Localizing Parts of Faces Using a Consensus of Exemplars

Peter N. Belhumeur, David W. Jacobs, *Member*, *IEEE*,
David J. Kriegman, *Senior Member*, *IEEE*, and Neeraj Kumar

Abstract—We present a novel approach to localizing parts in images of human faces. The approach combines the output of local detectors with a nonparametric set of global models for the part locations based on over 1,000 hand-labeled exemplar images. By assuming that the global models generate the part locations as hidden variables, we derive a Bayesian objective function. This function is optimized using a consensus of models for these hidden variables. The resulting localizer handles a much wider range of expression, pose, lighting, and occlusion than prior ones. We show excellent performance on real-world face datasets such as Labeled Faces in the Wild (LFW) and a new Labeled Face Parts in the Wild (LFPW) and show that our localizer achieves state-of-the-art performance on the less challenging BioID dataset.

Index Terms—Part localization, faces, biometrics, fiducial points

1 Introduction

Over the last decade, new applications in computer vision and computational photography have arisen due to earlier advances in methods for detecting human faces in images [27], [29]. These applications include face detection-based autofocus and white balancing in cameras, new methods for sorting and retrieving images in digital photo management software, anonymization of facial identity in digital photos, image editing software tailored for faces, and systems for automatic face recognition and verification.

Face detectors usually return the image location of a rectangular bounding box containing a face. This bounding box serves as the starting point for these applications. Yet, all of the above-mentioned applications, as well as numerous ones yet to be conceived, would benefit from the accurate detection and localization of face parts—for example, eyebrow corners, eye corners, tip of the nose, mouth corners, chin—within the specified bounding box. These parts are often referred to as facial feature points or fiducial points. However, unlike general interest or corner points, these part locations may not correspond to image locations with high gradients (e.g., tip of the nose), and their detection may require larger image support.

In addition, much of nonfrontal face processing has involved detecting multiple facial feature points (fiducial points) and using the correspondence between the detected features and stored exemplars. In [30], [32], 60 or more fiducial points are used to apply a 2D warp of the image to frontal pose. When using 3D morphable models for recognition, the fitting process was initialized using seven manually clicked fiducial points [24]. In [10], the face image is characterized by affine warped image patches about detected fiducial points. In [17], six detected fiducial points are used to compute qualitative attributes and similes, and these are used for face verification in unconstrained images demonstrating significant variability in pose, lighting, expression, facial hair, partial occlusion, glasses, and so on. Facial expression recognition often involves detecting facial features and tracking them through a video sequence [22].

There have been a number of recent works that have shown great accuracy in localizing parts in mostly frontal images and often in controlled settings. Our goal is to localize a large collection of prespecified parts in images taken under a variety of acquisition conditions. In this paper, we focus on applying our methodology to images of faces in which sources of variability include pose, lighting, expression, hairstyle, subject age, subject ethnicity, partial occlusion of the face, camera type, image compression, resolution, and focus.

To do this, we have acquired and labeled a data set called labeled face parts in the wild (LFPW) from Internet search sites using simple text queries. We have not intentionally filtered out faces due to poor image quality, keeping all faces that were detectable by our commercial, off-the-shelf (COTS) face detector.[1] Unlike datasets that are acquired systematically in the laboratory, there are few preconditions in our dataset that might aid detection—the eyes may be occluded by glasses, sunglasses, or hair; there may be heavy shadowing across features; the facial expression may be arbitrary; the face may have no makeup or be made up theatrically; the image may actually be an artistic rendering; the pose may be varied; there may be

- *P.N. Belhumeur is with the Department of Computer Science, Columbia University, New York, NY 10027.*
- *D.W. Jacobs is with the Department of Computer Science and UMIACS, University of Maryland, A.V. Williams Building, College Park, MD 20742.*
- *D.J. Kriegman is with the Department of Computer Science and Engineering, University of California, San Diego, EBU3B, Room 4120, 9500 Gilman Drive, La Jolla, CA 92093-0404.*
- *N. Kumar is with the University of Washington, Box 352350, Paul G. Allen Center 282, Seattle, WA 98195-2350.*
  E-mail: neeraj@cs.washington.edu.

Manuscript received 6 June 2012; revised 27 Nov. 2012; accepted 31 Dec. 2012; published online 14 Jan. 2013.
Recommended for acceptance by P. Felzenszwalb, D. Forsyth, P. Fua, and T.E. Boult.
For information on obtaining reprints of this article, please send e-mail to: tpami@computer.org, and reference IEEECS Log Number TPAMISI-2012-06-0431.
Digital Object Identifier no. 10.1109/TPAMI.2013.23.

0162-8828/13/$31.00 © 2013 IEEE Published by the IEEE Computer Society

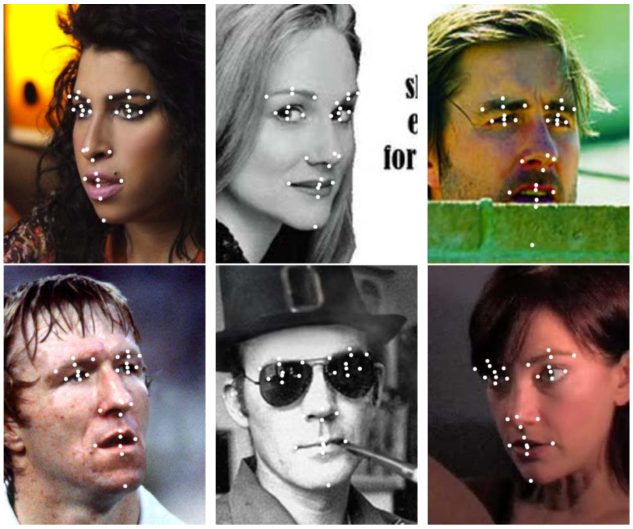

Fig. 1. Results of our face part localizer.

facial hair that occludes the fiducial points; and part of the face may be occluded by a hat, wall, cigarette, hand, or microphone. See Figs. 1 and 6. This dataset stands in contrast to datasets such as FERET or BioID that have been used for evaluating fiducial point detection in that the images are not restricted to frontal faces or collected in a controlled manner. Our dataset is more in line with recent face recognition datasets such as Labeled Faces in the Wild (LFW) [15] or PubFig [17] that contain images taken in uncontrolled settings.

We formulate part localization as a Bayesian inference that combines the output of local detectors with a prior model of face shape. Unlike previous work, our prior on the configuration of face parts is nonparametric, making use of our large collection of diverse, labeled exemplars. We then introduce hidden variables for the particular exemplar image and the similarity transformation applied to it that are assumed to generate fiducial locations in a new image. We marginalize out these hidden variables, but in doing so they provide us with valuable conditional independencies between different parts. To marginalize efficiently, we use a RANdom SAmple Consensus (RANSAC)-like process to sample likely values of the hidden variables. This ultimately leads to part localization as a combination of local detector output and the consensus of a variety of exemplars and poses that fit this data well.

Our process starts with an offline training phase, much of which is mirrored during online detection and captures the local appearance of the fiducials and models the global relationship of the fiducials. Faces are first detected with an off-the-shelf face detector. From the face box returned by the detector and our training data, we have bounds on the locations of each fiducial relative to the face box. Note that for datasets with only frontal images, the bounds are tighter than for datasets with large variations in in-plane and out-of-plane rotation. For each fiducial, a previously trained two-class classifier is scanned across the fiducial's corresponding bounding box, and at each location it returns a score that is proportional to the likelihood that the feature is at that location. Because this is multimodal and the highest mode of the local detector response may not correspond to the actual fiducial location, we also consider the global configuration of all the fiducials, which is a prior of the joint probability of fiducial locations. While it is well recognized in the literature that the geometric relationships of the fiducial location are important for robust fiducial detection, prior work generally modeled the constraints in an ad hoc manner by assuming independence or using a multivariate Gaussian distribution. In contrast, we model this nonparametrically using the labeled location of the fiducials in the training images. By formulating detection as a Bayesian estimation problem, the local appearance of each fiducial is naturally balanced with the nonparametric global prior on the configuration; the estimate is found with a Monte Carlo optimization procedure, based on RANSAC.

The method is evaluated on three datasets that are independent of the training set: The BioID dataset has been used to evaluate a number of existing methods, and it contains frontal, upright images of 35 people with a range of facial expressions taken with a single camera. In contrast, the Labeled Faces in the Wild (LFW) [15] dataset consists of 13,233 images of 5,749 public figures taken in unconstrained settings and downloaded from online news sources. We also introduce the new Labeled Face Parts in the Wild (LFPW) dataset in this paper, which is also unconstrained but contains a greater mix of image qualities. Experimental results demonstrate that our method is more accurate than existing methods on BioID and is just as accurate on the harder LFW and LFPW datasets. Furthermore, accuracy is comparable to that of human labeling, twice as accurate as a commercial detector and the detector of [10], and almost three times as accurate as the more recent method of Zhu and Ramanan [33]. This paper is based on work that was first described in [2].

2 Related Work

Early work on facial feature detection was often described as a component of a larger face processing task. For example, Burl et al. [4] take a bottom up approach to face detection and first detect candidate facial features over the whole image and then select the most face-like constellation using a statistical model of the distances between pairs of features. Other works detect large-scale facial parts such as each eye, the nose, and the mouth and return a contour or bounding box around these components [8], [12].

There is a long history of part-based object descriptions in computer vision and perceptual psychology. Recent approaches have shown a renewed emphasis on parts-based descriptions and attributes because one can learn descriptions of individual parts and then compose them, generalizing to an exponential number of combinations (e.g., [17], [1], [18]). In an influential example of such work, Felzenszwalb et al. [11] train a part-based model for object detection. As in our work, parts' appearance is modeled using trained, discriminative models. The configuration of parts is represented using a star model, in which a deformation cost penalizes deviations from the expected relative location of parts. This is quite unlike our approach, which uses a large number of exemplars nonparametrically to model possible part configurations. The recent Poselets work is especially related to our approach in its data-driven search for object parts [3]. They model human body shape as a configuration of parts in which each part is a cluster of

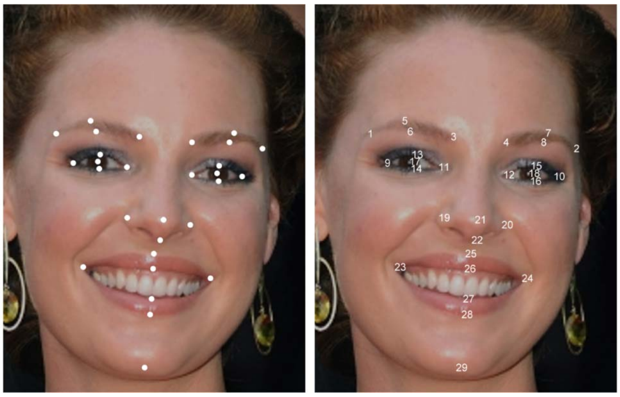

Fig. 2. One of the images in LFPW. Overlaid, we show hand-labeled points obtained using MTurk. Points are numbered to match Figs. 3 and 5.

examples that share a common appearance and configuration. By clustering examples of body parts that are nearby in configuration space, they create a nonparametric description of parts and their configurations across poses, which is related to our own nonparametric approach. We use RANSAC to apply our nonparametric shape model. While using a Gaussian shape model, Li et al. [19] use RANSAC to robustly initialize this model to find automobiles in cluttered environments.

In this paper, we provide a method for localizing parts by detecting finer scale fiducial points or microfeatures [23], as shown in Fig. 2. Many fiducial point detectors include classifiers that are trained to respond to a specific fiducial (e.g., left corner of the left eye). These classifiers take as input raw pixel intensities over a window or the output of a bank of filters (e.g., wavelets [5], Gaussian derivative filters [4], [12], Gabor filters [14], [28], or Haar-like features [7], [9]). These local detectors are scanned over a portion of the image and may return one or more candidate locations for the part or a "score" at each location. This local detector is often a binary classifier (feature or not feature). For example, Zhan et al. [31] have applied the Viola-Jones [27] style detector to facial features. Unlike face detection where the detector is scanned over the entire image area and where Viola and Jones [27] demonstrated efficient detection using a cascade of weak classifiers, bounds on the location of a fiducial are readily determined from the face detection box. Searching over a smaller region that includes the actual part location reduces the chance of false detections with minimal impact of missing fiducials [7]. In addition, since fewer locations are tested, more costly classifiers can be used. In this paper, we use support vector machines (SVMs) [5] with a radial basis function (RBF) kernel. Even so, false detections occur often, even for well-trained classifiers, because different portions of the image can have the appearance of the same fiducial under some imaging conditions. For example, a common error is for a "left corner of left eye" detector to respond to the left corner of the right eye. Eckhardt et al. [9] achieve robustness and handle greater pose variation by using a large area of support for the detector covering, for example, an entire eye or the nose with room to spare.

To better handle larger pose variation, constraints can be established about the relative location of parts to each other rather than the location of each part to the detector box. This can be expressed as predicted locations, bounding regions, or as a conditional probability distribution of one part location given another location [7]. Alternatively, the joint probability distribution of all the parts can be used, and one model is that they form a multivariate normal distribution whose mean is the average location of each part. This is the model underlying active appearance models and active shape models, which have been used for facial feature point detection in near frontal images [6], [7], [21]. Saragih et al. [25] extend this to use a Gaussian mixture model, whereas Everingham et al. [10] handle a wider range of pose, lighting, and expression by modeling the joint probability of the location of nine fiducials relative to the bounding box with a mixture of Gaussian trees. Zhu and Ramanan [33] also handle variations in pose using a mixture of trees model in which parts are shared. Like [10], we do not believe that a joint distribution of part locations over a wide range of poses is adequately modeled by a single Gaussian, but instead of a mixture model, we take a nonparametric approach and use the part locations in a large number of labeled exemplar images to model the joint distribution.

While a number of approaches balance local feature detector responses on the image with prior global information about the feature configurations [6], [7], [13], [21], [25], [26], optimizing the resulting objective function remains a challenge. The locations of some parts vary significantly with expression (e.g., the mouth, eyebrows), whereas others such as the eye corners and nose are more stable. Consequently, some detection methods organize their search to first identify the stable points; the locations of the mouth points are then constrained, possibly through a conditional probability, by the locations of stable points [26]. This approach fails when the stable points cannot be reliably detected, for example, when the eyes are hidden by sunglasses or occluded by hair (a very common occurrence). In contrast, our approach uses a RANSAC-like sampling to randomly select among the different types of parts and therefore tolerates occlusion of some facial features.

A few authors have released software implementations of their facial feature point detection method [10], [28], [33], and because of the utility of detected fiducial points, commercial products have become available by Betaface, face.com, Luxand, Omron, PittPatt, and others. While some of these systems can handle nonfrontal images and detect up to 40 fiducials, the underlying methods are not disclosed and evaluations of these methods have not been published.

3 Face Part Localization

In this section, we describe how we build our local and global detectors using a training image set, described in Section 5.1, with manually annotated part locations.

3.1 Local Detectors

For each part, we build a sliding window detector that can be scanned over a region of the image. These sliding window detectors are simply support vector machine (SVM) classifiers with grayscale scale-invariant feature transform (SIFT) [20] descriptors as features. We compute the SIFT descriptor window at two scales: roughly one-fourth and one-half the interocular distance. (In practice, this is computed relative to the size of the face detector's bounding box.) These two SIFT descriptors are then concatenated to form a single 256-dimensional feature vector for the SVM classifier.

For all of the training samples, we rescale the images so that the faces have an interocular distance of roughly 55 pixels. Positive samples are taken at the manually annotated part locations. Negative samples are taken at least one-fourth of the interocular distance away from the annotated locations. Additional samples are synthesized in two ways. First, we perform a left-right flip of all faces to double the number of training samples. Second, we perform five random rotations of each face where the rotation is selected uniformly from the interval $[-15$ degrees, $+15$ degrees].

Although we use local detectors, rather than regressors, the detectors return a score at each point $\mathbf{x}$ in the image (or in some smaller region around the face as inferred from an earlier face detection step), indicating the distance from a location to a separating hyperplane used in classification. The detector score $\mathbf{d}(\mathbf{x})$ indicates the likelihood that the desired part is located at point $\mathbf{x}$ in the image. This score is normalized to behave like a probability by dividing by the sum of the scores in the detector window. Once normalized, we write this score as $P(\mathbf{x} \mid \mathbf{d})$, i.e., the probability that the fiducial is at location $\mathbf{x}$ given all the scores in the detection window.

Nevertheless, as the local detectors are imperfect, the correct location will not always be at the location with the highest detector score. This can happen for many of the aforementioned reasons, including occlusions due to head pose and visual obstructions such as hair, glasses, hands, microphones, and so on. Yet these mistakes in the local detector almost always happen at places that are inconsistent with positions of the other—correctly detected—fiducial points. In the next section, we describe how we build our global detectors to better handle the cases where the local detectors are likely to go astray.

3.2 Global Detectors

Although faces come in different shapes, present themselves to the camera in many ways, and may possess often extreme facial expressions, there are strong anatomical and geometric constraints that govern the layout of face parts and their location in images. We do not try to model these constraints explicitly, but rather let our training data dictate this implicitly. Here, we need to consider all the part locations taken together to develop a global detector for a collection of fiducial points. To exploit this, we use a global model for a configuration of part locations.

We will let X denote the true location of the fiducial points in the image, while X_k will refer to a set of models that are used to generate the image fiducials. More specifically, let $X = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^n\}$ denote the locations of n parts, where $\mathbf{x}^i$ is the location of the ith part. Let $D = \{\mathbf{d}^1, \mathbf{d}^2, \ldots, \mathbf{d}^n\}$ denote the measured detector responses, where $\mathbf{d}^i$ is the window of scores returned by the ith local detector. We want to find the value of X that maximizes the probability of X given the measurements from our local detectors, i.e.,

$$X^* = \arg\max_X P(X \mid D). \tag{1}$$

Let X_k (where $k = 1, \ldots, m$) denote the locations of the n parts in the kth of m exemplars, and let $X_{k,t}$ be the locations of the parts in exemplar k transformed by some similarity transformation t; we call $X_{k,t}$ a global model.

If we suppose that each X is generated by one of our global models $X_{k,t}$, we can expand $P(X \mid D)$ as follows:

$$P(X \mid D) = \sum_{k=1}^{m} \int_{t \in T} P(X \mid X_{k,t}, D) P(X_{k,t} \mid D) dt, \tag{2}$$

where our collection of m exemplars X_k along with similarity transformations t have been introduced into the calculation of $P(X \mid D)$ and then marginalized out.

By conditioning on the global model $X_{k,t}$, we can now treat the locations of the parts $\mathbf{x}^i$ as conditionally independent of one another, given $X_{k,t}$. We write the parts of $X_{k,t}$ as $\mathbf{x}^i_{k,t}$, each denoting a part of exemplar k transformed by t. We assume that given $\mathbf{x}^i_{k,t}$ and $\mathbf{d}^i$, $\mathbf{x}^i$ will be conditionally independent of the other part locations and detector outputs, and rewrite the first term of (2) as

$$P(X \mid X_{k,t}, D) = \prod_{i=1}^{n} P\big(\mathbf{x}^i \mid \mathbf{x}^i_{k,t}, \mathbf{d}^i\big) \tag{3}$$

$$= \prod_{i=1}^{n} \frac{P\big(\mathbf{x}^i_{k,t} \mid \mathbf{x}^i, \mathbf{d}^i\big) P(\mathbf{x}^i \mid \mathbf{d}^i)}{P\big(\mathbf{x}^i_{k,t} \mid \mathbf{d}^i\big)}. \tag{4}$$

$\mathbf{x}^i_{k,t}$ represents the position of part i in exemplar k, supposing this exemplar has been used to generate the image, after being transformed by t. $P(\mathbf{x}^i_{k,t} \mid \mathbf{d}^i)$ represents the distribution of this part location, conditioned on the detector output for that part.

Since knowing the true location of the parts trumps any information provided by the detector about which exemplar and transformation were used to generate an image, $P(\mathbf{x}^i_{k,t} \mid \mathbf{x}^i, \mathbf{d}^i) = P(\mathbf{x}^i_{k,t} \mid \mathbf{x}^i)$. Also, since the relation between the transformed model fiducial and the true fiducial is translationally invariant, it should only depend on $\Delta\mathbf{x}^i_{k,t} = \mathbf{x}^i_{k,t} - \mathbf{x}^i$. With these observations, we can rewrite (4) as

$$P(X \mid X_{k,t}, D) = \prod_{i=1}^{n} \frac{P\big(\Delta\mathbf{x}^i_{k,t}\big) P(\mathbf{x}^i \mid \mathbf{d}^i)}{P\big(\mathbf{x}^i_{k,t} \mid \mathbf{d}^i\big)}. \tag{5}$$

Moving on to the second term in (2), we can use Bayes' rule to get

$$P(X_{k,t} \mid D) = \frac{P(D \mid X_{k,t}) P(X_{k,t})}{P(D)} \tag{6}$$

$$= \frac{P(X_{k,t})}{P(D)} \prod_{i=1}^{n} P(\mathbf{d}^i \mid \mathbf{x}^i_{k,t}), \tag{7}$$

where again conditioning on the global model $X_{k,t}$ allows us to treat the detector responses $\mathbf{d}^i$ as conditionally independent of one another.

A final application of Bayes' rule lets us rewrite (7) as

$$P(X_{k,t} \mid D) = \left[\frac{P(X_{k,t})}{P(D)} \frac{\prod_{i=1}^{n} P(\mathbf{d}^i)}{\prod_{i=1}^{n} P\big(\mathbf{x}^i_{k,t}\big)}\right] \prod_{i=1}^{n} P\big(\mathbf{x}^i_{k,t} \mid \mathbf{d}^i\big) \tag{8}$$

$$= C\prod_{i=1}^{n} P\big(\mathbf{x}^i_{k,t} \mid \mathbf{d}^i\big). \tag{9}$$

Note that the terms within the square bracket in (8) that depend only on D are constant given the image. Also note that the terms within the square bracket that depend only on $X_{k,t}$ are also constant because we assume a uniform distribution on our global models. Therefore, we may reduce all the terms within the square bracket to a single constant C.

Combining (1), (2), (5), and (9) yields

$$X^* = \arg\max_X \sum_{k=1}^{m} \int_{t\in T} \prod_{i=1}^{n} P\big(\Delta\mathbf{x}^i_{k,t}\big) P(\mathbf{x}^i \mid \mathbf{d}^i) dt, \tag{10}$$

where X^* is the estimate for the part locations.

The first term $P(\Delta\mathbf{x}^i_{k,t})$ is taken to be a 2D Gaussian distribution centered at the model location $\mathbf{x}^i_{k,t}$. Each part i has its own Gaussian distribution. These distributions model how well the part locations in the global model fit the true locations. If we had a large number of exemplars in our labeled dataset from which to construct these global models—i.e., if m were very large—then we would expect a close fit and low variances for these distributions. To estimate the covariance matrices for the part locations, we do the following. For each exemplar X_j from our labeled data set, we find a sample X_k from the remaining exemplars and a transformation t that gives the best L_2 fit to X_j. We compute the difference $X_j - X_{k,t}$ and normalize it by the interocular distance. These normalized differences are used to compute the covariance matrices for each part location.

The second term $P(\mathbf{x}^i \mid \mathbf{d}^i)$ is computed as follows: We take the estimated location $\mathbf{x}^i$ for part i and look up the response for the ith detector at that point in the image, i.e., $\mathbf{d}^i(\mathbf{x}^i)$. This value is then normalized to behave like a probability by dividing by the sum of $\mathbf{d}^i(\mathbf{x})$ for all $\mathbf{x}$ in the detector window.

4 Optimization

Computing the sum and integral in (10) is challenging, as they are taken over all global models k and all similarity transformations t. However, we note from (2) that if $P(X_{k,t} \mid D)$ is very small for a given k and t, it will be unlikely to contribute much to the overall sum and integration. Our strategy is thus to consider only those global models k with transformations t for which $P(X_{k,t} \mid D)$ is large.

In a sense, we wish to perform a Monte Carlo integration of (10) where the global models $X_{k,t}$ we choose are the ones that are likely to contribute to the sum and integral. In the following section, we describe how we select a list of k and t that are used to compute this integration.

4.1 Choosing the Global Models $X_{k,t}$

We wish to optimize $P(X_{k,t} \mid D)$ over the unknowns k and t. This optimization is nonlinear and not amenable to gradient descent-type algorithms. First, k is a discrete variable with a large number of possible values (in our experiments, we have about 1,000 possible exemplars). Second, we expect that even for a fixed k, different values of t will produce large numbers of local optima because our fiducial detectors usually produce a multimodal output. Transformations that align a model with any subsets of these modes are likely to produce local optima in our optimization function.

To cope with this, we adopt a RANSAC-like generate-and-test approach. We generate a large number of plausible values for k and t. We evaluate each of these using (9). We keep track of the m^* best global models, i.e., the m^* best pairs k and t. This is done in the following steps:

1. Select a random exemplar k.
2. Select two random parts. Randomly match each of these to one of the g highest modes of the detector output for that part.
3. Set t to be the similarity transformation that aligns the model fiducial points with the detector modes.
4. Evaluate (9) for this k, t.
5. Repeat Steps 1 to 4 r times.
6. Record in a set $\mathcal{M}$ the m^* pairs k and t for which (9) in Step 4 is the largest.

In our current experimental system, we use the values $r = 10{,}000$, $g = 2$, and $m^* = 100$.

4.2 Estimating X

In the previous section, we used a RANSAC-like procedure to find a list $\mathcal{M}$ of m^* global models $X_{k,t}$ for which $P(X_{k,t} \mid D)$ is the largest. With these in hand, we approximate the optimization for X in (10) as

$$X^* = \arg\max_X \sum_{k,t\in\mathcal{M}} \prod_{i=1}^{n} P\big(\Delta\mathbf{x}^i_{k,t}\big) P(\mathbf{x}^i \mid \mathbf{d}^i), \tag{11}$$

where the sum is now only taken over those $k, t \in \mathcal{M}$.

Equation (11) is essentially multilinear in $\mathbf{x}^i$, so we may not optimize each term independently. However, we can initialize a solution in this way. So, to find the best X^*, we first find an initial estimate $\mathbf{x}^i_0$ for each part i as

$$\mathbf{x}^i_0 = \arg\max_{\mathbf{x}^i} \sum_{k,t\in\mathcal{M}} P\big(\Delta\mathbf{x}^i_{k,t}\big) P(\mathbf{x}^i \mid \mathbf{d}^i). \tag{12}$$

This is equivalent to solving for x^i_0 by setting all $P(\Delta\mathbf{x}^j_{k,t})$ and $P(\mathbf{x}^j|\mathbf{d}^j)$ to a constant in (10) for all $j \neq i$. To compute each $\mathbf{x}^i_0$, we merely need to multiply the normalized detector output by a Gaussian function centered at $\mathbf{x}^i_{k,t}$ with the covariances calculated as described at the end of Section 3.2. Then, we find the image location $\mathbf{x}^i_0$ where the sum of the resulting products is maximized. The initial estimates, $\mathbf{x}^i_0$, $i \in 1, \ldots, n$, can then be used to initialize an optimization of (11) to find the final estimates $\mathbf{x}^{i*}$ that make up X^*.

In practice, we find that these initial estimates suffice, and further optimization is unnecessary. Therefore, following RANSAC to select transformed exemplars, we independently optimize the location of each fiducial point by averaging the product of the detector outputs and Gaussian distributions placed around the location of each transformed exemplar fiducial.

5 Experiments

Our work focuses on localizing parts in natural face images, taken under a wide range of poses, lighting conditions, and

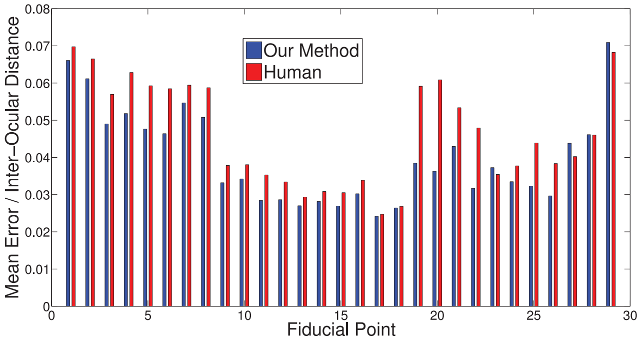

Fig. 3. Mean error of our fiducial detector on the LFPW dataset compared to the mean variation in human labeling. The fiducial labels are shown in Fig. 2, and the error is the fraction of interocular distance. Our detector is almost always more accurate.

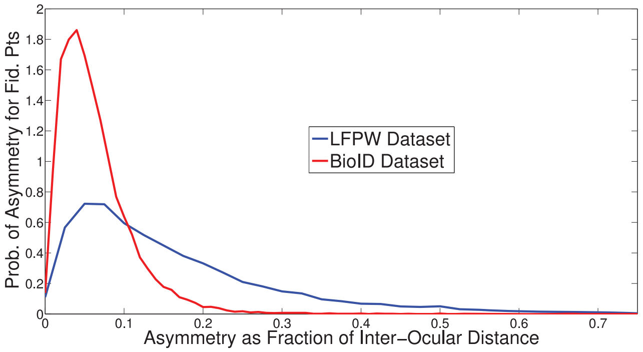

Fig. 4. The distribution of the asymmetry measure over images in the LFPW and BioID datasets. The BioID dataset consists mostly of frontal images, resulting in a sharp peak near the y-axis (i.e., nearly symmetric faces), whereas the LFPW dataset contains many more off-frontal faces, making it far more challenging.

facial expressions, in the presence of occluding objects such as sunglasses or microphones. Existing datasets for evaluating part localization do not contain the range of conditions that we aim to address in this paper, and so we show results on the Labeled Faces in the Wild (LFW) [15] dataset and on our new dataset, Labeled Face Parts in the Wild (LFPW). Our most significant results are on these datasets.

Since researchers have recently reported results on BioID, we present comparative results on BioID. Like most datasets used to evaluate part localization on face images, BioID contains near-frontal views and less variation in viewing conditions than LFPW.

5.1 Data Sets

Our new LFPW dataset consists of 3,000 faces from images downloaded from the web using simple text queries on sites such as google.com, flickr.com, and yahoo.com. The 3,000 faces were detected using a commercial, off-the-shelf (COTS) face detection system. Faces were excluded only if they were incorrectly detected by the COTS detector or if they contained text on the face. Note also that our COTS face detector does not detect faces in or near profile, and so these images are implicitly excluded from our dataset.

To obtain ground-truth data, 35 fiducial points on each face were labeled by workers on Amazon Mechanical Turk (MTurk). Of these 35 points, we only used 29 in this paper and excluded points associated with the ears. Fig. 2 illustrates the location of these points. Each point was labeled by three different MTurk workers. We used the average location as ground truth for the fiducial point. A subset of this data is made available at kbvt.com.

Fig. 6 shows example images from LFPW, along with our results. There is a degree of subjectivity in the way humans label the location of fiducial points in the images, and this is seen in Fig. 3, which shows the variation among the MTurk workers. Some parts like the eye corners are more consistently labeled, whereas the brows and chin are labeled less accurately.

Labeled Faces in the Wild (LFW) [15] is an existing large dataset of real-world images gathered from news sites. It consists of 13,233 images of 5,749 public figures, taken in unconstrained settings and with noncooperative subjects. It is commonly used for benchmarking face verification algorithms, and as such has been used by many researchers. It is qualitatively similar to our own LFPW dataset, but due to its age (collected several years ago), it contains slightly lower quality images in general.

The publicly available BioID dataset contains 1,521 images, each showing a frontal view of a face of one of 23 different subjects [16]. We used 17 fiducial points that had been marked for the FGNet project, and used in the me_{17} error measure as defined in [6]. This dataset has been widely used, allowing us to benchmark our results with prior work. Note that we trained using the LFPW dataset and tested on BioID in our experiments. There are considerable differences in the viewing conditions of these two datasets. Furthermore, the location of parts in LFPW do not always match those of BioID, and so we computed a fixed offset between parts that were defined differently (e.g., whereas the left and right nose points are outside of the nose in LFPW, they are below the nose in BioID). Fig. 8 shows some example images, along with our results.

To compare the challenge presented by different datasets, we created a measure of the asymmetry of the fiducials in an image. We reflect fiducials about a vertical line passing through their centroid and compute the mean distance between fiducial pairs that are symmetric in 3D (e.g., the outer corner of the left and right eyes). For a frontal image without occluded fiducials, the measure would be near zero. For faces that are rotated in 3D or about the optical axis, the asymmetry increases with the extent of rotation. Fig. 4 shows the distribution of the asymmetry measure for the BioID and LFPW datasets, and the distributions indicate that LFPW is truly a more challenging dataset.

5.2 Results

In our experiments with LFPW, we randomly split the dataset into 1,100 training images and 300 test images. (An additional 1,600 images have been held out for subsequent evaluations at future dates.) Training images were used to train our SVM-based fiducial detectors and also served as the exemplars for computing our global models X_k.

We evaluate the results of each localization by measuring the distance from each localized part to the average of three locations supplied by MTurk workers. Error is measured as a fraction of the interocular distance to normalize for image size. Fig. 3 shows the resulting error broken down by part.

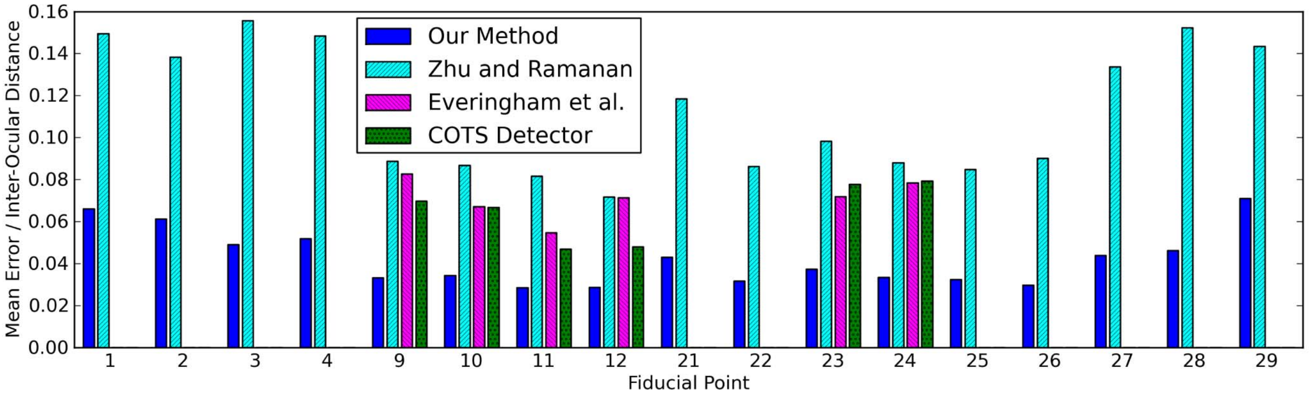

Fig. 5. Comparison of fiducial point detectors on the LFPW dataset. We show the mean error as a fraction of interocular distance for our method, Everingham et al. [10], Zhu and Ramanan [33], and a commercial off-the-shelf (COTS) system. Only the fiducial points shared in common with our method are shown. See Fig. 2 for locations of the fiducial points. Our detector is between two to three times as accurate as all other methods.

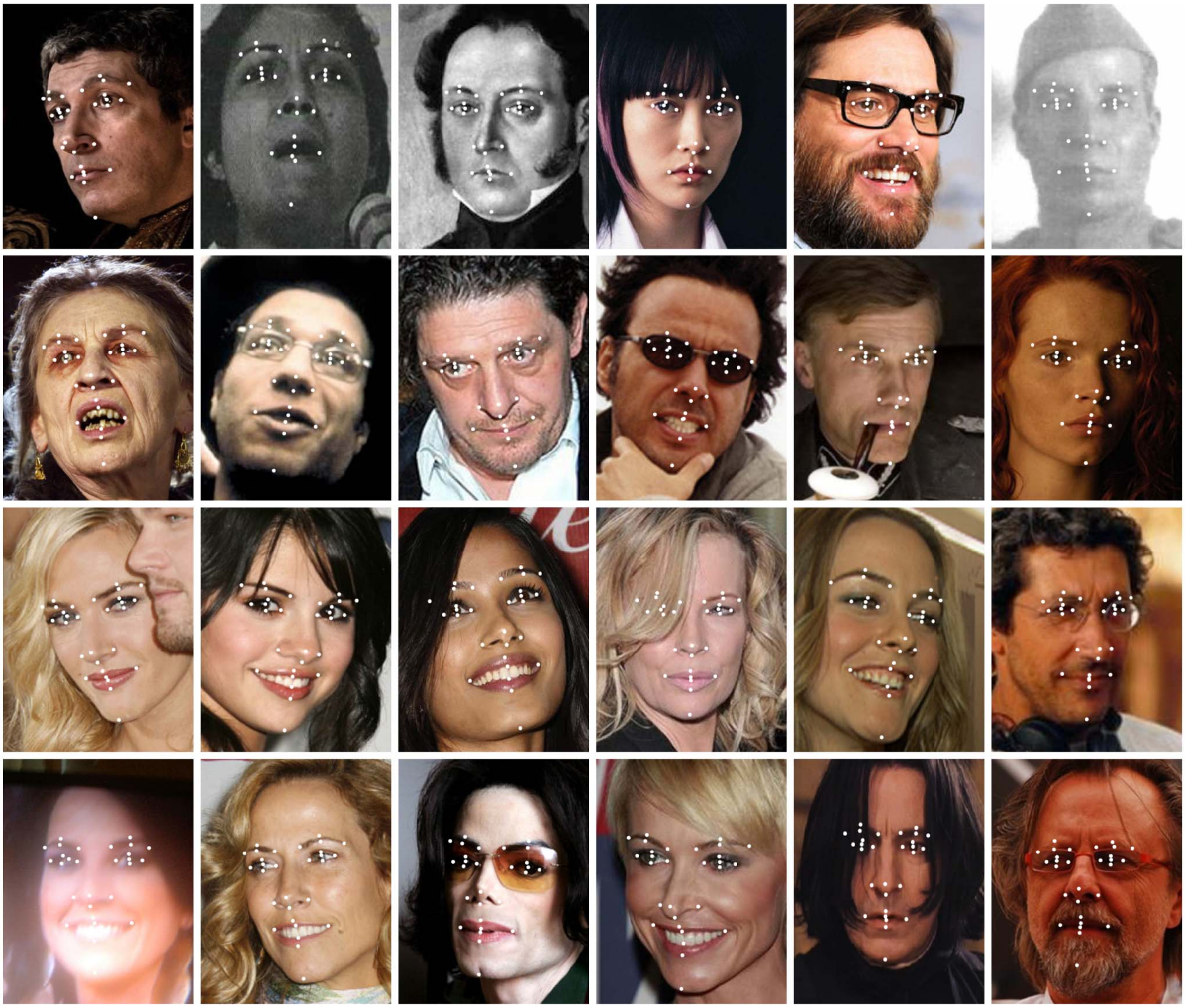

Fig. 6. Images from Labeled Face Parts in the Wild (LFPW), along with parts located by our detector.

This figure also compares the error in our system to the average distance between points marked by one MTurk worker and the average of the points marked by the other two. We can see that this distance almost always exceeds the distance from points localized by our system to the average of the points marked by humans. It is worth noting that the eye points (9-18) are the most accurate, the nose and mouth points (19-29) are slightly worse, and the chin and eyebrows (1-8, 29) are least accurate. This trend is consistent between human and automatic labeling.

Figs. 1 and 6 show results on some representative images. We highlight a few characteristics of these results. These

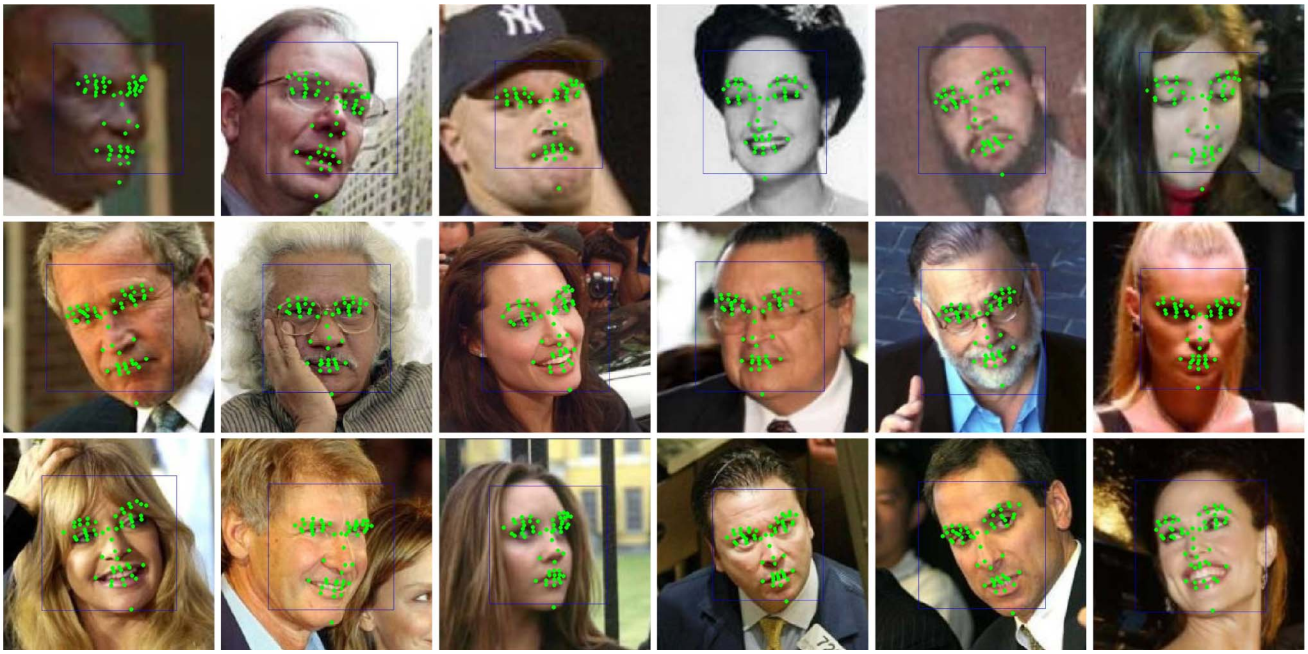

Fig. 7. Images from Labeled Faces in the Wild (LFW), along with OpenCV face detections (blue rectangles) and 55 parts localized by our detector (green dots). Our mean accuracy is 5.18 percent of the interocular distance.

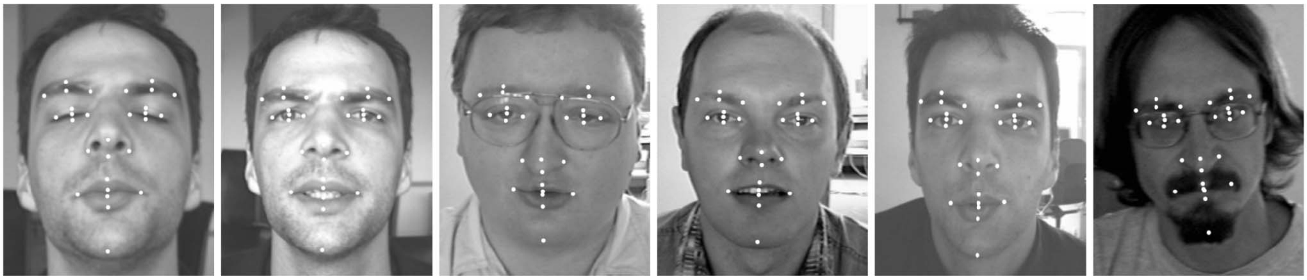

Fig. 8. Images from BioID, along with parts localized by our detector.

images include nonfrontal images including viewpoints from below (Row 1, Col. 2 and Row 2, Col. 2), difficult lighting (Row 4, Col. 1), glasses (Row 1, Col. 5), sunglasses (Row 2, Col. 4 and Row 4, Col. 3), partial occlusion (Row 2, Col. 5 by a pipe and Row 3, Col. 4 by hair), an artist drawing (Row 1, Col. 3), theatrical makeup (Row 2, Col. 1), and so on. The localizer requires 400 milliseconds per fiducial on an Intel Core i7 3.06-GHz machine; most of the time is spent evaluating the local detectors.

In Fig. 5, we compare our LFPW results to those of a commercial face and fiducial detector[2] and the detector of [10]. Since we had access to executables, we ran these detectors over the LFPW test set and used the same metric for evaluation. The commercial system locates six fiducials (both corners of each eye and the outer corners of the mouth), so we compare results on those fiducials only. At roughly 3 percent mean error rate, our results are roughly twice as accurate as the commercial system and [10].

We also ran the method for face detection, pose estimation, and landmark localization presented in [33] on the 300 test images in the LFPW dataset per a request by a reviewer and the editor, using the publicly available implementation by the authors. It should be noted that this method was published and code was released after submission of our paper to *TPAMI*. We used the pretrained model with 1,050 parts because the authors of [33] said that it gave the best performance on localization and was used for reporting the localization results in [33].

Fig. 5 shows the errors for the 17 face parts that are common between our method and theirs; these include fiducials on the brows, eyes, nose, mouth, and chin. Because the fiducials might not be defined as the same location for different methods, the bias between the detected locations and those of an average user was first removed. For 300 test images, the ground-truth face was not detected in 16 images by the detector in [33], and for 10 images, the detector returned near profile fiducial configurations with different fiducials than used in our method. These 26 images were removed from the evaluation because they would unfairly penalize the method of Zhu and Ramanan [33].

As shown in Fig. 5, the error for [33] was twice as large for the eyebrows and almost three times as large for the eyes. The sides and top of the mouth are twice as large, but the lower lip is even more poorly localized, and the error is nearly three times as large; lower lips tend to be harder for most methods, especially when the mouth is open. The average runtime was 107.6 s per image on a 3.06 Ghz Macbook Pro, which is about 10 times slower than our method.

[2] For contractual reasons, we may not identify the commercial system in this paper.

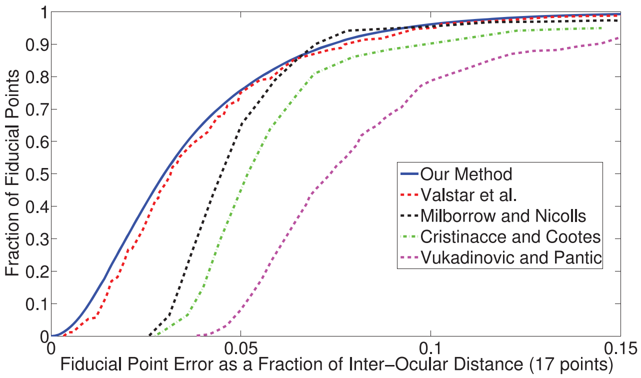

Fig. 9. Cumulative error distribution curves comparing our system to several others on the BioID dataset. All comparative results are from [26]. We outperform all previously published results.

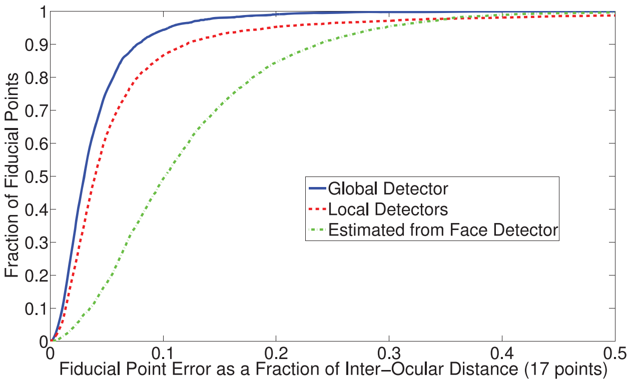

Fig. 10. Cumulative error distribution of our system on the LFPW dataset compared to locations predicted using the face detector box or found with just our local detectors. (Note the different x-axis scale from Fig. 9.)

Fig. 6 shows some examples of errors of our system. In Row 1, Cols. 2 and 5, local cues for the chin are indistinct, and the chin is not localized exactly. Row 2, Col. 4 shows an example in which the lower lip is incorrectly localized. This can happen when the mouth is open and a row of teeth are visible. We believe that these errors can be primarily attributed to the local detectors; in future work, we plan to make use of color-based representations that can more easily distinguish between lips and teeth. And in Row 4, Col. 1, the left corner of the left eyebrow is too low, presumably due to occlusion from the hair.

Fig. 7 shows results of our part localizers on images from the labeled faces in the wild (LFW) data set. These localizers were trained using 6,080 manually labeled images from Columbia's PubFig [17] dataset. Fifty-five points on the face were labeled, allowing for finer-grained localization on the face. Note the consistency of localization despite changes in the face box returned by OpenCV (blue rectangles). Our mean error across all images of LFW and all 55 fiducial points is 5.18 percent of the interocular distance.

We have also applied our part localizer to the BioID faces and show some example output images in Fig. 8. Results have been reported on this dataset by a number of authors. Fig. 9 shows the cumulative error distribution of the me_{17} error measure (mean error of 17 fiducials) defined in [6]. Fig. 9 compares the results of our method to those reported by authors in [6], [21], [26], [28] . Our results are similar to but slightly better than those of Valstar et al. [26], who, to our knowledge, report the best current results on this dataset. We note that we train on a very different dataset (LFPW) and use some fiducials whose locations are defined a bit differently.

Finally, in Fig. 10, we return to LFPW and show the cumulative error distribution of the me_{17} error measure for our method applied to LFPW. Even though LFPW is a more difficult dataset per Fig. 4, the cumulative error distribution curve on LFPW is almost identical to our cumulative error distribution curve on BioID. (Note that the figures have different scales along the x-axis.) Fig. 10 also shows the cumulative error distribution when only the local detectors are used and when locations are predicted solely from the face box. While the local detectors are effective for most fiducial points, there is a clear benefit from using the consensus of global models. Many of the occluded fiducial points are incorrectly located by the local detectors, as evidenced by the slow climb toward 1.0 of the red curve.

6 Conclusions

We have described a new approach to localizing parts in face images. Our primary innovation is a Bayesian model that combines local detector outputs with a consensus of nonparametric global models for part locations, computed from exemplars. Our localizer is accurate over a large range of real-world variations in pose, expression, lighting, makeup, and image quality. To train and test this system, we introduce LFPW, a large, real-world dataset of hand-labeled images. Our system demonstrates strong performance on this dataset, significantly outperforming previous research systems and a commercial system. We also demonstrate state-of-the-art performance on the LFW and BioID datasets.

Acknowledgments

This research was performed at Kriegman-Belhumeur Vision Technologies and funded all or in part by the US Government.

References

[1] First Int'l Workshop Parts and Attributes, 2010.
[2] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar, "Localizing Parts of Faces Using a Consensus of Exemplars," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.
[3] L. Bourdev and J. Malik, "Poselets: Body Part Detectors Trained Using 3D Human Pose Annotations," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 1365-1372, 2009.
[4] M. Burl, T. Leung, and P. Perona, "Face Localization via Shape Statistics," *Proc. Workshop Automatic Face and Gesture Recognition*, 1995.
[5] P. Campadelli, R. Lanzarotti, and G. Lipori, "Automatic Facial Feature Extraction for Face Recognition," *Face Recognition*, I-Tech Education and Publishing, 2007.
[6] D. Cristinacce and T. Cootes, "Feature Detection and Tracking with Constrained Local Models," *Proc. British Machine Vision Conf.*, pp. 929-938, 2006.
[7] D. Cristinacce, T. Cootes, and I. Scott, "A Multi-Stage Approach to Facial Feature Detection," *Proc. British Machine Vision Conf.*, pp. 231-240, 2004.
[8] L. Ding and A.M. Martinez, "Precise Detailed Detection of Faces and Facial Features," *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.

[9] M. Eckhardt, I. Fasel, and J. Movellan, "Towards Practical Facial Feature Detection," *Int'l J. Pattern Recognition and Artificial Intelligence,* vol. 23, no. 3, pp. 379-400, 2009.

[10] M. Everingham, J. Sivic, and A. Zisserman, "'Hello! My Name Is.. Buffy'—Automatic Naming of Characters in TV Video," *Proc. British Machine Vision Conf.,* 2006.

[11] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan, "Object Detection with Discriminatively Trained Part-Based Models," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 32, no. 9, pp. 1627-1645, Sept. 2010.

[12] N. Gourier, D. Hall, and J.L. Crowley, "Facial Features Detection Robust to Pose Illumination and Identity," *Proc. Int'l Conf. Systems, Man, and Cybernetics,* 2004.

[13] L. Gu and T. Kanade, "A Generative Shape Regularization Model for Robust Face Alignment," *Proc. European Conf. Computer Vision,* pp. 413-426, 2008.

[14] E. Holden and R. Owens, "Automatic Facial Point Detection," *Proc. Asian Conf. Computer Vision,* pp. 731-736, 2002.

[15] 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," Technical Report 07-49, Univ. of Massachusetts, Amherst, Oct. 2007.

[16] O. Jesorsky, K.J. Kirchberg, and R.W. Frischholz, "Robust Face Detection Using the Hausdorff Distance," *Proc. Conf. Audio- and Video-Based Biometric Person Authentication,* pp. 90-95, 2001.

[17] N. Kumar, A.C. Berg, P.N. Belhumeur, and S.K. Nayar, "Attribute and Simile Classifiers for Face Verification," *Proc. IEEE Int'l Conf. Computer Vision,* 2009.

[18] B. Leibe, A. Ettlin, and B. Schiele, "Learning Semantic Object Parts for Object Categorization," *Image and Vision Computing,* vol. 26, pp. 15-26, 1998.

[19] Y. Li, L. Gu, and T. Kanade, "Robustly Aligning a Shape Model and Its Application to Car Alignment of Unknown Pose," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 33, no. 9, pp. 1860-1876, Sept. 2011.

[20] D. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," *Int'l J. Computer Vision,* vol. 60, pp. 91-110, 2003.

[21] S. Milborrow and F. Nicolls, "Locating Facial Features with an Extended Active Shape Model," *Proc. European Conf. Computer Vision,* pp. 504-513, 2008.

[22] M. Pantic and L.J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 22, no. 12, pp. 1424-1445, Dec. 2000.

[23] M. Reinders, R.W.C. Koch, and J. Gerbrands, "Locating Facial Features in Image Sequences Using Neural Networks," *Proc. Conf. Automatic Face and Gesture Recognition,* pp. 230-235, 1997.

[24] S. Romdhani, V. Blanz, and T. Vetter, "Face Identification by Fitting a 3D Morphable Model Using Linear Shape and Texture Error Functions," *Proc. European Conf. Computer Vision,* pp. 3-19, 2002.

[25] J.M. Saragih, S. Lucey, and J. Cohn, "Face Alignment through Subspace Constrained Mean-Shifts," *Proc. IEEE Int'l Conf. Computer Vision,* Sept. 2009.

[26] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial Point Detection Using Boosted Regression and Graph Models," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2729-2736, 2010.

[27] P. Viola and M. Jones, "Robust Real-Time Face Detection," *Int'l J. Computer Vision,* vol. 57, pp. 137-154, 2004.

[28] D. Vukadinovic and M. Pantic, "Fully Automatic Facial Feature Point Detection Using Gabor Feature Based Boosted Classifiers," *Proc. Int'l Conf. Systems, Man, and Cybernetics,* pp. 1692-1698, 2005.

[29] M.-H. Yang, D.J. Kriegman, and N. Ahuja, "Detecting Faces in Images: A Survey," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 24, no. 1, pp. 34-58, Jan. 2002.

[30] Z. Yue, W. Zhao, and R. Chellappa, "Pose-Encoded Spherical Harmonics for Face Recognition and Synthesis Using a Single Image," *EURASIP J. Advance Signal Process,* vol. 2008, pp. 1-18, 2008.

[31] C. Zhan, W. Li, P. Ogunbona, and F. Safaei, "Real-Time Facial Feature Point Extraction," *Proc. Eighth Pacific Rim Conf. Advances in Multimedia Information Processing,* pp. 88-97, 2007.

[32] L. Zhang and D. Samaras, "Face Recognition from a Single Training Image under Arbitrary Unknown Lighting Using Spherical Harmonics," *IEEE Trans. Pattern Analysis and Machine Intelligence,* vol. 28, no. 3, pp. 351-363, Mar. 2006.

[33] X. Zhu and D. Ramanan, "Face Detection, Pose Estimation and Landmark Localization in the Wild," *Proc. IEEE Conf. Computer Vision and Pattern Recognition,* pp. 2879-2886, 2012.

Peter N. Belhumeur received the ScB degree in information sciences from Brown University in 1985 and the PhD degree in engineering sciences from Harvard University under the direction of David Mumford in 1993. He was a postdoctoral fellow at the University of Cambridge's Isaac Newton Institute for Mathematical Sciences in 1994. He became assistant, associate, and professor of electrical engineering at Yale University in 1994, 1998, and 2001, respectively. He joined Columbia University in 2002, where he is currently a professor in the Department of Computer Science and the director of the Laboratory for the Study of Visual Appearance. His main research focuses on problems in computer vision, biometrics, and machine learning. Applications include face and object recognition, content-based image retrieval, computer graphics, and automatic species identification. He received the Presidential Early Career Award for Scientists and Engineers, the US National Science Foundation Career Award, and the Yale University Junior Faculty Fellowship. His papers have received the Siemens Best Paper Award at CVPR 1996, the Olympus Prize at ECCV 1998, the Best Paper Honorable Mention Award at CVPR 2000, and the Mark Everingham Prize at BMVC 2012. He is also a co-creator of Leafsnap and recent corecipient of the 2011 Edward O. Wilson Award.

David W. Jacobs received the BA degree from Yale University in 1982. He is a professor in the Department of Computer Science at the University of Maryland with a joint appointment in the University's Institute for Advanced Computer Studies. From 1982 to 1985, he worked for Control Data Corporation on the development of database management systems, and attended graduate school in computer science at New York University. From 1985 to 1992, he attended MIT, where he received the MS and PhD degrees in computer science. From 1992 to 2002, he was a research scientist and then a senior research scientist at the NEC Research Institute. In 1998, he spent a sabbatical at the Royal Institute of Technology (KTH) in Stockholm, and in 2008 spent a sabbatical at the Ecole Normale Supérieure de Cachan. In 2002, he joined the CS Department at the University of Maryland. His research has focused on human and computer vision, especially in the areas of object recognition and perceptual organization. He has also published articles in the areas of motion understanding, memory and learning, computer graphics, human-computer interaction, and computational geometry. He has served as an associate editor of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* and has assisted in the organization of many workshops and conferences, including serving as a program co-chair for CVPR 2010. He and his coauthors received honorable mention for the Best Paper Award at CVPR 2000. He also coauthored a paper that received the Best Student Paper Award at UIST 2003. In collaboration with researchers at Columbia University and the Smithsonian Institution, he created Leafsnap, an app that uses computer vision for plant species identification, for which he and his collaborators have been awarded the 2011 Edward O. Wilson Biodiversity Technology Pioneer Award. He is a member of the IEEE.

David J. Kriegman received the BSE degree in electrical engineering and computer science from Princeton University in 1983. He received the MS degree in 1984 and the PhD degree in electrical engineering in 1989 from Stanford University. Since 2002, he has been a professor of computer science and engineering in the Jacobs School of Engineering, University of California, San Diego (UCSD). Prior to joining UCSD, he was an assistant and associate professor of electrical engineering and computer science at Yale University (1990-1998) and an associate professor with the Computer Science Department and Beckman Institute at the University of Illinois at Urbana-Champaign (1998-2002). He was founding CEO and presently serves as chief scientist of Taaz Inc., the leader in photorealistic virtual try on. He is also a partner of KBVT. His research is in computer vision with particular application to face recognition, robotics, computer graphics, microscopy, and coral reef ecology. He was chosen for the National Science Foundation Young Investigator Award, and has received Best Paper Awards at the 1996 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), the 1998 European Conference on Computer Vision, and the 2007 International Conference on Computer Vision (Marr Prize, runner up) as well as the 2003 Paper of the Year Award from the *Journal of Structural Biology*. He has served as program cochair of CVPR 2000 and general co-chair of CVPR 2005. He was the editor-in-chief of the *IEEE Transactions on Pattern Analysis and Machine Intelligence* from 2005 to 2008. He is a senior member of the IEEE.

Neeraj Kumar received the BSc degrees (both with highest honors) in computer science and aeronautical engineering from the Georgia Institute of Technology in 2005. He received the PhD degree in computer science from Columbia University in 2011, where he was co-advised by Professors P.N. Belhumeur and S.K. Nayar. He is currently a postdoctoral research scientist at the University of Washington, working with Professor Steven Seitz. His main research interests are the intersection of computer vision and machine learning, developing techniques that leverage large collections of real-world images for a variety of applications. He is particularly interested in the use of intermediate representations such as parts and attributes, both for improving performance on classical vision tasks such as search and recognition, as well as for creating novel applications such as automatic face replacement and exploration of image collections. He received a three-year National Defense Science and Engineering Graduate Fellowship by the American Society for Engineering Education in 2005.