Nonlinear Hierarchical Part-Based Regression for Unconstrained Face Alignment

Xiang Yu[†], Zhe Lin[‡], Shaoting Zhang[♯] and Dimitris N. Metaxas[§]
[†]NEC Laboratories America, Media Analytics
[‡]Adobe Research
[♯]University of North Carolina at Charlotte
[§]Rutgers, The State University of New Jersey
xiangyu@nec-labs.com, zlin@adobe.com, szhang16@uncc.edu, dnm@cs.rutgers.edu

Abstract

Non-linear regression is a fundamental and yet under-developing methodology in solving many problems in Artificial Intelligence. The canonical control and predictions mostly utilize linear models or multi-linear models. However, due to the high non-linearity of the systems, those linear prediction models cannot fully cover the complexity of the problems. In this paper, we propose a robust two-stage hierarchical regression approach, to solve a popular Human-Computer Interaction, the unconstrained face-in-the-wild keypoint detection problem for computers. The environment is the still images, videos and live camera streams from machine vision. We firstly propose a holistic regression model to initialize the face fiducial points under different head pose assumptions. Second, to reduce local shape variance, a hierarchical part-based regression method is further proposed to refine the global regression output. Experiments on several challenging faces-in-the-wild datasets demonstrate the consistently better accuracy of our method, when compared to the state-of-the-art.

1 Introduction

Face fiducial feature detection is a fundamental step in many Human-Computer Interaction (HCI) applications, such as face recognition, human emotion analysis, autonomous driving, etc. Many different algorithms have been proposed and have shown promising results both in accuracy and speed [Xiong and la Torre, 2013; Belhumeur *et al.*, 2011; Cootes *et al.*, 2012; Sagonas *et al.*, 2013]. They aim towards not only the near-frontal faces but also faces in the wild. However, due to large head pose variation, various types of occlusions, unpredictable illumination and some other factors, the fiducial point (landmark) localization task still remains challenging.

Early representative work, such as the Active Shape Model (ASM) [Cootes *et al.*, 1995; 1998; Matthews and Baker, 2004], uses a parametric model to represent a set of the face fiducial points and proposed an iterative framework for optimizing the fiducial positions. Following these efforts, researchers have attempted to build more robust and sophisticated models which can be robust to different types of interfering conditions, such as pose and expression variations exemplified in Fig. 1. The multi-view deformable part model [Zhu and Ramanan, 2012; Ghiasi and Fowlkes, 2014] alleviates the pose problem. However, discrete pose intervals and rigid shape modeling make it difficult to capture all possible facial variations. Recently, regression based methods [Cao *et al.*, 2012; Xiong and la Torre, 2013; Ren *et al.*, 2014] report accuracy approaching the human labeling level and their typical runtime can be within several milliseconds. However, the regression based methods may significantly suffer from uncommon part appearance due to extreme poses, lighting or expression. Noisy appearance results in bad features and the mapped landmark displacement is disturbed. On the other hand, faces in unconstrained environments are often affected by occlusion as shown in Fig. 1. Hence, occlusion handling becomes crucial for improving these methods. An ensemble of a set of occlusion-resistant regressors relying on the probabilistic inferring is proposed to implicitly overcome occlusion [Yu *et al.*, 2014]. Explicitly detecting the occlusion and incorporating the occlusion information into landmark

Figure 1: Results of our method on unconstrained face images with pose variations and occlusion. Detected occlusion landmarks are denoted in red dots and non-occluded landmarks are denoted in green dots.

detection also show some satisfactory [Ghiasi and Fowlkes, 2014]. Yet the interaction between the occlusion detection and landmark localization is under investigation.

Another important aspect for the regression-based methods is the number of iterations. Usually there is no sophisticated rule to set the "ideal" number of iterations. An empirical choice for the number of iterations may cause the regressor to under-fit the data or drift away from the correct solution. Establishing an online evaluation strategy would help to indicate how well the landmarks are localized by the current stage. Furthermore, if the occlusion information can be simultaneously retrieved with the localization process, the framework would be more unified and efficient.

In this paper, we propose a two-stage framework consisting of a pose-dependent holistic regression model and a hierarchical part-based regression model to robustly localize the face fiducial points. Faces with different poses are fed to different sets of pose-dependent regressors. Consequently, the shape variation inside each set is largely reduced. From the holistic regressors, the hierarchical part regressors are automatically learned by our proposed projection optimization algorithm. Based upon the hierarchical part-based structure, the alignment likelihood is firstly evaluated to determine whether further local regressions are needed and also to estimate the occlusion information simultaneously; then the hierarchical part-based regression models are applied to corresponding parts to further refine the landmarks. Occlusion status are propagated to all the landmarks from the previous occlusion detection during the part-based regression.

2 Related Work

Numerous methods have been proposed in the facial feature localization literature, e.g. deformable part model [Cristinacce and Cootes, 2007a; Huang *et al.*, 2007b; Saragih and Goecke, 2007; Saragih *et al.*, 2011; Tzimiropoulos and Pantic, 2013; Cheng *et al.*, 2013; Yu *et al.*, 2013; Medina and Zafeiriou, 2014; Xing *et al.*, 2014], regression [Liang *et al.*, 2008; Cao *et al.*, 2012; Xiong and la Torre, 2013; Ren *et al.*, 2014; Zhu *et al.*, 2015], convolutional neural network [Zhang *et al.*, 2014b; 2014a], etc.

To overcome pose variation, multi-view shape models [Cootes and Taylor, 1997; Zhu and Ramanan, 2012] were proposed either by local search to estimate the head pose or by combining models from different view-points. The regression-based methods can also handle certain pose variations, which are incorporated into the training data. However, too much pose variation increases the training complexity. Cascaded pose regression [Dollar *et al.*, 2010; Burgos-Artizzu *et al.*, 2013] and conditional regression forests [Dantone *et al.*, 2012] are the most similar works to ours. The former ones take pose as an explicit factor to regress, while ours treats the pose as a conditional hidden state. The latter one partitions the poses into subspaces before the regression fitting. Within each subspace, they aggregate many regression trees to predict landmarks, while in our method, we allow the pose state to change during each step of regressions.

A number of regression constructions have since been proposed, such as boosted regressions [Cristinacce and Cootes, 2007b; Valstar *et al.*, 2010; Martinez *et al.*, 2012], regression forests [Dantone *et al.*, 2012; Yang and Patras, 2013; Kazemi and Sullivan, 2014], linear regressions [Xiong and la Torre, 2013; Dollar *et al.*, 2010; Asthana *et al.*, 2014], regression ferns [Cao *et al.*, 2012; Burgos-Artizzu *et al.*, 2013], etc. Regression-based methods are fast but are sensitive to occlusion. There have been several works introduced for handling occlusion, for example, Artizzu et al. [Burgos-Artizzu *et al.*, 2013] proposed a block-wise statistical model to approximate the occlusion. Yu et al. [Yu *et al.*, 2014] introduced multiple regressors which are specially designed to infer the particular occlusions. A similar work [Ghiasi and Fowlkes, 2014] also used a hierarchical deformable part model to localize landmarks. This method is similar to [Zhu and Ramanan, 2012] in which it sets up multi-view shape models and adopts detection based strategies to vote for the positions. To infer the occlusion status, both [Ghiasi and Fowlkes, 2014] and ours use part-based models. But for alignment, instead of detection of facial features in a pictorial structure in [Ghiasi and Fowlkes, 2014], we model both holistic and local landmark update as a regression based strategy and jointly learn the part regressors from holistic regressors in a projection optimization framework.

3 Our Approach

we propose a hierarchical regression method with pose-dependent and part based modeling. As shown in Fig. 2, we first propose a conditional cascaded regression model to separate the regression manifold into several subspaces. Then a hierarchical part-based model is proposed to decompose the holistic structure into a more flexible part-based hierarchical structure. Each part represents a facial component, e.g. mouth. By inferring the occlusion conditions for the parts, the landmarks' update model is regularized.

3.1 Preliminary

Given the definition of N facial feature points, denoted as $\mathbf{s} = [\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N]$, and their starting position $\mathbf{s}_0$, the goal is to minimize the squared error in coordinates $\|(\mathbf{s}_0 + \Delta\mathbf{s}) - \mathbf{s}^*\|_2$, where $\mathbf{s}^*$ is the ground truth. The evidence we could observe is only the appearance feature. Thus, Eq. 1 minimizes the error in feature space instead of the coordinate space.

$$\arg\min_{\Delta\mathbf{s}} \|\mathbf{h}(I(\mathbf{s}_0 + \Delta\mathbf{s})) - \mathbf{h}(I(\mathbf{s}^*))\|_2^2 \tag{1}$$

$\mathbf{h}$ is the feature descriptor, i.e. SIFT feature [Lowe, 2004]. $I(s)$ are the facial image patches surrounding each fiducial point of $\mathbf{s}$. $\mathbf{h}(I(s))$ is the concatenated feature descriptor applied on each of the image patches $I(s)$. The cascaded regression based framework [Xiong and la Torre, 2013] has the shape update form as follows:

$$\mathbf{s}_{t+1} = \mathbf{s}_t + \mathbf{R}_t\phi_t + \mathbf{b}_t \tag{2}$$

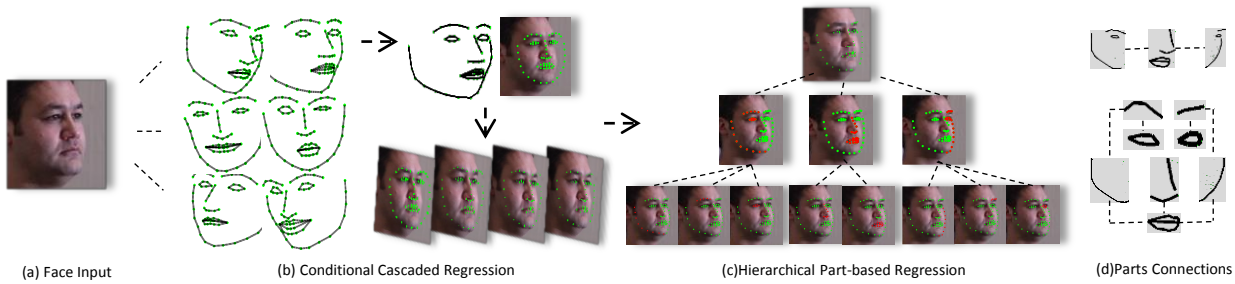


Figure 2: Graphical structure illustration of the proposed framework. (a) The input face image. (b) Conditioned by head poses, the face key points are initialized with different priors and a cascaded regression is applied as global shape fitting. (c) The holistic shape is split into parts hierarchically to effectively overcome the local shape variance, e.g. the shape is firstly divided into left part (left profile, left eyebrow and left eye), middle part (nose and mouth) and right part (right profile, right eyebrow and right eye). The second layer is derived from the first layer by further dividing the components. (d) The geometric connections of the two layer parts defined in (c).

where $\mathbf{R}_t$ is the regression matrix and $\mathbf{b}_t$ is the intercept. $\phi_t = \mathbf{h}(I(\mathbf{s}_t))$ denotes a local feature descriptor all through the work. Typically the number of iterations is fixed to 4 or 5. The cascaded regression attempts to apply a set of linear regressions sequentially to predict landmark positions. Given ground truth $\mathbf{s}^*$, the training process is to minimize the prediction error over all training samples $\mathcal{T}$ as follows:

$$\arg\min_{\mathbf{R}_t,\mathbf{b}_t} \sum_{z\in\mathcal{T}} \|\mathbf{s}^* - (\mathbf{s}_t + \mathbf{R}_t\phi_{z,\mathbf{s}_t} + \mathbf{b}_t)\|_2^2 \qquad (3)$$

3.2 Conditional Cascaded Regression

To reduce the complexity of the face shape manifold, we divide the manifold into several subspaces, as shown in Fig. 2 (b). Shapes are mainly clustered into three groups, the frontal view, the left view and the right view. We set the threshold angles to be $-22.5°$ and $22.5°$. Then given image I, by introducing head pose parameter θ, the regression problem becomes equivalent to solving Eq. 4.

$$\arg\max_{\Delta\mathbf{s},\theta\in\Theta} \; p(\Delta\mathbf{s}|I) = \frac{1}{\Xi} p(\Delta\mathbf{s}|\theta, I)p(\theta|I) \qquad (4)$$

where Θ is the set of discrete head pose intervals, Ξ is the distribution normalizer and the pose likelihood term is learned based on the logistic regression framework:

$$p(\theta|I) = \frac{1}{\Phi} \frac{\exp(w_\theta\psi + c_\theta)}{1 + \exp(w_\theta\psi + c_\theta)} \qquad (5)$$

where ψ is a holistic appearance feature, i.e. HoG and Φ is a normalization factor to make $p(\theta|I)$ a distribution.

The conditional alignment likelihood $p(\Delta\mathbf{s}|\theta, I)$ is modeled by the coordinate displacement in Eq. 6,

$$p(\Delta\mathbf{s}|\theta, I) = \frac{1}{\Gamma} \exp(-\beta\|\mathbf{R}_t(\theta)\phi + \mathbf{b}_t(\theta)\|_2) \qquad (6)$$

where Γ is again a normalization factor. Notice that $\Delta\mathbf{s} = \mathbf{R}_t(\theta)\phi + \mathbf{b}_t(\theta)$. We assume the alignment likelihood $p(\Delta\mathbf{s}|\theta, I)$ follows the exponential distribution. At each regression iteration, we maximize the alignment likelihood in Eq. 4 by conditioning on different head poses. The corresponding holistic regressors are applied to update the landmark positions. Such procedure largely reduces the shape complexity caused by head poses and are more likely to converge.

3.3 Hierarchical Part-based Regression

Holistic regression is effective in aligning the face as a whole, but it may not produce perfect fitting results at local parts due to appearance and shape deformation. For instance, assuming the same face with eyes fully open and half open, holistic regressions localize landmarks as a whole but may fail in the eye region due to the lack of local constraint from the holistic regression. A part-based regression step could alleviate this problem more by deformable local fitting.

Part-based Local Regression

From the holistic regression, each landmark's update utilizes exactly one row of R. By dividing the facial area into parts, we partition the regression matrix $\mathbf{R}$ into row-wise blocks. Recall from Fig. 2 (c), for the first layer, we divide the shape into left, middle and right parts. As shown in Eq. 7, the partition of $\mathbf{R}$ is denoted as $\mathbf{R} = [\mathbf{R}_l, \mathbf{R}_m, \mathbf{R}_r]^T$, where $\mathbf{R}_l, \mathbf{R}_m$ and $\mathbf{R}_r$ correspond to left, middle and right parts, respectively. Such division is recursively applied by further partitioning the previous layer's blocks into smaller units. Fig. 2 (c) shows the second layer of facial components, i.e. left profile, mouth, left eye, etc.

Notice that the partition still uses the holistic feature ϕ for update $\Delta\mathbf{s} = [\mathbf{R}_l, \mathbf{R}_m, \mathbf{R}_r]^T\phi$. In other words, inside the regressors $\mathbf{R}_l, \mathbf{R}_m, \mathbf{R}_r$ themselves, the correlation in between should be diminished. We aim to obtain local regressors from the holistic regressor $\mathbf{R}$ as shown in Eq. 7.

$$\mathbf{R} = \begin{bmatrix} \mathbf{R}_l \\ \mathbf{R}_m \\ \mathbf{R}_r \end{bmatrix} \rightarrow \hat{\mathbf{R}} = \begin{bmatrix} \hat{\mathbf{R}}_l & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \hat{\mathbf{R}}_m & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \hat{\mathbf{R}}_r \end{bmatrix} \qquad (7)$$

where each part regressor $\hat{\mathbf{R}}_i, i = l, m, r$ is a block-wise sub-matrix. The transformation optimization from $\mathbf{R}_i$ to $\hat{\mathbf{R}}_\mathbf{i}$ is introduced in Sec. 3.4. After generating the local part regressors directly from the holistic ones, the part-based

regression confirms to the same update rule, $\Delta \mathbf{s}_i = \hat{\mathbf{R}}_i \phi + \hat{\mathbf{b}}_i, i = l, r, m$.

Localization Evaluation

To determine when to halt the holistic and local regressions, we set up an evaluation function to validate the alignment. The function propagates each part's (a.k.a component's) alignment score to the upper layer and finally generate the overall alignment score. Given the k^{th} component $\mathcal{G}_k$ and its landmarks $\mathbf{s_i} \in \mathcal{G}_k$, the part score function can be defined as:

$$\mathbb{E}(I, \mathcal{G}_k) = \sum_i \mathbf{U}(I, \mathbf{s}_i) + \sum_{i,j} \mathbf{Q}(\mathbf{s_i}, \mathbf{s_j}), \mathbf{s_i}, \mathbf{s_j} \in \mathcal{G}_k \quad (8)$$

where $\mathbf{U}(I, \mathbf{s}_i)$ is the unary term defined in Eq. 9, i.e. the inner product of the feature and its corresponding weights. $\phi(I, \mathbf{s}_i)$ is the descriptor extracted at landmark $\mathbf{s}_i$. The relationship between ϕ appeared in the previous sections as descriptor and $\phi(I, \mathbf{s}_i)$ is $\phi = [\phi(I, \mathbf{s}_1), ..., \phi(I, \mathbf{s}_N)]$.

$$\begin{aligned} \mathbf{U}(I, s_i) &=< \boldsymbol{w}_i^u, \phi(I, s_i) > \\ \mathbf{Q}(\mathbf{s}_i, \mathbf{s}_j) &=< \boldsymbol{w}_{i,j}^b, \mathbf{q}(\mathbf{s}_i, \mathbf{s}_j) > \end{aligned} \quad (9)$$

The second pair-wise term $\mathbf{Q}(\mathbf{s}_i, \mathbf{s}_j)$ is defined as the geometric smoothness term of two landmarks in one component, i.e. $\mathbf{q}(\mathbf{s}_i, \mathbf{s}_j) = [|\mathbf{s}_i - \mathbf{s}_j|, \Delta|\mathbf{s}_i - \mathbf{s}_j|], \mathbf{s}_i, \mathbf{s}_j \in \mathcal{G}_k$, which is independent from the image I, the landmark's alignment likelihood and occlusion condition.

Occlusion Regularization

We then independently train a classifier $\mathcal{O}_i = \boldsymbol{w}_{o,i} \phi(I, \mathbf{s}_i) + c_{o,i}$ to provide the occlusion likelihood of each landmark at each regression step. This additional classification step produces little overhead due to the sharing of features for both regression and occlusion detection.

Misalignment is not necessarily caused by occlusion, while occlusion can adversely affect the alignment. Suppose landmark $\mathbf{s}_i$ is occluded. The alignment score $\mathbf{U}(I, \mathbf{s}_i)$ is close to 0, which does not contribute to the overall score. By detecting occlusion of $\mathbf{s}_i$, we can equivalently set the feature at $\mathbf{s}_i$ as $\phi(I, \mathbf{s}_i) = 0$. During the process, the landmarks' occlusion condition is confidently predicted by both the occlusion detector and the alignment score.

All the landmarks' occlusion states are then modulated by the neighboring landmarks with a Markov Random Field. Landmarks with sufficiently small and large scores of $\mathcal{O}_i$ and alignment are selected as negative and positive boundary conditions respectively. By setting up the connection weights among the landmarks, a label propagation algorithm [Zhu *et al.*, 2003] is applied to assign the unlabeled landmarks. Our system framework is illustrated in Algorithm 1.

Algorithm 1 The two-stage regression algorithm.

Input: I, $\mathbf{s}_0$, threshold d
Output: $\mathbf{s}$, $\mathcal{O}$
repeat
  run Eq. 2, Eq. 6 and Eq. 5, optimize Eq. 4.
  evaluate $\mathcal{O}_i = \boldsymbol{w}_{o,i} \phi(I, s_i) + c_{o,i}, i = 1, ..., N$, set $\phi(I, s_{\mathcal{O} \leq 0}) = \mathbf{0}$
until T_1 times
fix θ, evaluate Eq. 8, if $\mathbb{E}(\mathcal{G}) > d$, halt.
repeat
  for layers in hierarchical structure
    run part-based Eq. 2
    evaluate $\mathcal{O}_i = \boldsymbol{w}_{o,i} \phi(I, s_i) + c_{o,i}, i = 1, ..., N$, set $\phi(I, s_{\mathcal{O} \leq 0}) = \mathbf{0}$
    evaluate Eq. 8, if $\mathbb{E}(\mathcal{G}) > d$, halt.
  end
until T_2 times

3.4 Holistic and Part Regression Training

In this section we describe the training of the holistic regressors and how to derive the hierarchical part regressors directly from holistic regressors. We also introduce the training of the graphical model for evaluating the alignment likelihood.

Holistic Regressor Training: We firstly introduce the training details for holistic regressors. In experiments, we tried the gaussian random perturbation of each landmark, even if the perturbation step is small, the regression result returns jittering shapes. When the training samples are not sufficient, we augment the initialization by rotation and random perturbation of global translation. Meanwhile, to prevent overfitting, denoting $\Delta \mathbf{s} = \mathbf{s}^* - \mathbf{s}_0$, we modify Eq. 3 by adding the regularization terms, which is Eq. 10. The problem can be solved by splitting R into row pieces and each piece-wise sub-problem is convex.

$$\min_{\mathbf{R}, \mathbf{b}} \sum_{z \in \mathcal{T}} \sum_{\mathbf{s}_0} \|\Delta \mathbf{s} - \mathbf{R}\phi_z - \mathbf{b}\|_2^2 + \frac{\eta_1}{2} tr(\mathbf{R}\mathbf{R}^T) + \frac{\eta_2}{2} \mathbf{b}^T \mathbf{b} \quad (10)$$

Part-based Regressor Derivation: As we have introduced in Eq. 7, to convert holistic regressors $\mathbf{R}_i$ to part regressors $\hat{\mathbf{R}}_i$, we propose a projection matrix W to accomplish the transformation. The original partitioned regressor is projected onto a new subspace in which the correlation between parts is diminished, as shown in Eq. 11. Projection from holistic regressors is expected to preserve the global information for the local regression. Directly training part regressors only contains local information, which is sensitive to noise. Moreover, bounding the update error provides criterion for automatic update halting.

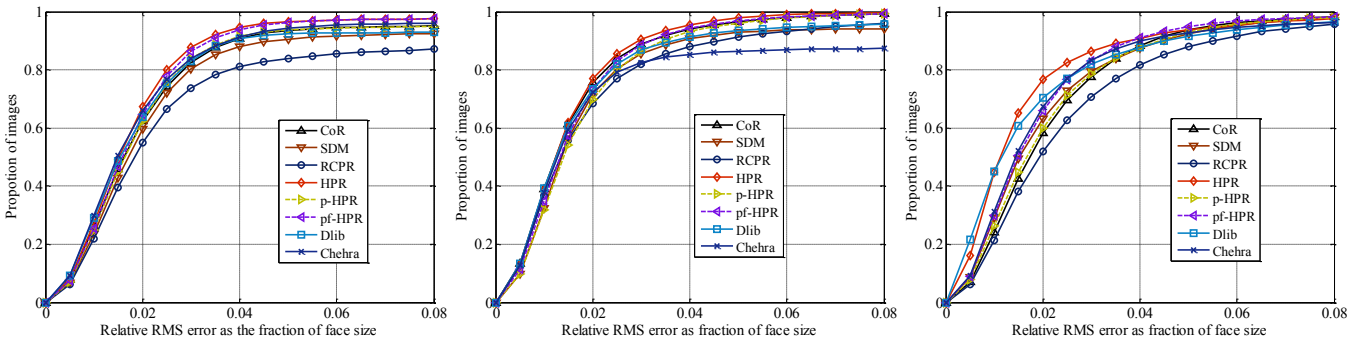
$$\hat{\mathbf{R}}_i = \mathbf{R}_i \mathbf{W}_i, i = l, r, m \quad (11)$$

We expect that after transforming the original $\mathbf{R}$ into block-wise $\hat{\mathbf{R}}$, by bounding the part regression error, the holistic regression error becomes a supreme of the part regression error in Eq. 12.

$$\Delta \mathbf{s} = \mathbf{R}_i \phi + \mathbf{b}_i = \sup_{\mathbf{W}_i} \left\{ [\mathbf{R}_i, \mathbf{b}_i] \mathbf{W}_i \left[\phi^T, 1\right]^T \right\} \quad (12)$$

Thus, it leads to an optimization over $\mathbf{W}_i, i = l, r, m$ such that the local part regression further reduces the update error based on the holistic result. The above optimization problem can be formulated as:

$$\arg\min_{\mathbf{W}_i} \ \|\tilde{\mathbf{R}}_i \mathbf{W}_i \tilde{\phi}\|_2^2 + \|\mathbf{W}_i\|_F^2, \mathbf{W}_i^T \mathbf{W}_i = \mathbb{I} \quad (13)$$



(a) LFW (b) LFPW (c) Helen

Figure 3: Cumulative distribution function curves of normalized error on LFW, LFPW and Helen, comparing the proposed method HPR with other state-of-the-art methods. The horizontal axis is the normalized error and the vertical axis is the image proportion of the volume of database. (a) Error CDF on LFW database. (b) Error CDF on LFPW database. (c) Error CDF on Helen database.

We simplify the notation of $\mathbf{R}_i, \mathbf{b}_i$ as $\tilde{\mathbf{R}}_i = [\mathbf{R}_i, \mathbf{b}_i]$ and the raw feature is rephrased as $\tilde{\phi} = \left[\phi^T, 1\right]^T$. $\tilde{\mathbf{R}}_i \in \mathbb{R}^{(m,n)}, \mathbf{W}_i \in \mathbb{R}^{(n,m)}$, m is the number of landmarks in the corresponding part and n is the original feature dimension plus one dimension of $\mathbf{b}_i$. $\mathbf{W}_i^T \mathbf{W}_i = \mathbb{I}$ constraints that the projection of each part should be orthogonal.

By solving independently each part's transformation matrix $\mathbf{W}_i$, we obtain each local part's regressor $\hat{\mathbf{R}}_\mathbf{i}$ which is guaranteed to further shrink the localization error because the optimization of Eq. 13 is to find the optimal $\mathbf{W}_i$ such that the displacement from the ground truth is minimized from the holistic step. For each part's regressor, a second layer regressor can be achieved under the same construction.

Localization Evaluation Model Training: The weights for score calculation of landmark localization evaluation in Eq. 9 are learned in the following. We first concatenate bottom layer unary weights $\boldsymbol{w}_i^u$ as $\boldsymbol{w}^u$, bottom layer pair-wise weights $\boldsymbol{w}_{i,j}^b$ as $\boldsymbol{w}^b$ and upper layer pair-wise weight $\boldsymbol{w}_{i,j}^g$ as $\boldsymbol{w}^g$. We denote $\mathbf{q}(\mathbf{s}) = [\mathbf{q}(\mathbf{s}_i, \mathbf{s}_j)]$ for all pairs (i, j), in which the pair-wise smoothness features for all landmarks are concatenated. Similarly, $\mathbf{q}(\mathcal{G}) = [\mathbf{q}(\mathcal{G}_i, \mathcal{G}_j)]$ denotes the upper level pair-wise feature for all the parts. Re-arranging all the weights as $\boldsymbol{w} = [\boldsymbol{w}^u, \boldsymbol{w}^b, \boldsymbol{w}^g]$ and all the features as $\mathbf{f} = [\phi, \mathbf{q}(\mathbf{s}), \mathbf{q}(\mathcal{G})]$, the evaluation score is $\boldsymbol{w}^T \mathbf{f}$. We set the loss function as hinge loss, which is the first term in Eq. 14. By regularizing $\boldsymbol{w}$ with l_2 norm, minimizing the loss function leads to solution of $\boldsymbol{w}$ for all the parts.

$$\arg\min_{\boldsymbol{w}} \sum_{\mathbf{f}_i \in \mathcal{C}} max(0, 1 - \alpha \cdot \boldsymbol{w}^T \mathbf{f}_i) + \frac{\lambda}{2} \|\boldsymbol{w}\|_2^2 \qquad (14)$$

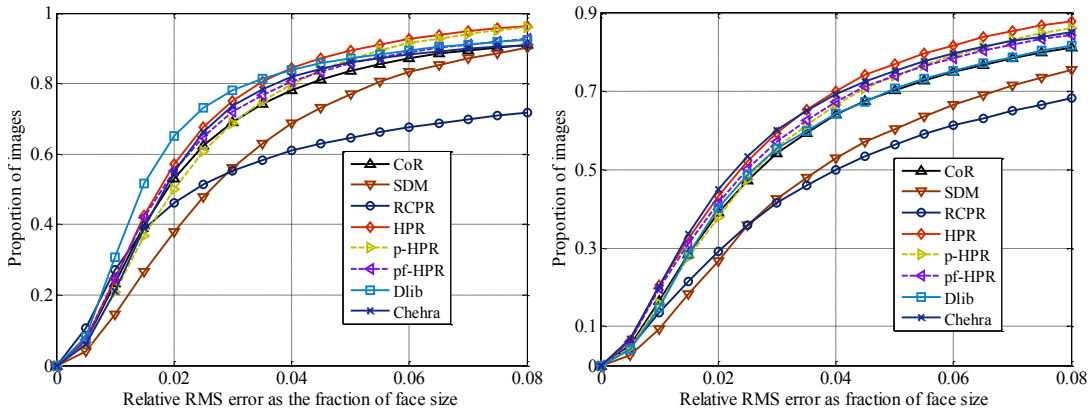
The training set $\mathcal{C}$ includes both positive and negative samples. The positive samples are the facial images with ground truth landmark positions while the negative samples are non-facial images with initialized landmarks or facial images with unaligned landmarks. α is the ground truth label taking 1 if it is positive sample and -1 if the sample is negative. The above problem can be efficiently solved by gradient descent approach.

4 Experiments

We evaluate our method on six challenging benchmarks, i.e., Labeled Faces in the Wild (LFW) [Huang *et al.*, 2007a], Labeled Facial Parts in the Wild (LFPW) [Belhumeur *et al.*, 2011], Helen [Le *et al.*, 2012], Annotated Faces-in-the-Wild (AFW) [Zhu and Ramanan, 2012], iBug [Sagonas *et al.*, 2013] and Caltech Occluded Faces in the Wild (COFW) [Burgos-Artizzu *et al.*, 2013]. To evaluate the localization performance under occlusion, subsets of LFPW and Helen are selected, which are denoted as LFPW-O and Helen-O. During all the experiments, LFPW and Helen refer to the whole datasets. Our occlusion detection method is evaluated on LFPW-O, Helen-O and COFW, with comparison to the state-of-the-art.

4.1 Experimental Setting

Evaluation: We compare our method, "hierarchical part-based regression" (HPR), with five state-of-the-art methods, Supervised Descent Method (SDM) [Xiong and la Torre, 2013], Robust Cascaded Pose Regression (RCPR) [Burgos-Artizzu *et al.*, 2013], Consensus of Regression (CoR) [Yu *et al.*, 2014], Dlib [Kazemi and Sullivan, 2014] and Chehra [Asthana *et al.*, 2014]. These methods report the top performance among the most recent regression-based methods. Methods with different experimental settings, e.g. Neural Network structures are currently not compared. The codes are provided by the authors from internet. RCPR provides its training code in which the annotation can be defined by input data. To make the comparison consistent, the training databases for HPR, RCPR and CoR are the same, which are LFPW and Helen. SDM is reported to be trained with MultiPIE and LFW. Since SDM and Chehra uses 49 points annotation for training and testing, we also select the overlapped 49 points from our 66 points training setup for testing on LFPW, Helen, AFW and iBug, by neglecting 17 points of face profile. For fair comparison, we use the images in which faces are successfully detected by a third-party face detector, i.e., Viola-Jones detector [Viola and Jones, 2004].



(a) AFW (b) iBug

Figure 4: Cumulative distribution function curves of normalized error on AFW and iBug, comparing the proposed method HPR with other state-of-the-art methods. (a) Error CDF on AFW database. (b) Error CDF on iBug database.

4.2 Localization on Wild Databases

As shown in Fig. 3, our method HPR performs consistently better, especially the performance on Helen database, which is 10% higher at relative error 0.02. The error is calculated by dividing root mean squal pixel error over the face size. Face size is calculated as the tight bounding box around the ground truth. The proposed method explicitly separates the landmarks into components which is more suitable for deforming the local shape variance. The average runtime on a 640 by 480 image is around 0.3s in Matlab with a dual core i7 3.4GHz CPU.

Furthermore, the AFW and iBug, two more challenging databases, are evaluated in Fig. 4. We provide each face in AFW and iBug with a bounding box according to the ground truth for fair comparison.The proposed method is still consistently on top of the other state-of-the-art methods.

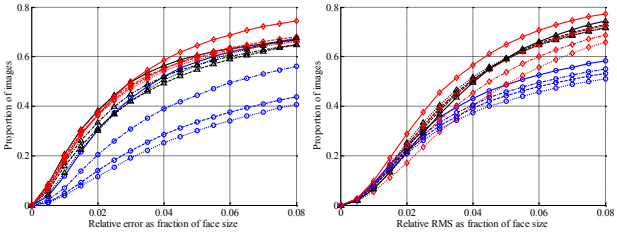
4.3 Localization on Occluded Datasets

For validating accuracy, an evaluation is conducted on the selected occlusion datasets, LFPW-O and Helen-O and the specific occlusion database COFW. Quantitative results from Table. 1 show that HPR achieves consistently better results on the occlusion datasets especially comparing the two occlusion-robust methods, CoR and RCPR. Note that RCPR is trained based on the COFW itself while other methods including ours are not trained on this database.

Further, we synthesize occlusion data from Helen and AFW. A black bounding box is centered at the randomly selected points. The occlusion ratio is controlled by the black bounding box area over the facial area, i.e. 5%, 10%, 15% and 20%, respectively.The CDFs comparing on the synthesized occlusion data is shown in Fig. 5. Regarding the increasing levels of occlusion, all the methods show accuracy decrease. RCPR is relatively more sensitive to the occlusion conditions since the accuracy of different occlusion levels varies largely. Our method performs better on Helen and on AFW at lower occlusion level, while it shows sensitive trend on AFW when the occlusion portion increases.

Table 1: RMS pixel error of CoR, SDM, RCPR, and proposed method HPR on LFPW-O, Helen-O and COFW databases.

	LFPW-O	Helen-O	COFW
CoR	3.33	7.10	3.46
SDM	4.49	9.52	3.63
RCPR	5.73	9.37	**3.03**
HPR	**3.17**	**7.03**	3.56



(a) Helen (b) AFW

Figure 5: Cumulative distribution function curves of normalized error on Helen and AFW at occlusion level 5%, 10%, 15% and 20%. Red: HPR; Black: CoR; Blue: RCPR.

5 Conclusion

We introduced a hierarchical non-linear regression model and facilitated it to the face landmark localization HCI. The non-linear regression model predicts quite well the highly non-linear manifold of the human face shape. With the proposed conditioned regressions, head pose variation is controlled within each separated subspace and the global shape is fast localized. With hierarchical part-based regression, the alignment is evaluated and occlusion information is fed back to the local regressions. Meanwhile, local shape variance is compensated by the part-based regression and the occlusion information is propagated to other landmarks at the last step. Demonstrated by the extensive experiments, the top localization accuracy and fast performance provide high potential for more applications such as human tracking.

References

[Asthana *et al.*, 2014] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Incremental face alignment in the wild. In *CVPR*, 2014.

[Belhumeur *et al.*, 2011] P.N. Belhumeur, D.W. Jacobs, D.J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, 2011.

[Burgos-Artizzu *et al.*, 2013] X. Burgos-Artizzu, P. Perona, and P. Dollar. Robust face landmark estimation under occlusion. In *ICCV*, 2013.

[Cao *et al.*, 2012] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, 2012.

[Cheng *et al.*, 2013] X. Cheng, S. Sridharan, J. Saragih, and S. Lucey. Rank minimization across appearance and shape for aam ensemble fitting. In *ICCV*, 2013.

[Cootes and Taylor, 1997] T.F. Cootes and C.J. Taylor. A mixture model for representing shape variation. In *BMVC*, 1997.

[Cootes *et al.*, 1995] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models-their training and application. *CVIU*, 1995.

[Cootes *et al.*, 1998] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. In *ECCV*, 1998.

[Cootes *et al.*, 2012] T.F. Cootes, M.C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *ECCV*, 2012.

[Cristinacce and Cootes, 2007a] D. Cristinacce and T. Cootes. Automatic feature localization with constrained local models. *PR*, 2007.

[Cristinacce and Cootes, 2007b] D. Cristinacce and T. Cootes. Boosted regression active shape models. In *BMVC*, 2007.

[Dantone *et al.*, 2012] M. Dantone, J. Gall, G. Fanelli, and L.V. Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, 2012.

[Dollar *et al.*, 2010] P. Dollar, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, 2010.

[Ghiasi and Fowlkes, 2014] G. Ghiasi and C. C. Fowlkes. Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model. In *CVPR*, 2014.

[Huang *et al.*, 2007a] G.B. Huang, M. Ramesh, T. Berg, and E.L. Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Technical Report*, 2007.

[Huang *et al.*, 2007b] Y. Huang, Q. Liu, and D.N. Metaxas. A component based deformable model for generalized face alignment. In *ICCV*, 2007.

[Kazemi and Sullivan, 2014] V. Kazemi and J. Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[Le *et al.*, 2012] V. Le, J. Brandt, and Z. Lin. Interactive facial feature localization. In *ECCV*, 2012.

[Liang *et al.*, 2008] L. Liang, R. Xiao, F. Wen, and J. Sun. Face alignment via component-based discriminative search. In *ECCV*, 2008.

[Lowe, 2004] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004.

[Martinez *et al.*, 2012] B. Martinez, M.F. Valstar, X. Binefa, and M. Pantic. Local evidence aggregation for regression-based faical point detection. *PAMI*, 2012.

[Matthews and Baker, 2004] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 2004.

[Medina and Zafeiriou, 2014] J. Medina and S. Zafeiriou. Bayesian active appearance models. In *CVPR*, 2014.

[Ren *et al.*, 2014] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, pages 1685–1692, 2014.

[Sagonas *et al.*, 2013] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCV Workshop*, 2013.

[Saragih and Goecke, 2007] J. Saragih and R. Goecke. A nonlinear discriminative approach to aam fitting. In *ICCV*, 2007.

[Saragih *et al.*, 2011] J. Saragih, S. Lucey, and J. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 2011.

[Tzimiropoulos and Pantic, 2013] G. Tzimiropoulos and M. Pantic. Optimization problems for fast aam fitting in-the-wild. In *ICCV*, 2013.

[Valstar *et al.*, 2010] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, 2010.

[Viola and Jones, 2004] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[Xing *et al.*, 2014] J. Xing, Z. Niu, J. Huang, W. Hu, and S. Yan. Towards multi-view and partially-occluded face alignment. In *CVPR*, 2014.

[Xiong and la Torre, 2013] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013.

[Yang and Patras, 2013] H. Yang and I. Patras. Sieving regression forest votes for facial feature detection in the wild. In *ICCV*, 2013.

[Yu *et al.*, 2013] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *ICCV*, pages 1944–1951, 2013.

[Yu *et al.*, 2014] X. Yu, Z. Lin, J. Brandt, and D. N. Metaxas. Consensus of regression for occlusion-robust facial feature localization. In *ECCV*, pages 105–118, 2014.

[Zhang *et al.*, 2014a] J. Zhang, S. Shan, M. Kan, and X. Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *ECCV*, 2014.

[Zhang *et al.*, 2014b] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. Facial landmark detection by deep multi-task learning. In *ECCV*, 2014.

[Zhu and Ramanan, 2012] X. Zhu and D. Ramanan. Face detection, pose estimation and landmark localization in the wild. In *CVPR*, 2012.

[Zhu *et al.*, 2003] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *ICML*, 2003.

[Zhu *et al.*, 2015] S. Zhu, C. Li, C. Loy, and X. Tang. Face alignment by coarse-to-fine shape searching. In *CVPR*, 2015.