Landmark Localisation in Radiographs Using Weighted Heatmap Displacement Voting

Adrian K. Davison[1(✉)], Claudia Lindner[1], Daniel C. Perry[2,3], Weisang Luo[3], Medical Student Annotation Collaborative[2,3], and Timothy F. Cootes[1]

[1] The University of Manchester, Centre for Imaging Sciences, UK
[2] University of Liverpool, UK
[3] Alder Hey Children's Hospital, UK
adrian.davison@manchester.ac.uk

Abstract. We propose a new method for fully automatic landmark localisation using Convolutional Neural Networks (CNNs). Training a CNN to estimate a Gaussian response ("heatmap") around each target point is known to be effective for this task. We show that better results can be obtained by training a CNN to predict the offset to the target point at every location, then using these predictions to vote for the point position. We show the advantages of the approach, including those of using a novel loss function and weighting scheme. We evaluate on a dataset of radiographs of child hips, including both normal and severely diseased cases. We show the effect of varying the training set size. Our results show significant improvements in accuracy and robustness for the proposed method compared to a standard heatmap prediction approach and comparable results with a traditional Random Forest method.

Keywords: Perthes disease, X-rays, paediatrics, convolutional neural network (CNN), fully convolutional network (FCN), deep learning, voting.

1 Introduction

Locating landmarks on medical images is an important first step in many analysis tasks, particularly those requiring geometric measurements of the shape of structures. Many methods have been proposed for this task, with some of the most effective using random forest regression-voting (RFRV) [1, 2] and, more recently, deep learning approaches [3–6].

Deep learning has been a popular method to extract information for classification, recognition and regression tasks. In various fields, convolutional neural networks (CNNs) have become the state-of-the-art, out-performing many traditional machine learning methods. For landmark localisation, including detecting anatomical landmarks in medical images [7, 8] and human pose estimation [4, 9], an effective technique has been to apply a CNN to estimate a new image with a Gaussian blob around each predicted landmark position (a so-called "heatmap").

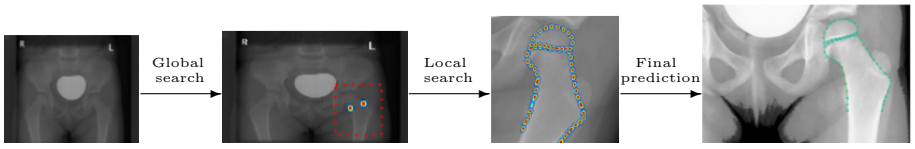


Fig. 1. Overview of our landmark localisation method in child hip radiographs: (a) A full pelvic radiograph. (b) The global searcher network locates two reference points to estimate the pose of the proximal femur. (c) A patch containing the approximately located femur is fed into the local search network to vote for the position of each landmark. (d) The fully automatically predicted landmark positions.

This has been shown to yield better results than directly regressing the landmark locations which tend to have a highly non-linear relationship to features [10].

We propose a novel voting-based scheme to identify landmark locations. We train a fully convolutional NN to estimate the displacement of every pixel from each target landmark, together with an associated weight. These displacements can then be used to vote for the landmark location, integrating information from the local area. We propose a novel loss function to train the CNN for this task, which focuses attention on the target regions. The combination of regressing the pixel offsets and heatmap weights adds further novelty to the approach.

We evaluate the proposed weighted heatmap displacement voting (WHDV) approach on the challenging problem of locating the outline of normal and badly diseased proximal femurs in radiographs of children, showing that WHDV significantly improves both accuracy and robustness compared to a standard heatmap prediction approach. We also show how the performance varies as the number of training examples increases. The overall pipeline can be seen in Fig. 1.

This paper makes three contributions: (i) We describe a novel method of landmark location which improves upon the widely used "heatmap" approach; (ii) we describe extensive experiments characterising the performance of the system as the size of the training set increases. This includes a detailed comparison with random forest regression-voting constrained local models (RFRV-CLMs) demonstrating that unless large numbers of examples are available the latter are to be preferred to CNN approaches; (iii) we demonstrate an automatic system for locating the outline of both normal and diseased femurs, showing that shape model-based systems can deal with considerable abnormalities in this case.

2 Related Work

Pfister et al. [4] used a CNN to regress heatmaps for each point, and dense optical flow to warp landmark positions onto videos for human pose estimation. The paper is one of the earliest to regress heatmaps through a deep network and to combine the results with an implicit spatial model.

To detect multiple landmarks on two-dimensional (2D) radiographs and three-dimensional (3D) magnetic resonance imaging (MRI) images of hands, Payer et al. [7] proposed a novel CNN (named SpatialConfiguration-Net) that was trained end-to-end to detect 37 landmarks in the radiographs and 28 in the MRI images. The new architecture could learn local features and imposed constraints on the spatial configuration of landmarks.

Bulat and Tzimiropoulos [9] proposed a CNN cascaded architecture that consisted of two components: a part detection network for detecting human body parts and a deep regression subnetwork that was able to regress the landmark locations using heatmaps, regardless of whether they were occluded or not.

Using the challenging COCO dataset for detecting keypoints, Papandreou et al. [11] used an RCNN detector to find people and estimate keypoints on each using heatmaps and offsets using a fully convolutional ResNet [12]. Both outputs were combined with a novel aggregation function to obtain localised keypoint predictions.

Belagiannis and Zisserman [6] estimated 2D human poses using a CNN with a recurrent module that combined intermediate feature representations to learn the image context and improve the final heatmap predictions in challenging datasets, including those classed as "in-the-wild".

Rather than using heatmaps, the relative position of landmarks can be predicted directly. The majority of such work has focused on medical images. Chen et al. [3] estimated displacements from randomly chosen patches to unknown landmark positions. These patches then voted on the final landmark position. The overall shape was regularised with a statistical shape model.

Aubert et al. [5] used a simple CNN to predict the 3D landmark of vertebral centres. The training used frontal and lateral hip patches to estimate the 2D displacement in the x plane for the frontal and lateral view and for the overall displacement in the y plane. The 3D landmark was determined using epipolar geometry.

Sofka et al. [13] used a fully convolutional network (FCN) to regress point locations. They created a center of mass layer that computed the mean position of the network prediction output. This had an advantage over direct heatmap regression as it could predict subpixel values and the objective function could penalise measurement length differences from the ground truth for their task. This differs from our approach as we calculate the landmark positions outside of the network (with a voting scheme) and we do not need a separate layer to specifically do this task.

Using limited medical image training data, Zhang et al. [8] extracted millions of images patches to be fed into a two-stage convolutional network that first output the predicted displacement vectors, and then directly predicted 1200 landmarks in 3D MRI brain scans and 7 landmarks from 3D tomography images of prostates.

Less common is a combination of heatmaps and displacements. Zhang et al. [14] proposed the use of displacement maps to explicitly model the spatial context information of cone-beam computed tomography scans. They used the

estimated displacement maps from the previous step as a guide to introduce a joint learning framework for bone segmentation and landmark localisation. The heatmaps were regressed in the second stage as the ground truth landmark areas.

3 Fully Convolutional Network with Global and Local Searchers

Our fully automated method has two stages: (i) a global search over the whole image for two reference points on the target object, which then define its position, orientation and scale; (ii) a local search in a region defined by these reference points to find n landmark points on the object. Both global and local search use the same approach to identify point positions.

We use two separate search stages as a full pelvic X-ray contains many similar features, especially when it comes to the opposite hip. The global search aims to find the position of the left-anatomical femur to then improve the local search performance. Using two reference points to crop the region of interest, in this case the femur, is an established technique to reduce the search area of a potentially cluttered radiograph [1]. To summarise the differences between the global and local searcher: the global searcher scans the whole pelvic X-ray for two key reference points and crops the detected femur; the local searcher uses the cropped image to locate 58 landmark points in a local region of the overall radiograph.

In each case we use a CNN to take the target image (for global search) or sampled region (for local search) and compute a set of output planes for each point. In the original "heatmap" approach one would compute a single image plane for each point. In our modified version we predict three planes per point, an x displacement, a y displacement and a weight plane. We use these to vote for the position of each point and take the maximum response in the accumulated vote image as the final point location.

3.1 Convolutional Network with Weighted Heatmap Loss

We use a modified version of the widely used U-Net architecture [15]. U-Net acts as a convolutional auto-encoder with added skip connections from encoder layers to decoder layers that are on the same level. Our modifications are in line with those in [7], where max pooling is replaced with average pooling and up-convolution layers are replaced with upsampling. Our method is similar to [11] in that it uses heatmaps and displacement vectors, however our approach differs by using a vote from every pixel to determine the landmark rather than using probability of being within a disk surrounding a keypoint. Further, we do not require pre-training and use a computationally simpler network architecture, U-net, over the ResNet-101 [12] pretrained on Imagenet.

Training For each input image (with known landmark positions, (x_p, y_p), $p = 1, \ldots, n$), we constructed three ground truth planes P_x^p, P_y^p, P_w^p as follows:

$$\begin{aligned} P_x^p(i,j) &= t(i - x_p), \\ P_y^p(i,j) &= t(j - y_p), \\ P_w^p(i,j) &= \exp(-|(i,j) - (x_p, y_p)|^2 / 2\sigma^2). \end{aligned} \tag{1}$$

The function $t(x)$ truncates the input to a fixed range:

$$t(x) = \begin{cases} -k & \text{if } x < -k, \\ k & \text{if } x > k, \\ x & \text{otherwise,} \end{cases} \tag{2}$$

where k is the displacement value chosen through empirical experiments. Note that P_w^p is the traditional "heatmap", a Gaussian blob centred on the landmark. P_x and P_y are displacement planes and σ is the standard deviation of the Gaussian function.

We trained the CNN to predict these $3n$ planes for each training image, using a loss function which encourages accurate displacement predictions near the points:

$$LossPerPixel(\hat{P_w^p}, \hat{P_x^p}, \hat{P_y^p}) = P_w^p((P_x^p - \hat{P_x^p})^2 + (P_y^p - \hat{P_y^p})^2)) + (P_w^p - \hat{P_w^p})^2, \tag{3}$$

where $\hat{P_w^p}, \hat{P_x^p}, \hat{P_y^p}$ are the outputs of the network. Note that scaling the first term by P_w^p down-weights the position prediction away from the points, where it is not needed.

Point Localisation To locate points on a new image, we feed the image to the CNN to generate the predicted planes. For each point p we then create a vote image, V_p, by scanning through all pixels (i, j), voting at $(i + \hat{P_x^p}(i,j), j + \hat{P_y^p}(i,j))$. The vote image is then multiplied (pixel-wise) by the weight image $\hat{P_w^p}(i,j)$. We smooth the vote image with a Gaussian with a SD = 4, which was chosen through experiments by changing the SD from 1...6 and choosing the best performing value. The maximum peak of the vote image is used to estimate the point positions.

4 Experiments

We performed a series of experiments to accurately locate landmarks along the proximal femur in radiographs of children's hips. To evaluate the performance of the proposed WHDV approach, we compare with two FCN heatmap-based approaches and a traditional machine learning method: RFRV [1,2].

4.1 Dataset

The dataset consists of 1,696 radiographs of hips from children aged between 2 and 11 years, with some affected by Perthes disease, where the blood supply to the growth plate of the bone at the end of the femur becomes inadequate [16]. This dataset is challenging as the hip is still growing during childhood, meaning the femur has growth areas such as the femoral head and greater trochanter, and because there is significant shape and appearance change due to disease (Fig. 2(b)).

We conducted 3-fold cross-validation experiments for a range of training set sizes, splitting the data into random subsets of 100, 200, 500 and 1000. The test data consists of 500 randomly chosen images (the same set used for all experiments). The test data does not overlap with the training data for any of the subsets. All images have been manually annotated with 58 points by two different people chosen randomly from a pool of ten trained annotators. The ground truth is then created by averaging the point positions between the two annotators.

For the deep learning based approaches, the data was augmented with random rotations (between 5° clockwise and 35° anti-clockwise) once for each image to allow for rotation variants (note that RFRV also includes random rotations as part of the training). The reason for the imbalance in rotation values is that rotating the hip too far clockwise would create an unrealistic pose for a pelvic X-ray.

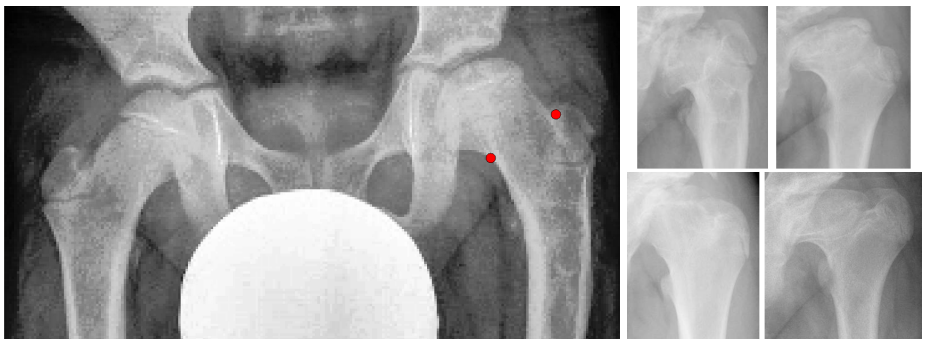
4.2 Network Parameters

Our FCN takes input images of size 256×192 (global search) or 224×224 (local search) and generates $3n$ output planes of the same size as described above. During training, 15% of the training set is used for validation. To ensure the validation set did not use a portion of the training set, we added 15% additional images to the training set. We performed 3-fold cross-validation experiments per method, where the reported results will show the average over all 3 folds.

We chose the Adam [17] optimiser through empirical experiments where all of the available optimisers in Keras (including stochastic gradient descent, Nadam and RMSProp) were tested with the network and the best performing chosen. We used the default parameters suggested in [17], where the learning rate was set to 0.001, the exponential decay rate for the first moment estimates (β_1) was set to 0.9 and the exponential decay rate for the second-moment estimates (β_2) was set to 0.999. To prevent division by zero, ϵ was set to 10^{-7}. The batch size was set to 10 and training was completed using an NVIDIA Titan Xp GPU. We use Keras [18] with a Tensorflow [19] backend.

4.3 Global Search

We focused on detecting the left proximal femur in full pelvic images. Each image was scaled to 192×256 along with 2 ground truth reference points (Fig. 2(a)).



(a) (b)

Fig. 2. (a) Two reference points were chosen to train the global searcher to locate the left proximal femur. Note that the input image is a full pelvic image, adding to the detection difficulty. (b) Sample images from the dataset showing the challenging nature of the diseased proximal femurs.

Each image was fed into the weighted heatmap loss network with the 3 ground truth elements (P_x^p, P_y^p, P_w^p). The network was trained to regress heatmaps and displacements for the two reference points, and landmark voting was applied to estimate their position. The latter was then used to sample the region of interest for the point localisation stage.

The two reference points were used to define the location, scale and orientation of a region of interest around the proximal femur which was sampled into a 224×224 patch. Such patches were used to train the second local search CNN to estimate the position of all 58 points.

4.4 Landmark Localisation Results

We investigated three CNN based methods: (i) The "Heatmap Only" (HO) approach where the network learns a heatmap centred on each landmark, trained using a mean squared error (MSE) loss; (ii) The "Heatmap with Displacement Voting" (HDV) method where we learned displacement and weight planes using an MSE loss; and (iii) the full WHDV approach with novel weighted loss function. The HO approach is based on the standard heatmap generation [10]. We note that other methods based around this, for example stacked hourglass networks [20], use heatmaps with a more sophisticated network structure, however we use the basic form of heatmap regression in this paper.

We report both mean point-to-curve and mean point-to-point errors measured as a percentage of the femoral shaft width defined by the distance between the bottom two landmark points (Fig. 1(d)). For comparison, we include results using the current state-of-the-art approach, a RFRV-CLM [1, 2] which uses random forests with Haar features to vote on the most likely landmark position, constrained using a shape model. We evaluated the accuracy with which

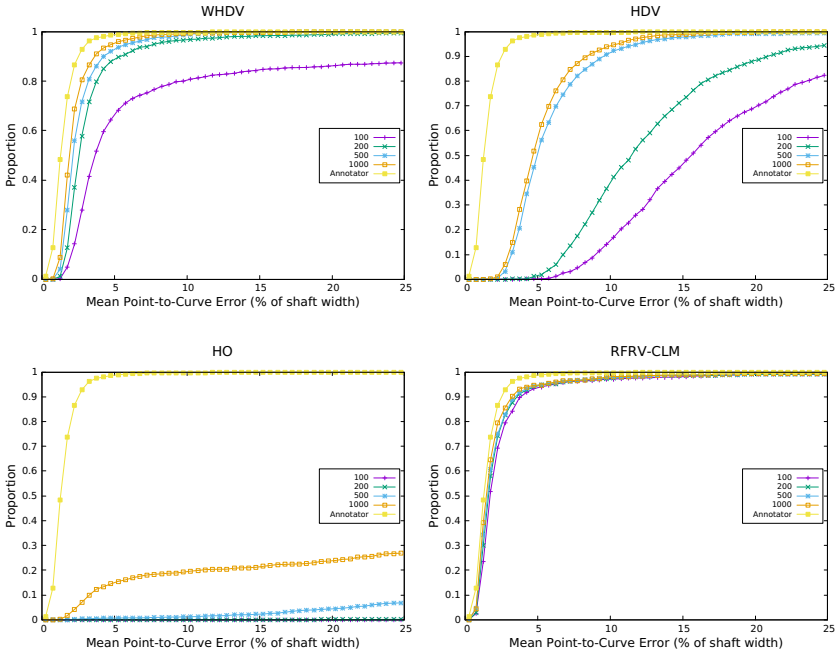


Fig. 3. The cumulative distribution functions of mean point-to-curve error for each method as a function of training set size.

the data was annotated by comparing results of 2 independent annotators on 1,696 images. We compute the average difference of each set of annotated points to the mean of the annotations for each image. This gives an indication of the maximum accuracy that may be achieved given the noise on the annotations (see curves marked "Annotator" on the graphs).

Firstly, we show the cumulative distribution function (CDF) graphs for all methods with the mean point-to-curve and point-to-point error for each training size in Figs. 3 and 4 respectively. For WHDV and HDV, the error is reduced as the training size grows, however WHDV performs better than the other 'heatmap'approaches even with a small training set, suggesting the novel loss function helps to stabilise the error, unlike in the similar HDV method. The increase in training data for the proposed method has a particular impact in the point-to-point error going from a median error of 12.5% in the 100 train set, to 6.71% in the 1000 train set.

In contrast, the RFRV-CLM method performs well for all training set sizes, however, unlike the other methods only shows small increases in performance, suggesting that adding more data would not effect the performance as much as it would in the proposed method. For example, the median error in the 100 train set and the 1000 train set was 6.92% and 5.85% respectively for the RFRV-CLM.

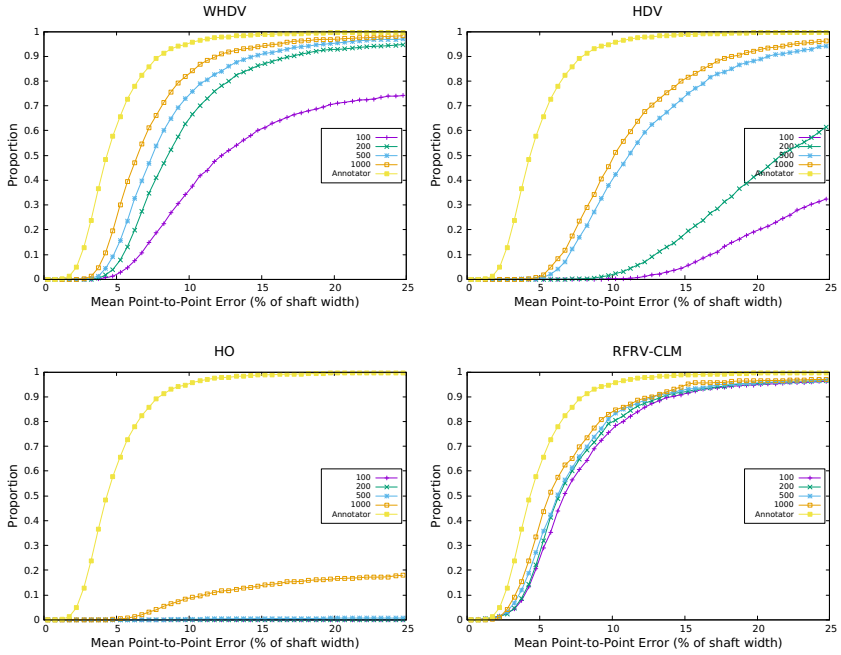


Fig. 4. The cumulative distribution functions of mean point-to-point error for each method as a function of training set size.

A comparison of each method, split into the four training set sizes for both point-to-curve and point-to-point error (Figs. 5 and 6, respectively). These results show that the HO approach performs poorly, regardless of the amount of training data, suggesting that the initial global search fails to locate the hip, which leads to poor performance of the local searcher.

The proposed method is outperformed by RFRV-CLM when trained on only 100 images. However the performance gap closes rapidly as more images are used for training. When trained with 500 examples WHDV outperforms RFRV-CLM significantly in the 99%ile with WHDV achieving a mean point-to-curve error of 10.6% and RFRV-CLM achieving 17.2%. With 1000 images WHDV and RFRV-CLM achieve a mean point-to-curve error of 9.02% and 17.1% respectively. Thus with larger training sets WHDV is more robust (making fewer large errors) than the RFRV-CLM.

5 Conclusion

We have described a novel voting-based heatmap method for training CNNs to identify the position of landmark points. Our results show that the proposed method leads to more accurate and robust results than the commonly used "standard heatmap" [10] approach on a challenging data set. One limitation

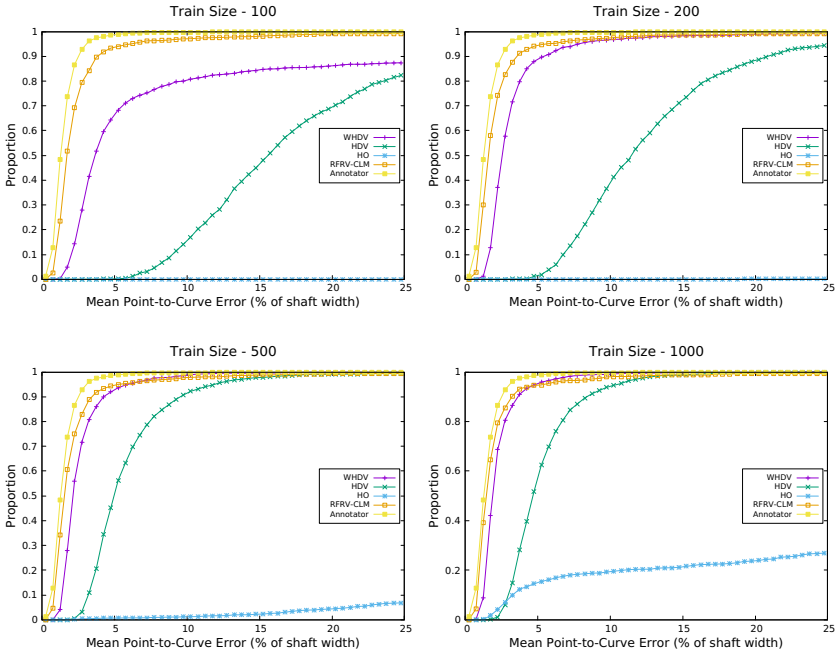


Fig. 5. The cumulative distribution function comparing the performance of each method by the training set size. The mean point-to-curve error is reported.

of the voting approach is that it cannot easily be differentiated. This would prohibit full end-to-end training of any system using this approach as its first stage. We showed extensive experiments in characterising the performance of the system as training set sizes increase, which included a comparison with the RFRV-CLM. The experiments showed that unless large numbers of training data can be used, the latter system is to be preferred over CNN approaches. Finally, we demonstrated an automatic system to locate the outline of both normal and diseased femurs, showing the effectiveness of shape-model systems when presented with considerable abnormalities.

RFRV-CLM is a mature technology and is known to work well even on relatively small datasets. It also has the advantage of constraining the points with an explicit (linear) shape model. However, it can be seen that as training data increases, RFRV-CLM has only modest increases in performance. The proposed WHDV method performs poorly when trained on few examples, but outperforms RFRV-CLM in the upper percentiles of the 500 and 1000 train set sizes. Splitting the data into disease and healthy cases would also be useful, but would require clinical expertise to classify the ground truth. Further work will include acquisition of larger datasets with a good representation of healthy and diseased cases, and more analysis on individual age groups and their affect on performance.

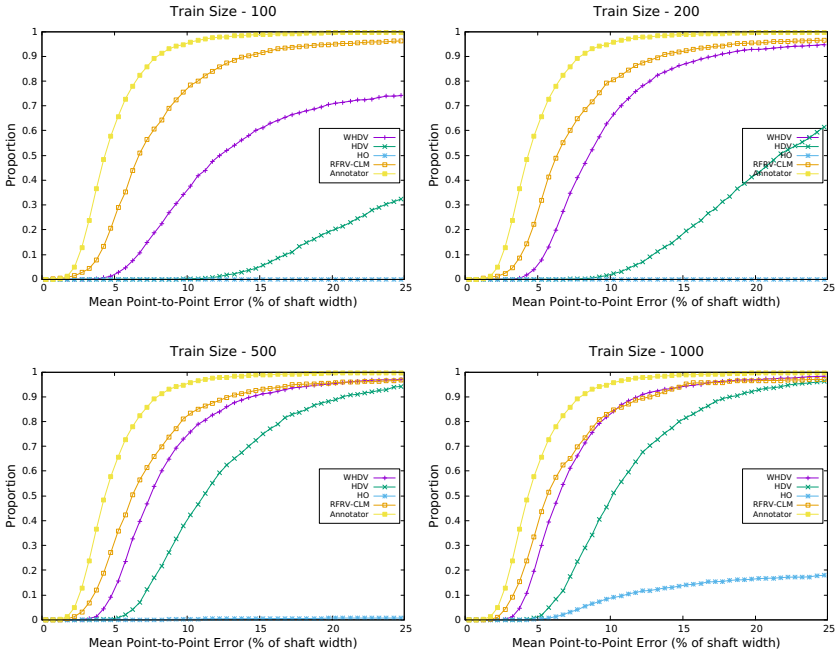


Fig. 6. The cumulative distribution function comparing the performance of each method by the training set size. The mean point-to-point error is reported.

The CNN, being trained on all points at once, should learn an implicit model, but some of the errors it makes suggest that this model may not be generalising as well as the traditional RF approach constrained with a shape model – this is something we continue to explore. We will also evaluate whether fitting a shape model to the voting images produces better results, though examination of the votes in the response images suggests that this might not be the case.

Both WHDV and RFRV-CLM perform well in automatically locating landmark points and the outline of the proximal femurs of children, both in cases with and without disease. When starting a new project of this nature, one will only have a few annotated images at first - the RFRV-CLM is much more suitable for helping annotators when building up the training set. This is the first step in the development of a system to quantify shape changes due to disease and to assist clinicians in the decision making on the best course of treatment.

Acknowledgements. A. K. Davison was funded by Arthritis Research UK as part of the ORCHiD project. C. Lindner was funded by the Engineering and Physical Sciences Research Council, UK (EP/M012611/1) and by the Medical Research Council, UK (MR/S00405X/1). Manual landmark annotations were provided by the Medical Student Annotation Collaborative (Grace Airey, Evan Araia, Aishwarya Avula, Emily

Gargan, Mihika Joshi, Muhammad Khan, Kantida Koysombat, Jason Lee, Sophie Munday and Allen Roby).

References

1. Lindner, C., Thiagarajah, S., Wilkinson, J., The arcOGEN Consortium, Wallis, G., Cootes, T.: Fully automatic segmentation of the proximal femur using random forest regression voting. IEEE Trans. Med. Imaging 32(8), 1462–1472 (2013), https://doi.org/10.1109/TMI.2013.2258030
2. Lindner, C., Bromiley, P., Ionita, M., Cootes, T.: Robust and accurate shape model matching using random forest regression-voting. IEEE Trans. Pattern Anal. Mach. Intell. 37(9), 1862–1874 (2015), https://doi.org/10.1109/TPAMI.2014.2382106
3. Chen, C., Xie, W., Franke, J., Grutzner, P., Nolte, L., Zheng, G.: Automatic x-ray landmark detection and shape segmentation via data-driven joint estimation of image displacements. Med. Image Anal. 18(3), 487–499 (2014), https://doi.org/10.1016/j.media.2014.01.002
4. Pfister, T., Charles, J., Zisserman, A.: Flowing ConvNets for human pose estimation in videos. In: International Conference on Computer Vision – ICCV 2015, pp. 1913–1921. IEEE (2015), https://doi.org/10.1109/ICCV.2015.222
5. Aubert, B., Vidal, P., Parent, S., Cresson, T., Vazquez, C., De Guise, J.: Convolutional neural network and in-painting techniques for the automatic assessment of scoliotic spine surgery from biplanar radiographs. In: Descoteaux, M., et al. (eds.) Proc. 20th International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2017, Lect. Notes Comput. Sc., vol. 10434, pp. 691–699. Springer (2017), https://doi.org/10.1007/978-3-319-66185-8_78
6. Belagiannis, V., Zisserman, A.: Recurrent human pose estimation. In: Proc. 12th International Conference on Automatic Face& Gesture Recognition – FG 2017, pp. 468–475. IEEE (2017), https://doi.org/10.1109/FG.2017.64
7. Payer, C., Štern, D., Bischof, H., Urschler, M.: Regressing heatmaps for multiple landmark localization using CNNs. In: Ourselin, S., et al. (eds.) Proc. 19th International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2016, Lect. Notes Comput. Sc., vol. 9901, pp. 230–238. Springer (2016), https://doi.org/10.1007/978-3-319-46723-8_27
8. Zhang, J., Liu, M., Shen, D.: Detecting anatomical landmarks from limited medical imaging data using two-stage task-oriented deep neural networks. IEEE Trans. Image Process. 26(10), 4753–4764 (2017), https://doi.org/10.1109/TIP.2017.2721106
9. Bulat, A., Tzimiropoulos, G.: Human pose estimation via convolutional part heatmap regression. In: Leibe, B., et al. (eds.) Proc. 14th European Conference on Computer Vision – ECCV 2016, Lect. Notes Comput. Sc., vol. 9911, pp. 717–732. Springer (2016), https://doi.org/10.1007/978-3-319-46478-7_44
10. Tompson, J., Jain, A., LeCun, Y., Bregler, C.: Joint training of a convolutional network and a graphical model for human pose estimation. In: Ghahramani, Z., et al. (eds.) Advances in Neural Information Processing Systems, vol. 27, pp. 1799–1807. NIPS Proceedings (2014)
11. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition – CVPR 2017, pp. 3711–3719. IEEE (2017), https://doi.org/10.1109/CVPR.2017.395

12. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition – CVPR 2016. pp. 770–778. IEEE (2016), https://doi.org/10.1109/CVPR.2016.90
13. Sofka, M., Milletari, F., Jia, J., Rothberg, A.: Fully convolutional regression network for accurate detection of measurement points. In: Cardoso, M., et al. (eds.) Proc. International Workshop on Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support – DLMIA 2017 & ML-CLS 2017, Lect. Notes Comput. Sc., vol. 10553, pp. 258–266. Springer (2017), https://doi.org/10.1007/978-3-319-67558-9_30
14. Zhang, J., Liu, M., Wang, L., Chen, S., Yuan, P., Li, J., Shen, S., Tang, Z., Chen, K., Xia, J., Shen, D.: Joint craniomaxillofacial bone segmentation and landmark digitization by context-guided fully convolutional networks. In: Descoteaux, M., et al. (eds.) Proc. 20th International Conference on Medical Image Computing and Computer Assisted Intervention – MICCAI 2017, Lect. Notes Comput. Sc., vol. 10434, pp. 720–728. Springer (2017), https://doi.org/10.1007/978-3-319-66185-8_81
15. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., et al. (eds.) Proc. 18th International Conference on Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lect. Notes Comput. Sc., vol. 9351, pp. 234–241. Springer (2015), https://doi.org/10.1007/978-3-319-24574-4_28
16. Perry, D., Hall, A.: The epidemiology and etiology of perthes disease. Orthop. Clin. North Am. 42(3), 279–283 (2011), https://doi.org/10.1016/j.ocl.2011.03.002
17. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv:1412.6980 (2014), http://arxiv.org/abs/1412.6980
18. Keras: deep learning for humans (2015), https://github.com/keras-team/keras
19. TensorFlow: large-scale machine learning on heterogeneous systems (2015), https://www.tensorflow.org
20. Newell, A., Yang, K., Deng, J.: Stacked hourglass networks for human pose estimation. In: Leibe, B., et al. (eds.) Proc. 14th European Conference on Computer Vision – ECCV 2016, Lect. Notes Comput. Sc., vol. 9912, pp. 483–499. Springer (2016), https://doi.org/10.1007/978-3-319-46484-8_29