Extended Supervised Descent Method for Robust Face Alignment

Liu Liu, Jiani Hu, Shuo Zhang, Weihong Deng

Beijing University of Posts and Telecommunications, Beijing, China

Abstract. Supervised Descent Method (SDM) is a highly efficient and accurate approach for facial landmark locating/face alignment. It learns a sequence of descent directions that minimize the difference between the estimated shape and the ground truth in HOG feature space during training, and utilize them in testing to predict shape increment iteratively. In this paper, we propose to modify SDM in three respects: 1) Multi-scale HOG features are applied orderly as a coarse-to-fine feature detector; 2) Global to local constraints of the facial features are considered orderly in regression cascade; 3) Rigid Regularization is applied to obtain more stable prediction results. Extensive experimental results demonstrate that each of the three modifications could improve the accuracy and robustness of the traditional SDM methods. Furthermore, enhanced by the three-fold improvements, the extended SDM compares favorably with other state-of-the-art methods on several challenging face data sets, including LFPW, HELEN and 300 Faces in-the-wild.

1 Introduction

Facial landmark locating is key to many visual tasks such as face recognition, face tracking, gaze detection, face animation, expression analysis etc. It is also a face alignment procedure, if we regard all landmarks together as a face shape. Previous works on face recognition has proven that the recognition performance highly depend on the preciseness of the image alignment process [1][2][3][4]. An excellent facial landmark locating approach should be fully automatic, efficient and robust in unconstrained environment with large variations in facial expression, appearance, poses, illuminations, etc. Also, the number of feature points that are needed to be located varies with the intended application. For example, for face recognition or detection, 10 points including nose tip, four eye corners and two mouth corners are enough. However, for higher level applications such as facial animation and 3D reconstruction, 68 or even more landmarks are preferred. To facilitate the assessment of locating algorithm in complex environment, datasets annotated with various number of landmarks have been released, such as LFPW(29) [5], HELEN(194) [6] and IBUG(68) [7]. The images in these datasets are almost real-world, cluttered images which are mainly collected from the Internet and exhibit large variations in pose, appearance, expression, resolution etc. When dealing with large amounts of landmarks, a global shape constraint is essential for the landmark detector, since some points such as those on the chin

and contours are difficult to be characterized alone by the local appearance and may need clues from the correlation between landmarks or the facial structure.

The normal process to achieve landmark locating is first learning some discriminative information such as principal components[8–10], static probability model[5, 11–13] or regressors [14, 15, 13, 16–19] from the training data, and then when given a test image, making use of the learned information or model as well as the property of the test image (feature or appearance) to predict the shape (increment) or parameter, aiming at aligning the estimated shape to the true shape. Always the estimation procedure could not be accomplished in one step, due to both intrinsic factors such as variability between individuals and extrinsic factors such as pose, partial occlusion, illumination and camera resolution differences and so forth, which means that usually feature point locating is an iterative coarse-to-fine procedure.

In recent years, discriminative shape regression has emerged as the leading approach for accurate and robust face alignment. Among them, SDM[15] proposed by Xiong et al. is a representative for its natural derivation from Newton Descent Method as well as its high efficiency and accuracy. Furthermore, the authors put forward the error function for discriminative methods, connecting it with parameterized appearance models. In this paper, based on SDM and the coarse-to-fine principle in face alignment procedure, we make three modifications to the original algorithm: 1) Multi-scale HOG features are applied orderly as a coarse-to-fine feature detector; 2) Global to local constraints of the facial features are considered orderly in regression cascade; 3) Rigid Regularization is applied to obtain more stable prediction results. We show the improvement in locating accuracy by testing these modifications on several challenging datasets.

2 Related Work

Pioneering work on landmark locating/shape estimation includes ASM and AAM[8]. These methods learn shape and appearance principal components from training data to guide the estimation of parameter such as reconstruction coefficients and geometric transformation to minimize the differences between the query and model face. The model face is reconstructed by projecting the query face to the trained principal component subspace. Due to the their elegant mathematical formulation and efficient computation, a number of such AAM/ASM-based approaches have been proposed. A representative example is STASM[9], it makes extensions to the original ASM including stacking two ASM in series, using two-dimensional profile, loosening up the shape model etc. Although various improvements have been proposed, the drawback of AAM/ASM-based methods remains apparent: the expressive power of the eigen space is still limited even when a much higher-dimensional appearance subspace is employed, therefore these methods show unsatisfactory performance on unseen images or images with large variation from the training data. Moreover, due to their use of gradient descent optimization, the result given by these approaches depends heavily on initialization, which is usually considered undesirable.

Other popular modern approaches prefer to perform alignment by maximizing the posterior likelihood of each facial part or point. They usually construct specific fiducial point detectors for each point or part, which predict their probability distributions given the local features, and combine them with a global shape prior to exert constraint. In [5], a Bayesian framework unifies the probability of a global shape in order to select similar samples from training data for query images. In [19], seven fiducial landmarks are firstly detected by priori probability, then MRF is employed to construct global shape to further estimate fewer fiducial landmarks. Besides, methods based on convolutional network[20] have also been evaluated and proposed in recent years.

In this study, we focus on the regression-based methods[14, 21, 22, 15, 17, 23, 24], which are the leading approaches in facial landmark locating in recent years because of their higher accuracy and efficiency over other methods. The good performance of regression-based methods mainly owe to their ability to adaptively enforce shape constraints, the strong learning capacity from large training datasets as well as the precise objective function.

In general, regression-based methods learn a mapping from image features to the face shape or shape increments.

Two representative regression-based approaches are explicit shape regression(ESR) and supervised descent method(SDM). In [22], Cao et al. make use of "shape indexed feature" and learn a weak regressor between the shape increment ΔS and the features in a cascade manner, and the features are regenerated in each iteration. During test, starting with an initial shape, shape increments are predicted step-by-step to refine the shape.

Xiong et al. proposed Supervised Descent Method(SDM)[15] which directly regresses the shape increment by applying linear regression on HOG feature map. Different from boosted regression based methods as ESR, they establish the objective function of aligning image in feature space and infer the approach from Newton Descent Method to give an explanation to linear regressor and feature re-generation, which is quite simple, effective and well-understood.

3 Method

To make this paper self-contained, we first give a brief introduction to the SDM approach[15]. Then we describe the modifications proposed in detail.

3.1 Supervised Descent Method(SDM)

Given an image $\mathbf{d} \in \Re^{m\times 1}$ of m pixels , $\mathbf{d}(\mathbf{x}) \in \Re^{2p\times 1}$ represents the vector of landmarks (face shape) coordinate, where p is the number of landmarks. $\mathbf{h}(\bullet)$ is feature extraction function. The objective function for face alignment is

$$f(x_0 + \Delta x) = \|\mathbf{h}(\mathbf{d}(x_0 + \Delta x)) - \phi_*\|_2^2 \tag{1}$$

where $\phi_* = \mathbf{h}(\mathbf{d}(x_*))$ represents the feature(HOG) of the true shape.

The target is to calculate the shape increment Δx recursively to minimize the difference between the feature of estimated shape and that of the true shape, i.e. align the test image w.r.t. a template ϕ_* in feature space. HOG feature extracted from patches around the landmarks is used to achieve robustness against illumination and appearance variations. After a series of derivation and approximation such as second order Taylor expansion, similar to Newton's method[25], the update for x is derived: $\Delta x_1 = -2H^{-1}J_h^T(\phi_0 - \phi_*)$. Considering that ϕ_* is unknown in testing, ϕ_* is not used in training, and the equation is rewritten as a generic combination of feature vector ϕ_0 plus a bias term b_0.

$$\Delta x_1 = \mathbf{R}_0\phi_0 + b_0 \tag{2}$$

where $\mathbf{R}_0$ is referred as a descent direction.

Due to the large variations of face, alignment can hardly be accomplished in a single step. To cope with the non-quadratic function, a sequence of generic descent directions are learned by applying the update rule in Eq. 3.

$$\begin{aligned} x_k &= x_{k-1} + \mathbf{R}_{k-1}\phi_{k-1} + b_{k-1} \\ \Delta x_k &= x_k - x_* \\ \phi_k &= \mathbf{h}(\mathbf{d}(x_k)) \end{aligned} \tag{3}$$

In training, for each iteration, Δx_* and ϕ_k are created applying the update rule in Eq. 3, then generic descent directions R_k and bias terms b_k are learned through minimizing the difference between the true shape and the predicted shape in feature space, which can be solved by linear regression, as expressed in Eq. 4. Then the current estimated shape is updated using Eq. 3. After several steps, the estimated shape x_k converges to the true shape x_* for all images in the training set, which means the training process is completed. The main step of training is summarized in Algorithm 1.

$$\arg\min_{R,b} \|\Delta x_*^i - R_k\phi_k - b_k\|_2^2 \tag{4}$$

During testing, given a test image, the learned regressors $\{R_i, b_i\}_{i=0}^N$ are employed to compute the shape increment recursively using Eq. 3, where N indexes the number of training stage. The main steps are summarized in Algorithm 2.

Apart from the above, training data augmentation (i.e. initializing each training image from different estimate) is also mentioned to enlarge the training set and improve the generalization ability.

3.2 Modifications of Original SDM Algorithm

The main step of SDM is to predict the shape increment making use of features extracted in current iteration by linear regression. Therefore, the performance depends on both the discrimination of extracted local features and the effectiveness of linear regression. Based on this, we propose three modifications to the original supervised descent method, which are described in detail below.

Algorithm 1 SDM training

Input:
a set of K face images $\{\mathbf{d}^i\}_{i=1}^K$
corresponding true shape $\{x_*^i\}$
initial estimate shape x_0
Output:
regressor $\{R_i, b_i\}_{i=0}^N$
while $i < N$ **do**
  compute regress target shape $\Delta x_* = x_* - x_{i-1}$
  extract HOG feature ϕ_i from current shape x_{i-1}
  learn R_k and b_k using Eq. 4
  update estimated shape x_i using Eq. 2
  $i \leftarrow i + 1$
end while

Algorithm 2 SDM testing

Input:
test image I
regressor $\{R_i, b_i\}_{i=0}^N$
initial shape S_0
Output:
final predicted shape S^N
while $i < N$ **do**
  extract feature ϕ_i in current shape S^i
  compute shape increment $\Delta S = R_i\phi_i + b_i$
  update shape S^{i+1} using $S^{i+1} = S^i + \Delta S$
  $i \leftarrow i + 1$
end while

Adaptive Feature Block In the original SDM algorithm[15], the SIFT features of landmarks are extracted in fixed-size blocks to predict the shape increment for each iteration. Intuitively, the feature extraction block size is related with the value of shape increment. If the shape increments of all training samples are scattered widely, a larger feature extraction block is preferred, thus our first modification is to replace the original fixed-size block by an adaptive feature extracting block.

In initial stages, when the estimated shape is far from the ground truth, it is favorable to extract features in bigger patches around predicted landmarks to utilize more useful information, which is beneficial for handling large shape variations and guaranteeing robustness, as the first image in Fig. 1 indicates. The red points stand for the true shape, green ones for currently estimated shape, and green circles are the range of feature extraction block, where we only draw five feature point for clarity. In training, the difference between the predicted shape and the true shape decreases step-by-step. Extracting features in a gradually-shrinking region helps to grasp the discriminative features effectively. Especially in late stages, where extracting feature in small-size blocks tends to reduce the

proportion of noise, thus ensuring accuracy, as is shown in the third picture in Fig. 1. Therefore, we recommend the progressively-reducing block size other than fixed size to achieve better feature extration performance.

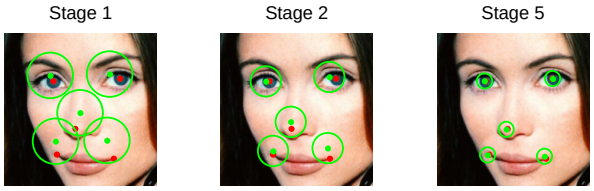


Fig. 1. An illustration of adaptive feature extraction block mechanism. The red points represents the true shape, green for estimated landmarks in corresponding stage, green circles are radius of extraction patch. Along with approximation to true shape, the radius shrinks.

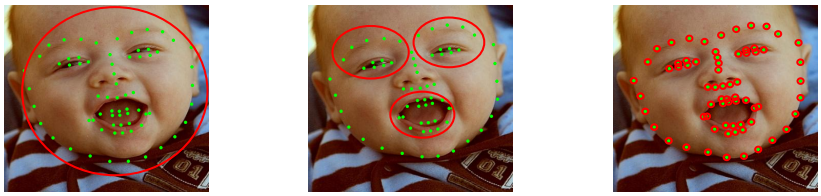
Adaptive Regression The second modification is intended to modify the regression mode.

During the training process, features of all fiducial points constitute the shape feature ϕ. In the same way, the displacement of all landmarks consist of the shape increment Δx. R and b are obtained by minimizing equation 2 through linear regression in 4 or 5 steps. ϕ and Δx are both an ensemble of all landmarks. We denote this mode as global regression. Global regression enforces a shape constraint to guarantee robustness, avoiding divergency in iteration, which is essential when the prediction is far from the target.

Though variations in face geometry and expression are considered, global regression would sacrifice the accuracy to guarantee robustness at times. Other regression methods involve part regression, which is regressing different regions of face separately, and local regression, to regress each landmark individually. By operating locally at each time, these methods could avoid influence from other part or landmark of the image.

To enhance the capability of handling large variation, we recommend replacing the original global regression in all stages by a more flexible mode of adaptive regression: global regression⇒part regression⇒local regression mechanism. The regression process starts from the mean shape. At initial stages, where the estimated shape is unreliable, global regression is preferred to enforce a shape constraint to guarantee robustness. At middle stages, however, thanks to the restricted model in previous stages, the majority of the landmarks have been at or near their optimal position except for some particular regions. For instance, as shown in Fig 2(a), in the current estimated shape, the mouth region is the furthest from the true shape than any other regions of the face. At this time, regressing different regions independently may improve the generalization ability of the regression model.

In late stages, on the premise of proper operations in the previous step, only small refinement on a few of the points is required, i.e., the shape constraint is now exhausted. Hence, regressing each landmark increment individually is beneficial, which means that employing local regression at late stages helps to further improve the locating accuracy.



(a) (b) (c)

Fig. 2. A graphical representation of adaptive regression mechanism.

The choice of landmarks in part regression is alterable, which depends on the distribution and amount of landmarks. A normal way is dividing the face into five parts: two eyes region, nose region, mouth region and contour, which is adopted in our method. Similarly, the configuration of regressions is also alterable. When the target is complicated, we tend to properly increase the proportion of global regression, otherwise, the number of local or part regression could be set higher to achieve better accuracy. In our experiment, we always adopt the setting of 2 *global* $\Rightarrow$ 2 *part* $\Rightarrow$ 3 *local* regression. It should be noticed that adaptive regression is compatible with the adaptive feature block scheme above, which jointly improve the locating accuracy.

Rigid Regularization Minimizing Eq. 4 is the well-known least squares problem, which can be solved by linear regression. A straightforward way to improve performance is to impose regularization, which can be expressed using Eq. 5, where λ controls the regularization strength, the value of which is adaptive. The dimension of HOG feature of a single point is 128; for a face shape consisting of 68 landmarks, the number could be as large as $10K+$. Without regularization, the linear regression model is subject to over-fitting, which can be observed clearly in the experiment. Furthermore, regularization gives rise to more iterations and slows down the convergence rate, thus making the first two modifications possible. The regularization strength should be adaptive in order to work with our coarse-to-fine principle. Concretely, in early stages, strong regularization is preferred to construct a robust model, while later on, a lower regularization strength is favorable to better fit the true shape S_*. In short, with regularization, we achieve a more flexible and robust model.

$$\arg\min_{R,b}(\|\Delta x_*^i - R_k\phi_k - b_k\|_2^2 + \lambda\|R_k\|_2^2) \tag{5}$$

3.3 Summary

To sum up, we propose three modifications to SDM in total: 1. adaptive feature extracting block; 2. adaptive regression mechanism; 3. rigid regularization. With these extensions, SDM better satisfies the coarse-to-fine criterion of landmark locating. Essentially, they are inter-related rather than independent. The involvement of regularization which slows down convergence, allows an adaptive regression mechanism as well as finer selection of block sizes. Also, the strength of regularization should decrease with the global-to-local regressions going on. It is worth mentioning that the computation complexity is reduced by applying local regression instead of simply global regression. When tackling faces with p landmarks, the dimension of one-level regressor R is $2p \times (128p + 1)$ (ignoring the bias term b), while the sum of all part (local) regressor is $\Sigma_{i=1}^{k} 2p_i \times (128p_i + 1), \Sigma_{i=1}^{k} p_i = p$.

4 Experiment and Comparison

In this section, we evaluate the performance of our method from two aspects. First of all, we validate the effectiveness of each modification on SDM [15], then compare that with state-of-the-art methods using two challenging datasets. The error normalized by the inter-pupil distance is used as an evaluation metric, as proposed in [5]. Note that the error is a percentage of the pupil-distance, and we omit the notation % in our reported results for clarity.

In the experiment, our modifications introduce a few adjustable parameters: feature block (normalized by the face size), regularization strength (represented as a vector of which the length equals the number of iterations) and adaptive regression parameter (noted as a three-dimensional vector $[a, b, c]$, which consists of the number of iterations in global regression, part regression and local regression respectively, where $a + b + c = \textit{iteration time}$). In the original algorithm, SDM converges in $4 - 5$ steps. The modifications we adopt would decrease the convergence rate, hence we adopt $1 - 2$ more steps for fair comparison in the experiment.

4.1 Comparison with SDM

In this section, we validate the effect of modifications we proposed using two common facial datasets: LFPW[5] and LFW[26], which correspond to different situations of landmark locating. Experimental configurations are described below.

LFW consists of 13232 images in total. We adopt 10-point annotations and face bounding boxes released by [17]. Among the 13232 images, 10000 are selected randomly as the training data, with the remaining 3232 used for testing. The images in LFW dataset are low-resolution, hence the amounts of landmarks are relatively small.

LFPW are released by Belhumeur et al[5], containing 1432 face images, 1132

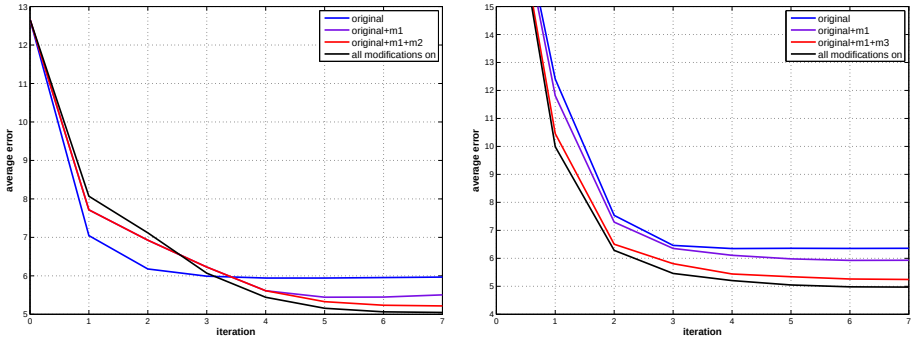
for training and 300 for testing, each of which has 35 fiducial landmark annotations. Since only image URLs are available and some links disappeared as time passed, we failed to download sufficient data for experiments. Thanks to [7], a version of 1035 images is available (each annotated with 68 landmarks), 881 of which are for training and the rest 224 for testing. Since the size of the training data is small, we employ data augmentation to improve generalization ability by initializing each training image multiple times. In this way, the descent direction varies to generate multiple training samples. We enlarge the training data 10 times in our experiment.

In our analysis, feature extracting block size and regularization strength decrease gradually. The adaptive regression parameter follows the global-to-local principle. All the parameter settings are adaptive to the coarse-to-fine criterion in landmark locating. The parameters we adopt are shown in Table 1 (the original SDM in not included). Specifically, for testing the original SDM, we adopt a fixed parameter which gives the best performance: a block size of 0.16 (normalized by face size) for feature extracting.

Table 1. Parameter configuration on two data sets.

data set	LFW	LFPW
number of landmarks	10	68
image size	$100 * 100$	$400 * 400$
iteration	7	7
adaptive block	$[0.48, 0.32, 0.16, 0.08, 0.04, 0.04, 0.02]$	$[0.24, 0.20, 0.16, 0.12, 0.08, 0.04, 0.02]$
adaptive regression	$[3, 0, 4]$	$[2, 2, 3]$

The two graphs in Fig. 3 illustrate the improvements brought by our three modifications on LFW and LFPW datasets respectively, from which we can observe several similarities. First of all, it is noticed that the original SDM converges faster than others (within 4 iterations), for its average error starts to remain stable earlier in both of the graphs. Also, on both datasets we achieve over 15% error reduction compared to the original SDM method. Fig. 3(a) alone shows that the effect of adaptive block is the most obvious on the LFW dataset, with the error declining from 5.96 to 5.5. In contrast, the curves of Fig. 3(b) indicate that, with the LFPW dataset of 68 landmarks, the rigid regularization applied contributes the most among the three modifications. An explanation is that as the amount of landmarks is large, the initially predicted shape is further from the ground truth, so a stronger global shape constraint is necessary. In this case, more iterations of global regression (instead of part or local regression) should be implemented and strong regularization is preferred in the mean time. Overall, it turns out that with each modification applied, the average error decreases more or less, which demonstrates that all three modifications proposed are effective in improving the performance of the original SDM.



(a) (b)

Fig. 3. Curves of average error versus iteration. 3(a) on LFW dataset, 3(b) on LFPW dataset. m1: modification of adaptive feature extracting block; m2: modification of adaptive regression mechanism;m3: modification of rigid regularization.

4.2 Comparison with State-of-the-art

In this section, we compare our method with state-of-the-art on two challenging datasets. These datasets show large variations in face shape, appearance, and number of landmarks, which is briefly introduced below. In the experiment, the average of normalized error of all landmarks and images is reported and compared. The adaptive regression parameter is set as$[2, 2, 3]$ and the block sizes (normalized by the length of face size) as $[0.24, 0.20, 0.16, 0.12, 0.08, 0.04, 0.02]$.

Helen (194 landmarks) [6] It contains $2,330$ high resolution, accurately labeled face images. In our experiment, we follow the setting in [6]: 2000 images for training and 330 images for testing. The most discriminative peculiarities that HENLEN dataset holds are its high resolution and the large number of facial landmarks.

300 Faces in-the-wild (68 landmarks) [7] It consists of existing datasets including LFPW[5], AFW [27], HELEN [6], XM2VTS [28] and IBUG. All images have been re-annotated by 68 landmarks. This dataset shows the largest variation in pose, expression, illumination, resolution due to cross-database annotation. To compare with state-of-the-art, we follow the setting in [14] which is dividing the training data into two parts in our own experiment. The training set is composed of the whole AFW dataset and the training data from LFPW and HELEN, with 3148 images in total.

The testing set is composed of IBUG and the test sets of HELEN and LFPW, with 689 images in total. The testing set are also divided into two subsets for further comparison: 1. the challenging IBUG subset, consisting of IBUG data; 2. the common subset, consisting of the testing sets from HELEN and LFPW.

Face bounding box is offered at the same time[1], produced by their in-house detector. All images are cropped and normalized into a fixed size of $400 * 400$ before operation.

Results Recently, a new approach based on boosted-regression called LBF [14] has been proposed. The accuracy appears to be the highest in the literature, outperforming SDM and ESR. Also, its speed is extremely fast. For comprehensive comparison, the performance of LBF is also reported, yet we should note that the aim of our method is to extend the simple, well-understood but efficient facial landmark location algorithm SDM.

From results illustrated in Table 2 and 3, we find that ESDM outperforms SDM by a large margin on all testing sets, again demonstrating the effectiveness of the modifications we proposed. In general, ESDM achieves comparable result with LBF.

Observing the average error on HELEN dataset reported in Table 2 alone, the performance of our method advances state-of-the-art over all. Regression-based methods[14, 21, 22, 15] significantly outperforms ASM-based[24, 29] approaches, proving the advantage of the flexible model of iterative alignment. During our experiment, it is also observed that our approach achieves slightly higher locating accuracy of certain parts than others, such as the contour of lip and the eyes. Table 3 illustrates the performance on 300 Faces in-the-wild dataset including two subsets. Similar to what is done in [15], we experiment on the original SDM algorithm and obtain results of $[7.32, 5.40, 15.32]$ on the fullset, common subset, challenging subset respectively. The rates are comparable with those reported in [15]. Our ESDM method shows superior results over SDM, ESR and LBF fast and is comparable to LBF. It is worth mentioning that on the common test subset, under the condition of identical training data, ESDM performs better than LBF, and we attribute this to the adequacy of training samples. Our method performs less desirably than LBF on the challenging test subset, mainly due to the drawback of the initialization mechanism of our method. In Fig. 4 and 5, some example images and comparison results from the common and the challenging subsets of 300 Faces in-the-wild are displayed.

5 Conclusion and Discussions

In this paper, we propose three modifications on SDM including: 1) Multi-scale HOG feature extraction in a coarse-to-fine manner; 2) Global to local regression of features in cascade; 3) Rigid Regularization applied to obtain more stable prediction results. Extensive experimental results demonstrate that each of the three modifications substantially improves the accuracy and robustness of the traditional SDM method. Furthermore, our method achieves favorable performance over most of the other state-of-the-art methods and comparable results

[1] http://ibug.doc.ic.ac.uk/resources/300-W

Table 2. Average error of several algorithm on HELEN dataset, results of other methods come from corresponding paper.

Method	HELEN (194 landmarks)
ESDM(Our method)	**5.32**
SDM[15]	5.85
ESR[22]	5.70
RCPR[21]	6.50
LBF[14]	5.41
LBF fast[14]	5.80
STASM[9]	11.1
CompASM[24]	9.10

Table 3. Error on 300-W(68 landmarks) data sets including two subsets, the result of SDM and ESR originate from [14].

Method	Full set	Common Subset	Challenging Subset
ESDM(our method)	6.44	**4.73**	13.92
SDM[15]	7.52	5.6	15.4
ESR[22]	7.58	5.28	17.0
LBF[14]	**6.32**	4.95	**11.98**
LBF fast[14]	7.37	5.38	15.5

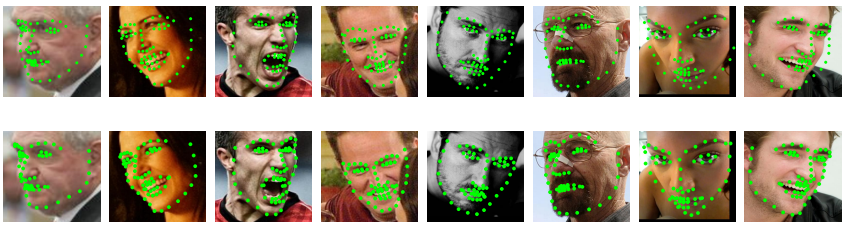


Fig. 4. Several examples from the Challenging Subset, some results of SDM are shown in the first row, corresponding results of our method are shown in the second line. Compared to SDM our method still achieve good performance on these extremely difficult images with large variation in pose, expression, partial occlusion.

Fig. 5. A comparative result on common subset from 300 Faces in-the-wild.

with the LBF method, on several challenging face datasets including LFPW, HELEN and 300 Faces in-the-wild.

Acknowledgement.

This work was partially sponsored by National Natural Science Foundation of China (NSFC) under Grant No. 61375031, No. 61471048, and No. 61273217. This work was also supported by the Fundamental Research Funds for the Central Universities, Beijing Higher Education Young Elite Teacher Project, and the Program for New Century Excellent Talents in University.

References

1. Deng, W., Hu, J., Guo, J., Cai, W., Feng, D.: Robust, accurate and efficient face recognition from a single training image: A uniform pursuit approach. Pattern Recognition **43** (2010) 1748–1762
2. Deng, W., Hu, J., Lu, J., Guo, J.: Transform-invariant pca: A unified approach to fully automatic face alignment, representation, and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **36** (2014) 1275–1284
3. Deng, W., Hu, J., Guo, J.: Extended src: Undersampled face recognition via intraclass variant dictionary. IEEE Trans. Pattern Anal. Mach. Intell. **34** (2012) 1864–1870
4. Deng, W., Hu, J., Zhou, X., Guo, J.: Equidistant prototypes embedding for single sample based face recognition with generic learning and incremental learning. Pattern Recognition **47** (2014) 3738–3749
5. Belhumeur, P.N., Jacobs, D.W., Kriegman, D., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE (2011) 545–552
6. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Computer Vision–ECCV 2012. Springer (2012) 679–692
7. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on, IEEE (2013) 896–903
8. Cootes, T.F., Edwards, G.J., Taylor, C.J., et al.: Active appearance models. IEEE Transactions on pattern analysis and machine intelligence **23** (2001) 681–685
9. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. In: Computer Vision–ECCV 2008. Springer (2008) 504–513
10. Saragih, J., Goecke, R.: A nonlinear discriminative approach to aam fitting. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, IEEE (2007) 1–8
11. Zhou, F., Brandt, J., Lin, Z.: (Exemplar-based graph matching for robust facial landmark localization)
12. Cristinacce, D., Cootes, T.: Automatic feature localisation with constrained local models. Pattern Recognition **41** (2008) 3054–3067
13. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: Computer Vision, 2009 IEEE 12th International Conference on, IEEE (2009) 1034–1041
14. Ren, S., Cao, X., Wei, Y., Sun, J.: (Face alignment at 3000 fps via regressing local binary features)

15. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 532–539
16. Sánchez-Lozano, E., De la Torre, F., González-Jiménez, D.: Continuous regression for non-rigid image alignment. In: Computer Vision–ECCV 2012. Springer (2012) 250–263
17. Dantone, M., Gall, J., Fanelli, G., Van Gool, L.: Real-time facial feature detection using conditional regression forests. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2578–2585
18. Cao, C., Weng, Y., Lin, S., Zhou, K.: 3d shape regression for real-time facial animation. ACM Trans. Graph. **32** (2013) 41
19. Valstar, M., Martinez, B., Binefa, X., Pantic, M.: Facial point detection using boosted regression and graph models. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 2729–2736
20. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE (2013) 3476–3483
21. Burgos-Artizzu, X.P., Perona, P., Dollár, P.: Robust face landmark estimation under occlusion, ICCV (2013)
22. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2887–2894
23. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, IEEE (2010) 1078–1085
24. Efraty, B., Huang, C., Shah, S.K., Kakadiaris, I.A.: Facial landmark detection in uncontrolled conditions. In: Biometrics (IJCB), 2011 International Joint Conference on, IEEE (2011) 1–8
25. Baker, S., Matthews, I.: Lucas-kanade 20 years on: A unifying framework. International journal of computer vision **56** (2004) 221–255
26. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst (2007)
27. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 2879–2886
28. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: Xm2vtsdb: The extended m2vts database. In: Second international conference on audio and video-based biometric person authentication. Volume 964., Citeseer (1999) 965–966
29. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer vision and image understanding **61** (1995) 38–59