2016 23rd International Conference on Pattern Recognition (ICPR)
Cancún Center, Cancún, México, December 4-8, 2016

The GIST of Aligning Faces

Siqi Yang, Arnold Wiliem, Brian C. Lovell
School of Information Technology and Electrical Engineering
The University of Queensland
siqi.yang@uq.edu.au, a.wiliem@uq.edu.au, lovell@itee.uq.edu.au

***Abstract*—We propose a novel supervised initialization scheme for cascaded face alignment by searching nearest neighbors based on global image descriptors. Unlike existing schemes which resort to additional large training data sets for learning features, our method does not require additional training steps; thus making our method low computational. Moreover, we found that it is sufficient to use a simple low-dimensional global image descriptor that is easy to extract. In particular, in this work we use the GIST features as our global image descriptor. The proposed initialization scheme outperforms existing initialization schemes for face alignment and improves on the state-of-the-art methods on two challenging datasets, 300-W and COFW.**

I. Introduction

Localizing facial landmarks, popularly referred as the face alignment problem, has been extensively studied in recent years. Amongst various face alignment methods, the ones adopting the cascaded regression approach [1]–[4] appear to be more popular as they achieve state-of-the-art results with extremely fast running times.

Unfortunately, it has been shown recently that methods utilizing regression are quite sensitive to poor initialization [1]–[4]. In earlier works, random initialization [2]–[4] and mean shapes [1], [5] are used as the primary initialization schemes. However, if the randomly chosen shape or mean shape initialization is far from the target shape, the final results of cascaded regression can be far from the target shape as well. As such, the regression will need more cascaded stages and/or regressors.

To that end, recent works of Yang *et al.* [6], [7] employed head pose information and estimated landmark location, respectively. More specifically, both works first extract this additional information and then calculate the K nearest neighbor set. The landmarks from this set will become the initialization. The head pose information is extracted using a Convolutional Network (ConvNet) [8] and the estimated landmark location is calculated using a Regression Forest [7]. To extract these, both schemes need to be trained using a set of labeled data. For instance, to train the ConvNet, one requires a set of face images with head pose information as the ground truth. Henceforth, we categorize these as learned feature based supervised initialization.

It has been shown that the learned feature based supervised initialization methods can reduce the randomness of the initial shapes which will lead to significant improvement. Unfortunately, these methods need additional training steps for feature learning and could have high computational complexity. For instance, the ConvNet training may require several days on a couple of powerful Graphical Processing Units (GPUs) [9]. In this work, we explore a relatively novel avenue to design a low-cost supervised initialization scheme which does not require such additional training for learning features and high computational complexity.

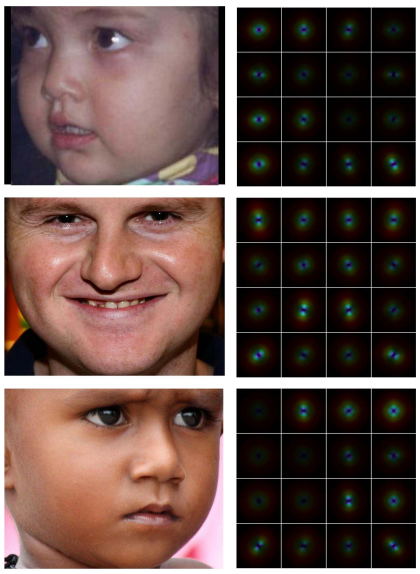

Fig. 1. **GIST features.** The right column shows the global descriptor, GIST features, of the images in the left column. GIST features are computed by convolving the image with multiscale-oriented filters and are displayed by splitting the images into 4×4 grid. The upper row shows that the face pointing to the left has higher magnitude of GIST features in the left regions.

To this end, we also use the K nearest neighbors idea from [6], [7]. However, we propose to use the low-dimensional GIST features proposed in [10]. More specifically, we show that this global descriptor is able to provide relevant spatial information for distinguishing faces in different poses. It is noteworthy to mention that our approach only uses simple descriptors which do not extract dense local descriptors as employed in [4], [7], [11] that requires much higher computational complexity.

To demonstrate the efficacy and adaptability of our proposed GIST initialization scheme, we employ our scheme on several state-of-the-art cascaded regression methods [1]–[4]. From our experiments in two challenging benchmarking datasets: (1) 300-W [12] and (2) COFW [3], we find that the proposed method generally improves on the performance of the state-of-the-art regression methods using random and mean shape initialization. In addition, the proposed approach is on par with head pose initialization which has extremely high computational complexity. The efficacy of our method can be observed from the ability of the global descriptor to capture the pose

978-1-5090-4846-5/16/$31.00 ©2016 IEEE

information (refer to Fig. 1).

We list our contributions as follows:

- We propose a low-cost supervised initialization scheme derived from GIST features, which can determine K initial shapes for the initialization.
- We show that by using a simple global image descriptor it is sufficient to encode information to select the best shapes for the initialization.
- In our evaluation, we show that our proposed initialization can improve on most recent state-of-the-art face alignment methods using random and mean shape initialization; suggesting that our method is generic for face alignment methods.

II. Related Work

Cascaded pose regression methods have gained much attention in recent years due to their excellent performance with low computational footprint. Technically, these methods use multiple cascaded stages of regressors. Given a set of reference landmark points as an input, that is generally represented by shape features, each regressor will use the features to determine the most likely landmark deformation. With a series of deformations produced from the cascaded regressors, the methods can infer the most likely landmark deformation given the initial landmark input. Li *et al.* [2] proposed a method named Explicit Shape Regression (ESR) that extracts shape indexed features from the given shape features and regresses selected discriminative pixel features using random ferns. The landmark deformations are predicted by learning shape indexed features in each regressor. We categorize this method as feature learning based methods. Despite their successes, the method is shown to struggle under occlusions and large shape variations. To this end, the Robust Cascaded Pose Regression (RCPR) method proposed by Burgos-Artizzu *et al.* [3] performs occlusion detection and landmark estimation at the same time. Additionally, they propose a smart restart scheme to avoid bad initializations.

Different from feature learning based methods [2], [3], the Supervised Descent Method (SDM) proposed by Xiong *et al.* [1], models the problem as a general Non-linear Least Square (NLS) optimization problem. In contrast to the feature learning methods, SDM directly regresses the landmark deformations by applying linear regression on the non-linear SIFT descriptors [13]. Following the structure of SDM, Zhu *et al.* [4] builds the face alignment framework based on coarse-to-fine shape searching (CFSS).

Clearly from the above discussions, these methods require initialization. Unequivocally, as shown in [4], [6], [7], the initialization step is indeed an important step for these methods as good initialization positively affects the alignment accuracy. Currently, there are two known initialization schemes: unsupervised and supervised schemes.

The unsupervised scheme does not require additional training steps nor a training dataset. Two popular schemes are: (1) random initialization and (2) mean shape initialization. Methods utilizing random initialization such as [2]–[4] simply randomly select the initial shapes from the available shapes in the training data. However, this scheme may lead a final result far from the target. In the other hand, mean shape initialization may suffer from the local minimum problem in the case of bad initializations [4]. Unfortunately, both initialization schemes are not reliable as they might produce initial shapes that are far away from the target ground truth shapes. To suppress the effects from bad initializations, Burgos-Artizzu *et al.* [3] and Yang *et al.* [14] adopt smart restarts techniques. Unfortunately, the initialization performed from each restart is still random. Hence, this does not fully address the issues caused by the random initialization.

TABLE I
Methods and their properties

Methods	ESR [2]	RCPR [3]	SDM [1]	CFSS [4]
Initialization	random	random	mean pose	random
Features	pixel diff	pixel diff	SIFT	SIFT
Regressor	random ferns	random ferns	linear	linear

Based on the assumption that a more reliable initialization would achieve better performance, several works propose supervised initialization. The initialization scheme of Yang *et al.* [6] is based on head pose information. Specifically, they train a convolutional network (ConvNet) model [8] for automatically extracting the head pose information. Yang *et al.* [7] propose an initialization scheme that uses the estimated landmark locations and their reliability is provided by the local based Regression Forest method. Both methods need ground truth information of facial landmarks for training and the training is enormously expensive. The above mentioned supervised initialization schemes are learned feature based methods.

In our work, our aim is to develop an extremely low computational supervised initialization scheme which utilizes non-learned features and has similar reliability to learned feature based initialization. This will give us two advantages: (1) to avoid issues suffered by the current unsupervised initialization schemes such as random and mean shape initializations and (2) to avoid computational and expensive additional training issues suffered by the learned feature based supervised initialization schemes. Perhaps the closest approach to our proposal is the approach proposed by Hasan *et al.* [11] which utilize non-learned based features. Unfortunately, their work uses a patch-level descriptors. More specifically, they extract HOG features from overlapping patches on the image grid which would significantly increase the computational time. In contrast, we use global image descriptors that are much simpler to compute.

III. Method

In this section, we first briefly present the general framework of the recent cascaded face alignment methods. Then we discuss the details of our proposed initialization scheme and how it can be used to improve the cascaded pose regression methods.

A. General Cascaded Face Alignment

As mentioned above, our method can be used by virtually all cascaded regression methods, such as ESR [2], RCPR [3], SDM [1] and CFSS [4]. These methods share a similar

framework of cascaded pose regression, as summarized in Algorithm 1.

Let shape $\boldsymbol{s} \in \mathbb{R}^{2n}$ be a series of coordinates of n landmarks $\boldsymbol{s} = [x_1, y_1, \cdots, x_n, y_n]$. In general, the cascaded pose regression starts from an initial shape $\boldsymbol{s^0} \in \mathbb{R}^{2n}$ and applies a cascade of T regressors, $\mathrm{R}^{1 \cdots T}$, to refine the shape until the last stage of the regression. Each regressor R^{t} can be designed using various methods such as random ferns [2], [3], random forest [15] and linear regression [1], [4]. At the t-th iteration, features $\boldsymbol{f^t}$ are extracted from the image $\boldsymbol{I} \in \mathcal{R}^{w \times h}$ based on the shape estimated at the previous iteration $\boldsymbol{s^{t-1}}$. Such features are denoted as shape-indexed features. Each regressor R^t takes shape indexed features as input and determines an update, $\triangle \boldsymbol{s} \in \mathbb{R}^{2n}$.

The current shape of the t-th iteration $\boldsymbol{s^t}$ is determined by adding the shape deformation $\triangle \boldsymbol{s}$ to the estimated shape from the previous iteration $\boldsymbol{s^{t-1}}$ via:

$\boldsymbol{s^t} = \boldsymbol{s^{t-1}} + \triangle \boldsymbol{s}$.

Algorithm 1 Cascaded Pose Regression (CPR)

Require: Image $\boldsymbol{I}$, initial shape $\boldsymbol{s^0}$
Ensure: Estimated pose $\boldsymbol{s^T}$
1: **for** $t = 1$ to T **do**
2: Compute the shape indexed features $\boldsymbol{f^t}$ from $\boldsymbol{I}$ based on $\boldsymbol{s^{t-1}}$
3: Apply regressor R^t and get $\triangle \boldsymbol{s}$
4: Update the current pose $\boldsymbol{s^t}$
5: **end for**

Our proposed initialization scheme automatically determines the initial shape, $\boldsymbol{s}^0$; thus, could be used by most existing cascaded pose regression approaches.

B. Initial Shape Determination Problem

The goal of face alignment is to refine the estimated shape $\boldsymbol{s^T}$ as close as possible to the ground truth shape $\boldsymbol{s}^*$. More specifically, the differences between the estimated shapes $\boldsymbol{s^T}$ and the ground truth shapes $\boldsymbol{s}^*$ can be denoted as alignment error e_A. Thus, the goal of face alignment can be regarded as minimizing the alignment error e_A:

$$e_A = \|\boldsymbol{s^T} - \boldsymbol{s}^*\|_2^2 \ . \tag{1}$$

As stated in Algorithm 1, given images and an initial shape $\boldsymbol{s}^0$, the regression based face alignment methods, $\mathrm{CPR}(\cdot)$, produces the estimated shapes $\boldsymbol{s^T}$:

$$\boldsymbol{s^T} = \mathrm{CPR}(\boldsymbol{I}, \boldsymbol{s^0}, \mathrm{R}, T) \ . \tag{2}$$

where $\boldsymbol{I} \in \mathcal{R}^{w \times h}$ is the input image; $\boldsymbol{s}^0$ is the initial shape; $\mathrm{R}(\cdot)$ and T are regressors and the number of cascaded levels, respectively.

Therefore, Equation (1) can be rewritten as:

$$\underset{\boldsymbol{s^0}, \mathrm{R}, T}{\arg\min}\, e_A = \|\mathrm{CPR}(\boldsymbol{I}, \boldsymbol{s^0}, \mathrm{R}, T) - \boldsymbol{s}^*\|_2^2 \ , \tag{3}$$

We note that although the above formulation only considers a single initial shape, it is easy to generalize this into multiple initial shapes problem.

As can be seen from Equation (3) that the initial shape $\boldsymbol{s^0}$, shape indexed features $\boldsymbol{f^t}$, regressors R and the number of regression levels T are factors that can influence the performance of cascaded regression based face alignment. In the light of this fact, the initial shape determination problem primarily aims to find the initial shape, $\boldsymbol{s}^0$ that ultimately reduces the alignment error, e_A:

$$\underset{\boldsymbol{s^0}}{\arg\min} \|\mathrm{CPR}(\boldsymbol{I}, \boldsymbol{s^0}, \mathrm{R}, T) - \boldsymbol{s}^*\|_2^2 \ . \tag{4}$$

The above optimization problem is difficult to study. To that end, we introduce a measure called initialization error e_I. In a similar way, the initialization error e_I is defined as the difference between the initial shapes $\boldsymbol{s^0}$ and the ground truth shapes $\boldsymbol{s}^*$:

$$e_I = \|\boldsymbol{s^0} - \boldsymbol{s}^*\|_2^2 \ . \tag{5}$$

Therefore, the goal of face alignment now is to minimize the initialization error e_I with respect to the initial shapes $\boldsymbol{s^0}$:

$$\underset{\boldsymbol{s^0}}{\arg\min}\, e_I = \|\boldsymbol{s^0} - \boldsymbol{s}^*\|_2^2 \ . \tag{6}$$

However, since it is impossible to obtain the ground truth shape $\boldsymbol{s}^*$ for a given test image $\boldsymbol{I}$, we opt to address Equation (6) by relaxing the problem into:

$$\underset{\boldsymbol{g^0}}{\arg\min} \|\boldsymbol{g}^0 - \boldsymbol{g}^*\|_2^2 \ , \tag{7}$$

where $\boldsymbol{g}^0$ and $\boldsymbol{g}^*$ are the global descriptors extracted from an image from the training set and the testing image, respectively. Notice that the above problem can be solved because in contrast to the ground truth shape, $\boldsymbol{s}^*$ which is impossible to be automatically determined, the $\boldsymbol{g}^*$ can be easily extracted from the test image.

C. GIST Initialization for Cascaded Face Alignment

1) GIST Features: The GIST features are computed by convolving the oriented filters with the image at different orientations and scales. The filters are Gabor filters tuned to 8 orientations at 4 different scales. We divide the image into a 4×4 grid, and compute the average energy of each channel in each grid cell, giving us $8 \times 4 \times 4 \times 4 = 512$ features. In this way, the high-frequency and low-frequency repetitive gradient directions of an image can be measured. Consequently, GIST descriptor is appropriate for selecting matching images that have a similar spatial composition. A visualization of GIST features can be seen from the right columns in Figure 1. The GIST features are displayed in the 4×4 grid and each grid cell shows the GIST features corresponding to 8 orientations and 4 scales. Figure 1 suggests that faces with similar poses would share similar spatial composition. Therefore, we use GIST features as the global descriptors to represent each face.

In order to eliminate the effects of the background, we crop the face by specifying a face bounding box $\boldsymbol{bb} \in \mathbb{R}^4$ around the face. Such a face bounding box is usually provided by a face detector. In addition, in order to be robust to illumination, we normalize the cropped-face by following the methods provided by [16] [17]. We denote the resulting global feature vector $\boldsymbol{g}^* \in \mathbb{R}^{512}$ derived from the normalized cropped-face $\boldsymbol{I_{crop}}$ by

$$\boldsymbol{g}^* = \mathrm{gist}(\boldsymbol{I_{crop}}) \ . \tag{8}$$

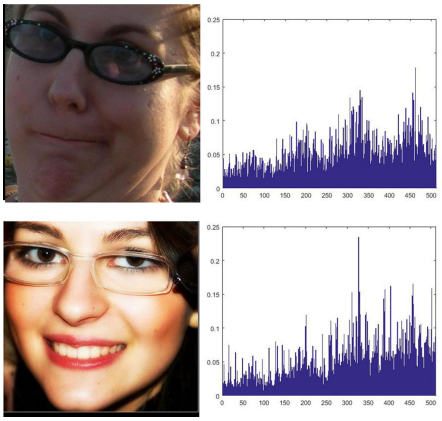

Fig. 2. GIST features of two images with similar head pose information are displayed on the right column. We see that the magnitude may be affected by illumination.

Fig. 3. Query faces (first column) and corresponding four nearest neighbors (columns: 2-5). This figure demonstrates that the nearest neighbors based on GIST features have similar pose to query faces. The 3rd column shows the furthest neighbor of each query, which often has the opposite pose.

2) GIST Initialization: We wish to improve the cascaded regression method by using a low-cost supervised initialization scheme. Our main idea is to utilize the GIST features to encode the head pose information. Assuming that we have a large set of images with their corresponding landmarks, we could simply employ nearest neighbor search in the GIST feature space. This step effectively minimizes the optimization problem presented in Equation (7).

However, from our empirical observations, we found that the ℓ_2 norm used in Equation (7) may not be accurate. This is because, the effect of the illumination between two faces with similar head pose could have high ℓ_2 distance due to the magnitudes of the GIST features. Figure 2 shows some examples. Thus, we opt to use the correlation metric that changes (1) into:

$$\arg\max_{\boldsymbol{g^0}} \operatorname{corr}(\boldsymbol{g}^0, \boldsymbol{g}^*) \ , \tag{9}$$

where $\operatorname{corr}(\boldsymbol{g}^0, \boldsymbol{g}^*)$ is defined via:

$$\operatorname{corr}(\boldsymbol{g}^0, \boldsymbol{g}^*) = \frac{\operatorname{cov}(\boldsymbol{g}^0, \boldsymbol{g}^*)}{\sigma_{\boldsymbol{g}^0}\sigma_{\boldsymbol{g}^*}} \ , \tag{10}$$

where cov means covariance and σ is the deviation. The correlation can predict how the test image and training images are related.

We first extract the GIST features from each image in the training set. Let, $\mathcal{G} = \{\boldsymbol{g}_1 \cdots \boldsymbol{g}_m\}$ be the set of GIST features extracted from each image in the training image. Inspired from [6], [7], [11], in order to increase the initialization reliability, we opt to use K nearest neighbors instead of the nearest neighbor:

$$\{\boldsymbol{s}_k^0\}_{k=1}^K = Knn(\{C_i\}_{i=1}^{i=m}, K) \ , \tag{11}$$

where C_i is defined as:

$$C_i = \operatorname{corr}(\boldsymbol{g}_i, \boldsymbol{g}^*), \boldsymbol{g}_i \in \mathcal{G} \ . \tag{12}$$

In Figure 3 , for each query image on the left, we show the 5 nearest neighbors on the right. From the figure, we can clearly see that the returned faces have similar pose and expressions to the query image to some extent. From this analysis, one can see that similar spatial compositions of images can infer similar shapes of faces.

After the KNN searching, we feed the CPR methods with multiple initial shapes and take the average of the predicted shapes as the final estimate:

$$\boldsymbol{s}^T = \frac{1}{K}\sum_{k=1}^{K} \operatorname{CPR}(\boldsymbol{I}, \boldsymbol{s}_k^0, R, T) \tag{13}$$

The proposed cascaded regression using GIST initialization scheme is summarized in Algorithm 2.

Algorithm 2 GIST Initialization for Cascaded Regression

Require: Query image $\boldsymbol{I}$, face bounding box $\boldsymbol{bb}$, GIST features of the training set, $\mathcal{G}$, number of nearest neighbors K, the cascaded regression function CPR and its pre-trained model including R and T

Ensure: Estimated pose $\boldsymbol{s^T}$

1: $\boldsymbol{I_{crop}} = \operatorname{crop}(\boldsymbol{I}, \boldsymbol{bb})$ ▷ Crop face
2: $\boldsymbol{g}^* = \operatorname{gist}(\boldsymbol{I_{crop}})$ ▷ Extract GIST features
3: $C_i = \operatorname{corr}(\boldsymbol{g}^*, \boldsymbol{g}_i), \boldsymbol{g}_i \in \mathcal{G}$ ▷ Calculate correlation of GIST features between $\boldsymbol{I}$ and each training exemplar
4: $\{\boldsymbol{s}_k^0\}_{k=1}^K = Knn(\{C_i\}_{i=1}^{i=m}, K)$ ▷ Choose multiple initial shapes by KNN
5: $\boldsymbol{s^T} = \frac{1}{K}\sum_{k=1}^{K} \operatorname{CPR}(\boldsymbol{I}, \boldsymbol{s}_k^0, R, T)$ ▷ Compute cascaded pose regression with chosen $\{\boldsymbol{s}_k^0\}_{k=1}^K$ as Algorithm 1

IV. Experiments and Results

We first describe dataset and implementation details. Then the experimental results and discussions are presented.

A. Datasets

In our work, we use two challenging face landmark datasets: (1) 300 Faces in-the-wild (300-W) [12] and (2) Caltech Occluded Faces in the Wild (COFW) [3].

300-W dataset [12] — is created for Automatic Facial Landmark Detection in-the-wild Challenge [12]. This dataset standardizes several popular alignment datasets, including AFW, LFPW, HELEN and XM2VTS with 68-point landmark annotations. In addition, the 300-W dataset contains a new challenging set called iBUG comprising 135 images. In order to compare with the recent methods, we follow the experiment setting of [4]. More specifically, we regard all the training images from HELEN, LFPW and the whole AFW as the training set (3148 images in total). The testing set consists of test images from HELEN (330 images), LFPW (224 images) and images in the iBUG subset (135 images), with 689 images in total.

COFW [3] — is a dataset designed to depict faces in real-world conditions with partial occlusions [3]. Due to differences in pose, expression, hairstyle, and use of accessories or interactions with other objects, the face images show large variation in shape and occlusions. The COFW dataset comprises 1007 images annotated by 29 landmarks. The training set includes 845 LFPW faces [18] + 500 COFW faces, that is 1345 images in total. The test set contains the remaining 507 COFW faces. The 29 landmarks of each image are labeled with their occluded/unoccluded state as well.

B. Implementation Details

We contrast our proposed initialization scheme, GIST initialization, with several recent initialization schemes: (1) random initialization; (2) mean shape initialization; (3) head pose initialization and (4) random forest initialization. We apply these schemes on the state-of-the-art cascaded regression methods such as ESR [2], RCPR [3], CFSS [4], SDM [1]. For RCPR and ESR, for which the code for training is available, we retrain the model on the training images of 300-W using all the 68 facial landmarks. For RCPR, we set the occlusion information as 0 since the occlusion status annotation of 300-W dataset is not available. CFSS provides both the training and test codes, and a pre-trained model on the 300-W dataset as well, so we apply it directly to the test images. For SDM, the original paper only provides the trained models with test codes. Fortunately, since CFSS shares a similar regression method to SDM, the CFSS implementation could be modified to evaluate the SDM performance.

For fair comparisons, all of the methods we evaluate use the same face bounding boxes for both training and testing. We use the face bounding boxes provided by both datasets.

For evaluation, we follow the popular evaluation scheme using the landmark mean error. More specifically, the alignment error and initialization error in (1) and (5) are normalized by the inter-ocular distance, i.e. the Euclidean distance between two eye centers. We do not report the result of CFSS using the mean shape initialization as CFSS requires multiple initial shapes.

In addition, we use default parameters for every method (*i.e.,* we set R and T according to the recommended values by the authors). The only parameter our system has is the number of nearest neighbors, K. In this case, we set $K = 10$ for all evaluated methods.

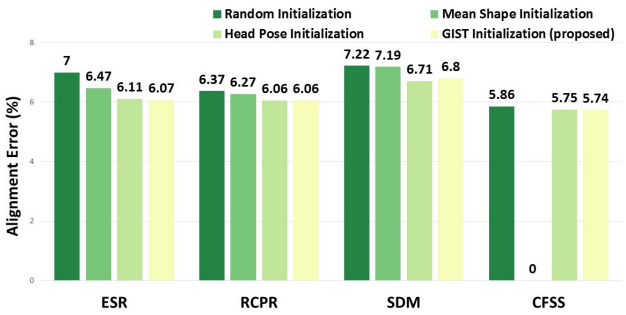

Fig. 4. Results on 300-W dataset, compared with ESR [2], RCPR [3], SDM [1] and CFSS [4].

Our proposed initialization scheme is contrasted with the existing initialization schemes: (1) random initialization; (2) mean shape initialization [1]; (3) head pose initialization [6] and (4) random forest initialization [7]. For head pose initialization, we directly use the ground truth head pose information of their work [6] as the features of the head pose initialization scheme. Based on this, the errors caused by their ConvNet pose estimator [6] may be eliminated.

As COFW dataset offers much more challenging problem related to occlusions, most systems do not perform well. To that end, we only perform the evaluation of RCPR which is specifically designed to address the posed COFW dataset problem. Since the COFW dataset randomly overlaps the face bounding boxes by 80% to simulate the output of a face detector, we use the ground truth bounding boxes for training images and leave the overlapped bounding boxes for test images as a fair comparison. To show that our proposed initialization scheme can be used on the occlusion dataset, COFW, we compare our work with the random forest initialization which is tested in COFW as well. Since head pose initialization [6] doesn't estimate results on COFW, we do not compare with this scheme here.

C. Evaluation on 300-W Dataset

On this dataset, we evaluate the performance of four methods from the cascaded pose regression family 1) the ESR [2]; 2) the RCPR [3]; 3) the SDM [1]; 4) the CFSS [4] using various initialization schemes. The experimental settings are configured as discussed above.

As can be seen from Figure 4, our proposed initialization scheme outperforms the unsupervised schemes such as random and mean pose initializations on all methods. It suggests that our proposed initialization approach can be used for general cascaded regression methods. It is worth noting that the performance of our GIST initialization scheme is competitive with head pose based initialization which requires enormous computing power as well as additional expensive training. This means our proposed initialization scheme can provide competitive results with markedly lower computation. In addition, our initialization scheme does not need the additional training and dataset collection.

TABLE II
INITIALIZATION ERROR ANALYSIS WITH RCPR [3]

	Initialization Error e_I (%)	Alignment Error e_A (%)
Random Initialization	38.56	6.37
Mean Shape Initialization	29.20	6.27
Head Pose Initialization	**18.54**	6.06
GIST Initialization (proposed)	19.24	**6.06**

TABLE III
COMPARISONS ON COFW DATASET WITH RCPR [3]

	Alignment Error e_A (%)
Random Initialization	8.51
Random Forest Initialization	8.62
Mean Shape Initialization	**7.615**
GIST Initialization (proposed)	**7.64**

It can also be seen from the results of ESR and RCPR that with GIST initialization, ESR can obtain a similar result with RCPR without using re-start scheme. It shows that with good initialization, the face alignment methods do not need the help of re-start.

In order to further study the influence of each initialization scheme, we evaluate the initialization error e_I according to (5). We present the results using RCPR in Table 2. We note that the results from other methods are similar. Hence, they are not shown. It can be clearly seen from Table II that when the initialization error e_I decreases, the alignment error e_A will decrease; showing the close relationship between initialization and alignment errors.

D. Evaluation on COFW

As can be seen from Table III, random forest initialization does not perform better than random initialization as reported in their original paper [7]. Whereas, the proposed GIST initialization scheme outperforms both random and random forest initializations. As can be seen, our proposed initialization scheme can also show competitive results on this very challenging dataset which is designed to depict faces in real-world conditions with partial occlusions. When compared with the mean shape initialization, we notice that our GIST initialization can get only a competitive result. It is because we average the occlusion information of the training set for the mean shape initialization, whereas our initialization scheme only provides head pose information without considering the occlusions. Additionally, in our case, the global image descriptor is used as feature space for KNN. Occlusions tend to modify this feature space such that similar head poses might not be close in the feature space, providing bad initializations.

E. Run-time Analysis

We record the run-time performance of our GIST initialization on a standard 3.40GHz CPU machine using non-optimized MATLAB implementation. For 300-W dataset which includes 3148 images, the run-time of extracting GIST features for all the images in the training set is 20 minutes. At test stage, the GIST initialization takes 0.381s for each query image. However, the head pose initialization, which achieves competitive performance as ours, completes their training of ConvNet in 2 hours on Tesla K40c GPU. The forward propagation of their network takes 0.3ms per image [6].

V. CONCLUSION

In this work, we presented a novel initialization scheme based on global descriptors, GIST features, to improve the state-of-the-art face alignment methods. The GIST initialization scheme chose K initial shapes by the nearest neighbor scheme. The searching was based on the correlation of GIST features between the query image and the candidate images in the training set. In our work, we found that GIST features are effective in capturing head pose information; thus, eliminating additional training steps to extract head pose information. We compared our proposed initialization scheme with several recent initialization schemes on top of the state-of-the-art face alignment methods. The results show that our proposed initialization scheme can outperform recent initialization schemes as well as achieve high efficiency.

In the future, we will study the performance of other global descriptors for this purpose.

REFERENCES

[1] X. Xiong and F. Torre, "Supervised descent method and its applications to face alignment," in *CVPR*, 2013, pp. 532–539.

[2] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," *IJCV*, vol. 107, no. 2, pp. 177–190, 2014.

[3] X. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *ICCV*, 2013, pp. 1513–1520.

[4] S. Zhu, C. Li, C. Change Loy, and X. Tang, "Face alignment by coarse-to-fine shape searching," in *CVPR*, 2015, pp. 4998–5006.

[5] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *CVPR*, 2014, pp. 1685–1692.

[6] H. Yang, W. Mou, Y. Zhang, I. Patras, H. Gunes, and P. Robinson, "Face alignment assisted by head pose estimation," *BMVC*, 2015.

[7] H. Yang, X. He, X. Jia, and I. Patras, "Robust face alignment under occlusion via regional predictive power estimation," *Image Processing, IEEE Transactions on*, vol. 24, no. 8, pp. 2393–2403, 2015.

[8] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[9] F. Perronnin and D. Larlus, "Fisher vectors meet neural networks: A hybrid classification architecture," in *CVPR*, 2015, pp. 3743–3752.

[10] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *IJCV*, vol. 42, no. 3, pp. 145–175, 2001.

[11] M. Hasan, C. Pal, and S. Moalem, "Localizing facial keypoints with global descriptor search, neighbour alignment and locally linear models," in *CVPR*, 2013, pp. 362–369.

[12] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 faces in-the-wild challenge: The first facial landmark localization challenge," in *CVPR*, 2013, pp. 397–403.

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[14] H. Yang and I. Patras, "Mirror, mirror on the wall, tell me, is the error small?" in *CVPR*. IEEE, 2015, pp. 4685–4693.

[15] V. Kazemi and J. Sullivan, "One millisecond face alignment with an ensemble of regression trees," in *CVPR*, 2014, pp. 1867–1874.

[16] V. Štruc and N. Pavešic, "Photometric normalization techniques for illumination invariance," *Advances in Face Image Analysis: Techniques and Technologies*, pp. 279–300, 2011.

[17] V. Štruc and N. Pavešić, "Gabor-based kernel partial-least-squares discrimination features for face recognition," *Informatica*, vol. 20, no. 1, pp. 115–138, 2009.

[18] P. N. Belhumeur, D. W. Jacobs, D. J. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 12, pp. 2930–2940, 2013.