Localizing Parts of Faces Using a Consensus of Exemplars

Peter N. Belhumeur[⋆,†] David W. Jacobs[⋆,‡] David J. Kriegman[⋆,§] Neeraj Kumar[⋆,†]
[⋆]Kriegman-Belhumeur Vision Technologies[*]
[†]Columbia University
[‡]University of Maryland, College Park
[§]University of California, San Diego

Abstract

We present a novel approach to localizing parts in images of human faces. The approach combines the output of local detectors with a non-parametric set of global models for the part locations based on over one thousand 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 a new dataset gathered from the internet and show that our localizer achieves state-of-the-art performance on the less challenging BioID dataset.

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 [21, 23]. 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 – *e.g.*, 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.

[*]This research was performed at Kriegman-Belhumeur Vision Technologies and was funded by the CIA through the Office of the Chief Scientist.

Figure 1. Results of our face part localizer.

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 pre-specified parts in images of human faces taken under a variety of acquisition conditions, including variability in 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 dataset 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. 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 facial hair that occludes the fiducial points; and part of the face may be occluded by a hat, wall, cigarette, hand, or microphone. See Figures 1

and 6. This dataset stands in contrast to datasets such as FERET or BioID which have been used for evaluating fiducial point detection in that the images are not restricted to frontal faces or collected in a controlled manner.

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 non-parametric, making use of our large collection of diverse, labeled exemplars. We then introduce hidden variables for the identity and location of the exemplar 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.

The method is evaluated on two 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. The LFPW dataset is introduced in this paper and is unconstrained. Experimental results demonstrate that our method is more accurate than existing methods on BioID and is just as accurate on the harder LFPW dataset. Furthermore, accuracy is comparable to that of human labeling and is twice as accurate as a commercial detector and the detector of [9].

2. Related Work

Early work on facial feature detection was often described as a component of a larger face processing task. For example, Burl, Leung and Perona [3] 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 [7, 10].

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., [14, 1, 15]). The recent Poselets work is especially related to our approach in its data-driven search for object parts [2].

In this paper, we provide a method for localizing parts by detecting finer-scale fiducial points or microfeatures [18], 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 [4], Gaussian Derivative filters [3, 10], Gabor filters [12, 22], or Haar-like features[6, 8]). 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, [24] has applied the Viola-Jones[21] style detector to facial features. False detections occur often, even for well-trained classifiers, because portions of the image have the appearance of a fiducial under some imaging condition. For example, a common error is for a "left corner of left eye" detector to respond to the left corner of the right eye. [8] achieves robustness and handles greater pose variation by using a large area of support for the detector covering, *e.g.*, an entire eye or the nose with room to spare. Searching over a smaller region that includes the actual part location reduces the chance of false detections with minimal impact of missing fiducials [6]. While this is somewhat effective for frontal fiducial point detection, the location of a part within the face detector box may vary significantly when the head rotates in 3-D. For example, while the left eye is in the upper-left side of the box when frontal, it can move to the right side when the face is seen in profile.

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 [6]. 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 [5, 6, 17]. [19] extend this to use a Gaussian Mixture Model whereas [9] 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. Like [9], 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 we take a non-parametric 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 [5, 6, 11, 17, 19, 20] 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 location of the mouth points are then constrained, possibly through a conditional probability, by the locations of stable points [20]. This approach fails when the stable points cannot be reliably detected, for example when the eyes are hidden by sunglasses. In contrast, our approach uses a RANSAC-like sampling to randomly select amongst 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 [9, 22], 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 non-frontal 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 Sec. 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) regressors with greyscale SIFT (Scale-Invariant Feature Transform) [16] descriptors as features. We compute the SIFT descriptor window at two scales: roughly $1/4$ and $1/2$ the inter-ocular distance. These two SIFT descriptors are then concatenated to form a single 256 dimensional feature vector for the SVM regressor.

For all of the training samples, we rescale the images so that the faces have an inter-ocular distance of roughly 55 pixels. Positive samples are taken at the manually annotated part locations. Negative samples are taken at least $1/4$ of the inter-ocular distance away from the annotated locations. In addition, random image plane rotations within $\pm 20^\circ$ are used to synthesize additional training samples.

These local 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). 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 $P(\mathbf{x}|\mathbf{d})$, *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, *etc*. 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 subsection, 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.

More formally, let $X = \{\mathbf{x}^1, \mathbf{x}^2, ...\mathbf{x}^n\}$ denote the locations of n parts, where $\mathbf{x}^i$ is the location of the i^{th} part. Let $D = \{\mathbf{d}^1, \mathbf{d}^2, ...\mathbf{d}^n\}$ denote the measured detector responses, where $\mathbf{d}^i$ is the window of scores returned by the i^{th} 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.*,

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

Let X_k (where $k = 1, ..., m$) denote the locations of the n parts in the k^{th} 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 $P(X|D)$ as follows:

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

where our collection of m exemplars X_k along with similarity transformations t have been introduced into the calculation of $P(X|D)$ 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 and rewrite the first term of Eq. 2 as

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

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

Since knowing the true location of the parts trumps any information provided by the detector, $P(\mathbf{x}^i_{k,t}|\mathbf{x}^i, \mathbf{d}^i) = P(\mathbf{x}^i_{k,t}|\mathbf{x}^i)$. 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 Eq. 4 as

$$P(X|X_{k,t},D) = \prod_{i=1}^{n} \frac{P(\Delta \mathbf{x}_{k,t}^i) P(\mathbf{x}^i|\mathbf{d}^i)}{P(\mathbf{x}_{k,t}^i|\mathbf{d}^i)}. \quad (5)$$

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

$$P(X_{k,t}|D) = \frac{P(D|X_{k,t})P(X_{k,t})}{P(D)} \quad (6)$$

$$= \frac{P(X_{k,t})}{P(D)} \prod_{i=1}^{n} P(\mathbf{d}^i|\mathbf{x}_{k,t}^i) \quad (7)$$

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 Eq. 7 as

$$P(X_{k,t}|D) = \left[\frac{P(X_{k,t})}{P(D)} \frac{\prod_{i=1}^{n} P(\mathbf{d}^i)}{\prod_{i=1}^{n} P(\mathbf{x}_{k,t}^i)} \right] \prod_{i=1}^{n} P(\mathbf{x}_{k,t}^i|\mathbf{d}^i) \quad (8)$$

$$= C \prod_{i=1}^{n} P(\mathbf{x}_{k,t}^i|\mathbf{d}^i) \quad (9)$$

Note that the terms within the square bracket in Eq. 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 Eqs. 1, 2, 5 and 9 yields

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

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

The first term $P(\Delta \mathbf{x}_{k,t}^i)$ is taken to be a $2D$ Gaussian distribution centered at the model location $\mathbf{x}_{k,t}^i$. 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 dataset, 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 inter-ocular distance. These normalized differences are used to compute the covariance matrices for each part location.

The second term $P(\mathbf{x}^i|\mathbf{d}^i)$ is computed as follows. We take the estimated location $\mathbf{x}^i$ for part i and look up the response for the i^{th} 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 Eq. 10 is challenging, as they are taken over all global models k and all similarity transformations t. However, we note from Eq. 2 that if $P(X_{k,t}|D)$ 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 to consider only those global models k with transformations t for which $P(X_{k,t}|D)$ is large.

In a sense, we wish to perform a Monte Carlo integration of Eq. 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 subsection, 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 $P(X_{k,t}|D)$ over the unknowns k and t. This optimization is non-linear, 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 multi-modal 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 Eq. 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 k.
2. Select two random parts. Randomly match each model part 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 Eq. 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 Eq. 9 in Step 4 is 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 subsection, we used a RANSAC-like procedure to find a list $\mathcal{M}$ of m^* global models $X_{k,t}$ for which $P(X_{k,t}|D)$ is largest. With these in hand, we approximate the optimization for X in Eq. 10 as

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

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

To find the best X^*, we first find an initial estimate $\mathbf{x}_0^i$ for each part i as

$$\mathbf{x}_0^i = \arg\max_{\mathbf{x}^i} \sum_{k,t \in \mathcal{M}} P(\Delta \mathbf{x}_{k,t}^i) P(\mathbf{x}^i | \mathbf{d}^i). \quad (12)$$

This is equivalent to solving for x_0^i by setting all $P(\Delta \mathbf{x}_{k,t}^j)$ and $P(\mathbf{x}^j | \mathbf{d}^j)$ to a constant in Eq. 10 for all $j \neq i$. To compute each $\mathbf{x}_0^i$ we merely need to multiply the detector output by a Gaussian function centered at $\mathbf{x}_{k,t}^i$, with the covariances calculated as described at the end of Subsection 3.2. Then we find the image location $\mathbf{x}_0^i$ where the sum of the resulting products is maximized. The initial estimates, $\mathbf{x}_0^i, i \in 1 \cdots n$ can then be used to initialize an optimization of Eq. 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.)

5. Experiments

Our work focuses on localizing parts in natural face images, taken under a wide range of poses, lighting conditions, and 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 our dataset, Labeled Face Parts in the Wild (LFPW). Our most significant results are on this new dataset.

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. Datasets

LFPW 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. Figure 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`.

Figure 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

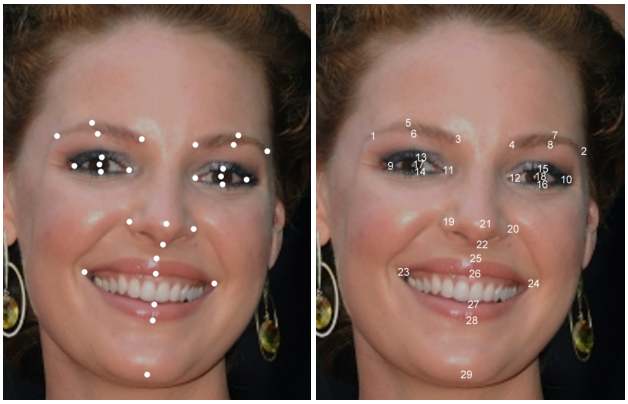

Figure 2. One of the images in LFPW. Overlaid, we show hand-labeled points obtained using MTurk. Points are numbered to match Figure 4.

this is seen in Figure 4, which shows the variation amongst the MTurk workers. Some parts like the eye corners are more consistently labeled whereas the brows and chin are labeled less accurately.

The publicly available BioID dataset contains 1,521 images, each showing a frontal view of a face of one of 23 different subjects [13]. We used 17 fiducial points that had been marked for the FGNet project, and used in the me_{17} error measure as defined in [5]. 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). Figure 7 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 3-D (*e.g.*, the outer corner of the left and right eyes). For a frontal image without occluded fiducials, the measure would be near

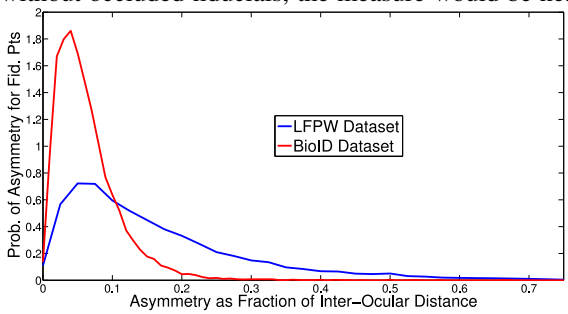

Figure 3. The distribution of the asymmetry measure over images in the LFPW and BioID datasets.

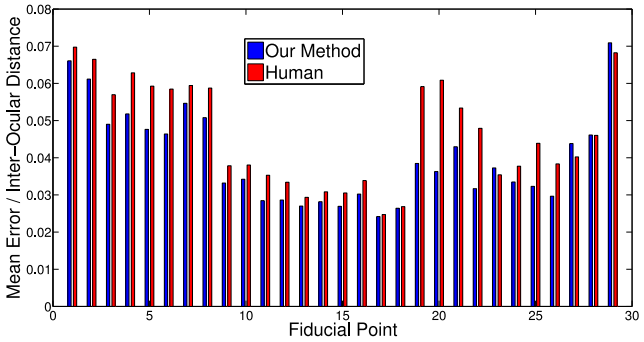

Figure 4. Mean error of our fiducial detector on the LFPW dataset compared to the mean variation in human labeling. The fiducial labels are shown in Fig. 2, and the error is the fraction of inter-ocular distance. Our detector is almost always more accurate.

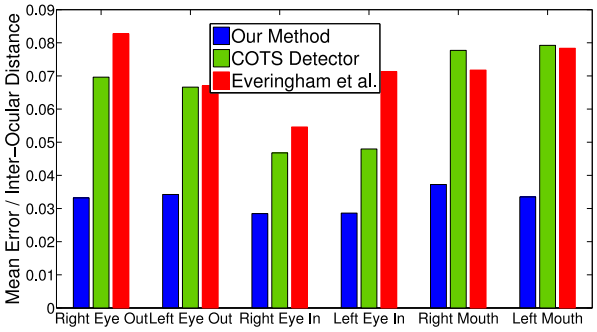

Figure 5. Comparison of our detector with a commercial off-the-shelf detector and the detector of [9]. Our detector is roughly twice as accurate as both.

zero. For faces that are rotated in 3-D or about the optical axis, the asymmetry increases with the extent of rotation. Figure 3 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 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 inter-ocular distance, to normalize for image size. Figure 4 shows the resulting error broken down by part. 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 eye brows (1-8, 29) are least accurate. This trend is consistent between human and automatic labeling.

Figures 1 and 6 show results on some representative images. We highlight a few characteristics of these results. These 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), *etc*. The localizer requires less than one second per fiducial on an Intel Core i7 3.06GHz machine; most of the time is spent evaluating the local detectors.

In Figure 5 we compare our LFPW results to those of a commercial face and fiducial detector[1] and the detector of [9]. 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, so we compare results on those fiducials only. At roughly 3% mean error rate, our results are roughly twice as accurate as the commercial system and [9].

Figure 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.

We have also applied our part localizer to the BioID faces and show some example output images in Figure 7. Results have been reported on this dataset by a number of authors. Figure 8 shows the cumulative error distribution of the me_{17} error measure (mean error of 17 fiducials) defined in [5]. Figure 8 compares the results of our method to those reported by [5, 17, 20, 22] . Our results are similar to but slightly better than those of [20], 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 Figure 9 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 Figure 3, the cumulative error distri-

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

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

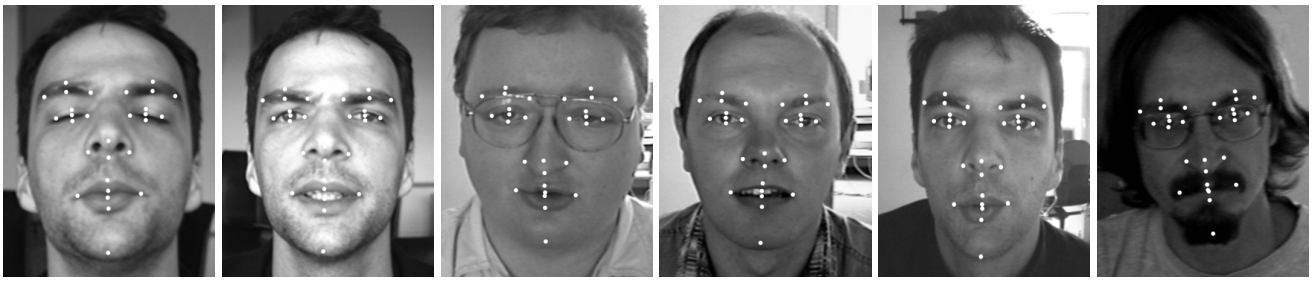

Figure 7. Images from BioID, along with parts localized by our detector.

bution 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.) Figure 9 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 non-parametric global models for part locations, computed

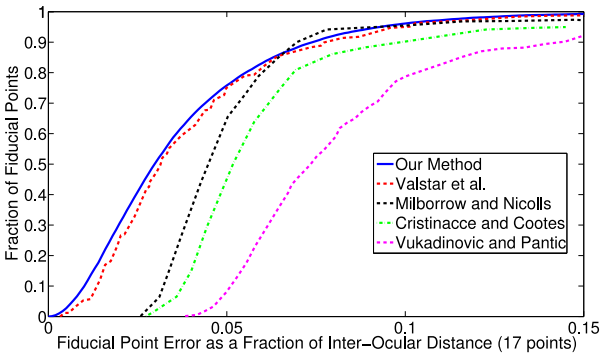

Figure 8. Cumulative error distribution curves comparing our system to several others on the BioID dataset. All comparative results are from [20]. We outperform all previously published results.

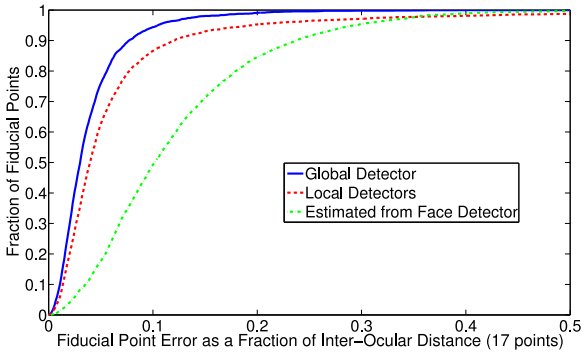

Figure 9. Cumulative error distribution of our system on the LFPW dataset compared to locations predicted using the face detector box or found with just our local detectors. (Note the different x-axis scale from Fig. 8.)

from exemplars. Our localizer is accurate over a large range of real-world variations in pose, expression, lighting, make-up, 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 both a previous research system and a commercial system. We also demonstrate state-of-the-art on the BioID dataset.

References

[1] *1st Intl. Workshop on Parts and Attributes*. 2010. 546

[2] L. Bourdev and J. Malik. Poselets: body part detectors trained using 3d human pose annotations. In *In IEEE Conference on Computer Vision and Pattern Recognition*, page 1365 1372, 2009. 546

[3] M. Burl, T. Leung, and P. Perona. Face localization via shape statistics. In *Workshop on Automatic Face and Gesture Recognition*, 1995. 546

[4] P. Campadelli, R. Lanzarotti, and G. Lipori. Automatic facial feature extraction for face recognition. In *Face Recognition*. I-Tech Education and Publishing, 2007. 546

[5] D. Cristinacce and T. Cootes. Feature detection and tracking with constrained local models. In *BMVC*, pages 929–938, 2006. 546, 549, 550

[6] D. Cristinacce, T. Cootes, and I. Scott. A multi-stage approach to facial feature detection. In *BMVC*, pages 231–240, 2004. 546

[7] L. Ding and A. M. Martinez. Precise detailed detection of faces and facial features. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2008. 546

[8] M. Eckhardt, I. Fasel, and J. Movellan. Towards practical facial feature detection. *Int. J. of Pattern Recognition and Artificial Intelligence*, 23(3):379–400, 2009. 546

[9] M. Everingham, J. Sivic, and A. Zisserman. "Hello! My name is... Buffy" – automatic naming of characters in TV video. In *BMVC*, 2006. 546, 547, 550

[10] N. Gourier, D. Hall, and J. L. Crowley. Facial features detection robust to pose, illumination and identity. In *Int. Conf. on Systems, Man and Cybernetics*, 2004. 546

[11] L. Gu and T. Kanade. A generative shape regularization model for robust face alignment. In *European Conference on Computer Vision (ECCV)*, pages 413–426, 2008. 546

[12] E. Holden and R. Owens. Automatic facial point detection. In *Asian Conf. Computer Vision*, pages 731–736, 2002. 546

[13] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz. Robust face detection using the Hausdorff distance. In *Conf. on Audio- and Video-Based Biometric Person Authentication*, pages 90–95. Springer, 2001. 549

[14] N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar. Attribute and simile classifiers for face verification. In *IEEE International Conference on Computer Vision*, 2009. 546

[15] B. Leibe, A. Ettlin, and B. Schiele. Learning semantic object parts for object categorization. *Image and Vision Computing*, 26:15–26, 1998. 546

[16] D. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision*, 2003. 547

[17] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *European Conf. on Computer Vision*, pages 504–513, 2008. 546, 550

[18] M. Reinders, R. W. C. Koch, and J. Gerbrands. Locating facial features in image sequences using neural networks. In *Conf. on Automatic Face and Gesture Recognition*, pages 230–235, 1997. 546

[19] J. M. Saragih, S. Lucey, and J. Cohn. Face alignment through subspace constrained mean-shifts. In *International Conference of Computer Vision (ICCV)*, September 2009. 546

[20] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *IEEE Computer Vision and Pattern Recognition (CVPR)*, pages 2729–2736, 2010. 546, 547, 550, 552

[21] P. Viola and M. Jones. Robust real-time face detection. *Intl. Journal of Computer Vision*, 57:137–154, 2004. 545, 546

[22] D. Vukadinovic and M. Pantic. Fully automatic facial feature point detection using Gabor feature based boosted classifiers. In *Int. Conf. on Systems, Man and Cybernetics*, pages 1692–1698, 2005. 546, 547, 550

[23] M.-H. Yang, D. J. Kriegman, and N. Ahuja. Detecting faces in images: A survey. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 24(1):34–58, 2002. 545

[24] C. Zhan, W. Li, P. Ogunbona, and F. Safaei. Real-time facial feature point extraction. In *Advances in multimedia information processing*, pages 88–97. Springer-Verlag, 2007. 546