the angle between the upright feature descriptors and the rotated face shape. We propose a 2-pass strategy in testing to mitigate the issue. In the first pass, we make use of the trained regressor to compute an approximate shape. Afterwards, the similarity transform to the upright mean shape is calculated and this temporary shape is discarded. Finally, the regressor is applied in the second pass to the features extracted from the pose-normalized image rotated and scaled subject to the similarity transform. Apparently, training regressors should also conform to the same procedure.

Fig. 4 reveals an example comparison of the baseline with the proposed fitting method.

Experiments and discussion In the last section, the RootSIFT feature loses a few percent in convergence of the cumulative error curve in y-axis. Fig. 5 demonstrates the loss is immediately reclaimed and improved when adapting the local patch size to different stages. The mean normalized error further decreases to the state-of-the-art level with compensation of in-plane rotation in both learning and fitting. In the next section, this fine-tuned landmark detector will be benchmarked on more challenging datasets.

7. Experiments

Datasets Apart from LFPW [4], two more recent and challenging datasets are employed.

300-W is created for the 300 Faces in-the-Wild Challenge [28], which combines several existing indoor and outdoor datasets and a new dataset (IBUG), with a unified 68-point markup. Since the original testing dataset is held for future challenges, we split the whole data into a training

Table 1: Mean normalized errors tested with different IRLS parameters γ and K on LFPW [4]

γ \ K	1	3	5	7	Mean	Std
20	3.74	3.87	3.90	3.94	3.86	0.07
50	3.60	3.72	3.75	3.77	3.71	0.06
100	3.60	3.63	3.65	3.67	3.64	0.03
200	3.68	3.60	3.61	3.62	3.63	0.03
20∼80	3.65	3.73	3.75	3.76	3.72	0.04
50∼200	3.65	3.64	3.66	3.68	3.66	0.02
100∼400	3.76	3.64	3.66	3.66	3.68	0.05
200∼800	4.01	3.70	3.69	3.70	3.77	0.13
80∼20	3.62	3.77	3.80	3.83	3.76	0.08
200∼50	**3.57**	3.63	3.66	3.68	3.63	0.04
400∼100	3.65	3.58	3.61	3.62	**3.61**	**0.02**
800∼200	3.90	3.61	3.61	3.62	3.68	0.12
Mean	3.70	**3.68**	3.69	3.71		
Std	0.13	**0.08**	0.09	0.10		

set of totally 3148 images, consisting of AFW [37] and the training sets of LFPW [4] and Helen [22], and a testing set of 689 images in total, composed of IBUG and the testing sets of LFPW and Helen. Following [27], we also divide the 300-W testing set into a common subset of LFPW and Helen, and a challenging subset of IBUG, which contains extremely large variations in pose, expression, occlusion and illumination (shortened to "Comm." and "Chlg." respectively in Tab. 2).

COFW is short for Caltech Occluded Faces in the Wild [6], which complements LFPW [4] with more occluded faces and occlusion annotation for landmarks. All 1345 training and 507 testing images have the same 29-point scheme as LFPW. Note that we do *not* use the occlusion mask to train an occlusion-aware model as in [6]. Instead, standard cascaded regression with exclusively the proposed improvements is exploited.

Evaluation criteria Standard average error normalized by interocular distance (IOD) and the corresponding cumulative error curves are reported on all datasets. On COFW, percentage of failure cases, meaning images with a larger error than 10%, is also provided as in [6]. The percentage sign for the error is dropped for the sake of clarity.

Implementation details The initial shape for the first cascade stage is initialized by the bounding box either detected by the standard Viola–Jones algorithm [30] for LFPW or included in 300-W and COFW. After squaring the bounding box at the same position, the face is cropped to 200×200 pixels and the mean shape is scaled and centered with reference to the normalized square. All feature descriptors are computed on 32×32 local patches and then projected to the PCA subspace with 98% variance to reduce dimensionality. To augment the training data [34], 10 perturbed samples per training image, with a standard deviation of 10 pixels for translation, 0.05 for scaling and 5° for rotation, are generated. Merely 4 cascade stages suffice to obtain satisfactory results on most images.

Table 2: Mean normalized errors and failures on 300-W [28] and COFW [6]

Methode	300-W			COFW	
	All	Comm.	Chlg.	Error	Failure
ESR [7]	7.58	5.28	17.00	11.2	36%
SDM [34]	7.52	5.60	15.40	—	—
LBF [27]	6.32	4.95	**11.98**	—	—
RCPR [6]	—	—	—	8.5	20%
Baseline[3]	7.40	5.90	13.57	9.9	37%
Proposed	**6.24**	**4.83**	12.02	**6.7**	**10%**
Human	—	—	—	5.6	0%

Comparison We first tune the IRLS parameters γ and K in Eqs. (5) and (9). γ for different cascade stages is specified by a sequence of choices, *i.e.*, same values, monotonically decreasing and increasing values. K controls the weighting matrix $\mathbf{W}^{(s)}$ in Eq. (9) as well as the convergence speed. Tab. 1 clearly reveals outstanding performance of 3.58 mean error on LFPW with parameters $\gamma = [400, 300, 200, 100]$ and $K = 3$, each having the least average error and highest stability through all test cases. We thus fix the values for 300-W and COFW in favor of those yielding the absolute best result of 3.57, which might be overfitted because $K = 1$ generally gives the worst performance. Overall, the proposed work (3.58) is on par with the best available methods reported on LFPW, *i.e.*, CE [4] (3.99), ESR [7] (3.43), SDM [34] (3.47), RCPR [6] (3.5), ERT [21] (3.8) and LBF [27] (3.35).

Note that because of the different size of retrieved data, face detection rate, and initialization strategies, *etc.*, the results on LFPW are not quite conclusive and comparable. Particularly, as pointed out in [31], some deployed face detectors struggle with difficult face images and they are removed from the evaluation, which questionably elevates the average score. On the contrary, 300-W and COFW both provide a fixed number of images and bounding boxes for initializing shapes. Tab. 2 lists the average errors and failure rates, if applicable, on these datasets. Results of the competing methods are reported in [27] and [6] for 300-W and COFW respectively. The similar results of our baseline SDM [34] on 300-W confirms the correctness of our implementation. On the full set and the common subset, we suc-

[3] Our implementation of [34]

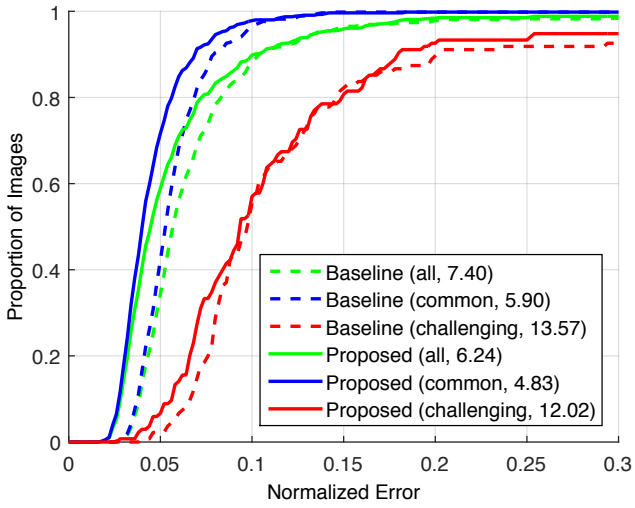


Figure 6: Cumulative error curves on 300-W [28]

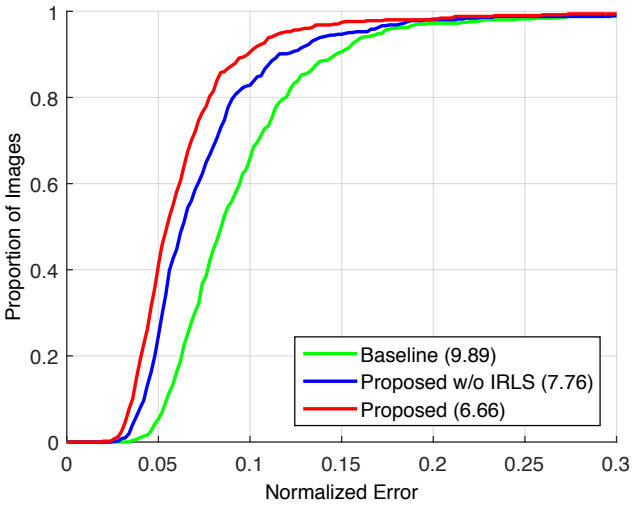


Figure 7: Cumulative error curves on COFW [6]

cessfully improve over the state-of-the-art LBF [27]. On the challenging IBUG dataset, only negligible deficit (12.02 *vs.* 11.98) is reported. On COFW, even without involving the occlusion annotation in training, our method outperforms RCPR [6] by significant error and failure reduction of 20% and 50% respectively. Larger contribution from IRLS is observed on COFW than on LFPW in Fig. 7. The overall performance falls to 7.8 mean error (6.45 on 300-W) and 17% failure when only ridge regression is applied alongside other proposed improvements, justifying the robustness against outliers. Example results from the challenging IBUG subset and COFW are presented in Figs. 8 and 9 respectively, indicating that despite recent advances in face alignment that nearly catch up with human precision, occlusion and large shape variations still remain a huge challenge for reliable and accurate landmark detection in the wild.

8. Conclusions

Following the design flow of the cascaded regression framework, we revisit the essential components and propose a proper regression method, feature descriptor and fitting strategies pursuing robust in-the-wild facial landmark localization. Extensive experiments help us identify the positive factors that most benefit the fitting performance over the baseline. On the challenging 300-W and COFW datasets, state-of-the-art results are achieved in spite of the straightforward approach, improving over more sophisticated algorithms with adaptive feature selection and occlusion handling. Nevertheless, our approach is non-excludable and we believe that combining those ideas may provide further boost in face alignment performance.

Acknowledgment

This study was partially supported by the MisPel project, co-funded by the German Federal Ministry of Education and Research (BMBF) under grant 13N12063, and by the MobilePass project, co-funded by the European Union under FP7 grant 608016.

References

[1] J. Alabort-i-Medina and S. Zafeiriou. Bayesian active appearance models. In *CVPR*, pp. 3438–3445, 2014. 2

[2] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *CVPR*, pp. 2911–2918, 2012. 1, 2, 4

[3] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, pp. 3444–3451, 2013. 2

[4] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, pp. 545–552, 2011. 1, 2, 4, 5, 6

[5] L. Breiman. Random forests. *Mach. Learn.*, 45(1):5–32, 2001. 2

[6] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. In *ICCV*, pp. 1513–1520, 2013. 1, 2, 3, 6, 7, 8

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, pp. 2887–2894, 2012. 1, 2, 3, 4, 6

[8] S. W. Chew, P. Lucey, S. Lucey, J. M. Saragih, J. F. Cohn, and S. Sridharan. Person-independent facial expression detection using constrained local models. In *FGR*, pp. 915–920, 2011. 1

[9] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In *ECCV*, vol. 1407, pp. 484–498, 1998. 1, 2

[10] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regres-

Figure 8: Example results from the challenging IBUG subset of the 300-W dataset [28] containing a mixture of unconstrained circumstances including pose, expression, illumination variations and occlusion. Successful fittings and some failure cases are highlighted with green and red frames respectively.

Figure 9: Example results from the COFW dataset [6] mainly comprised of partial occluded faces. Successful fittings and some failure cases are highlighted with green and red frames respectively.

sion voting. In *ECCV*, vol. 7578, pp. 278–291, 2012. 2

[11] T. F. Cootes and C. J. Taylor. Active shape models — 'smart snakes'. In *BMVC*, pp. 266–275, 1992. 1, 2

[12] D. Cristinacce and T. F. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, pp. 95.1–95.10, 2006. 2

[13] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, vol. 1, pp. 886–893, 2005. 4

[14] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, pp. 2578–2585, 2012. 2

[15] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, pp. 1078–1085, 2010. 1, 2, 3

[16] H. Gao, H. K. Ekenel, and R. Stiefelhagen. Pose normalization for local appearance-based face recognition. In *ICB*, pp. 32–41, 2009. 1

[17] H. Gao, H. K. Ekenel, and R. Stiefelhagen. Face alignment using a ranking model based on regression trees. In *BMVC*, pp. 118.1–118.11, 2012. 2

[18] P. J. Green. Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives. *J. R. Stat. Soc. B*, 46(2):149–192, 1984. 1, 3

[19] A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970. 3

[20] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz. Robust face detection using the Hausdorff distance. In *AVBPA*, vol. 2091, pp. 90–95, 2001. 1, 2

[21] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pp. 1867–1874, 2014. 2, 6

[22] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, vol. 7574, pp. 679–692, 2012. 6

[23] X. Liu. Discriminative face alignment. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(11):1941–1954, 2009. 2

[24] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.*, 60(2):91–110, 2004. 1, 4

[25] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):607–626, 2009. 1

[26] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002. 4, 5

[27] S. Ren, X. Cao, Y. Wei, and J. Sun. Face alignment at 3000 FPS via regressing local binary features. In *CVPR*, pp. 1685–1692, 2014. 1, 2, 3, 6, 7

[28] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, pp. 397–403, 2013. 1, 2, 5, 6, 7, 8

[29] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *J. Mach. Learn. Res.*, 6:1453–1484, 2005. 2

[30] P. Viola and M. J. Jones. Robust real-time face detection. *Int. J. Comput. Vis.*, 57(2):137–154, 2004. 3, 5, 6

[31] N. Wang, X. Gao, D. Tao, and X. Li. Facial feature point detection: A comprehensive survey. arXiv:1410.1037 [cs.CV], 2014. 2, 6

[32] N. Wang, D. Tao, X. Gao, X. Li, and J. Li. A comprehensive survey to face hallucination. *Int. J. Comput. Vis.*, 106(1):9–30, 2014. 1

[33] L. Wolf, T. Hassner, and Y. Taigman. Descriptor based methods in the wild. In *ECCVW*, 2008. 4

[34] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pp. 532–539, 2013. 1, 2, 3, 4, 6

[35] J. Yan, Z. Lei, D. Yi, and S. Z. Li. Learn to combine multiple hypotheses for accurate face alignment. In *ICCVW*, pp. 392–396, 2013. 1, 2, 5

[36] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *ECCV*, vol. 8690, pp. 1–16, 2014. 5

[37] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pp. 2879–2886, 2012. 6