Sign-Correlation Subspace For Face Alignment

Dansong Cheng, Yongqiang Zhang, Feng Tian, Ce Liu and Xiaofang Liu

Abstract—**Face alignment is an essential task for facial performance capture and expression analysis. Current methods such as random subspace Supervised Descent Method, Stage-wise Relational Dictionary and coarse-to-fine shape searching can ease multi-pose face alignment problem, but no method can deal with the multiple local minima problem directly. In this paper we propose a sign-correlation subspace method for domain partition in only one reduced low dimensional subspace. 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 their signs consistent in each dimension of the subspace, so that each hyper octant holds the condition that one general descent exists. Then a set of general descents are learned from the samples in different hyperoctants. Requiring only the feature projection for domain partition, our proposed method is effective for face alignment. We have validated our approach with the public face datasets which include a range of poses. The validation results show that our method can reveal their latent relationships to poses. The comparison with state-of-the-art methods demonstrates that our method outperforms them especially in uncontrolled conditions with various poses, while enjoying the comparable speed.**

Index Terms—**sign-correlation, Sparse Representation, Supervised Descent Method, Face alignment**

I. 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 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 positions of landmarks into optimal locations.

Since the ground truth of shape is unknown during test, its very difficult to predict the shape increment from initial shape to real shape. 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 many methods have been proposed for face alignment. Most of them can be categorized into two groups according to their 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], [4], [5], [6]. In this type of methods, the optimization target is model parameters. It means seeking the best parameters to generate the most fitting shape (facial landmarks). These methods mitigate the influence of various poses and illumination. Due to sub-optimization problem, they are however sensitive to initialization and often tend to fail in the wild condition.

Recently discriminative models have shown better performance for face alignment. Some discriminative methods are based on local classifiers or response maps for landmarks [7], [8]. These methods deal with each landmark independently and ignore the relationship between them. Different from these methods, the cascaded regression-based methods take all landmarks as a whole and solve a nonlinear optimization problem by the cascaded regression theory [9]. The main difference between the cascaded regression and related boosting regression is that the former uses shape-indexed features extracted from the image according to the current estimated shape. Cascaded Pose Regression (CPR) [10] for pose estimation has been widely extended into face alignment in recent works, represented by Explicit Shape Regression (ESR)[11] and Supervised Descent Method (SDM)[12]. It is noticed that SDM provides a theoretical explanation of the cascaded regression from the view point of optimizing a nonlinear problem, as a significant achievement in cascaded regression 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 [13], [14], [15]. Some works mainly aim at speeding up the prediction process while keeping high accuracy [14], [16]. The choice and learning of shape-indexed features are also studied [16], [17], [18]. A series of regression methods have been employed into the cascaded regression framework to deal with over-fitting and local minima problems in the wild condition, including ridge regression [19], Support Vector [20], Gaussian process [21], [22], Random Forest voting [23], [15], [24], Deep Neural Nets [16], [25], and project-out cascaded regression [26].

Although these works have produced remarkable results on nearly frontal face alignment, it remains 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. To address the issue, Xiong et al.[27] theoretically address 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 unable to find approximation due to lack of previous frame. A few recent

The authors D Cheng, Y Zhang and C Liu are with the School of Computer Science and Technology, Harbin Institute of Technology, 92 West Dazhi Street, Nan Gang District, Harbin, 150001, P.R.China.

The authors X Liu is with the School of Electrical Engineering and Automation, Harbin Institute of Technology, P.R.China.

The author F Tian is with the Faculty of Science & Technology, Bournemouth University, UK.

Xiaofang Liu, e-mail: liuxf@hit.edu.cn.

works[13], [14], [28], [29], [30], [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.

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. For decent domains partition, we only need 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 some challenging face datasets, which include face images from different poses. The results show that we can split complex sample space into homogeneous domains related to poses, thus a mutual manifold of feature and shape spaces is obtained. The experiments results demonstrate that our method achieves state-of-the-art performance for nearly frontal face images, and it is more robust on datasets with multiple poses, compared with current methods.

The rest of this paper is organized as follows. In Section II, we briefly introduce the related work, such as Cascaded Regression to Face Alignment and Multi-Pose Face Alignment. The key idea of the paper is given in Section III, where we describe our proposed approach. Experimental results and analysis are presented in Section IV, followed by the conclusion in Section V.

II. Related work

A. Cascaded Regression to Face Alignment

Both 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 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, the cascaded regression based methods have shown more promising performance than local classifiers or response map based methods [7], [8] and generative models. ESR[11] uses shape indexed intensity difference features for face alignment based on CPR [10]. Moreover, SDM extracts shape-indexed SIFT features and learns a sequence of general descent maps from supervised training data, providing a solution when it is hard to apply Newton Descent method for a not analytically differentiable nonlinear function or when 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 such as large poses, extreme expressions and partial occlusions.

Later research mainly focuses on performance improvement based on ESR and SDM. Burgosartizzu et al.[13] integrates partial visibility term into landmarks and presents interplolated shape-indexed features to tackle with occlusions and high shape variances. Kazemi et al.[32] estimates 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 has hardly been taken into account. Instead of least squares regression, Xing et al.[14] learns 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.[17] compares the performance of different local feature descriptors for face alignment, including SIFT, HOG, LBP and Gabor, and HOG shows best results in their experiments. Ren et al.[33] builds local binary features by learning regression random forest for each landmark independently, and then learns a global cascaded linear regression with pre-built binary features. Deep Neural Networks [25], [34], [16] 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 very challenging to tune many free parameters.

B. Multi-Pose Face Alignment

Xiong et al.[27] analyze the drawbacks of SDM and splits whole sample space into descent domains by PCA in both feature and shape spaces. However, the approach cannot 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] proposes random cascaded-regression copse, learning a set of cascaded strong regressors corresponding to different subsets of samples and averaging all predictions of them as final output. Similarly, Yang et al.[29] proposes random subspace SDM, randomly selecting a small number of dimensions from whole feature space and training an ensemble of regressors in several feature subspaces. Liu et al.[19] modifies 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. [20] to substitutes the commonly used least squares regressor, which improves the performance of face alignment across views. Gaussian process [21], [22] and Random Forest voting [23], [15], [24] are also introduced into cascaded regression framework. Zhang et al. [35] 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. [18] combine projective invariant characteristic number with appearance based constrains and solve a quadratic optimization by the standard gradient descent. Though their method shows 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 improvements on the challenging datasets.

Some methods among them have begun to deal with the impact of poses, like RPCR [13], SRD [14], CCR [31], hierarchical localization [35] 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 alleviate the problem by different strategies, but how to achieve robust and accurate face alignment across poses remains a challenging task. To tackle these limitations, our work focuses on analyzing the underlying relationship between feature space and shape space, finding a joint pose related subspace for global supervised descent domains partition, and finally improving the performance of face alignment under challenging conditions.

III. Sign-correlation Subspace for descent domains partition

A. Descent domain partition in SDM

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

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

where $\phi_* = \vec{h}(\mathbf{d}(\vec{x}_*))$ represents the feature extracted according to correct landmarks $\vec{x}_*$, which is known in the training images, but unknown in the testing images. For initial locations of landmarks $\vec{x}_0$, we solve $\Delta\vec{x}$, which minimizes the feature alignment error $f(\vec{x}_0 + \Delta\vec{x})$. Since the feature function is usually not analytically differentiable, it is hard to solve the problem with traditional Newtons descent methods. Alternatively, a general descent mapping can be learned from training datasets. The supervised descent method form is,

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

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

$$\Delta\vec{x} = \mathbf{R}_k(\overline{\phi_*} - \phi_k) \tag{3}$$

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 samplesare divided into T subsets $S = \{S_t\}_1^T$.

A generic DM exists under the two conditions: 1) $\mathbf{R}\vec{h}(\vec{x})$ is a strictly locally monotone operator anchored at the optimal solution; 2) $\vec{h}(\vec{x})$ is locally Lipschitz continuous anchored at $\vec{x}_*$. 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,

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

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

where $\Delta\vec{x}_*^i = \vec{x}_*^i - \vec{x}_k^i$, $\Delta\phi^{i,t} = \overline{\phi}_*^t - \phi^i$, and $\overline{\phi}_*^t-$ average all ϕ_* over the subset S_t. Eq.(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) are given:

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

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

where $\Delta\mathbf{X}_*^t = [\Delta\vec{x}_*^{1,t}, \dots, \Delta\vec{x}_*^{i,t}, \dots]$, each column is $\Delta\vec{x}_*^{i,t}$ 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 hyper octant (the generalization of quadrant) is positive, an ideal sufficient partition can be like that each subset S^t occupies a hyperoctant both in the parameter space $\Delta\vec{x}$ and feature space $\Delta\phi$. However, this leads to exponential number of DMs. Assuming $\Delta\vec{x}$ 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 $\Delta\vec{x}$ and $\Delta\phi$ are embedded in a lower dimensional manifold for human faces. So the dimension reduction methods(e.g. PCA) on the whole training set $\Delta\vec{x}$ and $\Delta\phi$ can be used for approximation. The Global SDM projects $\Delta\vec{x}$ onto the subspace expended by the first two components of $\Delta\vec{x}$ space, and projects $\Delta\phi$ 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. The 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.

B. Sign-correlation Subspace Partition

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

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

Then there must exist a DM $\mathbf{R}^t$ in one updating step. Eq.(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. Notice that there are two limitations of this partition strategy: 1) it cannot guarantee the samples lie in the same small neighborhood. In other words, even if $\{\Delta\vec{x}^{i,t}, \Delta\phi^{i,t}\}$, $\{\Delta\vec{x}^{k,t}, \Delta\phi^{k,t}\}$ keep Eq.(8), the $\Delta\vec{x}^{i,t}$, $\Delta\vec{x}^{k,t}$ may be very far from each other; 2) it only considers the dimension-to-dimension correlation of the first $\min(n, m)$ dimensions in the $\Delta\vec{x}$ space and $\Delta\phi$ space, and ignore other dimension. The correlation of any jth dimension of $\Delta\vec{x}$ with a non-corresponding $j^{'} - th$ dimension $j^{'} \neq j$ of $\Delta\phi$ is also ignored.

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

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

$\mathbf{Q}_{r\times n} = [\vec{q}_1, \ldots, \vec{q}_j, \ldots, \vec{q}_r]^T,\ \vec{q}_j \in \Re^n,$
$\mathbf{P}_{r\times m} = [\vec{p}_1, \ldots, \vec{p}_j, \ldots, \vec{p}_r]^T,\ \vec{p}_j \in \Re^m,$

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

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

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

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

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

For a fixed projected jth dimension, the constraint $\sum_{i=1}^{s} sign(v_j^i u_j^i) = s$ means that all the pairs $\{v_j^i, u_j^i\}$ of samples in jth dimension keeps the consistence of sign.There is a fact: if the angle θ_j between $\vec{w}_j$ and $\vec{z}_j$ is 0, the term $\vec{w}_j^T\vec{z}_j$ will reach maximum, so the sign condition must hold; and if the angle θ_j is $\pi/2$, the Eq.(12) 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 $\sum_{i=1}^{s} sign(v_j^i u_j^i)$ rises, and $\sum_{i=1}^{s} sign(v_j^i u_j^i)$ tends to go up 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,

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

When the samples $\{\vec{y}^i\}_{i=1:s}$ and $\{\vec{x}^i\}_{i=1:s}$ are normalized(by 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 $\{\vec{y}^i\}_{i=1:s}$ and $\{\vec{x}^i\}_{i=1:s}$ is,

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

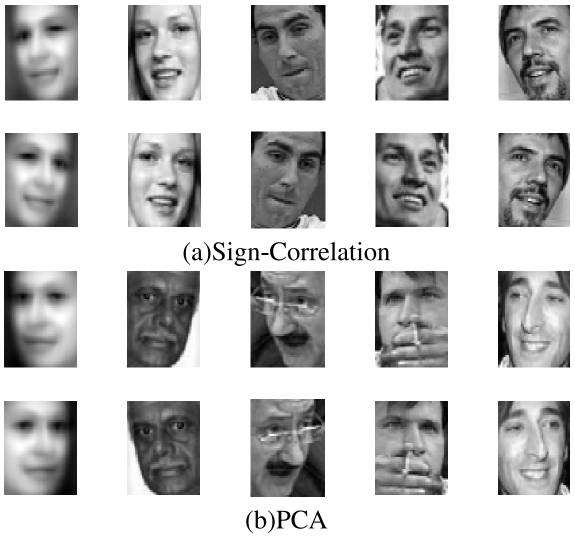
$$s.t.\ \vec{q}_j^T var(\mathbf{Y}, \mathbf{Y})\vec{q}_j = 1,\ \vec{p}_j^T var(\mathbf{X}, \mathbf{X})\vec{p}_j = 1 \tag{14}$$

Based on CCA, the max sign-correlation dimensions $\vec{p}_1$ and $\vec{q}_1$ are solved at first. Then one seeks $\vec{p}_2$ and $\vec{q}_2$ by maximizing the same correlation subject to the constraint that they are to be uncorrelated with the first pair $\vec{w}_1$, $\vec{z}_1$ of canonical variables. This procedure may be continued up to r times until $\vec{p}_r$ and $\vec{q}_r$ are solved.

After all $\vec{p}_j$ and $\vec{q}_j$ are solved, we only need the projection matrix $\mathbf{P}$ in $\Delta\vec{x}$ space. Subsequently we project each $\Delta\vec{x}^i$ into the sign-correlation subspace and get reduced feature $\vec{u}^i = \mathbf{P}\Delta\vec{x}^i$. Then we partition the whole sample space into independent descent domains by judging the sign of each dimension of $\vec{u}^i$ 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. To test 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.

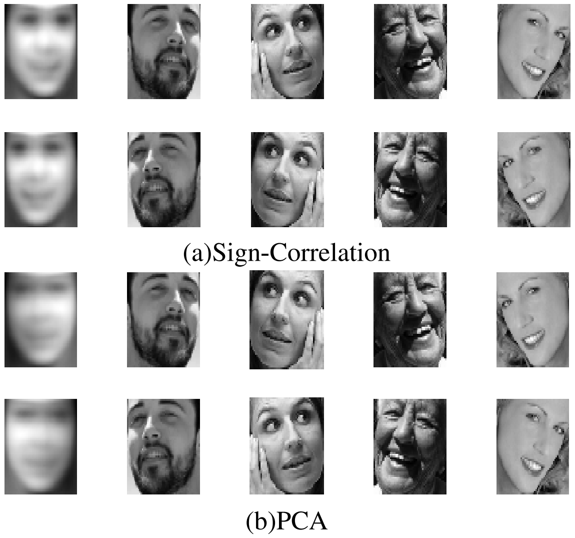
IV. Experiments and Evaluation

Since our work mainly focuses on face alignment across poses, we focus on our experiments especially on this task to analyze and evaluate our sign-correlation partition method. Firstly, we validate our method on multi-pose dataset and


(a)Sign-Correlation

(b)PCA

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.


(a)Sign-Correlation

(b)PCA

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.

compare our approach with the PCA partition scheme. Then we test our method on common datasets for general face alignment and compare it with state-of-the-art methods. According to Yans research [17], the multi-scale HOG outperforms multi-scale SIFT and other typical local descriptors HOG, SIFT, LBP and Gabor. Thus 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.

A. 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 [36] and 300W [37]. MTFL dataset contains labeled face images from AFLW [38], LFW [39] 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 the original MTFL dataset, and the number of left ones is not equal to that 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. The 300W dataset is mainly made up of images from LFPW [40], HELEN [41], AFW[42] with 68 re-annotated landmarks. The 3148 images from training dataset are chosen 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 which takes 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 Fig.1 and Fig.2 show that each sign-correlation domain mainly contains left or right pose images. The accuracy of pose partition is high, as shown in Table I. 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, the PCA partition only with face appearance features cannot capture the pose variation well, and the partition result is nearly random.

TABLE I
POSE ACCURACY VALIDATION ON MTFL AND 300W

Datasets	Left/Right Number	PCA	Sign-Corr
MTFW	2550	0.7780	**0.9275**
300W	3148	0.5179	**0.9319**

B. 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 [11], SDM [12], ERT [32], and LBF [33]. As mentioned above, there are 68 labeled landmarks in this dataset. Its training part contains 3148 images form AFW and training parts of LFPW, and HELEN dataset. The testing part of the dataset consists of 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

TABLE II
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	**5.88**	5.07	**10.79**

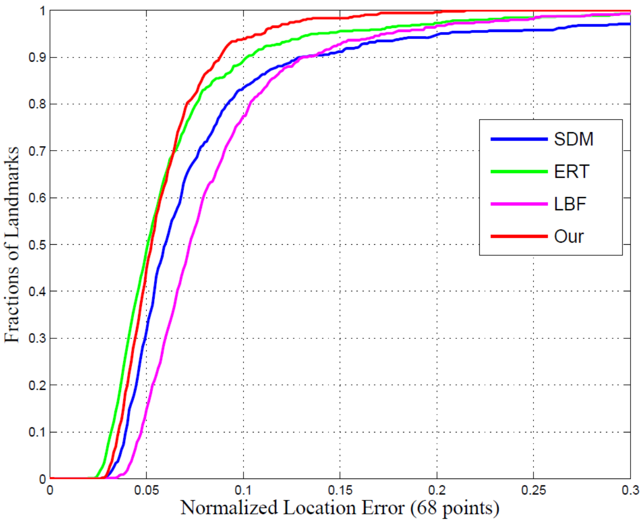


Fig. 3. CED curves over 300W

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 300W dataset: common subset (LFPW and HELEN), challenging dataset (IBUG) and full dataset. Following the standard [39], the normalized inner-pupil distance landmark error is used in our evaluation. The inner-pupil errors of different methods are given in Table II. The cumulative error distribution (CED) curves are also plotted, as shown in Fig.3.

The results has clearly illustrated that our method outperforms most of current methods over the full datasets, while achieving comparable results on common LFPW and HELEN datasets. In fact, our method works particularly well on the challenging IBUG dataset with large variations of poses.

C. Computation Complexity Analysis

Compared with a standard SDM [12], the additional computation load of our proposed approach mainly lies on the Canonical Correlation Analysis for learning sign-correlation projection matrix during training and the sign-correlation projection for descent domain partition during testing. Since each independent domain has its special descent mapping, the computational cost of learning multiple descent mappings at each stage during our training is T times as the cost of learning a global descent mapping at each stage during SDM training, assuming there are total T domains in our model. Note that our model will degrade to SDM when $T = 1$. During our testing, only is one descent mapping in its corresponding domain activated at each stage. In fact, given a test image, once its domain is decided, the subsequent prediction will run just like a standard SDM progress. Therefore, despite of extra costs during our training, the testing cost of our approach is just slightly higher than that of SDM, because the main operation for sign-correlation projection is the multiplication of the projection matrix with a feature vector, which is very fast.

TABLE III
Time costs comparison between our approach and SDM on 300W dataset(second)

Porgress	Train cca	Train dm	Train total	Test
SDM	0.0	1052.044	1052.044	0.0242728
Ours	82.8152	2074.7446	2157.5598	0.0248737

We compare the time costs of our method and SDM on the 300W dataset. As shown in Tab.III, for a $T = 2$ model, our training cost is over 2 times as SDM, while our test cost is at the same order of magnitude as SDM.

V. Conclusion

In this paper we propose a novel sign-correlation partition method for global SDM algorithm, and achieve promising results 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 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 hyper octants in the projected sign-correlation subspace. Due to the high consistence of sign between shapes and features in this subspace, the proposed approach can partition the descent domains only depending on features and learned sign-correlation projection components. Our method extends the global SDM method into face alignment task, the original partition scheme of which is not suitable for face alignment. The experiments on the widely used multi-pose dataset have demonstrated that our sign-correlation partition method can divide the global complex space into several pose-related descent domains only with appearance features rather than PCA partition in both shape and feature spaces. Our method also achieves noticeable improvements for face alignment on challenging datasets, compared with well-known methods. In our future work we will consider introducing the kernel method into sign-correlation analysis to further increase the partition acuracy.

Acknowledgment

This research was supported by the National Natural Science Foundation of China (Grant No. 61402133)

All Author declares that he/she has no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models&mdash;their training and application," *Computer Vision & Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[2] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *Pattern Analysis & Machine Intelligence IEEE Transactions on*, vol. 23, no. 6, pp. 681–685, 2001.

[3] S. Romdhani, "A multi-view nonlinear active shape model using kernel pca," in *British Machine Vision Conference*, 1999, pp. 483–492.

[4] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models," *BMVC*, vol. 41, pp. 929–938, 2006.

[5] J. Gonzalezmora, F. D. L. Torre, R. Murthi, N. Guil, and E. L. Zapata, "Bilinear active appearance models," in *IEEE International Conference on Computer Vision*, 2007, pp. 1–8.

[6] H. S. Lee and D. Kim, "Tensor-based aam with continuous variation estimation: application to variation-robust face recognition." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 31, no. 6, pp. 1102–16, 2009.

[7] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *International Journal of Computer Vision*, vol. 91, no. 2, pp. 200–215, 2011.

[8] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," 2010, pp. 2729–2736.

[9] E. Snchez-Lozano, B. Martinez, and M. F. Valstar, "Cascaded regression with sparsified feature covariance matrix for facial landmark detection ," *Pattern Recognition Letters*, vol. 73, no. C, pp. 19–25, 2016.

[10] P. Dollar, P. Welinder, and P. Perona, "Cascaded pose regression," *IEEE*, vol. 238, no. 6, pp. 1078–1085, 2010.

[11] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," Dec. 27 2012, uS Patent App. 13/728,584.

[12] X. Xiong and F. De, la Torre, "Supervised descent method and its applications to face alignment," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2013, pp. 532–539.

[13] X. P. Burgosartizzu, P. Perona, and P. Dollar, "Robust face landmark estimation under occlusion," in *IEEE International Conference on Computer Vision*, 2013, pp. 1513–1520.

[14] J. Xing, Z. Niu, J. Huang, W. Hu, and S. Yan, "Towards multi-view and partially-occluded face alignment," in *Computer Vision and Pattern Recognition*, 2014, pp. 1829–1836.

[15] H. Yang, X. He, X. Jia, and I. Patras, "Robust face alignment under occlusion via regional predictive power estimation," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2393–403, 2015.

[16] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment," in *Computer Vision–ECCV 2014*. Springer, 2014, pp. 1–16.

[17] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Learn to combine multiple hypotheses for accurate face alignment," in *IEEE International Conference on Computer Vision Workshops*, 2013, pp. 392–396.

[18] X. Fan, H. Wang, Z. Luo, Y. Li, W. Hu, and D. Luo, "Fiducial facial point extraction using a novel projective invariant." *Image Processing IEEE Transactions on*, vol. 24, no. 3, pp. 1164–77, 2015.

[19] L. Liu, J. Hu, S. Zhang, and W. Deng, *Extended Supervised Descent Method for Robust Face Alignment*. Springer International Publishing, 2014.

[20] B. Martinez and M. F. Valstar, "L 2,1 -based regression and prediction accumulation across views for robust facial landmark detection ," *Image & Vision Computing*, vol. 45, no. 4, pp. 371–382, 2015.

[21] D. Lee, H. Park, and C. D. Yoo, "Face alignment using cascade gaussian process regression trees," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4204–4212.

[22] B. Martinez and M. Pantic, "Facial landmarking for in-the-wild images with local inference based on global appearance ," *Image & Vision Computing*, vol. 36, no. C, pp. 40–50, 2015.

[23] C. Lindner, P. A. Bromiley, M. C. Ionita, and T. F. Cootes, "Robust and accurate shape model matching using random forest regression-voting," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 37, no. 9, pp. 1862–74, 2015.

[24] H. Yang and I. Patras, "Fine-tuning regression forests votes for object alignment in the wild," *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, vol. 24, no. 2, pp. 619–31, 2014.

[25] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Conference on Computer Vision & Pattern Recognition*, 2013, pp. 3476–3483.

[26] G. Tzimiropoulos, "Project-out cascaded regression with an application to face alignment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015.

[27] X. Xiong and F. D. la Torre, "Global supervised descent method," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 2664–2673.

[28] Z. H. Feng, P. Huber, J. Kittler, W. Christmas, and X. J. Wu, "Random cascaded-regression copse for robust facial landmark detection," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 76–80, 2015.

[29] H. Yang, X. Jia, I. Patras, and K. P. Chan, "Random subspace supervised descent method for regression problems in computer vision," *IEEE Signal Processing Letters*, vol. 22, no. 10, pp. 1816–1820, 2015.

[30] S. Zhu, C. Li, C. C. Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *CVPR*, 2015, pp. 4998–5006.

[31] Z. H. Feng, G. Hu, J. Kittler, W. Christmas, and X. J. Wu, "Cascaded collaborative regression for robust facial landmark detection trained using a mixture of synthetic and real images with dynamic weighting," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3425–3440, 2015.

[32] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Computer Vision and Pattern Recognition*, 2014, pp. 1867–1874.

[33] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," *IEEE Transactions on Image Processing*, pp. 1685–1692, 2014.

[34] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *IEEE International Conference on Computer Vision Workshops*, 2013, pp. 386–391.

[35] Z. Zhang, W. Zhang, H. Ding, J. Liu, and X. Tang, "Hierarchical facial landmark localization via cascaded random binary patterns," *Pattern Recognition*, vol. 48, no. 4, p. 1277C1288, 2014.

[36] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European Conference on Computer Vision*, 2014, pp. 94–108.

[37] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 896–903.

[38] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization," in *IEEE International Conference on Computer Vision Workshops, ICCV 2011 Workshops, Barcelona, Spain, November 6-13, 2011*, 2011, pp. 2144–2151.

[39] G. B. Huang, M. Mattar, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," 2008.

[40] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars." *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 12, pp. 2930–40, 2013.

[41] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European Conference on Computer Vision*, 2012, pp. 679–692.

[42] D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 31–37.