Joint Estimation of Pose and Face Landmark

Donghoon Lee, Junyoung Chung, Chang D. Yoo

Department of Electrical Engineering, KAIST, Daejeon, Korea

Abstract. This paper proposes a parallel joint boosting method that simultaneously estimates poses and face landmarks. The proposed method iteratively updates the poses and face landmarks through a cascade of parallel random ferns in a forward stage-wise manner. At each stage, the pose and face landmark estimates are updated: pose probabilities are updated based on previous face landmark estimates and face landmark estimates are updated based on previous pose probabilities. Both poses and face landmarks are simultaneously estimated through sharing parallel random ferns for the pose and face landmark estimations. This paper also proposes a triangular-indexed feature that references a pixel as a linear weighted sum of three chosen landmarks. This provides robustness against variations in scale, transition, and rotation. Compared with previous boosting methods, the proposed method reduces the face landmark error by 7.1% and 12.3% in the LFW and MultiPIE datasets, respectively, while it also achieves pose estimation accuracies of 78.6% and 94.0% in these datasets.

1 Introduction

Computer vision applications such as face recognition [29, 3, 28], facial expression recognition [8], age estimation [16, 17], and gaze estimation [24] are garnering significant attention, and achieving high performance in these applications remains difficult with variations in poses, expressions, and occlusions. In order to obtain more robustness, localizing the fiducial face landmark points is considered to be a key pre-processing step in many applications [29, 3, 28, 24]. However, accurate face landmark estimation itself is a difficult problem. Among the obstacles to obtaining accurate face landmarks, pose variations that involve 3D non-rigid deformation when projecting a 3D face on a 2D space is particularly difficult to manage, but it must be overcome for accurate estimation [9].

Most previous face landmark estimation methods do not use the pose information and, when they do, the pose is estimated prior to the face landmark estimation [9], which is referred to as the two-step method in this paper. The intuition behind the two-step method is that the face landmark depends on the pose, and a more precise location for the face landmark can be obtained using a pose-conditional landmark estimator. We observed that the face landmark information could also improve the accuracy of the pose estimate. In order to utilize this, the poses and face landmarks should be estimated simultaneously.

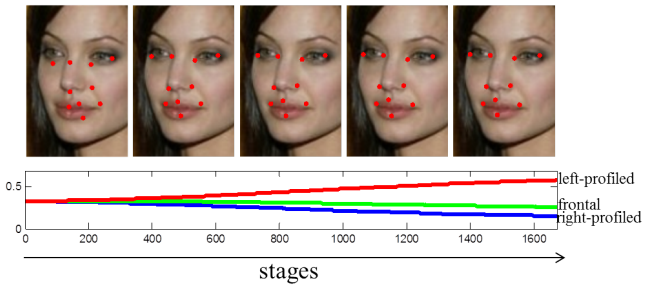

Fig. 1. Application results of parallel joint boosting to face landmarks and pose estimations in different stages. The pose probabilities and face landmark estimates are gradually updated with increases in the number of stages.

In this paper, a parallel joint boosting method to simultaneously estimate the pose and face landmarks is proposed. The proposed method has been motivated by gradient boosting [14] and LogitBoost [15], which have been used to estimate face landmarks [4] and poses, respectively. The proposed method combines gradient boosting and LogitBoost through a single weak regressor and then iteratively updates both estimates in a forward stage-wise manner as illustrated in Figure 1. In each stage, the face landmark estimates are updated based on the pose probability that is estimated in the previous stage, while the pose probabilities are updated based on the face landmarks obtained in the previous stage. Each stage consists of a number of weak regressors, and each weak regressor that assumes a particular pose is learned.

The remainder of this paper is organized as follows. Section 2 reviews selected related work, and the details of the proposed method are described in Section 3. The experimental and comparative results are reported in Section 4 and Section 5, respectively. The conclusions are presented in Section 6.

2 Related Work

Previous methods have attempted to manage large pose variations using appearance-based models [7, 5, 21, 20], morphable-model [2, 3], and template matching [25, 30] using parametric shape constraints. Although various strategies and improvements have been proposed over the past few decades, these methods have poor generalizability against unseen samples and suffer from slow training speeds. An alternative approach uses a part-based model [1, 23, 32] that considers the landmark estimation problem as a part detection problem.

Regression based methods [4, 6, 27] have been actively proposed where shape constraints are achieved in non-parametric manners, and they have exhibited promising results for both of accuracy and computation time.

Cao *et al.* [4] proposed a cascade regression method that uses random ferns as weak learners. The foundation of their method is representing the regressed

shape as a linear combination of all training shapes. It was demonstrated that the aggregation of random ferns results in a robust estimator with real-time operation.

Dantone *et al.* proposed a pose-inspired method with conditional random forests, where each conditional random forest is an expert regressor of each pose [9]. Their algorithm works in two-steps: estimate the head pose using regression forest and estimate the landmarks conditional to the head pose using conditional regression forests. The limitation of the two-step approach is that misclassification in poses cannot be managed well.

However, we found that poses and face landmarks should be estimated iteratively because they affect each other's improvement in accuracy. Furthermore, sharing parallel random ferns enables simultaneous estimation of poses and face landmarks in one structure.

3 Joint Estimation of Poses & Face Landmarks

This section provides a brief overview of two boosting methods: gradient boosting for face landmark estimation and LogitBoost for pose estimation. The proposed method for joint estimation of poses and face landmarks is also described.

3.1 Boosting methods

Gradient boosting. Gradient boosting provides a foundation for a number of face landmark estimation methods [4, 11]. Gradient boosting formulates the face landmark estimation problem as an additive cascade regression as follows:

$$\mathbf{s}^t = \mathbf{s}^{t-1} + r^t(\mathbf{I}; \alpha^t), \tag{1}$$

where $\mathbf{s}^t$ is the face landmark estimate at the t-th stage, $\mathbf{I}$ is an input image, and $r^t(\cdot;\cdot)$ is a weak regressor at the t-th stage parameterized by α^t. The final estimate at stage T is given by $\mathbf{s}^T = \mathbf{s}^0 + \sum_{t=1}^{T} r^t(\mathbf{I}; \alpha^t)$.

Given training samples, $\{\hat{\mathbf{s}}_i, \mathbf{I}_i\}_{i=1}^N$, and the regressor parameters, $\{\alpha^t\}_{t=1}^T$, are learned through minimizing the empirical loss, $\Psi(\cdot,\cdot)$, in a greedy forward stage-wise manner, as follows:

$$\alpha^t = \underset{\alpha^t}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \Psi(\hat{\mathbf{s}}_i, \mathbf{s}_i^t), \tag{2}$$

$$= \underset{\alpha^t}{\operatorname{argmin}} \frac{1}{N} \sum_{i=1}^{N} \Psi(\hat{\mathbf{s}}_i, \mathbf{s}_i^{t-1} + r^t(\mathbf{I}_i; \alpha^t)). \tag{3}$$

A reasonable choice of loss function, $\Psi(\cdot,\cdot)$, for the face landmark estimation is:

$$\Psi(\hat{\mathbf{s}}, \mathbf{s}) = ||\hat{\mathbf{s}} - \mathbf{s}||, \tag{4}$$

which is typically considered to be a performance measure.

In this paper, a random fern with L split functions, $\{f_l\}_{l=1}^{L}$, is considered to be a weak regressor, and it partitions the feature space into 2^L disjoint bins, $\{\mathbf{R}_b\}_{b=1}^{2^L}$. Equation 3 is reduced to:

$$\delta\bar{\mathbf{s}}_b = \underset{\delta\bar{\mathbf{s}}_b}{\operatorname{argmin}} \frac{1}{N}\sum_{i=1}^{N} ||\delta\mathbf{s}_i - r^t(\mathbf{I}_i; f_l, \delta\bar{\mathbf{s}}_b)||, \tag{5}$$

$$= \frac{1}{1+\beta/\sum_{i=1}^{N}\mathbf{1}_{R_b}(f(\mathbf{I}_i))} \frac{\sum_{i=1}^{N}\delta\mathbf{s}_i\mathbf{1}_{R_b}(f(\mathbf{I}_i))}{\sum_{i=1}^{N}\mathbf{1}_{R_b}(f(\mathbf{I}_i))}. \tag{6}$$

Here, $\alpha = \{\{f_l\}_{l=1}^{L}, \{\delta\bar{\mathbf{s}}_b\}_{b=1}^{2^L}\}$ and $\{\delta\bar{\mathbf{s}}_b\}_{b=1}^{2^L}$ are bin outputs, $\mathbf{1}_R(\cdot)$ is a binary indicator function, $\delta\mathbf{s} = ||\hat{\mathbf{s}} - \mathbf{s}||$, and β is a shrinkage parameter.

LogitBoost. LogitBoost [15] is a statistical boosting method for classification and has been used for pose estimations [31]. Given J number of pose classes, LogitBoost considers the following relationships between the pose probability, π_j, and the logistic function, H_j, as follows:

$$\pi_j = \frac{e^{H_j}}{\sum_{i=1}^{J} e^{H_i}}, \tag{7}$$

$$H_j = \log\pi_j - \frac{1}{J}\sum_{i=1}^{J}\log\pi_i. \tag{8}$$

Here, H_j has a cascade regression form similar to gradient boosting, which is given below:

$$H_j^t = H_j^{t-1} + h_j^t(\mathbf{I}; \beta^t), \tag{9}$$

which gives $H_j^T = H_j^0 + \sum_{t=1}^{T} h^t(\mathbf{I}; \beta^t)$ at the T-th stage. At each stage, LogitBoost fits an individual weak regressor, h_j^t, for each pose using the weighted least-squares of $z_j = \frac{y_j - \pi_j}{w_j}$ to feature with the weights $w_j = \pi_j(1-\pi_j)$. Here, y_j is a binary indicator that indicates the j-th pose.

The random fern is usually used as a weak regressor, and the solution to the weighted least-squares regression problem using the random fern is given as follows:

$$\delta\bar{h}_{b,j} = \alpha\frac{\sum_{i=1}^{N} w_{i,j}z_{i,j}\mathbf{1}_{R_b}(f(\mathbf{I}_i))}{\sum_{i=1}^{N} w_{i,j}\mathbf{1}_{R_b}(f(\mathbf{I}_i))}. \tag{10}$$

Here, $\{\delta\bar{h}_{b,j}\}_{b,j=1}^{2^L,J}$ are bin outputs and α is a shrinkage parameter. The shrinkage process based on β is omitted (refer to Equation 6)

For more details about LogitBoost, we suggest that readers refer to [15].

Algorithm 1 Joint boosting evaluation

1: **Input:** Image $\mathbf{I}$, weak regressors $\{f^t, \delta\bar{\mathbf{s}}^t, \delta\bar{\mathbf{h}}^t\}_{t=1}^T$
2: Initialize $\mathbf{s}^0$, H_j^0
3: **for** $t = 1$ to T **do**
4: $\quad b \leftarrow f^t(\mathbf{I})$ ▷ *Compute bin index*
5: $\quad \mathbf{s}^t \leftarrow \mathbf{s}^{t-1} + \delta\bar{\mathbf{s}}_b^t$
6: $\quad H_j^t \leftarrow H_j^{t-1} + \delta\bar{h}_{b,j}^t$ for $\forall j$
7: **end for**
8: $\pi_j^T \leftarrow \frac{e^{H_j^T}}{\sum_{j=1}^J e^{H_j^T}}$, for $\forall j$
9: **Output:** Face landmark estimates $\mathbf{s}^T$, pose probability π_j^T

3.2 Parallel joint boosting

Assuming that poses and face landmarks are closely coupled, we conjecture that the following procedures should be considered in order to improve both estimates: (1) estimation of face landmarks should use the pose information and (2) estimations of the poses should use face landmark information. Procedures (1) and (2) should be conducted in an iterative manner, such that both estimates are updated at each iteration. The proposed method includes both procedures. The details of the proposed method are described in the following.

Joint boosting. In the joint boosting method, gradient boosting and LogitBoost are combined through a single set of random ferns. When one set of random ferns to implement the gradient boosting and a separate set of random ferns to implement LogitBoost share a common split function, then the gradient boosting and LogitBoost can be combined using a single set of random ferns. Gradient boosting and LogitBoost, which are based on random ferns, have been proposed in the past for face landmark estimations [4] and pose estimations [31], respectively. However, neither has been used simultaneously or jointly.

Algorithm 1 describes an evaluation procedure using the joint boosting. The random fern of the joint boosting is parameterized by $\{f^t, \delta\bar{\mathbf{s}}^t, \delta\bar{h}^t\}_{t=1}^T$, which are the split functions, gradient boosting bin outputs, and LogitBoost bin outputs, respectively. Gradient boosting for face landmark estimations and LogitBoost for pose estimations share a common split function, f, and distinguish themselves through separate bin outputs. Consequently, the poses and face landmarks can be simultaneously estimated through carefully selecting the split function. Note that the bin outputs, $\delta\bar{\mathbf{s}}$ and $\delta\bar{h}$, can be obtained using Equations 6 and 10, respectively.

We adopted landmark-indexed features[1] [4, 13] and simple decision stumps in order to design the split functions. Landmark-indexed features have been successfully applied to both pose [13] and landmark [4] estimations. At each stage,

[1] Landmark-indexed features are also known as shape-indexed features [4] and pose-indexed features [13]. We use the term landmark-indexed feature for consistency in this paper.

Algorithm 2 Parallel joint boosting evaluation: soft

Input: Image $\mathbf{I}$, weak regressors $\{f^{t|k}, \delta\bar{\mathbf{s}}^{t|k}, \delta\bar{\mathbf{h}}^{t|k}\}_{t,k=1}^{T,J}$
Initialize $\mathbf{s}^0$, H_j^0
for $t = 1$ to T **do**
  for $k = 1$ to J **do**
    $b \leftarrow f^{t|k}(\mathbf{I})$ ▷ *Compute bin index*
    $\delta\mathbf{s}^{t|k} \leftarrow \delta\bar{\mathbf{s}}_b^{t|k}$
    $H_j^{t|k} \leftarrow H_k^{t-1} + \delta\bar{h}_{b,j}^{t|k}$, for $\forall j$
    $\pi_j^{t|k} \leftarrow \frac{e^{H_j^{t|k}}}{\sum_{j=1}^{J} e^{H_j^{t|k}}}$, for $\forall j$
  end for
  $\mathbf{s}^t \leftarrow \mathbf{s}^{t-1} + \sum_{k=1}^{J} \pi_k^{t-1}\delta\mathbf{s}^{t|k}$
  $\pi_j^t \leftarrow \sum_{k=1}^{J} \pi_k^{t-1}\pi_j^{t|k}$, for $\forall j$
  $H_j^t \leftarrow \log \pi_j^t - \frac{1}{J}\sum_{j=1}^{J} \log {\pi_j}^t$, for $\forall j$
end for
$\pi_j^T \leftarrow \frac{e^{H_j^T}}{\sum_{j=1}^{J} e^{H_j^T}}$, for $\forall j$
Output: Face landmark estimates $\mathbf{s}^T$, pose probability π_j^T

the landmark-indexed features are extracted based on previous face landmark estimates, and previous face landmark estimates are used for the pose estimations. The details of the landmark-indexed features that we used are described in Section 3.3.

Parallel expansion. The intuition behind the parallel joint boosting is to use the pose probabilities in the previous stage to improve the accuracies of the face landmark estimates. In the parallel joint boosting, each stage consists of J number of parallel random ferns, and each random fern assumes a particular pose and is pose-conditionally learned.

We model face landmarks, $\mathbf{s}$, as a mixture of J pose-conditional landmarks, $\{s^{|k}\}_{k=1}^{J}$, with the pose-probability weights, π^k, as follows:

$$\mathbf{s} = \sum_{k=1}^{J} \pi^k \mathbf{s}^{|k}. \tag{11}$$

The parallel joint boosting consists of $T \times J$ weak regressors with the parameters, $\{f^{t|k}, \delta\bar{\mathbf{s}}^{t|k}, \delta\bar{h}^{t|k}\}_{t,k=1}^{T,J}$, and are formulated based on the mixture model.

The overall procedure of the parallel joint boosting is described in Algorithm 2, and the details of each stage of the parallel joint boosting are described here.

1. The face landmark estimates, $\mathbf{s}^t$, are updated based on the previous face landmark estimates, $\mathbf{s}^{t-1}$, J number of pose-conditional face landmark updates, $\delta\bar{\mathbf{s}}^{t|k}$, and corresponding pose probabilities, π_k^{t-1}, in the previous stage using the equation: $\mathbf{s}^t = \mathbf{s}^{t-1} + \sum_{k=1}^{J} \pi_k^{t-1}\delta\bar{\mathbf{s}}^{t|k}$.

Algorithm 3 Parallel joint boosting evaluation: hard

Input: Image $\mathbf{I}$, weak regressors $\{f^{t|k}, \delta\bar{\mathbf{s}}^{t|k}, \delta\bar{\mathbf{h}}^{t|k}\}_{t,k=1}^{T,J}$
Initialize $\mathbf{s}^0$, H_j^0
for $t = 1$ to T **do**
  $k \leftarrow \text{argmax}_k\, \pi_k^{t-1}$
  $b \leftarrow f^{t|k}(\mathbf{I})$ ▷ *Compute bin index*
  $\mathbf{s}^t \leftarrow \mathbf{s}^{t-1} + \pi_k^{t-1}\delta\bar{\mathbf{s}}_b^{t|k}$
  $H_j^t \leftarrow H_j^{t-1} + \delta\bar{h}_{b,j}^{t|k}$, for $\forall j$
end for
$\pi_j^T \leftarrow \frac{e^{H_j^T}}{\sum_{j=1}^{J} e^{H_j^T}}$, for $\forall j$
Output: Face landmark estimates $\mathbf{s}^T$, pose probability π_j^T

2. The pose probabilities are simultaneously updated using the equation: $\pi_j^t = \sum_{k=1}^{J} \pi_k^{t-1}\pi_j^{t|k}$. Here, $\pi_j^{t|k}$ is obtained using the relationship between the pose probability and logistic function, $H_j^{t|k}$, given in Equation 7. $H_j^{t|k}$ can be computed using the equation, $H_j^{t|k} = H_j^{t-1} + \delta\bar{h}_{b,j}^{t|k}$.

This method is called the "soft" decision method, and it is distinguished from the "hard" decision method as follows.

The hard decision method is described in Algorithm 3. The computation cost in the evaluation procedure of the soft decision method lineally increases with J. The hard decision method updates the posse and face landmarks in a greedy manner, while the soft decision method processes all random ferns even when the associated probability is close to zero. Through choosing the most probable pose and corresponding pose-conditional random fern at every stage, the computational cost in the evaluation procedure is irrelevant to J.

The bin outputs for the face landmark estimations of the pose-conditional weak regressor can be obtained through solving the minimization problem weighted by the pose-probability estimates that were obtained in the previous stage, π_j^{t-1}, and these are reduced to:

$$\delta\bar{\mathbf{s}}_b^{t|k} = \frac{\sum_{i=1}^{N} \pi_k^{t-1}\delta\mathbf{s}_i^t \mathbf{1}_{R_b^{t|k}}(f^{t|k}(\mathbf{I}_i))}{\sum_{i=1}^{N} \pi_k^{t-1}\mathbf{1}_{R_b^{t|k}}(f^{t|k}(\mathbf{I}_i))}. \tag{12}$$

Here, the shrinkage process based on β is omitted (refer to Equation 6).

The bin outputs for the pose estimations can be obtained through applying Equation 10 to all parallel ferns.

$$\delta\bar{h}_j^{t|k} = \alpha\frac{\sum_{i=1}^{N} \pi_k^{t-1} w_{i,j} z_{i,j} \mathbf{1}_{R_b}(f^{t|k}(\mathbf{I}_i))}{\sum_{i=1}^{N} \pi_k^{t-1} w_{i,j}\mathbf{1}_{R_b}(f^{t|k}(\mathbf{I}_i))}. \tag{13}$$

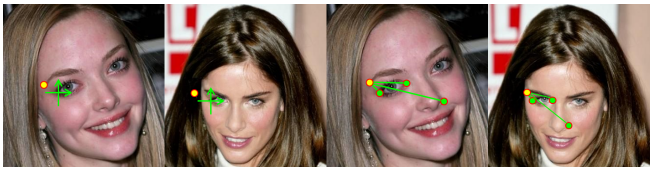

Fig. 2. Triangular-indexed features reference a point with a linear combination of randomly generated weight vectors that are constrained to $\sum_i w_i = 1$. Any point in the face region can be represented invariant to shape deformation and scaling.

3.3 Triangular-indexed features

Regression-based methods typically require a high number of iterations for accurate landmark estimations; thus, for real-time operations, the update should be based on features that require low computational costs, such as the pixel intensity difference between two points as used in [11, 22]. In order to gain geometric invariance, local coordinates are used to index a pixel point through determining its local coordinates references from the closest landmark [4]. This feature is invariant to scale variations face transitions, and its efficacy has been proven through achieving state-of-the-art performance in real-time operations [4]. However, this feature is limited by the large pose variations in yaw and roll axes that deform the pixel: single reference landmarks cannot counter pixel displacements due to large pose variations. This feature requires regular similarity transformations to the mean shape in the regression, which can hinder real-time operations. In order to overcome this problem, Cao *et al.* [4] used a two-step framework to obtain features that are robust to large pose variations.

In the proposed triangular-indexed feature, the pixel point is indexed as a linear weighted sum of three landmarks what form a triangular mesh on the face. The linear weighted sum of the arbitrarily chosen landmarks can represent almost every point in the face and it is invariant to large pose estimations. The triangular-indexed features do not require similarity transformations; hence, its computation is cheap. Triangular mesh templates can be generated through selecting three landmarks manually or randomly offline that will used throughout the iteration. Then, a weight vector, $\mathbf{w} \in \mathbb{R}^3$, is randomly sampled in the following manner: sample $w_1 \sim U(0,1)$ where $U(0,1)$ is the uniform distribution; and sample $w_2 \sim U(0,1)$ and $w_3 \sim U(0,1)$ such that $E[w_2] = E[w_3] = 0$; and then, normalize the weights to make the sum to 1. This procedure can be interpreted as randomly selecting a pixel point nearby landmark $\mathbf{l}_1$ in order to bind the location inside the face region and to index the coordinates to three landmarks including $\mathbf{l}_2$ and $\mathbf{l}_3$ to gain geometric invariance. Figure 2 depicts a triangular-indexed feature compared with [4] shown in the images on the left. Finally, a pixel point, $\mathbf{p} \in \mathbb{R}^2$, is generated through estimating $\mathbf{p} = \sum_i w_i \mathbf{l}_i$. In contrast to [4], we considered the pose probabilities, $\pi|_{j=1}^{J}$, as weights and computed the weighted correlation with $\delta\mathbf{s}$. The computational complexity was reduced from

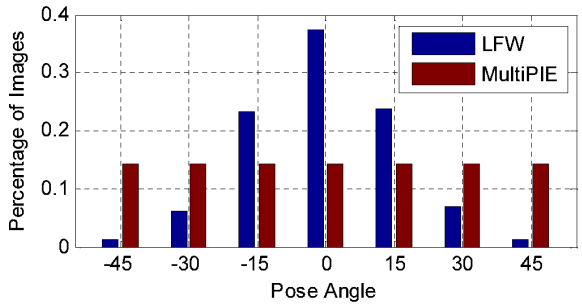

Fig. 3. Distributions of the poses in the LFW and MultiPIE. LFW consists of mostly frontal images, while MultiPIE consists of images in various poses.

$O(P^2)$ to $O(P)$ through adopting the fast correlation computation introduced in [4].

4 Experiments

The objectives of our experiments were two-fold: to compare the proposed method with the state-of-the-art methods and to demonstrate the efficacy of the joint estimation of poses and face landmarks. We conducted experiments on two benchmark datasets: Labeled Faces in the Wild (LFW) [19] and MultiPIE [18]. Our experiments focused on the evaluation of face landmark estimations because our key interest is in face landmark estimations.

4.1 Datasets

LFW. LFW contains 13,233 face images of 5,749 people and is remarkably challenging due to its constraint of the images of LFW being detected using the Viola-Jones face detector [26]. There were no pose annotations for the original LFW; therefore, we used the POSIT method [10] to obtain the approximate pose annotations, and they were quantized into three poses: left-profile, frontal, and right-profile.

MultiPIE. MultiPIE contains approximately 750,000 face images of 337 people with varying viewpoints, illumination conditions, and facial expressions. We considered 250 people collected from Session 1 with varying poses from $-45°$ to $45°$ with $15°$ intervals, under 19 illumination conditions and two facial expressions.

4.2 Implementation details

The benchmark methods [9, 27] often use the Viola Jones face detector to locate the face position. However, the detection often fails on profiled faces in Multi-PIE. Therefore, we simulated the output of the face detector through randomly providing a bounding box that overlaps a minimum of 80% of the ground truth.

Method	Error	Method	Error
Dantone [9]	0.0696	Cao-Ti	0.0563
Human [9]	0.0597	Soft-Li	0.0584
Ever. [12]	0.0963	Soft-Ti	**0.0552**
Yang [27]	0.0645	Hard-Li	0.0589
Cao-Li[4]	0.0594	Hard-Ti	**0.0552**

Table 1. Average error for landmark estimation on LFW.

Because the code used in [4] was not distributed to publicly, we developed and implemented it ourselves. The parameters of our implementation were set to be the same as [4], except the number of initial shapes for training data augmentation. We used 5 and 2 instead of 20 for the LFW and MultiPIE datasets, respectively, in order to adjust the number of training samples. In more detail, we set the number of stages to $T = 10$ and $K = 500$, the number of features to $P = 400$, the depth of the random fern to $F = 5$, and the shrinkage parameter to $\beta = 1000$.

In order to implement the parallel joint boosting, we adopted the two-stage cascade method proposed in [4], and we set the parameters to be the same as [4]. For the soft decision method, K was adaptively chosen as 166 and 71 for the LFW and MultiPIE, respectively, in order to to adjust the number of processed weak regressors in the evaluation. α was set to 0.005, and we manually designed 40 and 20 templates for the triangular-indexed features for LFW and MultiPIE, respectively.

5 Results

For the face landmark estimations, we measured the estimation errors as a fraction of inter-ocular distance, which is the distance between the ground truth and estimation normalized using the inter-ocular distance. For the pose estimations, the classification accuracies were reported. We performed five-fold cross validations, and we report the mean accuracy in both datasets.

5.1 Comparison using LFW

We compared the proposed methods with the following methods: Dantone *et al.* [9], human manual annotation [9], Everingham *et al.* [12], Yang and Patras [27], and Cao *et al.* [4]. Furthermore, we employed the results reported in [9]. For [27], we used the results of Figure 5 in [27], and [4] was implemented ourselves.

In order to evaluate the impact of the decision methods (soft or hard) and the triangular-indexed feature, we conducted experiments on all possible combinations: Cao-Li [4], Cao-Ti, Soft-Li, Soft-Ti, Hard-Li, and Hard-Ti. Here, Li and Ti indicate the landmark-indexed feature and triangular-indexed feature, respectively.

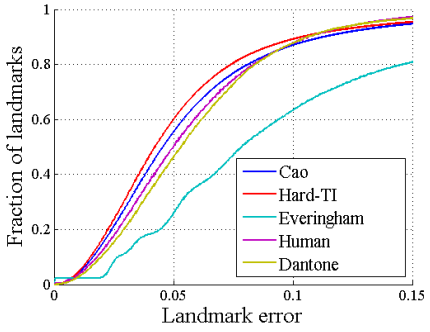

Fig. 4. Comparison between the benchmark methods and the proposed methods for the LFW.

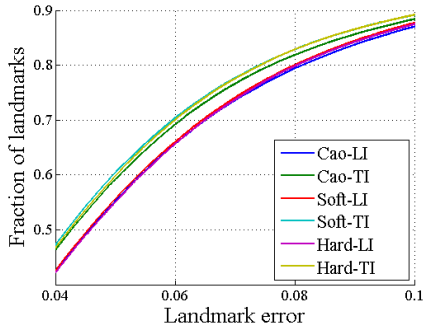

Fig. 5. Comparison with Cao *et al.* [4] and the proposed methods for the LFW.

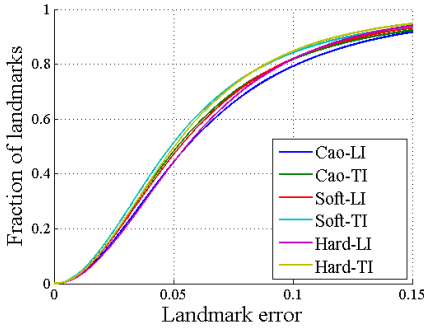

Fig. 6. Comparison with Cao *et al.* [4] and the proposed methods for the MultiPIE.

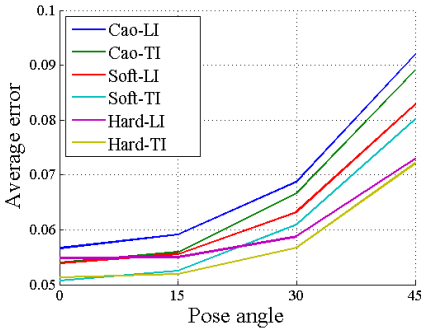

Fig. 7. Average error for the pose estimations (0°, 15°, 30°, and 45°) for the MultiPIE.

Table 1 and Figure 4 present the comparisons between the state-of-the-art methods and proposed methods. The proposed methods achieved remarkable performances and reduced the average error from the best current method [4] by 7% with 78.6% pose estimation accuracy. Surprisingly, both [4] and the proposed method outperformed the performance of human manual annotation performance.

Table 1 and Figure 5 illustrate the detailed comparisons with Cao *et al.* [4] and the proposed methods. The parallel joint boosting method was insignificant on the LFW. This resulted from the LFW consisting of nearly frontal images as illustrated in Figure 3. The triangular-indexed feature consistently improved the performance compared with the landmark-indexed feature [4].

Landmarks	Cao-Li[4]	Cao-Ti	Soft-Li	Soft-Ti	Hard-Li	Hard-Ti
Eyes	0.0547	0.0593	**0.0515**	0.0546	0.0526	0.0540
Nose	0.0859	0.0760	0.0794	0.0710	0.0787	**0.0703**
Mouth	0.0702	0.0647	0.0660	0.0605	0.0654	**0.0594**
Chin	0.0953	0.0922	0.0902	0.0880	0.0905	**0.0855**
Average	0.0715	0.0682	0.0671	**0.0638**	0.0670	**0.0627**

Table 2. Errors for each face landmark point and the average errors estimated on MultiPIE.

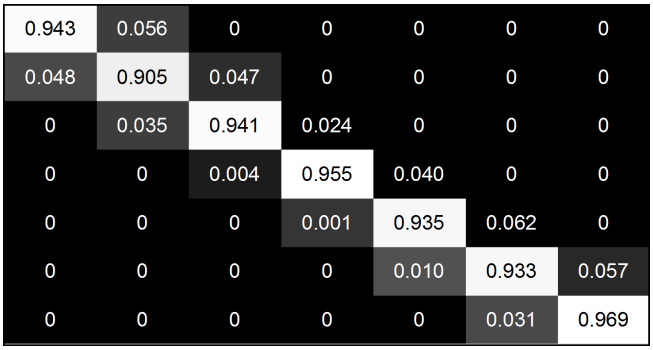

Fig. 8. The confusion matrix of the pose estimations for MultiPIE using Hard-Ti (values ≤ 0.001 have been omitted).

5.2 Comparison using MultiPIE

The primary objective of the experiments using MultiPIE was to verify the effectiveness of the joint estimation of the poses and face landmarks. We compared the landmark estimation results of the previous gradient boosting method from [4] and the proposed method.

The face landmark estimation results for MultiPIE are illustrated in Figure 6. The proposed hard and soft decision methods clearly outperformed Cao *et al.* [4] for both using the landmark-indexed features and the triangular-indexed features. Table 2 presents the errors for each face landmark point and the average error. The Hard-Ti method achieved the best performance for the nose, mouth, and chin, and it also achieved the minimum average error. The Hard-Ti method reduced the average error by 12.3% compared with Cao-Li. When the feature was fixed to a landmark-indexed feature, the soft and hard decision methods reduced the average error by 6.2% and 6.3%, respectively. Using the triangular-indexed features for the Cao, Soft, and Hard methods reduced the average error by 4.6%, 4.9%, and 6.4%, respectively. The most difficult point to estimate was the chin (0.0855 was the best result).

Figure 7 presents the average error for various poses. The Hard-Li and Hard-Ti methods exhibited more smooth curves compared with the Cao-Li and Cao-Ti

Fig. 9. Qualitative results for the LFW (top three rows) and MultiPIE (bottom three rows) datasets.

methods, which could be interpreted as the hard decision method being more robust to pose variations.

The confusion matrix of the pose estimations for the Hard-Ti method is depicted in Table 8. The average accuracy recorded was 94.6%. Although we did not compare this performance with the other pose estimation methods because our primary focus was on face landmark estimations, the proposed method achieved reliable accuracy.

6 Conclusion

We proposed a parallel boosted regression method for simultaneous estimation of poses and face landmarks. The proposed method enables the estimation of both poses and face landmarks simultaneously, and the method improved both estimations based on pose-conditional random ferns and triangular-indexed features. Experiments using the LFW database demonstrated that the proposed

method achieves high performance in face landmark estimation, even better than the performance of human manual annotations. The results from the MultiPIE database demonstrated that the proposed model improves the performance of face landmark estimations in large pose variations and sufficiently supports our intuitive idea. The pose estimation results have also demonstrated reliable accuracy.

Acknowledgement. This work was supported by ICT R&D program of MSIP/IITP [14-824-09-014, Basic Software Research in Human-level Lifelong Maching Learning (Machine Learning Center)].

References

1. Belhumeur, P., Jacobs, D., Kriegman, D., Kumar, N.: Localizing parts of faces using a consensus of exemplars. In: CVPR. (2011)
2. Blanz, V., Vetter, T.: A morphable model for the synthesis of 3d faces. In: Proceedings of the 26th annual conference on Computer graphics and interactive techniques, ACM (1999) 187–194
3. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. TPAMI **25** (2003) 1063–1074
4. Cao, X., Wei, Y., Wen, F., Sun, J.: Face alignment by explicit shape regression. IJCV **107** (2014) 177–190
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. TPAMI **23** (2001) 681–685
6. Cootes, T.F., Ionita, M.C., Lindner, C., Sauer, P.: Robust and accurate shape model fitting using random forest regression voting. In: ECCV. (2012)
7. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Computer Vision and Image Understanding **61** (1995) 38–59
8. Chew, S.W., Lucey, S., Lucey, P., Sridharan, S., Conn, J.F.: Improved facial expression recognition via uni-hyperplane classification. In: CVPR. (2012)
9. Dantone, M., Gall, J., Fanelli, G., Van Gool, L.: Real-time facial feature detection using conditional regression forests. In: CVPR. (2012)
10. Dementhon, D.F., Davis, L.S.: Model-based object pose in 25 lines of code. IJCV **15** (1995) 123–141
11. Dollár, P., Welinder, P., Perona, P.: Cascaded pose regression. In: CVPR. (2010)
12. Everingham, M., Sivic, J., Zisserman, A.: Hello! my name is... buffy–automatic naming of characters in tv video. In: BMVC. (2006)
13. Fleuret, F., Geman, D.: Stationary features and cat detection. Journal of Machine Learning Research **9** (2008) 1437
14. Friedman, J.: Greedy function approximation: a gradient boosting machine. The Annals of Statistics (2001) 1189–1232
15. Friedman, J., Hastie, T., Tibshirani, R.: Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). The Annals of Statistics **28** (2000) 337–407
16. Geng, X., Smith-Miles, K., Zhou, Z.H.: Facial age estimation by learning from label distributions. In: AAAI. (2010)

17. Geng, X., Zhou, Z.H., Zhang, Y., Li, G., Dai, H.: Learning from facial aging patterns for automatic age estimation. In: Proceedings of the 14th annual ACM international conference on Multimedia, ACM (2006) 307–316
18. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28** (2010) 807–813
19. Huang, G., Mattar, M., Berg, T., Learned-Miller, E., et al.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: ECCV workshop. (2008)
20. Matthews, I., Baker, S.: Active appearance models revisited. IJCV **60** (2004) 135–164
21. Milborrow, S., Nicolls, F.: Locating facial features with an extended active shape model. In: ECCV. (2008)
22. Ozuysal, M., Calonder, M., Lepetit, V., Fua, P.: Fast keypoint recognition using random ferns. TPAMI **32** (2010) 448–461
23. Smith, B.M., Zhang, L., Brandt, J., Lin, Z., Yang, J.: Exemplar-based face parsing. In: CVPR. (2013)
24. Sugano, Y., Matsushita, Y., Sato, Y., Koike, H.: An incremental learning method for unconstrained gaze estimation. In: ECCV. (2008)
25. Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: Robust and efficient parametric face alignment. In: ICCV. (2011)
26. Viola, P., Jones, M.J.: Robust real-time face detection. IJCV **57** (2004) 137–154
27. Yang, H., Patras, I.: Sieving regression forest votes for facial feature detection in the wild, ICCV (2013)
28. Yi, D., Lei, Z., Li, S.Z.: Towards pose robust face recognition. In: CVPR. (2013)
29. Yin, Q., Tang, X., Sun, J.: An associate-predict model for face recognition. In: CVPR. (2011)
30. Yuille, A.L., Hallinan, P.W., Cohen, D.S.: Feature extraction from faces using deformable templates. IJCV **8** (1992) 99–111
31. Zhang, J., Zhou, S.K., McMillan, L., Comaniciu, D.: Joint real-time object detection and pose estimation using probabilistic boosting network. In: CVPR. (2007)
32. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: CVPR, IEEE (2012) 2879–2886