Multi-landmark environment analysis with reinforcement learning for pelvic abnormality detection and quantification

Bekkouch, Imad Eddine Ibrahim; Maksudov, Bulat; Kiselev, Semen; Mustafaev, Tamerlan; Vrtovec, Tomaž; Ibragimov, Bulat

Published in:
Medical Image Analysis

DOI:
10.1016/j.media.2022.102417

Publication date:
2022

Document version
Publisher's PDF, also known as Version of record

Document license:
CC BY

Citation for published version (APA):
Multi-landmark environment analysis with reinforcement learning for pelvic abnormality detection and quantification

Imad Eddine Ibrahim Bekkouch a,b, Bulat Maksudov b,c, Semen Kiselev b, Tamerlan Mustafaev b,d, Tomaž Vrtovec e, Bulat Ibragimov e,f

a Sorbonne Center for Artificial Intelligence, Sorbonne University, Paris, France
b Institute of Data Science and Artificial Intelligence, Innopolis University, Innopolis, Russia
c Department of Computer Science, University College Dublin, Dublin, Ireland
d Public Hospital #2, Department of Radiology, Razan, Russia
e Faculty of Electrical Engineering, University of Ljubljana, Ljubljana, Slovenia
f Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

A R T I C L E   I N F O

Article history:
Received 9 August 2021
Revised 14 January 2022
Accepted 3 March 2022
Available online 12 March 2022

Keywords:
landmark detection
magnetic resonance
reinforcement learning
pelvic abnormalities

A B S T R A C T

Morphological abnormalities of the femoroacetabular (hip) joint are among the most common human musculoskeletal disorders and often develop asymptomatically at early easily treatable stages. In this paper, we propose an automated framework for landmark-based detection and quantification of hip abnormalities from magnetic resonance (MR) images. The framework relies on a novel idea of multi-landmark environment analysis with reinforcement learning. In particular, we merge the concepts of the graphical lasso and Morris sensitivity analysis with deep neural networks to quantitatively estimate the contribution of individual landmark and landmark subgroup locations to the other landmark locations. Convolutional neural networks for image segmentation are utilized to propose the initial landmark locations, and landmark detection is then formulated as a reinforcement learning (RL) problem, where each landmark agent can adjust its position by observing the local MR image neighborhood and the locations of the most-contributive landmarks. The framework was validated on T1-, T2- and proton density-weighted MR images of 260 patients with the aim to measure the lateral center-edge angle (LCEA), femoral neck-shaft angle (NSA), and the anterior and posterior acetabular sector angles (AASA and PASA) of the hip, and derive the quantitative abnormality metrics from these angles. The framework was successfully tested using the UNet and feature pyramid network (FPN) segmentation architectures for landmark proposal generation, and the deep Q-network (DeepQN), deep deterministic policy gradient (DDPG), twin delayed deep deterministic policy gradient (TD3), and actor-critic policy gradient (A2C) RL networks for landmark position optimization. The resulting overall landmark detection error of 1.5 mm and angle measurement error of 1.4° indicates a superior performance in comparison to existing methods. Moreover, the automatically estimated abnormality labels were in 95% agreement with those generated by an expert radiologist.

This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/)

1. Introduction

Morphological abnormalities of the femoroacetabular (hip) joint are among the most common human musculoskeletal disorders (Wilson and Furukawa, 2014), and their diagnosis is usually based on a combination of clinical and radiological examinations. Radiological examinations aim to quantitatively assess the shape and spatial relationships between the femur, femoral head, and acetabulum using anteroposterior X-rays computed tomography (CT), or magnetic resonance (MR) images of the joint. The orthopedic community has established a list of anatomical metrics that characterize different hip abnormalities (Iyer, 2018), out of which many are asymptomatic at the early, easily treatable stages. A systematic review by (Mascarenhas et al., 2016) has concluded that around 22% of 6,705 asymptomatic volunteers exhibited signs of femoroacetabular impingement that could be observed from medical images, while (Kim et al., 2019) demonstrated that up to 19% of the analyzed 10,342 hips of asymptomatic volunteers have undiagnosed hip dysplasia. The detection of silent or early-stage hip abnormalities can be achieved by examination of pelvic/abdominal images of patients that are acquired for reasons other than visualization of
the femoroacetabular joint. A greedy search for hip abnormalities in all such images is time-consuming and impractical, and should therefore not be performed by physicians. This makes room for computer-aided diagnosis (CAD) solutions that can routinely scan images of asymptomatic patients and alert the user about potential abnormalities. Such an approach is also supported by the Data Science Institute of the American College of Radiology, who included femoroacetabular impingement diagnosis from CT and MR images to the list of clinical challenges where artificial intelligence will be most impactful in providing better care for patients (“Define-Altern Case Directory,” n.d.).

Automated detection and segmentation of femur and acetabulum in medical images are usually the initial steps for the CAD of hip abnormalities. The earlier methods in this field were based on heuristics, and combined image thresholding, filtering, connected component decomposition, etc., for bone and bone edge enhancement (Cheng et al., 2013; Harischandra et al., 2020; Kim et al., 2018; Liu et al., 2016, 2014; Nishii et al., 2004; Yun et al., 2020; Zoroofi et al., 2003). Many of such algorithms relied on the relative sphericity of the femoral head that allows its reliable identification using spherical template matching (Ben Younes et al., 2014; Liu et al., 2014) or finding the intersection point for gradient vectors initialized at bone edges (Ibrajimov et al., 2012; Vrtovec et al., 2012b). Recently, (Lopes et al., 2020) presented a survey of papers that explicitly tested the idea that femoral heads and acetabular cups can be approximated with spherical shapes, and found that the sphere is already a good approximation of a healthy femoral head with the surface fitting error of around 0.87 mm, while a more complex superovoid can capture the bone surface with a 0.68 mm fitting error. It was, however, observed that heuristic-based methods fail on pathological hips (Gangwar et al., 2018), especially those affected by arthritis. As a result, methods such as graph analysis (Chu et al., 2015a; de Raedt et al., 2013; Krčah et al., 2011; Pauchard et al., 2016; Wang et al., 2019; Xia et al., 2014; Zheng, 2020) and anatomical atlasing (Arezoomand et al., 2015; Blesler et al., 2017; Carballido-Gamio et al., 2015; Chu et al., 2015a, 2015b; Ehhardt et al., 2001; Foroughi et al., 2008; Siversson et al., 2014; Xia et al., 2013; Zheng, 2020) were consequently applied for hip bone segmentation. One of the shortcomings of such methods is their inability to perform the morphometry of the hip structures, as even an accurate segmentation of the femoral head and acetabulum does not immediately indicate if there is an acetabular under- or over-coverage. On the other hand, the hip morphology can be automatically analyzed through statistical shape modeling, where the hip tissues are usually defined with meshes, and the acceptable deformations of such meshes that capture the anatomical variability of normal and pathological hips are learned from a training set of annotated images. Statistical shape models were used for segmentation of femurs (Almeida et al., 2016; Ben Younes et al., 2014; Schmid et al., 2011; Yokota et al., 2013, 2009; Yun et al., 2020; Zhang et al., 2016), femoral heads (Arezoomand et al., 2015; Chu et al., 2015b; Lindner et al., 2013; Väänänen et al., 2019; Xia et al., 2013), acetabulae (Yokota et al., 2013; Yun et al., 2020), innominate bones (Chu et al., 2015b; Schmid et al., 2011; Xia et al., 2013; Yokota et al., 2009) and the cartilage (Chandra et al., 2016; Xia et al., 2014). One of the inherent properties of statistical shape models is that they try to approximate the hip shape of a patient from the shapes in the training population, which can potentially complicate the segmentation of hips with previously unseen severe abnormalities. However, this inability to capture unseen abnormalities was used to pinpoint their presence. Several studies demonstrated that healthy pelvic meshes can be aligned to the patient with acetabular deformations to quantitatively assess the bone loss level, and then propose the optimal acetabular surface to plan the bone reconstruction for the revision total hip arthroplasty (Hettich et al., 2019; Meynen et al., 2021, 2020; Schierjott et al., 2019). Despite resulting in a more accurate segmentation than the previously described methods, statistical meshes can still fail in the case of unpredictable or difficult-to-capture abnormalities.

Hip abnormalities are clinically diagnosed using anatomical distances and angles, extracted using manually placed anchor points, i.e., landmarks. Automated landmarking of pelvic images mimics this diagnosis workflow and results in simple-to-interpret image descriptions. Landmarks can be automatically detected on the surface of statistical meshes aligned to the target pelvis, and such an approach was used for anthropometric pelvis analysis (Fischer et al., 2019), total hip arthroplasty planning (Seim et al., 2009), measurement of the Tonnis (Yang et al., 2020; S.-C. Zhang et al., 2020a), sharp (Yang et al., 2020), center-edge (de Raedt et al., 2013; Yang et al., 2020) and acetabular sector (de Raedt et al., 2013; Ibrajimov et al., 2012) angles, and measurement of the femoral head extrusion index (Yang et al., 2020). Similarly as for the other fields of diagnostic imaging, the rapid expansion of deep learning has also affected the field of CAD of hip abnormalities, and convolutional neural networks (CNNs) have been utilized for segmentation of the cartilage (Golan et al., 2016; Xia et al., 2017), femur (Chen et al., 2019; Wei et al., 2020; Zeng and Zheng, 2018) and pelvis (Rouzrok et al., 2021; Zeng and Zheng, 2018), as well as for hip landmark detection (Li et al., 2019; Yang et al., 2020; S.-C. Zhang et al., 2020a). Li et al. (2019) applied mask regional CNNs for the detection of box coordinates that were centered at the lower edge of the teardrop and posterior acetabulum edges with the aim to compute the sharp angle. Zhang et al. (S.-C. Zhang et al., 2020a) used a similar methodology for the detection of the triradiate cartilage, acetabular lateral edge points, and ossification center of the capital epiphysis to quantitatively evaluate the developmental hip dysplasia. Yang et al. (2020) reformulated landmark detection as a circle segmentation problem, where landmarks were first replaced with small circles and segmented by U-Net, and then the resulting circle centers were used to compute the center-edge, Tonnis and sharp angle, and the femoral head extrusion index. The authors also tested if the use of CNNs can enable a direct diagnosis of hip abnormalities from raw X-ray/CT images without the need to perform intermediate segmentation or landmarking steps. Such a direct approach proved is promising for the diagnosis of femoral and acetabular osteophytes (von Schacky et al., 2020), narrowing of the joint space (von Schacky et al., 2020), subchondral sclerosis and cysts (von Schacky et al., 2020), osteonecrosis (Chee et al., 2019), osteoarthritis (Xue et al., 2017), hip dysplasia (Park et al., 2021) and femoral fractures (Sato et al., 2020), and for evaluating the outcomes of hip arthroscopy (Kunze et al., 2021). Despite the methodological simplicity, such direct image analysis with CNNs often results in a slightly inferior performance, and requires more training images in comparison to the CAD solutions that mimic the clinical workflow.

The above-presented literature summary allows us to identify two major limitations in the CAD of hip abnormalities. While almost all recent work on pelvic segmentation has been performed on three-dimensional (3D) CT or MR images, most CAD solutions for pelvic landmark detection use two-dimensional (2D) anteroposterior X-ray images (Li et al., 2019; Yang et al., 2020; S.-C. Zhang et al., 2020b), with 3D landmarking used only as an auxiliary step for mesh-based segmentation (Arezoomand et al., 2015; Chu et al., 2015b; Rouzrok et al., 2021; Seim et al., 2009). Similarly, the CAD solutions that utilize deep learning are mainly focused on 2D X-rays (Chee et al., 2019; Kunze et al., 2021; Park et al., 2021; Sato et al., 2020; von Schacky et al., 2020), mostly because of the simplifications of the landmarking process and the corresponding metrics calculation. Many metrics describing pelvic anatomy can be computed from the anteroposterior and lateral view (Lim and Park, 2015; Vrtovec et al., 2012a), which significantly simplifies their
manual calculation, including the center-edge, Tonnis, sharp and femoral neck-shaft angle, but relies on the assumption that the position of the patient is optimal in the target X-ray. Bizdikian et al. (2018) observed the change of the femoral neck-shaft angle against the technique used for its calculation and the relative rotation of the patient to the X-ray detector. The authors reported that the manual angle computed from 2D X-rays on average differs for 3.9°–4.8° from the reference angle computed from 3D CT images, and that this error increases by up to 1° when a 5° patient rotation is present. The second major limitation in the CAD of hip abnormalities is the lack of morphometry information in the existing landmark detection solutions, despite the fact the inter-landmark spatial models were shown to significantly improve the detection accuracy and were, as a result, used in the top-ranked algorithms that were presented at international competitions on landmark detection in medical images (Wang et al., 2016, 2015).

In this paper, we propose a novel landmark-based CAD framework for automated hip abnormality detection from MR images. The framework investigates the possibility of reformulating hip landmarking as a reinforcement learning (RL) problem, where landmarks are treated as RL agents. The recent review on the use of RL in medical imaging has identified seven papers on the topic of RL-based landmarking (S. Kevin Zhou et al., 2021). One of the first attempts of using RL for landmark detection in medical images was proposed by (Ghesu et al., 2018), who relied on the idea that each landmark can represent a single agent, and demonstrated that RL models can use local image patches to iteratively improve landmark positions in 3D whole-body CT images having the landmark moving policy defined as a one-pixel shift in x-, y- and z-direction. Al and Yun (2020) extended this formulation using an actor-critic policy gradient (A2C) model that allowed the landmarks to move for an arbitrary number of pixels in x-, y- and z-direction, which resulted in a considerable acceleration of landmark detection. The proposed models were validated on various anatomical tasks including cardiac (Al and Yun, 2020; A. Alansary et al., 2019a; Zhang et al., 2020a), spine (Al and Yun, 2020), brain (Alansary et al., 2019b; Ghesu et al., 2018) and abdominal organ (Ghesu et al., 2018) landmark detection. One major shortcoming of most of the pioneering studies is the lack of spatial inter-landmark information in the detection process. Ghesu et al. (2018) acknowledged the need for spatial information and used a statistical shape model to refine the results of RL-based landmark detection. Vlontzos et al. (2019) proposed to separate landmarks into groups, and defined agents for simultaneous optimization of landmark positions in the group. Zhang et al. (2020c) incorporated the anatomical properties of fetuses into the RL-based landmark detection by first manually connecting landmarks belonging to the same limbs and pairing landmarks of the eyes or shoulders, and then training RL models using the resulting graph of connections. The inclusion of spatial information significantly improved the accuracy of landmark detection in comparison to pure intensity-based RL models. These approaches, however, do not offer a universal tool for developing optimal multi-landmark spatial models. We propose to address this issue and augment RL landmark detection framework by optimizing landmark/agent positions considering the positions of other agents and proposes a novel shape-based algorithm for the analysis of the spatial relationships among landmarks using the graphical lasso (Friedman et al., 2008) and Morris sensitivity analysis (Nori et al., 2019). The shape-based algorithm is built upon the idea that anatomical landmarks have pairwise and groupwise spatial dependencies, which can be extracted and used to improve the landmark detection accuracy. The framework was validated on one of the largest databases of 3D MR images used for CAD of hip abnormalities, consisting of 115 T1-, 109 T2- and 113 proton density (PD)-weighted MR images acquired from 260 patients suffering from different hip pathologies. The performance of the frame-work was compared to the performance of an expert radiologist in terms of measurement accuracy of 16 landmarks and four clinical angles and grading of hip abnormalities.

2. Methodology

Landmark detection can be significantly improved by inter-landmark relationship models as it was shown in international competitions (Wang et al., 2016, 2015). In contrast to 2D, the complexity of inter-landmark models is of critical importance for 3D landmark detection. Indeed, the number of landmarks in 3D can exceed thousands when these landmarks, for example, capture the surface of 3D organs (Ibragimov et al., 2014). In the proposed work, we formulate landmark detection as a multi-agent optimization problem, where each landmark is observed as an intelligent agent for RL, and the agent environment depends on inter-landmark relationships. We then propose to solve the obtained optimization problem by capitalizing on the ideas from the graphical lasso (Friedman et al., 2008) and Morris sensitivity analysis (Nori et al., 2019) to capture the most representative inter-landmark connections that best describe the target object.

2.1. Clinical application

An anonymized database of patient cases with pelvic abnormalities has been collected from local clinical institutions. Each patient case consists of one or more T1-, T2- or/and PD-weighted MR images, and the corresponding diagnosis notes. In each MR image, an experienced radiologist manually placed 16 anatomical landmarks for the purpose of evaluating hip pathologies, i.e. two landmarks at the femoral head centers, four landmarks at the axes of both femurs, four landmarks at both femoral head-neck axes, and six landmarks at the lateral, anterior and posterior rims of both acetabulae. The four most widely used clinical metrics describing hip pathologies were then computed from these landmarks: i) Lateral center-edge angle (LCEA): the angle between the longitudinal body axis and the line connecting the femoral head center with the lateral acetabular rim (Fig. 2a), which serves to detect and quantify the pincer type femoroacetabular impingement and acetabular dysplasia (Monazzam et al., 2013), ii) Femoral neck-shaft angle (NSA): the angle between the axis of the femur and the axis of the femoral head-neck (Fig. 2b), which serves to characterize the development of lower limbs and helps to diagnose various diseases (Boese et al., 2016), iii) and iv) Anterior and posterior acetabular sector angle (AASA and PASA, respectively): the angles between the line connecting the left and right femoral head centers, and, respectively, the anterior and posterior acetabular rim (Fig. 2c), which serve to quantitatively assess the acetabular under- and overcoverage (Anda et al., 1991).

Fig. 1, 3, 4

2.2. Landmark candidate proposal

The RL-based landmark detection is based on the idea of iterative landmark position optimization using the image intensity information and potentially the spatial relationships among landmarks. The iterative optimization requires some initialization in the form of landmark candidate proposals. We tested four different approaches for landmark candidate proposal generation. The first naive approach defined landmark candidate proposals as random points sampled in the image. Such an initialization allows us to understand the robustness of the RL-based landmark detection in the situation when the initial landmark proposals do not provide any information. The second naive approach defined landmark candidate proposals as the mean locations of the corresponding landmarks computed from the training MR images. For such an initial-
ization, the RL-based optimization starts from the average pelvic anatomy. The third and fourth initializations were based on candidate proposals detected with neural networks. Two deep learning models were tested, namely UNet and feature pyramid network (FPN) with a region proposal network (RPN). These networks learned the appearance of landmarks on the training MR images and then scanned a previously unseen MR to detect the most likely candidate locations of the landmarks. Note that, during network training and testing, landmarks and landmark proposal coordinates are normalized to the [0,1] range against the corresponding image dimensions. After the landmark positions are finalized on a testing image, their coordinates are returned to the original scale.

2.2.1. UNet for landmark candidate proposal

The UNet architecture was developed for image segmentation and is currently widely used in medical imaging due to its universality and highly accurate segmentation results (Cai et al., 2020). It consists of a downsampling encoder path that extracts image features at different scales and of different complexity, and of an upsampling decoder path that generates the segmentation results using the extracted features. The core idea of UNet is to use skip connections that pass from the encoder to the decoder, and go from the early levels of the encoder path, where fine image details are still preserved, to the late levels of the decoder path, which contain the information from coarse features required for object recognition and segmentation. The encoder path consists of four blocks, each composed of two $3 \times 3 \times 3$ convolutional layers followed by a rectified linear unit (ReLU) activation layer and a $2 \times 2 \times 2$ max-pooling layer. The decoder path consists of four blocks, each composed of an upsampling layer, a $2 \times 2 \times 2$ trans-pose convolution layer, and two $3 \times 3 \times 3$ convolution layers, each followed by a ReLU layer.

To obtain landmark proposals by UNet, the landmark detection problem is first converted into a segmentation problem. For each image $I$ of size $l \times m \times n$, a landmark mask $J$ of size $l \times m \times n$ with $|P|$ channels is generated, where $|P|$ is the number of landmarks. For landmark $p \in P$, a binary sphere of radius $r = 10$ mm is generated and inserted to the $p$-th channel of $J$ at the location of $p$. UNet is trained to generate masks $J$ for images $I$ using a linear combination of binary-cross entropy and Dice loss functions with the weighting factors of 0.5. For test image $\bar{I}$, the candidate proposal for landmark $p$ is computed as the center of mass of the largest connected component from the $p$-th channel of the corresponding thresholded UNet result $\bar{J}$. The UNet was trained using the Adam optimizer with a learning rate of 0.0002 and weight decay of 0.0001.

2.2.2. FPN for landmark candidate proposal

The FPN with RPN architecture is based on the idea of efficient extraction of multi-scale features, allowing to combine low-resolution and semantically robust features with high-resolution but semantically weak features. Such an approach has proved efficient for various object detection tasks (Lin et al., 2017). The FPN component extracts features from the input image, and during this
process, it exhibits certain similarities to UNet. The encoder part of FPN is a 3D-ResNet (He et al., 2015), pre-trained on the combined Kinetics-700 (Carreira et al., 2019) and Moments in Time (Monfort et al., 2019) datasets. The decoder part creates feature maps of higher resolution by upsampling the results of previous layers, and each decoder path block generates an individual output for network training. The skip connections consist of $1 \times 1$ convolution layers applied to the encoder blocks that are then added to the corresponding decoder blocks. As a result, FPN aims to generate features of variable resolution that will be able to describe the target object. The RPN component analyzes the outputs of the FPN encoder block to evaluate how descriptive the FPN features are, and mimics a sliding-window classagnostastic object detector, which performs object/non-object binary classification and bounding box regression on feature maps. For this purpose, RPN applies a set of 15 rectangular anchor filters to each voxel of the FPN output array and measures the probability of this voxel to correspond to a landmark from $\mathcal{P}$ using the features in the output array. If the anchor filters are computed for a voxel located in a close neighborhood of a true landmark, FPN is trained to enhance this voxel. The desired neighborhood is modeled by computing the intersection-over-union (IoU) between the anchor filters and a predefined rectangular bounding box encompassing the landmark, where IoU values above $\Omega = 0.7$ are considered to be positive landmark location examples. The outputs of FPNs at different scales are then analyzed by RPN to extract the voxels that most probably represent landmarks. The coordinates of these voxels are converted into a vector and then passed through two fully-connected layers with 1024 features that predict the coordinates of each landmark from $\mathcal{P}$, and the optimal coordinates are used as final landmark proposals. The FPN was trained using the Adam optimizer with a learning rate of 0.0002 and weight decay of 0.0001.

2.3. Multi-agent RL for landmark adjustment

We propose to iteratively adjust the obtained landmark candidate proposals by solving a multi-agent RL problem. To apply the RL methodology, landmark position optimization is reformulated as a multi-agent Markov decision process (Zhang et al., 2019), where each agent is associated with the corresponding landmarks from $\mathcal{P}$. The state space $\mathcal{S}^p$ observed by the landmark $p$ agent is defined as follows:

$$\mathcal{S}^p = \{l(p), \nu^p\},$$

where $l(p)$ is a local MR image patch extracted around proposal location $p = [x_p, y_p, z_p]$ for landmark $p$, $\nu^p$ is the vector of landmark coordinate proposals, and $\nu^p$ is a binary vector indicating which landmarks are visible to $p$ ($\nu^p = 1$ if all landmarks are visible, and their coordinates can influence those of $p$; $\nu^p$ contains at least one non-zero element, as $p$ can always see its own coordinates during position optimization). In other words, landmark $p$ agent can, during its position optimization, observe its current location $p$, the image intensities around $p$, and the locations of some other landmarks from $\mathcal{P}\backslash p$. 

Fig. 3. Summary of the proposed framework for landmarking pelvic MR images, which consist of three components, namely, segmentation neural network for detection of landmark proposals, a combination of graphical lasso, Morris sensitivity analysis and neural networks for identification of optimal connections between landmarks, and reinforcement learning for landmark position optimization using both landmark candidate points and connection-based shape model.
At each optimization step, the image patch \( I(p) \) and visible-to-\( p \) landmarks \( v^p \) are first analyzed by a neural network to generate a latent representation of the current state of landmark \( p \) agent. The image patch passes through three convolutional layers with the filter size of \( 3 \times 3 \times 3 \), each followed by a ReLU layer. The output of the convolutional layers is flattened into a 1D vector \( \mathbf{w}^p \). Vector \( \mathbf{w}^p \) of spatial information is defined as all non-zero elements of \( v^p \), i.e. as the current coordinates of all landmarks visible to \( p \). Vectors \( \mathbf{w}^p \) are \( \mathbf{w}^p \) and \( \mathbf{w}^p \) are concatenated into a current state representation \( \mathbf{w}^p \). The representation \( \mathbf{w}^p \) is then analyzed by one of the four tested RL models, namely the deep Q-network (DeepQ) (Cheng et al., 2016), deep deterministic policy gradient (DDPG) (Mnih et al., 2013), twin delayed deep deterministic policy gradient (TD3) (Lillicrap et al., 2019) and actor-critic policy gradient (A2C) (Mnih et al., 2016). The reward function \( R \) of each RL model is defined as a weighted sum between the quality of the current step of the agent and its movement history:

\[
R(p) = \frac{R^p - R^0}{||R^p - R^0||} + \lambda \frac{R^p - R^0}{||R^p - R^0||}
\]

where \( p^k \) is the landmark position at \( k \)-th optimization iteration, \( p^* \) is the correct landmark position, and \( \lambda \) is a weighting factor for balancing the terms of \( R \). The first term of \( R \) rewards the agent for moving closer to the correct landmark position, while the second term rewards the agent for continuing to move in the same direction and avoid zig-zagging.

The loss function \( L \) is, for all tested RL models, a weighted combination between the squared distance error and orientation.
penalty:

\[
L = \frac{1}{d} \sum_{k=0}^{d-1} \left( \frac{(p_k^p - p^q)^2}{\|p_k^p - p^q\|} - \lambda \frac{(p_k^p - p^{\max(0,k-1)})^2}{\|p_k^p - p^{\max(0,k-1)}\|} \right)^2
\]  

(3)

where \( d \) is the number of steps needed for the agent to move from its initial position \( p^0 \) to correct position \( p^* \). For DeepQm, DDPG, and TD3, an agent can move 1 mm in any of the x-, y-, and z-directions, while for A2C the agent movements are unrestricted in 3D. The RL models are trained to improve the positions of landmarks in \( P \) by considering the location of each other, i.e. the inter-landmark relationships.

The RL-based landmark detection framework is applied to detect landmarks on a new image as follows. For a testing image \( f \), the initial landmark candidate proposals \( P_0 \) were first generated. These proposals could be defined from the results of FPN, UNet, or using some other approach. The initial state \( \{f(p_0), vP_0\} \) for landmark \( p \) was converted into a state representation \( wP \) using the network described above, and passed through the corresponding RL model for landmark \( p \) to generate the optimal move that should improve the currently proposed location from \( p_0 \) to \( p_1 \). After finding the improved proposals for all landmarks, the set of landmark proposals is updated to \( P_1 \). The procedure is repeated until the movements for all landmarks are smaller than threshold \( \delta \) or a predefined number of iterations reached.

2.3.1. Sparse undirected graph of connections

Vectors \( vP \) (Eq. 1) are predefined for each landmark \( p \in P \) before the RL optimization, and represent one of the most critical components of the proposed framework. If \( vP = 1 \) (i.e. all landmarks are visible for \( p \)), the RL models can become too restrictive and move landmarks towards the average pelvic morphology. On the other hand, if \( vP \) contains a single non-zero element (i.e. self-visibility for \( p \)), the RL models will ignore the spatial relationships among landmarks. We want to find the optimal multi-landmark environment for each agent considering the spatial relationships among landmarks in \( P \). The 3D landmark coordinates for all \( M \) images from the training set populate the distance matrix \( D \) of size \( |P| \times M \). The values of \( D \) are normalized according to the values of each landmark coordinate from the training set. The covariance matrix \( \Sigma \) of \( D \) describes the degree of correlation of different landmark coordinates and consequently estimates the spatial dependency among these landmarks. To find a set of most correlated landmarks, we can optimize the sparsity of precision matrix \( \Theta = \Sigma^{-1} \) (Mazumder and Hastie, 2012). The optimal precision matrix \( \Theta^* \) can be found by minimizing a L1-regularized negative log-likelihood:

\[
\Theta^* = \arg \min_{\Theta > 0} \left( - \log(\Theta) + tr(S\Theta) + \gamma \|\Theta\|_1 \right).
\]  

(4)

where \( |\Theta| \) is the determinant of \( \Theta \), \( S \) is the empirical covariance matrix estimated from \( D \), trace function \( tr() \) computes the sum of the diagonal elements of its input, \( |\Theta|\) is the sum of absolute values of \( \Theta \), and parameter \( \gamma \) controls the sparsity of the resulting optimal estimation. Eq. (4) is called the graphical lasso problem and can be solved with the primal-dual or interior-point methods (Friedman et al., 2008; Mazumder and Hastie, 2012). The solution \( \Theta^* \) highlights the most correlated landmark coordinates in the training set. After aggregating values \( \Theta^* \) using 3x3 kernel, we obtain a sparse matrix of most correlated landmarks.

The inter-landmark relationship model computed from the graphical lasso solution has several inherent shortcomings. First, matrix \( \Theta^* \) is usually close to being symmetric (Rolls and Rajaratnam, 2013), which, in the case of inter-landmark relationships analysis, will translate in the situation when if landmark \( p \) spatially depends on landmark \( q \), landmark \( q \) spatially depends on \( p \) to a relatively similar degree, i.e. \( \Theta^*[p, q] = \Theta^*[q, p] \). This is not optimal for anatomical pelvic models. For example, a landmark at the acetabulum rim can strongly rely on the center of the femoral head, while the center of the femoral head is better to be estimated from landmarks on the femoral neck. Second, matrix \( \Theta^* \) only computes the dependencies for landmark pairs and is deficient for the multi-landmark relationship analysis, required to capture the anatomical variability and complex morphology of the human bones.

2.3.2. Neural networks for landmark position estimation

To perform multi-landmark relationship analysis, the predictability of landmark normalized coordinates using the locations of the remaining landmark is estimated by training \( |P| \) neural networks. The network \( F_\theta \) for landmark \( p \in P \) is defined as a multi-layer perceptron with four building blocks, where the first three blocks consist of dense layers of \( n_{rand} \) neurons, followed by batch normalization, ReLU activation, and dropout layers, while the last block is defined as a dense layer with three neurons followed by a softmax activation layer, and predicts the 3D coordinates of landmark \( p \). The network is trained with the mean squared error loss augmented with L1-regularization.

For landmark \( p \), the corresponding network \( F_\theta \) is trained on normalized landmark coordinates of landmarks \( P \setminus p \) to predict the coordinate for \( p \) using all the samples from the training set. When a plausible set of coordinates \( x_{p,\lambda} \) for landmarks \( P \setminus p \) passes through the network \( F_\theta \), it predicts coordinate \( x_p = F_\theta(x_{P \setminus p}) \) for landmark \( p \). Apart from predicting the coordinates of \( p \), network \( F_\theta \) can quantitatively estimate the relative influence of landmarks \( P \setminus p \) on \( p \). For example, if landmark coordinates \( x_{p,\lambda} \) are modified by \( \Delta x_{p,\lambda} \), then the resulting estimation changes to \( x'_p + \Delta x_p = F_\theta(x_{P \setminus p} + \Delta x_{P \setminus p}) \). This increment \( \Delta x_p \) allows us to assess how the coordinate changes of landmarks \( P \setminus p \) affect the coordinate of \( p \) and discover the most spatially dependent landmark subgroups.

2.3.3. Multi-landmark spatial relationship analysis

The aim of this step is to pinpoint landmarks that contribute most to the location of landmark \( p \in P \). First, the neighborhood set \( N(p) \) of all potentially related to \( p \) landmarks is populated with indices of non-zero elements in \( p \)-the raw and column of \( \Theta^* \). In order to find how changes in different subsets of landmarks from \( N(p) \) affect the coordinates of \( p \), a vector \( d_{p,\lambda} \) of total contributions of each landmark from \( N(p) \) is defined and initialized with zeros \( (d_{p,\lambda}(p) = 0) \). For a subset of landmarks \( S \) from \( N(p) \), a sample contribution vector \( d_S \) is computed as:

\[
d_S^p = |F_\theta(x_{S,\lambda} + y_S) - F_\theta(x_{P \setminus S})|
\]

where \( y_S[k] = \{ \begin{array}{ll} 0 & : k \notin S \; \text{rand}(0, \sigma_k) \; : k \in S \end{array} \}
\]

(7)

where \( \text{rand}(0, \sigma_k) \) is a random 3D displacement sampled from the normal distribution with zero mean and standard deviation \( \sigma_k \), so that values \( \sigma = \{\sigma_k\} \) define the degree of flexibility of landmark displacement. In other words, the contribution vector estimates how random changes of landmark position from \( S \) change the predicted position of landmark \( p \). To compute \( \sigma \), landmark \( p \) locations for each image \( I \) are normalized against translation and rotation:

\[
\hat{p}_I = R(l_{ref}, I) \cdot (\bar{p}_I - \bar{p})
\]

(8)

where \( \bar{p}_I \) is the centroid of landmarks \( p_I \) for image \( I \), and \( R(l_{ref}, I) \) is the optimal rotation matrix that brings normalized landmarks \( (\bar{p}_I - \bar{p}) \) for image \( I \) as close as possible to the normalized landmark \( (\bar{p}_I - \bar{p}_{ref}) \) for a randomly preselected reference image \( I_{ref} \). The 3D standard deviation \( \sigma_I \) of landmark \( p \in P \) location is computed from the normalizations \( \hat{p}_I \) for all training images.
The vector of total contributions \( \mathbf{d}_{p}^{N(p)} \) is incremented according to the sample contribution \( \mathbf{d}_{p}^{S} \):
\[
\mathbf{d}_{p}^{N(p)} \leftarrow \mathbf{d}_{p}^{N(p)} + |\mathbf{y}| \mathbf{d}_{p}^{S}
\]
so that the sample contribution \( \mathbf{d}_{p}^{S} \) is weighted according to the absolute displacement \( |\mathbf{y}| \). Eqs. (7)-(9) are repeated for a predefined number of times to evaluate the change of coordinates of \( p \) against different random displacements of landmarks from \( S \) while keeping the remaining landmark coordinates \( \mathbf{x}_{N \setminus \{p, S\}} \) unchanged. Such an analysis is repeated for landmark subsets \( S \) of different size, and the resulting vector \( \mathbf{d}_{p}^{N(p)} \) estimates the relative contribution of each landmark from the neighborhood \( N(p) \) on the coordinates of \( p \). The landmarks with the largest relative contributions defined non-zero elements of \( \mathbf{W} \) for the RL multi-agent optimization (Eq. 1). Note that there is no need to normalize vector \( \mathbf{d}_{p}^{N(p)} \), as landmarks with the largest contribution are selected, and each landmark from \( N(p) \) participates the same number of times in the generation of \( \mathbf{d}_{p}^{N(p)} \). The algorithm described in this subsection extends the concepts of the Morris sensitivity analysis (Nori et al., 2019) to the evaluation of the multi-landmark spatial relationship.

3. Experiments and results

3.1. Patient database

The proposed framework was evaluated on a database of 260 patients with a total of 337 MR images (115 T1-, 109 T2- and 113 PD-weighted), where 180 patients / 251 images (78 T1-, 92 T2-, and 81 PD-weighted) were used for training and 80 patients / 86 images (37 T1-, 17 T2-, and 32 PD-weighted) were used for testing. The images were coronally reconstructed to an isotropic resolution with the coronal cross-section size ranging from 320 × 320 mm\(^2\) to 370 × 370 mm\(^2\), and the number of cross-sections ranging from 58 to 180.

Patient age ranged from 5 to 86 years with the median of 56 years, and 63% of patients were females. The gender and age statistics agree with the clinical knowledge that pelvic diseases are more common among females. All patients suffered from some kind of pelvic abnormality, with cartilage thinning, and left- and right-sided coxarthrosis being the most prevalent diagnosis (Table 1). Other abnormalities included osteoarthritis, inflammations, fluid accumulations, cystic formations, femur deformations, aseptic necrosis, avascular necrosis, ligament abnormalities, subchondral edema, synovitis, and degenerative joint disease. Female-specific diseases included uterine leiomyomas or fibroids and myomatous uterine node, nodular adenomyosis, endometrioid tumors, and ovarian cysts. Apart from hip-related pathologies, some patients suffered also from other diseases, such as hematomas, bladder diverticulum, chronic myeloid leukemia, myelogenous leukemia, cyst, tumors, or inflammations.

3.2. Experimental design

The parameters of individual components of the framework were defined as follows. The RL models for optimizing landmark locations (DeepQN, DDGP, T3D, and A2C) used 3D MR image patches \( I(p) \) of size 10 mm\(^3\) to capture the appearance of landmark locations, and the loss functions had the orientation penalty weighted with \( \lambda = 0.3 \) against the mean squared error (Eq. 2) and Eq. (3)). The RL-based landmark position optimization continues if there is at least one landmark candidate proposal that moves to more than \( \delta = 0.1 \) mm from its location in the previous iteration. The graphical lasso algorithm was optimized for 100 iterations with the regularization parameter of \( \gamma = 0.01 \) that controls the result sparsity and was solved using the scikit-learn python package.

The neural networks for multi-landmark relationship analysis had \( N_{\text{land}} = 48,32, \text{and} 16 \) features for the first three network blocks. The maximum number of landmarks \( k = 3 \) defined a random combination of landmarks to test their joint influence on landmark \( p \in \mathcal{P} \) (Eq. 6), and 1000 random displacements were generated for each combination (Eq. 7).

The learning rate for the RL models and landmark shape network was 0.001. The RL optimization is terminated after 50 iterations if the landmark adjustment has not converged. The framework was trained on a server equipped with Tesla V100-SXM2-16GB (GPU)/ Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz 80 cores 503 GB RAM.

3.3. Results

The framework performance was evaluated by estimating the accuracy of landmark detection, anatomical angle calculation, and hip abnormality detection and quantification. The accuracy of landmark detection was computed as the Euclidean distance between automatically detected and reference landmarks. In the landmark detection framework, the FPN-based landmark proposals consistently outperformed those obtained by UNet, while the use of the proposed shape model outperformed the no-model configuration and the shape model defined with a complete set of inter-landmark connections (Table 2). When augmented with the FPN-based landmark proposals and the proposed shape model, the resulting landmark detection accuracy was of 1.9, 1.7, 1.7, and 1.5 mm for DeepQN, DDGP, T3D, and A2C RL models, respectively (Table 2). The observed superiority of A2C RL model is further reflected in the accuracy of anatomical pelvic angle estimation, with 2.12°, 0.95°, 1.12°, and 1.16° mean degree error for LCEA, NSA, AASA, and PASA, respectively (Table 3). For a more comprehensive evaluation, we implemented three state-of-the-art RL-based landmark detection models, namely DeepQN MS (Ghesu et al., 2018), Duel DeepQN (Amir Alansary et al., 2019a), and Collab DeepQN (Vlontzos et al., 2019). The performance comparison of these models to the proposed shape model is given in Table 4.

The performance of the optimal framework configuration, which includes FPN for landmark candidate proposals, the sparse neural network-based shape model for capturing inter-landmark spatial relationships, and the A2C RL model for optimizing land-
Table 2
The comparison of reinforcement learning (RL)-based landmark detection without a shape model, with the shape model defined from the complete set of inter-landmark connections, and with the proposed shape model for capturing optimal inter-landmark connections. Four approaches for initial landmark proposals, namely random, based on mean landmark locations in training images, as detected by UNet and as detected by feature pyramid networks (FPN), as well as four RL methods, namely the deep Q-network (DeepQN), deep deterministic policy gradient (DDPG), twin delayed deep deterministic policy gradient (TD3), and actor-critic policy gradient (A2C), are compared. The results are reported as the mean (± standard deviation) error of hip landmark detection and angle estimation errors measured between manually and automatically detected landmarks.

<table>
<thead>
<tr>
<th>Shape model</th>
<th>Landmark proposals</th>
<th>Landmark error (mm)</th>
<th>Angular error (°)</th>
<th>Landmark error (mm)</th>
<th>Angular error (°)</th>
<th>Landmark error (mm)</th>
<th>Angular error (°)</th>
<th>Landmark error (mm)</th>
<th>Angular error (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>Random</td>
<td>3.1 ± 1.8</td>
<td>4.0 ± 2.4</td>
<td>3.2 ± 2.0</td>
<td>4.2 ± 2.7</td>
<td>2.6 ± 1.4</td>
<td>4.3 ± 2.1</td>
<td>2.3 ± 1.8</td>
<td>3.8 ± 2.2</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>2.3 ± 1.3</td>
<td>3.2 ± 1.9</td>
<td>2.2 ± 1.2</td>
<td>3.6 ± 1.9</td>
<td>2.2 ± 1.2</td>
<td>3.5 ± 1.8</td>
<td>2.4 ± 1.3</td>
<td>3.6 ± 1.8</td>
</tr>
<tr>
<td></td>
<td>UNet</td>
<td>2.0 ± 1.1</td>
<td>2.9 ± 1.6</td>
<td>2.1 ± 1.1</td>
<td>3.1 ± 1.7</td>
<td>2.0 ± 1.0</td>
<td>2.9 ± 1.6</td>
<td>2.1 ± 1.1</td>
<td>2.3 ± 1.6</td>
</tr>
<tr>
<td></td>
<td>FPN</td>
<td>2.0 ± 1.0</td>
<td>2.9 ± 1.7</td>
<td>1.9 ± 0.9</td>
<td>3.0 ± 1.7</td>
<td>1.9 ± 0.9</td>
<td>2.7 ± 1.6</td>
<td>2.1 ± 1.1</td>
<td>2.1 ± 1.5</td>
</tr>
<tr>
<td>Complete set</td>
<td>Random</td>
<td>2.8 ± 1.6</td>
<td>3.5 ± 2.2</td>
<td>2.8 ± 1.6</td>
<td>3.9 ± 2.3</td>
<td>2.5 ± 1.5</td>
<td>3.3 ± 2.1</td>
<td>2.4 ± 1.3</td>
<td>2.7 ± 1.8</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>2.2 ± 1.1</td>
<td>3.0 ± 1.8</td>
<td>2.1 ± 1.1</td>
<td>3.4 ± 1.9</td>
<td>2.4 ± 1.1</td>
<td>3.0 ± 1.9</td>
<td>2.0 ± 1.1</td>
<td>2.4 ± 1.6</td>
</tr>
<tr>
<td></td>
<td>UNet</td>
<td>2.1 ± 1.1</td>
<td>2.8 ± 1.7</td>
<td>2.0 ± 1.1</td>
<td>2.8 ± 1.7</td>
<td>1.9 ± 0.9</td>
<td>2.5 ± 1.6</td>
<td>1.8 ± 0.8</td>
<td>2.1 ± 1.5</td>
</tr>
<tr>
<td></td>
<td>FPN</td>
<td>1.9 ± 0.9</td>
<td>2.7 ± 1.6</td>
<td>1.8 ± 0.8</td>
<td>2.6 ± 1.6</td>
<td>1.9 ± 1.0</td>
<td>2.3 ± 1.6</td>
<td>1.7 ± 0.8</td>
<td>1.8 ± 1.2</td>
</tr>
<tr>
<td>Proposed</td>
<td>Random</td>
<td>2.5 ± 1.4</td>
<td>3.5 ± 2.1</td>
<td>2.4 ± 1.3</td>
<td>3.3 ± 2.0</td>
<td>2.4 ± 1.3</td>
<td>2.9 ± 1.9</td>
<td>2.3 ± 1.3</td>
<td>2.4 ± 1.7</td>
</tr>
<tr>
<td></td>
<td>Mean</td>
<td>1.9 ± 1.0</td>
<td>2.9 ± 1.6</td>
<td>1.8 ± 0.9</td>
<td>2.8 ± 1.7</td>
<td>1.9 ± 0.9</td>
<td>2.5 ± 1.6</td>
<td>1.7 ± 0.8</td>
<td>2.3 ± 1.4</td>
</tr>
<tr>
<td></td>
<td>UNet</td>
<td>1.8 ± 0.9</td>
<td>2.7 ± 1.5</td>
<td>1.9 ± 0.8</td>
<td>2.3 ± 1.5</td>
<td>1.7 ± 0.8</td>
<td>2.1 ± 1.4</td>
<td>1.6 ± 0.7</td>
<td>1.7 ± 1.2</td>
</tr>
<tr>
<td></td>
<td>FPN</td>
<td>1.5 ± 0.9</td>
<td>2.7 ± 1.6</td>
<td>1.7 ± 0.8</td>
<td>2.2 ± 1.5</td>
<td>1.7 ± 0.7</td>
<td>2.1 ± 1.4</td>
<td>1.5 ± 0.3</td>
<td>1.3 ± 1.1</td>
</tr>
</tbody>
</table>

Table 3
The results for the best performing configuration of landmark detection, obtained by using feature pyramid networks for landmark candidate proposals, the proposed sparse neural network-based shape model for capturing inter-landmark spatial relationships, and the actor-critic policy gradient reinforcement learning model for optimizing landmark locations. The results are reported as the mean (± standard deviation) error of landmark detection and angle computation.

<table>
<thead>
<tr>
<th>Landmarks (# of landmarks)</th>
<th>Error (mm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Femoral head (2)</td>
<td>1.27 ± 0.24</td>
</tr>
<tr>
<td>Anterior acetabular rim (2)</td>
<td>1.45 ± 0.19</td>
</tr>
<tr>
<td>Posterior acetabular rim (2)</td>
<td>1.47 ± 0.21</td>
</tr>
<tr>
<td>Lateral acetabular rim (2)</td>
<td>1.37 ± 0.21</td>
</tr>
<tr>
<td>Femur axes (4)</td>
<td>1.56 ± 0.21</td>
</tr>
<tr>
<td>Femoral head-neck axes (4)</td>
<td>1.67 ± 0.19</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Angles, both right and left</th>
<th>Error (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LCEA</td>
<td>2.12 ± 1.85</td>
</tr>
<tr>
<td>NSA</td>
<td>0.95 ± 0.32</td>
</tr>
<tr>
<td>AASA</td>
<td>1.12 ± 0.36</td>
</tr>
<tr>
<td>PASA</td>
<td>1.16 ± 0.37</td>
</tr>
</tbody>
</table>

mark locations, was evaluated in terms of the accuracy rate of hip morphological abnormality detection and quantification, defined as a percentage of correctly classified abnormalities according to the LCEA, NSA, AASA, and PASA that were computed from the automatically detected landmarks. The femoral head coverage is reflected in LCEA values between 25° and 39° for healthy subjects, while values above 39° and below 25° are characteristic for the pincer type acetabular impingement and acetabular dysplasia, respectively (Tannast et al., 2007). The normal range for NSA is between 120° and 140°, with values below 120° or above 140° considered coxa vara and coxa valga femur deformities, respectively (Gilligan et al., 2013). Both AASA and PASA angles serve to quantify the acetabular coverage of the femoral head, and values below 50° and 90°, respectively, are associated with its under-coverage. The applied framework resulted in accuracy rates of 81%, 100%, 99%, and 100% for LCEA, NSA, AASA, and PASA, respectively, indicating that the proposed optimal framework configuration can successfully quantify the morphological abnormalities of the hip.

Table 4
The comparison of the proposed shape-based RL framework for landmark detection against three existing RL-based landmark detection solutions: Deep Q-network (DeepQN) MS, Duel DeepQN, and Collab DeepQN. The framework used four RL models: DeepQN, deep deterministic policy gradient (DDPG), twin delayed deep deterministic policy gradient (TD3), and actor-critic policy gradient (A2C). All solutions were initialized in the same way using (FPN)-based landmark proposals. The results for Unet and FPN networks for the generation of landmark candidate proposals (Section 2.2.1 and 2.2.2) was also shown for comparison. The results are reported in terms of the mean (± standard deviation) error for landmark detection and angle estimation.

<table>
<thead>
<tr>
<th>Landmark error (mm)</th>
<th>Angular error (°)</th>
</tr>
</thead>
<tbody>
<tr>
<td>No shape</td>
<td>DeepQN</td>
</tr>
<tr>
<td></td>
<td>DDPG</td>
</tr>
<tr>
<td></td>
<td>TD3</td>
</tr>
<tr>
<td></td>
<td>A2C</td>
</tr>
<tr>
<td>Complete set</td>
<td>DeepQN</td>
</tr>
<tr>
<td></td>
<td>DDPG</td>
</tr>
<tr>
<td></td>
<td>TD3</td>
</tr>
<tr>
<td></td>
<td>A2C</td>
</tr>
<tr>
<td>Proposed</td>
<td>DeepQN</td>
</tr>
<tr>
<td></td>
<td>DDPG</td>
</tr>
<tr>
<td></td>
<td>TD3</td>
</tr>
<tr>
<td></td>
<td>A2C</td>
</tr>
<tr>
<td>Collab DeepQN</td>
<td>Vlontzos et al.</td>
</tr>
<tr>
<td></td>
<td>DUEL DeepQN</td>
</tr>
<tr>
<td></td>
<td>Ghesu et al.</td>
</tr>
<tr>
<td></td>
<td>UNet only</td>
</tr>
<tr>
<td></td>
<td>FPN only</td>
</tr>
</tbody>
</table>
4. Discussion

Anteroposterior X-rays are the most accessible imaging modality for the diagnosis of femoral structural abnormalities due to the relatively low cost of imaging equipment (Clohisy et al., 2008). At the same time, the projective nature of X-rays introduces ambiguities into the visual assessment of femoral head measurements, for example, when the patient is imperfectly positioned or/and anatomical landmarks are obscured with pelvic tissues. The reliability of measurements from X-ray images has been of great interest to the orthopedic community (Blümel et al., 2021; Mast et al., 2011; Wilson et al., 2011). The intra- and inter-observer variability for various measurements was around ±2.5° and ±6°, respectively, for angle-based measurements (Wilson et al., 2011), and around 1.4 mm and 3.1 mm, respectively, for distance-based measurements (Mast et al., 2011). Although these errors may seem to be relatively low, measurements of healthy subjects also lie within relatively narrow intervals so that the intra- and inter-observer abnormality classification agreement could be less than 50% (Mast et al., 2011). Blümel et al. (2021) compared the metrics from hip X-rays and MR images and observed that MR images should be treated as the reference standard for computing femoral measurements and that the agreement between X-rays and MR images strongly correlates with visual signs of correct patient positioning. Wang et al. (2017) observed a mean deviation of 8° and 3.7° between, respectively, the anatomically and radiographically calculated inclination and anteversion of the acetabulum, by using CT images as the reference modality. Despite 3D MR and CT images offering all the information for computing the femoral metrics, it is more difficult to visually comprehend than the information from 2D X-rays. For example, it may be difficult to visually pinpoint the center of the acetabular fossa or the most lateral rim of the acetabulum when the femoral neck axis is not aligned with the MR/CT image axis. The observer needs a volumetric view of the femoral head for the identification of such landmarks, which is usually not offered by the MR/CT image viewers that can only display orthogonal image cross-sections. On the other hand, machine learning algorithms are not limited by the fields of view, and therefore have the potential to be more reliable for femoral abnormality quantification.

Four clinical measurements were targeted in this study, namely LCEA, NSA, AASA, and PASA. The core idea of LCEA is to quantitatively assess the bony coverage of the femoral head by the acetabulum. While angles in the range of 25°–39° characterize healthy hips, angles in the range of 15°–20° for children and 20°–25° for adults may suggest a borderline dysplasia, which can be treated non-operatively with physical therapy, intra-articular injections, or minimally invasive hip arthroscopy (Yeung et al., 2016). Manual LCEA measurements are subjected to human errors, and existing clinical studies observed that, due to the corresponding inter- and intra-observer errors, 3%–16% of patients can be classified to a wrong abnormality (Batailler et al., 2019; Neltitz et al., 1999; Tannast et al., 2007). Yang et al. (2020) obtained similar results when comparing deep learning-based evaluation of LCEA from X-rays with the performance of three radiologists, as they reported an abnormality misclassification rate of 7% and 14% for the left and right femur, respectively. The authors also reported a mean LCEA error of 6°, with 93% of LCEA landmarks automatically detected within a 3 mm neighborhood around their reference locations. de Raedt et al. (2013) used the graph cut algorithm for the segmentation of the femoral head and acetabulum, and then measured LCEA from the segmentation results with a mean error of 2.6°. In our study, the mean error of automatic LCEA measurements from MR images was 2.12°, which suggests that such an approach is more reliable in comparison to manual measurements of LCEA from X-rays. Despite a relatively low angular error, the corresponding accuracy of bony coverage grades derived from LCEA was 81%, which may be due to the fact that many pathological patients in our database have borderline-abnormal values of LCEA.

The inclination of the femoral head-neck against the femoral shaft is evaluated by NSA, also named the caput-collum-diaphyseal angle, and is mostly used for planning hip surgeries including total hip replacements. In addition, NSA is widely used as a diagnostic metric for children with cerebral palsy, Perthes disease, epiphyseal dysplasia (Boese et al., 2016). The normal range for NSA is 120°–140°, depending on the climate and population group (Gilligan et al., 2013), and the inter-observer error in its estimation is around ±2°–4° (Wilson et al., 2011). Such a moderate error can, however, be reflected in up to 25% misclassification rate of femur deformities because the normal range represents a relatively narrow interval (Bouittier et al., 2013; Doherty et al., 2008; Mast et al., 2011). The main challenge in measuring NSA is the correct estimation of the femoral head-neck axis, however, due to human anatomical variations and femoral abnormalities, the femoral head can be visually rotated along the femoral shaft, which changes the appearance of the femoral neck on anteroposterior X-rays. On the other hand, it was shown that the distance between the posteroomedial facet of the greater trochanter and piriformis fossa can be used to characterize the projection error and adjust the calculation of NSA (Blümel et al., 2021). Nevertheless, there have been several attempts to automate NSA calculation from X-rays. Wei et al. (2020) applied generative adversarial networks to segment the femur from X-rays, then aligned a circular template on the resulting segmentation to find the femoral head, and finally fitted an elongated strip to capture the femoral shaft. The obtained femur segmentation and anatomical parcellation allowed the authors to automatically compute NSA with an error of 2.3° against a human observer. Cerveri et al. (2010) automatically extracted femur meshes from manually segmented CT images and estimated the femoral neck and shaft axis by fitting circles to the extracted meshes. The 3D NSA, automatically computed from the circle fitting results, was on average for 3° from the manually defined reference. In our study, the automatically obtained NSA is on average 0.95° from the reference values, and therefore superior to the reported results for 3D NSA, suggesting that MR imaging is a more appropriate modality for accurate NSA measurement.

To diagnose the acetabular dysplasia and plan the acetabular reorientation surgery, AASA and PASA are measured, with values below 50° and 90° being associated with anterior and posterior acetabular under-coverage, respectively. Moreover, low AASA and PASA values indicate the development of hip dysplasia, while normally high values may suggest the development of hip arthritis (Valera et al., 2018). In contrast to LCEA and NSA that can be computed from anteroposterior X-rays, AASA and PASA require the acquisition of 3D CT or MR images, and the measurements are usually obtained from the axial image cross-section located one level above the greater trochanter. Manual measurement of AASA and PASA therefore highly depends on the correct identification of the target axial cross-section and patient positioning, as the resulting angles can fluctuate for up to 5° due to imperfect selection of the reference cross-section and/or identification of acetabular corners (Goronzy et al., 2020). A more robust approach is to rely on the geometric center of the femoral head and the acetabular axis (Hansen et al., 2012; Jóźwiak et al., 2015), which however requires the development of computational tools for visual femoral head and acetabulum reorientation in 3D to align image axis with the acetabular axis. To the best of our knowledge, so far there have been two attempts to automatically compute AASA and PASA. Ibragimov et al. (2012) combined the Canny edge detection with the gradient accumulation map for the identification and segmentation of femoral heads, while (de Raedt, et al., 2013) used the graph cut algorithm, with the resulting error of auto-
matically computed angles ranging from 3.7° to 5.4°. On the other hand, the mean AASA and PASA errors in our study are 1.12° and 1.16°, suggesting a considerably more accurate angle identification in comparison to the reported results, which may potentially be due to the fact that in our study, the angles were computed from reformatted MR cross-sections, i.e., by using the centers of femoral heads at the reference horizontal orientation.

We also evaluated the association of landmark detection accuracy with different hip abnormalities as observed from images by computing the Pearson correlation coefficient and its statistical significance for each pair of landmarks against different abnormalities (results with $p$-values below 0.05 were considered statistically significant; after the Bonferroni correction against the total number of landmarks, $p$-values below 0.05/16 = 0.003 were considered statistically significant). The abnormalities with a certain level of prevalence in the database were considered, including the narrowing of the hip-acetabulum space, cartilage thinning, osteophytes, subchondral sclerosis, and osteoarthritis and synovitis. In addition, we also computed the correlation coefficients against the patient gender and age, resulting in only one statistically significant value ($p = 0.0004$), i.e., between the patient age and the corresponding detection accuracy for the landmark located on the left femoral head that marked the femoral neck direction.

The intensity-based landmark detection with neural networks, which in our framework represents the initial step of landmark candidate proposal detection, has been of significant interest to the research community. The recent review on the topic (Schwendinge et al., 2021) indicates that the existing methods can be roughly subdivided into patch-based classification, global and local regressions, segmentation-based and region-based methods. The patch-based classification methods divide images into patches and estimate the probability of the landmark to be located inside each patch. Such methods are computationally expensive and therefore mainly used in 2D landmark detection (Arik et al., 2017). The regression-based methods output landmark locations directly from the input images, or compute maps of displacements to point at the location of each landmark (Noothout et al., 2020). The segmentation-based methods convert landmarks into heatmaps or binary masks and apply neural networks such as UNet, to segment these heatmaps/masks. The region-based methods are similar to the previous category but use networks that output rectangular regions, such as FPN- and YOLO. Segmentation and region-based methods account for the majority of the recent landmark detection studies most likely because they rely on well-established universal network architectures. As shown in our study, the UNet and FPN candidate point detectors (Table 4) exhibit similar performance. The use of RL for landmark position adjustment significantly improves the results of UNet and FPN in terms of landmark detection accuracy (Tables 2 and 4). At the same time, the angle detection accuracy is only improved when the top-performing RL and shape models are utilized. One of the reasons for such an observation is that there is a non-trivial dependency between the accuracy of landmarks and angles. For example, landmarks 3, 5, and 6 (Fig. 2) can move alongside the femoral neck and shaft, which will not affect the NSA.

Our results show that the A2C RL model (Mnih et al., 2016) outperformed all other evaluated RL models, namely the DeepQNN, DDPG, and TD3 models. The core advantage of A2C is in the reformulation of the RL task as a two-network problem, where the critic network is trained to evaluate the Q-value function, while the actor network updates the landmark optimization policy depending on the environment and in accordance with the critic. The superiority of A2C can be associated with its advantage function that subtracts the Q value (the expected reward of the next landmark movement) from the V value (the expected cumulative reward of the current landmark). The observed ranking of RL models is in agreement with the results achieved in the fields of active sensing and other applications (Yoon et al., 2019; S.K. Zhou et al., 2021b).

The performance of the proposed framework was compared to alternative recent RL models for landmark detection (Table 4). We observe that the performance of the Duel DeepQNN (Amir Alansary et al., 2019a) and DeepQN MS (Ghesu et al., 2018) is on a similar level as the performance of the DeepQNN with no shape model both in terms of landmark detection and anatomical angle estimation. At the same time, Collab DeepQNN, which captured the inter-landmark spatial relationships using the complete set of inter-landmark connections outperformed DeepQNN, DDPG, TD3, A2C, DeepQN MS, Duel DeeQNN with no shape models, and exhibited similar performance to DeepQNN with the model based on a complete set of connections. This observation confirms the benefit of the use of shape information for pelvic landmark detection. The proposed multi-landmark shape model improved all tested RL models and allowed augmented A2C to exhibit the best landmark detection and anatomical angle estimation performance. We can conclude that the proposed framework advances the current state-of-art of RL-based landmark detection, and proposes a universal approach for capturing the optimal inter-landmark spatial relationships and integrating them into modern RL models.

It is important to mention the limitations of this study. Despite the fact that the assembled database is among the largest databases used for pelvic landmark detection and pathology classification, 337 MR images from 260 patients are not sufficient to have a rich collection of rare abnormalities. For example, there were only 4 patients with femoral head deformations. Although the algorithm performance was not significantly worse for such patients in comparison to the overall database, the relatively low number of such cases does not allow us to make a significant conclusion for deformed femoral heads. The second limitation is associated with the fact that all patients in the database were older than 18 years. Some diseases are common in children and infants, which justifies the need for framework validation on younger patients. The future work will include both database extension and algorithmic improvements, for example, the reinforcement learning agents will be trained to observe not only the coordinates but also the intensity profiles of the correlated landmarks to understand which landmarks candidate locations can be trusted.

**Declaration of Competing Interest**

The authors have no conflict of interest to declare.

**Acknowledgments**

This work has been supported by grant from The Analytical Center for the Government of the Russian Federation (Agreement No. 70–2021–00143 dd. 01.11.2021, IGK 0000000D730321P5Q0002).

**References**


