Otto-von-Guericke-University Magdeburg
Faculty of Electrical Engineering and Information Technology
Institute of Medical Technology

Master Degree Thesis

Automated Hip Knee Ankle Angle Determination using Convolutional Neural Networks

submitted: December 21,2018

by: Henok Hagos Gidey
born on March 04, 1985
in Harar, Ethiopia

Abstract

Advanced osteoarthritis is a leading cause of knee replacement and loss of functionality. Early detection of risk factors plays an important role in the application of preventive measures. One of the risk factors is the leg alignment which influences the speed of knee cartilage degradation. The 'gold standard' measurement of leg alignment is done by determining the Hip Knee Ankle (HKA) angle from full lower limb radiographs. Convolutional Neural Networks (CNNs) have gained popularity recently in computer vision. In this thesis we developed methods using CNNs to determine HKA angles from full lower limb radiographs. We trained the CNNs using data from the Osteoarthritis Initiative (OAI). We evaluated our method's performance by evaluating its agreement to experts measurement and its reliability. Our best performing method shows excellent agreement and reliability levels.

Acknowledgements

Firstly, I would like to express my sincere gratitude to Dr. Stefan Zachow for giving me the opportunity to work on this thesis. I would like to extend my gratitude for the support and feedback I gained from the entire ZIB visual data analysis team. I would as well like to thank Prof. Dr. Bernhard Preim for the feedbacks you gave me. Special thanks to my advisors Alexander Tack and Felix Ambellan for the motivation, patience and guidance. Without your support, it would have been impossible. Last but not least, many thanks for my friend Ibe Samson Chibuike, Karolina Sucharska and my family for the constant support.

Task of the Thesis in the Original:

Declaration by the candidate

I hereby declare that this thesis is my own work and effort and that it has not been submitted anywhere for any award. Where other sources of information have been used, they have been marked.

The work has not been presented in the same or a similar form to any other testing authority and has not been made public.

Magdeburg, December 21,2018

Contents

List of Acronyms

OA Osteoarthritis

HKA Hip Knee Ankle

OAI Osteoarthritis Initiative

CNN Convolutional Neural Network

ANN Artificial Neural Network

AAM Active Appearance Model

CLMM Constrained Local Model Method

AAMM Active Appearance Model Method

RBM Regression Based Method

ASM Active Shape Model

PCA Principal Component Analysis

SIFT Scale Invariant Feature Transform

LBP Local Binary Pattern

SSM Statistical Shape Model

ReLU Rectified Linear Unit

TMA Tibial Mechanical Axis

FMA Femoral Mechanical Axis

LMA Leg Mechanical Axis

LBA Leg Bearing Axis

DICOM Digital Imaging and Communications in Medicine

ICC Intraclass Correlation Coefficient

List of Figures

List of Tables

1 Introduction

In this chapter, the motivation and goals of this thesis are described. This is followed by a brief introduction of the tasks performed and the structure of the thesis.

1.1 Motivation

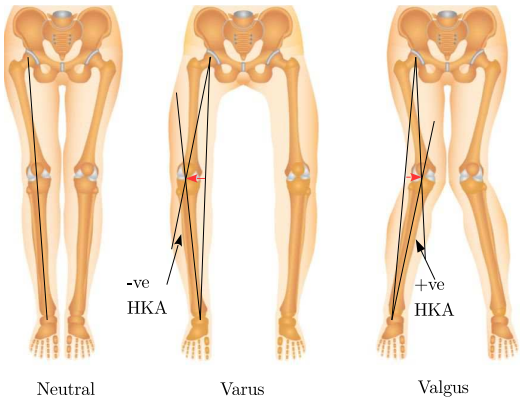
Knee osteoarthritis (OA) is one of the leading causes of loss of functionality [1]. With aging and increasingly obese global population, the prevalence of OA is expected to grow [1]. Nevertheless, there are no treatments that change the natural course of Knee Osteoarthritis [2]. As such, knee OA that advances to end-stage disease is the leading indication for total joint replacement surgery [2]. Without disease altering treatment, prevention strategies become paramount, and identification of the risk factors supports the development of prevention strategies.

One of the risk factors for OA is an increasing degree of coronal (frontal) plane leg misalignment. The association of knee alignment with OA is explained by the role knee alignment (varus and valgus alignment) plays on load distribution at the knee. An imbalance of load distribution at the knees causes more of the upper body weight to impact force on just one side of the knee causing more wear and tear of the knee joint cartilage (Fig. 1.1). The level of risk varies with the degree and orientation of misalignment.

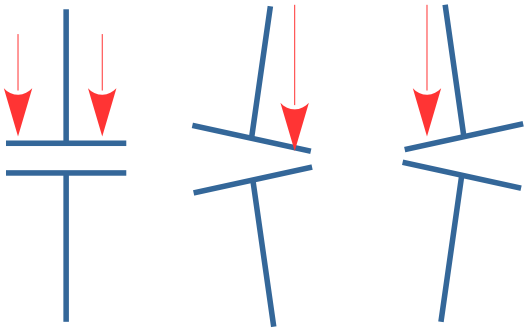
The association between leg alignment and load distribution means strategies to improve load distribution on misaligned knees (e.g. high tibial osteotomy) can play an important role towards management and better prognosis of knee OA. The first starting point of such strategies is, therefore, accurate measurement of leg alignment.

1.1.1 Measuring leg alignment

A number of methods can be used to assess and measure the coronal knee alignment: clinical deformity measuring device like a goniometer, standard knee radiographs, full lower limb radiographs, computer navigation systems, magnetic resonance scan, computerized tomographic scan or simply a surgeon's subjective measurement [4]. A goniometer does not expose the patient to radiation. However it can only be applied at the outer surface of the leg and can not accurately measure the geometry of the bony structure under various soft tissues. As standard knee radiographs are taken on just small portion of the leg anatomy above and below the knee joint, they expose the patient to a small amount of X-ray radiation in comparison to full leg radiographs. However the measurements of leg



(a) Leg alignment: A properly aligned leg, varus leg alignment and valgus alignment [3].



(b) Load distribution at the knee: The left side shows a neutrally aligned knee with weight equally distributed at the contact surface, in the middle and right image the most of the upper body weight imparts force on just small contact surface of the knee joint.

Figure 1.1: HKA angle determination and sign conventions.

alignment using standard knee radiographs are less accurate: this is due to the fact that standard knee X-rays does not allow the assessment of the entire leg anatomy.

An accurate estimation of leg alignment is done by measuring the Hip Knee Ankle (HKA) angle of a leg on a full lower limb radiograph [5,6]. A full lower limb radiograph is a frontal view X-ray image of the human anatomy containing the foot, thigh and hip. In such a radiographic image, the HKA angle is defined as the angle formed by the intersection of a line from the center of the femoral head to the center of the knee joint and another line from the center of the knee joint to the center of the ankle joint (Fig. 1.1a).

The determination of this angle is often done manually by locating landmarks at the center of the femoral head, knee and ankle joint. This manual process is prone to error. An automated implementation of such procedure can, therefore, produce a more accurate and reproducible HKA angle calculation without the need for X-ray reader training.

1.2 Goal of the thesis

Recently machine learning approaches to automate procedures have gained importance. This trend can also be seen in computer vision especially using CNN-based deep learning. The goal of this thesis is to develop a method to determine the HKA angle using CNNs and evaluate the method's performance. In this process we answer questions such as:

- What approaches can be used to determine higher level information such as the HKA angle from images using CNN?
- What factors influence the adaptability of CNN to determine HKA angle?
- What CNN architecture can be used to determine the HKA angle?

1.3 Project description

Convolutional Neural Networks (CNNs) are part of a broad class of machine learning approaches called deep learning which are inspired by the operation of the biological nervous system. Deep learning approaches learn a mapping between a representation of data and task output by finding not only the mapping function but also by learning the representation of the data which works best for particular task at hand.

In deep learning this is achieved by learning progressively higher and more abstract representation of the input data. Deep learning methods are particularly very useful in cases where it is hard to articulate computationally features of the input data that shall be used as basis to the task output. CNNs are implementations of deep learning for data that is arranged in grid like structure such as images. This makes CNNs particularly useful for computer vision tasks where data is organized as pixel values arranged in grids (arrays) and where it is hard to articulate features of an object in terms of its pixel values.

In this thesis two approaches based on CNNs are proposed to learn a mapping between the HKA angle and full lower limb radiographs (Fig. 1.2). The first investigates determining the HKA angle from raw images. In this approach we investigate if CNNs can learn a mapping between images and HKA angles. The second approach imitates the systematic approach human experts follow to determine the HKA angle. This systematic approach first determines landmarks which define the HKA angle and and determine the HKA angle of a leg using an algorithm which calculates the HKA angle from the landmarks determined by the CNN.

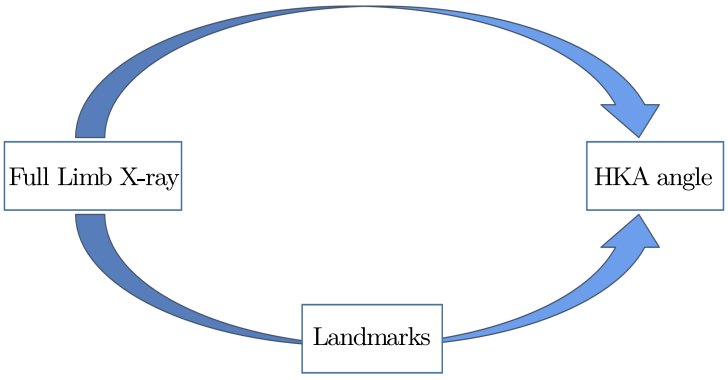


Figure 1.2: HKA determination approaches: HKA angle determination examines two approaches. The first directly predicts HKA angle the second determines landmarks first and calculates HKA angle from landmarks.

To achieve the goals of this thesis, data from the Osteoarthritis Initiative (OAI) is used. The data from OAI contains radiographic images and their associated HKA angle. The first approach is implemented in a data set of radiographic image with expert determined HKA angle (i.e., the OAI dataset as it is). For the second approach, a set of 900 images from the OAI dataset is labeled with landmark coordinate points by the author of this

report. This data is used with appropriate data augmentation techniques to generate enough training data for the CNN networks. The CNNs are evaluated to determine their accuracy in determining landmarks using the labeled data.

Considering that the medical images are of high image size (number of pixels), we explore ways to find smaller region of interest for landmark detection. In this regard, we explore two approaches using heuristic rules and automated ROI selection methods. The entire approaches performance is evaluated in determining the HKA angle on a much bigger testing data using the non landmarked image.

The second approach is finally tested as well on the dataset which is not labeled with landmark coordinates. The performance of both approaches is investigated by measuring the level of agreement between the HKA angle determined using the automated approaches proposed and the expert determined HKA angles.

1.4 Structure of the thesis

The thesis begins with an introduction of the theoretical background of CNNs and related work in regards to landmark detection in Chapter 2. Chapter 3 describe nomenclatures and conventions that are followed in the thesis. This is followed by the implementation details of the proposed methods to determine the HKA angle in an automated manner. The results of the experiments performed to evaluate the proposed methods is presented in Chapter 4 with explanation of the evaluation tools used. In Chapter 5, we discuss the challenges faced during the implementation of methods and the workarounds used together with the result of the experiments. Chapter 6 concludes the thesis with a summary of the thesis and discussion of limitation and directions for future work.

2 Literature Review

This chapter presents an overview of concepts involved in the thesis work. The automated determination of HKA angle is done using CNNs. Hence, an explanation of how CNNs work and how they are trained is given. One of the proposed approaches to determine HKA angle relies on accurate landmark detection. A brief introduction of landmark detection methods and the use of CNNs for landmark detection is given.

2.1 Convolutional neural networks

CNNs are a specialized form of machine learning approach called Artificial Neural Networks (ANNs). CNNs are particularly suitable to process data which is arranged in a grid like topology (e.g. arrays). They are used in various fields of applications such as computer vision, speech recognition, natural language processing. Since CNNs are specialized forms of ANNs, the discussion on CNNs starts with a brief introduction of ANNs and progresses to explaining CNNs.

2.1.1 Artificial neural networks

ANNs are a biologically inspired family of machine learning models that are used to learn patterns from data in order to map inputs to outputs. The simplest form of such a model is the perceptron. A perceptron makes a prediction based on a linear function defined by a set of weights and a vector of input elements (Fig. 2.1). This means that the decision boundary of a perceptron is linear, prohibiting its applicability for learning complicated patterns.

This limitation of perceptrons in modeling complicated patterns that can not be modeled using linear functions is solved by introducing a hidden layer (Fig. 2.2). The hidden layer is a collection of neurons which are connected to every neuron in the preceding and consecutive layer. The hidden layers transforms the input into a new form through using a linear transformation of the input followed by a nonlinear activation. This new transformation can be further processed by the neurons at the consecutive layer in order to achieve the desired output.

A series of hidden layers perform this transformation multiple times achieving a progressively more complex transformation of the input which can be easily evaluated at the output layer (for e.g. classified based on a linear combination of the elements of the transformations). This process is often termed as representation or feature learning.

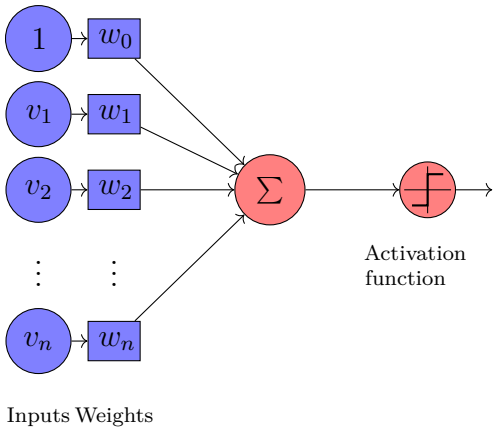


Figure 2.1: The perceptron: Each input element is weighted by a corresponding weights. The combined sum is used to make a decision.

Throughout the thesis the terms representation and feature are used interchangeably and refers to a transformation of the input which can be processed further with various machine learning approaches. More complex representation of the data are achieved by using many hidden layers which is known as deep learning.

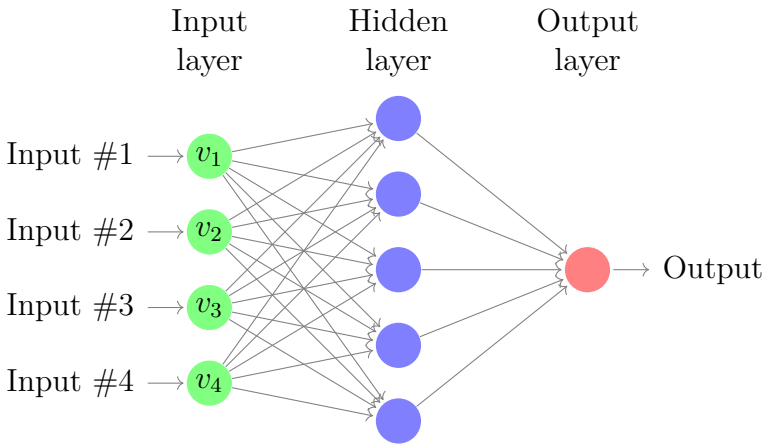


Figure 2.2: Multi layer ANN: The hidden layer with set of weights (not shown in the image corresponding to each neuron-neuron links) transforms the input linearly. This is fed into a nonlinear activation function. The output of the hidden layer becomes the input to the consecutive layers to be be further transformed into the desired output on the output layer.

The mapping of an input to an output using a multi layer ANN can be mathematically written as (e.g. for an ANN with only one hidden layer as given by) [7]:

$$y_k(\mathbf{v};\Theta) = f^2\left(\sum_{j=1}^{M} W_{kj}^{(2)} f^1\left(\sum_{i=1}^{D} W_{ji}^{(1)} v_i + b_j^{(1)}\right) + b_k^{(2)}\right), \tag{2.1}$$

where $\mathbf{v}$ is the input vector of D dimensions; $\Theta = \{\mathbf{W}^{(1)}, \mathbf{W}^{(2)}, \mathbf{b}^{(1)}, \mathbf{b}^{(2)}\}$ is a parametric set of weights and biases at each layer and y_k is the k^{th} element of the output vector.

$f^2(\cdot)$ and $f^1(\cdot)$ are non linear activations at the respective layers. As can be seen from Equation 2.1, ANNs model the mapping of input to output using a chain of linear and nonlinear transformation of the input.

Designing an ANN necessitates the appropriate selection of the width (number of neurons at a layer) and depth (number) of hidden layers; and the appropriate nonlinear activation function. The model parameters Θ are learned through an optimization process which minimizes the error function E of the training data, where E is defined as difference between the model output (Eq. 2.1) and the desired (observed) output for a particular input.

The error minimization is done iteratively using different algorithms which are based on the error gradient at the parameter set $\nabla E(\Theta)$. The gradient at each layer can be efficiently computed using the error back propagation algorithm [7]. Once the error gradient is calculated for all weights at each layer the corresponding parameters can be updated iteratively until convergence to a particular error threshold or iteration number using:

$$\Theta^{(t+1)} = \Theta^{(t)} - \eta \nabla E\left(\Theta^{(t)}\right), \tag{2.2}$$

where η is the learning rate and t represents the iteration number. Various approaches are employed to improve the iterative process such as stochastic and mini-batch gradient descent and momentum [8].

2.1.2 Convolutional neural networks

In ANNs the input is in a vector form in which the arrangement of individual input elements does not convey any information. In such cases, each neuron is connected to every neuron in the preceding layer. For variety of data modalities, the arrangement of individual units in a grid-like structure such as arrays conveys configurational information. In such data modalities, how the individual elements (e.g. pixels in images) are arranged in relation to each other determines the meaning of the data.

Some example of such data modalities are audio data and time series signals in the case of one dimensional arrays; images in the case of two dimensional arrays and videos and volumetric images in the case of three dimensional arrays. CNNs are designed to leverage such configurational information during data processing using a convolution operation (Eq. 2.3) at least on one of their layers which are called convolutional layers [9].

The convolution operation is defined in Equation 2.3 for continuous variable t and in Equation 2.4 for discrete variable t_n:

$$s(t) = (x * w)(t) = \int x(a)w(t-a)da, \tag{2.3}$$

$$s(t_n) = (x * w)(t_n) = \sum_{-\infty}^{\infty} x(a)w(t_n - a), \tag{2.4}$$

where s, x, w are functions on t and t_n. For two dimensional image the convolution operation takes the following form:

$$s(i,j) = (I * K)(i,j) = \sum_m \sum_n I(m,n)K(i-m, j-n), \tag{2.5}$$

where $s(i,j)$ denotes a convolution operation at pixel row i and column j; K represents the so-called convolutional kernel; m and n represent rows and columns in the convolution kernel. The convolution operation is often visualized as sliding a vertically upside down and horizontally left-right flipped convolution kernel and summing up the point-wise product of the convolution kernel and image. The convolution operation produces a two dimensional linear transformation of the input.

In addition to a convolution operation, a typical convolutional layer consists of additional operations. These operations are nonlinearly activation operation and a pooling operation. These are done in consecutive manner: a convolution stage, a non-linear activation stage and pooling stage (Fig. 2.3).

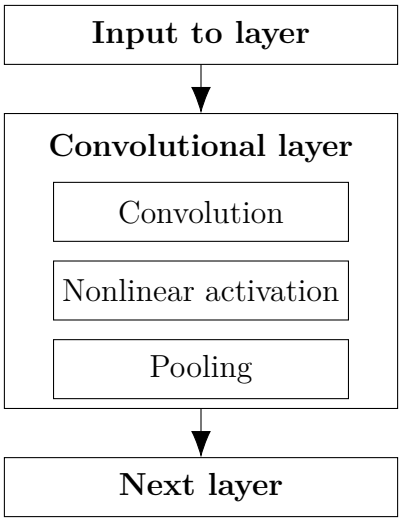


Figure 2.3: Components of a typical convolutional layer: Input image is transformed into an intermediary representation using a set of convolutions, nonlinear activation and pooling operation at a convolutional layer. Figure adapted from [9].

The parallel between ANNs and CNNs can be visualized by noticing that neurons in CNNs are arranged in a two dimensional grid (array). Since the elements of the input data are arranged in grid like structure, it is natural to arrange the neurons at the next layers that process the input are arranged in the same grid like structure. The neurons that are arranged in a grid like structure are often called feature maps (Fig. 2.4a). The process of extracting a set of feature maps is achieved using the operations of a convolutional layer.

This process can be seen in Fig. 2.4a. A convolution stage at layer l linearly transforms

the input using a convolution operation with kernel k. The convolution kernel size defines the region of close by neurons at the input that it has to be linked with. The weights of the convolutional kernel defines the linear transformation weights of this neurons. The convolution operation is followed by a nonlinear activation of the each neuron. This result is summarized using a pooling operation. The combined effect results in a new representation of the feature maps at preceding layer $l-1$. These operations are mathematically stated as follows:

$$\mathbf{A}_j^{(l)} = f\left(\sum_{i=1}^{M^{(l-1)}} \mathbf{A}_i^{(l-1)} * k_{ij}^{(l)} + b_j^{(l)}\right), \tag{2.6}$$

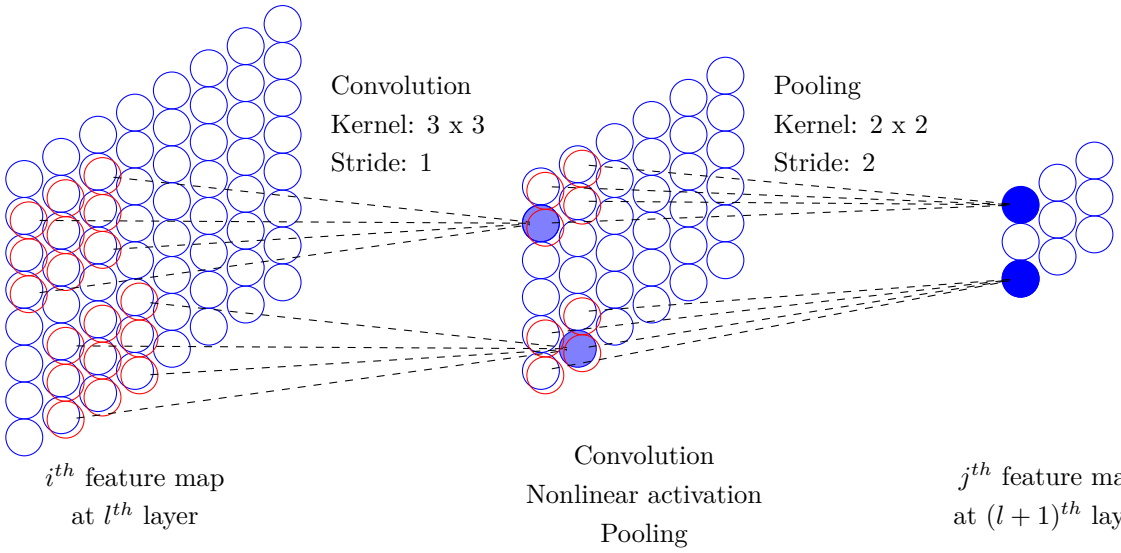
where $\mathbf{A}_j^{(l)}$ and $\mathbf{A}_i^{(l-1)}$ is the values of the j^{th} and i^{th} feature map at l^{th} layer and $(l-1)^{th}$ layer respectively. $*$ represents a convolution with convolution kernel $k_{ij}^{(l)}$. $b_j^{(l)}$ is a bias variable and f is a non linear activation function. Various nonlinear activation functions are used in CNNs. Some examples of such functions are sigmoid or logistic function, hyperbolic tangent function and rectified linear units (ReLU). Among these, ReLU units are popular due to their effect on improving the training time of a network [10]

A feature map at a certain level is a particular representation of the input which emphasizes some attributes of the input data. In CNNs, consecutive convolutional layers each composed of multiple feature maps extract an increasingly higher representation of the input (Fig. 2.4b). For instance a feature map at a first level may show edges which are extracted from pixel values of an input from a previous layer. The second layer can be a feature map which describes the arrangement of edges. This continues until the final output layer which is representation of the input that can be for example, linearly distinguishable for a particular prediction.

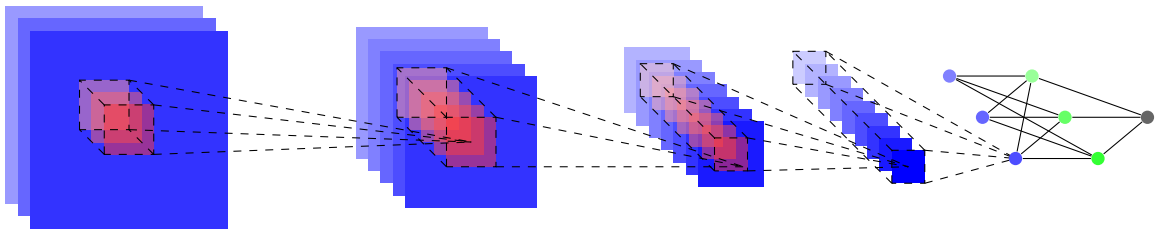
The pooling operation is used to generate a summary statistic of the output of the previous two stages of a convolutional layer. This is achieved by replacing the value of each unit in a feature map with a summary value corresponding to nearby regions. The summary is defined by a pooling function and the size of the neighborhood region with a kernel size (Fig. 2.4a). A popular pooling function is the max-pool function which gives the maximum pixel value of a region defined by the pooling kernel size. Other examples of pooling functions are average, weighted average based on a distance to central pixel.

Since pooling summarizes response of a neighborhood, neighboring pooling regions will be of same values when pooling regions are only a pixel apart. As a result, pooling regions are spaced more than one pixel apart. This can be seen in the pooling operation of Fig. 2.4a, which uses a stride of two pixels. The use of strides which are more than one pixel apart reduces the dimension of a feature map. Therefore, pooling stage is used to improve computational efficiency of the network and lower the memory requirement of a feature map.

In addition to summarizing the output of a convolutional layer, pooling stage makes a



(a) Convolutional layer operations: The input and output of a convolutional layer are arranged in an arrays conveying configurational information of the individual input elements.



(b) Typical CNN architecture: Each feature maps extracts a particular aspect of all the feature maps in the preceding layer using the operations of a convolutional layer. The red regions shows the region at multiple feature maps corresponding to the same area of the input image.

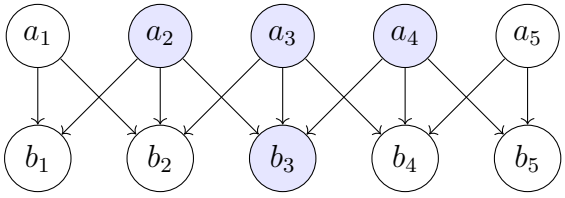
Figure 2.4: Operations of a convolutional layer and CNN architecture

representation to be invariant to small translations of the input. This is to mean that a small translation do not cause a change in most of the values a convolutional layer. As such, the pooling operation can be considered as prior that the function a convolutional layer learns will be a translation invariant.

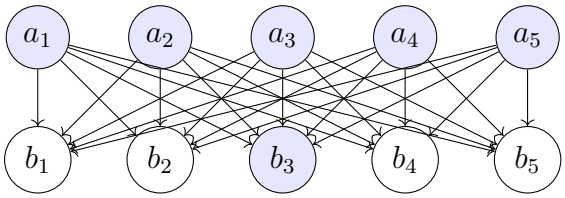
The convolution kernel weights which extract a particular feature map are learned during training. These weights connect the neurons at a particular feature map to a subset of spatially contiguous neurons at the preceding feature maps. This set of contiguous neurons can be the entire set of neurons at a preceding layer which is known as full connectivity; or it can be a neurons from a small neighborhood known as sparse connectivity (Fig. 2.5b, 2.5a).

The set of input units that are connected to an output neuron make the so called receptive field. Although direct connections of units among consecutive layers of a CNN are sparse, units at deeper layer have larger receptive fields in comparison to units at

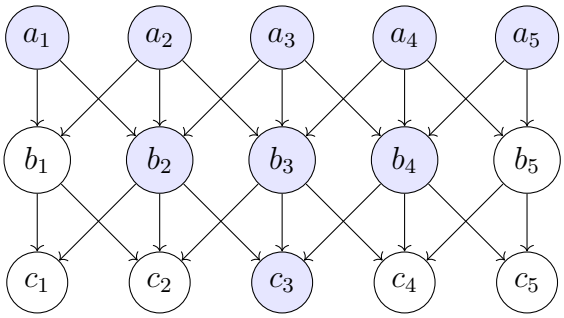
shallow layers. This is due to indirect connections to the input through intermediary layers. This is illustrated in Fig. 2.5c where b_3 has a receptive field of three units at the input layer: a_2, a_3 and a_4. The receptive field over the input layer is higher for deep layers: c_3 have access to all five units at the top layer through its connections to b_2, b_3 and b_4.



(a) Sparse connectivity: A neuron is connected to a subset of neurons in the preceding layer.



(b) Full connectivity: A neuron is connected to all neurons in the preceding layer.



(c) Receptive field: The effective receptive field of a neuron is the number of neurons to which it is connected to. The receptive field of deep layers is higher than the neurons at shallow layer.

Figure 2.5: CNN connectivity

Although each neuron at a feature map can have its own sparse connectivity defined by the convolution kernel weights, only one set of convolutional kernel weights is often used throughout a feature map which is known as parameter sharing. This condition improves memory requirement of a CNN as we only need one set of weights (kernel) to connect all neurons at a feature map to the neurons at a preceding layers. In addition to its effect on computational efficiency, parameter sharing causes a layer to be invariant to translation. This is to say that if the input changes (translation), the values of units in a feature map changes in the same way. This is due to the fact that the same convolutional kernel weights are used throughout a feature map. By using the same convolution kernel throughout a feature map a particular attribute can be detected no matter where in the image it is located.

2.2 Landmark detection

Landmark detection refers to the localization of key landmarks of an image. These landmarks are either visually distinguishable landmarks in an object such as corners and landmarks along contour lines or anatomically meaningful landmarks, e.g. tip of the nose, eye corners and landmarks along the contour of a jaw. Landmarks are often used as an input for other tasks performed on images. Few examples of such tasks are biometric measurements of anatomical structures, landmark based image registration,

image segmentation and image alignment.

HKA angle is defined as an angle between lines passing through biologically distinguishable landmarks. Therefore accurate determination of HKA angle requires the accurate determination of these landmarks. This section reviews scientific literature on landmark detection techniques. [1]

2.2.1 Landmark detection challenges

Although most medical images are acquired in a controlled situations, various issues complicate the process of accurate landmark detection on medical images. Some of these challenges encountered in medical images are the following:

Pose: A pose indicates the assumed stance or orientation of a subject in an image relative to the image acquisition instrument and the background. Therefore, a pose of an object influences the positioning of objects and shapes in an image in relation to each other. As landmark detection relies on evaluating such configurations, more variation of poses makes the landmark detection process more complex.

Even though medical image acquisition is often performed under a well defined guidelines of patient positioning and alignment, there is still a significant variation of pose of the patients. This can be due to inconsistencies in following the imaging guidelines or physical constraints making it difficult for patients to maintain appropriate pose during image acquisition.

Image quality: Medical image quality is defined by various parameters such as image contrast which describes the signal difference between two objects; noise which arises from stochastic nature of signals; and resolution which describes the smallest detail an image could depict. These quality factors are the byproducts of number of issues related to the imaging modality and procedure.

Some example of issues influencing the quality of medical image are random noise, signal strength, signal intensity, detector technologies and radiation dose considerations. The quality of image determines how much information about the object is contained in a discernible format inside the image. Wide variations in image quality plays a significant role in the effectiveness of any landmark detection process.

Occlusion: Occlusion indicates the presence of irrelevant objects in an image obstructing the visibility of an object of imaging interest. Despite guidelines describing medical imaging setups, it is common to find foreign objects in images. These can arise from misplaced radiation protection equipments, clothes of subjects and jewelery or the environment under which imaging is performed. In addition, it is as well common to find medical implants

[1] Most available landmark detection literature is on facial landmarks. This is due to the fact that facial landmarks are in depth researched due to their importance for facial recognition systems. As most of the concepts are transferable to landmark detection on medical images, this review relies on facial landmark detection literature.

and anatomical replacements in medical images. These can as well interfere with medical imaging consistency affecting the landmark detection process.

Artifacts: Artifacts are structures in a medical image which are not depiction of a natural object. Instead, they are artificial objects in an image which are caused by image reconstruction algorithms, signal acquisition errors, or other physical phenomena caused by interaction of the imaging signal with tissue. Such artifacts distort the natural shape of objects in an image making it challenging to determine landmarks on images with artifacts.

2.2.2 Landmark detection approaches

Landmark detection approaches can be grouped into three broad classes [11, 12]. These are Constrained Local Model Methods (CLMMs), Active Appearance Model Methods (AAMMs) and Regression-Based Methods (RBMs). The first two approaches which are also the earliest to be explored [11], are characterized by their explicit use of a parametric shape model to understand the shape of objects. In contrast, RBMs do not make explicit use of shape models. In the next subsections, these approaches are described briefly.

CLMMs

CLMMs predict landmark locations based on patches taken from an image using two components. The first component is called a local landmark expert. A local landmark determines a landmark given 'small' patch of an image around landmark. Here, a 'small' patch is supposed to indicate that the local landmark expert have a relatively local awareness (Fig. 2.6). The second component is a global shape model which models the spatial relationship between landmarks. Here, the shape model can be considered as prior knowledge of landmarks' configurations which is used to constrain the location of a landmark determined by local landmark expert (Fig. 2.6).

The shape model and the local landmark expert work iteratively by predicting each landmarks location independent of other landmarks on local patches using the local expert and then evaluating the validity of the shape created by combining all landmark. Mathematically, CLMM landmark detection can be articulated as optimization of the misalignment error (confidence score about the accuracy of an estimated landmark) over all landmarks subject to shape regularization:

$$Q(\mathbf{s}) = R(\mathbf{s}) + \sum_{i=1}^{N} D_i(\mathbf{x}_i; I_i), \tag{2.7}$$

where $\mathbf{s}$ represents shape created by combining N landmarks, R is a function of shape which penalizes implausible shapes and D_i denotes the measure of misalignment of landmark i at location $\mathbf{x}_i$ in the image patch I_i [13].

Local patches can be represented in various ways. It can be either simple pixel intensity

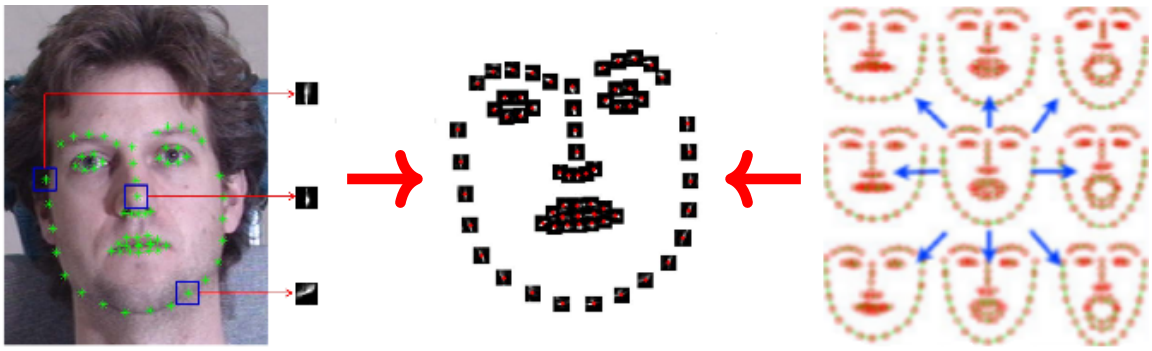


Figure 2.6: CLMM: Local landmark expert evaluates where a landmark is determines a new location of a landmark. The determined landmarks are combined into a shape which is evaluated by a shape model to check how likely it is. Figure adapted from [11, 13].

values or higher level features extracted from pixel values. The local landmark expert can be implemented using different approaches. For example, it can be a distance metric such as the Mahalanobis distance of landmark pixel value from the model pixel value of a landmark [14], a classifier such as linear support vector machine, a regressor which evaluates landmarks based on local patches around the landmark [15] or a convolutional neural network based local expert [16].

Shape information is usually deployed in the form of Statistical Shape Models (SSMs) [11, 12, 17]. An SSM describes shape variations of an object by calculating a mean shape and parameters of variation from a group of training examples [18]. This is achieved by representing shape as a distribution of landmark points across the surface of an object. A mean shape is calculated by averaging values for each coordinate. Principal Component Analysis (PCA) is carried out on the deviations of coordinates from the mean values to identify a small set of principal directions (modes) that best describe the observed variance in the training data [17]:

$$\mathbf{s} = \bar{\mathbf{s}} + \sum_{m=1}^{i} b_m \cdot p_m, \tag{2.8}$$

where p_m is a m^{th} mode of variation among i modes of variation, and b_m is the corresponding weight.

Although SSMs are used often, shape information can also be developed using a simple set of rules which describe landmark relationships or tree-based models which describe landmarks based on mutually dependent spatial positions [19].

Since landmarks are examined independently using local patches, CLMMs are fast in their landmark detection. They evaluate the validity of the shape relationships between landmarks using a shape model. The main challenge in this approach is the lack of global appearance awareness by the local landmark expert. By evaluating only the shape of the landmarks, they fail to model the relationship between appearance profiles of different

landmarks points. This information is valuable for accurate landmark detection.

AAMMs

AAMMs improves on the shortcomings of CLMMs discussed previously (Sec. 2.2.2). AAMMs are composed of both statistical shape and statistical appearance information of a deformable object. The statistical appearance model is developed in a similar way to SSMs albeit taking appearance representation (e.g. pixel values) on shape normalized training samples into account. Similar to SSM, a mean appearance and a reduced set of linear parameters is determined using PCA on the deviation of appearance from a mean value within training images [20].

$$\mathbf{a} = \bar{\mathbf{a}} + \sum_{n=1}^{j} c_n \cdot q_n, \tag{2.9}$$

where q_n is the n_{th} mode of variation among j modes, and c_n is the corresponding weight.

Although AAMMs contain both shape and appearance information, they are modeled using a unified parameter set controlling both shape and appearance [20]. Hence, AAMMs are used to generate a realistic looking template of an object. Landmark detection can then be formulated as a fitting problem which minimizes the difference between texture sampled from the testing image and texture synthesized by the template:

$$\Delta\mathbf{a} = \mathbf{a}_{model} - \mathbf{a}_{actual}, \tag{2.10}$$

where $\Delta\mathbf{a}$ represents the error between the appearance of model generated template ($\mathbf{a}_{model}$) and the actual appearance from an object ($\mathbf{a}_{actual}$). This is illustrated in Fig. 2.7. The model parameters which produces template with minimum fitting error would then be used to find the landmark locations.

Various approaches are deployed to fit model to a test image. The most straightforward approach is to use analytical solutions such as gradient descent on appropriate error function and adapt model parameters [21]. However, gradient descent based fitting is very slow and can not be used for real time applications. To improve the speed of convergence, regression functions are used to determine the model parameter's displacement during fitting (using linear regression [22] or nonlinear regression [11]).

Although regression based functions improve on speed of convergence compared to gradient descent methods, they are less accurate. For example, linear based models assume linear relationship between error image and update of model parameter which is not always the case. Although AAMMs improve on CLMM by introducing global appearance information, AAMM based methods are not robust towards inconsistencies in the image. These inconsistencies are due to variations in pose, image quality, as well as occlusion (Sec.2.2.1). In addition, AAMs (as well as SSMs in CLMMs) are inherently linear and

(a) Images are generated by changing the parameters of the AAM which are then compared with the image whose landmarks are to be detected.

(b) The best fitting model generated image (right image) is used to place the landmarks on an image of interst (left image).

Figure 2.7: AAMM. Figure adapted from [20].

assume linear shape and appearance variations. This assumptions is not always valid and can affects their suitability towards a particular problem.

RBMs

RBMs directly learn a mapping of landmark locations from image. They differ from the previous two methods in that they do not explicitly use a parametric shape model to find landmark positions. RBMs can be mathematically described as:

$$M : F(\mathbf{I}) \mapsto \mathbf{x}, \tag{2.11}$$

where M is mapping of feature $F(\mathbf{I})$ extracted from image $\mathbf{I}$ to landmarks $\mathbf{x}$.

The image appearance can be represented as raw pixel intensities or by a higher level representation (feature) extracted from pixel intensity values. Some examples of such features are Haar-like features [23], Scale Invariant Feature Transform (SIFT) [24], Local Binary Patterns (LBPs) [25] or Convolutional Neural Network (CNN) extracted higher level image representation [26]. Various regression functions can be used; such as random forest regression [27–29], linear regression on nonlinear features extracted from image (e.g. features extracted using a CNN [26], SIFT features [30]).

Regression can be applied on either a single-level or multi-level (in a cascaded manner). In

single level regression, landmarks are determined in one step without any prior initialization [28, 31]. In cascaded regression, the regression starts from an initialization and regresses to the actual landmarks by updating the landmarks from the previous stage [26, 32, 33].

Different cascading approaches are used for landmark detection. Xiong et al. [30] implemented a cascaded supervised descent algorithm to learn descent directions which minimize a L^2 loss function on SIFT features extracted around actual and initial landmarks [30]. Cao et al. [34] and Valstar et al. [35] use boosting algorithms which learn a series of weak learners in cascade which in combination make a strong learner. Zhou et al. [36] and Zhang et al. [37] used a coarse-to-fine cascade which starts with rough estimate of landmarks and refines the landmarks by taking subregions around landmarks. Dolar et al. [38] implements a cascading approach in order to implement a systematic landmark detection by simplifying the problem into groups which share the same structure. For example the first level identifies the pose and the next level in the cascade performs landmark detection based on the pose.

Recently RBMs in particular cascaded regressions have gained more popularity for landmark detection. This is particularly more pronounced for images with large variation on resolution, pose, occlusion, illumination (so called images in the wild). In addition, regression based approaches can be cascaded in various ways addressing the task of landmark detection in a structured way. RBMs can leverage a wide range of higher level features extracted from pixel values in order to determine landmarks from images. This particularly can be noticed by the recent rapid increase of scientific works which employ features automatically extracted by CNNs for landmark detection. Another advantage of RBMs is their ability to model shape implicitly without using parametric shape models.

2.2.3 CNN based landmark regression methods

As introduced in previous sections, CNNs have recently gained popularity in many computer vision fields achieving impressive results. Some of the application areas are object detection, classification, segmentation and localization. This popularity can be explained by the fact that many computer vision problems require manually engineered features as basis for further processing. In contrast, CNNs automatically extract features which are suitable to the task at hand. This makes the process of feature engineering unnecessary. Mirroring other trends in computer vision, there is an increase in the use of CNNs for landmark detection.

CNNs can be deployed on various landmark detection methods performing a specific task relevant for that approach (Sec. 2.2.2). For example, it can be deployed to regress directly coordinate values of landmarks from images [39–44] as in regression based methods, or a CNN can be deployed to learn parameters of a 3D appearance model [45]. It can be deployed as a local landmark expert in CLMM approaches [16]. This section focuses on the use of CNNs to regress landmark locations.

Landmark detection using CNN based regression relies on the power of CNNs to extract relevant representation of the input. Landmark detection process starts by defining individual landmarks as displacement vector (coordinate points) from a reference point in an image; or as a heat map indicating the landmark location [46]. A series of convolutional layers is used to transform the input into a transformation suitable to determine landmarks. This is followed by a fully connected layer to determine the landmarks. A loss function which calculates the error between the predicted and ground truth landmark functions is defined using mean squared error. The weights for the feature extractors and regression function will be learned by minimizing the loss function using gradient-based algorithms.

Although this landmark detection pipeline ca be implemented as a one stage process as explained on the previous paragraph, various modifications are employed to increase the accuracy of the pipeline by improving the feature extraction process or accuracy of the regression function. These approaches can be grouped in to two:

- Cascaded CNN regression
- Multi-task CNN regression

Cascaded CNN regression

In cascaded CNN regression, landmark detection is formulated as a regression task to be learned using CNN. However in order to improve the capacity of the landmark detection pipeline, the regression is done multiple times in a cascaded manner. The first regression predicts initial landmark positions. Each step afterwards improves the landmark locations by decreasing the input complexity (for example, decreasing the area of search).

A representative work from cascaded CNN regression is the work of Sun et al. [39]. In this work five facial landmarks are detected using three levels of CNN regressors following a coarse-to-fine approach. The CNN regressors in the first level in the cascade take the entire face which is in a bounding box and identify landmark locations. They use three CNN networks on the first level which operates on different parts of the whole image. One network takes the whole image; the second takes the upper facial region; and the third takes the lower facial region. All three networks are trained independently. During prediction, results from the three networks are averaged to decrease the variance of the result.

The next two levels refine a subset of landmarks predicted by the previous stage using patches of images around the predicted landmarks with decreasing area going from the second to the third level. Similar to the first cascade, multiple CNNs are trained independently per landmarks which take slightly varying size of patches around predicted landmarks from the previous level. During the prediction phase, the results from the group of CNNs at each level are averaged to reduce variance. Similarly, the region of interest for the next level is selected using this prediction and refined in the next levels.

Other works improve on aspects of the cascading process. Zhou et al. [40] uses cascaded architecture to predict 68 landmarks on facial images. Unlike Sun et al. [39], they use 4 cascades in which the first cascade predicts bounding boxes for subsets of landmarks. In addition, they incorporate a post-processing step going from the third level to the fourth in order to improve on the effects of rotation, translation and scaling of the patches around landmarks.

Zhang et al. [47] performs similar coarse-to-fine cascade of landmark detection using stacked auto-encoders. However, instead of detecting each landmark individually, they predict all landmarks together using one model at each cascade level. In order to do this, at the first cascade they train one model which predicts all landmarks from the whole image. At the following cascade levels, they take regions around landmarks predicted by the previous level and concatenate them to create one big input source which contains all patches extracted around each landmark. This approach maintains relationships between landmarks by predicting all landmarks using one model per cascade level.

Other works focus on improving the training procedure of the cascaded pipeline [41, 42]. They argue that the pipeline is not robust to poor predictions in the first levels as the cascaded models are trained individually. They aim to rectify this issues by training the whole cascade in an end-to-end manner instead of separately training individual networks for each level. Their cascaded pipeline contains three CNNs each taking a patch around a landmark which is predicted from the previous stage. During training the error between the predicted landmark and the ground truth is calculated at the end of the pipeline. The error gradient is back propagated throughout the pipeline by taking account of the interconnections between the networks at various stages. In addition, patches around landmarks are concatenated to predict all landmarks together.

Multi-task CNN regression

In multi-task CNN regression, landmark detection is considered as part of multiple heterogeneous; but related tasks such as landmark, pose, gender, facial expression identification. The accurate execution of all tasks is used to improve the accuracy of landmark detection by studying the relationship between individual tasks together can improve the performance of each task. The intuition behind this approach is that multiple tasks share the same representation learned by a CNN and jointly learning the tasks can improve extraction of accurate features; hence, improved individual task performance.

A representative work in this approach is that of Zhang et al. [43]. Here landmark detection is learned together with other auxiliary tasks such as identifying gender, pose, presence of a glass, and smile. All the tasks learn the same representation of the input by sharing the feature extraction (i.e., convolutional layers until the fully connected layer, Fig. 2.8). The feature extraction is followed by a fully connected network with a independent regression function. This optimizes the specific loss function for a particular

task using the shared feature extracted from the images.

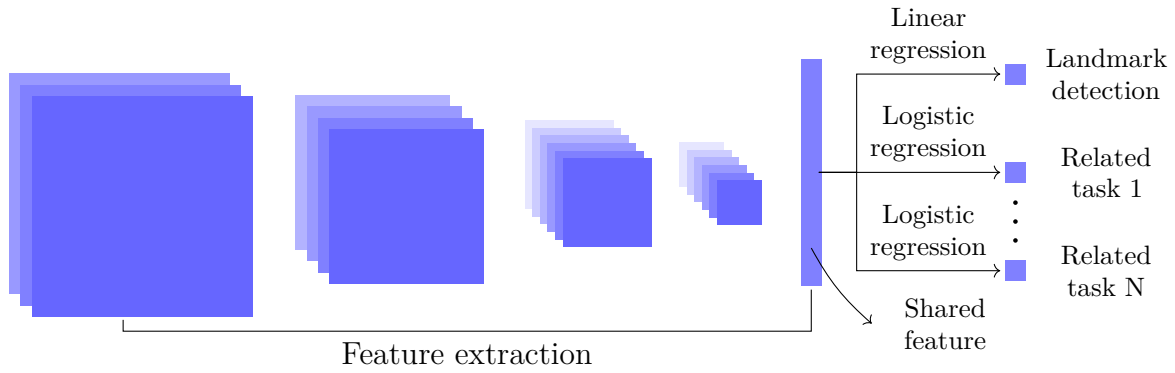


Figure 2.8: Multi task CNN: Multiple regression tasks are implemented using a shared common representation of input. The error of achieving a particular task is back-propagated to improve the shared representation of the input.

The error from each task's loss function is propagated separately until the shared feature extraction layer. The individual error gradients from each task are combined at the shared layer according to assigned weights. Zhang et al. [39] achieve a slightly better result than that of cascaded architecture discussed previously. The improvement is especially more pronounced on faces with severe occlusion (e.g. wearing glasses) and pose variations. In addition, it is demonstrated that multiple networks are not necessarily required. This improves on the computational time during training and prediction.

Other successive works build on this approach. Ranjan et al. [48] developed a framework for face detection, landmark localization, pose estimation and gender identification. In this work, instead of using shared representation only from the last convolutional layer, shared representations from multiple layers are extracted. The aim in this approach is to use global (higher level representations) and local (lower level representations) representations of the image for the shared tasks.

3 Implementation

In this chapter, a detailed descriptions of terminologies, the input data and preprocessing steps and a detailed description of two approaches followed to solve the problem of the HKA angle determination is given.

3.1 Definitions and implementation overview

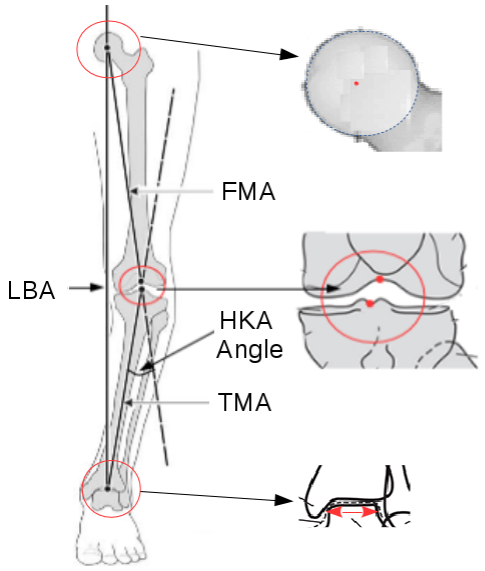
3.1.1 Definitions

The HKA angle is defined as the angle between the Femoral Mechanical Axis (FMA) and the Tibial Mechanical Axis (TBA) (Fig. 3.1a). The FMA is determined by a line drawn from the center of femoral head to the center of femoral notch [49–51]. The TMA is determined by a line drawn from the center of the tibial spine tips to the center of the ankle mortise [49–51]. The FMA and TMA together are referred as the Leg Mechanical Axis (LMA).

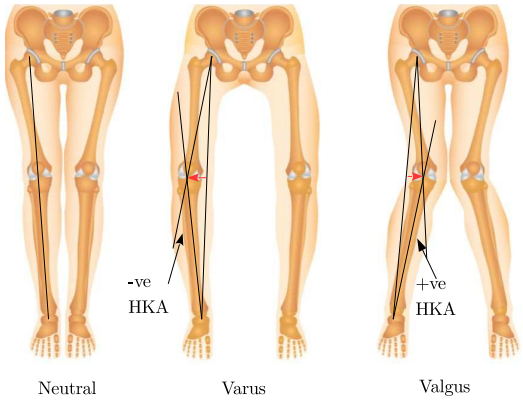
The Load Bearing Axis (LBA) is defined as the straight line from the center of the femoral head down to the center of the ankle mortise [52]. The LBA indicates the axial line through which the weight of the upper body exerts force to the leg anatomical structure. As seen in the Fig. 3.1b, this line ideally passes through the middle of the knee in which it is equally distributed across the knee. As the center of the knee (i.e the LMA) deviates away from the LBA, the distribution of load on anatomical structures of a knee becomes unbalanced.

As can be seen from Fig. 3.1a, the HKA angle measures the degree of misalignment of LMA from the LBA. The sign convention in which valgus is assigned positive and varus negative values can be used as a simple rule of thumb to convey the observation that varus deviations create more damaging imbalance of load distribution on the knee [50]. The widely used cut off values for clinical Varus-Neutral-Valgus classification are $2°$ and $-2°$ according to the following rule [50, 51]:

- Varus: HKA angle $\leq -2°$
- Neutral: $-2° <$ HKA angle $< 2°$
- Vargus: HKA angle $\geq 2°$.



(a) HKA angle landmarks and definitions: The femoral head is the center of a unique circle which fits the circular edge of the femoral head, the knee centers are defined by two landmarks at femoral notch and tibial spine tip, the center of talus bone is the midpoint between the lateral and medial edges of the talus bone.



(b) HKA angle conventions: A knee center which lies on the LBA indicates a neutral alignment. A knee center which is found laterally in relation to the LBA indicates varus deviation and is assigned a negative HKA angle value. A knee center which is found medially in relation to the LBA indicates valgus deviation and is assigned positive HKA angle value. Picture adapted from [3].

Figure 3.1: HKA angle determination and sign conventions.

3.1.2 Conventions followed in the thesis

Although the HKA angle is determined for both the left and the right angle, throughout this thesis the HKA angle refers, unless explicitly specified, to both the right and the left side. When left and right HKA is distinguished explicitly, the sides refer from the perspective of the subject in the image. This means that in anterior view radiographic image the right side is located on the left side of an image. In terms of implementation an HKA angle is determined using one model for both the left and the right angle.

Since human anatomy is symmetrical (at least in the case of radiographic full limb images), a method can be developed to determine one sided values such as an HKA angle of the right leg, but can be used for both legs. This is done by flipping the image sideways and determining the left sided HKA angle as if it is the right leg (Fig. 3.2). In this thesis, a method is developed to predict an HKA angle for the right leg in a particular image. The right leg's HKA angle is determined from original image, and left leg's HKA angle is determined by flipping the original image sideways.

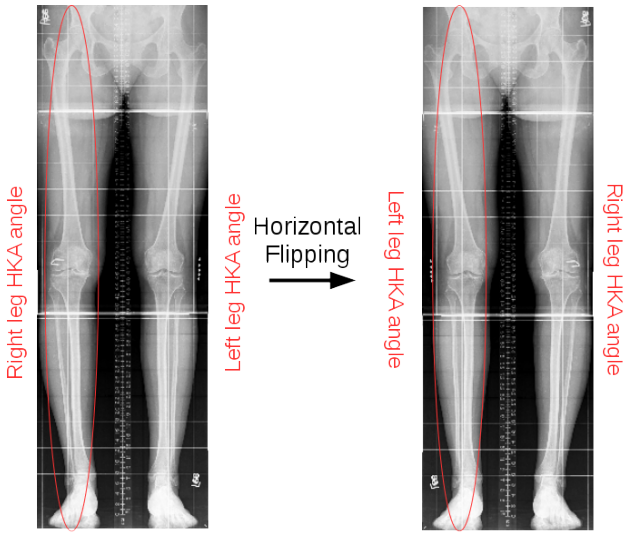


Figure 3.2: Right and left leg HKA angle determination: In the original image a method trained to determine the right leg's (the region inside the red ellipse) HKA angle given an image is used to find the right leg's HKA angle. The same method can be used to determine the left leg's HKA angle by flipping the image.

3.1.3 Automated HKA angle determination process overview

The flowchart in Fig. 3.3 gives an overview of the process pipeline designed to find an HKA angle. An input image is preprocessed to account for image variations which arise from image acquisition, processing and presentation (visualization) formats. This is followed by ROI selection to determine the relevant image section for further processing. To determine the HKA angle, two directions using CNNs are developed.

The first approach uses the power of CNNs to predict an HKA angle in a very direct manner by learning a mapping function between images and the HKA angles. In this approach the CNN model can be considered as a black box system which takes an image as input and calculates the HKA angle as an output.

The second approach uses a more systematic approach to reduce the complexity of the task in a way which imitates how experts determine the HKA angle. This approach first finds anatomical landmarks which define the femur and tibial mechanical axis using CNNs. The HKA angle is then measured as the angle created by the intersection of these two lines as is shown in Fig. 3.1b. In the following sections and chapters these two approaches are referred as HKA angle determination from images and HKA angle determination from landmarks, respectively.

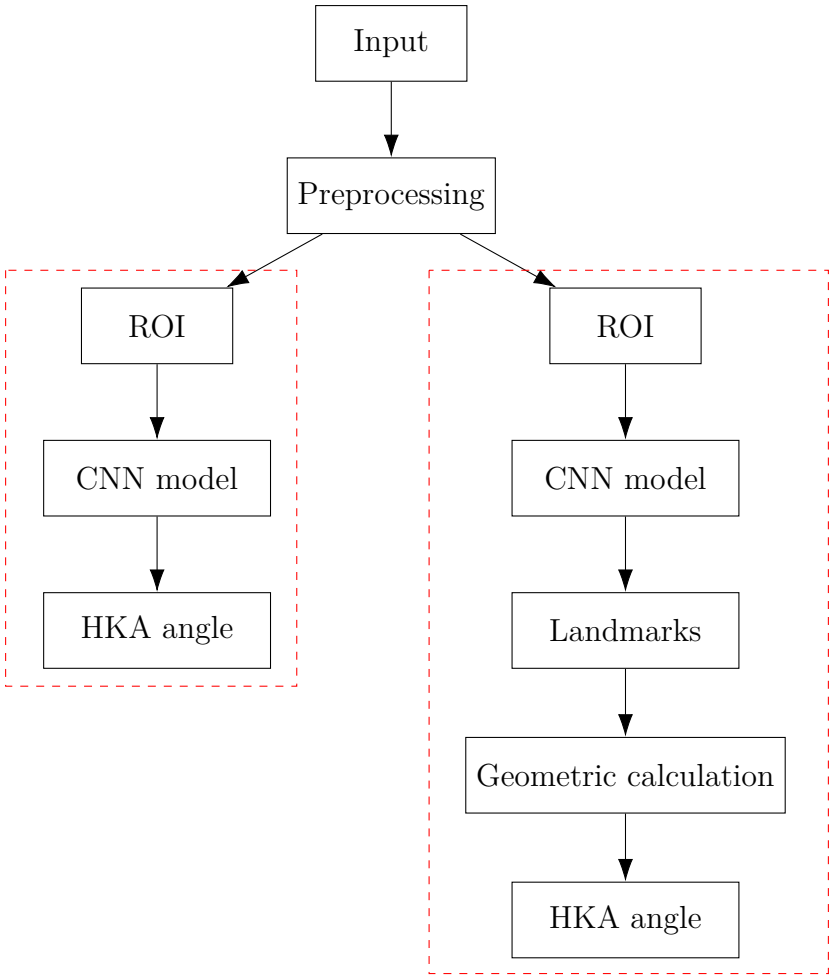


Figure 3.3: Process overview of HKA angle determination: Left branch indicates the process flow for determining HKA angle from images and the right branch indicates HKA angle determination from landmarks. The separation after preprocessing phase is to emphasize that preprocessing steps use the same tools, and that different approaches use different ROI as input for the CNN models.

3.2 Input and preprocessing

3.2.1 Input images

The input image is a frontal plane bilateral radiographic images of the lower limb in its entirety (Fig. 3.4). They are often referred as full limb radiograph. The radiographic images are taken with knees fully extended. Positional reliability of anatomical structures is ensured by following a detailed instruction manual about patient and X-ray tube positioning [53]. The radiographic image data are stored using Digital Imaging and Communications in Medicine (DICOM) standard.

Figure 3.4: Input image example: The image shows a typical lower limb radiograph.

3.2.2 Preprocessing

The training data consists of varying image dimensions and pixel value ranges. Therefore, the first step taken in the HKA angle determination pipeline is a preprocessing step to standardize the image dimensions and pixel values. For this purpose an analysis of image dimensions is performed. This analysis shows that there is a significant variation in the dimension of images (Appendix A). This can be attributed to both natural size variations of the subjects and systematic variations of image size caused by procedures in image acquisition and processing steps.

The systematic difference is particularly noticeable in the width of images. The majority of images have a dimension of around 34 to 40 cm, a width enough to cover the physical dimension of the subjects. However, a number of images have a width of around 110 cm. In such images the majority of the image is a non anatomical blank space, and the relevant anatomical section is in the middle. An example of such image can be seen in Fig. 3.5a. To alleviate this a heuristic rule is implemented to crop images to a maximum size of 40 cm (Fig. 3.5a). The heuristic rule defines this 40 cm image section as a region of the image 20 cm to the left and to the right from the middle point along the width of an image. This region contains all the relevant anatomical information inside an image.

The variation of height of images falls between 120 cm and 130 cm. This is often due to height variations of the subjects but as well some inconsistencies in image acquisition (some images contain anatomies up to the chest of a subject). There is no observable systematic variation along the height similar to the one along the width. Therefore there is no preprocessing done to remove the height variation in images.

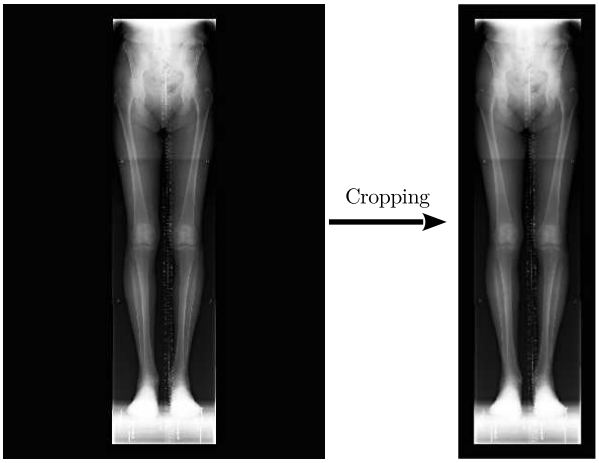
In addition to height variations, there is a variation of pixel value ranges within the input

data. This is to say that the minimum and maximum values in the pixel value vary from one image to another. This can be attributed to exposure variations in X-ray acquisition (x ray pixel values ranges relate to exposure levels) or X-ray instrument specifications such as its dynamic range.

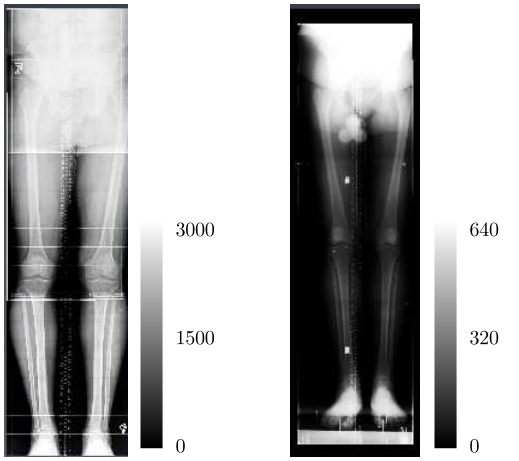
When viewing the X-ray using a medical image software, this can be easily alleviated by using appropriate gray value window which adjusts the range of pixel values which are mapped to the color range. CNN model training computations can be made more efficient by normalizing this variation. To this end, the pixel values are normalized to a minimum of zero and maximum of one using a linear interpolation of the pixel value ranges to a range between zero and one using the following functions,

$$V_{x,y}^{new} = (V_{x,y}^{old} - V_{min}^{old}) \cdot \left(\frac{1}{V_{max}^{old} - V_{min}^{old}} \right), \tag{3.1}$$

where $V_{x,y}^{new}$ and $V_{x,y}^{old}$ normalized (new) and pre-normalized (old) pixel values, V_{max}^{old} and V_{min}^{old} the maximum and minimum pixel values in the pre-normalized (old) image.



(a) Size normalization: The width of an image is cropped to only include the relevant anatomical part.



(b) Pixel value range: Variation in pixel value range is seen in the color-map legend. This range is linearly scaled to 0 and 1.

Figure 3.5: Image preprocessing

Another source of pixel value variation is the photometric interpretation attribute of the DICOM file. This attribute defines the relationship of pixel values with X-ray attenuation character of an object. Often the highly attenuating structures such as bones are given higher pixel values and soft tissues given low pixel values. The high pixel values and the low pixel values are mapped to a white and black color on visualization softwares. This means that in most images the bony structures will be white and the soft tissue and background will be gray.

However, in some images this mapping is reversed, i.e. bony structures are given low pixel value (darker color) in comparison with to soft tissues. This variation is removed

by making all images follow the common photometric interpretation of high pixel values for highly attenuating structures. This is done easily by controlling for photometric interpretation of image's DICOM header and inverting the pixel values. This means changing the smallest values to the maximum value and vice versa (Fig. 3.6). The pixel values in between are inversed proportionally. This is done using the following equation:

$$V_{x,y}^{new} = -1 \cdot (V_{x,y}^{old} - V_{max}^{old}). \tag{3.2}$$

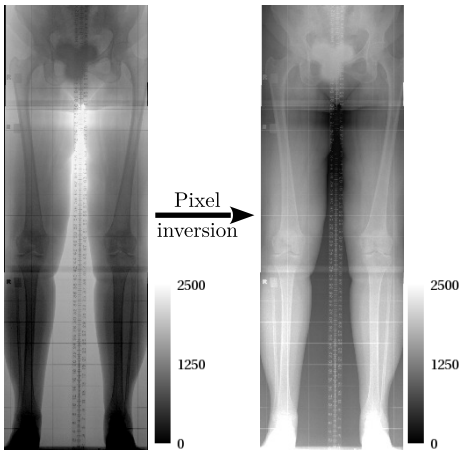


Figure 3.6: Pixel value inversion: The low pixel values are inverted to high and vice versa. the difference can be seen where in the first image the bone is gray (low pixel value) and in the second it is white (high pixel value).

3.3 HKA angle determination from images

This approach formulates the HKA angle determination as a regression problem with full limb X-ray images as input and an HKA angle as output. The regression function is modeled using a CNN. To achieve this task of HKA angle prediction from a full limb X-ray, a process pipeline is designed (see the left branch of Fig. 3.3).

3.3.1 ROI selection

Since the HKA angle depends on the configuration of femoral head, knee and talus bone, the ROI is selected to include all the aforementioned anatomical structures. This means that the network needs to see the entire full leg to determine the HKA angle, i.e. the full lower limb radiograph or at lease the full structure of one leg whose HKA angle is of interest (i.e. half side along the width of an image). This approach deploys CNNs in a black box manner, as a result it is not easy to know what information from an image a network uses to determine an HKA angle. This is due to the large number of parameters

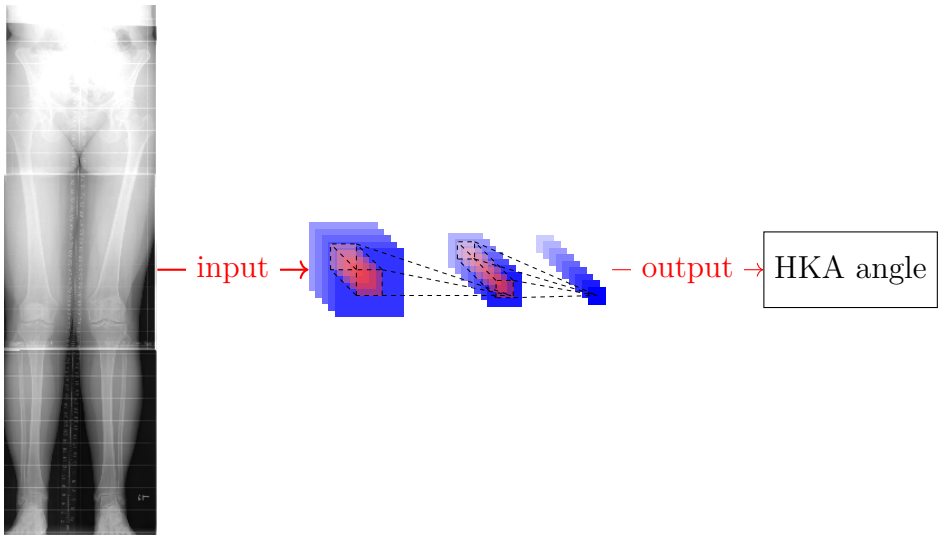


Figure 3.7: HKA determination from images: A CNN model is trained which takes frontal full limb X-ray as input and predicts an HKA angle for the image.

in a typical deep CNN model. Since many leg alignment issues usually occur similarly on both sides, it is possible a network can leverage information from both the left and the right side of a leg even if the prediction is made only for one side.

Accordingly, two ROIs are selected one containing a full size image and another half side of an image as input. Two models are trained one taking the full image as input and the second taking half of an image (along the width). The full size images are resized to a 256 by 1024 pixel and half sided images are resized to 128 by 1024 pixels to reduce the computational demand.

3.3.2 Training data labels: HKA angles

The HKA angle values are in degrees. The HKA angle values used for training are determined by two experts with a few images only labeled by one expert. In the case of both experts labeling a patient the arithmetic mean HKA angle from the two experts is used as a ground truth HKA angle value. If the HKA angle of an image is determined by only one expert we use that value. The choice to use the mean of the two experts' HKA angle is based on the assumption that in cases where there is an error the mean of the two evaluators will be closer to the actual HKA angle.

3.3.3 CNNs for HKA angle regression

To my best knowledge, there is no previous work which attempted to determine HKA angle using CNN. Therefore the network architecture design is performed independent of related work. For this approach a simple CNN architecture with multiple convolutional layers is chosen (Fig. 3.8).

This choice is based on the fact that this structure is the basic building block behind

the power of CNNs ability to learn highly nonlinear patterns. These bundle of convolution layers is followed by a fully connected layer that implements a regression function to determine HKA angle using the features extracted by preceding convolutional layers.

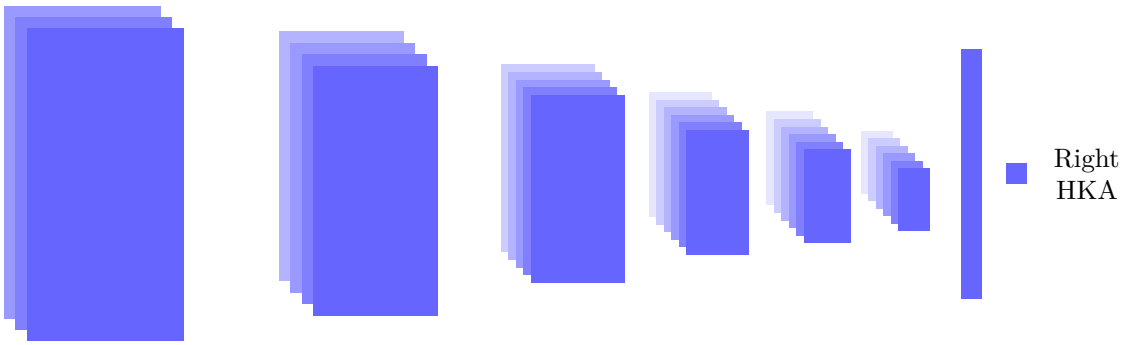


Figure 3.8: CNN model HKA angle prediction from images: The CNN model is composed of multiple layers of convolutional layers followed by a fully connected layer and regression function.

Table 3.1 presents the detailed description of the two CNN models deployed. Each convolution layer is composed of one convolution operation, ReLU activation and max pooling. In both models, there are six convolutional layers and one fully connected layer with a 40% dropout rate. This is followed by a single output neuron performing a linear regression function giving HKA angle. In the first convolution layer a 5×5 convolution kernel is deployed to sample a large local information (bigger field of view). The subsequent convolution operations use 3×3 convolutional kernel.

The size of CNN model parameters such as the number of feature maps and depth of network (e.g. multiple convolutions before pooling) architecture is minimized due to the limited amount of training data. In addition to the limited amount of original data, the shape of the input image played a role in the choice of parameters. For example, the dimensions of a full limb image are rectangular with height to width ratio of 4:1 in the a full image, and 8:1 in one half sided image. This imbalance on the width and height of image dimension influenced the choice of the pooling kernel size.

When using the popular 2×2 pooling, 128×1028 pixel image will be down sampled to 1×8 sized feature map after seven pooling operations. This means that there are eight times more parameters summarizing information across the height of an image compared to the width of an image. This also means each pixel at the last feature map have an effective field of view eight times lower along the height than along the width of an image (Fig. 2.5c). This decreases the representational capacity of the last feature maps and with it the accuracy of its predictive power. To compensate for the difference between the height and width, a rectangular (1×2) max pooling is used in the case of the second model.

Model 1 (Full Image)	$32 \times 5 \times 5$ C 2×2 M	$64 \times 3 \times 3$ C 2×2 M	$128 \times 3 \times 3$ C 2×2 M	$256 \times 3 \times 3$ C 2×2 M	$256 \times 3 \times 3$ C 2×2 M	$512 \times 3 \times 3$ C 2×2 M	512 FC 40% D	LR
Model 2 (Half Image)	32×5×5 con. **1×2 M**	64×3×3 con. **1×2 M**	$128 \times 3 \times 3$ C 2×2 M	$256 \times 3 \times 3$ C 2×2 M	$256 \times 3 \times 3$ C 2×2 M	$512 \times 3 \times 3$ C 2×2 M	512 FC. 40% D	LR

Table 3.1: Detailed network architecture (FC = fully connected layer, C = Convolution, M = Max pooling, D = dropout, LR = linear regression).

3.3.4 Training process

The network is trained using using SGD algorithm. Three learning rates are chosen 0.005, 0.001 and 0.0001 with each learning rate applied for 100 epochs with an early stopping after 10 epochs. This means that a training starts with a learning rate of 0.005 and continues for maximum of 100 epochs. If the network does not improve its performance over the validation set for 10 epochs the learning rate is decreased towards the next small learning rate. To improve convergence rate of training batch normalization is used and drop out at the last connected layer is used as a regularizer to reduce over-fitting.

3.4 HKA angle determination from landmarks

The second direction followed is a structured evaluation of the HKA angle which imitates the analysis of human experts to determine an HKA angle from images. The determination of the HKA angle by experts follows the following steps: find the center of the femoral head by finding the center of the femoral head, the knee and and the ankle. From these three points, the HKA angle is determined by measuring the angle between the lines connecting these points (Fig. 3.1).

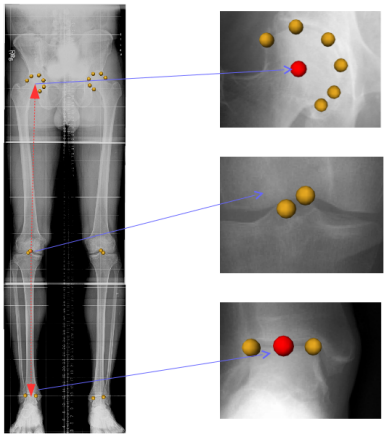


Figure 3.9: HKA angle determination from landmarks: In this approach landmarks are first determined defining the position of the relevant location for further processing to determine HKA angle.

Imitating these steps, six equally spaced landmarks are determined on the edges of the femoral head (Fig 3.9). The center of the femoral head is then determined by finding the center of a circle which best fits the landmarks on the edge of a femoral head, i.e. an

optimization task which finds a circle which best fits the circular edge of the femoral head. This task is defined mathematically as minimizing the euclidean distance of landmarks to a unique circle defined by radius r and a center with coordinates (x_c, y_c). The cost (error) function that is minimized E is calculated as follows:

$$E = \sum_{i=1}^{n} (d_i - r)^2 , \tag{3.3}$$

where d_i is the distance of a landmark i from the center, and it is defined by $d_i = \sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}$.

Various alternatives are found in literature defining the center of the knee [50]. Some use only one landmarks at center of the knee defining both the proximal and distal tips of the FMA and TMA, some others use two landmarks one for FMA and one for TMA. In this thesis work, two knee landmarks are used defining the distal and proximal points of the FMA and TMA tips. These points are visually identifiable structures at the knee radiograph (Fig 3.9).

Two landmarks are placed at the visually identifiable edge of the talus bone from which the center of the talus bone is calculated as the mean of the coordinates of these two landmarks. Two vectors connecting these four landmarks are determined. One connecting the center of the femoral head and the femoral notch of the knee, and another connecting the tibial notch of the knee and the center of talus bone. The angle from these four landmarks is computed using the following formula:

$$HKA\ angle = tan^{-1}\left(\frac{x_h - x_{kf}}{y_h - y_{kf}}\right) - tan^{-1}\left(\frac{x_a - x_{kt}}{y_a - y_{kt}}\right), \tag{3.4}$$

where (x_h, y_h), (x_{kf}, y_{kf}), (x_{kt}, y_{kt}) and (x_a, y_a) are the landmark coordinates of the femoral center, femoral notch of the knee, tibial tip of the knee and the center of the talus bone respectively.

The choice to use six landmarks on the edge of femoral head instead of directly determining the center of femoral head is based on the fact that there is a clearly defined edge on the femoral head. This means that accurate landmarks can be placed by a human expert. As early layers in CNNs detect edges this can be beneficial to predict landmarks on the edges instead of directly predicting the center of the femoral head. Similar logic was followed with two landmarks at the knee and at the ankle where landmarks are placed on visually defined edges on the corresponding body parts.

3.4.1 ROI selection

The fact that landmarks are determined first and HKA calculated from the landmarks means that the entire image does not need to be considered, rather only the region relevant for the landmarks. As landmark determination depends heavily on patterns found locally

around the landmark, the ROI relevant for each landmark group is only a subsection of the entire image around those group of landmarks. For example, to determine landmarks at the femoral edge the part of a full limb radiograph around the ankle does not play a significant role.

Therefore, if it is possible to find an approach to determine the region where landmarks are more likely to be found, we can narrow our landmark search region reducing variation which occurs with bigger region of interest and improving computational effort needed to find landmarks.

In this regard, two approaches are used to find initial ROIs around the femoral head, the knee and the ankle joint. One uses heuristic rules to define the initial ROI, and the second approach uses CNN based ROI detection. Using this ROI a cascaded landmark detection in a coarse to fine fashion is implemented. These two ROI determination methods are discussed in the following subsections.

Heuristic rule based initial ROI selection

The first approach exploits the observation that medical images are acquired in a standardized setting in order to make heuristic rule to define ROI for landmarks. Accordingly the top left quarter of the image contains one side of the femoral head, and the lower quarter of the image contains the ankle. The knee region is found in the middle section of the image which is defined as 25cm above the lower end of the image and including half size of the image height above (Fig. 3.10 for the case of hip landmark detection). This quarter of an image is resized to a 256×1024 pixel size and primary determination of landmarks is made on this ROI.

The second level of landmark determination re-samples a 512 (width) by 1024 (height) pixel region from the original image by taking center of the landmarks predicted on the first level. This is resized to a 256×512 pixel to decrease memory and time taken during training. Similarly, in the third level a 512×512 pixel is taken from the original image by taking the center of landmarks from the second level. The center of landmarks is defined as a mean value of (x, y) coordinates of landmarks.

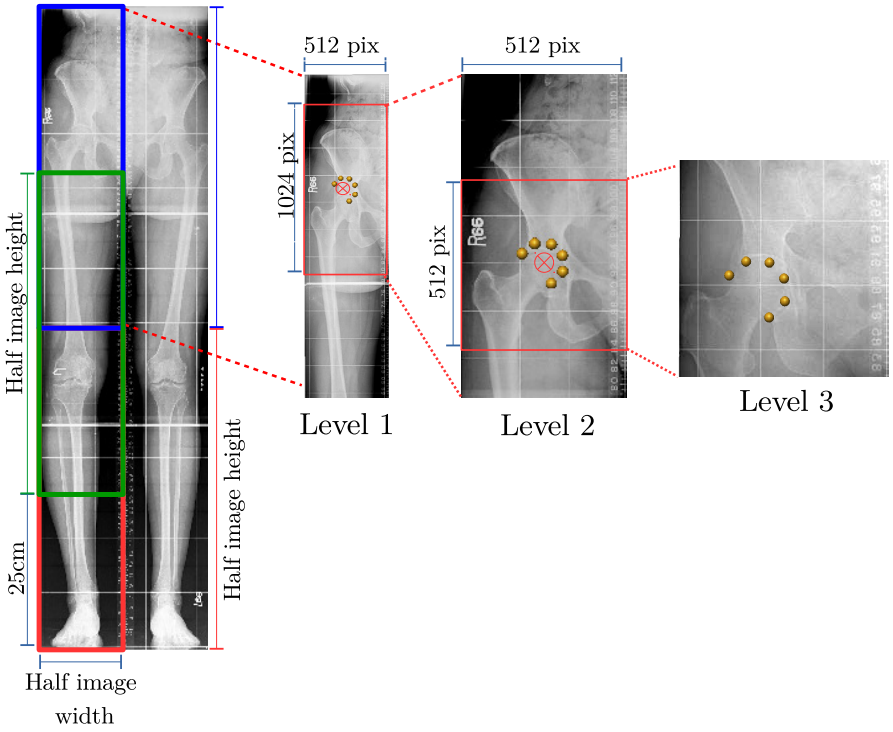


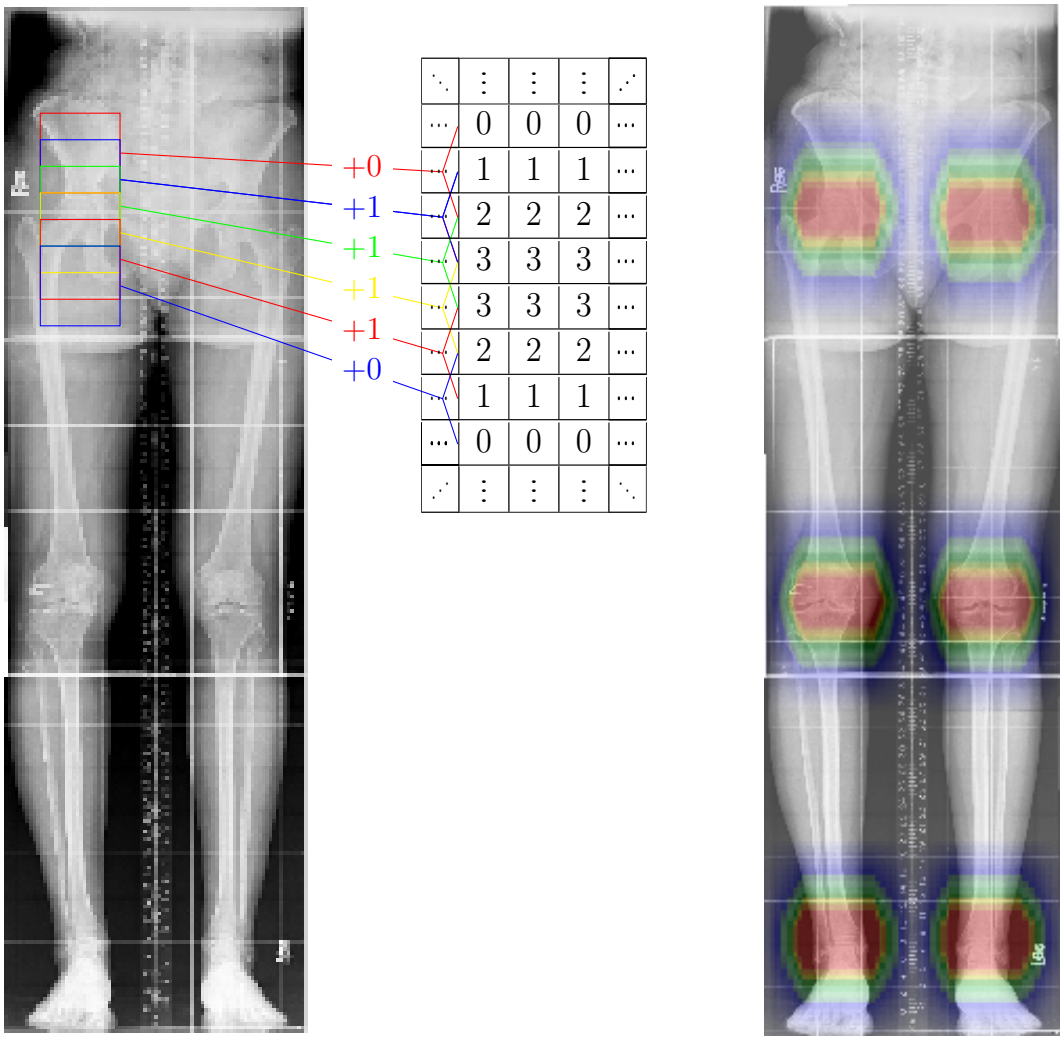
Figure 3.10: A three level cascaded hip landmark detection from heuristically defined initial ROI: On the left most image the blue, the green and the red lines define initial ROIs for the femoral head, the knee and the ankle landmarks respectively.

Automated initial ROI selection

The second approach detects ROI automatically using CNN. This is implemented as CNN based object classifier which is evaluated across the image using a sliding window approach. The classifier takes an area from an image defined by a sliding window and classifies as containing a femoral head, a knee joint or an ankle joint.

The result of the classification is aggregated into a heat-map by adding the result of classification pixel-wise (Fig. 3.11). For example, all the pixels corresponding to a window which is classified as containing a femoral head are added a value of one, and those in a window classified as non femoral window are added a value of zero. This sliding window is traversed across the image by sliding 100 pixels to define a new region.

The heat-map overlaid on the original image indicates the regions which are likely to contain a femoral head, a knee or an ankle joint (Fig. 3.12). The center of a ROI is determined by calculating the center of mass of each area labeled as containing femoral head, knee and ankle in the heat-map. The initial ROI is defined as a 512×512 pixels around the center of mass on the heat-map. Using this automatically detected ROI, landmarks are predicted in a cascaded manner.



Example image

Heat map overlayed

Figure 3.11: Sliding window object classifier: A window of size 3×3 (this is for illustration but in the actual implementation it is 324×324) is traversed across the image classifying if a window contains femoral head. The result is added into a heat-map which is equal in size to the image. The heat-map overlaid on the original image can be seen on the right image.

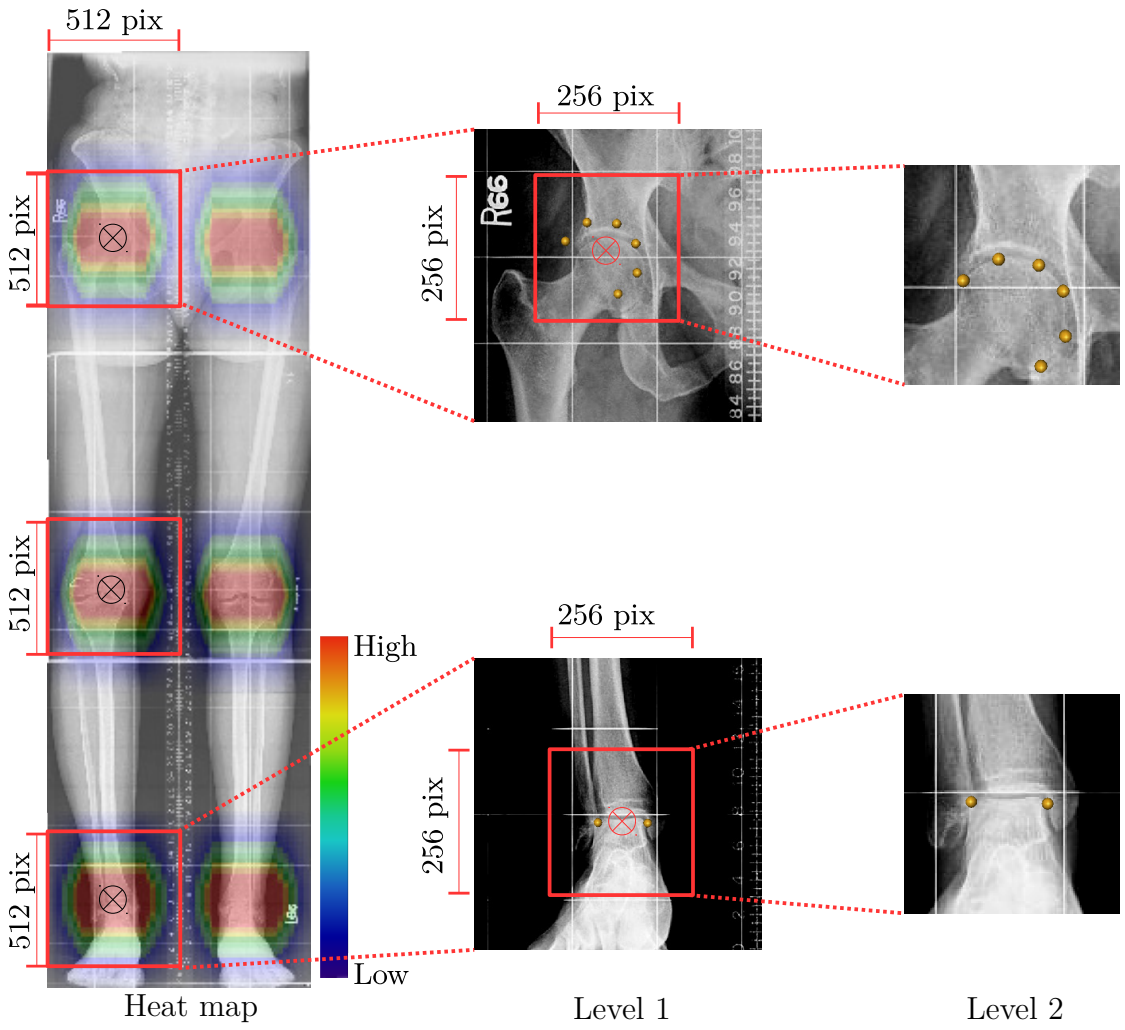


Figure 3.12: Automated intial ROI and cascaded landmark detection: A two level cascaded hip and ankle landmark detection from CNN based ROI selection; the left side shows heat-map result of ROI detection overlaid on the original image. Using a ROI determined a cascaded landmark detection is implemented.

In a similar fashion to Section 3.4.1, two level cascaded landmark detection for the femoral head and ankle joint landmarks is implemented. Knee landmarks are determined using a three level landmark detection. The first level samples 512×512 pixels around the center of mass of a heat-map and resizes it to 256×256 pixels. The second level samples 256×256 pixels around the center of landmarks determined in the first level and resizes it to 128×128 pixels. The third level for knee landmarks samples 128×128 pixels around the center of landmarks determined in the second level.

3.4.2 Training data labels: Landmarks

The training and testing data for landmark detection are sub-regions around the landmarks referred to as ROIs as described in Section 3.4.1. The size of a ROI depends on the approach (heuristic or automated initial ROI) and level of the cascaded landmark detection pipeline.

These ROIs are labeled with coordinate value of landmarks contained within it. The landmark coordinates indicate distance of a landmark on the x-axis and y-axis where the origin $(0, 0)$ is defined as the top-left corner of the ROIs (Fig. 3.13). The coordinate values are normalized to between 0 and 1 linearly.

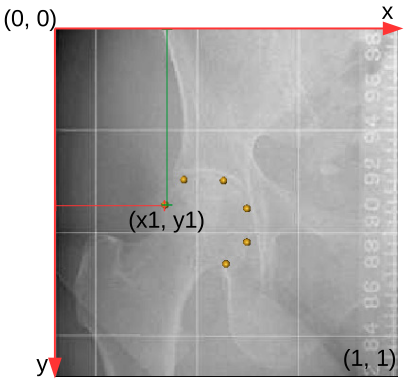


Figure 3.13: Labelling of training data for landmark detection: The coordinate values of landmarks are concatenated together to make a vector of coordinate values. (Vector size: 12 for the femoral head, 4 for the knee joint and 4 for the ankle.)

3.4.3 CNNs for landmark determination

The CNN models that are used for the landmark detection are similar to the CNN models that are used for the HKA angle determination from images. They are composed of the basic convolutional layers with increasing number of features maps followed by a one layer of fully connected network(Fig. 3.8).

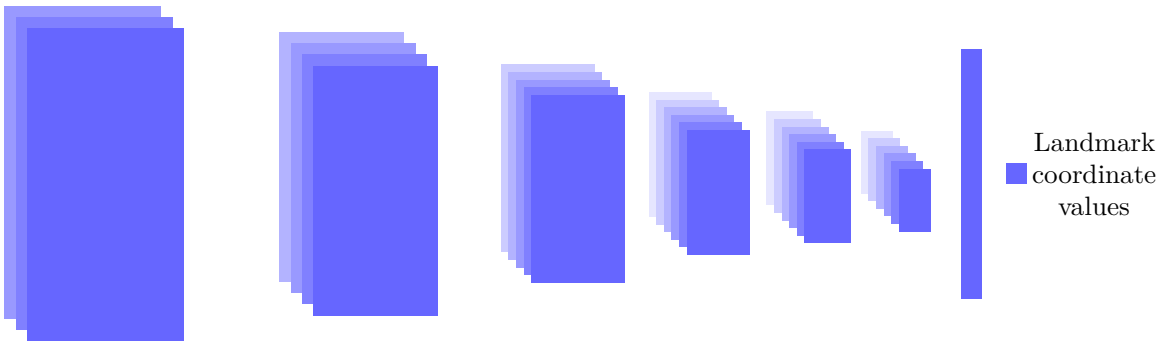


Figure 3.14: CNN model for landmark determination: The CNN model is composed of multiple layers of convolutional layers followed by a fully connected layer and regression function.

Table 3.2 presents the detailed description of the CNN models deployed.Each convolution layer performs two convolutions using 3×3 kernels before applying nonlinear (ReLU) activation. the last convolutional layer is followed by a fully connected layer with a 40%

to 50% dropout rate. The last layer performs linear regression on the output of the fully connected layer. Landmarks of each region are regressed together. This is to say that the last output is the coordinate values of 6 landmarks on the x-axis and y- axis (12 values in total). The CNN used for ROI classification follows the same architecture except it uses a soft max activation at the last layer.

Purpose	Convolutional Layer	Convolutional Layer	Convolutional Layer	Convolutional Layer	Convolutional Layer	Convolutional Layer	Fully Conected Layer	Output Layer
ROI classifier	$32 \times 3 \times 3$ C $32 \times 3 \times 3$ C 2×2 M	$64x3x3$ C $64x3x3$ C 2×2 M	$128x3x3$ C $128x3x3$ C 2×2 M	$256x3x3$ C $256x3x3$ C 2×2 M	$512x3x3$ C $512x3x3$ C 2×2 M	-	1024 FC 50%	SM
Heuristic L1 Landmarks	$32 \times 3 \times 3$ C $32 \times 3 \times 3$ C 2×2 M	$64 \times 3 \times 3$ C $64 \times 3 \times 3$ C 2×2 M	$128 \times 3 \times 3$ C $128 \times 3 \times 3$ C 2×2 M	$256 \times 3 \times 3$ C $256 \times 3 \times 3$ C 2×2 M	$512 \times 3 \times 3$ C $512 \times 3 \times 3$ C 2×2 M	$512 \times 3 \times 3$ C $512 \times 3 \times 3$ C 2×2 M	1024 FC 40% D	$N_l \times 2$ LR
Heuristic L2 Landmark	$32 \times 3 \times 3$ C $32 \times 3 \times 3$ C 2×2 M	$64 \times 3 \times 3$ C $64 \times 3 \times 3$ C 2×2 M	$128 \times 3 \times 3$ C $128 \times 3 \times 3$ C 2×2 M	$256 \times 3 \times 3$ C $256 \times 3 \times 3$ C 2×2 M	$512 \times 3 \times 3$ C $512 \times 3 \times 3$ C 2×2 M	$512 \times 3 \times 3$ C $512 \times 3 \times 3$ C 2×2 M	1024 FC 40% D	$N_l \times 2$ LR
Heuristic L3 Automated L1 Landmark	$32 \times 3 \times 3$ C $32 \times 3 \times 3$ C 2×2 M	$64 \times 3 \times 3$ C $64 \times 3 \times 3$ C 2×2 M	$128 \times 3 \times 3$ C $128 \times 3 \times 3$ C 2×2 M	$256 \times 3 \times 3$ C $256 \times 3 \times 3$ C 2×2 M	$512 \times 3 \times 3$ C $512 \times 3 \times 3$ C 2×2 M	$512 \times 3 \times 3$ C $512 \times 3 \times 3$ C 2×2 M	1024 FC 40% D	$N_l \times 2$ LR
Automated L2 Landmark	$32 \times 3 \times 3$ C $32 \times 3 \times 3$ C 2×2 M	$64 \times 3 \times 3$ C $64 \times 3 \times 3$ C 2×2 M	$128 \times 3 \times 3$ C $128 \times 3 \times 3$ C 2×2 M	$256 \times 3 \times 3$ C $256 \times 3 \times 3$ C 2×2 M	$512 \times 3 \times 3$ C $512 \times 3 \times 3$ C 2×2 M	-	1024 FC 40% D	$N_l \times 2$ LR
Automated L3 (only for the knee)	$32 \times 3 \times 3$ C $32 \times 3 \times 3$ C 2×2 M	$64 \times 3 \times 3$ C $64 \times 3 \times 3$ C 2×2 M	$128 \times 3 \times 3$ C $128 \times 3 \times 3$ C 2×2 M	$256 \times 3 \times 3$ C $256 \times 3 \times 3$ C 2×2 M	$512 \times 3 \times 3$ C $512 \times 3 \times 3$ C 2×2 M	-	1024 FC 40% D	$N_l \times 2$ LR

Table 3.2: Detailed architecture of Landmark determination CNN models: L = Level, C = Convolution, M = Max pooling, FC = Fully connected layer, D = Dropout, LR = Linear regression, SM = Softmax, N_l = Number of landmark coordinates.

3.4.4 Training process

The training processes follows the same scheme as the one for HKA angle models. The network is trained using using SGD algorithm. Three learning rates are chosen 0.005, 0.001 and 0.0001 with each learning rate applied for 100 epochs with an early stopping after 10 epochs.To improve convergence rate of training batch normalization is used and dropout at the fully connected layer is used as a regularizer to reduce over-fitting.

Training data augmentation

In order to increase the amount training data, multiple data augmentation technique are recommended [54]. Similarly, we use translation, rotation and scaling as our image augmentation technique. Translation refers to translating the window defining a ROI around the landmark randomly in order to create landmark locations that are distributed throughout ROIs. Rotation refers to rotating the original image by a small random rotation degree within -5° to 5° range before ROI is selected. This aims at creating training data that is representative of any inclined anatomical structures either due to natural variation or image acquisitions steps.

Scaling refers to zooming in and out of each image to a random amount of (between 90% and 110%) before ROI is selected. This introduces training data which reflects well

natural anatomical size differences. The cut of values of 90% and 110% value is used as limit for scaling, so as not to introduce very extreme magnification or reduction in size of anatomical structures than would have happened naturally. We use the augmentation amount for each level as shown in Table 3.3.

	Level 1	**Level 2**	**Level 3**	**Level 4**	**Level 5**
Translation	30	30	40	30	30
Rotation	10	10	10	10	30
Scaling	20	20	10	20	30

Table 3.3: Augmentation details for cascaded landmark detection: Each number indicates the number of augmented image generated using the augmentation technique specified in the row.

4 Experiments and Results

To test the performance of the methods proposed in Chapter 3, experiments are conducted. This chapter presents the results of these experiments and evaluation of the results.

4.1 Experimental setup

4.1.1 Experimental data

The experiments are conducted on full limb radiographs from the Osteoarthritis Initiative (OAI) [55, 56]. The OAI is a longitudinal observational study of knee osteoarthritis (OA) conducted by multiple study centers sponsored by the National Institutes of Health[1]. It aims to provide researchers with a database containing information on the progression of the knee OA using imaging and data on biochemical biomarkers in order to improve the understanding, prevention and treatment of the knee OA.

In the longitudinal study, one of the phenomena investigated in relation to knee OA is leg alignment. There are various ways to asses leg alignment but the 'gold standard' measure of leg alignment is done using a full lower limb radiograph (Section 1.1.1). Information on leg alignment of participants of the longitudinal study is collected by taking a full limb radiograph of the study participants once during the course of the study.

The time point of the image acquisition varies with most of the full limb X-ray taken at the early stages of the longitudinal study as seen in Table 4.1. Using the time of acquisition, the full limb X-rays in the database are structured as 12, 24, 36 and 48 month indicating the timespan within the longitudinal study in which the full limb X-ray is acquired. Unless there are cases where the quality of an initial image is assessed to be not good, in which case a second full limb X-ray is obtained on follow up visit, there is only one full limb radiograph per patient.

The participants of the study are divided as progression: showing symptoms of OA, incidence: high risk of developing OA and control sub-cohort by the study design [57]. The HKA angle determination from the radiographs is done by two reading centers: Dr. Derek Cooke's (OAI project 60) and Dr. Jeff Duryea's (OAI project 32). Table 4.1 shows the number of participants evaluated by the two reading centers and the time span in which the image is taken.

The number of participants evaluated by Dr. Duryea's reading center is higher than those by Dr. Cooke's because Dr. Duryea's reading center evaluated the HKA angle

[1] `https://www.nih.gov/`

for the control sub-cohort of the participants in the study. Each participant is given a unique anonymous identification number. Any repeat imaging can be found with the same number but with a different time of acquisition label (12, 24, 36, 48 months).

For our experiments we use only the HKA angle and its associated radiographic image of a participant. Therefore the timing of image acquisition and any associated information about the participants OA diagnosis is not important for this work. Similarly, for our purpose to which cohort a particular participant of the study belonged is not relevant.

Visit at which HKA angle is acquired	Dr. Duryea	Dr. Cooke
12-month visit	1432	1237
24-month visit	1261	1275
36-month visit	914	911
48-month visit	176	132
Total	3783	3555

Table 4.1: Overview of experiment data: The table shows the number of full lower limb radiographs that are evaluated by Dr. Duryea and Dr. Cooke, and when along the longitudinal study they are acquired.

Some observations on the data

The lower limb radiographs are taken according to the imaging manual for the acquisitions of the radiographs [53]. Despite the manual, there are some quality issues in the full-limb X-ray dataset. Some example of such issues can be seen in Fig. 4.1. One of the common issues is an improperly placed radiation protection. The imaging manual for the acquisitions of the radiographs [53] recommends the use of lead radiation protection to cover the gonads from radiation.

However, this protection is sometimes placed at the wrong location covering the femoral head completely in the worst cases. Another issue is that some images contain medical implants and other non anatomical objects. These implants can be found directly on the leg anatomy (hip and knee replacements and bone plates), or close to the leg anatomy such as medical implants which are not related to the leg anatomy, jewelery and other unidentified objects.

Some images having very low image contrast to the level that it is difficult to exactly spot the relevant anatomical landmarks. This variation in contrast can be due to imaging acquisition settings in consideration of radiation dose or due to variations in participant body weight which determines the level of radiation passing to detector.

In addition to the above mentioned quality issues, there are variations in image dimension, pixel size (resolution), pixel value range and photometric interpretation value which can introduce unnecessary variation in the training data affecting the accuracy and the speed

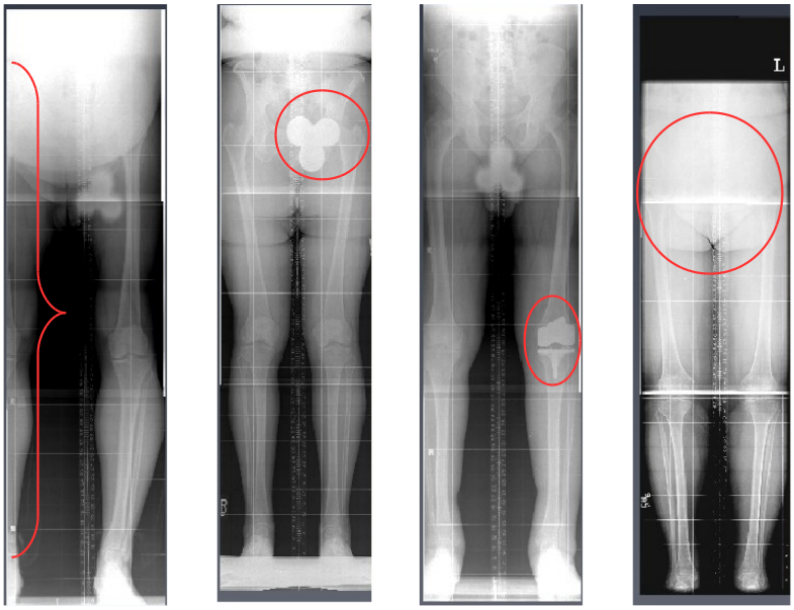


Figure 4.1: Image quality issues: The first image shows improperly positioned X-ray, the second image demonstrates improperly placed radiation protection, the third shows a knee replacement, and the fourth image shows a very low contrast X-ray at the femoral head.

of convergence during CNN training and application. These are addressed in preprocessing step to bring the training data into the same format (Section 3.2.2).

4.1.2 Computational setup

All data and image processing scripts are written in Python programming language on an Ubuntu desktop with Intel(R) Xeon(R) X5680, 3.33GHz CPU and GeForce GTX 980 GPU. CNN architecture was implemented using Keras with Tensorflow-backend. The training of CNN models is done using a Quad-Core AMD Opteron(R) 8384 CPU and a cluster of Nvidia Tesla P100 16GB SXM2 GPUs; testing is performed on Ubuntu desktop with Intel(R) Xeon(R) X5680, 3.33GHz CPU and GeForce GTX 980 GPU. All the visualization of images in this report and landmark data are generated using *AmiraZIBEdition-2018.11* - a system for visual data analysis [58].

4.1.3 Tools for evaluation of results

During the course of the thesis an automated method to determine HKA angle from full limb radiographs is developed. Developing a new measurement method for medical quantities necessitates validating its fitness by comparing its performance with established measurement techniques. Such validation requires showing that its measurements are reliable and that its measurements agree to other established methods of measuring the quantity in question.

In our case the established methods are the HKA angle measurements by medical

experts, and we validate the automated determination of HKA angle by comparing its measurements with experts' measurements. For this purpose, popular tools of analyzing accuracy, agreement between two measurements and reliability of a measurement are used. These are the mean absolute error, the Bland-Altman plot, the weighted kappa coefficient and the Intraclass Correlation Coefficient (ICC) which are described as follows:

Mean absolute error

Mean absolute error is a measure of accuracy. It can help us show how far away in average one set of measurements result is from another set of measurement consider as accurate. We use the mean error to evaluate landmark determination accuracy and HKA angle determination accuracy. The mean absolute error is defined as:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i|,$$

where x_i is the i^{th} measurement in question, and y_i is its corresponding accurate measurement.

Bland-Altman plot

Bland-Altman plot is a method to quantify agreement levels between two measurements [59, 60]. It is a scatter plot of the difference between two measurements (y-axis) against the mean of the measurements between two methods (x-axis). It quantifies the precision of agreement between two measurements using the 95% interval range of differences between two methods and the bias of a new method using the mean of the differences. These are depicted in the plot as can be seen in one of our Bland-Altman plot in Fig. 4.2b.

In addition to showing the bias and precision of agreement, Bland-Altman plots can also show some patterns between two measurements, for e.g., if the level of disagreement (difference between two measurements) increases as the magnitude of the quantity measured increases. Some representative examples of Bland-Altman plots showing various scenarios can be refereed on an article which discusses Bland-Altman plot from Giavarina [59].

Bland-Altman plots does not decide if two methods have an acceptable level of agreement or not. That decision is done by experts judging if the 95% interval range is acceptable range of difference between the new method and established method of measuring a certain quantity. In this thesis, we use the agreement interval of two experts as a rough guideline to discuss an acceptable range of agreement.

Weighted kappa coefficient (k)

A kappa coefficient of agreement is used to make evaluation of agreement between a pair of nominal measurements. The kappa coefficient evaluates the agreement between

the nominal variables by taking into account the possibility of agreement occurring by chance [61] and is defined mathematically as follows:

$$K = \frac{p_o - p_e}{1 - p_e},$$

where p_o is the observed probability of agreement and p_e is the probability of agreement occurring by chance.

The kappa coefficient varies between -1 and +1. A negative value indicates poorer than chance agreement, a value of zero shows agreement occurring by chance, and value of 1 shows a perfect agreement. We use the weighted kappa coefficient which is a derivative of the kappa coefficient for ordered nominal values to evaluate the agreement of classifying leg alignments based on HKA angles determined by our automated method and by experts.

Weighted kappa coefficient is calculated by taking into consideration how far away two classifications differ from each other. For example if one method classifies leg alignment as Varus alignment, and another method classifies a Neutral alignment, the disagreement is given a weight of one; However, if the classification is Valgus alignment then it is given a weight of two. The wighted kappa coefficient values can be interpreted as poor, slight, fair, moderate, substantial and almost perfect agreement if the weighted kappa value is less than 0, between 0 and 0.2, between 0.2 and 0.4, between 0.4 and 0.6, between 0.6 and 0.8 and between 0.8 and 1, respectively [61].

Intraclass Correlation Coefficient (ICC)

ICC is a popular measure of quantifying the reliability of a new method whereby a reliability of a method is defined as the extent to which its measurement can be replicated [62, 63]. Reliability of a new measurement method depends on the degree of its correlation and agreement to another measurement which is used as the basis of evaluation. Mathematically reliability is defined as a ratio of true variance of the quantity measured over true variance of the quantity measured plus the measurement error variance [62, 63]:

$$ICC = \frac{V_t}{V_t + V_e},$$

Where V_t is the true variance of the quantity measured and V_e is the variance caused by error [62]. For the exact formulas to calculate the variances we use the formula of ICC(3,1) in an article from Shrout et al [63].

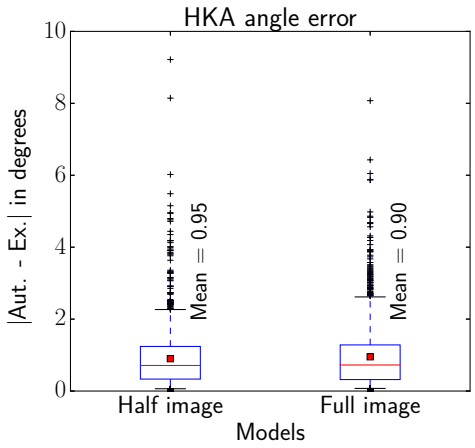
ICC values are interpreted as indicating poor, moderate, good and excellent reliability if the ICC values are less than 0.5, between 0.5 and 0.75, between 0.75 and 0.9 and greater than 0.9, respectively [62].

4.2 HKA angle determination from images experiments

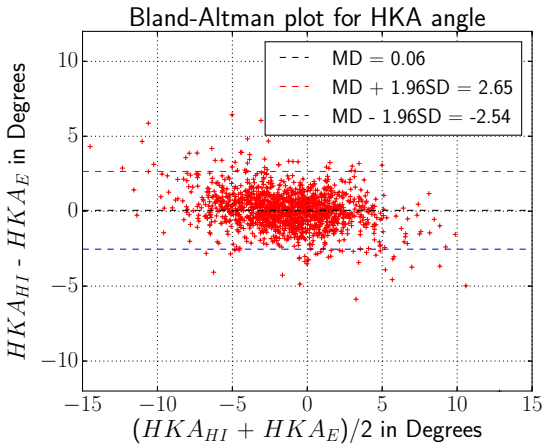
This experiment evaluate the performance of HKA angle determination from images. For this experiment two CNNs are trained according to the implementation and CNNs training description that can be found in Section 3.3. One of this CNNs takes in a full image and another takes half side of an image as input and gives us as output an HKA angle value. We use the entire (12, 24, 36 and 48 month) dataset of full limb X-rays with their associated HKA angle values. In cases where imaging is taken multiple times, the last image is taken for the purpose of training and testing. This strategy is chosen as images are taken multiple times per patient only when there is quality issue on the initial full limb radiograph.

After discounting for cases where multiple images are taken and images which did not pass preprocessing quality tests, there were a total of 3723 unique cases with HKA angles. Since we are only determining right leg HKA angle per image the image can be flipped sideways and used as training data (Section 3.2). This means a total of 7446 images and the associated right leg HKA angle were available for training and testing. All images are original images and no augmented image is used. The data is divided randomly into 60:20:20 (4474:1486:1486) ratio of training, validation and testing dataset. The validation set is used during training to tune hyper parameters such as over fitting and learning rate. This approach is tested on the test set of 1486 images by determining the HKA angle and comparing the result to experts' results. There are no augmented images in the test data set.

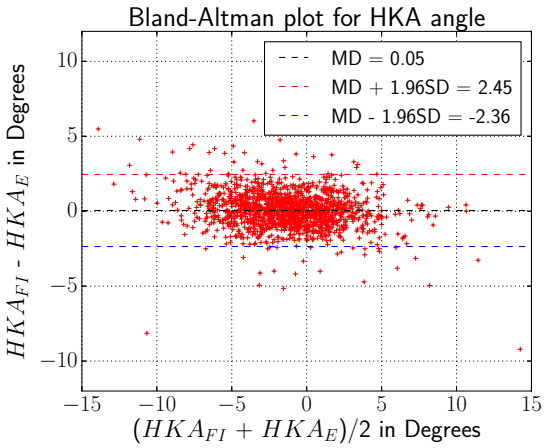
Figure 4.2a shows the HKA angle error using box-plots. This is the absolute value of the differences between the automatically determined HKA angle and the mean HKA angle from experts. We use the mean HKA angle of the two experts to make comparison in the plots as the mean HKA angle from two experts is used to make the HKA angle value used for training the models. Figure 4.2b and Fig. 4.2c show Bland-Altman plots showing HKA angle agreement between HKA angle from this approach and the mean of experts. A direct comparison of the HKA angle to individual experts can be found in Appendix B.1 and B.2. Leg alignment classification using the automatically determined HKA angle shows an agreement level of 0.76 and 0.74 as measured by weighted kappa coefficient for the full and half image model. The method reliability as measured by ICC value is 0.93 and 0.91 for the full image and half image model respectively. More visualization of the results using scatter plot and confusion matrix can be found Appendix B.4 and Appendix B.3, respectively.



(a) Absolute value of difference between model determined HKA angle and mean HKA angle determined by two experts.



(b) Bland-Altman plot of half image (containing one leg) CNN.



(c) Bland-Altman plot of full image (containing both legs) CNN.

Figure 4.2: HKA angle from images.

4.3 HKA angle determination from landmarks experiments

This approach first determines landmarks which define the HKA angle using CNNs and calculates the HKA angle as explained in Section 3.4. Therefore the experiments are designed to evaluate both landmark determination accuracy and the associated HKA angles agreement to experts' measurements. For the purpose of training and testing CNNs, 900 images are labeled with landmarks coordinate points. These images are divided into a training and a testing dataset of 575 and 325 images, respectively.

The CNNs proposed in Section 3.4.3 are trained according to the training process specified in Section 3.4.4. The training dataset is augmented according to the scheme in Table 3.2. The resulting CNNs are tasted on the test set of 325 images by determining landmarks first and then HKA angle. Testing is done for both left and right and left

leg's HKA angle which means the total test set consists of 650 images. Augmentation techniques are only used in the training dataset.

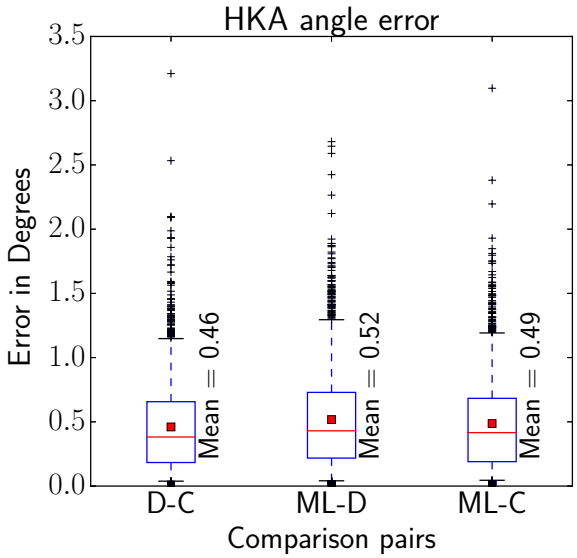
To test if the best performing approach can generalize to a larger test set, the best performing approach is tested on an extra dataset. This test is done on the set of full lower limb radiographs which is not labeled with landmarks. The test is performed by determining the HKA angles of this dataset and evaluating them on the basis of experts' HKA angles. The results of these experiments are presented in the following section beginning with an evaluation of the dataset which is manually labeled with landmark coordinates (900 full limb radiographs).

4.3.1 Generating manually labeled landmarks

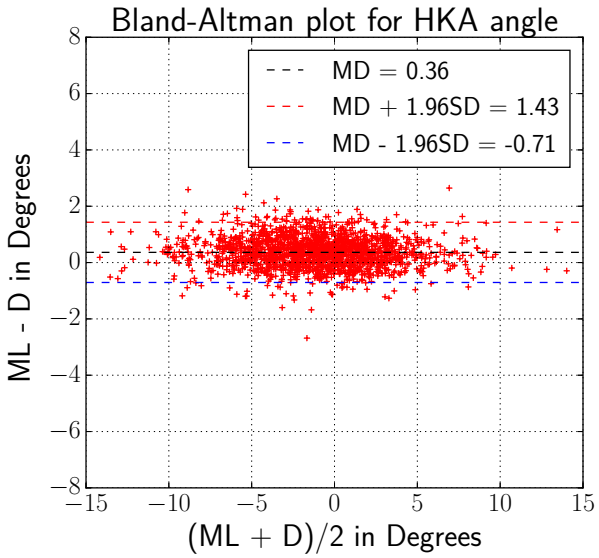
HKA angle from landmarks is based on the CNNs ability to predict landmarks. Since the dataset only contains HKA angles from experts and not the landmarks used by experts to determine the HKA angle, landmark data is generated by referring to journal articles that describe how the HKA angle is determined and the manual that accompanied the dataset [50, 51, 57].These landmarks are explained in Section 3.4.

During landmark generation, images in which there is very low contrast level to the extent that it is hard to detect either the femoral head, the knee or the ankle joint are excluded. The accuracy of these manually labeled landmarks' coordinate points is evaluated by comparing the HKA angle calculated from these landmarks with the HKA angles determined by the experts.

The result of this comparison using Bland-Altman plot and box plot of the absolute value of difference can be seen in Fig. 4.3. The Bland-Altman plot in the figure is plotted against the HKA angles from Dr. Duryea. The HKA angle calculated from these landmarks have reliability measured by ICC coefficient of 0.99, 0.99 against the angles determined by Dr. Duryea, and Dr. Cooke respectively. The weighted kappa coefficient of agreement for classifying leg alignment based on these HKA angles is 0.89 and 0,9 against that from HKA angles for Dr. Duryea, and Dr. Cooke respectively. Further visualization of the results using confusion matrix and scatter plot of correlation can be found in Appendix B.2.



(a) Box plots depicting absolute value of difference between HKA angle determined from manually labeled landmarks and HKA angle determined by experts.



(b) Bland-Altman plot depicting agreement level between HKA angle determined from manually labeled landmarks and that determined by Dr. Duryea.

Figure 4.3: Manually placed landmarks evaluation: ML= HKA from manually labeled landmarks, D = HKA from Dr. Duryea, C = HKA from Dr. Cooke.

4.3.2 Heuristically defined initial ROI

The first implementation of HKA angle determination from landmarks is done using heuristically defined rules to find initial ROI containing the femoral head, the knee and the ankle. This approach uses 3 cascaded pipelines to determine the femoral head landmarks, knee landmarks and ankle landmarks. Each cascaded pipeline uses three levels of landmark

determination starting from a heuristically defined ROI (Fig. 3.10). The entire HKA angle determination requires training 9 CNN models, three for each anatomical region.

The landmark detection results of this cascaded approach can be seen in Fig. 4.4. In this figure, the error of the automatically determined landmarks after the third level of cascaded landmark detection are shown using box plots for the landmarks at the ankle joint, the knee joint and the femoral head. The landmark detection error in each level of the cascade can be found in AppendixB.8.

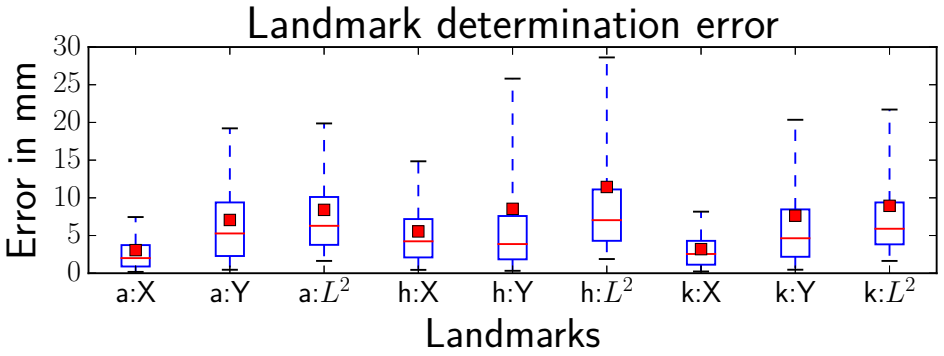
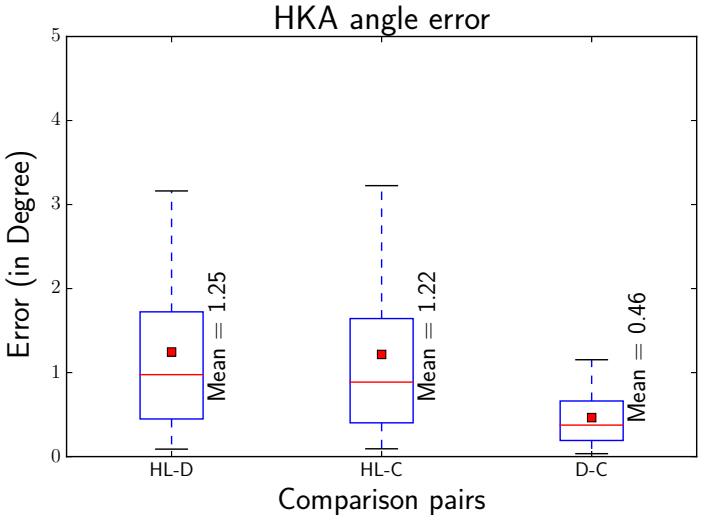
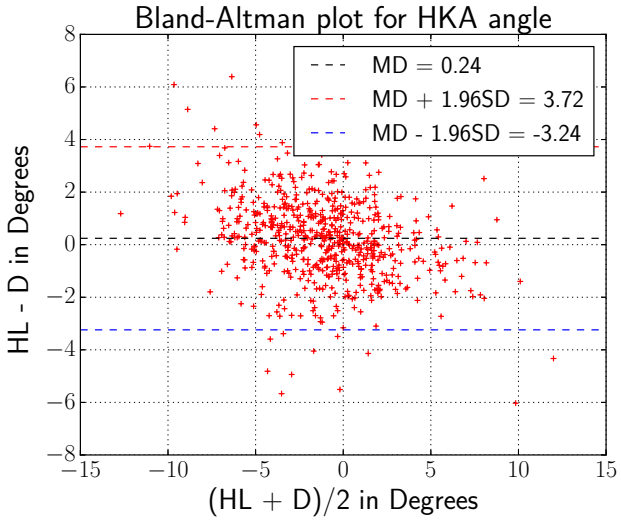


Figure 4.4: Evaluation of landmark determination using heuristically defined intial ROI: Results are from the last (third) level of landmark detection (a = ankle landmarks, h = femoral head, k= knee; X,Y,L^2 : x-axis, y-axis and euclidean norm, respectively).

The results of HKA angle determination using these landmarks is evaluated using Bland-Altman plot and HKA angle error as shown in Fig. 4.5. The HKA angle calculated from the predicted landmarks have an ICC value of 0.88 and 0.89 against the angles determined by Dr. Duryea, and Dr. Cooke respectively. The weighted kappa coefficient of agreement for classifying leg alignment based on these HKA angles is 0.72 and 0.73 against that from HKA angles for Dr. Duryea, and Dr. Cooke respectively. Further visualization of the results using confusion matrix and scatter plot of correlation can be found in appendix in Fig. B.3.



(a) Absolute value of HKA angle error



(b) Bland-altman plot of HKA angles.

Figure 4.5: Evaluation of HKA angle determination using heuristically defined intial ROI: from the last (third) level of landmark detection. (HL = HKA from landmarks determined at the last level of cascaded landmark detection on heuristically defined initial ROI, D = HKA from Dr. Duryea, C = HKA from Dr. Cooke).

Automatically defined initial ROI

An image classifier is trained and applied across the image in a sliding window fashion (Fig. 3.11). This landmark detection produced high accuracy ROI detector. In testing a sample of 325 cases, there were a 100% success rate in detecting a ROI of size 512×512 which contains all the relevant landmarks.

Using this initial ROI a 2 level cascade architecture is implemented as shown in figure 3.12 for the femoral head and ankle joint. For the knee landmark we added extra level of landmark prediction to improve the accuracy, as less accurate prediction of the knee impact the line defining the HKA angle more severely than that of the femoral head and ankle joint. Following the same evaluation tools used for heuristically defined initial ROI, the landmark detection error is evaluated. The result can be seen in Fig. 4.6.

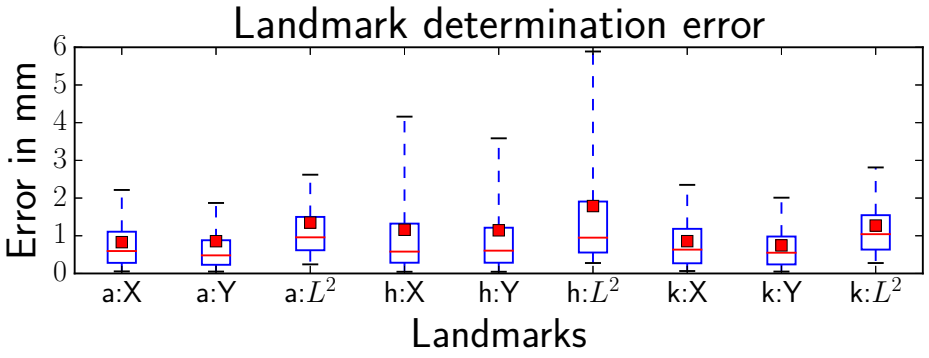
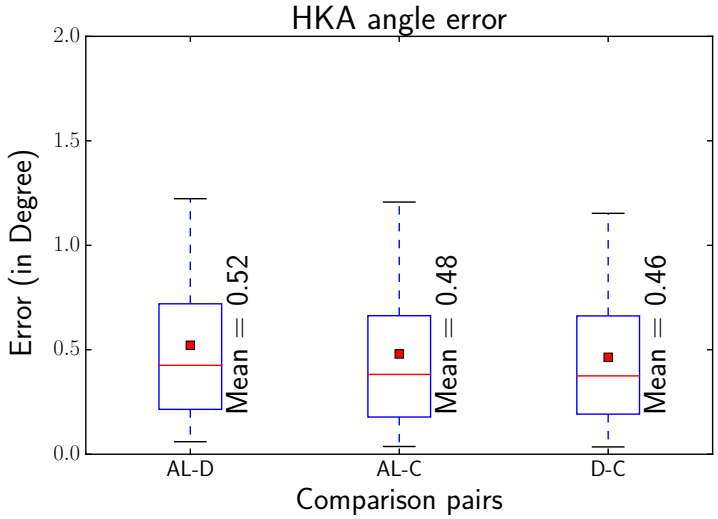
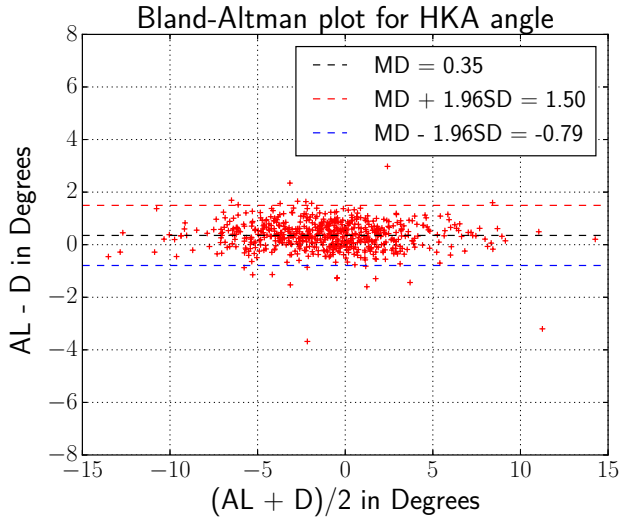


Figure 4.6: Evaluation of landmark detection using cascaded landmark detection with automatically defined initial ROI: The distance between automatically determined landmarks and manually labeled landmarks (a = ankle, h = femoral head, k= knee, X,Y,L^2 : x-axis, y-axis and euclidean norm).

The evaluation of HKA angle determination using these landmarks can be seen in Fig. 4.7 which shows the HKA angle agreement and accuracy. The HKA angle calculated from the determined landmarks have ICC coefficient of 0.98 and 0.99 against the angles determined by Dr. Duryea and Dr. Cooke respectively. The weighted kappa coefficient of agreement for classifying leg alignment based on these HKA angles is 0.87 and 0.89 against that from HKA angles for Dr. Duryea, and Dr. Cooke respectively. Further visualization of the results using confusion matrix and scatter plot can be found in appendix in Fig. B.3.



(a) HKA angle error



(b) Bland-Altman plot of HKA angles angles.

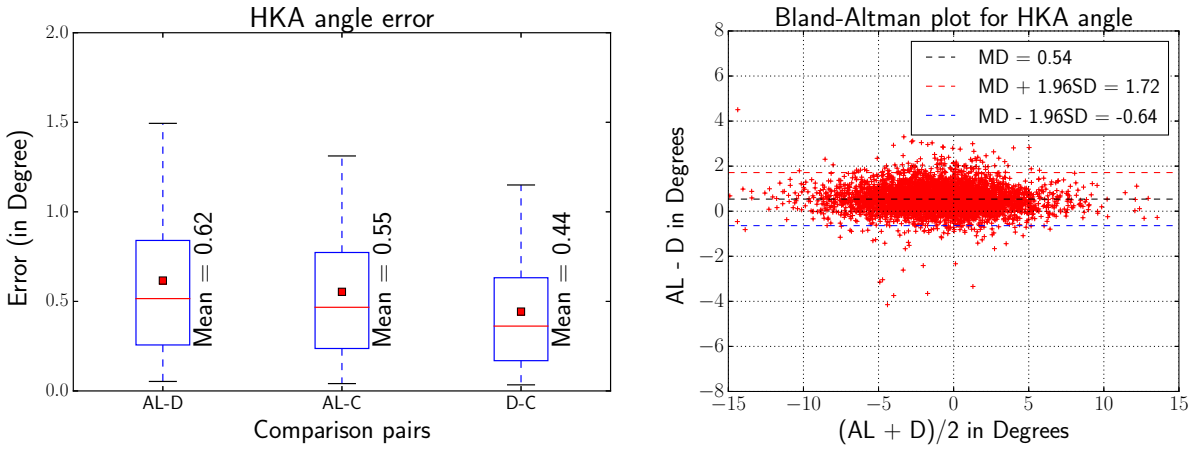
Figure 4.7: Results of cascaded landmark detection on automatically defined intial ROI: (AL = HKA determined using automatically defined initial ROI, D = Dr. Duryea, C = Dr. Cooke).

Automatically defined initial ROI: Test using large dataset

To test the generalizablity of our best performing approach, which is the cascaded landmark detection using the automatically selected ROI, it is tested on a dataset which is not labeled with landmarks. This contains a total of 2818 cases (5636 HKA angles). This test is performed by determining the HKA angle and comparing it with experts' HKA angle.

Following the same evaluation tools, the HKA angle agreement with experts is evaluated. The result can be seen in Figure 4.8. The weighted kappa coefficient of agreement for classifying leg alignment based on these HKA angles is 0.82 and 0.83 against that from HKA angles for Dr. Duryea, and Dr. Cooke respectively. The reliability of the HKA angle

determination on this dataset is demonstrated by an ICC value of 0.96 and 0.97 on the basis of Dr. Duryea and Dr. Cooke HKA angle measurement respectively.



(a) HKA angle error.

(b) Bland-altman plot of HKA angles angles.

Figure 4.8: Results of cascaded landmark detection on automatically defined initial ROI on a large test set: cases that are not labeled with landmark coordinates. AL = HKA determined using automatically defined initial ROI, D = Dr. Duryea, C = Dr. Cooke.

5 Discussion

In this chapter we discuss the results of the experiments. In this regard, the discussion starts with the challenges faced during the application of CNNs to determine HKA angle and the actions performed to tackle those challenges. This will be followed by the discussion of the results of the experiments.

5.1 Challenges and solutions

The challenges faced in order to apply a CNN-based method to determine HKA angle can be organized into four broad categories. These are issues related to the size and shape of the lower limb radiographs, the quality of the images, the choice of a CNN architecture and scarcity of training data for CNNs.

5.1.1 Size and shape of full lower limb radiographs

One challenge in the application of CNNs to lower limb radiographs is that the images are of high number of pixels. As medical radiographic images are acquired with the intention of capturing sub-millimeter details (resolution) of the anatomical structures, a radiographic image of the full lower limb is in the order of 10^3 pixels. As a result, the number of convolutional and pooling layers needed to perform deep learning in the context of the entire image area becomes high. However a very deep network has a large number of parameters which requires large computational resources and large number of training data. Because of these limitations, it is difficult to use the images as they are without facing issues when loading the model and the training data into memory. Two approaches are used to tackle this challenge.

The first is to narrow the relevant region with some preprocessing procedures before progressing into landmark or HKA angle determination. The second is to use image re-sampling tools in order to resize the image to a small number of pixels. Both of these methods to decrease the size of an image have their own limitations. In case of resizing the image, it lowers the resolution of the image, decreasing the detail that can be seen in the images. This is due to the fact that resizing the image combines multiple pixels into one pixel which will destroy the nuanced differences multiple pixels can contain. This becomes particularly severe if the original image is of bad quality (e.g. images with low contrast). On the other hand narrowing down the region of interest requires developing methods to find reliable and accurate ROI selection algorithms which is a difficult task.

In this regard two approaches are investigated to determine smaller than the entire image ROIs. One leverages the standardized setting of medical image acquisition by developing heuristic rules. The second uses the power of CNNs to detect ROIs containing femoral head, ankle and knee joints.

Another challenge is the rectangular shape of the radiographic images with close to 4:1 height to width ratio. Assuming pixel sizes to be equal across dimensions, the number of pixels across the height of an image is proportionally higher than that across the width of an image. This means that we have more variables across the height of an image than across the width of an image. With more variables comes more complexity in the respective axis. In addition, due to natural height variations of the participants, there is more variation in the y-axis than in the x-axis. This is particularly noticeable in image which have not been properly taken where the radiographs contain the anatomy of individuals up to their chest. These aspects make learning patterns in the y-axis more difficult than in the x-axis (e.g. less accurate landmark coordinates on the y-axis than on the x-axis as can be seen in Figure 5.1).

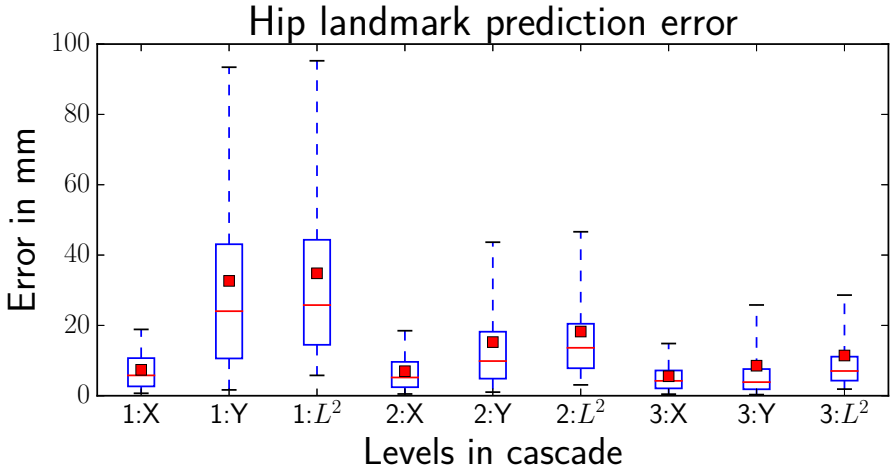


Figure 5.1: Hip landmark error on three levels of cascaded landmark detection using heuristically defined initial ROI. 1,2,3 are levels in cascaded pipeline with first level being the top left quarter of an image resized to 256×1024 pixels, 2nd level 512×1024 pixels around the center of landmarks predicted on the first level and the third level with 512×512 pixels around the predicted landmarks. X = x-axis, Y = y-axis, L^2 = Euclidean distance.

The issue of more variation on one axis can be countered with more image augmentation on that axis. Another approach that can be used to adjust shape of image is the use of non-squared pooling. As pooling is used to decrease number of variables by summarizing the output of a convolutional layer in neighboring pixels, it can be used to equalize the image by summarizing more in one dimension than in another. For example using 1×2 max poling which outputs the maximum of two pixels along the height and returns the pixel as it is along the width.

With this description of the challenges and approaches taken to address them, we will discuss the results of these actions individually in the next subsections.

HKA angle from landmarks

The approach of finding small ROI is particularly useful for landmark detection since landmark detection depends on the local information close to the anatomical structure of the landmarks [39]. A small ROI means less variables and complexity for the model to learn. In this regard, the first approach used is using heuristic rules by taking advantage of the fact that medical images have a clearly defined setup (Sec. 3.4.1). However the ROI found this way is still a very big image in the order of 500×2000 pixels.

In addition to the total size of the image, the ROI defined using this heuristic rule is a rectangular shaped image. This is to say that the height of the images is a multiple of times the width of the height. Such a rectangular images introduces a much bigger variation on the long dimension. The combined effect of such rectangular images and big image size can be seen by the higher error on the y-axis than on the x-axis in Fig. 5.1. Figure 5.1 shows as well that the mean error decreases and the difference between the error in the x-axis and y-axis becomes more equal as the ROI becomes smaller and more balanced in the number of pixels on the x and y-axis.

Using this observation, automatic ROIs selection was implemented to find initial ROI which covers a small region in a square window (512×512 pixels). The impact of this approach can be seen in Fig. 5.2 which shows the outliers and the mean errors of determined femoral head landmarks in the x-axis and y-axis using heuristic and automatic initial ROIs based landmark detection. It can be seen in the figure that there are many extreme outliers which are far away from the actual landmarks when using the heuristic rule. Such range of outliers restricts our scope on how much we can narrow our region of interest in the next level.

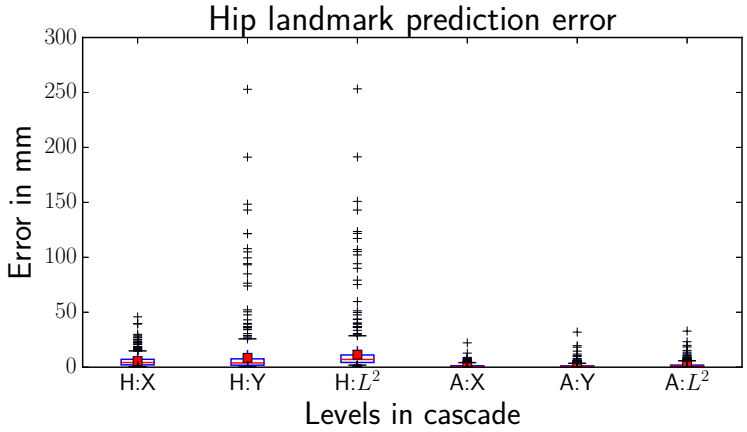


Figure 5.2: Comparison of outlier distribution of hip landmarks usign heuristically defined (H) initial ROI and automatically (A) defined ROI.

The use of automatically determined initial ROI necessitates accurately determining regions which contain the landmarks in question. In this regard the implementation described in Section 3.4.1 shows a 100% accurate object detection rate on the dataset of 900 cases for which landmark coordinate were labeled. In the dataset including for which landmark coordinates are not labeled, there were 4 cases out of 2818 cases where the automated ROI detection failed to detect the knee or ankle initial ROI[1]. The cases where the automatic ROI failed to define the ROI accurately were images with quality issues such as bad contrast and artifacts. Two examples can be seen in Appendix C.2.

The accuracy of this ROI detection can be explained by the fact that CNNs have been demonstrated for their ability in tasks that require object recognition [64]. The sliding window application which means the object classifier used to define ROI can sample the object multiple times aggregating the accuracy also contributes to the high accurate ROI detection. The 100% accuracy on the smaller labeled dataset can be explained by the fact that very low quality images were excluded as it is difficult to place landmarks at the appropriate location.

HKA angle from images

For the approach of direct HKA angle prediction from images narrowing down the region of interest approach is limited. This is due to the fact that the HKA angle depends on the entire configuration of the femoral head, the knee and ankle joint. As such it is difficult to decide the section which is relevant as an input to the CNN model that determines the HKA angle. Only one possibility is investigated to narrow the region for direct HKA angle determination which is to only consider an image part which contains one leg. This is easily done by dividing the images into two parts along the x-axis by finding the middle pixel. However this makes the ratio of height to width more severe to the tune of 8:1. Additionally highly rectangular images affects the networks ability. This affected the choices of architectural decision.

When learning high level information the network continually translates the image into more abstracts representations to learn patterns. For this it is necessary to have a full field of view to pick patterns from the entire image. This can be easily understood by referring to Fig. 2.5c where we discussed the receptive field of view of a neuron in a convolutional layer. In a highly rectangular image, when we achieve a full field of view on the x-axis we can only learn equivalent features on y-axis. More variation on the y-combined with lower field of view at the last convolutional layer makes learning features on the y-axis more difficult.

The resized image of a full limb X-ray has a size of 256×1024 pixels, half sided image has a size of 128×1024 pixels. As the series of convolutional layers increases and we

[1]We evaluated the cases which have HKA angle error greater than 2° manually.

achieve a full field of view on the x-axis, we achieve only $\frac{1}{4}$ and $\frac{1}{8}$ of the image on the y-axis for the full image and half image approach. This affects the level deep features we could learn in each dimension. To counter this we use a rectangular pooling of size 1×2 (i.e., pooling only on the y dimension) on the first two convolutional layers the half image model (Table 3.8). On the full image model we use the common 2×2 pooling on all convolutional layers.

Since our first assumption when using a small ROIs (half image) can improve the performance with lower number of parameters and less variation, on all methods of evaluation the model, using the full image performs better than the half image model (e.g. Fig 4.2). This can be due to the fact that the use of rectangular pooling can negate this advantage by pooling away more detail more quickly on the y-axis than in the x-axis.

5.1.2 Quality issues

Quality issues created a challenge in accurate determination of landmarks. The main quality issues are images with bad contrast, artifacts, non anatomical objects and not properly positioned patients. Bad contrast is particularly common at the hip landmarks which is often a result of bad exposure of the femoral head due to body fat blocking x-rays. In such cases the automated landmark prediction fails as it is difficult to differentiate between the pixel values. Artifacts are also another source of failure in accurately detecting landmarks. These artifacts are some unidentified lines and patterns in the image (Appendix C.2).

Improperly placed radiation protection and some medical implants were found in the cases where the HKA angle error were relatively big. The landmark detection was able to properly place landmarks in hip replacements despite the fact that there were only few of those in the training data. The medical implants at the knee were harder for our approach to place landmarks. This can be due to the diverse set of knee implant shapes and objects at the knee. In this regard more training data containing more types of replacements can improve the robustness of the approach.

The proper positioning of the patients also affected the outcome of the HKA angle determination. This is particularly an issue if the patients are not placed at the middle of the image. The implementation assumes the patients placed at the middle of the image to minimize the computational effort by only considering half side of the images. Therefore if the patient is positioned to the left or to the right of the center of an image, then the appropriate regions will fail to be in the region evaluated.

5.1.3 Choosing a CNN architecture

In this regard the challenge is in the number of parameters one can investigate when developing a CNN based model. This can be avoided by using other popular pre-trained CNN architectures. However, in our case, it was challenging to implement them for the

task of the HKA angle determination. The first of the reasons is that the popular networks' input image dimension which is smaller than the lower limb radiographs. The second is that other networks are trained for different tasks which means that they need to be trained again for the task of the HKA angle determination and landmark determination.

In this context, the architectural decisions were based by taking lessons from other architectures. We relied on other CNN based landmark detection architectures [39, 48] and the architectural design of VGG class of CNNs [65] and the U-net architecture [66] and adapted the architectures to our needs. For example the VGG19 and U-net architecture uses 4 and 2 convolutions before pooling operation the to improve the nonlinearity level the network can learn before pooling decreases the number of variables that can be used to learn such patterns.

We used these lessons to adapt the landmark detection architecture implemented by Sun et.al. [39] to our HKA angle determination. Similarly we increased the number of feature maps after each pooling operations number of convolutional layers to increase the representational capacity of our CNNs. Other parameters such as learning rate, dropout rate and other training parameters were adopted on a trial error fashion. For example batch normalization was found to improve the convergence of the CNN training and the speed of training and was used throughout the training process.

5.1.4 Scarcity of training data

The use of relevant image augmentation techniques to generate more training data is of paramount importance in the process of training CNN models when there are not enough original images [54]. However, the augmentation techniques shall be chosen carefully such that they do not introduce variations that are not represented in the input the network is likely to encounter.

The importance of image augmentation techniques is particularly useful in landmark detection and object detection (for ROI). This is due to the fact that landmark determination uses local information (Appendix D.1). This means the variance of the anatomical landmarks location can be easily imitated by shifting the relevant anatomical parts to different locations across the ROI. Similarly we used rotation and scaling. However, we limited the range of variation that can be introduced by these two techniques using some rules. For example the rotation of anatomical parts 90° does not introduce a variation that a CNN will encounter, as all pictures are taken with the individuals standing upright. Therefore the use of rotation was limited to between -5° and 5°. Similarly scaling was limited to within 90% and 110%.

From trial and error, on the start of the thesis, we found out that all the augmentation tools do not improve performance equally. For example translation were much more relevant than rotation and scaling on the first level of landmark detection. This can be explained by the fact that the proportion of an anatomical region (e.g. femoral head)is

smaller in relative to ROI. For example, in Appendix D.1, we can see that at the initial level the landmark detection depends on the entire pixels of the femoral head, in the second level they depend heavily on the edges. This means translating the whole femoral head can be much useful than small rotation of it in the early level.

Considering the amount of time and computational resources it takes to train a network of big starting ROI, gave each augmentation technique varying levels of importance. Considering that the landmark location can be found anywhere in the ROI, translation of the relevant anatomical objects within the ROI is given highest importance. In the second level of importance, is the scaling which is used to imitate natural size difference of anatomical objects. Scaling becomes more important as the ratio of the size of the anatomical object becomes higher in proportion to the size of the ROI (later levels of cascaded landmark detection). In third level of importance is the rotation of anatomical objects. This is also given more importance at later levels of cascaded landmark detection. This reasoning was used to develop the augmentation plan in Tabel 3.3.

Augmentation techniques were difficult to use in the case of HKA angle prediction from images. This is due to the fact that we do not know exactly what the relevant region is (Appendix D.2). Therefore we can not evaluate the appropriate augmentation tools without introducing unrealistic augmented images. For example, during application, we take a full lower limb radiograph and evaluate its HKA angle. However if we decided to use translation as a tool, we do not know which part of the lower limb radiograph to translate across the y-axis or across the x-axis. If we do that we can end up with an image which does not contain the hip, or the ankle.

5.2 Interpretation of results

The results of experiments used to determine HKA angle are summarized based on the evaluation tools.

5.2.1 Mean absolute error

Fig. 5.3 shows a summary of the difference between two sets of HKA angle determinations. As can be seen in the figure we achieve the highest accuracy of HKA angle determination using the automatically defined initial ROI. We achieved statistically significant (t-test : $p = 0.01$) accuracy improvements over all methods by using automatically detected initial ROI.

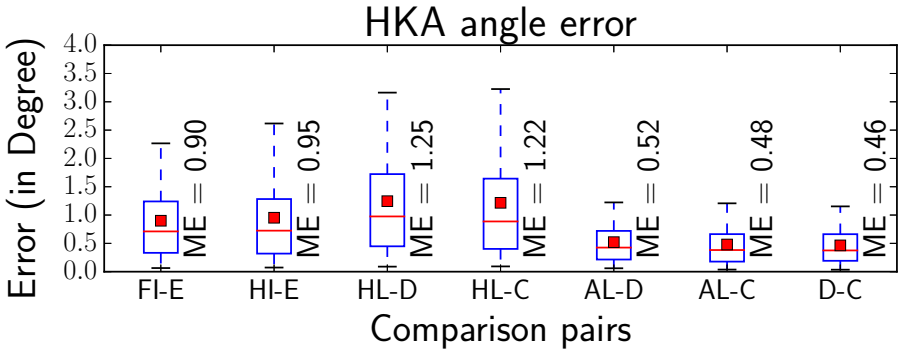


Figure 5.3: Comparison of absolute value of HKA angle error. FI, HI: HKA angle from images using full image and half image, respectively. E: the mean HKA angle from experts. HL, AL: HKA angle from landmarks using heuristically and automatically defined initial ROIs, respectively. D, C: HKA values from Dr. Duryea, and Dr. Cooke, respectively.

5.2.2 Agreement between automatically determined HKA angles and expert HKA angles

In this section we discuss the agreement levels of HKA angle values as evaluated by Bland-Altman plots. Figure 5.4 summarizes the agreement level of various methods in determining the HKA angle using the 95% range of difference collected from the bland alt man plots given in Chapter 4. It shows that we are able to achieve a precision level defined by the 95% interval range close to that of the experts but with a noticeable bias of 0.35° using the automated initial ROI angle.

Source of bias in automated method

Given the fact that the manually labeled landmarks have a bias of 0.3° (Fig. 5.4, HKA:ML), and that the HKA angle determination using the heuristic rule and automated ROI methods have a bias close to 0.3° as well (Fig. 5.4, HKA:HL, HKA:AL), we can argue that the bias comes from the manually labeled landmarks which are used as training data for landmark determination. In order to see if the bias indeed comes from how landmarks are placed, we reevaluated our landmark placement.

As discussed in Section 3.4, there is some variation in how the knee landmarks are defined. We previously placed the landmarks at the femoral and spinal notches of the knee joint which are visually easy to identify. However after further evaluation of more literature [67], a different landmark placement scheme at the knee is implemented on 50 images. The new landmarks are placed at the center of the femoral condyles and the center of tibial plateau paying more attention to placing the landmarks at the middle of the distal and medial edges of the knee joint (Fig. 5.5a).

The result of the investigation shows that the HKA angle measurement bias is reduced

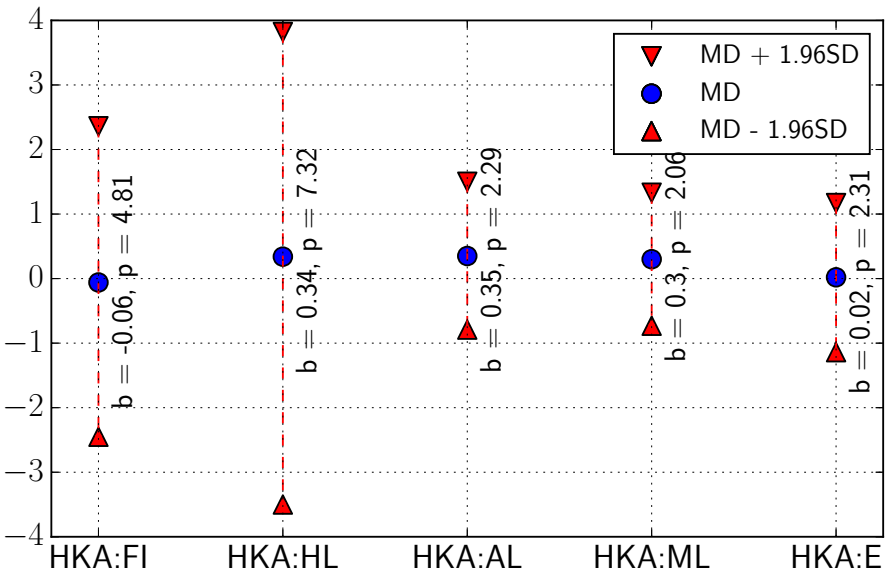


Figure 5.4: Agreement levels between automatically determined HKA angles and Dr. Duryea's HKA angles: The red vertical lines indicate the 95% difference range. The blue spots shows the mean difference. FI = Full image CNN, HL = Heuristically defined initial ROI, AL = Automatically defined initial ROI, ML = Manually labelled landmarks, E = Dr. Cooke.

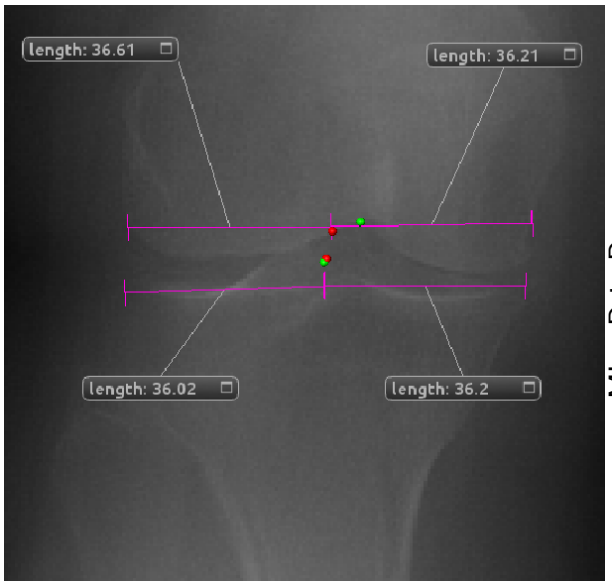
with the new scheme of landmark placement. This leads us to conclude that the bias comes from the manually labeled landmarks used for training. In this regard more consultation with experts can help to better implement a knee landmark placement scheme.

5.2.3 Leg alignment classification based on HKA angles

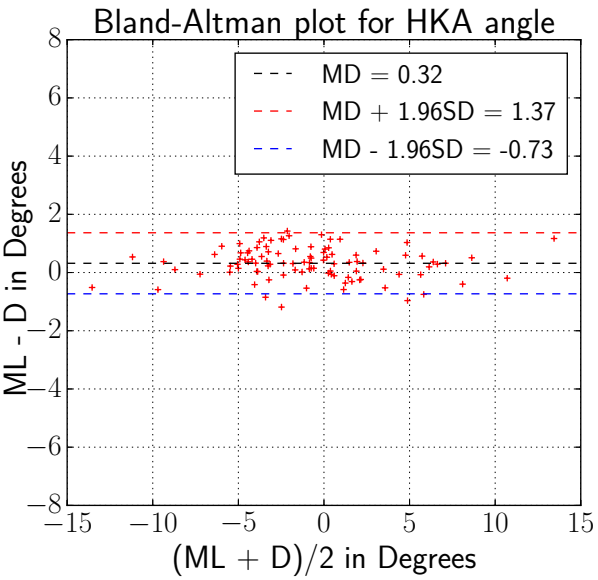
The agreement of classifications done based on various methods is illustrated in Table 5.1. Similar to previous evaluation we achieve agreement levels close to that of experts by determining HKA angle using landmarks determined from automatically defined initial ROIs. According to the interpretation guideline of weighted kappa coefficient this is almost perfect level of agreement, while the other methods tried show a substantial level of classification agreement.

5.2.4 Reliability of methods in determining HKA angle

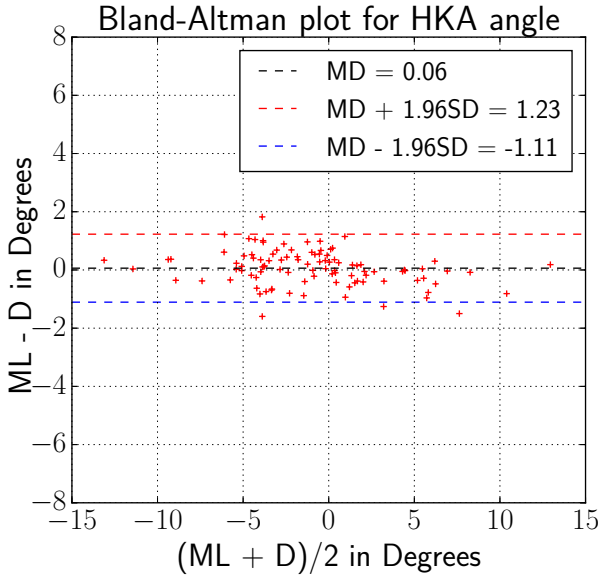
Reliability analysis using ICC evaluation show us that HKA angle determination using automatically defined initial ROI shows excellent reliability evaluated on the basis of the HKA angle determination by experts. The HKA angle determination using full image CNNs shows as well excellent reliability.



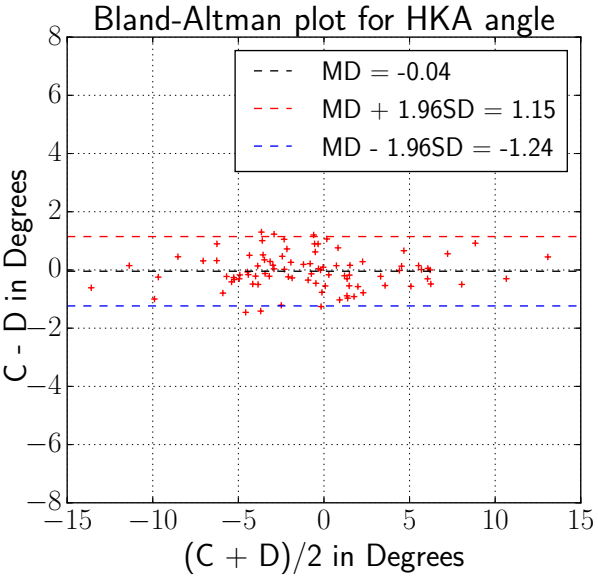
(a) Old (green) landmarks and new (red) landmarks.



(b) Bland-Altman plot of HKA angle from old landmarks



(c) Bland-Altman plot of HKA angle from new landmarks



(d) Bland-Altman plot of HKA angle from experts

Figure 5.5: HKA angle bias investigation: 50 cases randomly taken are labeled with two new landmarks at the the mid point of the tibial plateau and femoral condyles of the knee.

5.2.5 Generalizablity of the method

The result of our test using the entire data set shows a decrease in the agreement level of the determined the HKA angels as can be seen by comparing the plots in Fig. 4.8 and Fig. 4.7. This is particularly noticeable in the bias, which worsened from 0.35° to 0.54°. It is here important to notice that the bias between Dr. Cooke and Dr. Duryea increased

Weighted kappa Coefficent		Methods and Experts					
		FI	HL	AL	ML	D	C
Experts	D	0.77	0.72	0.87	0.89	1	0.9
	C	0.78	0.73	0.89	0.9	0.9	1

Table 5.1: Classification agreement evaluation using weighted kappa coefficient. (FI = HKA angle determination using full image, HL = HKA from landmarks determined using heuristically defined initial ROI, AL = HKA from landmarks determined using automatically defined initial ROI, ML = HKA from manually labeled landmarks, C = HKA angle from Dr. Cooke and D = HKA angle from Dr. Duryea).

ICC		Methods and Experts					
		FI	HL	AL	ML	D	C
Experts	D	0.92	0.88	0.98	0.99	-	0.99
	C	0.93	0.89	0.99	0.99	0.99	-

Table 5.2: Reliability of HKA angle determination methods. (FI = HKA angle determination using full image, HL = HKA from landmarks determined using heuristically defined initial ROI, AL = HKA from landmarks determined using automatically defined initial ROI, ML = HKA from manually labeled landmarks, C = HKA angle from Dr. Cooke and D = HKA angle from Dr. Duryea).

in the new dataset from 0.02° degree to 0.08° degree. Therefore, if we ignore this change of 0.06° then on the large test dataset, there was an increase of bias by 0.13°. The ICC and weighted kappa coefficient also decreased from 0.98 to 0.97 and from 0.876 to 0.82, respectively.

Despite the above described decrease in performance, the HKA determination method using the automatically defined initial ROI is performed with excellent reliability and excellent agreement of HKA classification as measured by ICC and weighted kappa coefficient, respectively. This shows that the method is generalizable to the entire OAI dataset.

6 Conclusion

This chapter concludes the thesis report with a summary of the thesis work to determine the HKA angle automatically using CNNs. It will elaborate on the limitation of the work and potential improvements and areas of further investigation

6.1 Summary

In this thesis we implemented an automated method to determine the HKA angle using CNNs. For this we investigated two approaches to determine the HKA angle from full lower limb radiographs. The first takes a full lower limb radiograph and determines the HKA angle using CNNs in a 'black box system' manner, i.e. take an image and regress to an HKA angle. The second approach imitates the HKA angle determination of a human experts. Accordingly the HKA angle determination is divided into multiple visual tasks, namely landmark determination at the femoral head, the knee and ankle joints. The HKA angle is then determined by an algorithm that calculates the HKA angle given the landmarks. Therefore, the second approach relies on accurate determination of landmarks that define the HKA angles. In this regard, we perform extensive literature review to explore current research status of landmark detection methods, which led us in the direction of CNNs which are a powerful machine learning tool to model patterns in data using a network of basic units called convolutional layers.

The main challenge in the application of CNNs for the HKA angle determination were the fact that the lower limb radiographs images were image size with unequal height and width. This was solved with a coarse to fine landmark determination approach and the use of accurate and small sized ROI determination method. The fact that a typical CNN network contain a large set of parameters which runs into millions means that it requires a large set of labeled training data to learn the weight of parameters. However gaining large amount of medical data is very expensive. Similarly in this thesis, the availability of training data was limited. We used 3783 full lower limb radiographs from the OAI whose HKA angles were determined by two experts. We manually labeled 900 images with landmark coordinates and used carefully image augmentation techniques to counter the shortage of enough training data.

CNNs for the HKA angle determination from images (approach 1) were trained and tested using the entire OAI dataset. CNN models for landmarks were trained and tested on 900 images with labeled coordinates. These methods of HKA angle determination were evaluated based on their accuracy, agreement and reliability in correspondence with

expert determined HKA angles. The HKA angle determination using landmarks that are determined with cascaded landmark detection on automatically defined initial ROIs have found to have a precise agreement to expert determined HKA angle. However, this approach has a significant bias. We demonstrated that the bias is caused by how the training landmarks at the knee joint were defined. The HKA angle determination from landmarks that are determined with cascaded landmark detection on automatically defined initial ROIs shows an excellent reliability with an ICC value close to that between experts. The classification of leg alignment using this approach shows as well excellent agreement to classification based on experts' HKA angles.

The generalizablty of this method is further evaluated by determining the HKA angle of the entire data set which has HKA angle value determined by experts. This showed a slight decrease of agreement and reliability level. This can be explained by the fact that the training data contained good quality images where it was possible to see the landmark locations, however the entire dataset contains images with bad quality as well. The thesis work demonstrated with the right approach, network architecture and wise use of image augmentation techniques CNNs can be used to determine the HKA angle.

6.2 Limitation and future work

Although we achieved a satisfactory evaluation of the HKA angle measurement methods, there are multiple points that can be considered as limitations. The landmarks were placed by a non-expert which created a significant bias. Although we showed a scheme which removed the bias, more consultations with experts to define the knee landmarks can be beneficial in this regard.

Any CNN training requires a large amount of data. Despite the use of image augmentation techniques to make the best of this limited data, the performance of the HKA angle measurement can decrease when applied to a diverse data set as shown by our experiment using more testing dataset. In addition to the limited dataset, we removed images with bad quality such as bad contrast, occlusion by radiation protection and medical replacements particularly at the knee from our training data. This means that our methods can perform badly if there are many implants and quality issues in an image.

We did not explore if further improvement can be gained by further levels of cascaded landmark detection. We performed two levels at the femoral head and ankle and three levels at the knee ROI (Fig. C.1 in Appendix). This was due to the fact that narrowing down the region of interest further means that we can not fit the entire anatomical object (femoral head or talus bone) in our ROIs (e.g. 128×128) without resizing the image into low resolution. We can see this by referring back to Fig. 3.12 where the second level of femoral head landmarks detection is done on a 256×256 pixel region. If we lower the ROI further to 128×128, we can no longer fit the femoral head inside the ROI. It was

possible at the knee to do three levels of landmark detection since the knee landmarks are just at the center of knee in a very small region. This means we do not need to have the knee in the full field of view of a ROI.

The fact that we can not fit the entire anatomical object means we can not determine all landmarks together from a ROI. It is possible to take a ROI around each landmark which is very small, but this requires training 6 CNNs for the femoral head only. In addition, determining landmarks together makes the network learn the configuration of the landmarks as well. Learning this configuration played a role in accurate landmark determination which is particularly important where the landmarks are used to find a center of a circle. This inability to contain the entire configuration of landmarks in one image can be considered a 'bottle neck' limitation of the coarse to fine approach as we implemented it.

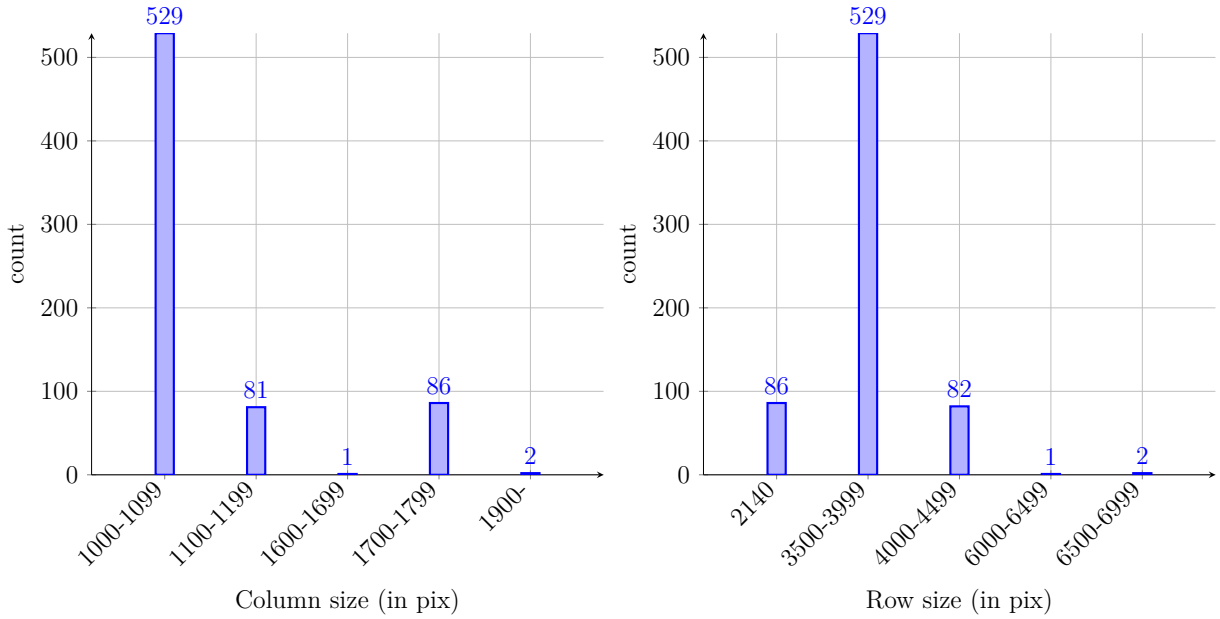
In addition, the coarse to fine approach requires many CNN models which would take a long time to train and to use. The method implemented uses 10 CNN models to determine landmarks in a cascaded way. This means that it takes a long time to train all the models and during the application of the models. This was one weakness of this approach where it took few minutes to find the ROIs for the first cascade. This is mainly due to the implementation of the ROI detector which traversed the image with a slide of 100 pixels along the height and the width, but as well to the numerous CNNs used to determine the landmarks.

An exploration of other network architectures can be beneficial in this regard to explore ways to predict all landmarks in a unified way using less number of models. In this regard one interesting work is the use of heat maps to predict landmarks [68]. In this approach landmarks are labeled on an image as heat maps. A U-net is then used to regress the heat maps. A challenge to this approach can be the size of the full limb radiographs. Another recent work uses iterative approach of landmark detection [69]. In this approach landmarks are determined iteratively by taking a patch of image around a staring landmark location and determining a displacement vector for a new more accurate landmark location using regression and classification functions. This can be an effective method given the fact that full limb radiographs are standardized images with a common pose. Another aspect of improvement that can be explored is the use of more advanced ROI (bounding box) selection approaches [70–72]. This can improve the ROI detection accuracy and the speed of finding the ROI.

The HKA angel prediction ability of the CNNs taking full image and half image and regressing was also very promising considering that HKA angle determination is a more abstract evaluation of the image than finding landmarks. In addition, when considering that there were limited images and less scope for image augmentation techniques, the approach's performance can be considered promising. In this regard it will be interesting if the performance can be further improved by increasing the training data. It shows

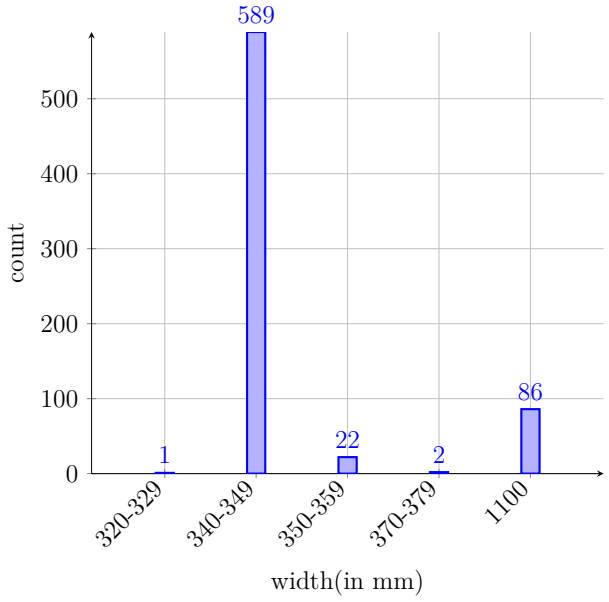
as well that CNNs can be used to predict higher level (more abstract) information from medical images. This can be transfered to areas such as disease progression evaluation and investigation of diseases from medical images.

A Input data statistics

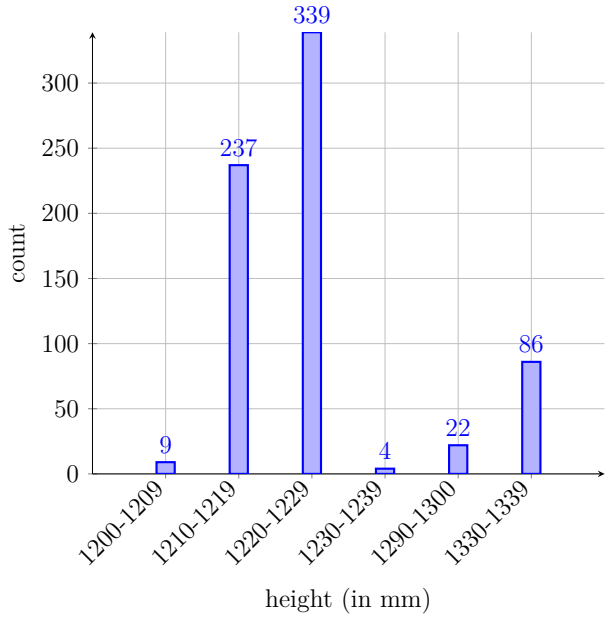


(a) Height (number of pixels) frequency

(b) Width (number of pixels) frequency



(c) Width (in mm) frequency



(d) Height (in mm) frequency

Figure A.1: Training data dimension statistics.

B Experiment results

B.1 Results from HKA angle from images experiment

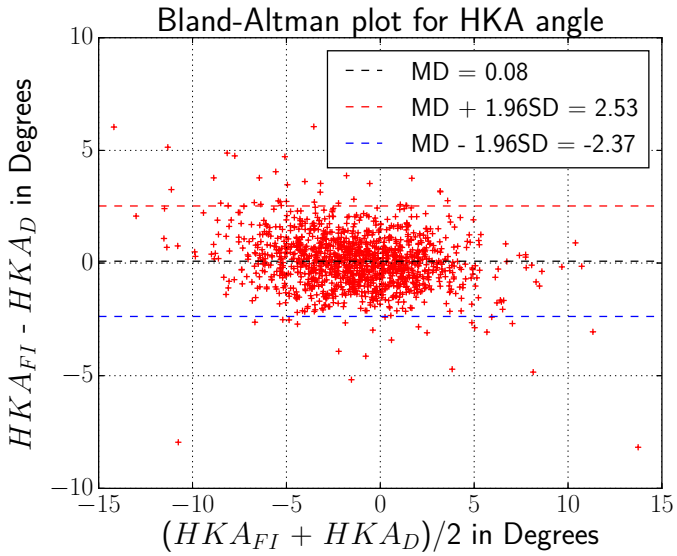


Figure B.1: Bland-Altman plot of HKA from full Full image CNN against HKA from Dr. Duryea.

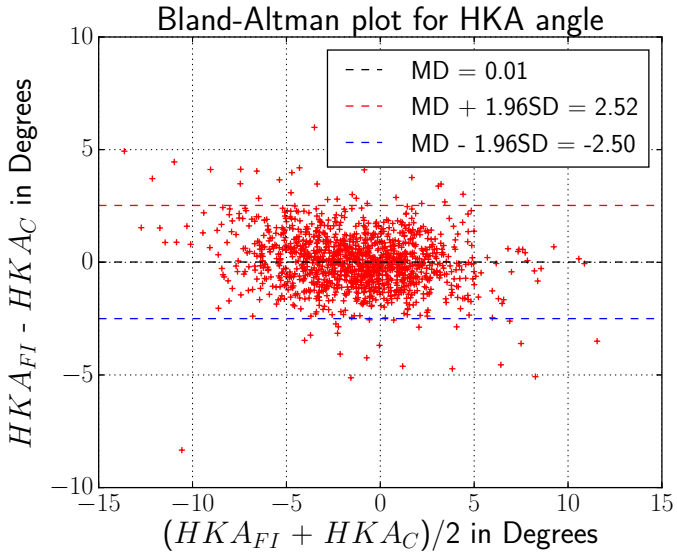


Figure B.2: Bland-Altman plot of HKA from full image CNN against HKA from Dr. Cooke.

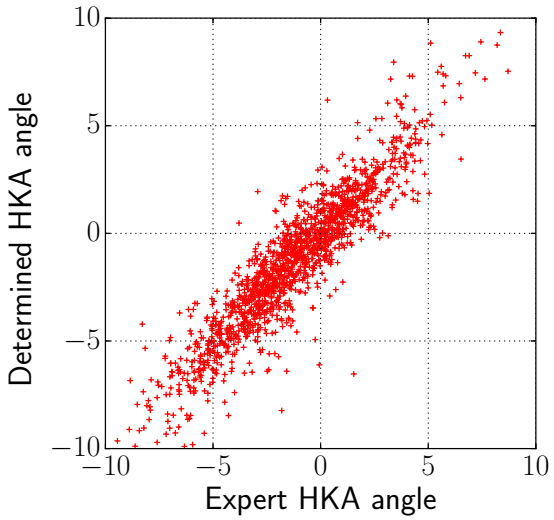
Predicted \ Expert	Varus	Neutral	Valgus
Varus	501	72	1
Neutral	84	550	64
Valgus	0	37	175

(a) Half image CNN (containing one leg)

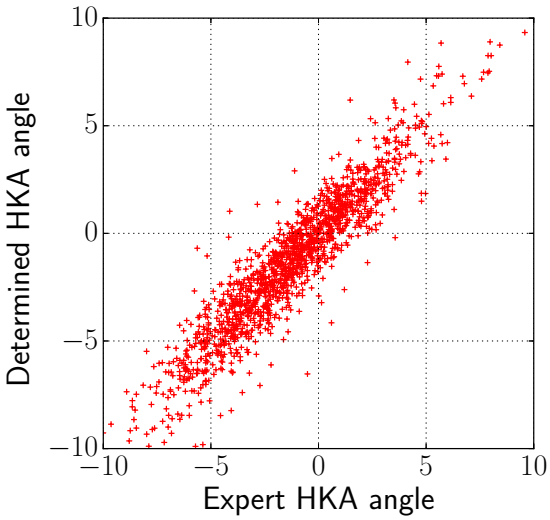
Predicted \ Expert	Varus	Neutral	Valgus
Varus	527	70	0
Neutral	79	560	49
Valgus	0	49	150

(b) Full image CNN (containing two legs)

Figure B.3: Confusion Matrix of leg alignment classification using outcome of HKA angle from images



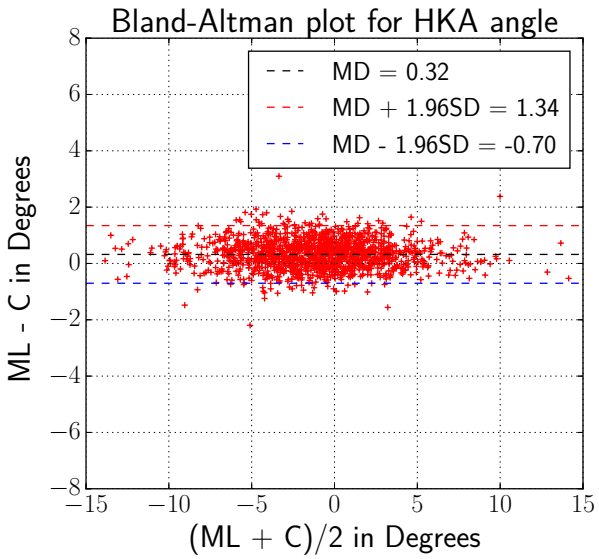
(a) Scatter plot: Half image CNN



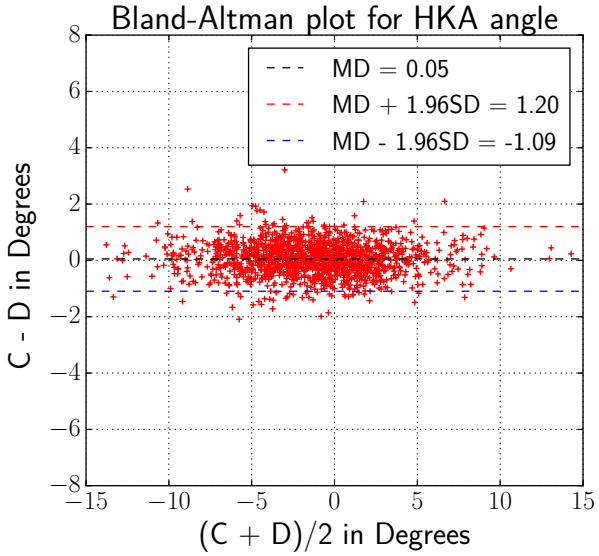
(b) Scatter plot: Full image CNN

Figure B.4: HKA from Images scatter plots.

B.2 Evaluation of manually labeled landmarks



(a)



(b)

Figure B.5: Evaluation of agreement between HKA angle from manually labelled landmarks and expert's HKA angle. ML = manually labelled, D = Dr. Duryea, C = Dr. Cooke.

Dr. Cooke \ Dr.Duryea	Varus	Neutral	Valgus
Varus	585	26	0
Neutral	41	508	23
Valgus	0	19	239

(a) Dr. Duryea and Dr. Cooke

ML \ Dr. Duryea	Varus	Neutral	Valgus
Varus	677	85	0
Neutral	12	639	51
Valgus	0	12	295

(b) Manual landmarks and Dr. Duryea

ML \ Dr. Cooke	Varus	Neutral	Valgus
Varus	550	65	0
Neutral	13	523	40
Valgus	0	7	254

(c) Manual landmarks and Dr. Cooke

Figure B.6: Confusion matrices of leg alignment classification based on HKA angle measurement from manually labeled landmarks.

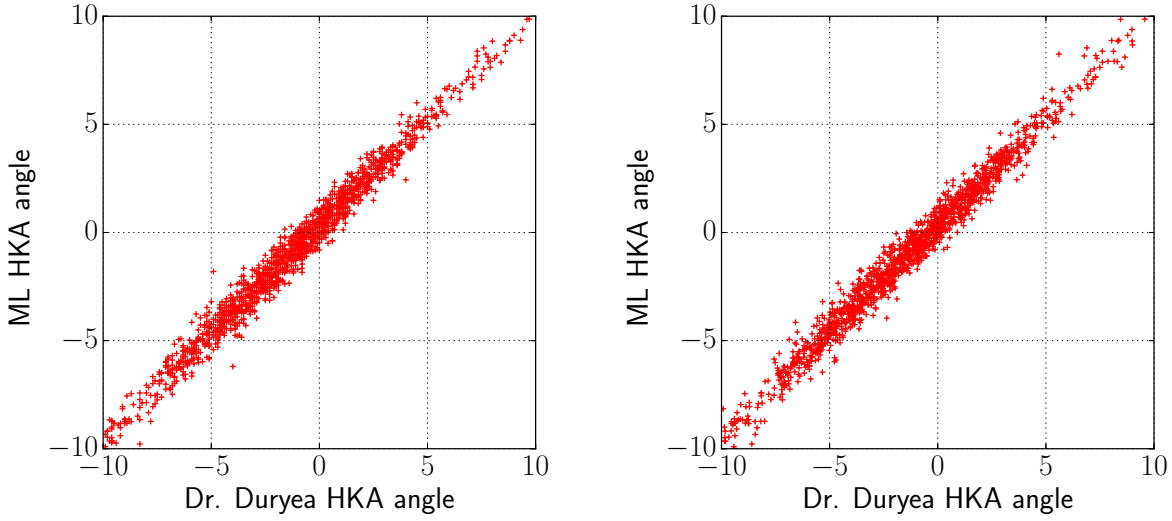
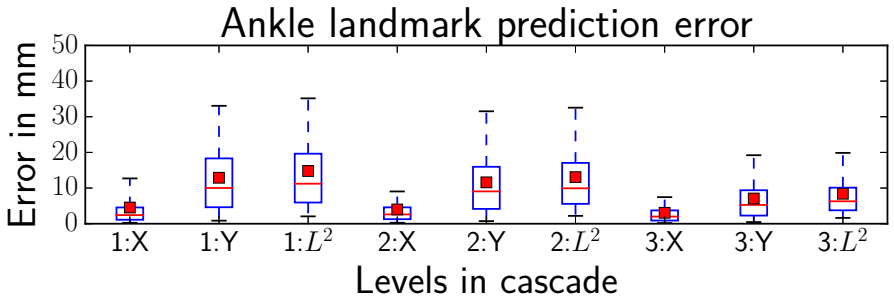
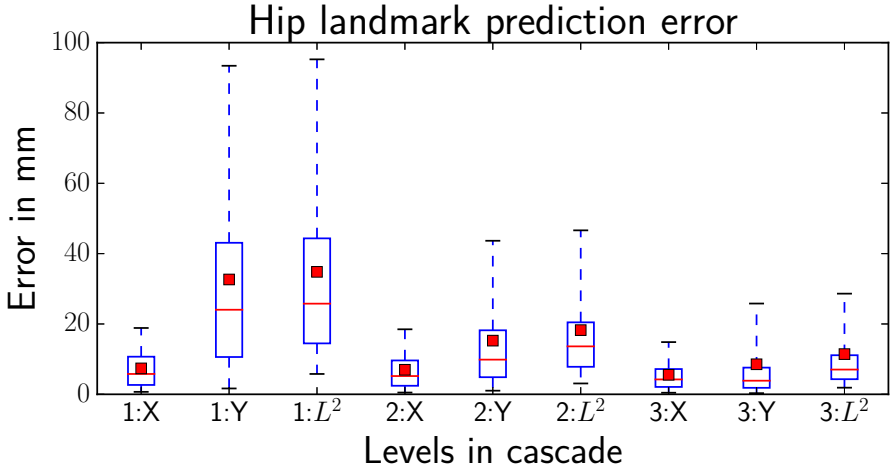


Figure B.7: Scatter plots of HKA angles of manually labeled HKA angles and experts angles

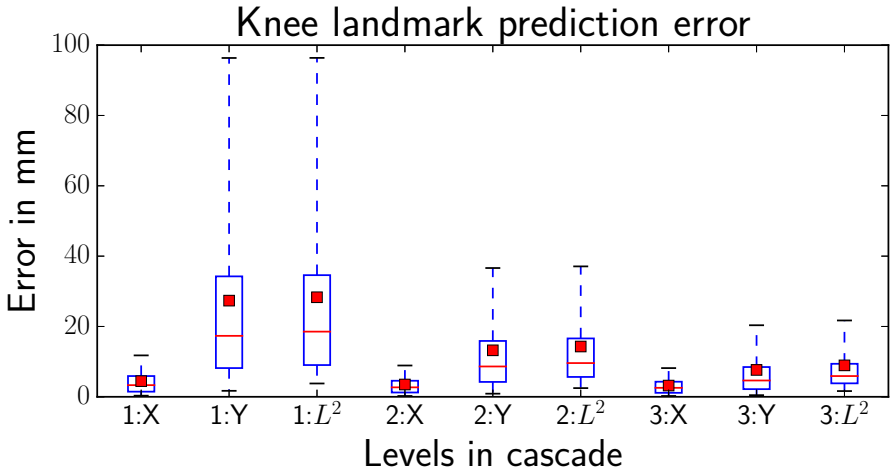
B.3 Evaluation HKA from landmarks using heuristic initial ROI



(a) Ankle landmark error



(b) Hip landmark error



(c) Knee landmark error

Figure B.8: Landmark detection accuracy using a heuristically defineid initial ROI: 1, 2, 3 correspond to the level in the cascade.

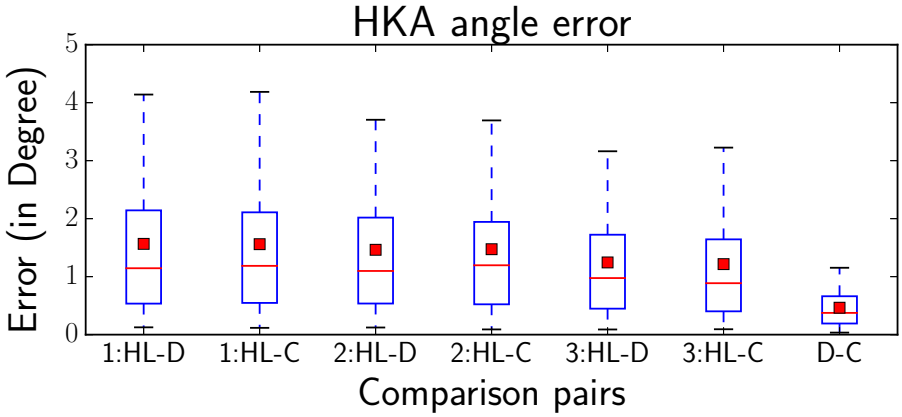


Figure B.9: HKA angle determination accuracy in differnt levels of the cascade: 1,2,3 = levels in cascaded pipeline, HL = HKA from landmarks using heuristic initial ROI, D = Dr. Duryea, C = Dr. Cooke.

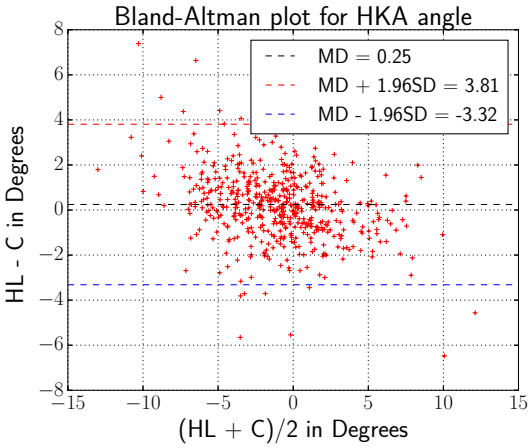
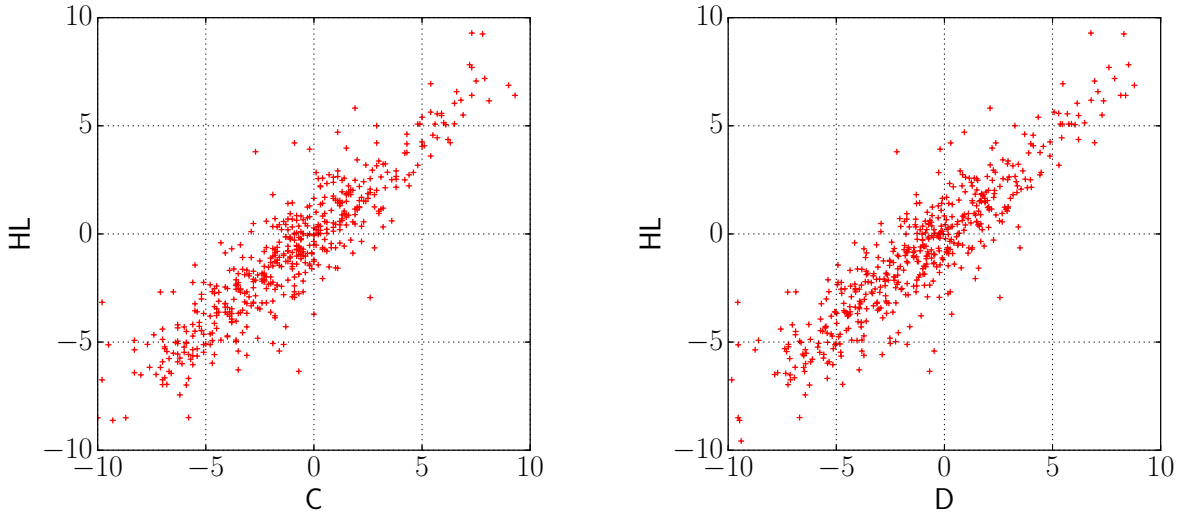


Figure B.10: Bland-Altman plot of HKA angle from automatically determined landmarks using heuristically defined initial ROI against Dr. Cooke.



(a) Automatically determined HKA angle and Dr. Duryea HKA angle.

(b) Automatically determined HKA angle and Dr. Cooke HKA angle.

Figure B.11: Evaluation of HKA from heuristiclly determined initial ROI using scatter plots.

Dr.Duryea

Dr. Cooke	Varus	Neutral	Valgus
Varus	198	8	0
Neutral	16	205	9
Valgus	0	6	87

(a) Confusion matrix: HKA classification based on Dr. Cooke and Dr. Duryea.

Dr. Duryea

HL	Varus	Neutral	Valgus
Varus	193	68	2
Neutral	22	232	22
Valgus	0	37	70

(b) Confusion matrix: HKA classification based on Manual landmarks and Dr. Duryea.

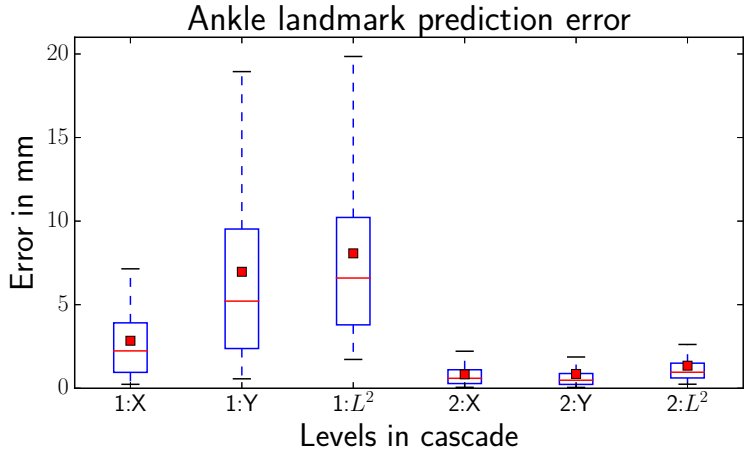
Dr. Cooke

HL	Varus	Neutral	Valgus
Varus	157	47	2
Neutral	16	193	21
Valgus	1	30	62

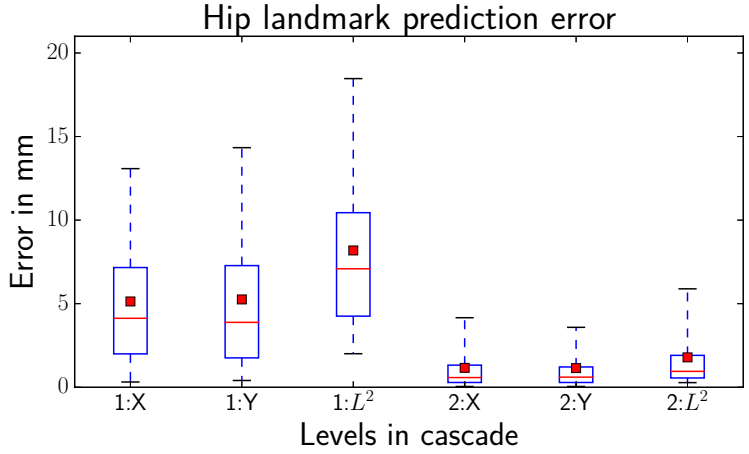
(c) Confusion matrix: HKA classification based on Manual lmrks and Dr. Cooke.

Figure B.12: Confusion matrices of leg alignment classification based on HKA angle measurement using heuristically defined initial ROI.

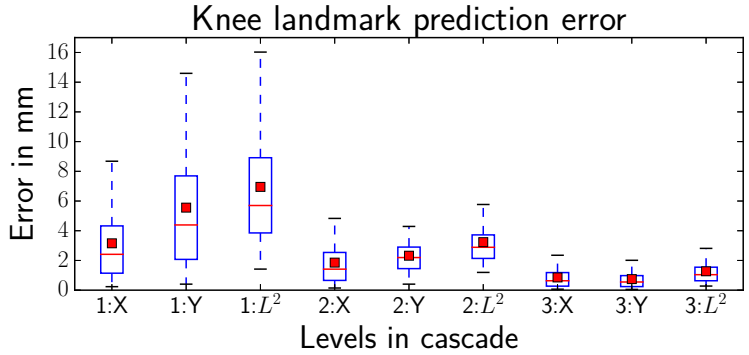
B.4 Evaluation of Automated ROI landmarks



(a) Ankle landmark error



(b) Hip landmark error



(c) Knee landmark error: Three level landmark detection was implemented.

Figure B.13: HKA angle determination accuracy in different levels of the cascaded pipeline: 1,2,3 = levels in cascade pipeline, C = Dr. Cooke, D = Dr. Duryea.

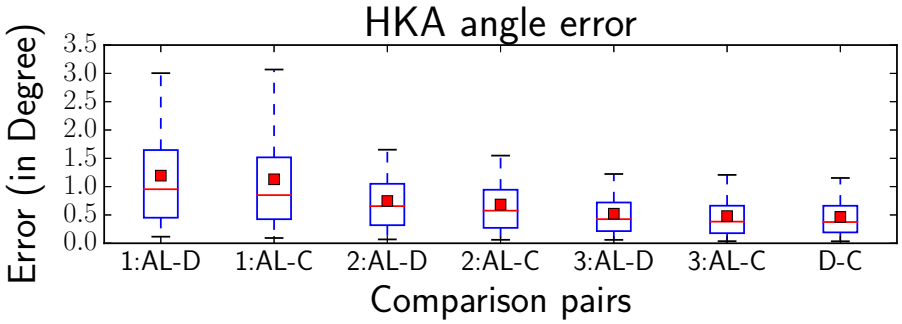
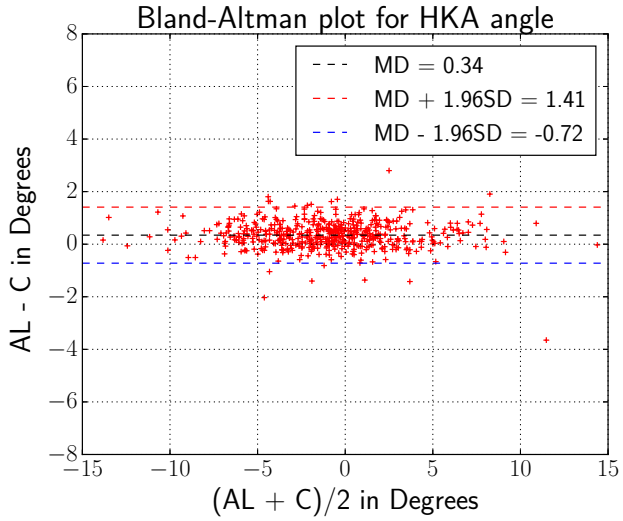
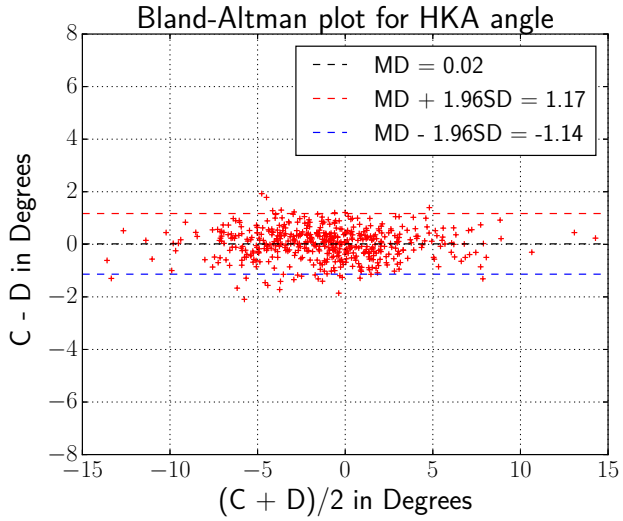


Figure B.14: Heuristic initial ROI HKA angle determination error : The absolute value of difference between automatically determined HKA angle and expert determined HKA angle (1,2,3 = levels in cascaded pipeline, AL = HKA from landmarks, D = Dr. Duryea, C = Dr. Cooke).

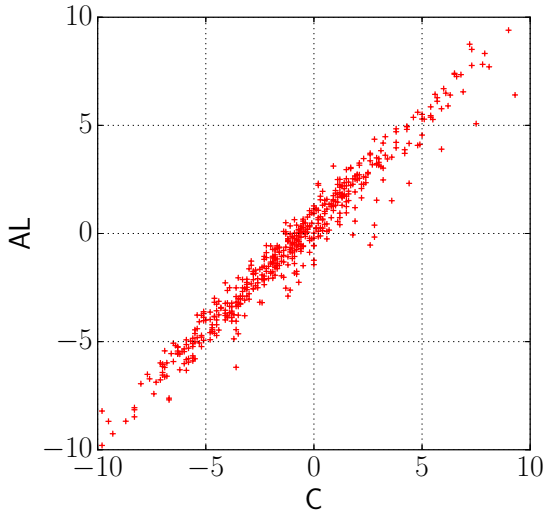


(a) Aut. determined HKA angle vs. Dr. Cooke.

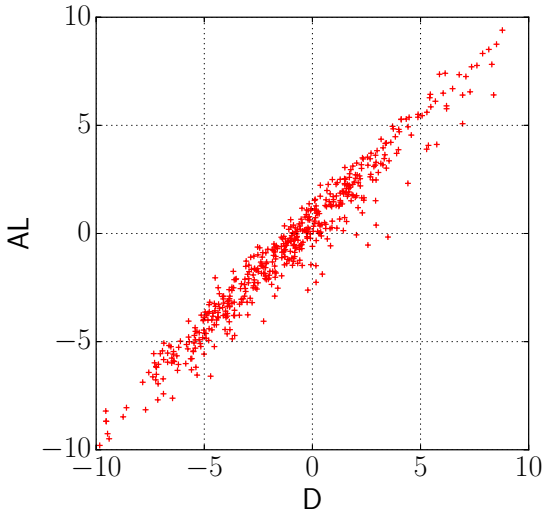


(b) Dr. Cooke vs. Dr. Duryea.

Figure B.15: Bland-Altman plot of HKA angles determined from landmarks determined using automatically defined initial ROI



(a) Aut. determined HKA angle and Dr. Duryea HKA angle.



(b) Aut. determined HKA angle and Dr. Cooke HKA angle.

Figure B.16: Evaluation of HKA from heuristiclly determined initial ROI

Dr. Cooke \ Dr.Duryea	Varus	Neutral	Valgus
Varus	198	8	0
Neutral	16	205	9
Valgus	0	6	87

(a) Confusion matrix: HKA classification based on Dr. Cooke and Dr. Duryea.

ML \ Dr. Duryea	Varus	Neutral	Valgus
Varus	232	31	0
Neutral	4	249	23
Valgus	0	7	100

(b) Confusion matrix: HKA classification based on Manual lmrks and Dr. Duryea.

ML \ Dr. Cooke	Varus	Neutral	Valgus
Varus	185	21	0
Neutral	4	207	19
Valgus	0	3	90

(c) Confusion matrix: HKA classification based on Manual landmarks and Dr. Cooke.

B.5 Evaluation of Automated ROI landmarks on the entire dataset

Dr. Cooke \ Dr.Duryea	Varus	Neutral	Valgus
Varus	1967	138	0
Neutral	148	2172	74
Valgus	0	116	719

(a) Confusion matrix: HKA classification based on Dr. Cooke and Dr. Duryea.

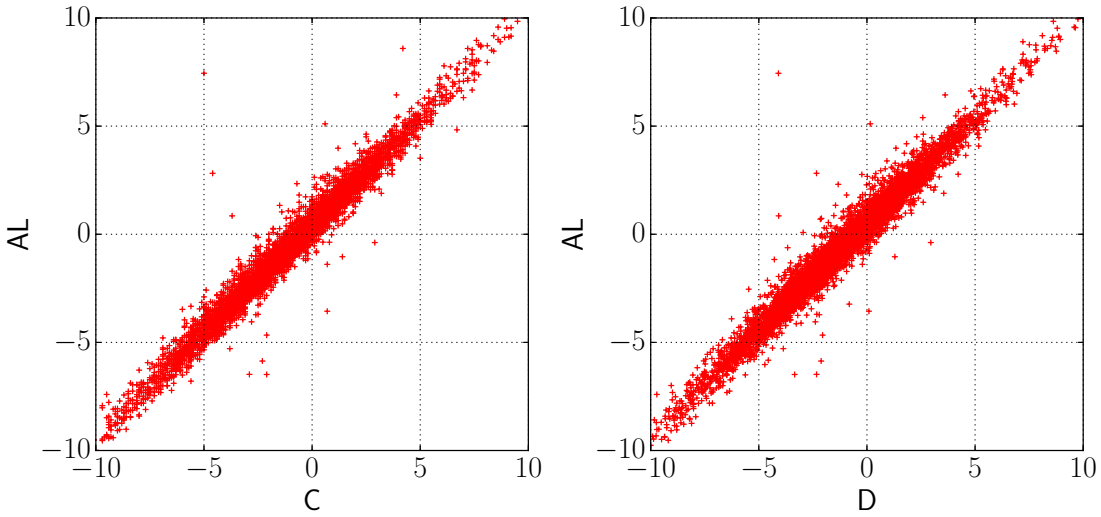
ML \ Dr. Duryea	Varus	Neutral	Valgus
Varus	1815	421	2
Neutral	25	2267	276
Valgus	0	25	805

(b) Confusion matrix: HKA classification based on automatically determined HKA angles and Dr. Duryea.

ML \ Dr. Cooke	Varus	Neutral	Valgus
Varus	1736	386	2
Neutral	17	2164	223
Valgus	0	321	819

(c) Confusion matrix: HKA classification based on automatically determined HKA angles and Dr. Cooke.

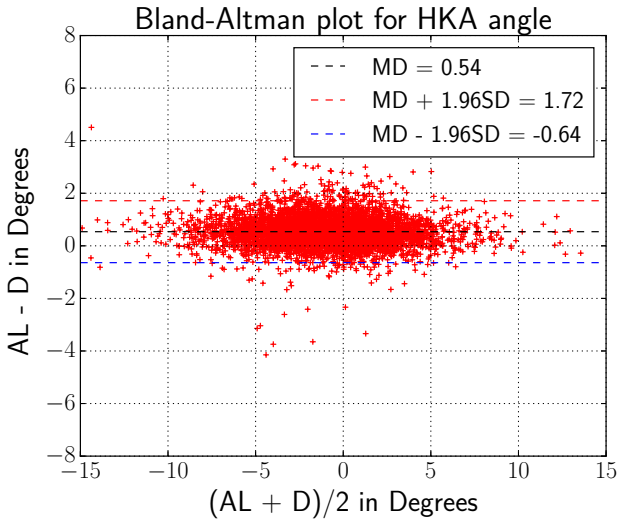
Figure B.18: Confusion matrix of HKA angle determination method using a large dataset which is not labeled with landmarks.



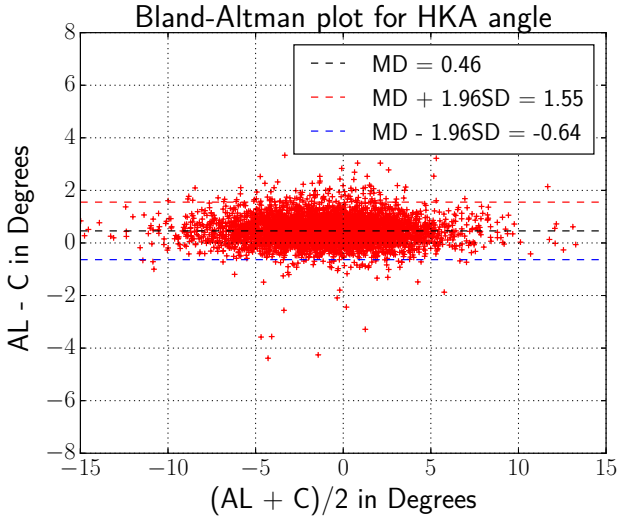
(a) Automatically determined HKA to Dr. Cooke.

(b) Automatically determined HKA to Dr. Duryea.

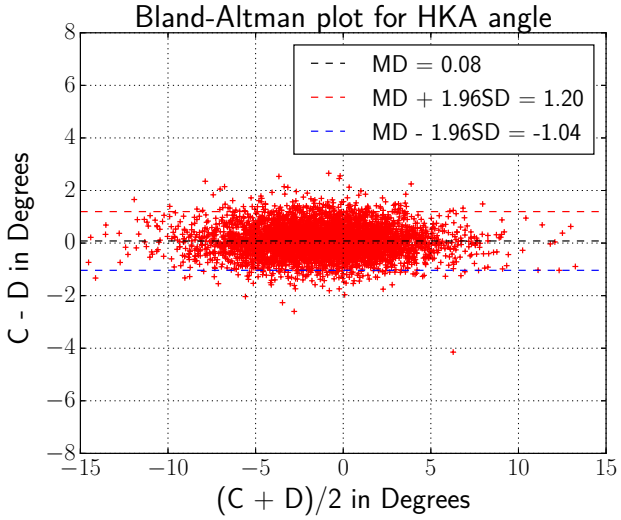
Figure B.19: Scatter plots of HKA angle determination method using a large dataset which is not labeled with landmarks.



(a) Automatically determined HKA angles and Dr. Dureyas.



(b) Automatically determined HKA angles and Dr. Cooke's.



(c) Dr. Duryea's and Dr. Cooke's.

Figure B.20: Bland-Altman evaluation of HKA angle determination method using a large dataset which is not labeled with landmarks.

C Example images

C.1 Knee landmark detection

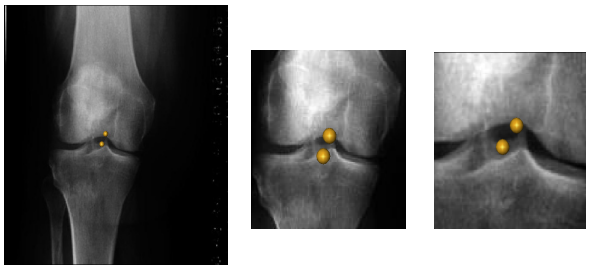


Figure C.1: Three level knee landmark detection with cascaded landmark detection with automatically defined initial ROI.

C.2 ROI detector failures

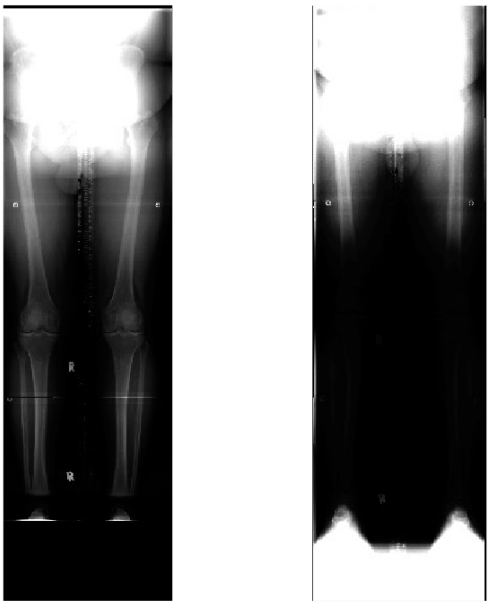


Figure C.2: ROI detector failure examples: The left shows artifact at the ankle joint and the right image shows bad contrast.

D What do the CNNs learn?

We used visualization tools [73], to get saliency maps. These maps show us the pixels which affect the output of a CNN.

D.1 Landmark determination CNNs

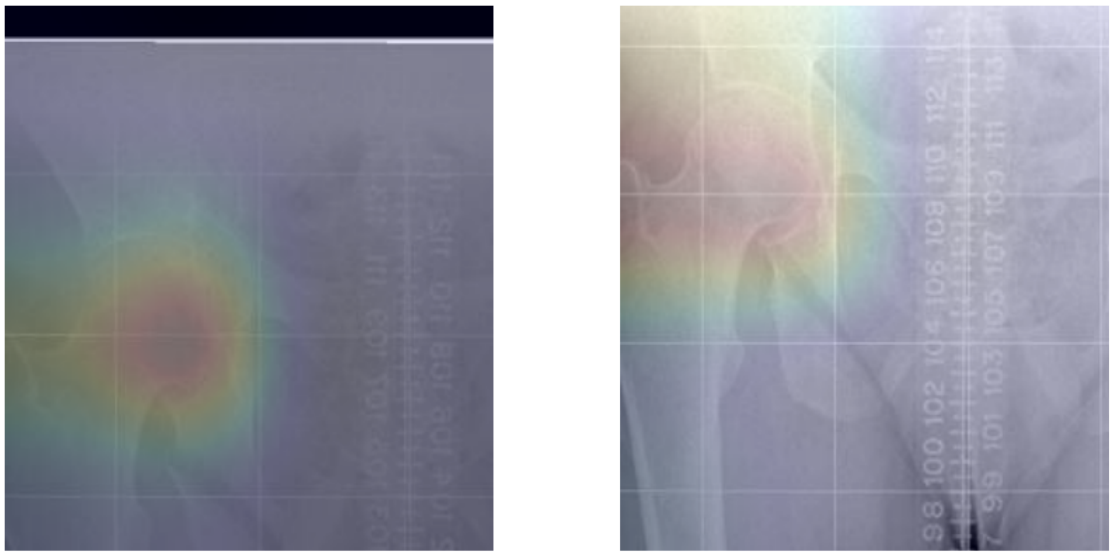


Figure D.1: Saliency map visualization of level 1 landmark detection of the femoral head.

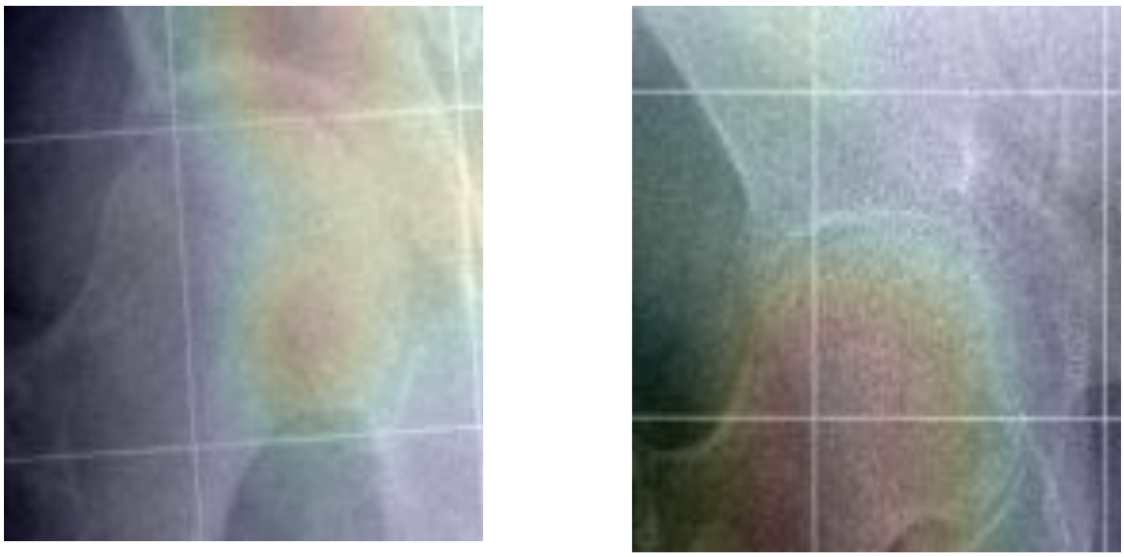


Figure D.2: Saliency map visualization of level 2 landmark detection of the femoral head.

D.2 HKA angle determination CNNs

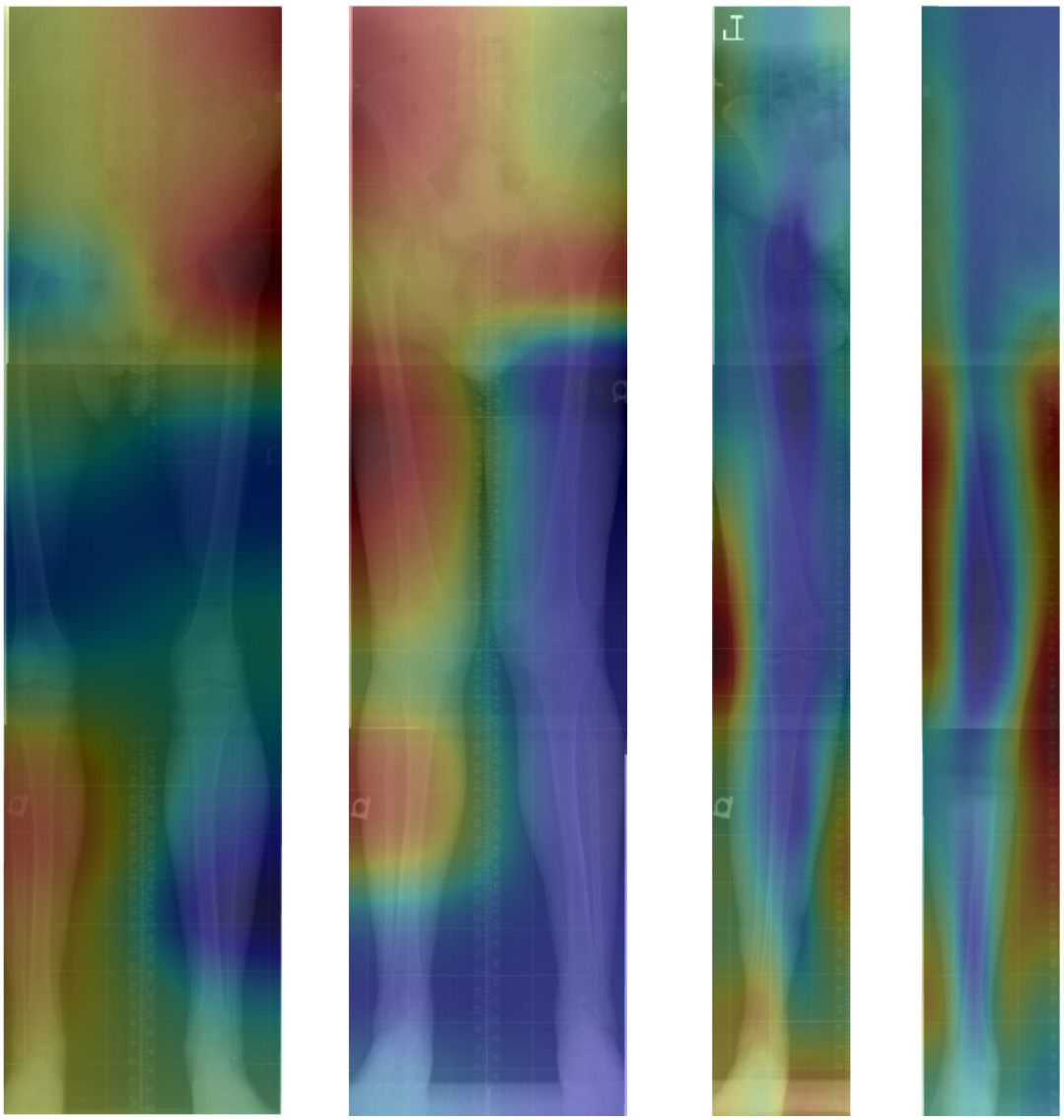


Figure D.3: Saliency map visualization of HKA angle determination: Two left images are for CNN whose input is full image and the two right are for CNNs whose input is half image.

Bibliography

[1] Marita Cross, Emma Smith, Damian Hoy, Sandra Nolte, Ilana Ackerman, Marlene Fransen, Lisa Bridgett, Sean Williams, Francis Guillemin, Catherine L Hill, et al. The global burden of hip and knee osteoarthritis: estimates from the global burden of disease 2010 study. *Annals of the rheumatic diseases*, 73(7):1323–1330, 2014.

[2] Leena Sharma. The role of varus and valgus alignment in knee osteoarthritis. *Arthritis & Rheumatism*, 56(4):1044–1047, 2007.

[3] Bandy legs shape. `https://www.shutterstock.com/image-vector/bandylegs-shape-legs-genu-varum-valgum-389876083`. Accessed: 2018-12-09.

[4] K Deep, P Nunag, N Willcox, AH Deakin, and F Picard. A comparison of three different methods of measure-ment of knee deformity in osteoarthritis. *J Orth Rhe Sp Med*, 1(1):107, 2016.

[5] Lisa Sheehy, David Felson, Y Zhang, Jingbo Niu, Y-M Lam, Neil Segal, John Lynch, and T Derek V Cooke. Does measurement of the anatomic axis consistently predict hip-knee-ankle angle (hka) for knee alignment studies in osteoarthritis? analysis of long limb radiographs from the multicenter osteoarthritis (most) study. *Osteoarthritis and cartilage*, 19(1):58–64, 2011.

[6] T Iranpour-Boroujeni, J Li, JA Lynch, M Nevitt, J Duryea, OAI Investigators, et al. A new method to measure anatomic knee alignment for large studies of oa: data from the osteoarthritis initiative. *Osteoarthritis and cartilage*, 22(10):1668–1674, 2014.

[7] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[8] Ryszard S Michalski, Jaime G Carbonell, and Tom M Mitchell. *Machine learning: An artificial intelligence approach*. Springer Science & Business Media, 2013.

[9] Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[10] Waseem Rawat and Zenghui Wang. Deep convolutional neural networks for image classification: A comprehensive review. *Neural computation*, 29(9):2352–2449, 2017.

[11] Nannan Wang, Xinbo Gao, Dacheng Tao, and Xuelong Li. Facial feature point detection: A comprehensive survey. *arXiv preprint arXiv:1410.1037*, 2014.

[12] Yue Wu and Qiang Ji. Facial landmark detection: a literature survey. *CoRR*, abs/1805.05563, 2018.

[13] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. *International Journal of Computer Vision*, 91(2):200–215, 2011.

[14] Tim Cootes, ER Baldock, and J Graham. An introduction to active shape models. In *Image processing and analysis*, chapter 7, pages 223–248. Oxford Univesity Press, 2000.

[15] Tim F Cootes, Mircea C Ionita, Claudia Lindner, and Patrick Sauer. Robust and accurate shape model fitting using random forest regression voting. In *European Conference on Computer Vision*, pages 278–291. Springer, 2012.

[16] Amir Zadeh, T Baltrusaitis, and Louis-Philippe Morency. Convolutional experts network for facial landmark detection. In *Proceedings of the International Conference on Computer Vision & Pattern Recognition (CVPRW), Faces-in-the-wild Workshop/Challenge*, volume 3, page 6, 2017.

[17] Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[18] Tobias Heimann and Hans-Peter Meinzer. Statistical shape models for 3d medical image segmentation: A review. *Medical Image Analysis*, 13(4):543 – 563, 2009.

[19] *Multi-view Facial Landmark Detection*. PhD thesis, Czech Technical University in Prague.

[20] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (6):681–685, 2001.

[21] Michael J Jones and Tomaso Poggio. Multidimensional morphable models: A framework for representing and matching object classes. *International Journal of Computer Vision*, 29(2):107–131, 1998.

[22] Gareth J Edwards, Christopher J Taylor, and Timothy F Cootes. Interpreting face images using active appearance models. In *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*, pages 300–305. IEEE, 1998.

[23] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57(2):137–154, 2004.

[24] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[25] Timo Ojala, Matti Pietikäinen, and David Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern recognition*, 29(1):51–59, 1996.

[26] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[27] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[28] Matthias Dantone, Juergen Gall, Gabriele Fanelli, and Luc Van Gool. Real-time facial feature detection using conditional regression forests. In *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on*, pages 2578–2585. IEEE, 2012.

[29] Shaoqing Ren, Xudong Cao, Yichen Wei, and Jian Sun. Face alignment at 3000 fps via regressing local binary features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1685–1692, 2014.

[30] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 532–539, 2013.

[31] Heng Yang and Ioannis Patras. Privileged information-based conditional regression forest for facial feature detection. In *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*, pages 1–6. IEEE, 2013.

[32] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[33] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503, 2016.

[34] Xudong Cao, Yichen Wei, Fang Wen, and Jian Sun. Face alignment by explicit shape regression. *International Journal of Computer Vision*, 107(2):177–190, 2014.

[35] Michel Valstar, Brais Martinez, Xavier Binefa, and Maja Pantic. Facial point detection using boosted regression and graph models. In *Computer Vision and*

Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 2729–2736. IEEE, 2010.

[36] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013.

[37] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014.

[38] Piotr Dollár, Peter Welinder, and Pietro Perona. Cascaded pose regression. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1078–1085. IEEE, 2010.

[39] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep convolutional network cascade for facial point detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3476–3483, 2013.

[40] Erjin Zhou, Haoqiang Fan, Zhimin Cao, Yuning Jiang, and Qi Yin. Extensive facial landmark localization with coarse-to-fine convolutional network cascade. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 386–391, 2013.

[41] George Trigeorgis, Patrick Snape, Mihalis A Nicolaou, Epameinondas Antonakos, and Stefanos Zafeiriou. Mnemonic descent method: A recurrent process applied for end-to-end face alignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4187, 2016.

[42] Zhenliang He, Meina Kan, Jie Zhang, Xilin Chen, and Shiguang Shan. A fully end-to-end cascaded cnn for facial landmark detection. In *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017)*, pages 200–207. IEEE, 2017.

[43] Zhanpeng Zhang, Ping Luo, Chen Change Loy, and Xiaoou Tang. Facial landmark detection by deep multi-task learning. In *European Conference on Computer Vision*, pages 94–108. Springer, 2014.

[44] Yue Wu, Tal Hassner, KangGeon Kim, Gerard Medioni, and Prem Natarajan. Facial landmark detection with tweaked convolutional neural networks. *IEEE transactions on pattern analysis and machine intelligence*, 2017.

[45] Anh Tuan Tran, Tal Hassner, Iacopo Masi, and Gérard Medioni. Regressing robust and discriminative 3d morphable models with a very deep neural network. In *Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on*, pages 1493–1502. IEEE, 2017.

[46] Zhujin Liang, Shengyong Ding, and Liang Lin. Unconstrained facial landmark localization with backbone-branches fully-convolutional networks. *arXiv preprint arXiv:1507.03409*, 2015.

[47] Jie Zhang, Shiguang Shan, Meina Kan, and Xilin Chen. Coarse-to-fine auto-encoder networks (cfan) for real-time face alignment. In *European Conference on Computer Vision*, pages 1–16. Springer, 2014.

[48] Rajeev Ranjan, Vishal M Patel, and Rama Chellappa. Hyperface: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017.

[49] CB Chang, J-Y Choi, IJ Koh, ES Seo, SC Seong, and TK Kim. What should be considered in using standard knee radiographs to estimate mechanical alignment of the knee? *Osteoarthritis and cartilage*, 18(4):530–538, 2010.

[50] T Derek V Cooke, Elizabeth A Sled, and R Allan Scudamore. Frontal plane knee alignment: a call for standardized measurement. *Journal of Rheumatology*, 34(9):1796, 2007.

[51] T Iranpour-Boroujeni, J Li, JA Lynch, M Nevitt, J Duryea, OAI Investigators, et al. A new method to measure anatomic knee alignment for large studies of oa: data from the osteoarthritis initiative. *Osteoarthritis and cartilage*, 22(10):1668–1674, 2014.

[52] Keith M Baumgarten, Stephen Fealy, Stephen Lyman, and Thomas L Wickiewicz. The coronal plane high tibial osteotomy. part 1: a clinical and radiographic analysis of intermediate term outcomes. *HSS Journal*, 3(2):147–154, 2007.

[53] Osteoarthritis Initiative. *Radiographic Procedure Manual for Examinations of the Knee, Hand, Pelvis and Lower Limbs*. Osteoarthritis Initiative, 2006.

[54] Luis Perez and Jason Wang. The effectiveness of data augmentation in image classification using deep learning. *arXiv preprint arXiv:1712.04621*, 2017.

[55] The osteoarthritis initiative. `https://data-archive.nimh.nih.gov/oai/`. Accessed: 2018-11-16.

[56] M Nevitt, D Felson, and G Lester. The osteoarthritis initiative: protocol for the cohort study. 2006. *URL: http://oai. epi-ucsf. org/datarelease/docs/StudyDesignProtocol. pdf*, 2016.

[57] Centeral assessment of full-limb x-rays for frontal plane lower limb alignment. OAI lower limb alignment dataset accompanying document, 2017.

[58] Amirazibedition-2018.11. `https://amira.zib.de/`, 2018.

[59] Davide Giavarina. Understanding bland altman analysis. *Biochemia medica: Biochemia medica*, 25(2):141–151, 2015.

[60] J Martin Bland and DouglasG Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The lancet*, 327(8476):307–310, 1986.

[61] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[62] Terry K Koo and Mae Y Li. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of chiropractic medicine*, 15(2):155–163, 2016.

[63] Patrick E Shrout and Joseph L Fleiss. Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*, 86(2):420, 1979.

[64] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014.

[65] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[66] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[67] R Moyer, W Wirth, J Duryea, and F Eckstein. Anatomical alignment, but not goniometry, predicts femorotibial cartilage loss as well as mechanical alignment: data from the osteoarthritis initiative. *Osteoarthritis and cartilage*, 24(2):254–261, 2016.

[68] Christian Payer, Darko Štern, Horst Bischof, and Martin Urschler. Regressing heatmaps for multiple landmark localization using cnns. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 230–238. Springer, 2016.

[69] Yuanwei Li, Amir Alansary, Juan J Cerrolaza, Bishesh Khanal, Matthew Sinclair, Jacqueline Matthew, Chandni Gupta, Caroline Knight, Bernhard Kainz, and Daniel Rueckert. Fast multiple landmark localisation using a patch-based iterative network. *arXiv preprint arXiv:1806.06987*, 2018.

[70] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.

[71] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[72] Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 580–587, 2014.

[73] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.