Cascaded Shape Space Pruning for Robust Facial Landmark Detection

Xiaowei Zhao, Shiguang Shan, Xiujuan Chai, Xilin Chen
Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing 100190, China
{xiaowei.zhao,shiguang.shan,xiujuan.chai,xilin.chen}@vipl.ict.ac.cn

Abstract

In this paper, we propose a novel cascaded face shape space pruning algorithm for robust facial landmark detection. Through progressively excluding the incorrect candidate shapes, our algorithm can accurately and efficiently achieve the globally optimal shape configuration. Specifically, individual landmark detectors are firstly applied to eliminate wrong candidates for each landmark. Then, the candidate shape space is further pruned by jointly removing incorrect shape configurations. To achieve this purpose, a discriminative structure classifier is designed to assess the candidate shape configurations. Based on the learned discriminative structure classifier, an efficient shape space pruning strategy is proposed to quickly reject most incorrect candidate shapes while preserve the true shape. The proposed algorithm is carefully evaluated on a large set of real world face images. In addition, comparison results on the publicly available BioID and LFW face databases demonstrate that our algorithm outperforms some state-of-the-art algorithms.

1. Introduction

Accurately detecting facial landmarks in images is essential for many computer vision tasks, such as face recognition, facial expression recognition, 3D face modeling and face animation. Yet, locating facial landmarks in face images captured under unconstrained real world environment remains challenging, due to tremendous variations in facial appearance caused by pose, lighting, partial occlusion and so on.

Generally speaking, given a face image, the goal of facial landmark detection is to find the most correct shape (in terms of the concatenation of the landmarks coordinates) from all possible landmarks configurations according to some criteria. To achieve this goal, lots of methods are proposed in recent years. One of the most popular methods is the cascaded AdaBoost framework [26, 28]. In this kind of method, the facial landmarks are detected separately. Typically, it learns a classification function and computes a confidence for each position in the image. The image position, which has the largest confidence, is determined as the target facial landmark. However, one drawback of this kind of method is that it is insufficient to reliably detect facial landmarks just using local texture information, especially under complex environment. For example, there might be many image positions which look locally like the mouth corner if they are not observed in a large context. As a result, it is inherently difficult to detect the corners of mouth due to this ambiguity of local image patches. To address this problem, the relationship among facial landmarks should be utilized to eliminate the false positive detections.

Another kind of very popular facial landmark detection methods is the Active Shape Model (ASM) [6] and Active Appearance Model (AAM) [4]. In addition, many variants of ASM and AAM [2, 5, 8, 16, 17, 20, 29] are proposed to further improve the accuracy of facial landmark detection. Typically, in these methods, the shape configuration of facial landmarks is required to satisfy some statistical constraints, which are characterized by the Point Distribution Model (PDM) [6]. In order to find the optimal shape parameters, varying optimization criteria are designed. For example, in original AAM [4], the shape parameters are estimated through minimizing the residual between the face appearance and the synthesized face template. In work [16], Liu *et al*. discriminatively learn a classification function based on the global face appearances. The optimal shape parameters are achieved through maximizing the classification score. In addition, works [5, 6, 8, 17, 20] essentially formulate the facial landmark detection as a posterior maximization problem and solve it through iteratively search or EM-like algorithms. However, it is well known that this kind of method is usually prone to local maximum due to the model expression power and the optimization strategies (*e.g.*, gradient descent or EM), which are sensitive to the initialization.

In recent years, there also appeared a few methods, which learn a regression function that directly maps the global or local image appearance to the target facial land-

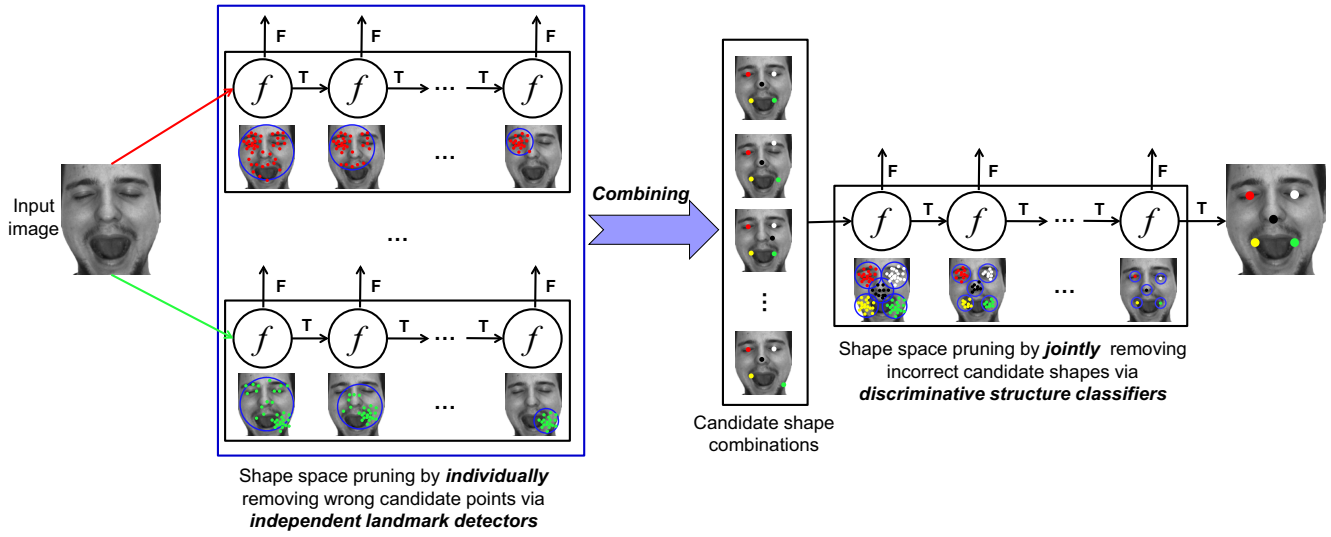


Figure 1. Overview of our robust facial landmark detection algorithm by cascaded shape space pruning. For the convenience of illustration, only five fiducial facial landmarks (e.g., the centers of eyes, nose tip and the corners of mouth) are used in this figure to describe the proposed algorithm.

marks [3, 7, 10, 19, 25]. For example, approaches in [7, 25] learn regressors which map the local image patches to the individual target landmarks. In addition, approaches in [3, 10] try to map the image appearance to the target shape configurations using conditional regression forest [10] or Boosting regression [3]. In these methods, the inherent geometric constraint among facial landmarks is implicitly encoded into the regressor. However, as demonstrated in [19], it is challenging to directly learn such an ideal regression function which can accurately predict a high dimensional shape from the image appearances, which usually present complex non-linear variations.

In short, above-mentioned facial landmark detection algorithms still have certain drawbacks in obtaining the optimal shape configurations, such as the local maximum problem of ASMs and AAMs and the ambiguity problem of individual landmark detectors, *etc*. To address these problems, in this paper, a novel facial landmark detection algorithm is proposed, which can *efficiently* achieve the *globally* optimal shape configuration from the *entire* candidate shape space. Specifically, instead of learning one strong criterion function, our algorithm learns a sequence of discriminative criterion functions in a cascaded structure. In each stage, part of the incorrect shape configurations is filtered out efficiently from the candidate shape space. In our implementation, the candidate shape space is firstly pruned by *individually* removing impossible positions of each landmark with separate landmark detector. Subsequently, in the later stages, all of the facial landmarks are considered as a whole and *jointly* evaluated by a discriminative structure classifier, which is learned using Structured Output SVM (SOSVM) [24]. Based on the discriminative structure classifier, an efficient pruning strategy is proposed to remove the incorrect candidate shape configurations quickly at the global shape level. Finally, in the remaining compact candidate shape space, the *globally* optimal shape configuration can be easily obtained via non-maximum suppression. We evaluate our algorithm in detail and compare it with a range of commonly used facial landmark detection algorithms. Experiments results show that our algorithm outperforms competitive algorithms on the LFW [14] and BioID [15] face databases.

Briefly speaking, the main contributions of this paper are:

- We propose a novel coarse-to-fine shape space pruning algorithm for robust facial landmark detection, which can *progressively* filter out the incorrect candidate shapes in a cascaded structure.
- We propose an algorithm to *jointly* assess the shape configurations and efficiently reject the incorrect candidate shapes based on a discriminative structure classifier.
- Our algorithm can efficiently achieve the target shape configuration from the entire candidate shape space without requiring *initialization*.

The remaining part of this paper is organized as follows. Section 2 gives a brief overview of our robust facial landmark detection algorithm. Section 3 presents more details of our algorithm, including shape space pruning by individually and jointly removing incorrect positions for facial landmarks. Section 4 reports the experimental results and

also the comparisons with the state-of-the-art methods. Section 5 concludes the paper.

2. Basic idea

In this section, we will first give a formulation of the facial landmark detection problem and then briefly describe the basic idea of our algorithm for solving the problem.

As described in Section 1, the task of facial landmark detection is to find the optimal shape $s^* = [x_1, y_1, x_2, y_2, \ldots, x_n, y_n]$ from the *entire* candidate shape space $\mathcal{S}$ according to some optimization criterion, where n is the number of facial landmarks. Specifically, the goal can be formulated as follows:

$$s^* = \arg\max_{s \in \mathcal{S}} f(s) \tag{1}$$

where f is the optimization criterion. It is hard to directly learn such an ideal optimization criterion f and efficiently estimate the *globally* optimal solution over a high dimensional shape space. Therefore, previous optimization-based methods are concerned with efficient *local* maximization from an initial guess [4, 5, 6, 8, 16, 17, 20].

In this paper, we try to estimate the *globally* optimal landmarks configuration through *progressively* filtering out the incorrect shapes. Specifically, instead of learning one strong criterion function f, we try to learn a sequence of criterion functions $f_1, f_2, \ldots, f_t$ in a cascaded structure. In each cascade stage i, most incorrect candidate shapes are fast rejected by criterion function f_i, *i.e.*, predicting a shape subspace $\mathcal{S}_i$ which satisfies that $\mathcal{S}_1 \supset \mathcal{S}_2 \supset \cdots \supset \mathcal{S}_i \supset \cdots \supset \mathcal{S}_t$. Ideally, the predicted shape subspaces are supposed to contain the true shapes and become as compact as possible.

The overview of our algorithm is shown in Figure 1. Given an input face image, the candidate shape space is firstly pruned by removing incorrect candidate points for each facial landmark. Specifically, for landmark l_i, if we remove one of its candidate points c_i, then the shapes with $l_i = c_i$ will be rejected. If there are N_i candidate points remained for landmark l_i, the size of candidate shape subspace is reduced to $\prod_{i=1}^{n} N_i$, which is still very huge. More implementation details are given in Section 3.1.

Subsequently, in the remaining cascade stages, the candidate shape space are further pruned by jointly removing the incorrect candidate shape configurations. Specifically, all of the facial landmarks are jointly modeled as a tree structure, as shown in Figure 2. In addition, the quality of shape configuration can be evaluated by a discriminative structure classifier. Based on the discriminative structure classifier, a very efficient shape space pruning strategy is applied to reject incorrect candidate shapes quickly. The implementation details of the discriminative structure classifier and the shape space pruning algorithm are described in Section 3.2.

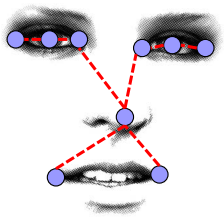


Figure 2. Tree structured model $T = (V, E)$ for jointly modeling the geometric constraints among facial landmarks. Each node in V represents a facial landmark, and the edges in E characterize the geometric shape deformation among facial landmarks. Nine facial landmarks are studied in this paper.

Finally, in the pruned and compact enough shape space, the optimal shape configuration can be easily obtained through the non-maximum suppression fusion technique.

3. Facial landmark detection by cascaded shape space pruning

In this section, we describe the details of our cascaded shape space pruning algorithm.

3.1. Shape space pruning by individually removing landmark candidates

In our algorithm, the candidate shape space is firstly pruned by individually removing candidate points for each landmark. To achieve this purpose, independent landmark detectors are trained to reject the incorrect candidate points.

For each facial landmark l, a landmark detector is trained using the Real AdaBoost classifier. Specifically, the Haar-like feature is used to characterize the local texture around the target facial landmark. The Look-Up-Table classifier is exploited as the weak classifier. Through combing the weak classifiers, a strong classifier is learned.

With the learned landmark detector, each candidate point c is assigned a confidence. If the confidence of candidate point c is lower than a certain threshold, it should be rejected. Correspondingly, the candidate shape configurations with $l = c$ should be rejected. In addition, several landmark detectors are cascaded for faster candidate points removing, as done in [26].

In this stage, although most of the candidate points can be removed through sliding window detection, there are still lots of false positive detections remained due to tremendous variations of image appearance. They will be further removed efficiently in the subsequent stages.

3.2. Shape space pruning by jointly removing candidate landmark positions

In this subsection, all of the facial landmarks are considered as a whole and jointly assessed by a discriminative

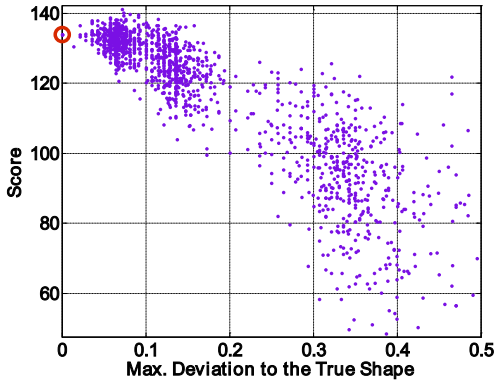


Figure 3. Score distribution of a real $F(I, s)$ on a face image. The x-axis represents the maximum deviation (normalized by eye distance) to the true shape, and the y-axis represents the score returned by F. The point marked with red circle is the true shape.

structure classifier. Besides, an efficient pruning strategy is proposed to fast reject the incorrect candidate shapes.

3.2.1 Face shape assessment via discriminative structure classifier

The basic idea of our algorithm evaluating the quality of candidate shape is to learn a discriminative structure classifier $F : \mathcal{X} \times \mathcal{S} \mapsto \mathbb{R}$ over image-shape pairs. For each image-shape pair (I, s), the learned classifier F outputs a score, with the constrains that the score of a true shape $\hat{s}$ should be greater than that of any other wrong shape $s \in \mathcal{S} \backslash \hat{s}$:

$$F(I, \hat{s}) > F(I, s). \tag{2}$$

In this paper, $F(I, s)$ is modeled as a linearly parameterized function:

$$F(I, s) = \langle w, \Psi(I, s) \rangle \tag{3}$$

where $\Psi(I, s)$ is the feature vector extracted from image I according to landmark configuration s, and w is the parameter vector.

As shown in Figure 2, the geometric relationship among facial landmarks is modeled as a tree-structured model $T = (V, E)$ [11, 30], where V is the set of facial landmarks, E is the set of edges connecting facial landmarks. More specifically, given a face image I and a shape s, $F(I, s) = \langle w, \Psi(I, s) \rangle$ is modeled as the combination of local textures and global shape deformations among facial landmarks:

$$\begin{aligned} &\langle w, \Psi(I, s) \rangle = \langle w_{tex}, \Psi_{tex}(I, s) \rangle + \langle w_{shape}, \Psi_{shape}(I, s) \rangle \\ &= \sum_{i \in V} \langle w^i_{tex}, \Psi^i_{tex}(I, s_i) \rangle + \sum_{jk \in E} \langle w^{jk}_{shape}, \Psi^{jk}_{shape}(I, s_j, s_k) \rangle \end{aligned} \tag{4}$$

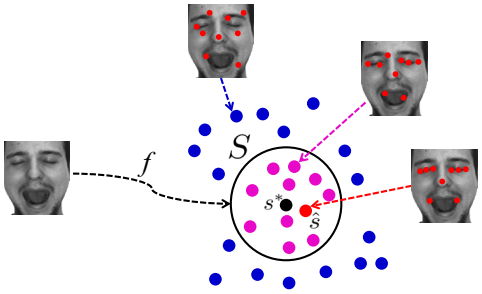


Figure 4. Illustration of the shape space pruning at the global shape level. f is a prediction function. It is expected that the true shape $\hat{s}$ lies in the pruned shape space S in a very high probability, and most incorrect shapes are excluded.

where $\Psi^i_{tex}(I, s_i)$ is the local texture feature (*e.g.*, Histograms of Oriented Gradients (HOG) [9], Local Binary Patterns (LBP) pyramid [27], *etc.*) extracted around the i-th landmark, and $\Psi^{jk}_{shape}(I, s_j, s_k)$ is the shape deformation between the j-th and k-th landmarks, which is defined as a deformation vector:

$$\Psi^{jk}_{shape}(I, s_j, s_k) = (dx, dy, dx^2, dy^2) \tag{5}$$

where $dx = (s_{j_x} - s_{k_x})$, and $dy = (s_{j_y} - s_{k_y})$.

Theoretically, the optimal shape configuration s^* can be obtained through maximizing Equation (3). Benefiting from the tree-structured model $T = (V, E)$, the *global* maximization of Equation (3) can be done *efficiently* with dynamic programming [11, 30].

However, in practice, it is hard to learn an ideal $F(I, s)$ such that the optimal shape s^* is exact or very close to the true shape $\hat{s}$. The reason mainly lies in the weak expression capability of image features and the linear modeling of $F(I, s)$. Figure 3 illustrates the score distribution of $F(I, s)$ on an example face image. Here, $F(I, s)$ is learned on a large set of real world face images. It can be observed that: (1) We can easily reject the shapes which are far away from the true shape (*e.g.*, the normalized maximum deviation is larger than 0.2); (2) The maximal score is not exactly at the true shape, but at the shape which is near to the true shape (*e.g.*, about 0.08 deviation).

3.2.2 Efficient shape space pruning

Based on the analysis in Section 3.2.1, in this subsection, we propose an approach to efficiently prune the candidate shape space at the global shape level.

As shown in Figure 4, instead of directly predicting the optimal shape configuration s^*, we turn to predict a shape subspace S, which excludes incorrect candidate shapes as much as possible and simultaneously includes the true shape $\hat{s}$ in a very high probability. The basic assumption

is that the optimal shape configuration s^* estimated according to $F(I, s)$ is not far away from the true shape $\hat{s}$. The predicted shape subspace can be expressed as:

$$S = \{s : | \ \|s - s^*\| < r\}, \tag{6}$$

which ensures that the true shape configuration $\hat{s}$ lies in S in a very high probability. Here, s^* is the center of shape space S, r is a threshold which constrains that the maximum deviation of $\hat{s}$ to the true shape s^*. The value of r balances the accuracy and efficiency of the shape space pruning. A smaller r means a more compact shape subspace and higher misdetection rate and vice visa. In practice, the value of r can be learned from the training set through setting the recall rate of the ground truth shapes.

It is important to note that in our strategy presented above, we do NOT filter out the incorrect candidate shapes simply by thresholding the score of $F(I, s)$, which although seems more natural for shape pruning. The most important reason is, in such an alternative strategy, every candidate shape have to be assessed by $\langle w, \Psi(I, s)\rangle$, which leads to high computation cost for a large shape space.

Similar to independent landmark detectors, the learned prediction functions are also cascaded for a coarse-to-fine shape space pruning. In each cascade stage, the optimal model parameters are learned in the pruned shape space. Along with the reduction of the solution space, we give more trust to the local texture rather than the global shape constraints among facial landmarks.

3.2.3 Learning of discriminative structure classifier

Our discriminative structure classifier is learned using the Structured Output SVM (SOSVM) algorithm [24].

Specifically, given M face images with labeled landmarks $\{(I^1, \hat{s}^1), (I^2, \hat{s}^2), \ldots, (I^M, \hat{s}^M)\}$, we can learn the model parameters w through minimization of a constrained quadratic optimization problem:

$$\begin{aligned} &\min_{w,\xi\geq 0} \frac{\lambda}{2}\|w\|^2 + \xi \\ &s.t. \forall (s^1, s^2, \ldots, s^M) \in \mathcal{S}^M : \\ &\frac{1}{M}\sum_{i=1}^{M}\langle w, \Psi(I^i, \hat{s}^i) - \Psi(I^i, s^i)\rangle \geq \frac{1}{M}\sum_{i=1}^{M}\Delta(\hat{s}^i, s^i) - \xi \end{aligned} \tag{7}$$

where $\Delta(\hat{s}^i, s^i)$ is a loss function which measures the loss of a shape s^i if the expected shape is $\hat{s}^i$, λ is a regularization term. Intuitively, the constraints in Equation (7) requires that for a training sample pair $(I^i, \hat{s}^i)$, $F(I^i, \hat{s}^i)$ has to produce a score that is higher than the score of any other pair (I^i, s^i) by at least $\Delta(\hat{s}^i, s^i)$. In our situation, the loss function $\Delta(\hat{s}^i, s^i)$ is defined as the average deviation over all facial landmarks:

$$\Delta(\hat{s}^i, s^i) = \frac{1}{n}\sum_{j=1}^{n}\|\hat{s}^i_j - s_j\|, \tag{8}$$

where n is the number of facial landmarks.

It is worth noting that, our model is different from [30], which only constrains the score of positive sample (face with true shape) greater than 1 while that of negative sample (non-face with any shape) smaller than -1. For the same face image, they do not have constrains on the relation between the true shape $\hat{s}$ and outlier shapes $s \in \mathcal{S}\backslash\hat{s}$ at all. However, this is crucial for learning a good landmark locator.

Equation (7) can be solved by Bundle Method for Regularized Risk Minimization (BMRM) optimization algorithm, which is a generic method for minimization of regularized convex functions [23].

4. Experiments

4.1. Datasets and evaluation metric

In this subsection, we describe the training set and the testing set for our algorithm. To train the individual landmark detectors and the structure classifier, we collect about 7,000 face images from multiple databases, such as CMU PIE [21], FRGC v1 [18], CAS-PEAL [12], FG-NET Aging [1], and CMU Multi-PIE [13].

To validate the effectiveness of the proposed algorithm, we compare it with the state-of-the-art methods on the BioID [15] and LFW [14] face databases, which are definitely excluded from our training set. Specifically, the BioID database consists of 1,521 images of frontal faces taken in uncontrolled conditions using a web camera. It features a large variety of illuminations, backgrounds and face sizes. The LFW database is collected from the wild conditions and varies in pose, lighting conditions, expression, partial occlusion and so on. Specifically, it contains 13,233 facial images of 5,749 subjects. The landmark annotations are available at http://www.dantone.me/?page_id=38.

In our experiments, the normalized root-mean-squared error (NRMSE) relative to the ground truth is adopted as the error measurement for the facial landmark detection. The NRMSE is given as a percentage, computed by dividing the root mean squared error by the distance between the two eye centers. The cumulative distribution function (CDF) of NRMSE is used to evaluate the performance of facial landmark detection algorithm.

4.2. Training

In our experiments, nine facial landmarks are localized and evaluated, which include two eye centers, four eye corners, two mouth corners and the nose tip. Specifically, to train the detectors for each individual facial landmark, we

Table 1. Analysis of cascaded shape space pruning.

Landmark	Candidates Size			Maximum Deviation			Detection Rate (NRMSE ≤ 0.10)		
	S1	S2	S3	S1	S2	S3	S1	S2	S3
Left eye center	90	52	22	0.46	0.14	0.10	99.2%	98.2%	96.5%
Right eye center	85	51	22	0.41	0.14	0.10	99.1%	98.0%	96.1%
Nose tip	90	53	22	0.44	0.15	0.10	99.2%	98.1%	96.2%
Left mouth corner	87	52	22	0.44	0.15	0.10	99.2%	98.3%	96.4%
Right mouth corner	84	50	22	0.45	0.14	0.10	99.2%	98.2%	96.3%
Outer corner of left eye	88	50	22	0.47	0.14	0.10	99.2%	98.0%	96.1%
Inner corner of left eye	94	52	22	0.49	0.15	0.10	99.0%	97.6%	95.7%
Inner corner of right eye	97	52	22	0.50	0.15	0.10	99.0%	97.5%	95.5%
Outer corner of right eye	91	52	22	0.47	0.15	0.10	99.1%	98.0%	96.1%

generate positive samples by cropping image patch which is centered at the ground truth landmarks, and synthesize more positive samples by some transformations. Negative samples are image patches shifted 5∼8 pixels away from the manually labeled ground truth position. The cascaded Real AdaBoost algorithm is exploited to train the individual facial landmark detectors.

To train the discriminative structure classifier, the normalized face images and the corresponding ground truth shapes are used as the input of the SOSVM learning algorithm. Specifically, in our implementation, the nine facial landmarks form a tree structure and two different structure classifiers are cascaded to progressively prune candidate shape space. The first one is trained to efficiently filter most outlier shapes, which are far away from the true shape, by using computational efficient Pyramid LBP feature. For the second one, more discriminative HOG feature is applied for more accurate prediction in the pruned shape space.

4.3. Algorithm analysis

In this subsection, we verify the effectiveness and efficiency of the proposed cascaded shape space pruning algorithm on 13,233 face images from the LFW face database.

Here, for the convenience of expression, we briefly note the procedure that pruning shape space by independent landmark detectors as "S1". In addition, the subsequent two discriminative structure classifiers as noted as "S2" and "S3" respectively.

4.3.1 Efficiency of shape space pruning

In order to analyze the efficiency of shape space pruning, three kinds of criteria are designed in our experiments:

- Candidates Size. This criterion represents the number of candidate points remained for each landmark after one stage.
- Maximum Deviation. It is the maximum deviation of the candidate points to the ground truth. A smaller value of this criterion represents a more compact candidates space.
- Detection Rate. This criterion represents the successful detection rate of the candidate point set. If the minimum deviation of the candidate points to the ground truth is smaller than a certain threshold (NRMSE = 0.10 is used in our experiments), we think the candidate points contains a successful detection.

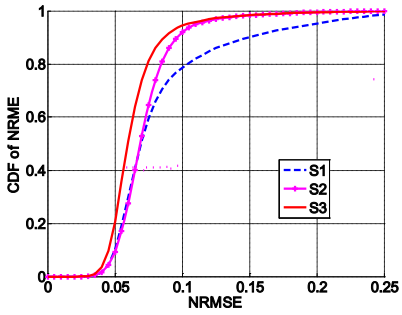
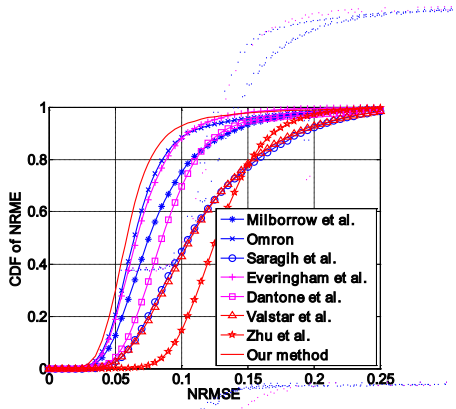


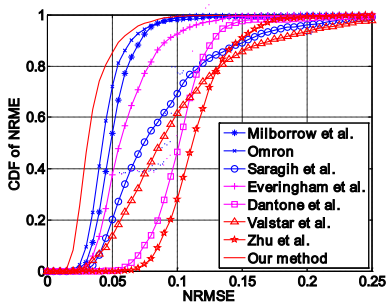
Figure 5. Performance improvement after each cascade stage on the LFW face database.

In addition, the deviation to the ground truth is normalized by dividing the eye distance. The criteria "Candidates Size" and "Maximum Deviation" are averaged over the whole test set.

The experimental results are shown in Table 1. It can be observed that: (1) The remaining candidate number is greatly reduced after each stage; (2) The maximum deviation of each landmark to the ground truth is reduced from about $(0.40 \sim 0.50)$ to 0.10; (3) There is only very small percentage (about $1\% \sim 2\%$) of ground truth are wrongly rejected in each stage along with the pruning of the shape space.



(a) Results on the LFW face database.



(b) Results on the BioID face face database.

Figure 6. Performance comparisons on two publicly available face databases.

4.3.2 Performance improvement of each stage

In this subsection, we will validate the effectiveness of the cascaded shape space pruning to improve the performance of facial landmark detection.

The experimental results are shown in Figure 5. The average detection accuracy of nine facial landmarks is explored. It can observed from the experimental results that detection accuracy improves after each stage. Especially, when NRMSE is 0.10, the detection accuracy of "S2" is about 12% higher than detection accuracy of "S1". In addition, when NRMSE is 0.05, the detection accuracy of "S3" is about 10% higher than detection accuracy of "S2", which shows that we can get more accurate landmark position in the pruned shape space.

In addition, in comparison to the algorithm which just uses independent landmark detectors, our algorithm dose not increase the runtime much. On an Intel(R) Core(TM)2 2.93GHz machine, the average runtime of our algorithm on 13,233 face images (250×250 pixels) from the LFW face database is 150.0ms (133.9ms for independent landmark detection).

4.4. Comparisons with state-of-the-art methods

In recent years, some promising methods for robust facial landmark detection emerge [10, 17, 20, 22, 25, 30]. For the convenience of comparison, we briefly denote them by Dantone *et al*. [10], Omron, Milborrow *et al*. [17], Saragih *et al*. [20], Everingham *et al*.[22], Valstar *et al*. [25], Zhu *et al*. [30]. In this subsection, we compare our method with these state-of-the-art methods on the BioID and LFW face databases. It is important to note that the source codes or executable programs of the competitive methods are available from the internet. We just use the *default* parameters given by the authors. In addition, in order to compare with these methods fairly, we just compare the common six facial landmarks, *i.e.*, four eye corners and two mouth corners.

The performance comparisons with these state-of-the-art methods are shown in Figure 6. It can be observed that our method outperforms these state-of-the-art methods on these two face databases.

Some localization results of our method on some challenging example images are shown in Figure 7 and Figure 8 respectively. It can be observed that our method can locate the facial landmarks robust and accurately on images with exaggerate facial expression and partial occlusion.

5. Conclusion and future work

In order to improve the accuracy of facial landmark detection, a cascaded shape space pruning algorithm is proposed in this paper. Through progressively filtering the incorrect shape configurations, our algorithm can accurately and efficiently achieve the globally optimal shape configuration from the entire candidate shape space. Specifically, the candidate shape space is pruned by not only individually removing candidate points for each landmark but also jointly removing incorrect candidate shapes. To jointly assess the shape configurations, a discriminative structure classifier is learned using SOSVM. The effectiveness of our algorithm is analyzed on the real world LFW face database. Moreover, experimental results on the BioID and the LFW face databases show that our algorithm outperforms some state-of-the-art methods.

In the future work, we will apply our algorithm to some other problems, such as the anatomic segmentation and hand radiographs localization in medical images.

Acknowledgments

This work is partially supported by Natural Science Foundation of China (NSFC) under contract Nos. 61025010, 61222211 and 61173065; and the FiDiPro program of Tekes.

References

[1] `http://www-prima.inrialpes.fr/FGnet/`, 2002.

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, pages 545–552, 2011.

Figure 7. Facial landmark detection results on example images from the LFW face database.

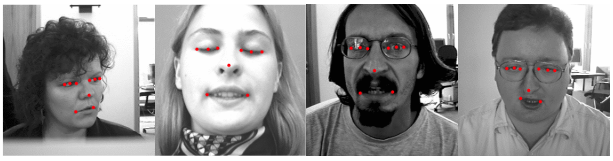


Figure 8. Facial landmark detection results on example images from the BioID face database.

[3] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, pages 2887–2894, 2012.

[4] T. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *IEEE TPAMI*, 23(6):681–685, 2001.

[5] T. Cootes, M. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *ECCV*, pages 278–291, 2012.

[6] T. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, et al. Active shape models-their training and application. *CVIU*, 61(1):38–59, 1995.

[7] D. Cristinacce and T. Cootes. Boosted regression active shape models. In *BMVC*, pages 880–889, 2007.

[8] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *PR*, 41(10):3054–3067, 2008.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005.

[10] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, pages 2578–2585, 2012.

[11] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005.

[12] W. Gao, B. Cao, S. Shan, X. Chen, D. Zhou, X. Zhang, and D. Zhao. The CAS-PEAL large-scale chinese face database and baseline evaluations. *IEEE TSMC-A*, 38(1):149–161, 2008.

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *IVC*, 28(5):807–813, 2010.

[14] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. In *Workshop on Faces in'Real-Life'Images: Detection, Alignment, and Recognition*, 2008.

[15] O. Jesorsky, K. Kirchberg, and R. Frischholz. Robust face detection using the hausdorff distance. In *AVBPA*, pages 90–95, 2001.

[16] X. Liu. Generic face alignment using boosted appearance model. In *CVPR*, pages 1–8, 2007.

[17] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *ECCV*, pages 504–513, 2008.

[18] P. J. Phillips, P. J. Flynn, T. Scruggs, K. W. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *CVPR*, pages 947–954, 2005.

[19] J. Saragih. Principal regression analysis. In *CVPR*, pages 2881–2888, 2011.

[20] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, 2011.

[21] T. Sim, S. Baker, and M. Bsat. The cmu pose, illumination, and expression (PIE) database. In *AFGR*, pages 46–51, 2002.

[22] J. Sivic, M. Everingham, and A. Zisserman. "Who are you?" – learning person specific classifiers from video. In *CVPR*, 2009.

[23] C. H. Teo, S. Vishwanthan, A. J. Smola, and Q. V. Le. Bundle methods for regularized risk minimization. *JMLR*, 11:311–365, 2010.

[24] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. *JMLR*, 6(2):1453, 2006.

[25] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, pages 2729–2736, 2010.

[26] P. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.

[27] W. Wang, W. Chen, and D. Xu. Pyramid-based multi-scale lbp features for face recognition. In *CMSP*, pages 151–155, 2011.

[28] X. Zhao, X. Chai, Z. Niu, C. Heng, and S. Shan. Context modeling for facial landmark detection based on non-adjacent rectangle (NAR) haar-like feature. *IVC*, 30(3):136–146, 2012.

[29] X. Zhao, S. Shan, X. Chai, and X. Chen. Locality-constrained active appearance model. In *ACCV*, pages 636–647. 2013.

[30] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012.