Consensus of Regression for Occlusion-Robust Facial Feature Localization

Xiang Yu[1], Zhe Lin[2], Jonathan Brandt[2], and Dimitris N. Metaxas[1]

[1] Rutgers University, Piscataway, NJ 08854, USA
[2] Adobe Research, San Jose, CA 95110, USA

Abstract. We address the problem of robust facial feature localization in the presence of occlusions, which remains a lingering problem in facial analysis despite intensive long-term studies. Recently, regression-based approaches to localization have produced accurate results in many cases, yet are still subject to significant error when portions of the face are occluded. To overcome this weakness, we propose an occlusion-robust regression method by forming a consensus from estimates arising from a set of occlusion-specific regressors. That is, each regressor is trained to estimate facial feature locations under the precondition that a particular pre-defined region of the face is occluded. The predictions from each regressor are robustly merged using a Bayesian model that models each regressor's prediction correctness likelihood based on local appearance and consistency with other regressors with overlapping occlusion regions. After localization, the occlusion state for each landmark point is estimated using a Gaussian MRF semi-supervised learning method. Experiments on both non-occluded and occluded face databases demonstrate that our approach achieves consistently better results over state-of-the-art methods for facial landmark localization and occlusion detection.

Keywords: Facial feature localization, Consensus of Regression, Occlusion detection, Face alignment.

1 Introduction

Facial feature localization is a longstanding active research topic due to its wide applicability in computer vision and graphics [2,4,8,20,26,33]. Accurate localization is crucial for many applications, including automated face editing, face recognition, tracking, and expression analysis. Recent state-of-the-art methods such as [2,26] have achieved impressive results, not only on near-frontal faces but also faces in the wild. Despite these advances, the problem remains challenging due to large viewpoint variation, severe illumination conditions, various types of occlusions, etc.

Early successes in facial feature localization, epitomized by the Active Shape Model(ASM) [7] and Active Appearance Model(AAM) [6,17], are characterized by a parametric template that is fit to a given image by optimizing over the template's parameter space. Although effective for many cases, these parametric

D. Fleet et al. (Eds.): ECCV 2014, Part IV, LNCS 8692, pp. 105–118, 2014.

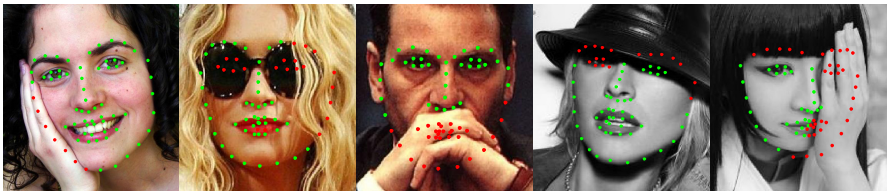


Fig. 1. Sample visual results from Helen, LFPW and COFW databases. Landmarks estimated by proposed method with occlusion detection (red: occluded, green: non-occluded).

approaches tend to break down under extreme pose, lighting and expression, due to lack of flexibility in the representation. Recently, regression-based methods [4,8,10,26] have been shown to overcome some of these difficulties, and have achieved high accuracy, largely due to their greater flexibility as compared to parametric methods, as well as effective sub-pixel localization capability. Despite these successes, a major weakness of the regression-based approach is occlusions, which occur often in faces in the wild (see, for example, Fig. 1). Regression-baded methods depend heavily on local appearance to obtain reliable feature location estimates. Occluded regions produce noisy features and result in erroneous location updates that not only affect the predicted locations of the occluded landmarks, but result in biased estimates of the visible landmarks as well.

In this paper, we propose to overcome the occlusion problem and improve the regression-based approach for facial feature localization. Our approach is based on the "consensus of experts" concept in machine learning. In our case, the "experts" are regressors that are each trained specifically to predict facial feature locations under the precondition that a particular region of the face is occluded. The occlusion region for each regressor is different from, yet overlapping with others. This enables a robust consensus to be formed using Bayesian inference. Note that regressor training requires no occlusion ground truth information because occlusion information is not used for each specific regressor. Once the landmark locations are determined, we employ a semi-supervised Gaussian MRF to smoothly propagate occlusion state labels from high-confident areas to the rest of the face.

Our contributions are as follows: 1) We propose a new regression-based facial feature localization method using a consensus of occlusion-specific regressors that effectively resists occlusions and achieves consistently better performance compared to state-of-the-art methods. The occlusion-specific regressors can be trained on standard landmark datasets without occlusion labels. 2) We propose a semi-supervised method using local occlusion detectors and a Gaussian MRF formulation to robustly identify coherent occluded regions. The resulting occlusion labels are shown to be competitive with the latest occlusion detection methods. 3) Extensive experiments on non-occlusion databases and occlusion databases are conducted to demonstrate the effectiveness of the proposed method.

2 Related Work

Facial landmark localization methods can be roughly divided into two major categories: parametric vs. non-parametric. Parametric methods are characterized by a model that attempts to capture facial appearance variations in terms of an underlying parameter space. Inference amounts to search in parameter space for the best-fitting model to the given image. In contrast, non-parametric methods learn to predict the face shape via training on a database, or by directly drawing exemplars in a data-driven manner.

The Active Shape Model (ASM) [7] and Active Appearance Model (AAM) [6] are both classical, seminal contributions to the parametric approach, with much follow-on work. Subsequently, the Constrained Local Model (CLM) [9,21] was introduced, which combines each local patch's alignment likelihood and predicts the optimal solution by maximizing the overall alignment likelihood. Component-wise ASM was proposed [13] to reduce the alignment error propagated among components. Le et al. [14] introduced a Viterbi process on facial contour fitting and user interaction model to improve the accuracy. Recently a fast AAM algorithm was presented for real time alignment [23], and an ensemble of AAM [5] was proposed to jointly register landmarks for image sequence. The combination of a part model and CLM [29] was proposed to alleviate pose variations, while other CLM frameworks focused on local patch expert learning [1].

In the category of non-parametric methods, Belhumeur et al. [2] proposed a data-driven method that employed RANSAC to robustly fit exemplar landmark configurations drawn from a database to a set of local landmark detections. Similar methods [22,31] either considered temporal feature similarity for joint face alignment or used graph matching to enhance the landmark localization. Notably, Zhu et al. [33] modeled the landmarks as a tree so that the positions could be efficiently optimized through dynamic programming.

Regression-based methods represent a significant sub-category of the non-parametric approach that have recently achieved high accuracy on standard benchmarks. An early contribution in this domain is Liang et al. [15], who proposed directional classifiers to predict the direction and step size of a landmark's update. Cristinacce and Cootes employed boosted regression [10] for local landmark alignment. Regression forest voting for accurate shape fitting was proposed by Cootes et al [8]. Valstar et al. [24] combined boosted regression with a graph model. Martinez et al. [16] proposed local evidence aggregation for regression based alignment. Dantone et al. [11] introduced conditional regression forests to treat faces with different poses separately. Dollar et al. [12] proposed cascaded pose regression to approximate 2D pose of objects. Rivera and Martinez [18] use kernel regression to handle low resolution images. Cao et al. [4] proposed a real-time explicit holistic shape regression method with robust shape indexed features. Xiong and De la Torre [26] proposed an efficient supervised descent method for regression training and inference. Yang et al. [28] employed dense interest points detection with sieving regression forests to obtain good results on faces in the wild.

In general, the regression-based approaches provide good accuracy with fast runtime. However, these methods suffer from the presence of occlusion due to the global nature of the regression which relies on the appearance around all the landmarks. To overcome this shortcoming, some researchers have proposed methods to cope with occlusion handling. For example, Roh et al. [19] used a large amount of facial feature detectors to provide over-sufficient landmark candidates and a RANSAC-based hypothesis and test method to robustly determine the whole shape. This method relies heavily on the facial feature detectors and is consequently computationally demanding. In [27], occlusion is modeled as a sparse outlier and the sparse constraint is applied during the optimization process. The sparse error could be from either occluded landmarks or perturbation of visible landmarks. Supervised occlusion detection methods are also proposed [25,30]. However, if a particular occlusion case is missing from the training set,these methods may fail. A recent work on face alignment with occlusion [3] attempts to use regression to predict the occlusion likelihood of landmarks. They divide the facial area into 3 by 3 blocks and use one non-occluded block each time to predict the landmark positions. The approach shows its positive effects but the statistical prior of each block's occlusion condition is fixed. In contrast, our approach uses all the features from the non-occluded regions. Though there is no occlusion prior, the proposed method applies Bayesian consensus over all the regressors to handle the occlusion.

3 Localization through Occlusion-Robust Regression

Linear regression has proven its effectiveness in facial landmark localization [4,26]. In order to tackle occlusion, we design a set of regressors which are designed specific to different occlusion conditions. For instance, a right eye regressor extracts features over all the landmarks except the landmarks of right eye, which we note as an occlusion-specific regressor. Then a Bayesian inference framework is introduced to predict the landmark positions by jointly considering all the regressor outputs and evidence of low-level appearance models. Encouraged by the regression results, we further apply SVM and Gaussian MRF regularization to identify the occluded landmarks.

3.1 Occlusion-Specific Regressors

As defined in [26], a linear regression based framework models the relationship between landmark displacement Δs and the local appearance Φ, shown in Eqn. 1.

$$\Delta s = A\Phi + b, R = (A, b) \tag{1}$$

where A is a regression matrix and b is the intercept. Φ is a feature vector concatenated by n feature vectors extracted at each fiducial point $(\tilde{x}_i, \tilde{y}_i)$, which indicates that each facial landmark's displacement is related to all other fiducial points' appearance.

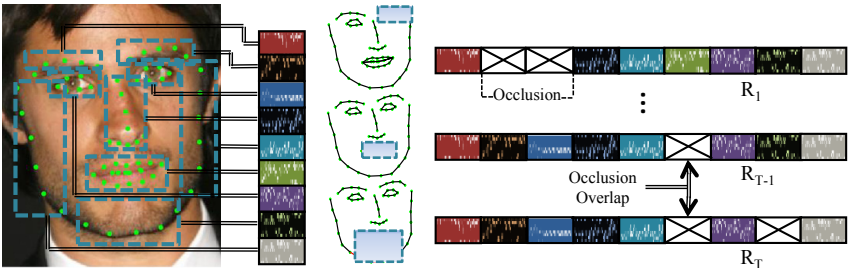


Fig. 2. Illustration of occlusion-specific regressors. Color blocks are regression weights for different components, i.e. left profile, mouth, etc. For different occlusion states, i.e. right eyebrow and right eye occlusion, the regressors are designed not to use the features from the occluded region. Those occlusion states are defined to have occlusion overlap with each other, e.g. mouth occlusion and mouth chin occlusion have overlap on the mouth area.

Based on this observation, we propose to train an ensemble of regressors, each of which handles one type of occlusions. The occlusions are combinations of different facial components, i.e. eyebrow, nose, left profile etc. The illustration is shown in Fig. 2. The training is almost the same as supervised descent method (SDM) [26]. The difference is that here we only extract features at those non-occluded landmarks, i.e. for training the mouth occlusion regressor, we only extract features at non-mouth landmarks. For robustness, the layouts of landmarks between different regressors overlap with each other. In this way, it is expected to be more than one regression result approaching optimal solution, which provides potential to conduct consensus of regressors.

Suppose there are T such regressors. (We define those T regressors as right-eyebrow-eye, right-eyebrow, right-eye, right-contour, left-eyebrow-eye, left-eye brow, left-eye, left-contour, chin, both-eyebrow, all-contour, both-eyes, chin-mouth, nose-mouth and mouth respectively). All of them are visually different because they are designed for different occlusions. In the training part, the goal is to minimize the regression error over all the training faces and all initialized landmark positions $s_t^k, t = 1, ..., T, k = 1, ..., K$. The superscript k means the k^{th} iteration of regressor R_t in the training. Advantageous to other occlusion detection methods, our method needs no occlusion information because the occlusion-specific regressors do not take the occlusion region into consideration. For instance, to train left-eye occlusion regressor, based on general non-occluded facial images, we only consider the features from all other areas except left eye, no matter left eye is occluded or not. Thus the general face image database is sufficient for training our method. As in SDM, practically four to five linear regression steps are needed to reach the convergence. We learn T regressors $R_1, ..., R_T$, each of which consists of K cascaded single regressor $R_t = \{R_t^1, ..., R_t^K\}$.

3.2 Consensus of Regression on Local Response Maps

Given the multiple landmark predictions resulting from the T cascaded regressors, it is necessary to select which of these is uncorrupted by occluded

features, and thereby determine the optimal landmark positions. To achieve this, we propose a Bayesian inference framework based on the local response maps $\mathcal{M} = \{M_j\}, j = 1...n$. The generation of response map M_j for a landmark is illustrated in Fig. 3. Firstly, a local region is cropped out as shown in Fig. 3, which is formed by bounding the estimated points (denoted as green dots) from all the regressors. For each point inside the local region, its likelihood of being the true landmark is evaluated by support vector regression trained off-line. After all points are calculated, the response map is formed as in Fig. 3.

Given the response maps, our objective function can be probabilistically formulated as Eqn. 2.

$$\arg\max_{s} p(s|\hat{s}_1, \hat{s}_2, ..., \hat{s}_T, \mathcal{M}), \tag{2}$$

where $\hat{s}_1, \hat{s}_2, ..., \hat{s}_T$ denote the shape predictions from the T regressors.

To handle occlusion, we introduce $v_i, i = 1, ..., T$, which is a binary variable that is true if regressor R_i's landmarks are non-occluded. Thus, the probability that the regression result $\hat{s}_i$ approximates the true position can be represented as $p(v_i = 1|\hat{s}_1, \hat{s}_2, ..., \hat{s}_T)$. Suppose there are sufficient such regressors, weighted mean is a straight forward way to estimate the optimum. But this naive method ignores the cue from the response maps. Our Bayesian framework takes the response maps into consideration by computing $p(s|v_i, \mathcal{M})$. Consequently, Eqn. 2 can be rewritten as:

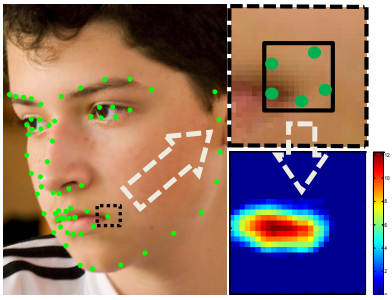


Fig. 3. Illustration of response map

$$p(s|\hat{s}_1, \hat{s}_2, ..., \hat{s}_t, \mathcal{M}) = \sum_{i=1}^{T} \sum_{v_i=\{0,1\}} p(s|v_i, \hat{s}_i, \mathcal{M}) p(v_i|\hat{s}_1, \hat{s}_2, ..., \hat{s}_T), \tag{3}$$

where the second term models the deviation of regressor R_i's output from the majority, which can be expressed as:

$$p(v_i = 1|\hat{s}_1, \hat{s}_2, ..., \hat{s}_T) = exp(-\eta \|\hat{s}_i - \bar{s}\|_2^2). \tag{4}$$

In the above model, we define $\bar{s}$ as the reference shape, which is obtained by an iterative outlier removal and averaging algorithm based on the T observations. The goal is to compute a robust mean while excluding the effect of outliers caused by none-compatible regressors.

Given conditional independence assumption of individual landmarks, the shape alignment probability (the first term) in the objective function can be modeled as:

$$p(s|v_i, \hat{s}_i, \mathcal{M}) = \prod_{j=1}^{n} p(x_j|\hat{x}_j^i, \mathcal{M}) \tag{5}$$

where $s = (x_1, x_2, ..., x_n)$ and $\hat{s}_i = (\hat{x}_1^i, \hat{x}_2^i, ..., \hat{x}_n^i)$, x_j denotes a landmark prediction and $\hat{x}_j^i$ denotes a landmark observation.

Each landmark's alignment probability can be modeled as a response map update problem:

$$p(x_j|\hat{x}_j^i, \mathcal{M}) = \sum_{y \in \phi \subset \mathcal{M}} p(x_j|y)p(y|\hat{x}_j^i). \tag{6}$$

Given the current estimate $\hat{x}_j^i$, we consider all the neighboring points y which forms neighborhood $\phi \subset \mathcal{M}$ of $\hat{x}_j^i$ to indicate the alignment likelihood of the next update position x_j. The posterior $p(x_j|y)$ is assumed Gaussian distribution $p(x_j|y) \sim N(x_j; y, \sigma_j I)$.

The probability map $p(y|\hat{x}_j^i)$ is obtained from the response map which is modeled by SVM from training data. Consequently, the response map update can be achieved by fitting a Mixture of Gaussian (MoG) model:

$$p(x_j|\hat{x}_j^i, \mathcal{M}) = \sum_{y \in \phi \subset \mathcal{M}} \gamma_y^i N(x_j; y, \sigma_j I) \tag{7}$$

where $\gamma_y^i = p(y|\hat{x}_j^i)$. The overall objective function now becomes:

$$\arg\max_s \sum_{i=1}^{T} p(v_i|\hat{s}_1, \hat{s}_2, ..., \hat{s}_T) \prod_{j=1}^{n} \sum_{y \in \phi \subset \mathcal{M}} \gamma_y^i N(x_j; y, \sigma_j I) \tag{8}$$

For optimization, we take an alternating scheme: fixing $p(x_k|\hat{x}_k^i, \mathcal{M})$ for all landmarks $k \neq j$, and optimize for the j^{th} landmark via the Expectation Maximization (EM) algorithm. We can iterate this alternating process multiple times until convergence.

3.3 Occlusion Inference

Compared to fully visible facial images, occluded faces are with one or several facial parts that are sheltered by obstacles. As we know, occlusion of landmarks is highly pose-dependent. The same landmark with different head poses may have different appearance. The head pose can be inferred by Procrustes Analysis over the predicted landmarks and the 3D reference face shape [29]. Then our inference process starts with classifying each landmark as occluded or non-occluded under different poses. By extracting pyramid SIFT descriptor $h(x)$, a standard linear SVM framework is applied to provide the detection score, $f(h(x)) = \omega^T h(x) + \beta$. In the training part, well-aligned landmark appearance and occluded appearance are collected with respect to three head poses, left head pose $(-45°, -15°)$, near-frontal head pose $(-15°, 15°)$ and right head pose $(15°, 45°)$. The testing examines the head pose first and apply the pose-dependent occlusion classifier.

Usually the classification might be sensitive and not consistent among landmarks. But we can obtain some detections with high confidence. These highly confident detections are labeled with occlusion state labels. We use a graph-based

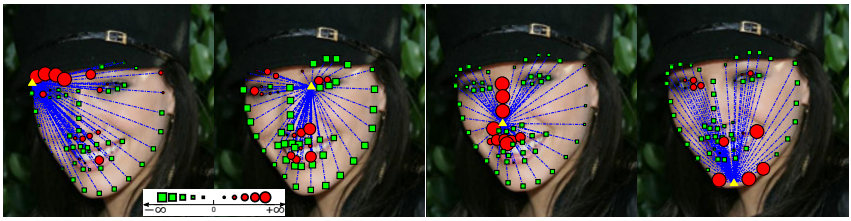


Fig. 4. Visualization of weights for label propagation. The size of a landmark is proportional to its weight. Yellow triangle is the central landmark being processed. Red landmarks are with positive weights which are similar to the central landmark while green landmarks are with negative weights which are dissimilar to the central landmark.

method to jointly infer the occlusion status for all the landmarks. Motivated by the work from Zhu et al. [32], assuming there are m labeled points $x_1, ..., x_m$, and $n-m$ unlabeled points $x_{m+1}, ..., x_n$, which constitutes the node set V. All those points are fully connected, which forms the edge set E. The weights between edges are defined by Eqn. 9.

$$w_{ij} = exp\left(-\|x_i - x_j\|^2_{\Sigma_d^{-1}} - \lambda\|h(x_i) - h(x_j)\|^2_{\Sigma_h^{-1}}\right) \tag{9}$$

The first term in the exponential represents the spatial distance, Σ_d is the covariance matrix of all the landmark positions. The second term measures the similarity of feature vectors, h denotes the feature extractor and Σ_h is the covariance matrix of all the features. λ is a balancing factor between the two terms. The similarity between different landmarks is visualized in Fig. 4.

Given such graph $G = (V, E)$, with the edges defined by the weights w_{ij}, the task becomes a label propagation problem on Graph G. By assuming the joint probability of the graph nodes a Gaussian distribution, we can use the closed form solution in [32] to predict occlusion confidence for all the landmarks jointly.

4 Results and Discussions

Our method is mainly focused on facial landmark localization under both non-occluded and occluded conditions. We evaluate our method on two challenging benchmarks, non-occluded images in Labeled Facial Parts in the Wild (LFPW) [2] and Helen facial feature database [14]. Moreover, we evaluate occluded images from both LFPW and Helen databases, denoted as LFPW-O and Helen-O. Together with Caltech Occluded Faces in the Wild (COFW) [3], we evaluate our method on the three occlusion datasets and compare with several state-of-the-art algorithms. We also evaluate the occlusion detection performance on COFW and compare it to [3].

4.1 Experimental Setup

In the experiments, we use the 66 points annotation from 300 Faces in-the-Wild challenge [20] for both training and testing, omitting two inner mouth

corner points. The annotation is consistent across different databases, e.g. LFPW and Helen. Since COFW uses the 29 points annotation same as the original annotation of LFPW, when evaluating on COFW, we use the overlapped 19 points which are defined by both 66 points annotation and 29 points annotation.

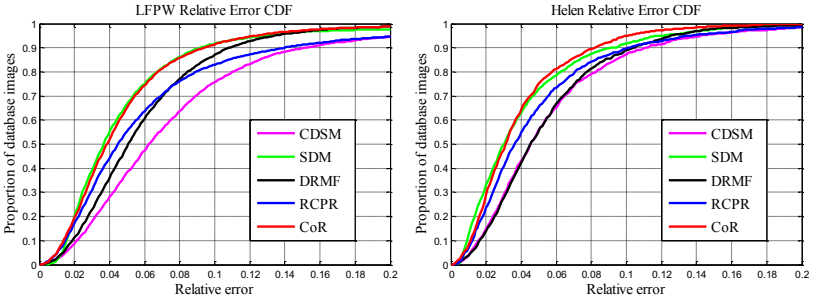
LFPW consists of face images under wild conditions. The images vary significantly in pose, illumination and occlusion. There are 811 training images and 224 testing images in this database. We selected all occluded images, which is 112 out of 224 testing images to form LFPW-O. Helen is another wild face database, consisting of faces under all kinds of natural conditions, both indoor and outdoor. Most of the images are of high resolution. The training set contains 2000 images and testing set contains 330 images. We randomly selected 290 occluded face images out of 2330 images to form Helen-O. In the training of our regressors, we select 402 Helen training images which are not included in Helen-O and 468 LFPW training images.

We compare our method Consensus of Regression (*CoR*) with 4 state-of-the-art methods, Supervised Descent Method (SDM) [26], Robust Cascaded Pose Regression (RCPR) [3], Discriminative Response Map Fitting (DRMF) [1] and Optimized Part Mixture with Cascaded Deformable Shape Model (CDSM) [29]. These methods report the top performance among the literature. SDM and RCPR are non-parametric methods while DRMF and CDSM are parametric methods. The codes used for this experiments are downloaded from internet provided by the authors. The DRMF and CDSM are 66 points annotation. RCPR's annotation is flexible since it provides the training code in which the annotation can be defined by users. To compare on Helen and LFPW, we re-trained RCPR model with the same training set which we used to train our occlusion-specific regressors. SDM only provides 49 points annotation, omitting 17 profile and jawline fiducial points. To make the comparison consistent, on LFPW and Helen, we adopt 49 points evaluation over all the methods. On COFW, we adopt the intersected 19 fiducial points which are defined by all the methods.

4.2 Evaluation on Facial Feature Localization

Non-occlusion Datasets: We compare the alignment accuracy on non-occluded images from LFPW and Helen databases with 4 state-of-the-art methods as shown in Fig. 5. The measurement is Cumulative Distribution Function (CDF).

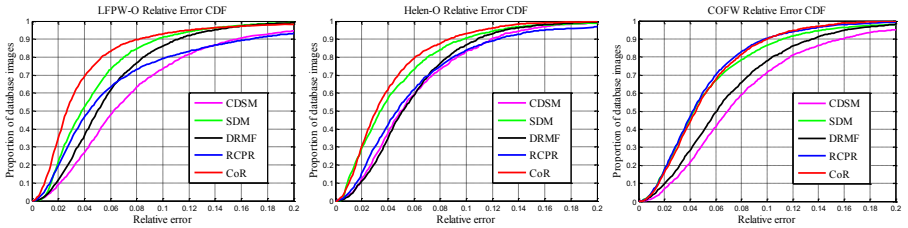
Almost all methods encounter failure during testing. It may be from the failure of face detection, improper initialization and the algorithm itself. For fairness, we compare on the images that encounter no failure by all the methods. In Fig. 5 (a), SDM and *CoR* (the proposed method) perform almost the same, which significantly outperform other methods with at least 10% proportion gap. In Fig. 5 (b), the proposed *CoR* method achieves better results than all other methods. Nevertheless, considering the failure cases, besides the face detection failure, our method achieves 9.7% and 33.3% failure rate on LFPW and Helen while SDM achieves 10.2% and 36.6% respectively. The non-occlusion evaluation over LFPW and Helen demonstrates that *CoR* is among the top level while marginally better than those methods.



(a) LFPW (b) Helen

Fig. 5. Relative error Cumulative Distribution Function curves for landmark localization on LFPW and Helen (non-occlusion images), comparing the proposed method *CoR* in Red curve with other state-of-the-art methods. (a) Error cumulative distribution tested on LFPW database. (b) Error cumulative distribution tested on Helen database.

Occlusion Datasets: When evaluating on occluded faces, traditional methods may have problems, i.e. SDM extracts every landmark's local appearance information for regression. The occluded landmarks' appearance which brings in error degrades the regression results significantly. We compare all the methods on the LFPW-O, Helen-O and COFW in Fig. 6.



(a) LFPW-O (b) Helen-O (c) COFW

Fig. 6. Relative error Cumulative Distribution Function curves for landmark localization on LFPW-O, Helen-O and COFW, comparing the proposed method *CoR* in Red curve with other state-of-the-art methods. (a) Error cumulative distribution tested on all occluded images from LFPW database. (b) Error cumulative distribution tested on occluded images selected from Helen database. (c) Error cumulative distribution tested on COFW database.

From all the plots, our method accomplishes significantly better accuracy than the rest of the methods especially on LFPW-O and Helen-O. For the COFW dataset, our method approaches the performance of RCPR and is significantly better than other methods. The RCPR result on COFW is trained based on COFW. But our method is trained on part of LFPW and Helen images. When

RCPR is trained on the same training set part of LFPW and Helen, the performance of RCPR on Helen, LFPW as well as Helen-O and LFPW-O is not as good as our method. Compared to non-occlusion results, the margin between the proposed method and SDM is larger when evaluating on LFPW-O and Helen-O. It is because our method is particularly designed with occlusion-specific regressors which shows the effectiveness in handling occlusion.

Quantitative results are evaluated in terms of Average RMSE in Table 1. *CoR* provides the most consistent and accurate performance against other methods on Helen, Helen-O and LFPW-O. It is very competitive to the state-of-the-arts on LFPW and COFW. As we know, the profile and jawline parts suffer the largest variance in face shape. The 49-point annotation in SDM omits the profile and jawline, which imports less variance. While in our method, we consider the profile and jawline and simultaneously optimize all the facial components, which needs to overcome more regression variance than SDM. Even so, *CoR* achieves the same while sometimes better performance than SDM.

Table 1. Average Root Mean Square Error (in pixels) of CDSM, DRMF, RCPR, SDM and proposed method *CoR* on LFPW, Helen, LFPW-O, Helen-O and COFW databases

Method	LFPW	Helen	LFPW-O	Helen-O	COFW
CDSM	6.33	9.57	5.81	10.28	5.17
DRMF	4.90	9.59	5.40	10.23	4.50
RCPR	5.49	8.75	6.32	10.62	**3.38**
SDM	**3.84**	8.16	4.62	8.93	3.80
CoR	3.96	**7.23**	**3.49**	**7.18**	3.51

4.3 Evaluation on *CoR* Framework

In this section, we investigate the effectiveness of the proposed *CoR* framework. We look into the comparison of *CoR*, wm-agg and gm-agg. In Table 2, wm-agg represents the weighted mean aggregation over all T regressors and gm-agg represents geometric mean over all regressors. The table shows that *CoR* consistently outperforms wm-agg and gm-agg with a significant margin, which

Table 2. Average Root Mean Square Error (in pixels) of *CoR*, weighted mean aggregation (wm-agg) and geometric mean aggregation (gm-agg) methods on LFPW, Helen, LFPW-O, Helen-O and COFW databases.

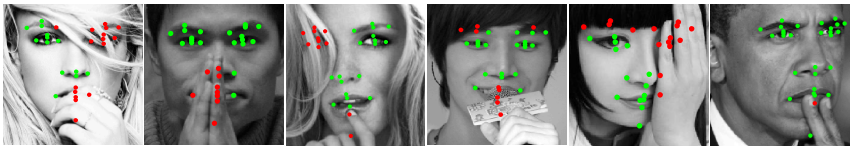
Method	LFPW	Helen	LFPW-O	Helen-O	COFW
CoR	**3.96**	**7.23**	**3.49**	**7.18**	**3.51**
wm-agg	4.32	7.34	3.61	7.41	3.63
gm-agg	4.43	7.66	3.65	7.53	3.66

indicates that the Bayesian consensus of regression scheme is a more robust and effective way in optimizing the positions.

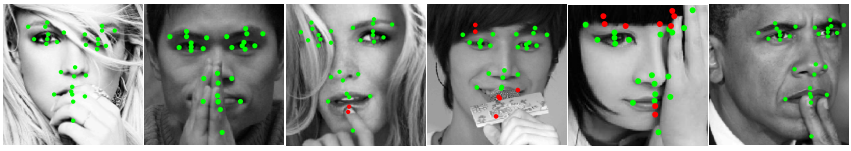
4.4 Evaluation on Occlusion Detection

Among the previous methods, only RCPR detects occlusion. Thus, we compare the performance of occlusion detection with RCPR. Since other databases do not provide occlusion ground truth, we only focus on COFW for evaluation. For RCPR, as the code published by the authors, we do not tune any parameter and simply use the default settings. In our method, we also fix the parameters for testing. The parameters are tuned via 3-fold cross validation. Fig. 7 shows some visual results on occlusion detection. Compared to ground truth, the RCPR results seem to miss out many occluded landmarks while our method hit more occluded ones.

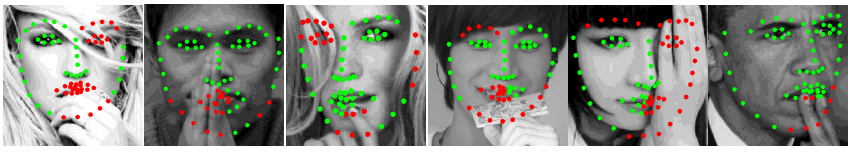
Quantitatively, by holding the false alarm at the same level, our method achieves 41.44% accuracy while RCPR is with 34.16%. Since the annotations of the two methods are different, if we count the component occlusion condition in which the component is labeled occluded if at least one landmark in a component is occluded,(landmarks are categorized into 7 components, left/right eyebrow, left/right eye, nose, mouth and chin), our method is with 47.18% and



(a) Ground truth landmarks of COFW



(b) Localization and occlusion detection result by RCPR



(c) Localization and occlusion detection result by proposed *CoR*

Fig. 7. Occlusion detection comparison of *CoR* and RCPR on COFW database (Red dots: occlusion, green dots: non-occlusion). (a) The first row shows ground truth from COFW. (b) The second row shows the results of RCPR with default parameters. (c) The third row shows the results of proposed *CoR* method.

RCPR is with 37.43%, which reveals that our method improves the detection precision by about 10%.

5 Conclusions

We proposed a new consensus of regression based approach which trains an ensemble of occlusion-specific regressors to handle occluded faces in the wild. Due to the non-existence of occlusion priors, we conduct the consensus of the occlusion-specific regressors under a Bayesian framework to optimize the inference. A graph-based semi-supervised learning is also utilized to explicitly detect the occlusion. Our method shows consistent improvement of facial feature localization on both non-occlusion and occlusion face databases. Additionally, our method demonstrates improvement on occlusion detection compared to the state-of-the-art.

References

1. Asthana, A., Cheng, S., Zafeiriou, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: CVPR (2013)
2. Belhumeur, P., Jacobs, D., Kriegman, D., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: CVPR (2011)
3. Burgos-Artizzu, X., Perona, P., Dollar, P.: Robust face landmark estimation under occlusion. In: ICCV (2013)
4. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: CVPR (2012)
5. Cheng, X., Sridharan, S., Saragih, J., Lucey, S.: Rank minimization across appearance and shape for aam ensemble fitting. In: ICCV (2013)
6. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
7. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active shape models-their training and application. Computer Vision and Image Understanding 61(1), 38–59 (1995)
8. Cootes, T.F., Ionita, M.C., Lindner, C., Sauer, P.: Robust and accurate shape model fitting using random forest regression voting. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part VII. LNCS, vol. 7578, pp. 278–291. Springer, Heidelberg (2012)
9. Cristinacce, D., Cootes, T.: Automatic feature localization with constrained local models. Pattern Recognition 41(10), 3054–3067 (2007)
10. Cristinacce, D., Cootes, T.: Boosted regression active shape models. In: BMVC (2007)
11. Dantone, M., Gall, J., Fanelli, G., Gool, L.: Real-time facial feature detection using conditional regression forests. In: CVPR (2012)
12. Dollar, P., Welinder, P., Perona, P.: Cascaded pose regression. In: CVPR (2010)
13. Huang, Y., Liu, Q., Metaxas, D.: A component based deformable model for generalized face alignment. In: ICCV (2007)
14. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 679–692. Springer, Heidelberg (2012)

15. Liang, L., Xiao, R., Wen, F., Sun, J.: Face alignment via component-based discriminative search. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 72–85. Springer, Heidelberg (2008)
16. Martinez, B., Valstar, M., Binefa, X., Pantic, M.: Local evidence aggregation for regression-based faical point detection. IEEE Transactions on Pattern Analysis and Machine Intelligence 35(5), 1149–1163 (2013)
17. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision 60(2), 135–164 (2004)
18. Rivera, S., Martinez, A.: Learning deformable shape manifolds. Pattern Recognition 45(4), 1792–1801 (2012)
19. Roh, M., Oguri, T., Kanade, T.: Face alignment robust to occlusion. In: Automatic Face and Gesture Recognition (2011)
20. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: ICCV Workshop (2013)
21. Saragih, J., Lucey, S., Cohn, J.: Deformable model fitting by regularized landmark mean-shift. International Journal of Computer Vision 91(2), 200–215 (2011)
22. Smith, B.M., Zhang, L.: Joint face alignment with non-parametric shape models. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 43–56. Springer, Heidelberg (2012)
23. Tzimiropoulos, G., Pantic, M.: Optimization problems for fast aam fitting in-the-wild. In: ICCV (2013)
24. Valstar, M., Martinez, B., Binefa, X., Pantic, M.: Facial point detection using boosted regression and graph models. In: CVPR (2010)
25. Wang, X., Han, T., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: ICCV (2011)
26. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: CVPR (2013)
27. Yang, F., Huang, J., Metaxas, D.N.: Sparse shape registration for occluded facial feature localization. In: Automatic Face and Gesture Recognition (2011)
28. Yang, H., Patras, I.: Sieving regression forest votes for facial feature detection in the wild. In: ICCV (2013)
29. Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N.: Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: ICCV (2013)
30. Yu, X., Yang, F., Huang, J., Metaxas, D.N.: Explicit occlusion detection based deformable fitting for facial landmark localization. In: Automatic Face and Gesture Recognition (2013)
31. Zhou, F., Brandt, J., Lin, Z.: Exemplar-based graph matching for robust facial landmark localization. In: ICCV (2013)
32. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: ICML (2003)
33. Zhu, X., Ramanan, D.: Face detection, pose estimation and landmark localization in the wild. In: CVPR (2012)