Computer Vision and Image Understanding 116 (2012) 922–935

Contents lists available at SciVerse ScienceDirect

Computer Vision and Image Understanding

journal homepage: www.elsevier.com/locate/cviu

Semi-supervised facial landmark annotation ☆

Yan Tong [a], Xiaoming Liu [b,*], Frederick W. Wheeler [b], Peter H. Tu [b]

[a] Department of Computer Science & Engineering, Univ. of South Carolina, Columbia, SC 29208, United States
[b] Visualization and Computer Vision Lab., GE Global Research, Niskayuna, NY 12309, United States

ARTICLE INFO

Article history:
Received 13 January 2011
Accepted 12 March 2012
Available online 24 April 2012

Keywords:
Face
Landmark
Annotation
Semi-supervised
Least-squares congealing
Image alignment
Image ensemble

ABSTRACT

Landmark annotation for training images is essential for many learning tasks in computer vision, such as object detection, tracking, and alignment. Image annotation is typically conducted manually, which is both labor-intensive and error-prone. To improve this process, this paper proposes a new approach to estimating the locations of a set of landmarks for a large image ensemble using manually annotated landmarks for only a small number of images in the ensemble. Our approach, named semi-supervised least-squares congealing, aims to minimize an objective function defined on both annotated and unannotated images. A shape model is learned online to constrain the landmark configuration. We employ an iterative coarse-to-fine patch-based scheme together with a greedy patch selection strategy for landmark location estimation. Extensive experiments on facial images show that our approach can reliably and accurately annotate landmarks for a large image ensemble starting with a small number of manually annotated images, under several challenging scenarios.

1. Introduction

Image annotation for training data is an essential step in many learning-based computer vision tasks. In general there are at least two types of prior knowledge represented by image annotation. One is *semantic* knowledge, such as a person's ID for face recognition, or an object's name for content-based image retrieval. The other is *geometric/landmark* knowledge. In learning-based object detection [1,2], for example, the position and size of the object (face/pedestrian/car) needs to be annotated for all training images. For supervised face alignment [3,4], each training image must be annotated with a set of landmarks, which describes the 2D location of the key facial features.

This paper focuses on geometric/landmark knowledge annotation, which is typically carried out *manually*. Practical applications, such as object detection, often require thousands of annotated images to achieve sufficient generalization capability. Hence, manual annotation becomes labor-intensive and time-consuming for these applications. Furthermore, image annotation is also an error-prone process due to annotator error, imperfect description of the objectives, and inconsistencies among different annotators.

To alleviate these problems, this paper presents an approach to automatically provide landmark annotation for a large set of images in a semi-supervised fashion. That is, using manually annotated landmark locations for a small number of images, our approach can automatically estimate the landmark locations for the entire set of images (see Fig. 1). In one example, we will demonstrate that 15 manually annotated images may be used to automatically annotate a complete set of 1176 images with the help of a face detector. The core of our algorithm, named *Semi-supervised Least-Squares Congealing (SLSC)*, is the minimization of an objective function defined as the summation of the pairwise L_2 distances between warped images. Two types of distances are used: the distance between the annotated and unannotated images, and the distance between the unannotated images. The objective function is iteratively minimized via the well-known and efficient inverse warping technique [5]. During the optimization process, we also constrain the estimated landmark locations by utilizing shape statistics that are learned in an online manner, which is shown to result in better convergence of landmark position estimates and hence improved robustness of the annotation.

Several prior work on joint alignment for an image ensemble [6–8] estimates global affine parameters for each image. However, most real-world objects exhibit non-rigid deformation that is not well-modeled by the affine transformation. Estimating more realistic deformations using a large set of landmarks is an important step towards accurately characterizing the shape variation within an object class. Motivated by this, we propose a hierarchical patch-based approach together with a greedy patch selection algorithm to estimating landmark positions. Starting from the whole face region, we iteratively select the patch with the greatest potential to minimize the objective function and conduct congealing for this patch simultaneously with its neighboring patch. These operations are consecutively applied to patches with gradually reduced size. In this strategy, the landmark annotation from the larger patch

☆ This paper has been recommended for acceptance by K.W. Bowyer.

* Corresponding author.
E-mail address: liux@research.ge.com (X. Liu).

1077-3142/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.cviu.2012.03.008

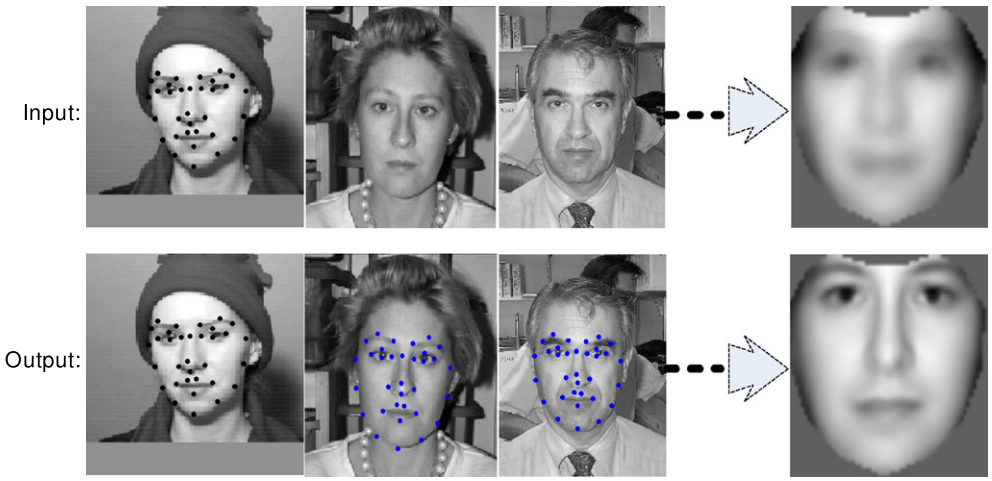

Fig. 1. Our approach takes an image ensemble as input with manually annotated landmark positions for only a small subset, and automatically estimates the landmarks for the remaining images. The mean warped face shown in the last column is an average of all warped faces in the image ensemble generated by a piecewise affine warp in a triangular face mesh based on the landmark positions. Note the improved sharpness in the mean warped face (the last column), an indicator of accurate landmark estimation by our algorithm.

can be propagated to smaller patches, which enhances the robustness of the annotation. Furthermore, congealing on small patches allows the locations of landmarks to be ultimately determined by local appearance information, which improves the precision of annotation. In addition, a joint congealing on two neighboring patches is proposed to enforce the geometrical consistency between them. Our applications on facial images show that even when manually annotating only a few images of the ensemble, the landmarks of the remaining images can be estimated accurately. An overview of the system is illustrated in Fig. 2.

Our proposed automatic image annotation framework has three primary contributions:

(1) A core algorithm is proposed for semi-supervised least-squares-based congealing of an image ensemble. We describe its efficient implementation using the inverse warping technique [5] and provide computational analysis.
(2) A statistical shape model learned online is integrated into the congealing process to reduce outliers of landmark estimation among the ensemble.
(3) A coarse-to-fine patch-based scheme together with a greedy patch selection strategy is proposed to improve the accuracy of landmark estimation. Furthermore, geometrical constraints are employed in the cost function to enforce the geometrical consistency between two neighboring patches, and thus improve the reliability of landmark annotation.

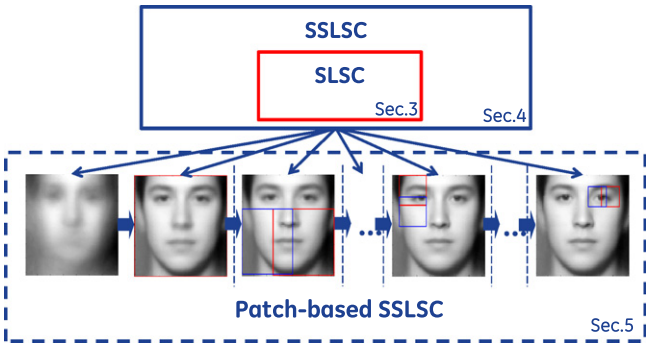

Fig. 2. System overview of the proposed landmark annotation approach. SLSC denotes the core algorithm of semi-supervised least-squares-based congealing; SSLSC represents the shape constrained SLSC; and patch-based SSLSC represents the coarse-to-fine landmark annotation system that employs SSLSC in each partitioning step.

(4) An end-to-end system is developed for automatic estimation of a set of landmarks in an ensemble of facial images with very few manually annotated images. Extensive experiments that qualitatively and quantitatively evaluate the performance and capabilities of the system and comparisons with the state-of-the-art techniques [9,8,10] have been conducted and are reported here.

The rest of the paper is organized as follows: After a brief description of related work in Section 2, this paper presents the semi-supervised least-squares-based congealing (SLSC) algorithm in Section 3. We then describe the shape constrained semi-supervised least-squares-based congealing (SSLSC) in Section 4, and the greedy patch selection scheme (i.e., patch-based SSLSC) in Section 5. Section 6 describes our extensive experimental results. The paper concludes in Section 7.

2. Prior work

In some notable and early work on *unsupervised* joint alignment, Learned-Miller [6,7] denotes the process as "congealing". The underlying idea is to minimize an entropy-based cost function by estimating the warping parameters of an ensemble. More recently, Cox et al. [8] propose a least-squares-based congealing (LSC) algorithm, which uses L_2 constraints to estimate the warping parameter of each image. An inverse compositional parameter updating strategy further improves the congealing performance in a larger database (500 images) in their updated work [11]. Vedaldi et al. [12] propose a joint data alignment approach based on the concept of lossy compression, which intends to generate a codebook optimal to the postulated structure of the data space. Storer et al. [13] propose a mutual information based cost function and formulate the congealing as a groupwise image registration problem. However, these approaches estimate only affine warping parameters for each image. Our work differs in that we estimate facial shape deformation described by a large set of landmarks, rather than a relatively simple global affine transformation.

Additional work on unsupervised image alignment has incorporated more general deformation models, though not with the use of a well-defined set of landmarks. Shelton [14] estimates the dense correspondence between a pair of n-dimensional surfaces. Through automatically establishing the correspondences between each sample in the surface ensemble and a reference sample, a

Morphable Surface Model is constructed. Balci et al. [15] extend the Learned-Miller's method [6] by including a free-form B-spline deformation model. Vetter et al. [16] have developed a bootstrapping algorithm to compute image correspondences and to learn a linear model based on optical flow. Guimond et al. [17] perform groupwise registration on brain images, where the affine transformation between the reference image and each image in the image set is removed beforehand. Baker et al. [18] use iterative Active Appearance Model (AAM) learning and fitting to estimate the location of mesh vertices, reporting results on images of the same person's face. Kokkinos and Yuille [19] formulate AAM learning as an EM algorithm and extend it to learning parts-based models for flexible objects. Cootes et al. [20–23] use a group-wise objective function to compute non-rigid registration. Sidorov et al. [24] further improve the efficiency of the groupwise registration by incrementally learning and accumulating the optimal deformation. Torre and Nguyen [25] improve manual facial landmark annotation based on parameterized kernel PCA. Langs et al. [26] employ an MDL-based cost function and estimate the correspondences for a set of control points. Asthana et al. [27] propagate the facial landmarks from frontal view images to arbitrary pose by initially learning the correspondence between frontal and varying pose images.

In general, we argue that for the discovery of non-rigid shape deformation using a specific set of physically defined landmarks, semi-supervised learning is more appropriate than unsupervised learning since prior knowledge of landmark location must be incorporated and this can be done easily via a few manually annotated examples. One could not rely upon an unsupervised learning algorithm to locate landmarks on physically meaningful facial features, such as mouth/eye corners or nose tip. This is the main difference between our work and the previously mentioned unsupervised image alignment approaches that do not utilize well-defined landmarks. Furthermore, with the guidance of the annotated samples, the parameter drift, which is a common issue unsupervised congealing approaches [8] suffer from, can be alleviated. Other unsupervised approaches on this topic are described by Cootes [28].

In contrast, there is a sizable literature for *supervised* face alignment, including Active Shape Model [4], AAM [3,9], Boosted Appearance Model [29]. Generally, a large number of annotated training images are required to train a statistical model so that it can generalize and fit unseen images well [30]. Hence, we are motivated to develop this semi-supervised approach to produce this training data more easily.

3. Semi-supervised least-squares congealing

In this section we will describe the objective function, detailed derivation, and computational analysis of our core algorithm, semi-supervised least-squares congealing (SLSC).

Similar to conventional congealing algorithms, SLSC takes an ensemble of images as input, among which we assume that there are K unannotated images $\mathbf{I} = \{\mathbf{I}_i\}_{i\in[1,K]}$, each of which is associated with an m-dimensional warping parameter vector $\mathbf{p}_i = [p_{i1}, p_{i2}, \ldots, p_{im}]^T$. $\mathbf{I}_i(\cdot)$ denotes a 1D vector containing the image intensity values at a given set of 2D pixel coordinates for image $\mathbf{I}_i$.

Each image $\mathbf{I}_i$ in the ensemble warps toward a predefined *common mean shape*[1] based on a warping function $\mathbf{W}(\mathbf{x};\mathbf{p}_i)$ that takes $\mathbf{x} = [x_1, y_1, x_2, y_2, \ldots, x_L, y_L]^T$, which is a collection of 2D coordinates of L pixels within the common mean shape, as input, and outputs the corresponding pixel coordinates in the coordinate space of image $\mathbf{I}_i$, according to the warping parameter vector $\mathbf{p}_i$. The warping function $\mathbf{W}(\mathbf{x};\mathbf{p}_i)$ can be a simple affine warp or a complex non-rigid warp such as the piecewise affine warp [9]; and $\mathbf{W}(\mathbf{x};\mathbf{0})$ represents an identical warp such that $\mathbf{W}(\mathbf{x};\mathbf{0}) = \mathbf{x}$. For example, $\mathbf{p}_i$ is defined as a 6-parameter affine warping parameter vector with $m = 6$ in this work, such that

$$\mathbf{W}(\mathbf{x};\mathbf{p}_i) = \begin{bmatrix} 1+\mathbf{p}_{i1}, & \mathbf{p}_{i2}, & \mathbf{p}_{i3} \\ \mathbf{p}_{i4}, & 1+\mathbf{p}_{i5}, & \mathbf{p}_{i6} \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{1}$$

Given this warping function, $\mathbf{I}_i(\mathbf{W}(\mathbf{x};\mathbf{p}_i))$ denotes the corresponding L-dimensional warped image vector obtained by bilinear interpolation of the image $\mathbf{I}_i$ using the warped coordinates $\mathbf{W}(\mathbf{x};\mathbf{p}_i)$, following the notation of [9,5]. The congealing process intends to iteratively align the warped image $\mathbf{I}_i(\mathbf{W}(\mathbf{x};\mathbf{p}_i))$ with the other images in the ensemble within the common mean shape.

Different from conventional congealing, our semi-supervised congealing further assumes there is a small set of $\widetilde{K}$ annotated images $\widetilde{\mathbf{I}} = \{\widetilde{\mathbf{I}}_n\}_{n\in[1,\widetilde{K}]}$, each with a set of manually annotated landmarks $\tilde{\mathbf{s}}_n$, where $\tilde{\mathbf{s}}_n$ is a concatenated vector of 2D landmark coordinates $\tilde{\mathbf{s}}_n = [x_{n1}, y_{n1}, x_{n2}, y_{n2}, \ldots, x_{nV}, y_{nV}]^T$ defined in the original image frame for V landmarks, which are located on biologically meaningful facial features. These biologically meaningful landmarks are usually of great interest and have distinguished appearance such as the eye corners and mouth corners. For each manually annotated image $\widetilde{\mathbf{I}}_n$, a warping parameter vector $\tilde{\mathbf{p}}_n$ can be calculated from $\tilde{\mathbf{s}}_n$ by solving a linear equation system

$$\begin{aligned} &\mathbf{W}(\mathbf{x}_{s1};\tilde{\mathbf{p}}_n) = \tilde{\mathbf{s}}_{n1} \\ &\ldots \\ &\mathbf{W}(\mathbf{x}_{sv};\tilde{\mathbf{p}}_n) = \tilde{\mathbf{s}}_{nv} \\ &\ldots \\ &\mathbf{W}(\mathbf{x}_{sV};\tilde{\mathbf{p}}_n) = \tilde{\mathbf{s}}_{nV} \end{aligned} \tag{2}$$

where $\tilde{\mathbf{s}}_{nv} = [x_{nv}, y_{nv}]^T$ denotes the 2D coordinates of the v^{th} landmarks in the original image frame, and $\mathbf{x}_{sv}$ represents the 2D coordinates of the corresponding landmark in the common mean shape. From Eq. (2), we can see that the warping parameter vector $\tilde{\mathbf{p}}_n$ is determined by the biologically meaningful landmarks. As a result, the content enclosed in the common mean shape represents what the well aligned face should look like, which will be utilized as a regulation during congealing and differentiates our proposed SLSC algorithm from other unsupervised approaches.

We denote the collection of all warping parameters to be estimated as $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_K]$. The goal of SLSC is to estimate $\mathbf{P}$ by minimizing a cost function defined on the entire ensemble:

$$\varepsilon(\mathbf{P}) = \sum_{i=1}^{K} \varepsilon_i(\mathbf{p}_i). \tag{3}$$

As we can see, the total cost $\varepsilon(\mathbf{P})$ is the summation of the individual cost of each unannotated image $\varepsilon_i(\mathbf{p}_i)$. We further define the individual cost as:

$$\begin{aligned} \varepsilon_i(\mathbf{p}_i) = &\frac{1-\alpha}{K-1} \sum_{j=1, j\neq i}^{K} \|\mathbf{I}_j(\mathbf{W}(\mathbf{x};\mathbf{p}_j)) - \mathbf{I}_i(\mathbf{W}(\mathbf{x};\mathbf{p}_i))\|^2 \\ &+ \frac{\alpha}{\widetilde{K}} \sum_{n=1}^{\widetilde{K}} \|\widetilde{\mathbf{I}}_n(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}_n)) - \mathbf{I}_i(\mathbf{W}(\mathbf{x};\mathbf{p}_i))\|^2, \end{aligned} \tag{4}$$

where $\mathbf{I}_i(\mathbf{W}(\mathbf{x};\mathbf{p}_i))$ denotes the L-dimensional warped image vector obtained by bilinear interpolation of the unannotated image $\mathbf{I}_i$ using the warped coordinates $\mathbf{W}(\mathbf{x};\mathbf{p}_i)$. Similarly, $\mathbf{I}_j(\mathbf{W}(\mathbf{x};\mathbf{p}_j))$ and

[1] Note that the common mean shape is defined as a reference frame of the region of interest(e.g., the face region in the face alignment application) consisting of L discrete image points. All images in the ensemble will warp toward the common mean shape such that each pair of corresponding pixels can be compared for a pair of warped images in the common mean shape. In our work, we use a pre-defined rectangular region as the common mean shape.

$\widetilde{\mathbf{I}}_n(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}_n))$ are the warping image vectors for the jth unannotated image and nth annotated image, respectively.

In the cost function, $\varepsilon_i(\mathbf{p}_i)$ equals the summation of the pairwise difference between each unannotated image $\mathbf{I}_i$ and all the other unannotated images $\mathbf{I}_j$ in the warped image space. On the one hand, minimizing the 1st term of Eq. (4) makes the warped image content of the ith unannotated image $\mathbf{I}_i(\mathbf{W}(\mathbf{x};\ \mathbf{p}_i))$ similar to that of the other *unannotated* images $\mathbf{I}_j(\mathbf{W}(\mathbf{x};\ \mathbf{p}_j))$, without regard for the physical meaning of the content. On the other hand, the 2nd term of Eq. (4) constrains the warped image content of the ith unannotated image $\mathbf{I}_i(\mathbf{W}(\mathbf{x};\ \mathbf{p}_i))$ to be similar to those of the *annotated* images $\widetilde{\mathbf{I}}_n(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}_n))$ and enforces the physical meaning of the content during alignment. Thus, the annotation of the annotated images $\widetilde{\mathbf{I}}$ are propagated to the unannotated images $\mathbf{I}$. Since the number of unannotated images is much greater than that of the annotated images ($K \gg \widetilde{K}$), a weighting coefficient α can balance the contributions of the two terms in the overall cost.

Since the total cost $\varepsilon(\mathbf{P})$ is difficult to optimize directly, we choose to iteratively minimize the individual cost $\varepsilon_i(\mathbf{p}_i)$ for each unannotated image $\mathbf{I}_i$. In order to estimate the warping parameter updates $\Delta\mathbf{p}_i$ for each unannotated image $\mathbf{I}_i$, instead of aligning $\mathbf{I}_i$ to the other images in the ensemble, we follows the approach described in [8], which aims to align each other images to $\mathbf{I}_i$ and has shown improved alignment performance by utilizing more image details of each image in the ensemble. To do this, we adopt the well-known inverse image warping technique [5] to minimize $\varepsilon_i(\mathbf{p}_i)$. We first estimate the warping parameter updates $\Delta\mathbf{p}_i$ by minimizing the following equation:

$$\varepsilon_i(\Delta\mathbf{p}_i) = \frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^{K}\|\mathbf{I}_j(\mathbf{W}(\mathbf{x};\mathbf{p}_j+\Delta\mathbf{p}_i)) - \mathbf{I}_i(\mathbf{W}(\mathbf{x};\mathbf{p}_i))\|^2 + \frac{\alpha}{\widetilde{K}}\sum_{n=1}^{\widetilde{K}}\|\widetilde{\mathbf{I}}_n(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}_n+\Delta\mathbf{p}_i)) - \mathbf{I}_i(\mathbf{W}(\mathbf{x};\mathbf{p}_i))\|^2, \quad (5)$$

where $\mathbf{W}(\mathbf{x};\ \mathbf{p}_j + \Delta\mathbf{p}_i)$ represents an image warping that aligns image $\mathbf{I}_j$ to image $\mathbf{I}_i$ with the warping parameter update $\Delta\mathbf{p}_i$.

Then the warping function is updated by:

$$\mathbf{W}(\mathbf{x};\mathbf{p}_i) \leftarrow \mathbf{W}(\mathbf{x};\mathbf{p}_i - \Delta\mathbf{p}_i). \quad (6)$$

It is not straightforward to optimize Eq. (5) directly because the function $\varepsilon_i(\Delta\mathbf{p}_i)$ is nonlinear w.r.t. $\Delta\mathbf{p}_i$. We choose to approximate this function by taking the first order Taylor expansion on $\mathbf{I}_j(\mathbf{W}(\mathbf{x};\ \mathbf{p}_j + \Delta\mathbf{p}_i))$ and $\widetilde{\mathbf{I}}_n(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}_n + \Delta\mathbf{p}_i))$:

$$\mathbf{I}_j(\mathbf{W}(\mathbf{x};\mathbf{p}_j+\Delta\mathbf{p}_i)) \approx \mathbf{I}_j(\mathbf{W}(\mathbf{x};\mathbf{p}_j)) + \frac{\partial\mathbf{I}_j(\mathbf{W}(\mathbf{x};\mathbf{p}_j))}{\partial\mathbf{p}_j}\Delta\mathbf{p}_i.$$

As a result, Eq. (5) is simplified to:

$$\varepsilon_i(\Delta\mathbf{p}_i) \approx \frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^{K}\|\mathbf{b}_j + \mathbf{c}_j\Delta\mathbf{p}_i\|^2 + \frac{\alpha}{\widetilde{K}}\sum_{n=1}^{\widetilde{K}}\|\tilde{\mathbf{b}}_n + \tilde{\mathbf{c}}_n\Delta\mathbf{p}_i\|^2, \quad (7)$$

where

$$\mathbf{b}_j = \mathbf{I}_j(\mathbf{W}(\mathbf{x};\mathbf{p}_j)) - \mathbf{I}_i(\mathbf{W}(\mathbf{x};\mathbf{p}_i)), \mathbf{c}_j = \frac{\partial\mathbf{I}_j(\mathbf{W}(\mathbf{x};\mathbf{p}_j))}{\partial\mathbf{p}_j},$$

$$\tilde{\mathbf{b}}_n = \widetilde{\mathbf{I}}_n(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}_n)) - \mathbf{I}_i(\mathbf{W}(\mathbf{x};\mathbf{p}_i)), \tilde{\mathbf{c}}_n = \frac{\partial\widetilde{\mathbf{I}}_n(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}_n))}{\partial\tilde{\mathbf{p}}_n}.$$

The minimization of Eq. (7) can be obtained by setting its partial derivative w.r.t. $\Delta\mathbf{p}_i$ to zero. We then have

$$\Delta\mathbf{p}_i = -\mathbf{H}^{-1}\left[\frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^{K}\mathbf{c}_j^T\mathbf{b}_j + \frac{\alpha}{\widetilde{K}}\sum_{n=1}^{\widetilde{K}}\tilde{\mathbf{c}}_n^T\tilde{\mathbf{b}}_n\right], \quad (8)$$

with

$$\mathbf{H} = \frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^{K}\mathbf{c}_j^T\mathbf{c}_j + \frac{\alpha}{\widetilde{K}}\sum_{n=1}^{\widetilde{K}}\tilde{\mathbf{c}}_n^T\tilde{\mathbf{c}}_n. \quad (9)$$

Joint alignment for an image ensemble can be a computationally intensive task. We have analyzed the computational cost of the SLSC method. Note that since $\widetilde{\mathbf{P}} = \{\tilde{\mathbf{p}}_n\}_{n\in[1,\widetilde{K}]}$ are known, $\tilde{\mathbf{c}}_n$ and part of the Hessian matrix $\mathbf{H}$ can be pre-computed and remain fixed during the iterations. As shown in Table 1, the computational cost for solving the second term of Eq. (7) is negligible. Therefore, semi-supervised congealing has a computational cost similar to that of unsupervised congealing.

4. Shape-constrained SLSC

In this section, we will introduce a shape-constrained SLSC, which improves the robustness of the congealing process by reducing outliers. Given the warping parameters for all images $\{\mathbf{P},\widetilde{\mathbf{P}}\} = [\mathbf{p}_1,\ldots,\mathbf{p}_K,\tilde{\mathbf{p}}_1,\ldots,\tilde{\mathbf{p}}_{\widetilde{K}}]$, and their corresponding landmark locations $\{\mathbf{S},\widetilde{\mathbf{S}}\} = [\mathbf{s}_1,\ldots,\mathbf{s}_K,\tilde{\mathbf{s}}_1,\ldots,\tilde{\mathbf{s}}_{\widetilde{K}}]$, where $\mathbf{s}$ is a concatenated vector of 2D landmarks $\mathbf{s} = [x_1,\ y_1,\ x_2,\ y_2,\ldots,x_V,\ y_V]^T$ for V landmarks, there are two ways of mapping between each pair of the warping parameter vector $\mathbf{p}_k \in \{\mathbf{P},\widetilde{\mathbf{P}}\}$ and its corresponding landmark locations $\mathbf{s}_k \in \{\mathbf{S},\widetilde{\mathbf{S}}\}$ for the k^{th} image in the ensemble. First, the warping parameter $\mathbf{p}_i$ can be obtained given the corresponding landmark locations $\mathbf{s}_i$ as described in Eq. (2). Consequently, improving the localization of the landmarks by a method independent of SLSC such as a PCA model can refine the estimation of the warping parameters obtained by SLSC. Second, the landmarks $\mathbf{s}_i$ can be obtained from the warping parameter $\mathbf{p}_i$ via $\mathbf{s}_i = \mathbf{W}(\mathbf{x}_s;\ \mathbf{p}_i)$, where $\mathbf{x}_s$ is a vector containing the coordinates of the target landmarks in the common mean shape. As a result, an incorrect warping parameter, which can result from an outlier in the congealing process, would produce a landmark set that is not a valid shape instance. Motivated by this, we develop an approach denoted *Shape-constrained SLSC (SSLSC)*, which integrates the shape constraints into each iteration of the appearance-based congealing process to improve the robustness of the SLSC.

Given that the objects in the ensemble have the same topological structure, we assume that the shape deformation of $\mathbf{s}_i$ satisfies a Point Distribution Model (PDM) [4]. Since only a few annotated images are available, the PDM is learned from both the annotated landmarks and an automatically selected *low-error* subset of the estimated landmarks in an online manner.

Table 1
The computational cost of major steps in SLSC. $\widetilde{\mathbf{I}}(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}))$ and $\mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p}))$ represent the image warping processes for all the annotated and unannotated images, respectively. $\frac{\partial\widetilde{\mathbf{I}}(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}))}{\partial\tilde{\mathbf{p}}}$ and $\frac{\partial\mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p}))}{\partial\mathbf{p}}$ denote the processes of calculating the partial derivatives for annotated and unannotated images, respectively. $\tilde{\mathbf{c}}_n = \frac{\partial\mathbf{I}_n(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}_n))}{\partial\tilde{\mathbf{p}}_n}$ and $\mathbf{c}_j = \frac{\partial\mathbf{I}_j(\mathbf{W}(\mathbf{x};\mathbf{p}_j))}{\partial\mathbf{p}_j}$ represent the partial derivative of the nth annotated and jth unannotated images, respectively. $\Delta\mathbf{P}$ is the warping parameter updates for all unannotated images. K and $\widetilde{K}$ are the total numbers of unannotated and annotated images, respectively; m is the dimension of the warping parameter vector p_i; and L is the number of pixels in the common mean shape.

Pre-comp.	$\widetilde{\mathbf{I}}(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}))$	$O(\widetilde{K}L)$
	$\frac{\partial\widetilde{\mathbf{I}}(\mathbf{W}(\mathbf{x};\tilde{\mathbf{p}}))}{\partial\tilde{\mathbf{p}}}$	$O(m\widetilde{K}L)$
	$\sum_{n=1}^{\widetilde{K}}\tilde{\mathbf{c}}_n^T\tilde{\mathbf{c}}_n$	$O(m^2\widetilde{K}L)$
Per-Iteration	$\mathbf{I}(\mathbf{W}(\mathbf{x};\ \mathbf{p}))$	$O(KL)$
	$\frac{\partial\mathbf{I}(\mathbf{W}(\mathbf{x};\mathbf{p}))}{\partial\mathbf{p}}$	$O(mKL)$
	$\sum_{j=1,j\neq i}^{K}\mathbf{c}_j^T\mathbf{c}_j$	$O(m^2KL)$
	Inverse Hessian $\mathbf{H}$	$O(m^2log(m)K)$
	Compute $\Delta\mathbf{P}$	$O(mK(K+\widetilde{K})L + m^2K)$
	Total	$O(mK(m(L+log(m)) + (K+\widetilde{K})L))$

Specifically, since $\varepsilon_i(\Delta\mathbf{p}_i)$ defined in Eq. (5) indicates the degree of misalignment between the ith unannotated image and the other images in the ensemble, we can rank $\varepsilon_1(\Delta\mathbf{p}_1),\ldots,\varepsilon_K(\Delta\mathbf{p}_K)$ in an ascending order and select the first K_M[2] images as a *low-error* subset, each of which has the corresponding landmark positions calculated as $\mathbf{s}_i=\mathbf{W}(\mathbf{x}_s;\mathbf{p}_i)$ for $i\in[1,K_M]$. Then, we can form a training set $\mathbf{S}_{\widetilde{K+K_M}}=[\tilde{\mathbf{s}}_1,\ldots,\tilde{\mathbf{s}}_{\widetilde{K}},\mathbf{s}_1,\ldots,\mathbf{s}_{K_M}]^T$ for online PDM learning. Next, the other *poor* estimates, i.e., the other images with higher $\varepsilon_i(\Delta\mathbf{p}_i)$, can be "corrected" through a PCA reconstruction as follows:

$$\hat{\mathbf{s}}_i=\bar{\mathbf{s}}+\mathbf{Q}\mathbf{z}_i, \tag{10}$$

where $\hat{\mathbf{s}}_i$ is the reconstructed shape vector for the ith image; $\bar{\mathbf{s}}$ and $\mathbf{Q}$ are the mean shape and the shape basis learned through the online PDM training; $\mathbf{z}_i$ is the projection of $\mathbf{s}_i$ on the PCA basis and is restricted in some range [4]. In this work, the number of shape basis vectors is automatically determined such that 90% shape variations in the training set are preserved. Finally, a new warping parameter vector $\hat{\mathbf{p}}_i$ is computed from the refined landmark positions $\hat{\mathbf{s}}_i$ by solving the inverted warping of $\hat{\mathbf{s}}_i=\mathbf{W}(\mathbf{x}_s;\hat{\mathbf{p}}_i)$. By doing so, the outliers of the congealing process are identified and constrained in a principled way.

Algorithm 1. Shape-constrained SLSC (SSLSC)

Input; $\mathbf{I},\widetilde{\mathbf{I}},\mathbf{P}^0,\widetilde{\mathbf{P}},\mathbf{x}$, and $\mathbf{x}_s$
Output: $\mathbf{P}^t$, $\mathbf{S}^t$, and ε
$t\leftarrow 0$;
Compute $\tilde{\mathbf{p}}_n$ from $\tilde{\mathbf{s}}_n$ by solving the inverted warping of $\tilde{\mathbf{s}}_n=\mathbf{W}(\mathbf{x}_s;\tilde{\mathbf{p}}_n)$ for $n\in[1,\widetilde{K}]$;
repeat
  for $i=1$ to K **do**
    $\varepsilon_i(\Delta\mathbf{p}_i),\mathbf{p}_i^{t+1}\leftarrow SLSC(\mathbf{I},\widetilde{\mathbf{I}},\mathbf{P}^t,\widetilde{\mathbf{P}},\mathbf{x})$;
  end for
  Rank $\varepsilon_1(\Delta\mathbf{p}_1),\ldots,\varepsilon_K(\Delta\mathbf{p}_K)$ in ascending order and select the first K_M images;
  Compute $\mathbf{s}_i^{t+1}=\mathbf{W}(\mathbf{x}_s;\mathbf{p}_i^{t+1})$ for $i\in[1,K_M]$;
  $\bar{\mathbf{s}},\mathbf{Q},\lambda\leftarrow$ PCA on $\mathbf{S}_{\widetilde{K+K_M}}=[\tilde{\mathbf{s}}_1,\ldots,\tilde{\mathbf{s}}_{\widetilde{K}},\mathbf{s}_1^{t+1},\ldots,\mathbf{s}_{K_M}^{t+1}]^T$;
  for $i=K_M+1$ to K **do**
    Reconstruct $\mathbf{s}_i^{t+1}$ as $\hat{\mathbf{s}}_i^{t+1}=\bar{\mathbf{s}}+\mathbf{Q}\mathbf{z}_i$, where $\mathbf{z}_i$ is restricted by some predefined range;
    Compute $\mathbf{p}_i^{t+1}$ from $\mathbf{s}_i^{t+1}$ by solving the inverted warping of $\mathbf{s}_i^{t+1}=\mathbf{W}(\mathbf{x}_s;\mathbf{p}_i^{t+1})$;
  end for
  $\mathbf{P}^{t+1}\leftarrow[\mathbf{p}_1^{t+1},\ldots,\mathbf{p}_K^{t+1}]$; $\mathbf{S}^{t+1}\leftarrow[\mathbf{s}_1^{t+1},\ldots,\mathbf{s}_K^{t+1}]^T$;
  $\varepsilon\leftarrow\sum_{i=1}^K\varepsilon_i(\Delta\mathbf{p}_i)$; $t\leftarrow t+1$.
until Converge

The SSLSC algorithm is summarized in Algorithm 1, where $\mathbf{P}^0$ is the initial warping parameter set for $\mathbf{I}$. The annotated landmarks are fully utilized in the sense that they not only contribute to the cost function minimization in Eq. (5), but also provide guidance for plausible shape deformation.

5. Iterative patch-based landmark annotation

Having dealt with the outliers during congealing process, we should improve the accuracy of the landmark annotation, which is crucial for practical applications. Since the shape deformation of a real-world object is often non-rigid due to inter-subject variations, object motions, and camera views, estimating the global and rigid transformation of the object is not sufficient to characterize the object. However, joint estimating the non-rigid transformation is difficult to solve since the warping parameter vector $\mathbf{p}_i$ resides in a high dimensional space.

In this section, we propose a patch-based approach to achieve an accurate estimate of the landmarks by searching for the optimal patch for congealing in the common mean shape. In the SSLSC algorithm, the warping function $\mathbf{W}(\mathbf{x};\mathbf{p})$ can be a simple global affine warp to model rigid transformation, or a piecewise affine warp to model non-rigid transformation. However, from our experience, the SSLSC algorithm does not perform satisfactorily when directly setting $\mathbf{W}(\mathbf{x};\mathbf{p})$ to be a piecewise affine warp. We attribute this difficulty to the high dimensionality of the warping parameter $\mathbf{p}_i$ in a piecewise affine warp.

To understand this issue, let us look at the piecewise affine warp closely. We note that in this case the warping function $\mathbf{W}(\cdot)$ is a concatenation of multiple affine transformations, each operating within a small triangular patch. On the one hand, the patch allows us to work in a space whose dimension is much smaller than the original space, and thus makes the problem easier to solve. On the other hand, *directly* applying the SSLSC on the small patches is not reliable due to poor initialization and limited information encoded in a single patch. Motivated by these observations, we developed a coarse-to-fine patch-based scheme to improve the precision of landmark annotation, where an affine warp is performed on each patch independently.

Our previous work [10] employs a brute-force partitioning method, which repeatedly partitions the common mean shape for a selected patch with the maximal congealing error (ϵ). Two equal-size and overlapped child patches are generated by each partitioning; and then two congealing processes are performed on the generated child patches, individually. Since the landmarks in the overlapped region may have inconsistent estimations caused by two different warping functions, [10] simply averages the estimated landmarks resulted from two individual SSLSC processes applied on the two patches independently. However, this brute-force partitioning strategy does not guarantee that the selected patch has the greatest potential to decrease ϵ. As a result, the process is stopped after a limited number of congealing rounds; and hence the accuracy of landmark annotation is limited.

In this work, we propose a novel greedy patch selection strategy by estimating the optimal region for alignment such that the congealing error (ϵ) decreases most significantly. To this end, we estimate the gradient of the congealing error w.r.t. the warping parameters for the image ensemble at the lth pixel of the common mean shape:

$$\begin{aligned}\frac{\partial\epsilon^l}{\partial\mathbf{P}}&=\frac{\partial\sum_{i=1}^K\varepsilon_i^l(\mathbf{p}_i)}{\partial\mathbf{P}}\\&=2\sum_{i=1}^K\left[\frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^K\left(\frac{\partial\mathbf{I}_j(\mathbf{W}(\mathbf{x}_l;\mathbf{p}_j))}{\partial\mathbf{P}}-\frac{\partial\mathbf{I}_i(\mathbf{W}(\mathbf{x}_l;\mathbf{p}_i))}{\partial\mathbf{P}}\right)\mathbf{b}_j(\mathbf{x}_l)\right.\\&\quad\left.-\frac{\alpha}{\widetilde{K}}\sum_{n=1}^{\widetilde{K}}\frac{\partial\mathbf{I}_i(\mathbf{W}(\mathbf{x}_l;\mathbf{p}_i))}{\partial\mathbf{P}}\tilde{\mathbf{b}}_n(\mathbf{x}_l)\right],\end{aligned} \tag{11}$$

where $\mathbf{x}_l$ is the coordinates of the lth pixel in the common mean shape; $\mathbf{b}_j$ and $\tilde{\mathbf{b}}_n$ are defined accordingly as in Eq. (7). There is similar prior work [31] where patch locations are searched in the optimization process of an objective function.

The process takes place in a series of congealing rounds. Starting from the whole common mean shape $\mathbf{x}$, the process is conducted, in each round, by searching for the square patch ($\mathbf{x}^{k*}$) in the whole common mean shape for congealing, which has the maximal value

[2] In this work, K_M is set as 50% of the unannotated images since we believe that the percentage of outliers is low.

of $\frac{1}{L_r}\left\|\sum_{l=1}^{L_r}\frac{\partial \epsilon^l}{\partial \mathbf{p}}\right\|^2$, where L_r is the total number of pixels within the patch for the current congealing round. Once Eq. (11) is computed for all pixels within the common mean shape, the aforementioned searching can be efficiently implemented by the integral image technique [32]. If the same patch is selected after congealing, we will search for and operate congealing on a new patch with a reduced patch size. In this work, the patch size is initialized to the size of the whole common mean shape and reduced by a factor of 2/3 in each dimension when no further improvement can be made at the current patch size. The process is stopped when the size of the patch reaches a limit. Here, the patch reaches its size limit if the number of target landmarks in the selected patch is less than the number of landmarks required for computing $\mathbf{p}_i$ in Eq. (2).

However, it is not reliable and robust to perform the congealing on a single patch alone, since the geometric relationships between the patch and the global context are neglected. In this work, besides performing congealing on the optimal patch $\mathbf{x}^{k*}$, an additional patch is chosen for alignment such that it overlaps with $\mathbf{x}^{k*}$, i.e., some target landmarks reside in both patches, and has the maximal value $\frac{1}{L'_r}\left\|\sum_{l=1}^{L'_r}\frac{\partial \epsilon^l}{\partial \mathbf{p}}\right\|^2$ among all possible neighbors[3] of $\mathbf{x}^{k*}$. We should note that the additional patch may have smaller size than L_r especially for the first several congealing rounds due to the boundary limit of whole common mean shape. An example of patch selection over multiple congealing rounds is shown in Fig. 3. In this work, we further enforce the geometric consistency between the two selected patches, i.e., the landmarks in the overlapped region should have consistent estimations under the two warping functions with the two patches. In order to utilize this geometrical constraint, the warping parameters of the ith unannotated image for the two patches are estimated simultaneously using a single cost function:

$$\varepsilon_i(\Delta\mathbf{p}_i^1, \Delta\mathbf{p}_i^2) = \left(\varepsilon_i(\Delta\mathbf{p}_i^1) + \varepsilon_i(\Delta\mathbf{p}_i^2)\right) + \beta\left\|\mathbf{s}_i^1 - \mathbf{s}_i^2\right\|^2, \tag{12}$$

where $\Delta\mathbf{p}_i^1$ and $\Delta\mathbf{p}_i^2$ represent the warping parameter updates for the two patches, respectively. As shown in Fig. 4, $\mathbf{x}_s^{1,2} = \mathbf{x}_s^1 \cap \mathbf{x}_s^2$ is a subset of $\mathbf{x}_s$ and consists of the 2D coordinates of the landmarks in the overlapped region in the common mean shape. $\mathbf{s}_i^1 = \mathbf{W}(\mathbf{x}_s^{1,2}; \mathbf{p}_i^1 - \Delta\mathbf{p}_i^1)$ contains the estimated landmark coordinates in the image $\mathbf{I}_i$ corresponding to $\mathbf{x}_s^{1,2}$ with the warping parameter $\mathbf{p}_i^1$ and warping parameter updates $\Delta\mathbf{p}_i^1$ for the first patch. $\mathbf{s}_i^2 = \mathbf{W}(\mathbf{x}_s^{1,2}; \mathbf{p}_i^2 - \Delta\mathbf{p}_i^2)$ is defined accordingly for the second patch. The first term of Eq. (12) is the summation of the congealing errors of the two patches, and the second term is a constraint to enforce the geometric consistency between the two patches. By doing this, the overlapped landmarks play a role as a joint connecting two patches, and directly contribute to the estimation of warping parameters for these two patches. Parameter β is a weighting coefficient that adjusts the strength of the constraint.

Since $\mathbf{W}(\mathbf{x}_s^{1,2}; \mathbf{p}_i^1 - \Delta\mathbf{p}_i^1)$ is nonlinear w.r.t. $\Delta\mathbf{p}_i^1$, we take the first order Taylor expansion on $\mathbf{W}(\mathbf{x}_s^{1,2}; \mathbf{p}_i^1 - \Delta\mathbf{p}_i^1)$:

$$\mathbf{W}(\mathbf{x}_s^{1,2}; \mathbf{p}_i^1 - \Delta\mathbf{p}_i^1) \approx \mathbf{W}(\mathbf{x}_s^{1,2}; \mathbf{p}_i^1) - \frac{\partial \mathbf{W}(\mathbf{x}_s^{1,2}; \mathbf{p}_i^1)}{\partial \mathbf{p}_i^1}\Delta\mathbf{p}_i^1. \tag{13}$$

Substituting Eqs. (7) to (12), the cost function becomes:

$$\begin{aligned}\varepsilon_i(\Delta\mathbf{p}_i^1, \Delta\mathbf{p}_i^2) = &\left(\frac{1-\alpha}{K-1}\sum_{j=1, j\neq i}^{K}\left\|\mathbf{b}_j^1 + \mathbf{c}_j^1\Delta\mathbf{p}_i^1\right\|^2 + \frac{\alpha}{\widetilde{K}}\sum_{n=1}^{\widetilde{K}}\left\|\tilde{\mathbf{b}}_n^1 + \tilde{\mathbf{c}}_n^1\Delta\mathbf{p}_i^1\right\|^2\right.\\ &\left.+\frac{1-\alpha}{K-1}\sum_{j=1, j\neq i}^{K}\left\|\mathbf{b}_j^2 + \mathbf{c}_j^2\Delta\mathbf{p}_i^2\right\|^2 + \frac{\alpha}{\widetilde{K}}\sum_{n=1}^{\widetilde{K}}\left\|\tilde{\mathbf{b}}_n^2 + \tilde{\mathbf{c}}_n^2\Delta\mathbf{p}_i^2\right\|^2\right)\\ &+\beta\left\|\mathbf{d}_i^1 - \mathbf{d}_i^2 - \mathbf{e}_i(\Delta\mathbf{p}_i^1 - \Delta\mathbf{p}_i^2)\right\|^2,\end{aligned} \tag{14}$$

[3] In this work, 4 neighbors are considered for four positions (left, right, up, and down) to the optimal patch.

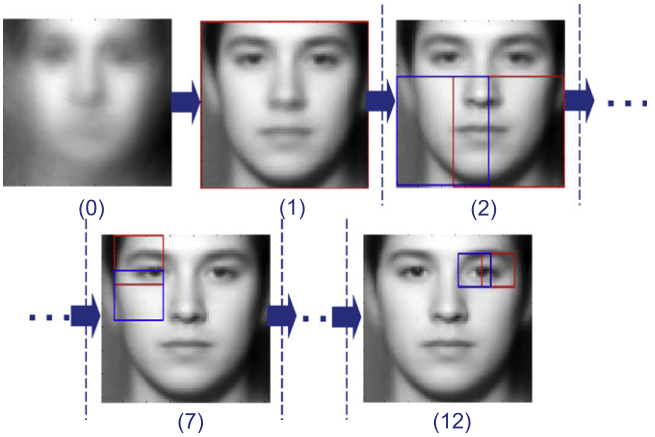

Fig. 3. An example shows patch selection for 12 congealing rounds in the common mean shape. The red rectangle encloses the region that has the maximal value of $\frac{1}{L_r}\left\|\sum_{l=1}^{L_r}\frac{\partial \epsilon^l}{\partial \mathbf{p}}\right\|^2$; and the blue rectangle is its selected neighbor. Note the improved sharpness in the patches that have been aligned: nose and mouth are sharper in (2) than those in (1); the left eye is clearer in (7) than that in (2); and the right eye is sharper in (12) than that in (7).

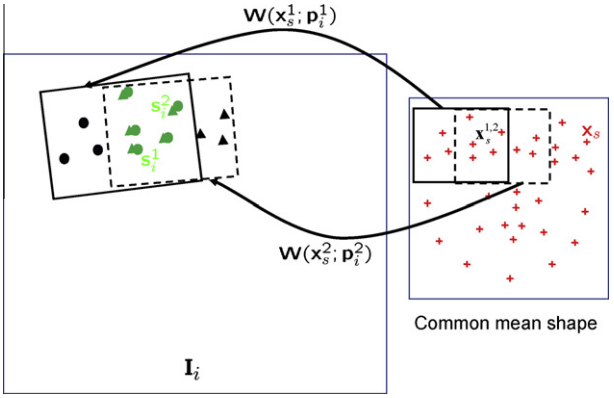

Fig. 4. Illustration of the geometric constraints in Eq. (12). In both the common mean shape (right) and the original image $\mathbf{I}_i$ (left), the solid and dashed rectangles represent the two patches being aligned. The crosses represent the target landmarks in the common mean shape with 2D coordinates denoted by $\mathbf{x}_s$, which are warped to the dots and triangles in the original image $\mathbf{I}_i$ according to two different warping functions $\mathbf{W}(\mathbf{x}_s^1; \mathbf{p}_i^1)$ and $\mathbf{W}(\mathbf{x}_s^2; \mathbf{p}_i^2)$, respectively. In the overlapped area of the two patches in $\mathbf{I}_i$, the green dots denoted by $\mathbf{s}_i^1$ and green triangles denoted by $\mathbf{s}_i^2$ are not overlapped. The geometrical constraints employed in Eq. (12) enforce $\mathbf{s}_i^1$ and $\mathbf{s}_i^2$ to be overlapped since they correspond to the same set of landmarks ($\mathbf{x}_s^{1,2} = \mathbf{x}_s^1 \cap \mathbf{x}_s^2$) in the common mean shape.

where $\mathbf{d}_i^1 = \mathbf{W}(\mathbf{x}_s^{1,2}; \mathbf{p}_i^1), \mathbf{d}_i^2 = \mathbf{W}(\mathbf{x}_s^{1,2}; \mathbf{p}_i^2)$, and $\mathbf{e}_i = \frac{\partial \mathbf{W}(\mathbf{x}_s^{1,2}; \mathbf{p}_i^1)}{\partial \mathbf{p}_i^1} = \frac{\partial \mathbf{W}(\mathbf{x}_s^{1,2}; \mathbf{p}_i^2)}{\partial \mathbf{p}_i^2}$.

The warping parameter updates $\Delta\mathbf{p}_i^1$ and $\Delta\mathbf{p}_i^2$ are estimated by solving a linear equation system as:

$$\begin{cases}\frac{\partial \varepsilon_i(\Delta\mathbf{p}_i^1, \Delta\mathbf{p}_i^2)}{\partial \Delta\mathbf{p}_i^1} = 0,\\ \frac{\partial \varepsilon_i(\Delta\mathbf{p}_i^1, \Delta\mathbf{p}_i^2)}{\partial \Delta\mathbf{p}_i^2} = 0.\end{cases} \tag{15}$$

Substituting Eqs. (12) to (15), we have:

$$\begin{bmatrix} \mathbf{A}_1, & \mathbf{B} \\ \mathbf{B}, & \mathbf{A}_2 \end{bmatrix} \begin{bmatrix} \Delta\mathbf{p}_i^1 \\ \Delta\mathbf{p}_i^2 \end{bmatrix} = -\begin{bmatrix} \mathbf{C}_1 \\ \mathbf{C}_2 \end{bmatrix}, \tag{16}$$

where

$$\mathbf{A}_1 = \frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^{K}(\mathbf{c}_j^1)^T\mathbf{c}_j^1 + \frac{\alpha}{\widetilde{K}}\sum_{n=1}^{\widetilde{K}}(\tilde{\mathbf{c}}_n^1)^T\tilde{\mathbf{c}}_n^1 + \beta\mathbf{e}_i^T\mathbf{e}_i, \tag{17}$$

$$\mathbf{A}_2 = \frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^{K}(\mathbf{c}_j^2)^T\mathbf{c}_j^2 + \frac{\alpha}{\widetilde{K}}\sum_{n=1}^{\widetilde{K}}(\tilde{\mathbf{c}}_n^2)^T\tilde{\mathbf{c}}_n^2 + \beta\mathbf{e}_i^T\mathbf{e}_i, \tag{18}$$

$$\mathbf{B} = -\beta\mathbf{e}_i^T\mathbf{e}_i, \tag{19}$$

$$\mathbf{C}_1 = \frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^{K}(\mathbf{c}_j^1)^T\mathbf{b}_j^1 + \frac{\alpha}{\widetilde{K}}\sum_{n=1}^{\widetilde{K}}(\tilde{\mathbf{c}}_n^1)^T\tilde{\mathbf{b}}_n^1 - \beta\mathbf{e}_i^T(\mathbf{d}_i^1 - \mathbf{d}_i^2), \tag{20}$$

$$\mathbf{C}_2 = \frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^{K}(\mathbf{c}_j^2)^T\mathbf{b}_j^2 + \frac{\alpha}{\widetilde{K}}\sum_{n=1}^{\widetilde{K}}(\tilde{\mathbf{c}}_n^2)^T\tilde{\mathbf{b}}_n^2 + \beta\mathbf{e}_i^T(\mathbf{d}_i^1 - \mathbf{d}_i^2). \tag{21}$$

Compared to congealing on the two patches separately based on Eq. (7), performing congealing jointly in this manner requires the computation of $\mathbf{e}_i, \mathbf{e}_i^T\mathbf{e}_i$, $\mathbf{d}_i^1$, and $\mathbf{d}_i^2$ additionally. Note that since $\mathbf{x}_s^{1,2}$ is known for each congealing process, $\mathbf{e}_i$ and $\mathbf{e}_i^T\mathbf{e}_i$ can be computed in advance. Since the dimension of $\mathbf{x}_s^{1,2}$ is small (less than the total number of target landmarks), the cost for computing $\mathbf{d}_i^1$ and $\mathbf{d}_i^2$ can be neglected. Furthermore, as we mentioned before, L_r' may be smaller than L_r, and thus, congealing at each round using the greedy patch selection method has a lower computational cost than performing congealing on two equal-size patches.

The partitioning strategy is summarized in Algorithm 2, where $\mathbf{S}_{init}$ is the initial guess of the landmark positions for the unannotated images $\mathbf{I}$. In the algorithm, besides the notation mentioned previously, *size* represents the size of the patches to be aligned in the current partition level; d represents the index of the patch being aligned. Starting from the initial common mean shape space $\mathbf{x}^1$, the process is conducted by repeatedly searching an optimal patch $(\mathbf{x}^{k*})$, which has the maximum potential to decrease the misalignment error (ε), in the common mean shape space. A second patch $(\mathbf{x}^{k**})$ is selected such that it can decrease ε the most significantly among the neighboring patches of $(\mathbf{x}^{k*})$. The patch will be shrunk by a factor of 2/3, if $\mathbf{x}^{k*}$ is the same as the one in the previous congealing run. Then, the SSLSC is applied on two selected patches simultaneously based on Eq. (12) to obtain the corresponding landmark positions. This process is stopped when the size of the patch is too small. Here, the patch reaches its size limit if the number of target landmarks in the patch is less than the number of landmarks required for computing $\mathbf{p}_i$ in Eq. (2). One example of multiple-level partition is shown in Fig. 3.

Our top-down congealing strategy performs a coarse-to-fine alignment for the entire image ensemble. The congealing on larger patches focuses on aligning the features that have the greatest variation in appearance among the image ensemble such as the facial boundary, whereas other features such as eyes are neglected. Hence, the landmark estimation on larger patches is often coarse, but provides a good initialization for the smaller patches. As the process progresses, finer details of the target object are revealed and aligned. As a result, the estimate of the landmark locations becomes more and more precise.

Algorithm 2. Landmark annotation by partition

Input: $\mathbf{I}, \widetilde{\mathbf{I}}, \mathbf{S}_{init}, \tilde{\mathbf{S}}, size_{init}$
Output: $\mathbf{S}$

$l_{min} \leftarrow$ minimum number of landmarks in a patch;
$d \leftarrow 1$;
$\mathbf{S} = \mathbf{S}_{init}$;
Compute $\mathbf{x}^1$ and $\mathbf{x}_s^1$ from $\tilde{\mathbf{S}}$; $\mathbf{x}_s \leftarrow \mathbf{x}_s^1$;
while $d <$ maximum number of patches **do**
  if $d = 1$ **then**
    Calculate $\mathbf{P}_{init}^1$ and $\tilde{\mathbf{P}}^1$ from $(\mathbf{S}_{init}^1, \mathbf{x}_s^1)$ and $(\tilde{\mathbf{S}}, \mathbf{x}_s^1)$, respectively;
    $\mathbf{P}^1, \mathbf{S}^1 \leftarrow SSLSC(\mathbf{I}, \widetilde{\mathbf{I}}, \mathbf{P}_{init}^1, \tilde{\mathbf{P}}^1, \mathbf{x}^1, \mathbf{x}_s^1)$; // for the alignment on the whole common mean shape based on Eq. (7);
  else
    $\hat{\mathbf{S}} = \mathbf{S} \setminus \mathbf{S}_{init}^d \bigcup \mathbf{S}_{init}^{d+1}$; // Obtain the landmarks that will not be updated in this congealing round;
    Calculate $\mathbf{P}_{init}^d$ and $\tilde{\mathbf{P}}^d$ from $(\mathbf{S}_{init}^d, \mathbf{x}_s^d)$ and $\left(\tilde{\mathbf{S}}, \mathbf{x}_s^d\right)$, respectively;
    Calculate $\mathbf{P}_{init}^{d+1}$ and $\tilde{\mathbf{P}}^{d+1}$ from $\left(\mathbf{S}_{init}^{d+1}, \mathbf{x}_s^{d+1}\right)$ and $\left(\tilde{\mathbf{S}}, \mathbf{x}_s^{d+1}\right)$, respectively;
    $\mathbf{P}^d, \mathbf{S}^d, \mathbf{P}^{d+1}, \mathbf{S}^{d+1} \leftarrow SSLSC2$
    $\left(\mathbf{I}, \widetilde{\mathbf{I}}, \mathbf{P}_{init}^d, \tilde{\mathbf{P}}^d, \mathbf{x}^d, \mathbf{x}_s^d, \mathbf{P}_{init}^{d+1}, \tilde{\mathbf{P}}^{d+1}, \mathbf{x}^{d+1}, \mathbf{x}_s^{d+1}\right)$; // for the simultaneous alignment on the two child patches based on Eq. (12);
  end if

  Compute $\frac{\partial\epsilon^l}{\partial\mathbf{P}}$ for each pixel $l \in \mathbf{x}_s$;
  if d=1 **then**
  $size \leftarrow size * 2/3$; //reduce the patch size by a ratio of 2/3;
  end if
  Find a region $\mathbf{x}^{k*}$ with maximum $\frac{1}{L_r}\left\|\sum_{l=1}^{L_r}\frac{\partial\epsilon^l}{\partial\mathbf{P}}\right\|^2$;
  if $\mathbf{x}^{k*} = \mathbf{x}^d$ **then**
  $size \leftarrow size * 2/3$; //reduce the patch size by a ratio of 2/3 if no new patch can be found at current patch size;
    Find a region $\mathbf{x}^{k*}$ with maximum $\frac{1}{L_r}\left\|\sum_{l=1}^{L_r}\frac{\partial\epsilon^l}{\partial\mathbf{P}}\right\|^2$;
  end if
  for each neighboring region of $\mathbf{x}^{k*}$ **do**
    Find a region $\mathbf{x}^{k**}$ with maximum $\frac{1}{L_r}\left\|\sum_{l=1}^{L_r}\frac{\partial\epsilon^l}{\partial\mathbf{P}}\right\|^2$;
  end for

  $\mathbf{S} = \hat{\mathbf{S}} \bigcup \mathbf{S}^d \bigcup \mathbf{S}^{d+1}$;
  $\mathbf{x}^{d+2} \leftarrow \mathbf{x}^{k*}$;$\mathbf{x}^{d+3} \leftarrow \mathbf{x}^{k**}$
  Calculate $\mathbf{x}_s^{d+2}$ from $\mathbf{x}_s$ and $\mathbf{x}^{d+2}$;
  Calculate $\mathbf{x}_s^{d+3}$ from $\mathbf{x}_s$ and $\mathbf{x}^{d+3}$;
  Calculate $\mathbf{S}_{init}^{d+2}$ from $\mathbf{x}^{d+2}$ and $\mathbf{S}$;
  Calculate $\mathbf{S}_{init}^{d+3}$ from $\mathbf{x}^{d+3}$ and $\mathbf{S}$;
  if $size(\mathbf{x}_s^{d+2}) < l_{min}$ **or** $size(\mathbf{x}_s^{d+3}) < l_{min}$ **then**
    return $\mathbf{S}$;
  end if
  $d \leftarrow d + 2$
end while

6. Experiments

In order to demonstrate the effectiveness of the proposed algorithm, we have performed extensive validation studies for

Table 2
Performance comparison of patch selection strategies: brute-force partitioning method [10] and the proposed greedy patch selection method in terms of accuracy (NRMSE), robustness (SOF), efficiency (time), and number of congealing rounds. $\tilde{K} = 1$ annotated images and $K = 200$ unannotated images from the first data set are used with the noise level $\eta_{max} = 30$.

	NRMSE (%)	SOF (%)	Time (s)	Congealing rounds
Tong, et al. [10]	7.05	1.5	3683	5
Greedy patch selection method	6.83	1.5	3262	7

Table 3
Performance comparison of SLSC and shape-constrained SLSC in terms of accuracy (NRMSE) and robustness (SOF) under varying noise levels. $\tilde{K} = 1$ annotated images and $K = 200$ unannotated images from the first data set are used.

	Noise Level					
	$\eta_{max} = 10$		$\eta_{max} = 20$		$\eta_{max} = 30$	
	NRMSE (%)	SOF (%)	NRMSE (%)	SOF (%)	NRMSE (%)	SOF (%)
SLSC	9.22	3.4	9.81	11.2	9.84	27.1
Shape constrained SLSC	9.16	1.7	9.69	9.3	9.83	22.9

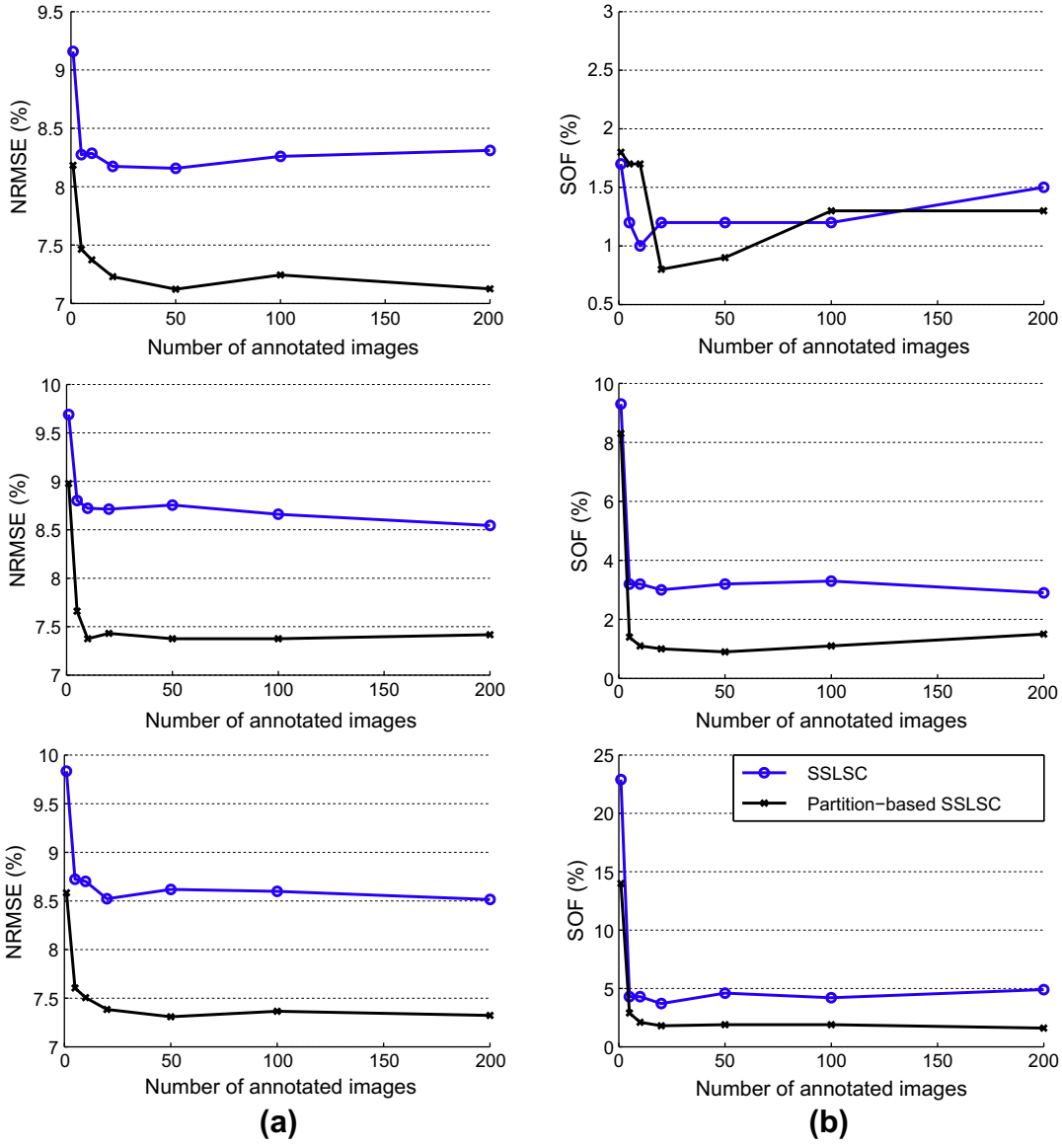

Fig. 5. Performance comparison for SSLSC (blue line with circles) and patch-based SSLSC (black line with crosses) in terms of (a) accuracy (NRMSE of landmarks excluding outliers) and (b) robustness (SOF). The results in each row correspond to a noise level ($\eta_{max} = 10$, 20, and 30 from top to bottom), respectively. The performance is evaluated on the first data set by varying the number of annotated images $\tilde{K}$ and the noise level η_{max} from an average of 5 random trials, where 200 unannotated images are used.

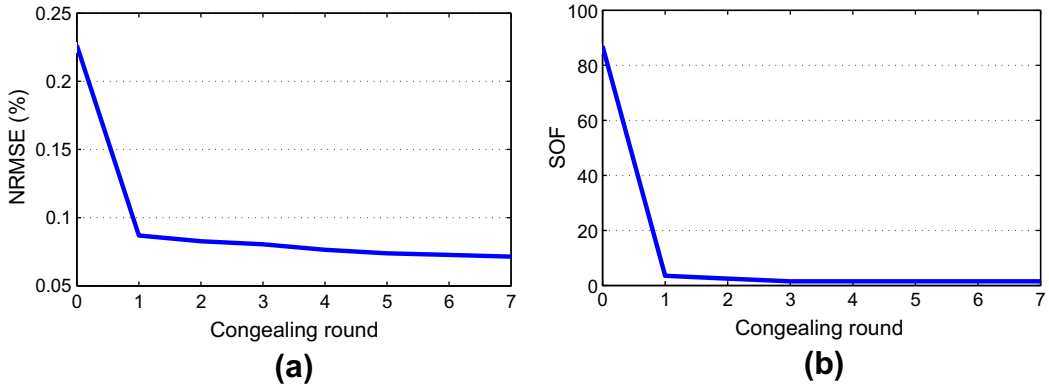

Fig. 6. Performance analysis by varying congealing rounds in terms of (a) accuracy (NRMSE) and (b) robustness (SOF) using $\tilde{K} = 1$ annotated images and $K = 200$ unannotated images with the noise level $\eta_{max} = 30$ from the first data set. The results of round-0 correspond to the initialization, and those of round-1 represent the congealing results on the whole common mean shape by SSLSC.

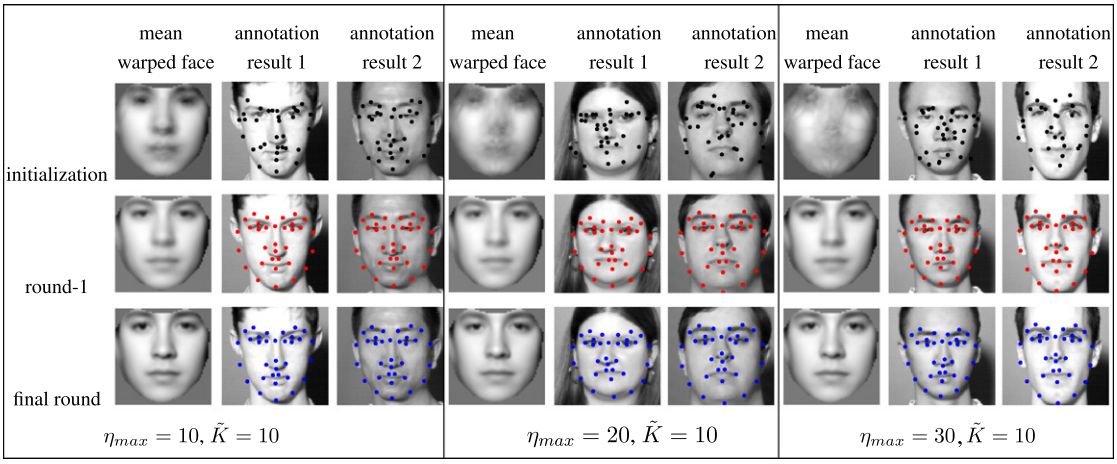

Fig. 7. Landmark annotation results for three noise levels. For each cell, the three rows illustrate the initial landmark positions, the round-1 results, and the final annotation results, respectively. In each row, the mean warped face region and two example images are shown.

the application of annotating facial landmarks. It is desirable to find results from a previous state-of-art approach to compare with ours. A fair comparison with supervised methods such as AAM [9] would be to train AAM using the same set of annotated images provided to our approach. In contrast, we aim to automatically annotate a set of specific landmarks around the facial features (i.e., eyes, eyebrows, nose, mouth, and facial boundary) for a large image set, given a few annotated example images. To the best of our knowledge, there is little prior work addressing this challenging problem, so we compare with [10], which uses a brute-force partitioning strategy for patch selection, [8], which is similar to our SLSC algorithm without using the annotated images, and an AAM-based supervised method [9].

For the experiments, we employed three test data sets: the first data set contains 400 images from the Notre Dame (ND1) database [33] with more than 100 subjects; the second data set consists of 255 images from Caltech 101 face database [34] and 50 images from the ND1 database; and the third data set contains 1176 images from FERET database [35] and ND1 database. We manually annotated 33 landmarks for each image in the first and the third data sets to establish a ground truth and to enable a quantitative evaluation for the automatic annotation performance. The experiments on the first data set aim to quantitatively evaluate the proposed annotation algorithm on various aspects including the robustness to the noise in initialization and the parameter setting, and to enable a quantitative comparison with other aforementioned methods [10,8]. The experiments on the other two data sets intend to demonstrate that the proposed annotation algorithm can generalize to a large population and deal with real-world challenges such as cluttered background and various illuminations (the second data set), and high dimensionality of the image ensemble (the third data set). Furthermore, the experiment on the third data set also provides a quantitative comparison with the supervised method [9] with same amount of training data.

Throughout the experiments, a six-parameter affine transformation is employed in each of the SLSC processes; and thus, each patch has an individual warping parameter vector with six elements in the partition-based SSLSC. To accommodate illumination changes, the warped face region undergoes a common normalization procedure, where we subtract the mean intensity, then divide by the standard deviation of intensity.

Our algorithm performance is evaluated by two criteria: (1) Normalized Root Mean Squared Error (NRMSE) of landmarks defined as the RMSE w.r.t. the ground truth divided by the eye-to-eye distance ρ_i, and expressed as a percentage; and (2) Sample "Outliers" Fraction (SOF) defined as the number of images, of which the NRMSE exceeds a threshold (10%), versus the number of unannotated images. A smaller NRMSE indicates a higher annotation accuracy, and a smaller SOF represents greater robustness.

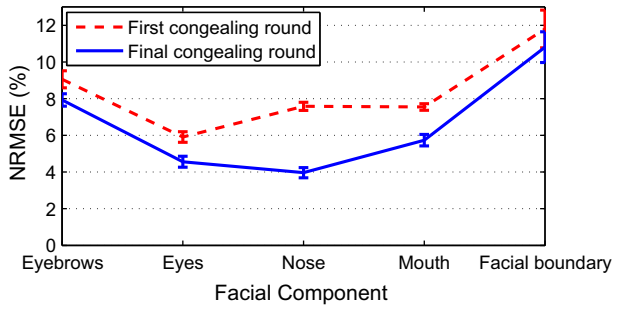

Fig. 8. Performance of facial components at the first and final congealing round in terms of accuracy (NRMSE) using $\widetilde{K} = 10$ annotated images and $K = 200$ unannotated images with the noise level $\eta_{max} = 20$ from the first data set. The center and the length of the error bar represent the mean and standard deviation of 5 random trials, respectively.

6.1. Experimental evaluation on the first data set

In the following, we will demonstrate that with only a few annotated images, robust and accurate landmark annotation can be obtained with the proposed algorithm. The images in the first data set are scaled to the same size (256×256) based on the ground-truth eye positions such that the face region of each image has nearly the same size and is located at approximately same position. We then divide the 400 annotated images into two non-overlapping sets: an annotated set with $\widetilde{K}$ images and an unannotated set with 200 images. A 72×72 square region, which encloses all the target landmarks, is used as the common mean shape in the experiments.

For quantitative evaluation, the initial value of the j^{th} element of $\mathbf{s}_i$ is generated by adding uniformly distributed random noise $\eta \in [-\eta_{max}, \eta_{max}]$ to the ground-truth value $\hat{\mathbf{s}}_{i,j}$ as follows,

$$\mathbf{s}_{i,j} = \hat{\mathbf{s}}_{i,j} + \frac{\eta \rho_i}{\bar{\rho}}, \tag{22}$$

where ρ_i is the interocular pixel distance of $\mathbf{I}_i$, and $\bar{\rho}$ is the average of ρ_i for all unannotated images ($\bar{\rho} \approx 105$ pixels in our experiments). By doing so, the level of deviation in the initialization is proportional to the interocular distance and face size.

In practical applications, the initial landmark positions can be obtained by placing a configuration of landmarks within a rectangular box returned from a face detector[4]. Hence, the proposed algorithm is fully automatic given a few annotated samples. This initialization strategy is employed in the experiments of the second and third data sets.

6.1.1. Evaluation on patch selection strategy

Table 2 compares the annotation performance by using the brute-force partitioning method as in [10] and the greedy patch selection method described in Section 5 in terms of accuracy (NRMSE), robustness (SOF), efficiency (computational complexity), and number of congealing rounds when $\widetilde{K} = 1$ and $\eta_{max} = 30$. We ensure that both algorithms are compared under the same condition and only differ in their patch selection strategies. For example, they use the *same* randomly selected annotated set and the *same* initialization. Note that, for this result, outliers are excluded from the computation of NRMSE. It can be seen that the greedy patch selection method slightly improves the annotation accuracy with the same performance of robustness. Furthermore, the greedy patch selection method is more efficient even with more congealing rounds, which is extremely important for annotating large image ensembles. In the remaining experiments we use the greedy patch selection method with $\beta = 0.225$ determined empirically.

6.1.2. Results of shape-constrained SLSC

Table 3 shows the comparison of SLSC and SSLSC under the effects of different noise levels $\eta_{max} \in \{10,20,30\}$, in terms of accuracy (NRMSE) and robustness (SOF). Note that, for this result, outliers are excluded from the computation of NRMSE. The results are computed from an average of five trials, where $\widetilde{K} = 1$ annotated images are randomly selected as the annotated set for each trial. Again we ensure that both algorithms are compared under the same conditions.

Comparing the results of SLSC and SSLSC in Table 3, we see that the shape constraints are effective in reducing the outliers significantly, especially when the congealing performance of SLSC is poor due to high initialization noise and a small number of annotated images. For example, the SOF decreases from 27.1% (SLSC) to 22.9% (SSLSC) with $\widetilde{K} = 1$ and $\eta_{max} = 30$, which is equivalent to removing 8.4 outliers. Since the shape constraints are not applied on those low-error estimations, there is no obvious improvement in the NRMSE excluding outliers.

6.1.3. Results of patch-based SSLSC

In this experiment, we demonstrate the improvement of annotation accuracy by patch-based SSLSC. Fig. 5 shows the comparison of SSLSC and patch-based SLSC under the effects of varying number of annotated images $\widetilde{K} \in \{1, 5, 10, 20, 50, 100, 200\}$ and different noise levels $\eta_{max} \in \{10,20,30\}$, in terms of NRMSE and SOF. Similar to the previous experiment, outliers are excluded from the computation of NRMSE and the performance is evaluated from an average of 5 random trials.

Comparing the results of SSLSC (blue[5] line with circles) and patch-based SSLSC (black line with crosses) in Fig. 5, it is clear that the patch-based approach further improves both precision and robustness in terms of reducing the NRMSE and SOF. For example, the SOF decreases from 22.9% (45.8 outliers) with SSLSC to 14% (28) with patch-based SSLSC, and the NRMSE decreases from 9.83% (≈10.32 pixels) using SSLSC to 8.58% (≈9.01 pixels) using patch-based SSLSC with $\widetilde{K} = 1$ and $\eta_{max} = 30$. In summary, an average of 1.1% (≈1.16 pixels) reduction of NRMSE is achieved for all noise levels, and an average of 3.2% (6.4 outliers) decrease of SOF is obtained for $\eta_{max} = 30$. From Fig. 5, we can see that there is no remarkable improvement when $\widetilde{K} \geqslant 20$, which means that even with only 9% (20/220) of the data manually annotated, we can estimate the landmark locations accurately and robustly.

Fig. 6 illustrates the performance improvement across different congealing rounds when $\widetilde{K} = 1$ and $\eta_{max} = 30$. The results of round-0 correspond to the initialization, and those of round-1 represent the congealing results on the whole common mean shape by SSLSC. We can see that as the number of congealing rounds increases, both the NRMSE and SOF decrease and converge at the last congealing round.

In Fig. 7, we also show exemplar annotation results under three initialization noise levels, respectively. To compare the overall annotation performance, a mean warped face region[6] is also displayed. It can be observed that the first round of the congealing (middle row) can roughly localize the landmarks, but fail to handle subtle facial appearance changes caused by individual difference, facial expression, and face pose. These round-1 results are basically the performance of the previous LSC approach [8] with additional shape constraints and annotated images. In contrast, the landmark annota-

[4] $\eta_{max} \approx 15$ when a commercial face detector developed by Pittsburgh Pattern Recognition [36] is used in our experiments.

[5] For interpretation of color in Figs. 1–14, the reader is referred to the web version of this article.

[6] Here, the mean warped face region is different from the common mean face $\mathbf{x}$ used in the congealing process. It is generated by piecewise affine warp in a triangular face mesh based on the landmark positions and only used for visualization purpose.

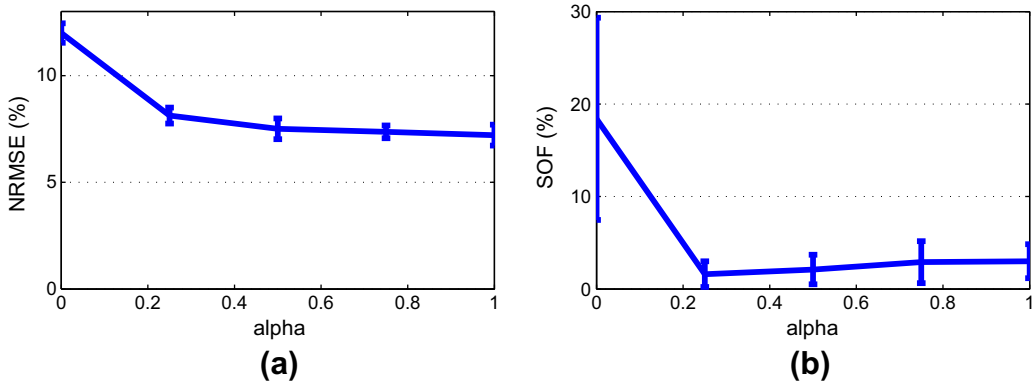

Fig. 9. Performance analysis by varying α in terms of (a) accuracy (NRMSE) and (b) robustness (SOF) using $\widetilde{K} = 10$ annotated images and $K = 200$ unannotated images with the noise level $\eta_{max} = 30$ from the first data set. $\alpha = 0$ implies an unsupervised congealing, and $\alpha = 1$ is supervised.

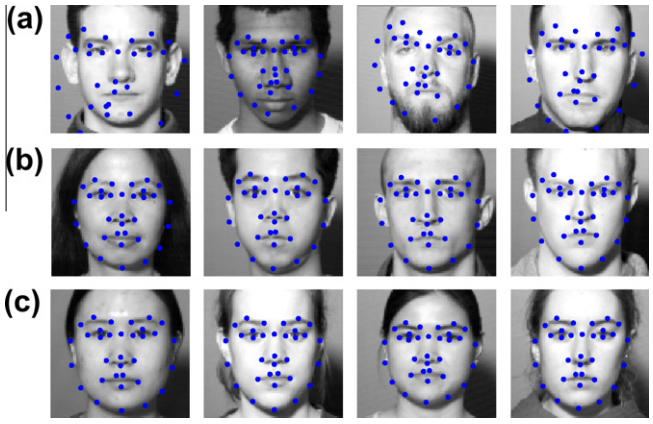

Fig. 10. Annotated images with the (a) lowest, (b) median, and (c) highest annotation confidences using $\widetilde{K} = 10$ annotated images and $K = 200$ unannotated images with the noise level $\eta_{max} = 30$ from the first data set.

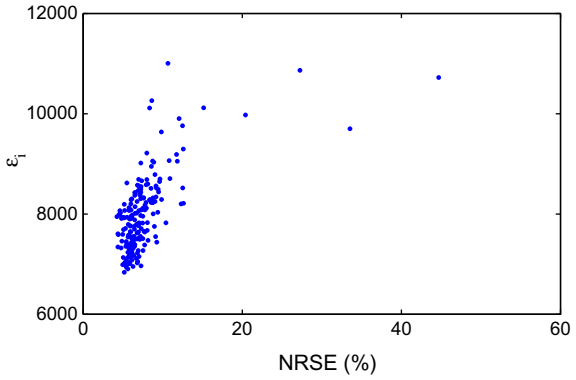

Fig. 11. Annotation confidence (ε_i) versus landmark annotation error (NRSE) using $\widetilde{K} = 10$ annotated images and $K = 200$ unannotated images with the noise level $\eta_{max} = 30$ from the first data set. The Pearson correlation coefficient between these two variables is 0.632.

tion in the final congealing round (last row) show a significant improvement of accuracy under slight changes in face pose (compare the noses and mouths in the 3rd, 5th, and 6th images of the last two rows) as well as individual differences (compare the noses and mouths in the 2nd, 8th and 9th images of the last two rows). This again shows the accuracy of our approach when compared to [8].

6.1.4. Analysis on different facial components

From Fig. 7, we see that the mean warped face regions in the last row have a much sharper appearance around the nose and mouth, compared with those in the second row. We can even see the philtrum clearly in the last row. However, the improvement around the facial boundary is not obvious. Therefore, we have conducted an analysis to study performance improvement using the patch-based approach for different facial components.

Using the case $\widetilde{K} = 10$ and $\eta_{max} = 20$ as an example, Fig. 8 shows that the patch-based algorithm improves the annotation accuracy on the different facial components at various degrees. On the one hand, the improvement on the facial boundary is the smallest among the facial components (a 0.99% reduction in terms of NRMSE), since the facial boundary has the most significant variations across images and attracts the most attention in the first round of congealing. On the other hand, the improvement on the inner-face components such as nose and mouth are significant, e.g., a 3.62% reduction for the nose and a 1.8% reduction for the mouth, since these inner-face components are involved in more congealing rounds as shown in Fig. 3. As the number of congealing rounds increases, more details of these facial components are revealed and contribute to the congealing process. This property of the proposed algorithm is valuable for practical applications, since an accurate estimation of landmarks for the inner-face component (eyes, eyebrows, nose, and mouth) are extremely useful for many applications such as face recognition and facial expression analysis.

6.1.5. Analysis of the weighting coefficient

Besides studying the effect of noise level and the number of annotated images on the congealing performance, we also analyze how the selection of the weighting coefficient α in Eq. (5) affects performance of the proposed algorithm. The larger α is, the more the algorithm relies on the annotated data.

In the extreme case, $\alpha = 0$ implies an unsupervised congealing, and $\alpha = 1$ is supervised. Fig. 9 illustrates the performance with various values of α with $\widetilde{K} = 10$ and $\eta_{max} = 30$. Although using a large α improves the accuracy of the algorithm, it tends to result in more outliers, especially with only a few annotated samples. This is because the shape/appearance variations in the large image ensemble cannot be well represented by only a small number of annotated samples. A good trade-off can be achieved by balancing the weights of the annotated and unannotated data. We used $\alpha = 0.5$ in the experiments reported in the previous discussions.

6.1.6. Annotation confidence

Often in computer vision, knowing when the algorithm fails is as important as how the algorithm performs. Hence, a confidence score is desirable for practical applications in order to evaluate the quality of annotation without ground truth. For this we use ε_i

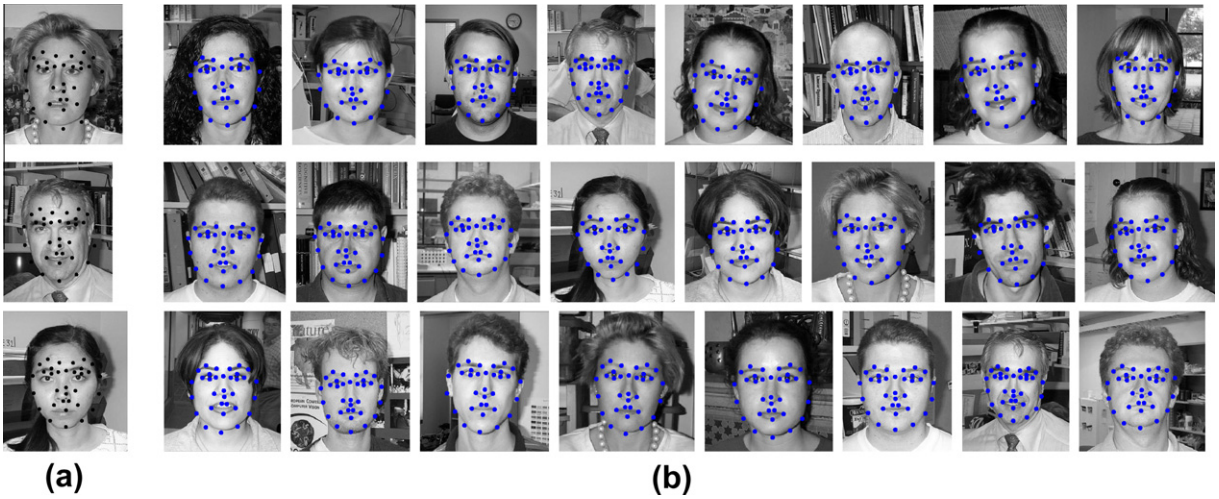

Fig. 12. Annotation results on the second data set: (a) the initialization of our algorithm and (b) the landmark annotation results. A face detector [36] is employed to detect the face region for initialization.

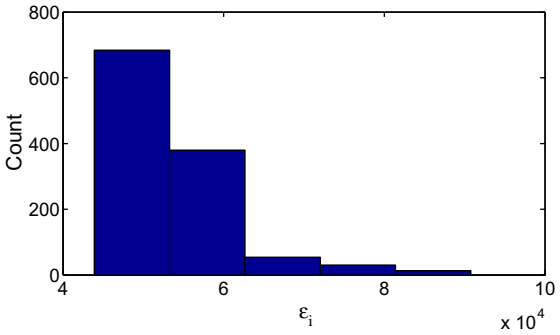

Fig. 13. Histogram of ε_i for the third data set.

in Eq. (5). A smaller ε_i indicates a higher-confidence in annotation. Fig. 10 shows annotated images with the lowest, median, and highest confidence scores, in an image ensemble ($\widetilde{K} = 10$ and $\eta_{max} = 30$), where the annotation is improving from the top to bottom row. Fig. 11 also illustrates the distribution of the estimated ε_i versus the real landmark annotation error represented by the Normalized Root Squared Error (NRSE) of landmarks. With the increase of the ε_i, the landmark annotation error increases significantly. Hence, it is clear that the confidence score is indicative of annotation performance. The linear correlation between ε_i and NRSE can also be shown by the computed Pearson correlation coefficient between them, 0.632. Similar phenomena have been observed for experiments on the other two data sets. In practice, after annotation, one can use this confidence score to select only well-anno-

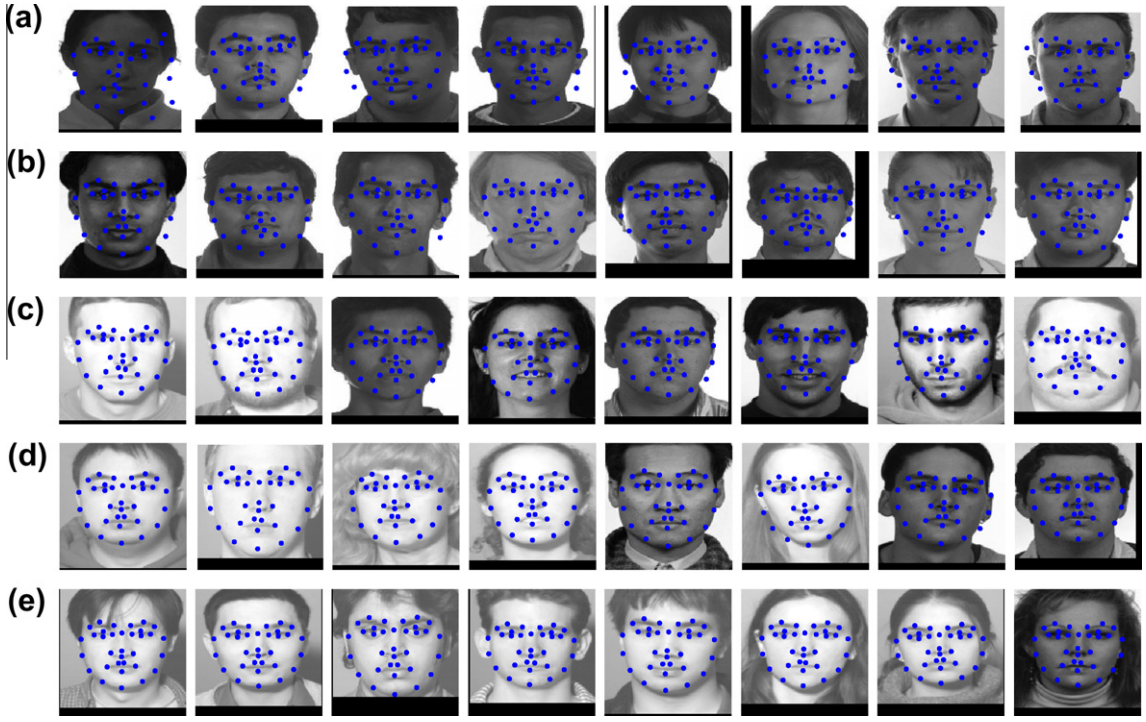

Fig. 14. Exemplar annotated images of the third data set at 5 annotation confidence levels corresponding to 5 bins in Fig. 13: ε_i decreases from the top to the bottom row. $\widetilde{K} = 15$ annotated images and $K = 1161$ unannotated images from the third data set are used with the detected face region for initialization.

Table 4
Performance comparison with supervised method [9] using $\widetilde{K} = 15$ annotated images and $K = 1161$ unannotated images from the third data set. A face detector [36] is employed to detect the face region for initialization.

	NRMSE (%)	SOF (%)
Method of [9]	6.94	17.09
Proposed method	7.63	0.85

tated samples for a training set, or to select samples for other appropriate additional processing.

6.2. Cross-database validation

Here, using the second and the third data sets, we will demonstrate the generalization capability of the proposed algorithm when dealing with an unseen imaging environment. It can be assumed that a face detector can perform reasonably well for the imaging conditions we are dealing with (i.e., the face occupies a large portion of the image and the face pose variation is moderate). For initialization of these two data sets, we first obtain the location of the face with a commercial face detector developed by Pittsburgh Pattern Recognition [36]. Then, as shown in Fig. 12a, the initial positions of landmarks are generated by adapting the target landmark coordinates on the common mean shape ($\mathbf{x}_s$) to the detected face region. Therefore, given a few annotated images, we can perform the landmark annotation in an automatic fashion.

For the second data set, we aim to automatically annotate the 33 landmarks on 255 images from Caltech 101 face database [34] with the help of 50 manually annotated images from the ND1 database. Since we do not have ground-truth landmark positions for this database, we can only perform qualitative evaluation by visual observation. Fig. 12b illustrates sample annotation results. Although the congealing on the Caltech 101 images is much more difficult than that of the first data set (ND1 images) due to cluttered backgrounds and challenging illumination conditions, our algorithm can still achieve satisfactory annotation performance.

Furthermore, we also obtain excellent annotation results on the third data set. We have manually annotated 33 landmarks on this combined database such that we can perform quantitative evaluation on the database.

As shown in Table 4, a 7.63% NRMSE of landmarks (excluding outliers) with 0.85% of SOF (10 outliers) is achieved for 1176 total images with only 15 (=1.3%) annotated images. Compared with the supervised method [9], which was trained using the same set of data and tested with the same initialization, the proposed method achieves similar annotation accuracy, and more importantly improves the robustness dramatically: the SOF decreases about 16.24%, equivalent to reducing 191 outliers. This further demonstrates that our system can accommodate a vast amount of data, without noticeable sacrifice in performance; while the conventional supervised alignment algorithms cannot handle the case with very few annotated training data. Fig. 13 shows the distribution of estimated ε_i after convergence for the third data set. We can see that the majority of the annotated images has high annotation confidence (low value of ε_i). Fig. 14 also gives some samples of annotated images selected from each of the five bins in Fig. 13.

Even though the main purpose of this work is the *one-time* off-line annotation of training data, the efficient run-time is still desirable for practical usage. Our method has very acceptable computational complexity. As shown in Table 2, a patch-based SSLSC experiment to produce 200 annotated images with a 72×72 common mean shape takes less than 1 h using a Matlab™ implementation with a 2 GHz CPU. In comparison, for an experienced human annotator, it takes at least 3 h to annotate 33 landmarks for 200 images. It is expected that the savings in time will be even greater when dealing with a larger ensemble.

7. Conclusions

Shape deformation of images of a real-world object is often non-rigid due to inter-instance variability, object motion, and changing camera view point. Automatically estimating non-rigid deformations for an object class is a critical step in characterizing the object and learning statistical shape models. Our proposed approach facilitates such a task by automatically producing annotated data sets with only a small number of manually annotated examples. Extensive experiments demonstrate that our system has achieved impressive annotation results on face images with nearly frontal view and moderate changes in expression, useful for many practical applications.

There are several future directions in which to extend this framework. First, although we have only applied our approach to facial images, no domain knowledge of faces is used in our work. Hence, the approach can be immediately applied to the task of annotating landmarks in images of other classes of objects such as vehicles, pedestrians, or objects in medical imaging. Second, we expect to extend our methods to handle large facial variations, such as face pose. Third, features such as Histogram of Oriented Gradients (HOG) [2] can also be utilized in our congealing approach in order to achieve improved robustness w.r.t. lighting and color. Finally, in this work, we intend to minimize the summation of the pairwise image difference among the image ensemble caused by deformation of the target region, while the illumination change and appearance variations within the target region are ignored. Since the illumination variation is not modeled in the cost function, the proposed SSLSC algorithm and the annotation confidence for evaluating the annotation performance will be affected by the illumination change especially the skin color variation and a high contrast between the face skin and the background. In the future, we are interested in extending the cost function to include the appearance variation parameters, which leads to a better modeling of the pairwise image difference.

Acknowledgments

This Project was supported by awards #2007-DE-BX-K191 and #2007-MU-CX-K001 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. 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.

References

[1] P. Viola, M. Jones, Robust real-time face detection, Int. J. Comput. Vision 57 (2) (2004) 137–154.
[2] N. Dalal, W. Triggs, Histograms of oriented gradients for human detection, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2005, pp. 886–893.
[3] T. Cootes, G. Edwards, C. Taylor, Active appearance models, IEEE Trans. Pattern Anal. Mach. Intell. 23 (6) (2001) 681–685.
[4] T.F. Cootes, C.J. Taylor, D.H. Cooper, J. Graham, Active shape models—their training and application, Comput. Vision Image Understand. 61 (1) (1995) 38–59.
[5] S. Baker, I. Matthews, Lucas–Kanade 20 years on: a unifying framework, Int. J. Comput. Vision 56 (3) (2004) 221–255.
[6] E. Learned-Miller, N. Matsakis, P. Viola, Learning from one example through shared densities on transforms, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2000, pp. 464–471.
[7] E. Learned-Miller, Data driven image models through continuous joint alignment, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2) (2006) 236–250.
[8] M. Cox, S. Sridharan, S. Lucey, J. Cohn, Least squares congealing for unsupervised alignment of images, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2008.
[9] I. Matthews, S. Baker, Active appearance models revisited, Int. J. Comput. Vision 60 (2) (2004) 135–164.

[10] Y. Tong, X. Liu, F.W. Wheeler, P. Tu, Automatic facial landmark labeling with minimal supervision, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2009.
[11] M. Cox, S. Sridharan, S. Lucey, J. Cohn, Least-squares congealing for large numbers of images, in: Proc. of the Intl. Conf. on Computer Vision (ICCV), 2009.
[12] A. Vedaldi, G. Guidi, S. Soatto, Joint data alignment up to lossy transformations, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2008.
[13] M. Storer, M. Urschler, H. Bischof, Intensity-based congealing for unsupervised joint image alignment, in: Proc. of the International Conference on Pattern Recognition (ICPR), 2010, pp. 1473–1476.
[14] C.R. Shelton, Morphable surface models, Int. J. Comput. Vision 38 (1) (2000) 75–91.
[15] S. Balci, P. Golland, M. Shenton, W. Wells, Free-form b-spline deformation model for groupwise registration, in: Proc. of Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI), 2007, pp. 23–30.
[16] T. Vetter, M.J. Jones, T. Poggio, A bootstrapping algorithm for learning linear models of object classes, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 1997, pp. 40–46.
[17] A. Guimond, J. Meunier, J.-P. Thirion, Automatic computation of average brain models, in: Proc. of Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI), 1998, pp. 631–640.
[18] S. Baker, I. Matthews, J. Schneider, Automatic construction of active appearance models as an image coding problem, IEEE Trans. Pattern Anal. Mach. Intell. 26 (10) (2004) 1380–1384.
[19] I. Kokkinos, A. Yuille, Unsupervised learning of object deformation models, in: Proc. of the Intl. Conf. on Computer Vision (ICCV), 2007.
[20] T. Cootes, S. Marsland, C. Twining, K. Smith, C. Taylor, Groupwise Diffeomorphic Non-rigid Registration for Automatic Model Building 4 (2004) 316–327.
[21] T. Cootes, C. Twining, V. Petrovic, R. Schestowitz, C. Taylor, Groupwise construction of appearance models using piece-wise affine deformations, in: Proc. of the British Machine Vision Conference (BMVC), vol. 2, 2005, pp. 879–888.
[22] D. Cristinacce, T. Cootes, Facial motion analysis using clustered shortest path tree registration, in: Proc. of the 1st Intl. Workshop on Machine Learning for Vision-based Motion Analysis with ECCV, 2008.
[23] T. Cootes, C. Twining, V. Petrovic, K. Babalola, C. Tylor, Computing accurate correspondences across groups of images, IEEE Trans. Pattern Anal. Mach. Intell. 32 (11) (2010) 1994–2005.
[24] K. Sidorov, S. Richmond, D. Marshall, An efficient stochastic approach to groupwise non-rigid image registration, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2009.
[25] F. Torre, M. Nguyen, Parameterized kernel principal component analysis: theory and applications to supervised and unsupervised image alignment, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2008.
[26] G. Langs, R. Donner, P. Peloschek, B. Horst, Robust autonomous model learning from 2D and 3D data sets, in: Proc. of Int. Conf. on Medical Image Computing and Computer Assisted Intervention (MICCAI), vol. 1, 2007, pp. 968–976.
[27] A. Asthana, R. Goecke, N. Quadrianto, T. Gedeon, Learning based automatic face annotation for arbitrary poses and expressions from frontal images only, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2009, pp. 1635–1642.
[28] T. Cootes, Timeline of developments in algorithms for finding correspondences across sets of shapes and images, Tech. Rep., University of Manchester, 2005.
[29] X. Liu, Discriminative face alignment, IEEE Trans. Pattern Anal. Mach. Intell. 31 (11) (2009) 1941–1954.
[30] R. Gross, I. Matthews, S. Baker, Generic vs. person specific active appearance models, J. Image Vision Comput. 23 (11) (2005) 1080–1093.
[31] P. Felzenszwalb, D. McAllester, D. Ramanan, A discriminatively trained, multiscale, deformable part model, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), 2008.
[32] P. Viola, M. Jones, Rapid object detection using a boosted cascade of simple features, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 1, 2001, pp. 511–518.
[33] K.I. Chang, K.W. Bowyer, P.J. Flynn, An evaluation of multi-modal 2D + 3D face biometrics, IEEE Trans. Pattern Anal. Mach. Intell. 27 (4) (2005) 619–624.
[34] L. Fei-Fei, R. Fergus, P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, Comput. Vision Image Understand. 106 (1) (2007) 59–70.
[35] P.J. Phillips, H. Moon, P.J. Rauss, S. Rizvi, The FERET evaluation methodology for face recognition algorithms, IEEE Trans. Pattern Anal. Mach. Intell. 22 (10) (2000) 1090–1104.
[36] H. Schneiderman, Feature-centric evaluation for efficient cascaded object detection, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), vol. 2, IEEE, 2004, pp. 29–36.