Face Alignment using Cascade Gaussian Process Regression Trees

Donghoon Lee, Hyunsin Park, and Chang D. Yoo
Korea Advanced institute of Science and Technology

Face alignment is a task to locate fiducial facial landmark points, such as eye corners, nose tip, mouth corners, and chin, in a face image. Shape regression has become an accurate, robust, and fast framework for face alignment [2, 4, 5]. In shape regression, face shape $\mathbf{s} = (x_1, y_1, \cdots, x_p, y_p)^\top$, that is a concatenation of p facial landmark coordinates $\{(x_i, y_i)\}_{i=1}^{p}$, is initialized and iteratively updated through a cascade regression trees (CRT) as shown in Figure 1. Each tree estimates the shape increment from the current shape estimate, and the final shape estimate is given by a cumulated sum of the outputs of the trees to the initial estimate as follows:

$$\hat{\mathbf{s}}^T = \hat{\mathbf{s}}^0 + \sum_{t=1}^{T} f^t(\mathbf{x}^t; \theta^t), \tag{1}$$

where T is the number of stages, t is an index that denotes the stage, $\hat{\mathbf{s}}^t$ is a shape estimate, $\mathbf{x}^t$ is a feature vector that is extracted from an input image I, and $f^t(\cdot;\cdot)$ is a tree that is parameterized by θ^t. Starting from the rough initial shape estimate $\hat{\mathbf{s}}^0$, each stage iteratively updates the shape estimate by $\hat{\mathbf{s}}^t = \hat{\mathbf{s}}^{t-1} + f^t(\mathbf{x}^t; \theta^t)$.

The two key elements of CRT-based shape regression that impact to the prediction performance are gradient boosting [3] for learning the CRT and the shape-indexed features [2] which the trees are based. In gradient boosting, each stage iteratively fits training data in a greedy stage-wise manner by reducing the regression residuals that are defined as the differences between the ground truth shapes and shape estimates. The shape-indexed features are extracted from the pixel coordinates referenced by the shape estimate. The shape-indexed features are extremely cheap to compute and are robust against geometric variations.

Instead of using gradient boosting, we propose cascade Gaussian process regression trees (cGPRT) that can be incorporated as a learning method for a CRT prediction framework. The cGPRT is constructed by combining Gaussian process regression trees (GPRT) in a cascade stage-wise manner. Given training samples $\mathbf{S} = (\mathbf{s}_1, \cdots, \mathbf{s}_N)^\top$ and $\mathbf{X}^t = (\mathbf{x}_1, \cdots, \mathbf{x}_N)^\top$, GPRT models the relationship between inputs and outputs by a regression function $f(\mathbf{x})$ drawn from a Gaussian process with independent additive noise ε_i,

$$\begin{aligned} s_i &= f(\mathbf{x}_i) + \varepsilon_i, \quad i = 1, \cdots, N, && (2)\\ f(\mathbf{x}) &\sim \mathcal{GP}(0, k(\mathbf{x}, \mathbf{x}')), && (3)\\ \varepsilon_i &\sim \mathcal{N}(0, \sigma_n^2). && (4) \end{aligned}$$

A kernel $k(\mathbf{x}, \mathbf{x}')$ in GPRT is defined by a set of M number of trees:

$$\begin{aligned} k(\mathbf{x}, \mathbf{x}') &= \sigma_k^2 \sum_{m=1}^{M} \kappa^m(\mathbf{x}, \mathbf{x}'), && (5)\\ \kappa^m(\mathbf{x}, \mathbf{x}') &= \begin{cases} 1 & \text{if } \tau^m(\mathbf{x}) = \tau^m(\mathbf{x}') \\ 0 & \text{otherwise,} \end{cases} && (6) \end{aligned}$$

where σ_k^2 is the scaling parameter that represents the kernel power, and τ is a split function takes an input $\mathbf{x}$ and computes the leaf index $b \in \{1, \cdots, B\}$.

Given an input $\mathbf{x}_*$, distribution over its predictive variable $\mathbf{f}_*$ is given as

$$\bar{\mathbf{f}}_* = \sum_{i=1}^{N} \alpha_i k(\mathbf{x}_i, \mathbf{x}_*), \tag{7}$$

where $\alpha = (\alpha_1, \cdots, \alpha_N)^\top$ is given by $\mathbf{K}_s^{-1}\mathbf{S}$. Here, $\mathbf{K}_s$ is given by $\mathbf{K} + \sigma_n^2 \mathbf{I}_N$, and $\mathbf{K}$ is a covariance matrix of which $\mathbf{K}(i, j)$ is computed from the i-th and j-th row vector of $\mathbf{X}$. Computation of Equation (7) is in $O(N)$; however, this can be more efficient as follows:

$$\bar{\mathbf{f}}_* = \sum_{m=1}^{M} \bar{\alpha}^{m, \tau^m(\mathbf{x}_*)}, \tag{8}$$

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

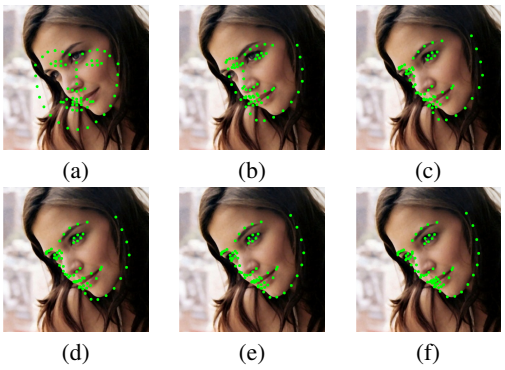

Figure 1: A selected prediction result on the 300-W dataset using cGPRT. The shape estimate is initialized and iteratively updated through a cascade of regression trees: (a) initial shape estimate, (b)–(f) shape estimates at different stages of cGPRT.

where $\bar{\alpha}^{m,b}$ is a predictive mean of the pseudo input that falls on leaf b of the m-th tree and does not fall on the other trees. The cGPRT is constructed by combining GPRTs in a stage-wise manner, and we propose a greedy stage-wise learning method for cGPRT and show that the prediction in cGPRT can be performed in the CRT framework.

Input features to cGPRT are designed through shape-indexed difference of Gaussian (DoG) features computed on local retinal patterns [1] referenced by shape estimates. The shape-indexed DoG features are extracted in three steps: (1) smoothing face images with Gaussian filters at various scales to reduce noise sensitivity, (2) extracting pixel values from Gaussian-smoothed face images indexed by local retinal sampling patterns, shape estimates, and smoothing scales, and (3) computing the differences of extracted pixel values. Smoothing scale of each local retinal sampling point is determined to be proportional to the distance between the sampling point and the center point. Thus, distant sampling points cover larger regions than nearby sampling points, and this leads to increasing stability of the distant sampling points against to shape estimate errors, while the nearby sampling points are more discriminative with an accurate shape estimate. In a learning procedure of cGPRT, this trade-off allows for each stage to select reliable features based on the current shape estimate errors.

In experiments on the 300-W dataset [6], the proposed cGPRT with shape-indexed DoG features achieves 5.71 mean error at 93 fps (accurate configuration) and 6.32 mean error at 871 fps (fast configuration) which are best performance compared with state-of-the-art methods.

[1] Alexandre Alahi, Raphael Ortiz, and Pierre Vandergheynst. Freak: Fast retina keypoint. In *CVPR*, pages 510–517. IEEE, 2012.

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

[3] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[4] Vahid Kazemi and Josephine Sullivan. One millisecond face alignment with an ensemble of regression trees. In *CVPR*, pages 1867–1874. IEEE, 2014.

[5] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *CVPR*, pages 1685–1692. IEEE, 2014.

[6] Christos Sagonas, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 300 faces in-the-wild challenge: The first facial landmark localization challenge. In *ICCVW*, pages 397–403. IEEE, 2013.