A Generative Shape Regularization Model for Robust Face Alignment

Leon Gu and Takeo Kanade

Computer Science Department
Carnegie Mellon University

Abstract. In this paper, we present a robust face alignment system that is capable of dealing with *exaggerating expressions*, *large occlusions*, and *a wide variety of image noises*. The robustness comes from our shape regularization model, which incorporates constrained nonlinear shape prior, geometric transformation, and likelihood of multiple candidate landmarks in a three-layered generative model. The inference algorithm iteratively examines the best candidate positions and updates face shape and pose. This model can effectively recover sufficient shape details from very noisy observations. We demonstrate the performance of this approach on two public domain databases and a large collection of real-world face photographs.

1 Introduction

Face alignment is a challenging problem especially when it comes to real-world images. The main difficulty arises from pervasive ambiguities in low-level image features. Consider the examples in Figure 1. While the main face structures are present in the feature maps, the boundaries of face components are frequently disrupted by gaps or corrupted by spurious fragments. Strong gradient responses could be due to reflectance, occlusion, image blur, or fine facial textures. In contrast, the boundaries of nose, jawline, and other subparts could be obscure or even barely perceptible. Detecting individual facial landmarks in such feature maps often yields noisy results. Yet, an interesting question is, on the basis of such "observations" what is the best hypothesis that one can make?

We believe that the ability to recover sufficient shape details from noisy image observations lies at the core of a *robust* face alignment system. In this paper, we address this problem by use of a three-layered generative model. The model allows multiple candidate positions to be generated for each facial landmark, and the image likelihood of seeing a landmark at one particular position is defined at the bottom layer of the model. The best candidate is treated as a hidden variable to be decided. The middle layer models global geometric transformation in which a noise term is assigned to each landmark to measure its fitting error. The top layer models prior shape distribution as Gaussian mixture, in which the number of free parameters in each Gaussian component is restricted to be small. This prior is compact, and also flexible to capture large shape deformations. We use expectation-maximization algorithm to iteratively select candidates and estimate

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part I, LNCS 5302, pp. 413–426, 2008.

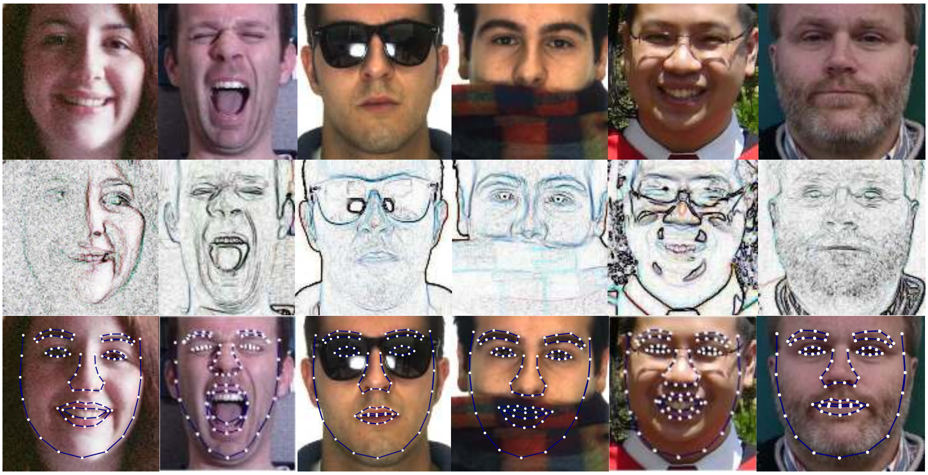

Fig. 1. Real-world face images could be extremely noisy in the eyes of computers. As shown in in the gradient feature maps (second row), face topologies could be significantly corrupted due to various kinds of factors. However, our model can still recover sufficient shape details from such noisy observations while using a simple gradient-based landmark detector. The third row shows our alignment results.

face shape and pose. We show that the model can deal with observation noises in a way such that the major trends of shape deformations can still be recovered, and the well-matched facial landmarks can be differentiated from outliers. One advantage of putting deformable matching in this generative setting and using EM for inference is that our treatment is guaranteed to be consistent.

The rest of this paper is organized as follows. The background and related works are introduced in section 2. We explain the details of the model including the inference algorithm in section 3, and show the experimental results and comparison in section 4. We conclude in section 5.

2 Background

Leveraging prior knowledge on low-level image interpretation has been the core idea of deformable matching since Eighties [1] [2] [3] [4] [5]. The forms of priors that have been imposed on images are diverse, including generic properties of parametric curves such as continuity and smoothness [1], specific object structures defined by assemblies of flexible curves [3], linear shape deformation subspace learned from landmarked training images [1], and shape dynamics [5]. Latter developments have been focused on modeling object appearance [6] [7]; exploiting nonlinear [8] [9] [10] and 3D [7] [11] [12] shape prior; refining alignment results using Markov network [13]; and seeking efficient matching algorithms by dynamic programming [14] [15] or belief propagation [16] when shape priors are defined on discrete spaces. Progresses have also been made to improve facial landmark encoding [17] and detection [18].

It is also worth mentioning that Gaussian mixture was used in [8] to refine the linear subspace model [4], but their shape regularization approach is completely deterministic. Our generative formulation for deformable matching is similar to the recently proposed methods by Zhou et al. [19] and Gu et al. [11]. These methods, however, are restricted to linear shape deformations, and likely to oversimplify shape observations - only one candidate position for each facial landmark.

3 The Generative Model for Shape Regularization

We follow the standard shape representation. A face Q consists of N landmark points, i.e., $Q = (Q_1^x, Q_1^y, \ldots, Q_N^x, Q_N^y)^t$. The landmarks are commonly placed along the boundaries of face components. The geometry information of Q decouples into two parts: a canonical shape S, and a rigid transformation $\mathcal{T}_\theta$ that maps S from a common reference frame to the coordinate plane of the target image I.

When low-level features are ambiguous, it makes sense to allow landmark detectors to produce multiple candidates. Suppose that for the n-th landmark, there are K candidate positions $Q_{nk} = (Q_{nk}^x, Q_{nk}^y)$ located on the image. Let $\mathcal{Q} = \{Q_{nk}\}$ denote the whole set of $N \times K$ candidates. Our goal is to estimate the shape S and the pose θ from $\mathcal{Q}$.

3.1 Model Structure

First a decision needs to be made for each landmark to select the "best" candidate. We introduce a latent variable h that assigns one candidate position to each landmark. h is a binary $N \times K$ matrix, in which each row contains only one "1" and all other entries are zeros. The image likelihood of seeing a landmark at one particular position Q_{nk} is measured by

$$p(I|h_{nk} = 1) = p(I|Q_{nk}) = \pi_{nk}, \tag{1}$$

and subject to the constraint $\sum_k \pi_{nk} = 1$. Let $\mathcal{Q}(h)$ denote the set of positions selected by h. Then we can write the n-th point of $\mathcal{Q}(h)$ as

$$\mathcal{Q}_n(h) = \left(\sum_{k=1}^{K} h_{nk} Q_{nk}^x, \sum_{k=1}^{K} h_{nk} Q_{nk}^y \right)^t . \tag{2}$$

Assume that $\mathcal{Q}(h)$ is generated from the canonical shape S and the pose θ by first transforming S according to θ, then adding an independent Gaussian noise to each landmark. We can write the conditional probability of $\mathcal{Q}(h)$ as

$$p(\mathcal{Q}(h)|S, \theta) = \mathcal{N}(\mathcal{T}(S, \theta); \Sigma), \tag{3}$$

where the covariance matrix Σ is diagonal, i.e., $\Sigma = \mathrm{diag}(\rho_1, \rho_1, \ldots, \rho_N, \rho_N)$. The independence assumption is valid because the detection of landmarks is often performed independently to each other. The variance ρ_n measures the noise level of the observation $\mathcal{Q}_n(h)$. A non-informative prior is put on the similarity transformation parameters $\theta = \{R, s, t\}$ to allow arbitrary rotation, translation and isotropic scaling.

We define the prior distribution over the shape S as a mixture of constrained Gaussian [20],

$$p(S|b) = \sum_{l=1}^{L} \pi_l \mathcal{N}(\Phi_l b_l + \mu_l; \sigma_l^2 I), \tag{4}$$

where $b = \{b_l\}$ denotes the deformation coefficients, and the model parameters associated with each Gaussian component are the mixing rate π_l, the linear principal subspace spanned by the columns of Φ_l, the mean shape μ_l, and the isotropic shape noise with zero mean and variance $\sigma_l^2 I$. Compared to general Gaussian mixtures, this mixture model is more compact and contains less free parameters. Each Gaussian component is restricted in a linear subspace spanned by Φ_l, and the dimension of the subspace is decided by the percentage of shape variance that is desired to be preserved. Within each subspace, b_l controls the amount of shape deformation, and σ_l^2 determines the variance of shape noise. The variance σ_l^2 is computed as the average residual shape variance outside of the subspace,

$$\sigma_l^2 = \frac{1}{N - M_l} \sum_{m=M_l+1}^{N} \lambda_{lm}. \tag{5}$$

Here $\{\lambda_{lm}\}$ denote the eigenvalues, which are arranged in a deceasing order, and M_l is the subspace dimension. Other model parameters $\{\pi_l, \mu_l, \Phi_l\}$ are also learned from training shapes (see [20] for details). The difference from [20] is that we model the deformation prior $p(b_l)$ as a diagonal Gaussian

$$p(b_l) = \mathcal{N}(0; Diag(\lambda_{l1}, \ldots, \lambda_{lM_l})), \tag{6}$$

and restrict the columns of matrix Φ_l to be orthogonal. We then rewrite (4) by introducing a latent component label z,

$$p(S|b, z) = \prod_{l} \mathcal{N}^{z_l}(\Phi_l b_l + \mu_l; \sigma_l^2 I), \tag{7}$$

and putting a multinomial distribution on z

$$p(z_l = 1) = \pi_l. \tag{8}$$

Combining (1) $\sim$ (8), we construct a hierarchical deformable model.

3.2 The Alignment Algorithm

Our problem now is to estimate the *deformation parameters* b and the *transformation parameters* θ from the candidate points set $\mathcal{Q}$. This is formulated as a

MAP problem, i.e., finding the optimum $\{b^*, \theta^*\}$ by maximizing the posterior $p(b, \theta|\mathcal{Q})$, and solved by EM. First we look at the joint distribution over the assignment variable h, the mixture component label z, the hidden shape vector S, the deformation parameters b, and the transformation parameters θ,

$$p(b,\theta,S,h,z|I) \propto p(S|b,z)p(b)p(z)p(\mathcal{Q}(h)|S,\theta)p(I|h) \tag{9}$$

Taking the expectation $\langle\cdot\rangle$ of the log of (9) over the posterior of the latent variables S, h, z, we obtain the so-called Q-function,

$$\begin{aligned}\langle \log p(b,\theta,S,h,z|I)\rangle &\propto \langle \log p(S|b,z)\rangle + \log p(b) \\ &+ \langle \log p(z)\rangle + \langle \log p(\mathcal{Q}(h)|S,\theta)\rangle + \langle \log p(I|h)\rangle .\end{aligned} \tag{10}$$

In the E-step, we compute the sufficient statistics that are required to evaluate (10); and in the M-step we maximize (10) to find the updated shape and pose.

Expectation Step. Substituting for the expectations on the right-hand side of (10) and absorbing terms that are independent of b and θ into an additive constant, we expand (10) as

$$\begin{aligned}&\langle \log p(b,\theta,S,h,z|\mathcal{Q})\rangle_{S,h} \propto \sum_{n,k} \langle h_{nk}\rangle \log \pi_{nk} \\ &+\sum_{l} \langle z_l\rangle \log \pi_l - \tfrac{1}{2}\sum_{n} \rho_n^{-2} \left\langle \|\mathcal{Q}(h_n) - \mathcal{T}(S_n,\theta)\|^2 \right\rangle \\ &-\tfrac{1}{2}\sum_{l} \left\langle z_l \sigma_l^{-2} \|S - \Phi_l b_l - \mu_l\|^2 \right\rangle - \tfrac{1}{2}\sum_{l,m} \lambda_{lm} b_{lm}^2 .\end{aligned} \tag{11}$$

In order to evaluate (11), we need the joint and marginal posteriors of all discrete and continuous latent variables. We first write down the joint posterior, which is given by

$$\begin{aligned}&p(S,h,z|I,b,\theta) \propto \\ &\quad p(S|b,z)p(z)p(\mathcal{Q}(h)|S,\theta)p(I|h).\end{aligned} \tag{12}$$

Because of the conditional independence assumptions made in (1) (3) (7), the right side of (12) can be further factorized between individual points

$$p(z) \prod_{n=1}^{N} \{p(S_n|b,z)p(\mathcal{Q}(h_n)|S_n,\theta)p(I|h_n)\} . \tag{13}$$

Thus we can evaluate $p(S_n, h_n|I, b, \theta, z)$ for each point separately. We factorize it by chain rule

$$p(S_n,h_n|I,b,\theta,z) = p(S_n|h_n,I,b,\theta,z)p(h_n|I,b,\theta,z). \tag{14}$$

According to the Bayes' rule, the first factor decomposes into a product of the prior $p(S_n|b, z)$ and the likelihood $p(\mathcal{Q}(h_n)|S_n, \theta)$. Because both distributions are Gaussian and similarity transform is linear, the posterior is still a Gaussian.

$$p(S_n|h_n, l, b, \theta, z) = \mathcal{N}(\bar{S}_{nkl}, c_{nl}^2 I_{2\times 2}). \tag{15}$$

Its mean and covariance are given by

$$\bar{S}_{nkl} = w_l^1 S_n(b_l) + w_l^2 \mathcal{T}^{-1}(h_n) \tag{16}$$

$$c_{nl}^2 = (\sigma_l^{-2} + s^2 \rho_n^{-2})^{-1} \tag{17}$$

where we have defined

$$w_l^1 = \frac{\sigma_l^{-2}}{\sigma_l^{-2}+s^2\rho_n^{-2}} \tag{18}$$

$$w_l^2 = \frac{s^2\rho_n^{-2}}{\sigma_l^{-2}+s^2\rho_n^{-2}} \tag{19}$$

$$S(b_l) = \Phi_l b_l + \mu_l \tag{20}$$

$$\mathcal{T}^{-1}(h_n) = \mathcal{T}^{-1}(\mathcal{Q}(h_n), \theta) \tag{21}$$

where $\mathcal{T}^{-1}$ denotes the inverse similarity transformation, and the subscript in S_n denotes the n-th landmark. For the second factor in (14) we marginalize the joint posterior $p(S_n, h_n | I, b, \theta, z)$ over S_n

$$\begin{aligned} p(h_n|I,b,\theta,z) &\propto p(I|h_n) \\ &\int p(S_n|b,z)p(\mathcal{Q}(h_n)|S_n,\theta)dS_n. \end{aligned} \tag{22}$$

The integral in (22) is a function of h_n, and its value measures a scaled distance between the model prediction $S(b_z)$ and the observed candidate position specified by h. Requiring that the distribution (22) be normalized, we obtain,

$$p(h_{nk} = 1 | I, b, \theta, z_l = 1) \propto \pi_{nk} r_{nkl} \tag{23}$$

where r_{nkl} is the exponential of the scaled distance, given by

$$r_{nkl} = \exp\{-\frac{||\sigma_l^{-2} S_n(b_l) - s^2 \rho_n^{-2} \mathcal{T}^{-1}(h_{nk})||^2}{2(\sigma_l^{-2} + s^{-2}\rho_n^2)(s/\rho_n)^8}\} \tag{24}$$

From (15) and (23), we shall see that the distribution $p(S_n|b, \theta, z)$ is a mixture of K Gaussian components, in which the mixing rate is given by (23) and the component mean and covariance are given by (16) and (17) respectively. We now integrate (13) over both S and h, and make use of (23) to compute $p(z|I, b, \theta)$

$$p(z_l = 1 | I, b, \theta) \propto \pi_l \prod_n \sum_k \pi_{nk} r_{nkl}. \tag{25}$$

Further decomposing the right-hand side of (11), we shall find this expectation depends on the posterior distributions only through the following statistics $\langle z_l \rangle$, $\langle h_{nk} \rangle$, $\langle S_n \rangle$, $\langle S_n^t S_n \rangle$, $\langle h_{nk} Q_{nk}^t R S_n \rangle$, $\langle z_l S_n \rangle$, $\langle z_l S_n^t S_n \rangle$. At this point, we shall find it convenient to define

$$\pi_l' := p(z_l = 1 | I, b, \theta) \tag{26}$$

$$\pi_{nkl}' := p(h_{nk} = 1 | I, b, \theta, z_l = 1) \tag{27}$$

$$\bar{S}_{nl} := \sum_k \pi_{nkl}' \bar{S}_{nkl}. \tag{28}$$

The sufficient statistics are easily evaluated from the distributions defined by (15) (23) (25), to give

$$\langle z_l \rangle = \pi_l' \tag{29}$$

$$\langle h_{nk} \rangle = \sum_l \pi_l' \pi_{nkl}' \tag{30}$$

$$\langle S_n \rangle = \sum_l \pi_l' \bar{S}_{nl} \tag{31}$$

$$\langle z_l S_n \rangle = \pi_l' \bar{S}_{nl} \tag{32}$$

$$\left\langle S_n^t S_n \right\rangle = \sum_{k,l} \pi_l' \pi_{nkl}' (\bar{S}_{nkl}^t \bar{S}_{nkl} + 2c_{nl}^2) \tag{33}$$

$$\left\langle h_{nk} Q_{nk}^t R S_n \right\rangle = Q_{nk}^t R \sum_l \pi_l' \pi_{nkl}' \bar{S}_{nkl} \tag{34}$$

$$\left\langle z_l S_n^t S_n \right\rangle = \pi_l' \sum_k \pi_{nkl}' (\bar{S}_{nkl}^t \bar{S}_{nkl} + 2c_{nl}^2) \tag{35}$$

Thus in E step we use old parameters $\{b^{\text{old}}, \theta^{\text{old}}\}$ to evaluate the posterior statistics (29∼35).

Maximization Step. In the M step, we maximize (11) with respect to b and θ using the sufficient statistics (29 ∼ 35). Note that b and θ are decoupled in (11), thus we can solve them separately. Taking the derivative of (11) with respect to b_l and setting it to zero and making use of the statistics (32) (29), we obtain the updating equation for the deformation parameters,

$$\tilde{b}_l = \frac{\langle z_l \rangle \sigma_l^{-2} (\Phi_l^t \bar{S}_l - \mu_l)}{\langle z_l \rangle \sigma_l^{-2} + \Lambda_l^{-1}}, \tag{36}$$

where $\bar{S}_l$ is a shape vector defined by $\bar{S}_l = (\bar{S}_{1l}, \ldots, \bar{S}_{Nl})^t$ and $\bar{S}_{nl}$ is defined in (28).

Substituting (30∼34) into (11), taking the derivatives with respect to s and t and setting them to zero, we obtain the updating equations for translation and scale,

$$\tilde{t} = \frac{1}{N} \sum_{n,k,l} \pi_l' \pi_{nkl}' Q_{nk} \tag{37}$$

$$\tilde{s} = \frac{\sum\limits_n \left(\tilde{t}^t \tilde{R} \langle S_n \rangle - \sum\limits_k \left\langle h_{nk} Q_{nk}^t \tilde{R} S_n \right\rangle \right)}{\sum\limits_n \langle (S_n^t S_n) \rangle}. \tag{38}$$

For rotation, maximizing (11) is equivalent to maximizing the trace

$$\text{Trace} \left\{ R \sum_n \left(\langle S_n \rangle \tilde{t}^t - \left\langle h_k S_n Q_{mk}^t \right\rangle \right) \right\}. \tag{39}$$

Therefore, the optimal rotation $\tilde{R}$ can be computed by polar decomposing the matrix $\sum\limits_n \left(\langle S_n \rangle \tilde{t}^t - \langle h_k S_n Q_{mk}^t \rangle \right)$.

3.3 Analysis

We first summarize the alignment algorithm. The inputs of the algorithm are the candidate position set $\{Q_{nk}\}$ and the image likelihood of each candidate $\{\pi_{nk}\}$; and the outputs are the optimal deformation and pose parameters $b = \{b_l\}, \theta = \{R, s, t\}$. Starting from an initial estimate $\{b^0, \theta^0\}$, the algorithm first computes a set of sufficient statistics (29) $\sim$ (35), then use them to update b (36) and θ (37) $\sim$ (39). Next we analyze the algorithm by looking into the details in (29) $\sim$ (39).

Making Use of Multiple Candidates: the selection of the best candidate is performed in a "soft" way by evaluating the posterior assignment probabilities (23) (30). Conditional on a particular mixture component l, the prior assignment π_{nk} is modulated by the similarity measure r_{nkl} between the position $S_n(b_l)$ predicted from the l-th component and the observed position Q_{nk}, to produce the conditional posterior assignment (23); the marginal posterior (30) is then computed by averaging (23) over all mixture components. These posteriors are used to weight candidate points to generate "averaged" observations in (28) - for a particular mixture component l, and (31) - for the whole mixture distribution. The posterior assignments are also used in pose estimation through (37) $\sim$ (39), to increase the contribution of good candidates.

Regularization by Multi-modal Prior: The averaged "observation" (28) is regularized by shape priors to produce a shape estimate. This regularization step is performed in (36), by first computing the subspace representation $\bar{b}_l = (\Phi_l^t \bar{S}_l - \mu_l)$ in each mixture component, then shrinking the deformation coefficients $\bar{b}_{lm}$ by a factor of $\frac{\langle z_l \rangle \sigma_l^{-2}}{\langle z_l \rangle \sigma_l^{-2} + \lambda_{lm}^{-1}}$. We see that the *degree of regularization* is determined in terms of three factors: the subspace responsibility $\langle z_l \rangle$, the shape variance λ_{lm} and the shape noise variance σ_l:

1. (between subspaces): a smaller $\langle z_l \rangle$, meaning that the observed shape is less likely to be generated from the l-th subspace, leads to a heavier regularization on b_l;
2. (within a subspace): a smaller shape variance λ_{lm} leads to a heavier penalization on the corresponding deformation component $\bar{b}_{lm}$ and vice versa.
3. (overall): the overall degree of regularization is controlled by the variance of shape noise σ_l^2 (5), which is determined by the percentage of shape variance that is preserve in each subspace. Reducing the percentage leads to larger regularization on b.

Identifying Outliers: the resistance to outliers is achieved by the observation noise model (3). Because observation noises are unpredictable, its variance ρ_n is unlikely to be learned as a prior. Therefore we set the initial ρ_n to be same for all landmark points, then change it according to the fitting error between the model prediction and the averaged candidate position

$$\rho_n = c \| \mathcal{T}(S_n(\tilde{b}), \tilde{\theta}) - \sum_k \pi_{nk} Q_{nk} \|, \tag{40}$$

where c is a constant. We update ρ_n whenever there exist new landmark detection results (see section 3.4), and used it to compute the weights (18) (19). These weights are in turn served for penalizing outliers (16) in both shape and pose estimation steps.

3.4 Initialization and Landmark Detection

The alignment program is initialized by a rotation-invariant face detector [21], which scans a target image and produces the initial guess of the pose $\theta^{(0)}$. The initial deformation parameter $b^{(0)}$ is simply set to be zero. The average training shape $\sum_l \pi_l \mu_l$ is transformed by $\theta^{(0)}$ and superimposed on the image.

We construct a simple gradient based landmark detector in a similar way to [4]. The neighboring image gradient centered on each landmark is normalized (L_1-distance equals to one) and modeled as a multivariate Gaussian whose mean and variance are learned from training data. For each landmark we search its nearby region and measure the probability of each pixel. That produces a response map, and the K largest local modes found in the map are used as the candidate positions. The response score of these candidates are further normalized and used as the image likelihood π_{nk}. This landmark detection procedure and the shape inference algorithm (29) $\sim$ (39) are performed recursively on a Gaussian image pyramid from the coarsest level to the finest.

4 Experiments

We first evaluated the proposed face alignment method on manually labeled frontal face images. The images are collected from two sources: CMU Multi-PIE database [22] and AR database [23]. All faces are frontal and non-occluded. Although captured in a controlled environment, this dataset covers large variations on subject identity, expression and illumination. We also extensively tested our program on a large collection of *real-world* images. Our goal in these experiments is to show that by putting deformable matching in our framework, we can improve the alignment accuracy, and more importantly, deal with a wide variety of situations in real-world face photographies.

Mis-Alignment Error When Images Are Clean. We compared our face alignment program with two previous techniques, namely, Active Appearance Model [6] and Bayesian Tangent Shape Model [19] on the labeled dataset. 800 randomly selected images are used for training, the rest 480 images are used for testing. For a fair comparison, same landmark detector is applied in all three methods; the initial shape parameters are set to zero; and the initial poses are generated by first computing a pose from ground-truth landmarks via procrustes analysis, then adding a small random permutation by translation ($-10\% \sim 10\%$ of the face size), rotation ($-15° \sim 15°$) and scaling ($0.9 \sim 1.1$). When computing the average mis-alignment error, we normalize the width of each testing face to be 120 pixels, keeping the aspect ratio unchanged and scaling the height accordingly. The overall mis-alignment error is 7.88 pixels for AAM, 5.90 pixels for BTSM

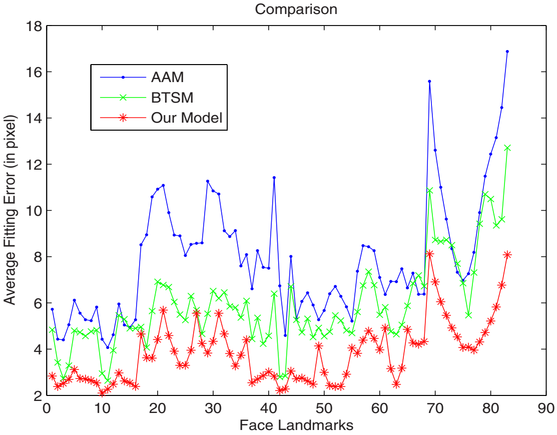

Fig. 2. We evaluate the performance of our model and compare it with AAM [6] and BTSM [19]. The graph plots the average fitting errors for every point: blue for AAM, green for BTSM, and red for our model. The landmarks represent left/right eyes (1 ∼ 8; 9 ∼ 16), left/right eyebrows (17 ∼ 26; 27 ∼ 36), nose (37 ∼ 48), mouth (49 ∼ 68) and silhouette (69 ∼ 83) respectively.

and 3.49 pixels for our model. Figure 2 compares the errors obtained by the three methods for each individual point. Our model consistently outperforms the other two techniques on all points.

Occlusions. In the presence of image clutters or occlusions, we expect that our model identifies the noisy points in the fitting process by measuring the discrepancy between model prediction and low-level image observation. Figure 3 shows a few partially occluded face images. Our model can deal with these cases, while traditional deformable models could easily fail. We plot the the estimated shape noise level (40) associated every point. When the points are occluded, such as those along eye contours (at the top row) or along mouth and jaw-line (at the bottom row), the corresponding observations are clearly considered as more ambiguous by the model. Also note that the predictions on the visible part are stable. As a result, small weights are assigned on the occluded points, and their positions are "hallucinated" by combining the information from reliable observations and shape priors according to the penalization rules (16)(31).

Facial Expression. The training images cover six types of expressions including neutral expression, smile, squint, surprise, disgust and scream. By use of a mixture shape prior, our deformable model is capable of capturing a larger range of expression variations than linear models. And the associated shape regularization rule ensures that the model can smooth out a noisy shape observation, and preserve dominant shape deformations. We select the number of mixture components as $L = 3$. Larger L does not improve the alignment performance but increases computational cost. The top row of Figure 4 shows the results on a

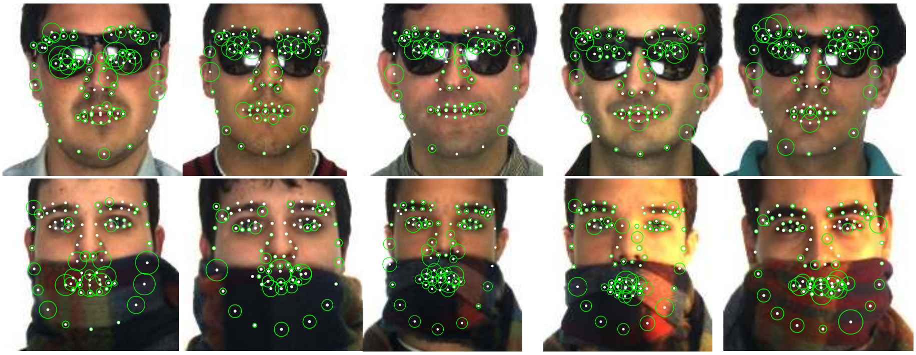

Fig. 3. Our model automatically distinguishes good landmarks from bad ones in the model matching process. White dots represent the alignment results; green circles represent the noise level of each landmark at the end of matching process. Larger observation variance implies a small contribution to the final estimates of shape and pose components.

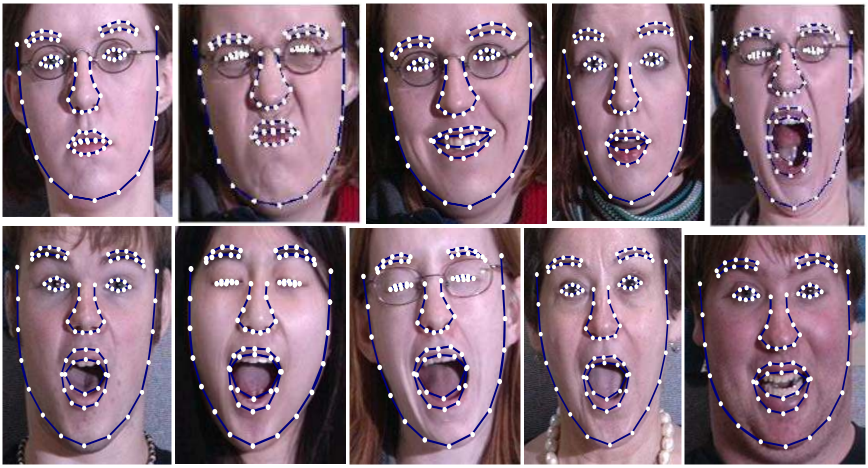

Fig. 4. Our model is capable of dealing with large expression changes. The top row shows the faces of a subject recorded in five different expressions: neutral, disgust, smile, surprise and scream; the bottom row highlights our alignment results on screaming faces.

testing subject in Multi-PIE dataset with different expressions, and the bottom row shows the results on a few images in our second dataset.

In-plane Rotation. The face detector is applied at 8 different orientations in increments of 35°. Our deformable model is capable of dealing with rotation in the range from $-25°$ to $25°$ degrees. Combining the model with the face detector our alignment program is capable of dealing with full $0° \sim 360°$ in-plane as shown in Figure 5.

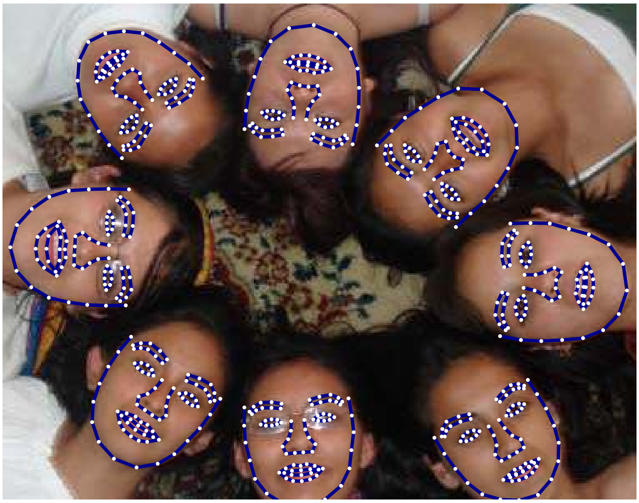

Fig. 5. Combining our deformable model with a rotation-invariant face detector our system is capable of dealing with the full 360 degree in-plane rotation

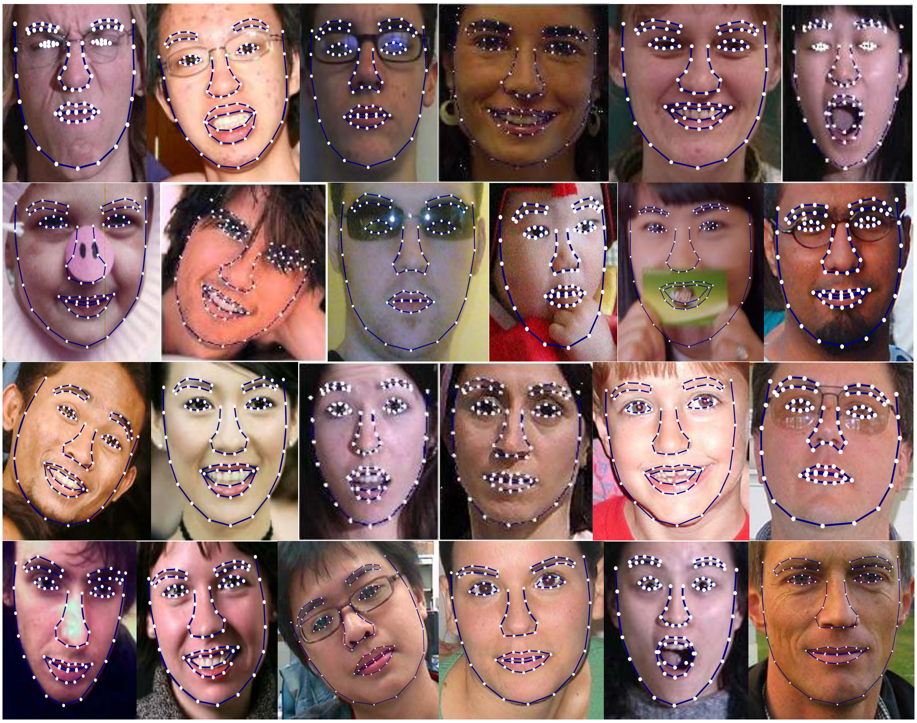

Fig. 6. The output of our face alignment program on some real-world photographs collected from Internet. More results are available on our website.

Real-World Photographs. Figure 6 shows the alignment results on a few real-world face image collected from Internet. More results are available on our website.

5 Conclusion and Discussion

In this paper, we have proposed a new approach for shape regularization by use of a multi-level generative model, and demonstrated its application in face alignment. We show our alignment system is capable of dealing with real-world images with a wide range of imaging conditions and appearance variations. Our model uses image gradients as the only low-level cues. By incorporating the model with other image cues or landmark detectors can potentially further improve the alignment accuracy.

References

1. Terzopoulos, D., Witkin, A., Kass, M.: Snakes: Active contour models. In: International Conference on Computer Vision, pp. 259–268 (1987)
2. Grenander, U., Chow, Y., Keenan, D.M.: Hands: a pattern theoretic study of biological shapes. Springer, New York (1991)
3. Yuille, A.L., Hallinan, P.W., Cohen, D.S.: Feature extraction from faces using deformable templates. International Journal of Computer Vision 8, 99–111 (1992)
4. Cootes, T.F., Taylor, C., Cooper, D., Graham, J.: Active shape models - their training and their applications. Computer Vision and Image Understanding (1995)
5. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. International Journal of Computer Vision 29, 5–28 (1998)
6. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. IEEE Trans. Pattern Anal. Mach. Intell. 23, 681–685 (2001)
7. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d-faces. In: ACM SIGGRAPH (1999)
8. Cootes, T., Taylor, C.: A mixture model for representing shape variation. IVC 17, 567–573 (1999)
9. Twining, C., Taylor, C.: Kernel principal component analysis and the construction of non-linear active shape models. In: British Machine Vision Conference (2001)
10. Zhou, Y., Zhang, W., Tang, X., Shum, H.: A bayesian mixture model for multiview face alignment. In: Proceedings of Computer Vision and Pattern Recognition (2005)
11. Gu, L., Kanande, T.: 3d alignment of face in a single image. In: Computer Vision and Pattern Recognition (2006)
12. Zhang, Z., Liu, Z., Adler, D., Cohen, M.F., Hanson, E., Shan, Y.: Robust and rapid generation of animated faces from video images - a model-based modeling approach. International Jornal of Computer Vision (2004)
13. Liang, L., Wen, F., Xu, Y.Q., tang, X., Shum, H.Y.: Accurate face alignment using shape constrained markov network. In: Computer Vision and Pattern Recognition (2006)
14. Amit, Y., Kong, A.: Graphical templates for model registration. IEEE Trans. Pattern Anal. Mach. Intell. 18, 225–236 (1996)

15. Coughlan, J., Yuille, A., English, C., Snow, D.: Efficient optimization of a deformable template using dynamic programming. In: Computer Vision and Pattern Recognition, pp. 747–752 (1998)
16. Coughlan, J., Ferreira, S.: Finding deformable shapes using loopy belief propagation. In: Tistarelli, M., Bigun, J., Jain, A.K. (eds.) ECCV 2002. LNCS, vol. 2359, pp. 453–468. Springer, Heidelberg (2002)
17. Ahonen, T., Hadid, A., Pietikainen, M.: Face recognition with local binary patterns. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3024, pp. 469–481. Springer, Heidelberg (2004)
18. Cristinacce, D., Cootes, T.: Boosted regression active shape models. In: British Machine Vision Conference, pp. 880–889 (2007)
19. Zhou, Y., Gu, L., Zhang, H.: Bayesian tangent shape model: Estimating shape and pose parameters via bayesian inference. In: Computer Vision and Pattern Recognition, pp. 109–116 (2003)
20. Tipping, M.E., Bishop, C.M.: Mixtures of probabilistic principal component analysers. Neural Computation 11, 443–482 (1999)
21. Rowley, H., Baluja, S., Kanade, T.: Rotation invariant neural network-based face detection. In: Computer Vision and Pattern Recognition, pp. 38–44 (1998)
22. Gross, R., Matthews, I., Cohn, J., Baker, S.: Guide to the cmu multi-pie face database. Technical report, CMU RI (2002)
23. Martinez, A., Benavente, R.: The ar face database. CVC Technical Report 24 (1998)