Robust and Accurate Shape Model Matching Using Random Forest Regression-Voting

Claudia Lindner, Paul A. Bromiley, Mircea C. Ionita, and Tim F. Cootes

Abstract—A widely used approach for locating points on deformable objects in images is to generate feature response images for each point, and then to fit a shape model to these response images. We demonstrate that Random Forest regression-voting can be used to generate high quality response images quickly. Rather than using a generative or a discriminative model to evaluate each pixel, a regressor is used to cast votes for the optimal position of each point. We show that this leads to fast and accurate shape model matching when applied in the Constrained Local Model framework. We evaluate the technique in detail, and compare it with a range of commonly used alternatives across application areas: the annotation of the joints of the hands in radiographs and the detection of feature points in facial images. We show that our approach outperforms alternative techniques, achieving what we believe to be the most accurate results yet published for hand joint annotation and state-of-the-art performance for facial feature point detection.

Index Terms—Computer vision, Random Forests, Constrained Local Models, statistical shape model, feature point detection

1 Introduction

The ability to accurately detect feature points of deformable models is important for a wide range of algorithms and applications. A widely used approach is to apply a statistical shape model to regularise the output of independent feature detectors trained to locate each point. Examples include Active Shape Models (ASMs) [1], [2], Pictorial Structures [3] and Constrained Local Models (CLMs) [4], [5], though there are many others.

The task of the feature point detector is to compute a (pseudo-) probability that the target point occurs at a particular position, given the image information $p(\mathbf{x}|I)$ (where techniques return a quality-of-fit measure, we assume these can be converted to a pseudo-probability with a suitable transformation). Local peaks in this correspond to candidate positions (e.g. in ASMs), or the probabilities for each point are combined with the shape model information to find the best overall match (e.g. CLMs and Pictorial Structures). A wide variety of feature point detectors have been used in such frameworks which can be broadly classified into three types:

Generative in which generative models are used, so $p(\mathbf{x}|I) \propto p(I|\mathbf{x})$.

Discriminative in which classifiers are trained to estimate $p(\mathbf{x}|I)$ directly.

Regression-Voting in which $p(\mathbf{x}|I)$ is estimated from accumulating votes for the position of the point given information in nearby regions.

Although there has been a great deal of work matching deformable models using the first two approaches, the regression-voting approach has only recently begun to be explored in this context. Work on Hough Forests [6] has shown that objects can be effectively located by pooling votes from Random Forest (RF) [7] regressors, and that facial feature points can be accurately located following a similar approach using kernel SVM-based regressors [8].

In the following, we show that regression-voting is a powerful technique, and that using RF regression-voting (RFRV) leads to fast, accurate and robust feature point detection results when used in the CLM framework. Preliminary outputs of this work were presented in [9] for introducing RFRV in the CLM framework and in [10] for analysing what properties of RFRV contribute to its highly accurate and robust performance. This paper expands on the latter in various ways: *(i)* We provide a more detailed description of RFRV and how it is used in the CLM framework. *(ii)* We apply the proposed feature point detection method as part of a fully automatic detection system for two application areas, the annotation of the joints of the hands in radiographs and facial feature point detection in the BioID and AFLW datasets. *(iii)* We perform additional in-depth experiments to investigate the nature of the performance of this approach, and to analyse the impact of the choice of parameters—making suggestions for best choices. *(iv)* We give significantly improved results for both hand joint annotation and facial feature point detection.

We demonstrate that our approach outperforms alternative feature point detection techniques across application areas, achieving what we believe to be the most accurate results yet published for hand joint annotation and state-of-the-art performance for facial feature point detection.

2 Related Work

Shape model matching. There is a wide range of literature on matching statistical shape models to images, starting with Active Shape Models [1] in which the shape model is fit to the results of searching around each feature point with a

• *The authors are with the Centre for Imaging Sciences, The University of Manchester, Manchester, United Kingdom. E-mail: claudia.lindner@manchester.ac.uk.*

Manuscript received 7 Feb. 2014; revised 4 Nov. 2014; accepted 24 Nov. 2014. Date of publication 17 Dec. 2014; date of current version 7 Aug. 2015. Recommended for acceptance by F. de la Torre. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below. Digital Object Identifier no. 10.1109/TPAMI.2014.2382106

suitably trained detector. Active Appearance Models (AAMs) [11] match combined models of shape and texture using an efficient parameter update scheme. Pictorial Structures [3] introduced an efficient method of matching part-based models to images, in which shape is encoded in the geometric relationships between pairs of parts. CLMs [4], [5] build on a framework in which response images are computed estimating the quality-of-fit of each feature point at a set of positions in the target image, then a shape model is matched to the data, selecting the overall best combination of points.

Belhumeur et al. [12] have shown impressive facial feature detection results using sliding window detectors (SVM classifiers trained on SIFT features) combined with a RANSAC approach to selecting good combinations of feature points. Recently, these results were marginally improved upon by combining feature point candidates and shape constraints using a graph matching algorithm where the position of each feature point is modelled as a weighted linear combination of all other feature points [13]. An alternative approach to combine candidate points and shape constraints using a graph matching algorithm was proposed in the context of detecting anatomical landmarks where a Markov Random Field is used to encode the mutual location information of the most predictive landmarks [14]. Recently, Nicolle et al. [15] have obtained excellent results by combining binary map cross-correlations with an implicit shape model based on the spatial relationships between point triplets.

Random Forests. RFs [7] have a number of favourable criteria that make them very suitable for application to many computer vision problems. They are fast to train and to evaluate, robust to noise, parallelisable and yield good performance for high-dimensional input data. RFs have been shown to be effective in a wide range of classification and regression problems [16], [17], [18].

Recently, a number of extensions to the standard RF regression approach have been suggested, including Alternating Decision Forests [19] and Alternating Regression Forests [20] where the aim is not only to minimise the uncertainty of the predictions locally on the node level of every tree independently but across all trees in the forest, as well as Neighbourhood Approximation Forests [21] which apply RFs to efficiently retrieve nearest-neighbour images.

Regression-based matching. One of the earliest examples of regression-based matching techniques was the work of Covell [22] who used linear regression to predict the positions of points on the face. The original AAM [23] algorithm uses linear regression to predict the updates to model parameters. Non-linear AAM extensions include the use of Boosted Regression [24], [25] and RF Regression [26]. The Shape Regression Machine [27] uses boosted regression to predict shape model parameters directly from the image (rather than the iterative approach used in AAMs). Alternative approaches are the use of sequences of Random Fern predictors to estimate the pose of an object or part [28], and the use of cascaded boosted fern regressors to predict the position of facial feature points without the need for an explicit shape model [29].

Xiong and De la Torre [30] have shown that effective facial feature point detection can be achieved by applying a sequence of updates to the point positions, with each update being calculated using linear regression on SIFT features around the model points.

Regression-based voting. Since the introduction of the Generalised Hough Transform [31] voting-based methods have been shown to be effective for locating shapes in images, and there have been many variants. For instance, the Implicit Shape Model [32] uses local patches located on an object to vote for the object position, and Poselets [33] match patches to detect human body parts.

Hough Forests [6] use RF regression from multiple subregions to vote for the position of an object. This includes an innovative training approach, in which regression and classification training are interleaved to deal with arbitrary backgrounds and where only votes believed to be coming from regions inside the object are counted. In contrast, our approach to feature point detection does not require a class label and allows all image structures to vote.

In the context of medical imaging, Shotton et al. [17] have shown that RFs can be used to vote for the position of joint centres when matching a human body model to a depth image, and Criminisi et al. [16] used RF regression to vote for the positions of the sides of bounding boxes around organs in CT images.

In [34], the annotation of facial features using Conditional RFs (e.g. head pose specific) was proposed. Our method differs from this in that we use an explicit shape model to find the best combination of points.

Martinez et al. [35] showed that facial feature points can be accurately located when combining a regression-based approach with a probabilistic graphical face shape model. They used Support Vector Regression to vote for each point position and constrained the search space using pair-wise point constraints. In [36], this facial point detection algorithm was used in a guided unsupervised learning framework where the training is partitioned according to a particular combination of head pose and facial expression. Sheerman-Chase et al. [37] have shown that hierarchical boosted regression can be used to track facial feature points across significant changes in pose without the need for explicit pose estimation.

3 Methods

We propose the application of RFRV in the CLM framework to vote for the optimal position of each feature point.

3.1 Random Forest Regression-Voting

In the regression-voting approach, we train a regressor from a set of images, each of which is annotated with the feature points of interest on the object, $\mathbf{x}$ (see e.g. Fig. 1). For every point in $\mathbf{x}$, a set of features $\mathbf{f}_j(\mathbf{x}+\mathbf{d}_j)$ is sampled at a set of random displacements $\mathbf{d}_j$ from the true position and a regressor $\delta = R(\mathbf{f}_j(\mathbf{x}+\mathbf{d}_j))$ is trained to predict the most likely position of the feature point relative to $(\mathbf{x}+\mathbf{d}_j)$. Based on an initial estimate of the position of a feature point l, we obtain predictions via evaluating a set of points in a grid over the region of interest. At each point $\mathbf{z}_l$ in the grid, the relevant feature values are extracted and used for regressor R to make predictions for the most likely position of feature point l. All predictions then vote for the best position in a 2D accumulator array. This gives a 2D

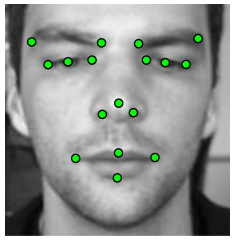


Fig. 1. Example 17-point model capturing features of the face.

histogram of votes V_l with $V_l(\mathbf{z}_l + \delta) \rightarrow V_l(\mathbf{z}_l + \delta) + v$ where v expresses the degree of confidence in the prediction (see Fig. 2). After scanning the whole region of interest, the histogram of votes can be smoothed to allow for uncertainty in the predictions.

One advantage of the regression approach is that it avoids the somewhat arbitrary cut-off radius sometimes required when selecting positive and negative examples for the training of classifiers. It also allows the integration of evidence from regions which may not even overlap the target point. Another advantage of using regression-voting, rather than classification, is that good results can be obtained by evaluating the region of interest on a sparse grid rather than at every pixel. Sampling every third pixel, for instance, significantly speeds up the process with minimal loss of accuracy (see Section 4.1.4).

A natural extension of regression-voting is to use multiple (independent) regressors, or to have each regressor predict multiple positions. Both ideas are combined in Hough Forests [6] which use sets of random trees whose leaves store multiple training samples. Thus each sample produces multiple weighted votes, allowing for arbitrary distributions to be encoded. In related work, [17] and [16] produce votes in higher dimensional spaces (3D or 6D), but work directly with the vector votes rather than accumulator arrays.

RFs consist of a set of multiple independent binary decision trees, each trained independently on a random subset of the data. Although any one tree may be somewhat over-trained, the randomness in the training process encourages the trees to give independent estimates, which can be combined to achieve accurate and robust results. In the following, we use RFRV which combines independent predictions from each tree in the RF and accumulates all predictions in a 2D histogram of votes (see Fig. 3). A key advantage of decision trees is that each leaf can store arbitrary information derived from the training samples which arrived at that leaf. For instance, this could be the mean $\bar{\mathbf{d}}$ and covariance $\mathbf{S}_d$ of these samples, or the full set of samples.

Fig. 2. Superposition of histograms of votes for a 17-point face model.

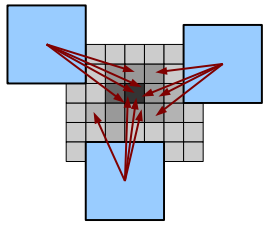


Fig. 3. During search each rectangular patch predicts (one or more) positions for the target point. Votes are accumulated over a grid.

When scanning the region of interest, a range of voting styles can be used:

1) A single, unit vote per tree at the mean offset.
2) A single, weighted vote per tree, using a weight of $|\mathbf{S}_d|^{-0.5}$. This penalises uncertain predictions.
3) A Gaussian spread of votes, $N(\bar{\mathbf{d}},\mathbf{S}_d)$.
4) Multiple votes from the training samples [6].

Following [9], we arrange for the leave nodes to only store the mean offset and covariance rather than the training samples. We explore the impact of the chosen voting style on performance in Section 4.2.3.

3.2 Constrained Local Models

CLMs provide a method for matching the points of a statistical shape model to an image. They combine global shape constraints with local models of the pattern of intensities. Here we summarise the key points of the approach; for details see [4], [5].

Based on a number of points for a set of images, a statistical shape model is trained by applying principal component analysis (PCA) to the aligned shapes [1]. This yields a linear model of shape variation which represents the position of each point l by

$$\mathbf{x}_l = T_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_l + \mathbf{P}_l\mathbf{b} + \mathbf{r}_l), \tag{1}$$

where $\bar{\mathbf{x}}_l$ is the mean position of the point in a suitable reference frame, $\mathbf{P}_l$ is a set of modes of variation, $\mathbf{b}$ are the shape model parameters, $\mathbf{r}_l$ allows for small deviations from the model, and $T_{\boldsymbol{\theta}}$ applies a global transformation (e.g. similarity) with parameters $\boldsymbol{\theta}$.

To match the model to a new image, $\mathbf{I}$, we seek the points, $\mathbf{x} = \{\mathbf{x}_l\}$, that optimise the overall quality-of-fit, Q, of the model to the image. More formally, we seek parameters $\mathbf{p} = \{\mathbf{b}, \boldsymbol{\theta}, \mathbf{r}_l\}$ that optimise

$$\begin{aligned} &Q(\mathbf{p}) = \Sigma_{l=1}^{n} C_l(T_{\boldsymbol{\theta}}(\bar{\mathbf{x}}_l + \mathbf{P}_l\mathbf{b} + \mathbf{r}_l)) \\ &\text{s.t.} \quad \mathbf{b}^T\mathbf{S}_b^{-1}\mathbf{b} \leq M_t \text{ and } |\mathbf{r}_l| < r_t, \end{aligned} \tag{2}$$

where C_l is a cost image for the fitting of feature point l, $\mathbf{S}_b$ is the covariance matrix of shape model parameters $\mathbf{b}$, M_t is a threshold on the Mahalanobis distance, and r_l is a threshold on the residuals. M_t is chosen using the cumulative distribution function (CDF) of the χ^2 distribution so that 98 percent of samples from a multivariate Gaussian of the appropriate dimension would fall within it. This ensures a plausible shape by assuming a flat distribution for model parameters $\mathbf{b}$ constrained within hyper-ellipsoidal bounds [1]. Examples of cost images include using normalised correlation

with a globally constrained patch model [4] and sliding window search with a range of classifiers [5], [12]. Below, we show that RFRV can be used to produce effective cost images, leading to fast, robust and accurate performance.

3.3 RFRV in the CLM Framework

We apply RFRV in the CLM framework to vote for the best position of every feature point. Based on the 2D histograms of votes V_l from an RF regressor (see Section 3.1), one for every point, we aim to combine the votes of all histograms given the shape constraints of the CLM via maximising Eq. (2) with $C_l = V_l$. In the following, we describe the training and shape model matching of an RFRV-CLM in more detail.

3.3.1 Training

We train the CLM and the RF regressors from a set of images, each of which is annotated with the feature points of interest on the object, $\mathbf{x}$. A statistical shape model is trained by applying PCA to the aligned shapes [1], see also Eq. (1). The shape model is used to assess the global pose, $\boldsymbol{\theta}$, of the object in each image by minimising $|T_{\boldsymbol{\theta}}(\bar{\mathbf{x}}) - \mathbf{x}|^2$. Each image is re-sampled into a standardised reference frame by applying the inverse of the pose, $\mathbf{I}_r(x, y) = \mathbf{I}(T_{\boldsymbol{\theta}}^{-1}(x, y))$. The model is scaled so that the width of the reference frame of the mean shape is a given value, w_{frame}.

To train the feature detector for a single feature point, we generate sample patches and extract features $\mathbf{f}_j$ at a set of random displacements $\mathbf{d}_j$ from the true position in the reference frame, $T_{\boldsymbol{\theta}}^{-1}(\mathbf{x}_l)$, where $\mathbf{x}_l$ is the position of the point in the image. Displacements $\mathbf{d}_j$ are drawn from a flat distribution in the range $[-d_{max}, +d_{max}]$ in x and y. To allow for inaccurate initial estimates of the pose and to make the detector locally pose-invariant, we repeat this process with random perturbations in scale and orientation of the estimate of the pose. We then train an RF [7] on N_s pairs $\{(\mathbf{f}_j, \mathbf{d}_j)\}$. To train a single tree, we take a bootstrap sample (drawing N_s examples with replacement) of the training set and use a standard greedy approach to construct the tree, recursively splitting the data at each node. Given the samples at a particular node, we seek to select a feature and a threshold to best split the data into two compact groups. Let f_{ji} be the value of one feature associated with sample j. The best threshold, t, for this feature at this node is the one that minimises

$$G_T(t) = G(\{\mathbf{d}_j : f_{ji} < t\}) + G(\{\mathbf{d}_j : f_{ji} \geq t\}), \tag{3}$$

where $G(S)$ is a function evaluating the set of vectors S. In the following, we aim to increase the compactness of the groups when splitting the nodes by minimising the following entropy measure

$$G(\{\mathbf{d}_j\}) = N \, log|\boldsymbol{\Sigma}|, \tag{4}$$

where N is the number of samples in $\{\mathbf{d}_j\}$ and $\boldsymbol{\Sigma}$ is the covariance matrix of the N samples.

In the experiments below, we use Haar-like features [38], randomly sampled in a patch around the current point, as they have been found to be effective for a range of applications and can be calculated efficiently from integral images. To select the feature and threshold that best splits the data at each node, we choose a random set of features (of size n_{feat}) from all possible Haar features and choose the feature and associated optimal threshold that minimise G_T. We stop splitting the nodes when the tree has either reached a maximal depth, D_{max}, or a minimum number of samples per node, N_{min}.

3.3.2 Shape Model Matching

Given an initial estimate of the pose of the model (either from an object detector or from the earlier application of a model), we seek to find parameters $\mathbf{p} = \{\mathbf{b}, \boldsymbol{\theta}, \mathbf{r}_l\}$ that optimise $Q(\mathbf{p})$ (see Eq. (2)). We apply the following technique to solve this optimisation problem. Given initial estimates of parameters $\mathbf{b}$ and $\boldsymbol{\theta}$, we first transform the image into the reference frame by re-sampling using the current pose: $\mathbf{I}_r(x, y) = \mathbf{I}(T_{\boldsymbol{\theta}}^{-1}(x, y))$. In the reference frame, we compute cost images C_l (i.e. 2D histograms of votes V_l from the RF regressor) by searching in $\mathbf{I}_r$ around the current estimate of each point (with displacements in the range $[-d_{search}, +d_{search}]$). This is done for all feature points independently. We then iteratively estimate the shape model and pose parameters using the simple but robust model matching approach outlined in Algorithm 1. In the experiments below, we set the initial disk radius r_{max} to match the search range d_{search}, r_t to 1.5 pixels (in the reference image) and $k_r = 0.7$.

Algorithm 1. Shape model and pose parameter optimisation: Iterative model matching procedure to estimate the shape and pose parameters in the reference frame, given a set of point based cost images C_l.

Input: r_{max}, r_t, k_r, $\mathbf{x}_l$ and $C_l \; \forall 1 \leq l \leq n$

1) Set $r \rightarrow r_{max}$, $\boldsymbol{\theta}_r \rightarrow \text{Identity}$, $\mathbf{x}_l \rightarrow \bar{\mathbf{x}}_l + \mathbf{P}_l\mathbf{b}$, $\mathbf{r}_l = 0$
2) While $r \geq r_t$
   a) For every feature point l, find the best point $\hat{\mathbf{y}}_l$ in a disk of radius r around the current estimate
      $\hat{\mathbf{y}}_l \rightarrow \arg\max_{\mathbf{y}_l : |\mathbf{y}_l - \mathbf{x}_l| < r} C_l(\mathbf{y}_l)$
   b) Fit the shape model to these best points to estimate shape and pose parameters $\{\mathbf{b}, \boldsymbol{\theta}_r\}$ by solving
      $\hat{\mathbf{y}}_l = T_{\boldsymbol{\theta}_r}(\bar{\mathbf{x}}_l + \mathbf{P}_l\mathbf{b})$
   c) If $\mathbf{b}^T\mathbf{S}_b^{-1}\mathbf{b} > M_t$ then move $\mathbf{b}$ to nearest valid point on limiting ellipsoid
   d) Update all feature point positions using
      $\mathbf{x}_l \rightarrow T_{\boldsymbol{\theta}_r}(\bar{\mathbf{x}}_l + \mathbf{P}_l\mathbf{b} + \mathbf{r}_l)$
   e) Set $r \rightarrow k_r r$ with $0 < k_r < 1$
3) Transform the resulting feature point positions into the image frame using $T_{\boldsymbol{\theta}}$ with $\boldsymbol{\theta} \rightarrow \boldsymbol{\theta} \circ \boldsymbol{\theta}_r$

If the initial estimate of the pose of the model is displaced by a large amount from its true position, we re-sample the image at the new pose, re-compute the cost images, repeat the parameter optimisation described in Algorithm 1 and update the point positions accordingly—we refer to this as *one search iteration*.

4 Experimental Evaluation

We performed a series of experiments to *(i)* analyse the performance of RFRV across application areas, *(ii)* identify which properties of the regressor contribute to its performance, and *(iii)* investigate the impact of the choice of parameters. We present the results of RFRV to annotate the joints of the hands in Section 4.1 and to annotate faces in Section 4.2.

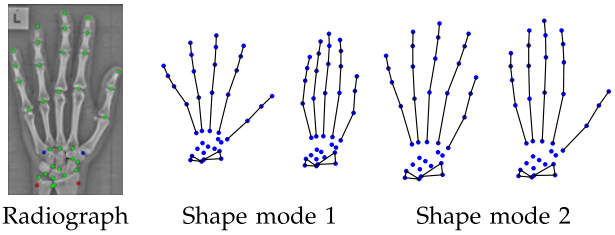


Fig. 4. Annotation example for the joints of the hand using 37 points, and the first two modes of the shape model. In the radiograph, blue points are the reference points returned by the object detector, and red points define the reference length used to give the mean point-to-point error as a percentage of the wrist width.

4.1 Evaluation of RFRV Using Hand Radiographs

Following [39], we applied RFRV as part of a *fully automatic* shape model matching system: We used our own implementation of Hough Forests [6] to estimate the position, orientation and scale of the object in the image, and used this to initialise the shape model matching (mean shape). We investigated the accuracy of RFRV and which properties of the regressor are important for achieving the best performance. In particular, we analysed the importance of:

1) The use of a coarse-to-fine strategy.
2) The RFs combining multiple independent estimates, one from each tree.
3) The integration of votes from multiple regions around the feature point.

We show that all three properties contribute both individually and collectively to the performance of RFRV, achieving what we believe to be the best yet published results on hand joint annotation.

We ran a series of systematic experiments on a set of 564 AP hand radiographs of children aged between five and eighteen years. Pixel size information was not available for any of the radiographs. All radiographs were manually annotated with 37 points, which gave the ground truth for all evaluations. Fig. 4 shows an annotation example and two of the shape modes of the resulting model. This indicates that there is significant shape variation, not only due to positioning during image acquisition but also because in the given age range the bones of the hand are still developing.

We initialised the model by automatically detecting nine points (four around the palm and one at the base of each finger). This was achieved by first detecting the object in the image and initialising two reference points within the detected bounding box (blue points in Fig. 4) as in [39]. We used these two points to initialise the mean shape of a 9-point RFRV-CLM and ran a single iteration to locate the nine points. The 37-point RFRV-CLM was then initialised using these nine points. Note that the positions of the points used for initialisation were refined during model matching.

The dataset was split randomly into a training and testing subset of equal size. Results are presented as cumulative density functions (CDFs) of the mean point-to-point error as a percentage of the wrist width (red points in Fig. 4), to provide invariance to image scaling. The wrist width can be used as a reference length to evaluate segmentation performance in mm as it tends to be relatively constant across individuals. We assumed an average wrist width of 50 mm.

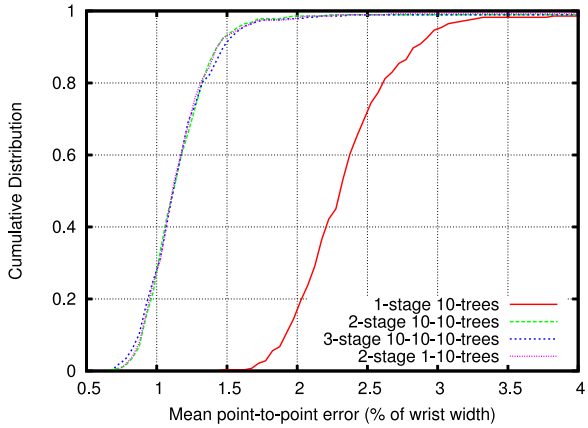


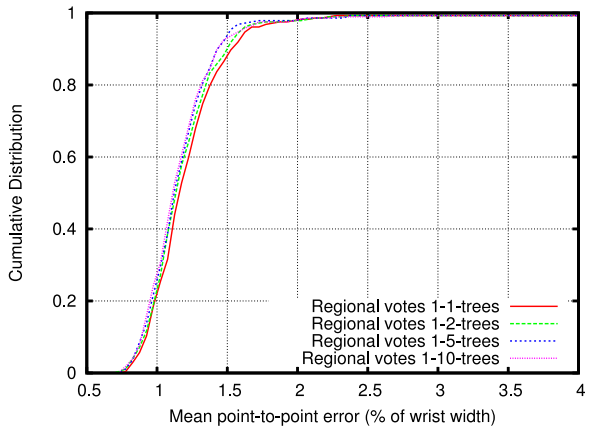
Fig. 5. Evaluation of the effect of introducing a second search stage to locally refine all point placements. Notation: x-y-trees refers to using x trees in the RFs of the first-stage model and y trees in the RFs of the second-stage model.

4.1.1 Effect of Using a Coarse-to-Fine Strategy

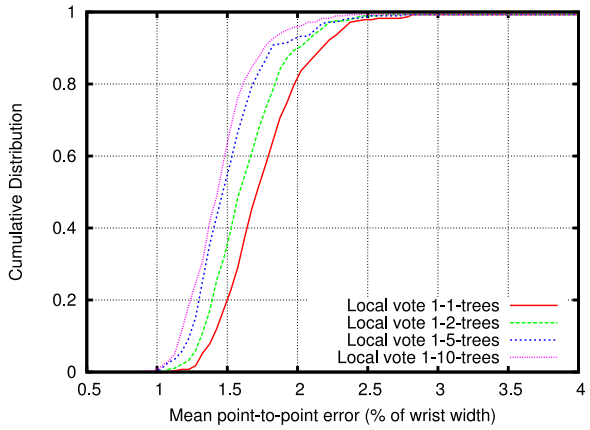
Significant improvements can be achieved when using the RFRV scheme in a coarse-to-fine, multi-stage approach. Here, the underlying hypothesis was that the first stage would perform a coarse search to locate the approximate position of the global optimum in a cost function exhibiting multiple local optima, and that the second stage would then refine this solution through a search across a smaller range within the global optimum. Note that this is somewhat similar to the framework of cascaded regression [28], [29], but in our approach the regressors are trained independently.

We empirically evaluated this hypothesis and found the following parameters to provide the best results. The optimised single-stage model used a reference frame of size $w_{frame} = 70$, and Haar features were sampled from patches of size 17×17. The regression functions for all points were trained using 10 random displacements in x and y with $d_{max} = 11$, as well as random displacements in scale (in the range of $\pm 5\%$) and rotation (in the range of $\pm 3°$) on each image. This rather coarse, low resolution model was then used as the first-stage model of a coarse-to-fine, multi-stage RFRV-CLM. The second-stage model used a fine, high resolution reference frame of size $w_{frame} = 210$, patches of size 17×17 and displacement range $d_{max} = 7$. The number of random displacements as well as scale and angle perturbation ranges were the same as for the coarse model. If not stated otherwise each RF had 10 trees, trained using about 300 random features per node. The search range, d_{search}, for both models was set to 15. We ran four search iterations (i.e. generating vote histograms and matching the shape model) with the coarse model, updating shape and pose parameters $\{\mathbf{b}, \boldsymbol{\theta}\}$ (see Eq. (2)). For the coarse-to-fine RFRV-CLM, we then refined the point positions using a single search iteration with the fine, second-stage model. Fig. 5 compares the performance of the fully automatic hand joint annotation system when using a single-stage vs a coarse-to-fine, two-stage RFRV strategy.

Fig. 5 shows that the results significantly improved by introducing a fine, second-stage model. Additional stages, however, did not lead to any further improvements for this application area. In Fig. 5, *10-10-trees* refers to how many



(a)



(b)

Fig. 6. Evaluation of the effect of combining multiple independent estimates of position by varying the number of trees in the RFs when every tree: (a) combines votes from multiple regions around the feature point; (b) votes only from the current position. Notation: x-y-trees refers to using x trees in the RFs of the first-stage model and y trees in the RFs of the second-stage model. Note the same scale of the plots.

trees were used in stages 1 and 2 of the RFRV-CLMs. We performed extensive experiments on varying the number of trees in the RFs in all but the last stage (including the 9-point initialisation model described above) while keeping the number of trees in the RFs of the fine, last-stage model fixed to 10 trees. This revealed that additional trees in earlier stages did not contribute to an increased overall performance. As is evident in Fig. 5, a two-stage model using only a single tree for the coarse model performed as well as a two-stage model that used 10 trees in the RFs of the coarse model. Hence, in all the experiments below we only used a single tree for the 9-point initialisation model as well as for the coarse, first-stage 37-point model.

4.1.2 *Effect of Combining Multiple Independent Votes for Each Feature Point*

To investigate the effect of multiple independent votes on the performance of the RF regressor, we varied the number of trees in the RFs of the second-stage 37-point RFRV-CLM. Fig. 6a shows the results when we varied the number of trees and combined multiple votes from different regions around the current estimate of the point (as in the proposed RFRV-CLM approach). This shows that when increasing the number of trees in the RFs of the final-stage model while using multiple regional votes, the main improvement happened when using more than one tree—additional trees only slightly improved the performance. Similar results were obtained when varying the number of trees for the first-stage 37-point model.

TABLE 1
Effect of Regional Sampling on a Coarse Grid Rather Than at Every Pixel

Step size 2nd stage	Search time (ms)	Median error (%WW)	90%ile (%WW)	95%ile (%WW)
1	333	1.13	1.46	1.57
2	143	1.14	1.46	1.59
3	111	1.15	1.45	1.61
4	89	1.15	1.50	1.60
5	82	1.17	1.50	1.65

Fully automatic annotation results using a two-stage 1-10-trees model where the step size of the first-stage is fixed to 3 and the step size of the second stage is varied. (%WW = mean point-to-point error as a percentage of wrist width).

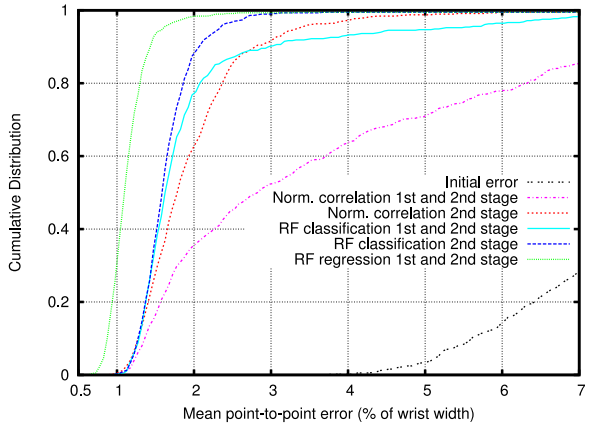
4.1.3 *Effect of Integrating Votes from Multiple Regions around the Feature Point*

In the suggested RFRV-CLM approach, we sparsely sample local patches in the area around the current estimate of the position of a feature point and obtain predictions from each patch (see also Fig. 3). To evaluate the effect on performance of this approach, we compared it to obtaining votes from only a single patch at the current estimated position. As is evident from Fig. 6a versus Fig. 6b, integrating multiple regional votes from an area around the current estimate significantly improved the model matching performance. Fig. 6b shows that increasing the number of trees while using only local votes from the patch at the current estimate of the feature point had a much more significant impact than varying the number of trees while using multiple regional votes as in Fig. 6a. However, Fig. 6 also demonstrates that using multiple regional votes around the feature point had a greater effect than increasing the number of trees. The best performance was achieved by combining multiple regional votes with votes from multiple independent trees in the RFs.

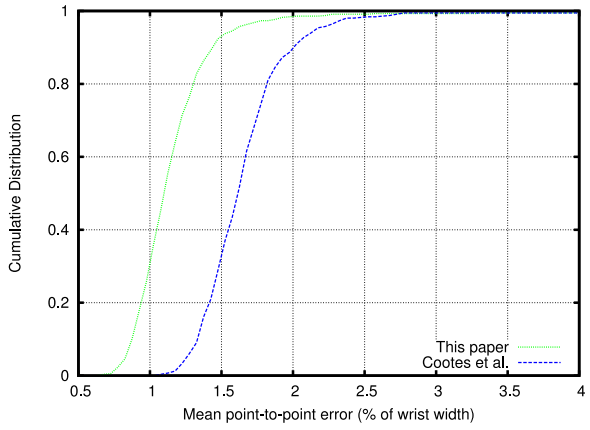
4.1.4 *Effect of Step Size*

When integrating votes from multiple regions around the feature point as described above, it is sufficient to sample patches on a sparse grid rather than at every pixel. Since the trees cast votes over a region, we can achieve high accuracy without having to sample everywhere. Table 1 shows the annotation performance as a function of step size (a step size of s means to only take one sample in each $s \times s$ square).

This shows that significant subsampling can be used without compromising accuracy. Assuming an average wrist width of 50 mm, a difference in error values of 0.1 percent related to 0.05 mm. All RFRV-CLM experiments described in this paper used a step size of 3 for all stages which significantly reduced the search time. It is worth pointing out that the results in Table 1 were obtained when applying a fully automatic system (including object



(a)



(b)

Fig. 7. Comparison of fully automatic hand annotation results with (a) different shape model matching techniques, and (b) previously published results from Cootes et al. [9]. All results are based on cross-validation experiments. Note the different scales of the two plots.

detection and 37-point initialisation) while only varying the step size of the second stage of the two-stage RFRV-CLM. This both explains the asymptotic behaviour of the reduction in total search time when increasing the step size, and indicates that the overall reduction in running time is even higher when increasing the step size for all stages of the fully automatic system.

4.1.5 Comparison with Other Shape Model Matching Techniques and Previously Published Results

In Fig. 7, we compare our results to other shape model matching techniques and previously obtained results. All search results are based on the application of a fully automatic annotation system as described above where each method started off from the initialisation of the 37-point model as a result of applying the 9-point RFRV-CLM (*Initial error* in Fig. 7a). We ran $4 + 1$ search iterations using a coarse-to-fine, two-stage model and evaluated the performance when using different methods of generating the cost images: a correlation-based CLM using normalised correlation and an RF-classification based CLM using an RF-based classifier with 10 trees each, trained on Haar-like features (as above). We show the results for using each of these methods for both stages of the 37-point model, as well as for only using these methods for the second-stage 37-point model while using the proposed RFRV-CLM for the first-stage. We performed two-fold cross-validation experiments training on one half of the data and testing on the other and then doing the same with the sets switched (using all 564 images). The error differences between both cross-validation experiments were less than 0.15 percent (0.07 mm) for the 50/90/99%iles. Fig. 7a shows that RF classification significantly outperformed normalised correlation, but that RFRV was by far the best overall.

Fig. 7b compares our two-stage 1-10-trees results to the results published in Cootes et al. [9] using the same experimental protocol as above. This shows that we significantly improved upon previous results achieving a mean point-to-point error of within 1.1 mm for 99 percent of all 564 images. Here, the main improvement was achieved by following [39] in relaxing the shape constraints in the final search step (i.e. the last iteration of Step 2 in Algorithm 1) of the last search iteration of the model matching approach. So in the final Step 2D, we set $\mathbf{r}_l = T^{-1}_{\hat{\boldsymbol{\theta}}_r}(\hat{\mathbf{y}}_l) - (\bar{\mathbf{x}}_l + \mathbf{P}_l\mathbf{b})$ and $\mathbf{x}_l = \hat{\mathbf{y}}_l$. In a locally restricted area, the latter gives priority to the RF votes over the shape constraints of the RFRV-CLM. We found this to significantly improve performance across bone annotation tasks (e.g. hands, hips, knees).

A direct comparison to other reported hand joint annotation results is difficult because of the different datasets and annotations used. However, along with the results in [9] the best reported results appear to be those by Donner et al. [40] who report a mean/median error of 0.77 mm/0.63 mm (tested on 20 images), compared to our 0.61 mm/0.56 mm (tested on 564 images).

All experiments on hand radiographs were run on a 3.33-GHz Intel Core 2 Duo PC in VMware using 2 GB RAM. No parallel computing was implemented.

4.2 Evaluation of RFRV Using Images of the Face

To compare the performance of the approach with alternatives, and to evaluate the effects of parameter choices, we trained several 17-point RFRV-CLMs. The markup scheme corresponded to the 20-point scheme used in the BioID dataset, after removal of the temple and chin-tip points. Training was performed with the AFLW dataset [41], which consists of real-world face images. We used the same subset of images as [9], consisting of 5081 images for which all frontal points were visible. The additional points in our 17-point scheme were manually annotated. All models were trained on 326 of these images. Testing was performed on the 1521-image BioID dataset, which consists of frontal images of 23 individuals in office environments. BioID has been used to test a range of algorithms, facilitating comparison to previous work. Further testing was performed using the remaining 4755 AFLW images, which exhibit a much greater variation in pose, illumination and background.

We present results from a fully automatic shape model matching system where a sequence of RFRV-CLMs was initialised using our own implementation of Hough Forests [6]. This detected the face in the image and estimated two reference points that corresponded to the pupil centres. In cases where the first-stage model consisted of more than two points, the remainder were initialised by fitting the mean model to the two reference points. In all cases, the

positions of the points used for initialisation were refined during model matching.

To investigate the dependence of the parameters on the relative strength of the data and shape model constraints, experiments were performed for models with 2 (pupils), 4 (pupils and mouth corners) and 17 feature points (see Fig. 1).

The performance was evaluated using the mean Euclidean distance between the automatically annotated points and manual annotations across each test image. Following common practice [4], the error was recorded as the mean point-to-point error as a percentage of the inter-ocular distance (IOD)

$$m_n = \frac{1}{n d_{eyes}} \sum_{i=1}^{n} |\mathbf{x}_i - \hat{\mathbf{x}}_i|, \tag{5}$$

where n was the number of points annotated by the model, $\mathbf{x}_i$ and $\hat{\mathbf{x}}_i$ were the model-annotated and manually annotated positions of the ith point, and d_{eyes} was the distance between the pupil centres i.e. $|\hat{\mathbf{x}}_{lefteye} - \hat{\mathbf{x}}_{righteye}|$. To account for the stochastic element of RF training and to support error estimation, we trained and tested five models for each experiment and report the mean CDF values across all repeats.

4.2.1 Parameter Optimisation

We performed a set of sequential experiments to optimise the parameters of multi-stage RFRV-CLMs. First, experiments were performed to optimise the parameters of the first-stage model (initialised with a Hough Forest face detector as described above). The optimised first-stage model was then included in a second set of experiments to optimise the parameters of the second-stage model. We ran a single search iteration for both the first- and second-stage models.

Due to the large number of parameters, we report complete results only for those showing the most influence and variation of performance. We divided the parameters of the algorithm, described in Section 3.3.1, into two sets. The first consisted of those parameters relating to the structure of the RFs: the number of trees in each forest, n_{trees}; the number of random features considered at each node, n_{feat}; the minimum number of training examples at each node, N_{min}; and the maximum depth of each tree, D_{max}. The second set consisted of those parameters relating to the extraction of data from the images: the width of the reference frame, w_{frame}; the size (width = height) of the sampled patches from which Haar-like features were generated, w_{patch}; the range of the uniform distribution used to generate random displacements for sampling Haar-like features, d_{max}; and the search range around the estimated position of each point, d_{search}. Parameters w_{patch}, d_{max} and d_{search} are defined in the reference image. The regression functions for all feature points were trained using 20 random displacements within the limits of $\pm d_{max}$, as well as random displacements in scale (in the range of $\pm 22\%$) and rotation (in the range of $\pm 13°$).

Constant performance improvements, subject to a law of diminishing returns, were found with the RF-based parameters in the expected direction i.e. increasing n_{trees}, n_{feat} and D_{max}, and reducing N_{min}. However, these parameters also had a significant effect on processing time. Therefore, when optimising the image-based parameters the RF-based parameters were set to sub-optimal values of using 10 trees in the RFs and considering 50-100 random features at each node. Table 2 shows the set of optimised parameters.

TABLE 2
Optimal Values for All Multi-Stage RFRV-CLM Parameters (c = coarse, first-stage model; f = fine, second-stage model)

Parameter	2-point		4-point		17-point	
	c	f	c	f	c	f
n_{trees}	15	15	15	15	15	15
n_{feat} ($\approx$)	1,000	2,000	1,000	1,000	1,000	1,000
N_{min}	1	1	1	1	1	1
D_{max}	25	25	25	25	25	25
w_{frame}	30	120	30	80	60	140
w_{patch}	18	18	18	18	18	18
d_{max}	15	12	15	15	15	15
d_{search}	10	10	10	10	10	10

Detailed results of the experiments on w_{frame} and w_{patch} are shown in Fig. 8. The CDFs for the experiments performed on each model-stage/point-number were clearly separable i.e. for each combination of parameters (a, b), if the CDF for parameter set a was greater than that for b at any point, it was greater than or equal to that for b at all points. Thus, to aid visualisation of the results across multiple parameters, they are given as the area under the CDF of m_n with the CDF truncated at a proportional error of $m_n = 0.25$; in all cases this corresponded to ≥ 0.99 on the CDF.

These results show that the reference frame width had a dominant effect on performance. In first-stage models, performance generally increased with increasing the frame width over the ranges tested. However, for a given patch width and search range, increasing the frame width decreased the capture range of the model in that a single patch will cover a smaller range and the relative search range in the image will be smaller since both of these parameters are defined in the reference frame. Given that the first-stage models were intended for combination into multi-stage models, it was desirable to use a low frame width. Therefore, optimal first-stage frame widths were selected as the point at which the improvement in performance ceased to be statistically significant, resulting in values of 30 for 2- and 4-point models and 60 for 17-point models. For the frame width of the second-stage models, optima emerged at 120 for the 2-point and at 80 for the 4-point models. The optimal frame width for the second-stage 17-point models was selected as the point at which performance improvements ceased to be statistically significant, giving a value of 140.

The variation in optimal frame widths across point numbers and stages indicates the varying influence of the shape model constraints and the varying information content of the feature points. The 2-point model had no effective shape model constraint, leading to a small frame width in the first-stage model. However, the 2-point models annotated the centres of the pupils, which were points of high information content, leading to a strong global optimum in the cost function and supporting a smaller search range in the second stage. Both the 4- and 17-point models included sufficient points to produce a shape model constraint, and in both cases the optimal ratio of first- to second-stage frame width was approximately 2.5.

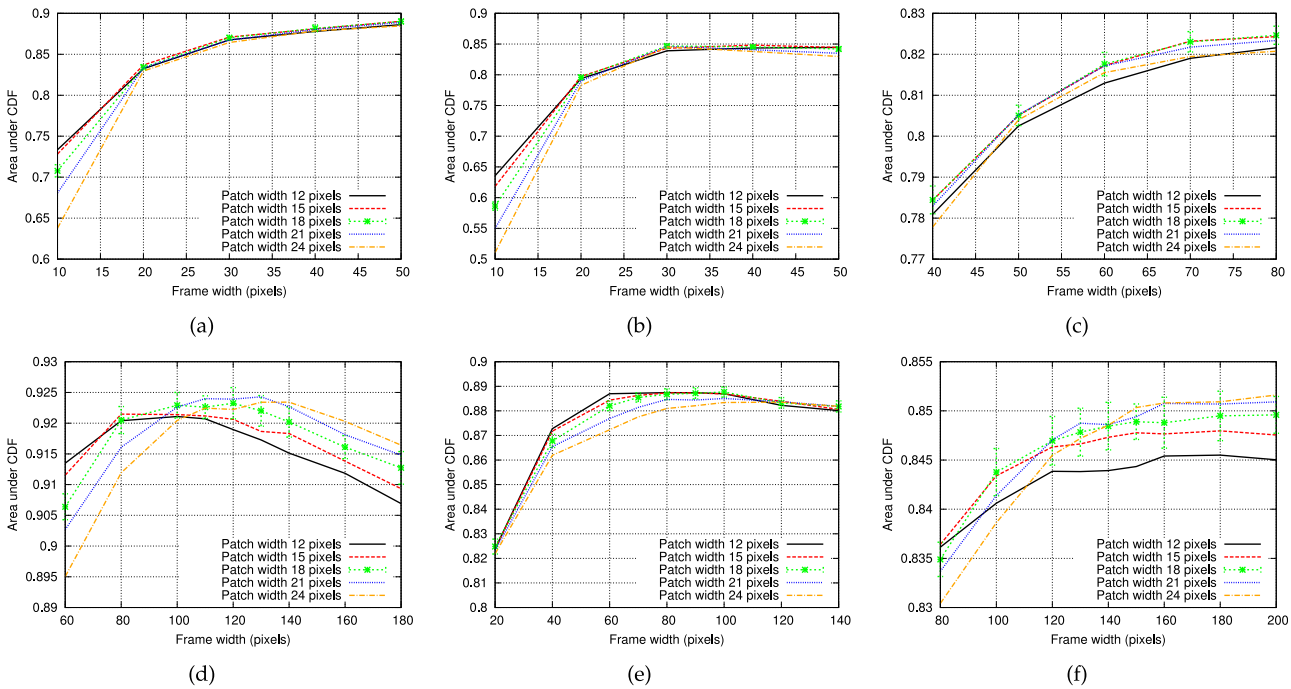


Fig. 8. Optimisation of the patch and frame width for (a,d) 2-, (b,e) 4- and (c,f) 17-point RFRV-CLMs with (a,b,c) one and (d,e,f) two stages. Performance was measured as the area under the CDF of m_n (see main text). Error bars are given as the standard deviation across five repeats of training and testing for each experiment. As all error bars across the curves of an experiment were roughly of the same size, we only show them for one curve per graph to aid visualisation.

Varying the patch width had a smaller effect and a width of 18 emerged as an approximate optimum across all numbers of points and stages—this value was either optimal or yielded a performance that was not statistically significantly worse than the optimum.

An identical set of experiments was then conducted using the 4755-image AFLW testing dataset described above. The same set of optimised parameters were obtained, and so detailed results are not shown here. However, the experiments indicated that the optimised parameters are generally applicable to facial images, and did not result in over-fitting to either the BioID or AFLW testing datasets. These results also show that, as for the hands (see Section 4.1.1), performance can be significantly improved when following a coarse-to-fine strategy (i.e. adding a second-stage model with a significantly higher frame width). Further experiments were performed to investigate the use of three-stage models. However, these generated no improvements in performance compared to two-stage models and so the results are not reported here.

4.2.2 *Effect of Chaining Multi-Stage Models*

So far we have considered applying RFRV-CLMs in a coarse-to-fine, multi-stage approach for a given set of feature points. Here, we analysed whether performance could be improved when chaining models with varying point numbers together. We trained 2-, 4-, and 17-point first- and second-stage models (as described above) using the optimal parameter values listed in Table 2 for both the RF-based and the image-based parameters. All first- and two-stage models were tested individually for all point numbers, as well as chains consisting of models with increasing point and stage numbers. We applied a single search iteration of every model. Selected results are shown in Fig. 9 as the CDF of m_n, where n was the number of points annotated by the final-stage model.

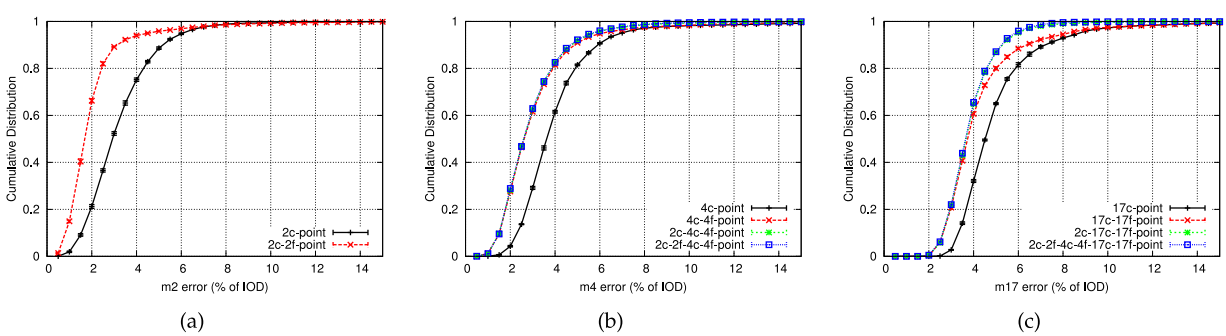


Fig. 9. Performance evaluation of combinations of optimal (a) 2-, (b) 4- and (c) 17-point models. Notation: xc-point refers to a coarse, first-stage model using x points, and xf-point refers to a fine, second-stage model using x points. For example, the 2c-17c-17f-point curve in (c) represents a model comprising a coarse, first-stage 2-point model followed by a coarse-to-fine, two-stage 17-point model. Error bars are given as the standard deviation across five repeats of training and testing for each experiment.

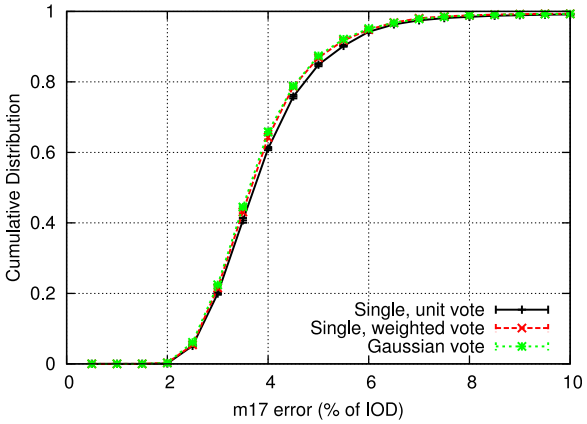


Fig. 10. Evaluation of voting styles for a chained 2c-17c-17f-point RFRV-CLM. Error bars are given as the standard deviation across five repeats of training and testing for each experiment.

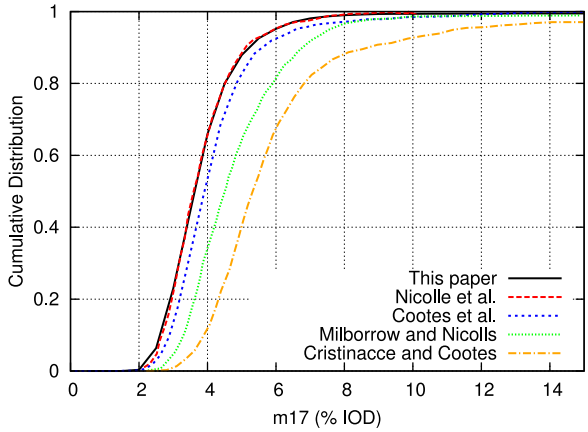


Fig. 11. Comparison of the performance of a chained 2c-17c-17f-point RFRV-CLM, evaluated on the BioID dataset, with previously published results from [4], [9], [15] and [42].

As before, addition of a fine, second-stage model for any given number of points resulted in a significant improvement in performance. Using the 90%ile point of the CDF as an example, the two-stage 2-point model had a 38 percent lower m_2 than the first-stage 2-point model. This performance gain decreased with increasing the number of points in the model. Adding a second stage resulted in m_n reductions of 17 percent for 4-point models and of 11 percent for 17-point models. For 4- and 17-point two-stage models further improvements in performance were achieved by including the first-stage 2-point model in the chain. For example, m_n at the 90%ile of the CDF was reduced by 3 and 17 percent for 4- and 17-point models, respectively. This indicates that the combination of a Hough Forest face detector (as above) with a coarse 2-point RFRV-CLM model trained on the pupil centres provides a significantly more accurate initialisation for the later stages than the Hough Forest face detector alone. However, inclusion of additional intermediate model stages resulted in no further significant improvement in performance.

In the following, xc-point refers to a coarse, first-stage model using x points, and xf-point refers to a fine, second-stage model using x points. We will use the notation *2c-17c-17f-point RFRV-CLM* to refer to the identified optimal multi-stage model comprising a coarse, first-stage 2-point model followed by a coarse-to-fine, two-stage 17-point model (as in Fig. 9c).

4.2.3 Effect of Voting Style

A series of experiments were conducted to evaluate the relative performance of the various voting styles (see Section 3.1). We applied a 2c-17c-17f-point RFRV-CLM as described above, i.e. with the optimal parameter values listed in Table 2, but using 10 trees in the RFs and considering 50-100 random features at each node. The results are shown in Fig. 10.

Probabilistic voting resulted in a small but significant improvement in performance. Using the 90%ile point of the CDF as an example, the weighted voting style reduced the m_{17} error by 3.3 percent, and the Gaussian spread voting style by 3.5 percent, of the value for single, unit voting. The difference between the two probabilistic voting strategies was not statistically significant. However, the Gaussian spread voting style significantly increased the mean time taken to fit the model per image to 530 ms, compared to 54 ms for both single, unit voting and single, weighted voting. The results therefore indicate that single, weighted voting is the optimal voting style.

For medical images, we found single, unit voting and single, weighted voting to perform equally well.

4.2.4 Comparison with Previously Published Results

The results from the optimal multi-stage model described in Sections 4.2.2 and 4.2.3 (i.e. a chained 2c-17c-17f-point RFRV-CLM with parameter values from Table 2 and using single, weighted voting) were compared with previously published results on the 17-point annotation of the BioID dataset using a range of algorithms [4], [9], [15], [42]. The results are shown in Fig. 11. Example images and fully automatically obtained 17-point annotations are shown in Fig. 12.

As the results show, the multi-stage RFRV-CLM outperformed all previously published results on the BioID dataset, with the exception of the results from [15], which gave similar performance. However, we note that the results from that paper were based on only a subset of 1083 images from the BioID database, filtered to remove those images for which the Viola-Jones face detector "did not give a relevant result". The images that were removed will necessarily tend to include the more difficult cases, for example, those where some features are occluded by facial hair as in Fig. 12c. In contrast, the results from the work presented here were obtained using all 1521 BioID images, including those for

(a) (b) (c)

Fig. 12. Example images from the BioID dataset showing the automatic annotations generated by a chained 2c-17c-17f-point RFRV-CLM, initialised using a Hough Forest to locate the pupil centres.

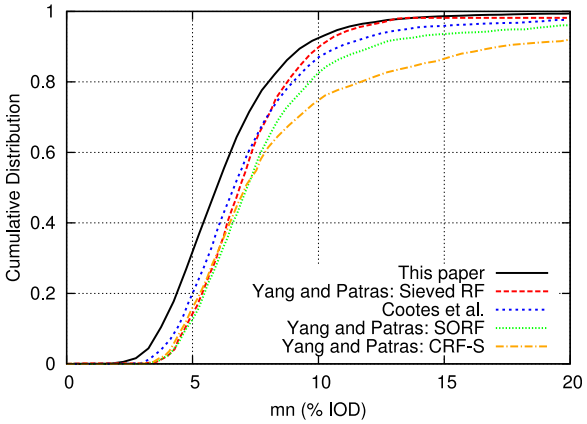


Fig. 13. Comparison of the performance of a chained 2c-17c-17f-point RFRV-CLM, evaluated on the AFLW dataset, with previously published results from [9], and from [43] for RF with vote sieving, Structured Output RF (SORF), and Conditional RF (CRF-S). Note that [43] used a 19-point markup (see main text).

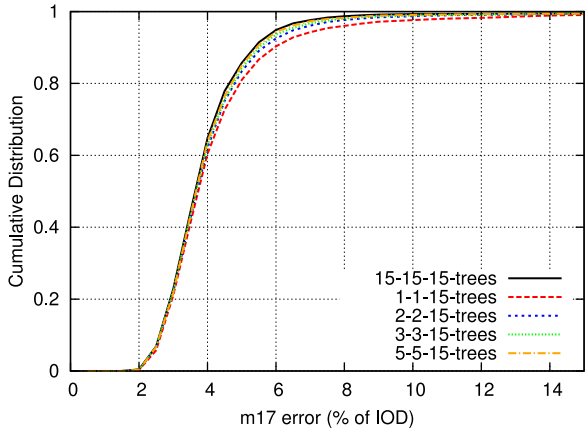


Fig. 14. Effect of varying the number of trees in the RFs for the 2c-point and 17c-point stages of a chained 2c-17c-17f-point RFRV-CLM. Notation: x-y-z-trees refers to using x trees in the RFs of the 2c-point stage, and y and z trees in the RFs of the 17c-point and 17f-point stages.

which the Hough Forest initialisation gave a poor result. This limits the extent to which the results can be compared.

Note that there seems to be some confusion in the literature regarding the quantities plotted on such graphs. In addition to the papers plotted in Fig. 11, results on the BioID dataset were also reported by [8], [12] and [29]. However, even though they state that their plots show CDFs of the mean m_{17} error for each image, scaled to the inter-ocular distance, their curves lack the expected sigmoid curve shape. These results may instead show the CDF of the errors on the individual points rather than of the mean across all 17 points on each image, and are thus not included in this comparison. Care should be taken in comparing such results if the expected sigmoid curve shape is absent. See also [9] and [15] for further discussion.

The optimised 2c-17c-17f RFRV-CLM was also tested on the 4755-image AFLW testing dataset. The results are shown in Fig. 13. For comparison, results from [9] are included which were obtained using identical testing and training datasets as well as the same 17-point markup scheme. Also shown are results from [43], obtained using a range of RF-based algorithms. These were tested on 1000 frontal images from AFLW, using all 19 frontal points. As with the experiments on BioID, the multi-stage RFRV-CLM outperformed the previously published results, although differences in markup scheme and test image subset must be considered in this comparison.

All experiments on facial images were performed on a Dell Precision workstation with 2 Intel Xeon 5670 processors and 24 GB RAM, running OpenSuse 11.3 x64 (Linux kernel 2.6.34). Note that during testing only a single processor core was used. The mean time taken to annotate the 17 points across the 1521 BioID images was 82 ms per image i.e. approximately 12 images/s or half of typical video frame rates.

4.2.5 Effect of Shape Constraints and Number of Trees

Unlike for medical images, relaxing the shape constraints in the final search step had no significant effect on performance for facial feature point detection.

When applying multi-stage RFRV-CLMs to radiographs, it was found in Section 4.1.1 that the number of trees in the first stages of multi-stage models could be reduced significantly compared to the number of trees in the final stage with no significant loss of accuracy. This effect was tested for facial feature point detection using the optimal multi-stage 2c-17c-17f-point RFRV-CLM described in Section 4.2.2, and the results are shown in Fig. 14. The results demonstrate that reducing the number of trees to below five in both the 2c-point and 17c-point stages had a significant effect on accuracy over the region of the CDF between the 85%ile and 100%ile. However, the model with five trees in both the 2c-point and 17c-point stages gave only a small and statistically insignificant reduction in performance compared to the model with 15 trees in all stages. This *reduced model* required significantly less running time, annotating the 17 points in 38 ms per image on the BioID dataset, compared to 82 ms per image for the model with 15 trees in each stage. This corresponds to 26 images per second i.e. is fast enough to perform the 17-point annotation at typical video frame rates. We believe this to be one of the fastest, highly accurate fully automatic facial feature point detection methods yet published.

These results contrast with the results from medical images (radiographs) where *(i)* the elimination of the shape model constraints resulted in a significant increase in performance, and *(ii)* having only a single tree in the first stages of multi-stage models led to no significant loss in accuracy. The results show that to achieve accurate point annotations in facial images, compared to medical images, requires both the shape constraints in the final search step and a greater number of trees in the RFs. We assume these differences to relate to the information content of the feature points themselves. In the medical images, all points were located on bony edges and could therefore be accurately located in at least one dimension using image data alone. In the facial images, however, the image features surrounding each point showed greater variety and the points were relatively poorly constrained by local intensity information (consider, for example, the tip of the nose).

5 Discussion and Conclusions

We have shown that voting with RF regressors trained to predict the offset of points from an evaluated patch is an

effective method for locating feature points. When incorporated into the CLM framework, it achieves excellent performance in a range of application areas. We applied RFRV-CLMs as part of a fully automatic shape model matching system, leading to very accurate and robust results for both hand joint annotation and facial feature point detection. We found that our method outperformed alternative approaches trained on the same data, achieving what we believe to be the best yet published results on hand joint annotation and state-of-the-art performance for facial feature point detection.

We conducted extensive experiments to investigate a range of properties and parameters that have an impact on the performance of our approach, and provided guidance on how to best use it. We have shown that using a single, weighted vote per tree gives good results, and is significantly faster than alternative approaches. The coarseness of the sampling step can be adjusted to balance speed against accuracy as required. We have demonstrated that applying RFRV-CLMs following a coarse-to-fine, multi-stage strategy leads to significant performance improvements. The number of trees in all but the final-stage RFs can be decreased considerably without a significant loss in performance, having a positive impact on running time. Overall the method is fast, allowing tracking of faces at typical video frame rates. Further speed improvements could be achieved via parallelisation since each tree in the RFs searches independently.

In [10], we demonstrated that the proposed method generalises well across medical application areas, achieving what we believe to be the best published results for proximal femur and knee joint segmentation. For facial feature point detection, we would like to point out that although the approach was demonstrated on frontal faces, it would be equally applicable to non-frontal faces, given a suitable training set.

Acknowledgments

The work of C. Lindner was supported by the Medical Research Council, UK (G1000399-1/1). The authors would like to thank K. Ward, R. Ashby, Z. Mughal and Prof. J. Adams for providing the hand radiographs, S. Adeshina for the hand annotations, and H. Yang for providing the data from [43] (Fig. 13). C. Lindner is the corresponding author of the article.

References

[1] T. Cootes, C. Taylor, D. Cooper, and J. Graham, "Active Shape Models - Their training and application," *Comput. Vis. Image Understanding*, vol. 61, no. 1, pp. 38–59, 1995.

[2] S. Milborrow and F. Nicolls, "Locating facial features with an extended Active Shape Model," in *Proc. ECCV 2008*, Springer *LNCS #5305*, 2008, pp. 504–513.

[3] P. Felzenszwalb and D. Huttenlocher, "Pictorial Structures for object recognition," *Int. J. Comput. Vis.*, vol. 61, no. 1, pp. 55–79, 2005.

[4] D. Cristinacce and T. Cootes, "Automatic feature localisation with Constrained Local Models," *J. Pattern Recognit.*, vol. 41, no. 10, pp. 3054–3067, 2008.

[5] J. Saragih, S. Lucey, and J. Cohn, "Deformable model fitting by regularized landmark mean-shift," *Int. J. Comput. Vis.*, vol. 91, no. 2, pp. 200–215, 2011.

[6] J. Gall and V. Lempitsky, "Class-specific Hough Forests for object detection," in *Proc. Comput. Vis. Pattern Recognit.*, 2009, pp. 1022–1029.

[7] L. Breiman, "Random Forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.

[8] M. Valstar, B. Martinez, X. Binefa, and M. Pantic, "Facial point detection using boosted regression and graph models," in *Proc. Comput. Vis. Pattern Recognit.*, 2010, pp. 2729–2736.

[9] T. Cootes, M. Ionita, C. Lindner, and P. Sauer, "Robust and accurate shape model fitting using Random Forest regression voting," in *Proc. ECCV 2012—Part VII*, Springer *LNCS #7578*, 2012, pp. 278–291.

[10] C. Lindner, S. Thiagarajah, M. Wilkinson, The arcOGEN Consortium, G. Wallis, and T. Cootes, "Accurate bone segmentation in 2D radiographs using fully automatic shape model matching based on regression-voting," in *Proc. MICCAI—Part II*, Springer *LNCS #8150*, 2013, pp. 181–1890.

[11] T. Cootes, G. Edwards, and C. Taylor, "Active Appearance Models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, Jun. 2001.

[12] P. Belhumeur, D. Jacobs, D. Kriegman, and N. Kumar, "Localizing parts of faces using a consensus of exemplars," in *Proc. Comput. Vis. Pattern Recognit.*, 2011, pp. 545–552.

[13] F. Zhou, J. Brandt, and Z. Lin, "Exemplar-based graph matching for robust facial landmark localization," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 1025–1032.

[14] R. Donner, B. Menze, H. Bischof, and G. Langs, "Global localization of 3D anatomical structures by pre-filtered Hough Forests and discrete optimization," *Med. Image Anal.*, vol. 17, no. 8, pp. 1304–1314, 2013.

[15] J. Nicolle, K. Bailly, V. Rapp, and M. Chetouani, "Locating facial landmarks with binary map cross-correlations," in *Proc. Int. Conf. Image Process.*, 2013, pp. 501–508.

[16] A. Criminisi, D. Robertson, E. Konukoglu, J. Shotton, S. Pathak, S. White, and K. Siddiqui, "Regression forests for efficient anatomy detection and localization in computed tomography scans," *Med. Image Anal.*, vol. 17, no. 8, pp. 1293–1303, 2013.

[17] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. M. P. Kohli, A. Criminisi, A. Kipman, and A. Blake, "Efficient human pose estimation from single depth images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 12, pp. 2821–2840, Dec. 2013.

[18] F. Schroff, A. Criminisi, and A. Zisserman, "Object class segmentation using Random Forests," in *Proc. Brit. Mach. Vis. Conf.*, 2008, pp. 54.1–54.10.

[19] S. Schulter, P. Wohlhart, C. Leistner, A. Saffari, P. Roth, and H. Bischof, "Alternating decision forests," in *Proc. Comput. Vis. Pattern Recognit.*, 2013, pp. 508–515.

[20] S. Schulter, C. Leistner, P. Wohlhart, P. Roth, and H. Bischof, "Alternating regression forests for object detection and pose estimation," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 417–424.

[21] E. Konukoglu, B. Glocker, D. Zikic, and A. Criminisi, "Neighbourhood approximation using randomized forests," *Med. Image Anal.*, vol. 17, no. 7, pp. 790–804, 2013.

[22] M. Covell, "Eigen-points: control-point location using principal component analyses," in *Proc. Int. Conf. Autom. Face Gesture Recognit.*, 1996, pp. 122–127.

[23] T. Cootes, G. Edwards, and C. Taylor, "Active Appearance Models," in *Proc. ECCV*, Springer *LNCS #1407*, 1998, pp. 484–498.

[24] J. Saragih and R. Goecke, "A nonlinear discriminative approach to AAM fitting," in *Proc. Int. Conf. Comput. Vis.*, 2007, pp. 1–8.

[25] P. Tresadern, P. Sauer, and T. Cootes, "Additive update predictors in Active Appearance Models," in *Proc. Brit. Mach. Vis. Conf.*, 2010, pp. 91.1–91.12.

[26] P. Sauer, T. Cootes, and C. Taylor, "Accurate regression procedures for Active Appearance Models," in *Proc. Brit. Mach. Vis. Conf.*, 2011, pp. 30.1–30.11.

[27] S. Zhou and D. Comaniciu, "Shape Regression Machine," in *Proc. Intell. Platform Manag. Interface*, 2007, pp. 13–25.

[28] P. Dollar, P. Welinder, and P. Perona, "Cascaded pose regression," in *Proc. Comput. Vis. Pattern Recognit.*, 2010, pp. 1078–1085.

[29] X. Cao, Y. Wei, F. Wen, and J. Sun, "Face alignment by explicit shape regression," in *Proc. Comput. Vis. Pattern Recognit.*, 2012, pp. 2887–2894.

[30] X. Xiong and F. De la Torre, "Supervised descent method and its application to face alignment," in *Proc. Comput. Vis. Pattern Recognit.*, 2013, pp. 532–539.

[31] D. Ballard, "Generalizing the Hough transform to detect arbitrary shapes," *Pattern Recognit.*, vol. 13, no. 2, pp. 111–122, 1981.
[32] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Proc. ECCV—Workshop Statist. Learn. Comput. Vis.*, 2004, pp. 17–32.
[33] L. Bourdev and J. Malik, "Poselets: Body part detectors trained using 3D human pose annotations," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 1365–1372.
[34] M. Dantone, J. Gall, G. Fanelli, and L. V. Gool, "Real-time facial feature detection using conditional regression forests," in *Proc. Comput. Vis. Pattern Recognit.*, 2012, pp. 2578–2585.
[35] B. Martinez, M. Valstar, X. Binefa, and M. Pantic, "Local evidence aggregation for regression-based facial point detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1149–1163, May 2013.
[36] S. Jaiswal, T. Almaev, and M. Valstar, "Guided unsupervised learning of mode specific models for facial point detection in the wild," in *Proc. Int. Conf. Comput. Vis. Workshops*, 2013, pp. 370–377.
[37] T. Sheerman-Chase, E.-J. Ong, and R. Bowden, "Non-linear predictors for facial feature tracking across pose and expression," in *Proc. FG*, 2013, pp. 1–8.
[38] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. Comput. Vis. Pattern Recognit.*, 2001, pp. 511–518.
[39] C. Lindner, S. Thiagarajah, M. Wilkinson, The arcOGEN Consortium, G. Wallis, and T. Cootes, "Fully automatic segmentation of the proximal femur using Random Forest regression voting," *IEEE Trans. Med. Imaging*, vol. 32, no. 8, pp. 1462–1472, 2013.
[40] R. Donner, B. Menze, H. Bischof, and G. Langs, "Fast anatomical structure localization using top-down image patch regression," in *Proc. MICCAI—Workshop MCV*, Springer *LNCS #7766*, 2013, pp. 133–141.
[41] M. Kostinger, P. Wohlhart, P. Roth, and H. Bischof, "Annotated Facial Landmarks in the Wild: A large-scale, real-world database for facial landmark localization," in *Proc. Int. Conf. Comput. Vis. Workshops*, 2011, pp. 2144–2151.
[42] S. Milborrow and F. Nicolls, "Active Shape Models with SIFT Descriptors and MARS," in *Proc. VISAPP*, 2014, pp. 5–12.
[43] H. Yang and I. Patras, "Sieving regression forest votes for facial feature detection in the wild," in *Proc. Int. Conf. Comput. Vis.*, 2013, pp. 1936–1943.

Claudia Lindner received the BSc degree and the MSc degree both in computer science with distinction from the Heinrich Heine University Düsseldorf, Düsseldorf, Germany, in 2005 and 2007, respectively. She received the PhD degree in medical computer science from the University of Manchester, Manchester, United Kingdom, in 2014, where she is currently a Research Associate. Her research interests include computer vision, machine learning and medical image computing.

Paul A. Bromiley received the MA degree in natural sciences from the University of Cambridge, Cambridge, United Kingdom, in 1994, the MSc degree in Astrophysics from Queen Mary University of London, London, United Kingdom, in 1995, and the PhD degree in Physics from University College London, London, in 2000. His research interests include statistically motivated approaches to medical image segmentation, registration, and automatic landmark annotation.

Mircea C. Ionita received the BSc and the MSc degrees in computer science and electrical engineering from the University Politehnica of Bucharest, Bucharest, Romania, in 2003 and 2004, respectively. He received the PhD degree in electrical and electronic engineering from the National University of Ireland, Galway, in 2009. He currently works as a Research Scientist, having interests in image processing, computer vision, and machine learning.

Tim F. Cootes received the BSc degree with honours in mathematics and physics from Exeter University, Exeter, United Kingdom, in 1986, and the PhD degree in Engineering from Sheffield City Polytechnic, Sheffield, United Kingdom, in 1991. He began work in computer vision at the University of Manchester, Manchester, United Kingdom, in 1991. He is currently a Professor of Computer Vision at the University of Manchester, with a focus on applications of statistical shape and appearance models.