Structured Semi-supervised Forest for Facial Landmarks Localization with Face Mask Reasoning

Xuhui Jia[1]
xhjia@cs.hku.hk
Heng Yang[2]
heng.yang@qmul.ac.uk
Angran Lin[1]
arlin@cs.hku.hk
Kwok-Ping Chan[1]
kpchan@cs.hku.hk
Ioannis Patras[2]
i.patras@qmul.ac.uk

[1] Department of Computer Science
The Univ. of Hong Kong, HK

[2] School of EECS
Queen Mary Univ. of London, UK

Abstract

Despite the great success of recent facial landmarks localization approaches, the presence of occlusions significantly degrades the performance of the systems. However, very few works have addressed this problem explicitly due to the high diversity of occlusion in real world. In this paper, we address the face mask reasoning and facial landmarks localization in an unified Structured Decision Forests framework. We first assign a portion of the face dataset with face masks, i.e., for each face image we give each pixel a label to indicate whether it belongs to the face or not. Then we incorporate such additional information of dense pixel labelling into training the Structured Classification-Regression Decision Forest. The classification nodes aim at decreasing the variance of the pixel labels of the patches by using our proposed structured criterion while the regression nodes aim at decreasing the variance of the displacements between the patches and the facial landmarks. The proposed framework allows us to predict the face mask and facial landmarks locations jointly. We test the model on face images from several datasets with significant occlusion. The proposed method 1) yields promising results in face mask reasoning; 2) improves the existing Decision Forests approaches in facial landmark localization, aided by the face mask reasoning.

1 Introduction

Accurate semantic facial landmarks localization is a well studied problem in computer vision as it is desirable for many facial analysis tasks including face recognition, facial expressions and facial animation. In recent years, remarkable progress has been made and some of them have reported close-to-human accuracy on face images in the wild [5, 9, 36, 42]. However,

This work was supported by GRF HKU 710412E

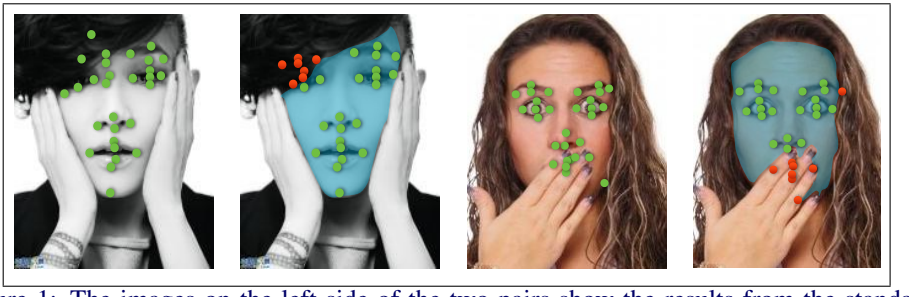

Figure 1: The images on the left side of the two pairs show the results from the standard Random Forests for facial landmarks localization [9], with failure cases under occlusion. The images on the right side of the two pairs show the results of our proposed method. It first explicitly predicts the face mask (the semi-transparent region), then use the face mask information to improve the localization and to predict the occlusion status of the landmarks.

these methods are prone to break down when confronting partial facial occlusions, which occur frequently in realistic scenarios (e.g. the use of scarf or sunglasses, hands or hair on the face). It is intractable to model the occlusion due to its high diversity. Thus very few work has been done [4, 43]. While [43] tried to model a few synthetic occlusion patterns, the recent method of [4] dealt with the occlusion problem in more realistic sceneries. Both of them only focused on modelling the occlusion in an unstructured way, i.e. treating the visibility of each landmark independently. However in realistic conditions, the occlusion patterns (or called occluders) often occupy a continuous region instead of an individual pixel location, as depicted in Fig 1. Thereby the whole occluded region will consistently affect the landmarks localization.

We in this work address the face mask reasoning and facial landmark localization in an unified random Decision Forests (DF) framework. The DF framework has shown powerful and efficient performance in various computer vision tasks such as human pose estimation [10], object detection [28], and facial landmark localization [9, 30, 40], which works in a way that local observations (patches) are extracted at several image locations, propagated into the forests and then cast votes for localization of targets (body joints or facial landmarks). However, as Yang & Patras stated in [40], not all votes from the forest are valid and the invalid votes degrade the localization accuracy. In our observation, these invalid votes are very likely from the occluded facial regions. Therefore, we model patch occlusion status explicitly, in a way similar to semantic image labelling [12, 22, 23], by encoding each pixel with a semantic label, face or non-face in our case. We propose a structured semi-supervised forest framework for face mask reasoning and landmarks localization. Specifically we make the following contributions:

1. We have built a rich face image dataset with face mask annotation. The dataset was built as an extension of the recent datasets: Caltech Occluded Faces in the Wild (COFW), Labeled Face Parts in the Wild (LFPW) and Labeled Face in the Wild (LFW). We manually annotate a portion of images in these datasets with face masks. The face mask indicates whether or not each pixel belongs to the face.

2. We propose a structured semi-supervised joint classification-regression forest with the following properties. First, semi-supervised, it uses training images from the above described augmented dataset, only a portion of which are with face masks. Second, structured, it has a novel structured criterion for split function selection for the pixel labelling (face

mask reasoning) problem. Third, joint classification-regression, it predicts face mask label for each pixel (classification) and the landmark locations (regression) at the same time, and more importantly it uses the face mask reasoning results to improve the accuracy of landmark localization.

The proposed method is evaluated on images with clear occlusions from LFPW, LFW and COFW datasets. The joint framework shows superior results in both facial landmark localization and face mask reasoning, over the existing Decision Forests methods, which were proposed to deal with one of these two tasks.

2 Related work

Facial landmark localization: There is a wide range of literature on facial landmark localization or face alignment. These methods can be roughly grouped into holistic based and local based. A typical holistic based method is the Active Appearance Model (AAM) [6] that regresses the shape (the locations of the landmarks) in an iterative way. At each iteration of AAM, an update of the current model parameters is estimated via a simple linear regression method. Recent methods in this category improve the original AAM by using better optimization strategies or different features for regression in each iteration [27, 33, 36]. A similar framework called Cascaded Pose Regression (CPR) with *random fern* as the primitive regressor has shown very good results [4, 5, 11, 13, 41] in both localization accuracy and efficiency. Convolutional neural networks [31] have been used for this problem as well. Local based methods often consist of local detection, classification [3] or regression [34], and shape constraints modelling, that includes Constrained Local Model (CLM) [2, 8, 19, 27], tree-structured model [42, 44, 46], Restricted Boltzmann Machines [35] and graph matching [45]. Our work is closely related to [7, 9, 38, 39] that use Regression Forest as a local detector for landmarks localization. Although these models achieved certain degree of success, they all struggle under partial face occlusion.

Face Mask Reasoning: Face mask reasoning, or explicit occlusion detection, has attracted very limited attention and only a few works have been proposed [26, 37, 43]. These works simulated only a few common occlusion patterns such as sunglasses, scarf and hands. The diversity of such synthetic patterns is far less than that in real world. A recent work [4] uses occlusion annotations in images collected from the real world when training a cascade of regressors (CPR). During testing it estimates the location of the landmarks and, for each one an occlusion label, that is, whether it is visible or not. All these methods treat the occlusion on landmarks in an independent way. However, occluders in real scenarios always cover a continuous region. Thereby, we in this work treat the face mask reasoning in a way similar to image labelling, i.e., to predict each pixel whether it belongs to a face or not. Our work is related to [17, 22, 23], which demonstrate the efficiency and effectiveness of random forests for semantic image labelling. Our approach is also related to the works for face parsing [24, 29] in terms of splitting the face into several non-overlapped regions and to [25] in terms of measuring the relevance of image observation for the target regression.

3 Methodology

Our structured semi-supervised forest performs classification and regression on their corresponding domains in one estimator as we believe these two tasks are mutually dependent. We

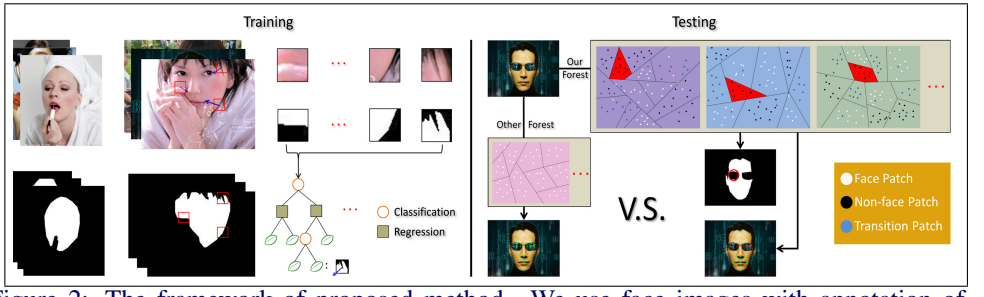

Figure 2: The framework of proposed method. We use face images with annotation of facial landmarks and face masks for training. By randomly switching the information gain function at the internal nodes, the decision trees are optimized with respect to both the offsets to landmarks (regression) and to the local structured label configuration (classification). The forest model is able to predict the face mask and landmark locations jointly. We exploit the face mask prediction to further improve the landmark localization.

start with a brief introduction of the augmented training data in Section 3.1. Then, we show how we encode both the landmarks locations and structured face/non-face labels within the decision forests in Section 3.2. Finally, we describe the inference procedure in Section 3.3.

3.1 Training data

A forest is an ensemble of trees $\mathcal{T} = \{T_t\}$. Each tree T_t is built on a randomly selected subset of the training images. In our semi-supervised setting, we have a portion of images with face mask labeling and the rest without. We randomly extract a set of training data (patches) from the training images. We denote it by $\mathcal{D} = \{\mathcal{P}_l, \mathcal{P}_u\}$, where $\mathcal{P}_l$ represents the patch extracted from training images with face mask label and $\mathcal{P}_u$ represents that from training images without face mask label. Without loss of generality we denote them by the same form $\mathcal{P} = (\mathcal{I}^{d\times d\times F}, \mathcal{V}^{2\times N}, \mathcal{M}^{d'\times d'})$, where $\mathcal{I}$ is the $d \times d$ sized image patch with F channels of features; $\mathcal{V}$ is a N $2D$ displacement vector from the patch centroid to each of the N facial landmarks; $\mathcal{M}$ consists of the $d' \times d'$ of class labels, i.e., $\mathcal{M} = \mathcal{Y}^{d'\times d'}$. Note that the size d' of label patch may differ from the size d of the image patch. For $\mathcal{P}_u$ where there is no face labels, $\mathcal{M}$ is a null matrix.

3.2 Structured decision forests

In this section, we demonstrate how to encode both the landmarks locations and structured face labels (face mask) in the learning procedure of decision forests. Of particular interest in this work is the case where $x \in \mathcal{X}$ represents input image patch and $y \in \mathcal{Y}$ encode the corresponding image annotation (in our case, $\mathcal{Y} = \mathcal{V} \times \mathcal{M}$, where $\mathcal{V}$ is the landmark offset vector and $\mathcal{M}$ is the face mask). Thus, we have two objectives: first, localization of the landmarks and second, the structured labels of different classes (face or non-face). Similar to the hybrid forests [32], we use two separate types of split nodes that optimize different objective functions. The first type of node is for regression and the second type is for classification.

Specifically, for a given node i and the training set $\mathcal{D}_i \subset \mathcal{X} \times \mathcal{Y}$, the goal is to find the best split function $h(x, \theta_i)$ with parameters $\theta_i = (f, k_1, k_2, \tau)$ from a pool of randomly generated candidates, where f is the feature channel, k_i is the sub-region within patch and τ is the

threshold,

$$h(x,\theta_i) = \begin{cases} 0 & \text{if} \quad x^f(k_1) < x^f(k_2) + \tau \\ 1 & \text{otherwise} \end{cases} . \tag{1}$$

that maximizes an objective function, in our case the information gain:

$$I(\mathcal{D}_i, \mathcal{D}_i^L, \mathcal{D}_i^R) = H(\mathcal{D}_i) - \sum_{j \in L,R} \frac{|\mathcal{D}_i^j|}{|\mathcal{D}_i|} H(\mathcal{D}_i^j). \tag{2}$$

where $H(\cdot)$ is the entropy function. The same procedure is applied recursively on the child nodes, $\mathcal{D}_i^L$ and $\mathcal{D}_i^R$, until a certain stopping criterion is met, for instance when a maximum depth is reached or the information gain or training data size fall below fixed thresholds.

For regression nodes, we need to adapt information gain calculation for continuous variables. In our case for $\mathcal{V}$, the aim is to cast precise votes concerning the landmarks location. Therefore, we follow the class-affiliation method proposed by [9] to measure the uncertainty which is defined as:

$$H_{\mathcal{V}}(\mathcal{D}_i) = -\sum_{n=1}^{N} \frac{\sum_{\mathcal{P} \in \mathcal{D}_i} (c_n|\mathcal{P})}{|\mathcal{D}_i|} log\left(\frac{\sum_{\mathcal{P} \in \mathcal{D}_i} p(c_n|\mathcal{P})}{|\mathcal{D}_i|} \right) \tag{3}$$

$$p(c_n|\mathcal{P}) \propto \exp\left(\frac{|v^n|}{\lambda} \right), \tag{4}$$

where $p(c_n|\mathcal{P})$ indicates the probability that the patch $\mathcal{P}$ is informative about the location of the landmark point n. The class affiliation assignment is based on $|v|$, the Euclidean distance between the patch and the landmark location. The variable λ is used to control the steepness of this function.

For classification nodes, we propose a structured way of calculating the entropy. A standard classification method can only deal with a single (atomic) label per input patch sample. It usually represents the patch center label with a finite set of discrete class labels ($y \in \mathcal{Z}$). Consequently, $H(\cdot)$ is defined as the Shannon entropy

$$H(\mathcal{D}_i) = -\sum_{y} p(y|x) log(p(y|x)) \tag{5}$$

where $p(\cdot)$ is the empirical class distribution estimated from the training set $\mathcal{D}_i$. However, the abandoning the structured labels and making the prediction independently will result in the inconsistency in the output spaces. For our face/non-face labeling problem, the unstructured prediction often results in inconsistent face mask reasoning. So far as $y \in \mathcal{Y}^{d' \times d'}$ is concerned, we face two main challenges: 1) information gain over structured label space is not well defined. 2) structured labels are often of high dimension, complex and prohibitively expensive to score numerous split candidates.

Inspired by recent works [17], we define a structured criterion for split function selection. We first discretize the structured labels by partitioning the label spaces, that is inspired by the structured edge detection work of [12]. We utilize a two-stage approach. First we map the structured space to an inter-median space $\mathcal{B}$, $\mathcal{Y} \rightarrow \mathcal{B}$. Then we map the space $\mathcal{B}$ to a discrete label space $\mathcal{Z}$, $\mathcal{B} \rightarrow \mathcal{Z}$. More specifically, $\mathcal{B} = \Pi(\mathcal{Y})$ is a long binary vector that

encodes whether every pair of pixels in $\mathcal{Y}$ belong to the same or different labels, such that we can approximately estimate the dissimilarity of $\mathcal{Y}$ by computing the hamming distance in space $\mathcal{B}$. Considering $\mathcal{B}$ may be high dimensional ($C_2^{d'\times d'}$ for a patch with $d'\times d'$ structured labels), dimensionality reduction is required for efficient computation. We first use a distinct and reduced mapping $\Pi_{\delta_i}:\mathcal{Y}\rightarrow\mathcal{B}$. Instead of using all pairs, we randomly generate m dimensions of $\mathcal{B}$, which is parametrized by δ_i and applied to the training set $\mathcal{D}_i$ at each node i. This not only contributes to fast computation but also introduces randomness into the learning process at the node level. After that, Principal Component Analysis (PCA) [20] is applied to further project the reduced $\mathcal{B}$ to T dimensions.

Finally we map the entry in space $\mathcal{B}$ to a label in space $\mathcal{Z}=\{1,...,k\}$, such that labels with similar $b\in\mathcal{B}$ are assigned to the same discrete labels z. We quantize b based on the top $log_2(k)$ PCA dimensions, assigning b a discrete label z according to the orthant (generalization of quadrant) into which b falls. To this end, mapping the structured label to space $\mathcal{Z}$ allows us to use the standard information gain criterion based on Shannon entropy as defined in Eq. (5). In practice, we use Π_δ with dimension $m=256$ and the discrete labels with $k=2$. In fact, even an approximate distance measure for $\mathcal{Y}$ like this suffice to train effective decision forests classifiers [16]. We note that, in our semi-supervised setting, there are both $\mathcal{P}_l$ and $\mathcal{P}_u$, thus at classification nodes, the information gain only evaluated on the data with labelled face mask. The entire learning procedure will be greatly benefited from the contribution of the ones with unlabelled mask at regression nodes.

Leaf Models. As in Hough forests [15], we assign certain levels of depth in the tree a fixed type of evaluation objective. We thus introduce a steering parameter γ which indicates from first levels up to depth $|\gamma|$, only those regression nodes are evaluated, such that the visual feature variation due to displacements to the facial points is first removed at top levels. Then, starting with depth $|\gamma|$ of the trees, classification nodes and regression nodes are selected randomly. Therefore, image patches reach one leaf node tend to have similar offsets to the facial points and exhibit similar structured face/non-face labels.

At each leaf node, e.g. leaf node l, we calculate: (i) the relative offsets to each facial point $O_l^n=(\Delta_l^n,\omega_l^n)$, similar to [14], where Δ_l^n is the mean value and $\omega_l^n=\frac{1}{trace(\Sigma_l^n)}$ with Σ_l^n the covariance matrix of the offsets to the nth facial landmark; (ii) a structured label y_l of size $d'\times d'$ based on $\mathcal{D}_l$ ($\mathcal{D}_l\subset\mathcal{D}$), which is a subset of training data at leaf node l. More specifically, we select the y_l ($y_l\in\mathcal{Y}$) whose value in the inter-median space $b_l\in\mathcal{B}$ is the medoid, i.e. the b_l that minimizes the sum distance to all other b in $\mathcal{D}_l$. This is equivalent to $min_l\sum_m(b_{lm}-\bar{b}_m)^2$, where $\bar{b}$ is the mean vector of all b in $\mathcal{D}_l$. We denote by $f_t^{\mathcal{C}}(x)$ the classification output of tree t cast by x and by $f_t^{\mathcal{R}}(x)$ the regression output.

3.3 Face mask reasoning and landmark localization

At testing time, image patches $x\in\mathcal{X}$ are densely extracted with a stride s and fed to the forest until they reach leaf nodes, where votes are cast for both the localization of facial points and the patch face/non-face label prediction. As opposed to standard classification algorithms, our classifier $f_t^{\mathcal{C}}(x)$ cast a prediction for the center pixel, as well as its neighbouring pixels. Hence, a predicted face mask $\mathcal{M}_p$ is obtained for each test image in a similar way to [12]. Specifically, each pixel gets $d'\times d'\times T/s^2$ predictions, where T is the number of trees and s is the stride size. Then we merge the multiple predictions by a simple average fusion to get the final face mask prediction.

Meanwhile, given the regression outputs of the forest, we can accumulate the Hough

score for facial landmark n as follows. Denote each image patch x_y by its location y, which ends in a set of leaf nodes $\mathcal{L}_{x_y}$ in the forest.

$$S(\hat{y}^n) \propto \sum_{x_y} \sum_{l \in \mathcal{L}_{xy}} \omega_l^n \exp\left(-\| \frac{\hat{y}_n - (y + \Delta_l^n)}{h^n} \|_2^2\right) \cdot \delta_1(f(\Delta_l^n) > \lambda_n) \cdot \delta_2(\mathcal{M}_p(y) > \tau') \quad (6)$$

where h^n is a learned per-point bandwidth. $f(\Delta)$ is the proximity metric defined in Eq. (4). $\delta_i(\cdot)$ is the Dirac delta function. $\delta_1(\cdot)$ only allows votes which fulfil the proximity test, using the proximity threshold λ^n.

The face mask term $\mathcal{M}_p(y)$ differentiates our method from the existing works as we believe that the patches from face region and non-face region contribute differently to the facial landmarks localization. The Dirac delta function $\delta_2(\cdot)$ isolates the effect of votes from non-face region which most likely correspond to the occluders. We note that, by setting $\tau' = 0$, we allow forests to collect votes from the entire image domain, while higher τ' only allows patches from face regions with higher face confidence.

Additionally, the ratio r^n: the sum of votes associated with each facial point n before and after the $\delta_2(\cdot)$ is applied is traced in our work. This is because for heavily occluded facial points, only few valid votes remain after $\delta_2(\cdot)$ is applied, so that the proximity threshold λ_n should be reduced to allow longer distant patches to cast their votes. Such votes essentially introduce stronger facial shape constraint. Finally a mean-shift mode finding algorithm is applied on the Hough map for final facial landmark localization.

4 Experiment

We evaluate the performance of our proposed framework for both landmark localization and face mask labelling on three augmented face image datasets.

COFW: Caltech Occluded Faces in the Wild. This dataset [4] consists of 1007 face images showing heavy occlusion and large shape variations, which was designed to benchmark face landmark algorithms in realistic conditions. All images were hand annotated with 29 landmarks and their corresponding visibility flag as well. Since they were obtained from a variety of sources, the faces are occluded by various patterns (e.g., hands, hats, hair, sunglasses, etc.) in different degrees. We augmented the dataset with a densely labelled face mask associated with each face image that is tightly fit to the face region. Meanwhile, we annotated the global head pose status as in [9], found that 903 out of 1007 images present nearly frontal head pose.

LFW: Labeled Faces in the Wild. This dataset [9, 18] contains low resolution face images of 5479 individuals, 1680 of which have more than one image, exhibiting a large variety of facial appearance as well as general imaging conditions. LFW consists of 13,233 images annotated with 10 landmarks.We provided the face mask annotation for 837 images.

LFPW: Labeled Face Parts in the Wild. This dataset [3] shares only 1300 image URLs on the web, all images annotated using the same 29 landmarks as in COFW but without visibility flag. They were also captured in unconstrained conditions. Only 811 of the 1000 training images and 224 of the 300 test images can be downloaded when we carried out the experiment. We provide 496 images with face mask annotation.

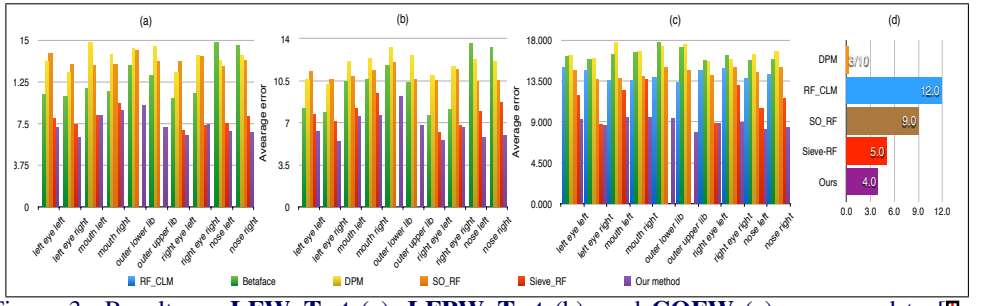

Figure 3: Results on **LFW_Test** (a), **LFPW_Test** (b), and **COFW** (c), compared to [1, 7, 33, 40, 46]. The error is measured as a fraction of the inter-ocular distance. Two lip landmarks are not detected by [40]. **LFW_Test** and **LFPW_Test** only contain 'difficult' image as described in Sec 4.1. (d) shows the run-time performance in fps.

4.1 Implementation details

Due to the performance saturation and the lack of occlusion on the LFPW and LFW dataset [4, 9], we cannot fully exploit the benefits of our face mask prediction and landmark localizations. Therefore, we only report the results on the 'difficult' subsets of LFW and LFPW. We obtained the difficult subsets in a similar way to [40], namely the face images are regarded as difficult if the average point localization error detected by the CRF-D [9] method is greater than 0.1 inter-ocular distance. 237 face images were obtained in the **LFW_Test** and 96 face images in the **LFPW_Test**. The number of resulting images is small due to the fact that face images on these two datasets are relatively easier. Only a few of them either contain occlusion caused by hair or sunglasses, or present large shape variation. We also randomly select 300 images from COFW dataset as test set. We note that, all images from the three test sets were annotated with mask.

We use all the remaining face images from the 3 datasets for model training, which consists of 6781 images and 1603 of them are with face mask labels. Each tree was built using 1200 images(nearly 600 of them with labelling mask) and 100 patches were extracted from each image with no labelling mask and 250 from the ones with labelling.

To build our forest model, we use similar experimental settings to [9] such as the face bounding box size, bandwidth parameter (6) and proximity threshold (6). Some other parameters are as follows: image patch size ($d = 24$), label patch size ($d' = 12$), 37 channels of image features (1 gray scale, 4 HOG-like features, 32 Gabor features), face confidence threshold ($\tau' = 0.78$). The macro forest parameters are: number of trees 10, steering parameter $\gamma = 7$, minimum number of samples 8, maximum depth 25.

4.2 Results for landmark localization

We compare our method with the recent Decision Forest methods for facial feature detection. They are Structured-Output Regression Forests (**SO_RF**) [33], Regression Forests with Constrained Local Model (**RF_CLM**) [7] and Regression Forests Sieving (**Sieve_RF**) [40]. We use the same experiment setting (image data, image feature and macro parameters of the forest) to re-train the Decision Forest models for SO_RF and RF_CLM. We use the trained model of Sieve_RF since the code is publicly available only recently and their model cannot detect the lip landmarks. We also compare the representative DPM+tree structure method

Method	BaselineRF	BalineRF +CRF	StructureRF +Simple Fusion	FullRF +Simple Fusion	FullRF Opt. Sel.	Ours
Global	68.8	81.7	73.6	74.8	78.8	83.9
Avg (face)	71.2	86.6	74.2	75.1	81.7	88.6

Table 1: Face mask reasoning results on the COFW dataset, compared to the related methods.

[16] (**DPM**) and a commercial system (**Betaface**) [1].

Fig. 3 shows the results of the 10 common facial landmarks on all three datasets. Our proposed method achieved better performance than the other methods on 'difficult' images from LFW LFPW and the challenging COFW datasets, despite the fact that all the benchmark Decision Forest methods have used shape models, explicitly (SO_RF and RF_CLM) or implicitly (Sieve_RF) while our method only works as a local detector. On the COFW dataset, the performance of our method still has a gap to the performances of human, due to the heavy occlusion. Note that we have focused on comparing the Regression Forests voting method proposed in recent years, rather than on producing the best facial landmarks detector as we aim to validates the effectiveness of our proposed scheme, i.e., to select reliable patches from face regions based on face mask prediction. As our method is still a local detector, it can be naturally further combined with face shape models, for instance it can be combined with CLM in a way similar to [7], in order to further boost the performance.

Our predicted face mask can intuitively reason the occluded regions on a face image, rather than just checking the visibility [4] of individual pixel. We propose a more reasonable method for landmarks visibility detection. We calculated the occlusion ratio over a small region (within 0.2 inter-ocular distance) surrounding the estimated landmark location, and obtained a 80/57% precision/recall for landmark visibility prediction, which is much better than 80/40% reported in [4].

4.3 Results for face mask reasoning

In this section, we evaluate face mask reasoning performance of our method on the COFW dataset. We compare to the methods that are used for general scene parsing: 1) the standard random forest which yield independent prediction (denoted by **Baseline RF**); 2) standard random forest + conditional random field post-processing (**BaselineRF+CRF**) [21]; 3) three structured forest variants from [22], namely: the **StructureRF+Simple Fusion**, the **FullRF+Simple Fusion** and the **FullRF+Optimized Selections**, all of which yield structured outputs. We followed the evaluation criteria as used in [22]. Specifically, two measurements are reported: *'Global'*, that refers to the percentage of all pixels that were correctly classified; *'Avg(face)'* that expresses the average recall over all classes (face and non-face).

We show the results in Table 1. First, we can clearly see a big margin between the standard RF and structured approaches, which enforce spatial consistency and yield plausible local configuration. Second, our structured approach outperforms the FullRF+Optimized selection and RF+CRF in terms of both *'Global'* and *'Avg(face)'*. The gain in performance validates the effectiveness of our proposed structured information gain criterion and the usefulness the joint classification and regression framework. Some results are shown in Fig 4.

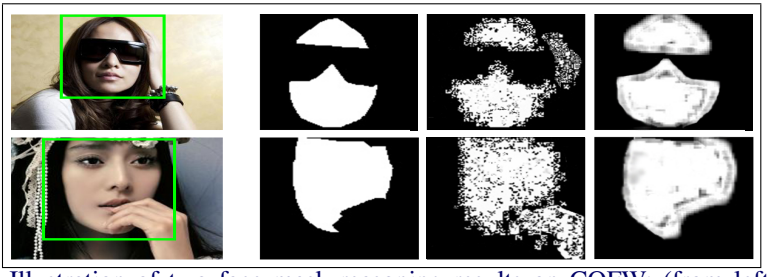

Figure 4: Illustration of two face mask reasoning results on COFW: (from left to right) original image, ground truth, result of the standard RF and result of our proposed method.

5 Conclusion

A structured semi-supervised forest model is presented in this paper for joint face mask reasoning and facial landmark localization under occlusion. We augmented a portion of training images with densely manually-labelled face masks that are used for structured output learning, based on our proposed structural information gain criterion. Experiments show that the proposed framework achieves accurate and spatial consistent face mask prediction, which further assists the landmark localization. We have focused on comparing to the Regression Forests based method and show competitive performance in both tasks. As our method is still a local facial feature detection, we believe that it could be incorporated into a range of model matching frameworks for facial landmarks localization performance boost.

Acknowledgement

The work of Xuhui Jia and Kwok-Ping Chan is supported by GRF HKU 710412E. The author Heng Yang would like to thank a CSC/QMUL joint scholarship and Yichi Zhang for discussion.

References

[1] http://www.betaface.com/.

[2] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, 2013.

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

[4] Xavier P Burgos-Artizzu, Pietro Perona, and Piotr Dollár. Robust face landmark estimation under occlusion. In *ICCV*, 2013.

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

[6] T F Cootes, G V Wheeler, K N Walker, and C J Taylor. View-based active appearance models. *Image and Vision Computing*, 20(9-10), 2002.

[7] T F Cootes, M C.Lindner Ionita, and Sauer P. Robust and Accurate Shape Model Fitting using Random Forest Regression Voting. In *ECCV*, 2012.

[8] D Cristinacce and T Cootes. Feature detection and tracking with constrained local models. In *BMVC*, 2006.

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

[10] Matthias Dantone, Juergen Gall, Christian Leistner, and Luc Van Gool. Human pose estimation using body parts dependent joint regressors. In *CVPR*, 2013.

[11] P Dollár, P Welinder, and P Perona. Cascaded pose regression. In *CVPR*, 2010.

[12] Piotr Dollár and C Lawrence Zitnick. Structured forests for fast edge detection. In *ICCV*, 2013.

[13] B Efraty, C Huang, S K Shah, and I A Kakadiaris. Facial landmark detection in uncontrolled conditions. In *IJCB*, 2011.

[14] G Fanelli, J Gall, and L Van Gool. Real time head pose estimation with random regression forests. In *CVPR*, 2011.

[15] Juergen Gall and Victor Lempitsky. Class-specific hough forests for object detection. In *Decision Forests for Computer Vision and Medical Image Analysis*. Springer, 2013.

[16] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine learning*, 2006.

[17] Ben Glocker, Olivier Pauly, Ender Konukoglu, and Antonio Criminisi. Joint classification-regression forests for spatially structured multi-object segmentation. In *ECCV*. 2012.

[18] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[19] Xuhui Jia, Xiaolong Zhu, Angran Lin, and K.P. Chan. Face alignment using structured random regressors combined with statistical shape model fitting. In *IVCNZ*, 2013.

[20] Ian Jolliffe. *Principal component analysis*. Wiley Online Library, 2005.

[21] Vladimir Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 2006.

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

[23] Peter Kontschieder, Pushmeet Kohli, Jamie Shotton, and Antonio Criminisi. Geof: Geodesic forests for learning coupled predictors. In *CVPR*, 2013.

[24] Ping Luo, Xiaogang Wang, and Xiaoou Tang. Hierarchical face parsing via deep learning. In *CVPR*, 2012.

[25] Ioannis Patras and Edwin R Hancock. Coupled prediction classification for robust visual tracking. *T-PAMI*, 32(9):1553–1567, 2010.

[26] Myung-Cheol Roh, Takaharu Oguri, and Takeo Kanade. Face alignment robust to occlusion. In *AFGR*, 2011.

[27] Jason Saragih and Roland Goecke. A nonlinear discriminative approach to aam fitting. In *ICCV*, 2007.

[28] Samuel Schulter, Christian Leistner, Paul Wohlhart, Peter M Roth, and Horst Bischof. Accurate object detection with joint classification-regression random forests. In *CVPR*, 2014.

[29] Brandon M Smith, Li Zhang, Jonathan Brandt, Zhe Lin, and Jianchao Yang. Exemplar-based face parsing. In *CVPR*, 2013.

[30] M Sun, P Kohli, and J Shotton. Conditional regression forests for human pose estimation. In *CVPR*, 2012.

[31] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *CVPR*, 2013.

[32] Danhang Tang, Tsz-Ho Yu, and Tae-Kyun Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *ICCV*, 2013.

[33] Philip A Tresadern, Patrick Sauer, and Timothy F Cootes. Additive update predictors in active appearance models. In *BMVC*, 2010.

[34] M Valstar, B Martinez, X Binefa, and M Pantic. Facial point detection using boosted regression and graph models. In *CVPR*, 2010.

[35] Yue Wu, Zuoguan Wang, and Qiang Ji. Facial feature tracking under varying facial expressions and face poses based on restricted boltzmann machines. In *CVPR*, 2013.

[36] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, 2013.

[37] Fei Yang, Junzhou Huang, and Dimitris Metaxas. Sparse shape registration for occluded facial feature localization. In *AFGR*, 2011.

[38] H. Yang and I. Patras. Face parts localization using structured-output regression forests. In *ACCV*, 2012.

[39] Heng Yang and Ioannis Patras. Privileged information-based conditional regression forests for facial feature detection. In *AFGR*, 2013.

[40] Heng Yang and Ioannis Patras. Sieving regression forests votes for facial feature detection in the wild. In *ICCV*, 2013.

[41] Heng Yang, C. Zou, and Ioannis Patras. Face sketch landmarks localization in the wild. *IEEE Signal Processing Letters*, 2014.

[42] Xiang Yu, Junzhou Huang, Shaoting Zhang, Wang Yan, and Dimitris N. Metaxas. Pose-free facial landmark fitting via optimized part mixtures and cascaded deformable shape model. In *ICCV*, 2013.

[43] Xiang Yu, Fei Yang, Junzhou Huang, and DN Metaxas. Explicit occlusion detection based deformable fitting for facial landmark localization. In *AFGRW*, 2013.

[44] Xiaowei Zhao, Shiguang Shan, Xiujuan Chai, and Xilin Chen. Cascaded shape space pruning for robust facial landmark detection. In *ICCV*, 2013.

[45] Feng Zhou, Jonathan Brandt, and Zhe Lin. Exemplar-based graph matching for robust facial landmark localization. In *ICCV*, 2013.

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