Privileged Information-based Conditional Structured Output Regression Forest for Facial Point Detection

Heng Yang, *Student Member, IEEE,* Ioannis Patras, *Senior Member, IEEE*

***Abstract*—This paper introduces a regression method, called Privileged Information-based Conditional Structured Output Regression Forest (PI-CSORF) for facial point detection. In order to train Regression Forest more efficiently, the method utilizes both privileged information, that is available only during training such as head pose or gender, and shape constraints on the location of the facial points. We propose to select the test functions at some randomly chosen internal tree nodes according to the information gain calculated on the privileged information. In this way the training patches arrive at leaves tend to have low variance both in terms of their displacements in relation to the facial points and in terms of the privileged information. At each leaf node, we learn three models: first, a probabilistic model of the pdf of the privileged information; second, a probabilistic regression model for the locations of the facial points; and third, shape models that model the interdependencies of the locations of neighbouring facial points in a predefined structure graph. Both of the latter two are conditioned on the privileged information. During testing, the marginal probability of the privileged information is estimated and the facial point locations are localized using the appropriate conditional regression and shape models. The proposed method is validated and compared with very recent methods, especially that use Regression Forests, on datasets recorded in controlled and uncontrolled environments, namely the BioID, the Labelled Faces in the Wild, the Labelled Face Parts in the Wild and the Annotated Facial Landmarks in the Wild.**

***Index Terms*—regression forest, facial point, privileged information, structured output.**

I. Introduction

DETECTING semantic facial points such as the mouth corners and the tip of the nose, is often the first step for many applications like face recognition and facial expression analysis. This has been a very active field in computer vision where considerable progress has been made over the last years[1], [2]. While most methods are tested on images recorded in constrained environments, some recent works in localizing facial landmarks have been extended to deal with more challenging face images collected "in the wild" [3], [4], [5], [6], [7]. However, detecting facial point in face images taken in uncontrolled conditions remains challenging due to high variations in facial appearance, pose, expression and also due to occlusions and illumination changes.

In recent years, Random Forests (RFs) have become increasingly popular for various high level computer vision tasks [8], [9] given their ability to handle large training datasets, their generalization power, their speed, and the relative ease of their implementation. Recently, random regression forests have been applied to the problems such as human pose estimation [10] and facial point detection [4]. In this framework, we make the following two contributions.

Our first contribution is that we learn higher quality decision trees using some additional information. That additional information, like the pose in [4], is only available at the training stage but not available at testing. To be consistent with the SVM-based LUPI (Learning Using Privileged Information) paradigm proposed by Vapnik & Vashist[11], this kind of additional information is called *privileged information*. Inspired by the LUPI paradigm, we propose a mechanism for regression forests which allows one to take advantage of the privileged information when training trees. A similar idea also appeared in methods that build regression forests that are conditioned on some global/additional information, such as [10] and [4]. Both of these models have shown that learning the probabilities of the target conditioned on global information can dramatically increase the detection accuracy while maintaining a low computational cost. However, neither [10] nor [4] exploited the privileged information when building the decision trees, but only utilized it at leaf nodes.

Our second contribution is that we model the shape constraints between the locations of the different points within the forest. In contract to the traditional methods that learn one or several statistical shape models using global parametric representation, our method builds shape models at each leaf node. In this way, the shape models are naturally conditioned on the test images. A recent work [12] also couples a shape model with random forests regression voting. However, that shape model is global and learned independently of the forest.

Similar to general random forests, our model is efficient to learn and to apply. In contrast to the classic random forest paradigm, the training process aims at decreasing the variance of image patches both in terms of privileged information and in terms of displacements relatively to the facial points. This goal is achieved by selecting the test functions at some randomly chosen internal tree nodes according to the information gain calculated from the privileged information. At each leaf node, we learn the probability of the privileged information, regression and shape models conditioned on it. During testing, the marginal probability of the privileged information is estimated and the facial point locations are localized using the appropriate conditional regression and shape models.

A preliminary version of part of this paper appeared in [13], where we introduced the concept of Structured Output Regression Forests and in [14], where we studied how privileged information can be used for tree induction. This paper presents

Heng Yang and Ioannis Patras are with the Department of Electrical Engineering and Computer Science, Queen Mary University of London, London, UK e-mail: {heng.yang,i.patras}@eecs.qmul.ac.uk.

Manuscript received 31st March 2014.

a general formulation that combines the two and includes a more in-depth discussion on the effectiveness of different types of privileged information and regression model selection. In this work we provide a thorough experimental evaluation to compare with the state-of-the-art methods on dataset recorded in controlled environment like BioID and also datasets with face images collected from the Internet, namely, Labelled Face in the Wild, Labelled Face Parts in the Wild and Annotated Facial Landmarks in the Wild.

The rest of the paper is organised as follows. We present related works in random forests and facial point detection in Section 2. In Section 3 we describes the proposed method. Experimental results and comparisons with the current state-of-the-art methods are given in Section 4. In Section 5 we draw some conclusions.

II. Related Work

In this section we first present a brief review of the random forests literature that is relevant to this work, and then present a review of related works on facial point detection.

A. Random Forests

Random forests have emerged as a powerful and versatile method successful in real-time human pose estimation, semantic segmentation, object detection and action recognition [8], [15], [16]. Recently, it has been applied to several problems on face analysis such as facial point detection [4], 3D head pose estimation and 3D facial point localization [17]. In what follows we only summarize very related methods. A more comprehensive introduction to decision forests and their applications in computer vision is given in [18].

In [17], in order to estimate the head pose, first the nose tip is localized the and then head pose is estimated. A voting framework is introduced to gather evidence from patches that are extracted from the whole depth image that can vote for the location of the nose tip and other key facial points. Since patches belonging to different parts of the image contain valuable global information, a particular point can be detected even when it is occluded. They have shown some success even when the nose tip is occluded.

So far as privileged information (i.e. additional information at training phase) is concerned, Sun *et al.* [10] propose a conditional regression forest model for human pose estimation. During training, at each leaf node, the probabilistic vote is decomposed into the distribution of 3D body joint locations for each leaf ID and the mapping probability. The latent variable can encode both known and unknown/uncertain properties of the pose estimation problems. When the global property is unknown, they propose to jointly estimate the body joint locations and the global property. Dantone *et al.* [4] also introduced a regression forests model conditioned on head pose for facial point detection. In their method, they divide the training set into subsets according to head pose yaw angle. An individual regression forest is trained on each subset and during testing a set of regression trees is selected according to the estimated probability of the head pose. The later is given by an additional forest trained to perform head pose estimation. Both [10] and [4] are called conditional regression forests. In this work we denote them by CRF (by Sun *et al.*) and C-RF (by Dantone *et al.*) respectively. From the perspective of training complexity, CRF proposes to share the tree structure instead of training a separate forest for each global property state.

B. Facial Point Detection

Facial point detection, or face parts localization, is a well-studied problem in computer vision as it is often the first step for further face analysis such as face recognition and facial expression recognition. We group them into local based and holistic based. The former involves local detection and usually combines with shape models. The latter treat the pose vector (locations of the facial points) as whole and regress it directly.

Local based Method: A wide variety of local feature detectors have been proposed which can be broadly classified into classification-based and regression-based. The classification-based approaches aim at designing discriminative classifiers for an individual facial points based on the texture information of the specific point and its surrounding region. Different types of classifiers and image features are employed. For instance, in [19], GentleBoost classifier based on Gabor features is proposed to detect 20 facial points separately. The classic Support Vector Machine (SVM) classifier is used as facial point detector in [20], [21], [22] and [6] with various image features such as Gabor,SIFT [23] and multichannel correlation filter responses [24]. Regression-based approaches to facial point detection have attracted the attention of researchers in recent years. Cristinacce & Cootes [25] presented a regression-based approach to facial point detection. It combines a GentleBoost regressor with an Active Shape Model (ASM) used to correct the estimates obtained. Another sequential regression-based approach was presented in [26], where Support Vector Regressors (SVRs) were combined with a probabilistic MRF-based shape model, that restricts the search to anthropomorphically consistent regions. Regression forests in recent years have also proven to be very powerful in detecting facial points [4]. The location of facial point is estimated by accumulating *votes* from nearby regions.

Since only a few facial points are discriminative, usually shape models are required to regularize the local detection outputs. Active Shape Model (ASM) [27] is one of the most common approaches to model the face shape. First, a mean shape is calculated as the concatenation of all the facial point coordinates. Then PCA is applied to find the basis of face variations. The Constrained Local Model (CLM) [28] learns a model of shape and texture in a similar manner as ASM, however, the texture is sampled in patches around individual features. The family of methods coined CLMs is shown to have better performance than ASM and AAM (e.g. [29]). Instead of using a densely connected spatial model, in [5] a tree model is proposed and the global optimal solution can be found through efficient dynamic programming algorithms. Further more, this work also proposes to build a mixture of tree-structured models to capture topological changes due to viewpoint and it has been used in [30], [31]. There are some

other shape models based on facial points such as the Pictorial Structure[32], Markov Random Fields [26], Restricted Boltzmann Machines [33], graph matching [34], and Regression Forests votes sieving [35].

Since the ratio of distance between co-linear points is fixed under affine transformations, line segments between facial points are also used to model the face shape such as [36]. Liang *et al*. [37] use a condensation algorithm modified with spatial constraints. It considers the segments forming the contours delimiting facial components, and a shape model is used to constrain consecutive segments to have coincident limits (closing the contour), and to keep a valid angle between them. The line segment is used in [7] as a type of geometry features for its cascade regression. Similar idea is employed in [26] but they go one step further and consider the relations between any two line segments connecting two pairs of facial points.

Holistic based Methods: Holistic based methods use global information (typically the whole facial image), and often try to align the shape in an iterative way. A typical method in this category is the Active Appearance Model (AAM) [1]. Such methods have difficulties with large variations in facial appearance due to head pose, illumination or expression. Their localization accuracy also degrades drastically on unseen faces [38] and low-resolution images. A recent attempt was made by [39] which shows improvement in memory and time requirements to train a discriminative appearance model. Instead of using a simple linear regression in each iteration of the AAM fitting, better optimizations are proposed in [40], [41], [42] and [43]. Noticeable progress in iterative holistic shape alignment has been made in recent years in the framework of Cascaded Pose Regression (CPR) [3], [7], [44]. Those methods directly learn a structural regression function to infer the whole facial shape (i.e., the location of the facial landmarks) from the image and explicitly minimize the alignment errors in the training data. The primitive random fern regressor at each iteration employs shape indexed features as input. Recent iterative approaches include the work by Xiong & de la Torre [43] based on SIFT features and convolutional neural networks [45]. Most of the iterative methods in this category depend on the initialization. Current CPR based methods like[3], [44], [46] attempt to deal with this issue by initializing the method with several shapes and then by selecting the median value of the outputs. [46] proposes a *smart restart* scheme to improve the robustness to random initialization. An user-assisted facial point localization algorithm is proposed recently [47]. Once the automatic fitting is performed, the user is instructed to pick the landmark with the largest error and move it to the correct location after which the algorithm adjusts the locations of the remaining landmarks to take into account the user input. The interaction round is repeated until the user is satisfied with the results. [48] is very close to our work that also uses regression forest for face analysis, which combines multiple tasks, face alignment, facial expression recognition and head pose estimation in a unified framework.

III. Proposed Method

In this section we describe the proposed method. The learning stage is illustrated in Fig. 1. This includes the privileged information-based tree induction (III-A) and models-learning at leaf nodes (III-B). As shown, by randomly selecting variable whose information gain is calculated, nodes decreasing the privileged information uncertainty and nodes decreasing displacement uncertainty, are interleaved in the decision tree. At each leaf node, three models are learned: First, a probabilistic model of the pdf of privileged information; Second, a regression model associated with each *base* feature point. A facial point is a base point for a certain leaf if the average relative offset of the patches that arrive at the leaf from the facial point in question is less than a threshold; Third, shape models related to the base feature point. Both of the latter two are conditioned on the privileged information.

During inference (described in Section III-C), the privileged information is firstly estimated and then it is used in the subsequent steps for calculating the regression voting map and the structure constraint voting map, as shown in Fig 2. The final detection is carried out on the product of these two maps.

A. Privileged Information-Based Tree Induction

We pose the facial point localization as a regression problem: given a set of input/output pairs (training data)

$$(x_1, y_1), ..., (x_M, y_M), x_m \in \mathcal{X}, y_m \in \mathcal{Y}, m \in 1, ...M,$$

the goal is to find a mapping function $f: x \rightarrow y$ from a set of mapping functions $F: \mathcal{X} \rightarrow \mathcal{Y}$ with a small error on the prediction $y = f(x)$. Similar to [11], in our method, additional privileged information $y^+ \in \mathcal{Y}^+$ is available during training as well. That is, the training set consists of triplets (x, y^+, y) instead of pairs (x, y). The privileged information $y^+ \in \mathcal{Y}^+$ belongs to a space that is different from the space $\mathcal{Y}$. The goal remains to find the best function $f: x \rightarrow y$ in the set of admissible functions.

In our case, a training sample is an image containing a face, the locations of facial points in the image and labels of privileged information, e.g., the head pose and subject's gender. Several fix-sized patches are randomly extracted from a training image, each represented by the image features $x = (x^1, x^2, ..., x^F) \in \mathcal{X}$ where F is the number of feature channels. Each patch is also annotated with a displacement vector $d = (d^1, ..., d^i, ...d^N) \in \mathcal{Y}$ to each of the N facial point and the privileged information label $y^+ \in \mathcal{Y}^+$. The set of training patches is therefore given by $\mathcal{P} = \{\mathcal{P}_m = (x_m, d_m, y_m^+)\}$. In this paper each tree considers only one type of privileged information.

1) General Tree Growing Procedure: A regression forest $\mathcal{T} = \{T_t\}$ is an ensemble of regression trees T_t. Each regression tree is most often induced greedily based on a randomly selected subset of the training data set $\mathcal{P} = \{\mathcal{P}_m\}$, in the following manner [49]. An empty tree starts with only one root node. Then, a number of test function candidates, ϕ, $\phi(x) \rightarrow \{0, 1\}$, defined over the image features x are sampled from a predefined distribution. Each patch is sent either to the left or to the right child depending on the test result. In this

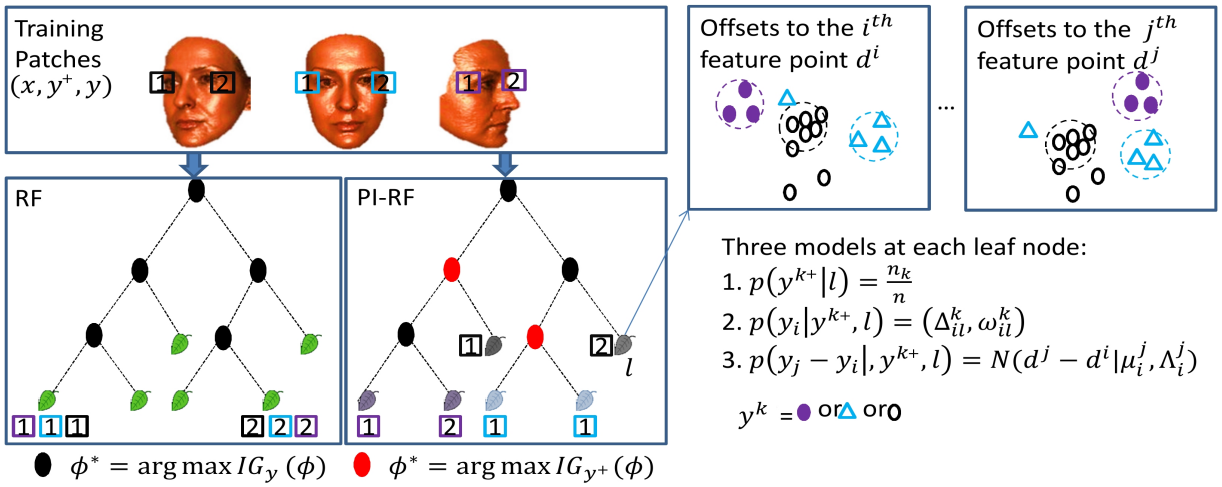


Fig. 1: **An illustration of our proposed learning stage.** An illustration of idealized tree induction for PI-RF and RF is shown on the left. The training patches are from face images with a large variety w.r.t. the Privileged Information (PI) (here the head pose). A classical RF attempts to guide patches that are located around the same facial point at the same leaf node. However, as the example shows, the visual features vary a lot due to changes in the PI and therefore it is difficult to guide them to the same leaf. On the contrary, in the PI-RF framework, the best split-function at some random internal nodes (in red) is selected directly according to the PI. As such, patches stored at the leaves tend to have low variation both in PI and in displacement. The information gain IG_y at dark nodes is calculated based on the entropy H_y, defined in (4) while at the color nodes, the information gain IG_{y^+} is calculated based on the entropy H_{y^+}, defined in (6). At each leaf node, one (or more) *base* feature point is defined and tree models are learned.

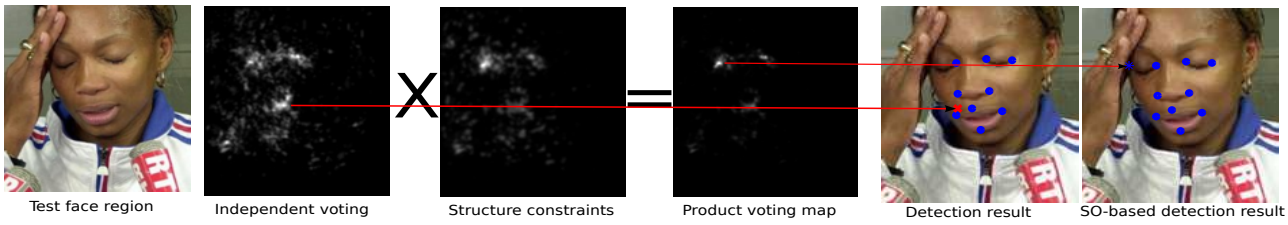


Fig. 2: An illustration of the structured output inference model. The face image shown here is *Laura_Flessel_0001.jpg* from LFW dataset.

way, a test function ϕ partitions the training set into two sets, $\mathcal{P}_L(\phi)$ and $\mathcal{P}_R(\phi)$. Each candidate test function is evaluated according to a certain scoring function, e.g. *information gain*, so that high scores are assigned to splits that aid in predicting the output well, i.e. those that reduce the average uncertainty about the target. The best test function, that is the one with the highest score, is selected and stored at the node in question. Then, the training set is partitioned according to this test into two subsets that are propagated to the two children nodes. The same procedure is recursively applied at each child node. The procedure stops when certain criteria are met, typically, when there are fewer than a minimum number of examples or a maximum tree depth is reached.

Our binary test function $\phi_{f,R_1,R_2,\tau}(x)$ is defined as in [8]:

$$\phi_{f,R_1,R_2,\tau}(x) = \begin{cases} 0 & \text{if} \quad x^f(R_1) < x^f(R_2) + \tau \\ 1 & \text{otherwise} \end{cases} \tag{1}$$

This is a comparison of the average value of the feature channel f in two asymmetric regions, R_1 and R_2, defined within the patch in question. $x^f(R)$ is the average value in region R and τ is a threshold.

Typically, the test function are randomly generated and the one that maximizes the information gain $IG(\phi)$ that is achieved by splitting the data is selected. That is,

$$\phi^* = \arg\max_\phi IG(\phi) \tag{2}$$

The information gain is a popular criterion used to determine the quality of a split and has been used for both classification, regression and density estimation[18]. The information gain is the *mutual information* between the local node decision (left or right) and the predicted output and it is defined as follows,

$$IG(\phi) = \mathcal{H}(\mathcal{P}) - \sum_{s \in \{L,R\}} \omega_s \mathcal{H}(\mathcal{P}_s(\phi)), \tag{3}$$

where $\omega_s = \frac{|\mathcal{P}_s(\phi)|}{|\mathcal{P}|}$ is the ratio of the patches sent to the child node. $\mathcal{H}(\mathcal{P})$ is a measure of uncertainty on the set $\mathcal{P}$ and it is usually related to the entropy of the labels of the elements in

the set. Depending on the nature of labels, $H(\mathcal{P})$ can either be a discrete entropy or a differential entropy. We will address this in next section.

2) Entropy Estimator: In our case, since $\mathcal{Y}$ and $\mathcal{Y}^+$ are different spaces, with different properties, appropriate entropy estimator is needed.

For $\mathcal{Y}$, we use the class-affiliation method proposed by [4] to measure the uncertainty, that is defined as:

$$H_{\mathcal{Y}}(\mathcal{P}) = -\sum_{i=1}^{N} \frac{\sum_m p(c_i|\mathcal{P}_m)}{|\mathcal{P}|} \log\left(\frac{\sum_m p(c_i|\mathcal{P}_m)}{|\mathcal{P}|}\right), \quad (4)$$

$$p(c_i|\mathcal{P}_m) \propto \exp\left(-\frac{|d_m^i|}{\lambda}\right), \quad (5)$$

where $p(c_i|\mathcal{P}_m)$ indicates the probability that the patch $\mathcal{P}_m$ is informative about the location of the feature point i. The class affiliation assignment is based on the Euclidean distance to the feature point. The constant λ is used to control the steepness of this function. In this way, we can avoid making a multivariate Normal distribution assumption on multiple feature points and calculate the differential entropy as in [18].

So far as $\mathcal{Y}^+$ is concerned, we only consider discrete privileged information because: 1) for our problem it is difficult to obtain the ground truth of the continuous head pose for each face image; 2) learning the model conditioned on continuous variable is still not well studied [10]. Therefore we discretise the head pose information by partitioning the pose space. In this context, head pose estimation becomes a multi-class classification problem. The finite set of privileged information classes is represented as $\mathcal{Y}^+ = \{1, 2, ..., K\}$. For each class, let h_k be the number of occurrences of the class, that is $h_k = \sum_{\mathcal{P}_m \in \mathcal{P}} \delta(y_i^+ = k)$. The empirical class probabilities $\hat{p}_k(\mathcal{P}) = \frac{h_k}{|\mathcal{P}|}$ (where $|\mathcal{P}| = \sum_k h_k$) are often used to calculate the entropy, i.e. $H_N(\mathcal{P}) = -\sum_{k=1}^{K} \hat{p}_k(\mathcal{P}) \log \hat{p}_k(\mathcal{P})$ (see e.g. [18] and references therein), however, it is pointed out by Nowozin [50] that the naive entropy estimator is biased and universally underestimates the true entropy. Therefore, as suggested in [50], we use the Grassberger entropy estimator [51], given as:

$$H_{\mathcal{Y}^+}(\mathcal{P}) = \log|\mathcal{P}| - \frac{1}{|\mathcal{P}|}\sum_{k=1}^{K} h_k G(h_k), \quad (6)$$

where the function $G(h)$ is given by $G(h) = \psi(h) + \frac{1}{2}(-1)^h \left(\psi(\frac{h+1}{2}) - \psi(\frac{h}{2})\right)$, and ψ is the *digamma* function. For large h the above function behaves like a logarithm and (6) is identical to naive entropy when $n \to \infty$. For small h, the estimation using (6) is shown to be more accurate.

In (4) and (6) we have designed the entropy estimator for both $\mathcal{Y}$ and $\mathcal{Y}^+$. During tree induction, at each internal node, the best split function is selected either based on (4) or (6). That is, the evaluation is either based on privileged information, or on the target. Note that in both cases the test itself is on the patch appearance, thus applicable both at training and test phase. When one of the stopping criteria of tree growing is met, several models will be learned at each leaf from patches that arrive there. An illustration of the tree induction process of our PI-based RF and of the traditional RF is in Fig. 1.

B. Models at Leaf Nodes

This section provides a description of our conditional regression model inspired by [10]. More specifically, three models are learned at each leaf: 1) a probabilistic model of the pdf of the privileged information at the leaf; 2) a probabilistic regression model for the locations of the *base* facial points; 3) shape models that model the interdependencies of the locations of facial points that are neighbors of the *base* point in a predefined structure graph.

1) Probabilistic Model of Privileged Information: First, at each leaf node, we calculate the pdf of the privileged information. Let n be the total number of training patches that arrive at a leaf node l, and let n_k be the number of patches belonging to class k. Then the probability for the class k at leaf l is

$$p(y^{k+}|l) = \frac{n_k}{n}, \quad (7)$$

where y^{k+} is a shorthand notation that $y^+ \in \mathcal{Y}^+$ belongs to the class k, i.e. $y^+ = k$.

2) Conditioned Regression Model: Second, at each leaf node, we learn the conditional regression model for the *base* feature point. Our model shares tree structures for all states of privileged information. This is similar to the *Partial* conditional regression model proposed in [10]. The samples are categorized into sub sets according to their privileged information labels and one conditional regression model is learned for each state.

Several regression models have been proposed in the literature. In our experiments we investigated two, both with one offset vector Δ and a weight ω, as the following.

1) A Mean Value model in which the offset vector Δ is the mean value Δ of the offsets and the voting weight ω is defined as $\omega = |S_\Delta|^{-\frac{1}{2}}$ where S_Δ is the covariance matrix.
2) A Mean-Shift model in which the offset vector Δ is the mode of the largest cluster returned from a Mean-Shift algorithm applied on the corresponding set of patches that arrive at leaf node in question. The weight w is assigned as the relative size of the largest cluster.

This greatly reduces the model complexity and training time since we do not need to train and store separate random forest for each state of privileged information as in [4]. Moreover, as shown in our experiments, it leads to better results.

The probability that the facial point i is located at y_i, given that a voting patch extracted at location z_x that arrive at leaf l is given by

$$p(y_i|y^{k+}, l) \propto \omega_{il}^k \cdot \delta(||\Delta_{il}^k||_2 \leq \gamma) \quad (8)$$

where $y_i = z_x + \Delta_{il}^k$, i and y^{k+} indicate the facial point number and privileged information state respectively. For notational clarity we will drop the facial point index i in the subsequent equations. γ is a threshold that prevents patches casting votes far away from place they are extracted. This factor avoids a bias towards an average face configuration as the votes from long distant patches are lack of accuracy. Thus at each leaf, the regression-voting models are only valid for those patches whose mean offset is less than the threshold γ.

In practice, each leaf is usually associated with one (in some cases two or more) facial point which we call a *base* point for the leaf in question.

3) Conditioned Shape Model: Third, at each leaf node, we learn the shape model for structured output regression. In contrast to the traditional face shape model such as ASM or CLM, our shape model is conditioned on the image information. Here we assume that the structure of the facial points can be organized in a graph, $G = (V, E)$, where V and E denote the sets of nodes and edges respectively. The nodes $i = 1, ..., N \in V$ correspond to facial points and the edges $(i, j) \in E$ capture their spatial relations. The graph can either be dense or sparse or a tree structured model as [5]. In this work, we assume the graph structure is already known and what needs to be done is to parameterise it. In practice, we manually define a sparse graph model according to the physical proximity of the facial points.

Recall that each leaf is associated with one (or more) base points. We proceed to model shape constraints between the base point and its neighbours in the predefined structure graph. More specifically, assuming j is one of the neighbouring nodes of node i in graph G, i.e. $j \in Ne(i)$, their *relative* position is modelled as a Gaussian,

$$p(y_j - y_i | y^{k+}, l) = N(d^j - d^i | \mu_i^j, \Lambda_i^j) \tag{9}$$

Note that y^{k+} is the privileged information state and that the shape model is conditioned on it. One model is learned for each state. Recall that d^j and d^i denote the patch offset to the j-th and i-th point respectively. μ_i^j and Λ_i^j denote the mean value and covariance matrix of the Gaussian model.

C. Inference

During testing, patches from the test image are densely sampled from the whole image and sent down through all trees in the forest. A stride parameter is set to control the density of the sampling. Each patch is guided by the binary tests stored at the internal nodes and will arrive at one leaf of each tree in the forest. In what follows, we use I denote the test image data and let X be the set of image patches x extracted from the image. Let L denote the set of leaf nodes in the forest.

We now describe how to estimate the facial point locations and the privileged information state based on the models at leaves defined in section III-B.

1) Privileged Information Inference: Similar to the ***MaxA*** approach in [10], the scoring function of privileged information state y^{k+} is defined as a sum of probabilistic votes contributed from all patches. Formally:

$$S(y^{k+}|I) = \sum_{x \in X} \sum_{l \in L} p(y^{k+}|l) p(l|x) \tag{10}$$

where $p(l|x)$ is delta function that a patch arrives at a leaf node l (referred to as the leaf ID mapping probability). We then estimate the most likely state of the privileged information $\hat{y}^+$ as:

$$\hat{y}^+ = \arg\max_{y^{k+} \in \mathcal{Y}^+} S(y^{k+}|I). \tag{11}$$

This estimate will be used as a known variable in subsequent steps.

2) Independent Regression: Firstly, we will describe the voting mechanism for independent estimation of locations of facial points, i.e. without considering the shape constraints. Similar to the *Partial Model* in [10], by expressing the probabilistic vote in terms of the distribution of each facial point for each *codeword* (leaf id) $p(y_i|l)$ and the probability $p(l|x)$ that the image patch is mapped to a codeword, the scoring function conditioned on the privileged information is defined as

$$S(y|y^{k+}, I) = \sum_{x \in X} \sum_{l \in L} p(y_i|y^{k+}, l) p(l|x) \tag{12}$$

Using the estimate $\hat{y}^+$ of y^+ given by (11), the best candidate of scoring functions over the privileged information state is selected as:

$$\hat{S}(y|I) = S(y|\hat{y}^+, I). \tag{13}$$

Then mean-shift mode finding algorithm can be applied on the selected scoring function for the corresponding facial point.

3) Structured Output Regression: Second, we will describe how to infer structured output based on the conditional shape model in (III-B3). Assume that a patch x that is extracted at z_x arrives at a leaf node l for which i is one of the base points. The vote for the i-th point is cast at $\bar{y}_i = z_x + \Delta_{il}$. Note that when privileged information is taken into account, Δ_{il}^k (instead of Δ_{il}) is used to estimate $\bar{y}_i^k$ where k is the state of the privileged information given in (8). Here we drop the index k to simplify the notation and make this model more general for regular regression forests. Recall (see III-B3) that at each leaf we maintain shape models that model the relative locations of the neighbours $j \in Ne(i)$ for each base point. Then. given the estimate $\bar{y}_i$ and the Gaussian model in (9), the structure constraint made on j is introduced in terms of the probability that the point j is located at y_j. The latter is modelled as $p_s(y_j|\bar{y}_i, l) = N(\bar{y}_i + \mu_i^j, \Lambda_i^j)$. Finally, the shape constraints on j given the estimated positions of all its neighbours i ($i \in Ne(j)$) are in the form of a scoring function S_s that gathers the votes cast by all the corresponding patches.

$$S_s(y_j|I) = \sum_{i \in Ne(j)} \sum_{x \in X} \sum_{l \in L} p_s(y_j|\bar{y}_i, l) p(l|x) \tag{14}$$

For each facial point, after accumulating votes cast from all patches in a test image, a local appearance evidence term like (13) and a structure constraint term like (14) are obtained. Then the structure constrained voting map is given as:

$$S_v(y|I) = S(y|I) \cdot S_s(y|I). \tag{15}$$

The mean-shift mode finding algorithm is applied on the final voting map to localize each facial point.

IV. Evaluation

In this section, we present results on public datasets and compare with a number of methods in the literature. In comparison to the recent state-of-the-art methods, our method shows better or comparable result in terms of location accuracy and training efficiency.

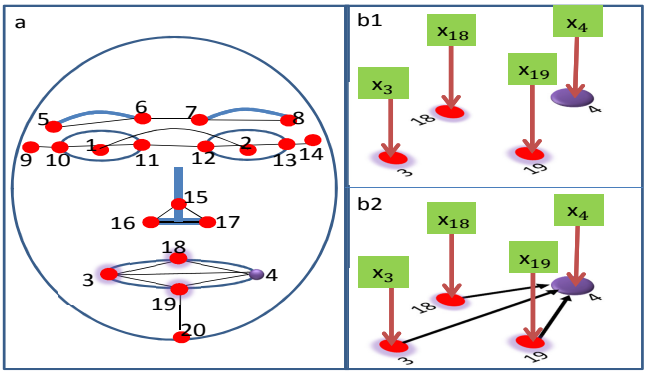


Fig. 3: Structured Output Regression. (a) shows manually defined sparse spatial relations of parts on face based on their physical locations. 20 selected face parts (dots) are displayed and their relations are represented by dark lines. The purple dot is one representative facial point and its neighbouring points are the red dots with purple shadow. (b1) illustrates an example of the independence assumption between points used in previous regression forests methods. Here we use x_i to represent the voting element x that is able to vote for part i, i.e. x arrive at leaves of which the i-th point is the base point. (b2) shows the spatial shape model of our method, in which the position of the 4^{th} point does not only depend on its voting patches x_4 but also on the estimated positions of its neighbouring points in the structure graph.

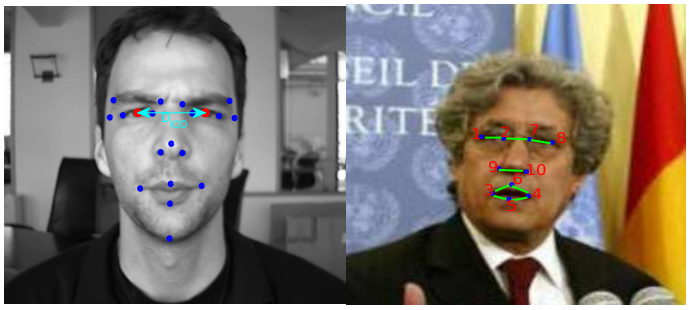


Fig. 4: Representative face images in BioID (left) and LFW (right) along with their facial point annotations. The green segments on the right face image represent our predefined graph model for the corresponding 10 facial points.

A. Datasets

In this paper we focus on datasets that contain face images that are recorded in uncontrolled environments, i.e. in the wild. One representative dataset obtained at laboratory conditions BioID is also used for comparison. Below we briefly describe the datasets that we used.

The *BioID dataset* [52] has been recorded in a laboratory environment using a low-cost web-cam. It consists of 1521 images, each depicting a frontal view of face of one of 23 different subjects with various facial expressions. One representative image from this dataset is shown in Fig. 4. Most of the previous methods in the topic of facial point detection have reported their results on this dataset. This allows us to compare our work with the state-of-the-art methods.

The *Labelled Face in the Wild (LFW) dataset* [53] has been designed for studying the problem of unconstrained face recognition. It contains more than 13,000 face images collected from the web. It consists of face images from 5749 individuals, 1680 of which have two or more distinct photos. Dantone *et al.* [4] have annotated 13,233 faces for this dataset with the location of 10 facial points. The images exhibit a large variation in face appearances (e.g., pose, expression, ethnicity, age, gender) as well as general imaging and environmental conditions (see Fig. 4 (right)).

The *Labelled Face Parts in the Wild (LFPW)* is also a dataset with face images in the wild. The images are downloaded from the Internet under a variety of acquisition conditions, including large variability in pose, illumination, expression, partial-occlusion of the face. This dataset shares only image URLs on web but some of them are no longer valid. Around 800 of the 1132 training images and 220 of the 300 test images could be downloaded when we carried out the experiment. In our experiment, we used the 220 testing images to test our trained model.

The *Annotated Face Landmarks in the Wild (AFLW)* [54] that contains real-world face images from Flickr. These images exhibit a very large variability in pose, lighting, expression as well as general imaging conditions. Many images exhibit partial occlusions that are caused by head pose, objects (e.g., glasses, scarf, mask), body parts (hair, hands) and shadows. We selected a subset in which all 19 frontal landmarks (i.e. excluding the two ear lobes) were annotated that consists of 6200 images.

B. Evaluation Methodology

Throughout the experimental section, we measure the localization performance using the inter-ocular distance (IOD)-normalized error. $e_i = \frac{||y_i^D - y_i^G||_2}{D_{IOD}}$. y_i^G is the ground truth location of point i, y_i^D is the estimated location of the point and D_{IOD} is the inter-ocular distance, defined as the distance between the eye centres. Since the locations of the eye centers are not annotated in LFW dataset, the inter-ocular distance is calculated as the distance between the midpoints of the ground truth eye corners. A point is regarded as a correct detection if $e_i < 0.1$. This measure is used to calculate the successful detection rate in the experiments.

To evaluate the overall performance of localization of multiple points on a face image, we use the m_{17} measure which defined in [55] as the mean error over all the internal points. Thus three of the 20 facial points, i.e. the chin and two temple points (i.e., P19, P9 and P14 in Fig. 3), are excluded when computing the m_{17}.

C. Experimental Settings

1) Setup: As in most of the previous face points detection approaches [4], [55], [12], our method assumes that the face bounding box is given both for training and testing images. The annotation of the LFW dataset already provides the face boxes for all face images. For the BioID dataset, we applied the Viola and Jones detector [56] in OpenCV to find the

face bounding boxes (all bounding boxes are then resized to 125×125 pixels). The height is enlarged by 20% in order to ensure all facial point points are enclosed. In order to ensure a fair comparison, we keep most of forest training setup in our experiments as similar as possible to the default setting of facial points detector described in [4]. The key setting parameters include: maximum depth of each tree (20), test candidates at split node (2500), patch size (0.25 $\times$ face box size), image features (one channel of normalized gray values, 35 channels of Gabor features and 2 channels of Sobel features), number of patches per image sample (100). Unless stated otherwise, those parameters were used for forest training in all of our experiments.

In order to illustrate the benefits of using privileged information, we consider three types of privileged information, namely, yaw head pose, roll head pose and gender status for the LFW dataset. More specifically, we constructed the privileged information as follows: we use the discrete head pose labels for the yaw angle (left profile (20.3%), left (7.9%), frontal (42.4%), right (9.4%), right profile (20.0%)) provided by [4]. Based on the locations of the facial points, we estimate the roll angles of head poses using the POSIT algorithm [57] and discretise them into 3 labels (left tilt, upright, right tilt). We discard the pitch angle because it is difficult to get the ground truth for the face images in the wild. We also annotate the gender status (male, female) for each face image.

2) Forests Description: In order to evaluate the contributions of each component of our methods, we have built 24 forests using variations of the methods and tested on the LFW dataset (I). Below we describe the way in which the different variants are built . *RF-MV* creates the tree in a classical manner and at each leaf node, one single Mean Value model is learned. *RF-MS* also builds the tree in in a classical way but at each leaf node, a single Mean-Shift model instead of mean value model is stored. *PI-RF-MV* and *PI-RF-MS* are created using head pose yaw as privileged information and their leaf node models are the same as *RF-MV* and *RF-MS*. *SORF-MV* and *SORF-MS* are the structured output variants of *RF-MV* and *RF-MS* respectively. Their privileged information-based versions are *PI-SORF-MV* and *PI-SORF-MS* respectively. *CRF-YAW*, *CRF-ROLL* and *CRF-GENDER* are forests that conditional regression models are learned based on corresponding privileged information, head pose yaw angle, roll angle and gender status respectively while their *PI-* counterparts (i.e., *PI-CRF-YAW*, *PI-CRF-ROLL*, *PI-CRF-GENDER*) use privileged information during the tree building process. The following 6 forest, from F15 to F20 are the corresponding versions with additional shape models. All the above forests have the same number of trees (10). Each tree is trained using 1500 randomly sampled face images. The same random number generator is used for the same tree index of all the forests in order to make the comparison fair. Finally we construct four hybrid forests, from F21 to F24, that are used to evaluate the effect of fusing different types of privileged information (see Section IV-D4). F25 shares the same forest from F24, however, during testing, it uses the ground truth privileged information to select the regression model at the leaf node.

In the BioID dataset, we randomly select 400 face images for testing and the remaining 1121 images are used for training. Two different forests are built, each with 10 trees, one (SORF) with structured output while the other not (RF). Each tree is trained using 600 randomly selected images. The structure graph for 20 facial points in BioID dataset is shown in Fig. 3. For this dataset at each leaf node we use the mean shift-based voting scheme.

D. Experimental Results

In what follows we summarize our results and discuss our findings from the experiments performed on the LFW and the BioID datasets. We evaluate the influence of the different components of our models and compare with the state-of-the-art methods.

1) Mean-Value vs. Mean-Shift: As stated in III-B2, we have developed two voting schemes for the base point at each leaf, i.e. Mean-Value Model and Mean-Shift Model. We have conducted experiments on the LFW dataset in order to compare their performance in localizing of the facial points. By comparing the pairs: (F1, F2), (F3, F4), (F5, F6) and (F7 F8) in Table I and in Table II, we conclude that Mean-Shift based voting scheme performs slightly better than Mean-Value model. On average, the difference is around 0.2% in terms of the mean localization error and 1.96% in terms of the successful detection rate. In the remaining experiments we used Mean Shift-based voting.

2) Effect of Privileged Information: In this part we will assess whether: 1) using the information gain on the privileged information as evaluation criterion at some internal nodes leads to better trained trees; 2) using regression model conditioned on the privileged information at leaf nodes is better. We assess the first by comparing forests trained using privileged information with their plain counterparts. We assess the second by comparing forests with conditional models at leaf nodes with their counterparts with single Mean Shift model at leaf node. In Table III we present results with and without using the head yew as privileged information.

TABLE III: Comparison of Mean Error (ME) and Successful Detection Rate (SDR) of forests that using and not head pose yaw privileged information (%).

	Plain Training		PI-Training	
	ME	SDR	ME	SDR
Single Model	base line	base line	$\downarrow$0.20	$\uparrow$1.14
Conditional	$\downarrow$0.62	$\uparrow$3.31	$\downarrow$0.76	$\uparrow$4.14

Furthermore, we assess the usefulness of three types of privileged information separately, i.e. head pose yaw angle, roll angle and gender status. As shown in Fig. 5, learning models conditioned on head pose privileged information considerably outperforms the single model approach. Similar improvements can also be seen in Table I and Table II by comparing the mean error and detection accuracy of *F18*, *F19* with *F6*. The improvement in the mean error when using a conditional model is 0.78% and 0.52% respectively and the corresponding increase in the detection rate is 4.43% and 2.53% respectively. When using gender as privileged information, there is a 0.33%

TABLE I: Mean error of each facial point in LFW dataset (%).

Forest ID	Short Description	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Avrg
F1	RF-MV	9.08	6.87	8.51	8.20	9.58	8.48	6.07	7.67	7.80	7.24	7.95
F2	RF-MS	9.29	6.72	8.23	7.85	9.40	8.23	5.65	7.84	6.95	6.89	7.70
F3	PI-RF-MV	8.35	6.65	8.15	7.78	9.37	8.22	5.93	7.51	7.51	7.12	7.66
F4	PI-RF-MS	8.39	6.28	7.96	7.76	9.44	7.92	5.83	8.09	6.67	6.71	7.51
F5	SORF-G	7.87	6.58	7.82	8.24	9.34	8.24	5.74	7.65	6.86	7.30	7.56
F6	SORF-MS	7.86	6.16	7.72	8.00	9.22	8.01	5.71	7.16	6.53	6.97	7.33
F7	PI-SORF-MV	7.72	6.45	7.61	7.93	9.04	7.96	5.67	7.41	6.77	7.20	7.37
F8	PI-SORF-MS	7.76	6.21	7.47	7.69	8.87	7.79	5.65	7.07	6.46	6.90	7.19
F9	CRF_YAW	7.70	5.30	7.90	7.90	9.40	7.10	5.60	7.50	6.20	6.40	7.10
F10	CRF_ROLL	8.60	5.40	7.80	7.80	10.10	7.60	5.60	7.70	6.70	6.70	7.40
F11	CRF_GENDER	8.10	5.40	8.50	8.10	9.70	8.20	5.70	7.30	7.20	7.00	7.52
F12	PI-CRF-YAW	7.50	**5.20**	7.70	7.60	9.30	6.90	5.50	7.20	6.10	6.30	6.93
F13	PI-CRF-ROLL	7.90	5.40	7.70	7.60	10.10	7.70	5.60	7.50	6.70	6.70	7.29
F14	PI-CRF-GENDER	8.10	5.40	8.40	8.10	9.80	8.00	5.70	7.40	7.10	7.10	7.51
F15	CSORF-YAW	7.00	5.30	7.30	7.60	8.20	6.60	5.30	7.10	6.00	6.50	6.69
F16	CSORF_ROLL	7.80	6.00	7.40	7.70	9.20	7.30	5.30	7.60	6.40	6.80	7.15
F17	CSORF-GENDER	7.80	6.20	8.10	8.30	9.10	7.90	6.00	7.70	6.70	7.30	7.51
F18	PI-CSORF-YAW	6.90	5.30	7.20	7.40	**8.00**	6.40	**5.00**	6.80	6.00	6.50	6.55
F19	PI-CSORF_ROLL	7.30	5.60	7.10	7.50	9.50	7.30	5.30	7.00	6.40	6.60	6.96
F20	PI-CSORF-GENDER	7.80	6.60	8.40	8.40	9.40	8.00	6.00	7.80	6.80	7.40	7.66
F21	PI-CSORF-Y+G	6.88	5.36	7.29	7.51	8.11	6.45	5.05	6.88	6.10	6.59	6.62
F22	PI-CSORF-R+G	6.91	5.47	7.18	7.29	8.87	7.42	5.27	7.00	6.35	6.67	6.84
F23	PI-CSORF-Y+R	**6.79**	5.22	**6.90**	**7.14**	8.10	**6.43**	5.19	**6.70**	**5.91**	**6.18**	**6.46**
F24	PI-CSORF-R+G+R	6.84	5.37	7.37	7.52	8.26	6.60	5.19	6.80	6.12	6.51	6.66
F25	*PI-CSORF-PIGT*	*6.70*	*5.30*	*6.70*	*7.14*	*7.76*	*6.32*	*5.00*	*6.60*	*5.71*	*6.11*	*6.33*

TABLE II: Successful detection rate of each facial point in LFW dataset (%).

Forest ID	Short Description	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	Avrg
F1	RF-MV	72.50	88.00	70.90	70.70	61.40	70.50	87.90	79.00	76.50	78.40	75.58
F2	RF-MS	77.20	88.40	72.30	73.00	63.70	72.90	90.80	79.60	82.00	80.70	78.06
F3	PI-RF-MV	77.20	90.00	73.20	73.00	62.50	71.70	89.20	80.70	78.90	80.00	77.64
F4	PI-RF-MS	79.50	91.60	75.30	74.90	63.20	74.90	90.90	79.60	83.50	80.60	79.40
F5	SORF-MV	79.60	90.40	74.60	70.50	62.50	73.00	91.60	77.50	82.60	78.50	78.08
F6	SORF-MS	81.50	91.00	75.80	73.20	65.40	73.30	91.70	79.10	85.30	80.80	79.71
F7	PI-SORF-MV	80.60	91.10	76.00	73.80	65.90	74.60	91.10	78.20	83.50	78.80	79.36
F8	PI-SORF-MS	81.50	91.30	75.30	74.90	63.20	74.90	90.90	79.60	83.50	80.80	79.59
F9	CRF_YAW	81.90	93.00	74.10	75.60	63.80	81.30	91.30	77.60	88.00	83.90	81.05
F10	CRF_ROLL	82.20	92.70	77.20	77.30	59.90	75.80	91.20	80.40	83.60	81.10	80.14
F11	CRF_GENDER	79.70	92.40	74.20	72.60	59.80	72.30	90.10	79.70	82.20	79.80	78.28
F12	PI-CRF-YAW	81.70	93.80	75.70	76.60	65.60	83.40	91.20	79.30	86.60	85.20	81.91
F13	PI-CRF-ROLL	81.70	92.80	77.50	76.80	62.10	76.70	91.00	80.00	83.60	81.70	80.39
F14	PI-CRF-GENDER	80.30	92.90	68.30	71.20	61.00	68.70	90.30	80.20	82.10	79.00	77.40
F15	CSORF-YAW	83.20	93.60	76.90	76.70	72.40	84.90	93.90	80.40	87.90	83.60	83.35
F16	CSORF_ROLL	80.90	92.30	78.80	75.60	65.80	79.30	93.50	77.90	84.40	80.90	80.94
F17	CSORF-GENDER	80.50	90.80	74.80	71.50	66.60	76.10	92.20	79.80	83.20	78.00	79.35
F18	PI-CSORF-YAW	83.40	**94.30**	78.80	77.20	74.10	**86.20**	94.30	81.70	87.50	83.90	84.14
F19	PI-CSORF_ROLL	83.60	92.90	80.70	78.70	65.20	78.50	94.30	82.70	84.90	80.90	82.24
F20	PI-CSORF-GENDER	79.80	91.40	75.10	71.50	65.20	74.90	91.80	79.80	82.90	78.80	79.12
F21	PI-CSORF-Y+G	83.20	93.60	77.70	75.90	**74.50**	85.00	**94.90**	81.80	87.70	83.30	83.76
F22	PI-CSORF-R+G	84.20	93.10	79.60	79.30	68.20	78.40	94.50	82.90	86.00	82.30	82.85
F23	PI-CSORF-Y+R	**85.10**	94.00	**82.20**	**79.40**	74.00	85.80	94.80	**83.50**	**88.40**	**85.80**	**85.30**
F24	PI-CSORF-R+G+R	84.80	92.80	79.40	76.20	72.40	84.50	94.40	83.00	87.80	83.80	83.91
F25	*PI-CSORF-PIGT*	*85.70*	*95.10*	*83.10*	*80.20*	*74.90*	*87.10*	*95.30*	*84.20*	*89.10*	*86.10*	*86.08*

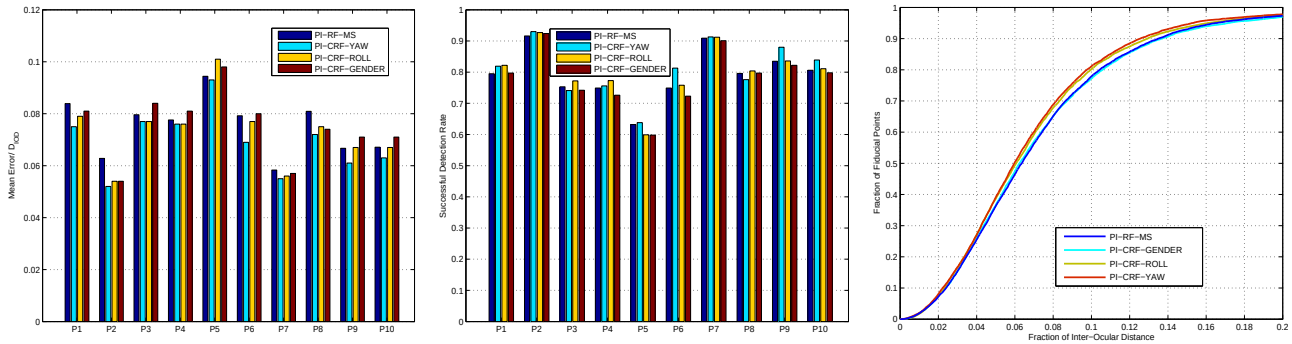
increase of the mean error and a 0.5% drop in the detection rate, however, for some facial points like P1 and P6, forests that use gender privileged information performs better. Further comparisons, as shown in Fig. 5 indicates that the gender privileged information does not have much impact on the model while the other two, i.e. head pose yaw and roll help to improve the performance.

3) Effect of Structured Output: To evaluate the effectiveness of our proposed structured output (SO) method, experiments are conducted both on the BioID dataset and on the LFW dataset. For the experiments in the BioID dataset we used the structured graph with 20 nodes that is illustrated in Fig. 3 while for the LFW with 10 nodes is illustrated in Fig. 4 (right).

TABLE IV: Estimation accuracy of privileged information.

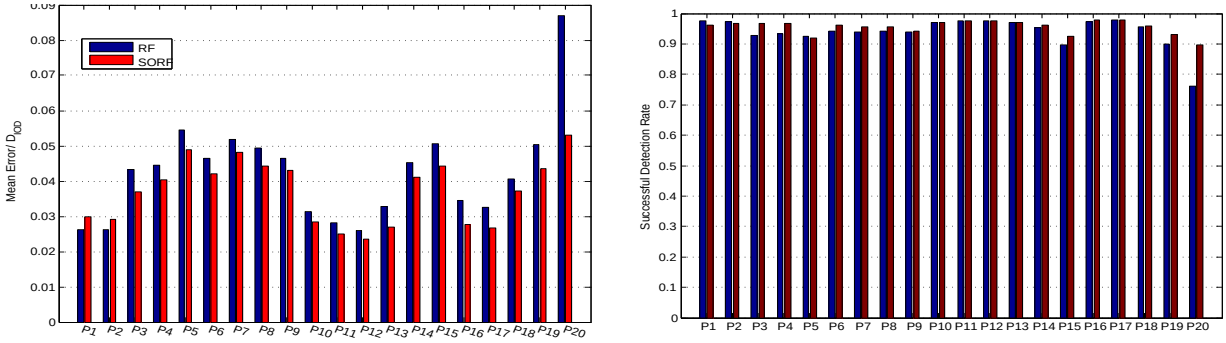
Property	yaw (5 classes)	roll (3 classes)	gender (2 classes)
Accuracy	68.25%	85.10%	87.5%

First, on the BioID dataset, we report the results from SO forests and non-SO forests, i.e. the comparison of the Regression Forest (RF) and Structured Output Regression Forest (SORF) in Fig. 6 in terms of the mean error and the detection rate. The comparison shows that our shape



(a) Mean error (b) Successful detection rate (c) Cumulative Distribution

Fig. 5: Conditional model vs. single model. Some representative results on LFW dataset.



(a) Mean error (b) Successful detection rate

Fig. 6: Overall performance and comparison of RF and SORF on BioID dataset.

model reduces the mean error and increase the successful detection rate for most of the facial points. Particularly, the improvements of the difficult points like the chin point and lower lip centre are more significant. This is expected since these points are not located at intensity edges and therefore there is inherent uncertainty.

We perform several experiments on the LFW dataset, in order to compare SO-forests and non-SO-forests for several variants of our method. The results are shown in Table I and Table II. We show the CDFs of the detection results for some representative forests in Fig. 7. More details can be seen in the tables. The result validates the efficiency of our proposed structured output model in the localization of the facial points.

4) Effect of Privileged Information Fusion: Finally, we perform experiments in which we fuse different types of privileged information. *PI-CSORF-Y+G*, *PI-CSORF-R+G* and *PI-CSORF-Y+R* randomly take trees from two of the corresponding forests, i.e., *PI-CSORF-YAW* (Y), *PI-CSORF-ROLL* (R) and *PI-CSORF-GENDER* (G), 5 from each. *PI-CSORF-R+G+R* randomly takes 3 trees from each of the three corresponding forests. The CDFs of detection accuracy of the hybrid forests are shown in Fig. 8. Except the Y+R combination, the other fusion types have very similar performances, better than *PI-CSORF-GENDER* but not better than *PI-CSORF-YAW* or *PI-CSORF-ROLL*. This implies that the hybrid forests with trees trained based on gender privileged information do not lead to performance improvement. On the contrary, the hybrid forest, *PI-CSORF-Y+R*, with trees from YAW and ROLL forests outperforms both the YAW and ROLL forests.

Finally, we assess the prediction accuracy of the privileged information as shown in Table IV. We can achieve high accuracy in predicting the three types of privileged information. We also note that, F25 is able to achieve the most accurate result, if all the privileged information can be perfectly predicted.

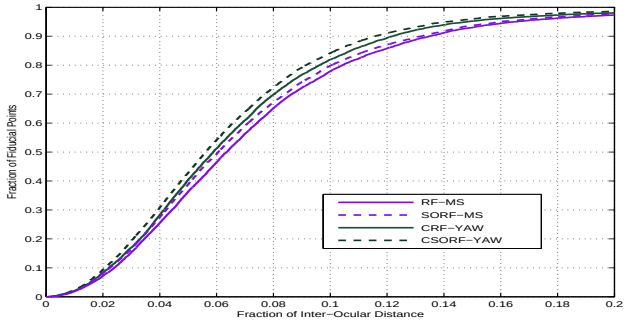


Fig. 7: Representative results from SO forests in LFW dataset, compared with their non-SO counterparts.

5) Run-time Performance: We record the run-time performance on a standard 3.30GHz CPU machine. Our full method

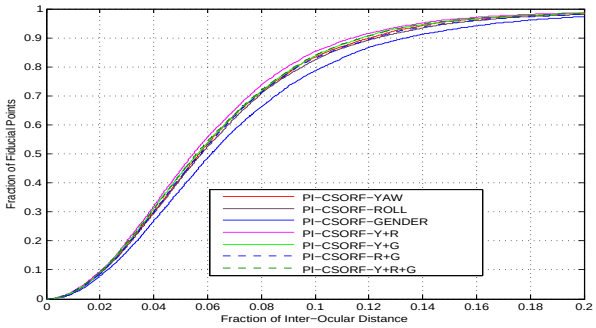


Fig. 8: The performances of hybrid forests on LFW dataset, compared with the original ones.

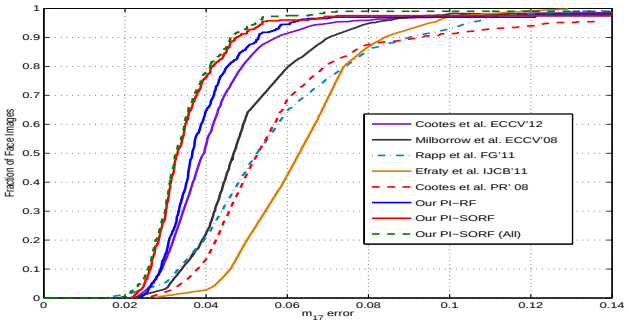


Fig. 9: CDFs of the m_{17} measure on BioID dataset, compared with reported results from [12], [21], [7], [55], [58]

performs on LFW dataset at a average speed of 22 FPS while that of the baseline C-RF method is 25 FPS. Though we have more models at leaf nodes than C-RF, we estimate the privileged information within the forests, which is in contrast to C-RF that uses additional forests to estimate the conditional/privileged information.

E. Comparison with State of the Art

Finally, we compare our proposed methods with state-of-the-art approaches facial point localization on the above mentioned datasets.

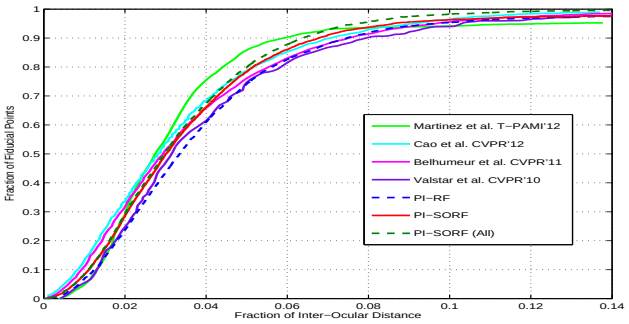


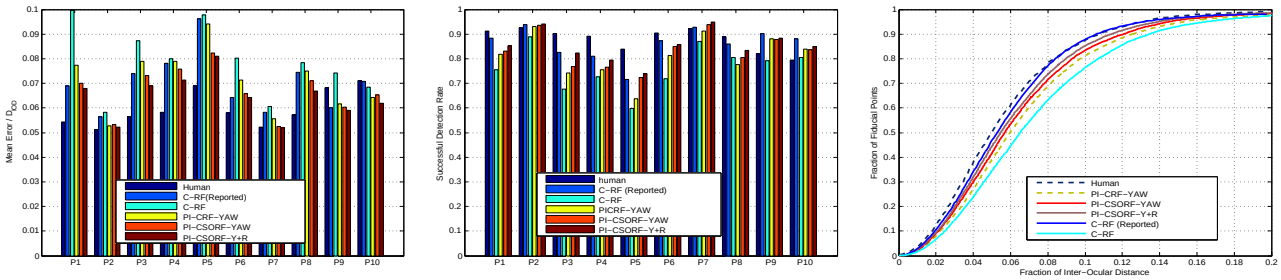
Fig. 10: CDFs over point error on BioID dataset, compared with [26], [3], [6], [2]. For fairness, only 17 internal facial points are used.

1) BioID Dataset: On BioID, we initialize the detection using the OpenCV Viola and Jones face detector. Since related methods that start from the face bounding box have not discussed how they treat the failure cases of face detection (around 10 out of 400), we report the results by 1) manually defining the bounding boxes in the face images in which the face detection failed (the corresponding curves are with "All" label in the figures); 2) treating them as failure cases in facial point detection when calculating the successful detection rate and the cumulative distribution curve. In the literature, two types of curves are used to measure the overall performance. One is the commutative distribution function (CDF) over point error (i.e., fraction of points) and the other is the CDF of m_{17} (i.e., fraction of face images). They are shown in Fig. 9 and Fig. 10 respectively, together with results on the same dataset published elsewhere. As shown in these two figures, our method achieves very promising results on this dataset. Compared to the related method [12] that has applied CLMs on the regression forest voting, our method performs better. This method has validated that its curve shape is consistent with the curve calculated from annotation with simulated Gaussian noise with around 1.5 pixels stand deviation. This implies that the root MSE of our method is smaller than 1.5 pixels. [12] points out the distinctive "S" shape of our curve suggests that the errors in the localization of different points are not correlated. The detection accuracy and the mean error for each of the 20 facial points is shown in Fig. 6.

2) LFW Dataset: We now focus on the more challenging dataset LFW and compare with the regression forest method presented in [4]. We use the publicly available implementation provided by the authors [1]. We have made a minor change, namely we changed the facial point data format from integer to float, in order to have a smoother error distribution. The CDFs of the error is shown in Fig. 11c. Note that the results that we obtained differ from what is reported in [4] possibly because the publicly available trained trees are a reimplementation. Different image features, parameter settings might affect the results. The close-to-human performance reported in [4] requires parameter optimization for each of the facial points and also training more than 10 trees in a sub-forest. The comparison here is based on the same experimental setting, namely the same number of training samples for each tree, the same image features used for training, and the same global parameters of a tree (maximum depth, number of testing candidates at each internal node). In this setting, our model outperforms the C-RF using the same yaw head pose privileged information. Furthermore, by incorporating the structure constraints and fusion of roll head pose information, the performance of our method is very close to human. As shown in Fig. 11a and Fig. 11b, the results are similar to results reported in [4] and very close to human performance.

We note that training our trees is computationally more efficient than training a C-RF. C-RF trains an additional forest for head pose estimation and also one forest for each head pose subset while only one forest is trained in our method. In the public implementation which we compare, 60 trees in

[1] http://www.dantone.me/projects-2/facial-feature-detection/



(a) Mean error (b) Successful detection rate (c) Cumulative distribution

Fig. 11: Overall performance of our method on LFW dataset, compared with [4].

total (10 trees for head pose estimation and 10 for each yaw pose) are built in C-RF while our method use only 10 trees in total. It also means that, many more training samples are used in their model despite a tree is trained using the same number of training samples.

Cao *et al.* [3] have reported results on LFW87 [59] This is a dataset that is not publicly available but which seems to be of similar difficulty. We also list our MRSE (Mean Root Square Error) evaluation metric in Table V in order to give an idea about the relative performance but note that the results are on different datasets with similar characteristics.

TABLE V: Percentages of test images with RMSE(Root Mean Square Error) less than the given thresholds on the LFW dataset, compared to [3], [59] on LFW87 dataset.

RMSE	< 5 pixels	< 7.5 pixels	< 10 pixels
Method in [3]	74.7%	93.5%	97.8%
Method in [59]	86.1%	95.2%	98.2%
PI-CSORF-Y+R	94.4%	96.3%	99.2%

3) LFPW Dataset: We compare our method and the C-RF detector on test images from the LFPW dataset to test whether the learned models can be transferred to a different dataset. Again, the OpenCV Viola and Jones face detector is applied first. The mean error of each facial point is shown in Fig. 12. Although our detector does not perform as well as [6] and [3], the average mean error, around 2 pixels, is very low. It is worth noting that neither our model nor C-RF is trained on LFPW and it is known that the image quality of LFW is much worse than that of LFPW. The performance of our detector and C-RF on LFPW is close to their performance on LFW. When the error fraction is less than 0.1, a detection is regarded as success. We reported the successful detection rate of each facial point in Fig. 13. As it can be seen, for most of the points, the successful detection rate is very high, more than 90%. The mouth corners and the outer lower lip are the most difficult points to localize. In Fig. 14, we show the detection results of our model and of the C-RF detector on some example images from LFPW . As it can be seen, under partial occlusion, both C-RF and our CRF method fail to localize all points at the correct positions since they are both local detectors. On the contrary, CSORF method is able to handle such cases since it takes the structure constraints into account.

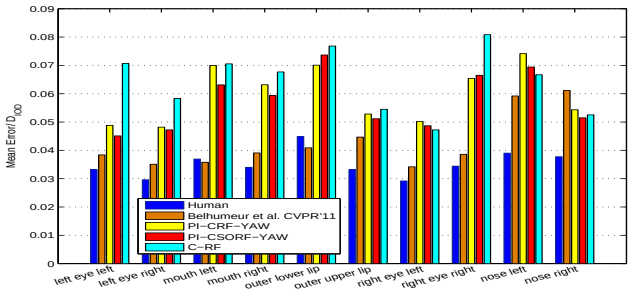


Fig. 12: Mean error of our model on LFPW dataset, compared to C-RF detector from [4].

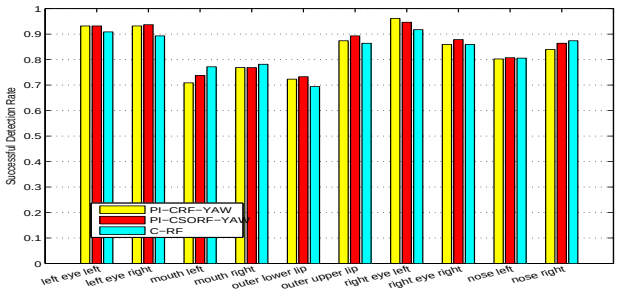


Fig. 13: Successful detection rate of our model on LFPW dataset, compared to C-RF detector from [4].

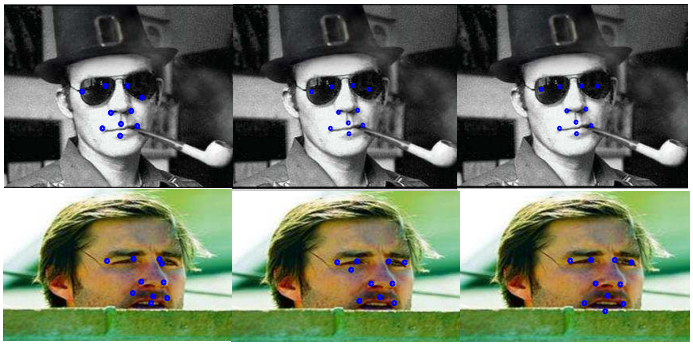


Fig. 14: Example Images from LFPW dataset. First column shows detected facial points by C-RF [4], second column the detection results by PI-CRF-YAW forest and the last column the detected facial points by PI-CSORF-YAW Forest.

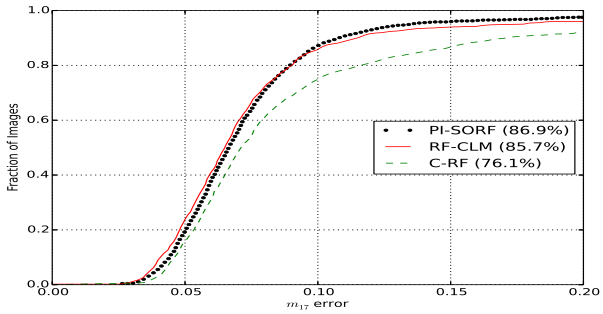


Fig. 15: Comparison to RF based methods (C-RF [4] and RF-CLM [12]) on AFLW.

4) AFLW Dataset: Finally we show the performance on the AFLW dataset and compare to recent Regression Forests based methods including the baseline C-RF [4], RF-CLM [12] (RF combined with CLM) as shown in Fig. 15. We select 1000 images from AFLW for testing and the rest of them for training the forests and repeat this process for four times and we report the average results as shown in Fig. 15. Our proposed method performs significantly better than the baseline RF and on par with the RF-CLM, which has explicitly shape models. Our method is able to further combine with other shape models for performance boost.

F. Sensitivity to Face Bounding Box Shift

In recent years, the cascaded methods have shown promising results in facial points detection. However, compared to the local-based methods as ours, they are more sensitive to initialization, which is often calculated from face detection. It is because the features are extracted around the initial estimate of the landmarks. Applying a different face detector influences the results of the cascaded methods - this is evident by the fact that [3], [46] rely on multiple initializations or to the so called 'smart starts'. By contrast, the method is this paper is a local-based one that does not rely on any initialization shape: patches from all over within the bounding box will be used and the RF will decide which ones will vote for which landmark. This decision is based on the patches appearance and not on their distance from a shape. Indeed, regions near the facial points give better predictions however such information (i.e. the true distance) is not known at test time. When the bounding box shifts due to an inaccurate face detection, then some patches fall out of the new bounding box, and some new patches fall in. However, all the patches that are in the intersection of the old and the new bounding box will vote in exactly the same way. This makes local methods more robust.

We perform experiments on LFW to demonstrate this. We apply the state of the art cascaded method, SDM [43] on the same test images from LFW that we have used and report the results of their common facial points. We shift the face Bounding Box (BB) by 5% to 20% and the results are as follows: Though SDM and our method have similar results given the ground truth face bounding box, when face bounding box shifts, the performance of SDM drops rapidly. On the contrary, until the shift is very huge (20%) and results in some facial points obviously fall out of the face bounding box, our method is fairly robust to the bounding box shifts.

Bounding Box shift	0%	5%	8%	10%	20%
SDM Mean Error	6.45	7.71	15.56	22.57	40.36
Our Mean Error	6.46	6.48	6.51	6.70	9.20

TABLE VI: SDM [43] vs. our method when face BB shifts.

V. Conclusion

In this paper, we have presented a novel method called privileged information based conditional structured output regression forest (PI-CSORF) and have applied it in the problem of facial point detection. We show how to utilize privileged information, i.e. information that is available only during training and how to incorporate structure information within the regression forests.

Extensive experimental evaluations on facial point detection on face images from both controlled and uncontrolled environments show the advantages of the proposed methods. We demonstrate state-of-the-art performance on the BioID dataset. On more challenging datasets (LFW, LFPW and AFLW) that consist of images that exhibit greater variability, our method considerably outperforms the recent conditional regression forest method and other regression forests related methods using the same experimental setting, despite the fact that we use much fewer training images and trees.

Although it does not perform better when comparing to the recent holistic-based methods, our method follows very different setting and we believe its advantage will be useful in certain circumstance for instance when the initialization is not reliable for cascaded holistic method. Also since our method does not use any explicit shape models, the performance can be boosted if we combine our method with the state of the art shape models like the CLM or the mixture of tree model.

References

[1] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, 2001.

[2] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[3] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face Alignment by Explicit Shape Regression," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[4] M. Dantone, J. Gall, G. Fanelli, and L. Van Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012, pp. 2578–2585.

[5] X. Zhu and D. Ramanan, "Face detection, pose estimation and landmark localization in the wild," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[6] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011.

[7] B. Efraty, C. Huang, S. K. Shah, and I. A. Kakadiaris, "Facial landmark detection in uncontrolled conditions," in *Biometrics (IJCB), 2011 International Joint Conference on*, 2011.

[8] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011.

[9] C. Leistner, A. Saffari, J. Santner, and H. Bischof, "Semi-supervised random forests," in *Proc. IEEE Conf. Computer Vision*, 2009, pp. 506–513.

[10] M. Sun, P. Kohli, and J. Shotton, "Conditional regression forests for human pose estimation," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2012.

[11] V. Vapnik and A. Vashist, "A new learning paradigm: Learning using privileged information," *Neural Networks*, vol. 22, no. 5, pp. 544–557, 2009.

[12] T. F. Cootes, M. C. Ionita, and S. P., "Robust and Accurate Shape Model Fitting using Random Forest Regression Voting," in *Proc. European Conf. Computer Vision*, 2012.

[13] H. Yang and I. Patras, "Face parts localization using structured-output regression forests," in *Proc. Assian Conf. Computer Vision*. Springer, 2012.

[14] ——, "Privileged information-based conditional regression forests for facial feature detection," in *Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, 2013.

[15] G. Fanelli, J. Gall, and L. Van Gool, "Real time head pose estimation with random regression forests," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2011, pp. 617–624.

[16] R. Girshick, J. Shotton, P. Kohli, A. Criminisi, and A. Fitzgibbon, "Efficient regression of general-activity human poses from depth images," in *Proc. IEEE Intl Conf. Computer Vision*, 2011.

[17] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. Van Gool, "Random forests for real time 3D face analysis," *Int'l J. of Computer Vision*, pp. 1–22, 2012.

[18] A. Criminisi, "Decision forests: A unified framework for classification, regression, density estimation, manifold learning and semi-supervised learning," *Foundations and Trends in Computer Graphics and Vision*, vol. 7, no. 2-3, pp. 81–227, 2011.

[19] D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using Gabor feature based boosted classifiers," in *Systems, Man and Cybernetics,IEEE International Conference on*, 2005.

[20] C. T. Liao, Y. K. Wu, and S. H. Lai, "Locating facial feature points using support vector machines," in *Cellular Neural Networks and Their Applications, 2005 9th International Workshop on*, 2005, pp. 296–299.

[21] V. Rapp, T. Senechal, K. Bailly, and L. Prevost, "Multiple kernel learning SVM and statistical validation for facial landmark detection," in *Proc. IEEE Int'l Conf. on Automatic Face and Gesture Recognition*, 2011.

[22] C. Du, Q. Wu, J. Yang, and Z. Wu, "SVM based ASM for facial landmarks location," in *Proc. IEEE Int'l Conf. Computer and Information Technology.*, 2008, pp. 321–326.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[24] H. K. Galoogahi, T. Sim, and S. Lucey, "Multi-channel correlation filters," in *Proc. IEEE Intl Conf. Computer Vision*, 2013.

[25] D. Cristinacce and T. Cootes, "Boosted regression active shape models," in *Proc. British Machine Vision Conference*, vol. 2, 2007, pp. 880–889.

[26] B. Martinez, M. Valstar, X. Binefa, and M. Pantic, "Local Evidence Aggregation for Regression Based Facial Point Detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2012.

[27] T. F. Cootes, C. J. Taylor, D. H. Cooper, J. Graham, and Others, "Active shape models-their training and application," *Computer vision and image understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[28] D. Cristinacce and T. Cootes, "Feature detection and tracking with constrained local models," in *Proc. British Machine Vision Conference*, 2006.

[29] J. M. Saragih, S. Lucey, and J. F. Cohn, "Face alignment through subspace constrained mean-shifts," in *Proc. IEEE Intl Conf. Computer Vision*, 2009.

[30] X. Yu, J. Huang, S. Zhang, W. Yan, and D. N. Metaxas, "Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model," in *Proc. IEEE Intl Conf. Computer Vision*, 2013.

[31] X. Zhao, S. Shan, X. Chai, and X. Chen, "Cascaded shape space pruning for robust facial landmark detection," in *Proc. IEEE Intl Conf. Computer Vision*, 2013.

[32] X. Tan, F. Song, Z. H. Zhou, and S. Chen, "Enhanced pictorial structures for precise eye localization under incontrolled conditions," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1621–1628.

[33] Y. Wu, Z. Wang, and Q. Ji, "Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[34] F. Zhou, J. Brandt, and Z. Lin, "Exemplar-based graph matching for robust facial landmark localization," in *Proc. IEEE Intl Conf. Computer Vision*, 2013.

[35] H. Yang and I. Patras, "Sieving regression forests votes for facial feature detection in the wild," in *Proc. IEEE Intl Conf. Computer Vision*, 2013.

[36] S. Cocsar and M. Çetin, "A graphical model based solution to the facial feature point tracking problem," *Image and Vision Computing*, vol. 29, no. 5, pp. 335–350, 2011.

[37] L. Liang, F. Wen, Y. Q. Xu, X. Tang, and H. Y. Shum, "Accurate face alignment using shape constrained Markov network," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006, pp. 1313–1319.

[38] R. Gross, I. Matthews, and S. Baker, "Generic vs. person specific active appearance models," *Image and Vision Computing*, vol. 23, no. 12, 2005.

[39] F. D. l. T. Enrique Sánchez-Lozano and D. González-Jiménez, "Continuous Regression for Non-Rigid Image Alignment," in *Proc. European Conf. Computer Vision*, Oct. 2012.

[40] J. Saragih and R. Goecke, "A nonlinear discriminative approach to aam fitting," in *Proc. IEEE Intl Conf. Computer Vision*, 2007.

[41] P. A. Tresadern, P. Sauer, and T. F. Cootes, "Additive update predictors in active appearance models." in *Proc. British Machine Vision Conference*, 2010.

[42] G. Tzimiropoulos and M. Pantic, "Optimization problems for fast aam fitting in-the-wild," in *Proc. IEEE Intl Conf. Computer Vision*, 2013.

[43] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[44] P. Dollar, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2010.

[45] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2013.

[46] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *Proc. IEEE Intl Conf. Computer Vision*, 2013.

[47] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Enhanced pictorial structures for precise eye localization under incontrolled conditions," in *Proc. European Conf. Computer Vision*, 2012.

[48] X. Zhao, T.-K. Kim, and W. Luo, "Unified face analysis by iterative multi-output random forests," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1765–1772.

[49] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001.

[50] S. Nowozin, "Improved information gain estimates for decision tree induction," *Proc. Int'l Conf. Machine Learning*, 2012.

[51] P. Grassberger, "Entropy estimates from insufficient samplings," *Arxiv preprint physics/0307138*, 2003.

[52] O. Jesorsky, K. Kirchberg, and R. Frischholz, "Robust face detection using the hausdorff distance," in *Audio-and Video-Based Biometric Person Authentication*. Springer, 2001.

[53] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments," University of Massachusetts, Amherst, Tech. Rep. 07-49, Oct. 2007.

[54] M. Kostinger, P. Wohlhart, P. M. Roth, and H. Bischof, "Annotated Facial Landmarks in the Wild: A Large-scale, Real-world Database for Facial Landmark Localization," in *Proc. IEEE Intl Conf. Computer Vision Workshops*, 2011, pp. 2144–2151.

[55] D. Cristinacce and T. Cootes, "Automatic feature localisation with constrained local models," *Pattern Recognition*, vol. 41, no. 10, pp. 3054–3067, 2008.

[56] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, 2004.

[57] D. F. DeMenthon and L. S. Davis, "Model-based object pose in 25 lines of code," *Int'l J. of Computer Vision*, vol. 15, no. 1, pp. 123–141, 1995.

[58] S. Milborrow and F. Nicolls, "Locating facial features with an extended active shape model," *Proc. European Conf. Computer Vision*, pp. 504–513, 2008.

[59] L. Liang, R. Xiao, F. Wen, and J. Sun, "Face alignment via component-based discriminative search," in *Proc. European Conf. Computer Vision*, 2008.