Joint Face Alignment with Non-Parametric Shape Models

Brandon M. Smith and Li Zhang

University of Wisconsin – Madison
http://www.cs.wisc.edu/~lizhang/projects/joint-align/

Abstract. We present a joint face alignment technique that takes a set of images as input and produces a set of shape- and appearance-consistent face alignments as output. Our method is an extension of the recent localization method of Belhumeur *et al.* [1], which combines the output of local detectors with a non-parametric set of face shape models. We are inspired by the recent joint alignment method of Zhao *et al.* [20], which employs a modified Active Appearance Model (AAM) approach to align a batch of images. We introduce an approach for simultaneously optimizing both a local appearance constraint, which couples the local estimates between multiple images, and a global shape constraint, which couples landmarks and images across the image set. In video sequences, our method greatly improves the temporal stability of landmark estimates without compromising accuracy relative to ground truth.

1 Introduction

Face alignment is an important problem in computer vision [1, 5, 9, 15, 16, 20] with many compelling applications, including performance-driven animation [10], automatic face replacement [2], and as a pre-process for face recognition and verification. As digital cameras become cheaper and more ubiquitous, and visual media-sharing websites like *Flickr*, *Picasa*, *Facebook*, and *YouTube* become more popular, it is increasingly convenient to perform face alignment on batches of images for other applications such as face image retrieval [17] and digital media management and exploration [11].

Intuitively, by using the additional information provided by multiple images of the same face, we can better handle a range of challenging conditions, such as partial occlusion, pose and illumination changes, image blur, and image noise. This intuition has been used in the face tracking literature, for example, to impose local appearance constraints across multiple video frames [14, 21].

Inspired by this intuition, we propose a joint face alignment technique that takes a set of images as input and produces a set of shape- and appearance-consistent face alignments as output. By *appearance-consistent* we mean that the local appearance at each landmark estimate is more similar across input images, and by *shape-consistent* we mean the spatial arrangements of landmarks on the input faces are more consistent. We might expect that the final set of alignments will drift somewhat from ground truth in order to achieve an internally consistent solution among input images. However, we show experimentally that our approach does not sacrifice alignment accuracy to achieve consistency.

2 Background

Non-rigid face alignment algorithms can be divided into two categories: *holistic* and *local* methods. Our approach is most directly inspired by a recent holistic method [20] that operates on a batch of input images simultaneously, and a recent local method [1] that combines the output of local detectors with a non-parametric set of shape models. In this section, we highlight some key differences between holistic and local methods, and give a brief overview of [20] and [1].

2.1 Holistic methods

In the domain of non-rigid alignment, Active Appearance Models (AAMs) [5] are among the most popular, with many recent works [3, 6, 7, 15, 20] relying on the AAM framework. Broadly, AAMs model both the overall appearance $\boldsymbol{\alpha}$ and shape X of the face linearly:

$$X = X_0 + \sum_{m=1}^{M} p_m X_m \qquad \boldsymbol{\alpha} = \boldsymbol{\alpha}_0 + \sum_{l=1}^{L} \lambda_l \boldsymbol{\alpha}_l, \tag{1}$$

where p_m and λ_l are the m-th shape and l-th appearance parameters (or loadings), respectively; X_m and $\boldsymbol{\alpha}_l$ capture the top M shape and L appearance modes, respectively; and X_0 and $\boldsymbol{\alpha}_0$ are the average shape and appearance vectors, respectively. The goal is to minimize the difference between the target image and the model:

$$\min_{p,\lambda} \left\| \boldsymbol{\alpha}_0 + \sum_{l=1}^{L} \lambda_l \boldsymbol{\alpha}_l - W(I; \mathbf{p}) \right\|_2^2, \tag{2}$$

where W warps the image I in a piecewise linear fashion using the shape parameters $\mathbf{p}$.

One of the major challenges in applying AAMs is that it is difficult to adequately synthesize the appearance of a new face using a linear model. As demonstrated experimentally by Gross *et al.* [6], it is much more difficult to build a generic appearance model than a person-specific one. This is known as the generalization problem.

To overcome this problem, Zhao *et al.* [20] introduced a *joint* alignment technique in which a batch of person-specific images are simultaneously aligned in a modified AAM setting. Intuitively, their approach aims to simultaneously (1) identify the person-specific appearance space of the input images, which are assumed to be linear and low-dimensional, and (2) align the person-specific appearance space with the generic appearance space, which are assumed to be proximate rather than distant. This joint approach was shown to produces excellent results on a wide variety "real-world" images from the internet. However, the approach breaks down under several common conditions, including significant occlusion or shadow, image degradation, and outliers. Like Zhao *et al.*, we operate on a set of input images jointly, but our approach is more robust to these conditions.

2.2 Overview of Zhao *et al.* [20]

Zhao *et al.* [20] *jointly* align a batch of images by extending the AAM framework:

$$\min_{p_j,\lambda_j} \rho \left\{ \sum_{j}^{J} \left\| \boldsymbol{\alpha}_0 + \sum_{l=1}^{L} \lambda_l \boldsymbol{\alpha}_l - W(I_j;\mathbf{p}) \right\|_2^2 \right\} + \text{rank}\left(\sum_{j}^{J} W(I_j;\mathbf{p})\mathbf{e_j}^{\mathsf{T}} \right), \quad (3)$$

where the sum is taken over J images, $\mathbf{e}_j$ is an indicator vector (*i.e.*, the j-th element is one, all others are zero), ρ is a weight parameter that balances the first term with the second, and all other parts are defined as in Eq. (2). The rank term encourages the appearance of the input images to be linearly correlated; intuitively, the rank term favors a solution in which the input images are well-aligned with one another (not just the global model). Unfortunately, the rank term is non-convex and non-continuous and minimizing it is an NP-hard problem. To make the problem more tractable, they borrow a trick from compressive sensing. That is, they replace the rank term with its tightest convex relaxation – the nuclear norm $||X||_*$, which is defined as the sum of singular values $||X||_* = \sum_k \sigma_x(X)$. See [20] for more details.

Like [20], we incorporate a rank term in our optimization. However, because we couple the local appearance of faces with global shape models, our rank term does not operate directly on the holistic appearance of faces. Instead, it operates on the simplified shape representation of faces, and encourages landmark estimates that are spatially consistent across images. We encourage the local appearance at each landmark estimate to be consistent across multiple images via a separate term, which we discuss in Section 3.

2.3 Local methods

Constrained Local Models (CLMs) [9, 14, 16, 19] overcome many of the problems inherent in holistic methods. For examples, CLMs have inherent computational advantages (*e.g.*, opportunities for parallelization) [14], and reduced modeling complexity and sensitivity to illumination changes [16, 19]. CLMs also generalize well to new face images, and can be made robust against other confounding effects such as reflectance, image blur, and occlusion [9].

CLMs model the appearance of the face locally via an ensemble of region experts, or local detectors. A variety of local detectors have been proposed for this purpose, including those based on discriminative classifiers that operate on local image patches [19], or feature descriptors such as SIFT [13]. These local detectors generate a likelihood map around the current estimate of each landmark location. The likelihood maps are then combined with an overall face shape model to jointly recover the location of landmarks. Like AAMs, CLMs typically model non-rigid shape variation linearly, as in Eq. (1).

Recently, Belhumeur *et al.* [1] proposed a different approach that avoids a *parametric* linear shape model, and instead combines the outputs of local detectors with a *non-parametric* set of global shape models. Belhumeur *et al.* showed excellent landmark localization results on *single* images of faces under challenging conditions, such as significant occlusions, shadows, and a wide range of pose and expression variation. According to their experiments on "real-world" images from the internet, their algorithm produced slightly more accurate results, on average, compared to humans assigned to the same task.

Despite its impressive accuracy, we found that Belhumeur *et al.*'s approach produces temporally inconsistent results in video. Our method makes use of the same non-parametric shape model approach proposed by Belhumeur *et al.*, but we further constrain the local appearance and face shape in a joint alignment setting. As our experiments show, we produce more consistent results, especially in video. At the same time, we do not sacrifice localization accuracy to achieve this consistency.

2.4 Overview of Belhumeur *et al.* [1]

Formally, [1] aims to solve the following problem:

$$X^* = \operatorname*{argmax}_X P(X|D), \tag{4}$$

where $X = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N\}$ gives the locations of N face landmarks (*i.e.*, the eye corners, nose tip, *etc.*), and $D = \{\mathbf{d}^1, \mathbf{d}^2, \ldots, \mathbf{d}^N\}$ are local detector responses. In words, the goal is to find the landmark locations X that maximize the probability of X given the measurements from the set of local detectors. In general, Eq. (4) is non-convex. Therefore, Belhumeur *et al.* employ a set of approximations to make the problem more tractable. The first approximation is to assume each X is generated by one of M shape exemplars, X_k, transformed by some similarity transformation t; they call $X_{k,t}$ a global model. This allows $P(X|D)$ to be expanded as

$$P(X|D) = \sum_{k=1}^{M} \int_{t \in T} P(X|X_{k,t}, D) P(X_{k,t}|D) dt, \tag{5}$$

where the collections of M global models $X_{k,t}$ have been introduced and then marginalized out. By conditioning on the global model $X_{k,t}$, the locations of the parts $\mathbf{x}^i$ are assumed to be conditionally independent of one another, *i.e.*, $P(X|X_{k,t}, D) = \prod_{i=1}^{N} P(\mathbf{x}^i|\mathbf{x}^i_{k,t}, \mathbf{d}^i)$. Bayes' rule is then applied to Eq. (5), and all probability terms that only depend on D and $X_{k,t}$ are reduced to a constant C. This yields the following optimization problem:

$$X^* = \operatorname*{argmax}_X \sum_{k=1}^{M} \int_{t \in T} C \prod_{i=1}^{N} P(\Delta \mathbf{x}^i_{k,t}) P(\mathbf{x}^i|\mathbf{d}^i) dt, \tag{6}$$

where $P(\Delta\mathbf{x}^i_{k,t})$ is a 2D Gaussian distribution that models how well the landmark location in the global model $\mathbf{x}^i_{k,t}$ fit the true location $\mathbf{x}^i$, *i.e.*, $\Delta\mathbf{x}^i_{k,t} = \mathbf{x}^i_{k,t} - \mathbf{x}^i$. $P(\mathbf{x}^i|\mathbf{d}^i)$ takes the form of a local response map for landmark i.

A RANSAC-like approach is used to generate a large number of global models $X_{k,t}$, which are evaluated using $P(X_{k,t}|D)$. The set $\mathcal{M}$ of top global models M^* for which $P(X_{k,t}|D)$ is greatest is used to approximate Eq. (6) as

$$X^* = \operatorname*{argmax}_X \sum_{k,t\in\mathcal{M}} \prod_{i=1}^{N} P(\Delta\mathbf{x}^i_{k,t})P(\mathbf{x}^i|\mathbf{d}^i). \tag{7}$$

X^* is found by choosing the image location where the following sum is maximized

$$\mathbf{x}^{i*} = \operatorname*{argmax}_{\mathbf{x}^i} \sum_{k,t\in\mathcal{M}} P(\Delta\mathbf{x}^i_{k,t})P(\mathbf{x}^i|\mathbf{d}^i), \tag{8}$$

which is equivalent to solving for $\mathbf{x}^{i*}$ by setting all $P(\Delta\mathbf{x}^{i'}_{k,t})$ and $P(\mathbf{x}^{i'}|\mathbf{d}^{i'})$ to a constant in Eq. (6) for all $i' \neq i$. For more details, please see [1].

3 Our Approach

This section provides details of how we couple the shape and appearance information from multiple related face images.

3.1 Single Image Alignment

Define $w^i(\mathbf{x}^i)$ as the accumulated response map for landmark i in Eq. (8),

$$w^i(\mathbf{x}^i) = \sum_{k,t\in\mathcal{M}} P(\Delta\mathbf{x}^i_{k,t})P(\mathbf{x}^i|\mathbf{d}^i). \tag{9}$$

As in [1], the value of $w^i(\mathbf{x}^i)$ is computed by multiplying the local detector output at $\mathbf{x}^i$ by a 2D Gaussian function centered at each $\mathbf{x}^i_{k,t}$, and summing the resulting products. The Gaussian function is parameterized by a 2×2 covariance matrix that captures the spatial uncertainty of the global shape model fitting the true landmark locations; the covariances are computed as described in [1]. Note that Eq. (9) couples the local appearance information with global shape information.

To estimate the location of each landmark, Belhumeur *et al.* [1] choose the location $\mathbf{x}^i_{k,t}$ that maximizes $w^i(\mathbf{x}^i)$. Because the global shape information is already incorporated via $P(\Delta\mathbf{x}^i_{k,t})$, the final selection of the best $\mathbf{x}^i$ can be done independent of all other landmark locations $\mathbf{x}^{i'}$ for $i' \neq i$. Therefore, assuming the global shape models $X_{k,t}$ are fixed after the generate-and-test procedure, maximizing the product of terms in Eq. (7) is equivalent to maximizing the sum of terms:

$$X^* = \operatorname*{argmax}_X \sum_{i=1}^{N} w^i(\mathbf{x}^i). \tag{10}$$

3.2 Multiple Image Alignment

In a multiple image setting, we can trivially modify Eq. (10) to incorporate all images:

$$\left\{X_j^*\right\}_{j=1,\ldots,J} = \underset{X_1,X_2,\ldots,X_J}{\operatorname{argmax}} \sum_{j=1}^{J}\sum_{i=1}^{N} w_j^i(\mathbf{x}_j^i), \tag{11}$$

where $w_j^i(\mathbf{x}_j^i)$ corresponds to landmark i in image j. Note that the selection of each set of global models $\mathcal{M}_j$ is still independent for each image j, and finding the best $\left\{X_j^*\right\}_{j=1,\ldots,J}$ in Eq. (11) is equivalent to solving Eq. (10) for each image separately.

Enforcing Joint Appearance Consistency Under this multiple image setting, we couple the appearance information of the input images by modifying Eq. (11) as follows:

$$\left\{X_j^*\right\}_{j=1,\ldots,J} = \underset{X_1,X_2,\ldots,X_J}{\operatorname{argmax}} \frac{1}{J}\sum_{j=1}^{J}\frac{1}{N}\sum_{i=1}^{N} w_j^i(\mathbf{x}_j^i)\cdot \varPhi\left(\mathbf{x}_j^i,\{\mathbf{x}_j^{i'}\}_{i'\neq i}\right), \tag{12}$$

where

$$\varPhi\left(\mathbf{x}_j^i,\{\mathbf{x}_j^{i'}\}_{i'\neq i}\right) = \frac{1}{J-1}\sum_{j'\in\mathcal{S}_j} \lambda_{j'}^i \exp\left\{-\gamma\|\mathbf{s}_j^i(\mathbf{x}_j^i)-\mathbf{s}_{j'}^i(\mathbf{x}_{j'}^i)\|^2\right\}, \tag{13}$$

$\mathcal{S}_j$ is the set of input images associated with, but not including, image j, and $\mathbf{s}_j^i(\mathbf{x}_j^i)$ is a local feature descriptor centered at $\mathbf{x}_j^i$. $\lambda_{j'}^i$ is a weight reflecting the confidence that $\mathbf{s}_{j'}^i(\mathbf{x}_{j'}^i)$ describes the correct landmark location. In practice, we use $\lambda_{j'}^i = w_{j'}^i(\mathbf{x}_{j'}^i)$. Intuitively, Eq. (12) is maximized when two conditions are simultaneously optimized: (1) the face shape coincides with positive measurements according to the local detectors *and* (2) the local descriptors are consistent across input images.

Enforcing Joint Appearance and Shape Consistency The optimization of Eq. (12) will yield landmark estimates that are *locally* more consistent. However, the estimated face shapes may not be consistent across the input images. Shape inconsistency can be due to several factors, including local appearance ambiguities and noise, and the randomness inherent in the search for global shape models. This inconsistency is especially noticeable in video sequences, where the landmark estimates appear to "swim" around their true location, or jitter along image contours such as the chin line. We could trivially remove these artifacts by forcing all of the face shape estimates to be the same, or by applying a low-pass filter to the landmark trajectories in a video. Unfortunately, this yields a poor solution that cannot adequately handle non-rigid deformations or fast non-rigid motion. Instead, we aim to find a set of face shapes that (1) provide locally consistent landmark estimates, and (2) are linearly correlated, but still flexible enough to handle a broad range of non-rigid deformations.

In order to incorporate a shape consistency constraint, we first change Eq. (12) to a *minimization* problem:

$$\{X_j^*\}_{j=1,\dots,J} = \underset{X_1,X_2,\dots,X_J}{\operatorname{argmin}} -\frac{1}{\mathcal{N}} \sum_i \sum_j \sum_{j' \in \mathcal{S}_j} \widetilde{w}_{jj'}^i(\mathbf{x}_j^i, \mathbf{x}_{j'}^i), \tag{14}$$

where

$$\widetilde{w}_{jj'}^i(\mathbf{x}_j^i, \mathbf{x}_{j'}^i) = w_j^i(\mathbf{x}_j^i) \cdot w_{j'}^i(\mathbf{x}_{j'}^i) \cdot \exp\left\{-\gamma \|\mathbf{s}_j^i(\mathbf{x}_j^i) - \mathbf{s}_{j'}^i(\mathbf{x}_{j'}^i)\|^2\right\} \tag{15}$$

and

$$\frac{1}{\mathcal{N}} = \frac{1}{N} \cdot \frac{1}{J} \cdot \frac{1}{J-1}. \tag{16}$$

We then add a rank term that encourages the recovered shapes to be linearly correlated:

$$\{X_j^*\}_{j=1,\dots,J} = \underset{X_1,X_2,\dots,X_J}{\operatorname{argmin}} \rho \left\{ -\frac{1}{\mathcal{N}} \sum_i \sum_j \sum_{j' \in \mathcal{S}_j} \widetilde{w}_{jj'}^i(\mathbf{x}_j^i, \mathbf{x}_{j'}^i) \right\} + \operatorname{rank}(\mathbf{X}), \tag{17}$$

where $\rho > 0$ is a scalar weight that balances the two terms, and $\mathbf{X}$ is formed by concatinating the face shape vectors into a $2N \times J$ matrix: $\mathbf{X} = [X_1(:), X_2(:), \cdots, X_J(:)]$. Eq. (17) is similar in form to Eq. (3). However, instead of encouraging the holistic face appearance to be linearly correlated across the input images, our rank term encourages the estimated face shapes to be linearly correlated.

Reformulation As Zhao *et al.* [20] point out, the difficulty in incorporating a rank term is that it is non-convex and non-continuous, which makes minimizing Eq. (17) NP-hard. Instead, we follow the same approach in [20] and replace the rank($\mathbf{X}$) term with its tightest convex relaxation – the nuclear norm $||\mathbf{X}||_*$.

At this point, it is convenient to introduce some additional notation. Let $h_{ij} = 1, 2, \dots, H_{ij}$ be an index that selects one of H_{ij} hypotheses for landmark i in image j, (*i.e.*, $\mathbf{x}_j^i[h_{ij} = 1]$ selects the first location hypothesis, $\mathbf{x}_j^i[h_{ij} = 2]$ selects the second, and so on). $\widetilde{w}_{jj'}^i(\mathbf{x}_j^i, \mathbf{x}_{j'}^i)$ in Eq. (17) can then be written $\widetilde{w}_{jj'}^i(\mathbf{x}_j^i[h_{ij}], \mathbf{x}_{j'}^i[h_{ij'}])$, or more simply $\widetilde{w}_{jj'}^i(h_{ij}, h_{ij'})$. Let $\alpha_j^i[h_{ij}] \in \{0, 1\}$ be an indicator variable that can either be on (1) or off (0) for each possible h_{ij}. Eq. (17) can then be reformulated into the following energy minimization problem.

$$\min_{\boldsymbol{\alpha}} \ \rho \cdot \Psi(\boldsymbol{\alpha}) + \|[\mathbf{X}, \mathbf{Z}]\|_* \tag{18}$$

$$\text{s.t.} \ \ \alpha_j^i[h_{ij}] \in \{0, 1\} \tag{19}$$

$$\sum_{h_{ij}} \alpha_j^i[h_{ij}] = 1 \tag{20}$$

$$\mathcal{A}(\boldsymbol{\alpha}) - \mathbf{X} = 0 \tag{21}$$

$$\mathbf{Z}_0 - \mathbf{Z} = 0, \tag{22}$$

where

$$\Psi(\boldsymbol{\alpha}) = -\frac{1}{\mathcal{N}} \sum_i \sum_j \sum_{j' \in \mathcal{S}_j} \sum_{h_{ij}} \sum_{h_{ij'}} \alpha_j^i[h_{ij}] \cdot \alpha_{j'}^i[h_{ij'}] \cdot \widetilde{w}_{jj'}^i(h_{ij}, h_{ij'}) \tag{23}$$

and $\boldsymbol{\alpha}$ contains all indicator variables $\alpha_j^i[h_{ij}]$ for $j = 1, \ldots, J$, $i = 1, \ldots, N$, and $h_{ij} = 1, \ldots, H_{ij}$; $\mathbf{Z}_0$ is formed by concatenating the global shape model vectors for all images into a $2N \times M$ matrix: $\mathbf{Z}_0 = [Z_1, Z_2, \ldots, Z_M]$, and $\mathcal{A}(\boldsymbol{\alpha})$ is a linear operator on the indicator variables $\boldsymbol{\alpha}$, which produces a $2N \times J$ matrix of landmark locations. The intuition behind including $\mathbf{Z}$, which is comprised of true shape vectors, is that we want to ensure $\mathbf{X}$ does not deviate significantly from the space of plausible face shapes.

To solve the optimization problem in Eq. (18), we use a strategy similar to [20]. That is, we adapt the augmented Lagrangian method to solve Eq. (18):

$$\begin{aligned} L(\boldsymbol{\alpha}, \mathbf{X}, \mathbf{Z}, Y_{\mathbf{X}}, Y_{\mathbf{Z}}) = \tau \left(\rho \cdot \varPsi(\boldsymbol{\alpha}) + \|[\mathbf{X}, \mathbf{Z}]\|_*\right) + \tfrac{1}{2} \left\|[\mathbf{X}, \mathbf{Z}]\right\|_F^2 \\ + \langle Y_{\mathbf{X}}, \mathcal{A}(\boldsymbol{\alpha}) - \mathbf{X}\rangle + \langle Y_{\mathbf{Z}}, \mathbf{Z}_0 - \mathbf{Z}\rangle, \end{aligned} \qquad (24)$$

where $\|U\|_F = \langle U, U\rangle$ is the Frobenius norm, $\langle U, V\rangle = \text{trace}(U^\mathsf{T} V)$, and $Y_{\mathbf{X}}$ and $Y_{\mathbf{Z}}$ are Lagrangian multipliers. Here we have also made use of a result from [4], namely that solving $min\ \tau\|U\|_* + \frac{1}{2}\|U\|_F^2$ subject to linear constraints converges to the same minimum as solving $min\ \|U\|_*$ for large enough τ. We iterate among four update steps to solve Eq. (24):

Step 1 $$\boldsymbol{\alpha}^{k+1} = \operatorname*{argmin}_{\boldsymbol{\alpha}} L(\boldsymbol{\alpha}, \mathbf{X}, \mathbf{Z}, Y_{\mathbf{X}}, Y_{\mathbf{Z}}) \qquad (25)$$

Step 2 $$\begin{aligned} \left[\mathbf{X}^{k+1}, \mathbf{Z}^{k+1}\right] &= \operatorname*{argmin}_{\mathbf{X},\mathbf{Z}} \tau \left\|[\mathbf{X}, \mathbf{Z}]\right\|_* + \left\|[\mathbf{X}, \mathbf{Z}] - [Y_{\mathbf{X}}^k, Y_{\mathbf{Z}}^k]\right\|_F^2 \\ &= \text{shrink}([Y_{\mathbf{X}}^k, Y_{\mathbf{Z}}^k], \tau) \end{aligned} \qquad (26)$$

Step 3 $$\begin{bmatrix} Y_{\mathbf{X}}^{k+1} \\ Y_{\mathbf{Z}}^{k+1} \end{bmatrix} = \begin{bmatrix} Y_{\mathbf{X}}^{k} \\ Y_{\mathbf{Z}}^{k} \end{bmatrix} + \delta_k \begin{bmatrix} \mathcal{A}(\boldsymbol{\alpha}^k) - \mathbf{X}^k \\ \mathbf{Z}_0 - \mathbf{Z}^k \end{bmatrix} \qquad (27)$$

Step 4 $$\delta_{k+1} = \eta \delta_k. \qquad (28)$$

The shrink operator is a nonlinear function which applies a soft-thresholding rule at level τ to the singular values of the input matrix. See [4] for more details. Steps $2 - 4$ can be easily implemented in MATLAB using a few lines of code. Step 1 is more complicated and so we give it some additional attention.

Solving for $\boldsymbol{\alpha}$ To find the $\boldsymbol{\alpha}^{k+1}$ that minimizes Eq. (25), we need consider only those terms in Eq. (24) that involve $\boldsymbol{\alpha}$:

$$\boldsymbol{\alpha}^{k+1} = \operatorname*{argmin}_{\boldsymbol{\alpha}} \left\{\tau\rho\ \varPsi(\boldsymbol{\alpha}) + \langle Y_{\mathbf{X}}, \mathcal{A}(\boldsymbol{\alpha})\rangle\right\}. \qquad (29)$$

We solve Eq. (29) by breaking it into independent subproblems, one for each landmark i. Each subproblem is then solved using an iterative greedy approach:

```
Loop until convergence
    For landmark $i = 1, 2, \ldots, N$
        For image $j = 1, 2, \ldots, J$
            1. Assume all $\alpha_{j'}^i[h_{ij'}]$ for $j' \in \mathcal{S}_j$ are known and assign a
               single hypothesis $h_{ij'}$ to each $j' \in \mathcal{S}_j$.
```

2. Extract the portion of $\langle Y_{\mathbf{X}}, \mathcal{A}(\boldsymbol{\alpha})\rangle$ involving landmark i and image j; denote this portion $\langle [Y_{\mathbf{X}}]_{ij}, [\mathcal{A}(\boldsymbol{\alpha})]_{ij}\rangle$.
3. Evaluate $\langle [Y_{\mathbf{X}}]_{ij}, [\mathcal{A}(\boldsymbol{\alpha})]_{ij}\rangle - \tau\rho\frac{1}{N}\sum_{j'\neq j}\widetilde{w}^{i}_{jj'}(h_{ij}, h_{ij'})$ (see Eq. (29)) for each $h_{ij} = 1, \ldots, H_{ij}$.
4. Choose the hypothesis h_{ij} that minimizes the sum above, and set the corresponding indicator variable $\alpha^{i}_{j}[h_{ij}]$ to one and all others to zero.

We remark that Step 1 above greatly simplifies Eq. (29), because we need only consider the small subset of terms in Eq. (23) for which $\alpha^{i}_{j'}[h_{ij'}] = 1$. Furthermore, evaluating $\langle [Y_{\mathbf{X}}]_{ij}, [\mathcal{A}(\boldsymbol{\alpha})]_{ij}\rangle$ involves relatively few operations.

To initialize $\boldsymbol{\alpha}$, we solve Eq. (6) for each landmark and image independently, and then set the corresponding indicator variables in $\boldsymbol{\alpha}$ to one (all others are zero). In our experiments, this initialization is close to the final solution for the majority of landmarks. As a result, the algorithm above typically converges quickly in $2-3$ iterations.

3.3 Implementation Details

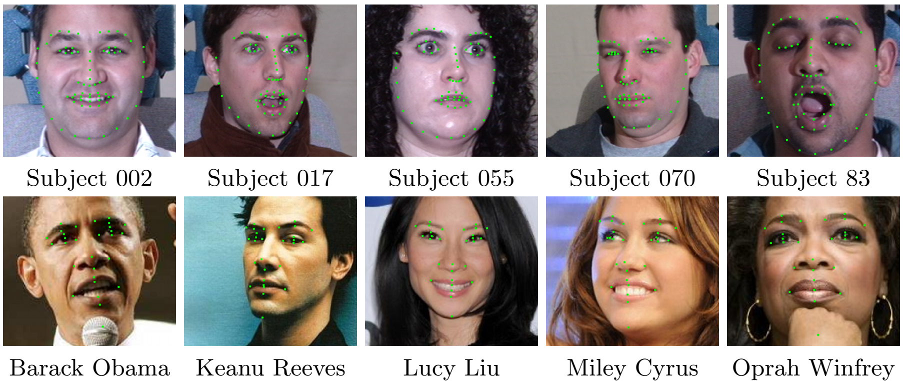
To obtain $w^{i}_{j}(\mathbf{x}^{i}_{j})$ in Eq. (9), we use the consensus of exemplars approach described in [1], with all parameters set according to [1] unless noted. Like [1], our landmark detectors are simple support vector machine (SVM) classifiers with grayscale SIFT [13] descriptors. We note that, for a typical 20-image input set, the detectors account for approximately 86% of our total computation. SIFT descriptors are also used to enforce joint appearance consistency in Eq. (15).

For efficiency reasons, we do not consider all possible hypotheses $\alpha^{i}_{j}[h_{ij}]$ when solving for $\boldsymbol{\alpha}$. Instead, we consider only the hypotheses that give the largest H_{ij} values according to $w^{i}_{j}(h_{ij})$; we typically choose $H_{ij} \approx 200$ in our experiments. This is acceptable because most hypotheses within each local search window correspond to very low values for $w^{i}_{j}(h_{ij})$.

Parameter τ is set such that the first $\sim 7-12$ singular values are preserved by the shrink operator in Eq. (26). The step size δ_0 is set to 1 in our experiments, with $\eta = 0.75$. ρ should depend on two things: (1) confidence in the joint appearance constraint (*i.e.*, ρ should increase with better image quality, fewer occlusions, *etc.*), and (2) confidence in the shape consistency constraint (*i.e.*, ρ should decrease as the local appearance becomes more corrupted). Roughly speaking, $\rho = C \cdot \left\|[\mathbf{X}^0, \mathbf{Z}^0]\right\|_* / \Psi(\boldsymbol{\alpha}^0)$, for an empirically determined constant C and initial states of $\mathbf{X}$, $\mathbf{Z}$, and $\boldsymbol{\alpha}$. We set $C \approx 10$ in our experiments.

The sum $\sum_{j'\in\mathcal{S}_j}$ in Eq. (23) is straightforward for static image sets: for a given image j, sum over all other images in the input set. For videos, it changes slightly: $\mathcal{S}_j$ becomes the *temporal* neighborhood around frame j; $\mathcal{S}_j = \{j-5, \ldots, j-1, j+1, \ldots, j+5\}$ in our experiments.

Our algorithm requires an initial estimate of the landmark locations in order to place the local search windows. For this purpose, we fit a canonical face shape to the bounding box generated by the OpenCV Viola-Jones face detector [18].



Subject 002 Subject 017 Subject 055 Subject 070 Subject 83

Barack Obama Keanu Reeves Lucy Liu Miley Cyrus Oprah Winfrey

Fig. 1. Selected images from two face datasets used for evaluation: Multi-PIE [8] on top and PubFig [12] on the bottom, with *ground truth* landmarks shown in green.

4 Results and Discussion

We now discuss our results on a variety of challenging image sets and video sequences, and compare them with state-of-the-art alignment methods, namely [1] and [16]. In general, we find that our method produces landmark localizations that are more consistent at both the local level (*i.e.*, the local appearance at each landmark estimate is more similar across input images) and the global level (*i.e.*, the arrangement of landmarks is more consistent across input faces). At the same time, we do not sacrifice alignment accuracy relative to ground truth.

4.1 Experimental Datasets

We evaluate our method on two static image datasets and one video dataset. As training data, we use the LFPW dataset [1], which includes many "real-world" images from the internet along with hand-labeled landmarks. For testing purposes, we use randomly selected images from the PubFig dataset [12]. In PubFig, unlike LFPW, each subject is depicted across many images. In order to perform a quantitative analysis, we manually labeled the subset of PubFig image according to the LFPW arrangement. We also perform an evaluation on the Multi-PIE dataset [8], some images of which are shown in Figure 1. Our video dataset is composed of several video sequences downloaded from YouTube.

4.2 Experiments on Static Image

Ten sets of 20 images were used to evaluate our algorithm on static image sets—five from Multi-PIE and five from PubFig. Each set includes images of the same subject under multiple expressions, illuminations, and poses; for reference, the subjects are shown in Figure 1. Figure 2 shows a comparison between our jointly

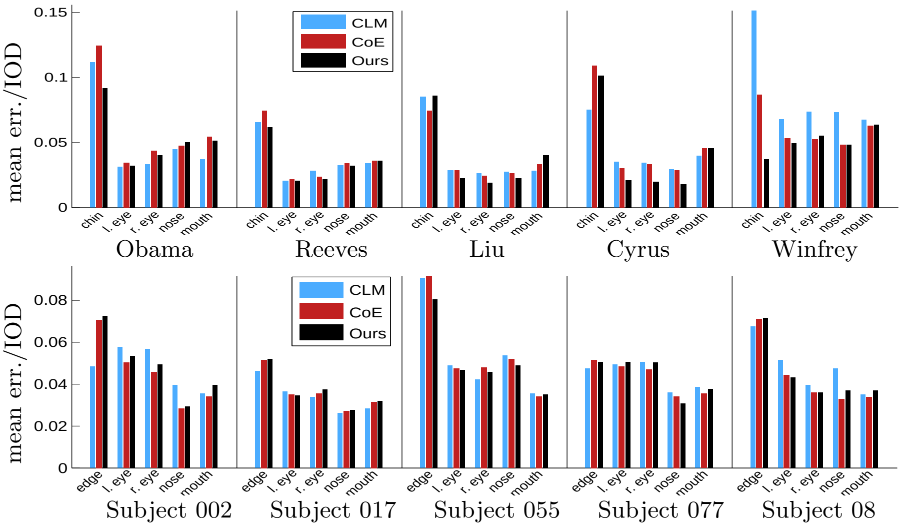


Fig. 2. This figure shows a comparison between our jointly aligned results, and the results generated by our implementations of Saragih *et al.*'s [16] algorithm (CLM), and Belhumeur *et al.*'s [1] algorithm (CoE) on five sets of images from the PubFig [12] dataset (*top*) and five sets of images from the Multi-PIE dataset [8] (*bottom*). Please see Section 4.2 for an explanation of these results.

aligned results, and the results generated by our implementations of Saragih *et al.*'s algorithm [16] and Belhumeur *et al.*'s Consensus of Exemplars (CoE) algorithm [1]. For a fair comparison, we used the same landmark detectors (described in Section 3.3) across all algorithms. The errors shown are the point-to-point distance from ground truth, normalized by the inter-ocular distance (IOD). As a reference, 0.05 is 2.75 pixels for a face with an IOD of 55 pixels.

Our method favors landmark localizations that are internally consistent among images in each set with respect to both the local appearance and the global arrangement of landmarks. To best satisfy these internal constraints, we might expect to lose some accuracy relative to ground truth. However, we maintain good localization accuracy similar to other state-of-the-art methods [2, 8] on challenging static images. Furthermore, as Figure 3 illustrates, our results are nearly identical even when an image is jointly aligned with entirely different sets of input images.

4.3 Experiments on Video Sequences

The improvements demonstrated by our joint alignment approach are most dramatically seen in video sequences. Figure 4 shows a comparison between our results, and the results generated by our implementations of Saragih *et al.*'s [16] algorithm (CLM), and Belhumeur *et al.*'s [1] algorithm (CoE), both applied to

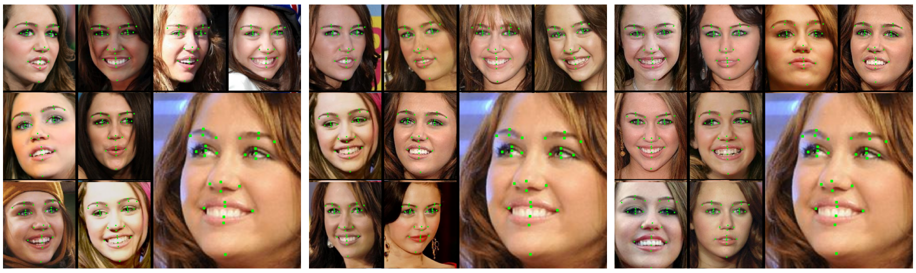


Fig. 3. Test results using our method on PubFig images. Our alignment results are nearly identical, even when the input image (shown larger than the rest) is jointly aligned with entirely different sets of input images.

each frame of the videos independently. In all cases, the same detectors were used. The top row shows the mean acceleration magnitude at each frame, computed as $\frac{1}{N}\sum_i \|\mathbf{x}_j^i - \frac{1}{2}(\mathbf{x}_{j-1}^i + \mathbf{x}_{j+1}^i)\|_2$, where $\mathbf{x}_j^i$ is the 2D location of landmark i in frame j. The bottom row shows the mean point-to-point error as a fraction of the inter-ocular distance (IOD). The errors were computed relative to ground truth locations, which were labeled manually at each 10th frame. Five challenging internet videos were used for this experiment.

We notice that the CLM and CoE methods generate landmark estimates that appear to "swim" or jump around their true location. Our approach generates landmark estimates that are noticeably more stable over time. Meanwhile, we do not sacrifice landmark accuracy relative to ground truth (*i.e.*, we do not over-smooth the trajectories); in fact, we often achieve slightly more accurate localizations compared to [1] and [16]. Figure 5 shows several video frames with our landmark estimates overlaid in green. For a clearer demonstration, please see our supplemental video.

5 Conclusions

We have described a new joint alignment approach that takes a batch of images as input and produces a set of alignments that are more shape- and appearance-consistent as output. In addition to introducing two constraints that favor shape and appearance consistency, we have outlined an approach to optimize both of them together in a joint alignment framework. Our method shows improvement most dramatically on video sequences. Despite producing more temporally consistent results, we do not sacrifice alignment accuracy relative to ground truth.

6 Acknowledgements

This work is supported in part by NSF IIS-0845916, NSF IIS-0916441, a Sloan Research Fellowship, and a Packard Fellowship for Science and Engineering. Brandon Smith is also supported by an NSF graduate fellowship.

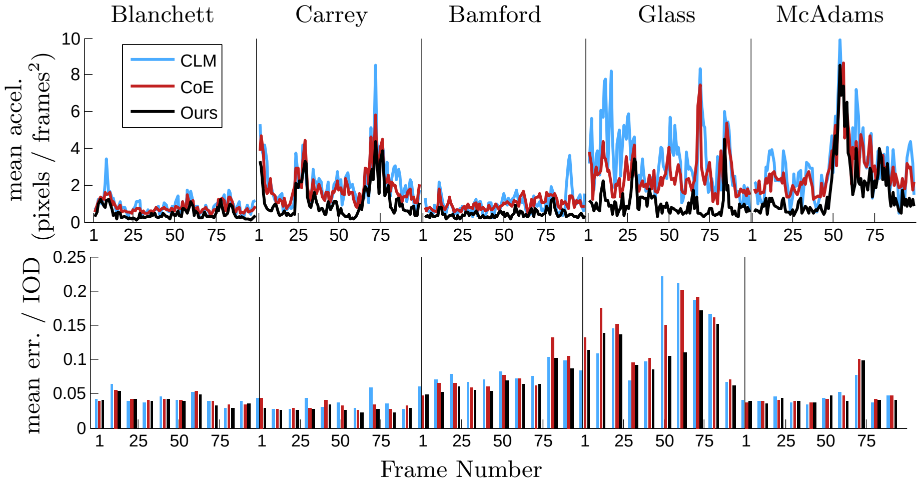
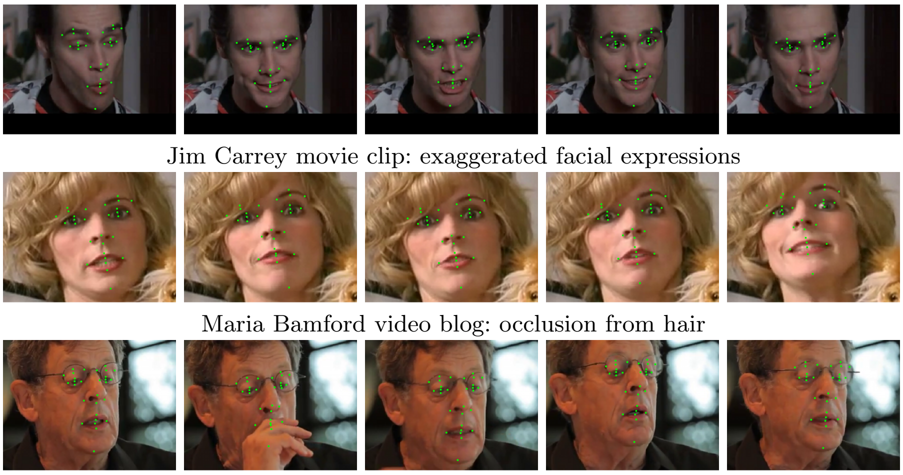


Fig. 4. This figure shows a comparison between our results, and the results generated by our implementations of Saragih *et al.*'s [16] algorithm (CLM), and Belhumeur *et al.*'s [1] algorithm (CoE), where both were applied to each frame of the videos independently. Our results are generally both more temporally stable (as measured by the average landmark acceleration in each frame), and slightly more accurate relative to ground truth. Please see Section 4.3 for an explanation of these results.


Jim Carrey movie clip: exaggerated facial expressions

Maria Bamford video blog: occlusion from hair

Philip Glass lecture: occlusion from hand gesture and glasses

Fig. 5. Selected frames from a subset of our experimental videos with landmarks estimated by our method shown in green. These video clips, downloaded from YouTube, were chosen with several key challenges in mind, including dramatic expressions, large head motion and pose variation, and occluded regions of the face.

References

1. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: Computer Vision and Pattern Recognition. pp. 545–552 (2011)
2. Bitouk, D., Kumar, N., Dhillon, S., Belhumeur, P., Nayar, S.K.: Face swapping: automatically replacing faces in photographs. In: ACM SIGGRAPH (2008)
3. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: ACM SIGGRAPH (1999)
4. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization 20(4), 1956–1982 (2008)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: European Conference on Computer Vision. pp. 484–498 (1998)
6. Gross, R., Matthews, I., Baker, S.: Generic vs. person specific active appearance models. Image and Vision Computing 23(11), 1080–1093 (2005)
7. Gross, R., Matthews, I., Baker, S.: Active appearance models with occlusion. Image and Vision Computing 24(6), 593–604 (2006)
8. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. In: International Conference on Automatic Face and Gesture Recognition (2008)
9. Gu, L., Kanade, T.: A generative shape regularization model for robust face alignment. In: European Conference on Computer Vision. pp. 413–426 (2008)
10. Kemelmacher-Shlizerman, I., Sankar, A., Shechtman, E., Seitz, S.M.: Being John Malkovich. In: European Conference on Computer Vision. pp. 341–353 (2010)
11. Kemelmacher-Shlizerman, I., Shechtman, E., Garg, R., Seitz, S.M.: Exploring photobios. In: ACM SIGGRAPH (2011)
12. Kumar, N., Berg, A.C., Belhumeur, P.N., Nayar, S.K.: Attribute and simile classifiers for face verification. In: International Conference on Computer Vision. pp. 365–372 (2009)
13. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
14. Lucey, S., Wang, Y., Saragih, J., Cohn, J.F.: Non-rigid face tracking with enforced convexity and local appearance consistency constraint. Image and Vision Computing 28(5), 781–789 (2010)
15. Matthews, I., Baker, S.: Active appearance models revisited. International Journal of Computer Vision 60(2), 135–164 (2004)
16. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: International Conference on Computer Vision. pp. 1034–1041 (2009)
17. Smith, B.M., Zhu, S., Zhang, L.: Face image retrieval by shape manipulation. In: Computer Vision and Pattern Recognition. pp. 769–776 (2011)
18. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
19. Wang, Y., Lucey, S., Cohn, J.F.: Enforcing convexity for improved alignment with constrained local models. In: Computer Vision and Pattern Recognition (2008)
20. Zhao, C., Cham, W.K., Wang, X.: Joint face alignment with a generic deformable face model. In: Computer Vision and Pattern Recognition. pp. 561–568 (2011)
21. Zhou, M., Liang, L., Sun, J., Wang, Y.: AAM based face tracking with temporal matching and face segmentation. In: Computer Vision and Pattern Recognition. pp. 701–708 (2010)