Fully Automated Shape Analysis for Detection of Osteoarthritis from Lateral Knee Radiographs

Luca Minciullo
Centre for Imaging Sciences
University of Manchester, UK
luca.minciullo@manchester.ac.uk

Tim Cootes
Centre for Imaging Sciences
University of Manchester, UK
timothy.f.cootes@manchester.ac.uk

Abstract—Osteoarthritis (OA) is the most common form of arthritis, affecting millions of people around the world. Since no cure has been discovered and considering the financial impact on health systems, any attempt to understand more of this disease could reveal new insights that would help develop new therapies.

Lateral knee radiographs are often ignored both by clinicians and the research community when trying to diagnose OA or other diseases that affect the knee joint. Our goal is to show that this view has a considerable potential. We present a fully automated method based on a Random Forest Regression Voting Constrained Local Model (RFCLM) to discriminate radiographs of people that have developed OA from people who have not. The experiments involved models built on different combinations of the four shapes (patella, tibia, medial and lateral femoral condyles) of the knee joint. We show that automated analysis of the lateral view achieves classification performance comparable if not better than similar techniques applied to the frontal view.

I. INTRODUCTION

Osteoarthritis (OA) is the most common form of arthritis, affecting millions of people around the world. It has been reported [13] that by 2030 around 20% of the American population will be above the age of 65, and that half of them (35 million of patients) will be at high risk of developing OA.

Since no cure has been discovered and considering the financial impact on health systems [1], any attempt to understand more of this disease could reveal new insights that would help develop new therapies.

OA is currently assessed from radiographs using the Kellgren and Lawrence (KL) [2] grades from 0 to 4, where 0 represents normality and 4 the most severe stage of OA. When a radiograph is taken clinicians assign a discrete KL grade based on features in the image. This is time consuming, subjective and there are shortages of suitably trained radiologists. There is an increasing need for reliable systems that can perform the grading automatically.

We describe the first fully automated system for classifying OA and KL grades from lateral radiographs of the knee. We evaluate its performance and show that analysing the shape of bones in lateral images gives better results than using the shape in PA (frontal) views. The lateral view is often ignored both by clinicians and the research community when trying to diagnose OA or other diseases that affect the knee joint. We show that this view has a considerable potential when trying to assess the state of OA from a radiograph.

We used a Random Forest Regression Voting Constrained Local Model (RFCLM) [3], [4], [5] to locate points in both single bones and combinations of bones. We used an object detector based on Random Forests (RF) to automatically initialise the RFCLM on each image.

The RFCLM returns the found points and the associated shape parameter vector. We used the components of these vectors as features on which to train a Random Forest classifier.

II. METHOD

Our model is made of four different sub-shapes: the patella (21 points), the lateral femoral condyle (24 points), the medial femoral condyle (25 points) and the tibia (32 points). The whole knee model is then made of 102 points (Figure 1).

We analysed different combinations of these shapes in order to understand which features are most informative.
A. Statistical Shape Model

A shape model [11] can be obtained by applying the Principal Component Analysis (PCA) to a set of aligned shapes (vectors). We used a linear model of shape variation, that represents each shape $x = (x_1, y_1, x_2, y_2, \ldots)^T$ in the following way

$$ x = T(\bar{x} + Pb; t), \quad (1) $$

where $\bar{x}$ is a representation of the mean shape in a suitable reference frame, $P$ is a matrix containing a set of modes of variation (Figure 2) and $T$ applies a global similarity transformation with parameters $t$.

The shape parameters $b$ can be calculated from $x$ using

$$ b = P^T(T^{-1}(x; t) - \hat{x}). \quad (2) $$

These parameters were the features in our classification tasks.

B. Shape Model Matching

The first step was to build a global searcher able to find the individual bones within each image. We used a Hough Forest approach [12]. We defined a bounding box starting from a pair of landmark points and then sampled from each image a set of $23 \times 23$ patches with different displacements, angles and scales with respect to the location of the bone of interest. We then trained a Random Forest to learn the functional relation between the pixel intensities in the image patches and the corresponding displacements. This RF is scanned over a new image at multiple scales and orientations, voting for likely knee locations. The output of the global search is a bounding box with two reference points, from which we initialise each model (Figure 3).

C. Classification

The approach above enables us to fully automatically locate the points of the outlines of the bones in new images. From both manual and automated annotation we can find shape parameters from the statistical shape model. The shape parameters are weights representing which modes of variation were found in the data instance and with what magnitude.

We train Random Forests on combinations of shape parameters, in order to predict: (a) OA vs non-OA, (b) the KL grade. Our implementation involves 100 trees per Forest and we use
two stopping criteria when building the trees: the maximum depth that a tree can have and the minimum entropy in the data. As soon as one of these two conditions is met we stop splitting.

### III. Results

#### A. Data

Our dataset is made of 300 lateral knee radiographs, 60 images per grade, from the MOST (Multicenter Osteoarthritis Study) dataset [13]. MOST is a longitudinal, prospective study of Knee OA run in the United States and involves 3026 participants, men and women aged 50 to 79. Each participant makes one visit per year, with about 5 visits in total. The dataset also contains various information recorded at each visit, including the KL grades from 0 to 4, indicating respectively: normal, doubtful, minimal, moderate, severe.

For the binary classification task the grades have been split into two groups: non-OA, KL (0,1), and the OA group, KL(2-4).

#### B. Landmark point detection

Finding Landmark points in a radiograph is challenging due to the way bones overlap in the projection. The lateral view is even more challenging than the frontal view since the two femoral condyles look almost identical, making them difficult to distinguish even for an experienced clinician. Table I contains the results of experiments evaluating the accuracy of the local search for each of the bony outlines shown in Figure 1. The latter results were obtained by training the model on 200 examples and testing on the remaining 100.

Our model performs well on the patella, where the error on 90% of the examples is less than 0.5 mm. This is probably due to the lower spatial variation of the landmark points describing this shape. The accuracy of detection of the other shapes is considerably worse, though we always have a median error of less than a millimeter.

#### C. OA Classification

The results shown in this section are obtained performing a 5-fold cross validation. For both binary and 5-class classification we trained a Random Forest made of 100 trees.

Our experiments have been performed first looking at the shape parameters from each shape, both using manual and fully automated annotation. We also concatenated different combinations of the points and then built the shape model on the concatenated points. In this way we can assess what shapes or combination of shapes contain more information and how big is the loss in performance when we move from a manual annotation to a fully automated one.

We also used the combinations of different shape parameters, independently obtained from different shape models, as features.

##### a) Binary Classification.

As we can see from Table II, the Tibia and the Patella are the two shapes whose features achieve the best individual classification accuracy. The best overall performance is obtained by the whole knee model in the manual annotation and by the individual tibia model in the fully automated system. The AUC for the Medial Femoral Condyle is higher in the fully automated system than in the model trained on features built on manual annotation.

The best results for the fully automated system were achieved when we concatenated the shape parameters of different shapes, calculated independently. In this way we ignore the relative position of the different shapes. The results corresponding to these experiments are shown in Table III. The overall best binary classification performance of our fully automated system is achieved by the knee model given by the concatenation of the shape parameters of the four sub-shapes. Figure 4 shows the results of the concatenation of the different shape parameters with manual annotation. If we add the lateral femoral condyle parameters to the patella we obtain a ROC curve that is consistently lower than the one related to just the patella. Figure 4 shows the results of the concatenation of the different shape parameters with manual annotation. If we add the lateral femoral condyle parameters to the patella we obtain a ROC curve that is consistently lower than the one related to just the patella. Figure 5 shows the ROC curves corresponding to the different combination of shapes obtained by concatenating the shape parameters. Concatenating shape parameters leads to improved ROC curves in each case.

We are not aware of any other work investigating the potential of the lateral view. Results using a similar approach
on 500 Posterior-Anterior (PA) radiographs from the OAI dataset [14] are given in [6].

The comparison of their best results using shape parameters with our performance is reported in Table IV. Although we are dealing with results coming from different views and different datasets this suggests that models trained on the lateral view show great promise.

b) 5-Class Classification.: In this section we describe the results corresponding to the 5-class classification task, a considerably more challenging problem. In this case we train a Random Forest classifier to predict the KL grade from the shape parameters. In Table V we show the results for this new task in terms of the proportion of data correctly classified.

<table>
<thead>
<tr>
<th>Shape</th>
<th>Our Method (manual)</th>
<th>[6]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual</td>
<td>0.842 ± 0.017</td>
<td>0.796</td>
</tr>
<tr>
<td>Fully Automated</td>
<td>0.794 ± 0.015</td>
<td>0.789</td>
</tr>
</tbody>
</table>

The individual shapes whose shape parameters perform best are the Patella from manual annotation and the Medial Femoral Condyle for the fully automated system. We see again two examples where the fully automated system has better performance than the manual system. However, the best overall performances for both annotations are achieved by the full knee model. For completeness we also report the confusion matrices for both knee models (Tables VI and VII).

With regard to the manual annotation (Table VI), radiographs with grade 0 were easiest to classify. It is encouraging to observe that in the vast majority of instances the mislabeled images were assigned to grades close to the ground truth.

Table VII is the confusion matrix corresponding to the fully automated model. In this case grade 0 turned out to be the hardest to classify, being often mislabeled with grade 1 and 2. Conversely, the fully automated model performs better than the manual one in classifying grade 1 and 2 and they achieve exactly the same accuracy on grade 3.

The proportion of data correctly classified by the model using concatenated shape parameters is reported in Table VIII. We observe an overall increasing trend when adding more shapes, but unlike what we saw with binary classification the results on automated annotation are consistently worse if
TABLE VIII

<table>
<thead>
<tr>
<th>Shape</th>
<th>Manual</th>
<th>Fully Automated</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patella+LCon</td>
<td>45.8 ± 2.8</td>
<td>32.4 ± 1.9</td>
</tr>
<tr>
<td>Patella+LCon+MCon</td>
<td>47.1 ± 0.7</td>
<td>35.6 ± 2.3</td>
</tr>
<tr>
<td>Knee</td>
<td>47 ± 1.6</td>
<td>39.2 ± 1.4</td>
</tr>
</tbody>
</table>

comparing to the ones obtained by concatenating the different annotations. The overall best results are achieved both by the manual and the automated system by the full knee models built on the concatenated point annotations, with a proportion of data correctly classified of respectively 47.9 ± 0.8% and 43.9 ± 1%.

IV. CONCLUSION

We have shown the first attempt at building a fully automated system to classify OA and OA grades using shape information from lateral knee radiographs. The results suggest that the lateral view contains very informative features that can achieve performance of the same level if not better to that of the PA view. One of the reasons for this is that one of the bones that is most affected by knee OA, the patella, is clearly visible in the lateral view, but it is obscured by the femur in PA images.

There is still a great room for improvement in fully automated OA diagnosis. In future work we will use a combination of shape and texture parameters for the lateral view. Informative texture features can be found near the tibial spines and, in general, in all the locations that are most likely to develop osteophytes. When concatenating shape parameters coming from different models it would be interesting to apply some sort of feature selection, in order minimise the noise in the data. Furthermore, it is worth investigating a way of concatenating features from both views.

Finally, we will be studying techniques able to quantify the risk for a patient to develop OA in the near future given the current state of the joint.

ACKNOWLEDGMENT

The research leading to these results has received funding from EPSRC Centre for Doctoral Training grant 1512584. The authors would like to thank Jessie Thomson and Raja Elsim for their useful comments.

REFERENCES