Error Estimation for Appearance Model Segmentation of Musculoskeletal Structures Using Multiple, Independent Sub-models

Paul A. Bromiley[1(✉)], Eleni P. Kariki[2], and Timothy F. Cootes[1]

[1] Centre for Imaging Sciences, School of Health Sciences, University of Manchester, Manchester, UK
{paul.bromiley,timothy.f.cootes}@manchester.ac.uk
[2] Radiology and Manchester Academic Health Science Centre, Manchester University Hospitals NHS Foundation Trust, Manchester, UK
eleni.kariki@mft.nhs.uk

Abstract. Segmentation of structures in clinical images is a precursor to computer-aided detection (CAD) for many musculoskeletal pathologies. Accurate CAD systems could considerably improve the efficiency and objectivity of radiological practice by providing clinicians with image-based biomarkers calculated with minimal human input. However, such systems rarely achieve human-level performance, so extensive manual checking may be required. Their practical utility could therefore be increased by accurate error estimation, focusing manual input on the images or structures where it is needed. Standard techniques such as the minimum variance bound can estimate random errors, but provide no way to estimate any systematic errors due to model fitting failure.

We describe the use of multiple, independent sub-models to estimate both systematic and random errors. The approach is evaluated on vertebral body segmentation in lateral spinal images, demonstrating large (up to 50%) and significant improvements in the accuracy of error classification with concurrent improvements in annotation accuracy. Whilst further work is required to elucidate the definition of "independence" in this context, we conclude that the approach provides a valuable component for appearance model based CAD systems.

1 Introduction

Standard statistical techniques exist to estimate errors on model fitting processes. For example, in maximum likelihood or equivalent techniques such as cross-correlation, the covariance matrix of the fitted model parameters is bounded by (e.g. [2])

$$C^{-1}_{\theta_{r,s}} \leq - < \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} >= - \frac{\partial^2 \log L}{\partial \theta_r \partial \theta_s} \bigg|_{\theta = \theta_{max}} \tag{1}$$

G. Zheng et al. (Eds.): CSI 2018, LNCS 11397, pp. 53–65, 2019.
https://doi.org/10.1007/978-3-030-13736-6_5

where L is the likelihood function, r, s index a vector of parameters θ, and the equality on the right-hand side is true in the large N limit. This is known as the Minimum Variance Bound (MVB) and has been successfully applied to estimate errors on registration, patch-matching and landmark localisation algorithms (e.g. [5,9,15]). However, Eq. 1 shows that the covariance matrix on the fitted model parameters is bounded by the width of the log-likelihood function about the fitted optimum. It is not sensitive to any systematic error introduced by unmodeled modes of variation in the data, or fit failure due to convergence on a local optimum, as shown in Fig. 1. In general, without either a prior distribution on the systematic errors or a perfect model, there is no way to estimate systematic errors since they cannot be randomly sampled.

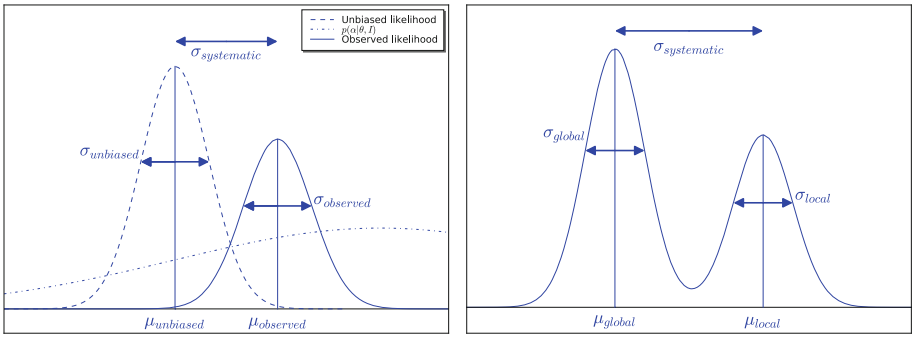


Fig. 1. (a) Unmodeled modes of variation in the data introduce a biasing parameter α and so a prior term $p(\alpha|\theta, I)$ where I is the query image and θ the model parameters. (b) Use of a local, rather than global, optimiser may allow fitting to converge on a local optimum. In either case, the minimum variance bound estimates the accuracy with which a given optimum has been found, which is dependent on the width of that optimum (σ_{biased} or σ_{local}) about its mean (μ_{biased} or μ_{local}), but is not sensitive to the systematic error $\sigma_{systematic}$.

It is often the case in medical image analysis that the most significant errors are systematic, and induced by the use of imperfect models that cannot account for all of the non-noise variation in the data. Such models can be considered as existing on sub-spaces in the space of the perfect model i.e. they span some, but not all, of the modes of variation of that model, and each has a systematic error on a given query image as a result. However, if multiple models could be produced, independent in the sense that they exist on different sub-spaces, their results would include random samples from the population of all possible systematic errors. The standard techniques for random error analysis could then be applied to estimate the systematic errors. This approach has been successfully applied to landmark annotation for Computed Tomography (CT) images using patch-based rigid registration [6]. Here, we explore its application to appearance model segmentation of musculoskeletal images.

As an exemplar task, Random Forest Regression Voting Constrained Local Models (RFRV-CLMs) [13] were applied to segment vertebrae in Dual-Energy X-ray Absorptiometry (DXA) spinal images, to support classification of osteoporotic vertebral fractures (VFs). This combination of method and application had several advantages. Osteoporosis is a common, degenerative disease that increases the risk of fragility fractures, which most frequently occur in the vertebrae, wrists and hips. Approximately 40% of postmenopausal Caucasian women are affected, increasing their lifetime fragility fracture risk to as much as 40% [14]. The impact of the disease is expected to grow as the population ages [7]. Early identification of osteoporotic VFs is therefore clinically important. However, the false negative rate for VF identification is high. A recent audit at a large UK hospital revealed a reporting rate of 36% on CT images [12], and similarly low rates have been reported elsewhere [1]. VF identification on CT images may be opportunistic. However, a recent multi-centre, multinational prospective study on VF reporting for lateral radiographs found a false negative rate of 34% [8]. The potential utility of computer-aided diagnostic (CAD) systems for VF identification in clinical images is therefore high. RFRV-CLMs have previously been applied to this task in both DXA [3] and CT [4] demonstrating state-of-the art annotation and classification accuracy. However, these publications showed that model fitting failure limited classification accuracy. A reliable method to identify such errors would considerably improve the practical utility of VF CAD systems based on RFRV-CLMs, avoiding much of the need for manual inspection and/or correction of the results.

2 Method

Random Forest Regression Voting Constrained Local Models. In the interests of brevity we provide only a summary of the RFRV-CLM and refer the reader to [13] for full details. RFRV-CLMs match a series of landmark points, described by a statistical shape model (SSM), to a query image. They consist of a SSM and a set of independent, local models of the image intensities around each point. The latter are aligned to the query image independently, with the SSM providing a global constraint. The training data consists of a set of images, each annotated with n homogeneous points $\mathbf{x}_l$, where $l = 1...n$. The sets of points are first aligned to remove non-shape variation using e.g. a similarity transformation. The shape in each aligned image is represented as a vector comprising the concatenated coordinates of the points in that image. Principal Component Analysis (PCA) is applied to these vectors to extract the main modes of variation $\mathbf{P}$. A linear model is then constructed giving $\mathbf{x}_l$ as the mean point position $\bar{\mathbf{x}}_l$ in a suitable reference frame, plus some proportion b of each of the modes of variation

$$\mathbf{x}_l = T_\theta(\bar{\mathbf{x}}_l + \mathbf{P}_l\mathbf{b} + \mathbf{r}_l) \tag{2}$$

where $\mathbf{P}_l$ is the sub-matrix of $\mathbf{P}$ relevant to l, and $\mathbf{b}$ are referred to as the shape parameters. T_θ is the transformation, with parameters θ, from the reference

frame to the query image, and $\mathbf{r}_l$ allows small deviations from the model. Fitting to a query image $\mathbf{I}$ proceeds by optimising a quality of fit Q over parameters $\mathbf{p} = \{\mathbf{b}, \theta, \mathbf{r}_l\}$, where

$$Q(\mathbf{p}) = \Sigma_{l=1}^{n} C_l(T_\theta(\bar{\mathbf{x}}_l + \mathbf{P}_l\mathbf{b} + \mathbf{r}_l)) \quad \text{s.t.} \quad \mathbf{b}^T\mathbf{S}_b^{-1}\mathbf{b} \leq M_t \quad \text{and} \quad |\mathbf{r}_l| < r_l \quad (3)$$

The threshold M_t is a shape constraint and is applied to the Mahalanobis distance of $\mathbf{b}$, using the covariance matrix $\mathbf{S}_b$ of the $\mathbf{b}$ from the training data, and r_l is a threshold on the residuals. The cost images C_l are produced by Random Forest (RF) regression voting. For each l, patches are sampled from the image at a set of random displacements from $\mathbf{x}_l$ in the reference frame, Haar-like features are derived from the patches, and a RF regressor is trained to predict the displacement from the features. During fitting each RF is scanned across the image around the current estimate of the point location and the predicted displacements are entered into a voting array C_l.

Data Collection and Manual Annotation. The dataset used in the evaluation consisted of 320 DXA VF assessment (VFA) images scanned on various Hologic (Bedford MA) scanners, with manual annotation of 33 landmarks on each vertebra from T7 to L4; see Fig. 2 for example images. Each vertebra was also classified by an expert radiologist into one of five groups (normal, deformed but not fractured, and grade 1 (mild), 2 (moderate), and 3 (severe) fractures as defined by Genant et al. [10].

RFRV-CLM Training and Fitting. The training procedures and parameters described in [3] were used. RFRV-CLMs were trained to model landmarks on triplets of neighbouring vertebrae, using training data from all levels between T7 and L4 such that the models could fit any level. Two-stage, coarse-to-fine models were used with two trees in the first stage and 15 in the second. Fitting to query images was initialised using manual annotations of vertebral body centroids. The shape constraint in Eq. 3 was removed in the last iteration of second-stage fitting to avoid correlated errors between the landmarks. The model was fitted to all triplets of centroids between T7 and L4, and landmarks from the central vertebrae of each (plus the extremal vertebrae on the first and last triplets) were concatenated to produce a segmentation of the vertebrae.

Error Estimation Methodology. The evaluation of the proposed approach to systematic error detection was based on comparing two model-training regimes. The first, referred to below as "multi-model", evaluated error estimators based on multiple, independent models. Weak independence was induced by training models on independent data sets; see Sect. 4 for comments on potential routes to formally inducing strong independence. The data set was divided into eighths and models were trained on each. Therefore, seven models trained on independent sets of images were available to fit each query image, producing seven independent estimates $\mathbf{x}_j$ for each landmark location. The final annotation was

produced by taking the centroid $\mathbf{x}_c$ of the multiple estimates. An error estimator sensitive to any fit failures across the set of models was calculated as the root-mean-square (RMS) of the Euclidean distances between the individual estimates and their centroid, and is referred to below as RMS goodness-of-fit (RMSGOF)

$$\mathbf{x}_c = \frac{1}{k}\sum_{j=0}^{k}\mathbf{x}_j \qquad RMSGOF = \sqrt{\frac{1}{k}\sum_{j=0}^{k}(\mathbf{x}_j - \mathbf{x}_c)^2} \tag{4}$$

For comparison, a standard four-fold cross validation was also performed and is referred to below as "single model". Here, models were trained on 3/4 of the data and tested on the remaining 1/4, such that one model was tested on each query image. Since RFRV-CLMs contain regressors capable of predicting the location of the landmark given patches of image data, a simple goodness-of-fit measure sensitive only to random errors was produced by applying the regressor at the optimised point position to estimate the residual; this is referred to below as RGOF (residual GOF). RGOF was also calculated for the multi-model approach by taking the mean of the RGOF from each of the multiple model fits for a given point, and a combined GOF, or CGOF, was produced by taking the product of the mean RGOF and the RMSGOF.

In both cases, the true error on the RFRV-CLM annotations was calculated as the Euclidean distance to the corresponding manual annotation. This is referred to below as point-to-point, or P2P, error. The RF parameters were kept consistent between the single model and multi-model approaches. In addition, multi-model training used all vertebral triplets from each training image whilst single-model training used only one; since T7 to L4 annotation provided eight triplets per image, this ensured that the number of training samples used for each model was consistent across both approaches.

Vertebral Fracture Classification. VF classification was performed using a simple approach based on six-point morphometry [11]. The anterior H_a, middle H_m, and posterior H_p heights of each vertebral body were calculated as the Euclidean distances between the relevant landmark pairs. The predicted posterior body height $H_{p'}$ was also calculated from the posterior heights of the closest four annotated vertebrae by taking the largest of the four values, since fractures decrease vertebral height. Three ratios were then calculated to measure the relative height reductions at the anterior (wedge ratio, H_a/H_p), middle (biconcavity ratio, H_m/H_p) and posterior (crush ratio, $H_p/H_{p'}$) positions. The data were whitened by subtracting the median and dividing by the square root of the covariance matrix, estimated using the median absolute deviation. Normal vertebrae predominated, and so this was equivalent to whitening to the mean and standard deviation of the normal class, without using manual classifications. A simple classifier was then constructed by placing a threshold t_c on the Euclidean distance from the origin to separate the data into fractured and non-fractured classes, the latter including both normal and deformed vertebrae. Error estimates for classification were derived from RMSGOF, CGOF

and RGOF by applying standard error propagation to the above calculation to produce a scaled estimate of the error HGOF on the length of the vector defined by the three whitened height ratios R_w, R_b and R_c, and calculating the ratio of this estimate to the distance of the data point from the decision boundary

$$ClassGOF = \frac{HGOF}{|t_c - \sqrt{R_w^2 + R_b^2 + R_c^2}|} \tag{5}$$

3 Evaluation

Figure 2 shows example images and serves as a flow diagram illustrating the method. Taking the original images (a, f) as input, together with manual annotations of vertebral body centres, RFRV-CLMs are fitted to produce high-resolution annotations of the vertebral bodies as a precursor to VF classification.

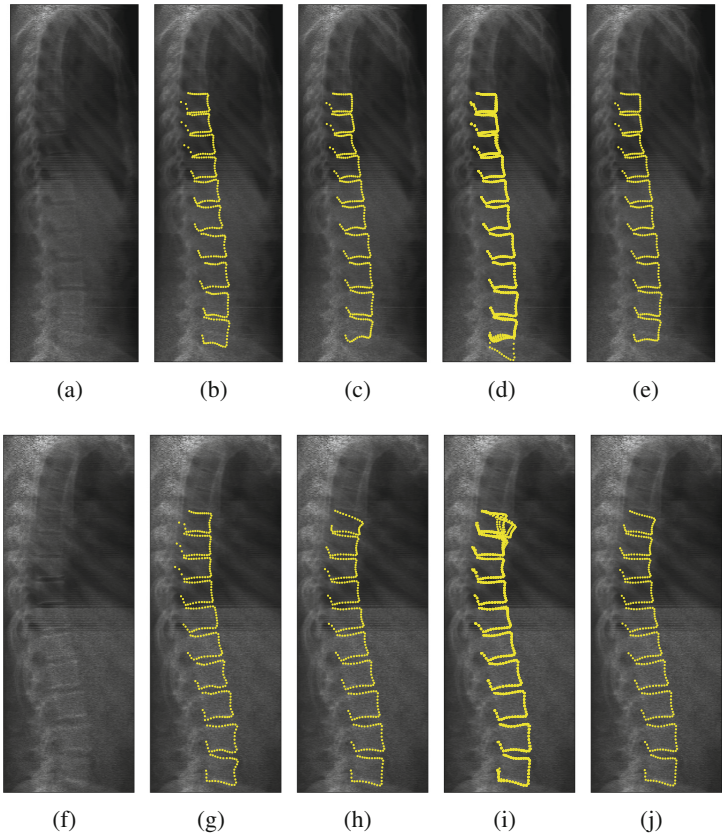
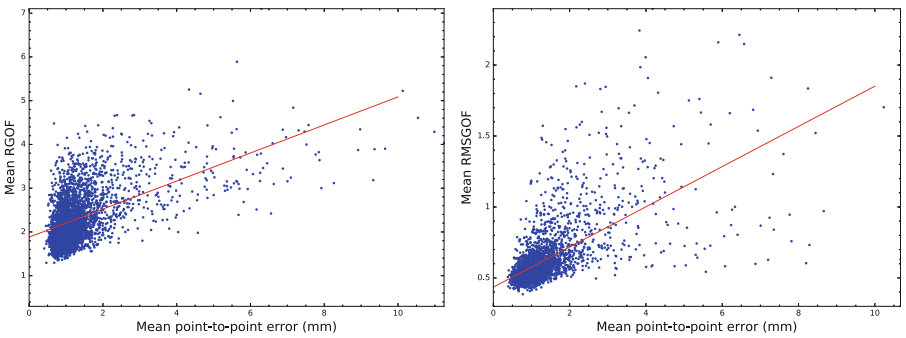


Fig. 2. Examples of image annotation using single and multiple models. (a, f) Original images. (b, g) Manual annotations of T7 to L4. (c, h) Automatic annotation using a single RFRV-CLM. (d, i) Automatic annotations from multiple, independent models. (e, j) Centroids of the multiple estimates for each landmark.

Comparing manual annotations (b, g) to automatic annotations produced by a single model (c, h), it can be seen that some vertebra (L4 in (c) and T7 in (h)) are poorly fitted, leading to the appearance of reduction in anterior vertebral body height that leads to misclassification of these normal vertebrae as fractured. Since these errors are systematic, rather than random, techniques based on the MVB will not identify them. However, if multiple, independent sub-models are fitted (d, i), they can serve to sample the systematic errors. The centroids of the multiple estimates for each point (e, j) serve as the final annotation.

The first stage of the evaluation focused on estimating the mean P2P error across each vertebral body. Figure 3 shows scatterplots of the mean single-model RGOF and multi-model RMSGOF for each vertebra against the vertebral mean P2P error. The correlation coefficient was 0.54 for the vertebral mean single model RGOF, 0.50 for the mean multi-model RGOF, 0.63 for the mean RMSGOF, and 0.67 for the mean CGOF, indicating that the RMSGOF is more strongly correlated to the P2P error than the RGOF. The CGOF resulted in a small improvement in correlation, indicating that there is some independent information between the RGOF and RMSGOF.

To provide a more quantitative interpretation of the various error estimators, they were used to construct binary classifiers. The ground truth classification for each vertebra was produced by imposing a threshold on mean P2P error, set to the 95th percentile of the error distribution, corresponding to 2.2 mm. Figure 4 shows ROC curves produced by applying a threshold to the error estimators and comparing the classification to the ground truth. Error estimators based on multiple models resulted in a large and significant increase in classification accuracy, e.g. raising the precision at 50% recall from 42.1% for a single-model RGOF to 59.3% for RMSGOF and 63.8% for CGOF.



(a) (b)

Fig. 3. Goodness-of-fit (GOF) measures vs. the P2P errors on automatically annotated points: (a) mean single-model RGOF; (b) RMSGOF. Each graph also shows a linear fit to the data.

The results discussed so far indicate that the use multiple, independent sub-models is helpful in error estimation. However, the effect on the accuracy of

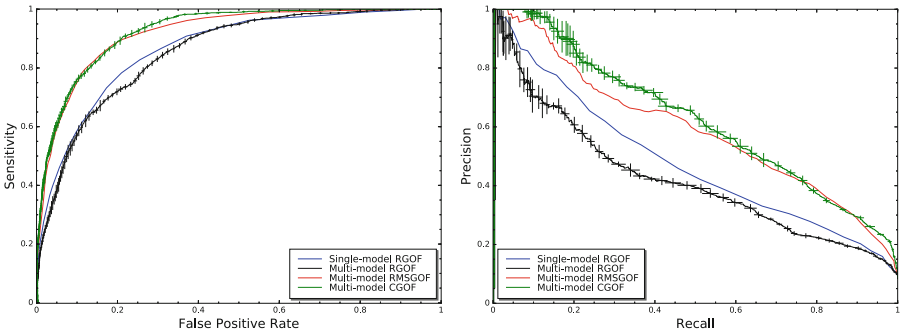


Fig. 4. ROC curves for binary classifiers of mean vertebral error using various error estimators. The ground truth was provided by a threshold on the mean vertebral P2P error corresponding to the 95th percentile of its distribution.

point localisation must also be evaluated. Since each of the multiple sub-models is trained on a smaller set of images, it might be expected that the resultant regressors would provide point location estimates with larger errors. In practice, the opposite was found. Figure 5 shows CDFs of the mean vertebral P2P errors divided by vertebral classification, for both single and multiple models. In general, multi-model annotation proved to be slightly more accurate than single model annotation, although the differences were small. This also accounts for the difference in accuracy between single and multi-model RGOF in Fig. 4; the multi-model annotation makes fewer errors and so they are more difficult to identify. However, Fig. 6 shows ROC curves for a six-point morphometry classifier applied to both the manual annotations and automated annotations from single and multi-models. Multi-model VF classification was significantly more accurate, approximately halving the difference compared to classification from manual annotations.

To investigate this difference more thoroughly, several additional model training strategies were applied, and the results are also shown in Fig. 6. As described in Sect. 2, in the experiments described up to this point single models were trained on one vertebral triplet from each of the training images whilst, when training multiple models, the training images were divided into eighths and one model was trained on each, using all of the vertebral triplets. Furthermore, the dimensions of the RFs were consistent, with two trees in the first stage and fifteen in the second stage. Therefore, the number of trees and training samples for each individual model was consistent but, as an ensemble, the multiple sub-models had seven times more trees and training samples available. To test whether this accounted for the difference in accuracy between the single and multi-model approaches, additional single models were trained with all vertebral triplets from all images, and with increased numbers of trees in the first and second stages, and the results are shown in Fig. 6. Increasing the training

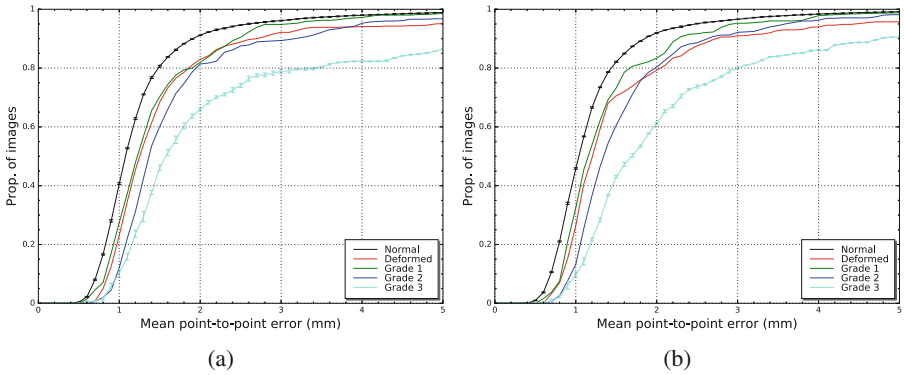


Fig. 5. Cumulative distribution functions of the vertebral mean P2P error on RFRV-CLM annotations using a single model (a) and multiple models (b), for each vertebral classification.

sample approximately halved the difference between the single and multi-model approaches, and increasing the number of trees in each stage produced further but smaller increases in accuracy. However, the single model still failed to achieve the accuracy of the multi-model approach. Conversely, the size of the model on disk increased dramatically. For example, the first stages of the multi-models were on average 290 MB; single-model first stages were 230 MB for two trees and one triplet per image and, when using all triplets, 1.2 GB for 2 trees and 5.0 GB for 8 trees, making the latter impractically large. This indicates that dividing training samples between multiple, independent models provides a more efficient way in which to use large data sets. The single models used in the remaining experiments reverted to the training strategy described in Sect. 2.

The final stage of the evaluation focused on identifying errors in the 6-point morphometry fracture classification. Classification and error estimation were performed as described in Sect. 2. A classification threshold t_c (an operating point in Fig. 6) giving 90% sensitivity was selected. A second threshold was applied to the ClassGOF (derived from CGOF) to classify the VF classification as accurate or erroneous. The ground truth was provided by the manual classification of each vertebra, and the threshold on ClassGOF was varied to produce the ROC curves shown in Fig. 7(a).

Comparison of the results from single and multi-model error estimation for VF classification is complicated by the fact that, as shown in Fig. 6, multi-model fracture classification is more accurate, and so the number of errors to be detected is smaller. However, in contrast to the results for classifying errors on vertebral mean errors, the multi-model approach did not provide significantly more accurate error estimation for VF classification compared to the single-model approach. To illustrate why this occurred, Fig. 8 shows the distributions of multi-model annotation error, across all vertebrae and images, for each of the 33 landmarks. However, instead of P2P error, the figure shows the

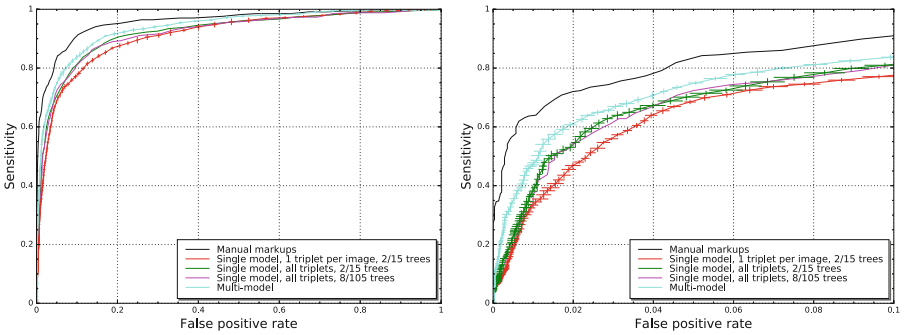


Fig. 6. ROC curves of osteoporotic VF classification using 6-point morphometry, for both manual landmarks and various automated annotations; (b) shows a detail from (a).

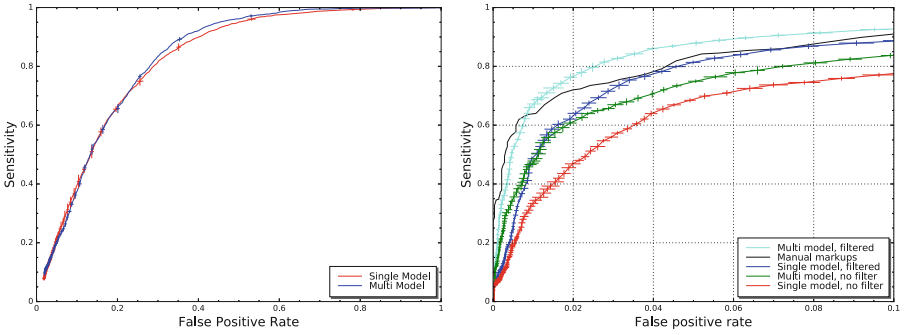
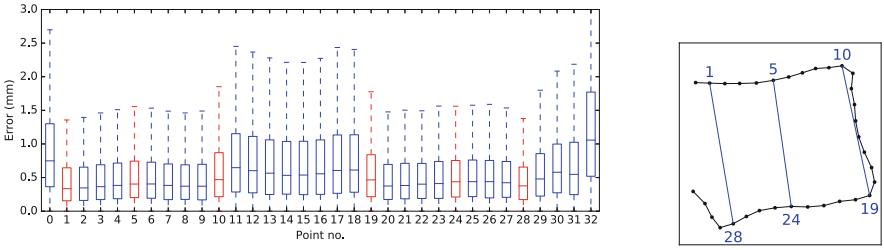


Fig. 7. (a) ROC curves showing the accuracy of error classification on the results of VF classification using 6-point morphometry. (b) ROC curves of osteoporotic VF classification using 6-point morphometry, for both manual landmarks, the single and multi-model automatic annotations, and these automatic annotations after filtering out results detected as erroneous by an error classifier.

point-to-curve (P2C) errors i.e. the minimum Euclidean distance between each point and a piecewise-linear curve through the manual annotations. Large P2C errors, where the points move away from the vertebral body edge, are predominantly found on the anterior side and the pedicle, whilst the points used in VF classification are more accurate, implying they are less subject to fit failure. Therefore, mean vertebral P2P error estimation benefits from the use of RMSGOF and the sensitivity of the technique to systematic error/fit failure, whilst error estimation for VF classification does not. However, the multi-model approach did not result in significantly worse error estimation.



(a) (b)

Fig. 8. (a) Box-and-whisker plots of the point-to-curve errors on the multi-model centroid estimates of each point. The points shown in red are those used to estimate heights for fracture classification, as shown in (b). (Color figure online)

To provide a more quantitative interpretation of the use of single and multi-model error estimation for VF classification, Fig. 7(b) shows ROC curves of the 6-point morphometry fracture classifier applied to the manual, single and multi-model annotations, and to the single and multi-model annotations after removal of all vertebrae that were classified as inaccurate by the error classifier. This reflects the use of the error estimation as a component of a CAD system, identifying potentially inaccurate classifications for manual checking and correction. The threshold used for error classification was set to the operating point that gave 10% false positive rate in Fig. 7(a). When combined with error classification, single-model fracture classification was more accurate than multi-model classification without error classification, and multi-model fracture classification was more accurate than classification based on manual annotations. At an operating point of 90% sensitivity in the filtered, multi-model ROC curve, fewer than 20% of the vertebrae were labeled for manual inspection and only 5.65% of the fractured vertebrae were misclassified both as normal and accurate i.e. 94.35% of fractured vertebrae were either correctly classified or identified as inaccurate.

4 Conclusion

The use of shape and appearance models to segment structures in clinical images is well established and has been proposed as the basis for clinical decision support systems for a number of musculoskeletal pathologies. However, these systems rarely achieve human-level accuracy. Reliable estimates of the errors on the results would significantly increase their practical utility by highlighting the images or structures requiring human input. However, this requires error estimation techniques sensitive not only to random errors but also to systematic errors such as model fitting failures.

This work has demonstrated the use of multiple, independent sub-models as a route to estimation of systematic errors on appearance model fitting. The underlying approach is not novel but we believe that this is the first time it has been applied to appearance models. Using vertebral body segmentation and

osteoporotic VF classification in DXA images as an example, the approach was shown to be as accurate as an RF regressor in estimating random errors, but significantly more accurate in estimating systematic errors. The use of multiple sub-models also resulted in improvements in annotation accuracy by allowing more efficient use of large training sets. The combination of these effects allowed multi-model VF classification based on 6-point morphometry with error filtering to exceed the accuracy of classification from manual annotations whilst rejecting fewer than 20% of the vertebral segmentations, implying that it could have practical utility in appearance model based CAD systems.

The work described here acts as a proof-of-concept but is preliminary. For example, we have not explored the variations in annotation and error estimation accuracy with varying numbers of sub-models. More significantly, the definition of independence of the sub-models was not explored. Some degree of independence was ensured by using independent training sets for each model. However, a true definition of independence would require that each model existed on a separate sub-space of the shape and appearance space. Independence might therefore be maximized by permuting the assignment of training samples to models to maximize the distances between the sub-spaces as measured using the Grassmanian. In the case of spinal images, constraints would be required to ensure this did not separate training samples by vertebral level and produce sub-models that could not fit the whole spine. We intend to explore this in future work.

Acknowledgments. This publication presents independent research supported by the NIHR Invention for Innovation (i4i) programme (grant no. II-LB_0216-20009). The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care. The authors acknowledge the invaluable assistance of Mrs Chrissie Alsop, Mr Stephen Capener, Mrs Imelda Hodgkinson, Mr Michael Machin, and Mrs Sue Roberts, who performed the manual annotations.

References

1. Adams, J.E.: Opportunistic identification of vertebral fractures. J. Clin. Densitom. **19**(1), 54–62 (2016)
2. Barlow, R.: Statistics: A Guide to the Use of Statistical Methods in the Physical Sciences. Wiley, Hoboken (1989)
3. Bromiley, P.A., Adams, J.E., Cootes, T.F.: Localisation of vertebrae on DXA images using constrained local models with random forest regression voting. In: Yao, J., Glocker, B., Klinder, T., Li, S. (eds.) Recent Advances in Computational Methods and Clinical Applications for Spine Imaging. LNCVB, vol. 20, pp. 159–171. Springer, Heidelberg (2015). https://doi.org/10.1007/978-3-319-14148-0_14
4. Bromiley, P.A., Kariki, E.P., Adams, J.E., Cootes, T.F.: Fully automatic localisation of vertebrae in CT images using random forest regression voting. In: Yao, J., Vrtovec, T., Zheng, G., Frangi, A., Glocker, B., Li, S. (eds.) CSI 2016. LNCS, vol. 10182, pp. 51–63. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-55050-3_5

5. Bromiley, P.A., Pokric, M., Thacker, N.A.: Empirical evaluation of covariance estimates for mutual information coregistration. In: Barillot, C., Haynor, D.R., Hellier, P. (eds.) MICCAI 2004. LNCS, vol. 3216, pp. 607–614. Springer, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30135-6_74
6. Bromiley, P.A., Schunke, A.C., Ragheb, H., Thacker, N.A., Tautz, D.: Semi-automatic landmark point annotation for geometric morphometrics. Front. Zool. **11**(61), 1–21 (2014)
7. Burge, R., Dawson-Hughes, B., Solomon, D.H., Wong, J.B., King, A., Tosteson, A.: Incidence and economic burden of osteoporosis-related fractures in the United States 2005–2025. J. Bone Miner. Res. **22**, 465–475 (2007)
8. Delmas, P.D., et al.: Underdiagnosis of vertebral fractures is a worldwide problem: the IMPACT study. J. Bone Miner. Res. **20**(4), 557–563 (2005)
9. Erdt, M., Steger, S., Wesarg, S.: Deformable registration of MR images using a hierarchical patch based approach with a normalized metric quality measure. In: 2012 9th IEEE International Symposium on Biomedical Imaging (ISBI), pp. 1347–1350 (2012)
10. Genant, H.K., Wu, C.Y., Kuijk, C.V., Nevitt, M.C.: Vertebral fracture assessment using a semi-quantitative technique. J. Bone Miner. Res. **8**(9), 1137–1148 (1993)
11. Jergas, M., Valentin, R.S.: Techniques for the assessment of vertebral dimensions in quantitative morphometry. In: Genant, H.K., Jergas, M., van Juijk, C. (eds.) Vertebral Fracture In Osteoporosis, pp. 163–188. University of California Osteoporosis Research Group, San Francisco (1995)
12. Kariki, E.P., Bromiley, P.A., Cootes, T.F., Adams, J.A.: Opportunistic identification of vertebral fractures on computed radiography: need for improvement. Osteoporos. Int. **27**(S2), 621 (2016)
13. Lindner, C., Bromiley, P.A., Ionita, M., Cootes, T.F.: Robust and accurate shape model matching using random forest regression-voting. IEEE TPAMI **37**(9), 1862–1874 (2015)
14. Rachner, T.D., Khosla, S., Hofbauer, L.C.: Osteoporosis: now and the future. Lancet **377**(9773), 1276–1287 (2011)
15. Söhn, M., et al.: Model-independent, multimodality deformable image registration by local matching of anatomical features and minimization of elastic energy. Med. Phys. **35**(3), 866–878 (2008)