Unified Face Analysis by Iterative Multi-Output Random Forests

Xiaowei Zhao Tae-Kyun Kim Wenhan Luo
Department of EEE, Imperial College London, United Kingdom
{x.zhao,tk.kim,w.luo12}@imperial.ac.uk

Abstract

In this paper, we present a unified method for joint face image analysis, i.e., simultaneously estimating head pose, facial expression and landmark positions in real-world face images. To achieve this goal, we propose a novel iterative Multi-Output Random Forests (iMORF) algorithm, which explicitly models the relations among multiple tasks and iteratively exploits such relations to boost the performance of all tasks. Specifically, a hierarchical face analysis forest is learned to perform classification of pose and expression at the top level, while performing landmark positions regression at the bottom level. On one hand, the estimated pose and expression provide strong shape prior to constrain the variation of landmark positions. On the other hand, more discriminative shape-related features could be extracted from the estimated landmark positions to further improve the predictions of pose and expression. This relatedness of face analysis tasks is iteratively exploited through several cascaded hierarchical face analysis forests until convergence. Experiments conducted on publicly available real-world face datasets demonstrate that the performance of all individual tasks are significantly improved by the proposed iMORF algorithm. In addition, our method outperforms state-of-the-arts for all three face analysis tasks.

1. Introduction

The problem of analyzing face images (*e.g.*, head pose estimation, expression recognition, and facial landmark detection) is a fundamental task in computer vision. It plays a central role in many real-world applications such as human-computer interaction, facial animation, video surveillance, *etc*. However, it still remains challenging due to complex variations of facial appearance, especially in the unconstrained in-the-wild scenarios.

In previous work [1, 2, 3, 7, 18, 20, 22, 24, 28], these face analysis tasks are usually approached as separate problems by a disparate set of techniques. However, these tasks are essentially *closely* related and can help each other. For example, the facial attribute information (*e.g.*, pose, expression, or facial phenotype) can provide a strong shape prior to constrain the variations of facial landmarks [7, 11]. On the other hand, the performance of pose estimation and expression recognition depend heavily on the localization accuracy of facial landmarks [12, 18, 22]. Therefore, in this paper, we propose to model the relatedness among multiple face analysis tasks and exploit such relatedness to *mutually* boost the performance of all tasks in a *unified* method.

To address this multi-task face analysis problem in a unified framework, we cast it as a joint probability estimation problem and tackle it using the powerful random forests algorithm [6]. Specifically, it can be formulated as follows, *i.e.*,

$$(\theta, \mathbf{e}, \mathbf{s})^* = \arg\max_{\theta,\mathbf{e},\mathbf{s}} p(\theta, \mathbf{e}, \mathbf{s}|\mathcal{I}, \mathbf{b}) \quad (1)$$

where $\mathcal{I}$ is the given face image, $\mathbf{b}$ is the corresponding face bounding box provided by a face detector, θ, e, $\mathbf{s}$ represent the pose, expression, and the concatenation of all landmark coordinates respectively. Different from the standard random forests, where the posterior of each task is usually estimated *independently*, we *explicitly* consider the dependencies among these tasks (*i.e.*, pose θ, expression e, and landmarks $\mathbf{s}$) and jointly predict them. Even though the recently proposed Structured-Output Random Forests algorithms also try to consider the dependencies among multiple output-variables, they mainly focus on the image segmentation problem, only considering the spatial consistency of pixels/objects in images [10, 14, 15, 17]. The relatedness among these *general* face tasks (pose estimation, expression recognition, and landmark localization *vs.* the pixels/objects within the same images) can not be *simply* characterized by such kind of spatial consistency.

Therefore, in this paper, a novel iterative Multi-Output Random Forests (*i*MORF) algorithm is proposed to characterize the dependencies among these tasks and *iteratively* exploit such dependencies to boost the performance of all tasks. Specifically, the *effect of pose and expression on landmark localization* is naturally encoded in a *hierarchical* face analysis forest. As shown in Fig. 1(a), we mainly focus on the *classification* of pose and expression at the top level of each tree (the function f), and the *regression* of landmark positions at bottom level (the function g). Through passing

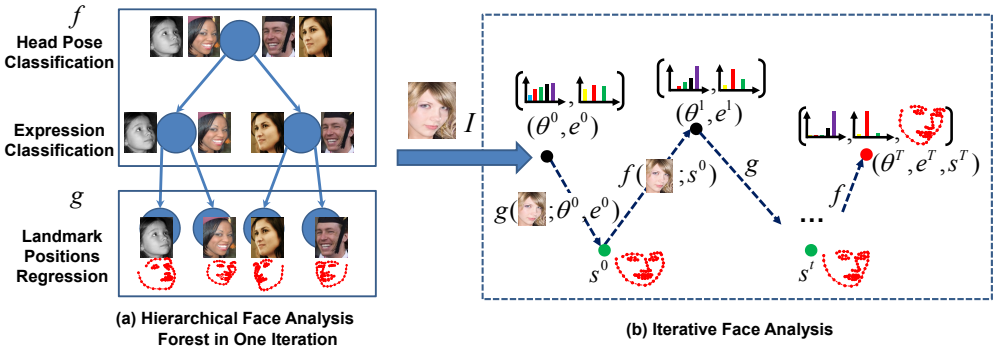

Figure 1: *Overview of our iterative Multi-Output Random Forests for unified face analysis. Here, θ, e, $\mathbf{s}$ represent the head pose, facial expression, and facial landmark positions respectively. f and g represent the classification and regression functions in the top and bottom layers. The superscripts of θ, e, $\mathbf{s}$ denote the iteration step.*

images down the top level trees in the forest, we can get the probability estimations of pose and expression, which provide strong prior to constrain the variations of landmark positions. Once the estimation of landmark positions is obtained, we can model its *effect on pose and expression* by encoding the shape information into shape-related features and feeding them to the subsequent *shape-aware* hierarchical face analysis forest. With the gradual refinement of the landmark positions, more discriminative shape-related features can be extracted to help the subsequent estimation of pose and expression. Through alternating these two steps for pose/expression (f) and landmark positions (g), as shown in Fig. 1(b), we can keep exploiting the relatedness among all tasks and mutually boost their performance. Our method is comprehensively evaluated on multiple real-world face databases. Experimental results demonstrate that the performance of all tasks are greatly improved through the proposed *i*MORF algorithm. Additionally, our method outperforms state-of-the-art methods on the evaluated databases for all three face analysis tasks.

The main contributions of this paper are two-fold:

- We *jointly* estimate multiple face analysis tasks in a unified framework. To our best knowledge, it is the first time that random forests is applied to jointly estimate head pose, facial expression and facial landmarks.
- We propose a novel *iterative* technology for Multi-Output Random Forests problem, where the dependencies among multiple outputs/tasks are *iteratively* modeled and utilized to boost the performance of all tasks.

2. Related work

In this section, we mainly review two related topics: face image analysis and random forests. There is a rich history of them in computer vision. Please refer to [4, 6, 18, 22] for comprehensive reviews. Due to the limit of space, we only focus on the most related methods.

2.1. Face image analysis

Facial landmark detection. Facial landmark detection has been an active topic in computer vision and lots of promising methods have emerged in recent years [1, 2, 3, 7, 20, 24, 27, 28, 30]. The work most related to ours is [7], which exploits conditional regression forests for facial landmark detection. In [7], the head pose is firstly estimated by a classification forest. With the estimated head pose as a conditional face property, they obtain more accurate facial landmark detection through regression forest. In our work, the *mutual* interaction among multiple face analysis tasks is considered, *i.e.*, further exploiting the shape information to improve the estimation of head pose and facial expression. Also, the training set for facial landmark detection is *automatically* determined through the hierarchical face analysis forest rather than being *manually* selected [7].

Head pose estimation. Two kinds of information are typically utilized for pose estimation, *i.e.*, the image appearance and the facial landmark configurations. For the former, they usually exploit the image appearance features, (*e.g.*, SIFT, LBP) and the discriminative learning methods (*e.g.*, SVM, Random Forests) for accurate pose estimation [9, 18]. For the latter, the pose is estimated using the correspondence between points in 2D shapes and points in 3D face models via the POSIT algorithm [8]. However, the estimated landmark points are often not accurate enough for reliable pose estimation.

Facial expression recognition. Similar to head pose estimation, traditional methods also exploit the image appearance features and some discriminative learning algorithms to estimate expression [22].

Recently, researchers also perform several face analysis tasks in an integrated system [23, 30]. For example, Zhu *et al.* [30] propose a tree-structured deformable part model to jointly perform face detection, pose estimation and facial landmark localization in real-world face images. However, it is computationally demanding in both training and test-

ing. Additionally, the *mutual* interaction among these tasks is also not *explicitly* modeled in [30].

2.2. Random Forests

Random forest techniques are very popular for their efficiency to handle multi-class classification and multivariate regression problem. They have demonstrated great performance in many computer vision tasks, such as action recognition [25], human pose estimation [26], *etc*. More recently, Structured-Output Random Forests is proposed to enable spatially consistent predictions in image segmentation problem [10, 14, 15, 17]. For example, Entangled Decision Forest (EDF) is presented for simultaneously segmenting multiple anatomical structures in CT scans [17]. The dependencies among pixels/objects are implicitly encoded by the long-range and context-rich features, which are extracted on the response map of random forests in earlier stages. On top of EDF, Geodesic Forests (GeoF) is proposed for structured output prediction by incorporating the Conditional Random Field (CRF) energy term into the random forest [15]. Moreover, joint classification-regression forests is introduced for simultaneous classification and regression, which is evaluated on the multi-object segmentation problem [10].

Different from the previous work, the proposed *i*MORF algorithm considers the relations among more general tasks (*i.e.*, pose estimation, expression recognition, and landmark detection) rather than the spatial consistency among pixels/objects within the same images (*e.g.*, segmenting organs in medical image).

3. Face analysis by iterative Multi-Output Random Forests

In this section, we first give a brief overview of our method. Then, we will give more implementation details about the training and testing for the proposed *i*MORF algorithm.

3.1. Method overview

As mentioned in Sec. 1, the unified face image analysis problem can be formulated as a joint probability estimation problem in Eq. 1. It is worth noting that we estimate discrete pose categories rather than continuous poses.

Specifically, as shown in Fig. 2(a), we first get an *initial* estimation of the head pose, expression, and facial landmark positions by a hierarchical face analysis forest. In this step, there is not any shape information that could be used. Therefore, we densely extract image patches within the face bounding box $\mathbf{b}$. At the top level of the hierarchical face analysis forest, it mainly focuses on the classification problem of head pose and facial expression, *i.e.*, *discriminatively* maximizing the posterior probabilities of θ and e:

$$(\theta, e)^0 = \arg\max_{\theta,e} p(\theta, e|\mathcal{I}, \mathbf{b}). \quad (2)$$

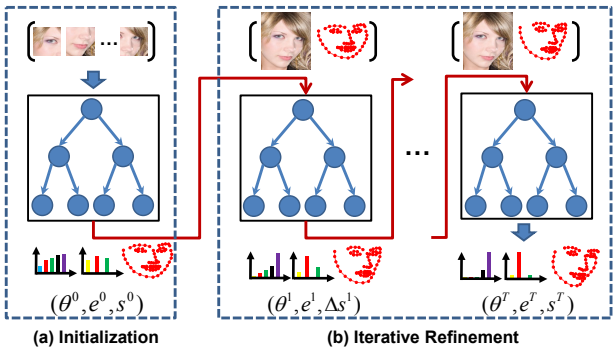

Figure 2: *Illustration of the iteration procedure of the proposed iMORF algorithm.*

Subsequently, at the bottom level, with the latent shape prior provided by the initially estimated pose and expression, the posterior probability of facial landmarks can be obtained through a regression forest, *i.e.*,

$$\mathbf{s}^0 = \arg\max_{\mathbf{s}} p(\mathbf{s}|\mathcal{I}, \mathbf{b}, \theta^0, e^0). \quad (3)$$

The implementation details will be described in Sec. 3.2.

For the following iteration steps, as shown in Fig. 2(b), the shape-related geometric features can be extracted from the previously estimated facial landmark positions together with the image appearance features. As demonstrated in [25], the shape related geometric features are more robust to lighting changes and occlusion. Therefore, based on the more expressive features, more accurate probability distribution of head pose and expression could be estimated through the hierarchical face analysis forest. Similarly, their posterior probabilities are predicted at the top level, *i.e.*,

$$(\theta, e)^t = \arg\max_{\theta,e} p(\theta, e|\mathcal{I}, \mathbf{b}, \mathbf{s}^{t-1}), \quad (4)$$

where t is the iteration index. Accordingly, more accurate positions of facial landmarks could be obtained based on the more accurate shape prior provided by the refined estimations of head pose and facial expression. It is worth noting that, we just predict the residual vectors $\Delta\mathbf{s}^t$ between the currently estimated facial landmark positions $\mathbf{s}^{t-1}$ and the target true shape $\mathbf{s}^*$, *i.e.*,

$$\Delta\mathbf{s}^t = \arg\max_{\Delta\mathbf{s}} p(\Delta\mathbf{s}|\mathcal{I}, \mathbf{b}, e^t, \theta^t, \mathbf{s}^{t-1}). \quad (5)$$

The newly estimated facial landmark positions $\mathbf{s}^t$ can be updated by $\mathbf{s}^t = \mathbf{s}^{t-1} + \Delta\mathbf{s}^t$. The implementation details will be described in Sec. 3.3.

Through iteratively performing Eq. 4 and Eq. 5 by the hierarchical face analysis forest, the prediction of each task could be progressively improved. We terminate the iteration process until the norm of shape residual $\Delta\mathbf{s}^t$ is smaller than a threshold ε, which is set to $0.04 \times d$ in the following experiments. Here, d is the distance between two outer eye corners.

3.2. Initialization by hierarchical face analysis forest

In this subsection, we firstly describe how to train a face analysis forest, which *hierarchically* performs multiple face analysis tasks in one forest rather than learning separate forests for each task. Then, we will show how to initialize the estimations of these tasks by the learned face analysis forest. In the current stage, no shape feature is available.

3.2.1 Training

To build the hierarchical face analysis forest, a set of training images are collected with ground truth labels. We randomly extract a set of image patches $\{\mathcal{P}_i\}$ with annotations $\{(\theta_i, e_i, \mathbf{d}_i)\}$ from each image. Here, $\mathbf{d}_i$ denotes the offset vector $(\Delta x_{i1}, \Delta y_{i1}, \Delta x_{i2}, \Delta y_{i2}, \ldots, \Delta x_{iK}, \Delta y_{iK})^{\mathrm{T}}$ from the centroid of patch $\mathcal{P}_i$ to all K landmarks.

We grow N decision trees by recursively splitting and passing the current training data to two child nodes. For each node, a group of candidate split functions are generated. Here, the split function is represented by a simple patch comparison test (f, τ) as in [7], where τ is the candidate threshold, and f is defined as the image appearance difference between two random rectangles within the patch. The best one is chosen by maximizing a quality function. Instead of using the information gain or label variance for the metric, a hybrid quality function is used in this paper:

$$Q_{face} = \alpha Q_\theta + (1-\alpha)\beta Q_e + (1-\alpha)(1-\beta)Q_{\mathbf{s}}, \quad (6)$$

where $\alpha \geq 0$ and $\beta \geq 0$ control the weights of these energy function terms. Specifically, the head pose term Q_θ and expression term Q_e are defined as the information gain used in standard classification forests [6]. These terms evaluate the classification performance of head pose and expression. The quality function term $Q_{\mathbf{s}}$ for landmark detection is defined as in [7]:

$$Q_{\mathbf{s}} = -\sum_{k=1}^{K} \frac{\sum_i p(c_k|\mathcal{P}_i)}{|\mathcal{P}|} \log\left(\frac{\sum_i p(c_k|\mathcal{P}_i)}{|\mathcal{P}|}\right), \quad (7)$$

where $p(c_k|\mathcal{P}_i)$ indicates the probability that patch $\mathcal{P}_i$ belongs to landmark k. The value of $p(c_k|\mathcal{P}_i)$ is determined by the distance between patch $\mathcal{P}_i$ and landmark k.

Similar to [26], the weights of these energy terms are *adaptively* switched according to the purity of head pose and facial expression. Let $\Delta(\cdot)$ denote the difference between the highest and the second highest posterior of a class in a node. $\Delta(\theta)$ and $\Delta(e)$ represent the margin measures of head pose labels θ and expression labels e. The purity of a node with respect to head pose and facial expression is measured by:

$$\alpha = \begin{cases} 1 & \text{if } \Delta(\theta) < t_\alpha \\ 0 & \text{otherwise} \end{cases}, \ \beta = \begin{cases} 1 & \text{if } \Delta(e) < t_\beta \\ 0 & \text{otherwise} \end{cases}. \quad (8)$$

The adaptive switch of α and β enables a coarse-to-fine learning for the face analysis tasks. At the initial top levels, Q_θ is dominant when class labels are evenly distributed. Therefore, the face analysis forest mainly focuses on the classification of head pose. As the purity of head pose becomes bigger than a threshold, β increases to shift the training objective to expression classification. At the bottom levels, the forest switches to landmark regression when the purities of pose and expression are high enough (Fig. 1). In our experiments, t_α and t_β are both set to 0.99.

Once the training is completed, for each leaf node l, the class (head pose or expression) posterior is obtained by:

$$p(c|l) = |\mathcal{S}_{lc}|/|\mathcal{S}_l|, \quad (9)$$

where $\mathcal{S}_l$ denotes an image patch set arriving at node l, and $\mathcal{S}_{lc}$ denotes an image patch set falling into node l with label c (pose or expression). Given the head pose and expression labels, the distribution over the offset vector is modeled by a multivariate Gaussian, $\mathcal{N}(\mathbf{d}; \overline{\mathbf{d}_l}, \Sigma_l)$, where $\overline{\mathbf{d}_l}$ and Σ_l are the mean and covariance matrix of the offset vectors of patches arriving at leaf node l.

The probabilities of the forest for a patch $\mathcal{P}_i$ over all tasks, *i.e.*, $p(\theta|\mathcal{P}_i), p(e|\mathcal{P}_i)$, and $p(\mathbf{d}|\mathcal{P}_i)$, are obtained by averaging over all N trees [6].

3.2.2 Testing (initialization)

Given an input image $\mathcal{I}$, we firstly normalize it by the automatically detected face bounding box. Then, as shown in Fig. 2(a), we densely extract image patches $\{\mathcal{P}_i\}, i = 1, \ldots, M$, within face and feed them to the hierarchical face analysis forest. By passing all image patches down all the trees in the forest, each patch ends in a set of leaf nodes. The *initial* estimation of head pose θ^0 is obtained by averaging the estimations of all patches, *i.e.*,

$$p(\theta|\mathcal{I}) = \frac{1}{M}\sum_{i=1}^{M} p(\theta|\mathcal{P}_i). \quad (10)$$

Same as head pose, the *initial* estimation of expression e^0 can also be predicted as in Eq. 10. The posterior of offset vector $\mathbf{d}$ could be obtained by:

$$p(\mathbf{d}|\mathcal{I}) = \frac{1}{M}\sum_{i=1}^{M} p(\mathbf{d}|\mathcal{P}_i). \quad (11)$$

Benefited from the hierarchical structure, samples with similar head pose and expression tend to group together, which makes a compact estimation of $p(\mathbf{d}|\mathcal{I})$. Then, the possible shape variations of image $\mathcal{I}$ are strongly constrained. Finally, the landmark positions $\mathbf{s}^0$ are obtained by performing mean-shift for each point.

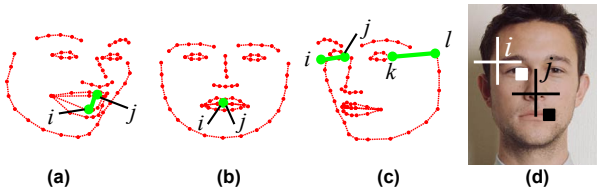

Figure 3: *Illustration of shape-related features, where (a) and (b) are landmark distance features, (c) is landmark distance ratio feature, and (d) is shape-related image appearance feature proposed in [3]. Here, i, j, k, l are indices of landmarks.*

3.3. Iterative refinement by shape-aware face analysis forest

Different from the face analysis forest in Sec. 3.2, more discriminative information can be extracted from the shape channel as well as the image appearance channel. In this subsection, we will demonstrate how to effectively encode the shape information into the training of the forest, and how to *iteratively* exploit the finer and finer shape information to improve the performance of all face analysis tasks.

Shape-related features. Compared to holistic shape used in the POSIT algorithm, in this paper, the shape information is encoded by two simple shape-related geometric features: the *landmark distance* feature and the *landmark distance ratio* feature. As shown in Fig. 3, the landmark distance feature is defined as the normalized distance between two landmarks, and the landmark distance ratio feature is defined as the ratio between two landmark distances. We can observe that the shape-related geometric features are effective to distinguish different expressions and poses. Their effectiveness is also validated in action recognition [25].

Besides the shape-related geometric features, as shown in Fig. 3(d), the shape-related image features are exploited to characterize the image appearance variation [3]. Here, the image patch is indexed relative to the currently estimated shape rather than the original image coordinates. As proved by [3], this feature achieves better geometric invariance than the one indexed by the original coordinates.

3.3.1 Training

Based on the currently estimated head pose θ^{t-1}, facial expression e^{t-1}, and landmark positions $\mathbf{s}^{t-1}$, the whole face images $\{\mathcal{I}_i\}$ and the associated annotations $\{(\theta_i^{t-1}, e_i^{t-1}, \mathbf{s}_i^{t-1}, \Delta\mathbf{s}_i^{t-1})\}$ are used for training the shape-aware hierarchical face analysis forest. Here, t is the iteration index, $\Delta\mathbf{s}_i^{t-1} = \mathbf{s}_i^{t-1} - \mathbf{s}_i^*$ represents the *shape residual vector*, where $\mathbf{s}_i^*$ is the ground truth shape for image $\mathcal{I}_i$. In order to increase the generalization capability of the face analysis forest, we augment the training samples by adding perturbations (*e.g.*, rotation, translation, scaling) to the *initial estimations* of landmark positions $\{\mathbf{s}_i^0\}$.

We grow N decision trees and the optimal shape-related features are automatically selected by maximizing the quality function, Q_{face}, which is the hybrid one defined in Eq. 6. The head pose term Q_θ and expression term Q_e are same as the definitions in Eq. 6. The facial landmark detection term $Q_\mathbf{s}$ is defined as:

$$Q_\mathbf{s} = -\Psi(\Sigma_l(\Delta\mathbf{s}^{t-1})) - \Psi(\Sigma_r(\Delta\mathbf{s}^{t-1})), \qquad (12)$$

where $\Psi(\cdot) = \log(\det(\cdot))$, Σ is the covariance matrix of shape residual vectors $\Delta\mathbf{s}^{t-1}$, and l, r denote the left and right split respectively. The quality functions Q_θ and Q_e enable the classification performance of head pose and expression at the top level. At the bottom level, the energy term $Q_\mathbf{s}$ is responsible for the *shape residual vector* regression. This term sends images with coherent $\Delta\mathbf{s}^{t-1}$ (in the direction and size of $\Delta\mathbf{s}^{t-1}$) into the same leaf node. The mean residual vector of a leaf node will be utilized to update the shape vector of a new test image arriving at the node. Images falling into the leaf nodes with larger $\Delta\mathbf{s}^{t-1}$ (*i.e.*, errors) would, therefore, be greatly updated towards their ground truth shapes (see Sec. 3.3.2). The energy term $Q_\mathbf{s}$ for residue regression enables quick convergence of our iterative algorithm.

When creating a leaf node l, the posteriors of l over the head pose, expression are computed as in Eq. 9. The distribution of shape residual vectors is modeled by a multivariate Gaussian, $\mathcal{N}(\mathbf{\Delta s}^{t-1}; \overline{\mathbf{\Delta s}}_l^{t-1}, \Sigma_l)$, where $\overline{\mathbf{\Delta s}}_l^{t-1}$ and Σ_l are the mean and covariance matrix of the shape residual vectors of images ending in node l. Through averaging the estimations of all trees in the shape-aware face analysis forest, we can obtain the estimations of pose θ_i^t, expression e_i^t, and shape residual vector $\Delta\mathbf{s}_i^t$ for an image $\mathcal{I}_i$.

By adding the shape residual predicted by the shape-aware forest, we update the shape of image $\mathcal{I}_i$ (see Sec. 3.3.2), and continue to extract shape-related features to train the next shape-aware face analysis forest. Through *iteratively* training several cascaded shape-aware face analysis forests, we can *gradually* approximate the ground truth shapes $\{\mathbf{s}_i^*\}$, as well as obtaining sharper and sharper posterior distributions for head pose and facial expression.

3.3.2 Testing (iterative refinement)

As shown in Fig. 2(b), starting from the initial estimation, $(\theta^0, e^0, \mathbf{s}^0)$, we extract shape-related features from the testing image $\mathcal{I}$ and shape $\mathbf{s}^0$, then feed them to the shape-aware face analysis forest learned in Sec. 3.3.1. By passing testing image $\mathcal{I}$ down the shape-aware face analysis forest, we can update the predictions of pose, expression, and the shape residual vector by voting the estimations of all trees. Then, the positions of facial landmarks are updated by adding the estimated shape residual vector.

Through iterating this step until convergence, *i.e.*, there is no change in estimated landmark positions or the maximum iteration number T is reached, we can achieve better performance for all tasks. We empirically set T to 5 in our experiments.

4. Experiments

4.1. Datasets and evaluation metrics

Our experiments are conducted on three publicly available face databases, *i.e.*, 300-W [1], Bosphorus [21], and CK+ [13].

The **300-W** is the first in-the-wild challenge for automatic facial landmark detection. It consists of several real-world data sets (*e.g.*, LFPW [2], AFW [30], HELEN [16], and IBUG [19]). All 6193 images are re-annotated with 68 points and the corresponding face bounding boxes are provided by an in-house detector. For a comprehensive analysis of our algorithm, we manually label the pose and expression for each image. Specifically, the pose is divided into 5 discrete categories, *i.e.*, Left2: $(-90°, -45°)$, Left1: $[-45°, -15°)$, Frontal: $[-15°, +15°]$, Right1: $(+15°, +45°]$, and Right2: $(+45°, +90°)$. Due to the ambiguity of expressions in real-world images, only three basic expressions (*i.e.*, Neutral, Happy and Others) are labeled.

The **Bosphorus** dataset is intended for research on face image processing. There are totally 105 subjects and 4666 faces, which are rich of expressions (*e.g.*, Neutral, Anger, Disgust, Fear, Happiness, Sadness, and Surprise) and pose variations (*i.e.*, 13 yaw and pitch rotations). For each image, at most 24 landmark points are manually annotated.

The **CK+** database is published for research in automatic expression recognition. It has 8 discrete facial expressions, *e.g.*, Neutral, Anger, Contempt, Disgust, Fear, Happiness, Sadness, and Surprise. Additionally, 68 landmark points, which are exactly the same as in 300-W, are manually annotated for each image.

In our experiments, the normalized root-mean-squared error (NRMSE) is adopted to measure the localization error of facial landmarks. It is given as a percentage, computed by dividing the root mean squared error by the inter-ocular distance. The cumulative distribution function (CDF) of NRMSE is used to evaluate the performance of facial landmark detection algorithms. The classification accuracy is exploited to measure the performance of pose estimation and expression recognition algorithms.

4.2. Algorithm analysis

In this section, we carefully analyze the details of our method on the above-mentioned 300-W and Bosphorus databases.

[1] http://ibug.doc.ic.ac.uk/resources/300-W/

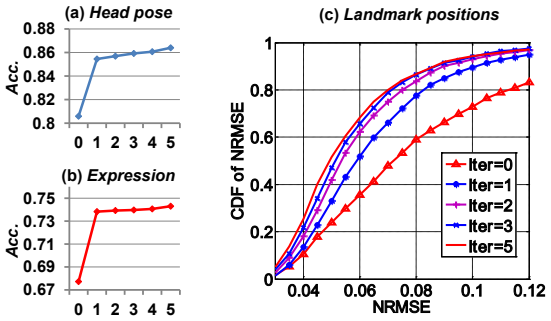

Figure 4: *Iterative performance improvement for each task on the 300-W evaluation set.*

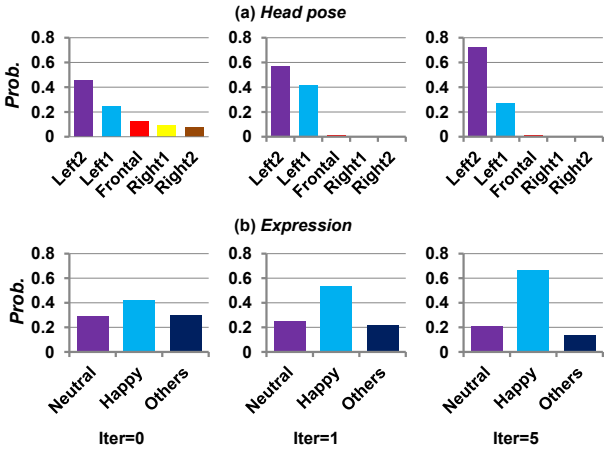

Figure 5: *Probability distribution variations for head pose ("Left2") and expression ("Happy") in different iterations (Iter = 0, 1, 5 respectively).*

We firstly evaluate our method on the 300-W database. To train the model of our algorithm, 80% images from each pose and expression sub-category are randomly selected as the training set. The remaining 20% images are used as the testing set. For each random forest, 10 trees are trained and the maximum depth of trees is set to 20. The minimum sample number in the leaf node is set to 10.

From experimental results, we can observe that: **1)** The performance of all face analysis tasks can be *mutually* boosted through our *i*MORF algorithm. As shown in Fig. 4, the performance of landmark detection (Fig. 4(c)), pose estimation (Fig. 4(a)), and expression recognition (Fig. 4(b)) are iteratively improved through our *i*MORF algorithm. Especially, their performance are significantly improved at the first iteration, demonstrating the effectiveness of the shape information; **2)** It is interesting to note that the iteration optimization strategy dose not increase the runtime much. The average end-to-end runtime on 100 images (720×576 pixels) is 350 ms (i7cpu@3.6GHz). Run-time (ms) for the first 6 iterations are 240, 30, 20, 20, 20, 20 respectively; **3)** The probability estimations of pose and expression become

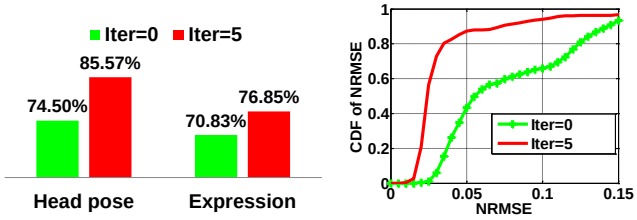

Figure 6: *Comparison results on the Bosphorus database.*

more and more accurate. Fig. 5(a) and Fig. 5(b) demonstrate the average posterior probabilities estimated for images with pose label "Left2" and images with expression label "Happy" respectively. Here, the probability estimation results in iteration 0, 1, 5 are shown. Through the iterations, it can be observed that the distributions become sharper and sharper. Therefore, more and more accurate shape prior is provided for the subsequent facial landmark localization.

To show the generalization capability of our method, we also evaluate it on the Bosphorus database with the similar setting on the 300-W database. Here, only the performance at the initial iteration (Iter = 0) and the fifth iteration (Iter = 5) are shown. As shown in Fig. 6, compared to the initial step, the performance of all face analysis tasks are significantly improved.

4.3. Comparison with state-of-the-art methods

In this subsection, we compare our method with state-of-the-art methods for these three tasks.

Facial landmark detection. We compare our method with the state-of-the-art landmark localization methods (*e.g.*, Asthana *et al.* [1], Dantone *et al.* [7], Zhu *et al.* [30], and Cox *et al.* [5, 20]) on the aforementioned 300-W dataset. It is important to note that, for Zhu *et al.*'s method, we use the model released by Asthana *et al.* This model is trained using the Multi-PIE and real-world LFPW training set (subset of the 300-W dataset). In addition, we run Asthana *et al.*'s matlab code based on Zhu *et al.*'s method, which can provide a better initialization. It is extremely slow but very accurate and represents the state-of-the-art on the 300-W database. The comparison results are shown in Fig. 7. It can be observed that our method outperforms these state-of-the-art methods on the 300-W evaluation set. Localization results of our method on some challenging example images with extreme pose, exaggerate expression, and partial occlusion are shown in Fig. 8.

Head pose estimation. We compare our method with the state-of-the-art image appearance-based ([9, 30]) and the shape-based ([1, 5, 20]) pose estimation methods on the 300-W evaluation set. Specifically, methods [1, 5, 20] exploit the detected facial landmarks and POSIT to estimate the continuous pose; methods [30] estimates discrete pose based on the tree-structured deformable part model, where

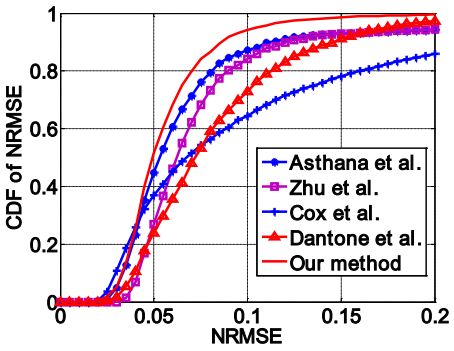

Figure 7: *Comparison with state-of-the-art landmark detection methods on the 300-W evaluation set.*

Table 1: *Comparisons with state-of-the-art head pose estimation and expression recognition methods.*

(a) Head pose estimation

Method	Acc.
Asthana *et al.* [1]	80.38%
Cox *et al.* [5, 20]	47.09%
Zhu *et al.* [30]	76.65%
Fanelli *et al.* [9]	80.59%
HOG+SVM	77.71%
Our method	***86.40%***

(b) Expression recognition

Method	Acc.
LBP+SVM	83.87%
SIFT+SVM	86.39%
HOG+SVM	89.53%
Gabor+SVM	88.61%
CSPL [29]	89.89%
Our method	***90.04%***

image appearance and the relative positions among landmarks are exploited. As shown in Table 1(a), our method outperforms these two kinds of methods on the 300-W evaluation set, demonstrating the effect of the gradually refined shape-related features.

Facial expression recognition. Finally, we compare our method with the state-of-the-art facial expression recognition methods on CK+ database with the same setting with [29], *i.e.*, for each sequence, the first image (neutral face) and three peak expression frames are used for prototype expression recognition. The experimental results are shown in Table 1(b), showing that our method obtains a better performance.

5. Conclusion and future work

In this paper, a novel *i*MORF algorithm is proposed for joint face analysis in a unified framework, where the relations among multiple tasks are *iteratively* exploited to *mutually* boost the performance of all tasks. Through encoding such relations into the hierarchical face analysis forests, more accurate prediction of facial landmarks could be obtained based on the stronger shape prior, *i.e.*, more compact probability estimation of head pose and facial expression. Simultaneously, through iteratively extracting finer and finer shape-related features (*cf.* low-level image features) from the improved estimation of shapes, the performance of pose estimation and expression recognition can also be signifi-

Figure 8: *Localization results on some challenging example images from 300-W evaluation set.*

cantly improved. The effectiveness and advantages of the proposed method are comprehensively evaluated on multiple real-world datasets.

In our future work, we will consider other facial attributes, such as facial phenotype, which plays an important role in facial beauty and gene-related medical diseases.

Acknowledgments

This work is supported by EPSRC grant (EP/J012106/1) 3D intrinsic shape recognition.

References

[1] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, 2013.

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

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

[4] O. Çeliktutan and B. Sankur. A comparative study of face landmarking techniques. *JIVP*, 2013(1):13, 2013.

[5] M. Cox, J. Nuevo, J. Saragih, and S. Lucey. CSIRO face analysis SDK. In *AFGR*, 2013.

[6] A. Criminisi and J. Shotton. *Decision Forests for Computer Vision and Medical Image Analysis*. 2013.

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

[8] D. F. Dementhon and L. S. Davis. Model-based object pose in 25 lines of code. *IJCV*, 15(1-2), June 1995.

[9] G. Fanelli, T. Weise, J. Gall, and L. Van Gool. Real time head pose estimation from consumer depth cameras. In *PR*, pages 101–110. 2011.

[10] B. Glocker, O. Pauly, E. Konukoglu, and A. Criminisi. Joint classification-regression forests for spatially structured multi-object segmentation. In *ECCV*. 2012.

[11] R. Gross and S. Baker. Generic vs. person specific active appearance models. *IVC*, 23(12):1080–1093, 2005.

[12] M. A. Haj, J. Gonzalez, and L. S. Davis. On partial least squares in head pose estimation: How to simultaneously deal with misalignment. In *CVPR*, 2012.

[13] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *AFGR*, 2000.

[14] P. Kontschieder, S. R. Bulo, H. Bischof, and M. Pelillo. Structured class-labels in random forests for semantic image labelling. In *ICCV*, 2011.

[15] P. Kontschieder, P. Kohli, J. Shotton, and A. Criminisi. GeoF: Geodesic forests for learning coupled predictors. In *CVPR*, 2013.

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

[17] A. Montillo, J. Shotton, J. Winn, D. Metaxas, and A. Criminisi. Entangled decision forests and their application for semantic segmentation of ct images. In *IPMI*, 2011.

[18] E. Murphy-Chutorian and M. M. Trivedi. Head pose estimation in computer vision: A survey. *IEEE TPAMI*, 31(4):607–626, 2009.

[19] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic. A semi-automatic methodology for facial landmark annotation. In *CVPRW*, 2013.

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

[21] A. Savran, H. Dibeklioğlu, O. Çeliktutan, B. Gökberk, B. Sankur, and L. Akarun. Bosphorus database for 3D face analysis. In *Biometrics and Identity Management*. 2008.

[22] C. Shan, S. Gong, and P. W. McOwan. Facial expression recognition based on local binary patterns: A comprehensive study. *IVC*, 27(6):803–816, 2009.

[23] Y. Wu, Z. Wang, and Q. Ji. Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In *CVPR*, 2013.

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

[25] A. Yao, J. Gall, and L. Van Gool. Coupled action recognition and pose estimation from multiple views. *IJCV*, 100(1):16–37, 2012.

[26] T.-H. Yu, T.-K. Kim, and R. Cipolla. Unconstrained monocular 3D human pose estimation by action detection and cross-modality regression forest. In *CVPR*, 2013.

[27] X. Zhao, X. Chai, and S. Shan. Joint face alignment: Rescue bad alignments with good ones by regularized re-fitting. In *ECCV*, 2012.

[28] X. Zhao, S. Shan, X. Chai, and X. Chen. Cascaded shape space pruning for robust facial landmark detection. In *ICCV*, 2013.

[29] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas. Learning active facial patches for expression analysis. In *CVPR*, 2012.

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