

# Sign-Correlation Partition Based on Global Supervised Descent Method for Face Alignment

Yongqiang Zhang<sup>1(✉)</sup>, Shuang Liu<sup>2</sup>, Xiaosong Yang<sup>2</sup>, Daming Shi<sup>1</sup>,  
and Jian Jun Zhang<sup>2</sup>

<sup>1</sup> Harbin Institute of Technology, Harbin, China  
[seekever@foxmail.com](mailto:seekever@foxmail.com)

<sup>2</sup> Bournemouth University, Poole, UK

**Abstract.** Face alignment is an essential task for facial performance capture and expression analysis. As a complex nonlinear problem in computer vision, face alignment across poses is still not studied well. Although the state-of-the-art Supervised Descent Method (SDM) has shown good performance, it learns conflict descent direction in the whole complex space due to various poses and expressions. Global SDM has been presented to deal with this case by domain partition in feature and shape PCA spaces for face tracking and pose estimation. However, it is not suitable for the face alignment problem due to unknown ground truth shapes. In this paper we propose a sign-correlation subspace method for the domain partition of global SDM. In our method only one reduced low dimensional subspace is enough for domain partition, thus adjusting the global SDM efficiently for face alignment. Unlike previous methods, we analyze the sign correlation between features and shapes, and project both of them into a mutual sign-correlation subspace. Each pair of projected shape and feature keep sign consistent in each dimension of the subspace, so that each hyperoctant holds the condition that one general descent exists. Then a set of general descent directions are learned from the samples in different hyperoctants. Our sign-correlation partition method is validated in the public face datasets, which includes a range of poses. It indicates that our methods can reveal their latent relationships to poses. The comparison with state-of-the-art methods for face alignment demonstrates that our method outperforms them especially in uncontrolled conditions with various poses, while keeping comparable speed.

## 1 Introduction

Face alignment is an important computer vision task, and plays a key role in many facial analysis applications, such as face recognition, performance-based facial animation, and expression analysis. It aims at locating predefined facial

---

**Electronic supplementary material** The online version of this chapter (doi:[10.1007/978-3-319-54187-7\\_19](https://doi.org/10.1007/978-3-319-54187-7_19)) contains supplementary material, which is available to authorized users.

landmarks (such as eye corners, nose tip, mouth corners) in face images automatically. Face alignment usually takes a face bounding box from a face detector as input, and fits initial landmarks positions into optimal locations.

Since the ground truth of shape is unknown during test, how to predict the shape increment from initial shape to real shape is a hard problem. Generally, The global or local face appearance is considered as extra constraints for optimization. Sufficient labeled face images are also very important for learning a reliable face alignment model. Recently there have been many methods proposed to face alignment. Most of them can be categorized into two groups according to the underlying model: generative models and discriminative models.

Typical generative models include Active Shape Model (ASM) [1], Active Appearance Model (AAM) [2], and their extensions [3–6]. In this type of methods, the optimization target is model parameters. It means searching the best parameters to generate the most fitting shape (facial landmarks). These methods mitigate the influence of various poses and illumination, but due to sub-optimization problem, they are sensitive to initialization and often tend to fail in the wild condition.

Recently discriminative models have shown better performance for face alignment. One discriminative method directly learns a mapping between facial appearance features and shapes. A kind of discriminative methods are based on local classifiers or response maps for landmarks [7, 8]. These methods deal with each landmark independently, and ignore the their relationship. As a popular discriminative model, current cascaded regression-based methods take all the landmarks as a whole, and solve a nonlinear optimization problem by cascaded regression theory. The main difference between cascaded regression and related boosting regression is that cascaded regression uses shape-indexed features extracted from the image according to the current estimated shape. Cascaded Pose Regression (CPR) [9] for pose estimation has been widely extended into face alignment in current works, represented by Explicit Shape Regression (ESR) [10], and Supervised Descent Method (SDM) [11]. It is noticed that SDM provides a theoretical explanation of the cascaded regression from the point of view of optimizing a nonlinear problem, as a significant achievement in cascaded regression-based methods.

Following the cascaded regression framework, many researchers focus on improving its efficiency and accuracy in uncontrolled conditions, including various poses, expressions, lighting and partial occlusions. Some of them can handle partial occlusions [12–14]. Some works mainly aim at speeding up the prediction process while keeping high accuracy [13, 15]. The choice and learning of shape-indexed features are also studied [15–17]. A series of regression methods have been employed into cascaded regression framework to deal with over-fitting and local minima problems in the wild condition, including ridge regression [18], Support Vector [19], Gaussian process [20, 21], Random Forest voting [14, 22, 23], Deep Neural Nets [16, 24, 25], and project-out cascaded regression [26].

Although these works have produced remarkable results on nearly frontal face alignment, it is still hard to locate landmarks across large poses and

expressions under uncontrolled conditions. The variation of poses leads to non-convex and multiple local minima problems. Especially Xiong et al. [27] theoretically addresses the limitation of SDM and proposes descent domain partition in feature and shape PCA space separately. Though their scheme works well for face tracking and pose estimation, it is not suitable for face alignment across various poses, because the ground truth of shapes or features are unknown and it is not able to find approximation due to lack of previous frame. A few recent works [12, 13, 28–31] begin to consider the influence of multiple poses. Most of them deal with the problem indirectly by random schemes or data augment, and they can only handle small changes in poses. How to solve non-convex and multiple local minima problems caused by large poses is still not well studied.

Inspired by Xiong's work [27], we proposed a novel sign-correlation subspace method for partitioning descent domains to achieve robust face alignment across poses. The main contributions of our work are: (1) The inherent relationship between poses space and appearance features or shapes space is explicitly obtained by sign-correlation reduced dimension strategy. The whole features and shapes spaces are projected into a mutual sign-correlation subspace, which mainly represents the variation of poses. (2) The decent domains partition is produced according to the signs of each dimension in this sign-correlation subspace. Since the face appearance features during cascaded regression, what we need to do for decent domains partition is to project features space into joint sign-correlation subspace and split whole sample space into different hyperoctants as decent domains. (3) Our method is validated on challenging face datasets, which includes face images from different poses. The results show that it can split complex sample space into homogeneous domains related to poses, thus a mutual manifold of feature and shape spaces are obtained. The experiments for face alignment indicate that our method get state-of-the-art performance for nearly frontal face images, and it is more robust on datasets with multiple poses, compared with current methods.

## 2 Related Work

### 2.1 Cascaded Regression to Face Alignment

Both of generative models and discriminative models have been studied for face alignment. As a typical generative model, ASM [1] is proposed to take advantage of prior knowledge from training datasets, which is one of the earliest data-driven model for shape fitting. PCA is used to build a linear combination model of major shape basis, and local textures around control points are also used for fitting the shape well. AAM [2] considers the global appearance rather than only local textures in ASM, and a PCA model is trained for global appearance while the shape PCA is trained at the same time. AAM can warp the initial shape and appearance into the current face very well due to both of its shape constraint and appearance constraint. There are also many methods based on them, like multi-view ASM [3], CLM [4], bilinear AAM [5] and tensor-based

AAM [6]. Since these methods are parametric models, it is hard to avoid a sub-optimization problem. Unexpected results often occur in the wild condition due to an inappropriate initialization.

Among discriminative models, cascaded regression based methods have shown more promising performance than local classifiers or response map based methods [7, 8] and generative models. ESR [10] uses shape indexed intensity difference features for face alignment based on CPR [9]. Moreover, SDM extracts shape-indexed SIFT features and learns a sequence of general descent maps from supervised training data, providing a solution when Newton Descent method is hard to be utilized for a not analytically differentiable nonlinear function or Hessian matrix is too large and not positive definite. Since SDM tends to average conflict descent directions over whole non-convex space, it is still limited in the wild scenes, like large poses, extreme expressions and partial occlusions.

Later research mainly focuses on improving performance based on ESR and SDM. Burgos-Artizzu et al. [12] integrate part visibility term into landmarks and presents interpolated shape-indexed features to tackle with occlusions and high shape variances. Kazemi et al. [32] estimate facial landmarks by learning an ensemble of regression trees (ERT) directly from a sparse subset of pixel intensities. Their ERT achieves millisecond performance and can handle partial or uncertain labels, but the correlation of shape parameters is little taken into account. Instead of least squares regression, Xing et al. [13] learn sparse Stage-wise Relational Dictionary (SRD) between facial appearances and shapes, which improves the robustness under different views and severe occlusions. Some recent research aims at choosing or learning shape-indexed features. Yan et al. [15] compare the performance of different local feature descriptors for face alignment, including SIFT, HOG, LBP and Gabor, and HOG shows the best results in their experiments. Ren et al. [33] build local binary features by learning regression random forest for each landmark independently, and then learn a global cascaded linear regressor with pre-built binary features. Deep Neural Networks [16, 24, 25] have also been studied for face landmark detection. DNNs-based methods fuse the feature description and networking training in a unified framework, but it is still a very hard task to tune many free parameters.

## 2.2 Multi-pose Face Alignment

More recently Xiong et al. [27] analyze the drawbacks of SDM and splits whole sample space into descent domains by PCA in both of feature and shape spaces, but it can not be used for face alignment across various poses due to unknown real shapes or features, which can be approximated by previous frame for face tracking and pose estimation tasks. There have been some works done to improve the SDM for face alignment. Feng et al. [28] propose random cascaded-regression copse, learning a sets of cascaded strong regressors corresponding to different subsets of samples and averaging all predictions of them as final output. Similarly, Yang et al. [29] propose random subspace SDM, randomly selecting a small number of dimensions from the whole feature space and training an ensemble of regressors in several feature subspaces. Liu et al. [18] modify traditional

SDM with multi-scale HOG features, global to local regression of features and rigid regularization to improve the accuracy and robustness. L2,1 norm based kernel SVR is presented by Martinez et al. [19] to substitute the commonly used least squares regressor, which improves the performance of face alignment across views. Gaussian process [20, 21] and Random Forest voting [14, 22, 23] are also introduced into cascaded regression framework. Zhang et al. [34] and Zhu et al. [30] further study hierarchical or coarse-to-fine searching for face alignment. Feng et al. [31] combine synthetic images with real images to train cascaded collaborative regression with dynamic weighting, handling the pose variations better. Fan et al. [17] combine projective invariant characteristic number with appearance based constraints and solve a quadratic optimization by the standard gradient descent. Though their method show well pose invariant, it can only handle a small number of landmarks. Tzimiropoulos presents project-out cascaded regression (PO-CR) [26] and extend the learn-based Newtons method further: Instead of learning directly a mapping from appearance features to non-parametric shapes, PO-CR learns a sequence of Jacobian and Hessian matrices based on parametric shape model. It shows noticeable results on the challenging datasets.

Some methods among them have begun to deal with the impact of poses, like RPCR [12], SRD [13], CCR [31], hierarchical localization [34] and coarse-to-fine searching [30]. However, few of them can give a clear interpretation for the correlation between poses and feature or shapes. Most of these methods loose the problem by different strategies, but how to achieve robust and accurate face alignment across poses is still a challenging task. Following previous methods, our work focuses on analyzing the underlying relationship between feature space and shape space, and finding a joint pose-related subspace for global supervised descent domains partition, to improve the performance of face alignment under challenging multi-pose conditions.

### 3 Sign-Correlation Subspace for Descent Domains Partition

#### 3.1 Descent Domain Partition Problem in SDM

In order to keep the completeness of our work, we first review the problem of SDM and the sign-correlation condition for existence of a supervised descent domain. According to SDM, setting one image  $\mathbf{d}$ ,  $p$  landmarks  $\mathbf{x} = [x_1, y_1, \dots, x_p, y_p]$ , a feature mapping function  $\mathbf{h}(\mathbf{d}(\mathbf{x}))$  corresponding to image  $\mathbf{d}$ , where  $\mathbf{d}(\mathbf{x})$  indexes landmarks in the image  $\mathbf{d}$ , the face alignment problem can be regarded as a optimization problem,

$$f(\mathbf{x}_0 + \Delta\mathbf{x}) = \|\mathbf{h}(\mathbf{d}(\mathbf{x}_0 + \Delta\mathbf{x})) - \phi_*\|_2^2 \quad (1)$$

where  $\phi_* = \mathbf{h}(\mathbf{d}(\mathbf{x}_*))$  represents the feature extracted according to correct landmarks  $\mathbf{x}_*$ , which is known in the training images, but unknown in the testing images. For initial locations of landmarks  $\mathbf{x}_0$ , we solve  $\Delta\mathbf{x}$ , which minimizes

the feature alignment error  $f(\mathbf{x}_0 + \Delta\mathbf{x})$ . Since the feature function is usually not analytically differentiable, it is hard to solve the problem with traditional Newtons methods. Alternatively, a general descent mapping can be learned from training datasets. The supervised descent method form is,

$$\mathbf{x}_k = \mathbf{x}_{k-1} - \mathbf{R}_{k-1}(\phi_{k-1} - \phi_*) \tag{2}$$

Since  $\phi_*$  of a testing image is unknown but constant, SDM modifies the objective to align with respect to the average one  $\bar{\phi}_*$  over training set, the update rule then is modified,

$$\Delta\mathbf{x} = \mathbf{R}_k(\bar{\phi}_* - \phi_k) \tag{3}$$

Instead of learning only one  $\mathbf{R}_k$  over all samples during one updating step, The Global SDM learns a series of  $\mathbf{R}^t$ , one for a subset of samples  $S^t$ , where the whole samples are divided into T subsets  $S = \{S^t\}_1^T$ .

A generic DM exists under these two conditions: (1)  $\mathbf{R}\mathbf{h}(\mathbf{x})$  is a strictly locally monotone operator anchored at the optimal solution (2)  $\mathbf{h}(\mathbf{x})$  is locally Lipschitz continuous anchored at  $\mathbf{x}_*$ . For a function with only one minimum, these normally hold. But a complex function might have several local minima in a relatively small neighborhood, thus the original SDM tends to average conflicting gradient directions. Therefore, the Global SDM proves that if the samples are properly partitioned into a series of subsets, there is a DM in each of the subsets. The  $\mathbf{R}_t$  for subset  $S_t$  can be solved with a constrained optimization form,

$$\min_{S, \mathbf{R}} \sum_{t=1}^T \sum_{i \in S^t} \|\Delta\mathbf{x}_* - \mathbf{R}_t \Delta\phi^{i,t}\|^2 \tag{4}$$

$$s.t. \Delta\mathbf{x}_*^i \mathbf{R}_t \Delta\phi^{i,t} > 0, \forall t, i \in S^t \tag{5}$$

where  $\Delta\mathbf{x}_*^i = \mathbf{x}_*^i - \mathbf{x}_k^i$ ,  $\Delta\phi^{i,t} = \bar{\phi}_*^t - \phi^i$ , and where  $\bar{\phi}_*^t$  - average all  $\phi_*$  over the subset  $S^t$ . Equation 5 guarantees that the solution satisfies DM condition 1. It is NP-hard to solve Eq. 4, so a deterministic scheme is proposed to approximate the solution. A set of sufficient conditions for Eq. 5 is given:

$$\Delta\mathbf{x}_*^i T \Delta\mathbf{X}_*^t > \mathbf{0}, \forall t, i \in S^t \tag{6}$$

$$\Delta\Phi^t T \Delta\phi^{i,t} > \mathbf{0}, \forall t, i \in S^t \tag{7}$$

where  $\Delta\mathbf{X}_*^t = [\Delta\mathbf{x}_*^{1,t}, \dots, \Delta\mathbf{x}_*^{i,t}, \dots]$ , each column is  $\Delta\mathbf{x}_*^{i,t}$  from the subset  $S^t$ ;  $\Delta\Phi^t = [\Delta\phi^{1,t}, \dots, \Delta\phi^{i,t}, \dots]$ , each column is  $\Delta\phi^{i,t}$  from the subset  $S^t$ .

Since the dot product of any two vectors within the same hyperoctant (the generalization of quadrant) is positive, an ideal sufficient partition can be regarded as that each subset  $S^t$  occupies a hyperoctant both in the parameter space  $\Delta\mathbf{x}$  and feature space  $\Delta\phi$ . However, this leads to exponential number of DMs. Assuming  $\Delta\mathbf{x}$  is  $n$ -dimension, and  $\Delta\phi$  is  $m$ -dimension, the number of subsets will be  $2^{n+m}$ . Moreover, if the number of all samples is small, there will be many empty subsets, and also the volume of some subsets will be too small to train.

It's known that as  $\Delta\mathbf{x}$  and  $\Delta\phi$  are embedded in a lower dimensional manifold for human faces. So dimension reduction methods (e.g. PCA) on the whole training set  $\Delta\mathbf{x}$  and  $\Delta\phi$  can be used for approximation. The Global SDM authors project  $\Delta\mathbf{x}$  onto the subspace expended by the first two components of  $\Delta\mathbf{x}$  space, and project  $\Delta\phi$  onto the subspace by the first component of  $\Delta\phi$  space. So there are  $2^{2+1}$  subsets in their work. It is a very naive scheme and not suitable for face alignment. Correlation-based dimension reduction theory can be introduced to develop a more practical and efficient strategy for low-dimension approximation of the high dimensional partition problem.

### 3.2 Sign-Correlation Subspace Partition

Interestingly, Xiong et al. [27] have proven that if one subset  $S^t$  satisfies: For any two samples  $\{\Delta\mathbf{x}^{i,t}, \Delta\phi^{i,t}\}, \{\Delta\mathbf{x}^{k,t}, \Delta\phi^{k,t}\}$  within  $S^t$ , the signs of each corresponding  $j - th$  dimension  $\{\Delta x_j^{i,t}, \Delta\phi_j^{i,t}\}$  between the samples keep the same,

$$\text{sign}(\Delta x_j^{i,t}, \Delta\phi_j^{i,t}) = \text{sign}(\Delta x_j^{k,t}, \Delta\phi_j^{k,t}), \forall i, k \in S^t, j = 1 : \min(n, m) \quad (8)$$

Then there must exist a DM  $\mathbf{R}^t$  in one updating step. Equation 8 provides a possible partition strategy: all the samples that follow Eq. 8 can be put into a subset, and there would be  $2^{\min(n,m)}$  subsets in total. It is noticed that there are two limits of this partition strategy: (1) It can not guarantee the samples lie in the same small neighborhood. In other words, even if  $\{\Delta\mathbf{x}^{i,t}, \Delta\phi^{i,t}\}, \{\Delta\mathbf{x}^{k,t}, \Delta\phi^{k,t}\}$  keep Eq. 8, the  $\Delta\mathbf{x}^{i,t}, \Delta\mathbf{x}^{k,t}$  may be very far from each other. (2) It only considers the dim-to-dim correlation of the first  $\min(n, m)$  dimensions in the  $\Delta\mathbf{x}$  space and  $\Delta\phi$  space, and other dimensions are ignored. The correlation of any  $j - th$  dimension of  $\Delta\mathbf{x}$  with a non-corresponding  $j' - th$  dimension  $j' \neq j$  of  $\Delta\phi$  is also ignored.

Considering the low dimensional manifold, the  $\Delta\mathbf{x}$  space and  $\Delta\phi$  space can be projected onto a medium low dimensional space with projection matrix  $\mathbf{Q}$  and  $\mathbf{P}$ , respectively, which keeps the projected vectors  $\mathbf{v} = \mathbf{Q}\Delta\mathbf{x}$ ,  $\mathbf{u} = \mathbf{P}\Delta\phi$  correlated enough: (1)  $\mathbf{v}$ ,  $\mathbf{u}$  lie in the same low dimensional space. (2) For each  $j - th$  dimension,  $\text{sign}(v_j, u_j) = 1$ . If the projection holds these two conditions, the projected samples  $\{\mathbf{u}^i, \mathbf{v}^i\}$  can be partitioned into different hyperoctants in the medium space only according to the signs of  $\mathbf{u}^i$ , due to condition 2. Since samples in a hyperoctant are close enough to each other, this partition can hold the small neighborhood better. It is also a compact low dimensional approximation of the high dimensional hyperoctant-based partition strategy in both  $\Delta\mathbf{x}$  space and  $\Delta\phi$  space, which is an sufficient condition for the existence of a generic DM, as mentioned above.

For convenience, we re-denote  $\Delta\mathbf{x}$  as  $\mathbf{y} \in \mathbb{R}^n$ , re-denote  $\Delta\phi$  as  $\mathbf{x} \in \mathbb{R}^m$ ,  $\mathbf{Y}_{s \times n} = [\mathbf{y}^1, \dots, \mathbf{y}^i, \dots, \mathbf{y}^s]$  is all the  $\mathbf{y}^i$  of training set.  $\mathbf{X}_{s \times m} = [\mathbf{x}^1, \dots, \mathbf{x}^i, \dots, \mathbf{x}^s]$  is all the  $\mathbf{x}^i$  of training set. The projection matrices are

$$\begin{aligned} \mathbf{Q}_{r \times n} &= [\mathbf{q}_1, \dots, \mathbf{q}_j, \dots, \mathbf{q}_r]^T, \mathbf{q}_j \in \mathbb{R}^n, \\ \mathbf{P}_{r \times m} &= [\mathbf{p}_1, \dots, \mathbf{p}_j, \dots, \mathbf{p}_r]^T, \mathbf{p}_j \in \mathbb{R}^m, \end{aligned}$$

Projection vectors are  $\mathbf{v} = \mathbf{Q}\mathbf{y}$ ,  $\mathbf{u} = \mathbf{P}\mathbf{x}$ . Here we denote projection vectors  $\mathbf{w}_j$ ,  $\mathbf{z}_j$  along the sample space:  $\mathbf{w}_j = \mathbf{Y}\mathbf{q}_j = [v_j^1, \dots, v_j^i, \dots, v_j^s]^T$ ,  $\mathbf{z}_j = \mathbf{X}\mathbf{p}_j = [u_j^1, \dots, u_j^i, \dots, u_j^s]^T$ . This problem can be formulated as a constrained optimization form,

$$\min_{\mathbf{P}, \mathbf{Q}} \sum_{j=1}^r \|\mathbf{Y}\mathbf{q}_j - \mathbf{X}\mathbf{p}_j\|^2 = \min_{\mathbf{P}, \mathbf{Q}} \sum_{j=1}^r \sum_{i=1}^s (v_j^i - u_j^i)^2 \tag{9}$$

$$s.t. \sum_{j=1}^r \sum_{i=1}^s \text{sign}(v_j^i u_j^i) = sr \tag{10}$$

It can be seen that  $\mathbf{w}_j$  and  $\mathbf{z}_j$  are the projected values of all the samples  $\mathbf{Y}$  or  $\mathbf{X}$  along a special direction  $\mathbf{q}_j$  or  $\mathbf{p}_j$ . For a fixed projected  $j$ -th dimension, assuming the  $\mathbf{w}_j$  and  $\mathbf{z}_j$  is normalized, which means that the mean of  $\{v_j^i\}_{i=1:s}$  is zero, and the standard deviation of it is  $1/s$ , so is  $\{u_j^i\}_{i=1:s}$ . Thus  $\mathbf{w}_j^T \mathbf{w}_j = 1$ ,  $\mathbf{z}_j^T \mathbf{z}_j = 1$ ,  $\mathbf{w}_j^T \mathbf{e} = 0$ ,  $\mathbf{z}_j^T \mathbf{e} = 0$ , where  $\mathbf{e} = [1, 1, \dots, 1]^T$ , then Eq.9 can be simplified as,

$$\min_{\mathbf{P}, \mathbf{Q}} \sum_{j=1}^r \|\mathbf{w}_j - \mathbf{z}_j\|^2 = \max_{\mathbf{P}, \mathbf{Q}} \sum_{j=1}^r \mathbf{w}_j^T \mathbf{z}_j \tag{11}$$

For a fixed projected  $j$ -th dimension, the constraint  $\sum_{i=1}^s \text{sign}(v_j^i u_j^i) = s$  means that all the pairs  $\{v_j^i, u_j^i\}$  of samples in  $j$ -th dimension keeps the consistence of sign. There is a fact: if the angle  $\theta_j$  between  $\mathbf{w}_j$  and  $\mathbf{z}_j$  is 0, the term  $\mathbf{w}_j^T \mathbf{z}_j$  will reach maximum, so the sign condition must hold; and if the angle  $\theta_j$  is  $\pi/2$ , the term  $\mathbf{w}_j^T \mathbf{z}_j$  will reach 0, so the sign condition will fail completely. Moreover, fixing the  $|v_j^i u_j^i|$ , the  $\cos \theta_j$  will get larger while the  $\sum_{i=1}^s \text{sign}(v_j^i u_j^i)$  rises, and  $\sum_{i=1}^s \text{sign}(v_j^i u_j^i)$  tends to be larger with the  $\cos \theta_j$  growing. Given some constraints, it can be proved that the  $\cos \theta_j$  can be taken as an approximation of the sign summation function for optimization,

$$\frac{1}{s} \sum_{i=1}^s \text{sign}(v_j^i u_j^i) \approx \cos \theta_j = \mathbf{w}_j^T \mathbf{z}_j \tag{12}$$

When the samples  $\{\mathbf{y}^i\}_{i=1:s}$  and  $\{\mathbf{x}^i\}_{i=1:s}$  are normalized (removing means and dividing standard deviation during pre-processing), the sign-correlation constrained optimization problem will be solved with the standard Canonical-Correlation Analysis (CCA). The CCA problem for normalized  $\{\mathbf{y}^i\}_{i=1:s}$  and  $\{\mathbf{x}^i\}_{i=1:s}$  is,

$$\max_{(\mathbf{p})_j, \mathbf{q}_j} \mathbf{q}_j^T \text{cov}(\mathbf{Y}, \mathbf{X})\mathbf{p}_j \tag{13}$$

$$s.t. \mathbf{q}_j^T \text{var}(\mathbf{Y}, \mathbf{Y})\mathbf{q}_j = 1, \mathbf{p}_j^T \text{var}(\mathbf{X}, \mathbf{X})\mathbf{p}_j = 1 \tag{14}$$

Following CCA algorithm, the max sign-correlation dimension  $\mathbf{p}_1$  and  $\mathbf{q}_1$  is solved at first. Then one seeks  $\mathbf{p}_2$  and  $\mathbf{q}_2$  by maximizing the same correlation subject to the constraint that they are to be uncorrelated with the first pair  $\mathbf{w}_1, \mathbf{z}_1$  of

canonical variables; This procedure may be continued up to  $r$  times until  $\mathbf{p}_r$  and  $\mathbf{q}_r$  is solved.

After all  $\mathbf{p}_j$  and  $\mathbf{q}_j$  is solved, we only need the projection matrix  $\mathbf{P}$  in  $\Delta\mathbf{x}$  space. Then we project each  $\Delta\mathbf{x}^i$  into the sign-correlation subspace and get reduced feature  $\mathbf{u}^i = \mathbf{P}\Delta\mathbf{x}^i$ . Then we partition the whole sample space into independent descent domains by judging the sign of each dimension of  $\mathbf{u}^i$  and group it into corresponding hyperoctant. Finally, in order to solve Eq. 4 at each iterative step, we learn a descent mapping for every subset at each iterative step with the ridge regression algorithm. When testing a face image, we also use the projection matrix  $\mathbf{P}$  to find its corresponding decent domain and predict its shape increment at each iterative step.

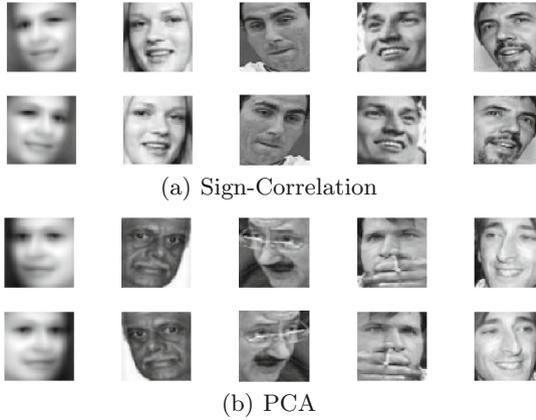
## 4 Experiments and Evaluation

Our work mainly focuses on face alignment across poses, so we conduct experiments especially for this task to analyze and evaluate our sign-correlation partition method. Firstly, we validate our method on multi-pose datasets, and the PCA partition scheme is also compared to our sign-correlation partition method. Then we also test our method on common datasets for general face alignment and compare it with state-of-the-art methods. According to Yans research [15], the multi-scale HOG outperforms multi-scale SIFT and other typical local descriptors HOG, SIFT, LBP and Gabor. We adopt multi-scale HOG as feature mapping function in our sign-correlation partition SDM algorithm. The two domains are enough for partition, so we only use the first sign-correlation projection component in appearance feature space.

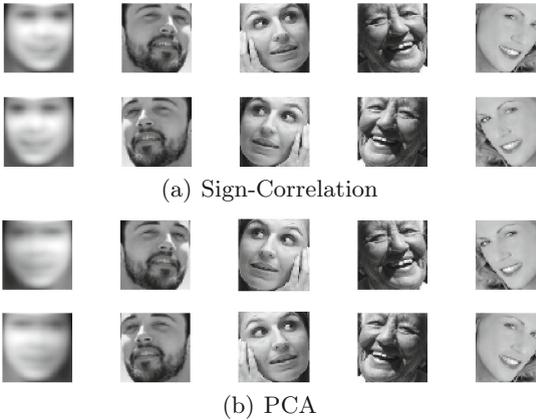
### 4.1 Sign-Correlation Partition Validation

In this section, we validate the underlying relationship between our sign-correlation partition and the variation of poses. Two widely used benchmark datasets are used in our validation: MTFL [35] and 300W [36]. MTFL dataset contains labeled face images from AFLW [37], LFW [38] and Internet. This dataset annotates 5 landmarks and labels 5 different left-right poses with the flags of gender, smile and glasses. Here we only focus on non-frontal poses to verify our partition method. There are 2550 non-frontal face images in original MTFL dataset, and the number of left ones is not equal to the number of right ones. For fairness, we augment these non-frontal images by a horizontal flip, so that we get the same numbers of left and right images. 300W dataset is mainly made up of images from LFPW [39], HELEN [40], AFW [41] with 68 re-annotated landmarks. The 3148 images from training dataset are selected in our validation. The flip augment is also used for obtaining the same number of left and right images. The left or right poses are estimated by a typical pose estimation taking known landmarks as input.

We partition the multi-poses images into two domains by the first sign-correlation projected dimension. The PCA partition by the first principle component is also tested as comparison. The results in Figs. 1 and 2 show that



**Fig. 1.** Pose validation on MTFL. In each subfigure: first column shows average faces of two subsets, and first or second row shows samples in subset 1 or 2.



**Fig. 2.** Pose validation on 300W. In each subfigure: first column shows average faces of two subsets, and first or second row shows samples in subset 1 or 2.

each sign-correlation domain mainly contain left or right pose images, and the accuracy of pose partition is high, as shown in Table 1. It indicates that our sign-correlation partition method can construct descent domains highly related to pose variations only with face appearance features. On the contrary, PCA partition only with face appearance features can not capture the pose variation well, and the partition result is nearly random.

### 4.2 Comparison of Face Alignment

We evaluate the proposed sign-correlation partition SDM method on the challenging 300W dataset, and compare it with state-of-the-art methods ESR [10],

**Table 1.** Pose accuracy validation on MTFW and 300W

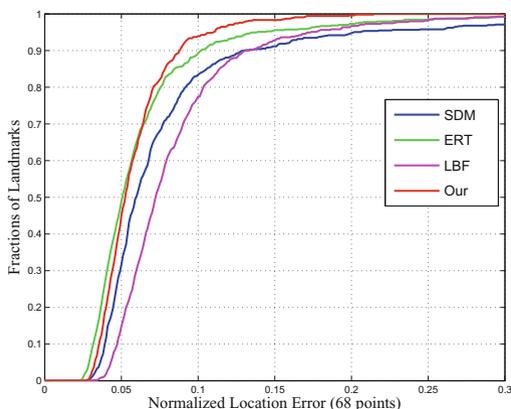
Datasets	Left/right number	PCA	Sign-Corr
MTFW	2550	0.7780	<b>0.9275</b>
300W	3148	0.5179	<b>0.9319</b>

**Table 2.** Comparison with current methods on 300W dataset

Datasets	Full	Common	Challenging
ESR	7.58	5.28	17.00
SDM	7.52	5.60	15.40
ERT	6.41	5.22	13.03
LBF	6.32	4.95	11.98
Ours	<b>5.88</b>	5.07	<b>10.79</b>

SDM [11], ERT [32], and LBF [33]. As mentioned above, there are 68 labeled landmarks in this dataset. Its training part contains 3148 images from AFW and training parts of LFPW, and HELEN dataset, and its testing part contains 689 images from testing parts of LFPW, and HELEN and IBUG. Among them, the LFPW dataset, although more challenging than other near-frontal datasets, is mainly made up of small pose variations, and the result on it nearly reaches limitation. The HELEN dataset contains faces of different genders, poses, and expressions. The IBUG testing dataset is the most challenging one due to extreme poses, expressions and lighting.

We conduct three experiments by testing different parts of the whole 300W dataset: common subset: LFPW and HELEN, challenging dataset: IBUG, and full

**Fig. 3.** CED curves over 300W

dataset. Following the standard [38], the normalized inner-pupil distance landmark error is used in our evaluation. The inner-pupil errors of different methods are given in Table 2. The cumulative error distribution (CED) curves are also plotted, as shown in Fig. 3. The results illustrate that our method outperforms most of current methods over the full datasets. We also get comparable results on common LFPW and HELEN datasets. Our method works better especially on the challenging IBUG dataset with large variations of poses.

## 5 Conclusion

In this paper we propose a novel sign-correlation partition method based on global SDM, which achieves promising performance for face alignment on the challenging datasets. We analyze the underlying relationship between shape/feature space and pose space by sign-correlation reduced dimensional projection. Taking the advantage of the inherent connection of shapes with features within a mutual pose-related subspace, the global descents partition can be operated according to different hyperoctants in the projected sign-correlation subspace. Due to the high consistence of sign between shapes and features in this subspace, it is able to partition the descent domains only depending on features and learned sign-correlation projection components. Our method adjusts the global SDM efficiently for face alignment, the original partition scheme of which is not suitable for face alignment. Moreover, we provide a clearer explanation about the influence of poses on the problems of multiple local minima and global descent direction conflict, and tackle with the robust face alignment across poses in a direct way. The experiment on the widely used multi-pose dataset indicates that our sign-correlation partition method can divide global complex space into several pose-related descent domains only with appearance features, while PCA-based partition only in feature space does not work. Our method also achieves noticeable Search Results performance for face alignment on challenging datasets, compared with popular methods. Although our sign-correlation subspace method improves the robustness in extreme conditions, there are still some parts we need to study in the future: The number of sign-correlation dimensions is needed to be chosen more carefully. On the other side, since the partition accuracy is limited by linear reduced dimensional projection, the kernel method can also be introduced into sign-correlation analysis.

**Acknowledgement.** This work was supported by Harbin Institute of Technology Scholarship Fund 2016 and National Centre for Computer Animation, Bournemouth University.

## References

1. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models—their training and application. *Comput. Vis. Image Underst.* **61**, 38–59 (1995)
2. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* **23**, 681–685 (2001)
3. Romdhani, S.: A multi-view nonlinear active shape model using kernel PCA. In: *British Machine Vision Conference*, pp. 483–492 (1999)
4. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. *BMVC* **41**, 929–938 (2006)
5. Gonzalez-mora, J., Torre, F.D.L., Murthi, R., Guil, N., Zapata, E.L.: Bilinear active appearance models. In: *IEEE International Conference on Computer Vision*, pp. 1–8 (2007)
6. Lee, H.S., Kim, D.: Tensor-based AAM with continuous variation estimation: application to variation-robust face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **31**, 1102–1116 (2009)
7. Saragih, J.M., Lucey, S., Cohn, J.F.: Deformable model fitting by regularized landmark mean-shift. *Int. J. Comput. Vision* **91**, 200–215 (2011)
8. Valstar, M., Martinez, B., Binefa, X., Pantic, M.: Facial point detection using boosted regression and graph models, pp. 2729–2736 (2010)
9. Dollar, P., Welinder, P., Perona, P.: Cascaded pose regression, vol. 238, pp. 1078–1085. *IEEE* (2010)
10. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression US Patent Application number 13/728,584 (2012)
11. Xiong, X., De, la Torre, F.: Supervised descent method and its applications to face alignment. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 532–539 (2013)
12. Burgosartizzu, X.P., Perona, P., Dollar, P.: Robust face landmark estimation under occlusion. In: *IEEE International Conference on Computer Vision*, pp. 1513–1520 (2013)
13. Xing, J., Niu, Z., Huang, J., Hu, W., Yan, S.: Towards multi-view and partially-occluded face alignment. In: *Computer Vision and Pattern Recognition*, pp. 1829–1836 (2014)
14. Yang, H., He, X., Jia, X., Patras, I.: Robust face alignment under occlusion via regional predictive power estimation. *IEEE Trans. Image Process.* **24**, 2393–2403 (2015)
15. Yan, J., Lei, Z., Yi, D., Li, S.Z.: Learn to combine multiple hypotheses for accurate face alignment. In: *IEEE International Conference on Computer Vision Workshops*, pp. 392–396 (2013)
16. Zhang, J., Shan, S., Kan, M., Chen, X.: Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) *ECCV 2014. LNCS*, vol. 8690, pp. 1–16. Springer, Heidelberg (2014). doi:[10.1007/978-3-319-10605-2\\_1](https://doi.org/10.1007/978-3-319-10605-2_1)
17. Fan, X., Wang, H., Luo, Z., Li, Y., Hu, W., Luo, D.: Fiducial facial point extraction using a novel projective invariant. *IEEE Trans. Image Process.* **24**, 1164–1177 (2015)
18. Liu, L., Hu, J., Zhang, S., Deng, W.: Extended supervised descent method for robust face alignment. In: Jawahar, C.V., Shan, S. (eds.) *ACCV 2014. LNCS*, vol. 9010, pp. 71–84. Springer, Cham (2015). doi:[10.1007/978-3-319-16634-6\\_6](https://doi.org/10.1007/978-3-319-16634-6_6)

19. Martinez, B., Valstar, M.F.: L 2,1-based regression and prediction accumulation across views for robust facial landmark detection. *Image Vis. Comput.* **45**, 371–382 (2015)
20. Lee, D., Park, H., Yoo, C.D.: Face alignment using cascade Gaussian process regression trees. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4204–4212 (2015)
21. Martinez, B., Pantic, M.: Facial landmarking for in-the-wild images with local inference based on global appearance. *Image Vis. Comput.* **36**, 40–50 (2015)
22. Lindner, C., Bromiley, P.A., Ionita, M.C., Cootes, T.F.: Robust and accurate shape model matching using random forest regression-voting. *IEEE Trans. Pattern Anal. Mach. Intell.* **37**, 1862–1874 (2015)
23. Yang, H., Patras, I.: Fine-tuning regression forests votes for object alignment in the wild. *IEEE Trans. Image Process.* **24**, 619–631 (2014). A Publication of the IEEE Signal Processing Society
24. Sun, Y., Wang, X., Tang, X.: Deep convolutional network cascade for facial point detection. In: *Conference on Computer Vision and Pattern Recognition*, pp. 3476–3483 (2013)
25. Zhou, E., Fan, H., Cao, Z., Jiang, Y., Yin, Q.: Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In: *IEEE International Conference on Computer Vision Workshops*, pp. 386–391 (2013)
26. Tzimiropoulos, G.: Project-out cascaded regression with an application to face alignment. In: *IEEE Conference on Computer Vision and Pattern Recognition (2015)*
27. Xiong, X., De la Torre, F.: Global supervised descent method. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2664–2673 (2015)
28. Feng, Z.H., Huber, P., Kittler, J., Christmas, W., Wu, X.J.: Random cascaded-regression copse for robust facial landmark detection. *IEEE Sig. Process. Lett.* **22**, 76–80 (2015)
29. Yang, H., Jia, X., Patras, I., Chan, K.P.: Random subspace supervised descent method for regression problems in computer vision. *IEEE Sig. Process. Lett.* **22**, 1816–1820 (2015)
30. Zhu, S., Li, C., Loy, C.C., Tang, X.: Face alignment by coarse-to-fine shape searching. In: *CVPR*, pp. 4998–5006 (2015)
31. Feng, Z.H., Hu, G., Kittler, J., Christmas, W., Wu, X.J.: Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting. *IEEE Trans. Image Process.* **24**, 3425–3440 (2015)
32. Kazemi, V., Sullivan, J.: One millisecond face alignment with an ensemble of regression trees. In: *Computer Vision and Pattern Recognition*, pp. 1867–1874 (2014)
33. Ren, S., Cao, X., Wei, Y., Sun, J.: Face alignment at 3000 fps via regressing local binary features. *IEEE Trans. Image Process.* **25**, 1685–1692 (2014)
34. Zhang, Z., Zhang, W., Ding, H., Liu, J., Tang, X.: Hierarchical facial landmark localization via cascaded random binary patterns. *Pattern Recogn.* **48**, 1277–1288 (2014)
35. Zhang, Z., Luo, P., Loy, C.C., Tang, X.: Facial landmark detection by deep multi-task learning. In: *European Conference on Computer Vision*, pp. 94–108 (2014)
36. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 896–903 (2013)

37. Kostinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, pp. 2144–2151, 6–13 November 2011
38. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: a database for studying face recognition in unconstrained environments (2008)
39. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 2930–2940 (2013)
40. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: European Conference on Computer Vision, pp. 679–692 (2012)
41. Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 31–37 (2015)