Occlusion Coherence: Localizing Occluded Faces with a Hierarchical Deformable Part Model

Golnaz Ghiasi Charless C. Fowlkes
Dept. of Computer Science, University of California, Irvine
{gghiasi,fowlkes}@ics.uci.edu

Abstract

The presence of occluders significantly impacts performance of systems for object recognition. However, occlusion is typically treated as an unstructured source of noise and explicit models for occluders have lagged behind those for object appearance and shape. In this paper we describe a hierarchical deformable part model for face detection and keypoint localization that explicitly models occlusions of parts. The proposed model structure makes it possible to augment positive training data with large numbers of synthetically occluded instances. This allows us to easily incorporate the statistics of occlusion patterns in a discriminatively trained model. We test the model on several benchmarks for keypoint localization including challenging sets featuring significant occlusion. We find that the addition of an explicit model of occlusion yields a system that outperforms existing approaches in keypoint localization accuracy.

1. Introduction

Accurate localization of detailed facial features provides an important building block for many applications including identification [3] and analysis of facial expressions [17]. Significant progress has been made in this task, aided in part by the fact that faces have less intra-category shape variation and limited articulation compared to other object categories of interest. However, feature point localization tends to break down when applied to faces in real scenes where other objects in the scene (hair, sunglasses, other people) are likely to occlude parts of the face. Fig. 1(a) depicts the output of a deformable template model [28] where the presence of occluders distorts the final alignment of the model. In this paper we propose a model that explicitly models occluded features in order to produce superior localization results (Fig. 1(b)).

A standard approach to handling occlusion in part-based models is to compete part feature scores against a generic background model or fixed threshold. However, setting such thresholds is fraught with difficulty since it is difficult to distinguish between parts that are present but simply hard to detect (e.g., due to unusual lighting) and those which are genuinely hidden behind another object.

(a) (b)

Figure 1. (a) A standard deformable part model [28] is distorted by the presence of occluders, disrupting localization even for parts that are far from the site of occlusion. (b) The output of our hierarchical part model explicitly models likely patterns of occlusion and improves localization as well as predicting which keypoints are occluded.

We believe that treating occlusions as an unstructured source of noise ignores a key aspect of the problem, namely that occlusions are induced by other objects and surfaces in the scene and hence should exhibit **occlusion coherence**. For example, it would seem very unlikely that every-other keypoint along an object's contours should be occluded. Yet many occlusion models make strong independence assumptions about occlusion that make it difficult to distinguish *a priori* likely from unlikely patterns. Furthermore, an occluder should not be inferred simply by the lack of evidence for object features, but rather by positive evidence for the occluding object that **explains away** the lack of object features.

The contribution of this paper is an efficient hierarchical deformable model that encodes these principles for modeling occlusion and achieves state-of-the-art performance on benchmarks for occluded face localization. The model describes the face by an arrangement of parts, each of which is in turn composed of local keypoint features. This two-layer model provides a compact, discriminative representation for the appearance and deformations of parts and the correlation between shapes of neighboring parts. In addition to representing the face shape, each part has an associated occlusion state chosen from a small set of possible occlusion patterns, enforcing coherence across neighboring keypoints and providing a sparse representation of the occluder shape where it intersects the part.

Specifying training data from which to learn feasible occlusion poses difficulties of its own. Practically speaking, existing datasets have focused primarily on fully visible faces. Moreover, it seems unlikely that any reasonable sized set of training images would serve to densely probe the space of possible occlusions. Beyond certain weak contextual constraints, the location and identity of the occluder itself are arbitrary and largely independent of the occluded object. To overcome this difficulty of training data, we propose a unique approach for generating synthetically occluded positive training examples. By exploiting the structural assumptions built into our model, we are able to include such examples as "virtual training data" without explicitly synthesizing new images.

2. Related Work

Landmark Localization: There is a huge literature on model alignment for facial landmark estimation. Classic approaches to 2D alignment include Deformable Templates [26], Active Appearance Models (AAMs) [6, 18, 19] and elastic graph matching [24]. Alignment with full 3D models provides even richer information [14, 3] at the cost of additional computation. One key difficulty in most of these approaches is the need to resort to iterative and local search techniques for optimizing model alignment. This typically results in high computational cost and the constant specter of local minima undermining system performance.

A more recent family of approaches makes use of constrained local models that detect candidate local features and then enforce constraints between parts [2]. Training regressors that learn to predict keypoint locations from appearance and other detector responses has also shown good performance [22, 9, 4, 5, 8]. A key advantage is that such regression models can be trained layer-wise in a discriminative fashion and thus sidestep the optimization problems of global model alignment as well as providing fast, feed-forward performance at test time.

Our model is most closely related to the recent work of [28] which applies discriminatively trained deformable part models (DPM) [10] to face analysis. This offers an intermediate between the extremes of model alignment and keypoint regression, by utilizing mixtures of simplified shape models that make efficient global optimization of part placements feasible while exploiting discriminative training criteria. Similar to [25], we use local part and keypoint mixtures to encode richer multi-modal distributions. The key difference in our model is the addition of hierarchical structure and occlusion to the model. We introduce intermediate part nodes that do not have an associated "root template" but instead serve to encode an intermediate representation of occlusion and shape state. The notion of hierarchical part models has been explored (e.g., [27, 12]) as a tool for compositional representations and parameter sharing. Intermediate state represented in such models can often be formally encoded in a non-hierarchical model with expanded state space. However, in our experiments the particular choice of model structure proves essential to efficient representation and inference.

Occlusion Modeling: Modeling part-level occlusion is a natural fit for recognition systems with an explicit representation of parts. Work on generative constellation models [23, 11] learned parameters of a full joint distribution over the probability of part occlusion and relied on brute force enumeration for inference, a strategy that doesn't scale to large numbers of keypoints. More commonly, part occlusions are treated independently which makes computation and representation more efficient. For example, the supervised detection model of [1] adds independent binary variables indicating occlusion of a part and learns a corresponding extra template. [12] imposes more structured distribution on the possible occlusion patterns by specifying grammar that generates a person detector as a variable length vertical chain of parts terminated by an occluder. Our approach provides a stronger model than full independence and has an advantage over the grammar approach that the occlusion patterns are not specified structurally but instead learned from data and encoded in the model weights.

Regression-based approaches have also incorporated occlusion. For example, the face model of [21] uses a robust m-estimator which serves to truncate part responses that fall below a certain threshold. We compare our results to the recent work of [4] which uses occlusion annotations when training a cascade of regressors where each layer predicts both part locations and occlusion states.

3. Hierarchical Part Model

Fig. 2 shows the model structure. The model has two layers with the face consisting of a collection of parts (nose, eyes, lips) each of which is in turn composed of a number of keypoints capturing the local edge features making up that part. All keypoints are connected to the part node with a star topology while the parts form a tree. In addition

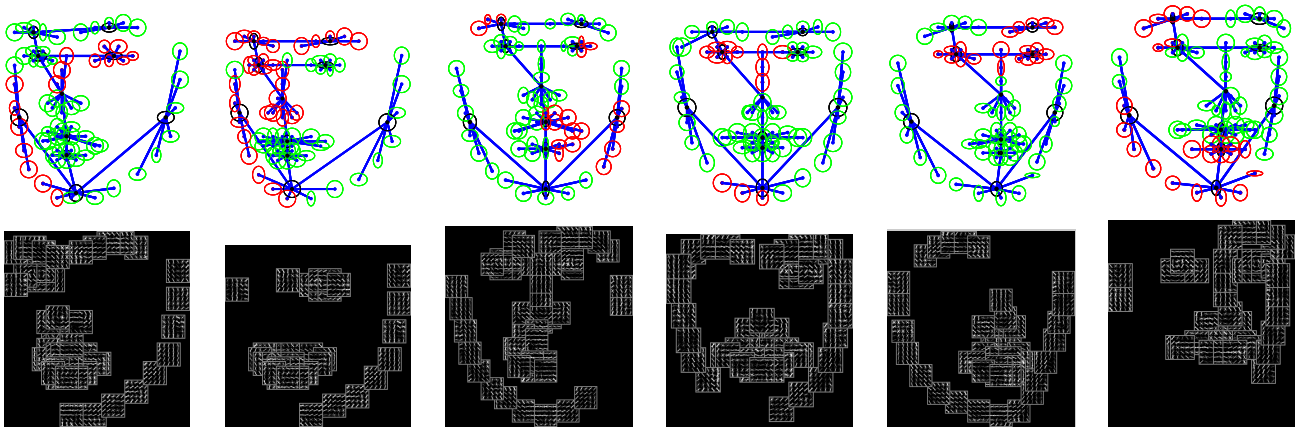


Figure 2. Our model consists of a tree of parts (black circles) each of which is connected to a set of keypoints (green or red) in a star topology. The examples here show templates corresponding to different choices of part shape and occlusion patterns. Red indicate occluded keypoints. Shape parameters are independent of occlusion state. Keypoint appearance is modeled with a small HOG template (2nd row) and occluded keypoints are constrained to have an appearance template fixed to 0. Note how the model produces a wide range of plausible shape configurations and occlusion patterns.

to location, each part takes one of a discrete set of shape states (corresponding to, e.g., different facial expressions) and occlusion states (corresponding to different patterns of occlusion). The grouping of facial features into parts was specified by hand while the shape and occlusion patterns are learned automatically from training data. This model, which we term a hierarchical part model (HPM) is similar to the DPM [10] and the flexible part model of [28]. It differs in the addition of part nodes that don't include any "root filter" and the use of mixtures to model occlusion patterns. In this section we introduce some formal notation to describe the model and some important algorithmic details for making the message passing used during inference efficient. In the following section we describe the details of training the model.

Model Structure: Let l, s, o denote the hypothesized locations, shape and occlusion of the face parts and keypoints. We define a tree structured scoring function by:

$$S(l,s,o) = \sum_i \phi_i(l_i, s_i, o_i) \qquad (1)$$
$$+ \sum_i \sum_{j \in child(i)} \psi_{ij}(s_i, s_j, l_i, l_j) + b_{ij}(s_i, s_j, o_i, o_j)$$

where the potential ϕ encodes local appearance at location l_i, and ψ is a quadratic shape/deformation penalty, and b is a co-occurrence bias.

The first (unary) term scores the appearance evidence. We linearly parameterize the unary appearance term with weights w_i where

$$\phi_i(l_i, s_i, o_i) = w_i^{s_i} \cdot \phi(l_i, o_i)$$

Appearance templates are only associated with the leaves (keypoints) in the model so the unary term only sums over those leaf nodes. The occlusion variables o_i for the keypoints are binary, corresponding to either occluded or visible. If the ith keypoint is unoccluded, the appearance feature ϕ is given by a HOG [7] feature extracted at location l_i, otherwise the feature is set to 0. This is natural on theoretical grounds since the appearance of the occluder is arbitrary and hence indistinguishable from background. Empirically we find that unconstrained occluder templates learned with sufficiently varied data do in fact have very small norms, further justifying this choice.

The second (pairwise) term in Eq. 1 scores the placement part j based on its location relative to its parent i. We parameterize this linearly

$$\psi_{ij}(s_i, s_j, l_i, l_j) = w_{ij}^{s_i, s_j} \cdot \psi(l_i - l_j)$$

where the feature ψ includes the x and y displacements and their cross-terms, allowing the weights w_{ij} to encode a standard quadratic "spring". We assume that the shape of the parts is independent of any occluder so the spring weights do not depend on the occlusion states. The pairwise parameter b_{ij} encodes a bias of particular occlusion patterns and shapes to co-occur. Each keypoint has two occlusion states and as many shape mixtures as its parent part, but the bias parameters learned between the part and its constituent keypoints are constrained to enforce a hard, 1-1 mapping between the mixture states of keypoints and parts.

Message Passing: The model above can be made formally equivalent to the FMP model used in [25] by introducing local mixture variables that live in the cross-product space of o_i and s_i. However, this reduction fails to exploit the structure of occlusion model. This is particularly important due to the large size of the model. Naive inference is quite slow due to the large number of keypoints and parts (N=78), and huge state space for each node which includes location, occlusion pattern and shape mixtures. Consider

the message passed from one part to another where each part has L possible locations, S shape mixtures and O occlusion patterns. In general this requires minimizing over functions of size $(LSO)^2$ or $L(SO)^2$ when using the distance transform. In the models we test, $SO = 27$ which poses a substantial computation and memory cost, particularly for high-resolution images where L is large.

While the factorization of shape and occlusion doesn't change the asymptotic complexity, we can reduce the runtime in practice by exploiting distributivity of the distance transform over $\max$ to share computations.

Standard message passing requires that we compute:

$$\mu_{j\to i}(l_i, s_i, o_i) = \max_{s_j, l_j, o_j} \Big[\psi_{ij}(l_i, l_j, s_i, s_j) + \sum_{k \in child(j)} \mu_{k\to j}(l_j, s_j, o_j) + b_{ij}(s_i, s_j, o_i, o_j) \Big]$$

We can move the maximization over occlusion patterns of part j inward, carrying out the computation in two steps:

$$\nu_{ij}(l_j, s_i, s_j, o_i) = \max_{o_j} \left[\sum_{k \in child(j)} \mu_{k\to j}(l_j, s_j, o_j) + b_{ij}(s_i, s_j, o_i, o_j) \right]$$

$$\mu_{j\to i}(l_i, s_i, o_i) = \max_{s_j, l_j} [\psi_{ij}(s_i, s_j, l_i, l_j) + \nu_{ij}(l_j, s_i, s_j, o_i)]$$

The second equation requires computing a distance transform for each value of s_i, s_j and o_i but is independent of o_j.

We also note that the keypoint occlusion and shape variables are determined completely by the parent part state which further simplifies the messages from keypoints to parts.

$$\mu_{k\to j}(l_j, s_j, o_j) = \begin{cases} 0 & \text{if k occluded in } o_j \\ \max_{l_k} \psi_{jk}(l_j, l_k, s_j) + \phi_k(l_k, s_j) \end{cases}$$

Since the score is known for an occluded keypoint in advance, it is not necessary to compute distance transforms for those components. In our models, this reduces the memory and inference time by roughly a factor of 2. This savings becomes increasingly significant as the number of occlusion mixtures grows.

4. Model Training and Inference

The potentials in our shape model are linearly parameterized, allowing efficient training using a standard SVM objective [10]. Face viewpoint, keypoint locations, shape and occlusion mixtures are completely specified by preclustering the training data so parameter learning is fully supervised. In this section we describe how these supervised labels are derived and how we synthesize "virtual" positive training examples that include additional occlusion.

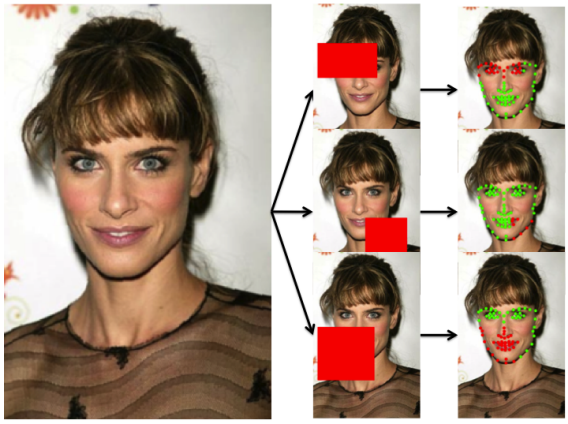


Figure 3. Virtual positive examples are generated synthetically by sampling random coherent occlusions from a given fully visible training example.

Viewpoint Mixtures: Viewpoint and scale are the largest sources of variability affecting the keypoint configurations. To cluster viewpoints in training data, we made use of the MultiPIE dataset which provides ground-truth viewpoint annotations. We perform Procrustes alignment between each training example and examples in the MultiPIE database and then transfer the viewpoint label from MultiPIE to the training example. This alignment also provides a standard scale normalization and removes in-plane rotations from the training set. In our experiments we used only three viewpoint clusters: center (+/-7.5 degrees), left, and right-facing (7.5-22.5 degrees).

Part Shape and Occlusion Mixtures: For each part and each viewpoint, we cluster the set of keypoint configurations in the training data in order to come up with a small number of shape mixtures for that part. The part shapes in the final model are represented by displacements relative to a parent node so we subtract off the centroid of the part keypoints from each training example prior to clustering. The vectors containing the coordinates of the centered keypoints are clustered using k-means. We imagine it would be efficient to allocate more mixtures to parts and viewpoints that show greater variation in shape, but in the final model tested here we use fixed allocation of $k = 3$ shape mixtures per part per viewpoint. Fig. 4 shows example clusterings of part shapes for the center view.

Synthetic Occlusion Patterns: In the model each keypoint is fully occluded or fully visible. The occlusion state of a part describes the occlusion of all its constituent keypoints. If there are N keypoints then there are 2^N possible occlusion patterns. However, many of these occlusions are quite unlikely (e.g. every other keypoint occluded) since occlusion is typically generated by an occluder object with a regular, compact shape.

To model spatial coherence among the keypoint occlusions, we synthetically generate "valid" occlusions patterns by first sampling mean part and keypoint locations from the model and then randomly sampling a quarter-plane shaped occluder and setting as occluded those keypoints that fall behind the occluder. Let a, b be uniformly sampled from a tight box surrounding the face. A keypoint i with location $l_i = (x, y)$ is occluded if $(\pm x \lesseqgtr a) \cap (\pm y \lesseqgtr b)$ where the quadrant is chosen at random. While our occluder is somewhat "boring", it is straightforward to incorporate more interesting shapes, e.g., by sampling from a database of segmented objects. Fig. 3 shows example occlusions generated for a training example.

In our experiments we generate 4 synthetically occluded examples for each original training example. For each part in the model we cluster the set of resulting binary vectors in order to generate a list of valid part occlusion patterns. The occlusion state for each keypoint in a training example is then set to be consistent with the assigned part occlusion pattern. In our experiments we utilized only $k = 3$ occlusion mixtures per part, typically corresponding to unoccluded and two half occluded states whose structure depended on the part shape and location within the face.

In-plane Rotation: In our experiments, we observed that part models with standard quadratic spring costs are surprisingly sensitive to in-plane rotation. Models that performed well on images with controlled acquisition (such as MultiPIE) fared poorly "in the wild" when faces were tilted. The alignment procedure above removes scale and in-plane rotations from the set of training examples. At test time detection, we perform an explicit search over scale and in plane rotations (+/-30 degrees).

Landmark Prediction: To benchmark keypoint localization in datasets which used different landmark points, we used linear regression to learn a mapping from the set of locations returned by our hierarchical part model. In our experiments, this prediction was important to accurately benchmarking localization performance. Using a heuristic approach of simply taking the closest keypoint reported (or the mean of the eye keypoints in the case of the LFPW29 eye center keypoint) performed significantly worse, in some cases doubling failure rates.

Let $l^i \in \mathbb{R}^{2N}$ be the vector of keypoint locations returned when running the model on a training example i and $\hat{l} \in \mathbb{R}^{2M}$ a vector of ground-truth keypoint location for that image. We train a linear regressor

$$\min_{\beta} \sum_{i} \|\hat{l^i} - \beta^T l^i\|^2 + \lambda \|\beta\|^2$$

where $\beta \in \mathbb{R}^{2N \times 2M}$ is the matrix of learned coefficients and λ is a regularization parameter. To prevent overfitting, we restrict β_{pq} to be zero unless the keypoint p belongs to the same part as q.

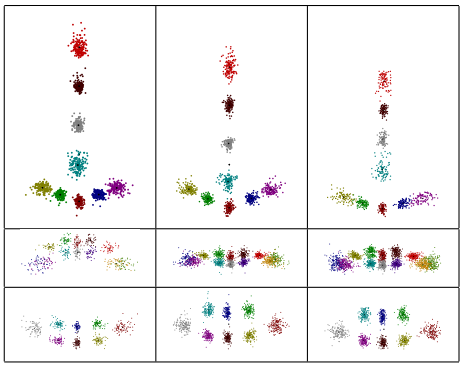


Figure 4. Example shape clusters for face parts (nose, upper lip, lower lip). Co-occurrence biases for combinations of part shapes are learned automatically from training data. Different colored points correspond to location of each keypoint relative to the part (centroid).

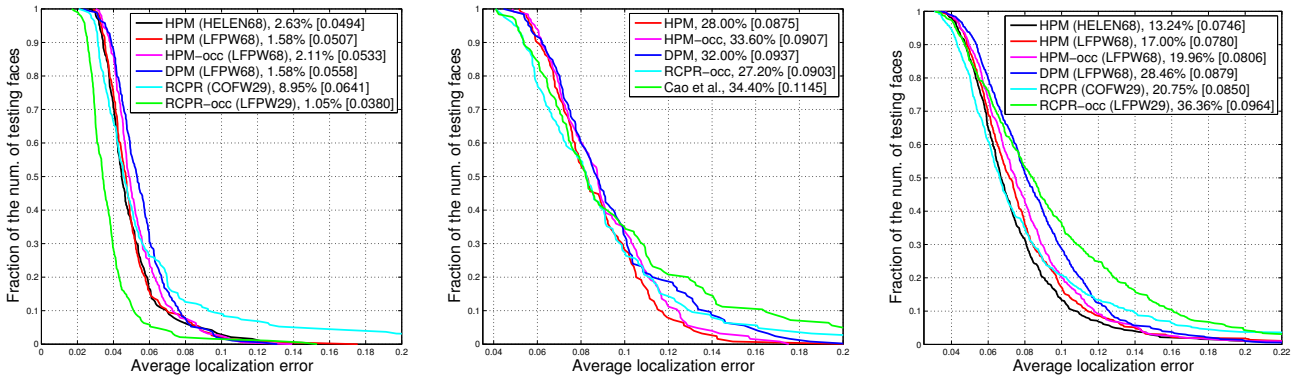
To predict keypoint occlusion, we carried out a similar mapping procedure using regularized logistic regression. However, we found that in practice a much simpler rule of specifying a correspondence between the two sets of keypoints based on their distance and transferring the occlusion flag from the model to benchmark keypoints achieved the same accuracy.

5. Experimental Evaluation

We evaluate performance of our method and related baselines on three benchmark datasets for localization: Labeled Face Parts in the Wild (LFPW) [15], a subset of the HELEN dataset [16] which contained occlusions, and the more difficult Caltech Occluded Faces in the Wild (COFW) [4] dataset. The latter two datasets were selected to highlight the performance of our model in the presence of occlusion and a wider variety of poses. The authors of [4] estimate that LFPW only contains 2% occluded keypoints compared to 23% for COFW.

Evaluation: There is a variety of keypoint annotation conventions across these different datasets. LFPW and COFW contain a set of 29 landmarks while HELEN includes a much denser set of 194 landmarks. The 300 Faces in-the-wild Challenge (300-W) [20] has also produced several unified benchmarks in which the LFPW and HELEN datasets have been re-annotated with a set of 68 standard keypoints. For the purposes of benchmarking, and to allow easy comparison to previously reported work, we utilize the 29 keypoints for the LFPW and COFW datasets and the 68 keypoints for HELEN.

To evaluate keypoint localization independent of detection accuracy, we assume that detection has already been performed and run the algorithm on cropped versions of the test images. We evaluate localization for the highest scoring detection that overlaps the ground-truth face bounding box



(a) LFPW (b) Occluded HELEN (c) COFW

Figure 5. Cumulative error distribution curves for landmark localization show the proportion of test images that have average localization error below a given threshold on LFPW, an occlusion rich subset of HELEN and on COFW. We compare the hierarchical part model (HPM) with and without occlusion mixtures to a baseline tree-structured DPM [28] and robust pose regression (RCPR) [4] trained on different training sets (in parentheses). Models tested on Occluded HELEN were all trained on LFPW68. The legend reports the failure rate % at a localization threshold of 0.1 and the average error (in brackets). The HPM shows good localization, especially in the presence of occlusion and robust cross-dataset generalization.

by at least 80%.

We report the average keypoint localization error across the entire test set as well as the proportion of "failures", test images that had average keypoint localization above a given threshold. Distances used in both quantities are expressed as a proportion of the inter-ocular (iod) distance specified by the ground-truth. Computing the failure rate across a range thresholds yields a cumulative error distribution curve (Fig. 5). When a single summary number is required we report the failure rate at a standard threshold of 0.1 iod.

Training: To train our model, we used a set of 682 near-frontal training images taken from LFPW using the 68 keypoint annotations provided by 300-W. From each training image we generate 4 synthetically occluded "virtual positives" yielding a final training set of 3410 positives. As mentioned previously, since we explicitly search over in-plane rotations and scales, standardize the pose of each training image prior to learning the model. To evaluate cross-dataset generalization, we also trained a version of our model on a portion the HELEN dataset consisting of 1758 frontal images annotated with 68 keypoints. To fit linear regression coefficients for mapping from the HPM and DPM predicted keypoint locations we ran the model on the COFW training data set which has 29 keypoint annotations.

For comparison, we trained baseline models including a version of our model without occlusion mixtures (HPM-occ) and the non-hierarchical part model (DPM) described by [28]. [1] We also evaluate the robust pose regression (RCPR) described in [4] and their implementation of explicit shape regression [5] using both pre-trained models provided by the authors and models retrained to predict 68 keypoints.

[1]The original DPM model of [28] was trained on the very constrained MultiPIE dataset [13]. Retraining the model and performing a similar search over in-plane rotations yielded significantly better performance which we report here. (c.f. [4])

5.1. Localization

Fig. 6 depicts selected results of running our detector on images from the HELEN and COFW test datasets. While the possible occlusion patterns are quite limited (3 occlusions per part shape), the final predicted occlusions (marked in red) are quite satisfying in highlighting the support of the occluder for many instances.

LFPW: Labeled Face Parts in the Wild (LFPW) [15], a commonly used dataset for evaluating landmark estimation consisting of 300 test images annotated with a standard set of 29 keypoints. The original LFPW test set is no longer completely available due to broken links, but we were able to download 194 of the test images. Fig. 5(a) shows the localization error distribution.

The HPM model achieves an average localization error of 0.0507 with a failure rate of 1.58%. By comparison, the DPM [28] has a higher average error 0.0558. Since there is very little occlusion in this dataset, we attribute the improved performance to the ability of the HPM to better capture the shape of facial features. This is verified by the similar performance of HPM-occ. While the tree structure used in [28] is optimal in the sense of Chow-Liu, the addition of extra nodes and mixture components representing part-level deformations yields a clear benefit in modeling shape.

The robust pose regression model of [4] performs exceedingly well on LFPW, particularly at high localization accuracy. However, we note that the training set used is slightly different (the RCPR model provided by the authors was trained directly with 29 keypoint annotations and boosts the training data to 2000 examples by introducing perturbed versions). Interestingly, RCPR trained on the

COFW dataset performs much worse, suggesting some degree of overfitting.

Occluded HELEN: We evaluated on a subset of the HELEN dataset [16] consisting of 126 images which were selected on the basis having some significant amount of occlusion. HELEN generally includes more difficult images than LFPW and our selected subset was harder still. These test images contain 68 keypoint annotations so we evaluated only models trained on LFPW68. We did not test HPM (HELEN68) on this dataset as there was overlap between training and testing images. Fig. 5(b) shows the error distribution. The HPM achieves an average error of 0.0875, beating out the DPM baseline and pose regression. For very small localization error thresholds (< 0.08) RCPR achieves a lower failure rate.

COFW: Finally, we tested on the 507 image test set from Caltech Occluded Faces in the Wild (COFW) [4] which contains internet photos depicting a wide variety of more difficult poses and includes a significant amount of occlusion. Since COFW training only contains 29 keypoints, we could not train the HPM model and instead evaluate models trained on LFPW68 and HELEN68. Fig. 5(c) shows results on COFW where HPM achieves a significantly lower average error than RCPR.

Occlusion Prediction: Since the COFW contains occlusion annotations, we can also evaluate prediction accuracy of occlusion on a per-keypoint basis. Our model outputs a hard decision on the occlusion state, so we can't easily generate a precision-recall curve. However, at the trained operating point, our model yields a precision of 80.8 and a recall of 37.0. This appears to be within 1% of the best precision-recall curves reported for RCPR in [4].

5.2. Detection

Pose regression requires good initialization provided by a face detector to accurately locate keypoints. In contrast, part-based models have the elegant advantage of performing detection and localization simultaneously. We evaluated the detection accuracy of the HPM model on the AFW dataset introduced in [28]. Since our model trained on LFPW68 only contains near-frontal views, we evaluated on near-frontal test images (+/-22.5 degrees). The model achieved an average precision of AP=0.997, indistinguishable from the DPM model performance with the same training and test split. HPM-occ performed slightly worse at AP=0.986.

Since there is relatively little occlusion in AFW, we also assembled a preliminary dataset for occluded face detection consisting of 61 images from Flickr containing 766 labeled faces. Of the faces in these images, 430 include some amount of occlusion. On these occluded faces we see a substantial boost in detection performance. HPM achieved AP=0.682, while HPM-occ and DPM had average precisions of 0.654 and 0.641 respectively. Fig. 6 shows example detection results.

6. Discussion

Our experimental results demonstrate that adding coherent occlusion and hierarchical structure allows for substantial gains in performance for keypoint localization and detection in part models. Our final HPM outperforms previously published results on the challenging COFW dataset in terms of keypoint localization accuracy and shows robust generalization across different training and test sets.

In comparing pose regression and part-based models, there seem to be several interesting tradeoffs. In our experiments, we see a general trend in which error distribution curves for pose regression and part models cross, suggesting that RCPR yields very accurate localization for a subset of images relative to the HPM but fails for some other proportion even at very large error thresholds. The run-time of our model implementation built on dynamic programming lags significantly behind those of regression-based, feed-forward approaches. Our model takes $\sim$30s to run on a typical COFW image, roughly 100x slower than RCPR. On the other hand, pose regression depends critically on having good initialization while the part model approach can be used for both simultaneous detection and localization.

Finally, we note that there are many avenues for future work. Performance depends on the graphical independence structure of the model which should ideally be learned from data. While our model implicitly represents the pattern of part occlusions, it does not integrate local image evidence for the occluder itself. A natural extension would be to add local filters which detect the presence of an occluding contour between the occluded and non-occluded keypoints. Such filters could be shared across parts to avoid increasing too much the overall computation cost while moving closer to our goal of explaining away missing object parts using positive evidence of coherent occlusion.

Acknowledgements: This work was supported by NSF IIS-1253538

References

[1] H. Azizpour and I. Laptev. Object detection using strongly-supervised deformable part models. In *ECCV*, pages 836–849. 2012. 2

[2] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*, pages 545–552, 2011. 2

[3] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *PAMI*, 25(9):1063–1074, 2003. 1, 2

[4] X. P. Burgos-Artizzu, P. Perona, and P. Dollár. Robust face landmark estimation under occlusion. ICCV, 2013. 2, 5, 6, 7

[5] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, pages 2887–2894, 2012. 2, 6

[6] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *PAMI*, 23(6):681–685, 2001. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893, 2005. 3

Figure 6. Right: Examples of landmark localization and occlusion estimation for images from the HELEN (row 1) and COFW (rows 2-3) test datasets. Left: Examples of detection and localization in scenes with substantial occlusion. Red indicates those keypoints which are predicted as being occluded by the HPM.

[8] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool. Real-time facial feature detection using conditional regression forests. In *CVPR*, pages 2578–2585, 2012. 2

[9] B. Efraty, C. Huang, S. K. Shah, and I. A. Kakadiaris. Facial landmark detection in uncontrolled conditions. In *IJCB*, pages 1–8, 2011. 2

[10] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010. 2, 3, 4

[11] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–264. IEEE, 2003. 2

[12] R. B. Girshick, P. F. Felzenszwalb, and D. A. Mcallester. Object detection with grammar models. In *NIPS*, pages 442–450, 2011. 2

[13] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-PIE. *IVC*, 28(5):807–813, 2010. 6

[14] L. Gu and T. Kanade. 3d alignment of face in a single image. In *CVPR*, pages 1305–1312, 2006. 2

[15] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCV Workshop*, pages 2144–2151, 2011. 5, 6

[16] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, pages 679–692. 2012. 5, 7

[17] A. Martinez and S. Du. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *JMLR*, 98888:1589–1608, 2012. 1

[18] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 60(2):135–164, 2004. 2

[19] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *ECCV*, pages 504–513. 2008. 2

[20] M. Pantic, G. Tzimiropoulos, and S. Zafeiriou. 300 faces in-the-wild challenge (300-w). In *ICCV Workshop*, 2013. 5

[21] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 91(2):200–215, 2011. 2

[22] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, pages 2729–2736, 2010. 2

[23] M. Weber, M. Welling, and P. Perona. Towards automatic discovery of object categories. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 2, pages 101–108. IEEE, 2000. 2

[24] L. Wiskott, J.-M. Fellous, N. Kuiger, and C. Von Der Malsburg. Face recognition by elastic bunch graph matching. *PAMI*, 19(7):775–779, 1997. 2

[25] Y. Yang and D. Ramanan. Articulated human detection with flexible mixtures-of-parts. *IEEE TPAMI*, 2013. 2, 3

[26] A. L. Yuille, P. W. Hallinan, and D. S. Cohen. Feature extraction from faces using deformable templates. *IJCV*, 8(2):99–111, 1992. 2

[27] L. Zhu, Y. Chen, and A. Yuille. Learning a hierarchical deformable template for rapid deformable object parsing. *PAMI*, 32(6):1029–1043, 2010. 2

[28] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, pages 2879–2886, 2012. 1, 2, 3, 6, 7