Face Alignment by Coarse-to-Fine Shape Searching

Shizhan Zhu[1,2], Cheng Li[2], Chen Change Loy[1,3], Xiaoou Tang[1,3]
[1]Department of Information Engineering, The Chinese University of Hong Kong. [2]SenseTime Group.
[3]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences.

Face alignment approaches typically include classic active appearance model [3], constrained local model [6], etc. Recently emerged regression based alignment methods [2, 4, 5, 7] have attracted great research interests and achieved state-of-the-art results. This type of method typically starts from a rough estimate of the shape (typically mean shape) and refine the shape by several iterations, as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + r_k(\phi(I;\mathbf{x}_k)), \tag{1}$$

where the $2n$ dimensional shape vector $\mathbf{x}_k$ represents the current estimate of (x,y) coordinates of the n landmarks after the k^{th} iteration. The local appearance patterns indexed by the shape $\mathbf{x}$ on the face image I is denoted as $\phi(I;\mathbf{x})$, and r_k is the k^{th} learned regressor. This purely discriminative regression method can encode explicit shape constraints in all steps and always predict reasonable results.

Despite its great discriminative power, the cascaded regression method still falls into limitations. One of the limitations comes from the initialisation nature of regressors. As demonstrated in our paper, the regressors tend to predict disparate shapes while given different initial shapes. This inconsistency is resulted from: *i)* shapes are predicted in an additive manner; and *ii)* features are indexed by the current estimated landmarks. Typically, cascaded regression applies the mean shape as the initial shapes of the first iteration since it produces the least initial error. However, if the target shape is far away from initial mean shape, the regressor might be caught in local optima, leading to inaccurate estimate. Our work aims to suppress such effects brought by the cascaded regressors, through switching from single-shape based regression to shape sub-region optimisation.

We divide the shape searching procedure into several stages, in which each stage is related to refining a shape sub-region. We form a $2n$ dimensional shape space, and denote N candidate shapes in the space as $\mathcal{S} = \{\mathbf{s}_1, \mathbf{s}_2, ..., \mathbf{s}_N\}$ ($N \gg 2n$). We denote a shape sub-region by two elements $\left(\bar{\mathbf{x}}_{(l)}, P^R_{(l)}\right)$, where $\bar{\mathbf{x}}_{(l)}$ denotes the center of the estimated shape sub-region, and $P^R_{(l)}$ represents the probability distribution that defines the scope of estimated sub-region around the center. Hence each searching stage contains two steps: searching for the sub-region center, and determine the probability of the candidate shapes (i.e. delineating the searching scope). While searching progresses through stages, the estimated sub-region tends to shrink its size and moving towards the target shape. Figure 1(a) serves as an illustration.

In each stage, we first determine the sub-region center $\bar{\mathbf{x}}_{(l)}$ when given the sub-region probability distribution $P^R_{(l-1)}$. We select candidate initial shapes according to the distribution $P^R_{(l-1)}$, and for the i^{th} sample we denote its j^{th} candidate as $\mathbf{x}_0^{ij}$. We then train K_l cascaded regressors based on this specific stage distribution, as

$$\begin{aligned} r_k &= \underset{r}{\operatorname{argmin}} \sum_{i=1}^{N} \sum_{j=1}^{N_l} \|\mathbf{x}^{i*} - \mathbf{x}_k^{ij} - r(\phi(I;\mathbf{x}_k^{ij}))\|_2^2 + \Phi(r), \\ \mathbf{x}_{k+1}^{ij} &= \mathbf{x}_k^{ij} + r_k(\phi(I;\mathbf{x}_k^{ij})) \qquad\qquad k = 0, \ldots, K_l - 1 \end{aligned} \tag{2}$$

where $\Phi(r)$ denotes the ℓ_2 regularisation term for each parameter in model r. During testing, we use the learned regressors for that stage to obtain the resulted shapes.To get the sub-region center based on these obtained resulted shapes, we apply dominant set approach to avoid the influence from the outliers. The goal of the this approach is to find a consistent shapes clique within all the resulted shapes and exclude outliers. Through replicator dynamics, we could obtain the estimated shape as the sub-region center.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.

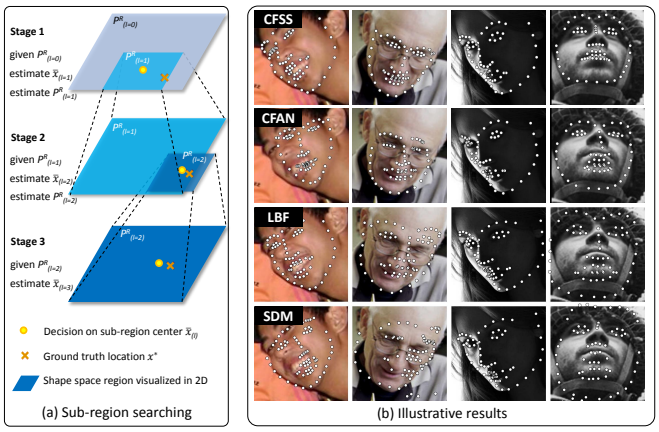

Figure 1: (a) The coarse-to-fine shape searching method for estimating the target shape. (b) Example images where the proposed CFSS outperforms CFAN [8], LBF [5] and SDM [7].

Next, we need to delineate the sub-region scope $P^R_{(l-1)}$ based on the estimated sub-region center $\bar{\mathbf{x}}_{(l)}$. We model the probabilistic distribution into two parts, as

$$P(\mathbf{s} - \bar{\mathbf{x}} | \phi(\bar{\mathbf{x}})) \propto P(\mathbf{s} - \bar{\mathbf{x}}) P(\phi(\bar{\mathbf{x}}) | \mathbf{s} - \bar{\mathbf{x}}), \tag{3}$$

where $\mathbf{s}$ represents the candidate shapes. The first part, $P(\mathbf{s} - \bar{\mathbf{x}})$ could be statistically learned based on the current error distribution between estimated sub-region center and ground-truth on all training samples. It approximately delineates the searching scope near $\phi(\bar{\mathbf{x}})$ and typically its distribution is more concentrated in later stages. For the second part $P(\phi(\bar{\mathbf{x}})|\mathbf{s} - \bar{\mathbf{x}})$, we divide it into different facial parts. This is because given the exemplar candidate shapes, the probability is conditionally independent for different facial parts [1]. The learned probabilistic distribution is used for sampling in next stage. Shape constraints are still strictly encoded throughout our method.

The proposed method achieves state-of-the-art results. Specifically, our method outperforms cascaded methods especially in cases with large pose variation. On the 300-W challenging dataset, we gain over 16% of error reduction compared to previous state-of-the-art. Illustrative examples can be found in Fig. 1(b) and our paper.

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

[2] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *IJCV*, 107(2):177–190, 2014.

[3] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *TPAMI*, 23(6):681–685, 2001.

[4] Vahid Kazemi and Sullivan Josephine. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, 2014.

[5] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, 2014.

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

[7] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013.

[8] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (CFAN) for real-time face alignment. In *ECCV*, pages 1–16. 2014.