Learn to Combine Multiple Hypotheses for Accurate Face Alignment

Junjie Yan Zhen Lei Dong Yi Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, China
{jjyan,zlei,dyi,szli}@nlpr.ia.ac.cn

Abstract

In this paper, we present the details of our method in attending the 300 Faces in-the-wild (300W) challenge. We build our method on cascade regression framework, where a series of regressors are utilized to progressively refine the shape initialized by face detector. In cascade regression, we use the HOG feature in a multi-scale manner, where the large pose validation is handled in early stages by HOG feature at large scale, and then shape is refined at later stages with HOG feature at small scale. We observe that the performance of the cascade regression method decreases when the initialization provided by face detector is not accurate enough (for faces with large appearance variations, face detection is still a challenging problem). To handle the problem, we propose to generate multiple hypotheses, and then learn to rank or combine these hypotheses to get the final result. The parameters in both learn to rank and learn to combine can be learned in a structural SVM framework. Despite the simplicity of our method, it achieves state-of-the-art performance on LFPW, and dramatically outperforms the baseline AAM on the 300-W challenge.

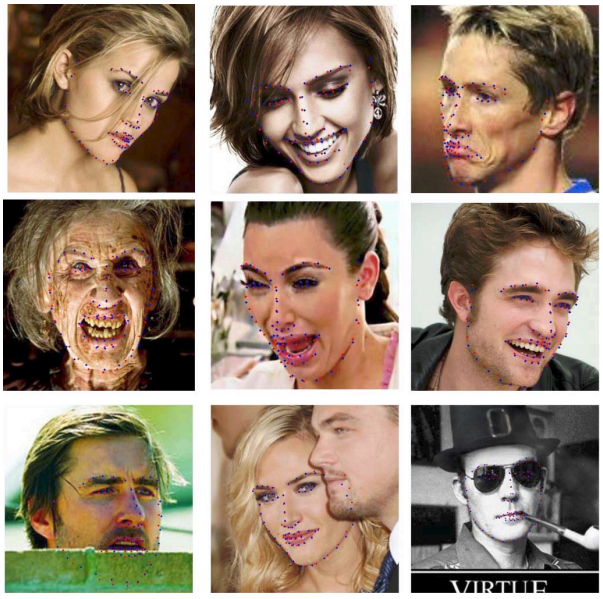


Figure 1. Qualitative results of our method for face alignment with pose, expression and occlusion. The red points are the groudtruth and the blue points are the output of our method. (Best viewed in color.)

1. Introduction

Face alignment is one important component in face based applications, such as face attribute and expression analysis. Recent works also showed the importance of face alignment in real world face recognition task on LFW [12]. Due to its importance, a lot of works were proposed to advance face alignment, and achieved remarkable improvements on standard benchmarks, such as BioID [14] and LFPW [1]. However, face alignment is still an unsolved problem, especially for automatical real world applications, where large appearance variations exist.

Traditional face alignment methods are usually based on an assumption that a reliable initialization from face detector is provided and the face alignment methods iteratively optimize the shape. Although the assumption is always hold for faces in constrained setting, it is probably not hold for more wild faces in unconstrained environment where large appearance variations such as pose, expression, and occlusion exist, which greatly affects the performance of face detector. For example, recent good performance face detectors [25, 24] still has some biased detections in the 300-W train dataset. In face detection field, performance is usually measured by the 50% overlap metric [13], which is not accurate enough for the initialization of face alignment algorithm. The biased face detection would greatly decrease the face alignment performance. Despite the importance in real applications, however, the problem does not attend so much attention, mainly due to that current face alignment benchmarks are generally collected from easily detected faces.

In this paper, we try to make face alignment algorithm robust to the initialization of face detector. In the initial-

ization step, we generate multiple hypotheses by randomly rescale and shift the bounding box provided by the face detector, and then estimate face landmark location independently from these initializations. Two strategies are proposed to fuse these hypotheses to the final shape output. The first is learn to rank, where a function is learnt to rank multiple hypotheses. The second is learn to combine, which automatically determines a combination of shape parts from multiple hypotheses. We show that these two strategies can be handled in a unified structural SVM framework.

For each hypothesis, the shape is estimated by cascade regression, originally proposed in [10] for pose estimation task in image sequence. In the learning phase, the model learns a series of regressors. In each iteration of the test phase, the shape bias is estimated by the specific regressor in this iteration according to the feature around landmarks. Specially in our implementation, we use simple linear regressor and HOG [8] feature for the generalization in faces in the wild. The regression matrix can be effectively learned by least square regression method. We use HOG feature in a multi-scale manner, where in the first stages the HOG feature extracted at large scales can handle large poses, and at the later stages the HOG feature extracted at small scales can further refine the shape.

The following of the paper is organized as follows. In section 2, we review the related work. The cascade regression and learn to combine are discussed in section 3 and section 4, respectively. Experimental results are shown in section 5, and finally the paper is concluded in section 6.

2. Related Work

Generally speaking, previous works on face alignment can be divided into *model based approach* and *regression based approach*. Here we briefly discuss these two approaches.

Model based approach The model based approach learns a shape model in the training phase and uses the shape model to fit the novel faces in the test phase. The pioneering works such as active shape model (ASM) [6] and active appearance model (AAM) [3] were built on PCA constraints of shape and appearance. The ASM has been adopted by a lot of works. In [7], the ASM is generalized to be constrained local model (CLM), where every landmark has a descriptor to capture appearance, and these descriptors were constrained by a shape model. In [19], more sophisticated local model and a mean-shift matching strategy were used to get good results. In [4], impressive performance was achieved by using random forest to vote for the best position of each landmark.

Regression based approach Since face alignment is naturally a regression problem, regression based approach has achieved great progress in recent years. Regression based approach benefits from robust local descriptors and regressors. [10] proposed to use cascade regression to estimate pose in image sequence with pose-indexed feature. Maybe the most interesting idea in this paper is to use cascade simple regressors to approximate complex mapping. [2] extended the work for face alignment task and achieve very promising result on face alignment task. [23] used the cascade regression with SIFT feature, and interpreted the cascade regression procedure from a gradient descent view. In [9], random forest was used to learn the map between the image patch and landmark position, and all the sampled patches were used to vote landmark locations. In [20], a deep neural network was conducted to directly learn the regression function between the original image and landmark position. Different from the model based approach, these regression based methods did not relay a parametric constraint on the shape, and were proved to be more suitable for face alignment in the wild.

3. Face Alignment by Cascade Regression

Face alignment is naturally a regression problem, where the input is a face image (and a rough bounding box for initialization), and the output is a shape parameterized by the coordinates of each face landmarks. In this part, we describe a simple yet effective cascade regression based method for face alignment.

In training, we have N training samples $\{I_i, S_i, S_i^0\}$, where I_i is the image, S_i the grountruth shape of face in I_i and S_i^0 is the initialization of S_i. We want to learn a regression function f to minimize the mean square error:

$$f = \arg\min_f \sum_{i=1}^{N} \|f(S_i^0, I_i) - S_i)\|_2, \tag{1}$$

where f returns a new shape based on the initial shape S_i^0 for each image I_i. Here we set the initial shape to be mean face shape of training image, and place it on each face according to the predication of face detection. The direct regression, however, can be very complex, due to the high dimensional output and complex nonlinear relationship. Instead, we use the cascade regression approach by dividing f into a series of simpler regression function $\{f_1, f_2, \cdots, f_T\}$, which satisfies that:

$$f = f_T \circ f_{T-1} \circ \cdots \circ f_1, \tag{2}$$

where the input of f_t is the output of f_{t-1}. By combining f_1 to f_T, the cascade approach can approximate complex nonlinear mapping between the initial shape and the true shape.

While the basic framework of cascade regression is very simple and basic, the critical problem is in how to design the sub-regression function, which dramatically affects the performance for face alignment. To keep f_t to be a nonlinear

function, we add a feature transform procedure:

$$S_i^t = f_t(S_i^{t-1}, I_i) = W_t \cdot \Phi(S_i^{t-1}, I_i), \quad (3)$$

where $\Phi(S_i^{t-1}, I_i)$ is the feature transform to encode the appearance information of image I_i around shape S_i^{t-1}, and W_t is a linear transformation to map appearance feature $\Phi(S_i^{t-1}, I_i)$ to a new shape. The whole function f_t is nonlinear once the feature transform is nonlinear. The linear transform matrix W_t can be learned in the training phase:

$$W_t = \arg\min_{W_t} \sum_{i=1}^{N} \|S_i - W_t \cdot \Phi(S_i^{t-1}, I_i)\|_2, \quad (4)$$

where W_t has closed-form solution by solving a least-square problem. It is worth noting that the regularization term is important for the final face alignment performance. We determine the number of iterations T on the training set, where the cascade regression stop once the objective function defined in Eq. 1 does not increase any more. The final face alignment hypothesis $\widehat{S}_i$ is set to be S_i^T.

Although arbitrary nonlinear feature descriptors such as LBP, Gabor, SIFT and HOG can be adopted in Eq. 4, different descriptors have quite different performance on face alignment task. In our experiments, we validate these features and find that HOG performs the best in this task. Moreover, we use HOG feature in a multi-scale manner. We find that HOG feature at large scale is useful to handle large shape variations in early stages, and HOG feature at small scale is useful to refine the local shape.

The above procedure is so simple that it can be implemented in a few lines of Matlab code. Surprisingly, it can achieve the state-of-the-art performance on LFPW. For the 29 landmarks annotated in [1], it achieves a mean error ($\times 10^{-2}$, normalized by the inter-ocular distance) of 2.79.

4. Learn to Combine Multiple Hypotheses

Although the cascade regression based face alignment achieves good performance, we find its output is very sensitive to the shape initialization. This phenomenon is especially remarkable for challenging faces with large pose, occlusion and expression, where the face detector is still not perfect. We argue that the current face detection task measured by 50% overlap radio is not accurate enough for face alignment task.

In order to reduce the influence of the initialization by face detection, we propose to first generate multiple hypotheses by randomly shifting and re-scaling the bounding box provided by face detector. The cascade regression model is applied to these different initializations respectively and a series of shape hypotheses are obtained. The question becomes how to fuse these hypotheses for the final output. Here we propose two strategies.

4.1. Learn to Rank

Given an image I_i and rough face bounding box provided by face detector b_i, we resize and shift b_i to get a series of bounding boxes $B_i = \{b_{i1}, \cdots, b_{iM}\}$. Here we can safely assume that at least one bounding box localizes closer to the groundtruth than the original bounding box. We use each b_{ij} to generate the initial shape S_{ij}^0, and then use the regression procedure described above to get a set of hypotheses $\{\widehat{S}_{i1}, \cdots, \widehat{S}_{iM}\}$.

To select the best one from multiple hypotheses, we define a rank function g to rank each hypothesis, and want g to ensure that the output of the best hypothesis is larger than any other hypothesis:

$$\widehat{S}_i = \arg\max_{i=1,\cdots,M} \{g(\widehat{S}_{i1}, I_i), \cdots, g(\widehat{S}_{iM}, I_i)\}, \quad (5)$$

where $\widehat{S}_i$ is the final output. In our application g is defined as a linear function that satisfies:

$$g(\widehat{S}_{ij}, I_i) = w_g^T \cdot \Phi(\widehat{S}_{ij}, I_i). \quad (6)$$

Here we use the same notation Φ as the cascade regression step, but note that they are not necessary the same feature.

4.2. Learn to Combine

Another observation is that even the best hypothesis does not always generate satisfying result and there is complementary information in different hypotheses. For example, one hypothesis corresponds to good landmark locations for border of a face, while another hypothesis results in a good alignment to landmarks exclude the border. This observation motivates us to learn a function to automatically select the optimal combination of multiple hypotheses. We name this procedure as learn to combine.

Similar to learn to rank, we define a function h to evaluate the combination. Given multiple hypotheses of a face, we optimizes the following problem:

$$\max_{\{\delta_1,\cdots,\delta_M\}} \quad h(\sum_j \delta_j \widehat{S}_{ij}, I_i) \quad (7)$$

$$s.t. \quad \sum_{k=1}^{K} \delta_{jk} = 1, \delta_{ij} \in \{0, 1\}, \quad (8)$$

where δ_{jk} belongs to $\{0, 1\}$. $\delta_{jk} = 1$ indicates that the j-th landmark is selected from the k-the hypothesis. $\sum_{k=1}^{K} \delta_{jk} = 1$ is used to constrain that one landmark is selected only once from multiple hypotheses. The final output $\widehat{S}_i$ is $\sum_j \delta_j \widehat{S}_{ij}$. Similar to g, the function f is assumed to be a linear function:

$$h(\widehat{S}_{ij}, I_i) = w_h^T \cdot \Phi(\sum_j \delta_j \widehat{S}_{ij}, I_i), \quad (9)$$

where w_h is the parameter of h, which is learned in the training phase.

Actually in our experiments, we find that the above constraint is not enough, since the latent non-parametric constraint shape in cascade regression is ignored in this formulation, which can result in a lot of wrong hypotheses. To this end, we further add a region constraint to Eq. 10:

$$\delta_{jk} = \delta_{ik} \quad if \quad R_i = R_j, \tag{10}$$

where R_i is the region index of the i-th landmark. With this constraint the final face alignment is forced to have a locality property, where the landmarks in a local region is selected from the same hypothesis. It is easy to see that "learn to rank" is a special case of "learn to combine" by setting the region constraint in learn to combine as the whole region.

4.3. Inference and Learning

The inference problem in both learn to rank and learn to combine are very trivial. In learn to rank, we evaluate each hypothesis independently, and select the one with the largest output. In learn to combine, we find the best fit for each region, and then combine different regions.

Now we discuss how to learn the parameters in ranking function g and combining function h. Since learn to rank is just a special case of learn to combine, here we just describe how to learn the parameters w_h in combining function h. Once an estimation of w_h is derived, it can be used in inference procedure and estimate the output of each face. The optimal parameter w_h should satisfies that:

$$w_h = \arg\min_{w_h} \frac{1}{2}\|w_h\|_2^2 + C\sum_{i=1}^{N}\Delta(S_i, \widehat{S}_i), \tag{11}$$

where the first term is used to regularize $\|w_h\|$, and the later term is used to measure the loss between the estimated shape $\widehat{S}_i$ by w_h and groundtruth shape S_i. Here the we define the loss function as $\Delta(S_i, \widehat{S}_i) = \|S_i - \widehat{S}_i\|_2^2$.

The above problem is a standard structural SVM problem [21], and can be learned by standard SVM package, such as [22].

5. Experiments

In this part, we first conduct experiments on publicly available LFPW dataset to examine different settings of the proposed method, and then report the final result on the 300-w challenge.

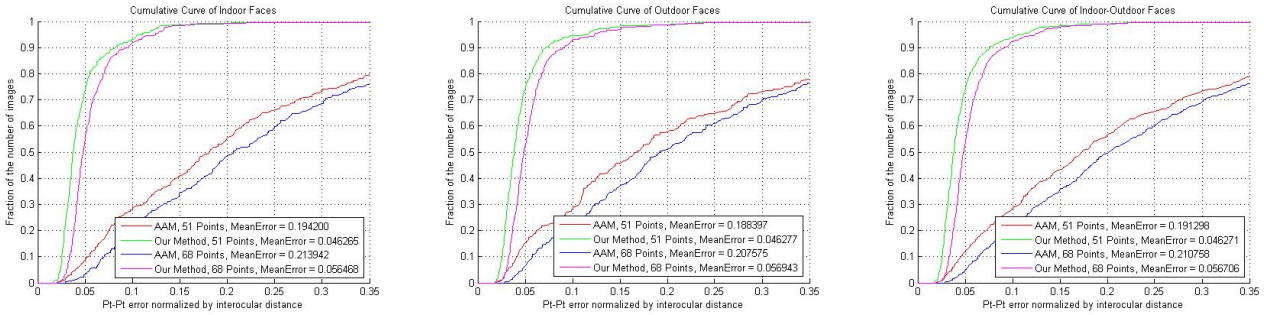
The basic framework of the cascade regression is quite simple, and actually it only has two parameters to tune: the number of iteration and the regularization term in least squares step. In our experiments, we find that satisfying performance can generally be achieved when the iteration number is 7 and the regularization term is set as the number of training samples.

While the basic cascade regression framework is easy to tune, the selection of feature is critical to the performance. In our experiments, we compare four features: SIFT, LBP, Gabor and HOG, and finally find that HOG performs the best on LFPW, and use it for the all the experiments. Fusing multiple features is perhaps able to further improve the performance, but we leave it as the future work.

The training and testing of the cascade regressors are very efficient. On LFPW, a 68 landmark model takes about 15 minutes for training, and 2 minutes for testing with non-optimized Matlab code runs at a PC with Intel X5650 CPU. For learn to rank and learn to combine, we use HOG features extracted around landmarks of each hypothesis. Since learn to rank is a specific form of learn to combine, we only use learn to combine for the final result. The features of each hypothesis can be cached in SVM training, so that this step is also very efficient in training.

Finally, we report the result of our submissions on the 300-W challenge. We strictly follow the protocol, and only use the provided training annotations generated by [11, 25, 15, 17, 18] to train our model. The face alignment algorithm is initialized by face detector from [25]. We submitted four methods, "demo1","demo2","demo3" and "demo4". The "demo1" and "demo3" are trained on the provided 7 datasets, "demo2" and "demo4" are trained on the provided 7 datasets and their mirrors. For "demo3" and "demo4", we use a three-view mixture model, where the a pose estimator is used to determine which view to use in the test phase. Since our four submissions generated similar performance, we only report the result of "demo4", which is slightly better than other three submissions.

The 51 and 68 point cumulative curves for our submission "demo4" and the baseline are shown in Figure 2. The baseline is a project-out inverse compositional AAM method proposed in [16] with the edge-structure feature used in [5]. Following the suggestions of the organizers, the faces are divided into "Indoor","Outdoor" and "Indoor-Outdoor", and the cumulative curves are reported for the three partitions independently. Based on the cumulative curves provided by the organizers, we also calculate the mean error of each method (which corresponds to the area above the cumulative curve). For the 51 landmarks without border, our method achieve a 0.045 mean error. Similar to the observations on LFPW, the mean error increases when border landmarks included, which is 0.056. The performance our method on all the three partitions have similar mean error, which indicates that our method is robust to different settings. We observe that our method decreases seriously on this dataset compared with the LFPW dataset, due to the more "wild" setting of 300-W. We are going to have a more in-depth analysis of the reason once the test data is available.



(a) Cumulative Curve of Indoor Faces (b) Cumulative Curve of Outdoor Faces (c) Cumulative Curve of Indoor-Outdoor Faces

Figure 2. Quantitative results on 300W dataset.

6. Conclusion

We present the details of our method in attending the 300 Faces in-the-wild challenge. Our method is built on a cascade regression framework with multi-scale HOG feature. We observe the performance of the cascade regression method decreases when the initialization provided by face detector is not accurate enough, and propose to handle this problem by learn to rank and learn to combine. Both of these two methods can be learned in the structural SVM framework. The proposed method achieves state-of-the-art performance on LFPW dataset, and dramatically outperforms the baseline AAM on the 300-W challenge. Currently, we do not know why the performance of our method decreases so seriously on the 300-W dataset compared with LFPW, and plan to find the exact reason and further improve it once the testset is available.

Acknowledgement

We thank the organisers of 300-W challenge for providing data and evaluating our submission. This work is supported by the Chinese National Natural Science Foundation Project #61070146, #61105023, #61103156, #61105037, #61203267, #61375037, National IoT R&D Project #2150510, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, European Union FP7 Project #257289 (TABULA RASA), and AuthenMetric R&D Funds.

References

[1] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar. Localizing parts of faces using a consensus of exemplars. In *CVPR*. IEEE, 2011. 1, 3

[2] X. Cao, Y. Wei, F. Wen, and J. Sun. Face alignment by explicit shape regression. In *CVPR*. IEEE, 2012. 2

[3] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. *TPAMI*, 2001. 2

[4] T. F. Cootes, M. C. Ionita, C. Lindner, and P. Sauer. Robust and accurate shape model fitting using random forest regression voting. In *ECCV*. Springer, 2012. 2

[5] T. F. Cootes and C. J. Taylor. On representing edge structure for model matching. In *CVPR*. IEEE, 2001. 4

[6] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *CVIU*, 1995. 2

[7] D. Cristinacce and T. Cootes. Automatic feature localisation with constrained local models. *Pattern Recognition*, 2008. 2

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*. IEEE, 2005. 2

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

[10] P. Dollár, P. Welinder, and P. Perona. Cascaded pose regression. In *CVPR*. IEEE, 2010. 2

[11] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker. Multi-pie. *Image and Vision Computing*, 2010. 4

[12] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 1

[13] V. Jain and E. G. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. *UMass Amherst Technical Report*, 2010. 1

[14] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz. Robust face detection using the hausdorff distance. In *Audio-and video-based biometric person authentication*. Springer, 2001. 1

[15] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang. Interactive facial feature localization. In *ECCV*. Springer, 2012. 4

[16] I. Matthews and S. Baker. Active appearance models revisited. *IJCV*, 2004. 4

[17] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. Xm2vtsdb: The extended m2vts database. In *Second international conference on audio and video-based biometric person authentication*. Citeseer, 1999. 4

[18] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPR Workshop on AMFG*. IEEE, 2013. 4

[19] J. M. Saragih, S. Lucey, and J. F. Cohn. Deformable model fitting by regularized landmark mean-shift. *IJCV*, 2011. 2

[20] Y. Sun, X. Wang, and X. Tang. Deep convolutional network cascade for facial point detection. In *CVPR*. IEEE, 2013. 2

[21] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun. Large margin methods for structured and interdependent output variables. In *Journal of Machine Learning Research*, pages 1453–1484, 2005. 4

[22] A. Vedaldi. A MATLAB wrapper of SVM$^{\text{struct}}$. `http://www.vlfeat.org/~vedaldi/code/svm-struct-matlab.html`, 2011. 4

[23] X. Xiong and F. De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*. IEEE, 2013. 2

[24] J. Yan, X. Zhang, Z. Lei, D. Yi, and S. Li. Structural models for face detection. In *Automatic Face & Gesture Recognition(FG 2013), 2013 IEEE International Conference on*. IEEE, 2013. 1

[25] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*. IEEE, 2012. 1, 4