Dual Sparse Constrained Cascade Regression for Robust Face Alignment

Qingshan Liu, *Senior Member, IEEE*, Jiankang Deng, and Dacheng Tao, *Fellow, IEEE*

***Abstract*— Localizing facial landmarks is a fundamental step in facial image analysis. However, the problem continues to be challenging due to the large variability in expression, illumination, pose, and the existence of occlusions in real-world face images. In this paper, we present a dual sparse constrained cascade regression model for robust face alignment. Instead of using the least-squares method during the training process of regressors, sparse constraint is introduced to select robust features and compress the size of the model. Moreover, sparse shape constraint is incorporated between each cascade regression, and the explicit shape constraints are able to suppress the ambiguity in local features. To improve the model's adaptation to large pose variation, face pose is estimated by five fiducial landmarks located by deep convolutional neuron network, which is used to adaptively design the cascade regression model. To the best of our best knowledge, this is the first attempt to fuse explicit shape constraint (sparse shape constraint) and implicit context information (sparse feature selection) for robust face alignment in the framework of cascade regression. Extensive experiments on nine challenging wild data sets demonstrate the advantages of the proposed method over the state-of-the-art methods.**

***Index Terms*— Robust face alignment, sparse feature selection, sparse shape constraint, cascade regression.**

I. Introduction

FACE alignment, that is, locating the facial landmarks on face images, plays an important role in a face analysis system [1]–[4]. A large number of facial landmark localization methods have been proposed in the past two decades [5], and the most popular solution is to take the ensemble of facial landmarks as a whole shape and learn a general face shape model from labeled training images [6]. In respect of this shape model, the previous works can be categorized as explicit shape model-based methods and implicit shape model-based methods.

Manuscript received January 4, 2015; revised May 24, 2015 and September 1, 2015; accepted October 7, 2015. Date of publication November 20, 2015; date of current version January 5, 2016. This work was supported in part by the Natural Science Foundation of Jiangsu under Grant BK2012045, in part by the National Natural Science Foundation of China under Grant 61272223 and Grant 61532009, and in part by the Australian Research Council under Project DP-140102164 and Project FT-130101457. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Shiguang Shan.

Q. Liu and J. Deng are with the B-DAT Laboratory, School of Information and Control, Nanjing University of Information and Technology, Nanjing 210014, China (e-mail: qsliu@nuist.edu.cn; jiankangdeng@gmail.com).

D. Tao is with the Centre for Quantum Computation & Intelligent Systems and the Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway Street, Ultimo, NSW 2007, Australia (e-mail: dacheng.tao@uts.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2015.2502485

Explicit shape model-based methods aim to learn a parametric shape model from the labeled training data. The representative parametric models are Active Shape Model (ASM) [7], Active Appearance Model (AAM) [8]–[10], and Constrained Local Model (CLM) [11], [12], which usually optimize the shape configuration by expensive iterations. There are also a number of other works. For example, in [13]–[18], the elastic deformation model of the human face is built by Markov Random Fields (MRF) or Graph model. Exemplar-based methods try to represent face shapes by a set of similar exemplars generated from the training data [18], [19]. In the context of medial image analysis, Zhang *et al.* [20], [21] developed an Adaptive Shape Composition method (ASC) to model shapes and implicitly incorporate the shape prior constraint effectively. It is based on sparse representation and is able to handle non-Gaussian errors, model multi-modal distribution of shapes and recover local details. The problem is efficiently solved by an EM type of framework and an efficient convex optimization algorithm.

In contrast to explicit shape model-based methods, implicit shape model-based methods try to learn a regression function that directly maps the appearance of an image to the landmark coordinates without a parametric shape model, and the relationship between the landmarks is incorporated flexibly. Since there is no iterative fitting step and no sliding window search, implicit shape models are more efficient and have attracted much attention in recent years. There are two popular ways to learn such a regression function, one of which is based on deep network. Sun *et al.* [22] pioneered the use of Deep Convolutional Neural Network (DCNN) in the regression framework to locate five fiducial facial landmarks. Zhou *et al.* [23] proposed a four-level convolutional network cascade, where each level is trained to locally refine the outputs of the previous network levels. Zhang *et al.* [24] proposed Coarse-to-Fine Auto-encoder Networks (CFAN) for real-time face alignment. Zhang *et al.* [25] formulated a novel task-constrained deep model, with task-wise early stopping to facilitate learning convergence.

The cascade regression model is another popular implicit shape model, which depends on shape index feature and stacked regressors. In [26], Cascade Pose Regression (CPR) was first proposed to estimate pose in an image sequence with pose-indexed features. Explicit Shape Regression (ESR) [27] combines two-level boosted regression, shape-indexed features and correlation-based feature selection. The Supervised Descent Method (SDM) [28] uses cascade regression with fast SIFT feature, and interprets the cascade regression

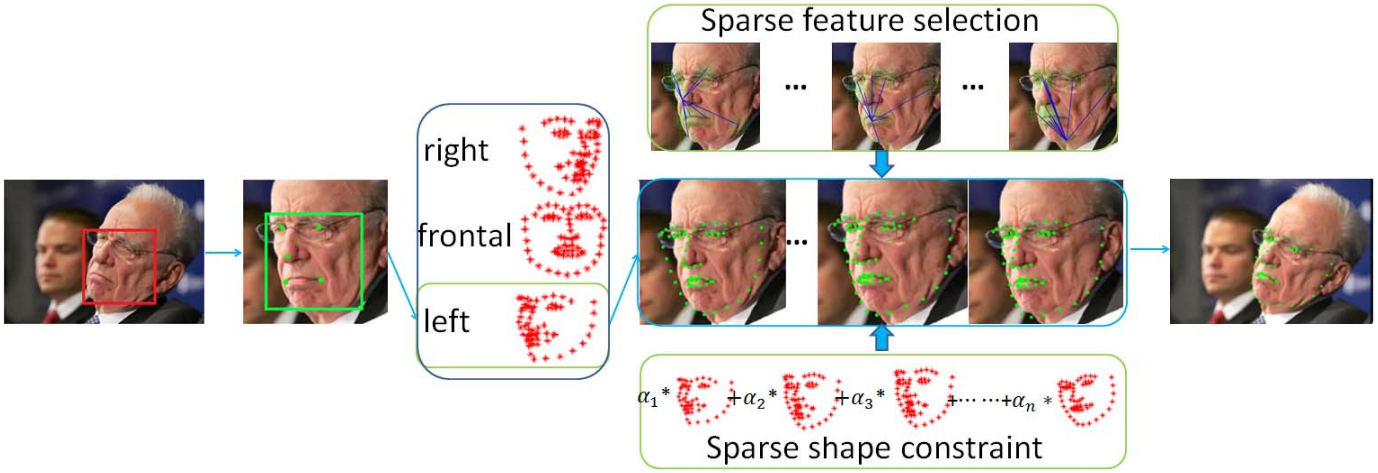


Fig. 1. The overview of P-DSC-CR.

procedure from a gradient descent view [29]. Yan *et al.* [30] utilized the strategy of "learn to rank" and "learn to combine" from multiple hypotheses in a structural SVM framework to handle inaccurate initializations from the face detector. Local Binary Features (LBF) [31] utilizes local binary features to accelerate the alignment. [32] proposed an Incremental Parallel Cascade Linear Regression (iPar-CLR) method, which incrementally updates all linear regressors in a parallel way instead of the traditional sequential manner. Each level is trained independently by using only the statistics of the previous level, and the generative model is gradually turned to a person-specific model by the recursive linear least-squares method. [33] proposed an ℓ_1-induced Stagewise Relational Dictionary (SRD) model to learn consistent and coherent relationships between face appearance and shape for face images with large view variations. Yu *et al.* [34] proposed an occlusion-robust regression method by forming a consensus estimation arising from a set of occlusion-specific regressors. Robust Cascade Pose Regression (RCPR) [35] reduces exposure to outliers by explicitly detecting occlusions on the training set marked with occlusion annotations.

Both deep network methods and cascade regression methods have achieved success in face alignment [5]. Cascade regression methods are effective [28] and efficient [31], and have comparable results compared to deep network methods [5], because deep network methods tend to over fit on limited training data. However, cascade regression-based approaches often depend on local feature descriptors (e.g. Haar wavelets [36], random ferns [37], SIFT [28] and HoG [30]) and regression algorithms (e.g. boosting [36], random forest [26] and least squares [29]). Local features extracted from local patches around the landmarks are sensitive to occlusion, which often occurs in real-world conditions. Additionally, as Cao *et al.* [27] has pointed out, local evidence is sufficiently strong for only a few prominent landmarks, but most of the others are not sufficiently conspicuous and cannot be reliably characterized by their image appearance. It is therefore essential to introduce shape constraint for robust face alignment.

In this paper, we propose a pose-insensitive dual sparse constrained cascade regression model (P-DSC-CR) for robust facial landmark localization. Dual sparse constraints are introduced into the process of cascade regression. During the regressor training, a sparse constraint is incorporated by Lasso [38], which can select the robust features and compress the size of the model. Another sparse shape constraint is incorporated between the regressors to suppress the ambiguity in the local features. In our experiments, the sparse shape constraint largely contributes to the performance improvement under occlusions. The sparse feature selection improves the performance slightly, but it can greatly compress the model's size. The wild datasets usually exhibit large pose variations, which is challenging to face alignment. In [39], projective geometry across different views is used to deal with facial images taken under wide viewpoint variations. In [40], a novel projective invariant named the characteristic number is proposed to derive strong shape priors on frontal upright faces, which can characterize the intrinsic geometries shared by human faces. To improve the model's adaptation to large pose variation, we utilize DCNN [22] to predict five facial landmarks on the pupil, the tip of nose, and the corner of the mouth for face pose estimation. Face pose is then used to guide the regression model selection, which significantly improves the model's adaptation to large pose variation. Compared to the conventional cascade regression model, the proposed P-DSC-CR is more robust to occlusion and pose variation. Extensive experiments on nine challenging wild data sets demonstrate the advantages of P-DSC-CR over the state-of-the-art methods.

II. Overview

As shown in Figure 1, the proposed P-DSC-CR model first uses DCNN to estimate face pose, and adaptively selects different cascade regression models according to face pose estimation. In each cascade regression model, dual sparse constrains are introduced to further improve the performance of the face alignment.

In [30], it was demonstrated that face shape initialization has an effect on the performance of cascade regression.

The authors generated multiple hypotheses and proceeded to learn to rank or combine the hypotheses to handle inaccurate initializations from the face detector. In contrast to [30], we propose to refine the face initialization by pose estimation and divide the face pose into three classes, i.e., a right-profile, frontal, and left-profile pose respectively. Pose estimation is based on five fiducial facial landmarks on the pupil, the tip of nose, and the corner of mouth, which are localized by DCNN [22]. Pose is represented by the roll, yaw, and pitch angle of the face. The roll angle is used to make the face upright, and the face is normalized by the interocular distance. As a result, there is no need to rotate each local feature to the principal direction and estimate the face scale during the process of feature extraction. Yaw angle is used to select the corresponding cascade regression model. If the yaw angle is larger than 30°, the profile model is selected, otherwise, the frontal model is used to obtain the final result.

Following shape initialization, we use a novel cascade regression model with dual sparse constraints for facial landmark localization. Multi-scale HoG features [30] are used to represent the local information of the face shape. Instead of using the least-squares method [28], the Lasso regression method [38] is used to obtain a sparse regression matrix. Lasso can not only select the robust local features, but can also compress the size of the model. Moreover, an explicit sparse shape constraint [20], [21] is introduced to reduce the uncertainty of the gradient descent direction during cascade regression. The idea is to ensure that the regressed shape always resides in the subspace constructed by the learnt training shapes. This guarantees the plausibility of the shape as well as global consistency. To some extent, the sparse regression matrix takes account of context information in an implicit way, and the sparse shape constraint incorporates the relationship between the landmarks in an explicit way. To our best knowledge, this is the first attempt to fuse the explicit shape constraint (sparse shape constraint) and the implicit context information (sparse feature selection) for robust facial landmark localization in the framework of cascade regression.

III. Dual Sparse Constrained Cascade Regression

Cascade regression means to combine a sequence of regressors in an additive manner. During training, the main idea is to learn a regression function f to minimize the mean square error:

$$f = \arg\min_{f} \sum_{i=1}^{N} \left\| f(I_i, X_i^0) - X_i^* \right\|_2^2,$$

where N is the number of training samples, X_i^* is the ground truth shape, X_i^0 is the initialization of face shape. Due to the complex variation in the human face, one regression step is insufficient. A series of simpler regression functions $\{f_1, f_2, \cdots, f_T\}$ can be combined to approximate complex nonlinear mapping between the initial shape and the true shape. In [28], a linear function is utilised as f, and demonstrated to be effective and efficient.

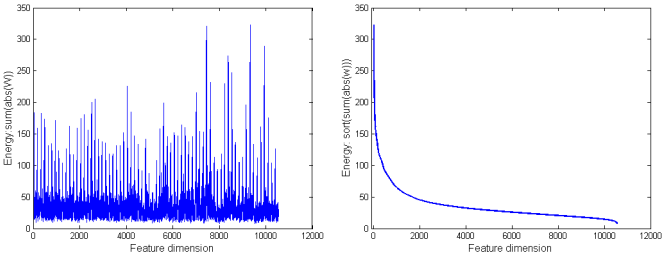


Fig. 2. The energy distribution of W_t.

$$\arg\min_{W_t} \sum_{i=1}^{N} \left\| \left(X_i{}^* - X_i^{t-1} \right) - W_t \Phi(I_i, X_i^{t-1}) \right\|_2^2,$$

where W_t is the linear transform matrix, which maps the current feature vectors to the landmark location update. This is a linear least squares problem, and W_t has a close-form solution. Φ is a nonlinear feature descriptor. Different descriptors demonstrate quite different levels of performance. Both SIFT [28] and HoG [30] show good performance on the task of facial landmark localization, because most of the landmarks are on the edge of facial organs, and gradient orientation histogram features are very suitable for this task. We adopt a multi-scale HoG feature descriptor [30] in this paper.

To improve the model's robustness to occlusion, we incorporate dual sparse constraints into the cascade framework as follows:

$$\arg\min_{\alpha,\gamma,W} \left\| X^* - \Psi(D\alpha, \gamma) - W\Phi(I, \Psi(D\alpha, \gamma)) \right\|_2^2 + \lambda_1 \|W\|_1 + \lambda_2 \|\alpha\|_1,$$

where D is the shape dictionary, α is the sparse shape coefficient, γ is the similarity transformation coefficient, and Ψ is the similarity transformation. X^* is the label shape, I is the face image, Φ is the feature descriptor, and W is the linear regression matrix. This objective function requires W to be sparse and the iteration shapes to be in the subspace constructed by the learnt training shapes. This optimization problem can be solved by iteration.

Fix the shape coefficient α and γ,

$$\arg\min_{W} \left\| X^* - Y - W\Phi(I, Y) \right\|_2^2 + \lambda_1 \|W\|_1,$$

where $Y = \Psi(D\alpha, \gamma)$. This can be solved by Lasso [38], and the learnt W is sparse, which corresponds to the sparse feature selection.

Fix the regression matrix W,

$$\arg\min_{\alpha,\gamma} \left\| X^* - \Psi(D\alpha, \gamma) - W\Phi(I, \Psi(D\alpha, \gamma)) \right\|_2^2 + \lambda_2 \|\alpha\|_1,$$

To facilitate the optimization procedure, we transform the equation as follows:

$$\arg\min_{\alpha,\gamma} \|\Psi(D\alpha, \gamma) - Y - W\Phi(I, Y)\|_2^2 + \lambda_2 \|\alpha\|_1,$$

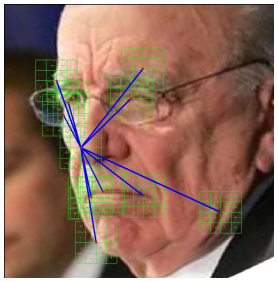


Fig. 3. The schematic diagram of sparse feature selection.

which incorporates the shape constraint into the update shape $Y + W\Phi(I, Y)$.

A. Sparse Feature Selection

Linear cascade regression model utilizes least squares to obtain the regression matrix W_t. Through Figure 2, we analyze the energy distribution of W_t, and find that each dimension of the shape-indexed feature is not the same importance. As a result, we can use the training data to choose relatively robust and stable features under various occlusion patterns. We introduce a sparse constraint during learning as:

$$\arg\min_{W_t} \sum_{i=1}^{N} \left\| \left(X_i{}^* - X_i^{t-1} \right) - W_t \Phi(I_i, X_i^{t-1}) \right\|_2^2 + \lambda_1 \| W_t \|_1,$$

The optimization problem can be solved by Lasso [38], and most of the elements in W_t are zero. Compared to SDM, the sparse constraint can select the robust features from holistic observation, besides compressing the size of the model. Note that an adequate number of training data with partial occlusion are essential for the purpose of selecting robust features. As is shown in Figure 3, the displacement of the nose tip is only associated with parts of the landmark features. If there is a partial occlusion on the face image, there will be a sparse error e on the feature vector $\Phi(I_i, X_i^{t-1})$, which will ultimately affect the shape update. Through sparse constraint, the regression matrix W_t is sparse, and it will decrease the influence on shape update by decreasing $W_t * e$.

B. Sparse Shape Constraint

The cascade regression is generally an open loop during testing [28], and it depends on shape-indexed local features. When the face image is partially-occluded, it fails to locate the landmarks corrupted by occlusion. Inspired by [21], we incorporate a sparse shape constraint to deal with this issue, which is represented by exemplar shapes or a shape dictionary learnt from the training data in a sparse way. From the gradient descent view, as in [28], the process of cascade regression is embedded in the space of sparse shape constraint, which reduces the probability of convergence to local optimum and accelerates the convergence speed. In addition, the sparse shape constraint is able to correct the gross errors of the input shape and can preserve shape details, even if they are not statistically significant in the training set. As a result, the sparse shape constraint and the sparse constrained cascade regression method are complementary.

$$\arg\min_{\alpha,\gamma} \left\| \Psi\left(D\alpha, \gamma\right) - \mathrm{X}^t \right\|_2^2 + \lambda_2 \|\alpha\|_1,$$

where D is the shape dictionary, α is the sparse coefficient, γ is estimated by using Procrustes Analysis, and Ψ is the similarity transformation. As is shown in Figure 4, the landmark localizations on the right eye are influenced by the occlusion, and we introduce the sparse shape constraint from the exemplars to refine the alignment result. To solve the problem, we need to optimize multiple variables simultaneously. Traditionally, an Expectation-Maximization (EM)-like algorithm (or alternating minimization) is needed to solve the above mentioned problem iteratively until α and γ converge, but this is time-consuming. We design an approximate way called "variation rank" to estimate γ, because the occlusions are often partial. Taking the 68 landmark configuration as an example, we divide the facial landmarks into five parts, as shown in Figure 4. We select three of the five parts and randomly select five landmarks from each part to estimate γ by Procrustes Analysis. We do this ten times ($C_5^3 = 10$), and obtain ten results. The variances of each landmark are computed, and the more stable landmarks are used to estimate the final γ. This variation rank strategy has proven to be effective in our experiments. Most of the ten γs are close to the true transformation, because occlusions or large pose variations usually affect no more than three parts. With a fixed γ, α can be solved by Lasso [38].

Shape priors are incorporated on-the-fly by sparse shape composition, however, the computational complexity of the interior-point convex optimization solver is $(\mathrm{N}^2\mathrm{K})$ given the shape dictionary $\mathrm{D} \in \mathrm{R}^{2\mathrm{N}\times\mathrm{K}}$. When the training data are small, the shape dictionary can be directly constructed by the exemplar shapes and the sparse shape composition can be solved efficiently. When the number of training data is very large, it is infeasible to simply stack all the shapes into the data matrix, since sparse shape composition can not handle them efficiently. Owing to the fact that there is significant redundant information among face shape instances, we can use a compact shape dictionary instead of including all the shape instances. We employ an algorithm of K-SVD [41] to learn such a shape dictionary. This compact dictionary greatly improves computational efficiency without sacrificing location accuracy.

C. Dual Sparse Constraints

Both the sparse feature selection and the sparse shape constraint are based on the optimization of Lasso, and many methods have been proposed to solve Lasso in the literature [38] (e.g., the CVX tool).

$$\arg\min_{\alpha} \frac{1}{2} \|X - D\alpha\|_2^2 + \lambda \|\alpha\|_1,$$

For efficiency, we choose the frequently-used method of Augmented Lagrange Multiplier (ALM) [42]. Note that it is inconvenient to solve Lasso by directly applying the

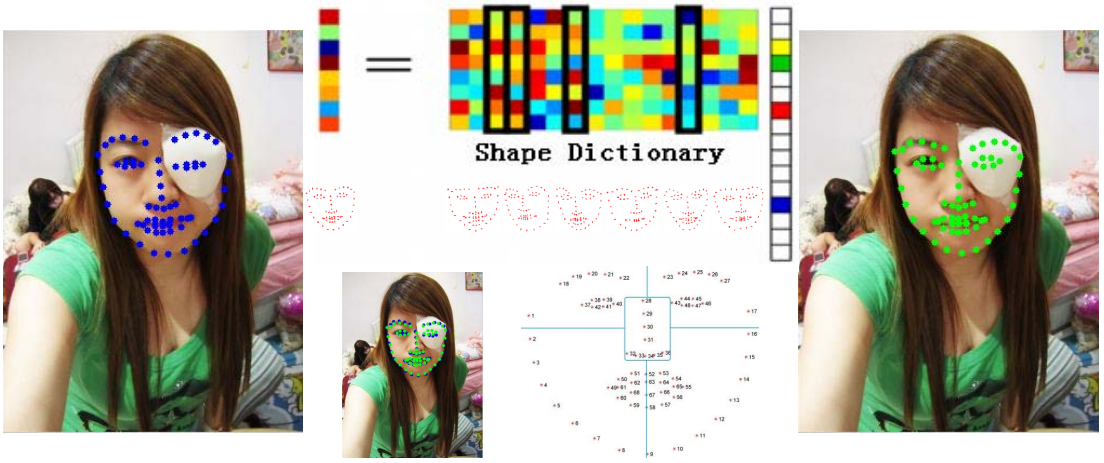


Fig. 4. Sparse shape constraint demo on the exemplar dictionary (68 landmarks are divided into five parts to estimate the transformation parameter γ).

Algorithm 1 Solving Lasso by Inexact ALM

Input: X, D and λ.
Initialization: $\beta = 0, y = 0, \mu = 10^{-6}, \rho = 1.3$.
while not converged **do**

1. fix the others and update α by

$$\alpha = (D^T D + \mu I)^{-1}(\mu\beta - y + D^T X).$$

2. fix the others and update β by

$$\beta = \arg\min_{\beta} \frac{\lambda}{\mu}\|\beta\|_1 + \frac{1}{2}\|\beta - (\alpha + \frac{y}{\mu})\|_2^2.$$

3. update the multiplier and penalty parameter:

$$y = y + \mu(\alpha - \beta) \text{ and } \mu = \min(\mu\rho, 10^{10}).$$

end while
output: α.

ALM method. We therefore convert the optimization of Lasso to the following equivalent problem:

$$\arg\min_{\alpha,\beta} \frac{1}{2}\|X - D\alpha\|_2^2 + \lambda\|\beta\|_1, \text{ s.t. } \alpha = \beta.$$

This problem can be conveniently solved by using ALM, which minimizes the augmented Lagrange function,

$$\frac{1}{2}\|X - D\alpha\|_2^2 + \lambda\|\beta\|_1 + \langle\alpha - \beta, y\rangle + \frac{\mu}{2}\|\alpha - \beta\|_2^2,$$

with respect to α and β, respectively, by fixing the other variables and then updating the Lagrange multiplier y and enlarging the penalty parameter μ. The detailed procedures for solving the Lasso problem are outlined in Algorithm 1.

We finally combine the sparse feature selection and the sparse shape constraint in the united framework. The proposed pose-insensitive dual sparse constrained cascade regression algorithm is summarized in **Algorithm** 2.

Algorithm 2 Pose-Insensitive Dual Sparse Constrained Cascade Regression

Input: Face image I_i, face rect R_i, landmark labels X_i^*.

1) Predict five fiducial landmarks by DCNN and estimate the face pose.
2) Rotate and resize the face image.
3) Divide the training data to train the profile model and frontal model separately.
4) Compute the normalized mean shape and train the shape dictionary from example shapes.
5) Generate ten different initializations for each training data by sampling.
6) **for** $i = 1$ to 3 **do**
7) **for** $t = 1$ to T **do**
8) Extract multi-scale HoG features around each landmark $\Phi(I_i, X_i^{t-1})$.
9) Compute the sparse regression matrix W_t by Lasso, $\arg\min_{W_t} \sum_{i=1}^{N} \left\|(X_i^* - X_i^{t-1}) - W_t\Phi(I_i, X_i^{t-1})\right\|_2^2 + \lambda_1\|W_t\|_1$
10) Update the shape $X_i^t = X_i^{t-1} + W_t\Phi(I_i, X_i^{t-1})$
11) Sparse shape constraint, $\arg\min_{\alpha_i,\gamma_i} \|\Psi(D\alpha_i, \gamma_i) - X_i^t\|_2^2 + \lambda_2\|\alpha_i\|_1$, γ_i is estimated by variation rank based on face component.
12) Generate the constrained shape, $X_i^t = \Psi(D\alpha_i, \gamma_i)$
13) **end for**
14) Save each model.
15) **end for**

Output: Frontal and profile cascade regression model.

IV. Experiments

A. Datasets

There are quite a number of datasets for evaluating face alignment algorithms [43]. Different datasets present different variations in face pose, expression, illumination and occlusion, and the landmark configurations are also different. We select nine well-known and challenging datasets for our experiments.

Extended LFW (5 landmarks)[1] [22] contains 5590 images from LFW [44] and 7876 images are downloaded from the web. Each face is labeled with the positions of five landmarks. We test DCNN for pose estimation on the extended LFW [22].

LFW (10 landmarks)[2] [45] contains the facial images of 5749 individuals, 1680 of which have more than one image in the database. The dataset consists of 13233 images collected in the wild. Instead of using a fixed training and testing set, the evaluation procedure proposed in [45] consists of a ten-fold cross validation using 1500 training images each time and the rest for testing.

BioID (17 landmarks)[3] has 1521 images of 23 subjects. All the faces are frontal, with moderate variations in illumination and expression.

LFPW (29 landmarks) [19] contains 1432 face images (1132 training images and 300 test images). This dataset is intended to test face alignment in unconstrained conditions, and faces show large variations in pose, illumination, expression, and occlusion. This dataset shares only image URLs and some image links are no longer valid. We only downloaded 873 training images and 233 test images.

COFW (29 landmarks) [35] is designed to present faces in real-world conditions, and shows large variations in shape and occlusion due to differences in pose, expression, accessories such as sunglasses and hats, and interactions with objects (e.g. food, hands, microphones, etc.). The training set of COFW includes 845 LFPW faces and 500 COFW faces (1345 in total), and the test set of COFW includes 507 face images. The faces in COFW are occluded to varying degrees, with large variations in the types of occlusion encountered. COFW has an average occlusion of over 23%, and the training set of LFPW contains only 2% occlusion. Note that there are two incorrect annotations in the COFW training set, so we re-annotate these two face images.

300-W (68 landmarks) [6] is created from the existing datasets LFPW [19], AFW [17], Helen [46], XM2VTS [47], FRGC [48] and a new dataset called IBUG. It is created as a challenge and provides only training data. [33] uniformly selected faces with different view angles from the **300-W** dataset to construct a new multi-view wild face dataset called **MVFW**. **MVFW** contains 2050 training samples and 450 testing samples. [33] also built an occlusion dataset called **OCFW**, which contains 2591 training samples and 1246 testing samples. The objective of building **OCFW** was to evaluate the generalization ability of a face alignment model to deal with occlusions when trained on samples without occlusions. In [31], **300-W** is also split into their own training and testing subset. The training set consists of AFW, the training set of LFPW, and the training set of Helen, with 3148 images in total. The testing set consists of IBUG and is called the **300-W Challenging Subset**, while the combination of the testing set of LFPW and Helen is called the **300-W Common Subset**, with 689 images in total.

[1]http://mmlab.ie.cuhk.edu.hk/archive/CNN_FacePoint.htm

[2]http://www.dantone.me/projects-2/facial-feature-detection/

[3]https://www.bioid.com/

Helen (194 landmarks) [46] contains 2300 high resolution web images. We follow the same setting as in [46], with 2000 images for training and 330 images for testing.

B. Experimental Setting

Based on the face detector [49], we train the DCNN model as [22] to predict the five fiducial facial landmarks and use them to estimate the face pose. To achieve a general DCNN-based face pose estimator, the training data are collected from the public dataset AFLW [50], multi-PIE [51], PUT [52], MUCT [53] and about 80000 manually labeled face images from the web. We randomly enlarge ($1.1 \sim 1.3$ times) and translate ($0 \sim 10\%$ of the face width/height) the original face box to generate more training samples. All the face samples are resized to $64 * 64$ for training. The deep convolutional network is designed based on cuda-convnet.[4] During testing, we randomly generate five different patches around the detected face and predict the coordinates of five fiducial landmarks by means of those different patches.

When the five fiducial facial landmarks are predicted, face pose can be estimated to select the corresponding model, and a more exquisite and stable initialization is available. The normalized mean error of [22] on the extended LFW is 0.0439, and the normalized mean error of DCNN designed in this paper is 0.0465, which is accurate enough for the following pose estimation. For the quantitative evaluation of face pose estimation, we use the pose estimator in [54] to label the 300-W data set, and we test the proposed pose estimator on the 300-W data set. The accuracy of the proposed pose estimator is 97.4%. To further evaluate the performance of the proposed pose estimator, we implement the proposed cascade regression model in two ways in the following experiments. One is named P-DSC-CR (Pose-insensitive Dual Sparse Constrained Cascade Regression), in which the DCNN-based pose estimation is used, while the other is named DSC-CR (Dual Sparse Constrained Cascade Regression), in which the pose estimation model is not integrated.

To enlarge the training data, we flip the training image and the landmarks to obtain double training data, as in [27], but in the LFW and Helen datasets, only the original training data is used, since the landmarks are not absolutely symmetrical and the alignment result is even worse when the flipped training data is considered. To achieve a better generalization ability, we randomly sample 10 initial shapes from each training image. The differences between the initial and true landmark locations of each landmark are computed, and their means and variances are used to generate Monte Carlo samples. Since only about 0.5% of the face images have a yaw angle larger than 30° in 300-W (3837 faces), we train the profile model using the face images having a yaw angle larger than 20°. The frontal model is trained on the face images having a yaw angle between $-30°$ and 30°. The overlap of the division makes the model selection more robust. The alignment error monotonically decreases as a function of the number of

[4]http://code.google.com/p/cuda-convnet/

TABLE I
THE REGULARIZATION PARAMETER λ_1 UNDER DIFFERENT LANDMARK CONFIGURATION

	10	17	29	68	194
λ_1	0.2	0.2	0.2	0.3	0.3
sparsity	91.15%	94.88%	93.84%	94.68%	98.04%

regressors added. In all the experiments, we set the iteration number of cascade regression as seven.

We use multi-scale HOG for the feature descriptor, as in [30]. For sparse feature selection, we use Lasso [38] to train the model, and the regularization parameter λ_1 is tuned by cross-validation. As shown in Table I, we use different regularization parameters under different landmark configurations. For the sparse shape constraint, we use K-SVD [41] to train a compact shape dictionary from the training sets. The regularization parameter λ_2, which controls the sparsity of the coefficient, is also tuned by cross-validation. Lastly, we choose 0.01 as λ_2, and the number of example shapes is about 20.

Given the ground-truth, the localization performance can be evaluated by normalized mean error (NME), and the normalization is typically carried out with respect to Inter-Ocular Distance (IOD).

$$err = \frac{1}{N}\sum_{i=1}^{N}\frac{\frac{1}{M}\sum_{j=1}^{M}\left|p_{i,j}-g_{i,j}\right|_2}{|l_i - r_i|_2},$$

where N is the number of images in the test set, M is the number of landmarks, p is the prediction, g is the ground truth, and l and r are the positions of the left eye and right eye.

Based on NME, we plot two kinds of cumulative error distribution (CED) curve. One is calculated from the normalized mean errors over each image, and the other is calculated from the normalized mean errors over each landmark. The permissible error (localization threshold) is taken as a percentage of the interocular distance IOD, typically 10% or below of IOD.

With the face detection results, the cost of the proposed method is only about 0.1 seconds to align a normalized face image (the interocular distance is 100 pixels). It takes about 40ms to estimate face pose, and about 60ms for sparse constrained cascade regression.

C. Experiment on LFW

LFW only has ten labeled landmarks for face alignment [45], and the robust cascaded pose regression (RCPR) [35] has obtained state-of-the-art results in this data set. Thus, we compare the proposed method with RCPR, and following [35], conditional regression forests (CRF) [45], and the explicit shape regression method (ESR) [27] are compared. We also compare the method proposed by Zhao *et al.* [55], which achieves high localization accuracy on LFW. The comparison results are listed in Table II. We can see that P-DSC-CR outperforms all the other methods and DSC-CR is better than CRF, ESR, RCPR and the method proposed by Zhao. This indicates

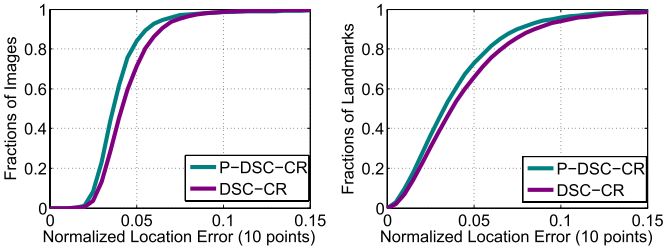


Fig. 5. CED curves on the LFW dataset.

Fig. 6. Normalized mean error on each landmark of DSC-CR and P-DSC-CR on the LFW dataset.

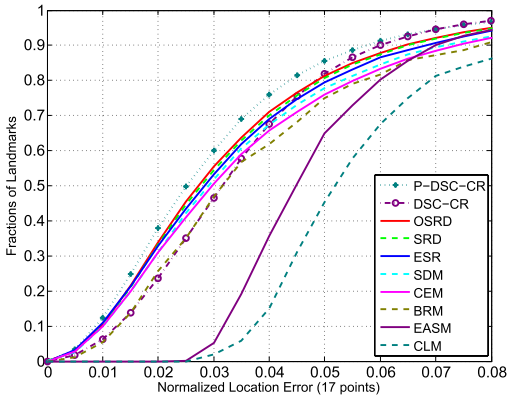


Fig. 7. CED curves on the BioID dataset.

that pose-based strategy and sparse constraints are of great importance for cascade regression. Figure 5 and Figure 6 illustrate detailed comparison results between P-DSC-CR and DSC-CR.

D. Experiment on BioID

On this dataset, the latest results were reported in [33]. We test the proposed method on this dataset and the experimental setting is the same as that of [33]. In addition to the stage-wise relational dictionary model (SRD) and the occlusion stage-wise relational dictionary (OSRD) proposed in [33], the explicit shape regression method (ESR) [27], the supervised descent model (SDM) [29], the consensus exemplar method (CEM) [19], the constrained local models (CLM) [11], the extended ASM method (EASM) [56], and the boosted regression method (BRM) [14] are compared. As is shown in Figure 7, the performance gaps of the most recent methods are not obvious. However, P-DSC-CR still improves the localization accuracy. Sparse feature selection (SFS) and sparse

TABLE II
MEAN ALIGNMENT ERRORS ON THE LFW DATASET

Algorithm	CRF	ESR	Zhao	RCPR	DSC-CR	P-DSC-CR
NME(%)	7.0	5.9	5.8	5.3	4.53	**4.03**

TABLE III
MEAN ALIGNMENT ERRORS ON THE LFPW DATASET

Algorithm	CEM	ESR	EGM	RCPR	SDM	SRD	OSRD	S-SR-DPM	P-DSC-CR
NME(%)	3.99	3.43	3.98	3.50	3.49	3.24	3.19	**2.98**	3.12

shape constraint (SSC) are complementary, and the normalized localization error is more centralized because the explicit shape constraint tends to average the errors. The advantages of sparse feature selection and sparse shape constraint are not obvious because all the faces in the BioID data set are frontal without occlusion. P-DSC-CR has ultimately improves the localization accuracy because DCNN is able to achieve the subsequent cascade regression with better initialization and a more concentrated frontal model. The normalized mean error of DSC-CR is 0.0345, and the normalized mean error of P-DSC-CR is 0.0307.

E. Experiment on LFPW and COFW

LFPW [19] is widely used for face alignment because it is the first wild face alignment database. As with the latest report in [33], we compare our algorithm with other state-of-the-art methods in Table III. The eight baseline methods are the consensus exemplar method (CEM) [19], the explicit shape regression method (ESR) [27], the exemplar-based graph matching method [18], the robust cascaded pose regression [35], the supervised descent model (SDM) [29], the stagewise relational dictionary model (SRD) [33], the occlusion stagewise relational dictionary method (OSRD) [33], and the stacked multiple deformable part model with shape regression (S-SR-DPM) [57]. The result of S-SR-DPM is slightly better than P-DSC-CR because the training data is insufficient to train the profile models for P-DSC-CR. We extend the training set of LFPW with COFW (1007 faces) [35] and Helen (348 faces) [18],[5] and the normalized mean error of the retrained P-DSC-CR is 0.030, which is comparable to S-SR-DPM. The slight performance difference may be caused by the different constitutions of the training data. The performance on LFPW is almost saturated because the human performance is 0.0328, as reported in [35].

The faces in COFW are occluded to different degrees, with large variations in the types of occlusion encountered. Many alignment methods failed to locate facial landmarks accurately due to the large variation in occlusions, and the robust cascaded pose regression (RCPR) [35] obtained state-of-the-art results on this data set. In order to verify the effectiveness of the proposed sparse feature selection and sparse shape constraint, we compare cascade regression (CR), sparse feature selection based cascade regression (SFS-CR), sparse shape constrained cascade regression (SSC-CR), dual sparse constrained cascade regression (DSC-CR), and pose-insensitive dual sparse constrained cascade regression (P-DSC-CR) on the test set of COFW. Like [35], we report the performance of each method on COFW in terms of the normalized mean error and the failure rate, and compare it with other state-of-the-art methods in Figure 9. As shown in Figure 8,

[5]http://www.f-zhou.com/fa.html

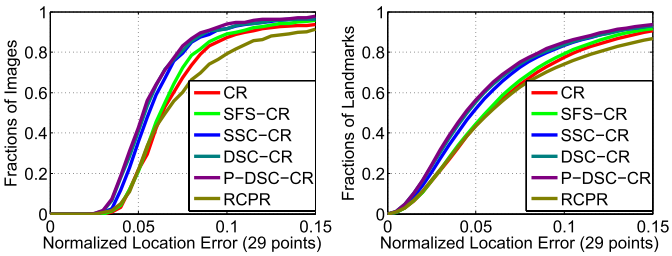


Fig. 8. CED curves on the COFW dataset.

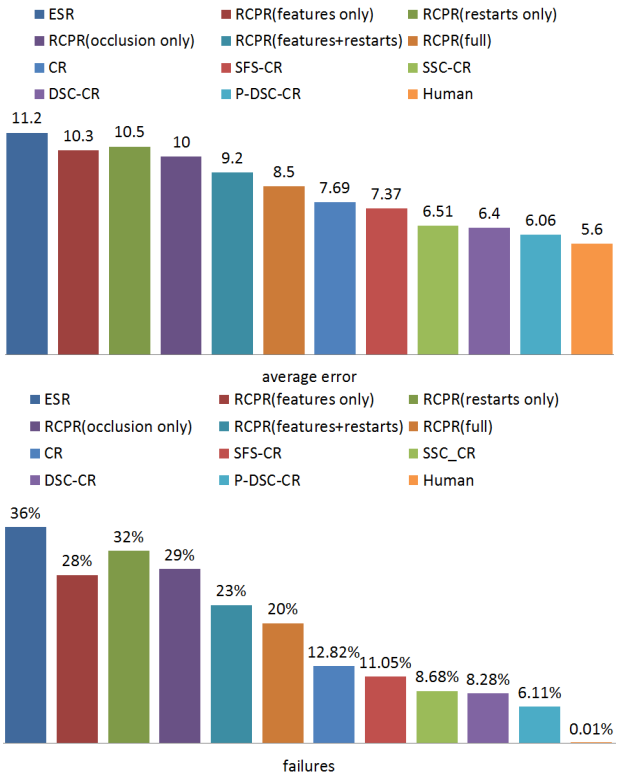


Fig. 9. Normalized mean error and the failure rate on the COFW dataset.

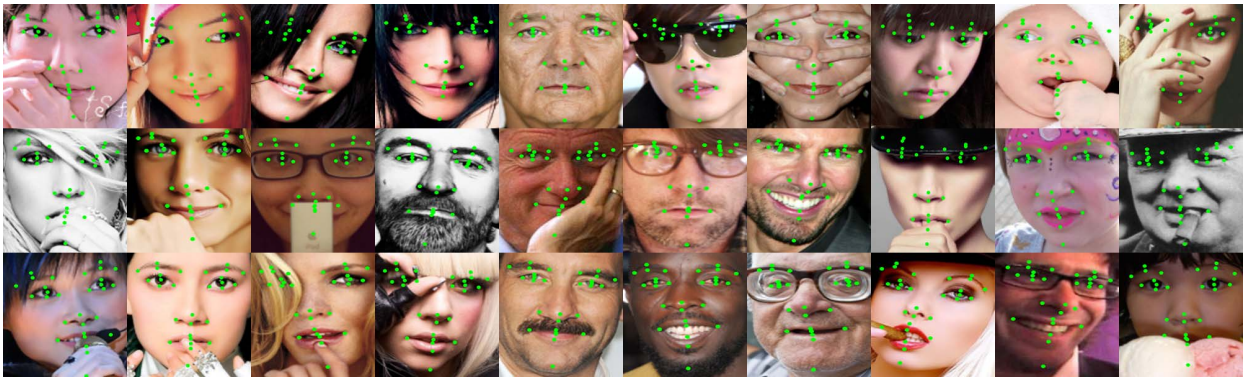


Fig. 10. Example results of P-DSC-CR on the COFW dataset.

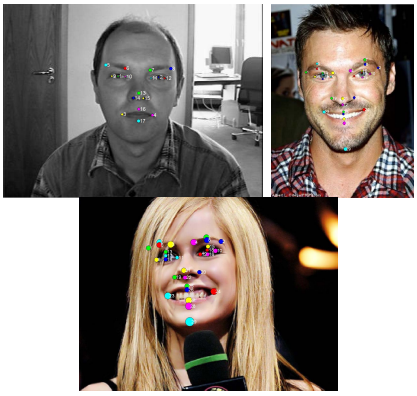


Fig. 11. Mean error of each landmarks on BioID, LFPW, and COFW.

the proposed P-DSC-CR and DSC-CR outperform RCPR by a large margin. The normalized mean error of RCPR on COFW is 0.085 [35]. CR is based on multi-scale HoG features, and its normalized mean error is 0.0769. SFS-CR slightly improves the accuracy to 0.0737, and SSC-CR is able to improve the result to 0.0651. DSC-CR combines sparse feature selection and sparse shape constraint and improves localization accuracy to 0.0640. The alignment result of P-DSC-CR finally reaches 0.0606. The main challenge on COFW is occlusion, and pose estimation just improves the alignment results slightly. Sparse shape constraint is more effective than sparse feature selection to deal with occlusion, however, sparse feature selection is able to greatly compress the model size.

We randomly select results from the test set of COFW images aligned by P-DSC-CR in Figure 10. P-DSC-CR improves the robustness of previous cascade regression methods against occlusion and high shape variance. However, when occlusions are not partial, or exaggerated facial expressions, large pose variations and partial occlusion occur together, P-DSC-CR sometimes fails to locate facial landmarks accurately. In Figure 11, we present the normalized mean error of each landmark on BioID, LFPW, and COFW. The radius of each landmark is the mean error, and it is evident that the alignment result on the eyebrow and chin tip is worse than the others.

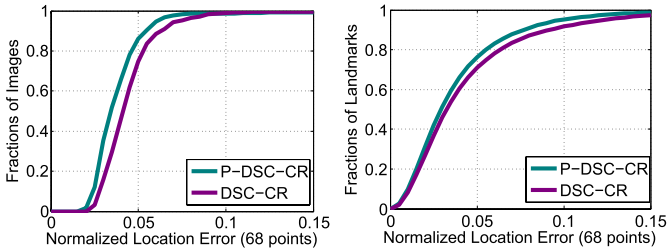


Fig. 12. CED curves on MVFW.

TABLE IV
MEAN ALIGNMENT ERRORS ON THE OCFW

Algorithm	SDM	SRD	OSRD	DSC-CR	P-DSC-CR
NME(%)	9.65	6.37	4.98	4.99	**4.36**

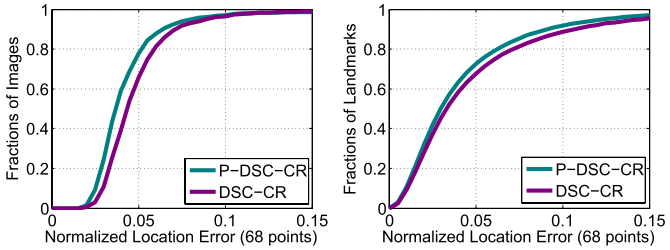


Fig. 13. CED curves on OCFW.

F. Experiment on 300-W

1) Experiment on MVFW and OCFW: The authors in [33] uniformly selected faces with different view angles from the 300-W data set to construct the MVFW data set. SRD [33] obtained the state-of-the-art results on this data set, and the normalized mean error of SRD is 0.0397. The mean error of DSC-CR is 0.0447, and the result of P-DSC-CR is 0.0390, which indicates that the pose-based strategy can greatly decrease pose variation and improve the accuracy of landmark localization. We give the detailed results of DSC-CR and P-DSC-CR on MVFW in Figure 12.

The objective of building OCFW is to evaluate the generalization ability of a face alignment model to deal with occlusion

TABLE V
MEAN ALIGNMENT ERRORS ON THE 300-W COMMON SUBSET, CHALLENGING SUBSET AND FULLSET (*0.01)

	ESR	SDM	LBF	LBF fast	TCDCN	TCDCN-Averaged	DSC-CR	P-DSC-CR
Common Subset	5.28	5.60	4.95	5.38	6.10	5.59	4.88	**3.83**
Challenging Subset	17.00	15.40	11.98	15.50	9.88	9.15	11.49	**6.93**
Fullset	7.58	7.52	6.32	7.37	6.83	6.29	6.04	**4.38**

TABLE VI
MEAN ALIGNMENT ERRORS ON THE HELEN DATASET

Algorithm	EASM	CompASM	ESR	RCPR	SDM	LBF(fast)	DSC-CR	P-DSC-CR
NME(%)	11.1	9.10	5.70	6.50	5.85	5.41(5.80)	5.00	**3.90**

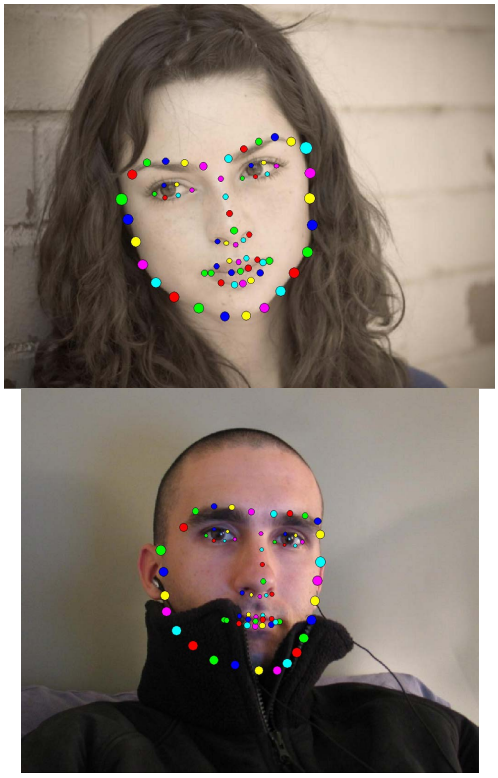


Fig. 14. Mean error of each landmark of P-DSC-CR on MVFW and OCFW.

when trained on samples without occlusions [33]. There are no occluded face images in the training set of OCFW, so we also generate partially occluded training data as in [33]. OSRD [33] obtained state-of-the-art results on this data set. As with the experimental setting in [33], we report the alignment results of DSC-CR and P-DSC-CR, and compare them with the other three baseline methods in Table IV. In Figure 13, we give the detailed result of DSC-CR and P-DSC-CR on OCFW. In Figure 14, we present the IOD normalized mean error of each landmark on MVFW and OCFW. The landmark localization on face contour is clearly worse than those inner landmarks. Without generating the occluded training data, the alignment result for DSC-CR is just 0.0624. Therefore, the occluded training data is of great importance to sparse feature selection.

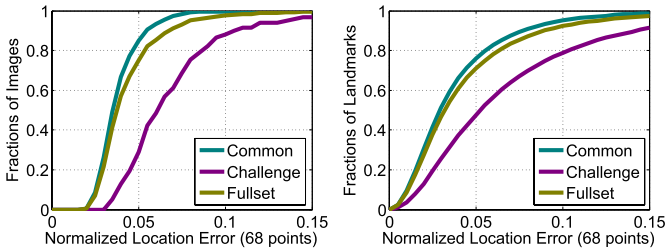


Fig. 15. CED curves of P-DSC-CR on the 300-W dataset.

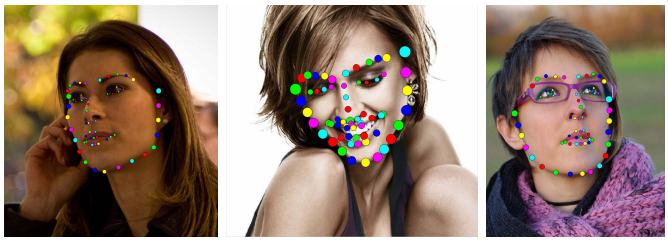


Fig. 16. Mean error of each landmark on the Common Subset, Challenge Subset, and Fullset of 300-W.

2) Experiment on the Common and Challenging Subset of 300-W: The common and challenging subset of 300-W is generated from 300-W, and LBF [31] obtained state-of-the-art results on these data sets. As in [31], we report the alignment results of DSC-CR and P-DSC-CR, and compare them with the other six baseline methods in Table V. P-DSC-CR outperforms all other methods by a large margin and is even better than the deep-based TCDCN [58], especially on the Challenging Subset of 300-W. In Figure 15, we give the detailed results of P-DSC-CR on the 300-W data sets. Example P-DSC-CR results from the 300-W Challenging Subset are given in Figure 17. As can be seen from the results, P-DSC-CR is robust under various conditions. The normalized mean error of each landmark is also given in Figure 16.

G. Experiment on Helen

Helen contains 2300 high resolution web images with 194 landmark annotations for each face. Like [31], we report the performance of our algorithm on Helen, and compare it with other six methods in Table VI. The six baseline methods

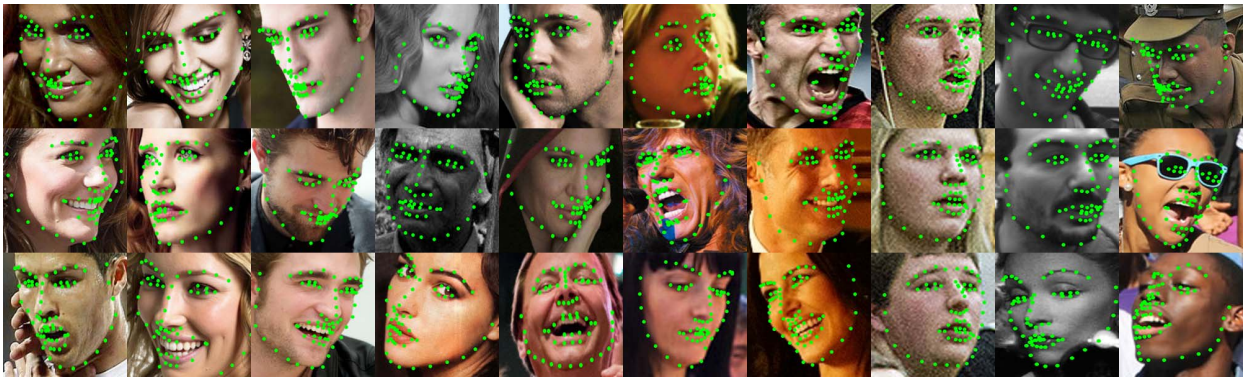


Fig. 17. Example results of P-DSC-CR on 300-W Challenging Subset.

are EASM [56], CompASM [46], ESR [27], RCPR [35], SDM [29], LBF and LBF fast [31]. The human performance on the test set of Helen is 0.033, as reported in [35]. The annotation on Helen is very dense, which is helpful for introducing shape constraint. More landmarks mean a larger size model, and sparse feature selection can clearly compress the model size.

V. Conclusion

This paper presents a pose-insensitive dual sparse constrained cascade regression model for robust facial landmark location. Cascade regression methods for face alignment are usually sensitive to initialization and occlusion. Compared to conventional algorithms, P-DSC-CR has three novel steps: (1) Face pose is first estimated by five fiducial points which are located by deep convolutional neuron network. Face pose is then used to rotate the face image and select the corresponding regression model, which significantly improves the model's adaptation to large pose variation. (2) During the regressor training process, sparse constraint is incorporated by Lasso, which can select robust features and compress the size of model, instead of using least-squares method (3) Sparse shape constraint is incorporated between each cascade regression, and the explicit shape constraints suppress the ambiguity in local features. The implicit shape constraint complements the explicit shape constraint, and the process of cascade regression is embedded in the subspace constructed from the shape of the exemplar, which increases the robustness of the cascade regression model. To improve the computational efficiency, we employ K-SVD to learn a compact shape dictionary without sacrificing location accuracy, and utilize inexact ALM to accelerate the optimization of Lasso. Extensive experiments on nine challenging wild data sets demonstrate the advantages of P-DSC-CR over the state-of-the-art methods.

References

[1] A. Yao and S. Yu, "Robust face representation using hybrid spatial feature interdependence matrix," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3247–3259, Aug. 2013.

[2] J. Qian, J. Yang, and Y. Xu, "Local structure-based image decomposition for feature extraction with applications to face recognition," *IEEE Trans. Image Process.*, vol. 22, no. 9, pp. 3591–3603, Sep. 2013.

[3] A. Ramirez Rivera, R. Castillo, and O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1740–1752, May 2013.

[4] Y. Li, S. Wang, Y. Zhao, and Q. Ji, "Simultaneous facial feature tracking and facial expression recognition," *IEEE Trans. Image Process.*, vol. 22, no. 7, pp. 2559–2573, Jul. 2013.

[5] O. Çeliktutan, S. Ulukaya, and B. Sankur, "A comparative study of face landmarking techniques," *EURASIP J. Image Video Process.*, vol. 13, no. 1, p. 13, 2013.

[6] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops (ICCVW)*, Dec. 2013, pp. 397–403.

[7] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models-their training and application," *Comput. Vis. Image Understand.*, vol. 61, no. 1, pp. 38–59, Jan. 1995.

[8] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," in *Proc. Eur. Conf. Comput. Vis.*. 1998, pp. 484–498.

[9] I. Matthews and S. Baker, "Active appearance models revisited," *Int. J. Comput. Vis.*, vol. 60, no. 2, pp. 135–164, Nov. 2004.

[10] J. Alabort-i-Medina and S. Zafeiriou, "Bayesian active appearance models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2014, pp. 3438–3445.

[11] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," in *Proc. Brit. Mach. Vis. Conf.*, vol. 17. 2006, pp. 929–938.

[12] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, Jan. 2011.

[13] L. Liang, F. Wen, Y.-Q. Xu, X. Tang, and H.-Y. Shum, "Accurate face alignment using shape constrained Markov network," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2006, pp. 1313–1319.

[14] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 2729–2736.

[15] Y. Huang, Q. Liu, and D. N. Metaxas, "A component based deformable model for generalized face alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, Oct. 2007, pp. 1–8.

[16] Y. Huang, Q. Liu, and D. N. Metaxas, "A component-based framework for generalized face alignment," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 41, no. 1, pp. 287–298, Feb. 2011.

[17] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2879–2886.

[18] F. Zhou, J. Brandt, and Z. Lin, "Exemplar-based graph matching for robust facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2013, pp. 1025–1032.

[19] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 545–552.

[20] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. N. Metaxas, and X. S. Zhou, "Sparse shape composition: A new framework for shape prior modeling," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 1025–1032.

[21] S. Zhang, Y. Zhan, M. Dewan, J. Huang, D. N. Metaxas, and X. S. Zhou, "Towards robust and effective shape modeling: Sparse shape composition," *Med. Image Anal.*, vol. 16, no. 1, pp. 265–277, Jan. 2012.

[22] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3476–3483.

[23] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 386–391.

[24] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 1–16.

[25] Z. Zhang, P. Luo, C. Change, and X. Tang, "Facial landmark detection by deep multi-task learning," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 94–108.

[26] P. Dollar, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2010, pp. 1078–1085.

[27] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2887–2894.

[28] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 532–539.

[29] X. Xiong and F. De la Torre. (2014). "Supervised descent method for solving nonlinear least squares problems in computer vision." [Online]. Available: http://arxiv.org/abs/1405.0601

[30] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Learn to combine multiple hypotheses for accurate face alignment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 392–396.

[31] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1685–1692.

[32] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1859–1866.

[33] J. Xing, Z. Niu, J. Huang, W. Hu, and S. Yan, "Towards multi-view and partially-occluded face alignment," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1829–1836.

[34] X. Yu, Z. Lin, J. Brandt, and D. N. Metaxas, "Consensus of regression for occlusion-robust facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 105–118.

[35] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1513–1520.

[36] D. Cristinacce and T. F. Cootes, "Boosted regression active shape models," in *Proc. Brit. Mach. Vis. Conf.*, 2007, pp. 1–10.

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

[38] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statist. Soc. B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[39] F. M. Sukno, J. J. Guerrero, and A. F. Frangi, "Projective active shape models for pose-variant image analysis of quasi-planar objects: Application to facial analysis," *Pattern Recognit.*, vol. 43, no. 3, pp. 835–849, Mar. 2010.

[40] X. Fan, H. Wang, Z. Luo, Y. Li, W. Hu, and D. Luo, "Fiducial facial point extraction using a novel projective invariant," *IEEE Trans. Image Process.*, vol. 24, no. 3, pp. 1164–1177, Mar. 2015.

[41] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Trans. Signal Process.*, vol. 54, no. 11, pp. 4311–4322, Nov. 2006.

[42] Z. Lin, M. Chen, and Y. Ma. (2010). "The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices." [Online]. Available: http://arxiv.org/abs/1009.5055

[43] R. Dhananjay, A. Vinay, S. S. Shylaja, and S. Natarajan, "Facial landmark localization—A literature survey," *Int. J. Current Eng. Technol.*, vol. 4, no. 3, pp. 1–7, Jun. 2014.

[44] G. B. Huang, M. Marwan, B. Tamara, and L. Erik, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, Tech. Rep. 07-49, 2007.

[45] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 2578–2585.

[46] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 679–692.

[47] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre, "XM2VTSDB: The extended M2VTS database," in *Proc. 2nd Int. Conf. Audio Video-Based Biometric Person Authentication*, vol. 964. 1999, pp. 965–966.

[48] P. J. Phillips *et al.*, "Overview of the face recognition grand challenge," in *Proc. IEEE Comput. Vis. Pattern Recognit.*, vol. 1. Jun. 2005, pp. 947–954.

[49] J. Yan, X. Zhang, Z. Lei, D. Yi, and S. Z. Li, "Structural models for face detection," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit.*, Apr. 2013, pp. 1–6.

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

[51] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.

[52] A. Kasinski, A. Florek, and A. Schmidt. "The PUT face database," *Image Process. Commun.*, vol. 13, nos. 3–4, pp. 59–64, 2008.

[53] S. Milborrow, J. Morkel, and F. Nicolls, "The MUCT landmarked face database," in *Proc. Pattern Recognit. Assoc. South Africa*, 2010, pp. 1–6.

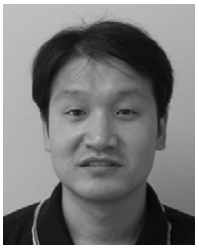
[54] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3444–3451.

[55] X. Zhao, S. Shan, X. Chai, and X. Chen, "Cascaded shape space pruning for robust facial landmark detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1033–1040.

[56] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," in *Proc. Eur. Conf. Comput. Vis.*, 2008, pp. 504–513.

[57] J. Yan, Z. Lei, Y. Yang, and S. Z. Li, "Stacked deformable part model with shape regression for object part localization," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 568–583.

[58] Z. Zhang, P. Luo, C. C. Loy, and X. Tang. (2014). "Learning deep representation for face alignment with auxiliary attributes." [Online]. Available: http://arxiv.org/abs/1408.3967



Qingshan Liu received the Ph.D. degree from the National Laboratory of Pattern Recognition, Chinese Academic of Science, in 2003, and the M.S. degree from the Department of Auto Control, South-East University, in 2000. He was an Assistant Research Professor with the Department of Computer Science, Computational Biomedicine Imaging and Modeling Center, Rutgers, the State University of New Jersey from 2010 to 2011. Before he joined Rutgers University, he was an Associate Professor with the National Laboratory of Pattern Recognition, Chinese Academic of Sciences, and an Associate Researcher with the Multimedia Laboratory, Chinese University of Hong Kong, in 2004 and 2005. He is a Professor with the School of Information and Control Engineering, Nanjing University of Information Science and Technology, China. His research interests are image and vision analysis including face image analysis, graph and hypergraph-based image and video understanding, medical image analysis, and event-based video analysis. He received the President Scholarship of the Chinese Academy of Sciences in 2003.

Jiankang Deng received the bachelor's degree from the Nanjing University of Information Science and Technology, in 2012, where he is currently pursuing the master's degree with the School of Information and Control Engineering. His research interests are face detection and face alignment.

Dacheng Tao (F'15) is a Professor of Computer Science with the Centre for Quantum Computation & Intelligent Systems, and the Faculty of Engineering and Information Technology in the University of Technology Sydney. He mainly applies statistics and mathematics to data analytics problems and his research interests spread across computer vision, data science, image processing, machine learning, and video surveillance. His research results have expounded in one monograph and 200+ publications at prestigious journals and prominent conferences, such as the IEEE T-PAMI, T-NNLS, T-IP, JMLR, IJCV, NIPS, ICML, CVPR, ICCV, ECCV, AISTATS, ICDM; and ACM SIGKDD, with several best paper awards, such as the best theory/algorithm paper runner up award in IEEE ICDM'07, the best student paper award in IEEE ICDM'13, and the 2014 ICDM 10-year highest-impact paper award. He received the 2015 Australian Scopus-Eureka Prize, the 2015 ACS Gold Disruptor Award and the UTS Vice-Chancellor's Medal for Exceptional Research. He is a fellow of the IEEE, OSA, IAPR, and SPIE.