Towards a Unified Framework for Pose, Expression, and Occlusion Tolerant Automatic Facial Alignment

Keshav Seshadri, *Student Member, IEEE* and Marios Savvides, *Member, IEEE*

Abstract—We propose a facial alignment algorithm that is able to jointly deal with the presence of facial pose variation, partial occlusion of the face, and varying illumination and expressions. Our approach proceeds from sparse to dense landmarking steps using a set of specific models trained to best account for the shape and texture variation manifested by facial landmarks and facial shapes across pose and various expressions. We also propose the use of a novel ℓ_1-regularized least squares approach that we incorporate into our shape model, which is an improvement over the shape model used by several prior Active Shape Model (ASM) based facial landmark localization algorithms. Our approach is compared against several state-of-the-art methods on many challenging test datasets and exhibits a higher fitting accuracy on all of them.

Index Terms—Facial alignment, automatic facial landmark localization, ℓ_1-regularized least squares, active shape models (ASMs)

1 Introduction

The localization of facial landmarks, also referred to as facial landmarking or facial alignment, is a key preprocessing step that can aid in carrying out facial recognition [1], [2], [3], [4], generation of 3D facial models [5], expression analysis [6], [7], age estimation [8], facial hair segmentation [9], and a variety of other facial analytic tasks. With the strides made in all of these areas over the past few years there has been a shift towards harnessing local information in regions around key facial landmarks. This in turn has motivated the need for extremely precise automatic facial alignment algorithms that can generalize well enough to handle variations in pose, illumination, expression, and large levels of occlusion in unseen test images.

We present an approach that overcomes the previously mentioned challenges. Since facial shape and the local texture around the landmarks that constitute them vary dramatically with facial pose and expressions, it is necessary to build not one, but multiple models that can best account for these variations. In order to combat the problem of facial occlusions, we explicitly account for them during our training process. Many existing facial alignment algorithms also rely heavily on consistent facial detection results, something that is seldom guaranteed when dealing with real-world images data as facial bounding box results produced by the same detector vary in size and location even for a similar set of images and do not always account for in-plane rotation (roll) of the face. It is for these reasons that we structure our approach in a different fashion compared to many existing ones.

Our approach proceeds in a stage wise fashion, as shown in Fig. 1. In our first step, we use a sliding window approach to only localize a few key landmarks, such as the centers of the eyes, corners of the mouth, tip of the nose, etc., that we refer to as seed landmarks. Our next step involves aligning a denser set of landmarks (a canonical mean facial shape specific to a particular yaw angle range that is obtained during our training stage) using just two of the seed landmark candidates at a time and evaluating the goodness of fit of this dense set of points. We are thus able to generate the best fitting initial shape for each of our pose specific models while accounting for any in-plane rotation of the face (roll) and the presence of occlusions. The final stage involves the refinement of the top ranked shapes and the selection of a single set of final landmarks that best model the shape and texture of a given face.

Our contributions are as follows: (1) the formulation of a framework (pipeline) for dense facial landmark localization to jointly deal with the problems posed by facial pose variation (with yaw variation from -90 to $+90$ degree being handled), varying facial expressions, and partial occlusion of the face, (2) a novel method to constrain shape coefficients to avoid the generation of implausible facial shapes, and (3) a comparison of our approach against many state-of-the-art approaches on several challenging real-world datasets.

2 Related Work

Traditionally facial landmark localization has been carried using deformable template (parametric) based models, such as Active Shape Models (ASMs) [10], [11] (that belong to a class of methods that can be broadly referred to as Constrained Local Models (CLMs) [12], [13], [14]) and Active Appearance Models (AAMs) [15]. Both ASMs and AAMs build shape models (also referred to as Point Distribution

• *The authors are with the CyLab Biometrics Center and the Department of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA 15213.*
E-mail: kseshadr@andrew.cmu.edu, msavvid@ri.cmu.edu.

Manuscript received 18 Oct. 2014; revised 22 Oct. 2015; accepted 17 Nov. 2015. Date of publication 2 Dec. 2015; date of current version 12 Sept. 2016. Recommended for acceptance by T. Cootes.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TPAMI.2015.2505301

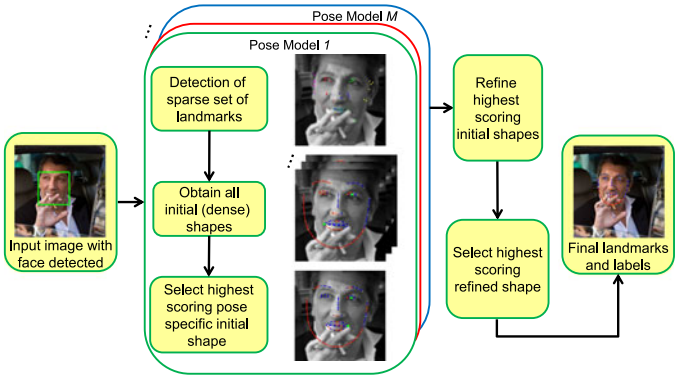


Fig. 1. An overview of our approach showing how each stage in the process works on an image from the test set partition of the LFPW dataset. In all facial images with landmarks overlaid on them, yellow dots are used to indicate the locations of facial landmarks, blue line segments indicate that the landmark at their center is accurately localized, and red line segments indicate that the landmark at their center is misaligned or potentially occluded. This same color scheme is maintained in all figures containing fitting results produced by our approach.

Models (PDMs)), that model the shape of a typical face that is represented by a set of constituent landmarks, and texture models of what the region enclosed by these landmarks looks like. The difference between the two is that ASMs build local texture models of what small 1D or 2D regions around each of landmarks look like, while AAMs build global texture models of the entire convex hull bounded by the landmarks. Recently, Tzimiropoulos and Pantic [16] proposed new optimizations for fast and accurate AAM fitting and demonstrated better fitting results on unseen images using a more unconstrained training set drawn from the Labeled Face Parts in the Wild (LFPW) dataset [17]. This is one of the approaches that we benchmark our approach against in Section 4.2.

There has been a recent increase in literature dealing with the automatic landmarking of non-frontal faces using various unique approaches. In their recent seminal work, Zhu and Ramanan [18] proposed an elegant framework that built on the previously developed idea of using mixtures of Deformable Part Models (DPMs) for object detection [19] to simultaneously detect faces, localize a dense set of landmarks, and provide a course estimate of facial pose (yaw) in challenging images. The approach is quite effective at handling a wide range of yaw variation but does not account for excessive in-plane rotation of faces or large occlusion levels. Ghiasi and Fowlkes [20] built on this work and proposed a hierarchical deformable part model for face detection and landmark localization to explicitly model the occlusion of parts and hence achieved more accurate results on challenging occluded images in the wild. However, their approach modeled three clusters and handled a yaw variation of only ± 22.5 degree. Yu et al. [21] also built on the Zhu and Ramanan's work and proposed two-stage cascaded deformable shape model with a 3D facial shape model to simultaneously detect faces and localize landmarks on faces with large head pose variations. Asthana et al. [22] developed a discriminative regression based approach for the CLM framework that they referred to as Discriminative Response Map Fitting (DRMF). Xiong and De la Torre [23] formulated the Supervised Descent Method (SDM) for minimizing a Nonlinear Least Squares (NLS) function and applied it to the task of detecting facial landmarks. We benchmark our approach against all three of these approaches in Section 4.2.

Recently, a variety of approaches that can be broadly grouped under the category of regression based approaches [24], [25], [26], [27] have emerged. To explicitly deal with occluded faces and provide feedback on which landmarks were occluded Burgos-Artizzu et al. [25] proposed the Robust Cascaded Pose Regression (RCPR) algorithm to localize a set of 29 facial landmarks (the same as those localized in [17]). They incorporated occlusion directly into the learning stage, using facial images that were both manually annotated and provided with occlusion labels, to improve shape estimation.

Many of the previously developed approaches, such as those proposed in [28] and [29], only localize a sparse set of landmarks which is unsuitable for many applications, such as expression analysis or the building of 3D facial models. Most of the approaches, with the exception of [18] and [21], do not yet handle faces with absolute yaw angles in greater than 60 degree and are thus presently incapable of dealing with profile faces. Finally, only a few of the approaches, such as [25], provide feedback in the form of occlusion labels for the detected landmarks. It is our intention to draw attention to the all of these challenges (or desirable attributes in a facial alignment algorithm) and provide details on our own algorithm that is able to generalize well enough to handle these variations.

3 Our Approach

This section outlines our approach to pose, expression, and occlusion tolerant facial alignment. It is to be noted that while it is possible for our approach to also deal with some slight variations in pitch, we do not explicitly train pitch models and hence in our context the use of the phrase "pose specific" is meant to be synonymous with "yaw specific".

3.1 Sparse Landmark Detection

As we have previously mentioned, our initial step in the alignment process is the detection of a sparse set of key facial landmarks that we refer to as seed landmarks. We train our models (see Section 4.1 for details) using a subset of images from the CMU Multi-PIE (MPIE) database [30], [31] with manual annotations available for images using a 68 point landmarking scheme for frontal faces (the definition of frontal in our context includes faces with a yaw angle between -45 and $+45$ degree) and a 39 point landmarking scheme for non-frontal (faces with an absolute yaw angle greater than 45 degree) faces. These landmarking schemes are shown in Fig. 2. For frontal faces, we search for eight seed landmarks include the centers of the two eyes, tip of the nose, corners of the mouth, tip of the chin, and two opposite points on the facial boundary close to the ears (landmarks 2 and 16 in frontal faces and landmark 38 in profile faces), as depicted in the landmarking schemes in Fig. 2. The same corresponding set of seed landmarks is searched for in profile faces (faces that exhibit an absolute yaw angle greater than 45 degree), however, the number of seed landmarks in such cases is 5, due to hidden landmarks on one side of the face.

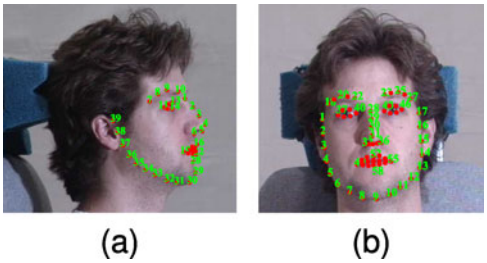


Fig. 2. MPIE landmarking (markup) schemes for (a) profile faces (39 facial landmarks), (b) frontal faces (68 facial landmarks). The facial images in the figure are from the MPIE database.

During our training stage we construct landmark, expression, and pose specific local appearance (texture) models for each landmark, including the seed landmarks. It is to be noted that we build and use $M = 10$ models for the various yaw angle ranges used in our approach. We also build six models for frontal yaw angles with open mouth expressions (scream and surprise). However, we do not use them at this stage, or our subsequent dense landmarking stage, to ensure a higher a fitting speed and since we found that only our final refinement step demanded the use of expression specific models to obtain high fitting accuracy. Section 4.1 provides details on how these models are built and the parameters used in their construction.

The first step in a model's construction is to generate a crop of a fixed size around the ground truth landmark locations. For each landmark in every model, a classifier is built to distinguish the local texture around the landmark in a particular feature space from the local texture of a different landmark or an occlusion. This is carried out by extracting features for positive samples, at the exact locations of the ground truth coordinates and from a small region around these locations to add variation to the training set and account for a small amount of variance in human labeling, and negative samples (local texture descriptors for incorrect locations for a landmark) at various random locations close to and far away from the ground truths using all images for a specific yaw angle range and expression. We also construct separate linear subspaces (for the positive and negative classes using samples from the respective classes) using Principal Component Analysis (PCA) [32], [33], [34] as our dimensionality reduction technique. These subspaces are used in the next stage of our facial alignment pipeline (the dense landmark alignment stage). We use Histogram of Oriented Gradients (HOG) [35] as our feature descriptors as they have been proven to be quite discriminative in prior facial alignment algorithms, such as [18], [36], [37], [38], and are tolerant to illumination variation.

Our local texture classifiers are constructed using an ensemble of classifiers (decision stumps) in a Real AdaBoost [39] framework. We chose the Real AdaBoost framework due to the minimal parameters that need to be determined for such a classifier (only the number of boosting rounds or number of classifiers in the ensemble need to be specified) and its resistance to overfitting [40], [41]. The Real AdaBoost framework not only allows for the classification of a feature vector as positive or negative (misaligned or possibly occluded), but also returns a confidence score for the prediction. This allows us to greedily retain the highest scoring locations in the response map for a particular seed landmark

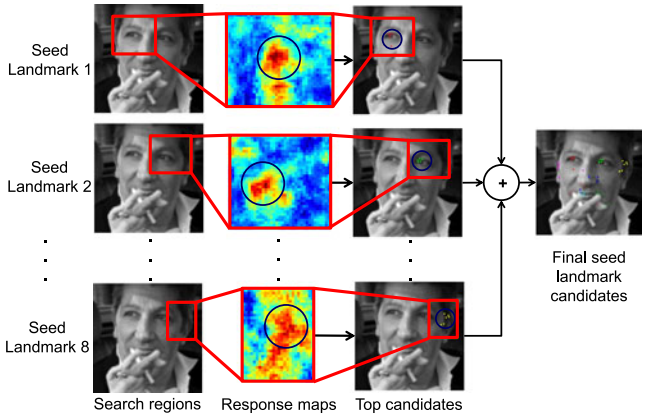


Fig. 3. Process by which seed landmark candidates are retained by our approach when fitting a facial image from the test set partition of the LFPW dataset. The process is shown only for one of the pose models (for a yaw angle of 0 to +15 degree), but is repeated to retain seed landmark candidates specific to each pose model.

when a search over the typical region where the landmark is likely to lie is performed on a test face crop. The search is repeated for rotated versions of the crop (typically for rotation angles between -30 and $+30$ degree in 15 degree increments) with clustering used to reduce the number of candidates if a number of them are found to lie within a small bandwidth. Fig. 3 shows how we retain candidates for the various seed landmarks for a particular pose specific model.

3.2 Dense Landmark Alignment and Optimal Shape Initialization

Once we have pose specific seed landmark candidates, our task becomes one of selecting a single combination of candidates for two different seed landmarks that allows for the optimal initialization of a pose specific mean facial shape $\bar{\mathbf{s}}^m \ (m = 1, \ldots, M)$ consisting of the full set of facial landmarks for that pose model, i.e., alignment of a dense set of pose specific landmarks. By aligning each pose specific mean shape with a combination of seed landmarks we end up with a total set of J_m dense shapes $\mathbf{s}^{j,m} \ (j = 1, \ldots, J_m)$ that must be ranked using a scoring function that numerically assesses their goodness of fit. This step is extremely important to the fitting process because poor initialization is something a facial alignment algorithm can seldom recover from [42]. Thus, our contribution in providing a framework to transition from a set of sparse landmarks (possibly containing some spurious detections) to a dense set of initial landmarks is quite important.

At this point it becomes necessary to provide details on how shape models (also sometimes referred to as Point Distribution Models) for CLMs work. Each facial shape $\mathbf{s}$ in the training set is represented by its N x and y coordinates in vectorial form as $\mathbf{s} = [x_1 \ x_2 \ldots x_N \ y_1 \ y_2 \ldots y_N]^{\mathrm{T}}$. These shapes are aligned using Generalized Procrustes Analysis (GPA) [43], [44], to normalize for scale, rotation, and translation effects and bring them into a common reference frame. In this reference frame, conventional PDMs are built by obtaining a mean facial shape $\bar{\mathbf{s}}$ and by constructing a PCA subspace $\mathbf{\Phi}$ of facial shape variation. A facial shape can now be represented using (1):

$$\mathbf{s} = T_{s,\theta,x_t,y_t}(\bar{\mathbf{s}} + \mathbf{\Phi}\mathbf{b}). \tag{1}$$

In (1), T is a similarity transformation parametrized by a scaling factor s, a rotation angle θ, and translation parameters x_t and y_t. The result obtained when the transformation T is applied to a single point (x, y) is shown in (2):

$$T_{s,\theta,x_t,y_t}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{bmatrix} s\cos\theta & -s\,sin\,\theta \\ s\sin\theta & s\,cos\,\theta \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} x_t \\ y_t \end{bmatrix}. \quad (2)$$

The entire shape fitting process centers around the determination of the optimal vector shape coefficients **b** that best represents (and regularizes) the current set of landmarks whose locations are determined using local texture based methods.

Since facial shape varies dramatically with pose and expressions, we construct 16 pose and expression specific PDMs in our approach. However, our approach to determine the shape coefficients is a novel method that does not use the conventional shape model equation in (1). Instead of using PCA to determine our set of basis vectors, we retain the entire set of shapes in a pose and expression specific dictionary that we later use in an ℓ_1-regularized least squares approach to determine the shape coefficients. However, we do retain the building of pose and expression specific mean shapes in our approach as well. It is the $M = 10$ pose specific mean shapes $\bar{\mathbf{s}}^m \ (m = 1, \ldots, M)$ that we use to determine the best initialization that can be provided for each pose range (the best fitting mean shape roughly aligned over the face for each pose range). It is to be noted that before the scoring is performed, the region around the shape is cropped and de-rotated (since the angle of rotation required can be calculated using two fixed landmarks) in order to match our training crops. For profile poses, a fewer set of shapes need to be evaluated as there are fewer seed landmarks. All of these shapes are now scored on their "goodness of fit", which is assessed by determining how well a landmark's surrounding local texture matches pre-trained models of what this local texture looks like in a particular feature space.

In order to numerically determine this, we first project the feature vector $\mathbf{t}_n^{j,m}$ (obtained using the local texture around $\mathbf{x}_n^{j,m}$) onto the positive subspace $\mathbf{\Psi}_{\mathrm{pos}_n}{}^m$ (after subtracting the mean texture vector $\bar{\mathbf{t}}_{\mathrm{pos}_n}{}^m$ for the subspace) for that landmark and pose to obtain coefficients $\mathbf{c}_{\mathrm{pos}_n}{}^m$, using (3):

$$\mathbf{c}_{\mathrm{pos}_n}{}^m = (\mathbf{\Psi}_{\mathrm{pos}_n}{}^m)^{\mathrm{T}}(\mathbf{t}_n^{j,m} - \bar{\mathbf{t}}_{\mathrm{pos}_n}{}^m). \quad (3)$$

These coefficients are used to generate a reconstruction $\mathbf{t}'_{\mathrm{pos}_n}{}^{j,m}$, using (4):

$$\mathbf{t}'_{\mathrm{pos}_n}{}^{j,m} = \bar{\mathbf{t}}_{\mathrm{pos}_n}{}^m + \mathbf{\Psi}_{\mathrm{pos}_n}{}^m \mathbf{c}_{\mathrm{pos}_n}{}^m. \quad (4)$$

The reconstruction is in turn used to compute a reconstruction error vector, whose norm $r_{\mathrm{pos}}(\mathbf{x}_n^{j,m})$ is given by (5):

$$r_{\mathrm{pos}}(\mathbf{x}_n^{j,m}) = \|(\mathbf{t}'_{\mathrm{pos}_n}{}^{j,m} - \mathbf{t}_n^{j,m})\|_2. \quad (5)$$

The same process is followed using the negative subspace for the specific landmark to obtain $r_{\mathrm{neg}}(\mathbf{x}_n^{j,m})$. We then calculate the ratio of the reconstruction error norms $r_{\mathrm{pos}}(\mathbf{x}_n^{j,m})$ and $r_{\mathrm{neg}}(\mathbf{x}_n^{j,m})$ for a particular landmark, using (6):

$$r(\mathbf{x}_n^{j,m}) = \frac{r_{\mathrm{pos}}(\mathbf{x}_n^{j,m})}{r_{\mathrm{neg}}(\mathbf{x}_n^{j,m})}. \quad (6)$$

Fig. 4. The highest scoring aligned initial shapes for each of the $M = 10$ pose models for a facial image from the test set partition of the LFPW dataset.

Next, the mean of these ratios over all N_m landmarks in shape $\mathbf{s}^{j,m}$ is calculated using (7):

$$R^{j,m} = \frac{1}{N_m}\sum_{n=1}^{N_m} r(\mathbf{x}_n^{j,m}). \quad (7)$$

Finally, we use this reconstruction error based metric in combination with knowledge of the number of inliers $N_{\mathrm{inliers}}^{j,m}$ in shape $\mathbf{s}^{j,m}$, i.e., the number of landmarks in the shape that are classified as accurately aligned by our local texture based classifier in our shape scoring function $f(\mathbf{s}^{j,m})$. $f(\mathbf{s}^{j,m})$ is determined using (8) and is the final metric we use to determine the "best" aligned initial shape from among the total set of J_m shapes for pose m:

$$f(\mathbf{s}^{j,m}) = \frac{N_{\mathrm{inliers}}^{j,m}}{R^{j,m}}. \quad (8)$$

The idea here is that a well aligned shape will contain more inliers than a poorly aligned one and hence will end up with a high value for the numerator and a low value for the denominator in (8). The highest scoring aligned shape $\mathbf{s}_{\mathrm{init}}^m$ for each pose from among the J_m evaluated shapes can be determined, using (9) and (10), and used as initialization for the final step in our alignment process:

$$j_0 = \arg\max_j \ f(\mathbf{s}^{j,m}), \quad (9)$$

$$\mathbf{s}_{\mathrm{init}}^m = \mathbf{s}^{j_0,m}. \quad (10)$$

Fig. 4 shows these highest scoring aligned shapes for each pose for a sample test image.

3.3 Shape Refinement

The last stage of our alignment algorithm involves the refining (deforming and regularizing of a shape) of the highest scoring initial shapes that were obtained using the previous stage and the selection of one of these refined shapes as the final locations of the facial landmarks. To carry this out we use an iterative fitting process that has it roots in ASMs and CLMs. In practice, to allow for a gain in fitting speed, only a few $(M' < M)$ of the highest scoring fitting M initial shapes $\mathbf{s}_{\mathrm{init}}^m \ (m = 1, \ldots, M)$ are selected for refinement to obtain shapes $\mathbf{s}_{\mathrm{ref}}^{m'} \ (m' = 1, \ldots, M')$. It is also to be noted that during the refinement process we also score results produced using the open mouth expression shape and texture models for the frontal pose ranges and the higher scoring of the open mouth and closed mouth fitted shapes are retained for each pose m'.

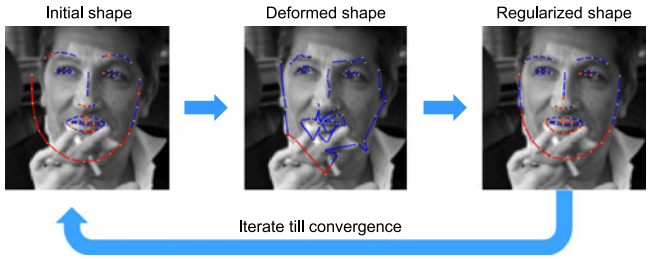


Fig. 5. Iterative process used in our shape refinement step demonstrated on a facial image from the test set partition of the LFPW dataset.

A window around each landmark's current location is generated and the local texture around each pixel in the window is scored and classified using our local texture classifiers. The landmarks are independently moved into the highest scoring locations for them. The process is repeated for a few iterations until the landmarks converge. However, between each iteration, the facial shape produced as a result of landmark motion must be regularized in order to generate a shape that is consistent with what a typical facial contour looks like. We carry out this regularization using a novel technique that allows for a higher fitting accuracy compared to the regularization method employed by ASMs. Fig. 5 illustrates how one iteration of this process is carried out. Finally, the highest scoring shape from among the refined shapes is identified and returned.

Algorithm 1. Overview of Our Approach

Input: Image $\mathbf{I}$, face detection bounding box $\mathbf{b}_b$, pre-trained yaw and expression specific models $\{\mathbf{M}_m\}_{m=1}^{16}$
Output: Final landmarks (shape) $\mathbf{s}_{\text{fin}}$, landmark occlusion/misalignment labels for N landmarks $\{\mathbf{o}_n\}_{n=1}^{N}$
for $m = 1, \ldots, M$ **do**
  Retain top seed landmark candidates $\{\mathbf{p}_1^m, \mathbf{p}_2^m, \ldots, \mathbf{p}_{10}^m\}_{i=1}^{N_m^l}$ for each of the $N_m^l = 5$ or $N_m^l = 8$ seed landmarks for pose m
end for
for $m = 1, \ldots, M$ **do**
  Score all dense shapes $\mathbf{s}^{j,m}$ $(j = 1, \ldots, J_m)$ using (6)-(8)
  Retain highest scoring initial shape $\mathbf{s}_{\text{init}}^m$ using (9) and (10)
end for
Retain top scoring $M' = 4$ pose specific initial shapes
for $m' = 1, , \ldots, M'$ **do**
  Refine $\mathbf{s}_{\text{init}}^{m'}$ to obtain $\mathbf{s}_{\text{ref}}^{m'}$ using model $\{\mathbf{M}_{m'}\}$
end for
Retain highest scoring refined shape $\mathbf{s}_{\text{fin}}$ using (19)-(21)
Output final landmarks (shape) $\mathbf{s}_{\text{fin}}$ and landmark occlusion/misalignment labels for N landmarks $\{\mathbf{o}_n\}_{n=1}^{N}$

3.3.1 ℓ_1-Regularized Least Squares Based Shape Coefficients Determination

The task of shape regularization involves the determining and updating of a vector of shape coefficients. Consider an initial shape $\mathbf{s}_{\text{init}}^{m'}$ (we drop the superscript m' in this section for the sake of simpler notation). After each of the landmarks in the shape have been allowed to independently move into the optimal locations for them, the new shape obtained is denoted by $\mathbf{s}_{\text{def}}$. In an ASM based approach, the inverse of the similarity transformation T that best aligns the mean shape $\bar{\mathbf{s}}$ with $\mathbf{s}_{\text{def}}$ is applied to $\mathbf{s}_{\text{def}}$ (in the image space) to generate $\mathbf{s}'_{\text{def}}$ (in the model space). The problem becomes one of determining the optimal set of shape coefficients $\mathbf{b}_{\text{init}}$ to minimize (11):

$$\mathbf{b}_{\text{init}} = \arg\min_{\mathbf{b}} \ \|\mathbf{\Phi}\mathbf{b} - (\mathbf{s}'_{\text{def}} - \bar{\mathbf{s}})\|_2^2. \tag{11}$$

In (11), $\mathbf{\Phi}$ is a previously trained orthonormal PCA subspace of shape variation (all shapes being aligned using Procrustes analysis before the building of the subspace) with dimensions $d \times u$ $(d > u)$ where $d = 2N$ is the dimension of each shape vector, u is the number of eigenvectors retained in order to account for 95-97 percent of the shape variance (and also the dimensionality of the shape coefficients vector), and $\bar{\mathbf{s}}$ is the mean shape. The solution to the overdetermined least squares problem (LSP) in (11) is given by (12):

$$\mathbf{b}_{\text{init}} = \mathbf{\Phi}^{+}(\mathbf{s}'_{\text{def}} - \bar{\mathbf{s}}). \tag{12}$$

In (12) $\mathbf{\Phi}^{+}$ denotes the left Moore-Penrose pseudoinverse of $\mathbf{\Phi}$. Since $\mathbf{\Phi}$ is an orthonormal basis, $\mathbf{\Phi}^{+} = (\mathbf{\Phi}^{\mathrm{T}}\mathbf{\Phi})^{-1}\mathbf{\Phi}^{\mathrm{T}} = \mathbf{\Phi}^{\mathrm{T}}$ and (12) gets simplified to (13):

$$\mathbf{b}_{\text{init}} = \mathbf{\Phi}^{\mathrm{T}}(\mathbf{s}'_{\text{def}} - \bar{\mathbf{s}}). \tag{13}$$

The values of $\mathbf{b}_{\text{init}}$ are constrained to not lie beyond three standard deviations of their zero mean values (assuming a Gaussian distribution for the coefficient values) in order to generate plausible shapes (regularization) and this results in a new vector of shape coefficients denoted by $\mathbf{b}_{\text{mod}}$. In practice, the similarity transformation parameters and the shape coefficients are determined simultaneously using an iterative procedure that is outlined in [45] and the process we have described is repeated for a few iterations until the shape parameter values do not change by much between iterations. A regularized shape $\mathbf{s}_{\text{reg}}$ is obtained when the final set of shape coefficients are applied and the resulting shape is aligned back into the image space using the transformation T, as shown in (14):

$$\mathbf{s}_{\text{reg}} = T(\bar{\mathbf{s}} + \mathbf{\Phi}\mathbf{b}_{\text{mod}}). \tag{14}$$

In our approach, rather than constructing a PCA subspace to model shape variation, we retain the entire dictionary of shapes for each pose model. Thus, the analogue to the previously defined $\mathbf{\Phi}$ is a dictionary of shape variation $\mathbf{D}$ of size $d \times v$ $(d < v)$, where $d = 2N$ is the dimension of each shape vector in the dictionary and v is the number of such training shapes for a specific yaw model and also the dimensionality of the shape coefficients vector. We recast the problem of shape regularization using (15), in which λ is a regularization parameter, and generate a regularized shape using (16):

$$\hat{\mathbf{b}} = \arg\min_{\mathbf{b}} \ \|\mathbf{D}\mathbf{b} - \mathbf{s}'_{\text{def}}\|_2^2 + \lambda\|\mathbf{b}\|_1 \tag{15}$$

$$\mathbf{s}_{\text{reg}} = T(\mathbf{D}\hat{\mathbf{b}}). \tag{16}$$

What we achieve by formulating the problem in this fashion is that simultaneous determination and regularization of shapes is now possible using a single objective

function without the need for the additional step involved in ASMs to modify the shape coefficients based on the Gaussian assumption. Our formulation makes no assumptions about the distribution of the coefficients, is not a linear function of $\mathbf{s}'_{\text{def}}$ (as is the case in (11)), and allows for a data driven framework to achieve regularization, which is a key area of focus in in [17] and [46] as well.

The problem in (15) is commonly called the ℓ_1-regularized least squares problem whose general form is given by (17):

$$\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1. \tag{17}$$

In (17), $\mathbf{A} \in \mathbb{R}^{p\times q}$ is a data matrix, $\mathbf{y} \in \mathbb{R}^p$ is a vector of observations, $\mathbf{x} \in \mathbb{R}^q$ is a vector of unknowns, and $\lambda > 0$ is the regularization parameter. The problem in (17) is convex but not differentiable. It always has a solution, but there is no closed form equation to obtain it. However, it is possible to compute a solution numerically by transforming the problem into a convex quadratic problem with linear inequality constraints and solving it by standard convex quadratic methods, such as interior-point methods or homotopy methods and variants. However, some of these solvers can be quite slow and also only efficient when the solution is very sparse. It is to be noted that going into details of how these of these solvers work is beyond the scope of this paper, however, we found that a custom interior point based method for solving large scale ℓ_1-regularized LSPs that was developed by Kim et al. [47] was ideally suited for our purposes and is the solver we use in our work. A key result that is outlined in [47] that governs the choice of the regularization parameter λ is that for $\lambda \geq \lambda_{\max} = \|2\mathbf{A}^{\mathrm{T}}\mathbf{y}\|_\infty$ ($\|2\mathbf{D}^{\mathrm{T}}\mathbf{s}'_{\text{def}}\|_\infty$ in our problem setup) an all zero vector becomes the optimal solution. A value of $\lambda = 10^{-4}\lambda_{\max}$ was recommended by Kim et al. (when using an open source MATLAB implementation of their code [48]) and such a value was empirically found to be suitable for the purposes of our problem as well.

Shape regularization can be carried out more accurately if only inliers are used in the process. Since this is possible in our approach, using the results produced by local texture classifiers, we exclude all outliers from participating in the shape regularization process and only use the rows of $\mathbf{D}$ ($\mathbf{\Phi}$ in the case of the previously described PCA based approach that is used by ASMs) that correspond to these inlier landmarks. The shape coefficients obtained using this process can be used to reconstruct a full set of landmarks and hallucinate the locations of the outliers.

An important set of results that we highlight in Section 4.3 is that even when only the inliers are used for shape regularization, our ℓ_1-regularized approach outperforms the previously outlined approach used in ASMs to obtain more accurate fitting results on several datasets. In addition, we also demonstrate that the ℓ_1-regularized approach provides a higher level of accuracy than using an ℓ_2-regularized (Tikhanov regularization) based approach (details on this problem can also be found in [47]), when the same value of λ is used. In such an ℓ_2-regularized approach, a closed form solution to the problem in (18) is provided by $\mathbf{x} = (\mathbf{A}^{\mathrm{T}}\mathbf{A} + \lambda\mathbf{I})^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y}$.

$$\underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{A}\mathbf{x} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_2^2. \tag{18}$$

Our intuition behind these results is that the ℓ_1-regularized approach results in a deformed shape either being modeled using a smaller set of training shapes (in the training shapes dictionary) that best approximate their locations or by using smaller weights for unsuitable training shapes in the dictionary. This sparsity promotion results in a deformed shape better approximated using a set of coefficients that is not a linear function of the deformed shape itself, and thus in higher fitting accuracies than those obtained using an ℓ_2-regularized approach.

3.3.2 Final Shape Scoring and Selection

The last step in our alignment process is the selection of a single shape from among the set of $M' = 4$ (in our implementation) refined shapes that best fits the underlying facial structure. The shape $\mathbf{s}_{\text{fin}}$ with the highest percentage of inliers, expressed as a percentage of the total number of landmarks in the shape using the scoring function $g(\mathbf{s}^{m'}_{\text{ref}})$ in (19), is chosen to obtain a final set of landmark coordinates using the following equations:

$$g(\mathbf{s}^{m'}_{\text{ref}}) = \frac{N^{\text{inliers}}_{m'}}{N_{m'}} \quad \text{where } m' = 1, \ldots, M', \tag{19}$$

$$m'_0 = \arg\max_{m'} \; g(\mathbf{s}^{m'}_{\text{ref}}), \tag{20}$$

$$\mathbf{s}_{\text{fin}} = \mathbf{s}^{m'_0}_{\text{ref}}. \tag{21}$$

The same scoring function is also used to determine the best fitting expression specific shape for each pose m'. Algorithm 1 summarizes the steps in our approach.

4 Experiments and Results

In this section we provide details on how our approach was trained and describe the experiments that we carried out in order to demonstrate the effectiveness of our algorithm when it was tested on challenging real-world datasets.

4.1 Training Our Algorithm

We trained a set of models using a subset of the CMU Multi-PIE (MPIE) database. Our shape and texture models were trained using a total of $6{,}495$ images of various subjects drawn from across all four sessions that spanned 13 sets of viewpoints from -90 to $+90$ degree in steps of 15 degree and various expressions (neutral, disgust, smile, squint, scream, and surprise). Manually annotated ground truths for all these images were available to us as a small subset of the MPIE database was annotated using 68 landmarks for frontal faces and 39 landmarks for profile faces by the database's curators. We clustered the data into $M = 10$ bins with overlapping yaw ranges and the same number of facial landmarks for every image in the bin, i.e., -90 to -75 degree, -75 to -60 degree, -45 to -30 degree, -30 to -15 degree, -15 to 0 degree, and five more similar bins for the positive yaw angles. These 10 partitions were created using facial images with the mouth slightly open or closed (neutral, disgust, smile, and squint expressions). A similar set of six partitions (for frontal poses with yaw angles from -45 to $+45$ degree only) were created to model the shape and texture of facial landmarks across pose in expressions

when the mouth is completely open (scream and surprise expressions).

Our models were built using the process described in Section 3 with Real AdaBoost as our choice of classifier. However, when we tested on a subset of the MPIE dataset, we only trained on three-fourths of the training data with a test set drawn from the remaining images. Thus, we always tested our algorithm on unseen images and subjects. We used an open source MATLAB toolbox [49] to extract the HOG features and implement the training and testing of the Real AdaBoost framework. A standard facial crop size of 100×100 and a patch size of 15×15 were used by us to extract HOG feature descriptors and build the local texture models (classifiers).

4.2 Benchmarking Our Approach

We compared the fitting accuracy of our approach against that of various existing state-of-the-art methods (an overview of these methods has been provided in Section 2) on several challenging real-world datasets. The approaches (all of which have open source code available) we compared ours against are those proposed by Tzimiropoulos and Pantic [16] (henceforth denoted/abbreviated as AAM-Wild), Zhu and Ramanan [18] (henceforth denoted/abbreviated as Tree-DPMs) using their pre-trained and best performing *Independent-1050* model, Yu et al. [21] (henceforth denoted/abbreviated as CDSM), Asthana et al. [22] (henceforth denoted/abbreviated as DRMF), Xiong and De la Torre [23] (henceforth denoted/abbreviated as SDM), and Burgos-Artizzu et al. [25] (henceforth denoted/abbreviated as RCPR).

Details on the various datasets we used to benchmark all of these approaches are provided below.

1) MPIE: A set of 850 images were held back from our training set and served as a test set of MPIE images containing faces with varying expressions and yaw angles in the range from -90 to $+90$ degree. This test set was created to demonstrate that our algorithm could deal with such variations in unseen images from outside its training set and was also used to benchmark our approach against the Tree-DPMs algorithm, which could also handle this range of yaw variation.
2) LFPW: The Labeled Face Parts in the Wild [17] dataset originally consisted of $1{,}132$ training images and 300 test images of various people (mainly celebrities) that were collected from the Internet and manually annotated with 29 landmarks. Many of the URLs for the images in the dataset have expired, however, a set of 811 training images and 224 test images were recently made available along with landmark annotations [50], [51] for the 68 landmarks in the MPIE markup as part of the the 300 Faces in-the-wild (300-W 2013) challenge [52], [53]. All algorithms were tested on the 224 images in the test set partition of the dataset.
3) AFW: The Annotated Faces in-the-Wild (AFW) dataset [18] consists of 205 images with 468 faces (some images contain multiple faces) drawn from Flickr images. Facial bounding boxes, manual annotations for 6 landmarks (the centers of the eyes, the tip of the nose, and the two corners and center of the mouth), and discretized pose information were originally made available along with the images. As part of the 300-W 2013 challenge, 68 point annotations for 337 faces in the images were made available [50], [51], which served as a test set.
4) ibug: The ibug dataset [53], [54] consists of 135 facial images with annotations for 68 landmarks for each of the faces. The dataset was made publicly available as part of the 300-W 2013 challenge.
5) COFW: The Caltech Occluded Faces in the Wild dataset [25] consists of 500 training and 507 test images that were downloaded from the Internet. All images were manually annotated with the same 29 landmarks that were used in the exemplar based facial alignment method proposed by Belhumeur et al. [17]. The faces in the images exhibit slight pose variation (absolute yaw angles of up to 30-45 degree and sometimes severe in-plane rotation), varying expressions, and large levels of occlusion (the average level of occlusion of faces due to hats, sunglasses, food, etc. in the dataset is 23 percent). The dataset was mainly proposed to push the boundaries of occlusion tolerance by facial alignment algorithms and thus also provides occlusion labels for each landmark. All algorithms were evaluated on the 507 images in the test set partition of the dataset.

It is to be noted that while some of the approaches (Tree-DPMs, DRMF, and CDSM) were trained on MPIE images, similar to the data our approach was trained on, some of the other approaches were at an advantage as they were trained on more unconstrained real-world data. For example, the AAM-Wild approach was trained on the training set partition of the LFPW dataset, RCPR was trained on the training set partitions of the LFPW and COFW datasets (to ensure the best fitting results on the challenging test set partition of the COFW dataset), and the SDM implementation we used was trained on images from the MPIE and LFW datasets [55]. Thus, in order to perform a fair comparison and to demonstrate the importance of training data when dealing with real-world images, we report results obtained by our approach when it was trained using the MPIE images we previously mentioned, the 811 images in the training set partition of the LFPW dataset (when fitting images from the LFPW test set, AFW, and ibug datasets), and the 500 COFW and 845 LFPW training set images with the 29 landmarks and occlusion labels that RCPR was trained on (when testing on the COFW test set images). Our models were built in the same fashion as previously described (using clustering of images into appropriate pose and expression groups) with appropriate changes to account for a different set of landmarks and a lack of images to model absolute yaw variation in excess of 45 degree. We also report results obtained by training the RCPR algorithm (trained using the optimal parameter values specified by the authors when training on un-occluded images) on the same set of MPIE images (images with yaw angles between -45 and $+45$ degree and 68 ground truth annotations) as our approach and using the 68 point landmarking scheme. As we will show, our approach performs admirably when

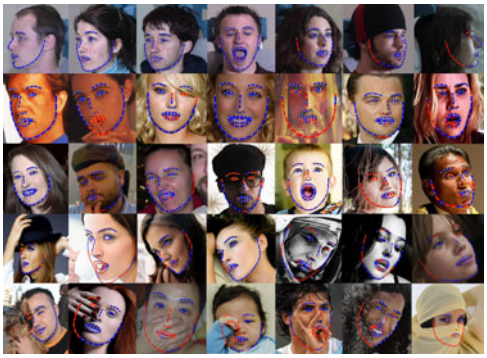
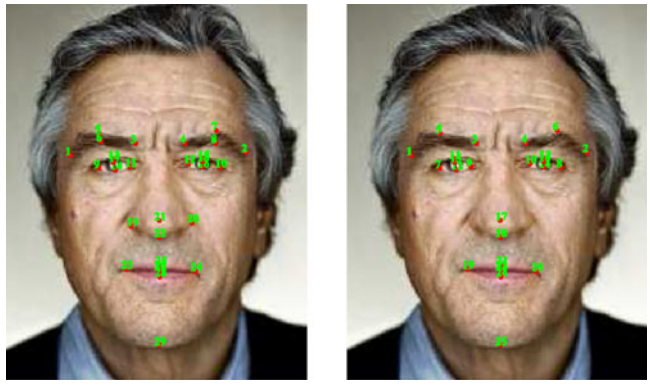


Fig. 6. Qualitative landmark localization results produced by our approach (when trained on images from the MPIE dataset) on some images from the MPIE, LFPW, AFW, ibug, and COFW datasets from the top to bottom rows, respectively.



(a) (b)

Fig. 7. (a) The 29 point landmarking scheme for the COFW dataset and (b) The 25 landmarks common to both the 68 point MPIE landmarking scheme and the ground truth annotations available for the COFW dataset. The facial image in the figure is from the training set partition of the COFW dataset.

trained on only MPIE images (sample qualitative results produced by our facial alignment algorithm trained on MPIE images on images from the various datasets are shown in Fig. 6) and provides much higher accuracy levels than all the other approaches when trained on the real-world LFPW and COFW training set images.

The other key aspect to consider when reporting a fair performance comparison of facial alignment algorithms is initialization (facial bounding boxes provided as input). The DRMF and SDM algorithms are extremely sensitive to these inputs and it was observed that they produced extremely poor results when we used the bounding box initializations for the AFW, LFPW test, and ibug datasets [53] provided by the organizers of the 300-W 2013 challenge that were obtained using what was referred to as their "in-house face detector." Thus, for these approaches we used an OpenCV implementation of the Viola-Jones [56] face detection algorithm to provide bounding box initializations for these datasets whenever the face detection results produced were in close agreement with the provided bounding boxes and by reverting to suitably modified versions (square bounding boxes) of the provided bounding boxes whenever spurious/no detections were made by the OpenCV face detector. For initializing the RCPR algorithm (trained on MPIE images), that is also sensitive to bounding boxes provided, on these three datasets, we used the same process as during training and provided bounding boxes that were crops around then ground truth landmark locations grown by 15 percent. For the AAM-Wild algorithm, our approach, CDSM, and Tree-DPMs, we used the used the bounding box initializations provided by the organizers of the 300-W 2013 challenge with the crop grown by a factor of 1.5 to enclose the facial region in the latter three cases. This was carried out because though the CDSM and Tree-DPMs approach function as face detectors, a fair comparison of landmark fitting accuracy demands that an appropriate region of interest be provided. Our approach does not detect faces and can fail in the event of extremely poor face detection results, however we did not train our approach assuming specific details about the bounding boxes available during testing. Thus, specifying a large region of interest for our initial seed landmark detection stage is sufficient to deal with slight scale and translation differences in face detection bounding boxes and we were thus in a position to use bounding boxes that we had no information about during our training stage. For the COFW test set we observed that the facial bounding boxes provided along with the ground truth landmarks were generally not optimal for many of the alignment algorithms (DRMF, AAM-Wild, and SDM) as they only partially enclosed the facial region in many cases. In order to ensure better initialization, a square crop was generated to enclose the ground truth landmark locations and then grown by 15 percent before being provided as initialization (for the DRMF and SDM algorithms, these bounding boxes were only used when OpenCV implementation of the Viola-Jones face detection algorithm could not provide accurate bounding boxes). For CDSM, Tree-DPMs, and our approach, this crop region was further expanded by a factor of 1.5 to enclose the facial region and yet not provide any initialization advantage. For testing the RCPR approach (trained on MPIE images) on the COFW test set images we used the same initialization protocol as we did when evaluating it on the other test sets. We also evaluated our approach trained in identical fashion to the RCPR approach on the same set of COFW and LFPW images with 29 landmarks with both algorithms initialized using the bounding boxes provided along with the COFW test set images. Finally, for the MPIE dataset, a square crop around the convex hull of the ground truth landmark locations was extracted and then grown by a factor of 1.5 before being provided as initialization to Tree-DPMs and our approach.

Most of the approaches use the same 68 point landmarking scheme to annotate frontal facial images, making a fair comparison possible on the LFPW, AFW, and ibug test datasets. However, the SDM algorithm localizes 49 facial landmarks (does not localize landmarks $1-17$ (facial boundary points) and landmarks 61 and 65 (interior points near the corners of the mouth) in Fig. 2(b). Thus, our results are reported for both these cases and by utilizing the maximum possible common landmarks localized by the various algorithms. For the COFW dataset, where only 29 manually annotated landmarks are available, we measured the fitting accuracy of the algorithms using a set of 25 landmarks (and 24 for the SDM method) which are common to both the 29 point and 68 point markups. This set of landmarks is shown in Fig. 7(b). However, we also provide results that

TABLE 1
Performance of Various Algorithms on Various Test Sets with Statistics Computed Using 68 (or 39 for the MPIE Test Set) Common Landmarks (Except in Cases Where an Alternative Number Is Indicated in Brackets)

Algorithm	MPIE		LFPW		AFW		ibug		COFW	
	MNFE (%)	Failure (%)	MNFE (%)	Failure (%)	MNFE (%)	Failure (%)	MNFE (%)	Failure (%)	MNFE (%)	Failure (%)
Ours (Best Models)			**4.98**	**3.17**	**7.10**	**9.42**	**9.95**	**25.58**	**6.00 (29)**	**6.31 (29)**
Ours (MPIE Tr Set)	5.37	2.88	6.68	5.88	8.79	19.16	13.18	51.16	8.53 (25)	20.92 (25)
DRMF	—	—	6.77	9.95	11.81	26.95	19.40	66.28	10.32 (25)	30.33 (25)
CDSM			7.63	13.18	10.30	34.10	19.57	81.48	9.73 (25)	23.76 (25)
Tree-DPMs	6.68	9.77	8.99	30.32	10.72	35.71	25.46	72.09	9.58 (25)	30.13 (25)
RCPR (COFW + LFPW Tr Sets)	—	—							6.16 (29)	9.27 (29)
RCPR (MPIE Tr Set)	—	—	8.10	18.55	12.54	34.74	20.14	70.93	12.47 (25)	36.61 (25)
AAM-Wild	—	—	12.41	40.27	17.75	60.71	27.88	87.21	12.03 (25)	53.56 (25)

compare our approach against the RCPR approach on the full set of 29 landmarks, shown in Fig. 7(a). The MPIE dataset is the only one that contains facial images with absolute yaw in excess of 45 degree and is hence used to only compare our algorithm against Tree-DPMs, as none of the other algorithms provide localization results using the same 39 point landmarking scheme that are shown in Fig. 2(a) or handle such yaw variation.

To compare the fitting accuracy of the various algorithms, the fitting error (the euclidean distance between the automatically fitted landmarks and their corresponding manually annotated ground truth locations) was normalized for each image using the distance between the corners of the eyes (landmarks 37 and 46 for the frontal landmarking scheme in Fig. 2(b), landmarks 9 and 10 in Fig. 7(a), and landmarks 7 and 8 in Fig. 7(b), as was carried out in the 300-W 2013 challenge, to enable a fair comparison across all images (of varying resolution and facial sizes) in the datasets. For the MPIE dataset the average eye center to mouth corner distance was used for normalization as this dataset contained images with profile views. These distances were averaged over all landmarks to produce a normalized fitting error for each image in the dataset. The Mean Normalized Fitting Error (MNFE) of these fitting errors, calculated by averaging the normalized fitting error over all images in the test dataset (and expressed as a percentage), is the common metric commonly employed to determine the accuracy of a facial alignment algorithm. Another metric that is used to compare the approaches is the failure rate. This is computed as the percentage of the total images fitted that have an MNFE value of over 10 percent of the normalization distance, a measure that was proposed in [57].

Due to a lack of correspondence between landmarks in the 68 and 39 point landmarking schemes, we report results over (average over) only those images where the Tree-DPMs method determined a set of 68 landmarks. Table 1 lists the MNFE and failure rate values obtained by the various approaches on the different test sets over the largest number (most commonly 68 landmarks) of common landmarks while Table 2 lists the same values when only interior facial landmarks (mostly commonly 49 landmarks) are considered and also has results obtained by the SDM algorithm implementation, which does not localize landmarks along the facial boundary. For the COFW dataset case, this corresponded to the exclusion of just the tip of the chin. Predictably, all methods demonstrated a higher accuracy when localizing only the interior landmarks. As can be seen from the tables, our approach performs quite well when trained only on images from the MPIE database. However, the best performance (indicated by best models) for our approach is achieved on the LFPW, AFW, and ibug test sets when trained on LFPW training set images and on the COFW test set when trained on COFW and LFPW training set images (in a similar fashion to the RCPR algorithm). Our best performing models provide more accurate results than the other algorithms across all the test sets and demonstrate the

TABLE 2
Performance of Various Algorithms on Test Sets with Statistics Computed Using 24 (for the COFW Test Set), 49 or 27 (for the MPIE Test Set), and 49 (for All Other Test Sets) Common Interior Landmarks Localized by the Various Algorithms

Algorithm	MPIE		LFPW		AFW		ibug		COFW	
	MNFE (%)	Failure (%)	MNFE (%)	Failure (%)	MNFE (%)	Failure (%)	MNFE (%)	Failure (%)	MNFE (%)	Failure (%)
Ours (Best Models)			**4.19**	**0.90**	**6.00**	**5.20**	**8.34**	**19.77**	**5.90**	**5.33**
Ours (MPIE Tr Set)	4.38	1.25	6.09	3.62	8.02	14.94	11.97	36.05	8.34	19.87
SDM	—	—	5.00	4.07	9.83	15.26	15.90	36.05	6.65	11.92
DRMF	—	—	5.92	8.14	10.68	18.51	16.80	54.65	10.14	29.71
CDSM			6.10	6.36	8.40	19.02	16.78	54.32	9.49	22.25
Tree-DPMs	5.08	2.63	7.30	14.48	9.19	22.73	23.11	58.14	9.27	27.62
RCPR (COFW + LFPW Tr Sets)	—	—							6.00	8.68
RCPR (MPIE Tr Set)	—	—	7.79	16.29	12.67	31.17	20.11	65.12	12.33	35.98
AAM-Wild	—	—	12.07	39.82	17.80	58.12	28.42	83.72	11.57	50.84

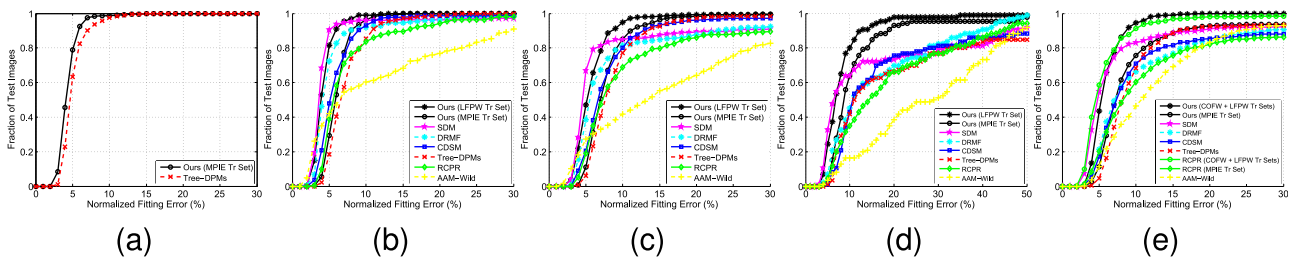


Fig. 8. Cumulative Error Distribution curves for various algorithms obtained by averaging normalized fitting errors (percent) over common interior landmarks (24 for the COFW test set, 49 or 27 for the MPIE test set, and 49 for all other test sets) on the (a) MPIE, (b) LFPW, (c) AFW, (d) ibug, and (e) COFW test sets.

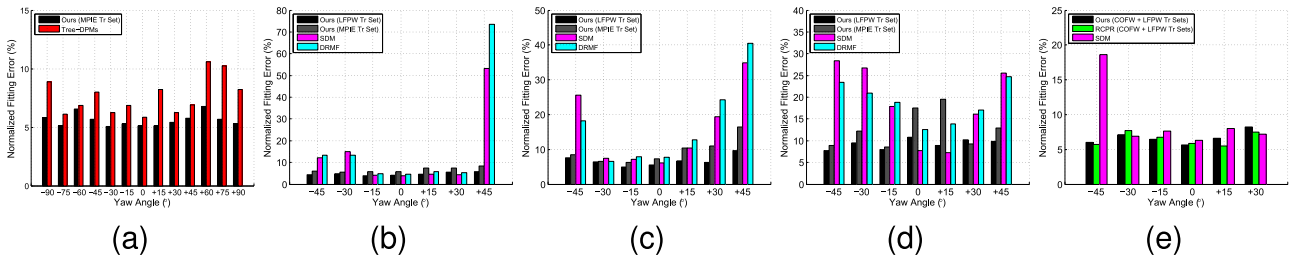


Fig. 9. Normalized fitting errors (percent) as a function of yaw angle for the top performing algorithms calculated using a common set of landmarks (24 for the COFW test set, 68 or 39 for the MPIE test set, and 49 for all other test sets) on the (a) MPIE, (b) LFPW, (c) AFW, (d) ibug, and (e) COFW test sets.

efficacy of our approach as well as the role the training set plays when reporting such accuracy rates.

An alternative way of comparing the accuracy of the methods is using Cumulative Error Distribution (CED) curves that plot the fraction of facial images (plotted along the y-axis) found to have an normalized fitting error (percent) value lower than a certain value (plotted along the x-axis). CED curves summarizing the performance of the various methods on the various datasets are shown in Fig. 8 using the same number of landmarks as those in Table 2. From Fig. 8 it is again clear that our approach (best performing models) localizes landmarks more accurately than all the other algorithms on all the test sets.

We also determined how pose, expression, and occlusion factors influenced the fitting results of the top performing algorithms on each test set. Fig. 9 shows how the normalized fitting error (percent) values vary as a function of pose for the various algorithms across al images in the test sets. The images in each test set were clustered into various pose (yaw angle) bins (by comparing the ground truth landmarks to those of images in the MPIE database) and the average of the normalized fitting errors for all the images belonging to that pose bin were calculated and plotted as a function of the yaw angle. In similar fashion, graphs (see Fig. 10) were obtained to determine how fitting errors varied by expression on the MPIE test set (for which expression labels were available) and by level of occlusion on the COFW test dataset (for which individual landmark occlusion labels were available). As can be seen in Fig. 9, our approach (best models or models trained using MPIE images) provides a consistent level of performance across the various yaw angles and demonstrates a higher tolerance to faces with yaw angles of ± 45 degree than SDM and DRMF. This tolerance to pose is particularly evident on the ibug and AFW datasets that have a larger number of images with high yaw angles when compared to the LFPW and COFW test sets. Similarly, Fig. 10(a) shows how our approach provides consistent results on the MPIE test set for varying expressions, while Fig. 10(b) serves in making the same point for the varying level of occlusion in the COFW test set.

We also provide details on the occlusion prediction performance (over 29 landmarks) of RCPR and our approach (when both were trained on the same set of LFPW and COFW training set images and provided real valued occlusion labels that had to be thresholded to produce binary occlusion labels) on the COFW test set in Table 3. The metrics used in the table are the accuracy $((TP + TN)/(P + N)$,

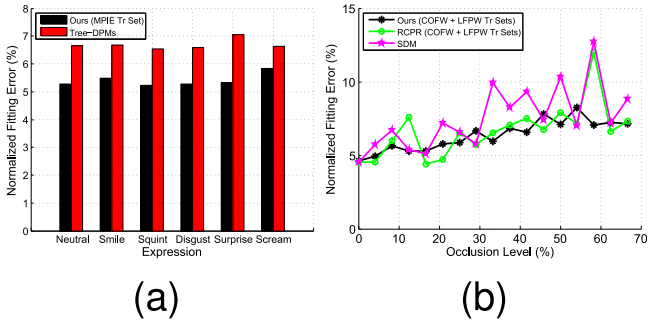


Fig. 10. Normalized fitting errors (percent) obtained (using a common set of 68 or 39 landmarks on the MPIE test set and 24 landmarks on the COFW test set) using various algorithms on faces with various expressions and occlusion levels on the (a) MPIE and (b) COFW test sets, respectively.

TABLE 3
Occlusion Prediction Performance of RCPR and Our Algorithm (Both Trained UsingCOFW and LFPW Training Set Images) When Localizing 29 Landmarks on the COFW Test Set

Algorithm	Accuracy (%)	True Positive Rate (%)	False Positive Rate (%)
RCPR (thresh = 0.6100)	80.62	20.32	1.51
RCPR (thresh = 0.2445)	82.88	59.25	10.11
Ours	81.25	52.08	10.11

TABLE 4
Average Fitting Times per Image for Various Algorithms

Algorithm	Avg. fitting time per image (secs)
SDM	≈ 0.10
RCPR	≈ 1
DRMF	≈ 2
AAM-Wild	≈ 4
CDSM	≈ 5
Tree-DPMs	≈ 18
Ours	≈ 33

where TP is the number of true positive detections, TN is the number of true negative detections, P is the total number of positive samples, N is the total number of negative samples), true positive (TP/P), and false positive (FP/N) rates. As can be seen, our approach provides a higher accuracy rate than RCPR when the default (precomputed using the training data) occlusion detection threshold value of 0.61000 is used for RCPR. However, RCPR provides marginally more accurate results for a different threshold value of 0.24445 (determined from a Receiver Operating Characteristic (ROC) curve using the same false positive rate as our approach).

Finally, to complete the benchmarking analysis, we also provide a timing analysis of the various approaches in Table 4. The table lists the average time required for the various approaches to fit an image on a desktop computer with an Intel Xeon X5680 processor with a clock rate of 3.33 GHz running Windows 7. While our approach presently requires a larger amount of time to process an image, it is to be noted that this our implementation is currently purely MATLAB based and is not heavily optimized for speed and is something that we plan on addressing. Our approach also contains many tasks that lend themselves to parallelization that we haven't taken advantage of yet. GPUs and GPU programming could be used to obtain massive speedups to our implementation and is an area of work we wish to address in the future.

4.3 Evaluation of Our Approach

Table 5 shows a time breakdown and the MNFE (percent) values obtained at each stage of our approach across all datasets when fitting errors were averaged across all landmarks localized at each respective stage. Stage 1 corresponds to the seed landmark localization stage (when only eight landmarks are localized for frontal images and five for profile images) and the best localized candidates for each seed point are compared to their respective ground truth locations, Stage 2 corresponds to the dense landmark alignment and optimal shape initialization stage, and finally Stage 3 is the refinement stage. The seed landmarks detection stage provides much higher accuracy rates on the test sets with lower occlusion levels and is impacted most heavily on the COFW test set, when large occlusion levels result in high fitting errors for the occluded seed landmarks. Stage 2 transitions from a sparse set to a dense set of landmarks. However, this stage provides only a course alignment using a two seed landmark candidates and is thus substantially improved upon by the refinement stage, that provides the final landmark coordinates as output. The importance of the role of Stage 2 in the fitting pipeline (that we alluded to in Section 3.2) is easy to understand using Table 5, especially in cases where large occlusions levels are present and when it is not apparent which seed landmarks have been accurately localized. We have also provided numbers on the best case results that can be obtained using our approach. These values were obtained by comparing all finally refined shapes (not just the finally picked highest scoring one) against the ground truths and choosing the shape with the lowest fitting error. This indicates that a common error that occurs in our approach is at the final step when a single set of landmarks has to be chosen from among the $M^{'}$ refined shapes and is an area for possible improvement. In a semi-supervised setting, a human in the loop would be able to inspect the $M^{'}$ refined shapes and pick an appropriate set of landmarks in such cases.

In closing, we provide some justification for our novel addition to the shape regularization stage. For each of the datasets, we selected all images with an MNFE (percent) lower than 10 percent (over 68 landmarks) and refit these images at the final shape refinement stage using the shape regularization technique that has been used in prior ASM implementations and described in Section 3, as well as by using an ℓ_2-regularized approach to solve the problem in (18), instead of the problem in (17). It is evident from Table 6 that our ℓ_1-regularized approach consistently provides more accurate results, albeit at an increased computational cost, than both these approaches and serves to justify why our ℓ_1 based shape fitting approach is an important contribution to the facial landmark localization procedure.

TABLE 5
MNFE (Percent) Values Obtained on Test Sets by Each Stage of Our Approach by Averaging Over the Maximum Number of Landmarks Localized at that Stage

Test Set	Ours (MPIE Tr Set)				Ours (Best Models)			
	Stage 1 (≈ 3 sec)	Stage 2 (≈ 20 sec)	Stage 3 (≈ 10 sec)	Best Result	Stage 1	Stage 2	Stage 3	Best Result
	MNFE (%)	MNFE (%)	MNFE (%)	MNFE (%)	MNFE (%)	MNFE (%)	MNFE (%)	MNFE (%)
MPIE	5.48	9.21	5.40	5.17				
LFPW	7.43	9.23	6.70	6.12	$5.\overline{43}$	$7.\overline{41}$	$5.\overline{00}$	$4.\overline{55}$
AFW	10.16	11.90	9.25	8.09	7.75	10.15	7.23	6.53
ibug	14.58	16.19	13.38	10.96	12.00	14.42	10.85	9.34
COFW	10.91	10.21	8.74	7.39	8.61	7.67	6.06	5.88

The average time (in secs) taken by each stage when fitting an image is also indicated.

TABLE 6
Comparison of MNFE (Percent) Values Obtained by Averaging Over 25 Landmarks on the COFW Test Set, 68 or 39 Landmarks on the MPIE Test Set, and 68 Landmarks on All Other Test Sets Using Various Shape Regularization Approaches with Models Trained on MPIE Images

Test Set	**Ours (ℓ_1-regularized)**	Ours (ℓ_2-regularized)	ASM Method (PCA based)
	MNFE (%)	MNFE (%)	MNFE (%)
MPIE	**5.15**	5.24	5.81
LFPW	**6.36**	6.92	7.24
AFW	**7.06**	7.70	8.01
ibug	**8.03**	8.20	8.64
COFW	**6.69**	7.33	7.50

5 Conclusion

We have presented a new facial alignment framework to jointly deal with the problems posed by facial pose, expressions, and occlusions that included a novel ℓ_1-regularized least squares approach to the shape fitting stage that ensured higher accuracy rates than those achieved by previously used shape models. Our approach also provides misalignment/occlusion labels for each fitted facial landmark, which is something many existing approaches do not do. We demonstrated the superiority of our approach over several existing state-of-the-art algorithms on challenging real-world datasets and also provided proof of its consistent performance across varying facial pose, expressions, and occlusion levels.

Acknowledgments

This work was supported by the Federal Highway Administration's Exploratory Advanced Research Program under contract DTFH61-14-C-00006. The authors would also like to thank Dr. Craig Thor for his support and feedback on this work.

References

[1] G. J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," in *Proc. 5th Eur. Conf. Comput. Vis.*, Jun. 1998, pp. 581–595.
[2] N. Faggian, A. Paplinski, and T. J. Chin, "Face recognition from video using active appearance model segmentation," in *Proc. 18th Int. Conf. Pattern Recog.*, Aug. 2006, vol. 1, pp. 287–290.
[3] R. Abiantun, U. Prabhu, K. Seshadri, J. Heo, and M. Savvides, "An analysis of facial shape and texture for recognition: A large scale evaluation on FRGC ver2.0," in *Proc. IEEE Workshop Appl. Comput. Vis.*, Jan. 2011, pp. 212–219.
[4] T. H. N. Le, K. Luu, K. Seshadri, and M. Savvides, "A facial aging approach to identification of identical twins," in *Proc. IEEE Int. Conf. Biometrics: Theory, Appl. Syst.*, Sep. 2012, pp. 1–8.
[5] J. Heo and M. Savvides, "3-D generic elastic models for fast and texture preserving 2-D novel pose synthesis," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 2, pp. 563–576, Apr. 2012.
[6] O. Rudovic, I. Patras, and M. Pantic, "Regression-based multi-view facial expression recognition," in *Proc. Int. Conf. Pattern Recog.*, Aug. 2010, pp. 4121–4124.
[7] H. C. Choi and S. Y. Oh, "Real-time recognition of facial expression using active appearance model with second order minimization and neural network," in *Proc. IEEE Int. Conf. Syst., Man, Cybern.*, Oct. 2006, pp. 1559–1564.
[8] K. Luu, K. Seshadri, M. Savvides, T. D. Bui, and C. Y. Suen, "Contourlet appearance model for facial age estimation," in *Proc. IEEE Int. Joint Conf. Biometrics*, Oct. 2011, pp. 1–8.
[9] T. H. N. Le, K. Luu, K. Seshadri, and M. Savvides, "Beard and mustache segmentation using sparse classifiers on self-quotient images," in *Proc. IEEE 19th Int. Conf. Image Process.*, Sep. 2011, pp. 165–168.
[10] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham, "Active shape models—their training and application," *Comput. Vis. Image Understanding*, vol. 61, no. 1, pp. 38–59, Jan. 1995.
[11] T. F. Cootes, C. J. Taylor, and A. Lanitis, "Active shape models: Evaluation of a multi-resolution method for improving image search," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 1994, pp. 32.1–32.10.
[12] D. Cristinacce and T. Cootes, "Feature detection and tracking with constrained local models," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2006, pp. 929–938.
[13] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recog.*, vol. 41, no. 10, pp. 3054–3067, Oct. 2008.
[14] J. M. Saragih, S. Lucey, and J. F. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, Jan. 2011.
[15] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.
[16] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast AAM fitting in-the-wild," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 593–600.
[17] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2011, pp. 545–552.
[18] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2879–2886.
[19] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part based models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1627–1645, Sep. 2010.
[20] G. Ghiasi and C. C. Fowlkes, "Occlusion coherence: Localizing occluded faces with a hierarchical deformable part model," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1899–1906.
[21] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1944–1951.
[22] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3444–3451.
[23] X. Xiong and F. De la Torre, "Supervised descent method and its application to face alignment," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 532–539.
[24] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2887–2894.
[25] X. P. Burgos-Artizzu, P. Perona, and P. Dollar, "Robust face landmark estimation under occlusion," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1513–1520.
[26] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 FPS via regressing local binary features," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1685–1692.
[27] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1867–1874.
[28] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 3476–3483.
[29] M. Everingham, J. Sivic, and A. Zisserman, "'Hello! my name is... Buffy' – automatic naming of characters in TV video," in *Proc. Brit. Mach. Vis. Conf.*, Sep. 2006, pp. 889–908.
[30] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," in *Proc. IEEE Int. Conf. Face Gesture Recog.*, Sep. 2008, pp. 1–8.
[31] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.
[32] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Mag.*, vol. 2, no. 6, pp. 559–572, 1901.
[33] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *J. Educational Psychol.*, vol. 24, no. 6, pp. 417–441, 498–520, Oct. 1933.

[34] H. Hotelling, "Relations between two sets of variates," *Biometrika*, vol. 28, nos. 3/4, pp. 321–377, Dec. 1936.
[35] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2005, vol. 1, pp. 886–893.
[36] B. Tamersoy, C. Hu, and J. K. Agarwal, "Nonparametric facial feature localization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2013, pp. 838–845.
[37] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Learn to combine multiple hypotheses for accurate face alignment," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 392–396.
[38] X. Zhao, S. Shan, X. Chai, and X. Chen, "Cascaded shape space pruning for robust facial landmark detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1033–1040.
[39] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: A statistical view of boosting," *Ann. Statist.*, vol. 28, no. 2, pp. 337–407, Dec. 2000.
[40] Y. Freund and R. E. Schapire, "A short introduction to boosting," *J. Japanese Soc. Artif. Intell.*, vol. 14, no. 5, pp. 771–780, Sep. 1999.
[41] R. E. Schapire and Y. Singer, "Improved boosting algorithms using confidence-rated predictions," *Mach. Learn.*, vol. 37, no. 3, pp. 297–336, Dec. 1999.
[42] K. Seshadri and M. Savvides, "An analysis of the sensitivity of active shape models to initialization when applied to automatic facial landmarking," *IEEE Trans. Inf. Forensics Security*, vol. 7, no. 4, pp. 1255–1269, Aug. 2012.
[43] J. C. Gower, "Generalized procrustes analysis," *Psychometrika*, vol. 40, no. 1, pp. 33–51, Mar. 1975.
[44] C. Goodall, "Procrustes methods in the statistical analysis of shape," *J. Roy. Statistical Soc.. Series B (Methodological)*, vol. 53, no. 2, pp. 285–339, 1991.
[45] T. F. Cootes and C. J. Taylor, "Statistical models of appearance for computer vision," Imag. Sci. Biomed. Eng., Univ. Manchester, Manchester, U.K., Mar. 2004, www.face-rec.org/algorithms/AAM/app_models.pdf
[46] B. M. Smith, J. Brandt, Z. Lin, and L. Zhang, "Nonparametric context modeling of local appearance for pose and expression-robust facial landmark localization," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 1741–1748.
[47] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky, "An interior-point method for large-scale ℓ_1-regularized least squares," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 4, pp. 606–617, Dec. 2007.
[48] S. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. (2007). l1_ls: Simple matlab solver for l1-regularized least squares problems [Online]. Available: http://www.stanford.edu/~boyd/l1_ls/
[49] P. Dollár. (2005). Piotr's computer vision mAtlab toolbox (PMT) [Online]. Available: http://vision.ucsd.edu/~pdollar/toolbox/doc/index.html
[50] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "A semi-automatic methodology for facial landmark annotation," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog. Workshops*, Jun. 2013, pp. 896–903.
[51] (2013). Facial Point Annotations [Online]. Available: http://ibug.doc.ic.ac.uk/resources/facial-point-annotations/
[52] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *Proc. IEEE Int. Conf. Comput. Vis. Workshops*, Dec. 2013, pp. 397–403.
[53] (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge [Online]. Available: http://ibug.doc.ic.ac.uk/resources/300-W/
[54] (2013). Intelligent behaviour understanding group (iBUG) dataset [Online]. Available: http://ibug.doc.ic.ac.uk/download/annotations/ibug.zip/
[55] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Univ. Massachusetts, Amherst, MA, Tech. Rep. 07-49, Oct. 2007.
[56] P. Viola and M. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, May 2004.
[57] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 2578–2585.

Keshav Seshadri received the BTech degree in electronics and communication engineering from Visvesvaraya National Institute of Technology (VNIT), Nagpur, India, in 2007, and the MS degree in electrical and computer engineering (ECE) from Carnegie Mellon University, in 2008. He is currently working toward the PhD degree at Carnegie Mellon's ECE Department. His research is focused on applying machine learning and pattern recognition techniques to the tasks of automatic facial landmark localization and facial recognition. He is a student member of the IEEE.

Marios Savvides is a research professor at the Electrical and Computer Engineering (ECE) Department, Carnegie Mellon University, and is the director in the CMU CyLab Biometrics Center. His research is mainly focused on developing algorithms for robust face and iris recognition as well as in using pattern recognition, machine vision, and computer image understanding for enhancing the performance of real-time biometric systems. He is a program committee member on several biometric conferences such as IEEE BTAS, SPIE Biometric Identification, and IEEE AutoID. He has organized and co-chaired the Robust Biometrics Understanding the Science and Technology (ROBUST) 2008 conference. He has authored and coauthored more than 120 journal and conference publications, including several book chapters, on the field of biometrics and is an area editor of *Springer's Encyclopedia of Biometrics*. He is a member of the IEEE.