AUTOMATICALLY DIAGNOSING HIP CONDITIONS FROM X-RAYS USING LANDMARK DETECTION

James McCouat [1,2] *Irina Voiculescu* [1] *Siôn Glyn-Jones* [2]

[1] Department of Computer Science, University of Oxford, UK
[2] NDORMS, University of Oxford, UK.

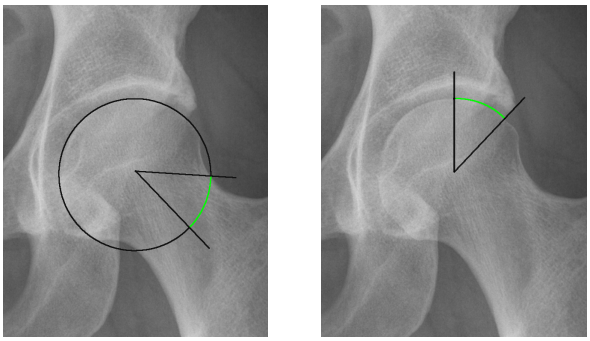
ABSTRACT

When patients present with symptoms of hip pain a clinician might diagnose a condition called femoroacetabular impingement (FAI), where the ball and socket of the hip joint rub together during movement. To diagnose FAI a doctor inspects an x-ray, and records the angles between certain key points in the image. If the angles are 'too big' then FAI is diagnosed. We anticipate that these key points can be located in an x-ray using deep learning and thus the angles measured and FAI diagnosed automatically. In this paper we deploy a stacked hourglass network to automatically locate key-points in hip x-rays, which we then use to automatically diagnose FAI in a patient. On a test set of 112 hips our algorithm diagnoses cam impingement, one of two types of FAI, correctly 90% of the time. To our knowledge this is the first time any kind of FAI has been automatically diagnosed.

***Index Terms*—** Landmark Detection, X-ray, FAI, Deep Learning

1. INTRODUCTION

Femoroacetabular impingement (FAI) is a condition where the ball and socket of the hip joint rub together during movement [1]. The ball is the head of the femur bone and the socket is called the acetabulum, a concave surface of the pelvis where the femur head fits. FAI is an important condition to diagnose because the patient can then receive treatment such as physiotherapy or keyhole surgery to repair the joint. There are two types of FAI: cam impingement, where there is an abnormal bump on the femur head and pincer impingement, where the acetabulum extends too far and encases too much of the femur head thus restricting movement. A cam impingement is diagnosed by measuring what is called the α-angle, the angle between the neck axis of the femur and a point where the femur head stops being spherical [2, 3] as shown in Fig 1a. A pincer impingement is diagnosed by measuring the Lateral Center Edge (LCE) Angle between the vertical axis and the lateral edge of the acetabulum [2] as shown in Fig 1b. In both cases the larger the angle the worse the impingement is. In the case of α-angle a threshold of $> 70^\circ$ for men and $> 61^\circ$ for women has been proposed to diagnose a cam impingement [4]. An LCE-angle threshold of $> 40^\circ$ has been proposed to diagnose pincer impingement [5].



(a) α-angle illustration (b) LCE-angle illustration

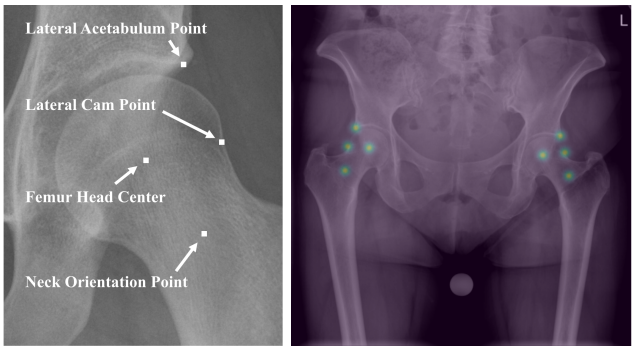
Fig. 1. Annotation performed by a doctor on an anteroposterior (AP) x-ray. For this hip $\alpha \approx 42^\circ$ and LCE $\approx 44^\circ$.

We formulate this as a supervised learning problem. When we compare the angles measured by our algorithm to angles measured by an expert clinician over a test set of 112 hips we find a median difference of 4.00° with an intraclass correlation coefficient (ICC) score 0.76 for α-angle and a median difference of 1.89° with an ICC score 0.93 for LCE-angle. We then threshold the α-angle to automatically diagnose cam impingement. Our diagnosis is the same as an expert clinician's on 90% of the studied hips.

2. METHOD

To measure α-angle for a hip we need to locate the center of the femur head, a point along the neck axis and the point where a circle, plotted to represent the curve of the femur head, intersects the femur head closest to the neck. In this paper we call these points 'femur head center', 'neck orientation point' and 'lateral cam point' respectively. To measure LCE-angle we need to locate the center of the femur head and a point which represents the most lateral point on the acetabulum rim, which we call 'lateral acetabulum point'. In total we locate 4 key-points (landmarks) per hip, shown in Fig 2a.

In our supervised learning approach we make use of



(a) The key-points (b) Heatmaps used for training.

Fig. 2. (a) shows the key-points our algorithm locates for each hip, these are needed to measure α-angle and LCE-angle. (b) shows the heatmaps which are passed into the network to train it, we can see there are bright spots for 8 key-point landmarks. Each landmark has its own heatmap but we squashed 8 heatmaps into one image here for clarity.

antero-posterior (AP) x-rays, each with an associated mask of $2 \times 4 = 8$ key-points. We do not train our deep learning model directly on the key-point masks but instead convert these masks to a set of heatmaps (Fig 2b). We use a mean squared error (MSE) loss function per pixel. At test time, as is common in landmark detection problems [6, 7], we find the hottest point on each heatmap and use that as our key-point.

3. DATASET

The x-ray data was collected at Oxford University as part of FAIT (Femoroacetabular Impingement Trial) [8]. The x-rays were taken at multiple sites in the UK and contain multiple images for the same patient taken at different time points. This is a pathological dataset: for an 18 to 60 year old to be eligible for the trial they must have presented with evidence of FAI. They must not have significant osteoarthritis or hip dysplasia. The dataset comprises 375 AP x-rays (hence 750 hips) taken from 179 different patients. The `dcm` format files were converted to `png`. The x-rays had an average size of 2794×2961 pixels but were downsampled using interpolation to either 512×512 or 256×256 pixels, depending on the the experiment (section 4), before being fed into our models.

3.1. Choosing the Training, Validation and Test Set

We choose a 70:15:15 split for our training, validation and test sets. We made absolutely sure that these sets were as representative of the overall dataset as possible. We choose 'representative' to mean that the distributions of the α-angles of the hips contained in each of these subsets were similar. To do this we produced 10^6 potential subset splits, ensuring that no images from the same patient were in different sets; then a 3–sample Kolmogorov–Smirnov (KS) [9] statistic based on the α-angles contained in each set was calculated, choosing the split which resulted in the lowest KS test score: 0.0458 (Fig 3). The final training set contained 264 images from 127 patients, validation contained 55 images from 28 patients and test contained 56 images from 24 patients.

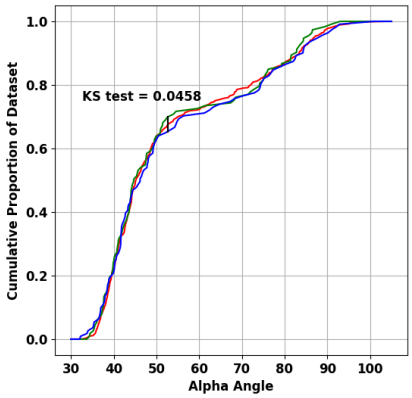
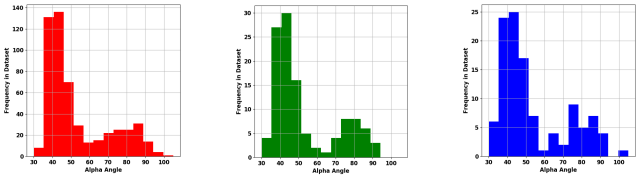


Fig. 3. Shows the lowest KS statistic computed for any training (red line), validation (green line) and test (blue line) split.



(a) Training Set (b) Validation Set (c) Testing Set

Fig. 4. Distribution of α-angles in each subset corresponding to the split from which the KS statistic in Fig 3 was computed. Note that the distributions are similar.

4. EXPERIMENTS

We experimented with two architectures, a simple U-net architecture [10] with 512×512 images as input and output and a stacked hourglass network [11] with 8 hourglasses and 256×256 images. The stacked hourglass network was proposed for human-pose detection but has been applied to find landmarks in medical images [12]. We used a batch size of 4 and an ADAM optimizer with a learning rate of 0.001 stepped down by a factor of 0.1 after epoch 40. We used the following data augmentation steps during training: random horizontal flipping, random vertical shift up or down at most 10% of the image height, a random rotation clockwise or anticlockwise of at most $5°$ and a multiplication of all pixels by a random factor between 0.75 and 1.25. We trained for 80 epochs.

Table 1. Results for α-Angle difference over the test set.

Network	Mean $\pm$ STD	Median	ICC
U-Net	$7.67^\circ \pm 10.29^\circ$	**3.98°**	0.71
Stacked Hourglass	**7.53° ± 9.42°**	4.00°	**0.76**

Table 2. Cumulative percentage of α-angle differences. E.g. 22% of the time our algorithm predicted an α-angle which was at least 10° different from the clinician.

α-angle discrepancy	$> 10^\circ$	$> 20^\circ$	$> 30^\circ$	$> 40^\circ$	$> 50^\circ$
Percentage	22%	7%	5%	3%	1%

We used the official stacked hourglass implementation [13] and ran our experiments on an Intel Xeon E5-2698@2.20GHz with a 32GB GPU.

5. RESULTS AND DISCUSSION

Results for automatic α-angle calculation are displayed in Table 1. We find that the stacked hourglass network performs the best, achieving a mean and median average difference from the clinician of 7.53° and 4.00° respectively; it achieves a 0.76 ICC when compared with the clinician measurements. This confirms our algorithm is comparable to another human operator because Agricola *et al.* [3] found that the ICC score between clinicians at the task of recording α-angles was 0.73.

In Table 2 we show how often our model disagrees with the clinician to various degrees. In one image our algorithm produces an α-angle over 50° different from the clinician (Fig 6f). In this case it misses a small bump on the lateral side of the femur head. This is the most common mistake our algorithm makes, another example of which is shown in Fig 6e. Our algorithm accurately measures α-angle for 'pistol-grip' femurs, where instead of a small bump, the whole femur head is convex (Fig 6c). Overall our algorithm does well at measuring small α-angles, such as those in Fig 6a,b.

A cam impingement is diagnosed when the α-angle for a hip is $> 65^\circ$: since previous work has proposed thresholds of $> 70^\circ$ for men and $> 61^\circ$ for women [4], we assumed there were equal numbers of men and women in the test set and averaged 70° and 61° to get the 65°. For this chosen threshold we diagnose correctly in 90% of hips. Other statistics for this threshold are shown in Table 3. We focused on diagnosing cam impingement because it is considered a clinically harder problem, we could also have diagnosed pincer impingement using a threshold on the LCE-angle.

Table 3. Comparison of algorithm diagnosis using an $\alpha > 65$ threshold against ground truth clinician diagnosis.

Total number of of hips in test set	112
Correctly diagnosed with no cam impingement	80
Correctly diagnosed with cam impingement	21
Missed cam impingement	9
Incorrectly diagnosed with cam impingement	2
Percentage of hips correctly diagnosed	90%

Table 4. Results for LCE-Angle difference over the test set.

Network	Mean $\pm$ STD	Median	ICC
U-Net	**2.26° ± 1.61°**	**1.89°**	**0.93**
Stacked Hourglass	$2.60^\circ \pm 1.85^\circ$	2.30°	0.91

Table 5. Localisation differences (in mm) for each key-point when compared to a clinician over the test set.

Key-point	Mean $\pm$ STD	Median
Lateral Ace. Point	1.17mm $\pm$ 0.71mm	1.18mm
Lateral Cam Point	3.60mm $\pm$ 3.89mm	2.35mm
Femur Head Center	1.41mm $\pm$ 0.79mm	1.18mm
Neck Orientation Point	1.68mm $\pm$ 1.00mm	1.86mm

(a) (b) (c)

Fig. 5. Comparison between **clinician LCE-angle shown in red** and **automatic LCE-angle shown in blue**. (a) and (b) are accurate measurements. (c) shows the hip our model performs worst on: it 'believes' the acetabulum rim extends further around the femur head than our clinician.

Table 4 shows the results for the LCE-angle calculation. LCE-angle is much easier to measure accurately. We obtain the best result for LCE-angle measuring using a U-Net architecture which achieves a mean and median average difference from the clinician of 2.26° and 1.89° respectively, and an 0.93 ICC score. Two accurate measurements are shown in Fig 5a, b. The worst measurement, with a 6.4° difference, is shown in 5c: our algorithm believes that the acetabulum rim extends further round the femur head than our clinician.

The reason we see better results for LCE-angle than α-angle is because the key-points upon which the angle depends are easier to locate in the image. This is shown in Table 5. LCE-angle depends on locating the 'femur head center', the 'neck orientation point' and the 'lateral acetabulum point',

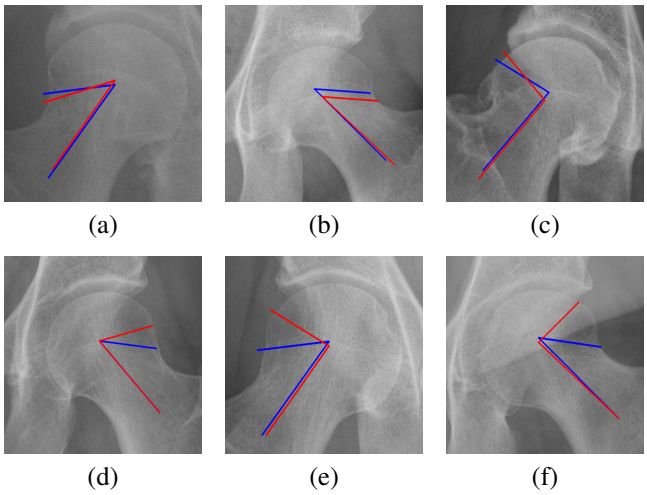


Fig. 6. Comparison between **clinician α-angle shown in red** and **automatic α-angle shown in blue**. (a) and (b) are examples of accurate measurement for small α, (c) is an acceptable measurement of a large α. In (d) our algorithm slightly underestimates α. In (e) and (f) our model makes a large error by not spotting a small bump on the femur head.

each of which is located more accurately that the 'lateral cam point' which is required for α-angle. intuitively, the 'lateral cam point' is harder to place for human operators as well which is why the inter-operator error between clinicians measuring α-angle is so high (ICC score of 0.73 in [3]). We also found that the U-net was better at locating 'easier' key-points because it trains on 512×512 images hence it obtains the best scores for LCE-angle measurement. The stacked hourglass network is better at locating the 'lateral cam point' despite training on 256×256 images. We hypothesise that this is because it has a higher capacity than the U-Net and can hence 'understands' how to place the non-trivial 'lateral cam point'.

6. CONCLUSION AND FUTURE WORK

To our knowledge this the first automated α-angle calculation and the first automated FAI diagnosis. In future we will also compare our algorithm against multiple clinicians. When a larger dataset is available, we will construct a network which only takes in appropriate parts of the input image, thereby achieving more accurate localisation of the 'non-trivial' key-points. This approach can be used on a variety of x-rays to produce various measurements for different body parts.

7. COMPLIANCE WITH ETHICAL STANDARDS

This research study was conducted retrospectively using human subject data made available to us by the FAIT study [8] which had ethical approval.

8. ACKNOWLEDGMENTS

We acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work. http://dx.doi.org/10.5281/zenodo.22558. No conflicts of interest to declare.

9. REFERENCES

[1] R Ganz et al., "Femoroacetabular impingement: a cause for osteoarthritis of the hip," *Clinical Orthopaedics and Related Research®*, vol. 417, pp. 112–120, 2003.

[2] R Santiago et al., "Imaging of hip pain: from radiography to cross-sectional imaging techniques," *Radiology research and practice*, 2016.

[3] R Agricola et al., "Cam impingement of the hip—a risk factor for hip osteoarthritis," *Nature Reviews Rheumatology*, vol. 9, no. 10, pp. 630, 2013.

[4] CR Fraitzl et al., "Femoral head-neck offset measurements in 339 subjects," *Knee Surgery, Sports Trauma, Arthroscopy*, vol. 21, no. 5, pp. 1212–1217, 2013.

[5] S Kutty et al., "Reliability and predictability of the centre-edge angle," *Int orthopaedics*, vol. 36, no. 3, pp. 505–510, 2012.

[6] K Sun, "Deep hi-res representation learning for human pose estimation," in *IEEE CVPR*, 2019, pp. 5693–5703.

[7] B Bier et al., "X-ray-transform invariant anatomical landmark detection," in *Int Conf Med Im Comp & Comp-Assist Interv*, 2018, pp. 55–63.

[8] AJR Palmer et al., "Protocol for the femoroacetabular impingement trial (FAIT)," *Bone & joint research*, vol. 3, no. 11, pp. 321–327, 2014.

[9] FJ Massey Jr, "Kolmogorov-Smirnov test for goodness of fit," *J Amer Stat Assoc*, vol. 46, no. 253, pp. 68–78, 1951.

[10] O Ronneberger et al., "U-net: Convolutional networks for biomedical image segmentation," in *Int Conf Med Im Comp & Comp-Assist Interv*, 2015, pp. 234–241.

[11] A Newell et al., "Stacked hourglass networks for human pose estimation," in *ECCV*, 2016, pp. 483–499.

[12] F Kordon et al., "Multi-task localization and segm for x-ray guided planning in knee surgery," in *Int Conf Med Im Comp & Comp-Assist Interv*, 2019, pp. 622–630.

[13] ," `github.com/princeton-vl/pytorch_stacked_hourglass`.