Image and Vision Computing 47 (2016) 19–26

Contents lists available at ScienceDirect

Image and Vision Computing

journal homepage: www.elsevier.com/locate/imavis

M^3 CSR: Multi-view, multi-scale and multi-component cascade shape regression☆

Jiankang Deng [a], Qingshan Liu [a,*], Jing Yang [a], Dacheng Tao [b]

[a] B-DAT Laboratory, School of Information and Control, Nanjing University of Information and Technology, Nanjing 210044, China
[b] QCIS Laboratory, Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway Street, Ultimo, NSW 2007, Australia

ARTICLE INFO

Article history:
Received 27 February 2015
Received in revised form 17 August 2015
Accepted 30 November 2015
Available online 15 December 2015

Keywords:
Face alignment
Cascade shape regression
Multi-view
Multi-scale
Multi-component

ABSTRACT

Automatic face alignment is a fundamental step in facial image analysis. However, this problem continues to be challenging due to the large variability of expression, illumination, occlusion, pose, and detection drift in the real-world face images. In this paper, we present a multi-view, multi-scale and multi-component cascade shape regression (M^3CSR) model for robust face alignment. Firstly, face view is estimated according to the deformable facial parts for learning view specified CSR, which can decrease the shape variance, alleviate the drift of face detection and accelerate shape convergence. Secondly, multi-scale HoG features are used as the shape-index features to incorporate local structure information implicitly, and a multi-scale optimization strategy is adopted to avoid trapping in local optimum. Finally, a component-based shape refinement process is developed to further improve the performance of face alignment. Extensive experiments on the IBUG dataset and the 300-W challenge dataset demonstrate the superiority of the proposed method over the state-of-the-art methods.

1. Introduction

Automatic facial landmark localization plays an important role in facial image analysis [1–4]. A lot of methods [5–10] have been proposed, achieving remarkable improvements [11,12] on standard benchmarks in the past two decades. Existing methods can be roughly divided into three categories: generative methods, discriminative methods and statistical methods [13]. *Generative methods* attempt to optimize the shape parameter configuration by maximizing the probability of a face image being reconstructed by a facial deformable model. Active Shape Model (ASM) [14] and Active Appearance Model (AAM) [15–18] are two representative generative methods. *Discriminative methods* try to infer a face shape through a discriminative regression function, which directly maps a face image to the landmark coordinates [19–24]. There are two popular ways to learn such a regression function. One is based on deep neural network learning [25–28], the other is the well-known cascade shape regression model, which aims to learn a set of regressors to approximate complex nonlinear mapping between the initial shape and the ground truth [29–31]. The idea of *Statistical methods* is to combine both generative and discriminative methods, trying to fit the shape model on a statistical way after learning patch experts. The most notable example is probably the Constrained Local Model (CLM) [32–34] paradigm, which represents the face via a set of local image patches cropped around the landmark points. Recent research efforts have been made on the collection, annotation and alignment of face images captured in-the-wild [12]. However, face alignment is still challenging due to the large variability of expression, illumination, occlusion and pose in the real-world face images [11].

An automatic face alignment system also suffers from the performance of face detector, because its initialization is usually based on the output of face detector. Challenging factors such as pose, illumination, expression and occlusion also have great effects on the performance of face detection [35]. Moreover, face detection is often determined by the criterion [36] that the ratio of the intersection of a detected region with an annotated face region is greater than 0.5. As shown in Fig. 1, all of the faces are detected according to the criterion of 0.5 overlap, but there is more or less drift with the detection results. When the detection result is largely drifted from the ground truth, it is actually not accurate enough for the initialization of face landmark localization algorithm.

In this paper, we propose a robust face landmark localization algorithm. The proposed method is based on the popular cascade shape regression model, and we try to further improve its robustness from three aspects. Firstly, we develop a robust deformable parts model (DPM) [37,35] based face detector to provide a good shape initialization for face alignment. We also utilize the deformable parts information to predict the face view, so as to select the view-specific shape model. View based shape model is not only able to decrease the shape variance, but also can accelerate the shape convergence. Secondly, we develop a multi-scale cascade shape regression with multi-scale HOG features [38]. Multi-scale HOG features can incorporate local structure information implicitly, and multi-scale cascade shape regression helps to

☆ This paper has been recommended for acceptance by Stefanos Zafeiriou.
* Corresponding author.
E-mail addresses: jiankangdeng@gmail.com (J. Deng), qsliu@nuist.edu.cn (Q. Liu), yang.xiaojing00@gmail.com (J. Yang), Dacheng.Tao@uts.edu.au (D. Tao).

http://dx.doi.org/10.1016/j.imavis.2015.11.005
0262-8856/

Fig. 1. Successful face detection results with more or less drift.

avoid trapping in local optimum. To further improve the performance of face alignment, a refinement process is conducted on facial components, such as mouth. The proposed methods achieve the state-of-the-art performance on the challenging benchmarks including the IBUG dataset and the 300-W dataset.

The rest of the paper is organized as follows. The related work is reviewed in Section 2. Cascade shape regression model with multi-view, multi-scale, and multi-component are presented in Section 3. Experimental results are shown in Section 4, and finally the conclusion is drawn in Section 5.

2. Related work

The cascade shape regression model (CSR) has attracted much attention in recent years, because it has achieved much success in face alignment under uncontrolled environment [13]. In [29], Cascade Pose Regression (CPR) is first proposed to estimate pose with pose-indexed features, which iteratively estimates object pose update from the features on current pose. Explicit Shape Regression (ESR) [30] improves CPR by using a two-level boosted regression and correlation-based feature selection. The Supervised Descent Method (SDM) [31] uses linear cascade shape regressions with fast SIFT features, and interprets the cascade shape regression procedure from a gradient descent view [39]. Global SDM (GSDM) extends SDM by dividing the search space into regions of similar gradient directions and obtains better and more efficient convergence [40], which indicates that decreasing shape variation is helpful for cascade shape regressions. Yan et al. [38] utilize 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. In [41], highly discriminative local binary features are used to jointly learn a linear regression. Because extracting and regressing local binary features is computationally cheap, this method achieves over 3000 fps on a desktop. [13] proposes an Incremental Parallel Cascade Linear Regression (iPar-CLR) method, which incrementally updates all the linear regressors in a parallel way instead of the traditional sequential manner. Each level is trained independently by using only the statistics of the former level, and the generative model is gradually turned to a person-specific model by the recursive linear least-squares method. [42] proposes 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. [43] propose an occlusion-robust regression method by forming a consensus estimation arising from a set of occlusion-specific regressors. Robust Cascade Pose Regression (RCPR) [44] reduces exposure to outliers by explicitly detecting occlusion on the training set marked with occlusion annotations. Substantially, CSR is a procedure of shape variance decreasing. In this paper, we develop a robust CSR for face alignment, in which multi-view, multi-scale and multi-component strategies are carefully designed to decrease shape variance.

3. M^3CSR model

Although the cascade shape regression model has achieved much success in face alignment [31], it is still sensitive to some large variations, such as illumination, pose, expression, and occlusion which often exist in real-world images, as well as shape initialization from face detector [38]. In this paper, we propose a new M^3CSR model to make CSR more robust to the real-world variations. Its work flow is illustrated in Fig. 2, in which we enrich the system from three steps. The first step is to develop a reliable face detection and view estimation algorithm to provide a view specified initialization and a view specified cascade shape regression. The second step is to design multi-scale cascade shape regressions with multi-scale HOG features. The last step is to refine facial components to obtain more accurate results.

3.1. Cascade shape regression

The main idea of CSR is to combine a sequence of regressors in an additive manner to approximate complex nonlinear mapping between the initial shape and the ground truth. Specifically, in [31,38], a linear regression function is iteratively used to minimize the mean square error:

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

where N is the number of training samples, $t = 1, \cdots, T$ is the iteration number, X_i^* is the ground truth shape, X_i^0 is the initialization of face shape, Φ is the shape-index feature descriptor, W^t is the linear transform matrix, which maps the shape-indexed features to the shape update. This is a linear least squares problem, and W^t has a close-form solution. During testing, the shape update is iteratively calculated by linear regressions based on the shape-indexed features.

Fig. 2. The work-flow of M^3CSR.

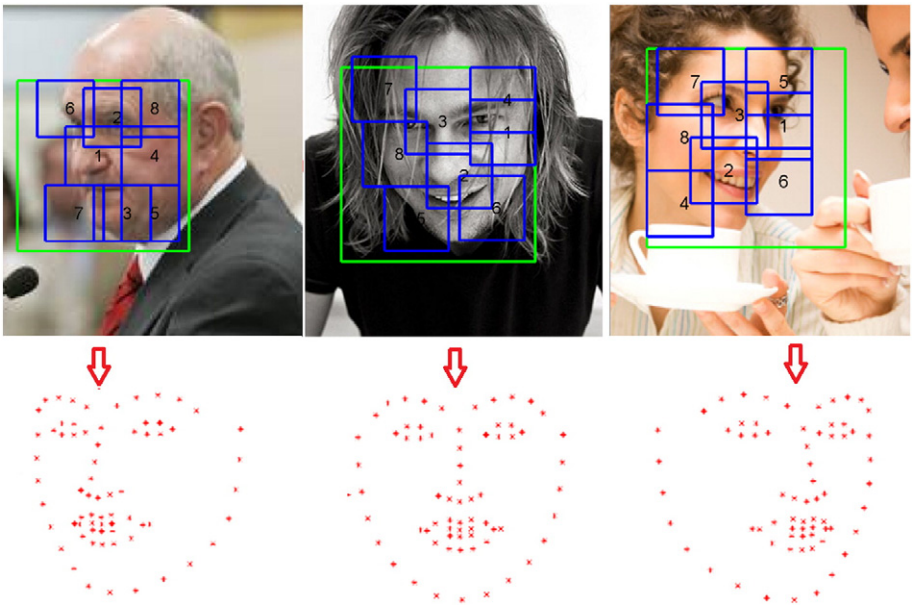


Fig. 3. Illustrations of DPM-based view estimation.

To further improve the CSR's robustness in real-world applications, we enrich the CSR model from three factors including view-specified initialization and regression, multi-scale feature description and multi-scale regression, and component-based final refinement as in the following.

3.2. Multi-view

Inspired by subspace regression [40], we divide the training data into three views, that is, right, frontal, and left views, and we train the CSR model specifically in each view. The view estimation is based on our robust deformable part model (DPM) based face detector.

The DPM-based face detector uses multiple components and deformable parts to handle face variance [37,35]. We define the number of DPM components as six, and each component has one root template and eight parts. We keep all the other training parameters as in [35]. Even though these deformable parts are not exactly corresponding to facial organs, they indeed obey a specific distribution, which is automatic learned as latent variable.[1] As shown in Fig. 3, deformable parts of DPM indicate the face layout, so we use the location of deformable parts to estimate the view status by

$$\arg\min_{R} \sum_{i=1}^{N} \|V_i - RP_i\|_2^2,$$

where V_i is the view status. $P_i \in \mathbb{R}^{32\times 1}$ is the locations of eight deformable parts of DPM. R is the regression matrix, which can be solved by least square method. In the experiments, we only categorize the face views into the frontal ($-15^\circ - 15^\circ$), left ($-30^\circ - 0^\circ$), and right ($0^\circ - 30^\circ$) views, which cover all of the face poses from the 300-W training dataset.[2] The overlaps between the frontal view and the profile views are used to make view estimation more robust.

With the view information, we can train the view-specified CSR model. Because the shape variance of each view set is much smaller than that of the whole set, and the mean shape of each view is much closer to the expected result, so view based shape model is not only able to decrease the shape variance, but also it can accelerate the shape convergence.

[1] http://www.cs.berkeley.edu/rbg/latent/.

[2] Face pose of each training image is estimated by http://www.humansensing.cs.cmu.edu/intraface/download.html.

3.3. Multi-scale

As described above in Section 3.1, the shape indexed feature descriptor is another key issue to the CSR model. Most popular local feature descriptors are successfully applied in the CSR model, such as Haar wavelets [45], random ferns [19], SIFT [31] and HOG [38]. In this paper, we use the HOG descriptor as the shape indexed feature descriptor, because it has achieved good performances in face alignment [39]. The HOG descriptor contains three main steps: pixel-wise feature mapping, spatial aggregation, and normalization as in [46]. In the pixel-wise feature mapping step, the gradient of each pixel is discretized into different partitions according to the orientation. In the spatial aggregation step, the gradient magnitude of each pixel is added to its corresponding bins in four cells around it. Finally, the normalization step is applied to gain the invariance by normalizing the feature vector with energy of four cells around. Fig. 4 shows an illustration, so the final feature map has a 31-dimensional vector, where 27 elements are corresponding to different orientation channels (9 contrast insensitive and 18 contrast sensitive), and 4 elements represent the overall gradient energy in square blocks of four cells.

To further enhance the robustness of HOG, we follow [38] to extract the multi-scale HOG feature, in which local structure information is incorporated implicitly. As shown in Fig. 5, there is one coarse grained region and four fine grained regions. The size of the coarse grained region is 48*48 pixels, and the size of the fine grained region is $24 * 24$ pixels. Thus, we can extract a 31×5-dimensional multi-scale HOG feature around each landmark to construct the shape indexed features.

Additionally, we conduct the cascade shape regression in coarse-to-fine manner. We first conduct it on the small face (we set the face width as 100 pixels in this scale). Then, we double the face size and implement it again to get the final result. This two-scale CSR optimization strategy will help to accelerate the shape convergency and avoid trapping in local optimum.

3.4. Multi-component

Different facial components pose different levels of shape variation [11]. For example, the landmarks on nose are usually more stable than the landmarks on mouth. According to our experiments, in the case of the exaggerate expression, the landmarks on the mouth are usually not accurate enough (Normalized mean error of 68 points is even beyond

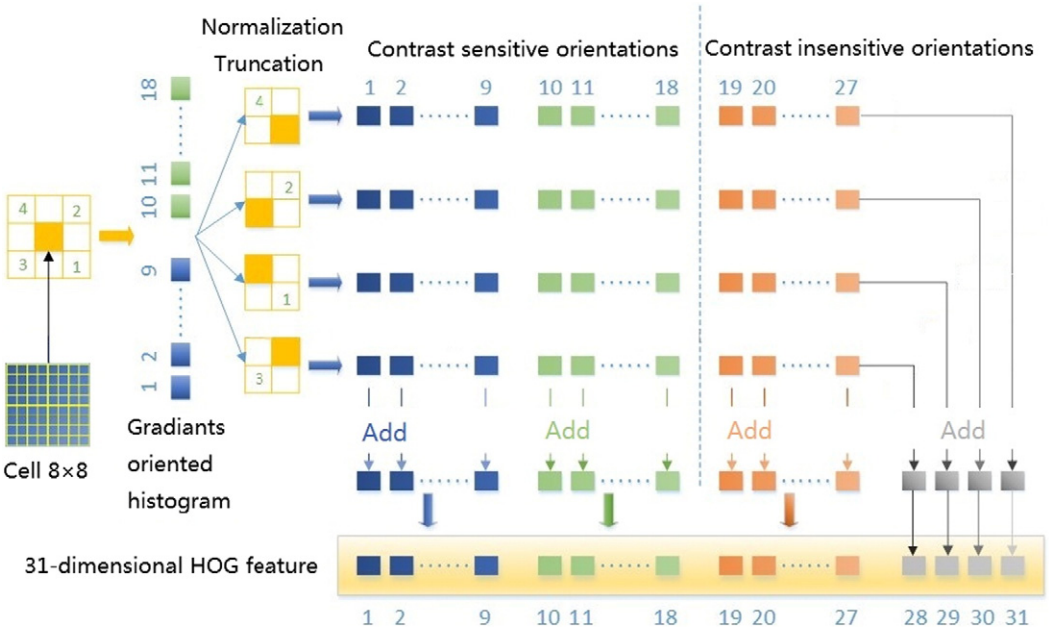


Fig. 4. 31-Dimensional HOG feature.

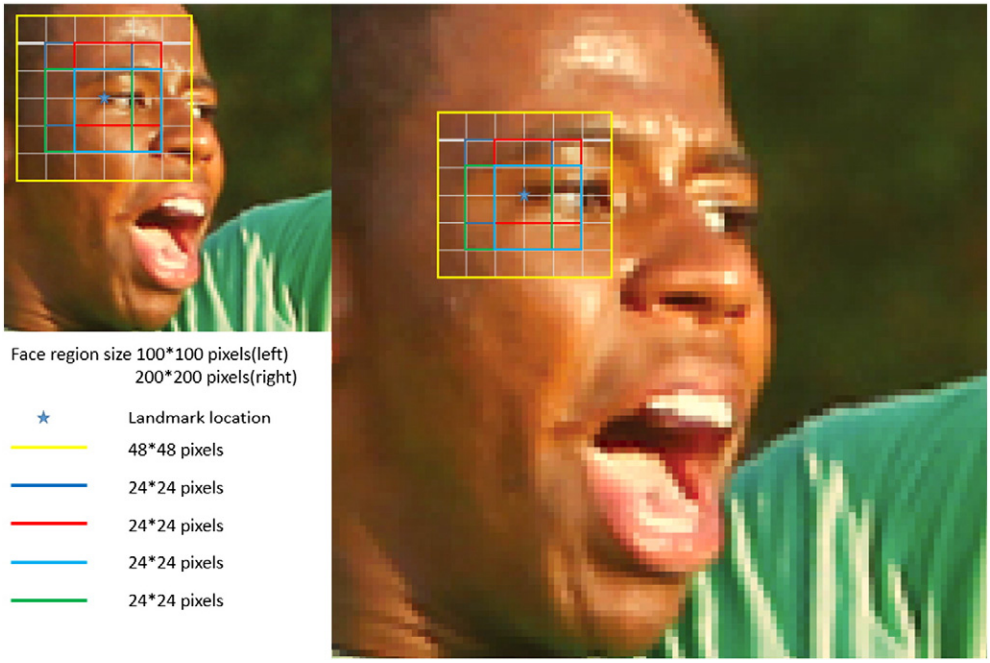


Fig. 5. Multi-scale HOG feature.

7.0%). To handle this issue and to further improve the alignment accuracy, we finally refine the alignment result on each facial component as:

$$\arg\min_{W_j^t}\sum_{i=1}^{N}\left\|\left(X_{j,i}^{*}-X_{j,i}^{t-1}\right)-W_j^t\Phi\left(I_{j,i},X_{j,i}^{t-1}\right)\right\|_2^2,$$

where N is the number of the training samples, $X_{j,i}^{*}$ is the ground truth shape of facial component j, $X_{j,i}^{0}$ is the initialization shape of facial component j, and W_j^t is the regression matrix of facial component j. After component based refinement, we can obtain more accurate alignment results, especially on the large shape variation facial components.

3.5. Algorithm and acceleration

The training process of the M^3CSR is summarized in Algorithm 1. M^3CSR has three sub-view models, each with seven steps of iteration.

Table 1
Face detection results on 300-W dataset.

Dataset	Indoor	Outdoor	All
FDR	99.00%	99.67%	99.33%
Missing faces	3	1	4

Table 2
Face alignment results of CSR based on different face detectors.

Detectors	Generated	DPM	OpenCV ov0.5	OpenCV ov0.7	300-W
NME	6.99%	7.14%	8.50%	7.89%	7.72%

Table 3
Face pose estimation on IBUG.

Face pose (Yaw)	Left	Frontal (−15 ° − 15 °)	Right
Face number	54	38	43
Accuracy	93.6%	95.3%	94.1%

The first four steps of iteration are performed on the 100 × 100 pixels face regions, and the rest three steps of iteration are performed on the 200 × 200 pixels face regions. The eye, nose and mouth refinement models are trained by three steps of iteration. During testing, the DPM-based face detector predicts the location of the face box and estimates the face pose by the deformable parts. Then, the corresponding sub-view model is selected, and linear cascade shape regressions are performed by alternately extracting HOG features and calculate the shape update. Finally, facial component refinement is performed to obtain more accurate alignment results on eyes, nose and mouth.

Algorithm 1. M^3CSR

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

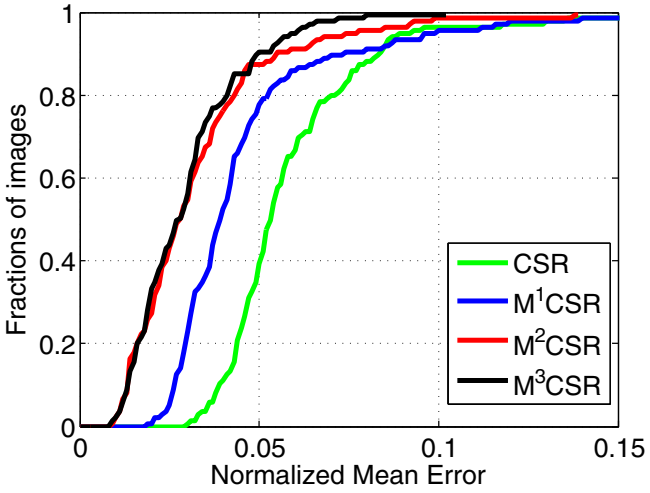
1. Estimate face pose by landmark labels X_i^*.
2. Divide the training data into profile and frontal training data separately.
3. Face detection by DPM and estimate the face view by deformable parts.
4. Resize all of the face regions into 100 × 100 pixels and 200 × 200 pixels.
5. Compute the normalized mean shape for each sub-view.
6. Estimate the error distribution between the mean shape initialization and annotation.
7. Generate ten different initializations for each training data by sampling.
8. **for** each sub-view **do**
9. **for** $t = 1$ to T **do**
10. Extract multi-scale HoG features around each landmark $\Phi(I_i, X_i^{t-1})$.
11. Compute the regression matrix W_t, $\arg\min_{W_t} \sum_{i=1}^{N} \left\| \left(X_i^* - X_i^{t-1}\right) - W_t \Phi(I_i, X_i^{t-1}) \right\|_2^2 + \lambda \|W_t\|_2^2$, λ is set as the number of the training data.
12. Update the shape $X_i^t = X_i^{t-1} + W_t \Phi(I_i, X_i^{t-1})$.
13. **end for**
14. **for** $t = 1$ to 3 **do**
15. Extract multi-scale HoG features on each component $\Phi(I_{j,i}, X_{j,i}^{t-1})$.
16. Compute the regression matrix for each component W_j^t, $\arg\min_{W_j^t} \sum_{i=1}^{N} \left\| \left(X_{j,i}^* - X_{j,i}^{t-1}\right) - W_j^t \Phi(I_{j,i}, X_{j,i}^{t-1}) \right\|_2^2 + \lambda \left\| W_j^t \right\|_2^2$, λ is set as the number of the training data.
17. Update the shape for each component $X_{j,i}^t = X_{j,i}^{t-1} + W_j^t \Phi(I_{j,i}, X_{j,i}^{t-1})$.
18. **end for**
19. Save each model.
20. **end for**

Output: M^3CSR model.

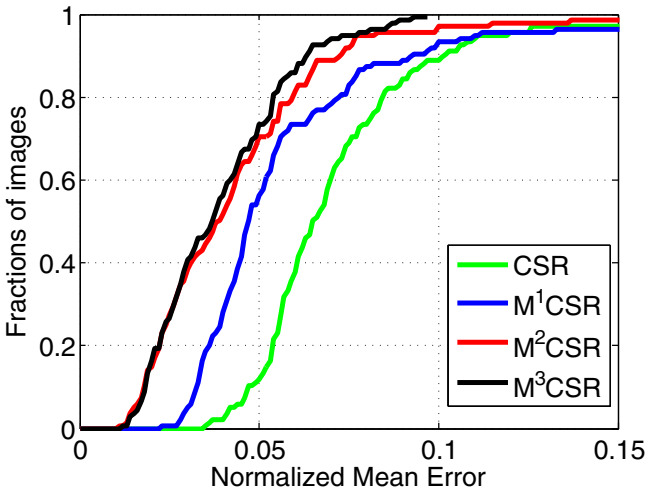
During testing, the linear regression is a simple matrix multiplication, which can be accelerated by Intel Math Kernel Library (MKL). The main computational cost of the proposed method is in the multi-scale HOG feature extraction during cascade shape regression. We use look-up tables (LUT) to accelerate gradients computation. Since the gray value of each pixel is within [0, 255], the gradients at x and y directions are in range of 511 integral numbers [−255, 255]. We pre-calculate 511 × 511 look-up tables, which store all of the possible gradient combinations in the x and y directions. In runtime, these values for each pixel can be indexed from LUT instead of explicit computation. Through the above acceleration strategies, the testing time for each face is within 50 ms, and the training time on the 300-W data sets is about 2 h on a PC with Intel Core i7 CPU. We download the DPM code v5[3] [46,35] to train and test our face detector without any optimization. The running time for the DPM-based face detector is about 1 s on a 640 × 480 face images.

Table 4
Normalized mean error on IBUG dataset.

	CSR	M^1CSR	M^2CSR	M^3CSR
NME51	5.86%	4.60%	3.27%	2.97%
NME68	7.14%	5.73%	4.30%	3.92%



(a) 51 *points*



(b) 68 *points*

Fig. 6. Cumulative error curve on IBUG dataset.

4. Experiments

4.1. Experimental data and setting

A number of face datasets [9,47,7] with different facial expression, pose, illumination and occlusion variations have been collected for

[3] Code http://markusmathias.bitbucket.org/2014_eccv_face_detection/.

Table 5
Eye center distance normalized mean error on IBUG dataset.

Algorithm	DRMF	RCPR	ESR	CFAN	SDM	LBF	TCDCN	Linkface	M^1CSR	M^2CSR	M^3CSR
NME68	19.79%	17.26%	17.00%	16.78%	15.40%	11.98%	9.15%	8.60%	8.25%	6.20%	5.65%

evaluating face alignment algorithms. In [12], some in-the-wild datasets including AFW [7], LFPW [9], and HELEN [47] are re-annotated[4] using semi-supervised methodology [48], and the well established landmark configuration of Multi-PIE [49]. A new dataset called IBUG is also created by [12], which has 135 images with highly expressive faces. The test dataset of 300-W challenge contains 300 Indoor and another 300 Outdoor images, respectively. This test set covers different variations like unseen subjects, pose, expression, illumination, background, occlusion, and image quality, which is aimed to examine the ability of face alignment methods to handle naturalistic, unconstrained face images.

We test the alignment methods on two benchmarks: the IBUG dataset [12] and the 300-W challenge dataset [12]. On the IBUG testset, we use AFW [7], LFPW [9], and HELEN [47] datasets to train the models. The baseline method is CSR, which is trained from the face boxes generated from the DPM-based face detector without view categorization. We evaluate the performance gain of M^3CSR compared to M^1CSR (multi-view cascade shape regression) and M^2CSR (multi-view and multi-scale cascade shape regression). On the 300-W challenge test, we use four datasets (AFW, LFPW, HELEN, and IBUG) to train the M^2CSR model,[5] and obtain the state-of-the-art results on this challenging test set.

The normalized mean error (NME) is adopted to measure the localization performance,

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

where M is the number of landmarks, p is the prediction, g is the ground truth, l and r are the positions of the left eye corner and right eye corner.

The distance between eye corners is used to normalize the error as in [12]. The allowed error (localization threshold) is taken as some percentage of the inter-ocular distance (IOD), typically ≤10%. The normalization is able to make the performance measure independent of the actual face size or the camera zoom factor. Following the evaluation criteria of the 300-W challenge, we use the cumulative error curve of the percentage of images against NME to evaluate the algorithms.

4.2. Face detection and view estimation

For an automatic face alignment system, the shape initialization generally depends on the performance of face detector. Thus, we first evaluate the performance of our face detector. We utilize the AFLW dataset [50] and another one million face images downloaded from the Internet to train our face detector. All the training data for face detection is annotated with five landmarks (left/right eye center, left/right mouth corner, and tip of nose). Face box of each training data is generated from these five landmarks, and the nose tip is taken as the center of the face box, and the scale of the face box is three times of the yaw angle normalized distance between left and right eye center. Table 1 reports face detection results on the 300-W challenge dataset, where the face detection rate (FDR) is the proportion of faces that the NME of face alignment is little than 20%. We can only get the detection results on the 300-W challenge dataset by analyzing the face alignment results. Each image from the 300-W dataset only contains one face, and the proposed face detector only missed four faces. As 300-W challenge dataset, we crop each annotated face from the original image on IBUG with the largest non-face region, and our face detector obtains the detection rate of 100%.

In order to investigate the influence of face detector on the subsequent alignment, we compare the DPM-based face detector with the OpenCV face detector,[6] and 300-W face detector[7] on the IBUG dataset. We also generate face rectangles from five facial landmarks like our face box annotation. We implement the linear cascade shape regression model, using HOG [39] as shape indexed feature, with seven steps of iteration. For the DPM-based face detector and the OpenCV face detector, we select the rectangles with the largest overlap with the bounding box of the annotated landmarks. Moreover, for the OpenCV face detector, we drop the rectangles, which have smaller overlap than particular thresholds. We adopt two thresholds, 0.5 and 0.7, which are named as OpenCV ov0.5 and OpenCV ov0.7 in Table 2. We train the CSR models based on these different kinds of face rectangles, and the alignment results are shown in Table 2. The face rectangles generated from five facial landmarks are better than the face rectangles predicted by the normal face detectors. The key information of the generated rectangles is the locations of the facial components, because they are more semantically stable. Compared with the performance of 300-W official detector, the DPM-based face detector improves CSR by 7.5%, which indicates similar potential gain by the locations of the facial components. The rectangles predicted by OpenCV face detector are most unstable, which increase the shape variance in initialization and give the worst alignment results. Compared to the OpenCV ov0.5, the OpenCV ov0.7 is able to decrease the NME by 7.2%, which indicates that the drift of face detection generates great influence on the following face alignment and more accurate detection results can greatly improve the alignment accuracy.

The DPM-based face detector is able to predict stable face rectangles, which are helpful for the subsequent face alignment. To further decrease the shape variance in initialization, we utilize deformable parts to predict face pose and give the view-specific initial shape. The SDM code implemented by Xiong [31] is able to accurately predict face pose from the facial landmarks. We utilize this code to generate face pose as the ground truth from the landmark annotation for each training data of 300-W challenge. Table 3 reports the view estimation results on the IBUG dataset. When we take into account of the overlap between adjacent views (frontal (−15 ° — 15 °), left (−30 ° — 0 °), and right (0 ° — 30 °)), the accuracy of pose estimation is 99.2%. The normalized mean error of the mean shape initialization on IBUG is 27.51%, which is decreased to 20.69% by the three view-specific shape initialization.

4.3. Face alignment on IBUG

We first investigate the performance gain of M^1CSR, M^2CSR, and M^3CSR on the IBUG dataset. The experimental results are shown in Table 4 and Fig. 6. For the 51 landmarks, M^3CSR, M^2CSR and M^1CSR achieve the mean errors of 2.97%, 3.27% and 4.6% respectively. In the 68 landmarks, where the contour landmarks are included, they obtain the mean errors of 3.92%, 4.30% and 5.73% respectively. It can be seen that multi-view and multi-scale strategies greatly improve the alignment results compared to the baseline of CSR. Multi-component refinement can slightly improve the alignment results too. From Fig. 6(a) and (b), we can see that multi-component refinement obviously improve the alignment results when the mean errors are between 5% to 10%.

[4] http://ibug.doc.ic.ac.uk/resources/facial-point-annotations/.
[5] Because only one submission is admitted for the contest, we can not obtain the performance of M^3CSR on the 300-W dataset.
[6] haarcascade_frontalface_alt.xml.
[7] Rectangles marked as "bb_detector" in http://ibug.doc.ic.ac.uk/media/uploads/competitions/bounding_boxes.zip.

Fig. 7. Example alignment results from IBUG by M^3CSR.

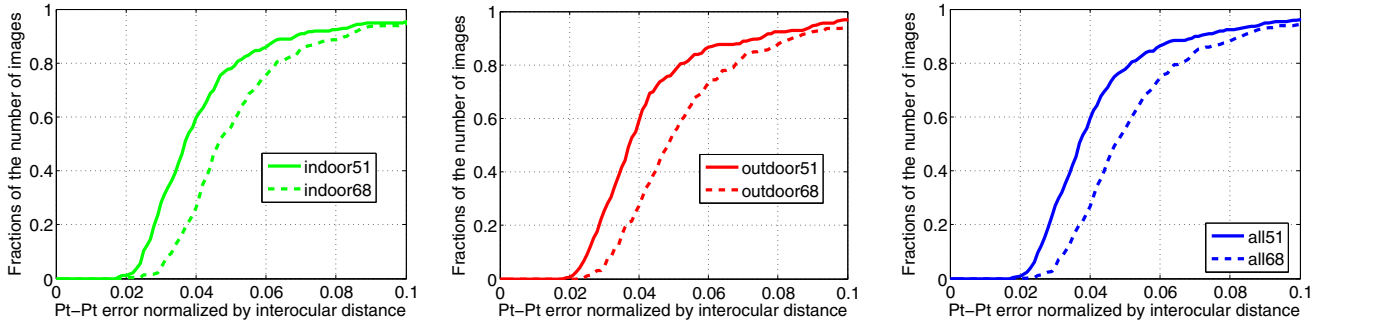


Fig. 8. Cumulative error curve of M^2CSR on 300-W dataset.

Table 6
Normalized mean error (Image numbers) of M^2CSR under different localization thresholds on 300-W dataset.

Dataset	Indoor51	Indoor68	Outdoor51	Outdoor68	In-Out51	In-Out68
≤ 5 %	3.73 % (233)	4.36 % (165)	3.75 % (228)	4.37 % (156)	3.74 % (461)	4.36 % (321)
≤ 10 %	4.23 % (285)	5.19 % (283)	4.27 % (291)	5.30 % (282)	4.25 % (576)	5.24 % (565)
≤ 15 %	4.38 % (296)	5.37 % (294)	4.34 % (298)	5.44 % (297)	4.36 % (594)	5.40 % (591)
≤ 20 %	4.43 % (297)	5.44 % (297)	4.35 % (299)	5.46 % (299)	4.39 % (596)	5.45 % (596)

We further compare M^1CSR, M^2CSR and M^3CSR with the other eight state-of-the-art methods reported in [41], including DRMF [34], RCPR [44], ESR [30], CFAN [27], SDM [39], LBF [41], TCDCN [51], and Linkface.[8] Table 5 lists the experimental results, and we can see that M^1CSR, M^2CSR and M^3CSR are better than all the other eight methods, especially M^3CSR outperforms them by a large margin. Fig. 7 illustrates some example results of M^3CSR on the IBUG dataset. It can be seen that M^3CSR is robust under various conditions.

4.4. Face alignment on 300-W challenge

Following the contest rule, the face images on the 300-W challenge dataset are categorized into "Indoor", "Outdoor" and "Indoor-Outdoor" images. Fig. 8 shows the experimental results of M^2CSR on three kinds of images respectively. Besides the cumulative error curves provided by the contest, we also calculate the normalized mean error in Table 6, which corresponds to the area above the cumulative error curve. For the 51 landmarks, which do not contain the landmarks on face contour, M^2CSR achieves the mean error of 4.39% on all the 596 detected face images, and the mean error of M^2CSR is 5.45% in the 68 landmarks, where the contour landmarks are included. Due to the ambiguity of the landmark definition, the alignment results on face contour are often not accurate enough. M^2CSR obtains similar performances on all the three kinds of images. It indicates that M^2CSR is robust to different "wild" settings.

5. Conclusion

In this paper, we present a M^3CSR model for robust face alignment. Firstly, we develop a robust DPM-based face detector, and we estimate face view based on the locations of deformable facial parts for specifying the view-based CSR models. Secondly, we use the multi-scale HOG features for CSR. Finally, a process of facial component refinement is conducted to obtain more accurate results on the facial components. Extensive experiments on the IBUG dataset and the 300-W challenge

[8] http://www.linkface.cn/index.html.

dataset demonstrate the advantages of the proposed method over the state-of-the-art methods.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 61532009 and Grant 61272223, in part by the Graduate Education Innovation Project of Jiangsu under Grant KYLX15_0881.

References

[1] A. Yao, S. Yu, Robust face representation using hybrid spatial feature interdependence matrix, IEEE Trans. Image Process. 22 (8) (2013) 3247–3259.
[2] J. Qian, J. Yang, Y. Xu, Local structure-based image decomposition for feature extraction with applications to face recognition, IEEE Trans. Image Process. 22 (9) (2013) 3591–3603.
[3] R. Ramirez, C. Rojas, O. Chae, Local directional number pattern for face analysis: face and expression recognition, IEEE Trans. Image Process. 22 (5) (2013) 1740–1752.
[4] Y. Li, S. Wang, Y. Zhao, Q. Ji, Simultaneous facial feature tracking and facial expression recognition, IEEE Trans. Image Process. 22 (7) (2013) 2559–2573.
[5] L. Liang, F. Wen, Y. Xu, X. Tang, H. Shum, Accurate face alignment using shape constrained Markov network, Computer Vision and Pattern Recognition, vol. 1, IEEE 2006, pp. 1313–1319.
[6] M. Valstar, B. Martinez, X. Binefa, M. Pantic, Facial point detection using boosted regression and graph models, Computer Vision and Pattern Recognition, IEEE 2010, pp. 2729–2736.
[7] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, Computer Vision and Pattern Recognition, IEEE 2012, pp. 2879–2886.
[8] F. Zhou, J. Brandt, Z. Lin, Exemplar-based Graph Matching for Robust Facial Landmark Localization, 2013 1025–1032.
[9] P.N. Belhumeur, D.W. Jacobs, D. Kriegman, N. Kumar, Localizing parts of faces using a consensus of exemplars, Computer Vision and Pattern Recognition, IEEE 2011, pp. 545–552.
[10] J. Saragih, Principal regression analysis, Computer Vision and Pattern Recognition, IEEE 2011, pp. 2881–2888.
[11] O. Çeliktutan, S. Ulukaya, B. Sankur, A comparative study of face landmarking techniques, EURASIP J. Image Video Process. 2013 (1) (2013) 13.
[12] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, 300 Faces in-the-wild Challenge: the First Facial Landmark Localization, Challenge, 2013 397–403.
[13] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Incremental face alignment in the wild, Computer Vision and Pattern Recognition 2013, pp. 1859–1866.
[14] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models: their training and application, Comput. Vis. Image Underst. 61 (1) (1995) 38–59.
[15] T.F. Cootes, G.J. Edwards, C.J. Taylor, Active appearance models, European Conference on Computer Vision, Springer 1998, pp. 484–498.
[16] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vis. 60 (2) (2004) 135–164.
[17] J. Alabort, S. Zafeiriou, Bayesian active appearance models, International Conference on Computer Vision, IEEE 2014, pp. 3438–3445.
[18] G. Tzimiropoulos, M. Pantic, Optimization problems for fast AAM fitting in-the-wild, International Conference on Computer Vision, IEEE 2013, pp. 593–600.
[19] T.F. Cootes, M.C. Ionita, C. Lindner, P. Sauer, Robust and accurate shape model fitting using random forest regression voting, European Conference on Computer Vision, Springer 2012, pp. 278–291.
[20] H. Yang, I. Patras, Sieving regression forest votes for facial feature detection in the wild, International Conference on Computer Vision, IEEE 2013, pp. 1936–1943.
[21] B. Martinez, M. Valstar, X. Binefa, M. Pantic, Local evidence aggregation for regression-based facial point detection, IEEE Trans. Pattern Anal. Mach. Intell. 35 (5) (2013) 1149–1163.
[22] G. Tzimiropoulos, M. Pantic, Gauss–Newton deformable part models for face alignment in-the-wild, Computer Vision and Pattern Recognition, IEEE 2014, pp. 1851–1858.
[23] V. Kazemi, J. Sullivan, One millisecond face alignment with an ensemble of regression trees, Computer Vision and Pattern Recognition, IEEE 2014, pp. 1867–1874.
[24] A. Asthana, S. Zafeiriou, G. Tzimiropoulos, S. Cheng, M. Pantic, From pixels to response maps: discriminative image filtering for face alignment in the wild. IEEE Trans. Pattern Anal. Mach. Intell. 37 (6) (2015) 1312–1320.
[25] Y. Sun, X. Wang, X. Tang, Deep convolutional network cascade for facial point detection, Computer Vision and Pattern Recognition, IEEE 2013, pp. 3476–3483.
[26] E. Zhou, H. Fan, Z. Cao, Y. Jiang, Q. Yin, Extensive facial landmark localization with coarse-to-fine convolutional network cascade, International Conference on Computer Vision Workshops 2013, pp. 386–391.
[27] J. Zhang, S. Shan, K. Meina, X. Chen, Coarse-to-Fine Auto-Encoder Networks (CFAN) for real-time face alignment, European Conference on Computer Vision, Springer 2014, pp. 1–16.
[28] Z. Zhang, P. Luo, L.C. Change, X. Tang, Facial landmark detection by deep multi-task learning, European Conference on Computer Vision, Springer 2014, pp. 94–108.
[29] P. Dollar, P. Welinder, P. Perona, Cascaded pose regression, Computer Vision and Pattern Recognition, IEEE 2010, pp. 1078–1085.
[30] X. Cao, Y. Wei, F. Wen, J. Sun, Face alignment by explicit shape regression, Computer Vision and Pattern Recognition, IEEE 2012, pp. 2887–2894.
[31] X. Xiong, F.D. la Torre, Supervised descent method and its applications to face alignment, Computer Vision and Pattern Recognition, IEEE 2013, pp. 532–539.
[32] D. Cristinacce, T.F. Cootes, Feature detection and tracking with constrained local models, British Machine Vision Conference, vol. 17 2006, pp. 929–938.
[33] J.M. Saragih, S. Lucey, J.F. Cohn, Deformable model fitting by regularized landmark mean-shift, Int. J. Comput. Vis. 91 (2) (2011) 200–215.
[34] A. Asthana, S. Zafeiriou, S. Cheng, M. Pantic, Robust discriminative response map fitting with constrained local models, Computer Vision and Pattern Recognition, IEEE 2013, pp. 3444–3451.
[35] M. Mathias, R. Benenson, M. Pedersoli, L.V. Gool, Face detection without bells and whistles, European Conference on Computer Vision, Springer 2014, pp. 720–735.
[36] V. Jain, E.G. Learned-Miller, Fddb: A benchmark for face detection in unconstrained settings, UMass Amherst Technical Report (2010).
[37] J. Yan, Z. Lei, L. Wen, S. Li, The Fastest Deformable Part Model for Object Detection, 2014 2497–2504.
[38] J. Yan, Z. Lei, D. Yi, S.Z. Li, Learn to combine multiple hypotheses for accurate face alignment, International Conference on Computer Vision Workshops (ICCVW), IEEE 2013, pp. 392–396.
[39] X. Xiong, F. De la Torre, Supervised descent method for solving nonlinear least squares problems in computer vision, 2014 arXiv:1405.0601.
[40] X. Xiong, F.D. la Torre, Global supervised descent method, Computer Vision and Pattern Recognition 2015, pp. 2664–2673.
[41] S. Ren, X. Cao, Y. Wei, J. Sun, Face alignment at 3000 fps via regressing local binary features, Computer Vision and Pattern Recognition, IEEE 2014, pp. 1685–1692.
[42] J. Xing, Z. Niu, J. Huang, W. Hu, S. Yan, Towards multi-view and partially-occluded face alignment, Computer Vision and Pattern Recognition 2013, pp. 1829–1836.
[43] X. Yu, Z. Lin, J. Brandt, D.N. Metaxas, Consensus of regression for occlusion-robust facial feature localization, European Conference on Computer Vision, Springer 2014, pp. 105–118.
[44] X.P. Burgos-Artizzu, P. Perona, P. Dollar, Robust face landmark estimation under occlusion, International Conference on Computer Vision, IEEE 2013, pp. 1513–1520.
[45] D. Cristinacce, T.F. Cootes, Boosted regression active shape models, British Machine Vision Conference 2007, pp. 1–10.
[46] F. Pedro, B. Ross, M. David, R. Deva, Object detection with discriminatively trained part based models, IEEE Transactions on Pattern Analysis and Machine Intelligence, IEEE, 2010.
[47] V. Le, J. Brandt, Z. Lin, L. Bourdev, T.S. Huang, Interactive facial feature localization, European Conference on Computer Vision, Springer 2012, pp. 679–692.
[48] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, M. Pantic, A semi-automatic methodology for facial landmark annotation, Computer Vision and Pattern Recognition Workshops (CVPRW), IEEE 2013, pp. 896–903.
[49] R. Gross, I. Matthews, J. Cohn, T. Kanade, S. Baker, Multi-pie, Image Vis. Comput. 28 (5) (2010) 807–813.
[50] M. Kostinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, International Conference on Computer Vision Workshops (ICCVW), IEEE 2011, pp. 2144–2151.
[51] Z. Zhang, P. Luo, C. Loy, X. Tang, Learning and transferring multi-task deep representation for face alignment, 2014 arXiv:1408.3967.