Cascaded Collaborative Regression for Robust Facial Landmark Detection Trained using a Mixture of Synthetic and Real Images with Dynamic Weighting

Zhen-Hua Feng, *Student Member, IEEE,* Guosheng Hu, *Student Member, IEEE,* Josef Kittler, *Life Member, IEEE,* William Christmas, and Xiao-Jun Wu

***Abstract*—A large amount of training data is usually crucial for successful supervised learning. However, the task of providing training samples is often time-consuming, involving a considerable amount of tedious manual work. Also the amount of training data available is often limited. As an alternative, in this paper, we discuss how best to augment the available data for the application of automatic facial landmark detection (FLD). We propose the use of a 3D morphable face model to generate synthesised faces for a regression-based detector training. Benefiting from the large synthetic training data, the learned detector is shown to exhibit a better capability to detect the landmarks of a face with pose variations. Furthermore, the synthesised training dataset provides accurate and consistent landmarks as compared to using manual landmarks, especially for occluded facial parts.**

The synthetic data and real data are from different domains; hence the detector trained using only synthesised faces does not generalise well to real faces. To deal with this problem, we propose a *cascaded collaborative regression* (CCR) algorithm, which generates a cascaded shape updater that has the ability to overcome the difficulties caused by pose variations, as well as achieving better accuracy when applied to real faces. The training is based on a mix of synthetic and real image data with the mixing controlled by a dynamic mixture weighting schedule. Initially the training uses heavily the synthetic data, as this can model the gross variations between the various poses. As the training proceeds, progressively more of the natural images are incorporated, as these can model finer detail. To improve the performance of the proposed algorithm further, we designed a dynamic multi-scale local feature extraction method, which captures more informative local features for detector training. An extensive evaluation on both controlled and uncontrolled face datasets demonstrates the merit of the proposed algorithm.

***Index Terms*—Facial landmark detection, 3D morphable model, cascaded collaborative regression, dynamic multi-scale local feature extraction.**

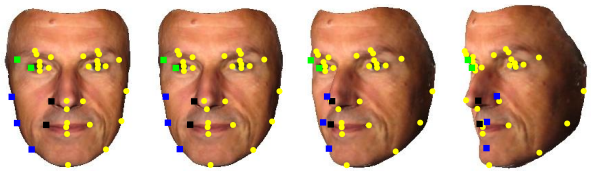
I. Introduction

This work was supported in part by the Key Grant Project of Chinese Ministry of Education under Grant 311024, the Fundamental Research Funds for the Central Universities under Grant JUDCF09032, the UK EPSRC Project EP/K014307/1, European Commission project BEAT under Grant 284989, the National Natural Science Foundation of China under Grants 61373055 and 61103128, the Specialized Research Fund for the Doctoral Program of Higher Education of China under Grant 20130093110009, and the Natural Science Foundation of Jiangsu Province under Grants BK20140419 and BK2012700.

Z.-H. Feng is with the School of Internet of Things, Jiangnan University, Wuxi 214122, China, and also with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: z.feng@surrey.ac.uk).

G. Hu, J. Kittler and W. Christmas are with the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford GU2 7XH, U.K. (e-mail: g.hu, j.kittler, w.christmas@surrey.ac.uk).

X.-J. Wu is with the School of Internet of Things, Jiangnan University, Wuxi 214122, China (e-mail: xiaojun_wu_jnu@163.com).



(a) 0° (b) 10° (c) 30° (d) 50°

Fig. 1. Self-occluded facial landmarks (square points) of the same identity rendered from its 3D face scan, with yaw rotations 0°, 10°, 30° and 50°. As the yaw rotation is increased, the ground truth landmarks of the occluded face parts are getting harder for a human to determine.

FACIAL landmark detection (FLD), or localisation, is an essential preprocessing step in any automatic face analysis system [1]–[3]. According to the image type, FLD algorithms can be categorised as either 2D- [3]–[9] or 3D-based methods [10]–[13]. Typically, a 2D FLD algorithm is applied to the content of a face bounding box output by a face detector, and attempts to locate the positions of a set of pre-defined landmarks (key points), *e.g.* eyebrows, eye centres, nose tip or mouth corners, in a 2D facial image. In contrast, a 3D FLD algorithm performs the task on 3D face data, such as 3D face meshes or range images that are usually in the form of point clouds [12]. In this paper, we focus on designing a robust 2D FLD algorithm trained from a mixed training data with both synthetic and real images. The detected 2D facial landmarks can be used either as geometric features directly, or to extract meaningful local or global face texture features for subsequent face analysis procedures. One common step is to perform face normalisation based on the detected landmarks prior to the feature extraction procedure [2], [3], [14]. Another popular technique is to directly extract local features from the neighbourhoods around all landmarks [1], [2].

There are two main types of algorithms, either based on generative or discriminative models (more details in Section II), for FLD. The most well-known generative models are Active Shape Models (ASM) [4] and Active Appearance Models (AAM) [5]; those have been successfully and widely used for face modelling and landmark detection during the past 20 years, especially in controlled scenarios. However, robust and accurate FLD in uncontrolled scenarios is very challenging. ASM and AAM often fail to accurately estimate the landmarks for 'faces in the wild', in the presence of pose, expression, illumination, and occlusion. Recently, the focus has been on discriminative models [8], [9], [15]–[26] which have been

shown to offer better performance in FLD, especially for faces exhibiting greater variability in appearance.

As most of the discriminative models are supervised, the availability of a large quantity of training samples is crucial [27], [28]. In general, data acquisition involves collecting and labelling data manually, which is laborious and expensive. Moreover, not all annotators are well-motivated to perform the annotation accurately. Belhumeur *et al.* [29] measured three Amazon Mechanical Turk workers' landmarking results on more than 1000 images and found a significant diversity in human annotations. Fortunately, with the growth of digital content on the Internet, the automatic acquisition of a huge face dataset has become a possibility, as exemplified by the well-known LFW face database [30] collected from the Internet using a Viola-Jones face detector [31]. However, the collected data then has to be annotated with meaningful tags. This task is hard to accomplish using existing automatic techniques, especially for the specific application of FLD. One promising alternative is to expand an existing annotated/labelled dataset using either 2D [8], [9], [17], [32], [33] or 3D techniques [34]–[36]. In this paper, we advocate the use of a 3D morphable face model (3DMM) [37], [38] to generate a large number of synthetic faces to augment the training data. Compared to manually annotated training images, the use of synthetic faces from a 3D face model has several advantages:

a) : It is inexpensive to obtain a large quantity of training samples. As the synthesised faces are projections of the shape and texture of an annotated 3D face scan from 3D to 2D, one annotated 3D face can render a number of 2D faces under a variety of pose variations without re-annotating.

b) : The landmarks of the synthesised faces with different pose variations are accurate and consistent due to the direct projection from 3D to 2D. In contrast, manually annotated landmarks often exhibit bias across different individuals and images, especially for the landmarks located at weak corners or along edges [29].

c) : A 3D face scan is naturally non-self-occluded, which enables the true positions of the occluded landmarks in a 2D face to be accurately estimated by mapping the 3D coordinates to 2D, as shown in Fig. 1. With this information, we can teach a 2D facial landmark detector to accurately estimate the landmarks of the occluded facial parts in a 2D image. In previous work, one often-used technique is to omit the self-occluded landmarks and use different numbers of landmarks for faces in different poses [2], [39]. But the use of different landmark schemes builds multiple shape models and cannot give a unified shape representation. Thus some algorithms only consider the problem of annotating near-frontal faces [3], [20]. However, the landmarks of the occluded facial parts in a 2D face image are important for some practical applications, such as 3D face reconstruction from a single 2D image [37], [38], [40]–[42]. In this task, 2D landmarks are usually required to initialise and constrain the 3D face model fitting.

Through our early experimental investigation, we found that a facial landmark detector trained merely from synthesised faces often failed to adapt to real faces. This confirms the conclusion in [28] that simply increasing the training data can sometimes results in over-training. The main reason is that the synthesised faces belong to a different domain of real faces. Furthermore, the synthesised faces from a size-limited 3D face scan dataset lack realistic variations in appearance, such as variations in lighting, make-up, skin colour, occlusion and sophisticated background (Fig. 3). This is possibly why there are only a few previous papers using synthesised faces for FLD. To tackle this problem, we propose a cascaded collaborative regression (CCR) algorithm to efficiently exploit synthesised and real faces in a compound training scheme with a dynamic mixture weighting schedule. We present an important innovation whereby in our CCR training scheme we progressively adapt the relative contribution of synthesised and natural images. At first, the synthetic data dominates the training; the impact of the synthetic data is then progressively reduced as the training proceeds. Thus the proposed CCR is first trained on a mixed dataset with a large number of synthesised images to improve the generalisation capacity followed by adaptation using a small number of real faces.

Note that a discriminative model is a mapping function from a feature space to a shape space; the choice of features affects the final performance of the trained model to a great extent. Generally, we use local features to train a discriminative model due to their robustness to appearance variations. Both hand-crafted and learning-based local descriptors have been employed in previous work, *e.g.* SIFT [9], HOG [17], shape-indexed features [8] and the sparse auto-encoder [23], [26]. Yan *et al.* compared HOG, SIFT, LBP and Gabor features and found that the HOG descriptor performed best on the task of FLD [17]. However, due to the complexity of the variations in appearance of human faces, a single HOG descriptor does not adequately represent face image properties for FLD. Therefore, to extract informative local features, we present a dynamic multi-scale local feature extraction scheme that has the capacity to provide a rich face representation.

In summary, the proposed algorithm has three main contributions: First, we include a large number of synthesised samples from a 3DMM when we train a facial landmark detector in 2D. Second and the most important, we adapt the trained model from synthesised faces to real faces, for which we advocate a CCR approach to optimally exploit the synthetic data in tandem with real face training samples. In the proposed CCR, the synthesised faces collaborate with the real faces to build a cascaded facial landmark detector using a dynamic mixture weighting scheme. Last, we propose a dynamic multi-scale local feature extraction strategy, which uses a dynamic window size combined with a multi-scale representation to extract the local features around a landmark. By design, the dynamic adaptation of the image window size (from big to small) naturally works together with the proposed CCR algorithm that is based on a coarse-to-fine landmark detection process. Moreover, the use of the multi-scale strategy helps to extract more informative local features in model training.

The remainder of the paper is organised as follows. We first give a brief review of related work in Section II. Then we overview the cascaded regression method in Section III, and present proposed CCR and dynamic multi-scale local feature extraction strategy in Section IV. Evaluation and conclusion are demonstrated in Section V and VI respectively.

TABLE I
A SUMMARY OF CASCADED-REGRESSION-BASED METHODS.

method	structure	learning algorithm	features
Cao *et al.* [8]	cascaded regression	two-level boosted regression	shape-indexed pixel difference
Xiong *et al.* [9]	cascaded regression	linear regression	SIFT
Yan *et al.* [17]	cascaded regression	linear regression	HOG
Chen *et al.* [24]	joint cascaded regression	random forests	shape-indexed pixel difference
Zhang *et al.* [25]	multi-task learning (single regressor)	deep convolutional network	deep convolutional network
Ren *et al.* [22]	cascaded regression	random forests	local binary features
Burgos-Artizzu *et al.* [21]	cascaded regression	two-level boosted regression	shape-indexed pixel difference
our CCR	cascaded collaborative regression	weighted ridge regression	dynamic multi-scale HOG

II. RELATED WORK

Existing FLD methods can be divided, according to the underlying model, into two main categories: generative and discriminative. A generative model generates different face instances by adjusting the model parameters, and matches these generated instances to an input image to optimise the model parameters. In contrast, a discriminative model directly predicts the shape update, using a mapping function from a feature space to the shape space, to guide the evolution from an initial shape estimate to the true positions.

A. Generative models

Typical generative models are ASM [4], AAM [5] and their extensions [3], [6], [7], [39], [43], [44]. A common characteristic of ASM and AAM is a parametric PCA-based shape model that is constrained by the corresponding eigenvalues when fitting the models to an input image. In general, during the fitting phase, the model parameters are optimised to minimise a cost function gauging the error between the generated instance and the input image. In ASM, the optimisation was accomplished by first searching for the strongest edge along the normal to the boundary passing through a landmark [4], and then minimising the Mahalanobis distance between a statistical local intensity profile model and the corresponding intensities of the pixels along the normal [45]. AAM minimises the difference between the generated face appearance and the input facial image by adjusting the parameters of the shape and appearance models. The cost function in AAM is usually optimised using a gradient descent algorithm, such as the well-known project-out inverse compositional algorithm [6].

Although generative models have been successfully used in many scenarios, they often fail on faces in the presence of varied pose, expression, illumination and occlusion. The main reason is that the gradient-descent-based algorithms are often trapped in local minima, and the use of intensity value is not robust to those variations. Some improved algorithms, such as view-based AAM [39], bilinear AAM [43], tensor-based AAM [3], and kernel ASM [46], [47], have addressed these problems to some extent, but the task of robust and accurate FLD is still very challenging, especially for faces in the wild.

B. Discriminative models

Compared with the generative models, a discriminative model builds a mapping function that predicts the shape or model parameter updates using the features extracted from an image. In fact, the first discriminative method was adopted in the original AAM [4], [48]. It used a linear regression model to build a mapping function between the texture residuals and the parameter update of a combined appearance model. The combined appearance model explicitly represents a face including its shape and appearance. It accomplishes the tasks of FLD and face texture reconstruction in a unified framework. The assumption behind linear regression is that there is a constant linear relationship between texture residuals and parameter updates. However, this assumption is unrealistic: linear regression is incapable of solving such a complex non-linear multi-variable optimization problem. A common way to overcome this difficulty is to replace linear regression using more advanced regression schemes, such as canonical correlation analysis [49], random forests [22], random ferns [8], [16] and deep neural networks [18], [19], [23].

More recently, cascaded regression (CR)-based methods have been shown to be more successful in overcoming the difficulties posed by variations in appearance and have demonstrated impressive performance in both controlled and uncontrolled scenarios [8], [9], [16]–[19], [21]–[25]. The key to the success of these models is, first of all, to train a strong regressor which is composed of many weak regressors in cascade. It has been demonstrated that even a simple set of cascaded linear regressors achieves promising landmarking results [9], [17]. Second, a discriminative approach predicts the landmarks directly using a non-parametric shape model, which exhibits a better representation capacity compared to a PCA-based parametric shape model [8]. Benefiting from the inherent shape constraint of the cascade structure, the non-parametric models can update the face shape explicitly and accurately. Last, the local feature descriptors used in CR-based methods extract more robust local features than using conventional pixel intensities.

As the main idea of cascaded regression is to use a sequence of weak regressors to estimate the highly complicated non-linear relationship between robust local features and shape updates, the development of cascaded regression has focused on designing new cascade structures, learning algorithms and local feature extraction methods. A summary of a set of state-of-the-art CR-based algorithms is shown in Table I. Dollàr *et al.* described the cascaded pose regression for 2D pose estimation using shape-indexed pixel difference features with random ferns [16], and further proposed a robust cascaded

regression method to improve the performance of FLD when images contain occlusions and large pose variations [21]. Cao *et al.* [8] introduced the concept of shape constraint inheritance in cascaded regression and designed a two-level boosted regression in cascade for FLD. In [9], Xiong and De la Torre proposed a theoretical explanation of the cascaded regression approach as a supervised descent method (SDM) and used SIFT features for successful FLD. Yan *et al.* [17] compared different local features, *e.g.* HOG, SIFT, LBP and Gabor, for cascaded regression and proposed a ranking scheme to select the best location from many candidates. Ren *et al.* [22] proposed a local binary feature extraction method for CR-based FLD and achieved state-of-the-art results. Sun *et al.* [18], Zhou *et al.* [19] and Zhang *et al.* [25] proposed the use of cascaded deep convolutional neural networks (CNN) and showed promising results for facial landmarks detection. The advantage of the use of CNN is that it combines the feature extraction and regression steps in a unified framework. However, it is a very difficult task to tune the parameters of deep neural networks. For a data-driven approach, the issue of training data size and representativeness for learning cascaded regression is of paramount importance. Unfortunately, none of these methods has addressed this issue systemically. In contrast, the key innovation of this paper is to explore how best to augment the available data with a larger quantity of synthetic training samples from a 3D face model. To meet this requirement, a cascaded collaborative regression algorithm is proposed to adapt the model from the domain of synthesised images to the domain of real images, along with an innovative dynamic multi-scale local feature extraction strategy.

C. Training data augmentation

As stated above, the discriminative models are usually supervised, which requires a large training dataset to cover all possible variations likely to be encountered in practice. However, manually landmarking a training dataset is tedious, repetitive and error-prone work. To obtain a large training set, a standard way is to augment the available training dataset by using different initial shape estimates [8], [9], [17]. To be more specific, a training sample consists of an image, the ground truth face shape and the initial shape estimate. The aim of a discriminative model is to learn a mapping function that estimates a shape update so as to guide the initial shape estimate towards the ground truth shape. As the number of training images and the corresponding ground-truth shapes is fixed for a specific training dataset, we can augment the volume of training data using different initial shape estimates for one particular training image, *i.e.* putting a mean shape or a randomly selected shape around the ground truth shape. However, this approach is inadequate because it only samples a limited range of similarity transformations of a training set.

Another popular way is to synthesise some virtual samples using either 2D or 3D techniques. For instance, Tang *et al.* [50] proposed the use of a set of synthesised training samples from an articulated hand model for real-time hand pose estimation. Ghiasi and Fowlkes [33] generated additional training samples by the means of adding artificial occlusions to original training images for FLD. Pishchulin *et al.* [35] proposed using a 3D shape model to synthesise realistic human body images with random backgrounds and demonstrated superior performance in human detection and pose estimation. Pepik *et al.* [36] generated a set of virtual images with a 3D CAD model for fine-grained object detection and pose estimation using deformable part models constrained by 3D geometric information. As in [35] and [36], we propose the use of a 3DMM to synthesise 2D faces. The 3DMM is a powerful tool that can be used to obtain 3D shape and texture representations of a human face and generate arbitrarily varied 2D faces. For instance, Rätsch *et al.* [34] used the 3DMM to render a set of 2D faces and learnt support vector regression for pose estimation. However, simply using 3D synthesised images has some drawbacks as stated in previous sections. The synthesised images often lack realistic appearance variations, such as illumination, expression, occlusion and background. To synthesise realistic human body images, randomly selected backgrounds were used in [32], [35]. However, this technique cannot address the problem fundamentally due to a variety of changes in appearance of the synthesised objects. In this paper, we consider the problem as a domain adaptation problem and correspondingly propose a cascaded collaborative regression algorithm to adapt the model trained on synthesised images to real ones.

III. Backgrounds: cascaded regression

In discriminative FLD, a 2D face shape is represented by a set of pre-defined landmarks $\mathbf{s} = [x_1, y_1, \cdots, x_L, y_L]^{\mathbf{T}}$, where L is the number of landmarks and $[x_l, y_l]^{\mathbf{T}}$ are the coordinates of the lth landmark. Given N training images $\{\mathbf{I}(1), \cdots, \mathbf{I}(N)\}$ with the initialised face shape estimates $\{\mathbf{s}_0(1), \cdots, \mathbf{s}_0(N)\}$, we first extract the shape-related features $\{\mathbf{f}(\mathbf{I}(1), \mathbf{s}_0(1)), \cdots, \mathbf{f}(\mathbf{I}(N), \mathbf{s}_0(N))\}$ using a feature mapping function $\{\mathbf{f}(\mathbf{I}, \mathbf{s})\}$. Then the discriminative model, *i.e.* a mapping function,

$$\mathbf{\Phi} : \mathbf{f}(\mathbf{I}, \mathbf{s}_0) \mapsto \delta\mathbf{s}, \tag{1}$$

is trained by minimising the cost function,

$$\frac{1}{2N}\sum_{n=1}^{N}\|\mathbf{s}_0(n) + \delta\mathbf{s}(n) - \mathbf{s}^*(n)\|_2^2, \tag{2}$$

where $\mathbf{s}^*(n)$ is the ground truth shape of the nth training image and $\delta\mathbf{s}(n) = \mathbf{\Phi}(\mathbf{f}(\mathbf{I}, \mathbf{s}_0))$ is the corresponding shape update. The training of $\mathbf{\Phi}$, in essence, is a data-driven process requiring a set of annotated facial images. Generally, the mapping function in equation (1) can be obtained using any regression method, such as linear regression [9], [17], random forests [22] and artificial neural networks [18]. However, using only one regressor is insufficient, because the task of FLD is a highly non-linear optimisation problem, made difficult by the numerous variations in face appearance. In this paper, we use a state-of-the-art CR structure [9], [16], [17], because of its robustness and effectiveness. In CR, by linking several linear regression steps, we can realise a non-linear mapping needed for the task in hand. Also, the cascade structure provides an essential framework for the proposed CCR approach, because

Algorithm 1 Cascaded regression.

Input: Image $\mathbf{I}$, initial shape estimate $\mathbf{s}_0$ and the pre-trained cascaded regressors $\mathbf{\Phi} = \{\mathbf{R}_1, \cdots, \mathbf{R}_M\}$.
Output: Estimated face shape $\mathbf{s}_M$.
Repeat:
for $m = 1 \cdots M$ **do**
  Obtain shape-related features $\mathbf{f}(\mathbf{I}, \mathbf{s}_{m-1})$.
  Estimate the shape update: $\delta\mathbf{s} = \mathbf{A}_m\mathbf{f}(\mathbf{I}, \mathbf{s}_{m-1}) + \mathbf{b}_m$.
  Update shape: $\mathbf{s}_m = \mathbf{s}_{m-1} + \delta\mathbf{s}$.
end for

different types of training data can be used in designing the respective stages of the cascade.

In CR, the mapping function $\mathbf{\Phi}$ is a strong regressor formed by a sequence of weak regressors in cascade:

$$\mathbf{\Phi} = \mathbf{R}_1 \circ \cdots \circ \mathbf{R}_M, \tag{3}$$

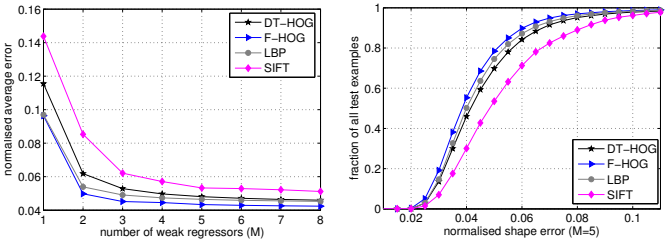
where $\mathbf{R}_m = \{\mathbf{A}_m, \mathbf{b}_m\}$ $(m = 1 \cdots M)$ is the mth weak regressor, $\mathbf{A}_m \in \mathbb{R}^{2L \times F}$ is the projection matrix, $\mathbf{b}_m \in \mathbb{R}^{2L \times 1}$ is the offset, F is the dimensionality of the shape-related feature vector $\mathbf{f}(\mathbf{I}, \mathbf{s})$. The CR is a coarse-to-fine process, in which the first few weak regressors cover gross variations and the subsequent weak regressors refine the roughly estimated shapes. It has been shown in [8] that the first weak regressor in the cascade mainly performs affine transformations dealing with large-scale pose variations, and so can roughly update the landmarks of the initial shape closer to the ground truth shape. The last weak regressor covers small-scale variations. It performs fine tuning to drive the roughly updated shape estimate to a more accurate position. Details of training the cascaded regressors, *i.e.* $\mathbf{A}_m$ and $\mathbf{b}_m$, using our CCR are discussed later in Section IV-B.

Assuming that a strong regressor $\mathbf{\Phi}$ has already been pre-trained, given an input image $\mathbf{I}$ and a rough initial shape estimate $\mathbf{s}_0$, we apply the first weak regressor to update this shape estimate to $\mathbf{s}_1$ and then pass $\mathbf{s}_1$ to the subsequent weak regressors until the final shape is obtained (Algorithm 1). To be more specific, this is a recursive algorithm in which the mth shape is obtained by:

$$\mathbf{s}_m = \mathbf{s}_{m-1} + \mathbf{A}_m\mathbf{f}(\mathbf{I}, \mathbf{s}_{m-1}) + \mathbf{b}_m. \tag{4}$$

Note that the shape-related feature vector $\mathbf{f}(\mathbf{I}, \mathbf{s}_{m-1})$ changes after each shape update, because the features are extracted from local regions around the landmarks of the updated shape. Typically, the shape-related features are computed by applying a local feature descriptor to the neighbourhoods around all the landmarks and then concatenating them into a single feature vector. We compared the performance of a standard linear-regression-based CR method using different local features. As indicated by Fig. 2, the F-HOG [51] descriptor performs best in terms of accuracy, which confirms the conclusion of [17]. Furthermore, to obtain a more informative representation of each local region, we extract local features at multiple scales with a dynamic window size. The details of the proposed local feature extraction approach are discussed in Section IV-C.

Fig. 2 also demonstrates that the use of cascaded regression greatly improves the performance of a FLD detector in accuracy. Despite different local features, the accuracy of the use of multiple linear regressors in cascade is much better than that of a single linear regressor. However, augmenting the number of weak regressors leads to a linear increase in computational and memory costs and hence slows down the whole system. The use of too many weak regressors in cascade is time-consuming and brings little benefit. The improvement is marginal after cascading more than 5 weak regressors, as indicated by Fig. 2.



(a) (b)

Fig. 2. A comparison of different local descriptors (F-HOG, DT-HOG, LBP and SIFT), in terms of accuracy on the synthesised face database using a standard linear-regression-based CR method: (a) normalised average shape errors of different descriptors with respect to the number of cascaded weak regressors (M); (b) detailed cumulative distribution curves of different descriptors with $M = 5$. The errors are normalised by the inter-ocular distance, and the details of the synthetic dataset are described in Section V-A1.

To set a rough initial shape estimate for FLD, the most widely used way is to use the face bounding box output by a face detector, as presented in [8], [9], [17], [26]. More specifically, in this paper, we firstly calculate the mean face shape across all the training samples. Secondly, for a training image, we can estimate the translation (t_x, t_y) between the geometric centres of the detected face bounding box and the ground truth shape. The scale ratio s, both on X- and Y-axis, between the face bounding box and the ground truth shape can also be estimated. Hence, for each training image, we have a parameter vector $\mathbf{p} = [t_x, t_y, s_x, s_y]^T$ that indicates the relative position between the face bounding box and the ground truth shape. Finally, we calculate the average value of $\mathbf{p}$ across all the training samples, and use this to put the mean shape inside of the detected face bounding box as the initial shape estimate of an input face image.

IV. Cascaded Collaborative Regression (CCR)

It is worth noting that CR [16] is crucial for our CCR. The basic idea of CR is to form a strong regressor from a sequence of weak regressors in series, while the main purpose of our proposed CCR is to optimally combine the complementary information from the synthesised and real face datasets. This involves mixing these two types of data in designing successive stages of the cascade via a dynamic mixture weighting scheme imposing an iterative adjustment of the weights during the training phase. This can only be accomplished when we have a multiple regressor system, *e.g.* CR. In our proposed CCR, the synthesised training dataset dominates the training of the first few weak regressors, whereas the real face dataset dominates the last few weak regressors. Benefiting from the large synthesised training dataset, the first few weak regressors

Fig. 3. Facial images rendered from one 3D face scan, with yaw rotation from -50° to 50° at 10° intervals and with pitch rotation from -30° to 30° at 15° intervals.

in CCR are capable of overcoming the difficulty caused by pose variation. The last few weak regressors mainly trained on a relatively small number of real faces refine the rough shape estimates output by the first few elements, to create a more versatile model.

A. Face synthesis using 3DMM

As a means of augmenting the training dataset, we generate training samples synthetically, using a 3D morphable face model (3DMM) [37], [38], [41], [52]. We captured 163 3D face scans [53], [54] and registered them using the Iterative Multi-resolution Dense 3D Registration (IMDR) approach [55]. After registration, the 3D shape information of a 3D face scan can be expressed by V 3D vertices $\mathbf{v} = [x_1, y_1, z_1, \cdots x_V, y_V, z_V]^\mathbf{T}$, where $\mathbf{v}_v = [x_v, y_v, z_v]^\mathbf{T}$ are the coordinates of the vth vertex. The corresponding vertex colour information is represented by $\mathbf{t} = [r_1, g_1, b_1, \cdots, r_V, g_V, b_V]^\mathbf{T}$, where $[r_v, g_v, b_v]^\mathbf{T}$ are the RGB intensities of the vth vertex.

To project the 3D vertices to a 2D image plane, a perspective camera is used. Specifically, each vertex $\mathbf{v}_v$ can be mapped to the position $\mathbf{w}_v = [w_{v_x}, w_{v_y}, w_{v_z}]^\mathbf{T}$ in a camera-centred 3D coordinate system by the rigid transformation:

$$\mathbf{w}_v = \mathbf{R}_{\theta_z}\mathbf{R}_{\theta_y}\mathbf{R}_{\theta_x}\mathbf{v}_v + \boldsymbol{\tau}, \tag{5}$$

where $\mathbf{R}_{\theta_x}$, $\mathbf{R}_{\theta_y}$ and $\mathbf{R}_{\theta_z}$ denote the 3D rotation matrices with the Euler angles θ_x, θ_y and θ_z around the X, Y and Z axes of the virtual camera coordinate system, and $\boldsymbol{\tau} \in \mathbb{R}^3$ defines the spatial translation of the camera with respect to the model. Then $\mathbf{w}_v$ is projected on the 2D image plane coordinates $\mathbf{p}_v = [p_{v_x}, p_{v_y}]^\mathbf{T}$ by a perspective projection:

$$p_{v_x} = o_x + f\frac{w_{v_x}}{w_{v_z}}, \quad p_{v_y} = o_y - f\frac{w_{v_y}}{w_{v_z}}, \tag{6}$$

where f is the focal length in the camera-centred coordinate system, and $[o_x, o_y]^\mathbf{T}$ is the image-plane position of the optical axis. Given a registered 3D face scan ($\mathbf{v}$ and $\mathbf{t}$) and the rotation parameters θ_x, θ_y and θ_z, we can calculate the 2D coordinates of each vertex on the 2D image plane coordinate system and obtain the rendered 2D image using the corresponding 3D RGB intensities. In general, a registered 3D face scan can be rendered as 2D faces with arbitrary poses by changing the rotation parameters θ_x, θ_y and θ_z. Fig. 3 shows some rendered 2D faces under different pose variations from one 3D face scan. Although the number of the vertices of a 3D face scan exceeds 30,000 in our model, we only need a small number of 2D landmarks for FLD. In our case, we select 34 landmarks (Fig. 8c) from the registered 3D scans and use their projected 2D coordinates as the ground truth shapes.

Note that the texture details of synthesised faces are not as good as real ones, *e.g.* the synthesised facial images have a uniform background. Also, the synthesised faces lack variety of appearance compared with real faces, such as expression, illumination and occlusion by other artefacts. Thus simply using only synthesised facial images as a training dataset is insufficient. To overcome this problem, one straightforward solution is to synthesise more realistic facial images. For example, we can use randomly selected background as suggested by [32], [35], and add a Phong model to generate virtual facial images with illumination variations [37], [38]. However, neither of them can solve the problem fundamentally due to the rich variations in appearance of human faces. Another way is to directly train a CR-based model on a mixed training dataset including both real and synthesised faces. We use the term 'one-off' training for this approach. However, the model obtained by this one-off training does not fit to a real face very accurately, due to the preponderance of the synthesised faces in the mixed training dataset. The synthesised faces dominate the trained model, especially when the size of the synthetic dataset is much bigger than that of the real face dataset. Thus we propose a CCR approach trained using the mixed dataset with dynamic mixture weighting, so that the synthetic and real face datasets are complementary. In our proposed CCR, the use of dynamic multi-scale HOG features leads to further robustness in these variations, especially in illumination, because it operates on image gradient orientations rather than raw pixel values.

B. CCR training

Given a mixed training dataset with T synthesised images $\{\tilde{\mathbf{I}}(1), \cdots, \tilde{\mathbf{I}}(T)\}$ and R real images $\{\mathbf{I}(1), \cdots, \mathbf{I}(R)\}$, and the corresponding ground truth shapes $\{\tilde{\mathbf{s}}^*(1), \cdots, \tilde{\mathbf{s}}^*(T)\}$ and $\{\mathbf{s}^*(1), \cdots, \mathbf{s}^*(R)\}$, we first generate the initial shape estimates $\{\tilde{\mathbf{s}}_0(1), \cdots, \tilde{\mathbf{s}}_0(T)\}$ and $\{\mathbf{s}_0(1), \cdots, \mathbf{s}_0(R)\}$ by putting a reference shape in the detected face bounding box, similar to [9], [17]. The reference shape is either the mean shape or a randomly selected face shape across all the training shapes.

Then we recursively learn the weak regressors from $m = 1$ to M. In the training phase, the initial shape estimates $\mathbf{s}_0$ are used to obtain the first weak regressor $\mathbf{R}_1 = \{\mathbf{A}_1, \mathbf{b}_1\}$, and then we apply this trained weak regressor to update all initial shapes to train the next weak regressor, until all the weak regressors in $\boldsymbol{\Phi} = \{\mathbf{R}_1, \cdots, \mathbf{R}_M\}$ are obtained. To be more specific, the cost function of learning the mth weak regressor $\mathbf{R}_m = \{\mathbf{A}_m, \mathbf{b}_m\}$ by CCR is:

$$\mathbf{J}(\mathbf{A}_m, \mathbf{b}_m) = \frac{\omega(m)\mathbf{J}_t + (1-\omega(m))\mathbf{J}_r}{2N} + \lambda\|\mathbf{A}_m\|_F^2, \quad 0 < \omega < 1, \tag{7}$$

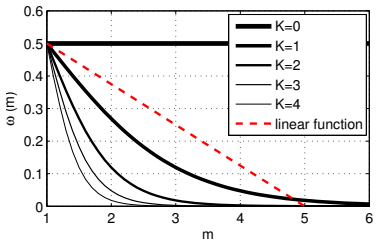


Fig. 4. The curve of $\omega(m)$ as m is increased, with different shrinking rates K in equation (10). For comparison, a simple linear function is also used.

where $\omega(m)$ is a dynamic mixing parameter, $N = T + R$ is the total number of the training samples, λ is the weight of the regularisation term and $\|\cdot\|_F$ is the Frobenius norm. $\mathbf{J}_t$ is the cost function calculated from the synthesised training samples:

$$\mathbf{J}_t = \sum_{t=1}^{T} \|\delta\tilde{\mathbf{s}}_{m-1}(t) - \mathbf{A}_m \mathbf{f}(\tilde{\mathbf{I}}(t), \tilde{\mathbf{s}}_{m-1}(t)) - \mathbf{b}_m\|_2^2, \quad (8)$$

and $\mathbf{J}_r$ is the cost function calculated from the real training samples:

$$\mathbf{J}_r = \sum_{r=1}^{R} \|\delta\mathbf{s}_{m-1}(r) - \mathbf{A}_m \mathbf{f}(\mathbf{I}(r), \mathbf{s}_{m-1}(r)) - \mathbf{b}_m\|_2^2, \quad (9)$$

where $\mathbf{f}(\mathbf{I}, \mathbf{s}_{m-1})$ is the updated-shape-related local feature, $\delta\mathbf{s}_{m-1} = \mathbf{s}^* - \mathbf{s}_{m-1}$ is the shape difference between the ground truth shape $\mathbf{s}^*$ and the updated shape $\mathbf{s}_{m-1}$. The updated shape $\mathbf{s}_{m-1}$ is obtained by applying the previous $m-1$ weak regressors to the initial shape estimate $\mathbf{s}_0$, as described in Algorithm 1.

The synthesised image dataset interplays with the real image dataset via the dynamic mixing parameter $\omega(m)$. The first few regressors in the cascaded structure generate gross shape updates to accommodate pose variations, while the subsequent regressors generate more precise shape refinement. The annealing schedule of $\omega(m)$ as a function of m is set by:

$$\omega(m) = \frac{1}{1 + e^{K(m-1)}}, \quad (10)$$

where K is a shrinking rate. Fig. 4 shows how the function $\omega(m)$ varies when we set K to different values. In the strong cascaded regressor trained by CCR, the first few weak regressors are trained on a mixed dataset with a large number of synthesised images and hence have a good generalisation capacity to pose variations, while the last few weak regressors are trained mainly on real images and hence can fine-tune the shape estimate of a real face.

Note that although the overall mapping function in (3) is non-linear, each stage in the cascade implements a linear regressor (4). Thus the mth weak regressor can be efficiently solved by:

$$\mathbf{A}_m^{\mathbf{T}} = (\mathbf{F}_{m-1}\mathbf{\Omega}_m\mathbf{F}_{m-1}^{\mathbf{T}} + \lambda\mathbf{I})^{-1}\mathbf{F}_{m-1}\mathbf{\Omega}_m\delta\mathbf{S}_{m-1}^{\mathbf{T}}, \quad (11)$$

where $\mathbf{\Omega}_m \in \mathbb{R}^{N\times N}$ is a diagonal matrix:

$$\mathbf{\Omega}_m(i,j) = \begin{cases} \omega(m) & \text{if } i = j \text{ AND } i,j \le T \\ 1-\omega(m) & \text{if } i = j \text{ AND } i,j > T \\ 0 & \text{others} \end{cases}, \quad (12)$$

with the first T non-zero values $\omega(m)$ and the last R non-zero values $1-\omega(m)$ on the main diagonal. $\mathbf{F}_{m-1} \in \mathbb{R}^{F\times N}$ is the concatenated feature matrix of the synthesised training samples and the real training samples:

$$[\tilde{\mathbf{f}}_{m-1}(1), \cdots, \tilde{\mathbf{f}}_{m-1}(T), \mathbf{f}_{m-1}(1), \cdots, \mathbf{f}_{m-1}(R)], \quad (13)$$

where $\mathbf{f}_{m-1}(n) = \mathbf{f}(\mathbf{I}(n), \mathbf{s}_{m-1}(n))$ stands for the shape-related feature vector extracted from the nth training image $\mathbf{I}(n)$ using the corresponding updated shape $\mathbf{s}_{m-1}(n)$, and F is the dimensionality of the extracted local features. $\delta\mathbf{S}_{m-1} \in \mathbb{R}^{2L\times N}$ is the concatenated shape difference matrix of the synthesised training samples and the real training samples:

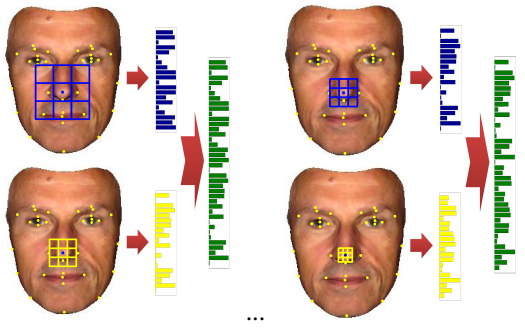
$$[\delta\tilde{\mathbf{s}}_{m-1}(1), \cdots, \delta\tilde{\mathbf{s}}_{m-1}(T), \delta\mathbf{s}_{m-1}(1), \cdots, \delta\mathbf{s}_{m-1}(R)], \quad (14)$$

and L is the number of landmarks. To obtain $\mathbf{b}$, we can simply add one more element with the value of 1 at the beginning of each shape-related feature vector; hence the first column of the solved matrix $\mathbf{A}$ becomes $\mathbf{b}$.

C. Dynamic multi-scale local feature extraction

The CCR algorithm is presented above and the only remaining task is feature extraction, *i.e.* how to obtain $\mathbf{f}(\mathbf{I}, \mathbf{s})$ for the CCR training. In this section, we present a dynamic multi-scale local feature extraction scheme to extract more informative local features. As suggested by [17], the HOG descriptor performs best in CR-based FLD. To validate this conclusion, we tested the accuracy of HOG, LBP and SIFT descriptors on our synthetic dataset using a standard linear-regressor-based cascaded regression method. For HOG, we used both the classical Dalal-Triggs HOG [56] and the extended Felzenszwalb HOG [51]. The experimental results confirm the conclusion of [17], as demonstrated in Fig. 2. Hence, in this paper, we use HOG as the baseline to demonstrate the superiority of the proposed dynamic multi-scale local feature extraction strategy.

We first extract multi-scale HOG features from multiple neighbourhoods of each landmark, similar to Chen *et al.* [1]. They showed that the use of multi-scale local feature descriptors was beneficial for the task of face recognition, so we combine this multi-scale feature extraction method with the variable-scale strategy [17], [23], *i.e.* using a larger window for the first regressor and progressively smaller windows for the following regressors. In contrast to [1], [23] that generate a set of pyramid images of the original image and use a fixed basic window, we apply the multi-scale feature extraction strategy by varying the sizes of windows around a landmark for each regressor. This leads to a reduction in storage. The variable-scale feature extraction method benefits the proposed CCR, because the extracted local features using the relatively small windows of the last few weak regressors are more robust to pose variations that have already been tackled by the first few weak regressors.



(a) (b)

Fig. 5. A schematic overview of the proposed dynamic multi-scale local feature extraction approach, when we set the number of sub-regressors in cascade $M = 5$ and the number of scales for multi-scale local feature extraction $C = 2$ (equation (15)): (a) dynamic multi-scale windows for the 1st sub-regressor; and (b) dynamic multi-scale windows for the 5th sub-regressor.

To obtain dynamic multi-scale local features, we denote the window size of the cth scale of the mth sub-regressor in the cascade as $S_{(m,c)}(m = 1, \cdots, M; c = 1, \cdots, C)$, where C is the number of scales for local feature extraction and M is the number of sub-regressors. We set

$$S_{(m,c)} = \frac{(1 + \frac{1}{m})S_{face}}{c + 1}, \tag{15}$$

where m reduces the window size as the order of the sub-regressor grows, c generates different scales for each sub-regressor, and S_{face} is the size of the face. The face size can be obtained using the maximum of the inter-ocular distance and the height between eye and mouth. The neighbourhoods with dynamic multi-scale window sizes of the 1st and 5th regressors are shown in Fig. 5, when we set $M = 5$ and $C = 2$. Then we split each neighbourhood into 3×3 cells and resize each cell to 10×10 to extracted HOG features on each cell. Finally, we concatenate all the outputs from the HOG descriptor into one feature vector. We could further rescale the basic windows to more sub-windows or split the neighbourhood into more cells, to extract more informative local features. However, the dimensionality of the extracted local feature vector would be higher, leading to increased storage and computation costs.

V. Evaluation

We extensively evaluated the proposed algorithm on a number of face datasets: a synthetic face dataset rendered from 163 3D face scans [53]; the CMU Multi-PIE face dataset [57]; the BioID dataset [58]; the HELEN dataset [59]; the LFPW dataset [29]; and the COFW dataset [21].

In this section, we firstly introduce these datasets and detail our experimental settings. Secondly, we evaluate the performance of a basic linear-regression-based CR algorithm with our proposed dynamic local feature extraction approach on the synthetic dataset, both in the FLD accuracy and 3D face reconstruction accuracy. Thirdly, we investigate the proposed CCR training strategy on the Multi-PIE dataset using different settings. Lastly, we compare our proposed FLD pipeline with several recently proposed state-of-the-art algorithms [8], [9], [20], [21], [26], [29], [33], [60]–[62] on BioID, HELEN, LFPW and COFW.

(a) near-frontal faces

(b) half-profile faces

Fig. 6. Pose variations in the selected subset of Multi-PIE, including (a) 5 near-frontal poses (05_1, 05_0, 14_0, 04_1, 13_0) with 16 IDIAP ground truth landmarks, and (b) 4 half-profile poses (19_0, 19_1, 08_0, 08_1) with 8 IDIAP ground truth landmarks.

A. Experimental setup

1) Datasets:

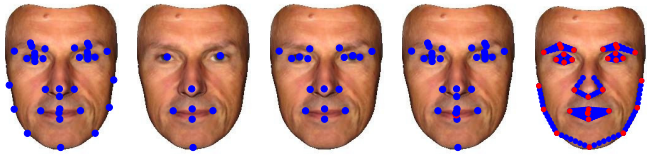
a) Synthetic dataset: The images of this dataset were rendered from a 3D face scan database with 163 identities [53], [54]. For each identity, we rendered 11×5 2D faces with 11 evenly-distributed yaw rotations from $-50°$ to $50°$ and 5 pitch rotations at $-30°, -15°, 0°, 15°, 30°$, as shown in Fig. 3. Thus, we generated 8965 2D facial images of 163 identities in total. The ground truth landmarks of the synthesised face dataset were obtained by projecting the 3D vertices to 2D directly. For model training and test, we generated 34 2D landmarks in total, as shown in Fig. 7a and Fig. 8c.

b) Multi-PIE: The Multi-PIE database contains more than 750000 images of 337 people captured in 4 time sessions, with a wide range of pose, expression and illumination variations. We chose all 249 identities in Session-01, with 9 different pose variations, all 20 illumination variations and neutral expression. The total number of the images in the selected subset is 44820. The pose variations of the selected subset are shown in Fig. 6. We used the ground truth landmarks manually annotated by the IDIAP research institute [2]. As the IDIAP ground truth has 16 landmarks for near-frontal faces and 8 landmarks for half-profile faces, we measured the FLD accuracy using the overlapped 8 landmarks (Fig. 6b).

c) BioID: The BioID [58] face dataset has 1521 near-frontal faces with slight pose and expression variations, collected under a lab environment. The dataset was firstly used for face detection and recently for facial landmark detection. Each BioID face has 20 manually annotated landmarks and 17 of them are usually used to test a FLD algorithm [7], as shown in Fig. 12a.

d) HELEN: The HELEN face dataset [59] is a high resolution dataset consisting of 2000 training and 330 test images. Each image in HELEN has 194 annotated landmarks.

e) LFPW: LFPW [29] is a standard FLD benchmark that has 1100 training images and 300 test images collected from the Internet. Each LFPW face has 29 manually annotated landmarks. However, LFPW provides only hyperlinks to the original web images. We were only able to download 797



(a) (b) (c) (d) (e)

Fig. 7. Landmarks for different datasets: (a) the original 34 landmarks of our synthetic dataset; (b) the selected 8 landmarks for Multi-PIE; (c) the selected 17 landmarks for BioID; (d) the selected 29 landmarks for LFPW and COFW; and (e) the expanded 194 landmarks for HELEN.

training and 237 test images because some of the hyperlinks have expired. This is a common problem for experiments on LFPW. All results in [8], [9], [20], [21], [29], [61] are based on different training and test images. This is also the main reason for also using the newly proposed COFW benchmark.

f) COFW: COFW [21] is an expanded version of LFPW. The COFW dataset contains 1345 training images and 507 test images in the wild with a variety of pose, expression, illumination and occlusion variations. But COFW is much more challenging than LFPW, due to a large number of very varied and occluded faces. As in LFPW, each COFW face has 29 manually annotated landmarks.

Note that, to use the proposed CCR algorithm, we have to train our model on a mixed dataset with both synthetic and real images. However, the synthetic and real images used in our evaluation have different numbers of landmarks. To address this issue, we use a subset of our 34 landmarks of the synthetic data for Multi-PIE, BioID, LFPW and COFW; and expand our 34 landmarks to 194 landmarks by linear interpolation for HELEN, as shown in Fig. 7.

2) Measurement metric: We evaluated the proposed algorithm in terms of landmark detection accuracy on all three datasets and also in terms of 3D face reconstruction accuracy on the synthetic dataset, because the ground truth 3D shape and appearance required for the measurement were not available for the other datasets.

As a metric of accuracy of landmark detection, we used the average Euclidean distance between the detected and the ground truth landmarks, normalised by inter-ocular distance:

$$e = \frac{1}{L \cdot E} \sum_{l=1}^{L} \sqrt{(x_l^* - x_l')^2 + (y_l^* - y_l')^2}, \qquad (16)$$

where L is the number of landmarks, E is the inter-ocular distance of the ground truth shape, $[x_l^*, y_l^*]$ are the coordinates of the lth landmark of the ground truth shape, and $[x_l', y_l']$ are the coordinates of the lth landmark of the detected shape.

The accuracy of 3D face reconstruction is measured by the cosine distance in shape, texture and shape plus texture. Taking texture as an example, given a ground-truth 3D texture $\mathbf{t}^*$ and the reconstructed 3D texture $\mathbf{t}'$ from a single 2D image, the similarity between them is measured by the cosine distance:

$$cos(\theta) = \frac{\mathbf{t}^* \cdot \mathbf{t}'}{\|\mathbf{t}^*\| \|\mathbf{t}'\|}, \qquad (17)$$

where θ is the angle between these two vectors.

3) Implementation details: The parameters of the CCR were tuned using on the validation set. The regularisation term λ was set to 1000, the number of weak regressors M in the cascade was set to 5, the shrinking parameter K in the proposed dynamic weighting scheme was set to 2, and the number of scales used for dynamic multi-scale local feature extraction C was set to 2. For COFW, the detected face bounding boxes are provided along with the database. For HELEN and LFPW, we used the face bounding boxes from iBUG [63]. For BioID, a Viola-Jones face detector was applied to obtain the face bounding boxes. For the synthetic dataset and the Multi-PIE dataset, it is hard to detect all the faces due to their wide pose variations. Thus we synthesised the face bounding boxes. We first calculated the ground truth face bounding boxes using the corresponding ground truth shapes. Then we performed random displacements between $[-15\%, 15\%]$ of the width and height of a ground truth face bounding box to its left-upper corner, and then resized its width and height randomly between $[0.85, 1.15]$. These face bounding boxes were used to obtain the initial shape estimate $\mathbf{s}_0$ both in training and test. We used the same initialisation approach as in [9]. The details are also described at the end of Section III. The proposed algorithm was tested on a 3.5GHz CPU with a MATLAB implementation.

B. Experiments on the synthetic dataset

The reason we use this synthetic dataset is mainly to demonstrate the effectiveness of the use of 3D synthesised faces in detecting self-occluded landmarks, and how this influences 3D face reconstruction using a 3DMM. In this part, we did not use the proposed CCR algorithm, because it was developed for a mixed training data. We used a standard linear-regression-based cascade regression (CR) algorithm with different HOG descriptors and their dynamic multi-scale versions, compared with a standard generative view-based AAM using the inverse-compositional fitting algorithm [6]. We compared the landmark detection accuracy using different methods to validate the superiority of the discriminative model to generative model, and also to demonstrate the merit of the proposed dynamic multi-scale local feature extraction approach compared to the standard way. In addition, we evaluated the detection accuracy for self-occluded landmarks and demonstrated how it affects the 3D face reconstruction accuracy from a single 2D facial image. We repeated a 2-fold cross-validation 10 times on the synthetic dataset and used the average value for measurement. The images of the training and test sets were all synthesised faces.

1) Facial landmark detection: In this part, different FLD algorithms were used. We used a standard linear-regression-based CR algorithm in this experiment, with two different HOG features, and compared it with a standard generative view-based AAM [39] and human annotated results. To obtain the human annotated results, we randomly selected 200 test images and manually annotated them. Note that the superiority of the CR-based algorithms has already been validated both in controlled and uncontrolled scenarios by many recent papers [9], [16]–[18], [20]–[22], [26]. Despite the differences

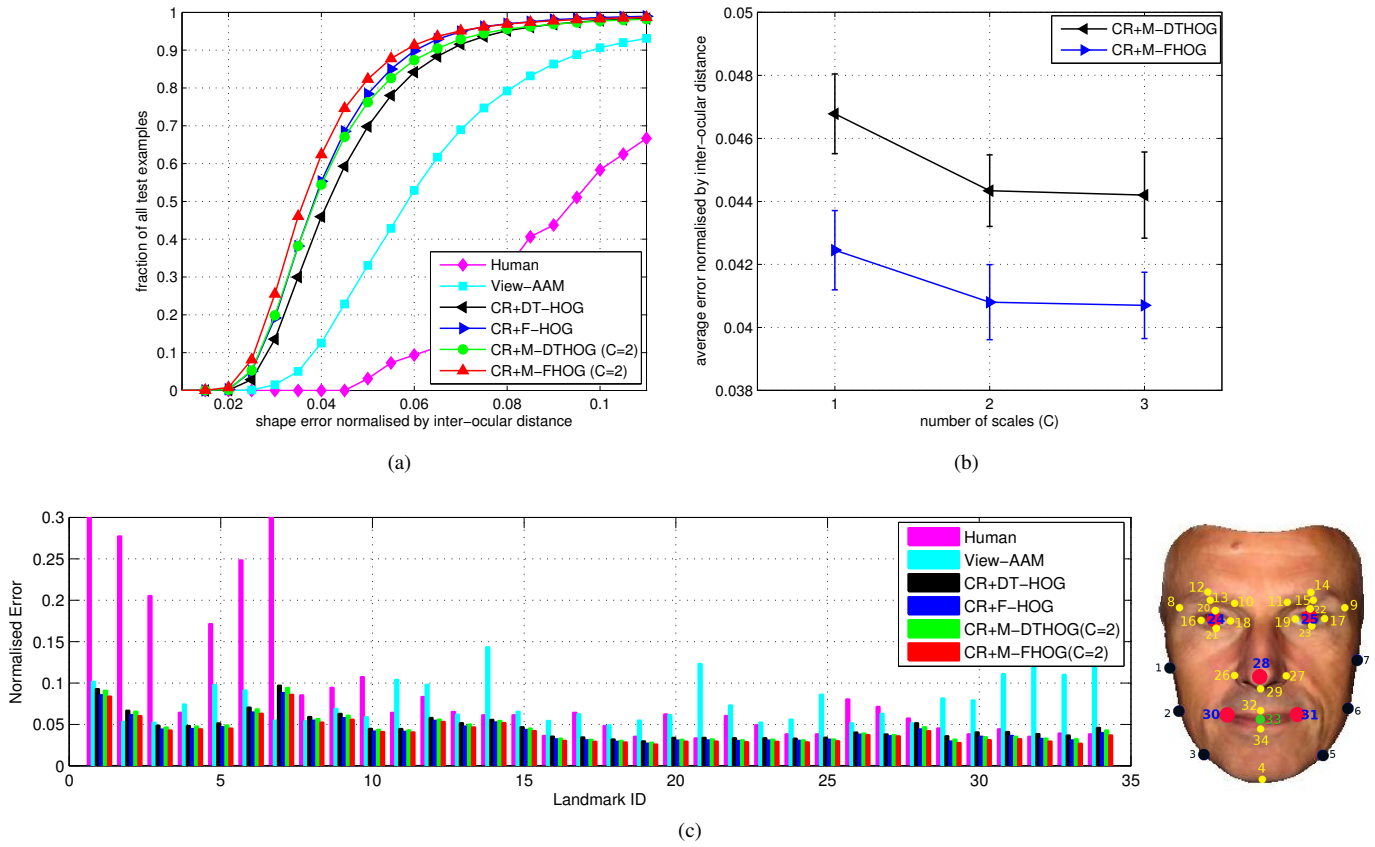


Fig. 8. A comparison of different facial landmark detectors in terms of accuracy on the synthetic dataset: (a) the cumulative distribution curves of different facial landmark detectors, including manually annotated results, view-based AAM and the linear-regression-based CR methods using DT-HOG, F-HOG and their dynamic multi-scale versions; (b) the average error of the CR-based approach with respect to the number of scales for our proposed dynamic multi-scale local feature extraction strategy; (c) performance measured in terms of the accuracy of different landmarks.

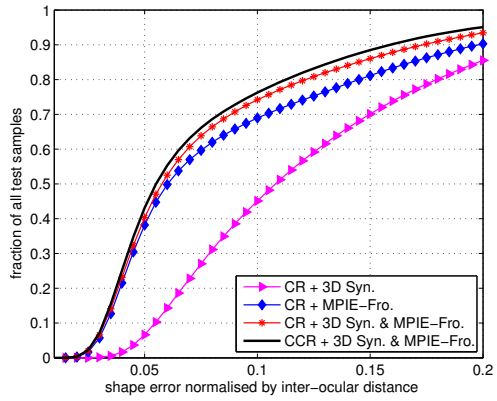
in the type of regression and in the local feature extraction methods we claim that the key to the success of these algorithms is the use of the CR framework [16]. Furthermore, we did not use the proposed cascaded collaborative regression in this part; hence the only difference between the basic CR approach used here and the state-of-the-art SDM [9] is that we use our proposed dynamic multi-scale HOG features rather than SIFT. From our experience, the choice of HOG results in more accurate landmark detection, as demonstrated in Fig. 2.

The landmark detection performance in terms of accuracy of different algorithms is shown in Fig. 8. Firstly, all the automatic detected results exceed the human performance in accuracy, including the view-based AAM. This is because the majority of the test images are non-frontal faces and it is hard to estimate the landmarks of the self-occluded facial parts manually, especially for the landmarks on the outer boundary of a human face (ID 1-7 in Fig. 8c). The human annotation results would be much better than those of the view-based AAM if we were to exclude some occluded landmarks, as shown in Table II. The performance of the manually annotated results across 5 rotation-robust landmarks (eye centres, nose tip and mouth corners) is much better than that measured on all 34 landmarks. Fig. 8c shows the detection accuracy of different algorithms for all 34 landmarks. It is obvious that the landmarks at the occluding boundary of a face contribute much more than the others to the final detection error. Secondly, we measured the contribution of the number of feature scales C to the system. Both Table II and Fig. 8b demonstrate the superiority of the proposed dynamic multi-scale local feature extraction strategy. However, the improvement is minor when we use more than 3 scales for local feature extraction, as indicated by Fig. 8b. The main reason is twofold: 1) the local features extracted from a small window size provide less information for a discriminative FLD learning; 2) the information of the local features extracted from different scales is somewhat redundant. Lastly, the discriminative linear-regression-based CR algorithm performs much better than the generative view-based AAM.

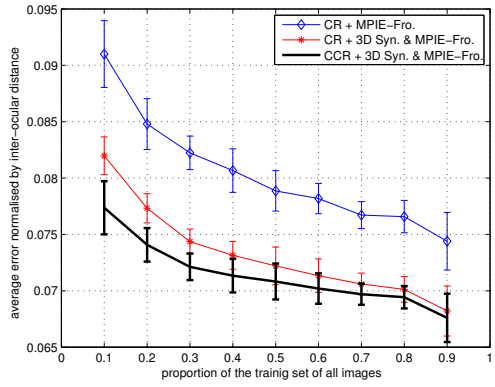
2) 3D face reconstruction: We evaluated the 3D reconstruction performance by fitting a 3DMM to a 2D facial image [54]. The landmarks detected by different algorithms described in the previous subsection, as well as the ground truth landmarks, were used for 3DMM initialisation. Table III shows the cosine correlations between the reconstructed and ground truth 3D information in shape, texture and shape plus texture. Now 3D face reconstruction from a single 2D facial image is a very challenging task. Hence the correlation of shape is very low even when we use the ground truth 2D landmarks for 3DMM initialisation. Similarly to the landmark detection accuracy shown above, the 3D face reconstruction

TABLE II
A COMPARISON OF DIFFERENT FLD ALGORITHMS IN TERMS OF ACCURACY, MEASURED BY THE AVERAGE ERROR USING 34 LANDMARKS AND 5 ROBUST LANDMARKS (ID 24, 25, 28, 30, 31 IN FIG. 8C), NORMALISED BY INTER-OCULAR DISTANCE.

	34-landmark (%)	5-landmark (%)
View-AAM	7.83 ± 0.73	6.08 ± 0.56
CR + DT-HOG	4.68 ± 0.13	3.97 ± 0.12
CR + F-HOG	4.25 ± 0.13	3.52 ± 0.12
CR + M-DTHOG (C=2)	4.43 ± 0.11	3.56 ± 0.11
CR + M-FHOG (C=2)	4.08 ± 0.12	3.25 ± 0.12
Human	*9.62*	*4.31*



(a)



(b)

Fig. 9. A comparison of the detectors trained from the proposed CCR and the one-off CR in terms of accuracy on Multi-PIE, using three different training datasets (3D Syn.: 3D synthesised faces, MPIE-Fro.: Multi-PIE faces with only 5 near-frontal pose variations and 3D Syn. & MPIE-Fro.: a combined one with all above): (a) the cumulative distribution curves of different algorithms when 10% of selected Multi-PIE faces are used as the training subset; (b) the average normalised errors of different algorithms with respect to the proportion of the training subset of all selected Multi-PIE faces.

accuracy initialised by the dynamic multi-scale F-HOG CR is superior to all the others, especially to the manually annotated 2D facial landmarks. The main reason is that the trained CR-based model estimates the landmarks of the self-occluded face parts better. In contrast, it is very hard to estimate the landmarks of the occluded facial parts manually.

C. Experiments on Multi-PIE

The experiments on Multi-PIE explore the effectiveness of the use of the synthesised faces as a complementary training set for our proposed CCR algorithm. We evaluated the algorithms in confirmation with different types of training data varying both in quantity and quality. Here, the term 'quantity' stands for the number of the training samples and the term 'quality' stands for the variety of the poses and face appearance in the training set. To vary the quantity and examine the quantitative relationship between the size of training data and the algorithm performance, we split the selected subset into training and test sets with different proportions. We randomly selected a training subset with different proportions $(10\%, 20\%, \cdots, 90\%)$ of the available identities and used the remaining identities as the test set. To vary the quality, we designed two different protocols: one had a training subset with incomplete pose variations and the other had a training subset with all pose variations. The first protocol used 5 near-frontal poses in the training set for landmark detector training, whereas the second protocol used all 9 poses. Thus the size of the real training dataset varies from around 2500 to 22500 images for the first protocol, and from around 4500 to 40500 for the second protocol. We repeated each random selection 10 times and reported the average value. As the relative performance of different HOG descriptors has already been shown in the last section, we only use the proposed dynamic multi-scale F-HOG in this part.

1) Training set with incomplete pose variations: In this protocol, to examine the effectiveness of the use of 3D synthesised faces, we generate three different training datasets using: a) only 3D synthesised facial images; b) only Multi-PIE faces with 5 near-frontal poses; c) a combination of a) and b) with mixed training samples. For the first two training sets, we use the classical CR training, because they only contain either real or synthesised faces. For the last one with mixed training images, we use the one-off CR and the proposed CCR methods.

First, note that the detection results obtained using only 3D synthesised faces do not adapt to real faces well (Fig. 9a), because the latter contains a wide range of variations in appearance exhibited by real faces. Second, as expected, increasing the number of training samples improves the accuracy of the FLD (Fig. 9b). This confirms that a large amount of training data is crucial to the success of a regression-based facial landmark detector training. However, this large training data is not always available in practical applications. Third, the use of the 3D synthesised faces as additional training samples improves the performance of the existing linear-regression-based CR approach. Last, the proposed CCR algorithm improves the performance in accuracy even further, especially when we have a small number of training samples (Fig. 9b).

In this part, we also evaluated the performance of the proposed CCR method using different dynamic weighting functions and parameters, as shown in Fig. 10. We investigated the effectiveness of the proposed dynamic weighting function $\omega(m)$ as a function of the shrinking rate parameter K in

TABLE III
A COMPARISON OF THE ACCURACY OF 3D FACE RECONSTRUCTION USING DIFFERENT 2D FACIAL LANDMARKS AS INITIALISATION, MEASURED IN THE COSINE DISTANCE BETWEEN RECONSTRUCTED AND GROUND TRUTH 3D FACES IN SHAPE, TEXTURE AND THE CONCATENATED 3D SHAPE AND TEXTURE.

	shape	texture	shape + texture
View-AAM	-0.0412 ± 0.0078	0.4887 ± 0.0145	0.4194 ± 0.0128
CR + DT-HOG	0.1866 ± 0.0068	0.5978 ± 0.0106	0.5239 ± 0.0094
CR + F-HOG	0.2159 ± 0.0097	0.6096 ± 0.0109	0.5365 ± 0.0096
CR + M-DTHOG (C=2)	0.2127 ± 0.0084	0.6118 ± 0.0103	0.5387 ± 0.0095
CR + M-FHOG (C=2)	**0.2343 ± 0.0096**	**0.6216 ± 0.0101**	**0.5482 ± 0.0090**
Ground Truth	*0.4771*	*0.7000*	*0.6388*
Human	*0.1593*	*0.4659*	*0.4051*

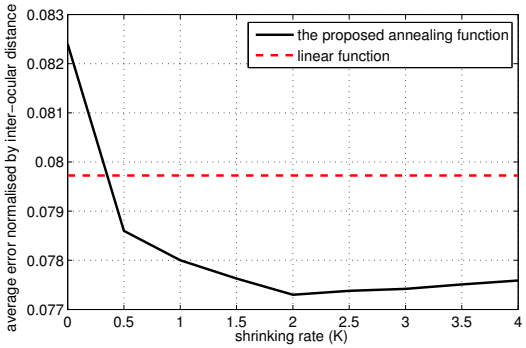
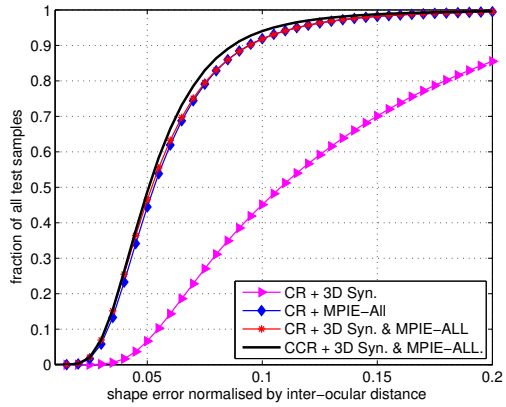


Fig. 10. A comparison of the accuracy of the proposed CCR algorithm with respect to the shrinking parameter K. A simple decreasing linear function (shown in Fig. 4) is also used for comparison. The result is evaluated on Multi-PIE using a mixed training dataset. For CCR training, both the 3D synthesised faces and 10% 2D frontal real images from Multi-PIE are used. The remaining 90% real images are used for test.

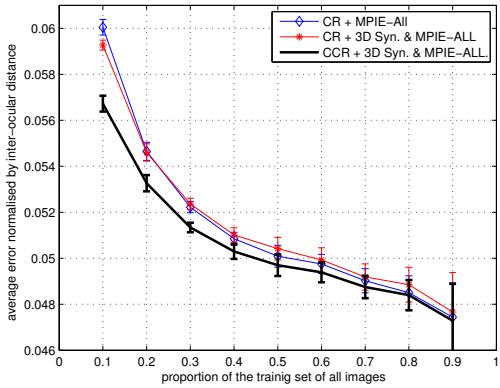
equation (10). Also, we used a linear function (shown in Fig. 4) as a baseline to further validate the superiority of our proposed dynamic weighting function. The proposed CCR algorithm degenerates to a classical CR method simply trained on a mixed dataset when we set K to 0; hence the system cannot achieve very high accuracy. As K increases, the average error of the proposed CCR firstly goes down and then slightly up. This result indicates that the weights of the synthetic and real datasets should be balanced carefully. The system cannot generalise well when the real image data prematurely dominates the system, and cannot adapt to real images well when the real image data engage in the training too late. Also, using the proposed annealing function $\omega(m)$ is better than simply using a decreasing linear function.

2) Training set with a complete range of pose variations: As stated above, in spite of the number of training samples, the use of 3D synthesised faces as a complementary training subset improves the final results, especially in conjunction with the proposed CCR algorithm. Thus, it is interesting to ask why the use of the 3D synthesised faces helps to train the facial landmark detector even when the original training set already contains wide ranging pose variations.

In this protocol, we generate three different training sets using: a) only 3D synthesised facial images; b) only Multi-PIE faces with all pose variations; c) a combination of a) and b). As in the previous protocol, we only use the one-off CR method for the first two training sets, and use both the one-off CR and the proposed CCR methods for the last one with mixed training samples.



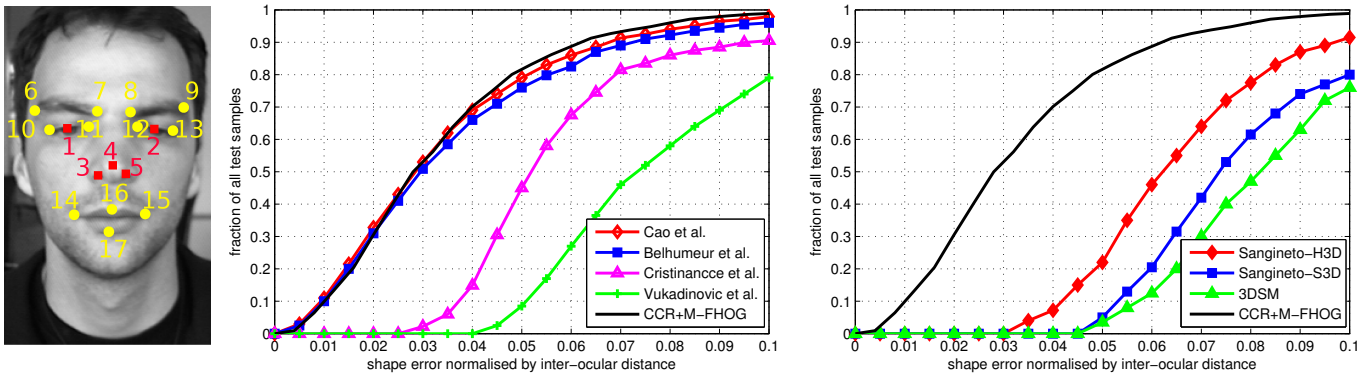
(a)



(b)

Fig. 11. A comparison of the detectors trained from the proposed CCR and the one-off CR in terms of accuracy on Multi-PIE, using three different training datasets (3D Syn.: 3D synthesised faces, MPIE-ALL: Multi-PIE faces with all pose variations and 3D Syn. & MPIE-ALL: a combined one with all above): (a) the cumulative distribution curves of different algorithms when 10% of selected Multi-PIE faces are used as the training subset; (b) the average normalised errors of different algorithms with respect to the proportion of the training subset of all selected Multi-PIE faces.

First, similar to the first protocol, increasing the number of training samples improves the performance of a trained facial landmark detector (Fig. 11b). Second, the performance of the landmark detector trained on all Multi-PIE faces is much better than that trained on only near-frontal Multi-PIE faces (Fig. 9 vs. Fig. 11), because the training set used in this protocol covers all pose variations in the test set. Third, using the



(a) (b) (c)

Fig. 12. A comparison of the proposed method to a set of state-of-the-art algorithms (Cao *et al.* [8], Belhumeur *et al.* [29], Cristinacce *et al.* [7], Vukadinovic *et al.* [64], Sangineto [62] and 3DSM [65]) on BioID: (a) the 17 (ID 1-17) and 12 (ID 6-17) landmarks used for comparison; (b) the cumulative distribution curves of different algorithms measured on 17 landmarks; and (c) the cumulative distribution curves of different algorithms measured on 12 landmarks.

one-off CR training improves the trained detector marginally only when we have a small number of training samples. In contrast, the proposed CCR algorithm improves the detection performance even when the variations in the training set cover all possible variations in the test set. Last, when the training set is large enough, *e.g.* when 90% of all images are selected as the training set (Fig. 11b), the improvement is marginal. This is because the size of the training set is already very large ($44820 * 90\% = 40338$). However, in practical applications, it is hard to obtain such a large training data set with different pose variations. Thus, the use of 3D synthesised faces as a complementary training dataset is an excellent alternative.

D. Experiments on BioID, HELEN, LFPW and COFW

To validate the superiority of the proposed CCR algorithm, we compared it with a set of state-of-the-art algorithms on the BioID, LFPW, HELEN and COFW datasets. However, it is a very difficult work to compare with different algorithms across these datasets due to different measurement metrics used in previous work. For example, [7], [8], [29], [64] evaluated the performance of their methods in accuracy using 17 landmarks on BioID, whereas [62], [65] measured their approaches using 12 landmarks. Also, different presentation styles, such as a figure with cumulative distribution curve and a table with average error, have been used. In this part, to make an comprehensive comparison, we compared the performance of our algorithm in accuracy on the BioID dataset using both 17 and 12 landmarks with a cumulative distribution curve figure. On HELEN, LFPW and COFW, we compared our approach to the other algorithms using a table with average error, failure rate and speed, as used in [8], [21], [22]. The average error was normalised by the inter-ocular distance across 194 landmarks on HELEN, 17 and 29 landmarks on LFPW, and 29 landmarks on COFW. The failure rate is defined by the percentage of the failed detected images whose average error is bigger than 10% of the inter-ocular distance, as in [21]. The speed is measured in fps (frames-per-second). We also discuss the efficiency of the proposed CCR algorithm at the end of this subsection.

TABLE IV
COMPARISON ON **HELEN**. THE ERROR WAS MEASURED ON 194 LANDMARKS AND NORMALISED BY THE INTER-OCULAR DISTANCE.

method	error (%)	failures (%)	speed (fps)
ST-ASM [66]	11.1	-	-
Le *et al.* [59]	9.1	-	-
Cao *et al.* [8]	7.1	13	2
RCPR [21]	6.5	8	6
SDM [9]	5.85	-	21
LBF [22]	5.41	-	**200**
CCR + M-FHOG	**5.25**	**7.1**	24

1) BioID: To make a comparison on BioID, we trained our CCR model on our synthetic dataset and the COFW dataset. The performance of different algorithms in accuracy is presented in Fig. 12, using both 17 and 12 landmarks. As the main purpose of the use of our synthetic data is to deal with the difficulty caused by pose variations and the BioID dataset only consists of near-frontal faces, the superiority of the proposed CCR is not very significant compared to some state-of-the-art methods, *e.g.* [8] and [29]. However, CCR outperforms all the other algorithms in accuracy.

2) HELEN: A face image in the HELEN dataset contains 194 landmarks, whereas the synthesised face has only 34 landmarks. To address this problem, we regenerated 194 landmarks using linear interpolation on our existing 34 landmarks, as shown in Fig. 7e. The results on HELEN are demonstrated in Table IV. Note that, the results of [8] were provided by [21] using their re-implementation. The proposed CCR performs best both in the average error and the failure rate, and only slower than LBF [22]. However, the LBF method was tested on a more powerful CPU than us. Also, the speed of the proposed CCR is still competitive compared with all the others and is sufficient for real-time applications.

3) LFPW & COFW: To use the synthetic dataset as a complementary training set on LFPW and COFW, we first chose a landmark subset with 28 landmarks from our all 34 landmarks (ID 4, 8-34 in Fig. 8c). Then we used the landmark with ID 33 twice to obtain the final landmark subset. The

TABLE V
COMPARISON ON **LFPW**. THE ERROR WAS MEASURED BOTH ON 17 AND 29 LANDMARKS AND NORMALISED BY THE INTER-OCULAR DISTANCE.

method	error (%)		failures (%)	speed (fps)
	me17	*me29*		
Zhou *et al.* [61]	3.89	3.92	-	25
Cao *et al.* [8]	-	3.43	-	20
SDM [9]	-	3.47	-	30
RCPR (full) [21]	-	3.50	2.00	12
DRMF *et al.* [20]	6.50	-	5.74	1
Belhumeur *et al.* [29]	3.96	3.99	≈6	1
LBF [22]	-	3.35	-	**460**
RCRC [26]	3.29	3.31	0.84	21
CCR + M-FHOG	**3.28**	**3.29**	**0.84**	69

TABLE VI
COMPARISON ON **COFW**. THE ERROR WAS MEASURED ON 29 LANDMARKS AND NORMALISED BY THE INTER-OCULAR DISTANCE.

method	error (%)	failures (%)	speed (fps)
Zhu *et al.* [60]	14.4	80	0.1
Cao *et al.* [8]	11.2	36	4
RCPR (full) [21]	8.5	20	3
HPM (LFPW68) [33]	7.5	13	0.03
HPM (HELEN68) [33]	7.8	17	0.03
RCRC [26]	7.3	12	22
CCR + M-FHOG	**7.03**	**10.9**	**69**

results on LFPW and COFW are shown in Table V and Table VI, respectively. The proposed CCR beats all the others both in the average error and the failure rate, along with a competitive speed measured in fps (frames per second). The average errors are measured both on 29 landmarks and 17 landmarks for LFPW, and only on 29 landmarks for COFW. The selected 17 landmarks are the same as on BioID (Fig. 12a). The LFPW dataset is less changing than COFW, due to many images with very varied poses and occlusions in COFW; hence the superiority of the proposed CCR on COFW is more significant than that on LFPW. Some examples of the detected landmarks on COFW are shown in Fig. 13.

It is worth noting that the comparisons above are not totally fair because we used a synthetic dataset as a complementary training dataset. However, the other algorithms also augmented the training data in different ways. For instance, HPM [33] generated 4 extra images per training image by putting some artificial occlusions to the original images, and the RCRC method augmented the training data of COFW 10 times (1345×10) by giving different initialised shapes that is even larger than the number of our mixed training samples (8965). This demonstrates that the use of synthesised faces as an augmentation method is more helpful.

4) The efficiency of CCR: As discussed above, the speed of the proposed CCR on HELEN is almost half of that on LFPW and COFW. The reason is that the speed of the proposed algorithm is highly related to many factors, such as the number of landmarks (L). To speed up our algorithm on HELEN, we only extracted local features around a subset of all 194 landmarks. Also, the speed of proposed algorithm is related to some parameters used in our system, including the number of cascaded weak regressors (M), and the number

Fig. 13. Detected landmarks on COFW by the proposed CCR algorithm.

of the scales used to extract local features (C). However, the shrinking rate (K) used in the weighting function is not relevant to the speed of the algorithm, but does affect the accuracy seriously as shown in Fig. 10. Note that the increase of C and L leads to a quadratic increase in the computational and memory costs of the training step. Although this can be carried out offline, it also leads to a linear increase in the computational and memory costs of online testing. The increase of M leads to a linear increase in the computational and memory costs of the offline training and online testing steps. In practical applications, these parameters should be tuned carefully according to the running environment and the requirements of a specific task.

VI. CONCLUSION

We have presented a supervised cascaded collaborative regression (CCR) algorithm that exploits synthesised faces for robust FLD in 2D. The use of synthesised faces generated from a 3DMM greatly improves the generalisation capability of the trained facial landmark detector. A major advantage of using synthesised faces is that they do not need to be manually annotated. Furthermore, the 2D landmarks of the synthesised faces by direct projection from 3D to 2D are accurate, especially for the self-occluded facial parts with large pitch and yaw rotations.

However, confining the training of the facial landmark detector to merely synthesised faces improves the performance in real applications only marginally. To effectively exploit the 3D synthesised training samples and adapt the trained facial landmark detector to realistic facial images, we proposed the CCR algorithm. Here, the mixed training data dominates the training of the first few sub-regressors in the cascaded strong regressor, which enables the landmark detector to accommodate pose variations; and the smaller set of real faces dominates the training of the last few sub-regressors, which enhances the accuracy of the final shape estimates. To extract more informative local features, we designed a

dynamic multi-scale local feature extraction scheme, which further improved the accuracy of the learned regressor. The results of a number of datasets demonstrated the superiority of the proposed algorithm.

REFERENCES

[1] D. Chen, X. Cao, F. Wen, and J. Sun, "Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3025–3032.

[2] L. El Shafey, C. McCool, R. Wallace, and S. Marcel, "A Scalable Formulation of Probabilistic Linear Discriminant Analysis: Applied to Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1788–1794, Jul. 2013.

[3] H.-S. Lee and D. Kim, "Tensor-Based AAM with Continuous Variation Estimation: Application to Variation-Robust Face Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 6, pp. 1102–1116, 2009.

[4] T. Cootes, C. Taylor, D. Cooper, J. Graham *et al.*, "Active shape models-their training and application," *Computer Vision and Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[5] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," in *European Conference on Computer Vision*, 1998, pp. 484–498.

[6] I. Matthews and S. Baker, "Active Appearance Models Revisited," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 135–164, 2004.

[7] D. Cristinacce and T. F. Cootes, "Feature Detection and Tracking with Constrained Local Models," in *British Machine Vision Conference*, 2006, pp. 929–938.

[8] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2012, pp. 2887–2894.

[9] X. Xiong and F. De la Torre, "Supervised Descent Method and Its Applications to Face Alignment," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 532–539.

[10] P. Nair and A. Cavallaro, "3-D face detection, landmark localization, and registration using a point distribution model," *IEEE Transactions on Multimedia*, vol. 11, no. 4, pp. 611–623, 2009.

[11] M. Pamplona Segundo, L. Silva, O. R. P. Bellon, and C. C. Queirolo, "Automatic face segmentation and facial landmark detection in range images," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 40, no. 5, pp. 1319–1330, 2010.

[12] M. Song, D. Tao, S. Sun, C. Chen, and S. Maybank, "Robust 3D Face Landmark Localization Based on Local Coordinate Coding," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5108–5122, Dec 2014.

[13] P. Perakis, G. Passalis, T. Theoharis, and I. A. Kakadiaris, "3D facial landmark detection under large yaw and expression variations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 7, pp. 1552–1564, 2013.

[14] G. J. Edwards, T. F. Cootes, and C. J. Taylor, "Face recognition using active appearance models," in *European Conference on Computer Vision*. Springer, 1998, pp. 581–595.

[15] X. Liu, "Discriminative face alignment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 11, pp. 1941–1954, 2009.

[16] P. Dollár, P. Welinder, and P. Perona, "Cascaded pose regression," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 1078–1085.

[17] J. Yan, Z. Lei, D. Yi, and S. Z. Li, "Learn to Combine Multiple Hypotheses for Accurate Face Alignment," in *International Conference on Computer Vision Workshops*, 2013.

[18] Y. Sun, X. Wang, and X. Tang, "Deep Convolutional Network Cascade for Facial Point Detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3476–3483.

[19] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive Facial Landmark Localization with Coarse-to-Fine Convolutional Network Cascade," in *International Conference on Computer Vision Workshops*, 2013, pp. 386–391.

[20] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3444–3451.

[21] X. P. Burgos-Artizzu, P. Perona, and P. Dollár, "Robust face landmark estimation under occlusion," in *International Conference on Computer Vision*, 2013.

[22] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face Alignment at 3000 FPS via Regressing Local Binary Features," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[23] J. Zhang, S. Shan, M. Kan, and X. Chen, "Coarse-to-Fine Auto-Encoder Networks (CFAN) for Real-Time Face Alignment," in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2014, vol. 8690, pp. 1–16.

[24] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint Cascade Face Detection and Alignment," in *European Conference on Computer Vision*. Springer, 2014, pp. 109–122.

[25] Z. Zhang, P. Luo, C. C. Loy, and X. Tang, "Facial landmark detection by deep multi-task learning," in *European Conference on Computer Vision*. Springer, 2014, pp. 94–108.

[26] Z.-H. Feng, P. Huber, J. Kittler, W. Christmas, and X.-J. Wu, "Random Cascaded-Regression Copse for Robust Facial Landmark Detection," *IEEE Signal Processing Letters*, vol. 22, no. 1, pp. 76–80, Jan 2015.

[27] M. Banko and E. Brill, "Scaling to very very large corpora for natural language disambiguation," in *the 39th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2001, pp. 26–33.

[28] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes, "Do We Need More Training Data or Better Models for Object Detection?." in *British Machine Vision Conference*, vol. 3. Citeseer, 2012, p. 5.

[29] P. N. Belhumeur, D. W. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 545–552.

[30] G. B. Huang, M. Mattar, T. Berg, E. Learned-Miller *et al.*, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition*, 2008.

[31] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[32] M. Enzweiler and D. M. Gavrila, "A mixed generative-discriminative framework for pedestrian classification," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[33] G. Ghiasi and C. C. Fowlkes, "Occlusion Coherence: Localizing Occluded Faces with a Hierarchical Deformable Part Model," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, June 2014.

[34] M. Rätsch, P. Quick, P. Huber, T. Frank, and T. Vetter, "Wavelet reduced support vector regression for efficient and robust head pose estimation," in *IEEE Conference on Computer and Robotics Vision*, 2012, pp. 260–267.

[35] L. Pishchulin, A. Jain, M. Andriluka, T. Thormahlen, and B. Schiele, "Articulated people detection and pose estimation: Reshaping the future," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3178–3185.

[36] B. Pepik, M. Stark, P. Gehler, and B. Schiele, "Teaching 3d geometry to deformable part models," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 3362–3369.

[37] V. Blanz and T. Vetter, "A morphable model for the synthesis of 3D faces," in *the 26th Annual Conference on Computer Graphics and Interactive Techniques*. ACM, 1999, pp. 187–194.

[38] ——, "Face recognition based on fitting a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, no. 9, pp. 1063–1074, 2003.

[39] T. Cootes, K. Walker, and C. Taylor, "View-based active appearance models," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2000, pp. 227–232.

[40] P. Mortazavian, J. Kittler, and W. Christmas, "3D morphable model fitting for low-resolution facial images," in *International Conference on Biometrics*. IEEE, 2012, pp. 132–138.

[41] O. Aldrian and W. A. Smith, "Inverse rendering of faces with a 3D morphable model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 5, pp. 1080–1093, 2013.

[42] S. Schönborn, A. Forster, B. Egger, and T. Vetter, "A Monte Carlo Strategy to Integrate Detection and Model-Based Face Analysis," in *Pattern Recognition*. Springer, 2013, pp. 101–110.

[43] J. Gonzalez-Mora, F. De la Torre, R. Murthi, N. Guil, and E. Zapata, "Bilinear Active Appearance Models," in *International Conference on Computer Vision*, 2007, pp. 1–8.

[44] Z.-H. Feng, J. Kittler, W. Christmas, X.-J. Wu, and S. Pfeiffer, "Automatic face annotation by multilinear AAM with Missing Values," in *International Conference on Pattern Recognition*, 2012, pp. 2586–2589.

[45] T. Cootes, E. Baldock, and J. Graham, "An introduction to active shape models," *Image Processing and Analysis*, pp. 223–248, 2000.

[46] S. Romdhani, S. Gong, A. Psarrou *et al.*, "A Multi-View Nonlinear Active Shape Model Using Kernel PCA." in *British Machine Vision Conference*, vol. 10, 1999, pp. 483–492.

[47] C. J. Twining and C. J. Taylor, "Kernel principal component analysis and the construction of non-linear active shape models." in *British Machine Vision Conference*, 2001, pp. 1–10.

[48] T. Cootes, G. Edwards, and C. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 6, pp. 681–685, 2001.

[49] R. Donner, M. Reiter, G. Langs, P. Peloschek, and H. Bischof, "Fast active appearance model search using canonical correlation analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, p. 1690, 2006.

[50] D. Tang, T.-H. Yu, and T.-K. Kim, "Real-time articulated hand pose estimation using semi-supervised transductive regression forests," in *International Conference on Computer Vision Workshops*. IEEE, 2013, pp. 3224–3231.

[51] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[52] V. Blanz, P. Grother, P. J. Phillips, and T. Vetter, "Face recognition based on frontal views generated from non-frontal images," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 454–461.

[53] G. Hu, C.-H. Chan, J. Kittler, and B. Christmas, "Resolution-Aware 3D Morphable Model," in *British Machine Vision Conference*, vol. 12, 2012, pp. 1–10.

[54] G. Hu, P. Mortazavian, J. Kittler, and W. Christmas, "A facial symmetry prior for improved illumination fitting of 3D morphable model," in *International Conference on Biometrics*. IEEE, 2013, pp. 1–6.

[55] J. T. Rodriguez, "3D Face Modelling for 2D+3D Face Recognition," Ph.D. dissertation, University of Surrey, Guildford, UK, 2007.

[56] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[57] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-PIE," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[58] O. Jesorsky, K. J. Kirchberg, and R. W. Frischholz, "Robust face detection using the hausdorff distance," in *Audio-and video-based biometric person authentication*. Springer, 2001, pp. 90–95.

[59] V. Le, J. Brandt, Z. Lin, L. Bourdev, and T. S. Huang, "Interactive facial feature localization," in *European Conference on Computer Vision*. Springer, 2012, pp. 679–692.

[60] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 2879–2886.

[61] F. Zhou, J. Brandt, and Z. Lin, "Exemplar-based Graph Matching for Robust Facial Landmark Localization," in *International Conference on Computer Vision*, December 2013.

[62] E. Sangineto, "Pose and expression independent facial landmark localization using dense-SURF and the Hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 624–638, 2013.

[63] C. Sagonas, G. Tzimiropoulos, S. Zafeiriou, and M. Pantic, "300 Faces in-the-Wild Challenge: The first facial landmark localization Challenge," in *International Conference on Computer Vision Workshops*. IEEE, 2013, pp. 397–403.

[64] D. Vukadinovic and M. Pantic, "Fully automatic facial feature point detection using Gabor feature based boosted classifiers," in *IEEE International Conference on Systems, Man and Cybernetics*, vol. 2, Oct 2005, pp. 1692–1698.

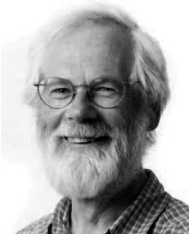
[65] A. Caunce, D. Cristinacce, C. J. Taylor, and T. F. Cootes, "Locating Facial Features and Pose Estimation Using a 3D Shape Model." in *ISVC (1)*, ser. Lecture Notes in Computer Science, vol. 5875. Springer, 2009, pp. 750–761.

[66] S. Milborrow and F. Nicolls, "Locating Facial Features with an Extended Active Shape Model." in *European Conference on Computer Vision*, ser. Lecture Notes in Computer Science, vol. 5305. Springer, 2008, pp. 504–513.

Zhen-Hua Feng (S'13) received the B.Sc. degree in Mathematics and the M.Eng. degree in Computer Science both from Jiangnan University, Wuxi, China, in 2007 and 2009, respectively. He is now a PhD student at the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. His research topic is 'Pose invariant face recognition using 2D morphable models', and his research interests include image processing, pattern recognition, machine learning and computer vision.

Guosheng Hu (S'13) received the B.S degree in Computer Science from Dalian Nationalities University, Dalian, China in 2007, and the M.Sc. degree in Computer Science from Jilin University, Jilin, China, in 2010. He is currently a PhD student at the Centre for Vision, Speech and Signal Processing, University of Surrey, Guildford, U.K. His research topic is 'Face analysis using 3D morphable models', and his research interests include image understanding, computer vision and machine intelligence.

Josef Kittler (M'74-LM'12) received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively. He is Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering, University of Surrey, Guildford, U.K. He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He published the textbook Pattern Recognition: A Statistical Approach (Englewood Cliffs, NJ, USA: Prentice-Hall, 1982) and over 600 scientific papers. He serves on the Editorial Board of several scientific journals in pattern recognition and computer vision.



William Christmas received the B.A. degree from the University of Oxford, Oxford, U.K., and the Ph.D. degree from the University of Surrey, Guildford, U.K. After graduating, he held posts with BBC and BP Research International, researching on hardware and software aspects of parallel processing, real-time image processing, and computer vision. He is currently a university fellow in technology transfer at the University of Surrey. His current research interests include the integration of machine vision algorithms to create complete applications.

Xiao-Jun Wu received the B.Sc. degree in mathematics from Nanjing Normal University, Nanjing, China, in 1991. He received the M.S. degree and the Ph.D. degree in pattern recognition and intelligent systems from Nanjing University of Science and Technology, Nanjing, China, in 1996 and 2002, respectively. He is currently a Professor in artificial intelligent and pattern recognition at Jiangnan University, Wuxi, China. His current research interests include pattern recognition, computer vision, fuzzy systems, neural networks, and intelligent systems.