Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Computer Science 00 (2016) 000–000

Procedia Computer Science

www.elsevier.com/locate/procedia

International Conference On Medical Imaging Understanding and Analysis 2016, MIUA 2016, 6-8 July 2016, Loughborough, UK

Multi-point Regression Voting for Shape Model Matching

P.A. Bromiley*, C. Lindner, J. Thomson, M. Wrigley, T.F. Cootes

Centre for Imaging Sciences, University of Manchester, Stopford Building, Oxford Road, Manchester M13 9PT, U.K.

Abstract

Regression-based schemes have proven effective for locating landmarks on images. Most previous approaches either predict the positions of all points simultaneously, or use regressors that predict individual points combined with a global shape constraint. The former can be efficient, but such models tend to be less robust. Conversely, Random Forest (RF) voting methods for individual points have been shown to be robust and accurate, but can lead to very large models. We explore the continuum between these two approaches by training RF regressors to predict subsets of points.

Multi-point regression voting was implemented within the Random Forest Regression Voting Constrained Local Model framework and evaluated on clinical and facial images. Significant model size reductions were achieved with little difference in accuracy. The approach may therefore be useful where high numbers of points, and limitations on memory or disk space, make single-point models impractically large.

Peer-review under responsibility of the Organizing Committee of MIUA 2016.

Keywords: Random Forests ; Constrained Local Models ; Landmark Annotation ; DXA Imaging

1. Introduction

We propose an automated method for annotating landmarks on deformable structures by voting for their positions using multiple, overlapping sub-models, each trained to predict the positions of subsets of the points. Algorithms such as the Active Appearance Model (AAM)[1], Shape Regression Machine[2] and others[3] use a sequence of regressors to fit the entire shape simultaneously. However, holistic methods tend to generalise poorly[4]. An alternative is to use sequences of regressors for individual points. The ambiguity inherent in the use of local image patches may be dealt with by imposing a global shape constraint using, for example, statistical shape models (SSMs)[5], pictorial structures[6], or Markov random fields (MRFs)[7]. In particular, regression voting (RV) methods[8], especially those[4 9 10] based on Random Forests (RFs)[11], tend to be robust. The RFRV Constrained Local Model (RFRV-CLM)[9], which uses a RF regressor for each point constrained by a global shape model, has been successfully applied to both clinical and facial images[9], and has shown superior generalisation capability compared to the AAM[12].

* Corresponding author. Tel.: +44-161-275-5175.
E-mail address: paul.bromiley@manchester.ac.uk

Peer-review under responsibility of the Organizing Committee of MIUA 2016.

One drawback of such methods is that RFs increase in size linearly with the number of trees and exponentially with tree depth. Clinical image annotation often requires many, densely annotated points, which can make RF-based models impractically large. Some authors [13] have proposed alternative RF structures to reduce size. Here we investigate the use of RF regressors to localise multiple points simultaneously, reducing the number of RFs needed. The process of learning mappings between arbitrarily complex input and output spaces is generally referred to as structured learning [14]. Several authors have investigated the application of RFs within this framework, often using standard input and structured output spaces. For example, Dollár and Zitnick [15] performed edge detection using RFs trained to predict local edge maps from image features. Ebner et al. [16] applied RFs to localise entire sets of points simultaneously in hand MRI. Our novel contribution is to use structured RFs to predict the parameters of SSMs that cover subsets of points. An accumulator array for each point is used to collect votes from each sub-model covering that point, retaining the robustness of RF voting methods. Fitting is constrained using a global SSM. The approach has advantages in terms of flexibility; each sub-model can cover any subset of the points, arbitrary numbers of sub-models can be used, and they can overlap such that several sub-models localise a point using information from different image regions. We briefly describe RFRV and CLMs in Section 2 (see Lindner et al. [9] for details), and then describe how multi-point sub-modelling (MP) is implemented within this framework. We report results on clinical and facial images in Section 3; the latter are included to show the importance of correlations between points in the sub-models.

2. Method

2.1. RF Regression Voting

RFRV uses a single RF regressor for each point, trained to predict the offset to that point based on local patches of image features. The training data consists of a set of images $\mathbf{I}$ with manual annotations $\mathbf{x}_l$ of a set of N points $l = 1...N$ on each. The images are first aligned into a standardised reference frame using a similarity registration, giving a transformation T with parameters θ, and then resampled into this frame by applying $\mathbf{I}_r(m,n) = \mathbf{I}(T_\theta^{-1}(m,n))$, where (m,n) specify pixel coordinates. The reference frame width, in pixels, is controlled by a parameter w_{frame}, allowing variation of the resolution of the resampled images. Random displacements $\mathbf{d}_j$ are generated by sampling from a uniform distribution with apothem d_{max} and the same dimensionality as the images. For each point, image patches of area w_{patch}^2 are extracted at these displacements from each resampled training image, and features $\mathbf{f}_j$ are derived from them. Haar-like features [17] are used, as they have proven effective for a range of applications and can be calculated efficiently from integral images. To allow for inaccurate initial estimates of pose during model fitting, and to make the detector locally pose-invariant, the process is repeated with random perturbations in scale and orientation. A RF is then constructed; each tree is trained on a bootstrap sample of pairs $\{(\mathbf{f}_j, \mathbf{d}_j)\}$ from the training data using a standard, greedy approach. At each node, a random set of n_{feat} features is chosen, and a feature f_i and threshold t that best split the data into two compact groups are selected by minimising an entropy measure [9]. The process is terminated at a maximum depth D_{max} or minimum number of samples N_{min}, and repeated to generate a forest of n_{trees}.

2.2. Constrained Local Models

The CLM [5] uses a SSM to constrain the fitting of models for individual points. The concatenated, reference-frame coordinates of the points in each training image define its shape; the SSM is generated by applying principal component analysis (PCA) to the set of training shapes [1]. This yields a linear model of shape variation, giving the position of point l

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

where $\bar{\mathbf{x}}_l$ is the mean point position in the reference frame, $\mathbf{P}_l$ is a set of modes of variation, $\mathbf{b}$ encodes the shape model parameters, and $\mathbf{r}_l$ allows small deviations from the model.

2.3. RFRV-CLM Fitting

The fitting of a RFRV-CLM to a query image $\mathbf{I}_q$ is initialised via an estimate of pose from a previous model or a manual initialisation, providing estimates for $\mathbf{b}$ and θ. The image is resampled in the reference frame using the current

pose $\mathbf{I}_{qr}(m,n) = \mathbf{I}_q(T_\theta^{-1}(m,n))$. For each point l, a grid of locations $\mathbf{z}_l$ is defined covering a search range of apothem d_{search} around the initial estimate of its position. Regressor R_l is applied to the image features extracted from the local patch around each grid location. Each tree in R_l predicts the offset to the true point position, and casts a vote into an accumulator array C_l at the predicted position. This is performed independently for each point. The shape model places a constraint on the results from all regressors. The quality of fit Q is given by

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

where $\mathbf{S}_b$ is the covariance matrix of shape model parameters $\mathbf{b}$, M_t is a threshold on the Mahalanobis distance, and r_t is a threshold on the residuals. M_t is chosen using the cumulative distribution function (CDF) of the χ^2 distribution so that 98% 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[18]. Q is iteratively optimised, over parameters $\mathbf{p} = \{\mathbf{b}, \theta, \mathbf{r}_l\}$, as described in[9].

2.4. Multi-Point Sub-modelling

In the RFRV-CLM, each RF regressor predicts the offset of a single point using local image features. In the multi-point (MP) algorithm, the RFs are trained to predict the parameters of a function that describes the positions of multiple points. The function used here is a SSM. Let S_k be a set of integers indicating a subset of N_k points (each element of S_k is an integer in the range $[1, N]$). A SSM is built, representing a shape $\mathbf{x}_k$ in the reference frame as

$$\mathbf{x}_k = \bar{\mathbf{x}}_k + \mathbf{P}_k\mathbf{b}_k \quad (3)$$

where $\bar{\mathbf{x}}_k$ is the mean shape, $\mathbf{P}_k$ the modes of variation, and $\mathbf{b}_k$ the shape parameters. The first two columns of $\mathbf{P}_k$ correspond to translation; it is assumed that the scale and rotation components are small and thus well approximated by a linear model. A set of perturbations to the shape parameters $\{d\mathbf{b}_k\}$ are sampled from a uniform distribution of range $[-d\mathbf{b}_{max}, d\mathbf{b}_{max}]$, and applied to each training image. Patches of image data around the perturbed points are then drawn from the images, and features are extracted. The displacements $d\mathbf{b}_k$ and corresponding features are then used to train a RF regressor $R_k()$ that, given the current position of a point subset $\mathbf{x}'_k$, estimates the shape parameter offsets $d\mathbf{b}_k$ required to improve the points

$$d\mathbf{b}_k = R_k(\mathbf{f}_k(\mathbf{x}'_k)) \quad \Rightarrow \quad \hat{\mathbf{x}}_k = \mathbf{x}'_k + \mathbf{P}_k d\mathbf{b}_k \quad (4)$$

Fitting proceeds as within the RFRV-CLM framework. Separate voting arrays are used for each point, and each regressor that includes a given point in its sub-model votes into the corresponding array, allowing the combination of results from multiple, overlapping sub-models. The resulting voting arrays are used as the C_l in Eq. 2.

3. Results

3.1. Segmentation of Vertebrae in DXA Images

A series of experiments was performed to evaluate the effect of the number of points in the sub-models on the performance of the MP algorithm, using 320 dual-energy X-ray absorptiometry (DXA) images of the spine, split randomly into halves for training and testing. Sub-models were constructed from contiguous sets of points along boundaries and, to reduce the size of the parameter space, overlap was set to $\lfloor (n_{points}/2) \rfloor$ i. e. each point was contained within a maximum of two sub-models. The dimensionality of $\mathbf{b}_k$ was chosen by removing PCA modes that varied point positions by less than 0.5 pixels. The experiments focused on a single vertebra (L2) to limit CPU time and, as in Roberts et al.[19], the model covered the target vertebra and both of its neighbours (L1-L3). Manual annotation of 130 points on each image was performed by an expert radiographer (see Fig. 1(a)). Errors were calculated as the mean, over the 38 points on L2, of the distances both between the automatic and manual points (point-to-point (P-to-P) error) and, to compensate for the aperture problem, of the minimum distances between the automatic points and a Bezier spline through the manual points (point-to-curve (P-to-C) error). A 2-stage, coarse-to-fine model was used and all free parameters were set to the values given in Bromiley et al.[20].

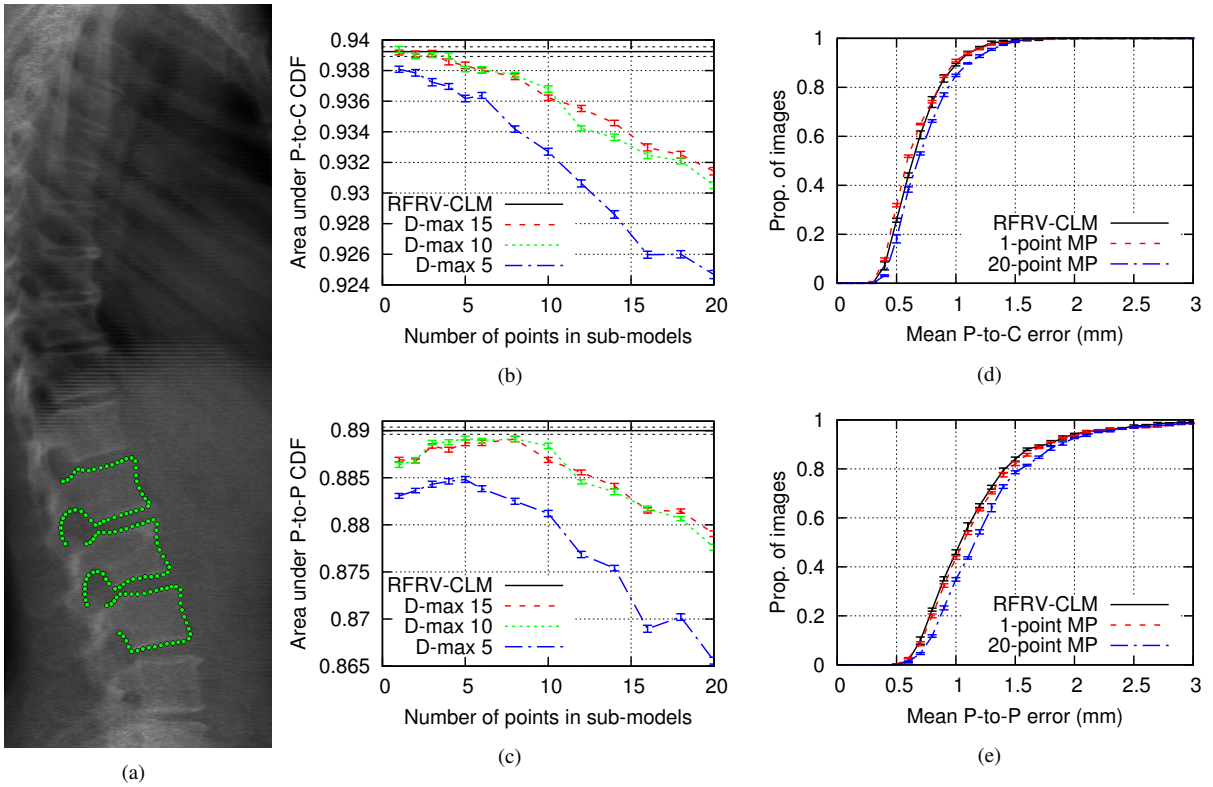


Fig. 1. (a) Example 130-point annotation of the L1-L3 vertebrae on a DXA spinal image. (b,c,d,e) Performance of MP in annotation of 38 points on the L2 vertebra in 160 DXA images of the spine: (b,c) The proportional area under the CDF of mean P-to-C and P-to-P error with varying N_k, together with the results from (non-MP) RFRV-CLM with $D_{max} = 15$, $N_{min} = 1$; (d,e) CDFs of errors for MP with $N_k = 1$ and 20 with $D_{max} = 15$, compared to the results from RFRV-CLM. The error bars show the standard error on the mean of five repeat experiments.

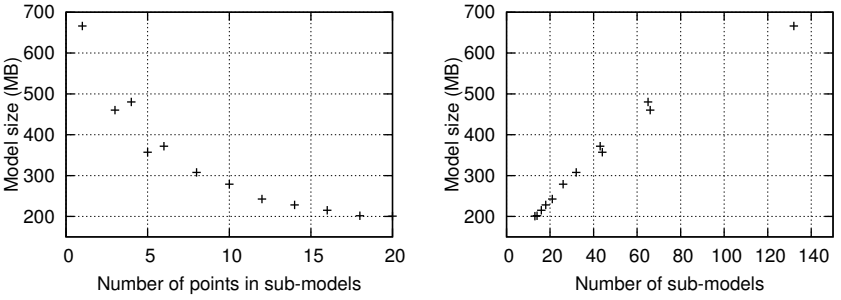
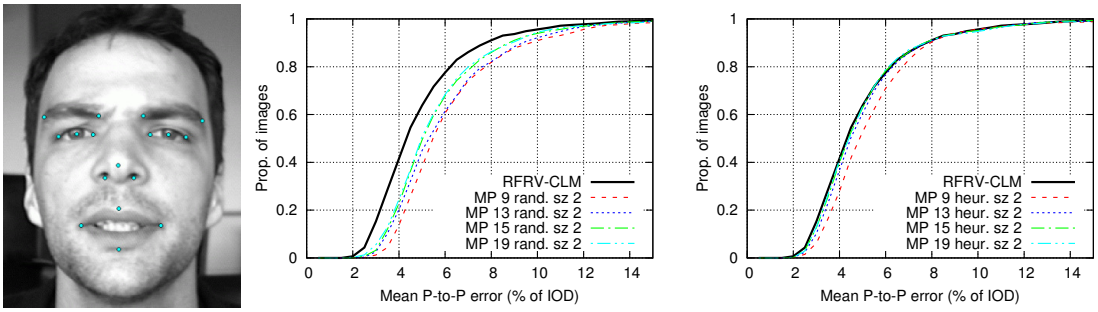


Fig. 2. MP size on disk compared to the number of points in each sub-model and the number of sub-models, for $D_{max} = 15$ and $N_{min} = 1$.

Three sets of experiments were performed to evaluate the dependence of MP annotation accuracy on sub-model size N_k, varying N_k from 1 to 20 for each of $D_{max} = 15, 10$ and 5. Figs. 1(b,c) show the results as the proportional area under the cumulative distribution function (CDF) of P-to-P and P-to-C errors. Examples of the CDFs for the experiments with $D_{max} = 15$, $N_{min} = 1$ and N_k of 1 and 20, and results from a 2-stage, coarse-to-fine RFRV-CLM, are shown in Figs. 1(d,e). Considering P-to-C error first, performance fell with N_k regardless of D_{max}, but the reduction was greater at lower D_{max}. This implies that the reduction in performance was due to the dimensionality of the output space increasing with N_k whilst the number of samples available to populate it was constant, since constraining RF depth limits the ability to model more complex output spaces. The points were densely annotated along bone edges, and so most of the single-point regressors had limited information about the spacing of points along the edge. The



(a) (b) (c)

Fig. 3. (a) Example facial image from the BioID data set, with 17-point annotation. (b,c) CDFs of the mean P-to-P error on BioID, relative to the IOD, comparing RFRV-CLM to MP with 19, 15, 13 and 9 random (b) and heuristically chosen (c) sub-sets of size 2.

use of MP sub-models imposed stronger constraints on inter-point spacing, and this was reflected in the P-to-P error, which reduced with N_k up to 6. At this value, neither P-to-P nor P-to-C errors were significantly larger for the MP algorithm compared to the single-point RFRV-CLM. The increase in P-to-C error across the range of N_k tested was small ($\approx 0.2mm$) compared to the error itself, and all models attained 100% on the CDF at the same value i.e. were equally robust. Figure 2 shows that the model size on disk depended linearly on the number of regressors i.e. had a significant (a factor of 3.5 difference between $N_k = 1$ and 20) inverse linear relationship to N_k. The CPU time required for model fitting showed a similar, but weak (a 7.1% drop from $N_k = 1$ to 20), dependence on N_k.

3.2. Facial Landmark Annotation in Natural Images

The points in the spinal images were densely aligned along the edges of bony structures, providing a natural ordering in which to select sub-sets and significant correlations between neighbouring points. To demonstrate sub-model point selection in images that did not have this property, and applicability to a wide range of image types, the method was also applied to natural images of faces. Models were trained to annotate 17 points, as shown in Fig. 3(a). A multi-stage model was used as described in Lindner et al.[9]. The models were trained on 267 images from the AFLW data set[21], and tested on 1476 images from the BioID data set[22]. Performance was evaluated using the mean P-to-P distance relative to the inter-ocular distance (IOD)[5].

Two sets of experiments were performed using $N_k = 2$ and varying the number of sub-models. In the first, the points in each sub-model were chosen randomly, i. e. were dissimilar in location and appearance. In the second, they were chosen heuristically, to minimise the Euclidean distance between them. A RFRV-CLM with identical parameters was used for comparison. The results are shown in Fig. 3. When points were chosen heuristically, to maximise their correlations, performance of models containing between 13 and 19 regressors was not significantly different to that of the single-point method. The 19-regressor model, containing more than one regressor per point, did not result in improved performance compared to the RFRV-CLM. However, only when the number of regressors was reduced to 9, the smallest number of 2-point regressors that could annotate all of the points, was a reduction in performance observed. In contrast, when the points were chosen randomly, a reduction in performance was observed regardless of the number of regressors used. This demonstrates that the MP method extracts information from the correlations of points within the sub-models. As with the spinal images, the differences in performance were relatively small, and all models attained 100% on the CDF at the same value i.e. were equally robust.

4. Conclusion

Most previous regression-based approaches for automatic landmark annotation fall into two groups; holistic appearance models such as the AAM, where all points are predicted from a single model, and atomistic models such as the RFRV-CLM, where local intensity is modelled separately for each point, and a global shape constraint is used during fitting. The aim of this paper was to explore the continuum between these approaches, by training structured regressors to predict the parameters of shape models that, in turn, predicted the positions of subsets of points.

Evaluation on densely annotated DXA images indicated that, at optimal sub-model size, there was no significant difference in performance compared to RFRV-CLM. However, the corresponding reduction in the size of the model on disk was highly significant, being linear in the number of regressors. Performance dropped when the sub-model size was increased beyond the optimal value, due to the increasing output space dimensionality. However, the magnitude of the reduction was small, and RFRV-CLM and the MP algorithm were equally robust, achieving 100% on the CDF at the same point. The method was also used to segment the proximal femur in 839 anteroposterior (AP) pelvic radiographs showing unilateral hip osteoarthritis, and the tibia in 500 AP knee radiographs showing varying stages of osteoarthritis, with identical conclusions; these results are not included here due to space constraints.

Experiments on natural images of faces demonstrated that the method is applicable to a wide variety of image types, and showed the importance of selecting correlated points to train the sub-models. This also proved that the MP method compensates for the smaller number of regressors used, compared to RFRV-CLM, by extracting additional information from the correlations between points in each sub-model. We conclude that MP can significantly reduce model size with little or no loss in accuracy, and so allows larger numbers of points to be modelled without exceeding available memory or disk space. Furthermore, since MP does not focus on reducing the size of each RF, it could be combined with methods that do, such as global retraining and pruning [23], to produce further size reductions.

References

1. Cootes, T.F., Edwards, G.J., Taylor, C.J.. Active appearance models. *IEEE TPAMI* 2001;**23**:681–685.
2. Zhou, S.K., Comaniciu, D.. Shape regression machine. *Inf Process Med Imaging* 2007;**20**:13–25.
3. Cao, X., Wei, Y., Wen, F., Sun, J.. Face alignment by explicit shape regression. In: *Proc. CVPR*. 2012, p. 2887–2894.
4. Dantone, M., Gall, J., Fanelli, G., Van Gool, L.. Real-time facial feature detection using conditional regression forests. In: *Proc. CVPR*. 2012, p. 2578–2585.
5. Cristinacce, D., Cootes, T.. Automatic Feature Localisation with Constrained Local Models. *Journal of Pattern Recognition* 2008; **41**(10):3054–3067.
6. Felzenswalb, P., Huttenlocher, D.. Pictorial structures for object recognition. *International Journal of Computer Vision* 2005;**61**(1):55–79.
7. Donner, R., Micusik, B., Langs, G., Bischof, H.. Sparse MRF appearance models for fast anatomical structure localisation. In: Rajpoot, N., Bhalerao, A., editors. *Proc. BMVC*. 2007, p. 1080–1089.
8. Valstar, M.F., Martinez, B., Binefa, X., Pantic, M.. Facial point detection using boosted regression and graph models. In: *Proc. CVPR*. 2010, p. 2729–2736.
9. Lindner, C., Bromiley, P.A., Ionita, M., Cootes, T.F.. Robust and Accurate Shape Model Matching using Random Forest Regression Voting. *IEEE TPAMI* 2015;**37**(9):1862–1874.
10. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E.. Regression Forests for Efficient Anatomy Detection and Localization in CT Studies. In: Menze, B., Langs, G., Tu, Z., Criminisi, A., editors. *MICCAI 2010 Workshop MCV*; vol. 6533 of *LNCS*. Springer, Heidelberg; 2001, p. 106–117.
11. Breiman, L.. Random Forests. *Machine Learning* 2001;**45**:5–32.
12. Sauer, P., Cootes, T., Taylor, C.. Accurate Regression Procedures for Active Appearance Models. In: *Proc. BMVC*. 2011, p. 30.1–30.11.
13. Shotton, J., Sharp, T., Kohli, P., Nowozin, S., Winn, J., Criminisi, A.. Decision Jungles: Compact and Rich Models for Classification. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q., editors. *Advances in Neural Information Processing Systems 26*. 2013, p. 234–242.
14. Nowozin, S., Lampert, C.H.. Structured learning and prediction in computer vision. *Foundations and Trends in Computer Graphics and Vision* 2011;**6**:185365.
15. Dollár, P., Zitnick, C.L.. Fast edge detection using structured forests. *PAMI* 2015;**37**(8):1558 – 1570.
16. Ebner, T., Stern, D., Donner, R., Bischof, H., Urschler, M.. Towards Automatic Bone Age Estimation from MRI: Localization of 3D Anatomical Landmarks. In: *Proc. MICCAI 2014*; vol. 8674 of *LNCS*. Springer-Verlag, Berlin; 2014, p. 421–428.
17. Viola, P., Jones, M.. Rapid object detection using a boosted cascade of simple features. In: *Proc. CVPR*. IEEE Computer Society; 2001, p. 511–518.
18. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.. Active Shape Models - Their training and application. *Comput Vis Image Und* 1995; **61**(1):38–59.
19. Roberts, M.G., Cootes, T.F., Adams, J.E.. Automatic Location of Vertebrae on DXA Images Using Random Forest Regression. In: *Proc. MICCAI 2012*; vol. 7512 of *LNCS*. Springer-Verlag, Berlin; 2012, p. 361–368.
20. Bromiley, P.A., Adams, J., Cootes, T.F.. Localisation of Vertebrae on DXA Images using Constrained Local Models with Random Forest Regression Voting. In: *Recent Advances in Computational Methods and Clinical Applications for Spine Imaging*. Springer International Publishing Switzerland; 2015, p. 159–171.
21. Kostinger, M., Wohlhart, P., Roth, P.M., Bischof, H.. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In: *ICCV Workshops*. 2011, p. 2144–2151.
22. Jesorsky, O., Kirchberg, K.J., Frischholz, R.. Robust face detection using the Hausdorff distance. In: *Proc. 3rd Conference on Audio- and Video-based Biometric Person Authentication*. 2001, p. 90–95.
23. Ren, S., Cao, X., Wei, Y., Sun, J.. Global refinement of random forest. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, p. 723–730.