Pose-Robust 3D Facial Landmark Estimation from a Single 2D Image

Brandon M. Smith Charles R. Dyer
University of Wisconsin-Madison

Sponsors: This work was supported in part by the United States Department of Transportation Federal Highway Administration DTFH61-14-C-00011.

Motivation

Despite much research interest in facial landmark estimation in recent years, **relatively little work has been done to handle the full range of head poses encountered in the real world** (e.g., beyond 45° rotation). Large head pose variation is challenging for several reasons:

1. The 2D shapes of profile faces and frontal faces are significantly different;
2. Many landmarks become self-occluded on profile faces; and
3. Even when visible, landmark appearance changes significantly with head pose.

As a result, the large majority of face alignment algorithms are limited to near fronto-parallel faces, and break down on profile faces.

Conventional Cascaded Shape Regression (CSR)

Conventional CSRs learn a sequence of $t=1,\dots,T$ descent maps $\{\mathbf{Q}^t\}_{t=1}^T$ that minimize the following:

$$\hat{\mathbf{Q}}^t = \underset{\mathbf{Q}^t}{\operatorname{argmin}} \sum_i \|\Delta\mathbf{s}_i^t - \mathbf{Q}^t\mathbf{d}^t(I_i, \hat{\mathbf{s}}_i^{t-1})\|_2^2 \qquad \Delta\mathbf{s}_i^t = \mathbf{s}_i - \hat{\mathbf{s}}_i^{t-1}$$

where $\mathbf{d}(I,\mathbf{s})$ is a feature descriptor that captures the local appearance in image I relative to shape $\mathbf{s}$, $\mathbf{s}_i$ is the ground truth set of landmarks for training example i, and $\hat{\mathbf{s}}_i$ is an estimate of $\mathbf{s}_i$. At test time, starting with a mean face shape initialization $\hat{\mathbf{s}}^0=\mu$, $\hat{\mathbf{s}}$ is updated over $t=1,\dots,T$ iterations:

$$\begin{aligned}\Delta\hat{\mathbf{s}}^t &= \hat{\mathbf{Q}}^t\mathbf{d}^t(I,\hat{\mathbf{s}}^{t-1})\\ \hat{\mathbf{s}}^t &= \hat{\mathbf{s}}^{t-1}+\Delta\hat{\mathbf{s}}^t.\end{aligned}$$

3DPDM Viewpoint Domain CSR

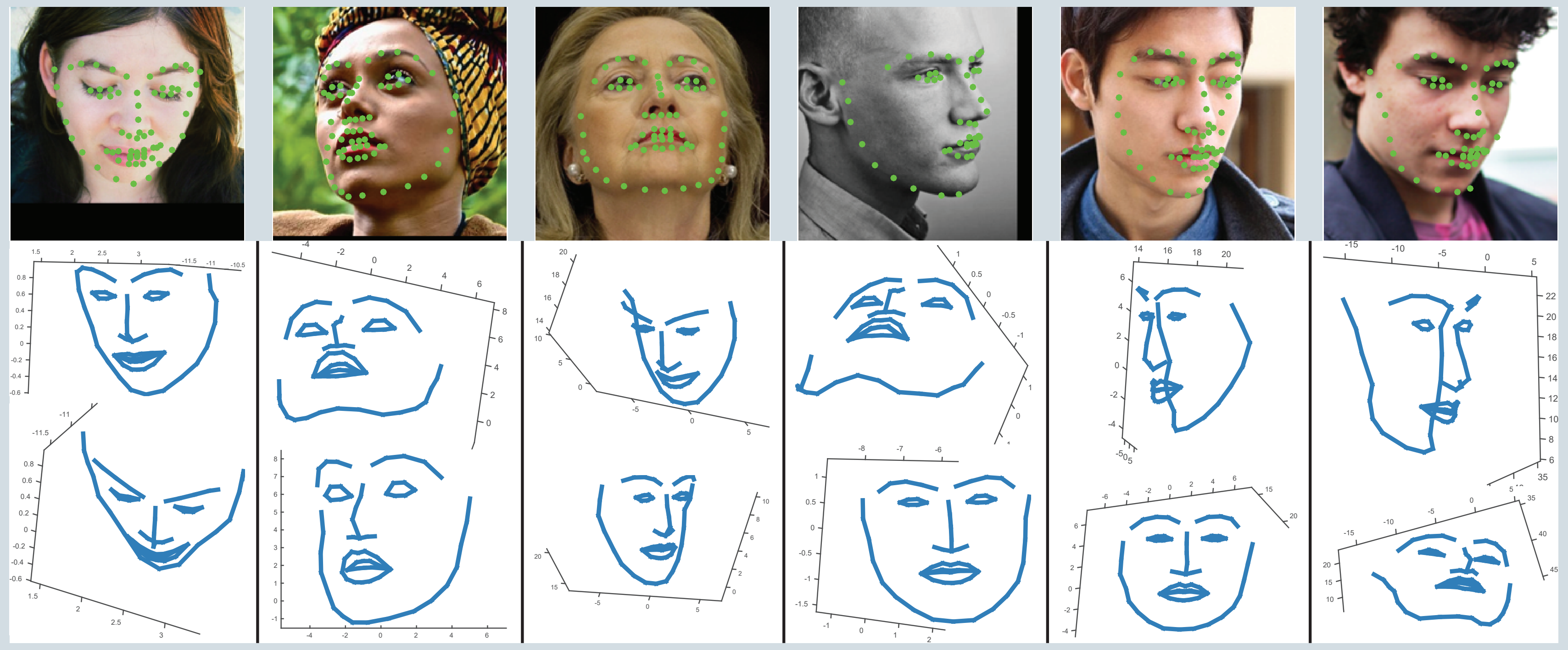
A single sequence of $\{\mathbf{Q}^t\}_{t=1}^T$ can result in undesirable performance because the descent maps average conflicting gradient directions, which is increasingly problematic with more head pose variation. Therefore, we partition the conventional CSR objective into $v=1,\dots,V$ *viewpoint domains* and learn a separate CSR for each one.

$$\hat{\mathbf{Q}}^{t,v} = \underset{\mathbf{Q}^{t,v}}{\operatorname{argmin}} \sum_{i\in\Phi^v} \|\Delta\hat{\mathbf{p}}_i^t - \mathbf{Q}^{t,v}\mathbf{d}^{t,v}(I_i,\mathbf{s}^{t-1})\|_2^2 + \alpha\|\mathbf{Q}^{t,v}\|_2^2,$$

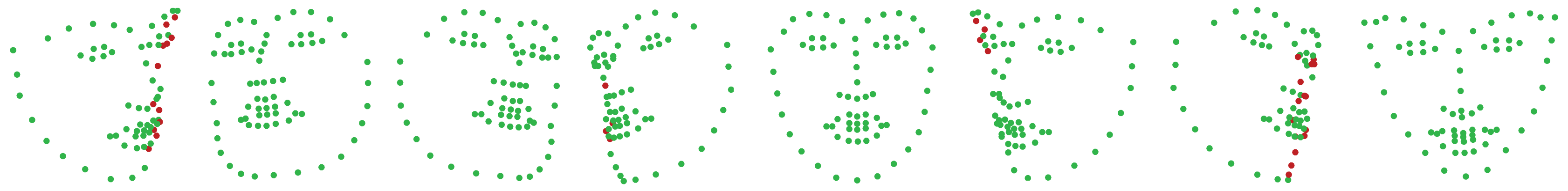
where $\Delta\hat{\mathbf{p}}_i^t$ is the ideal parameter update for face i, and Φ^v is the subset of training instances that belong to viewpoint domain v. At test time, the shape is updated for $t=1,\dots,T$:

$$\begin{aligned}\Delta\hat{\mathbf{p}}^t &= \mathbf{Q}^{t,v}\mathbf{d}^{t,v}(I,\mathbf{s}^{t-1})\\ \hat{\mathbf{p}}^t &= \hat{\mathbf{p}}^{t-1}+\Delta\hat{\mathbf{p}}^t\\ \hat{\mathbf{s}}^t &= \mu^v+\mathbf{B}^v\hat{\mathbf{p}}^t.\end{aligned}$$

3D Face Shape Estimation In The Wild

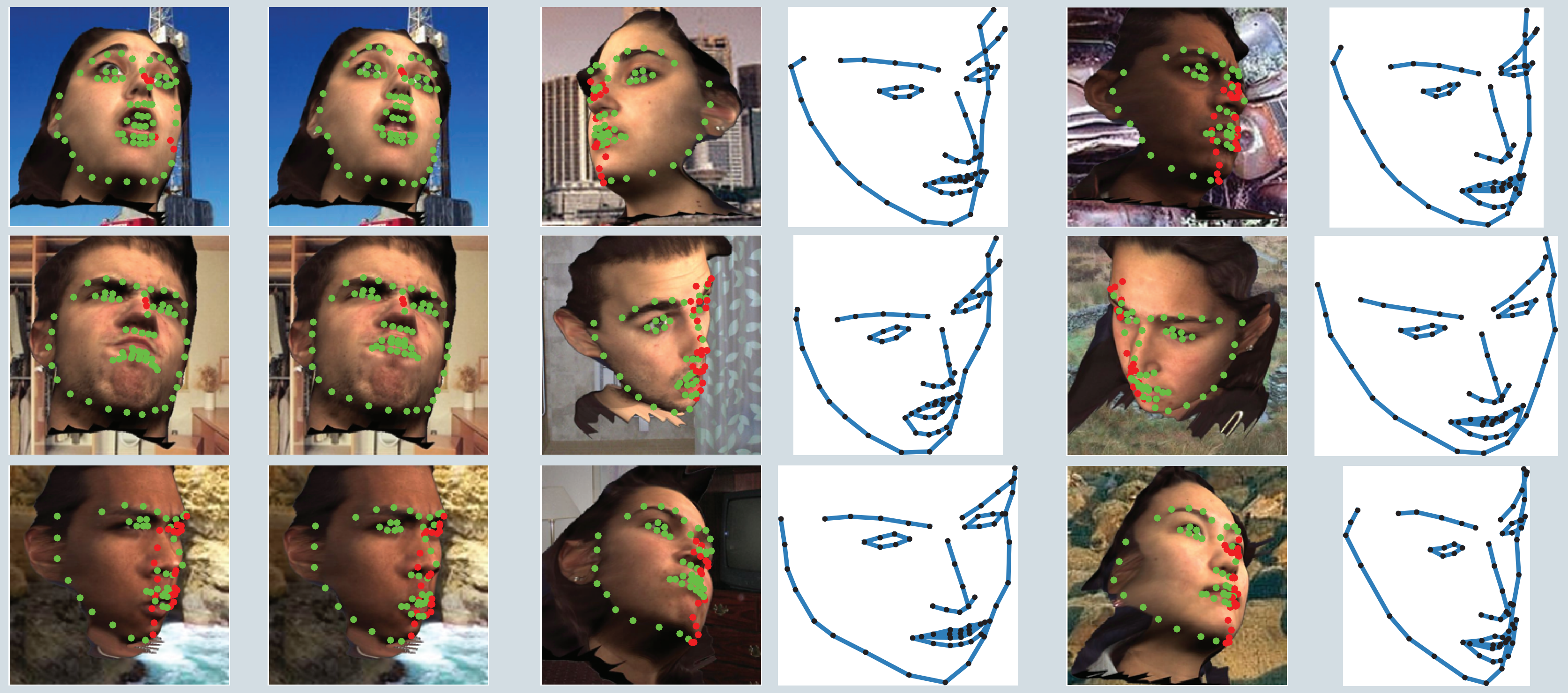


Viewpoint Domains



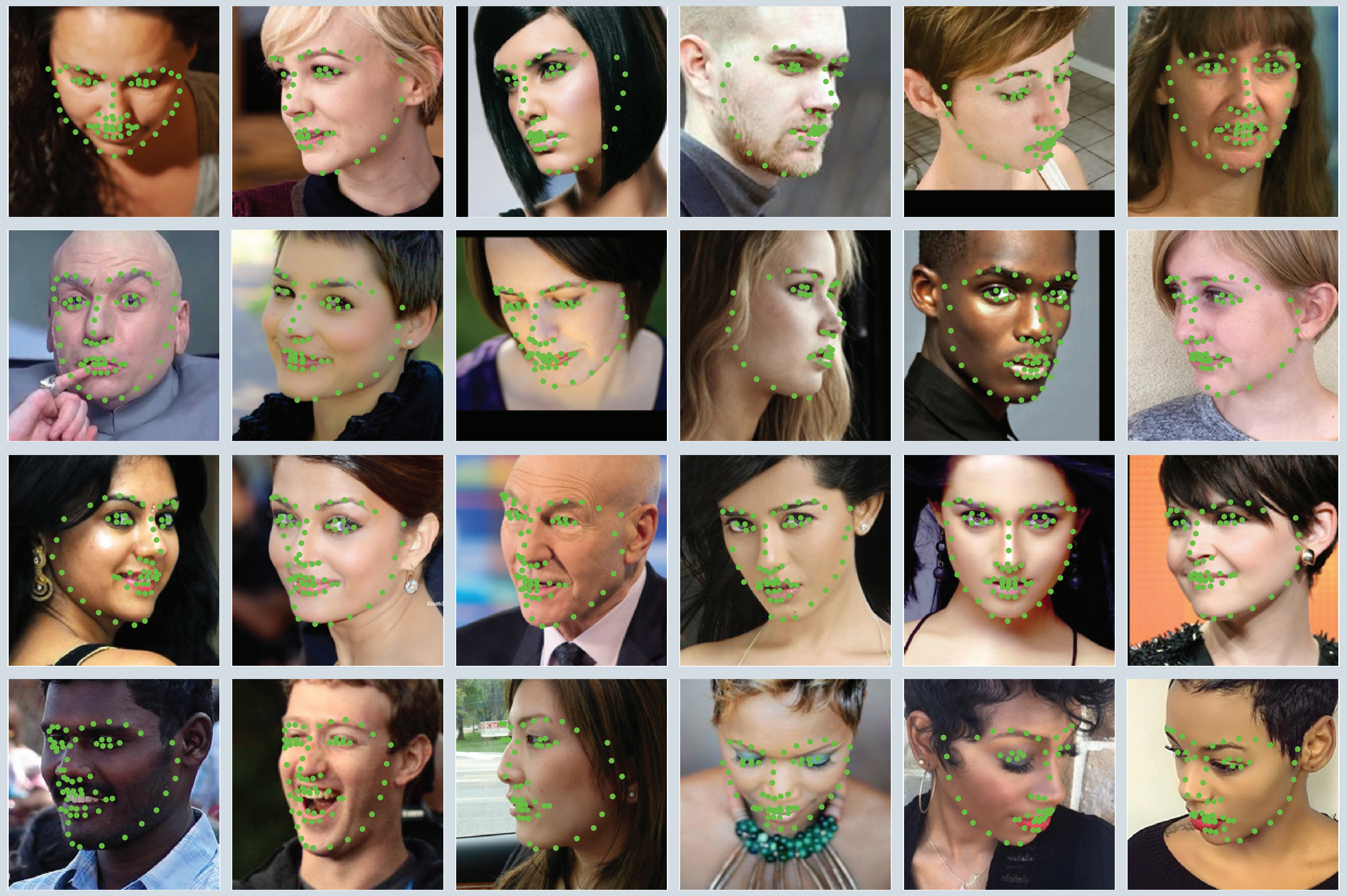
The modal viewpoints *automatically* found for V = 8 viewpoint domains. The modal occlusion state of the landmarks is stored for each viewpoint domain (**green** is visible, **red** is occluded). Features are extracted around only visible landmarks.

BU-4DFE Dataset



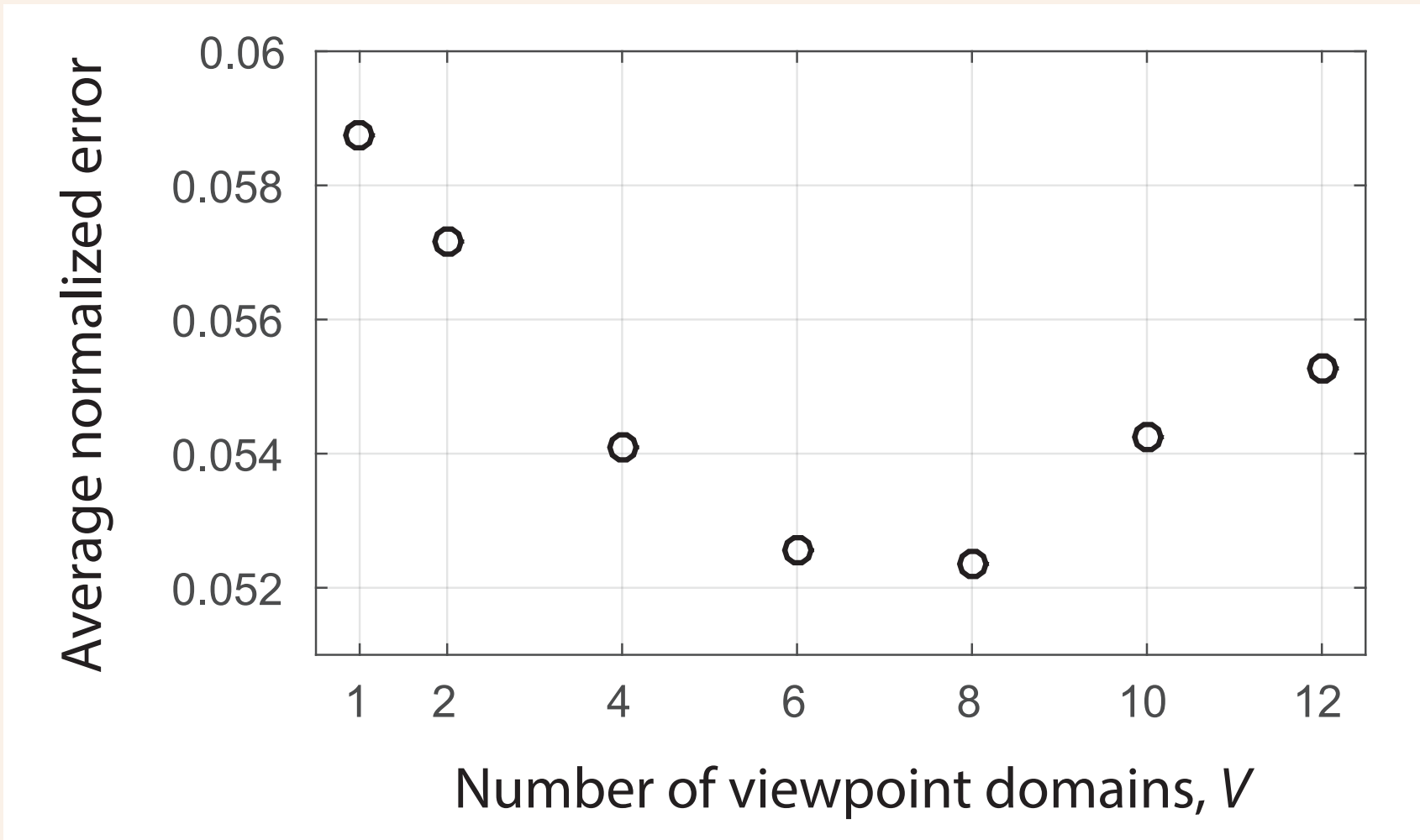
The estimated visibility of each landmark is shown in **green** (visible) and **red** (occluded). Results in (a) are from our baseline algorithm, and results in (b) are from our algorithm with V = 6. Notice the improvements from (a) to (b). In (c) and (e), a selection of results with estimated 3D shapes in (d) and (f). Note that the 3D shapes were rotated to a common orientation for fair comparison.

In-The-Wild Faces

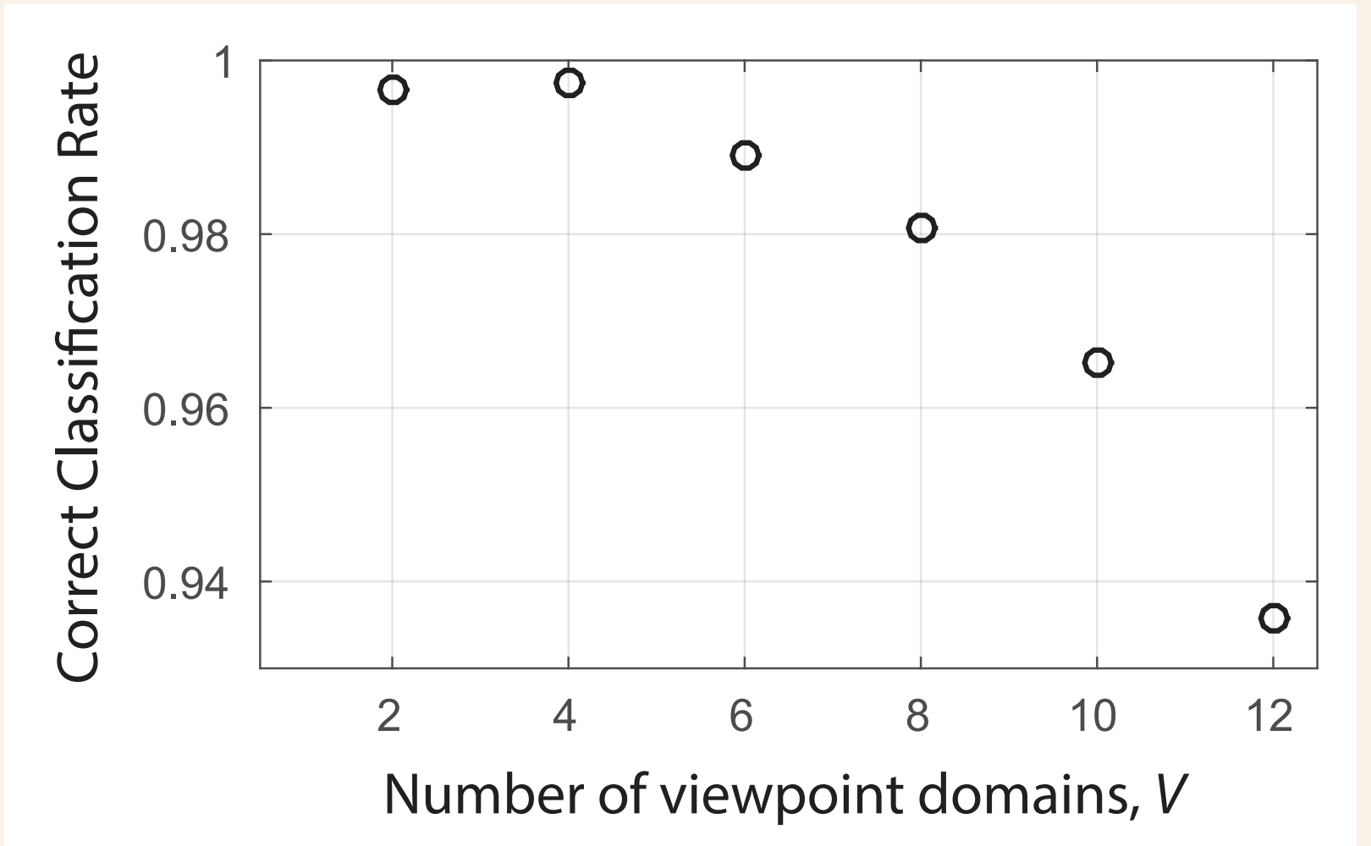


Although our model was trained on a laboratory dataset, BU-4DFE, it generalizes well to in-the-wild faces across a wide variety of poses, expressions, and other variations.

Quantitative Results



Face alignment accuracy vs. number of viewpoint domains. Observe that V =8 produces the most accurate results overall. However, V = 6 produces a good tradeoff between accuracy and efficiency (model size and runtime).

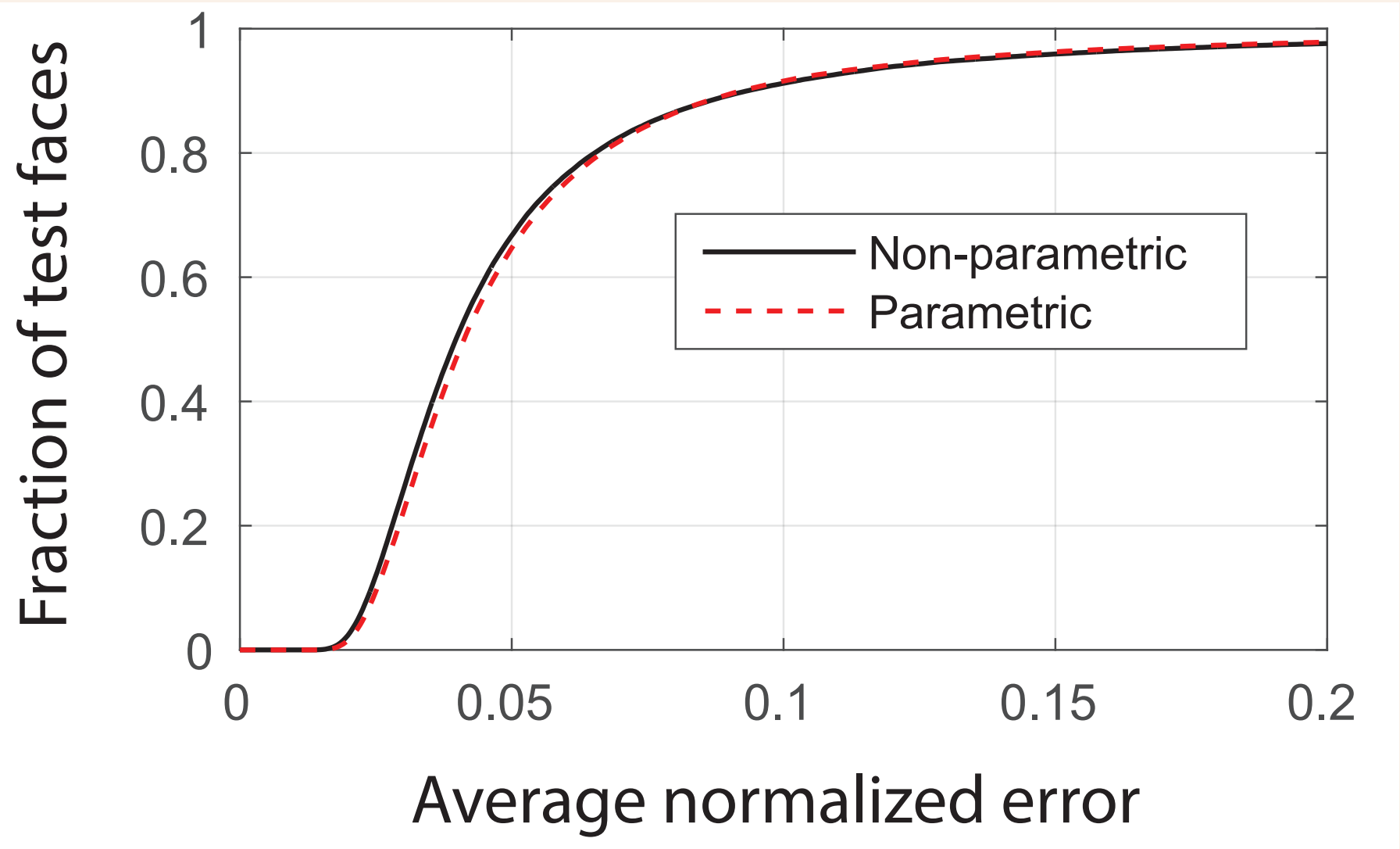


Viewpoint domain assignment accuracy vs. number of viewpoint domains. By "correct" we mean the test face is correctly assigned to either its true viewpoint domain or an overlapping domain.

Method	Average Normalized 3D Errors
Tulyakov and Sebe, Baseline indexing	0.0610
Tulyakov and Sebe, 3D Transform	0.0607
Tulyakov and Sebe, Basis Transform	0.0592
Our Baseline (V = 1), Nonparametric	0.0586
Our Baseline (V = 1), Parametric	0.0588
Ours, V = 6	0.0535
Ours, V = 8	**0.0524**

Comparison of average normalized 3D landmark localization errors (average point-topoint error between estimated and ground truth landmarks divided by inter-ocular distance). The top three rows are copied from S. Tulyakov and N. Sebe. *Regressing a 3D face shape from a single image*, ICCV 2015.

Parametric vs. Nonparametric Shape Models



Comparison between two baseline algorithms, one with landmark coordinates used directly as regression targets (nonparametric), and the other with 3DPDM shape parameters used as regression targets (parametric). Observe that performance is almost identical.

A recent trend has been to model face shape nonparametrically, and directly update landmark coordinates. However, parametric point distribution models (PDMs) have several desirable qualities:

1. There are fewer parameters to optimize, which results in a smaller set of regression coefficients.
2. They generalize well to unfamiliar faces.
3. All landmarks are optimized simultaneously.

In fact, we show empirically that there are no significant differences in accuracy between parametric and nonparametric shape models when used in otherwise identical systems.