Learning-based shape model matching: training accurate models with minimal manual input

Claudia Lindner, Jessie Thomson, The arcOGEN Consortium, and Tim F. Cootes

Centre for Imaging Sciences, University of Manchester, U.K.

Abstract. Recent work has shown that statistical model-based methods lead to accurate and robust results when applied to the segmentation of bone shapes from radiographs. To achieve good performance, model-based matching systems require large numbers of annotations, which can be very time-consuming to obtain. Non-rigid registration can be applied to unlabelled images to obtain correspondences from which models can be built. However, such models are rarely as effective as those built from careful manual annotations, and the accuracy of the registration is hard to measure. In this paper, we show that small numbers of manually annotated points can be used to guide the registration, leading to significant improvements in performance of the resulting model matching system, and achieving results close to those of a model built from dense manual annotations. Placing such sparse points manually is much less time-consuming than a full dense annotation, allowing good models to be built for new bone shapes more quickly than before. We describe detailed experiments on varying the number of sparse points, and demonstrate that manually annotating fewer than 30% of the points is sufficient to create robust and accurate models for segmenting hip and knee bones in radiographs. The proposed method includes a very effective and novel way of estimating registration accuracy in the absence of ground truth.

Keywords: Random Forest regression-voting, bone segmentation, radiographs, statistical shape models, machine learning

1 Introduction

There are many research questions which can be answered by analysing the large databases of tens of thousands of medical images which are now becoming available, both from large studies and from growing electronic archives of clinical data. A key first step in such analysis is often locating the outlines of the structures of interest. Recently, it has been shown that robust and accurate annotations can be automatically obtained using shape-based model matching algorithms [2, 4, 8, 9]. Unfortunately, building such models requires accurate annotation of large numbers of points on several hundred images. This creates a significant bottleneck, hampering the ability to analyse large datasets efficiently.

This paper addresses the problem of building effective shape model matching systems from large sets of images with as little manual intervention as possible.

We use a pragmatic, but effective, approach in which we use a *small number* of manually annotated points on each image to initialise a dense groupwise non-rigid registration (GNR) [10] to establish correspondences across the set of images, allowing us to propagate a dense annotation from one image to all the rest and then to use these to build a detailed model to be matched to new images.

One option to also consider for the generation of dense annotations is fully automatic GNR. However, while this can work well for some reasonably homogeneous datasets, problems often occur where images have been gathered from multiple sources. This tends to be the case in large epidemiological studies where images will have been collected retrospectively, from various centres and without consistent imaging protocols in place. Furthermore, even when the registration works well, models built from such automated correspondences are generally less effective than those built from careful manual annotations. This is because the registration is likely to smooth out details and fail to match to unusual variations. Fully automatic GNR often works well on examples close to the average but performs poorly on outliers. Many failures in the registration are one of two types, (a) gross failures where the registration has converged to the wrong place completely, or (b) localised failures where one part of the object has been poorly matched. Both types can be substantially mitigated if the user supplies a small number of manually annotated landmarks, integrating human expertise into the annotation procedure (see e. g. [6, 11]).

Although registration techniques are widely used to establish correspondences and build shape models (particularly in 3D data) [1, 3, 5], we make a number of key contributions. We show on the challenging, but representative, datasets that we examine that *(i)* model matching systems built from GNR initialised only with the correct initial pose (from two manually annotated points) perform significantly worse than those built from dense manual annotations, particularly on the harder images; *(ii)* the robustness of the resulting model matching system improves monotonically with the number of additional manual points used for initialising the GNR; *(iii)* only small numbers of manually annotated points are required to achieve model matching results very similar to those of the best dense manual models; and *(iv)* we propose a novel method of estimating GNR accuracy in the absence of ground truth annotations.

Some of the most robust and accurate shape model matching results have been achieved using Random Forest regression-voting in the Constrained Local Model (RFRV-CLM) framework [7]. It was shown that RFRV-CLMs can successfully be applied as part of a fully automatic segmentation system to accurately and robustly segment bone shapes in 2D radiographs [8, 9]. In this paper, we explore the effect of different annotation schemes on RFRV-CLM performance.

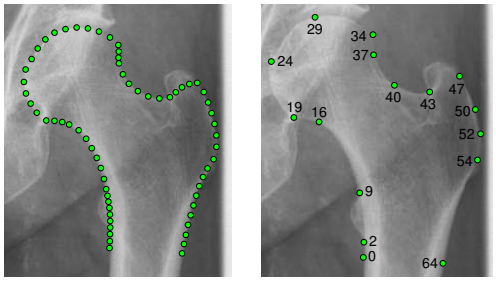
2 Methods

2.1 Groupwise non-rigid registration

The aim of GNR is to identify dense correspondences across images. We use a coarse-to-fine approach similar to that suggested in [3]. At each stage, the

deformation field is represented by the location of the nodes of a triangulated mesh, and piece-wise affine interpolation is used to estimate the correspondence of any internal point. Early stages use coarse meshes with few points. Later stages use denser point sampling. Each stage involves two steps: 1) Build a GNR-model by warping the target images into the reference frame using the current points; and 2) Optimise the positions of the points on each image in turn to better match to the current GNR-model.

The process is usually initialised with an affine registration. However, if some manual annotations are available, they can be used to give a better start. In the following, we use a Thin Plate Spline to deform a reference mesh (defined on one image) to each of the other images, guided by the available manual sparse annotations. This gives a better initialisation for the GNR process. A dense annotation of a bone in one image, as in Fig. 1(a), is propagated to all other images using the deformated meshes from the GNR, and the resulting dense points are used to build an RFRV-CLM.



(a) (b)

Fig. 1. Points used for a model of the femur using 65 points: (a) manual ground truth, and (b) sparse subset points.

2.2 RF regression-voting in the CLM framework

Recent work has shown that one of the most effective approaches to matching a statistical shape model to a new image is to train RFs to vote for the likely position of each model point and then to find the shape model parameters which optimise the total votes over all point positions. Full details are given in [7].

3 Experiments and evaluation

We performed a set of experiments to analyse GNR accuracy and the performance of RFRV-CLMs built on annotations generated via GNR, as we vary the number of manually annotated points used to initialise the GNR. Extensive experiments were run on the segmentation of the proximal femur from pelvic radiographs by outlining its contour using a dense annotation of 65 points as shown in Fig. 1(a). We also provide results for a subset of experiments performed on the segmentation of the proximal tibia from knee radiographs using 37 points.

Datasets: The hip dataset comprised anteroposterior (AP) pelvic radiographs of 420 subjects suffering from unilateral hip osteoarthritis. All images were provided by the arcOGEN Consortium and were collected under relevant ethical approvals. The knee dataset comprised AP knee radiographs of 500 subjects from the Multicenter Osteoarthritis Study and the Osteoarthritis Initiative public use datasets.

For each test, we performed two-fold cross-validation experiments, averaging results from training on each half of the data and testing on the other half. All evaluations are based on manually annotated ground truth. The performance is summarised by the cumulative distribution function (CDF) for the the point-to-curve error (i. e. the shortest distance between each model point and the curve through the manual points, averaged over all points per image).

3.1 Guided groupwise non-rigid registration and refinement

On the hip dataset, we first investigate the effect of using increasing numbers of manual landmark points to initialise the GNR. We assume that between 2 and 16 points are available on each image. These are used to initialise the mesh used for the registration. The registration involves six stages, each refining the result of the previous stage. On completion of the GNR, the 65 points from one image are projected to all the others using the resulting deformation fields.[1]

We performed experiments for subsets of 2 (sparse2: 29,54), 4 (sparse4: sparse2 + 2,19), 6 (sparse6: sparse4 + 34,47), 8 (sparse8: sparse6 + 9,43), 10 (sparse10: sparse8 + 24,37), and 12 (sparse12: sparse10 + 0,64) 14 (sparse14: sparse12 + 16,50), and 16 (sparse16: sparse14 + 40,52) points. Fig. 1(b) shows the positions of all sparse points. The first two points (sparse2), 29 and 54, were chosen to define the location, size and orientation of the proximal femur. Subsequent pairs of points were chosen so as to sparsely capture the outline of the structure and the main areas of variation. This was assisted by identifying which regions were least well registered using the current set of points, qualitatively evaluating the "crispness" of the contours of the mean reference image (Fig. 2).

The quantitative GNR results in Fig. 3(a) show that with only two points for initialisation, GNR performs fairly poorly, with a median error of over 2mm and significant numbers of images with large errors. Thus, the 2-point GNR results are unlikely to be good enough for building an effective shape model matching system. As more points are added the performance gets better, particularly with the dramatic reduction in outliers. Gains beyond 12 points are marginal.

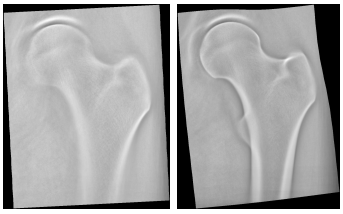
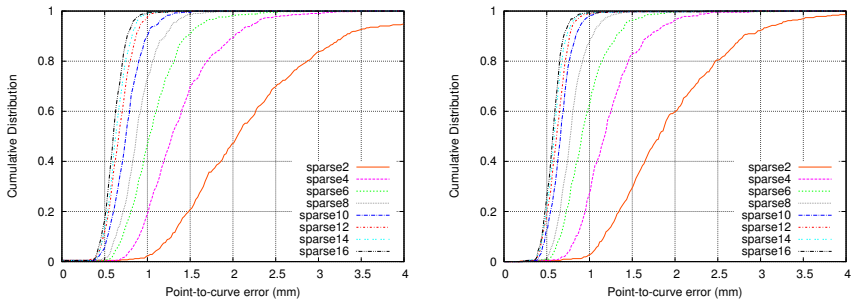


Fig. 2. Mean reference images after GNR for one of the two proximal femur cross-validation sets containing 210 images, for sparse2 and sparse12.

The results of the GNR and landmark propagation are a set of 65 points placed on every image. These can be used to train a two-stage RFRV-CLM; we

[1] In these experiments, all points were propagated from one of the training images. Better results are achievable when propagating the annotations from the reference image as this will be the best "match" to each of the images. This was not done here so as to have consistent points across the subset experiments (the reference image changes for each subset) and because for the smaller subsets the reference image was not good enough to reliably place points in all areas (e.g. sparse2 and the medial femoral head area, see Fig. 2).



(a) Groupwise registration result (GNR) (b) GNR + RFRV-CLM

Fig. 3. Registration accuracy on the proximal femur training set compared to manual ground truth annotations: (a) without, and (b) with additional RFRV-CLM search.

used the same model parameters as those used in [8]. Fig. 3(b) shows the results of applying this RFRV-CLM to the *training set* itself, initialising the search with the sparse points used for GNR and comparing the search results with the ground truth labels. This demonstrates that, in every case, refining the points resulting from GNR using the RFRV-CLM actually leads to improvements in accuracy; though the improvement is largest for the cases with fewer landmarks for initialisation. Our experiments also showed that training another RFRV-CLM on these results led to no further improvement in accuracy (data not shown).

3.2 Estimation of GNR accuracy in the absence of ground truth

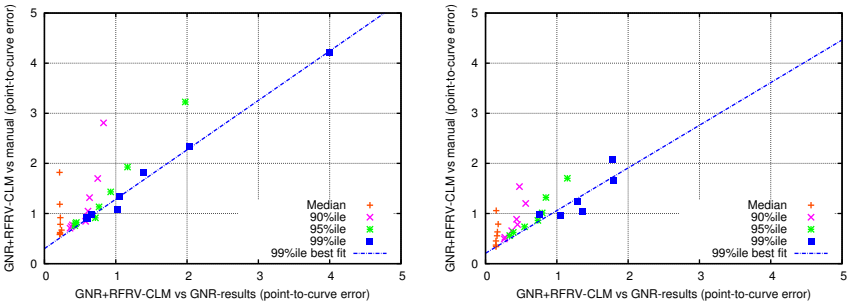
In the above, we trained on points which we know to be incorrect, because the registration is not accurate. In order to evaluate the performance of the registration, we also compared the above search results of the GNR-based RFRV-CLM with the dense points generated from the GNR stage.

Fig. 4(a) shows the quantitative evaluation of the performance of the registration by comparing the performance of the RFRV-CLM on the manual ground truth with its performance on the points generated by GNR (i. e. the points the RFRV-CLM was trained on), for sparse2 to sparse16 as above. Here the line gives the best fit to the 99%ile results (blue squares), demonstrating a strong correlation. This suggests that the performance of the GNR-based RFRV-CLM on the training set (when compared to the GNR results) can be used to estimate the overall accuracy of the registration for various subsets *in the absence of dense ground truth annotations*.

We obtain similar results when repeating the experiments on segmenting the proximal tibia in knee radiographs as shown in Fig. 4(b).

3.3 Performance of GNR-trained RFRV-CLM on new images

To evaluate the segmentation performance of the RFRV-CLMs trained on the results of the GNR, we use them to search *unseen* test images. We incorporate



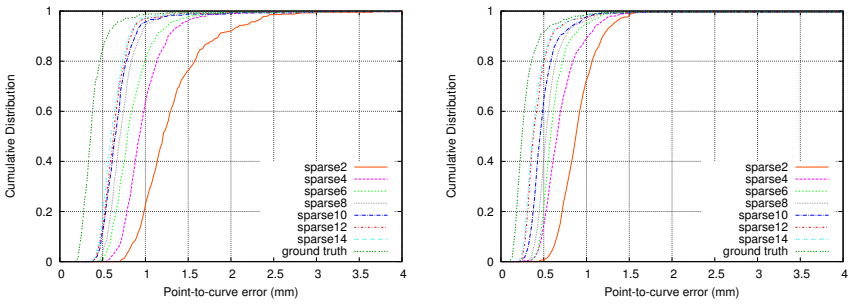
(a) Proximal femur (b) Proximal tibia

Fig. 4. Comparison of the GNR + RFRV-CLM search results vs manual ground truth with the GNR + RFRV-CLM search results vs the points generated by GNR (one marker per sparse subset per evaluated percentile). The strong correlation provides a means for the estimation of GNR accuracy in the absence of ground truth annotations.

the RFRV-CLMs built on the GNR results into a fully automatic shape model matching system, which uses a Hough Forest to perform a global search of the whole image, followed by local refinement with the RFRV-CLM (similar to [8]).

Fig. 5(a) shows the results of running the hip system over the test images, using different training sets. The "ground truth" curve corresponds to the results of a model trained from the full set of 65 manually placed points. The results demonstrate that the fully automatic system works well, and that the results are strongly effected by the quality of the annotations on the training set. As more manual landmarks are added to initialise the GNR of the training set, the performance of the resulting model matching system improves, though adding more points beyond 12 leads to only very small further improvements. In terms of robustness, the RFRV-CLMs trained on the GNR annotations are similar to the models built on the manual ground truth annotations for subsets with at least 10 points, achieving sub-millimetre accuracy for at least 95% of all images. However, there remains a small (0.2mm) gap between the best median result from the GNR-based models and the dense manual model. This is partially explained by the choice of reference image used for propagating the dense points based on the GNR results. Annotating the mean reference image instead of one of the training images leads to a considerable improvement in performance (data not shown). For example, for sparse12 this achieves a point-to-curve error of within 1.1mm for 99% of all images and a median accuracy of less than 0.5mm, reducing the gap between medians to about 0.1mm. Though this does not quite reach the performance of the fully automatic system trained on the dense manual ground truth, the robustness of the two systems is comparable and the GNR-based models require significantly less manual input.

Fig. 5(b) gives the results of the fully automatic tibia segmentation system tested on *unseen* images. Building a fully automatic system based on just 10 manually annotated points achieved a point-to-curve-error of within 1.1mm for 99% of all images and a median accuracy of less than 0.5mm.



(a) Proximal femur (b) Proximal tibia

Fig. 5. Search results of GNR-trained fully automatic RFRV-CLM model matching systems on unseen test images compared to manual ground truth annotations.

4 Discussion and conclusions

We have shown that it is possible to use guided groupwise non-rigid registration to build segmentation models from large sets of images which are almost (but not quite) as effective as those built from careful dense manual annotations.

For segmenting the bone contours of the proximal femur, annotating 12 points on each image reduces the workload by over 80% but leads to a system whose median point-to-curve accuracy is only about 0.1mm worse than that of one built from dense 65-points manual annotations per image, and is almost as robust (both achieve over 98% of results with less than 1mm error). It should be noted that because the 12 points were chosen as well-defined landmarks, they are easier for a human to place than the intermediate points along a boundary – so in practice this will reduce the manual annotation time by even more than 80%. Applying the same approach to segmenting the contours of the proximal tibia shows that 10 points are sufficient to train a system that achieves a median accuracy of less than 0.5mm and that is as robust (99%ile $\leq$ 1.1mm) as one built from dense 37-points manual annotations, reducing the workload by over 70%.

Given a set of accurately annotated sparse points, this approach is equally applicable to the training of 3D models (e. g. using GNR methods as in [3]). Since the sparse points are used only to initialise the registration, small annotation errors in the sparse subsets are likely to be corrected for during the GNR process.

Fig. 4 shows a strong correlation between the ability of the RFRV-CLM to relocate the training points and the accuracy of those points compared to the ground truth. This holds the tantalising prospect of allowing an estimate of the accuracy of groupwise registration on new data in the absence of dense ground truth annotations. Furthermore, it provides an estimate of the improvement in accuracy by adding additional sparse points during GNR and hence is supportive in deciding on whether adding additional sparse points are likely to lead to significant improvements in overall accuracy.

For very large datasets (several hundreds to thousands of images) a cascaded approach might be followed to further minimise the manual input required: train

a RFRV-CLM based on sparse annotations for a subset of all images and then use this to find the sparse annotations in all remaining images to initialise the GNR. Further work is needed to automate the selection of *suitable* sparse points to be used for the GNR initialisation, and to investigate the impact of a cascaded approach on performance. Overall this method will make it much easier to build effective models of different bone shapes from large datasets. This opens up opportunities to use bone shape analyses in more medical studies, allowing analysis of the change in shape across populations and of links between shape and disease.

Acknowledgments. arcOGEN is funded by Arthritis Research UK, J.T. by the Manchester Musculoskeletal Biomedical Research Unit (NIHR) and C.L. by the Engineering and Physical Sciences Research Council, UK (EP/M012611/1).

References

1. Baker, S., Matthews, I., J.Schneider: Automatic construction of active appearance models as an image coding problem. IEEE TPAMI 26(10), 1380–84 (2004)
2. Chen, C., Xie, W., Franke, J., Grutzner, P., Nolte, L.P., Zheng, G.: Automatic X-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements. Medical Image Analysis 18(3), 487–499 (2014)
3. Cootes, T., Twining, C., Petrovic, V., Babalola, K., Taylor, C.: Computing accurate correspondences across groups of images. IEEE TPAMI 32(11), 1994–2005 (2010)
4. Donner, R., Menze, B., Bischof, H., Langs, G.: Fast anatomical structure localization using top-down image patch regression. In: Menze, B., Langs, G., Lu, L., Montillo, A., Tu, Z., Criminisi, A. (eds.) MICCAI 2012 - Workshop MCV. LNCS, vol. 7766, pp. 133–141. Springer (2013)
5. Frangi, A., Rueckert, D., Schnabel, J., Niessen, W.: Automatic 3D ASM construction via atlas-based landmarking and volumetric elastic registration. In: Insana, M., Leahy, R. (eds.) IPMI 2001. LNCS, vol. 2082, pp. 78–91. Springer (2001)
6. Langs, G., Donner, R., Peloschek, P., Bischof, H.: Robust Autonomous Model Learning from 2D and 3D Data Sets. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part I. LNCS, vol. 4791, pp. 968–976 (2007)
7. Lindner, C., Bromiley, P., Ionita, M., Cootes, T.: Robust and Accurate Shape Model Matching using Random Forest Regression-Voting. IEEE TPAMI (2014), http://dx.doi.org/10.1109/TPAMI.2014.2382106
8. Lindner, C., Thiagarajah, S., Wilkinson, M., The arcOGEN Consortium, Wallis, G., Cootes, T.: Fully Automatic Segmentation of the Proximal Femur Using Random Forest Regression Voting. IEEE TMI 32(8), 1462–1472 (2013)
9. Lindner, C., Thiagarajah, S., Wilkinson, M., The arcOGEN Consortium, Wallis, G., Cootes, T.: Accurate bone segmentation in 2D radiographs using fully automatic shape model matching based on regression-voting. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part II. LNCS, vol. 8150, pp. 181–189. Springer (2013)
10. Sotiras, A., Davatzikos, C., Paragios, N.: Deformable Medical Image Registration: A Survey. IEEE TMI 32(7), 1153–1190 (2013)
11. Zhang, P., Adeshina, S., Cootes, T.: Automatic learning sparse correspondences for initialising groupwise registration. In: Jiang, T., Navab, N., Pluim, J., Viergever, M. (eds.) MICCAI 2010, Part II. LNCS, vol. 6362, pp. 635–642 (2010)