Collaborative Facial Landmark Localization for Transferring Annotations Across Datasets

Brandon M. Smith and Li Zhang

University of Wisconsin – Madison
http://www.cs.wisc.edu/~lizhang/projects/collab-face-landmarks/

Abstract. In this paper we make the first effort, to the best of our knowledge, to combine multiple face landmark datasets with different landmark definitions into a super dataset, with a union of all landmark types computed in each image as output. Our approach is flexible, and our system can optionally use known landmarks in the target dataset to constrain the localization. Our novel pipeline is built upon variants of state-of-the-art facial landmark localization methods. Specifically, we propose to label images in the target dataset jointly rather than independently and exploit exemplars from both the source datasets and the target dataset. This approach integrates nonparametric appearance and shape modeling and graph matching together to achieve our goal.

1 Introduction

Facial landmark localization is a popular and extensively studied area in computer vision. Many approaches have been proposed over the years, from classic methods like Active Shape Models (ASMs) [4], Active Appearance Models (AAMs) [3], and Constrained Local Models (CLMs) [6] to more recent exemplar-based [2], voting-based [26], and supervised descend-based methods [25]. Many datasets have also been proposed to evaluate these methods, from early datasets collected in the lab like CMU PIE [21], Multi-PIE [7], AR [14], and XM2VTSDB [15], to more recent in-the-wild datasets like LFPW [2], AFLW [10], AFW [30], Helen [11], and IBUG [17].

On one hand, new datasets pose new challenges to the research community and foster new ideas. On the other hand, as researchers, we must choose specific datasets for evaluation to publish our work, which becomes increasingly difficult because different datasets have different landmark definitions (for example, AFLW uses a 21-landmark markup, while Helen uses 194 contour points). As a result, models trained on one dataset often cannot be evaluated on other datasets. Furthermore, inconsistencies between datasets make it difficult to train robust landmark localization models that combine many different datasets.

Ideally, it would be desirable to have a common and unified definition of landmarks and collect datasets following the same definition. However, this goal is challenging in practice because the speed of collecting labels will always lag the speed of collecting face data. Furthermore, it is difficult to predict which landmark definitions (*e.g.*, ears) new applications will find useful.

D. Fleet et al. (Eds.): ECCV 2014, Part VI, LNCS 8694, pp. 78–93, 2014.

In this paper we make the first effort, to the best of our knowledge, to combine multiple face landmark datasets with different landmark definitions into a super dataset, with a union of all landmark types computed in each image as output. Specifically, we present a novel pipeline built upon variants of state-of-the-art facial landmark localization methods that transfers landmarks from multiple datasets to a target dataset. Our system labels images in the target dataset jointly rather than independently and exploits exemplars from both the source datasets and the target dataset. This approach allows us to integrate nonparametric appearance and shape modeling and graph matching together to transfer annotations across datasets. Toward this goal, our paper makes the following contributions:

1. A pipeline that transfers landmark annotations from multiple source datasets to never-before-labeled datasets.
2. An algorithm that takes multiple source datasets as input and labels a partially labeled target dataset using a union of landmarks defined in the source datasets. Our system can optionally use known landmarks in the target dataset as constraints.
3. 64 supplementary landmarks for faces in the AFLW database [10], for a total of 85 landmarks. AFLW is significant in that, to the best of our knowledge, it is currently the largest publicly available in-the-wild face dataset with 25,000 annotated faces.

2 Related Work

We are aware of no other works that explicitly address the problem of *automatically* combining multiple datasets that have different landmark annotations. However, components of our system are inspired by and/or are built upon existing methods in the literature, which we summarize below.

Like Smith *et al*. [22] and Shen *et al*. [20], we use a Hough voting approach to generate landmark response maps in Stage 2 of our system. Yang and Patras [26] also rely on a Hough voting scheme for facial feature detection; they use several 'sieves' to filter out votes that are not relevant. In our approach, we adjust the weight of each vote by considering how well it agrees with other votes from matched features in other images.

Our landmark detection algorithm optionally uses known landmarks in the target image as constraints. Cootes and Taylor [5] proposed a constrained AAM that utilizes some known landmarks in the target image; AAMs are parametric models, while our approach is nonparametric and exemplar-based. Sagonas *et al*. [18] proposed a semi-automatic method for creating facial landmark annotations using person-specific models. Their process is iterative: users label results as 'good' or 'bad', and good results are used in later iterations as training data. Sagonas *et al*. used this approach to re-annotate several facial landmark datasets according to a consistent set of landmark definitions for the 300 Faces in-the-Wild Challenge (300-W) [17]. However, because their procedure is semi-automatic, it

does not scale well to very large datasets like AFLW [10]. Further, their procedure requires a consistent training dataset and ignores existing landmarks in the target datasets, *i.e.*, it completely overwrites them. In contrast, our method is fully automatic (our system has the ability to take user input, but we do not consider it in our experiments) and our pipeline transfers all existing landmarks across different datasets so that previous annotation efforts are utilized rather than wasted.

Exemplar-based approaches have been popular since Belhumeur *et al.*'s pioneering work [2]. Zhao *et al.* [28] use grayscale pixel values and HOG features to select k-nearest neighbor training faces, from which they construct a target-specific AAM at runtime. Smith *et al.* [22] and Shen *et al.* [20] perform Hough voting using k-NN exemplar faces; we use the same basic approach in our system. Finally, Zhou *et al.* [29] combine an exemplar-based approach with graph matching for robust facial landmark localization. We extend Zhou *et al.*'s approach to integrate different landmarks from multiple source datasets.

3 Our Approach

In this section we first give a brief overview of our system followed by a more detailed explanation of each stage in subsequent sections.

3.1 Overview

The input to our system is one or more *source* face datasets, and one *target* face dataset. We assume that each source dataset consists of a set of face images, in which each image is labeled with a set of facial landmarks, *e.g.*, eye centers, mouth corners, nose tip. Importantly, we do not require the landmark definitions to be consistent between source datasets. Optionally, each target image can have known landmarks, which our system uses as additional constraints. The output of our system is a combined set of landmark estimates (*i.e.*, the union set of landmark types from all source datasets) for each target face.

Stage 0: Preprocessing. We first rotate and scale all faces such that the eyes are level and the size is approximately the same across all face instances.[1] We then extract dense SIFT [13] features across each face at multiple scales. Following the approach in [20], we quantize each SIFT descriptor using fast approximate k-means [16], which efficiently maps each descriptor to a visual word.

Stage 1: Selection of Top Source Faces. For each target face, retrieve a separate subset of top k similar faces from each source dataset. The goal is to retrieve source faces that are similar to the target face in appearance, shape, expression, and pose so that features in the source images will produce accurate landmark votes in the target image.

[1] Eyes are easier to locate than other parts of the face, and so we assume they can be located accurately beforehand to rectify the face, *i.e.*, using an eye detector as in [28]. However, our method is not that sensitive to eye localization accuracy.

Stage 2: Weighted Landmark Voting. For each target face, independently compute a separate voting map for each landmark type from each source dataset using a generalized Hough transform [12]. Each feature from the top k source faces casts a vote for a possible landmark locations in the target image.

Stage 3: Shape Regularization. For each target face, compute a separate set of landmark estimates from each source dataset. Due to local ambiguities, occlusions, *etc*. each voting map may contain multiple peaks. We employ a robust nonparametric shape regularization technique [2] that avoids false peaks and estimates a globally optimized set of landmarks from each source dataset.

Stage 4: Final Landmark Estimation and Integration. For each target face, retrieve the top m most similar faces from the *target* dataset. The goal is to exploit the correlation between landmark estimates from Stage 3 among similar target faces to consistently label all target images. We combine estimates for landmarks common to multiple source datasets, and we optionally use known landmarks in each target image to constrain the optimization. We extend the graph matching technique in [29] for landmark integration from multiple source datasets. The final output for each target face is a full set of landmark estimates; by 'full' we mean the union of landmark types from all source datasets.

Input: One or more source face datasets, and one target dataset

Output: A combined set of landmark estimates for each target face

Stage 0: Preprocessing
for all target faces **do**
  Stage 1: Selection of Top Source Faces
  Stage 2: Weighted Landmark Voting
  Stage 3: Shape Regularization
end
for all target faces **do**
  Stage 4: Final Landmark Estimation and Integration
end

Fig. 1. Overview of our pipeline. Stage 4 is in a separate loop because it uses all the target face results from Stage 3 to help constrain and consistently estimate the final landmark results.

3.2 Stage 1: Selection of Top Source Faces

To transfer landmarks from each source dataset to the target image, the shape and appearance of the source faces and the target face should be similar. For example, a frontal face has much different appearance and shape than a profile face; there are few geometric feature-landmark correlations between the two. We therefore select a top subset of source faces for further processing.

Many strategies exist for retrieving similar face images from a database. In our system, we use a generalized Hough transform framework to score each source face. Specifically, we use the features on the target face to vote for the center of the face in each source image. The final score for each source face is the height of the maximum peak in the voting map associated with each source image. The intuition is that source faces with many shared features in similar geometric layouts with the target image will produce many consistent votes for the center of the face. We sort the scores and select the top $k = 200$. Shen *et al.* [20] adopt a similar strategy for retrieving exemplar faces in the validation step of their face detection algorithm.

3.3 Stage 2: Weighted Landmark Voting

For efficiency, rather than exhaustively sliding each source landmark region over the target image, we use quantized features and employ an inverted index file to efficiently retrieve matched features (*i.e.*, features in the same quantization bin) from the top k source images. When a feature in the target image is matched to a feature in a source image, the feature-to-landmark offset in the source image is transferred to the target image. The offset vector extends from the feature in the target image toward a potential landmark location and produces a vote. After many such votes, a voting map is formed, where the votes tend to cluster at landmark locations.

In practice, due to errors in the feature quantization step, image noise, occlusions, locally ambiguous image regions, *etc.*, many of the votes are incorrect, which can significantly impact overall voting accuracy. Yang and Patras [26] eliminate bad votes via a cascade of 'sieves.' Shen *et al.* [20] attempt to down-weight potentially bad votes using a heuristic from object retrieval: $\frac{\mathsf{idf}^2(k)}{\mathsf{tf}_Q(k)\mathsf{tf}_D(k)}$, where $\mathsf{idf}^2(k)$ is the squared inverse document frequency of visual word k, and $\mathsf{tf}_Q(k)$ and $\mathsf{tf}_D(k)$ are the term frequencies of k in the query image and the database image, respectively.

We instead compute a weight for each vote online as follows. For a given feature in the target image, retrieve all features in the top k source images that share the same quantization bin. For each of these features, compute their offset from landmark l. After rejecting outlier votes (*i.e.*, by measuring the distribution and rejecting vote offsets outside the inter-quartile range), we compute the variance σ_v^2 of the remaining offsets. We then cast a "fuzzy" vote from each offset using a 2D Gaussian $\mathcal{N}(v; \sigma^2)$ centered on the vote location v. Intuitively, this rewards matched features that produce consistent voting offsets and suppresses features that disagree. Our weighting scheme is similar to [22] and is less heuristic than [20]. Because our weights are computed online, we can easily add additional faces to the source dataset. In contrast, [22] and [26] require retraining when the training dataset changes.

We note that the Hough voting strategy is sensitive to scale and rotation differences between source and target faces. Shen *et al.* [20] address this problem by performing Hough voting over multiple scales. We instead normalize the scale

and orientation of each face in Stage 0, which eliminates the need to search for scale and rotation parameters.

3.4 Stage 3: Shape Regularization

There are many strategies for enforcing shape constraints, *e.g.*, [2,19,25,29,30]. However, in our case, we use an exemplar-based approach to shape regularization [2], which fits nicely with our exemplar-based Hough voting strategy for generating landmark response maps.

Belhumeur *et al.* [2] use SVM-based landmark detectors to establish an initial set of landmark location hypotheses, which forms the input to their final shape optimization algorithm. Each SVM attempts to capture all the local appearance variation around each landmark within a single model. This works well on faces with limited head pose variation. In contrast, our Hough voting strategy creates a nonparametric appearance model for each landmark, specific to each target face, which works well on faces with extreme head pose variation. Also, by aggregating the votes from many features, our method takes advantage of the larger appearance context around each landmark, which provides much more robustness to local noise, occlusions, *etc.* We therefore use our landmark voting maps in place of the local detector response maps used in [2].

Additionally, rather than using the entire set of exemplar face shapes as input, which is the approach taken in [2], we use only the top k source faces retrieved in Stage 1 of our pipeline. The top k source faces tend to be better tailored to the target face than the general set of faces, which further aids the optimization.

3.5 Stage 4: Final Landmark Estimation and Integration

The goal of this stage is to combine the individual landmark estimates from each source dataset into a single result for each target image. We incorporate several constraints into the optimization:

1. We model each landmark location as a linear combination of the other landmarks, which provides an affine-invariant shape constraint [29].
2. If available, known landmarks in the target image are fixed and help steer nearby landmark estimates to their correct locations.
3. Only one estimate is allowed for each landmark type irrespective of the number of source datasets contributing estimates for each type.

We address the shape constraint first. Let $\mathbf{P} = [\mathbf{p}_1, \ldots, \mathbf{p}_{N_P}]$ be a face shape composed of N_P landmarks. Following [29], we assume that the c-th landmark location $\mathbf{p}_c$ can be reconstructed by a linear combination of neighboring landmarks: $\mathbf{p}_c = \mathbf{P}\mathbf{w}_c$, where $\mathbf{w}_c \in \mathbb{R}^{N_P}$ is a vector of weights for the other $N_P - 1$ landmarks (the c-th entry of $\mathbf{w}_c$ is fixed to zero).

Suppose we have N_S source datasets and therefore N_S {target image t, source dataset s} pairs. Each pair has a union set of landmark types, $L_{ts} = \{L_t \cup L_s\}$,

composed of landmark types L_t defined in the target image[2] or landmark types L_s defined in the source dataset. Each L_{ts} contains all landmark types either known *a priori* in target image t, or estimated from source s or both. We aim to compute an optimal $\mathbf{w}_{tsc}$ for $c \in L_{ts}$ for each {target image t, source dataset s} pair (we subsequently omit t and s subscripts in $\mathbf{w}_{tsc}$ for simplicity).

To accomplish this task we need a set of example shapes that include all landmark types in L_{ts}. Given a target image t, we retrieve the m most similar face shapes among the target face images; the face shapes for the target images come from the regularized landmark localization results from Stage 3. As a distance metric, we simply use the mean Euclidean error between shapes after similarity transformation alignment. Using [29], we compute the $\mathbf{w}_c$ for image t that minimizes the sum of reconstruction errors among the top m most similar shape results from Stage 3:

$$\min_{\mathbf{w}_c} = \sum_{j}^{m} ||\mathbf{P}^j \mathbf{w}_c - \mathbf{p}_c^j||_2^2 + \eta ||\mathbf{w}_c||_2^2 \tag{1}$$

$$\text{s.t. } \mathbf{w}_c{}^{\mathsf{T}} \mathbf{1}_{N_P} = 1, \qquad w_{cc} = 0, \qquad w_{cr} = 0 \ \ \forall r \notin L_{ts},$$

where $\mathbf{P}^j$ is the j-th most similar face shape relative to t among other results from Stage 3; the constraint $w_{cr} = 0 \ \forall r \notin L_{ts}$ means that we force weights to zero if the r-th landmark is undefined in L_{ts}; and $\eta ||\mathbf{w}_c||_2^2$ is a regularization term that penalizes the sparsity of the weight vector, *i.e.*, it promotes more uniformity in the weights, which means that non-local landmarks can also carry importance in determining the c-th landmark location. Eq. (1) is a small convex quadratic problem, which we solve independently for each $\mathbf{w}_c$. The formulation of Eq. (1) is the same as [29] except for our added third constraint.

We compose the joint weight matrix as $\mathbf{W}_s = [\mathbf{w}_1, \ldots, \mathbf{w}_{N_P}]$, and we repeat the process for each source dataset s to create a set of N_S joint weight matrices $\mathbf{W}_1, \ldots, \mathbf{W}_{N_S}$ specific to target image t. Note that undefined columns in each $\mathbf{W}_s$ (corresponding to landmarks not defined in L_{ts}) are set to zero.

Following [29], let us now define a global coordinate matrix $\mathbf{Q} = [\mathbf{Q}_1, \ldots, \mathbf{Q}_{N_P}] \in \mathbb{R}^{2 \times N}$, where $\mathbf{Q}_c \in \mathbb{R}^{2 \times N_c}$ denotes candidate locations for the c-th landmark and $N = \sum_c N_c$. Let $\mathbf{G} \in \{0,1\}^{N_P \times N}$ be a binary association matrix, where $g_{ci} = 1$ if the i-th point belongs to the c-th landmark. Note that the candidate locations are the locations of the local peaks in the landmark response maps in Stage 3. When a landmark is common in multiple source datasets, we average the response maps from different source datasets before finding the local peaks. Let $\mathbf{A} \in \mathbb{R}^{N_P \times N}$ denote the assignment cost matrix, *i.e.*, $a_{ci} = -\log(R_c(\mathbf{q}_i))$, where $R_c(\mathbf{q}_i)$ is the height value in the c-th voting map at $\mathbf{q}_i$ after the voting map is normalized to sum to 1.

Given the candidates $\mathbf{Q}$, $\mathbf{G}$, $\mathbf{A}$ and the shape constraints $\mathbf{W}_1, \ldots, \mathbf{W}_{N_S}$, the problem consists of finding the optimal correspondence $\mathbf{X}$ that minimized the following error:

[2] L_t can be empty, in which case the target dataset has no known landmarks.

$$\min_{\mathbf{X}} \quad \lambda \mathrm{tr}(\mathbf{A}\mathbf{X}^{\mathsf{T}}) + \sum_{s}^{N_S} \|\mathbf{Q}\mathbf{X}^{\mathsf{T}}(\mathbf{I}_s - \mathbf{W}_s)\|_1 \tag{2}$$

$$\text{s.t. } \mathbf{X}\mathbf{1}_N = \mathbf{1}_{N_P}, \quad \mathbf{X} \in [0,1]^{N_P \times N}, \quad x_{ci} = 0, \quad [c,i] \in \{[c,i] | g_{ci} = 0\},$$

where $\mathbf{I}_s$ is an $N_P \times N_P$ identity matrix except that we set $\mathbf{I}_s(r,r) = 0\ \forall r \notin L_{ts}$ (*i.e.*, the r-th diagonal element is set to zero if landmark r is not defined in the target image or in dataset s). Eq. (2) is inspired from [29] except here we sum over multiple shape constraint terms instead of just one. Due to the integer constraint on $\mathbf{X}$, optimizing Eq. (2) is NP-hard. Like [29], we solve Eq. (2) by relaxing the integer constraint with a continuous one, and by reformulating the problem to incorporate two auxiliary variables that replace the non-smooth ℓ_1 norm with a smooth term and a linear constraint. Please see [29] for more details.

Incorporating known landmarks in the target image as constraints in Eq. (2) is straightforward. We simply provide a single candidate location for each of the known landmarks via the matrices $\mathbf{Q}$ and $\mathbf{G}$.

Because we use the same correspondence matrix $\mathbf{X}$ for all terms in Eq. (2), we obtain only one estimate for each landmark type, regardless of how many source datasets contribute to the estimate.

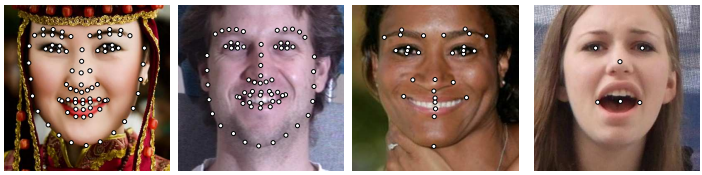
3.6 Implementation Details and Runtime

For quantizing SIFT features we use fast approximate k-means [16] with $k = 10^5$ clusters. For efficiency, we quantize the spatial variance σ_v^2 measurement of each vote cluster in Section 3.3 and convolve each voting map after all voting is complete using a set of precomputed Gaussian kernels. We also threshold σ_v to prevent erroneous spikes in the voting maps: we do not allow σ_v to fall below 3 pixels. In Stage 4, we set $\eta = 1000$, $\lambda = 100$, and we use about 200 candidates for each landmark.

Because our system operates on face datasets, we consider our pipeline to be entirely 'offline.' However, it is not prohibitively slow despite the number of steps involved. All tests were conducted on an Intel Xeon E5-2670 workstation. For each 480×480 image in our evaluation set, feature extraction and quantization takes less than a second. For each {target image, source dataset} pair, top exemplar selection (Stage 1) takes approximately 2.5 seconds, landmark voting (Stage 2) across 84 landmarks takes approximately 15 seconds, and shape regularization (Stage 3) takes approximately 10 seconds using our MATLAB implementation. The final stage is the most expensive (approximately 30 seconds per image) in part because MATLAB's linear program solver is relatively slow with many landmarks and candidate locations. We remark that most parts of our pipeline can be easily parallelized.

4 Results and Discussion

In this section we present two groups of experiments to evaluate the accuracy of our approach. First, we compare our accuracy with recent facial landmark



(a) Helen (b) Multi-PIE (c) LFPW (d) IBUG, AFW

Fig. 2. Experimental datasets. When each dataset is acting as source, we use the landmark annotations shown above. There are 85 landmark types across all datasets.

localization methods [1, 2, 23, 25, 27–30]. For fair comparison, we assume that no landmarks are known in the target images, and we measure accuracy over a common subset of landmarks computed across all methods. We show that our algorithm generally outperforms recent methods on especially challenging in-the-wild faces. Second, we measure the accuracy of our algorithm using different numbers of known landmarks in the target dataset to show that our method exploits additional known landmarks as constraints to further significantly improve accuracy. For all experiments we use multiple source datasets, each with a different set of landmark definitions.

4.1 Experimental Datasets

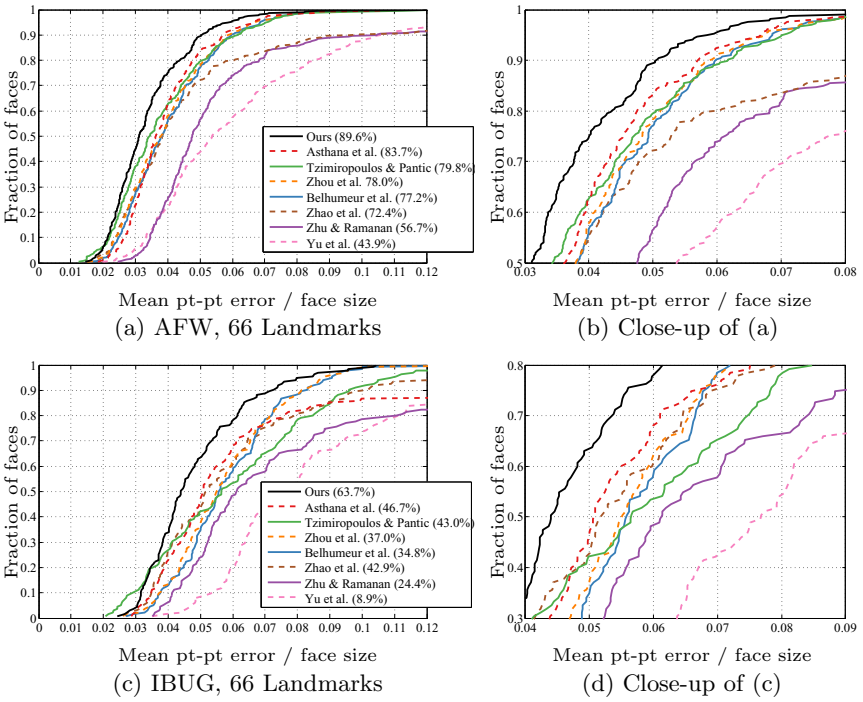
We used five face datasets for our quantitative evaluation: Multi-PIE [7], Helen [11], LFPW [2], AFW [30], and IBUG [17]. In the literature, there are two versions of landmark annotations for Helen, LFPW, and AFW: (1) the annotations provided when the datasets were originally released, which we refer to as 'original' hereafter; and (2) the recent annotations provided as part of the 300 Faces in-the-wild Challenge (300-W) [17], which we refer to as '300-W' hereafter. We use both versions of the landmarks; details are described in the context of individual experiments.

As in [23, 27, 30], we measure the size of the face as the average of the height and width of the rectangular hull around the ground truth landmarks. We favor this size measurement over inter-ocular distance (IOD) because it is more robust to yaw head rotation. Prior to evaluating all algorithms, we rescaled all test faces to a canonical size (200 pixels) and rotated them to make the eyes level.

4.2 Comparisons with Recent Works

In this section we quantitatively compare our algorithm with recent works [1, 2, 23, 25, 27–30]. The source datasets for training consist of Multi-PIE, Helen, and LFPW. For our algorithm, we used the ground truth landmark annotations shown in Figure 2 for Multi-PIE, Helen, and LFPW as training. The ground truth landmarks come from both the original annotations and the 300-W annotations (300-W annotations are favored in cases of redundant definitions). We note that there are 85 unique landmark types across all datasets.

Our target datasets for testing are AFW [30] and IBUG [17]. We use these two datasets for evaluation because they are particularly challenging, *e.g.*, they



(a) AFW, 66 Landmarks (b) Close-up of (a)

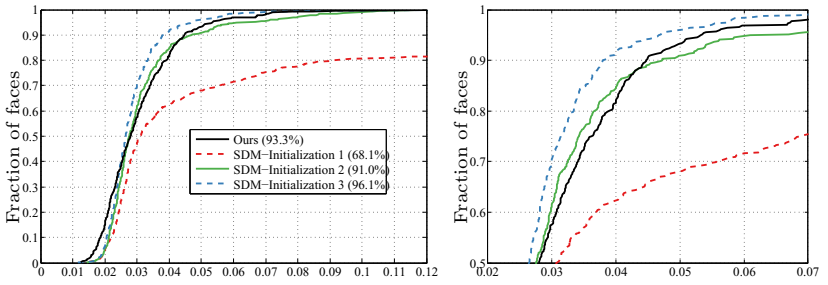
(c) IBUG, 66 Landmarks (d) Close-up of (c)

Fig. 3. Two sets of cumulative error distribution (CED) curves on AFW and IBUG face datasets. In all cases, the average localization error is normalized by the face size as defined in [30]. The numbers in parantheses are the fraction of faces at 0.05 error. Here we compare the accuracy of our approach with several recent works: Asthana *et al.* [1], Tzimiropoulos and Pantic [23], Zhou *et al.* [29], Belhumeur *et al.* [2], Zhao *et al.* [28], Zhu and Ramanan [30], and Yu *et al.* [27]. We see that our approach generally produces significantly more accurate results among those evaluated above. **Best viewed in color.**

include a large percentage of faces with extreme facial expression and/or head pose. In contrast, other popular datasets like BioID [9], Helen [11], LFW [8], and LFPW [2] contain faces with less challenging variations, which are consequently well addressed by current methods.

We made every effort to implement Belhumeur *et al.* [2] and Zhou *et al.* [29] algorithms faithfully; we trained them on the source dataset (Multi-PIE, Helen, and LFPW) using only 300-W annotations. For all other algorithms, we used the original authors' implementations. We used the off-the-shelf models provided with each implementation, with the exception of Zhao *et al.* [28]. Zhao *et al.* compute target face-specific models online from a given training database; for their algorithm, like Belhumeur *et al.* [2] and Zhou *et al.* [29], we provided Multi-PIE, Helen, and LFPW faces as training data using only 300-W annotations.

Initialization. For Belhumeur *et al.* [2] and Zhou *et al.* [29] we initialized the position of each landmark detector using a mean face shape aligned to the face. The diameter of each detector window was set to the larger of 33% of the face



(a) AFW & IBUG, 49 Landmarks (b) Close-up of (a)

Fig. 4. CED curves comparing our accuracy with SDM [25]. We observe that SDM is sensitive to initialization, especially for non-frontal faces, and so we evaluate it using three different initialization strategies. Initialization 1 and 3 follow the authors' strategy: fit a mean shape to the face detection rectangle. For 1 we follow [17] and use the rectangular hull around all 68 ground truth landmarks, and for 3 we use a much tighter rectangular hull around the interior 49 ground truth landmarks. Initialization 2 follows the strategy of [28]: fit a mean shape to the target face using ground truth eye centers. Our performance is similar to SDM–Initialization 2, and is slightly lower on average than SDM–Initialization 3. However, we remark that Initialization 3 provides an artificially favorable initialization to SDM because it is much tighter than Initialization 1. In contrast, our approach is not sensitive to the initialization: we initialize our algorithm using a 25% larger bounding rectangle than Initialization 1 (the least reliable but most realistic initialization here), and we do not rely on an initial shape. Unlike SDM, our full pipeline can use known landmarks in the target image as constraints to further significantly improve accuracy, as shown in Figure 6. **Best viewed in color.**

size or large enough to overlap the true landmark location. Zhao *et al.*'s implementation [28] is initialized via eye detectors; we provided their algorithm with ground truth eye centers. Tzimiropoulos and Pantic's [23] algorithm requires a face bounding box for initialization; for this we provided the ground truth bounding boxes as defined by [17].

Zhu and Ramanan's [30] algorithm is tied to their detection algorithm, and so we do not provide it with an initialization. We set their detection threshold to $-\infty$ to avoid missing faces. For each ground truth face annotation, we select the output face with the largest bounding box overlap (the area of intersection divided by the area of union), and we ignore all false positives. Zhu and Ramanan provide three models with their implementation. We used their *Independent-1050* model for all of our experiments since it generally performs best.

Asthana *et al.* [1] and Yu *et al.* [27] each rely on a version of [30] for initialization, and so we do not provide one separately. However, since Yu *et al.*'s implementation only returns landmark estimates for the highest scoring face in each image, we isolated the true face by cropped it out (the crop window was centered on the true face and set to approximately twice the face height/width).

Xiong and De la Torre's Supervised Descent Method (SDM) [25] is considered the current state of the art. The authors use the Viola-Jones face detector [24] for initialization. Unfortunately, Viola-Jones fails to detect 10% and 26% of

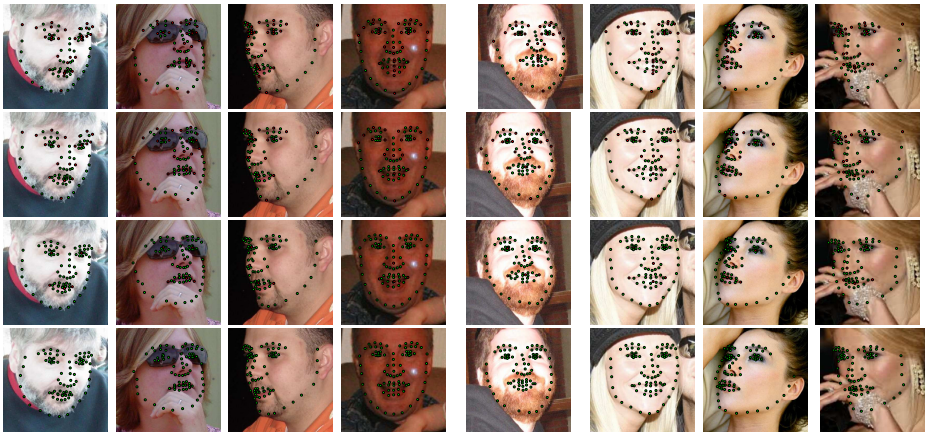


Fig. 5. Qualitative results on AFW faces (a)-(f) and IBUG faces (g)-(h) with varying numbers of known landmarks in the target images. Green points are estimated landmark locations, and red points are known landmark locations. From the top row to the bottom row, results were computed with 32, 21, 6, and 0 known landmarks. We see that errors are corrected with additional known landmarks, *e.g.*, the eyebrows in (a) and (f), and the lips in (d) and (g). Even with no known landmarks (bottom row), our algorithm performs well on challenging faces, including those with significant head pitch rotation (a, e, g, and h), head yaw rotation (b, c, g, and h), occlusion (b and h), and facial hair (a and e). Figure 3 shows quantitative results with no known landmarks in the target images. **Best viewed electronically in color.**

AFW and IBUG faces, respectively, despite our pre-rectification step. We instead initialized SDM using three different strategies, described in Figure 4.

Quantitative Results. Figure 3 shows two sets of cumulative error distribution (CED) curves, which compare the accuracy of our approach with others. Using Multi-PIE, Helen, and LFPW as source datasets, our algorithm produces 84 landmark estimates (a union of both 300-W and original annotations from the three source datasets).[3] We evaluated the accuracy of 66 landmarks in Figure 3 because [1] estimates 66 landmarks. Errors are computed relative to the 300-W ground truth landmarks as the mean point-to-point error normalized by the face size. We compare with SDM [25] separately in Figure 4 because the authors' implementation estimates 49 landmarks instead of 66. We see that our approach generally outperforms recent methods on AFW and IBUG faces.

[3] We supplemented the 300-W annotations on Helen with 10 landmarks from the original annotations (three on each eyebrow, four on the nose). When we use LFPW as a source dataset, we use only the 29 landmarks from the 300-W annotations that coincide with the original annotations. When we use AFW and IBUG as source datasets, we use only the six 300-W annotations that coincide with the original AFW annotation. Figure 2 shows the layout of landmarks for each source dataset.

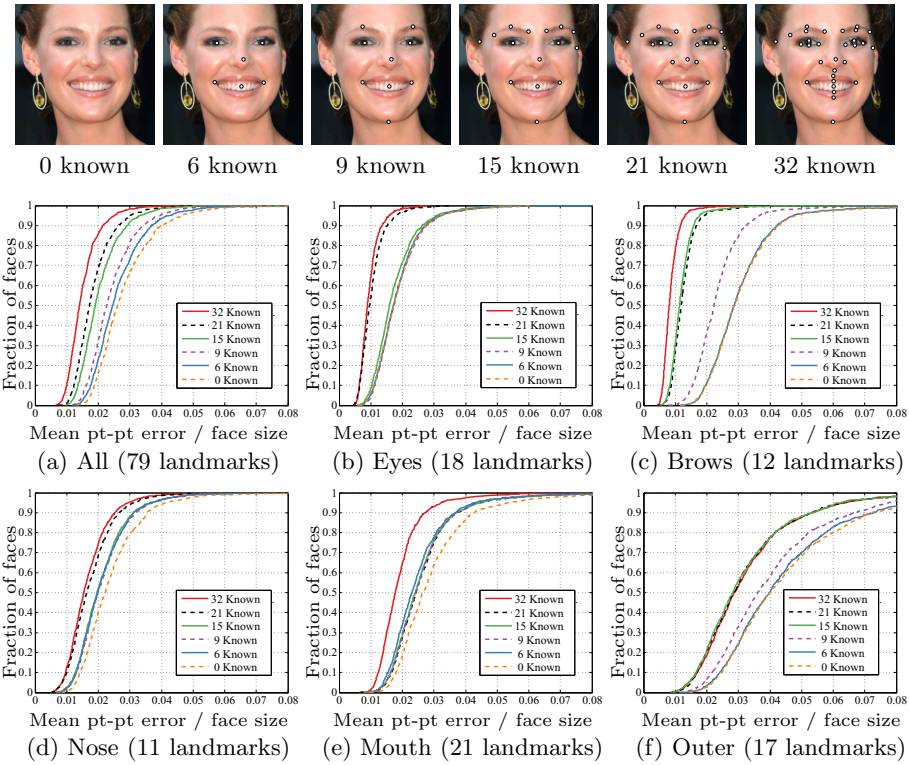


Fig. 6. Quantitative evaluation of our full pipeline with six different trials, each assuming a different number of known landmarks. The top row shows the arrangement of known landmarks for each trial. "6 known" corresponds to the original AFW layout; "21 known" corresponds to AFLW; and "32 known" closely resembles the original LFPW annotations. In (a) we see the overall mean accuracy is high with 0 known landmarks (96.5% at 0.05), and the accuracy continues to improve significantly as additional landmarks become known. For reference, Belhumeur *et al.* [2] showed that their algorithm surpasses the average accuracy of human labelers on most landmarks, and our algorithm further improves Belhumeur *et al.*'s localization accuracy (see Figure 3) even with 0 known landmarks. We note that inherent ambiguities exist on the face, especially on longer contours such as the lips and the outer face contour. For example, a landmark estimate on the outer contour may be qualitatively correct, but in disagreement with "ground truth" in terms of its location along the contour. This phenomenon partly explains the lower CED curves in (c), (e), and (f). In general, we see that our approach correctly estimates landmarks on a large majority of faces, especially with 21 or 32 known landmarks. This suggests that our approach is well-suited for automatically supplementing the landmarks in large, sparsely annotated datasets like AFLW. **Best viewed in color.**

4.3 Evaluation with Known Target Landmarks

We have quantitatively evaluated our full pipeline using 1035 images from LFPW as the target dataset, and using Multi-PIE, Helen, AFW, and IBUG as source datasets. The union of the different annotation definitions results in a total of 85 landmarks. We performed six different trials, with each trial assuming a different number of known landmarks in the target dataset: 0, 6, 9, 15, 21, and 32. Among these different numbers, we chose 6 because it corresponds to the original AFW annotations; we chose 21 because it corresponds to annotations provided in the AFLW dataset [10]; and we chose 32 because it closely resembles the original annotations in LFPW. The top of Figure 6 shows the arrangement of known landmarks for each trial. For each face, we measure the accuracy of 79 landmarks (out of 85 estimated) relative to ground truth annotations from 300-W and the original LFPW dataset; the ground truth of the remaining 6 landmarks are not available for LFPW.

The CED curves in Figure 6 show the accuracy of our algorithm on each of these trials across 79 landmarks. We see that the accuracy of our algorithm is high with 0 known landmarks (96.5% at 0.05 average overall), and the accuracy continues to improve with additional known landmarks.

A prime target dataset for our approach is AFLW [10], which contains 25,000 in-the-wild face images from Flickr, each manually annotated with up to 21 sparse landmarks. Our approach is well-suited to automatically supplementing AFLW with additional landmarks from source datasets like Multi-PIE [7] and Helen [11]. Our supplementary AFLW landmarks are available at our project website: http://www.cs.wisc.edu/~lizhang/projects/collab-face-landmarks/.

5 Conclusions

Our quantitative comparison shows that our approach generally significantly outperforms recent methods, and achieves accuracy comparable to the current state of the art on challenging in-the-wild faces, even with zero known landmarks in the target dataset. Our evaluation using different numbers of known landmarks in the target dataset show that our approach is well-suited to automatically supplementing an existing dataset with additional landmarks from other source datasets. However, our algorithm is not perfect and occasionally makes mistakes. For infrequent problem cases, our system naturally allows the user to provide a few additional landmarks as constraints. For these reasons, we want to build upon our system to include humans in the loop as part of a crowdsourcing platform for efficiently adding landmarks to large face datasets.

Acknowledgements. This work is supported by NSF IIS-0845916, NSF IIS-0916441, a Sloan Research Fellowship, and a Packard Fellowship for Science and Engineering.

References

1. Asthana, A., Zafeiriou, S., Cheng, S., Pantic, M.: Robust discriminative response map fitting with constrained local models. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
2. Belhumeur, P.N., Jacobs, D.W., Kriegman, D.J., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)
3. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. In: Burkhardt, H., Neumann, B. (eds.) ECCV 1998. LNCS, vol. 1407, pp. 484–498. Springer, Heidelberg (1998)
4. Cootes, T.F., Taylor, C.J.: Active shape models – 'smart snakes'. In: British Machine Vision Conference (1992)
5. Cootes, T.F., Taylor, C.J.: Constrained active appearance models. In: IEEE International Conference on Computer Vision (2001)
6. Cristinacce, D., Cootes, T.F.: Feature detection and tracking with constrained local models. In: British Machine Vision Conference, pp. 929–938 (2006)
7. Gross, R., Matthews, I., Cohn, J.F., Kanade, T., Baker, S.: Multi-PIE. Image and Vision Computing 28(5), 807–813 (2010)
8. Huang, G.B., Ramesh, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Tech. Rep. 07-49, University of Massachusetts, Amherst (October 2007)
9. Jesorsky, O., Kirchberg, K.J., Frischholz, R.W.: Robust face detection using the hausdorff distance. In: Bigun, J., Smeraldi, F. (eds.) AVBPA 2001. LNCS, vol. 2091, p. 90. Springer, Heidelberg (2001)
10. Koestinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: First IEEE International Workshop on Benchmarking Facial Image Analysis Technologies (2011)
11. Le, V., Brandt, J., Lin, Z., Bourdev, L., Huang, T.S.: Interactive facial feature localization. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 679–692. Springer, Heidelberg (2012)
12. Leibe, B., Leonardis, A., Schiele, B.: Combined object categorization and segmentation with an implicit shape model. In: European Conference on Computer Vision Workshop on Statistical Learning in Computer Vision (2004)
13. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
14. Martinez, A., Benavente, R.: The AR Face Database. Tech. Rep. 24, CVC (1998)
15. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTSDB: The extended m2vts database. In: 2nd International Conference on Audio and Video-Based Biometric Person Authentication (1999)
16. Muja, M., Lowe, D.G.: Fast approximate nearest neighbors with automatic algorithm configuration. In: VISAPP International Conference on Computer Vision Theory and Applications (2009)
17. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: The first facial landmark localization challenge. In: IEEE International Conference on Computer Vision (ICCV-W 2013), 300 Faces in-the-Wild Challenge (300-W) (2013)

18. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: A semi-automatic methodology for facial landmark annotation. In: IEEE Conference on Computer Vision and Pattern Recognition Workshop (2013)
19. Saragih, J.M., Lucey, S., Cohn, J.F.: Face alignment through subspace constrained mean-shifts. In: IEEE International Conference on Computer Vision (2009)
20. Shen, X., Lin, Z., Brandt, J., Wu, Y.: Detecting and aligning faces by image retrieval. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
21. Sim, T., Baker, S., Bsat, M.: The CMU pose, illumination, and expression database. IEEE Transactions on Pattern Analysis and Machine Intelligence 25(12), 1615–1618 (2003)
22. Smith, B.M., Brandt, J., Lin, Z., Zhang, L.: Nonparametric context modeling of local appearance for pose- and expression-robust facial landmark localization. In: IEEE Conference on Computer Vision and Pattern Recognition (2014)
23. Tzimiropoulos, G., Pantic, M.: Optimization problems for fast AAM fitting in-the-wild. In: IEEE International Conference on Computer Vision (2013)
24. Viola, P., Jones, M.J.: Robust real-time face detection. International Journal of Computer Vision 57(2), 137–154 (2004)
25. Xiong, X., De la Torre, F.: Supervised descent method and its applications to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
26. Yang, H., Patras, I.: Sieving regression forest votes for facial feature detection in the wild. In: IEEE International Conference on Computer Vision (2013)
27. Yu, X., Huang, J., Zhang, S., Yan, W., Metaxas, D.N.: Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In: IEEE International Conference on Computer Vision (2013)
28. Zhao, X., Shan, S., Chai, X., Chen, X.: Locality-constrained active appearance model. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 636–647. Springer, Heidelberg (2013)
29. Zhou, F., Brandt, J., Lin, Z.: Exemplar-based graph matching for robust facial landmark localization. In: IEEE International Conference on Computer Vision (2013)
30. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: IEEE Conference on Computer Vision and Pattern Recognition (2012)