Optimal Gradient Pursuit for Face Alignment

Xiaoming Liu
Visualization and Computer Vision Lab
GE Global Research, Niskayuna, NY 12309
liux@research.ge.com

Abstract— **Face alignment aims to fit a deformable landmark-based mesh to a facial image so that all facial features can be located accurately. In discriminative face alignment, an alignment score function, which is treated as the appearance model, is learned such that moving along its gradient direction can improve the alignment. This paper proposes a new face model named "Optimal Gradient Pursuit Model", where the objective is to minimize the angle between the gradient direction and the vector pointing toward the ground-truth shape parameter. We formulate an iterative approach to solve this minimization problem. With extensive experiments in generic face alignment, we show that our model improves the alignment accuracy and speed compared to the state-of-the-art discriminative face alignment approach.**

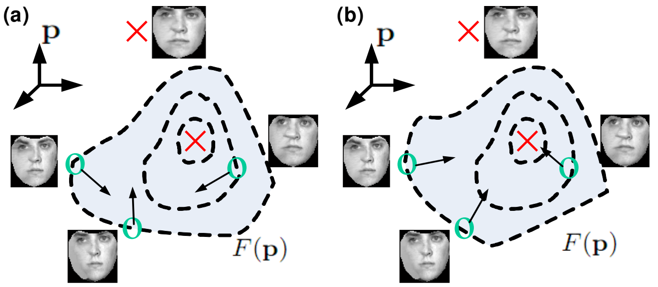

Fig. 1. **Alignment Score Function Learning.** (a) A concave function learned via BRM [31] has sub-optimal gradient directions. (b) Our model attempts to enforce all gradient directions of the random samples on the surface pointing toward the ground-truth alignment.

I. Introduction

Model-based image registration/alignment is a fundamental topic in computer vision, where a model is deformed such that its distance to an image is minimized. In particular, face alignment is receiving considerable attention, because it not only enables various practical capabilities such as facial feature detection, pose rectification, face animation, etc, but also poses many scientific challenges due to facial appearance variations in pose, illumination, expression, and occlusions.

There have been many successful studies on face alignment. Active Shape Model (ASM) [3] is one of the early methods that fit a statistical shape model to an object class. It was extended to Active Appearance Model (AAM) [1], [4], which has become a popular approach for face alignment. During AAM-based model fitting, the Mean-Square-Error between the appearance instance synthesized from the appearance model and the warped appearance from the input image is minimized by iteratively updating the shape and/or appearance parameters. Although AAM performs well while learning and fitting on a small set of subjects, its performance degrades quickly when it is trained on a large dataset [17] and/or fit to subjects that were not seen during the model learning [9].

In addition to the generative model based approaches such as AAM, there are also discriminative model based alignment approaches. Boosted Appearance Model (BAM) [13], [14] utilizes the same shape model as AAM, but an entirely different appearance model that is essentially a two-class classifier and learned discriminatively from a set of correctly and incorrectly warped images. During model fitting, BAM aims to maximize the classifier score by updating the shape parameter along the gradient direction. Though BAM has shown to generalize better in fitting to unseen images compare to AAM, one limitation is that the learned binary classifier can not guarantee a concave score surface while perturbing the shape parameter, i.e., moving along the gradient direction does not always improve the alignment. Boosted Ranking Model (BRM) [31] alleviates this problem by enforcing the convexity through learning. Using pairs of warped images, where one is a better alignment than the other, BAM learns a score function that attempts to correctly rank the two warped images within all training pairs. However, in BRM the gradient direction can still have a relative large angle with respect to the vector pointing to the ground-truth shape parameter starting from the current shape parameter. Hence, the alignment process may take a *convoluted* path during the optimization, which not only increases the chances of divergence, but also slows down the alignment.

To address this limitation, as shown in Figure 1, this paper proposes a novel approach to learn a discriminative face model, named *Optimal Gradient Pursuit Model (OGPM)*. Using the same shape representation as BAM and BRM, the learning of our appearance model, which is also an alignment score function, is formulated with a very different objective. That is, we aim to learn a score function, whose gradients at various perturbed shape parameters have the minimal angle with respect to the *ideal travel direction*,

This work was supported by the National Institute of Justice, Office of Justice Programs, US Department of Justice, under the award #2007-DE-BX-K191. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

i.e., the vector pointing directly to the ground-truth shape parameter. The score function is composed of a set of weak functions, each operating on one local feature in the warped image domain. We formulate the objective function such that each weak function can be estimated in an incremental manner from a large pool of feature candidates. During the model fitting, given an image with initial shape parameter, we perform gradient ascent by updating the shape parameter in the gradient direction, which is hopefully similar to the ideal travel direction. Experiments on a large set of facial images demonstrate the superior performance compared to BRM.

II. Prior Art

Image alignment is a fundamental problem in computer vision. The most popular work in face alignment are ASM, AAM or their variations [5]–[8], [11], [22], [28]. In ASM, the local appearance model for each landmark has been trained generatively [5] or discriminatively [6]. In contrast, most AAM-relevant work use the generative shape and appearance models. For example, the inverse composition method [18] greatly improves the efficiency of the AAM-based face alignment. There are some AAM variations using discriminative fitting methods [8], [22]. Other notable works in face alignment are [12], [33].

In contrast, there are discriminative image alignment approaches where the training data of the appearance model includes the *incorrect* images warped from perturbed shape parameters. Notable examples are BAM [14] and BRM [31], whose difference to our approach has been discussed in Section I. We will also experimentally compare with BRM in Section VI. In object tracking domain, there are work learning to predict the motion vector using regression. For example, Williams *et al.* [29] build a displacement expert, which takes an image as input and returns the displacement, by using Relevance Vector Machine. Our work differs in that we estimate shape parameters in the high dimensional space, which is much harder than that in the 2D space. In [32], Zhou and Comaniciu propose to learning a regressor with multidimensional output to predict the landmarks locations from the image content. In contrast, we train a statistical shape model where the shape parameter is the unknown parameter to be estimated.

In face alignment, there are also prior work on learning a discriminative appearance model for each landmark [30] in the ASM framework, or one local-minima-free appearance model [19] in the AAM framework. The work done by Nguyen and la Torre has similar set-up as ours while the main difference is in the specific learning approach being employed. Note that image alignment problem is in the context of registering between one image and a supervisely-learned model, which is different to the conventional image registration/tracking problem being solved between two images [10], [24], [26].

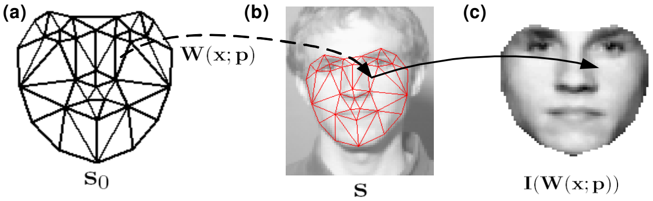
(a)
(b)
(c)
$\mathbf{W}(\mathbf{x};\mathbf{p})$
$\mathbf{s}_0$
$\mathbf{s}$
$\mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p}))$

Fig. 2. **Shape Model and Warping Function.** (a) Representation of the mean shape. (b) The face image with a superimposed shape. (c) The face image warped to the mean shape domain.

III. Face Model

Similar to BAM and BRM, our face model is composed of a generative shape model component and a discriminative appearance model component. In this section, we will present the model representation for both the shape and appearance model components.

A. Shape Model

Landmark-based shape representation is a popular way to describe the facial shape of an image. That is, a set of 2D landmarks, $\{x_i, y_i\}_{i=1,\cdots,v}$, are placed on top of key facial features, such as eye corner, mouth corner, nose tip, etc. The concatenation of these landmarks forms a shape observations of an image, $\mathbf{s} = [x_1, y_1, x_2, y_2, ..., x_v, y_v]^T$. Given a face database where each image is manually labeled with landmarks, the entire set of shape observations are treated as the training data for the shape model. In our approach, we use the same shape model as AAM, BAM and RAM, i.e., the Point Distribution Model (PDM) [3] learned via Principal Component Analysis (PCA) on the observation set. Thus, the learned generative PDM can represent a particular shape instance as,

$$\mathbf{s}(\mathbf{p}) = \mathbf{s}_0 + \sum_{i=1}^{n} p_i \mathbf{s}_i, \quad (1)$$

where $\mathbf{s}_0$ and $\mathbf{s}_i$ are the mean shape and i^{th} shape basis, respectively. Both of them are the results of the PDM learning. $\mathbf{p} = [p_1, p_2, ..., p_n]^T$ is the shape parameter. Similar to the shape component of AAM [18], the first four shape bases are trained to represent global translation and rotation, while the remaining shape bases represent the non-rigid deformation of facial shapes.

As shown in Figure 2(b), a warping function from the mean shape coordinate system to the coordinates in the image observation is defined as a piece-wise affine warp:

$$\mathbf{W}(x^0, y^0; \mathbf{p}) = [1 \; x^0 \; y^0]\mathbf{a}(\mathbf{p}), \quad (2)$$

where (x^0, y^0) is a pixel coordinate within the mean shape domain, and $\mathbf{a}(\mathbf{p}) = [\mathbf{a}_1(\mathbf{p}) \; \mathbf{a}_2(\mathbf{p})]$ is a unique 3×2 affine transformation matrix that relates each triangle pair in $\mathbf{s}_0$ and $\mathbf{s}(\mathbf{p})$. Given a shape parameter $\mathbf{p}$, $\mathbf{a}(\mathbf{p})$ needs to be computed for each triangle. However, since the knowledge of which triangle each pixel (x^0, y^0) belongs to is known a priori, the warp can be efficiently performed via a simple table lookup (see [18] for detailed description). Using this warping function, any face image can be warped into the

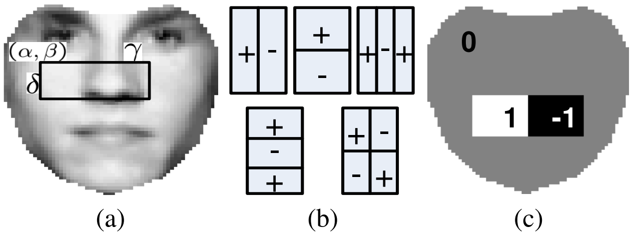

Fig. 3. **Appearance Features.** (a) Warped face image with feature parametrization. (b) Representation of the five feature types used by the appearance model. (c) Notional template $\mathbf{A}$.

mean shape and results in a shape-normalized face image $\mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p}))$ (see Figure 2(c)), from which the appearance model is learned.

B. Appearance Model

While the shape model learning can be simply conducted via PCA, the appearance model learning is more complex and also the focus of this paper. Hence, in this section, we will only introduce the representation of our appearance model, and its learning approach will be presented in Section IV.

Following the same representation as BAM and BRM, our appearance model is described by a collection of m local features $\{\varphi_i\}_{i=1,\cdots,m}$ that are computed on the shape-normalized face image $\mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p}))$. We adopt the Haar-like rectangular feature [20], [27] in our appearance representation for the following reasons: (a) the computational efficiency due to the integral image technique [27]; (b) successful experiences in facial image processing; and (c) making a fair comparison between our approach and BRM.

A rectangular feature can be computed as follows

$$\varphi \doteq \mathbf{A}^T \mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p})) \ , \tag{3}$$

where $\mathbf{A}$ is an image template (Figure 3(c)). The inner product between the template and the warped image is equivalent to computing the rectangular feature using the integral image. As shown in Figure 3(a), the image template $\mathbf{A}$ can be parameterized by $(\alpha, \beta, \gamma, \delta, \tau)$, where (α, β) is the top-left corner, γ and δ are the width and height, and τ is the feature type. Figure 3(b) shows the five feature types used in our model.

IV. Alignment Learning Problem

Having introduced the appearance model representation, in this section we describe in detail how to learn our appearance model, which is essentially an alignment score function that will be used during the model fitting stage.

To begin with, let us denote $\mathbf{p}$ as the shape parameter of a given image that represents the current alignment of the shape model (1). The goal of our appearance model learning is the following. *From labeled training data, we aim to learn a score function $F(\mathbf{p})$, such that, when maximized with respect to $\mathbf{p}$, it will result in the shape parameter of the correct alignment.* Specifically, if $\mathbf{p}_0$ is the shape parameter corresponding to the correct alignment of an image, F has to be such that

$$\mathbf{p}_0 = \arg\max_{\mathbf{p}} F(\mathbf{p}) \ . \tag{4}$$

A. Objective Function for Learning F

Given the above equation, we choose to optimize $F(\mathbf{p})$ via *gradient ascent*. That is, by assuming that F is differentiable, the shape parameter is iteratively updated in each alignment iteration starting from an initial parameter $\mathbf{p}^{(0)}$

$$\mathbf{p}^{(i+1)} = \mathbf{p}^{(i)} + \lambda \frac{\partial F}{\partial \mathbf{p}} \ , \tag{5}$$

where λ is a step size. After k iterations when the alignment process converges, the alignment is considered successful if the Euclidean distance $\|\mathbf{p}^{(k)} - \mathbf{p}_0\|$ is less than a pre-defined threshold.

From Equation (5), it is clear that $\frac{\partial F}{\partial \mathbf{p}}$ indicates the *travel direction* of the shape parameter $\mathbf{p}$. Because the final destination of such traveling is $\mathbf{p}_0$, the *ideal travel direction* should be the vector that points to $\mathbf{p}_0$ starting from $\mathbf{p}$, which is denoted as $\vec{\mathbf{p}}$:

$$\vec{\mathbf{p}}^+ \doteq \frac{\mathbf{p}_0 - \mathbf{p}}{\|\mathbf{p}_0 - \mathbf{p}\|}. \tag{6}$$

Similarly, the worst travel direction is the opposite direction of $\vec{\mathbf{p}}^+$, i.e., $\vec{\mathbf{p}}^- = -\vec{\mathbf{p}}^+$. Hence, during the learning of the score function F, we would like to see $\frac{\partial F}{\partial \mathbf{p}}$ has a direction that is as similar to the ideal travel direction $\vec{\mathbf{p}}^+$ as possible, or equivalently, as dissimilar to the worst travel direction $\vec{\mathbf{p}}^-$ as possible. Specifically, if we define a classifier

$$H(\mathbf{p};\vec{\mathbf{p}}) = \frac{\frac{\partial F}{\partial \mathbf{p}}}{\|\frac{\partial F}{\partial \mathbf{p}}\|}\vec{\mathbf{p}}, \tag{7}$$

which is the inner product between two unit vectors and is also the cosine response of the angle between these two vectors, then we have

$$H(\mathbf{p};\vec{\mathbf{p}}) = \begin{cases} +1 & if \ \ \vec{\mathbf{p}} = \vec{\mathbf{p}}^+ \ , \\ -1 & if \ \ \vec{\mathbf{p}} = \vec{\mathbf{p}}^- \ . \end{cases} \tag{8}$$

In practice, it is hard to expect $H(\mathbf{p})$ can always equal to 1 or -1 as shown in the above equation. Thus, we formulate the objective function of learning the H classifier as,

$$\arg\min_F \sum_{\mathbf{p}} (H(\mathbf{p};\vec{\mathbf{p}}^+) - 1)^2, \tag{9}$$

where only the ideal travel direction $\vec{\mathbf{p}}^+$ is used since it can represent the constraint from $\vec{\mathbf{p}}^-$ as well. From now on, we will simplify $\vec{\mathbf{p}}^+$ as $\vec{\mathbf{p}}$ for the clarity. This objective function essentially aims to estimate a function F such that its gradient direction has minimal angle with respect to the ideal travel direction, at all possible shape parameters $\mathbf{p}$ for all training data.

B. Solution for the Objective Function

In this section, we will describe our solution in minimizing the objective function (9). First, let us assume our alignment score function uses a simple additive model:

$$F(\mathbf{p}; m) \doteq \sum_{i=1}^{m} f_i(\mathbf{p}) \ , \qquad (10)$$

where $f_i(\mathbf{p})$ is a weak function that operates on one rectangular feature φ_i. Therefore, the gradient of F is also in an additive form: $\frac{\partial F(\mathbf{p};m)}{\partial \mathbf{p}} = \sum_{i=1}^{m} \frac{\partial f_i}{\partial \mathbf{p}}$. By plugging this into Equation (7), we have

$$\begin{aligned} H(\mathbf{p}; \vec{\mathbf{p}}, m) &= \frac{\sum_{i=1}^{m} \frac{\partial f_i}{\partial \mathbf{p}}}{\| \sum_{i=1}^{m} \frac{\partial f_i}{\partial \mathbf{p}} \|} \vec{\mathbf{p}} \\ &= \frac{H(\mathbf{p}; \vec{\mathbf{p}}, m-1) \| \frac{\partial F(\mathbf{p};m-1)}{\partial \mathbf{p}} \| + \frac{\partial f_m}{\partial \mathbf{p}} \vec{\mathbf{p}}}{\| \frac{\partial F(\mathbf{p};m-1)}{\partial \mathbf{p}} + \frac{\partial f_m}{\partial \mathbf{p}} \|} . \end{aligned} \qquad (11)$$

Given the fact that H function can be written in a recursive fashion, a natural way to minimizing the objective function (9) is to use an incremental estimation. That is, by defining a set of training samples and a hypothesis space where the rectangle feature can be chosen from, we can iteratively estimate each weak function f_i and incrementally add it into the target function F. We will now describe each part of the learning process as follows.

a) Training samples: In our appearance learning, a training sample is a N-dimensional warped image $\mathbf{I}(\mathbf{W}(\mathbf{x}; \mathbf{p}))$. Given a face database $\{\mathbf{I}_i\}_{i \in [1,K]}$ with manually labeled landmarks $\{\mathbf{s}_i\}$, for each face image $\mathbf{I}_i$, we use Equation (1) to compute the ground-truth shape parameter $\mathbf{p}_{0,i}$, and then synthesize a number of "incorrect" shape parameters $\{\mathbf{p}_{j,i}\}_{j \in [1,U]}$ by random perturbation. Equation (12) describes our perturbation, where $\boldsymbol{\nu}$ is a n-dimensional vector with each element uniformly distributed within $[-1, 1]$, $\boldsymbol{\mu}$ is the vectorized eigenvalues of all shape bases in the PDM, and perturbation index σ is a constant scale controls the range of perturbation. Note that $\circ$ represents the entrywise product of two equal-length vectors.

$$\mathbf{p}_{j,i} = \mathbf{p}_i + \sigma \boldsymbol{\nu} \circ \boldsymbol{\mu}. \qquad (12)$$

Then, the set of warped images $\mathbf{I}_i(\mathbf{W}(\mathbf{x}; \mathbf{p}_{j,i}))$ are treated as *positive training samples* ($y_i = 1$) for the learning. Together with the ideal travel direction, this constitutes our training set:

$$\mathfrak{P} \doteq \{\mathbf{I}_i(\mathbf{W}(\mathbf{x}; \mathbf{p}_{j,i})), \vec{\mathbf{p}}_i\}_{i=1,\cdots,K;\, j=1,\cdots,U} \ . \qquad (13)$$

b) Weak function: In this work, we define the weak function f_i as

$$f_i(\mathbf{p}) \doteq \frac{2}{\pi} \arctan(g_i \varphi_i(\mathbf{p}) - t_i) \ , \qquad (14)$$

where $g_i = \pm 1$, and the normalizing constant ensures that f_i stays within the range of $[-1, 1]$. This choice is based on a few considerations. First, f_i has to be differentiable because we assume F is a differentiable function. Second, we would like each function f_i operates on one rectangular feature φ_i only. Within the mean shape space, all possible locations, sizes, and types of the rectangular features form the hypothesis space $\mathcal{F} = \{\alpha, \beta, \gamma, \delta, \tau\}$, from which the best feature can be chosen at each iteration.

Algorithm 1: Model learning of OGPM

```
  Data: Positive samples $\mathfrak{P}$ from Equation (13)
  Result: The alignment score function $F$
1 Initialize the score function $F = 0$
2 foreach $t = 1, \cdots, m$ do
3     Fit $f_t$ in the weighted least squares sense, such that
          $$f_t = \underset{f}{\operatorname{argmin}} \sum_{ij} (1 - H(\mathbf{p}_{j,i}; \vec{\mathbf{p}}_i, t))^2 \qquad (15)$$
4     Update $H(\mathbf{p}_{j,i}; \vec{\mathbf{p}}_i, t)$ with $f_t$
5     $F \longleftarrow F + f_t$
6 return $F = \sum_{t=1}^{m} f_t$.
```

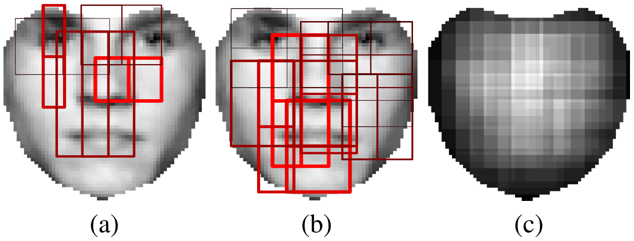

Fig. 4. **Top Appearance Features.** (a) Top 5 Haar features selected by the our learning algorithm. (b) Top 6-15 Haar features. (c) Spatial density map of the top 100 Haar features.

c) Learning procedure: Algorithm 1 describes the procedure for learning the alignment score function (10). Note that Step 3 is the most computationally intensive step since the entire hypothesis space needs to be exhaustively searched. Our incremental estimation of the score function F is very similar to the boosting algorithm. The reason that boosting is not used here is the function H can not be simply represented in an additive form. Hence, in Step 3, the best feature is chosen based on the L^2 distance of H with respect to 1, rather than that of the weak classifier in boosting-based learning.

Basically learning the score function F is equivalent to learning the set of features $\{\varphi_i\}$, the thresholds $\{t_i\}$, and the feature signs $\{g_i\}$. In practical implementation, we set $g_i = +1$, and $g_i = -1$ respectively and estimate the optimal threshold for both cases. Eventually g_i will be set based on which case has a smaller error (Equation 15). The optimal threshold is estimated by binary searching in the range of feature values φ_i such that the error is minimized.

The final set of triples $\{(\varphi_i, g_i, t_i)\}_{i=1,\cdots,m}$, together with the shape model $\{\mathbf{s}_i\}_{i=0,\cdots,n}$ is called an *Optimal Gradient Pursuit Model* (OGPM). Figure 4 shows the top 15 features selected by the learning algorithm, as well as the spatial density map of the top 100 features. Notice that many selected features are aligned with the boundaries of the facial features.

Fig. 5. **Face Dataset Samples.** ND1 database [2] (left), FERET database [21] (center), and BioID database [23] (right).

V. Face Alignment

In this section, we will describe how to fit an OGPM to the face of a given image $\mathbf{I}$, with an initial shape parameter $\mathbf{p}^{(0)}$ (at the 0-th iteration). As shown in Equation (5), the alignment is iteratively performed by using the gradient ascent approach. From Equation (3), (10), and (14), we can see that the derivative of F with respect to $\mathbf{p}$ is

$$\frac{\partial F}{\partial \mathbf{p}} = \frac{2}{\pi}\sum_{i=1}^{m}\frac{g_i\left(\nabla \mathbf{I}\frac{\partial \mathbf{W}}{\partial \mathbf{p}}\right)^T \mathbf{A}_i}{1+\left(g_i\mathbf{A}_i^T\mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p}))-t_i\right)^2}\ , \qquad (16)$$

where $\nabla \mathbf{I}$ is the gradient of the image evaluated at $\mathbf{W}(\mathbf{x};\mathbf{p})$, and $\frac{\partial \mathbf{W}}{\partial \mathbf{p}}$ is the Jacobian of the warp evaluated at $\mathbf{p}$. The BAM [14] has a detailed discussion on the alignment procedure, and the computational complexity, and efficient implementation of $\frac{\partial F}{\partial \mathbf{p}}$. Compare to the BAM-based fitting, one improvement we have is that the step size λ is dynamically determined via line searching, rather that a simple static constant. That is, at each iteration, we search for the optimal λ within certain range such that the updated shape parameter can maximally increase the current score function value $F(\mathbf{p})$.

VI. Experiments

In this section we will present the various experiments to demonstrate the properties of the proposed approach. We begin with the description on the dataset used in our experiments and then introduce each experiment.

A. Experimental Setup

TABLE I
Summary of the dataset.

	ND1	FERET	BioID
Images	534	200	230
Subjects	200	200	23
Variations	Frontal view	Pose	Background, lighting
Set 1	200	200	
Set 2	334		
Set 3			230

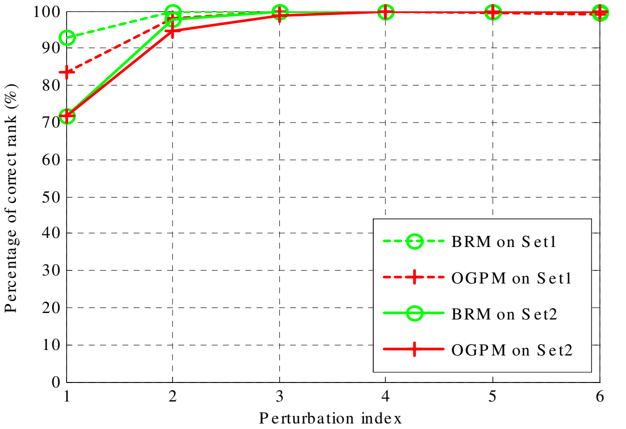

Fig. 6. **Ranking Performance.** For both Set 1 and 2, OGPM achieves similar ranking performance as BRM.

Our experimental dataset contains 964 images from three public available databases: the ND1 database [2], FERET database [21] and BioID database [23]. There are 33 manually labeled landmarks for each of the 964 images. To speed up the training process, we down-sample the image set such that the facial width is roughly 40 pixels across the set. Sample images of these databases are illustrated in Figure 5. As shown in Table I, we partition all images into three non-overlapping datasets. Set 1 includes 400 images (one image per subject) from two databases. Set 2 includes 334 images from the *same* subjects but different images as the ND1 database in Set 1. Set 3 includes 230 images from 23 subjects in the BioID database that were never used in the training. Set 1 is used as the training set for the model learning and all three sets are used for testing the model fitting. The motivation for such a partition is to experiment various levels of generalization capability. For example, Set 2 can be tested as the unseen data of seen subjects; Set 3 can be tested as the unseen data of unseen subjects, which is the most challenging case and most similar to the scenario in practical applications.

In the experiments we compare our OGPM algorithm with BRM based on two considerations. First, our proposed algorithm is a direct extension of BRM. Second, it has been shown the BRM outperforms other discriminative image alignment such as BAM. During the model learning, both BRM and OGPM are trained from 400 images of Set 1. BRM uses 24000 $(= 400 \times 10 \times 6)$ training samples synthesized from Set 1, where each image synthesizes 10 *profile lines* and each line has 6 evenly spaced samples. In comparison, OGPM uses 12000 training samples, where each image synthesizes 30 samples according to Equation (12). The reason we use less samples for OGPM is that all synthesized samples are randomly spread out, rather than multiple samples selected from one profile line as in BRM. Hence, OGPM is likely to achieve good performance with less training samples. The manually labeled landmarks of Set 1 images are improved using the automatical model refinement approach in [17]. After model learning, the shape model component of both BRM and OGPM is a PDM with 9 shape bases, and their appearance model (i.e., the alignment score function) has 100 weak classifiers/functions.

TABLE II

ALIGNMENT PERFORMANCE (PIXELS) COMPARISON ON THREE SETS. σ INDICATES THE AMOUNT OF PERTURBATION.

	σ	2	4	6	8
Set 1	BRM	0.50	1.12	1.30	1.45
	OGPM	0.47	0.57	0.70	0.87
Set 2	BRM	0.88	0.94	1.02	1.12
	OGPM	0.58	0.72	0.81	0.93
Set 3	BRM	0.85	1.34	1.59	1.94
	OGPM	0.80	1.12	1.35	1.60

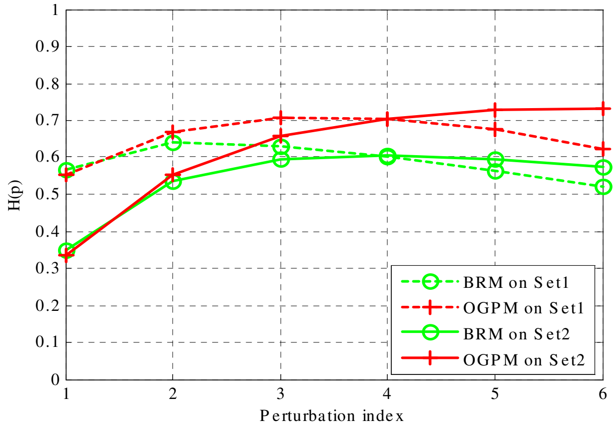

Fig. 7. **Angle Estimation Performance.** Compared to BRM, OGPM clearly can estimate the gradient direction that is more similar to the ideal travel direction for model fitting.

B. Experimental Results

BRM aims to improve the convexity of the learned score function by correctly ranking pair of warped images. OGPM extents BRM in the sense that the score function should not only be concave, but also have minimal angle between the gradient direction and the vector pointing to the ground-truth shape parameter. Hence, convexity is a good metric for evaluating the score functions for both BRM and OGPM. Similar to BRM, the convexity is measured by computing the percentage of correctly ranked pairs of warped images. Given Set 1 and Set 2, we synthesize two sets of pairs respectively and test the ranking performance of BRM and OGPM. As shown in Figure 6, the perturbation index σ controls the amount of perturbation of the image pair (see Equation 12). We can see that for both sets, OGPM achieves very similar ranking performance as BRM, despite the fact that, unlike BRM, OGPM does not utilize ranking in its objective function directly. The only exception is the slight better performance of BRM when the perturbation is very small ($\sigma = 1$). We attribute this mostly to the labeling error in the training data, since a small perturbation of labeled landmark can also be treated as a fairly good alignment, which makes the ranking harder.

In addition to the convexity measure, we also validate the estimation of the angle between the gradient direction and the vector pointing to the ground-truth shape parameter. The minimization of this angle is the objective function of OGPM, as represented by the $H(\mathbf{p})$ function. Similar to the aforementioned ranking experiments, given the Set 1, we randomly synthesize six sets of warped images using various perturbation index σ. Then for each image in a set, we compute the $H(\mathbf{p})$ score, and plot the average score of each set in Figure 7. Similar experiments are conducted for Set 2 as well. It is obvious that even though OGPM and BRM have similar ranking performance, OGPM achieves larger function score for both Set 1 and 2, hence smaller gradient angle. This shows that using ranking performance as the objective, as done by BRM, does not guarantee the optimal angle estimation. Rather, we should directly use the gradient angle as the objective function in order to obtain a better alignment score function, as done by OGPM.

In the alignment experiments, we run the model fitting algorithm on each image with a number of initial landmarks and evaluate the alignment results. The initial landmarks are generated using Equation (12), i.e., randomly perturbing the ground-truth landmarks by an independent uniform distribution whose range equals to a multiple (σ) of the eigenvalue of shape basis during PDM training. Once fitting on one image terminates, the alignment performance is measured by the resultant Root Mean Square Error (RMSE) between the aligned landmarks and the ground-truth landmarks.

We conduct the alignment experiments for all three sets using both OGPM and BRM. Table II shows the RMSE results in terms of pixels, where each element is an average of more than 2000 trials at one particular perturbation index σ. Hence, each image in Set 1, 2, and 3 is tested with 5, 6, and 9 random trials, respectively. OGPM and BRM are tested under the same condition. For example, both algorithms are initialized with the same random trails and the termination condition is the same as well. That is, the alignment iteration exits if the alignment score $F(\mathbf{p})$ can not increase further, or the landmark difference (RMSE) between consecutive iterations is less than a pre-defined threshold, which is 0.05 pixel in our work.

From Table II, we can see that for all three sets, OGPM is able to achieve better alignment performance than BRM. Note that the performance gain is more when the initial perturbation is relatively large, such as σ=6 or 8, which are the most challenging cases in practical applications. Given the fact that our test images are in very low resolution, we consider this is substantial performance improvement. Comparing among the three data sets, the performance gain in the training set (Set 1) is larger compared to the other two data sets.

One obvious strength of smaller gradient angles is the ability to converge in less iterations during the alignment. In Figure 8, we show the histogram of the number of iterations that OGPM and BRM requires to converge on Set 3, when $\sigma = 8$. We can see that on average OGPM can converge faster than BRM. Here, the average number of iterations of OGPM is 5.47, while that of BRM is 6.40. Similarly, on Set 1, the average number of iterations of OGPM is 5.08, and

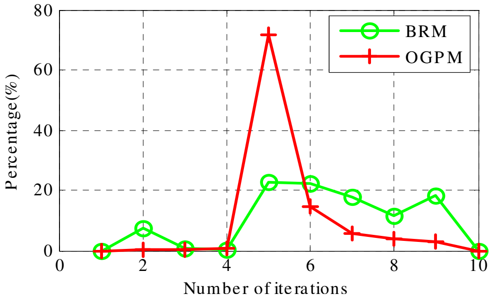

Fig. 8. **Alignment Speed.** Histogram of the number of iterations for fitting on Set 3 when σ=8.

that of BRM is 6.09, when $\sigma = 8$.

VII. Conclusions

This paper proposes a novel face model, "Optimal Gradient Pursuit Model", for facial image alignment. Motivated by the fact that a high dimensional concave score function can have sub-optimal gradient directions, the objective of our face model learning lies in the minimization of the angle between the gradient direction and the vector pointing toward the ground-truth shape parameter. We formulate an iterative approach to solve this minimization problem. Through extensive experiments, we show that our model can improve the alignment accuracy and speed compared to the BRM approach.

Future work includes applying this image alignment framework to objects other than faces since no prior knowledge of human faces is used in our approach. Also, since discriminative face alignment can take advantage of infinite number of training samples ($\mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p}))$) through synthesis, there is a possibility that model learning can be conducted with a small number of labeled face images, which may leverage the work of semi-supervised image alignment [15], [16], [25].

References

[1] S. Baker and I. Matthews. Lucas-Kanade 20 years on: A unifying framework. *Int. J. Computer Vision*, 56(3):221–255, March 2004.

[2] K. Chang, K. Bowyer, and P. Flynn. Face recognition using 2D and 3D facial data. In *Proc. ACM Workshop on Multimodal User Authentication*, pages 25–32, December 2003.

[3] T. Cootes, D. Cooper, C. Tylor, and J. Graham. A trainable method of parametric shape description. In *Proc. of the British Machine Vision Conference (BMVC)*, pages 54–61, Glasgow, UK, Sept. 1991.

[4] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 23(6):681–685, June 2001.

[5] D. Cristinacce and T. Cootes. Facial feature detection and tracking with automatic template selection. In *Proc. of Int. Conf. on Automatic Face and Gesture Recognition (FG)*, pages 429–434, Southampton, UK, 2006.

[6] D. Cristinacce and T. Cootes. Boosted regression active shape models. In *Proc. of the British Machine Vision Conference (BMVC)*, volume 2, pages 880–889, University of Warwick, UK, 2007.

[7] G. Dedeoglu, T. Kanade, and S. Baker. The asymmetry of image registration and its application to face tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(5):807–823, May 2007.

[8] R. Donner, M. Reiter, G. Langs, P. Peloschek, and H. Bischof. Fast active appearance model search using canonical correlation analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 28(10):1690–1694, 2006.

[9] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *J. Image and Vision Computing*, 23(11):1080–1093, Nov. 2005.

[10] G. Hager and P. Belhumeur. Efficient region tracking with parametric models of geometry and illumination. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 20(10):1025–1039, 1998.

[11] A. Kanaujia and D. Metaxas. Large scale learning of active shape models. In *Proc. of the International Conference on Image Processing (ICIP)*, volume 1, pages 265–268, San Antonio, Texas, 2007.

[12] L. Liang, F. Wen, Y. Xu, X. Tang, and H. Shum. Accurate face alignment using shape constrained Markov network. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, New York, NY, June 2006.

[13] X. Liu. Generic face alignment using boosted appearance model. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Minneapolis, MI, 2007.

[14] X. Liu. Discriminative face alignment. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(11):1941–1954, November 2009.

[15] X. Liu, Y. Tong, and F. W. Wheeler. Simultaneous alignment and clustering for an image ensemble. In *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, Kyoto, Japan, Oct. 2009.

[16] X. Liu, Y. Tong, F. W. Wheeler, and P. H. Tu. Facial contour labeling via congealing. In *Proc. of the European Conf. on Computer Vision (ECCV)*, Crete, Greece, Sept. 2010.

[17] X. Liu, P. Tu, and F. Wheeler. Face model fitting on low resolution images. In *Proc. of the British Machine Vision Conference (BMVC)*, volume 3, pages 1079–1088, 2006.

[18] I. Matthews and S. Baker. Active appearance models revisited. *Int. J. Computer Vision*, 60(2):135–164, November 2004.

[19] M. H. Nguyen and F. D. la Torre Frade. Learning image alignment without local minima for face detection and tracking. In *Proc. of the British Machine Vision Conference (BMVC)*, Leeds, UK, 2008.

[20] C. P. Papageorgiou, M. Oren, and T. Poggio. A general framework for object detection. In *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, pages 555–562, 1998.

[21] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi. The FERET evaluation methodology for face recognition algorithms. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 22(10):1090–1104, October 2000.

[22] J. Saragih and R. Goecke. A nonlinear discriminative approach to AAM fitting. In *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, Oct. 2007.

[23] M. B. Stegmann, B. K. Ersboll, and R. Larsen. FAME - A flexible appearance modeling environment. *IEEE Trans. on Medical Imaging*, 22(10):1319–1331, October 2003.

[24] R. Szeliski and J. Coughlan. spline-based image registration. *Int. J. Computer Vision*, 22(3):199–218, 1997.

[25] Y. Tong, X. Liu, F. W. Wheeler, and P. Tu. Automatic facial landmark labeling with minimal supervision. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Miami, FL, June 2009.

[26] B. C. Vemuri, S. Huang, S. Sahni, C. M. Leonard, C. Mohr, R. Gilmore, and J. Fitzsimmons. An efficient motion estimator with application to medical image registration. *Medical Image Analysis*, 2(1):79–98, March 1998.

[27] P. Viola and M. Jones. Robust real-time face detection. *Int. J. Computer Vision*, 57(2):137–154, May 2004.

[28] C. Vogler, Z. Li, A. Kanaujia, S. Goldenstein, and D. Metaxas. The best of both worlds: Combining 3D deformable models with active shape models. In *Proc. of the Intl. Conf. on Computer Vision (ICCV)*, Rio de Janeiro, Brazil, Oct. 2007.

[29] O. Williams, A. Blake, and R. Cipolla. Sparse Bayesian learning for efficient visual tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 27(8):1292–1304, 2005.

[30] M. Wimmer, F. Stulp, S. Pietzsch, and B. Radig. Learning local objective functions for robust face model fitting. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 30(18):1357–1370, August 2008.

[31] H. Wu, X. Liu, and G. Doretto. Face alignment via boosted ranking models. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, Anchorage, Alaska, 2008.

[32] S. Zhou and D. Comaniciu. Shape regression machine. In *Proc. of Int. Conf. on Information Processing in Medical Imaging (IPMI)*, pages 13–25, Kerkrade, The Netherlands, July 2007.

[33] Y. Zhou, L. Gu, and H. Zhang. Bayesian tangent shape model: Estimating shape and pose parameters via Bayesian inference. In *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, volume 1, pages 109–116, Madison, WI, 2003.