Face Alignment by Supervised Descent Method with Head Pose Estimation

Zheng Zheng

Nanjing, Jiangsu Province, China

zhzheng2014@gmail.com

Abstract. Face alignment, which aims at locating facial key points automatically, is an important topic in computer vision community. And many works have been done to solve this problem. The most well-known solution is Supervised Decent Method(SDM). However, SDM has been designed to use mean shape as initial shape, which is vulnerable to large pose variation. In this paper, we present a novel approach for detection of the facial key points, getting initial shape from a special shape according to the head pose of the data. Experiments show that our approach achieves significant improvement. In both 21 points and 68 points detection cases, our method achieves nearly 50% improvement on challenging dataset IBUG, and about 1% improvement in HELEN and LFPW test set.

1. Introduction

Face alignment is an essential preprocessing step for many face analysis tasks, e.g. face recognition[1], expression recognition[2], and facial attribute analysis[3]. Among most face alignment algorithms, cascades methods[4, 5] have becoming one of the most popular and state-of-the-art methods. The method starts from an initial shape, e.g. mean shape of training samples, and refine the shape through several trained regression stages. Although, cascaded method can achieve good result in test samples, like 300W, it has been blamed for its initialization dependency[6]. Many researchers show that without suitable initial shape, the detection accuracy of cascaded model will be degraded.

In this paper, we re-design the initial part of the cascaded method, to make sure it can perform better when it encounters large pose variation. Since many works[5, 6] have suggests that, using mean shape of all training sample as the initial shape of all the training data to train SDM model will trapped into local optimal. So we consider to use pose-directed shape as initial shape, with considering that different data with similar head pose will have similar shape. The head pose of pose-directed shape is close to the real head pose of the input image. To achieve this goal, we introduce head pose estimation as the first step of the face alignment, and then classified each train sample into several clusters according to head pose of the image. Then we use the mean shape of each cluster as the initial shape to train several regressors. In the test process, we also choose the initial shape according to the head pose of the image.

We use trainset 300W[7] as the evaluation samples, including AFW, HELEN, LFPW, and the challenging testset IBUG. The experimental results show that our approach performs much better than the traditional SDM method in these datasets, especially nearly 50% error reduction in the challenging IBUG dataset. As the proposed initialization-optimization is generalized, it can significantly benefit other cascaded methods.

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

2. Related work

A number of methods have been proposed to solve face alignment, including cascaded regression[4, 5] and coarse-to-fine methods[6]. Supervised descent method (SDM) is one of the successful cascaded regression methods. It is proposed to solve nonlinear least squares optimization problem. By using training data, SDM learns a series of parameter updates to minimize the overall errors of whole training samples. But SDM strongly depends on suitable initial shape. If the initial shape is far from the target, the final solution may be trapped in local optimal. Coarse-to-Fine shape searching(CFSS), on the contrary, relaxes the need of shape initialization. It constructs shape space from training data, and draw several shapes in shape space as initial shapes. However, CFSS needs to maintain whole shape space of all the trainset and perform several stages of shape searching in the whole space, which limits the initial shape ranges and regression speed.

In our work, we evaluate both the advantages and disadvantages of SDM and CFSS. SDM is vulnerable to large pose variation because initial shape (mean shape) would be far from target shape and CFSS, which contains the shape space with large size, provides a promising direction of solving this problem. We first construct shape space based on head pose from training data. We divide all the training data into several parts according to the head pose to keep the angular range in each part the same. In practice, we divide all training data into 18 parts according to the head poses, so angular range of each part is 10 degrees. We then take the mean shape of all the shapes in each part as the initial shape with the corresponding angular range. When dealing with a training data, its head pose is first determined and then mean shape of the corresponding angular range is taken as the initial shape. Our method can avoid bad initial shape caused by large pose variation and only need to maintain less shape space than CFSS's.

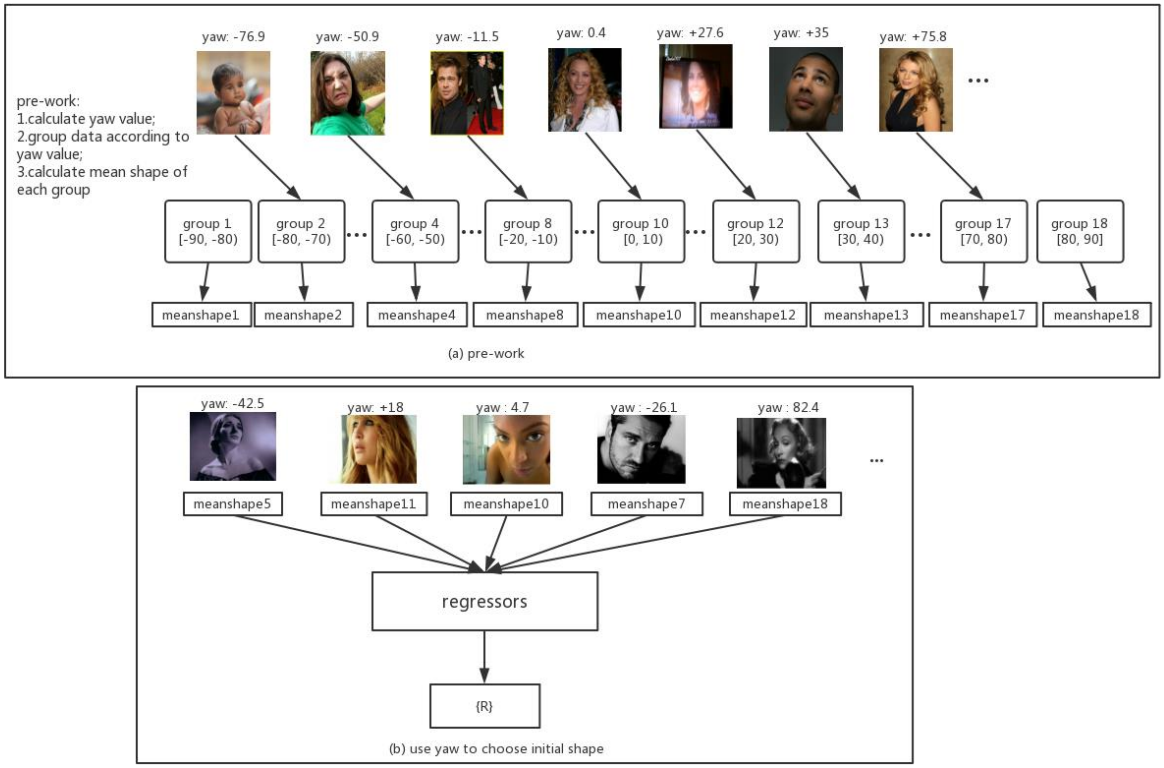


Figure 1. (a) a diagram illustrated the progress of classifying training data into 18 groups. (b) a diagram illustrate how to choose initial shape to train several regressors.

3. Supervised Descent Method based on Head Pose Estimation

Conventional SDM refines a shape via sequentially regressing local features indexed by the current estimated shape. In particular,

$$x_{k+1} = x_k + r_k \phi_k(I, x_k) + b_k \quad (1)$$

where the 2*n* dimensional shape vector $\boldsymbol{x_k}$ represents the current estimate of {*x*, *y*} coordinates of the *n* facial key points after the *k*th iteration. The local features indexed by the shape *x* on the face image *I* is denoted as $\phi_k(I, x_k)$. r_k is the *k*th learned descent direction, b_k is the *k*th learned bias, which are typically learned from a new linear regressor in the training set by minimizing

$$\arg\min_{r_k,\, b_k} \sum_{d^i,\ x_k^i} \left\| \Delta x_\cdot^{ki} - r_k \phi_k^i - b_k \right\|^2 \quad (2)$$

where $\Delta x_\cdot^{ki} = x_\cdot^i - x_k^i$, is the ground truth of *i*th facial key point and x_k^i is the current estimated *i*th facial key point, and ϕ_k^i is the feature of *i*th facial key point, d^i represents *i*th facial key points.

The estimation of conventional SDM can be easily trapped into the local optimal with poor initial shape. In this section, we will introduce the pose-estimation-based SDM to release this problem, More details about how to get suitable shape initialization is described as follows.

3.1. Data clustering based on head pose

Unlike the traditional SDM, an initialization optimization process is supplemented. As shown in the Figure.1, head poses of training data are first estimated and samples are clustered into *n* different groups according to estimated head poses.

[8] proposes head pose estimation method. Given the face region, this work constructs a model to estimate head pose, e.g. yaw and pitch, with HOG-based region descriptors as its input. We first use this model on data set to estimate samples' head poses.

In this paper, we classify training data by its yaw value, and we know that the yaw value is range from -90 degree to 90 degree (face from left to the right), as figure.1(a) shows. All the training data can be classified into 18 groups. In each group, angle range is 10 degree, e.g. in *i*th group, yaw angle range is {-90+*i**10, -90+*i**10-1|, *i* = 1,2,...,18}. After classifying all training data, we then calculate the mean shape of all the 18 groups.

3.2. Regressors training with initial shape selection

All training data is classified into 18 groups, according to their head pose. For each training sample, we choose the mean shape of its group as the initial shape. As we assume that with similar yaw value, their shape would be similar. Like conventional SDM, we use the equation 3 to learn r_0, b_0,

$$\Delta x_1 = r_0 \phi_0 + b_0 \quad (3)$$

where $\Delta x_1 = x_\cdot - x_0$, x_0 is the mean shape selected based on data's yaw angle, as Figure 1(b) shows. And ϕ_0 is feature vector of the data. Using training data to train the first linear regressor, we can get r_0, b_0. We then use the equation 1 to get the next initial shapes $x_1, x_2, \ldots, x_k, x_{k+1}$, and using equation 2 to learn a sequence of descent direction $\{r_k\}$ and bias $\{b_k\}$.

4. Experiments

Dataset Evaluations are performed on four widely used benchmark datasets. These datasets are challenging as images are mostly with large head pose variation.

300-W dataset: This dataset standardizes various alignment databases, include AFW, LFPW, HELEN, and challenging 135-image IBUG set, with 68-point annotation. For fair comparison, we regard all the training sample from LPFW, HELEN and the whole AFW as the training set (3148 images in total), and perform testing on three parts: the testing set from LFPW and HELEN, the 135-image IBUG set as the challenging set, and union of them as the full set (689 images in total).

HELEN dataset: it contains 2000 training and 330 testing samples. We conduct evaluation on 68/21 points.

LFPW dataset: it contains 1100 training and 300 testing samples. However due to some invalid URLs, we only employ 811 training and 224 testing samples. We perform evaluations on 68/21 points.

E**valuation** We use the standard landmarks mean error normalized by the inter-pupil distance to evaluate the alignment accuracy for each sample. We omit the '%' symbol for simplicity. The overall accuracy is reported based on either the averaged errors or cumulative errors distribution (CED) curve.

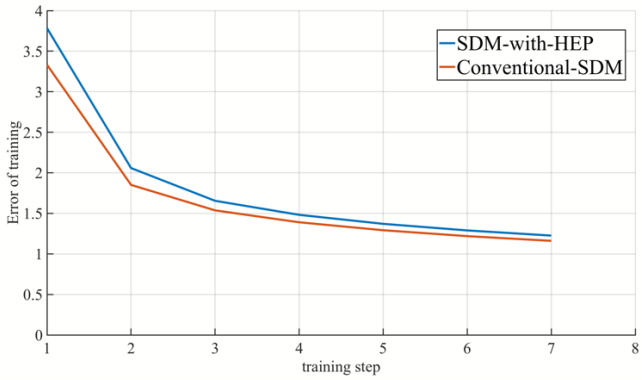


Figure 2. training error of SDM with HEP and conventional SDM.

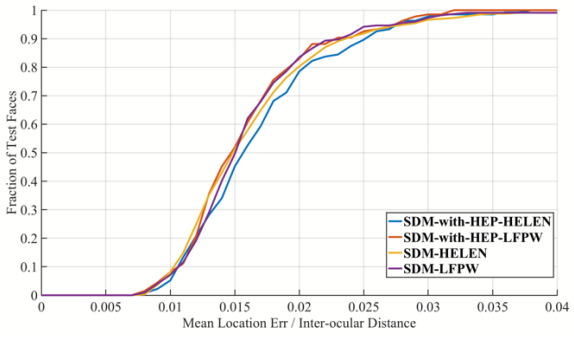
4.1. Comparison with conventional SDM

Training error comparison First, we use model from [8] to estimate head pose of images, with error 7.5 ± 7.28 in degree. We trained our model only using training set without external sources. Figure 2 shows the training error of SDM with HEP and conventional SDM. SDM with HEP starts with a little higher training error(3.78) and become nearly same training error(1.22) with conventional SDM after converge.

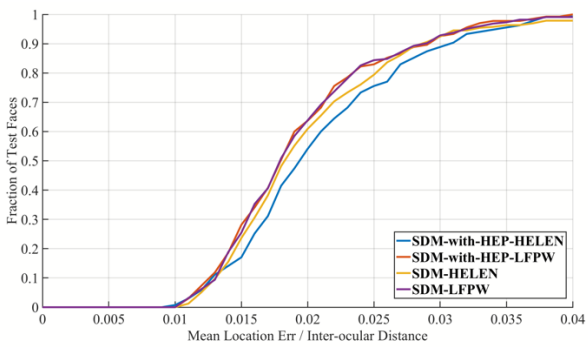
Table 1. comparison of averaged error between conventional SDM and SDM with HEP. Note that our method outperforms conventional SDM specially in challenging dataset IBUG.

Dataset	Conventional SDM		SDM with HEP	
	21-point	68-point	21-point	68-point
LFPW	1.62	1.95	1.60	1.92
HELEN	1.68	2.05	1.67	2.08
IBUG	5.72	6.08	2.73	3.06

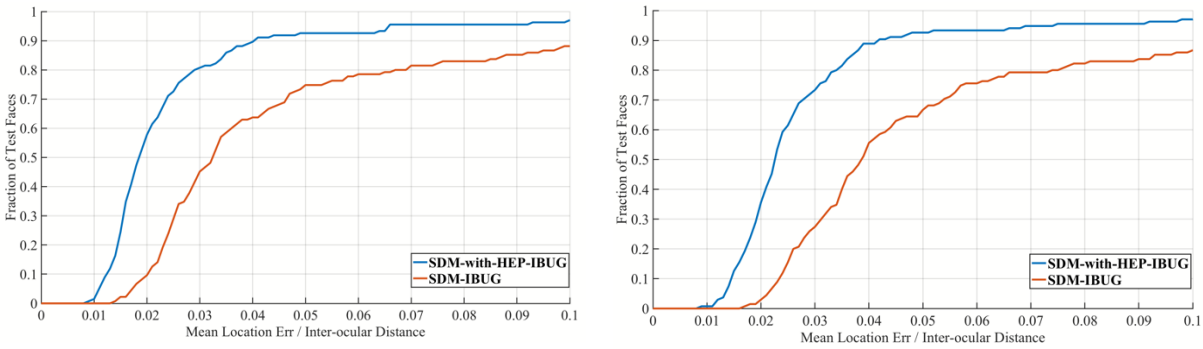
Averaged error comparison We summarize the comparison results in Table 1. We compare performance of conventional SDM and SDM with HEP on 21-point and 68-point evaluation respectively. It can be found that in LFPW and HELEN, SDM with HEP has a slightly improvement compared with conventional SDM, and in challenging testing set IBUG,SDM with HEP almost improves the accuracy by 2x over conventional SDM.



(a)



(b)



(c) (d)

Figure 3. Comparison of cumulative errors distribution (CED) curves.

The proposed method outperforms the conventional SDM. (a) CED for 21-point HELEN and LFPW test set, (b) CED for 68-point HELEN and LFPW test set, (c) CED for 21-point IBUG test set, (d) CED for 68-point IBUG test set.

Cumulative error distribution comparison To further compare the results with CED performance, we use the CED curves to evaluate the performance of conventional SDM and SDM with HEP on different dataset with 68-point and 21-point cases. The results are shown in Figure 3. Again, we can see that the proposed SDM with HEP is a slightly better on HELEN and LFPW test set. Especially, thanks to the head-pose-based initialization optimization, much better results have been achieved by our proposed method on IBUG test set.

5. Conclusion and future work

In the paper, we propose a novel face alignment method which clusters training samples into 18 different groups based on their head poses. Initial shapes are constructed based on face's head pose in the first cascade of SDM. As we assume that if different images with similar head pose angle, they would have similar shape. The experiment shows that the proposed method is better than conventional SDM in widely used datasets, especially in IBUG with large head pose variation images. In the future, we plan to achieve more accurate mean shape in the classified dataset group to improve the accuracy of training and testing.

References

[1] Chen, C., Dantcheva, A., & Ross, A. (2013, June). Automatic facial makeup detection with application in face recognition. In *2013 international conference on biometrics (ICB)* (pp. 1-8). IEEE.

[2] Ashraf, A. B., Lucey, S., Cohn, J. F., Chen, T., Ambadar, Z., Prkachin, K. M., & Solomon, P. E. (2009). The painful face–pain expression recognition using active appearance models. *Image and vision computing*, *27*(12), 1788-1796.

[3] Datta, A., Feris, R., & Vaquero, D. (2011, March). Hierarchical ranking of facial attributes. In *Face and Gesture 2011* (pp. 36-42). IEEE.

[4] Xiong, X., & De la Torre, F. (2013). Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 532-539).

[5] Mo, H., Liu, L., Zhu, W., Yin, S., & Wei, S. (2018). Face alignment with expression-and pose-based adaptive initialization. *IEEE Transactions on Multimedia*, *21*(4), 943-956.

[6] Zhu, S., Li, C., Change Loy, C., & Tang, X. (2015). Face alignment by coarse-to-fine shape searching. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4998-5006).

[7] Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., & Pantic, M. (2013). 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *Proceedings of the IEEE International Conference on Computer Vision Workshops* (pp. 397-403).
[8] Drouard, V., Ba, S., Evangelidis, G., Deleforge, A., & Horaud, R. (2015, September). Head pose estimation via probabilistic high-dimensional regression. In *2015 IEEE International Conference on Image Processing (ICIP)* (pp. 4624-4628). IEEE.