Robust face landmark estimation under occlusion

Xavier P. Burgos-Artizzu* Pietro Perona* Piotr Dollár†
*California Institute of Technology †Microsoft Research, Redmond
xpburgos,perona@caltech.edu pdollar@microsoft.com

Abstract

Human faces captured in real-world conditions present large variations in shape and occlusions due to differences in pose, expression, use of accessories such as sunglasses and hats and interactions with objects (e.g. *food). Current face landmark estimation approaches struggle under such conditions since they fail to provide a principled way of handling outliers. We propose a novel method, called Robust Cascaded Pose Regression (RCPR) which reduces exposure to outliers by detecting occlusions explicitly and using robust shape-indexed features. We show that RCPR improves on previous landmark estimation methods on three popular face datasets (LFPW, LFW and HELEN). We further explore RCPR's performance by introducing a novel face dataset focused on occlusion, composed of 1,007 faces presenting a wide range of occlusion patterns. RCPR reduces failure cases by half on all four datasets, at the same time as it detects face occlusions with a* $80/40\%$ *precision/recall.*

1. Introduction

Accurate object shape computation is key to many visual tasks; for instance, classification of facial expression [17, 30], facial identity [38], action analysis [4], and fine-grained categorization [34, 3, 18, 40] require accurate registration of a (deformable, part-based) model to the image. The shape of human bodies and human faces has attracted particular attention [43, 46, 33, 2]. By *shape* here we mean the parameters of a model that describe the configuration of an object in the image or, alternatively, the location of a number of parts or landmarks in the image or in 3D space. The complexity and parametrization of shape depends on the object type and on the task at hand.

Cascaded Pose Regression (CPR) [14] is capable of estimating shape using any parametrized variation of the object's appearance. Recently, it has emerged as a particularly effective and accurate approach for estimating face landmarks [7]. However, face landmark estimation "in the wild" remains a very challenging task. We find that CPR struggles under occlusions and large shape variations.

We propose a novel method inspired by CPR, called Robust Cascaded Pose Regression (RCPR). RCPR improves performance by increasing robustness to occlusions and large shape variations, which occur often in real-world conditions. RCPR is able to detect face occlusion at the same time as it estimates the landmark positions.

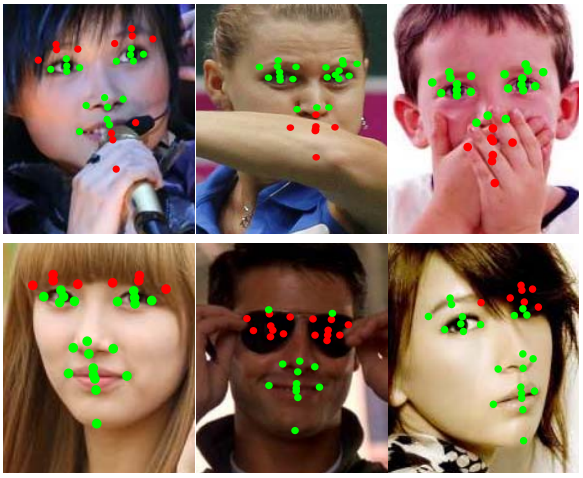


Figure 1. Example results. RCPR estimates landmark positions as well as their occlusion state (red=occluded, green=unoccluded).

The occlusion information helps during learning to select unoccluded features and is exploited dynamically through robust statistics to reduce errors inside the cascade. This results in an overall improvement as well as a reduction of failure cases by half when faced with difficult images.

The main contributions of this work are:
1 – A novel cascaded regression method, called Robust Cascaded Pose Regression (RCPR). As we show in Section 5, RCPR outperforms previous landmark estimation work on four different, varied face datasets. RCPR is more robust to bad initializations, large shape deformations and occlusion. Moreover, RCPR is the first approach capable of detecting occlusions at the same time as it estimates landmarks, see Figure 1. Code is available online.
2 – The introduction of a challenging face landmark dataset: Caltech Occluded Faces in the Wild (**COFW**). This dataset is designed to benchmark face landmark algorithms in realistic conditions, which include heavy occlusions and large shape variations. The dataset is available online.

2. Related work

Early work on shape estimation includes Active Contours Models [28], Template Matching [45], Active Shape Models (ASM) [10] and Active Appearance Models (AAM) [9]. In recent years, improvements over AAM and ASM have been proposed [31, 11, 32, 13, 37, 36]. In general, these methods suffer from poor generalization performance and slow training. Although some of these issues have been mitigated, none of these approaches reach state-of-the-art performance on "in the wild" datasets.

Other popular modern approaches to detect the pose or parts of an object involve first detecting the object parts independently and then estimating pose and/or shape through flexible parts models [5, 27, 2, 20] or directly from detections [21, 19, 8]. These methods are effective at detecting articulated objects [44] and localizing objects from multiple views in difficult scenarios [41, 47]. However, our experimental results seem to indicate that they are less suited for high accuracy landmark estimation, see Section 5.2.

Another option is to tackle shape estimation as a regression problem, learning regressors that directly predict the object shape or the location of its parts, starting from a raw estimate of its position [39, 14, 16, 12, 7, 42, 25, 6]. These methods generally use boosted regression [23, 15] and random fern regressors [35]. Also key to iterative regression methods are *shape-indexed features* first introduced by Fleuret et al. [22], whose output depends on both the image and the current shape estimate. Current regression methods are fast and tolerate a small amount of shape variations but are not robust to occlusions and large shape variations.

We find that occlusions and large shape variations are quite common in real-world faces. In Section 3 we propose a novel regression method designed to be robust to both. In Section 4 we introduce a new, more realistic face dataset, collected with a focus on real-world occlusions and a variety of expressions. We benchmark our method against several of the methods mentioned above both in pre-existing datasets and our new dataset, see Section 5.

3. Method

To make this paper self-contained we first review the original CPR approach [14] and the improved variant proposed in [7]. Then, in Section 3.2 we describe our approach: Robust Cascaded Pose Regression (RCPR).

3.1. Cascaded Pose Regression (CPR)

Algorithm 1 shows the main steps of the CPR [14] evaluation procedure. CPR is formed by a cascade of T regressors $R^{1..T}$ that start from a raw initial shape guess $\mathcal{S}^0$ and progressively refine estimation, outputting final shape estimation $\mathcal{S}^T$. Shape $\mathcal{S}$ is represented as a series of P part locations $\mathcal{S}_p = [x_p, y_p], p \in 1..P$. At each iteration, regressors R^t produce an update $\delta\mathcal{S}$, which is then combined with previous iteration's estimate $\mathcal{S}^{t-1}$ to form a new shape.

input : Image I, initial guess $\mathcal{S}^0$, regressors $R^{1..T}$, shape-indexed features $h^{1..T}$
for $t = 1$ *to* T **do**
  `// compute shape-indexed features`
  $x^t = h^t(\mathcal{S}^{t-1}, I)$
  `// evaluate regressor`
  $\delta\mathcal{S} = R^t(x^t)$
  `// Update estimation`
  $\mathcal{S}^t = \mathcal{S}^{t-1} + \delta\mathcal{S}$
end
output: final estimation $\mathcal{S}^T$

Algorithm 1: CPR evaluation given an image I, initial raw shape estimation $\mathcal{S}^0$, and trained cascade regressors $R^{1..T}$ with shape-indexed features $h^{1..T}$.

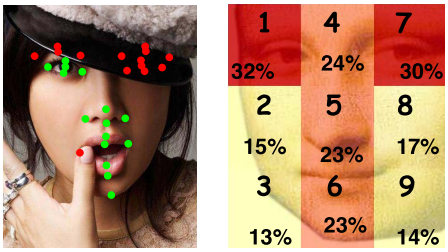
During learning, each regressor R^t is trained to attempt to minimize the difference between the true shape and the shape estimate of the previous iteration $\mathcal{S}^{t-1}$. The available features depend on the current shape estimate and therefore change in every iteration of the algorithm; such features are known as pose-indexed or shape-indexed features. The key to CPR lies on computing robust shape-indexed features and training regressors able to progressively reduce the estimation error at each iteration.

Both [14, 7] use depth 5 random fern regressors as regressors R^t and shape-indexed control point features [35]. Each fern selects which 5 features to use from a large pool of F features via either a random-step optimization [14] or a correlation-based evaluation [7] which is faster and improves performance.

Cao et al. [7] proposed a number of improvements over CPR [14]. To speed-up training convergence and improve overall performance, [7] performs regression on all shape parameters at once instead of one parameter at a time, effectively exploiting shape constraints. To strengthen regressors, Cao et al. use two-level *boosted* regression [23, 15]. Finally, to improve feature invariance to shape variations, features are referenced *locally* with respect to their closest landmark instead of globally with respect to global shape as [14] originally proposed.

3.2. Robust Cascaded Pose Regression (RCPR)

Both the original CPR [14] and the variant proposed in [7] struggle when faced with occlusions and large shape variations. Boosted regressors are unable to handle outliers in a principled way, causing a propagation of errors inside the cascade, harming the whole process. Occlusions and large shape variations are very common in real-world faces. We propose a new method, called Robust Cascaded Pose Regression (RCPR), which improves robustness to both.



(a) (b)

Figure 2. **COFW** dataset. (a) example annotation with occlusion information (red=occluded, green=unoccluded) (b) Dataset occlusion statistics, grouped in 9 zones. Average face occlusion is 23%. Eyes and eyebrows are most often occluded (hair, hats, sunglasses), as well as center of the face (object interactions).

3.2.1 Robustness to occlusion

Current approaches struggle under occlusion because they do not treat it in a principled way. We propose to incorporate occlusion directly during learning to improve shape estimation. Our method requires ground-truth annotations for occlusion in the training set. This information can be added with minor cost during the annotation procedure, adding a flag to each landmark encoding its visibility, see Figure 2(a).

Shape $\mathcal{S}$ is traditionally represented as a series of P part locations $\mathcal{S}_p = [x_p, y_p], p \in 1..P$. We propose to extend this definition, incorporating also part visibility $\mathcal{S}_p = [x_p, y_p, v_p]$, where $v_p \in \{0, 1\}$. We propose to learn this third dimension directly at the same time as the part locations. As we will show, this information is not only a richer representation of the object shape, it can also be of great use to better handle occlusions during shape estimation.

In our framework, part locations are initialized randomly by re-projecting training shapes into a raw guess of the object location, as usual. Then, all three dimensions are learnt simultaneously using cascaded regression (treating visibility as a continuous, non-binary variable).

CPR's coarse-to-fine nature implies that occlusion estimation starts to be accurate from early in the cascade. This suggests that occlusion information can be used at the same time as it is being estimated to help shape estimation. We introduce a novel occlusion-centered approach which leverages occlusion information to improve the robustness of shape updates $\delta\mathcal{S}$ at each iteration.

Given an image, the face (whose location is provided by a face detector) is divided into a $3x3$ grid, see Figure 2(b). At each iteration t, the amount of occlusion present in each one of the 9 zones can be estimated by projecting the current estimate $\mathcal{S}^{t-1} = [x_{1..P}, y_{1..P}, v_{1..P}]$ in the image. Then, instead of training a single boosted regressor R^t at each iteration t, we propose to train S_{tot} regressors $R^t_{1..S_{tot}}$, ensuring they are "visually different". More precisely, each regressor is allowed to draw features only from 1 of the 9 pre-defined zones (sampled randomly for each regressor).

Finally, each of the regressor's proposed updates $\delta\mathcal{S}_{1..S_{tot}}$ is combined through a weighted mean voting, where weight is inversely proportional to the total amount of occlusion present in the zones from which the regressor drew features. We found that good results are achieved using as little as $S_{tot} = 3$ regressors (see Supp. Material for impact of using different number of face zones, sampling areas and S_{tot}). Shape-indexed feature computation $x = h(\mathcal{S}, I)$ is unchanged, allowing features to be computed around occluded landmarks. This allows regressors to learn image occlusions properly.

At the end of the cascade, predicted visibility $v^i_p \in \mathbb{R}$ for all images $i = 1..N$ and parts $p = 1..P$, needs to be converted back to binary. Given ground-truth $v^i_p \in \{0, 1\}$ visibility, a threshold τ is selected by computing the precision/recall curve, and selecting the most desirable performance point (task-dependent). In our case, we chose to tune performance towards high precision, due the difficulty of the task, selecting as threshold that achieves 90% precision during training. During evaluation, the same process is carried out, using the same threshold value τ found during training. With this setting, RCPR is still able to predict occlusion with high precision in test images.

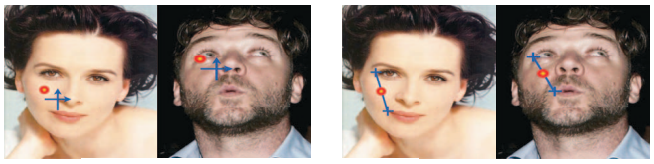
The key behind our approach is that we enforce "visually different" regressors to reach consensus, trusting more those using features from non-occluded areas of the image. As we show in Section 5.2, adding our occlusion reasoning results in a win-win scenario: it improves both landmark estimation and occlusion detection.

3.2.2 Robustness to shape variations

Interpolated shape-indexed features. Real-world faces present large variations, due to differences in expressions and pose. Shape-indexed features invariant to face scales and poses are key to shape estimation success under these conditions. With this goal in mind, [7] proposed to compute a similarity transform to normalize the current shape to a mean shape, and reference pixels by its local coordinates δ_x, δ_y with respect to its closest landmark.

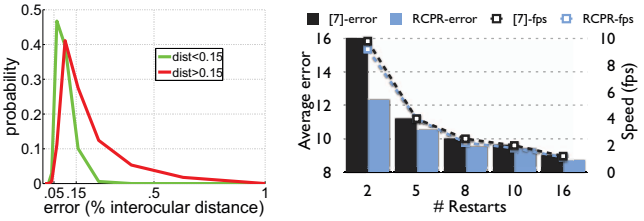
This is more robust than referencing features directly with respect to the global shape as originally proposed in [14]. However, these features are still not robust enough against large pose variations and shape deformations. If the number of annotated landmarks is low in a given region, it is likely that many of the features fall far from any landmark, becoming increasingly subject to small variations.

To overcome this issue, we propose to reference features by linear interpolation between two landmarks. These new features are much more robust to shape variations as shown in Figure 3. In Section 5 we show that the new features improve overall performance and greatly reduce failure cases. Furthermore, since there is no longer need to find the closest landmark in the current estimate of the shape for each feature, computation is considerably faster (3x speedup).



(a) (b)

Figure 3. Referencing shape-indexed features as points in the line between two landmarks increases feature invariance to large pose variations. (a) Features from [7]. (b) Our features.



(a) (b)

Figure 4. Smart restarts. (a) Final error distribution of parallel RCPR runs on **COFW** presenting a low/high (green/red) variance after 10% of the iterations. The difference in shape of the red curve (longer tailed distribution) suggests that variance can be used to predict failure cases early on. (b) Average error (bars) and speed (dashed lines) of the smart restarts (blue) compared with the traditional approach (black) on **COFW** as a function of the number of initializations used.

Smart restarts. CPR is initialization dependent. Both [14, 7] propose to run the algorithm several times using different initializations and take as final output the median of all predictions. Given an image, each restart is run independently, ignoring estimations until the very end. This approach fails to leverage an important fact: due to the natural coarse-to-fine nature of CPR, even if initialized differently, after just a few iterations each run should reach similar predictions. If they don't, its a clear sign that the regression is failing, see Figure 4(a).

We propose a novel restart scheme. Given an image and a number of different initializations, first only 10% of the cascade is applied to each. Then, the variance between their predictions is checked. If the variance is below a certain threshold, the remaining 90% of the cascade is applied as usual. If, however, the variance is higher than the threshold, the process is restarted with a different set of initializations.

We found by cross-validation that the threshold can be set to 0.15, and that three iterations are a good trade-off between performance and speed. This approach improves performance compared with the classical approach, especially at low number of restarts, while maintaining similar speed, see Figure 4(b). More sophisticated approaches are possible, however, the straightforward scheme described above is already quite effective.

4. Datasets and implementation details

4.1. Datasets

We first report the performance of our method on three popular datasets: Labeled Face Parts in the Wild (**LFPW**) [1], **HELEN** [29] and Labeled Faces in the Wild (**LFW**) [26, 12].

LFPW is one of the most used datasets to benchmark face landmark estimation in unconstrained conditions, and is composed of 1300 images, annotated with 29 landmarks. **HELEN** is a collection of 2,330 high resolution face portraits downloaded from *Flickr*. Faces are densely annotated using 194 landmarks, representing a benchmark for high detail face landmark localization. **LFW** contains facial images of 5749 individuals, 1680 of which have more than one image in the database. It consists of 13,233 images, collected "in the wild" and annotated with 10 landmarks.

Our approach outperforms all previous work on these datasets, see Table 1. However, we could not exploit all the benefits of our method due to the lack of occlusions in these datasets and performance saturation (RCPR reaches results almost on par with humans on **LFPW** and **LFW**). These datasets are not challenging enough since they do not contain faces showing high variations in pose, expressions and occlusions which are typical in real-world images. Therefore, we produced a new and significantly more challenging dataset, which we call Caltech Occluded Faces in the Wild (**COFW**).

Our face dataset is designed to present faces in real-world conditions. We wanted faces showing large variations in shape and occlusions due to differences in pose, expression, use of accessories such as sunglasses and hats and interactions with objects (e.g. food, hands, microphones, etc.). We asked four people with different levels of computer vision knowledge to each collect 250 faces representative of typical real-world images, with the clear goal of challenging computer vision methods.

The result is 1,007 images of faces obtained from a variety of sources. All images were hand annotated in our lab using the same 29 landmarks as in **LFPW**. 150 images were annotated twice by different people to measure human performance. We annotated both the landmark positions as well as their occluded/unoccluded state, see Figure 2. The faces are occluded to different degrees, with large variations in the type of occlusions encountered. **COFW** has an average occlusion of over 23%.

To increase the number of training images, and since **COFW** has the exact same landmarks as **LFPW**, for training we use the original non-augmented 845 **LFPW** faces + 500 **COFW** faces (1345 total), and for testing the remaining 507 **COFW** faces. To make sure all images had occlusion labels, we annotated occlusion on the available 845 **LFPW** training images, finding an average of only 2% occlusion.

4.2. Implementation details

In all experiments, to best replicate results of [7, 1], we simulate the output of a face detector providing the bounding box location and scale of the face with a minimum 80% random overlap with ground truth. Bounding boxes are used to project training shapes into test images as initialization to the algorithm(s) (also for data augmentation during training). For methods that already include their own detector (methods in [47]) we avoid false alarms and gross detection errors by cropping the image around the face. All reported run times correspond to a standard 3.47 GHz CPU using Matlab.

Code for [7] is not publicly available; we reimplemented the method ourselves starting from the original CPR code [14] available online. In our implementation, fewer boosted regressors and more iterations ($T = 100, K = 50$) perform better than what originally reported by the authors as optimal ($T = 10, K = 500$). We therefore use those values in all experiments for both [7] and RCPR.

When using $S_{tot} > 1$ regressors for robustness to occlusions, we reduce K accordingly to have approximately the same total number of regressors in the cascade (*e.g.* $S_{tot} = 3, K = 15$). The rest of the parameters are set to the original values recommended by [7]: number of features $F = 400$, data augmentation factor of 20 during training, and 5 restarts during testing. We also use depth 5 random fern regressors.

5. Results

In all cases we report the average error, percentage of failure cases, and speed which is measured in frames per second (fps). Errors in all datasets are measured as the average landmark distance to ground-truth, normalized as percentages with respect to interocular distance. We consider any error above 10% to be a failure, as proposed in [12].

5.1. LFPW, HELEN and LFW

Since the main component of RCPR is occlusion-centered regression and occlusion is virtually non-existent in **LFPW**, **HELEN** and **LFW** datasets, in this section we benchmark a version of RCPR which uses only the new shape-indexed features and smart restarts.

For **LFPW**, some URLs are no longer valid, so we were only able to download 845 of the 1,000 training images and 194 of the 300 test images, resulting in different training/test sets compared to [1, 7]. Due to these missing images our error for [7] is slightly different from what was originally reported. For **LFW**, instead of using a fixed training/test set as in the other two, the evaluation procedure proposed in [12] consists of a ten-fold cross validation using each time 1,500 training images and the rest for testing.

Table 1 shows the results on all three datasets. Both [7] and RCPR outperform previous work on all datasets, proving the efficiency of cascaded regression. RCPR improves [7]'s results in all cases, reducing failure cases by half, proving its higher robustness to outliers. RCPR is also between 1.5 to 4 times faster than competing approaches.

(a) LFPW

(b) HELEN

(c) LFW

Figure 5. Example RCPR results on **LFPW**, **Helen** and **LFW**. RCPR reduces failure rate by nearly half compared to the state-of-the-art to just 2-8%, see Table 1.

Usefulness of **LFPW** and **LFW** is reaching saturation, as the state of the art is already very similar to human performance. Figure 5 shows some example results. See Supp. Material for more detailed evaluation on these datasets.

5.2. Caltech Occluded Faces in the Wild (COFW)

We benchmark RCPR against [7] as well as the pre-trained DPM methods available online from Zhu et al. [47]. Both [7] and RCPR are trained on **COFW**. Both methods from [47] were trained using 900 positive images from CMU Multi-PIE dataset [24] and 1218 negative images. Apart from full RCPR, we also benchmark separately each of its components to study their individual contribution. Figure 6 shows landmark estimation results and Figure 7 shows occlusion detection results for each RCPR variant.

Each of RCPR's components contributes to improve landmark estimation. The new features reduce the number of failures by 8% and speedup computation by a factor of 3x. The smart restarts also reduce errors while maintaining similar speed performance. The occlusion-centered regression further improves landmark estimation at some cost in speed. Combining different regressors and weighting them

LFPW				HELEN				LFW			
Method	error	failures	fps	Method	error	failures	fps	Method	error	failures	fps
				[32]	11.1	-	-	[19]	12.0	-	-
[1]	3.90	-	-	[29]	9.1	-	-	[12]	7.0	-	10
[7]	3.8	4%	3	[7]	7.1	13%	2	[7]	5.9	7%	11
RCPR	**3.5**	**2%**	**12**	**RCPR**	**6.5**	**8%**	**6**	**RCPR**	**5.3**	**4%**	**15**
Human	3.28	0%	-	Human	3.3	-	-	Human	4.5	-	-

Table 1. Results on **LFPW**, **HELEN** and **LFW** datasets. Errors are measured as percentage of the interocular distance. We report both the mean error and the failure rates. **RCPR reduces failure cases by half compared to best performing method on all three datasets**.

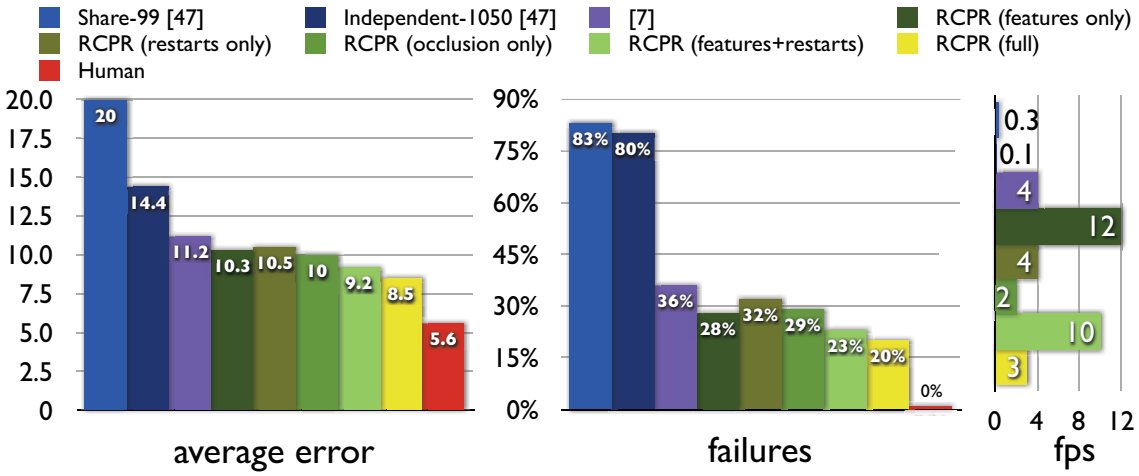


Figure 6. Results on **COFW**. Errors are measured as percentage of the interocular distance. See Section 5.2 for dataset/methods details. We report both the mean error and the failure rate to better capture each method's performance. We also report speed measured in frames-per-second (fps). Note that methods from [47] were trained on a different dataset. **RCPR reduces failure cases 16% (almost half) compared to best performing method.**

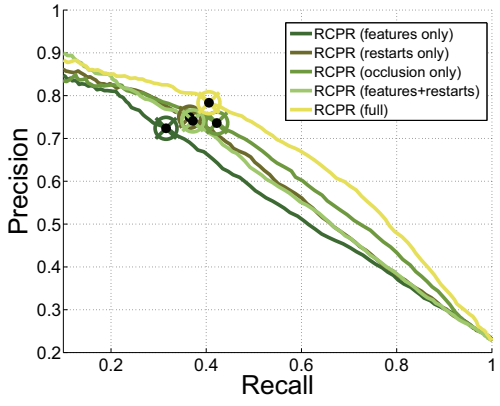


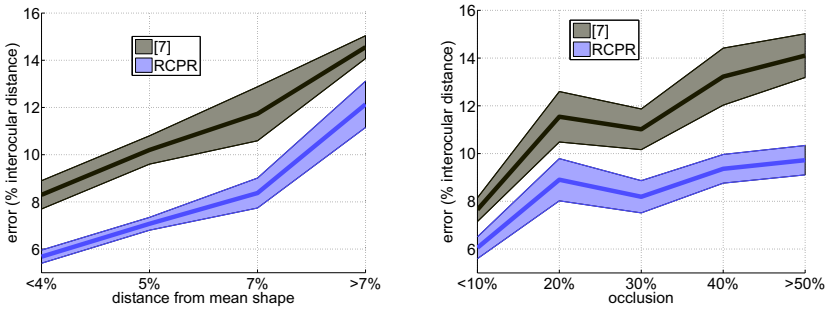
Figure 7. Occlusion detection precision/recall curves of the different variants of RCPR on **COFW**. Crosses mark selected threshold for occlusion classification. Using occlusion-centered regression clearly improves area under the curve for occlusion detection.

according to their occlusion also improves the area under the curve for occlusion detection around 10%.

Full RCPR improves on previous cascaded regression approaches [7] by a large margin, especially improving on difficult images, reducing the number of failure cases by 16% (almost half). As shown in Figure 8, RCPR improves results overall, and especially when faced with high occlusions, making it more suited for real-world applications.

RCPR also outperforms methods in [47]. These methods struggle with occlusion because they weren't trained on it. However, by looking at results on the 90 non-occluded images present in the dataset (occlusion$< 5\%$), both methods in [47] have average errors above 10.9, compared to RCPR's 5.5. Possibly these results could be improved, but this is outside of our scope.

Overall, performance of all methods is much lower in our more realistic and challenging **COFW** dataset. Figure 9 shows some example RCPR results. See Supp. material for a comparison between human and machine.



(a) shape variation (b) face occlusion

Figure 8. RCPR vs. [7] error comparison on **COFW**. Both average error and standard deviations are shown. (a) Error vs. increasing shape distance from average. (b) Error vs. amount of occlusion. RCPR improves [7] overall, and most significantly under high occlusions, where [7] struggles.

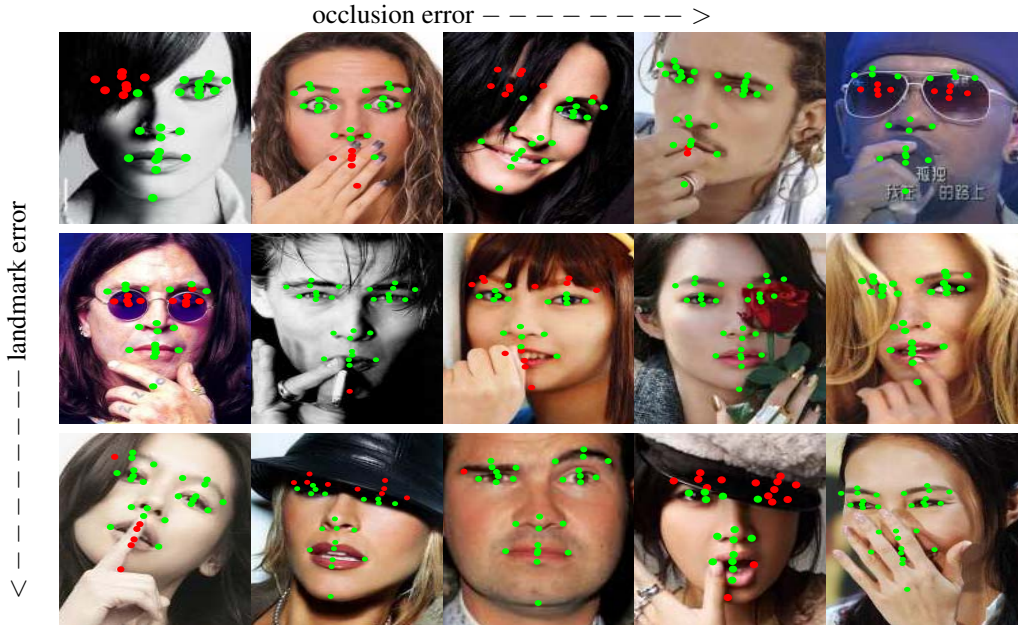


Figure 9. Example RCPR results from **COFW**. Results are ordered by increasing landmark estimation error (Y axis) and occlusion error (X axis). RCPR succeeds at localizing face landmarks within 10% of their true location in 80% of **COFW** images, and detects occlusion with an 80/40% precision/recall.

6. Discussion and conclusions

Occlusions and high shape variances are a difficult challenge for current face landmark estimation methods. We proposed a novel method, called Robust Cascaded Pose Regression (RCPR), that improves robustness of previous work against both. RCPR is capable of detecting occlusions explicitly, estimating both the landmark positions and their occlusion. We first benchmarked RCPR against several state-of-the art approaches on three pre-existing challenging face datasets (**LFPW**, **HELEN** and **LFW**). RCPR improves previous methods on all three datasets, while being faster. However, we could not exploit all of RCPR's capabilities due to dataset saturation (RCPR performs almost on par with humans in **LFPW** and **LFW**) and lack of occlusions. Therefore, we produced a much more challenging face dataset, called Caltech Occluded Faces in the Wild (**COFW**). This dataset represents a very challenging task due to the large amount and variety of occlusions and large shape variations. We show that our method clearly improves previous work on this dataset, thanks to its unique occlusion-centered reasoning. The COFW dataset and RCPR code are available online.

Acknowledgments

This work is funded by the Gordon and Betty Moore Foundation and ONR MURI Grant N00014-10-1-0933.

References

[1] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar. Localizing parts of faces using a concensus of exemplars. In *CVPR*, 2011.

[2] L. Bourdev and J. Malik. Poselets:body part detectors trained using 3D human pose annotations. In *ICCV*, 2009.

[3] S. Branson, C. Wah, F. Babenko, B. Schroff, P. Welinder, P. Perona, and S. Belongie. Visual recognition with humans in the loop. In *ECCV*, 2010.

[4] X. Burgos-Artizzu, P. Dollár, D. Lin, D. Anderson, and P. Perona. Social behavior recognition in continuous videos. In *CVPR*, 2012.

[5] M. Burl, M. Weber, and P. Perona. A probabilistic approach to object recognition using local photometry and global geometry. In *ECCV*, 1998.

[6] C. Cao, Y. Weng, S. Lin, and K. Zhou. 3D shape regression for real-time facial animation. In *SIGGRAPH*, 2013.

[7] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*, 2012.

[8] H. Cevikalp, B. Triggs, and V. Franc. Face and landmark detection by using cascade of classifiers. In *FG*, 2013.

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

[10] T. Cootes and C. Taylor. Active shape models. In *BMVC*, 1992.

[11] D. Cristinacce and T. Cootes. Boosted regression active shape models. In *BMVC*, 2007.

[12] M. Dantone, J. Gall, G. Fanelli, and L. VanGool. Real-time facial feature detection using conditional regression forests. In *CVPR*, 2012.

[13] L. Ding and A. Martinez. Features vs. context: An approach for precise and detailed det. and delineation of faces and facial features. *PAMI*, 32(11):2022–2038, 2010.

[14] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*, 2010.

[15] N. Duffy and D. P. Helmbold. Boosting methods for regression. *Machine Learning*, 47(2-3):153–200, 2002.

[16] B. Efraty, C. Huang, S. Shah, and I. Kakadiaris. Facial landmark det. in uncontrolled conditions. In *IJCB*, 2011.

[17] P. Ekman and W. Friesen. *Facial action coding system*. 1977.

[18] N. K. et al. Leafsnap: A computer vision system for automatic plant species identification. In *ECCV*, 2012.

[19] M. Everingham, J. Sivic, and A. Zisserman. Hello! My name is... Buffy - automatic naming of characters in tv video. In *BMVC*, 2006.

[20] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *PAMI*, 32(9):1627–1645, 2010.

[21] F. Fleuret and D. Geman. Fast face detection with precise pose estimation. In *ICPR*, 2002.

[22] F. Fleuret and D. Geman. Stationary feat. and cat detection. *J. of Machine Learning Research*, 9:2549–2578, 2008.

[23] J. H. Friedman. Greedy function approximation:a gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[24] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. In *FG*, 2008.

[25] I. P. H. Yang. Privileged information-based conditional regression forest for facial feature detection. In *FG*, 2013.

[26] G. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face rec. in unconstr. environments. Technical report, Amherst, 2007.

[27] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz. Robust face detection using the Hausdorff dist. In *AVBPA*, 2001.

[28] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *IJCV*, 1(4):321–331, 1988.

[29] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*, 2012.

[30] A. Martinez and S. Du. A model of the perception of facial expressions of emotion by humans: Research overview and perspectives. *JMLR*, 13:1589–1608, 2012.

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

[32] S. Milborrow and F. Nicolls. Locating facial features with an extended active shape model. In *ECCV*, 2008.

[33] E. Murphy-Chutorian and M. Trivedi. Head pose estimation in computer vision:a survey. *PAMI*, 31(4):607–626, 2009.

[34] M.-E. Nilsback and A. Zisserman. Automated flower classification over a large num. of classes. In *ICVGIP*, 2008.

[35] M. Ozuysal, M. Calonder, V. Lepetit, and P. Fua. Fast keypoint recognition using random ferns. *PAMI*, 32(3):448–461, 2010.

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

[37] C. P. Sauer and T. Cootes. Accurate regression procedures for active appearance models. In *BMVC*, 2011.

[38] J. Sivic, M. Everingham, and A. Zisserman. Who are you? Learning person specific classifiers from video. In *CVPR*, 2009.

[39] M. Valstar, B. Martinez, X. Binefa, and M. Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, 2010.

[40] C. Wah, S. Branson, P. Perona, and B. S. Multiclass recog. and part localiz. with humans in the loop. In *ICCV*, 2011.

[41] M. Weber, W. Einhauser, M. Welling, and P. Perona. Viewpoint-invariant learning and detection of human heads. In *FG*, 2000.

[42] H. Yang and I. Patras. Face parts localization using structured-output regression forests. In *ACCV*, 2012.

[43] M.-H. Yang, D. Kriegman, and N. Ahuja. Detecting faces in images: a survey. *PAMI*, 24(1):34–58, 2002.

[44] Y. Yang, S. Baker, A. Kannan, and D. Ramanan. Recognizing proxemics in personal photos. In *CVPR*, 2012.

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

[46] W. Zhao, R. Chellappa, P. Phillips, and A. Rosenfeld. Face recognition: A literature survey. *ACM Computing Surveys*, 35(4):399–458, 2003.

[47] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localiz. in the wild. In *CVPR*, 2012.