# Automatic Facial Landmark Labeling with Minimal Supervision *

Yan Tong Xiaoming Liu Frederick W. Wheeler Peter Tu
Visualization and Computer Vision Lab
GE Global Research, Niskayuna, NY 12309
{tongyan,liux,wheeler,tu}@research.ge.com

## Abstract

*Landmark labeling of training images is essential for many learning tasks in computer vision, such as object detection, tracking, and alignment. Image labeling is typically conducted manually, which is both labor-intensive and error-prone. To improve this process, this paper proposes a new approach to estimate a set of landmarks for a large image ensemble with only a small number of manually labeled images from the ensemble. Our approach, named semi-supervised least-squares congealing, aims to minimize an objective function defined on both labeled and unlabeled images. A shape model is learnt on-line to constrain the landmark configuration. We also employ a partitioning strategy to allow coarse-to-fine landmark estimation. Extensive experiments on facial images show that our approach can reliably and accurately label landmarks for a large image ensemble starting from a small number of manually labeled images, under various challenging scenarios.*

## 1. Introduction

Image labeling for training data is an essential step in many learning-based vision tasks. There are at least two types of prior knowledge represented by image labeling. One is *semantic* knowledge, such as person IDs for face recognition, or an object's name for content-based image retrieval. The other is *geometric/landmark* knowledge. In learning-based object detection [25, 12], for example, the position of the object (face/pedestrian/car) needs to be labeled for all training images. For supervised face alignment [6], each training image must be labeled with a set of landmarks, which describe the shape of the face.

This paper focuses on geometric/landmark knowledge labeling, which is typically carried out *manually*. Practical applications, such as object detection, often require thousands of labeled images to achieve sufficient generalization capability. However, manual labeling is labor-intensive and time-consuming. Furthermore, image labeling is an error-prone process due to labeler error, imperfect description of the objectives, and inconsistencies among different labelers.

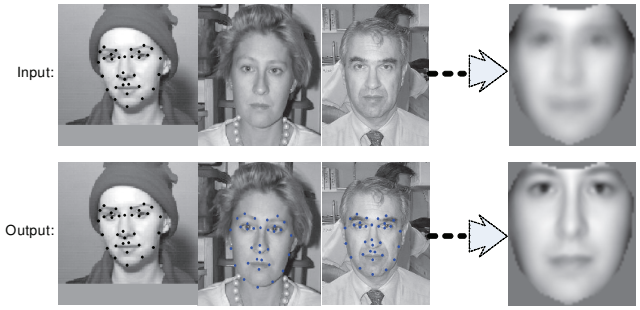



Figure 1. Our approach takes an image ensemble as input with manually labeled landmark positions for only a small subset, and automatically estimates the landmarks for the remaining images such that the non-rigid shape deformation is discovered. Note the improved sharpness in the average warped face (the last column), an indicator of accurate landmark estimation by our algorithm.

To alleviate these problems, this paper presents an approach to automatically provide landmark labeling for a large set of images in a semi-supervised fashion. That is, using manually labeled landmark locations for a few images, our approach can automatically estimate the landmark locations for the full set of images (see Fig. 1). In one example, we will demonstrate that 10 manually labeled images may be used to label the complete training set of 300 images. 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 utilized: the distance between the labeled and unlabeled images, and the distance between the unlabeled images. The objective function is iteratively minimized via the well-known and efficient inverse warping technique [1]. During the optimization process, we also constrain the estimated landmark locations by utilizing shape statistics learnt from the relatively *low-error* estimations in an on-line manner, which is shown to result in better convergence of land-

*This work 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.

mark position estimates.

Prior work on joint alignment for an image ensemble [20, 17, 10] mainly 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 toward accurately characterizing the shape variation within an object class. Motivated by this, we propose a hierarchical patch-based approach to estimate landmark positions. Starting from the whole face region, treated as the first level patch, we iteratively partition patches into smaller child patches, in which initial landmark locations are obtained from the parent patch and whose refined landmark estimations result in an accurate landmark labeling based on the local patch appearance. Applications in facial images show that by labeling only 3% of the ensemble, the landmarks of the remaining images can be estimated accurately.

The proposed automatic image labeling framework has three main contributions:

◇ A core algorithm is proposed for semi-supervised least-squares-based alignment of an image ensemble. We describe its efficient implementation using the inverse warping technique [1] and provide computational analysis.

◇ Two additional techniques are introduced for improving landmark estimation. One is to use a statistical shape model learnt on-line to reduce outliers among the ensemble. The other is to utilize patch-based partitioning to improve the precision of landmark estimation.

◇ 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 labeled images. Extensive experiments to evaluate the performance and capabilities of the system have been conducted and are reported here.

## 2. Prior Work

In some notable and early work on *unsupervised* joint alignment, Learned-Miller [20, 17] denotes the process as "congealing". The underlying idea is to minimize an entropy-based cost function by estimating the warping parameter of an ensemble. More recently, Cox et al. [10] propose a least-squares congealing (LSC) algorithm, which uses $L_2$ constraints to estimate each warping parameter. However, both approaches estimate 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 the 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. Balci et al. [3] extend the Learned-Miller's method [20] by including a free-form B-spline deformation model. Vetter et al. [24] have developed a bootstrapping algorithm to compute image correspondences and to learn a linear model based on optical flow. Baker et al. [2] 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 [16] formulate AAM learning as an EM algorithm and extend it to learning parts-based models for flexible objects. Cootes et al. [7, 9, 11] use a group-wise objective function to compute non-rigid registration. Torre and Nguyen [23] improve manual facial landmark labeling based on parameterized kernel PCA. Langs et al. [14] employ an MDL-based cost function and estimate the correspondences for a set of control points. Saragih and Goecke [22] tackle the alignment problem by tracking the image sequence with an adaptive template.

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 can be incorporated easily via a few manually labeled 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. Other approaches on this topic are described by Cootes [5].

In contrast, there is a sizable literature for *supervised* face alignment, including Active Shape Model [8], AAM [6, 19], Boosted Appearance Model [18]. Generally, a large number of labeled training images are required to train a statistical model so that it can generalize and fit unseen images well [15]. Hence, we are motivated to develop this semi-supervised approach to produce this training data more easily.

## 3. Semi-supervised Least-Squares Congealing

Given an ensemble of images, where a few manually labeled images have known warping parameters, our SLSC approach aims to estimate the warping parameters of the remaining unlabeled images using the cost function:

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

where $\tilde{K}$ is the number of labeled images $\tilde{\mathbf{I}} = \{\tilde{\mathbf{I}}_n\}_{n\in[1,\tilde{K}]}$, and $K$ is the number of unlabeled images $\mathbf{I} = \{\mathbf{I}_i\}_{i\in[1,K]}$. $\mathbf{p}_i$ is an $m$-dimensional warping parameter vector to warp $\mathbf{I}_i$ to a common mean shape using a warping function $\mathbf{W}(\mathbf{x};\mathbf{p}_i)$, which can be a simple affine warp or a complex non-rigid warp such as the piecewise affine warp [19]. $\mathbf{I}_i(\mathbf{W}(\mathbf{x};\mathbf{p}_i))$ is the corresponding $N$-dimensional warped image vector. $\tilde{\mathbf{p}}_n$ is the known warping parameter vector

for $\tilde{\mathbf{I}}_n$. $\mathbf{x}$ is a collection of $N$ pixel coordinates within the mean shape. $\mathbf{P} = [\mathbf{p}_1, \cdots, \mathbf{p}_K]$ contains all the warping parameters for $\mathbf{I}$ that need to be estimated by minimizing $\varepsilon(\mathbf{P})$. Since $\varepsilon(\mathbf{P})$ is difficult to optimize directly, we choose to iteratively minimize $\varepsilon_i(\mathbf{p}_i)$ for each $\mathbf{I}_i$.

In the cost function, $\varepsilon_i(\mathbf{p}_i)$ equals the summation of the pairwise difference between $\mathbf{I}_i$ and all the other images in the warped image space. On the one hand, minimizing the $1^{st}$ term of Eqn. (1) makes the warped image content of the $i^{th}$ unlabeled image similar to that of the other *unlabeled* images, without regard for the physical meaning of the content. On the other hand, the $2^{nd}$ term of Eqn. (1) constrains $\mathbf{I}_i(\mathbf{W}(\mathbf{x};\mathbf{p}_i))$ to be similar to those of the *labeled* images and enforces the physical meaning of the content during alignment. Thus, the labels of $\tilde{\mathbf{I}}$ are propagated to $\mathbf{I}$. Since $K \gg \tilde{K}$, a weighting coefficient $\alpha$ can balance the contributions of the two terms in the overall cost.

We adopt the inverse warping technique [1] to minimize $\varepsilon_i(\mathbf{p}_i)$. We first estimate the warping parameter updates $\Delta\mathbf{p}_i$ by minimizing the following equation:

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

and then update the warping function by:

$$\mathbf{W}(\mathbf{x};\mathbf{p}_i) \leftarrow \mathbf{W}(\mathbf{x};\mathbf{p}_i) \circ \mathbf{W}(\mathbf{x};\Delta\mathbf{p}_i)^{-1}. \quad (3)$$

The function $\varepsilon_i(\Delta\mathbf{p}_i)$ is nonlinear with respect to $\Delta\mathbf{p}_i$. To support numeric optimization of this function, we take the first order Taylor expansion on $\mathbf{I}_j(\mathbf{W}(\mathbf{W}(\mathbf{x};\Delta\mathbf{p}_i);\mathbf{p}_j))$ and $\tilde{\mathbf{I}}_n(\mathbf{W}(\mathbf{W}(\mathbf{x};\Delta\mathbf{p}_i);\tilde{\mathbf{p}}_n))$:

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

As a result, Eqn. (2) is simplified to:

$$\frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^{K}\|\mathbf{b}_j + \mathbf{c}_j\Delta\mathbf{p}_i\|^2 + \frac{\alpha}{\tilde{K}}\sum_{n=1}^{\tilde{K}}\|\tilde{\mathbf{b}}_n + \tilde{\mathbf{c}}_n\Delta\mathbf{p}_i\|^2, \quad (4)$$

where

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

The least-squares solution of Eqn. (4) is given as:

$$\Delta\mathbf{p}_i = -\mathbf{H}^{-1}\left[\frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^{K}\mathbf{c}_j^T\mathbf{b}_j + \frac{\alpha}{\tilde{K}}\sum_{n=1}^{\tilde{K}}\tilde{\mathbf{c}}_n^T\tilde{\mathbf{b}}_n\right], \quad (5)$$

with

$$\mathbf{H} = \frac{1-\alpha}{K-1}\sum_{j=1,j\neq i}^{K}\mathbf{c}_j^T\mathbf{c}_j + \frac{\alpha}{\tilde{K}}\sum_{n=1}^{\tilde{K}}\tilde{\mathbf{c}}_n^T\tilde{\mathbf{c}}_n. \quad (6)$$

Joint alignment for an image ensemble can be a computationally intensive task, so we have analyzed the computational cost of the SLSC method. Note that since $\tilde{\mathbf{P}} = \{\tilde{\mathbf{p}}_n\}_{n\in[1,\tilde{K}]}$ 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 Eqn. (4) is negligible. Therefore, semi-supervised congealing has a computational cost similar to that of unsupervised congealing.

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

Table 1. The computational cost of major steps in SLSC.

# 4. Automatic Landmark Labeling with SLSC

In this section, we will first present a shape-constrained SLSC, which improves the robustness of the congealing process by reducing outliers, and then extend it to a patch-based approach to achieve an accurate estimate of the landmarks by partitioning the mean shape space.

## 4.1. Shape-constrained SLSC

Given the warping parameters for all images $\{\mathbf{P},\tilde{\mathbf{P}}\} = [\mathbf{p}_1,\cdots,\mathbf{p}_K,\tilde{\mathbf{p}}_1,\cdots,\tilde{\mathbf{p}}_{\tilde{K}}]$, and their corresponding landmark locations $\{\mathbf{S},\tilde{\mathbf{S}}\} = [\mathbf{s}_1,\cdots,\mathbf{s}_K,\tilde{\mathbf{s}}_1,\cdots,\tilde{\mathbf{s}}_{\tilde{K}}]$, where $\mathbf{s}$ is a concatenated vector of a set of 2D landmark coordinates $\mathbf{s} = [x_1,y_1,x_2,y_2,...,x_v,y_v]^T$, there are two ways of mapping between $\{\mathbf{P},\tilde{\mathbf{P}}\}$ and $\{\mathbf{S},\tilde{\mathbf{S}}\}$. First, 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 mean shape space. 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. Second, the warping parameter $\mathbf{p}_i$ can be obtained given the corresponding landmark pairs ($\mathbf{x}_s$ and $\mathbf{s}_i$). Consequently, refining the positions of the landmarks can improve the estimation of the warping parameters. Motivated by this, we develop an approach denoted *Shape-constrained SLSC (SSLSC)*, which integrates the shape constraints with 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) [8]. Since only a few labeled images are available, the PDM is learnt from both the labeled landmarks and an automatically chosen *low-error* subset of the estimated landmarks in an on-line manner. Then, the other *bad* estimations can be "corrected" through a PCA reconstruction as follows:

**Algorithm 1** Shape-constrained SLSC (SSLSC)

**Input:** $\mathbf{I}$, $\tilde{\mathbf{I}}$, $\mathbf{P}^0$, $\tilde{\mathbf{P}}$, $\mathbf{x}$, and $\mathbf{x}_s$
**Output:** $\mathbf{P}^t$, $\mathbf{S}^t$, and $\varepsilon$

```
t ← 0;
Compute s̃_i = W(x_s; p̃_i) for i ∈ [1, K̃];
repeat
    for i = 1 to K do
        ε_i(Δp_i), p_i^{t+1} ← SLSC(I, Ĩ, P^t, P̃, x);
    end for
    Rank ε_1(Δp_1), ··· , ε_K(Δp_K) in ascending order and pick the first
    K_M images;
    Compute s_i^{t+1} = W(x_s; p_i^{t+1}) for i ∈ [1, K_M];
    s̄,Q, λ ← PCA on S = [s̃_1, ··· , s̃_K̃, s_1^{t+1}, ··· , s_{K_M}^{t+1}]^T;
    for i = K_M + 1 to K do
        Reconstruct s_i^{t+1} as Eqn. (7);
        Compute p_i^{t+1} from s_i^{t+1};
    end for
    P^{t+1} ← [p_1^{t+1}, ··· , p_K^{t+1}]; S^{t+1} ← [s_1^{t+1}, ··· , s_K^{t+1}]^T;
    ε ← Σ_{i=1}^K ε_i(Δp_i); t ← t + 1.
until Converge
```

$$\hat{\mathbf{s}}_i = \bar{\mathbf{s}} + \mathbf{Q}\mathbf{z}, \tag{7}$$

where $\hat{\mathbf{s}}_i$ is the reconstructed shape vector for the $i^{th}$ image; $\bar{\mathbf{s}}$ and $\mathbf{Q}$ are the mean shape [1] and the shape basis obtained through the on-line training; $\mathbf{z}$ is the shape parameter vector that is restricted in some range [8]. Finally, a new warping parameter vector $\hat{\mathbf{p}}_i$ is computed from the refined landmark positions $\hat{\mathbf{s}}_i$. By doing so, the outliers of the congealing process are discovered and constrained in a principled way.

The SSLSC is summarized in Algorithm 1, where $\mathbf{P}^0$ is the initial warping parameters for $\mathbf{I}$. The labeled landmarks are fully utilized in the sense that they not only contribute to the cost minimization in Eqn. (2), but also provide guidance for shape deformation.

### 4.2. Landmark Labeling by Partition

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 with direct use of the piecewise affine warp. We attribute this difficulty to the high dimensionality of the warping parameter $\mathbf{p}_i$.

To address this issue, let us look at the piecewise affine warp closely. We note that in this case the warping function $\mathbf{W}()$ is a series of 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 the poor initialization and limited information encoded in the patch. Based on these observations, we propose a coarse-to-fine partition-based congealing method to improve the precision of landmark labeling.

[1] $\bar{\mathbf{s}}$ is different from $\mathbf{x}_s$. The former is learnt from the on-line training, whereas the latter is predefined and represents a generic structure of the target object.

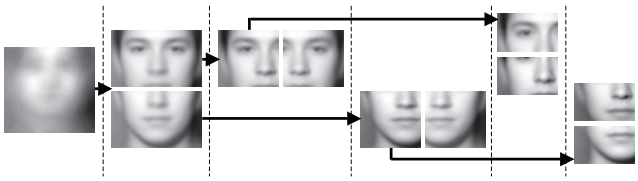


Figure 2. An example of 6-level partition in the mean shape space.

The partitioning strategy is summarized in Algorithm 2, where $\mathbf{S}_{init}$ is the initial guess of the landmark positions for $\mathbf{I}$. In the algorithm, besides the notation mentioned previously, $R$ represents the indices of the patches to be aligned in the current partition level; $d$ represents the index of the patch. Starting from the initial mean shape space $\mathbf{x}^1$, the process is conducted by repeatedly partitioning the mean shape space for a selected patch ($\mathbf{x}^{k*}$), which has the maximal congealing error ($\varepsilon$), into multiple child patches [2]. After the partitioning, the SSLSC is applied on each child patch, independently, to obtain the corresponding landmark positions within the patch itself. The partitioning is stopped when no cost reduction is achieved or the size of the patch is too small. Here, the patch reaches its size limit if the number of target landmarks in $\mathbf{x}^d$ is less than the number of corresponding landmarks required for computing $\mathbf{p}_i$. One example of multiple level partition is shown in Fig. 2.

Our top-down congealing strategy performs a coarse-to-fine alignment for the entire image ensemble. The congealing in the coarse-level partition focuses on aligning the features that are most similar among the image ensemble such as eyes, whereas the other features like nose and mouth are neglected. Hence, the landmark estimation on the larger patches is often coarse and used to provide a good initialization for the latter levels. With the increasing of the partition level, more details of the target object are revealed. As a result, the estimation of the landmarks becomes more and more precise.

## 5. Experiments

In order to demonstrate the effectiveness of the proposed algorithm, we have performed extensive validation studies for the application of labeling facial landmarks. We aim to automatically annotate 33 specific landmarks around the facial features (i.e. eyes, eyebrows, nose, mouth, and contour) for a large image set, given a few labeled example images. To the best of our knowledge, there is no prior work addressing this challenging problem, so we only compare with our own specific work as described above. For this purpose, we collect 300 images from the Notre Dame (ND1) database [4]. We manually label 33 landmarks for

[2] In this work, two equal sized child patches are generated by each partitioning. To enforce a geometrical relationship between the child patches, they are overlapped such that some landmarks reside in both of them. Positions of these landmarks are estimated as averages of the SSLSC results of the two child patches.

**Algorithm 2** Landmark Labeling by Partition

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

$l_{min} \leftarrow$ minimum number of landmarks in a patch;
$d \leftarrow 1$; $R \leftarrow \{1\}$; $\mathbf{S}^1_{init} \leftarrow \mathbf{S}_{init}$;
Compute $\mathbf{x}^1$ and $\mathbf{x}^1_s$ from $\tilde{\mathbf{S}}$;
**while** $d <$ maximum number of patches **do**
  **for** each $r \in R$ **do**
    Calculate $\mathbf{P}^r_{init}$ and $\tilde{\mathbf{P}}^r$ from $(\mathbf{S}^r_{init}, \mathbf{x}^r_s)$ and $(\tilde{\mathbf{S}}, \mathbf{x}^r_s)$, respectively;
    $\mathbf{P}^r, \mathbf{S}^r, \varepsilon^r \leftarrow SSLSC(\mathbf{I}, \tilde{\mathbf{I}}, \mathbf{P}^r_{init}, \tilde{\mathbf{P}}^r, \mathbf{x}^r, \mathbf{x}^r_s)$;
    **if** no cost reduction is achieved in $r$ over its parent patch **then**
      **return** $\mathbf{S}^{d-2}$ // return labeling results of last partition level;
    **end if**
  **end for**
  $k* = \underset{k}{\operatorname{argmax}}\, \varepsilon^k, k \in [1, d]$;
  child patches $\mathbf{x}^{d+1}, \mathbf{x}^{d+2}, \mathbf{x}^{d+1}_s, \mathbf{x}^{d+2}_s \leftarrow$ Partition the $k*^{th}$ patch;
  $\mathbf{S}^{d+1}_{init} \leftarrow \mathbf{S}^{k*}$; $\mathbf{S}^{d+2}_{init} \leftarrow \mathbf{S}^{k*}$; $\varepsilon^{k*} \leftarrow 0$;
  **if** $size(\mathbf{x}^{d+1}_s) < l_{min}$ **or** $size(\mathbf{x}^{d+2}_s) < l_{min}$ **then**
    **return** $\mathbf{S}^d$;
  **end if**
  $d \leftarrow d + 2$; $R \leftarrow \{d + 1, d + 2\}$.
**end while**

each image to establish a ground truth and to enable a quantitative evaluation for the labeling performance.

In the following, we will demonstrate that with only a few labeled images, robust and accurate landmark labels can be obtained with the proposed algorithm. We divide the 300 labeled ND1 images into two non-overlapping sets: a labeled set with $\tilde{K}$ images and an unlabeled set with $300 - \tilde{K}$ images. For quantitative evaluation, the initial value of the $j^{th}$ element of $\mathbf{S}_i$ is generated by adding a uniformly distributed random noise $\eta \in [-\eta_{max}, \eta_{max}]$ to the ground-truth value $\hat{\mathbf{S}}_{i,j}$ as follows,

$$\mathbf{S}_{i,j} = \hat{\mathbf{S}}_{i,j} + \frac{\eta \rho_i}{\bar{\rho}}. \tag{8}$$

where $\rho_i$ is the eye-to-eye pixel distance of $\mathbf{I}_i$, and $\bar{\rho}$ is the average of $\rho_i$ for all unlabeled images ($\bar{\rho} \approx 130\ pixels$ in our experiments). By doing so, the level of deviation in the initialization is relative to the face size. In practical applications, the initial landmark positions can be obtained from a face detector. A 6-parameter affine transformation is employed in the congealing process, and a $72 \times 72$ square region, which encloses all the target landmarks, is used as the common mean shape in the experiments. To accommodate illumination changes, the warped face region is normalized by subtracting its mean intensity value and then divided by its standard deviation.

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 $\rho_i$, 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 unlabeled images. A smaller NRMSE indicates a higher labeling accuracy, and a smaller SOF represents greater robustness.

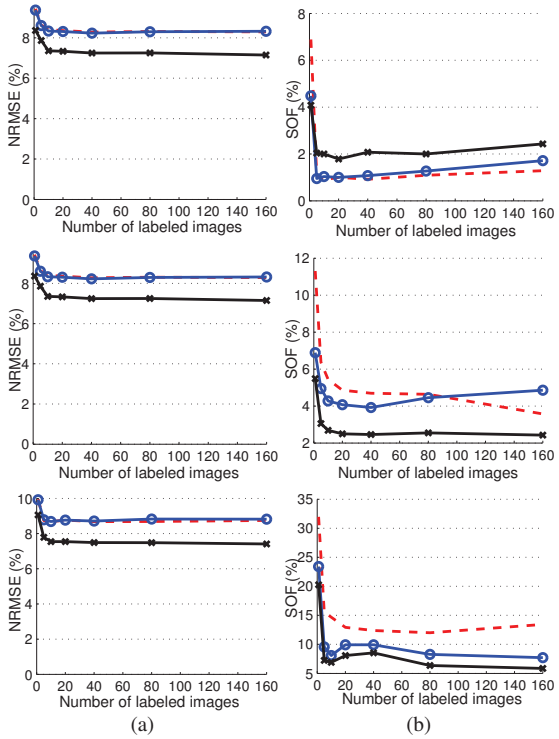
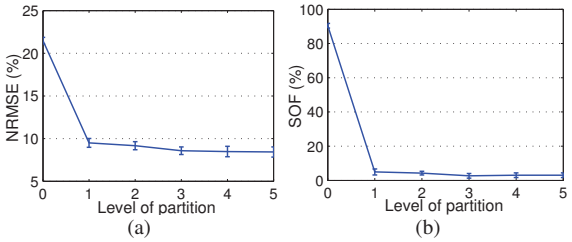



Figure 3. Performance comparison for SLSC (red dashed line), SSLSC (blue line with circles), and partition-based SSLSC (black line with crosses) in terms of (a) NRMSE of landmarks excluding outliers and (b) SOF. The results in each row correspond to a noise level ($\eta_{max} = 10,\ 30,\ 50$ from top to bottom), respectively.

## 5.1. Results of Shape-constrained SLSC

Fig. 3 shows the comparison of SSLSC and SLSC under the effects of varying number of labeled images $\tilde{K} \in \{1, 5, 10, 20, 40, 80, 160\}$ and different noise levels $\eta_{max} \in \{10, 30, 50\}$, in terms of NRMSE and SOF. Note that, for this result, outliers are excluded from the computation of NRMSE. The results are computed from an average of 5 trials, where $\tilde{K}$ images are randomly selected as the labeled set for each trial. We ensure both algorithms are compared under the same condition. For example, they use the *same* randomly selected labeled set and the *same* initialization.

Comparing the results of SLSC (red dashed line) and SSLSC (blue line with circles) in Fig. 3, we see that the shape constraints are effective in reducing the outliers significantly, even when the congealing performance of SLSC is poor due to a high initialization noise level and a small number of labeled images. For example, the SOF decreases from 32% (SLSC) to 23.4% (SSLSC) with $\tilde{K} = 1$ and $\eta_{max} = 50$, which is equivalent to removing 26 outliers. Furthermore, an average of 5.2% reduction of SOF is obtained when $\eta_{max} = 50$. Since the shape constraints are not applied on those low-error estimations, there is no improvement in the NRMSE excluding outliers.



(a) (b)

Figure 4. Performance analysis by varying partition levels in terms of (a) NRMSE of landmarks and (b) SOF, where the center and the length of the error bar represent the mean and standard deviation of 5 random trials, respectively.

## 5.2. Results of Partition-based SSLSC

In this experiment, we demonstrate the improvement of labeling accuracy by partition-based SSLSC. Similar to the previous experiment, the performance is evaluated under varying $\tilde{K}$ and $\eta_{max}$ from an average of 5 random trials.

Comparing the results of SSLSC (blue line with circles) and partition-based SSLSC (black line with crosses) in Fig. 3, it is obvious that the partition-based approach further improves both precision and robustness in terms of reducing the NRMSE and SOF. For example, the SOF decreases from $23.4\%$ (SSLSC) to $20.3\%$ (partition-based SSLSC), and the NRMSE decreases from $9.92\%$ (SSLSC) to $9.06\%$ (partition-based SSLSC) with $\tilde{K}$=1 and $\eta_{max}$=50. In summary, an average of $1\%$ reduction of NRMSE is achieved for all noise levels, and an average of $2\%$ decrease of SOF is obtained for $\eta_{max}=30, 50$. From Fig. 3, we can find that there is no remarkable improvement when $\tilde{K}$>=10, which means that only using $3\%$ $(10/300)$ labeled data, we can estimate the landmarks accurately and robustly.

Fig. 4 illustrates the performance improvement across different partition levels when $\tilde{K} = 5$ and $\eta_{max} = 30$. The results of level-0 correspond to the initialization, and those of level-1 represent the congealing results on the whole mean shape space by SSLSC. We can see that with the increasing levels of partition, both the NRMSE and SOF decrease and converge at the last partition level.

In Fig. 5, we also show exemplar labeling results under three initialization noise levels, respectively. To compare the overall labeling performance, a mean warped face region [3] is also displayed. It can be observed that the first level of the congealing (middle row) can roughly localize the landmarks, but fail to handle the subtle facial appearance change caused by individual difference, facial expression, and face pose. In contrast, the landmark labels in the final partition level (last row) show a significant improvement of accuracy under slight changes in face pose (compare the noses and mouths in the $2^{nd}$, $3^{rd}$, $5^{th}$, and $6^{th}$ images of the last two rows) as well as individual differences (compare the contour in the $8^{th}$ image, and the nose and mouth in the $9^{th}$ image of the last two rows) [4].

[3] 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.

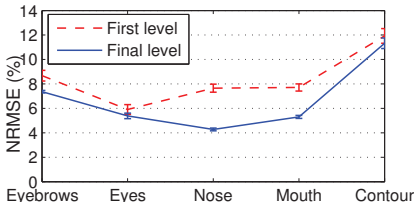



Figure 6. Performance of facial components at the first and final partition level. The center and the length of the error bar represent the mean and standard deviation of 5 random trials, respectively.

## 5.3. Analysis on Different Facial Components

From Fig. 5, 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 face contour is not obvious. Therefore, we conduct an analysis to study the performance improvement using the partition-based approach for different facial components.

Using the case $\tilde{K} = 10$ and $\eta_{max} = 30$ as an example, Fig. 6 shows that the partition-based algorithm improves the labeling accuracy on the different facial components at various degrees. On one hand, the improvements on the eyes and contour are subtle: a $0.5\%$ deduction of NRMSE for the eyes and a $0.6\%$ deduction for the contour, since these two facial components are the most distinguished features on the face, and attract the most attention in the first level of congealing. To get accurate alignment for the eyes and contour, the accuracy of other facial components is sacrificed. On the other hand, the improvement on the nose and mouth are significant: a $3.4\%$ reduction for the nose and a $2.4\%$ reduction for the mouth, as well as the remarkable reduction in their corresponding variances, because the nose and upper lip are involved in almost all partition levels as shown in Fig. 2. With the increasing of partition levels, more details raised from these facial components are revealed and contribute to the congealing process.

## 5.4. Analysis on the Weighting Coefficient

Besides studying the effect of noise level and the number of labeled images on the congealing performance, we also analyze how the selection of the weighting coefficient $\alpha$ in Eqn. (2) affects the proposed algorithm. The larger $\alpha$, the more the algorithm relies on the labeled data.

In the extreme case, $\alpha = 0$ implies an unsupervised congealing, and $\alpha = 1$ is supervised. Fig. 7 illustrates the performance with various values of $\alpha$ with $\tilde{K} = 5$ and $\eta_{max} = 30$. Although using a large $\alpha$ improves the accuracy of the algorithm, it tends to result in more outliers, especially with only a few labeled data. This is because the shape/appearance

[4] More iteratively landmark labeling results can be found in the supplemental materials (iteratively-landmark-labeling-demo.avi).

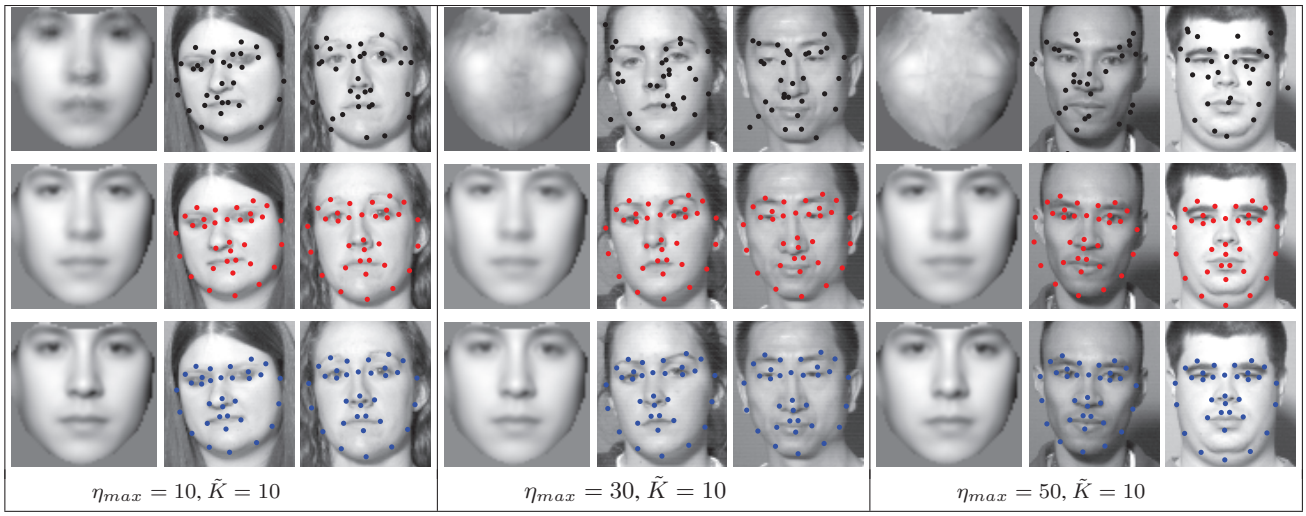
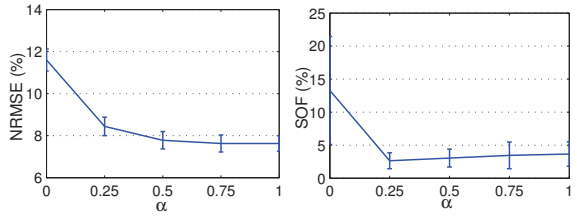



Figure 5. Landmark labeling results for three noise levels. For each cell, the three rows illustrate the initial landmark positions, the level-1 results, and the final labeling results, respectively. In each row, the mean warped face region and two example images are shown.




(a) (b)

Figure 7. Performance analysis by varying $\alpha$ in terms of (a) NRMSE of landmarks and (b) SOF.

variations in the large image ensemble cannot be well represented by only a small number of labeled data. A good trade-off can be achieved by balancing the weights of the labeled and unlabeled data. We use $\alpha = 0.5$ in the experiments reported in the previous discussions.

### 5.5. Labeling Confidence

A confidence score is desirable for practical applications in order to evaluate the quality of labeling without ground truth. For this we use $\varepsilon_i$ in Eqn. (2). A smaller $\varepsilon_i$ indicates a higher-confidence in labeling. Fig. 8 shows labeled images with the lowest, median, and highest confidence scores, in an image ensemble ($\tilde{K} = 10$ and $\eta_{max} = 30$). It is clear that the confidence score is indicative of labeling performance in that the labeling is improving from the top to bottom row. Similar phenomena have been observed for experiments on other databases such as Caltech 101 face database [13]. In practice, after labeling, one can use this confidence score to select only well-labeled samples for a training set, or to select samples for other appropriate additional processing.

### 5.6. Cross-Database Validation

Here, we demonstrate the generalization ability of the proposed algorithm through two cross-database validations. First, we automatically label the 33 landmarks on 255 images from Caltech 101 face database [13] with the help of 50 manually labeled images from the ND1 database. For initialization, we first obtain the location of the face with a face detector. Then, as shown in Fig. 9(a), the initial positions of landmarks are generated by adapting the mean shape vector $\mathbf{x}_s$ to the detected face region. Since we do not have ground-truth labels for this database, we can only perform qualitative evaluation by visual observation. Fig. 9(b) illustrates sample labeling results. Although the congealing on the Caltech 101 face database is much more difficult than that on the ND1 database due to cluttered backgrounds and challenging illumination conditions, our algorithm can still achieve satisfactory labeling performance.

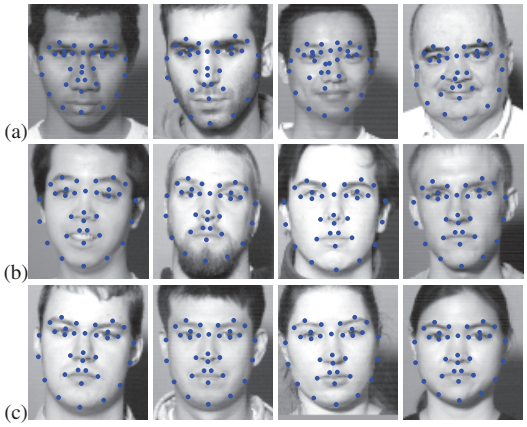



Figure 8. Labeled images with the (a) lowest, (b) median, and (c) highest labeling confidences.

We also obtain excellent label results on a very large dataset (combining ND1 and FERET databases [21]), 1176 total images with only 15 (= 1.3%) labeled images. This shows that our system can accommodate a vast amount of data, without noticeable sacrifice in performance [5].

[5] Please see supplemental material: labeling-results-large-database.avi.

(a) (b)

Figure 9. Results on the Caltech 101 face database: (a) the initialization of our algorithm and (b) the landmark labeling results.

The main purpose of this work is the *one-time* off-line annotation of training data, so efficient run-time is not a critical requirement. Our method has acceptable computational complexity. A partition-based SSLSC experiment to produce 300 labeled images takes about 2.5 hours using a Matlab$^{TM}$ implementation with a 2 GHz CPU.

## 6. Conclusions

Shape deformation of images of a real-world object is often non-rigid due to inter-subject variability, object motion, and camera view point. Automatically estimating non-rigid deformations for an object class is a critical step in characterizing the object and learning statistical models. Our proposed approach facilitates such a task by automatically producing labeled data sets. Extensive experiments demonstrate that our system has achieved impressive labeling results on face images with nearly frontal view and moderate changes in expression, useful for many practical applications. In the future, we expect to extend the methods to handle large facial variations.

Although we apply our approach on facial images, no domain knowledge of faces is used here. Hence, the approach can be immediately applied to the task of labeling landmarks in images of other classes of objects such as vehicles or pedestrians.

## References


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

[2] S. Baker, I. Matthews, and J. Schneider. Automatic construction of active appearance models as an image coding problem. *IEEE TPAMI*, 26(10):1380–1384, 2004.

[3] S. Balci, P. Golland, M. Shenton, and W. Wells. Free-form b-spline deformation model for groupwise registration. In *MICCAI*, pages 23–30, 2007.

[4] K. Chang, K. Bowyer, and P. Flynn. An evaluation of multi-modal 2D+3D face biometrics. *IEEE TPAMI*, 27(4):619–624, 2005.

[5] T. Cootes. Timeline of developments in algorithms for finding correspondences across sets of shapes and images. Technical report, University of Manchester, July 2005.

[6] T. Cootes, G. Edwards, and C. Taylor. Active appearance models. *IEEE TPAMI*, 23(6):681–685, 2001.

[7] T. Cootes, S. Marsland, C. Twining, K. Smith, and C. Taylor. Groupwise diffeomorphic non-rigid registration for automatic model building. In *ECCV*, volume 4, pages 316–327, 2004.

[8] T. Cootes, C. Taylor, D. Cooper, and J. Graham. Active shape models — their training and application. *CVIU*, 61(1):38–59, 1995.

[9] T. Cootes, C. Twining, V. Petrovic, R. Schestowitz, and C. Taylor. Groupwise construction of appearance models using piece-wise affine deformations. In *BMVC*, volume 2, pages 879–888, 2005.

[10] M. Cox, S. Sridharan, S. Lucey, and J. Cohn. Least squares congealing for unsupervised alignment of images. In *CVPR*, 2008.

[11] D. Cristinacce and T. Cootes. Facial motion analysis using clustered shortest path tree registration. In *Proc. of the 1st Int. Workshop on Machine Learning for Vision-based Motion Analysis with ECCV*, 2008.

[12] N. Dalal and W. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 1, pages 886–893, 2005.

[13] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *CVIU*, 106(1):59–70, 2007.

[14] L. Georg, P. Peloschek, R. Donner, and B. Horst. Annotation propagation by MDL based correspondences. In *Proc. of Computer Vision Winter Workshop*, pages 11–16, 2006.

[15] R. Gross, I. Matthews, and S. Baker. Generic vs. person specific active appearance models. *IVC*, 23(11):1080–1093, 2005.

[16] I. Kokkinos and A. Yuille. Unsupervised learning of object deformation models. In *ICCV*, 2007.

[17] E. Learned-Miller. Data driven image models through continuous joint alignment. *IEEE TPAMI*, 28(2):236–250, 2006.

[18] X. Liu. Generic face alignment using boosted appearance model. In *CVPR*, 2007.

[19] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004.

[20] E. Miller, N. Matsakis, and P. Viola. Learning from one example through shared densities on transforms. In *CVPR*, volume 1, pages 464–471, 2000.

[21] P. Phillips, H. Moon, P. Rauss, and S. Rizvi. The FERET evaluation methodology for face recognition algorithms. *IEEE TPAMI*, 22(10):1090–1104, 2000.

[22] J. Saragih and R. Goecke. A nonlinear discriminative approach to AAM fitting. In *ICCV*, 2007.

[23] F. Torre and M. Nguyen. Parameterized kernel principal component analysis: Theory and applications to supervised and unsupervised image alignment. In *CVPR*, 2008.

[24] T. Vetter, M. Jones, and T. Poggio. A bootstrapping algorithm for learning linear models of object classes. In *CVPR*, pages 40–46, 1997.

[25] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 57(2):137–154, 2004.