Neil Cronin

AUTOMATED ANALYSIS OF MUSCULOSKELETAL ULTRASOUND IMAGES USING DEEP LEARNING

UNIVERSITY OF JYVÄSKYLÄ
FACULTY OF INFORMATION TECHNOLOGY
2020

ABSTRACT

Cronin, Neil
Automated Analysis of Musculoskeletal Ultrasound Images Using Deep Learning
University of Jyväskylä, 2020, 56 pp.
Cognitive Computing and Collective Intelligence, Master's Thesis
Supervisor: Terziyan, Vagan

Ultrasound is widely used to image musculoskeletal tissues, and offers many benefits including low cost, portability, and non-invasiveness. However, the analysis of ultrasound images remains an area in need of development. Until very recently, all analyses were performed manually, which is very subjective and time-consuming. There are currently very few open source methods designed for this purpose. Current approaches also tend to rely on rules-based analyses, and such methods tend to fail if images exhibit large deviations from those that were used to develop the method. In recent years, deep learning approaches to medical image analysis have been shown to yield excellent segmentation results on various imaging modalities, but such approaches have not yet been broadly applied to musculoskeletal ultrasound images. In this thesis I present a deep learning-based automated method that computes muscle architectural parameters from ultrasound images and videos. The method is based on the U-net architecture, and models were trained using hand-labelled images of muscle fascicles and aponeuroses. These trained models were then used to perform pixel-wise semantic segmentation of new images, classifying each pixel as one of three possible classes (fascicle, aponeurosis, other). The results were compared to manual analysis performed by two independent researchers, as well as to an existing semi-automated method. In general, the method performed very favourably when compared to manual and semi-automated analysis, and was robust to images from different muscles and those obtained with different ultrasound systems and settings. The method is also able to detect multiple muscle fascicles in a given image. The approach presented here offers an objective, time-efficient method of segmenting ultrasound images that does not require any user input. The method and all labelled training data are available under an open source license, allowing others to use and extend this work.

Keywords: Artificial intelligence, Deep learning, convolutional neural network, Keras, musculoskeletal ultrasound, muscle mechanics, aponeurosis, muscle fascicle, Python, semantic segmentation, Tensorflow, U-net

SUOMENKIELINEN TIIVISTELMÄ

Ultraääntä käytetään paljon tuki- ja liikuntaelimistön kudosten kuvaamiseen, etuina ollen matalat kustannukset, siirrettävyys sekä ei-invasiivisuus. Ultraäänikuvien analysointi vaatii edelleen kehitystä. Viime aikoihin saakka kaikki analyysit on tehty manuaalisesti, mikä on hyvin subjektiivista ja aikaa vievää. Tällä hetkellä on olemassa muutamia tähän tarkoitukseen suunniteltuja avoimen lähdekoodin menetelmiä. Tämänhetkiset lähestymistavat perustuvat yleensä sääntöpohjaisiin analyyseihin. Useat niistä tekevät virheitä, jos analysoitavissa kuvissa on suuria poikkeamia menetelmän kehityksessä käytettyihin kuviin verrattuna. Viime vuosina syväoppimisen käyttäminen lääketieteellisissä kuva-analyyseissä on tuottanut erinomaisia segmentointituloksia eri kuvantamistavoilla, mutta useimpia niistä ei ole vielä sovellettu laajemmin tuki- ja liikuntaelimistön ultraäänikuviin. Tässä tutkielmassa esittelen syväoppimiseen perustuvan automatisoidun menetelmän, joka laskee lihasten rakenteeseen liittyviä parametreja ultraäänikuvista ja -videoista. Tämä menetelmä perustuu U-net arkkitehtuuriin, ja mallit on opetettu käsin merkityillä lihassolukimppujen ja aponeuroosien kuvilla. Näiden opetettujen mallien avulla tehtiin pikselitasoinen semanttinen segmentointi uusille kuville luokittelemalla jokainen pikseli yhteen kolmesta mahdollisesta luokasta (lihassolukimppu, aponeuroosi, muu). Tuloksia verrattiin kahden eri tutkijan tekemiin manuaalisiin analyyseihin sekä olemassa olevaan puoliautomaattiseen menetelmään. Kehitetty menetelmä suoriutui yleisesti hyvin verrattuna manuaaliseen ja puoliautomaattiseen analyysiin, ja toimi vakaasti eri ultraäänilaitteilla ja asetuksilla otetuilla kuvilla. Menetelmä kykeni myös havaitsemaan monia lihassolukimppuja samassa kuvassa. Esitelty lähestymistapa tarjoaa objektiivisen, aikaa säästävän menetelmän ultraäänikuvien segmentointiin ilman käyttäjän syötettä. Menetelmä ja kaikki merkitty data ovat saatavilla avoimen lähdekoodin lisenssillä, mahdollistaen toisten hyödyntää ja laajentaa tätä työtä.

Avainsanat: Tekoäly, syväoppiminen, konvoluutioneuroverkko, Keras, muskuloskeletaalinen ultraääni, lihasmekaniikka, aponeuroosi, lihassyy, Python, semanttinen segmentointi, Tensorflow, U-net

GLOSSARY

CNN, convolutional neural network
CPU, central processing unit
CT, computed tomography
CV, coefficient of variation
DL, deep learning
FL, muscle fascicle length
GPU, graphical processing unit
IoU, intersection-over-union
MRI, magnetic resonance imaging
MVC, maximal isometric voluntary contraction
Penn, pennation angle
ReLu, rectified linear unit
ResNet, residual neural network
RPC, reproducibility coefficient
SSE, sum of squared errors
Thick, muscle thickness
U-net, U-shaped neural network architecture
VGG, neural network architecture by the Oxford Visual Geometry Group

FIGURES

TABLES

TABLE OF CONTENTS

1 INTRODUCTION

Among many other applications, ultrasound has been used to examine muscle and tendon function for the past three decades. This imaging modality holds many advantages over other techniques, including its relatively low cost, portability, and the ability to display real-time images non-invasively. It is currently the only technique that can be used to image dynamic muscle activity, e.g. during muscle contraction or more complex human movements like walking and running (Cronin & Lichtwark, 2013; Leitner et al., 2019).

Despite the widespread usage of the ultrasound method, the analysis of ultrasound images remains an area in need of development. Until very recently, all analyses were performed manually. This requires the user to place multiple markers on an image, and to repeat this for each image in a sequence. Given that many systems sample data at 50-100 Hz, and that a typical walking stride can last around 1 s, this process rapidly becomes unfeasible as the size of the dataset increases.

There are several possible reasons why analysis approaches for ultrasound images lag behind those of other imaging modalities. In general, the spatial resolution and visual quality of ultrasound images is not very high. Moreover, unlike other imaging modalities such as MRI and CT scans, ultrasound imaging is usually performed with a freehand scanner, which means that there is a lack of standardisation between different users. The method can also be used to examine different muscles, which have quite different architecture. These factors all make it challenging to develop a single, robust analysis approach.

To date, a few attempts have been made to release open source methods to semi- or fully automatically analyse ultrasound images (Cronin, Carty, Barrett, & Lichtwark, 2011; Drazan, Hullfish, & Baxter, 2019; Farris & Lichtwark, 2016). However, these approaches usually focus on a single parameter of interest and/or are developed for use with a specific muscle. Moreover, current approaches often rely on rules-based analyses, and such methods tend to fail if

images exhibit large deviations from those that were used to develop the method. In recent years, deep learning approaches to medical image analysis have been shown to yield excellent segmentation results on various imaging modalities (e.g. Ronneberger, Fischer, & Brox, 2015). However, such approaches have not yet been applied to musculoskeletal ultrasound images.

1.1 Research questions

In this thesis I address the following specific research questions:

1. Can an automated method be developed that uses deep learning to accurately extract key muscle architectural parameters (muscle thickness, muscle fascicle length and pennation angle) from ultrasound images?
2. More specifically, can a modified implementation of the U-net architecture (and a unique combination of hyperparameters) be used to perform accurate semantic segmentation of this type of image?
2. Is this method robust enough to work with images from a range of different muscles and with different image settings?
3. How does the performance of this approach compare to the current gold standard methods in this field?

1.2 Thesis layout

This thesis is at the interface between medical imaging, image processing, and artificial intelligence. The literature review gives an overview of the most relevant themes, and details the current state-of-the-art in these areas. The methods section outlines the approach that was developed and tested in this thesis project. The results section details various metrics related to the training and subsequent inference performance of the approach, and comparisons are made with other similar methods. Finally, the discussion section places this work in the context of the field, and outlines future steps that could be taken on the basis of this work.

2 LITERATURE REVIEW

In this section I first outline the ultrasound method, and describe the history of the analysis process. I then introduce the field of deep learning, and summarise recent developments in this area that are relevant to the topic of this thesis. Finally, I outline potential applications of deep learning-based analysis methods in relation to musculoskeletal ultrasound images.

2.1 Musculoskeletal ultrasound

The ultrasound method has been used in medicine since the mid-1950's (Nicolson & Fleming, 2013). Traditionally, the main use of this method has been to examine a developing foetus in the maternal womb, helping to identify any obvious signs of developmental disorders at different stages of the pregnancy.

More recently, the value of this method was recognised in the musculoskeletal field, and with very little modification, essentially the same technology began to be used to study muscle and tendon function *in vivo*. Whereas previous approaches relied on static imaging or estimates of muscle function based on joint kinematics, the first studies in this area provided vital new information about how muscles actually function during dynamic settings.

Muscle architecture was first examined in vivo during rest (Narici et al., 1996; Rutherford & Jones, 1992) and different levels of voluntary contraction (Fukashiro, Itoh, Ichinose, Kawakami, & Fukunaga, 1995; Herbert & Gandevia, 1995). Since then, a vast array of studies have extended the analysis to more complex tasks such as walking and running (see below). Although the key findings of all of these studies are beyond the scope of this thesis, the common thread among all studies is a desire to quantify muscle architectural parameters (figure 1).

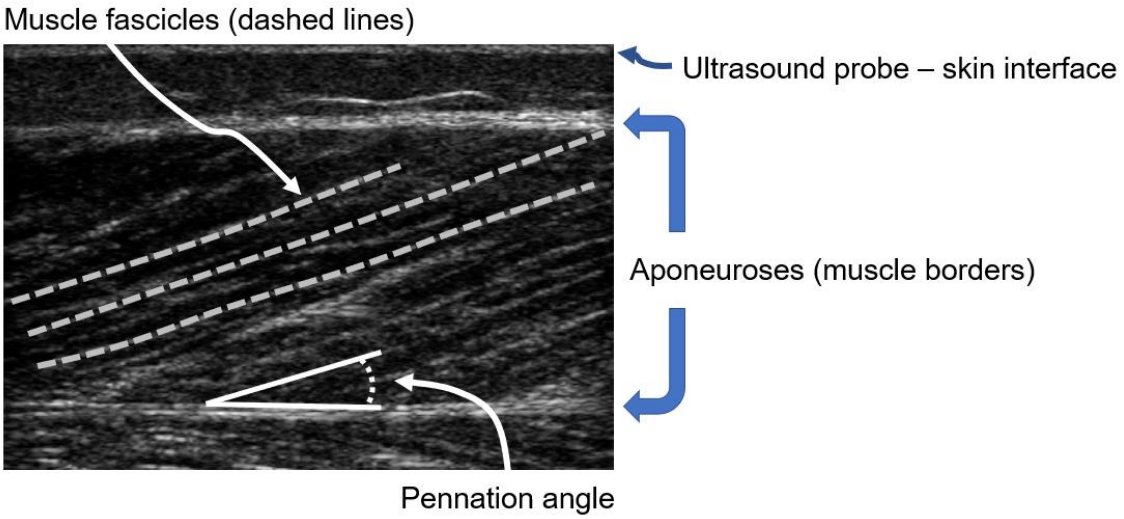

FIGURE 1 Example of an ultrasound image taken from the medial gastrocnemius muscle of the lower leg.

In figure 1 the key structures of interest are all visible. Depending on the study aims, one or more of the following parameters may be computed from this type of image:

- **Muscle fascicle length:** This usually involves visually identifying a striation that runs between the two aponeuroses, and where necessary, extrapolating it so that its ends meet the aponeuroses. In the case of figure 1, the lower end of each of the three identified fascicles extends beyond the edge of the image. In such cases it is necessary to extrapolate both the fascicle and lower aponeurosis trajectories, and estimate where they coincide.
- **Pennation angle:** This angle represents the angle at which muscle force is produced relative to the tendon that the muscle attaches to. It is usually obtained by identifying a muscle fascicle, and then calculating the angle of this fascicle relative to the corresponding aponeurosis (the lower aponeurosis in figure 1).
- **Muscle thickness:** This is determined as the distance between the two aponeuroses of a muscle, and can be determined in various ways, e.g. as the mean distance between aponeuroses across the width of the image, or taken only from the centre of the image.

The musculoskeletal ultrasound method has many application areas. For example, numerous studies have examined muscle architectural parameters before and after some kind of intervention, such as resistance training (Seynnes, De Boer, & Narici, 2007) or as a result of chronic activity over many months or years (Sipila & Suominen, 1991). The method has also been used to examine the effects of ageing on muscle (and tendon) architecture (Narici, Maganaris, Reeves, & Capodaglio, 2003; Reeves, Narici, & Maganaris, 2004), and studies are

beginning to appear involving other human populations (e.g. Körting et al., 2019).

In addition to static assessments of muscle architecture, many studies have sought to examine the dynamic function of a muscle. As an ultrasound probe can be secured to the skin, it is possible to examine muscle mechanics at reasonably high temporal resolution during tasks such as walking (Fukunaga et al., 2001; Lichtwark & Wilson, 2006), running (Lai et al., 2015; Lichtwark, Bougoulias, & Wilson, 2007) and different types of jumping (Sousa, Ishikawa, Vilas-Boas, & Komi, 2007).

It should be noted that in addition to muscle architecture, ultrasound can also be used to track tendon and aponeurosis behaviour. These applications are not covered in detail in this thesis, since the main focus is on muscle architecture assessment. However, the interested reader is referred to several publications that address the applications of the method in this area (Franz & Thelen, 2016; Seynnes et al., 2015), as well as associated methods of analysing the images (Karamanidis, Travlou, Krauss, & Jaekel, 2016; Zhou et al., 2018).

Given the broad scope of application areas of the ultrasound method, it is clear that there is a need for efficient analysis methods, in order to cope with the ever-increasing volumes of data. In particular, when ultrasound is used to assess dynamic muscle behaviour, the resulting video sequences require each individual frame to be processed, and it is often infeasible to perform this process manually, as has traditionally been done in this field.

2.2 Analysis of musculoskeletal ultrasound images

Until quite recently, the only way to analyse musculoskeletal ultrasound images was to manually place tracking markers in each image. This process is extremely time-consuming, but also introduces a lot of potential bias and inter-individual variability into the results. The reliability of the analysis process has been examined in several studies. As shown in table 1 below, the results of such studies generally yield high levels of reliability for the metrics of interest (see also Kwah, Pinto, Diong, & Herbert, 2013).

TABLE 1 Measurements of reliability/reproducibility of the ultrasound technique.

Reference	Parameter(s)	Coefficient(s)	Comments
Kawakami, Abe, & Fukunaga, 1993	thick, penn	r = 0.978, r = 0.906	Measurement error: <1mm, <1˚
Narici et al., 1996	thick, penn, FL	CV 4.8%, 9.8%, 5.9%	Variation: <2mm, <2º, <4mm
Caresio, Salvi,	thick	ICC 0.99	4 muscles tested, mean difference

Molinari, Meiburger, & Minetto, 2017			-0.05±0.22 mm
Kurokawa, Fukunaga, & Fukashiro, 2001	FL, penn	r = 0.994, r = 0.998	Variation: 0-3%
Kawakami, Ichinose, & Fukunaga, 1998	FL, penn	CV 0-2%	
Ito, Kawakami, Ichinose, Fukashiro, & Fukunaga, 1998	FL, penn	CV 2%, 7%	
Fukunaga, Ichinose, & Ito, 1997	FL, penn	CV 0-6.8%, 0-3.8%	
Kubo et al., 2000	FL, penn	CV 0-3.8%, 0-4.5%	Inter-trial variation: 0.2-6.2%, 0.4-5.4%
Cronin et al., 2009; Cronin et al., 2009, 2008	FL	CV 4-6%	

thick = muscle thickness; penn = pennation angle; FL = fascicle length; CV = coefficient of variation; r = correlation coefficient.

However, one difficulty of such analyses is that they are invariably performed by a single researcher. It seems likely that if the same images were to be manually interpreted by a group of different people (e.g. from different labs), the variability of the results would increase, making the method appear less reliable. In this respect, methods that allow some degree of automation could potentially help to reduce the variability of results between different research groups, making it easier to compare the results of different studies.

Early attempts to automate (or at least semi-automate) the analysis of ultrasound images tended to focus on a single parameter. For example, Magnusson et al. (2003) used a pyramidal implementation of the Lukas-Kanade feature tracking method (Bouguet & Bouguet, 2000) to track the 2D displacement of tendinous tissue during passive joint rotation and isometric voluntary contractions. Loram et al. (Loram, Maganaris, & Lakie, 2004, 2006) introduced a method based on spatial cross-correlation that was designed to track small longitudinal displacements in muscle tissue, e.g. during quiet standing. Both of the above-mentioned methods provide excellent tracking within the context that they were developed. However, in both cases, there is a requirement that the structures being tracked are visible in all images. This is often not the case during dynamic tasks like walking or running.

Rana et al. (2009) used multiscale vessel enhancement in combination with either wavelet analysis or Radon transform to automatically detect the orientation of muscle fascicles. Both methods yielded similar results to manual analysis

(difference less than 0.02°). Thus, this method has potential to automate muscle architectural measurements, but would need to be combined with additional processing in order to do so.

Cronin et al. (2011) developed a Lucas-Kanade optical flow algorithm with an affine optic flow extension, and used the method to estimate muscle fascicle length (see also Farris & Lichtwark, 2016). One important advantage of this method is that the specific structures being tracked do not necessarily need to be within the image field of view, since the affine optic flow tracking determines flow within a region of interest that is always visible. Thus, this method is well suited to tracking images where large displacements of the muscle occur, as is often the case in human movement. With this approach it is possible to process videos efficiently, but importantly, this method does require the user to manually label the trajectory of a single muscle fascicle in the first image of the sequence (see also Drazan et al., 2019). Thus, this method is not fully automated, nor does it provide all of the possible parameters of interest.

As a general rule, most methods assume that muscle fascicles are linear structures, and that they always visibly extend between the superficial and deep aponeuroses. However, in practice, this is often not the case. Marzilger et al. (2018) used an approach whereby their algorithm identified any visible fascicle-like structures, regardless of whether they fully extended between the two aponeuroses or not. The structures that were identified, referred to as 'snippets', could then be extrapolated once the aponeuroses had been identified. However, this latter step was performed manually, so again, the process of identifying all relevant parameters was not fully automated.

Very few studies have documented full automation of the analysis process. Caresio et al. (2017) reported a promising method that involved a multi-stage filtering approach to identify the aponeuroses, and used this information to compute muscle thickness in several different muscles. Although they also detected the orientation of muscle fascicles automatically, they did not use this information to determine muscle fascicle length or pennation angle. Conversely, Zhou et al. (2015) detailed an approach involving multi-resolution analysis and line feature extraction (using Gabor wavelets) to determine both fascicle length and pennation angle fully automatically. In theory, either of the above-mentioned approaches could potentially be expanded to provide all parameters of interest.

Recently, Seynnes and Cronin (2019) detailed an open-source approach built in ImageJ/Fiji, which was designed to compute all of the relevant parameters (muscle thickness, fascicle length and pennation angle) without any user involvement. This approach solves many of the limitations in this field, but it was primarily designed for processing single images rather than videos. Although video analysis is possible, the algorithm treats individual images as being inde-

pendent, so processing videos- where there is a strong correlation between consecutive frames- can result in inconsistent results. Jahanandish et al. (2019) also developed an automated approach to compute all relevant parameters. However, this method was developed for a single muscle (rectus femoris) and makes assumptions about the spatial location of tissues that may not always be valid. Moreover, the approach assumes straight muscle fascicles, which is not always the case. Thus, the robustness of this approach to other muscles and test conditions may not be sufficient.

Clearly there are numerous methods in existence that go some way to solving the problem of fully automated analysis of musculoskeletal ultrasound images. However, there are several limitations that serve to hinder development in this field. Firstly, many of the methods outlined above require at least some degree of user input, either to identify a structure at the start of a sequence, or to correct mistakes due to erroneous tracking performance. Secondly, most algorithms are developed to track a single parameter, so for a comprehensive assessment of muscle architecture, it would be necessary to use multiple algorithms. Many methods are developed using images from a single muscle or a single ultrasound device, and both of these factors could limit the robustness of the method to different types of images. Moreover, the methods outlined above generally rely on a rules-based approach, e.g. detecting pixel values that exceed a certain threshold, and may thus not be robust to images recorded with different settings. In the case of fascicle detection, most methods also do not take into the curved nature of the fascicles. Although this is unlikely to yield large errors at low force levels, it is known that fascicle curvature can be quite high in some conditions such as maximal force contractions (see Darby, Li, Costen, Loram, & Hodson-Tole, 2013 for one attempt to solve this problem). Finally, and crucially, the vast majority of published studies have not released any code that would allow other researchers to build upon these approaches. As well as hindering any efforts to replicate previous work, this also likely slows down development in this field.

Thus, one of the aims of this thesis was to develop a method that could solve as many of the issues outlined above as possible, resulting in an open source approach that could be used and extended by other researchers. However, based on the above-mentioned limitations, it may be that a different conceptual framework is required in order to achieve these goals. One potential solution to the ongoing challenge in this area that has not yet been introduced is the use of artificial intelligence-based approaches. This issue will be examined in the following sections.

2.3 Introduction to deep learning

In recent years, the concept of machine learning has gained widespread recognition in broad society, thanks to numerous success stories spanning various fields. Machine learning involves the use of mathematical principles to derive patterns from data. One of the most commonly used tools in this area is the so-called artificial neural network, which was originally inspired by networks of nerve cells in the human brain (Schmidhuber, 2015).

Until quite recently, most neural networks used a feed forward architecture that included a single hidden layer, where each node (or neuron) in the input layer is connected to each node in the hidden layer, and each node in the hidden layer is in turn connected to each node in the output layer (figure 2A). A non-linear activation function (e.g. sigmoid) is then applied to the output of each node. If the network is being used for classification, the number of nodes in the output layer is equal to the number of possible classes. In this case, a softmax function is used to assign a probability to each of the possible classes. If instead the goal is to predict a continuous output, a single linear output neuron is typically used.

In the case of deep neural networks, the same principles are essentially just extended across multiple hidden layers instead of just one (Hinton & Salakhutdinov, 2006), whereby each node of a given layer is connected to each node of both the previous and subsequent layers (figure 2B). In this kind of architecture, hidden layers are often also referred to as 'dense' layers, because each layer is fully connected to the previous one.

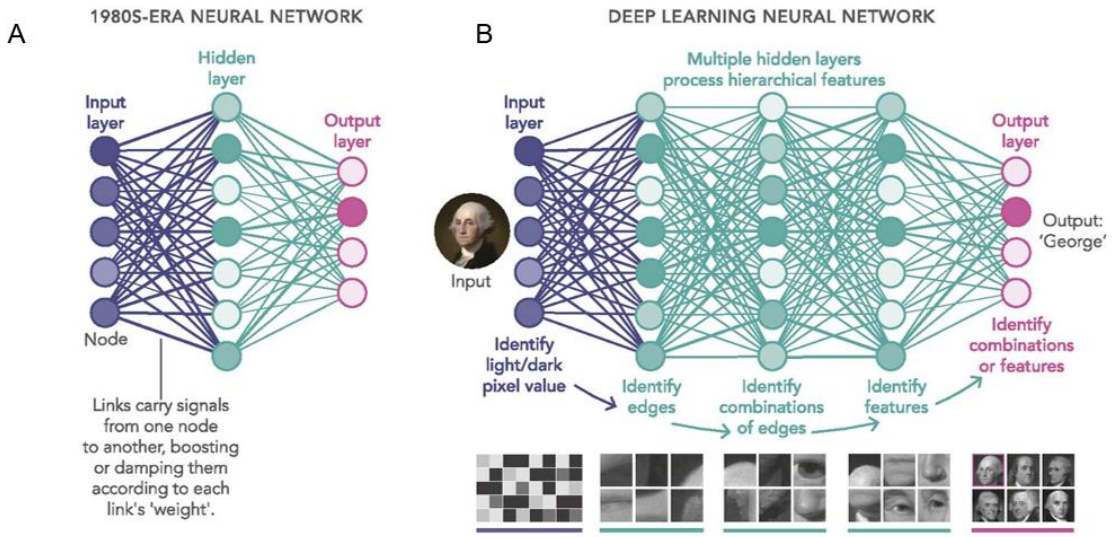

FIGURE 2. A: A simple neural network that contains a single hidden layer. B: A 'deep' neural network containing multiple hidden layers.

During training, data are fed into the network (see the left-hand side of figure 2; modified from Mitchell Waldrop, 2019), and the hidden layers 'learn' common features of the training images. This 'knowledge' is then used to predict the existence of the learned features in new, previously unseen images. Note that in figure 2B, the features become gradually more abstract in later layers, as simpler features learned in previous layers are gradually merged.

The popularity of deep neural networks has increased in recent years, in part because of the increased availability of large datasets, but also because of improvements in hardware capabilities (e.g. GPUs) that have made it feasible to train larger and deeper neural networks within a reasonable timeframe. One particularly popular approach involves convolutional neural networks (CNNs), which were first used in 2012 to outperform previous approaches on image recognition tasks (Krizhevsky, Sutskever, & Hinton, 2012) (figure 3). CNNs offer several potential advantages for image processing, namely the use of local receptive fields, weights sharing, and subsampling. These mechanisms combine to enable CNNs to learn spatially invariant features, meaning that the learned features can be detected in new images regardless of where they appear within the image. Moreover, the CNN approach requires fewer parameters to be learned, thus reducing the computational requirements.

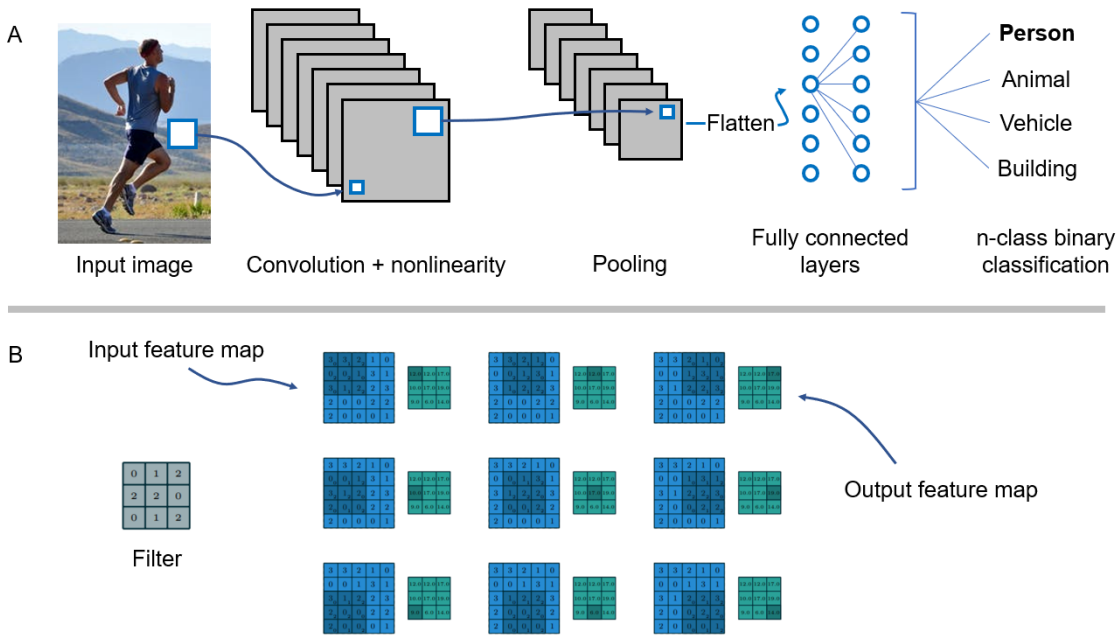

FIGURE 3. A convolutional neural network.

The general concept of a CNN is shown in part A of figure 3. The number of convolution/pooling layers and the number of filters are modifiable. The mathematical details of convolution are shown in part B of the same figure (modified

from Dumoulin & Visin, 2016). Each filter is applied to a patch of the input image (matrix multiplication) and the sum of the result is added to the output feature map. The filter is then moved to the next image patch (in this case with a stride of 1). In this example, the filter is applied to 9 different image patches, resulting in a single output map with shape 3 x 3.

In CNNs, information still flows in a feed forward manner, but the nodes of the network are not fully connected. Instead, filters with different weights are moved across the input (usually an image), and the resulting feature maps are fed to the following layer (figure 3). The filters are thereby able to identify features in the input images regardless of precisely where those features appear spatially. Pooling such as max or mean pooling (averaging of nearby positions in the feature maps) enables the network to become invariant to small local deformations in the input. A CNN architecture will generally include several filters and may also have multiple convolution and pooling layers, allowing more abstract image features to be identified (Lecun, Bengio, & Hinton, 2015).

Mathematically, the output of any given node or layer in a standard neural network can be summarised as follows:

$$\text{Output} = \sigma(xW + b)$$

where σ is a nonlinear activation function applied to the result of the calculation in brackets, x is the input (or for a hidden layer, the output from the previous layer), W represents the weight(s), and b is a bias value that is often set to a constant. In the case of a CNN, the same logic applies, but the main difference is that a convolutional operator is added between the input/previous layer and the weight instead of a matrix multiplication, i.e.

$$\text{Output} = \sigma(x*W + b)$$

where * denotes a convolution operation.

Methods built on CNN architectures have consistently been shown to perform well at a wide range of feature detection tasks, including cases where the input images are not well standardised (Girshick, Donahue, Darrell, & Malik, 2014; Szegedy et al., 2015). The success of CNN-based approaches lies partly in the depth of the network architecture. This allows features to be identified at multiple levels of abstraction, which is important in complex images that don't always contain clearly identifiable structures. Accordingly, this kind of approach may be well suited to the segmentation and analysis of medical images.

2.4 Deep learning for medical image processing

In recent years, CNN-based methods have been applied extensively in the medical domain (Litjens et al., 2017; Shen, Wu, & Suk, 2017). The success of these applications is often staggering, with several algorithms demonstrating human or even superhuman performance (Topol, 2019). The breadth of application areas is beyond the scope of this thesis. Instead, the following sections outline broad methodological approaches and associated issues, and then detail applications of deep learning relevant to the specific topic of this thesis.

2.4.1 Deep CNN architectures

One of the earliest examples of a CNN architecture gaining traction in the broader scientific community was that of AlexNet (Krizhevsky et al., 2012), which used five convolutional layers, kernels with large receptive fields in layers close to the input and smaller kernels closer to the output, as well as rectified linear (ReLu) activation function. This paper began a hugely productive era that is still ongoing, whereby new architectures were developed, and nowadays the newest models almost always use far deeper architectures than AlexNet. Newer models also tend to use smaller kernels stacked together, rather than a single layer with a large receptive field. This offers the advantage of fewer learned parameters, and thus lower memory requirements during inference.

Simonyan and Zisserman (2015) were among the first to use an architecture with many hidden layers and smaller, fixed-size kernels. This model, referred to as VGG19, won the 2014 ImageNet challenge. Subsequent models such as GoogLeNet (Szegedy et al., 2015) also used a deep architecture but added novel layer structures designed to improve efficiency, namely the so-called Inception module, which uses multiple sets of convolutions of different sizes to again reduce the number of learned parameters. Similarly, He et al. (2015) won the ImageNet challenge in 2015 using an architecture that contained so-called ResNet-blocks, which learn residuals rather than entire functions, thereby tending to learn mappings that are close to the identity function values.

Importantly, in many medical applications, the speed of inference is often secondary to performance, since most analysis has traditionally been done manually, so any accurate computational approach would likely offer time savings. As a result, there is less of a need to always adopt the latest and fastest approach, and somewhat older models such as VGG are still widely used in this field.

In addition to standard CNN architectures, it is sometimes advantageous to employ a so-called multistream or dual pathway approach (Kamnitsas et al., 2016), e.g. for the purpose of multi-scale image analysis. In many medical appli-

cations, contextual information may be necessary to enable accurate image segmentation. Context can be provided by feeding larger image patches through the network, although this generally also increases computational requirements. Alternatively, context can be added by combining a down-scaled representation with high-resolution local information, as has been achieved in several medical imaging studies (e.g. Kamnitsas et al., 2016). For a review of performance metrics of different deep learning architectures, see Canziani, Paszke, & Culurciello, 2016).

2.4.2 Image segmentation

The task of interest in this thesis is essentially a segmentation problem. In the medical domain, many tasks are based on a binary classification of each image, e.g. diseased versus healthy. In such cases, the entire image may be scanned, but a single decision (prediction) is made based on the image as a whole. However, in the application being addressed in this thesis, the task is a labelling of features, which is more suited to an approach like semantic segmentation, where the network's task is essentially to apply a binary label to each pixel of an image. In doing so, each pixel is labelled as either belonging to or not belonging to a specific structure.

CNNs can be used for pixel-wise labelling of an image, for example using a sliding-window approach whereby the network focuses on a small patch of the image at one time, with each patch centred around the target pixel. In 2015, the first attempts to use fully convolutional neural networks for semantic segmentation began to appear. Long et al. (2015) successfully replaced fully connected layers with convolutional layers, and included an upsampling layer to allow dense inference and pixel-wise labelling. Shelhamer et al. (2017) also used a 'fully convolutional' approach that takes in images of any size and returns a same-sized output that includes pixel-wise labels. Their model included a skip architecture that combines semantic information from a deep, coarse layer with appearance information from a shallow, fine layer. This approach yielded superior results to the previous state-of-the-art whilst also improving inference time.

Many recent advances have built upon the concept of fully convolutional semantic segmentation (e.g. Chen, Papandreou, Kokkinos, Murphy, & Yuille, 2018; Wu, Zhang, Huang, Liang, & Yu, 2019; for review see Garcia-Garcia, Orts-Escolano, Oprea, Villena-Martinez, & Garcia-Rodriguez, 2017). Of particular interest to this thesis is the work of Ronneberger et al. (2015), who proposed the U-net architecture, comprising a fully convolutional neural network (termed the contractive pathway) and a subsequent upsampling (or expansive) pathway where upsampling convolutions were used to increase the image size. This network also included so-called skip-connections to directly connect contracting and expanding convolutional layers. The U-net method was particularly tailored for and tested with medical images, providing the major advantage of

accurate segmentation using only small datasets (see section 'U-net' below). More recently, Milletari et al (2016) proposed a variant of the U-net architecture for use with 3D images, called V-net, although the current thesis will only consider 2D image segmentation, since this is currently far more common in ultrasound imaging.

2.4.3 CNN limitations: Overfitting

In spite of the obvious power of CNN approaches, there are several limitations associated with the basic CNN approach (Greenspan, Van Ginneken, & Summers, 2016) within this field. Firstly, as is often the case in deep learning, this type of network tends to require a large labelled dataset. This can be a major obstacle in medical domains, where the labelling process may be very expensive, there may be a lack of data due to the rarity of a condition, and/or there may be ethical restrictions on the availability of data. Moreover, in the case of ultrasound imaging, there may be large variability within the dataset because the scanning process is not well standardised. This tends to require an even larger labelled dataset.

One common side effect of using smaller datasets is the issue of overfitting, which essentially means that the trained model performs well on the training dataset, but often fails to make accurate predictions when it is used to make inferences on new, previously unseen images. To overcome this issue, various tools have been developed recently, including transfer learning and fine-tuning. One benefit of transfer learning in particular is that it is often possible to get good results using a much smaller dataset than with a standard CNN approach. As already stated, this is useful in domains where large labelled datasets are not common. Fine-tuning allows supervised correction of image labelling, whereby the new labels are fed back into the network to improve the quality of the final trained network. These two approaches, transfer learning and fine-tuning, have been shown to be important when dealing with relatively small medical image datasets (e.g. Tajbakhsh et al., 2016).

Various other approaches are now commonly used in medical deep learning applications, with the aim of helping to reduce overfitting. For example, appropriate initialisation and use of momentum can serve to improve the efficiency of the training process (Sutskever, Martens, Dahl, & Hinton, 2013). Dropout is another common technique, whereby a portion of the nodes are effectively removed from a neural network, in an attempt to help improve the robustness of the learned features by decreasing the reliance of the network on individual nodes (Wan, Zeiler, Zhang, Lecun, & Fergus, 2013). Batch normalisation (Ioffe & Szegedy, 2015) is also used to reduce the amount of variation in the weights of hidden layers, which can help to make training more robust and reduce overfitting.

2.4.4 Deep learning with ultrasound images

A surprising variety of model types and architectures have already been applied to the broader field of ultrasound imaging. This includes recurrent neural networks that take into account the influence of time (e.g. video sequences), unsupervised approaches such as autoencoders and deep belief networks, and CNN-based approaches (for review see Liu et al., 2019). These approaches are generally used to perform classification, detection or segmentation within an image, and they have been applied to a wide range of tissues (e.g. breast, liver, heart, fetus). In this section I outline work done specifically in segmentation.

Ultrasound imaging is a modality that would strongly benefit from automated analyses, because there is commonly variability between different analysers, and because the contrast between structures within an image (as well as overall image quality) is not necessarily very high. As a result, the manual labelling process is often slow and costly, making it difficult to access large labelled datasets. Another difficulty is the issue of boundary incompleteness, whereby a structure is only partially visible, making automated analyses more difficult. Two solutions to this problem have been offered: a bottom-up, supervised approach of classifying each pixel (binary); and a top-down approach that uses prior shape information to guide the segmentation. In fact, pixel-wise segmentation has been achieved for various tissues in the human body, including lymph nodes (Zhang, Ying, Yang, Ahuja, & Chen, 2017) and bone (Baka, Leenstra, & Van Walsum, 2017), and such approaches have been found to be superior compared to previous state-of-the-art methods, whilst also offering improvements in computational speed.

Ravishankar et al. (2017) combined the bottom-up and top-down methods by using a previously learned shape to refine the predicted segmentation result obtained from a fully convolution segmentation network. The results on ultrasound images of kidneys demonstrated that the addition of the prior shape information improved segmentation accuracy by around 5%. Wu et al. (2017) used an auto-context scheme combined with a fully convolutional architecture to take advantage of local contextual information. The helped to deal with boundary incompleteness and improved segmentation accuracy. Ma et al. (2017) used a different approach, whereby an image was divided into patches and each patch was classified, with the potential advantage of decreased computational and memory requirements.

Although the focus of this thesis is the analysis of 2D images, some attempts to analyse 3D data are relevant, since 3D sequences represent a stack of 2D images, and thus some methods may be applicable to both data types. Ghesu et al. (2016) used a typical non-rigid segmentation method (rigid object localisation and

non-rigid object boundary estimation) to detect and segment the aortic valve in 3D US volumes (essentially 2D images stacked together). They combined deep learning with marginal space learning. After detecting an object, the non-rigid shape was estimated, followed by a sparse adaptive deep neural network-based active shape model to detect the shape deformation. The results on almost 3000 3D transesophageal echocardiogram images demonstrated the efficiency and robustness of the approach for 3D detection and segmentation of the aortic valve, with a significant improvement of up to 42.5% over the state of the art.

Milletari et al. (2016) used a patch-wise CNN method (Hough-CNN) to segment transcranial ultrasound volumes. The approach combined the CNN predictions with voting by exploiting the features learned by the later layers of the network. They also examined the importance of training data volume and data dimensionality. The proposed method was shown to be superior to voxel-wise semantic segmentation of 3D images. This method could thus offer potential for the analysis of 2D images.

In the context of this thesis, the most relevant study performed to date is that of Cunningham et al. (2018). In this work, the authors used a deep CNN approach combined with deconvolutional layers to predict muscle fibre pennation angle in lower limb muscles. The authors compared their results to the ground truth, i.e. hand-labelled images. The results indicated an improvement in the predictive accuracy of pennation angle compared to a previous approach based on wavelet analysis. However, the mean difference between the results of their method and that of the ground truth was quite large, ranging between about -6 to almost 11°. Whilst this attempt represents a big step forward in the analysis of musculoskeletal ultrasound images, this error amplitude is too large for the approach to be useful for longitudinal studies. Moreover, this method was only designed to compute pennation angle values, and not all possible architectural parameters (see section 2.1 above). Thus, there is still scope for a deep learning approach that provides all of the necessary information with sufficient accuracy.

3 METHODS

3.1.1 Experimental approach

For the present thesis, I chose to implement the U-net method (Ronneberger et al., 2015). Although there are many possible approaches that could be used, U-net was specifically developed for use with biomedical images, and is well suited to the current application, where large annotated datasets are not currently available. U-net makes extensive use of data augmentation by applying elastic deformations to the training images, helping to reduce the size of the training dataset.

As shown in figure 4, the network architecture is made up of a contracting path on the left and an expansive path on the right. The contracting path consists of blocks of two 3x3 convolutions, each of which is followed by a rectified linear unit (ReLU) and 2x2 max pooling to downsample the feature maps. Each downsampling stage doubles the number of feature channels. The expansive path serves to upsample the feature channels, and in correspondence with the contracting path, each upsampling layer is followed by 2x2 up-convolution that halves the number of feature channels. Each upsampling layer is then concatenated with the corresponding feature map from the contracting path (grey arrows in figure 4), and followed by two 3x3 convolutions and a ReLU. The final layer is a 1x1 convolution that is used to map the component feature vectors to the required number of classes, in this case two per model (see details below). Training is performed in a supervised manner by inputting original images and corresponding hand-crafted segmentation maps.

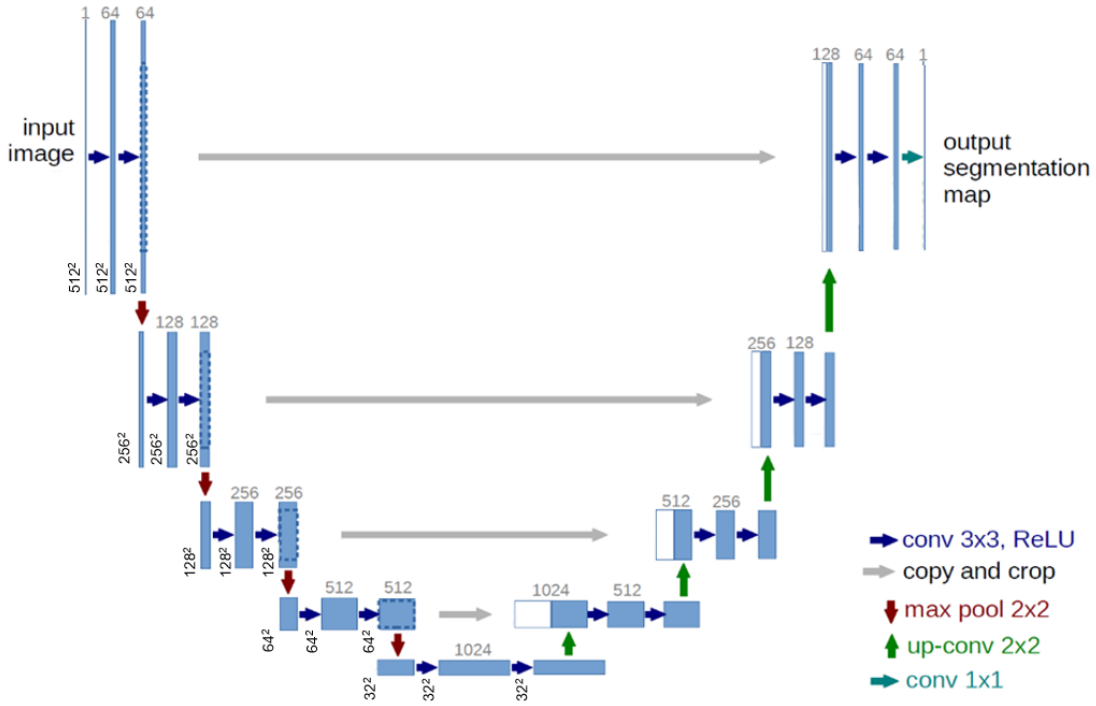

FIGURE 4. The modified U-net architecture used in this thesis.

In figure 4, which is modified from Ronneberger et al. (2015), convolutional layers are denoted by 'conv'. Blue boxes represent multi-channel feature maps; the number of channels is denoted above each box. The x-y sizes are shown at the lower left edge of each box on the contracting side, and these are identical for the expanding side. White boxes represent copied feature maps from the contracting side, which are concatenated with those from the expanding side.

The output of the U-net model is a pixelwise binary label, i.e. every pixel of an image is predicted to belong to one of two possible classes. In my approach these classes were aponeurosis/not aponeurosis, and fascicle/not a fascicle, since I trained two separate models designed to detect different tissues (see below).

3.1.2 Data

I compiled a large volume of anonymised single image and video data from different leg muscles (medial and lateral gastrocnemius, vastus lateralis, tibialis anterior) and with 4 different ultrasound devices, as well as from different human populations (athletes, older people, young healthy individuals) and different movements/muscle contraction types. Individual frames were extracted from this dataset at random using a custom-written function in Python (Python Software Foundation, v3.6), resulting in a set of around 570 images for aponeurosis training and 310 for fascicle training. All data were acquired in previous

studies, all of which received ethical approval from the relevant committees. I manually annotated all images, generating binary masks that denoted the two aponeuroses (for the aponeurosis model), or all instances of muscle fascicles (for the fascicle model).

3.1.3 Neural network training

I trained two separate models, each using the same U-net architecture. Images were imported and resized to 512*512 pixels for training. In general, neural networks perform faster with smaller images, but in this case I chose the largest possible image size given RAM limitations, since the quality of ultrasound images is typically quite low, and further reductions in spatial resolution due to image downsampling would likely compromise the ability to successfully train a neural network for pixelwise labelling. I used a 90/10% training/validation data split. Training was performed using an RTX2070 GPU, and took less than one hour per model using 50 epochs and a batch size of 1, with Adam optimizer and the binary cross-entropy loss function. Training was stopped early if overfitting was evident, as characterised by a decrease in the training error and a concomitant stagnation or increase in the test error. The code runs in Python and uses Keras frontend with a Tensorflow backend.

All code and data from this project are available from Github (https://github.com/njcronin/DL_Track).

For each of the aponeurosis training images, I manually identified all instances of aponeuroses using the polygon tool in Fiji software (Schindelin et al., 2012) to create a binary mask, whereby individual pixels belonging to an aponeurosis were white, and all other pixels were black. This process was repeated separately for the fascicle training set, where all instances of muscle fascicles (or parts of fascicles) were identified (1-20 per image). The binary masks were used as ground truth labels to train two deep neural networks; one to identify aponeuroses and one to identify muscle fascicles (figure 5).

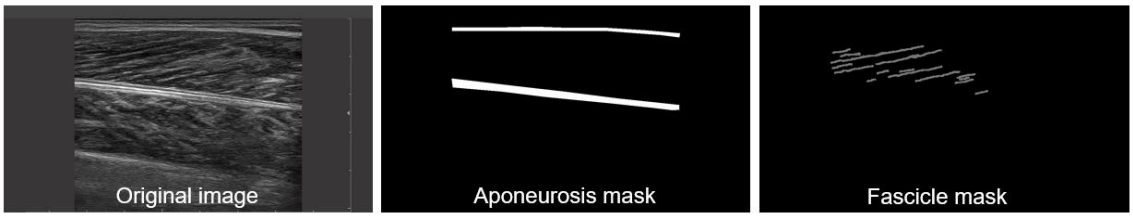

FIGURE 5. Examples of the input to the two U-net models.

3.1.4 Post-processing

When processing a new image or video, the trained neural networks identify aponeuroses and muscle fascicles. Aponeuroses below a user-defined threshold length are removed, and where necessary, those that satisfy the length constraint are extrapolated laterally, since this can assist in finding the intersection with muscle fascicles. The trained network very rarely identifies muscle fascicles that extend from superficial to deep aponeuroses (this is to be expected given the complex anatomy of muscles and the 2D nature of imaging). Instead it identifies portions of fascicles, referred to as 'snippets' by Marzilger et al. (2018). Those snippets that are beyond a threshold length are extrapolated proximally and distally using a 1st (straight) or 3rd (curved) order polynomial fitted to the identified structure. The intersection points between aponeuroses and fascicles are identified, and fascicle length is determined. Pennation angle is computed as the difference between the local slope of the lower aponeurosis and each fascicle.

As multiple 'fascicles' are usually detected per image, the median fascicle length and pennation angle are computed for each image, although other metrics such as mean or maximum can easily be used instead or in addition. In addition, data are shown where all detected fascicles are included, rather than computing an average metric. Muscle thickness is determined from the central portion of the image, as the perpendicular distance from the superficial to deep aponeurosis. These metrics are saved for every frame, converted to a Python dataframe and exported to excel. By choosing a higher order polynomial fit to each fascicle, it is possible to take fascicle curvature into account, although this will naturally sometimes result in an inaccurate fit to the 'actual' fascicle that can be seen visually. It may be possible to address this issue using more sophisticated analyses (e.g. Darby et al., 2013), but this was not implemented here.

3.1.5 Analysis metrics

To determine the overlap between manually created labels from the training sets and the labels predicted by the neural network, I used a custom implementation of intersection over union (IoU). A set of 35 test images unseen during training represented a test set, all of which were processed using the trained networks, to estimate muscle thickness, and median muscle fascicle length and pennation angle. The same test set and parameters were also analysed manually by 2 individuals using Fiji software. Comparisons between the neural network and human results for this test set were done using Bland-Altman plots (Bland & Altman, 1986; https://www.mathworks.com/matlabcentral/fileexchange/45049-bland-altman-and-correlation-plot). For a set of videos, I also compared the results of my method with those of Ultratrack (Cronin et al., 2011; Farris & Lichtwark,

2016), which is arguably the most commonly used (and open source) semi-automated approach for analysing ultrasound videos. As Ultratrack does not take fascicle curvature into account, I analysed these videos with my approach using both straight and curved fascicle models.

4 RESULTS

The results section is broken down into several sub-sections. First, some metrics related to the trained neural networks are presented. Then, some example images and analysed results are shown, followed by the results of the comparison between algorithm- and human-labelled images. The performance of the trained algorithms is then examined for video data, and the results are compared to an existing semi-automated algorithm. Finally, some failure cases are demonstrated.

4.1 Neural network training

Figure 6 shows the loss function and IoU values as a function of epoch number during training. For both models, the optimum was reached after around 20 epochs. Training time for each model was less than one hour.

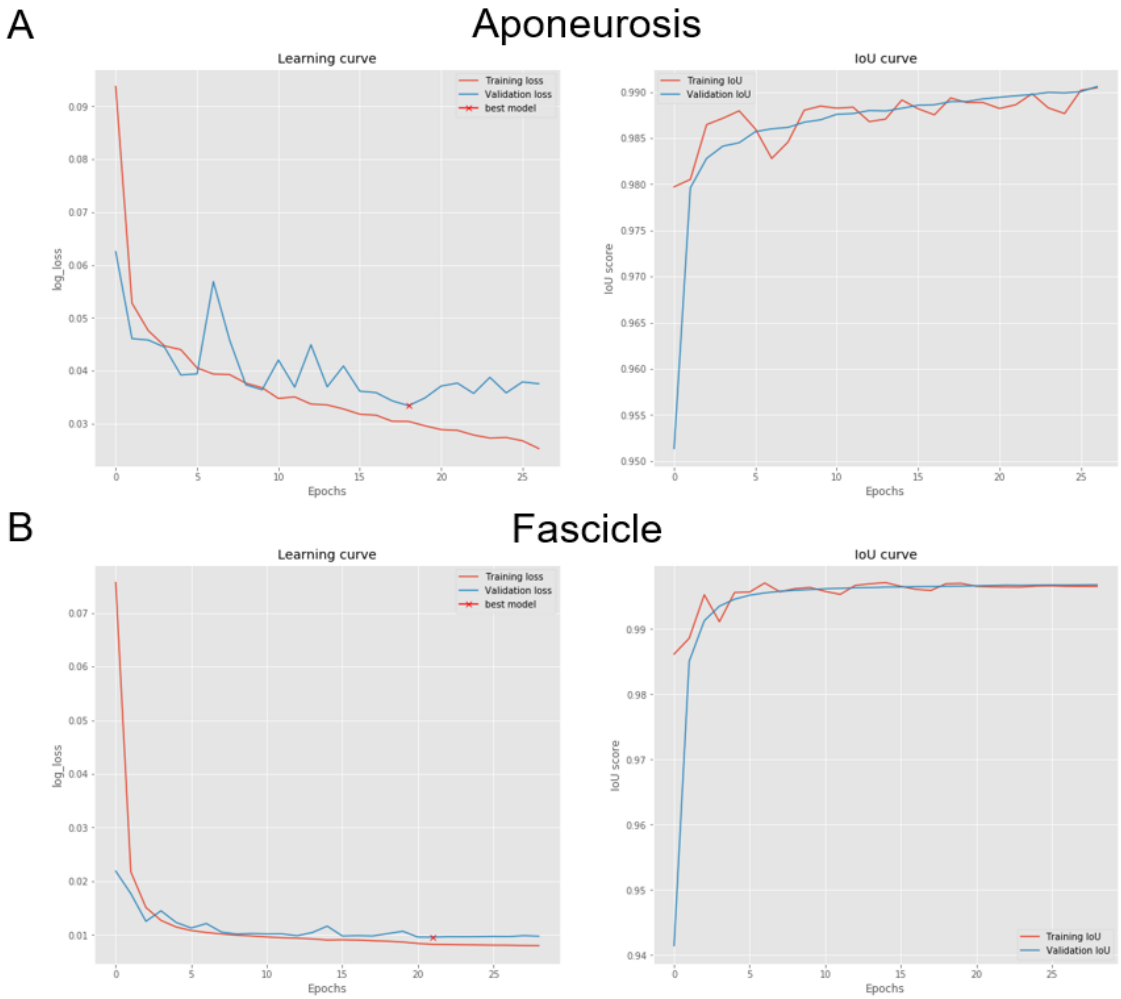

FIGURE 6. Loss function/IoU curves for the two trained models.

Using the trained models, inference time for a single image with a CPU was around 4.6s, compared to around 0.7s with GPU.

4.2 Analysis of single images

In the post-processing phase, I combined the predictions of the two models so that all of the relevant tissues could be detected in a single image. An example of the initial predictions of the neural network models is shown in figure 7.

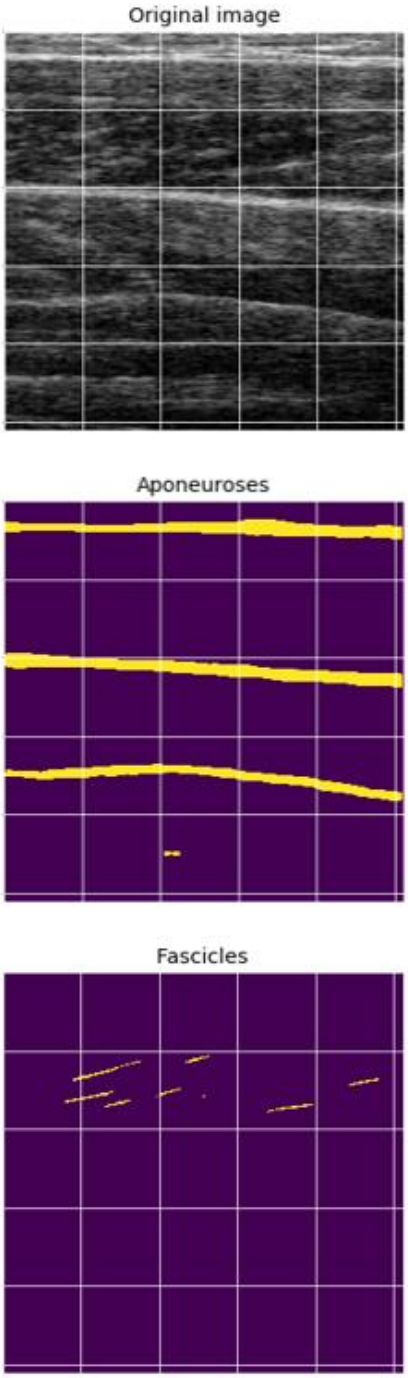

FIGURE 7. Example of trained neural network predictions for a single image.

In this example, the aponeurosis detection is essentially perfect, and a third aponeurosis is even detected below the two target aponeuroses (this corresponds to the border of a deeper muscle). For the example in figure 7, the IoU value for the algorithm-generated aponeuroses versus human manual analysis was 0.82. Only a few candidate muscle fascicles are detected, but several of these could be extrapolated in order to estimate the muscle fascicle parameters. This image is also of particularly poor quality, so the fascicle results are quite encouraging.

Figure 8 shows examples of single images successfully tracked with the trained models, after post-processing. These images were obtained from three different muscles: medial gastrocnemius (A), tibialis anterior (B) and vastus lateralis (C). None of these images were seen by the algorithm during the training process. It is noteworthy that in two of these examples, some of the fascicle endpoints extend beyond the field of view of the image, and their locations must be extrapolated. The analysis results are displayed on each image, and saved simultaneously to an excel file. Fascicle length and thickness results are displayed in pixels, but there is an option to scale the results.

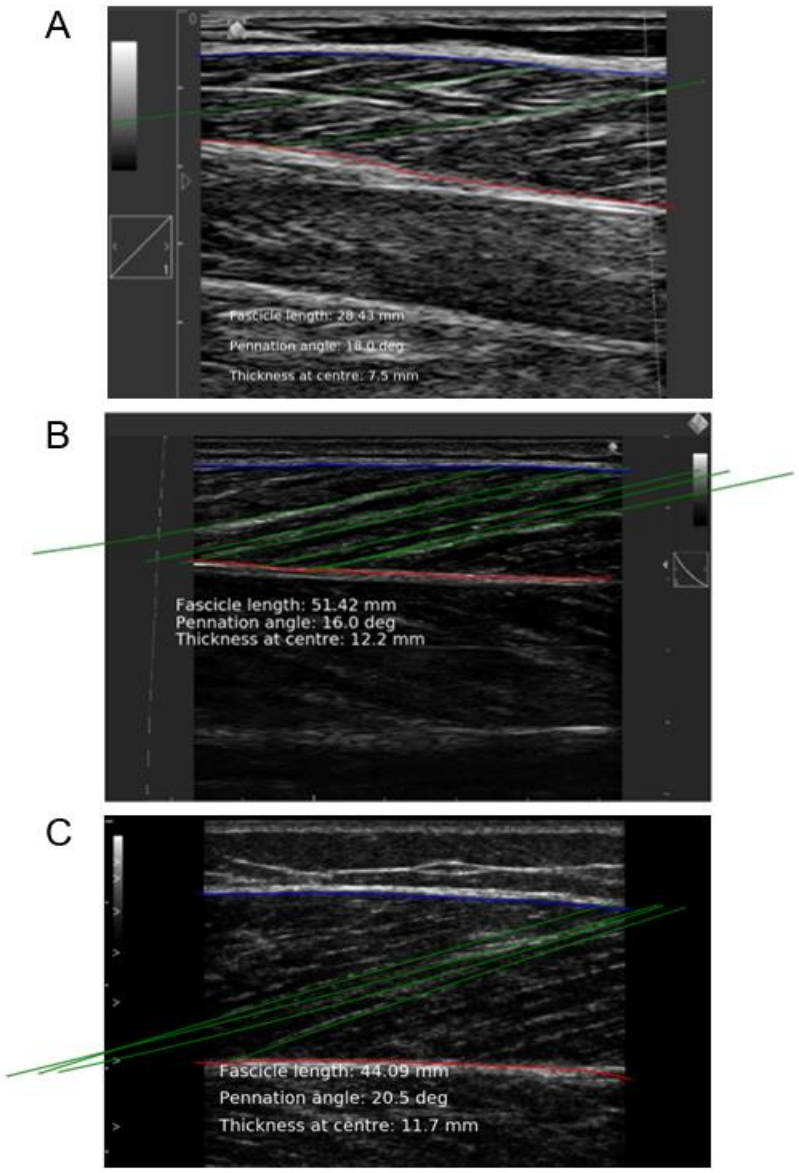

FIGURE 8. Examples of images annotated by the trained neural networks.

4.3 Comparison between neural network and human tracking

Figure 9 shows correlation and Bland-Altman plots of fascicle length data, demonstrating the similarity in predictive performance of two researchers and the newly developed deep learning approach (based on a set of 35 test images). On the correlation plots (left column), the following metrics are displayed as text: slope and intercept equation, Pearson r-squared value, sum of squared errors (SSE), and the number of data points.

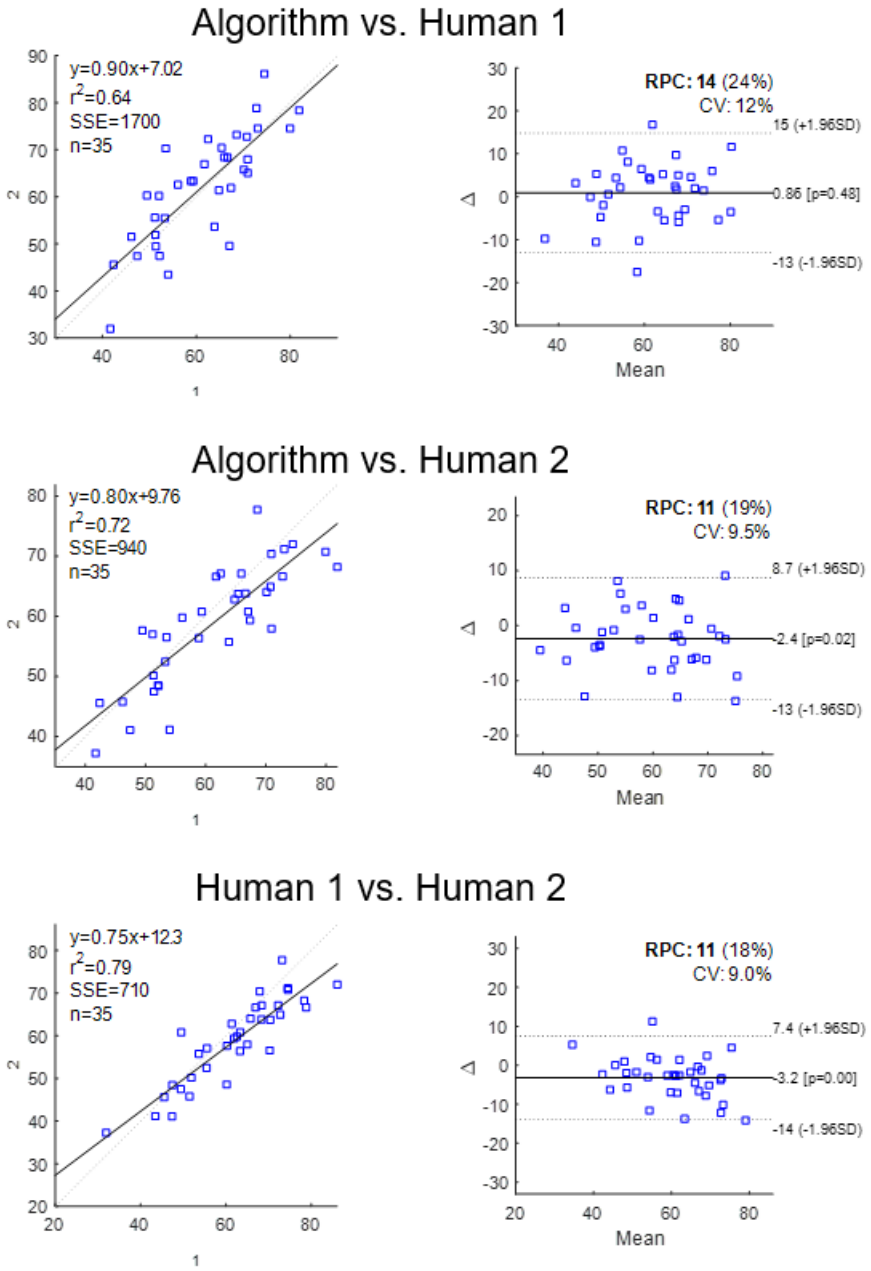

FIGURE 9. Correlation and Bland-Altman plots of the fascicle length data comparing human and algorithm labelling (values are in mm).

In the Bland-Altman plots shown in Figure 9 (right column), the reproducibility coefficients (RPC, 1.96 * standard deviation) and percentages are displayed, along with the coefficient of variation (CV).

Figure 10 shows the same parameters for pennation angle, and figure 11 for muscle thickness. It can be seen that the RPC values are generally highest for fascicle length estimates and lowest for muscle thickness. CV values were highest for pennation angle, and again lowest for muscle thickness.

For fascicle length, there is no evidence of a systematic basis in fascicle length estimates, and for all correlation plots in figure 9, the r^2 values are still quite high (0.64-0.79). In terms of pennation angle, the variation in estimates between human and algorithm values are sometimes up to about 5°. However, again, there is no evidence of systematic bias, although interestingly the two human observers showed greater similarity to each other's estimates than either compared to the algorithm results.

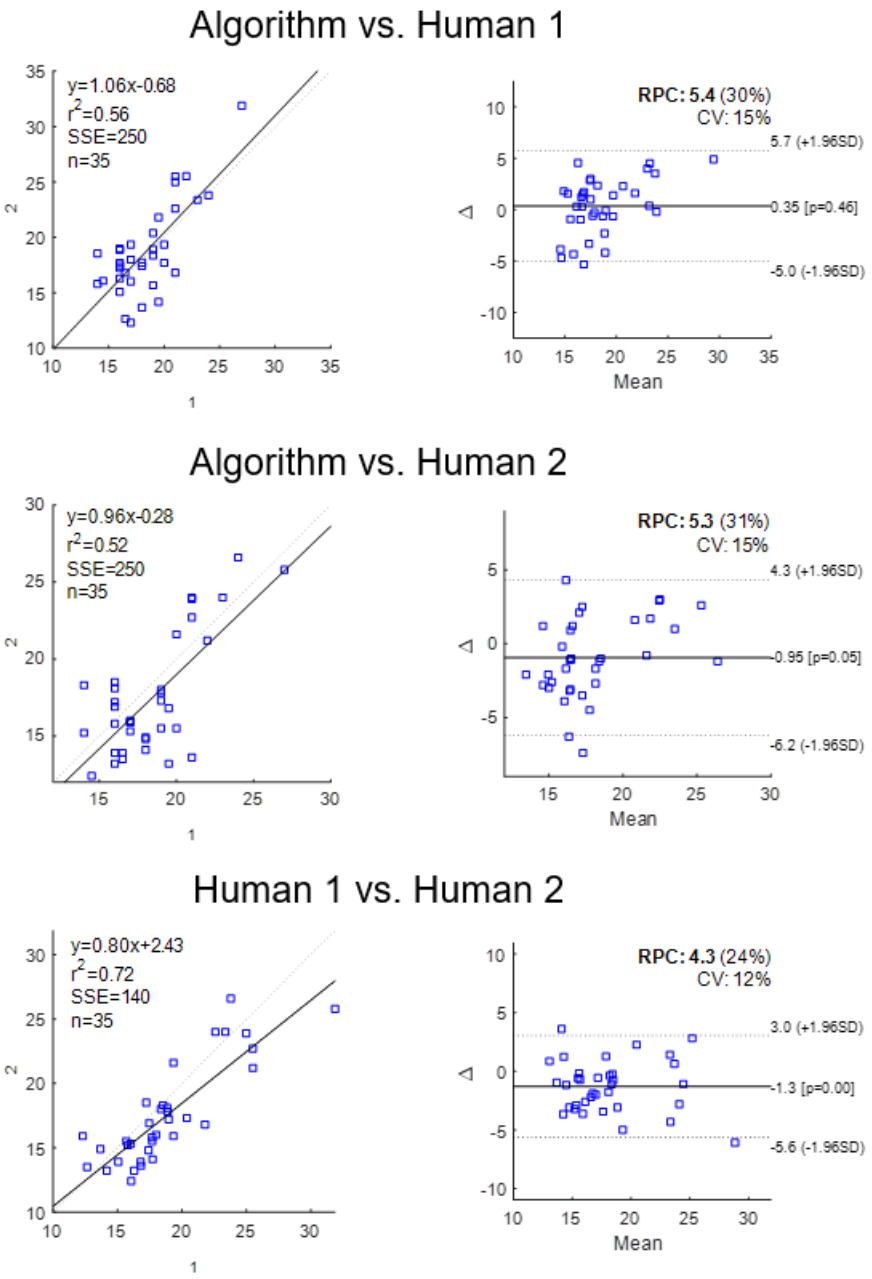

FIGURE 10. Correlation and Bland-Altman plots of the pennation angle data comparing human and algorithm labelling (values are in °).

The parameter yielding the most consistent results between humans and the algorithm was muscle thickness. In fact, differences were always less than or equal to 1.2 mm (CV: 1.7-2.9%), and the slope of the correlation plots was always close to 1 (0.96-1), with r^2 values of at least 0.98.

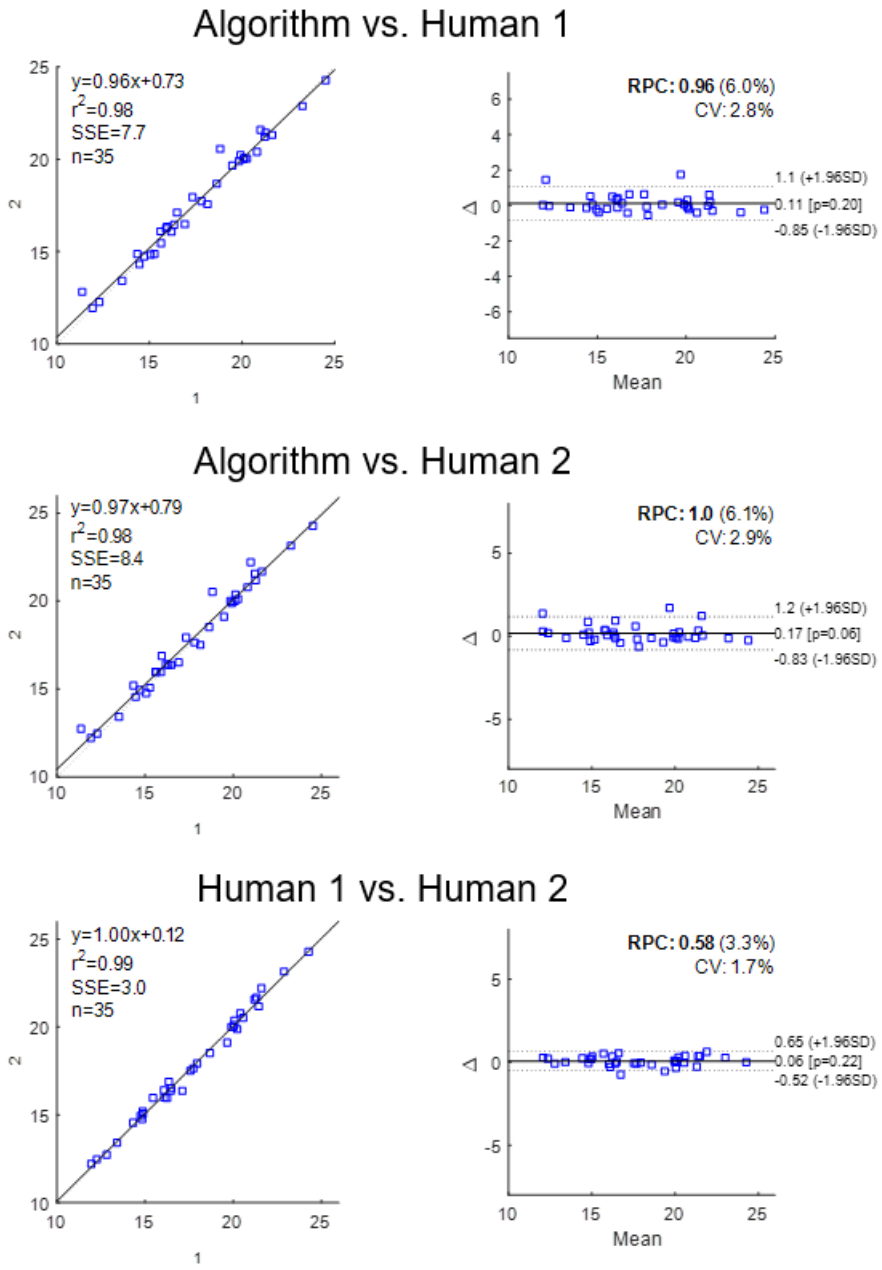

FIGURE 11. Correlation and Bland-Altman plots of the muscle thickness data comparing human and algorithm labelling (values are in mm).

One additional (experimental) feature of my approach is the ability to take muscle fascicle curvature into account, as this is often not done with other methods. In figure 12, the same images are analysed manually (using Fiji), as well as using the deep learning approach with both a straight and a curved fascicle model, and the resulting fascicle lengths are displayed on each panel. All of these images were selected on the basis that fascicle curvature was visually evident in some portions of the image.

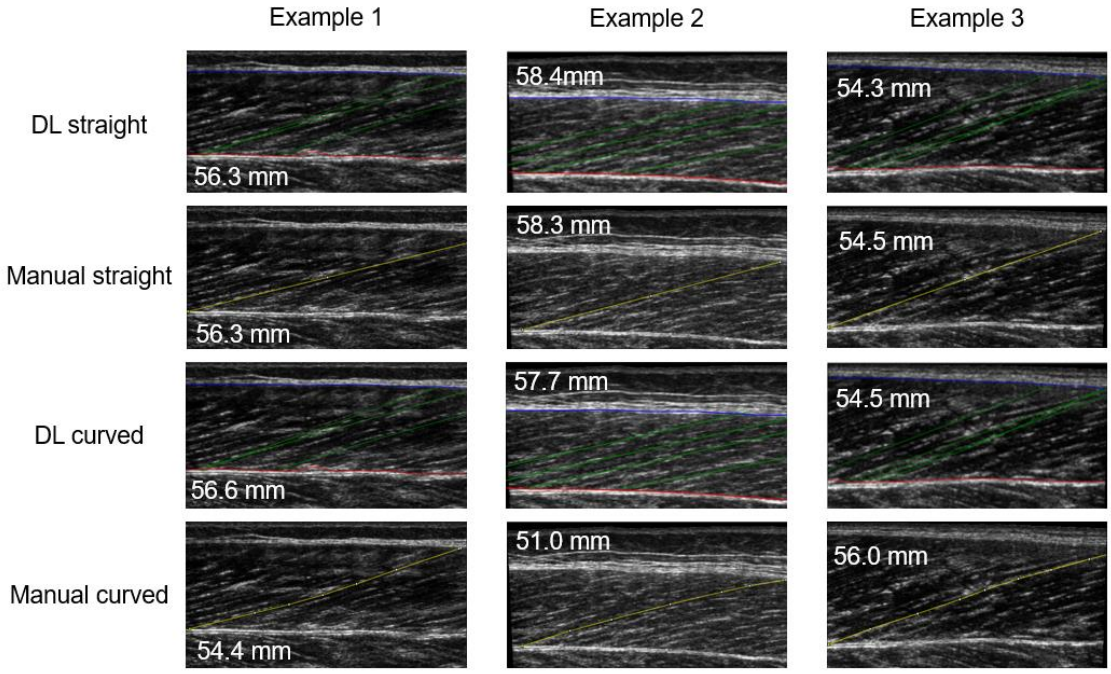

FIGURE 12. Effects of straight versus curved fascicle model on fascicle length.

For the three examples in figure 12, using a curved model with the algorithm (DL curved) resulted in very similar values to those produced using a straight-line model (DL straight), with relative differences between -1.2 and 0.5%. The manual estimates with a straight model were also very similar to those generated by the algorithm (relative differences: -0.2 to 0.4%). However, comparisons of the two manual models resulted in much larger differences, between -14.3 to 2.7%. The direction of differences was also not consistent, i.e. using a curved model resulted in a longer or a shorter fascicle length compared to a straight-line model, depending on the image being analysed.

4.4 Comparison of the new approach with Ultratrack

Ultratrack (Cronin et al., 2011; Farris & Lichtwark, 2016) is arguably the current gold standard in (semi-)automated analyses of muscle fascicle length in video sequences. Here I compared the performance of my algorithm with Ultratrack on several different videos involving different types of muscle contraction.

Figure 13 shows time-series traces of fascicle length tracked using the two methods during human walking at preferred speed (note that Ultratrack does not reliably track pennation angle or muscle thickness, so these parameters were not compared). All of the data in this section are from the medial gastrocnemius muscle.

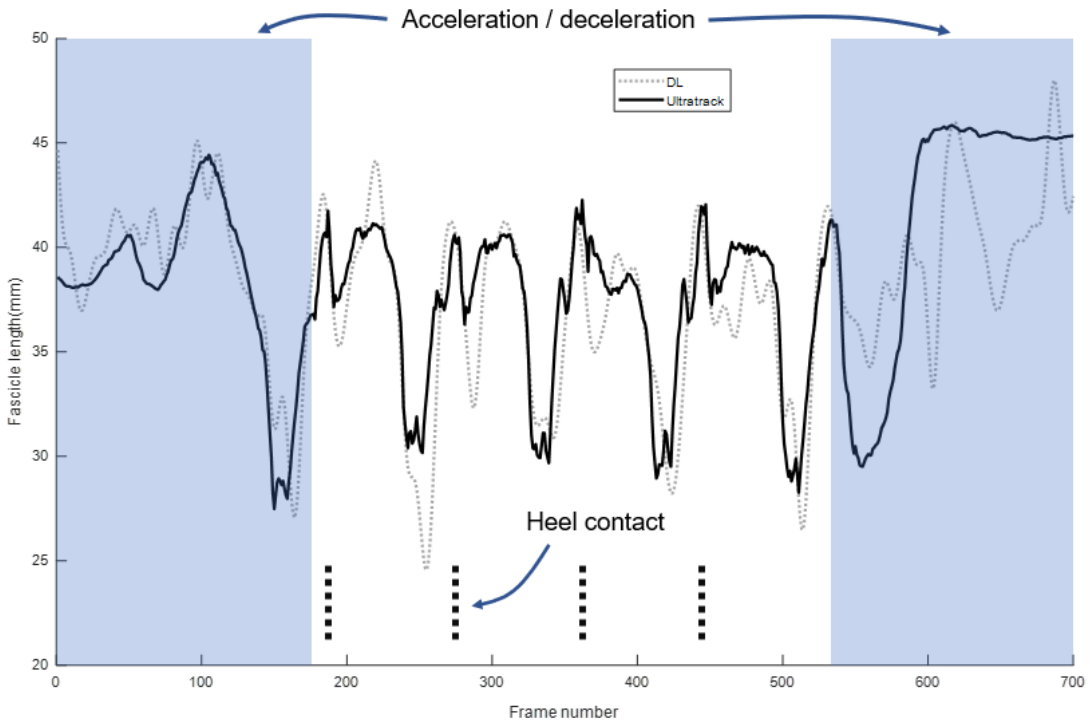

FIGURE 13. Fascicle length traces from walking obtained via the deep learning approach (DL) and using Ultratrack.

In Figure 13, the overlap of the two traces is generally quite good, especially considering that Ultratrack is designed to track a single fascicle, whereas my approach tracks all fascicles visible in the image, and in this case outputs the mean length. In this figure, four full walking strides are shown (approximate heel contact of each stride is denoted by a vertical dotted line). The mean difference between traces within the white region of figure 13 was -0.91 mm, with a correlation value of r = 0.82.

Figure 14 shows a similar comparison but this time for a passive rotation of the ankle joint while the participant sat in an ankle dynamometer with a fully extended knee and the foot strapped to the foot plate of the dynamometer. The ankle joint starts from 10° of plantar flexion, and is then driven through plantar-dorsiflexion cycles. Note that the discrepancy between methods is most evident at the longest muscle lengths. The mean absolute difference between traces in this figure was -0.40 mm, with a correlation between traces of r = 0.98.

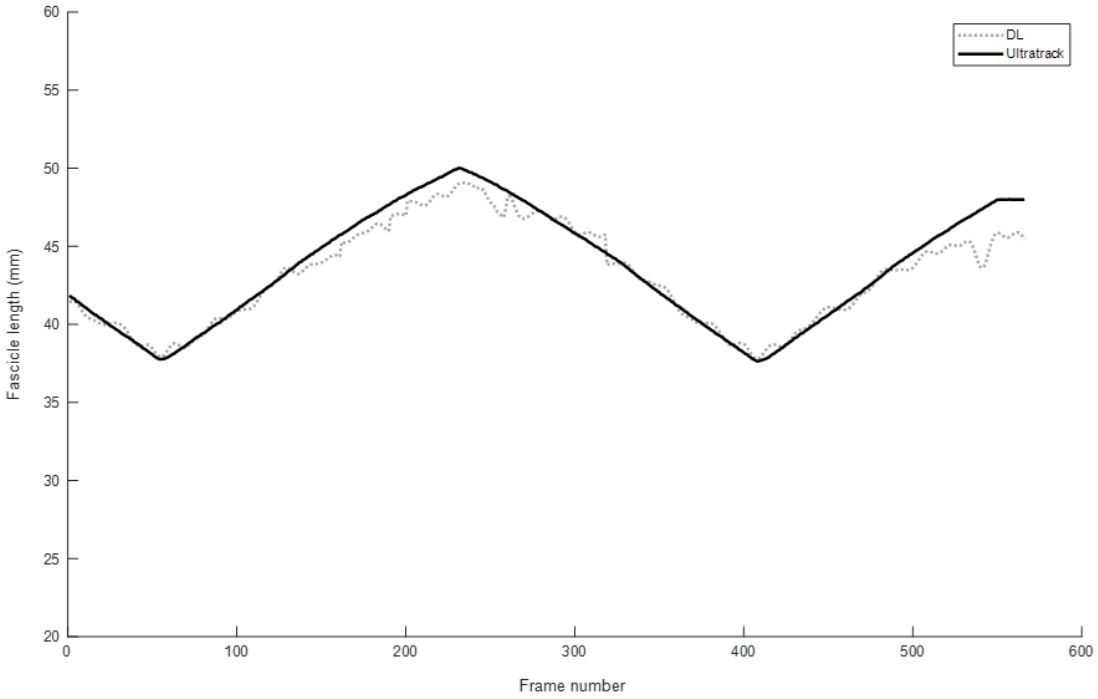

FIGURE 14. Fascicle length traces during passive ankle rotation obtained via the deep learning approach (DL) and using Ultratrack.

In figure 15, the two methods are again compared, but this time for a maximal isometric voluntary contraction, which usually requires large deformations of the muscle. For the data in this figure, the mean absolute difference between traces was -0.37 mm, with a correlation between traces of r = 0.96.

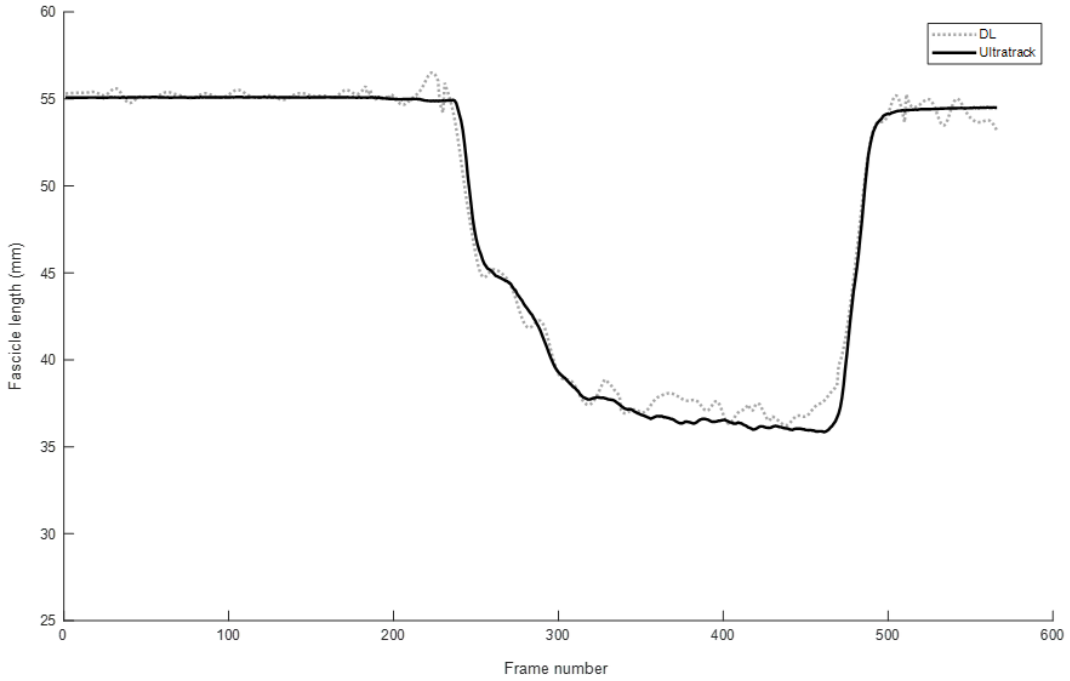

FIGURE 15. Fascicle lengths during a maximal voluntary isometric contraction obtained via the deep learning approach (DL) and using Ultratrack.

4.5 Tracking multiple fascicles instead of a mean value

With the DL method, the trained algorithm usually detects more than one fascicle. For simplicity, the previous data and figures show the median values (fascicle length, pennation) computed from all detected fascicles in a given frame. However, in some cases, it may be desirable to retain information about all detected fascicles. In figure 16, three different conditions are analysed, this time showing the fascicle length (left column) and x location of the start of each fascicle (right column). The software also records the x location of the end point of each fascicle, but this information is not shown in Figure 16. In the left column of this figure, each grey point denotes the length (in pixels; y-axis) of an individual fascicle within a given frame (x-axis). The black traces denote the mean of all detected fascicles for a given frame.

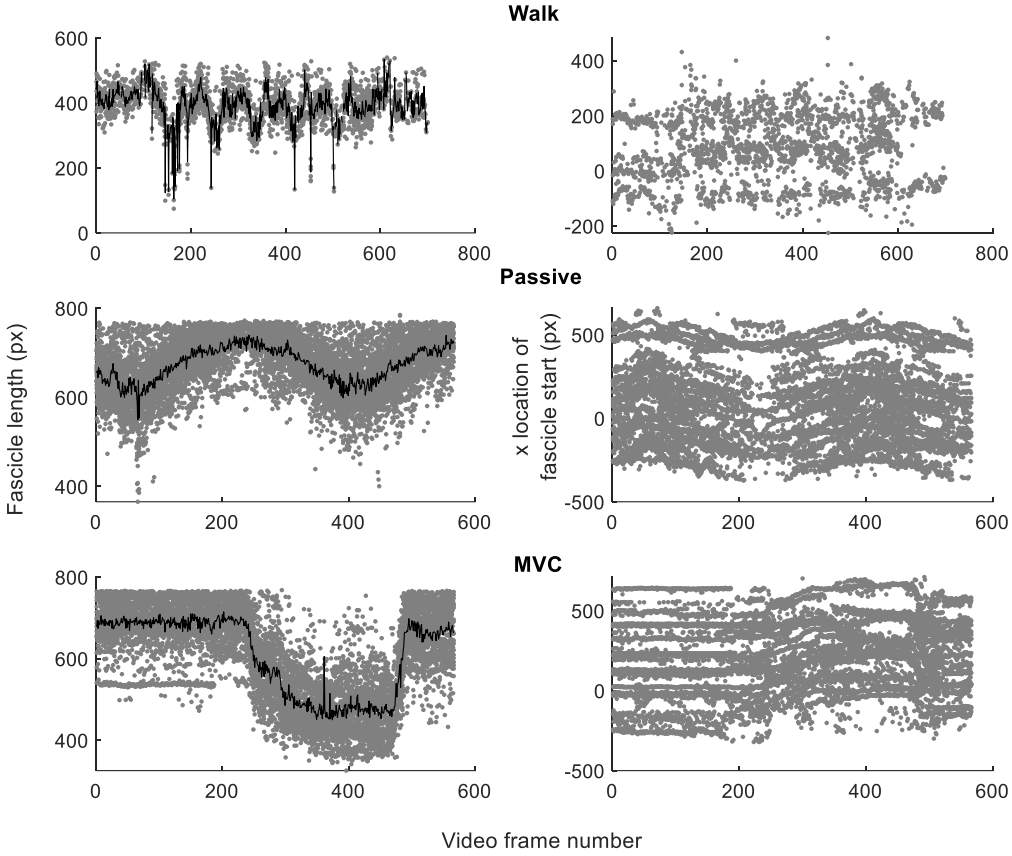

FIGURE 16. Analysis of walking, passive and maximal isometric voluntary contraction (MVC), showing data from all detected fascicles.

As can be seen in the figure, a large number of fascicles are often detected, with fascicle starting points at a wide range of different locations on the x axis of the image (negative pixel values in the right column of figure 16 denote a start point outside of the visible image). This figure also reveals why the median (or mean) value of all detected fascicles may not be the most appropriate metric; in the walking trial in particular, there are several frames where only a few fascicles are detected, and this has a big effect on any average metric, as shown by the large deviations in the black mean trace.

4.6 Failure cases

The deep learning approach was trained to perform a very narrow task within certain constraints. For example, all of the training images were oriented so that the fascicles extend in the same direction, since it proved difficult to train a model that was indifferent to the horizontal orientation of the image. In figure 16, an example is given of the tracking results when an image is both correctly

and incorrectly flipped on the horizontal axis and then analysed with the trained model.

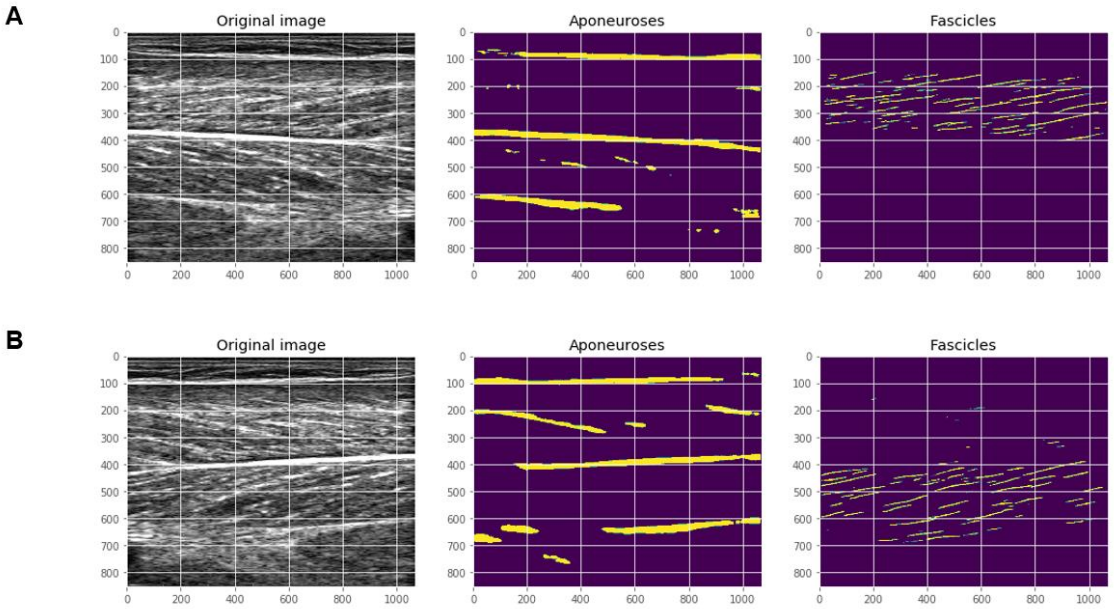

FIGURE 17. Using the trained model to analyse the same image that has been correctly (A) and incorrectly (B) flipped along the horizontal axis.

Figure 17A shows muscle fascicles with the correct orientation appearing in the most superficial muscle (medial gastrocnemius), which is consistent with the training data. In figure 17B, the same image has been flipped horizontally so that the fascicles are oriented in the opposite direction to those in the training images. In this case, the aponeurosis tracking is similar (but reversed), whereas fascicles are only detected in a deeper muscle (soleus), since the orientation of these fascicles is now consistent with those identified in the training set.

Figure 18 shows two further possible failure cases. In A, only a single fascicle is detected by the trained model. This fascicle could be extrapolated and used to denote fascicle length, but this result may not be representative, and it is preferable to detect as many fascicles as possible. In figure 18B, several candidate aponeuroses are detected, and in cases like this, it is surprisingly difficult to filter the unwanted candidates out.

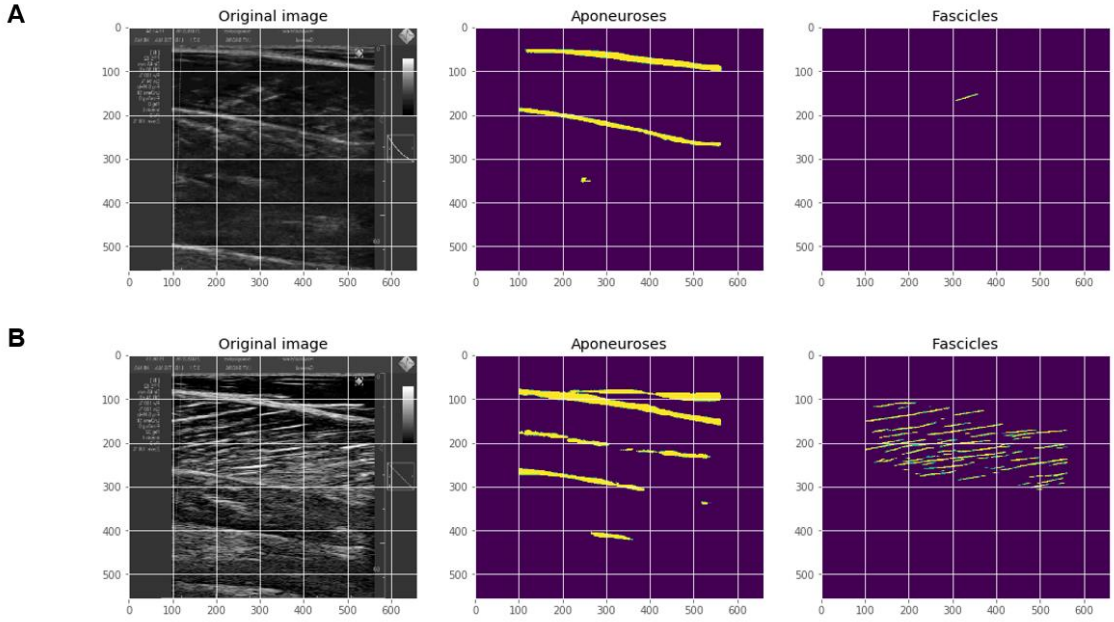

FIGURE 18. Failure cases where too few fascicles are detected (A) or too many aponeuroses are detected (B).

Additional failures may occur during tracking of videos, especially when the task is dynamic in nature. In some frames, the trained model was unable to identify any fascicle or aponeurosis candidates. These failures can easily be filtered if they occur in 1 or 2 consecutive frames, but if large sections of video are untrackable, then the approach will not yield satisfactory results.

5 DISCUSSION

In this work I present a deep learning-based method that provides full automation of muscle architecture from single ultrasound images as well as videos. In general, the method performs very favourably when compared to manual analysis and existing semi-automated methods, and is robust to images from different muscles and those obtained with different ultrasound systems and settings. This method therefore offers a more objective, time-efficient method of segmenting ultrasound images. The method and all labelled training data will also be released under an open source license, allowing others to use and extend this work.

5.1 Comparisons with manual analysis

Fascicle length estimates were generally well correlated between the DL method and manually-derived results. Naturally however, the results between the two methods were not identical for several reasons. Firstly, the DL method computed the fascicle trajectories of multiple fascicles across the width of each image, but for the comparison with manual analysis, I only used the median value output by the DL method. This median value thus includes a lot of variability because of differences in fascicle lengths in different parts of the image (or muscle). On the contrary, for the manual analysis, only a single fascicle was detected from the central part of the image, because it would not be feasible to manually analyse as many fascicles per image as the DL algorithm. An alternative approach to using a median (or mean) value is to use the data computed from all fascicles, regardless of their location in the image, as discussed below.

Out of the three architectural parameters that were assessed, fascicle length estimates varied the most between computer and human. Fascicle length estimates are indeed likely to vary the most because even very slight differences in the interpretation of fascicle angle can have a substantial effect on the length of

an extrapolated fascicle, especially for long fascicles. Furthermore, parts of a fascicle often extend beyond the visible image, so different methods of extrapolation could have large effects on length estimates.

Similarly, this variation is also seen in pennation angle, with differences of up to 5° in some cases. A difference of this size could be very significant, since changes in pennation angle due to long-term training interventions would likely be much smaller than this (Aagaard et al., 2001). However, this comparison does not imply that the deep learning approach is not sufficient for determining this parameter. Rather, it likely represents the difference in approach, as mentioned above. When determining pennation angle from multiple fascicles in different parts of a muscle, it is very likely that the values will differ in different regions. The results presented here simply give the mean pennation angle of all fascicles taken into consideration, which is not the same as only selecting a single fascicle from the middle of the image, as the human labelers did. A simple solution to this problem would be to constrain the deep learning method to only select fascicles from the central portion of the image. However, in some images, e.g. those where there are no clear fascicle-like structures in the middle of the image, this may result in the algorithm failing to detect anything.

Thickness measures were very similar between the two human analysers, as well as between humans and the algorithm. This is unsurprising because the analysis was constrained to the central region of the image, and so similar results would be expected as long as the algorithm is able to successfully detect the aponeuroses.

Curvature tracking with deep learning doesn't currently work well enough based on this implementation. Although the use of a curved model often resulted in different fascicle lengths compared to the straight-line model, these differences were usually small. Moreover, when comparing the output to manually analysed data, it is clear that the algorithm often underestimates the degree of curvature. In the results shown in figure 12, this led to an under- or overestimation of fascicle length of up to about 14% by the algorithm. Clearly, for cases where curvature is a significant problem, a more sophisticated approach of quantifying the curvature is needed than a polynomial fit to the identified points (Darby et al., 2013).

5.2 Comparisons with Ultratrack

In comparison to the results from Ultratrack, my approach yielded traces that were more variable (i.e. less smooth), probably because multiple fascicles were being tracked, and fascicles from slightly different regions are detected in different images. This differs from the approach of Ultratrack, where a single fascicle is first identified, and this location information is then used to update the

fascicle endpoints in subsequent images of the sequence. In other words, Ultratrack does not treat each image in the sequence as being independent, whereas my approach does. It is debatable which approach is superior, but it would be of interest to build in the effect of time to my approach, using some form of recurrent model. This would allow a model to be trained that takes into account the fascicle length (or other parameter) from *n* previous frames, and use this to help inform the result of the current frame.

An additional possibility would be to constrain the identification of fascicles to only the middle portion of the image, since this is a rule that is commonly used when manually analysing the data. However, this could result in failure in cases where few or no candidate fascicles are identified in the middle portion of the image. Moreover, in some cases it may be a specific goal to identify differences in muscle architecture within different muscle regions. With the current approach, this could be achieved by retaining the information from all detected fascicles (see following sub-section), and then averaging the values from specific image regions.

Regarding the noisiness of the fascicle length traces, it would of course be possible to smooth the traces. However, when computing a simple mean of all detected fascicles, smoothing the trace would result in some loss of data, and it would not be immediately clear how to define the boundary between signal and noise. Therefore, I refrained from filtering the traces in the results of this thesis. As stated below, one option is to simply output the lengths of all detected fascicles, allowing a range of lengths to be identified.

It should be noted that the approach presented here is likely slower than Ultratrack when analysing videos, partly due to the post-processing rather than the actual inference time of the neural network models. All of the work presented in this thesis was done using a Geforce GTX 2070 GPU, whereas the average user would likely run this software using a CPU. Nonetheless, given the time requirements of manual analysis, my approach is still likely to offer huge time savings in comparison.

5.3 Tracking multiple fascicles

During analysis, the current DL approach treats each individual frame of a video as being independent from other frames. As a result of this, combined with the fact that multiple fascicles are usually detected, the mean fascicle traces obtained when computing the mean or median per frame are often much less smooth than the equivalent trace produced by Ultratrack. As an alternative, I also output data from all detected fascicles. This approach allows the user to determine the metric used to quantify fascicle length, as well as filtering and/or filling missing data if they so choose.

Given the rather large variations in fascicle lengths in different muscle regions, it may often be preferable to include all data, providing a range of values across both the spatial direction of the muscle and the time dimension of the activity. Moreover, in cases where it is a specific study aim to examine region-dependent muscle architecture, this method is clearly superior to the manual labelling of individual fascicles in different parts of the image. In principle, the current approach could be used to analyse extended field of view scans, where a series of longitudinal images are stitched together to allow imaging of a region that is longer than the ultrasound probe.

5.4 Limitations

One limitation that is common to many applications of this kind is that the approach does not actually exhibit intelligence in the sense that we associate with human thinking (Lake, Ullman, Tenenbaum, & Gershman, 2016). This is demonstrated by some of the common failure cases presented here. These cases would be generally simple to solve for a human, and demonstrate that the approach can be quite brittle when presented with data that is even slightly different to what is present in the training set. The likelihood of some of these failures could be somewhat offset by increasing the size and diversity of the training set, but without any built-in ability to generalise knowledge or contextual information, this kind of approach is unlikely to ever achieve perfect tracking in all cases.

An additional limitation of the current implementation is the lack of an option to manually correct the tracking, e.g. when the human researcher sees an obvious tracking error. In theory, this could be implemented in a future version. I made no attempt to do so here because one of the main goals was to determine whether it is feasible to develop a fully automated approach built on deep learning principles. Moreover, the implications of allowing manual corrections should be considered; the more human intervention there is, the higher the risk of biasing the results, since the human doing the analysis is usually aware of the study context, aims and hypotheses.

Finally, it should be noted that this approach was only trained to analyse data from superficial muscles, and to provide results for a single muscle. This was done because it is sufficient for the majority of studies performed in this field. In future, this could be overcome by training a broader model that is able to detect fascicles and other structures anywhere in an image, regardless of the orientation of the structures or their spatial locations. Alternatively, the user could define a region of interest, allowing the analyses to be localised.

6 CONCLUSION

In this work I present a deep learning approach that allows full automation of the analysis of muscle architecture from ultrasound images and videos. The results produced by this method generally correspond well with those from manual analysis or existing algorithmic approaches. The new method provides fast and objective results, and can be used to analyse different superficial muscles. Certain aspects of this approach, such as the ability to analyse curved muscle fascicles, require further development, as the current implementation often underestimates the degree of curvature (compared to manual analysis). Nonetheless, the code and training data from this project are publicly available (https://github.com/njcronin/DL_Track), allowing other users to benefit from and potentially extend upon this work.

REFERENCES

Aagaard, P., Andersen, J. L., Dyhre-Poulsen, P., Leffers, A.-M., Wagner, A., Magnusson, S. P., … Simonsen, E. B. (2001). A mechanism for increased contractile strength of human pennate muscle in response to strength training: changes in muscle architecture. *The Journal of Physiology*, *534*(2), 613–623. https://doi.org/10.1111/j.1469-7793.2001.t01-1-00613.x

Baka, N., Leenstra, S., & Van Walsum, T. (2017). Ultrasound Aided Vertebral Level Localization for Lumbar Surgery. *IEEE Transactions on Medical Imaging*, *36*(10), 2138–2147. https://doi.org/10.1109/TMI.2017.2738612

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet (London, England)*, *1*(8476), 307–310. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2868172

Bouguet, J., & Bouguet, J. (2000). Pyramidal implementation of the Lucas Kanade feature tracker. *INTEL CORPORATION, MICROPROCESSOR RESEARCH LABS*. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.185.585

Canziani, A., Paszke, A., & Culurciello, E. (2016). *An Analysis of Deep Neural Network Models for Practical Applications*. Retrieved from http://arxiv.org/abs/1605.07678

Caresio, C., Salvi, M., Molinari, F., Meiburger, K. M., & Minetto, M. A. (2017). Fully Automated Muscle Ultrasound Analysis (MUSA): Robust and Accurate Muscle Thickness Measurement. *Ultrasound in Medicine and Biology*, *43*(1), 195–205. https://doi.org/10.1016/j.ultrasmedbio.2016.08.032

Chen, L. C., Papandreou, G., Kokkinos, I., Murphy, K., & Yuille, A. L. (2018). DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(4), 834–848. https://doi.org/10.1109/TPAMI.2017.2699184

Cronin, N.J., Ishikawa, M., af Klint, R., Komi, P. V., Avela, J., Sinkjaer, T., & Voigt, M. (2009). Effects of prolonged walking on neural and mechanical components of stretch responses in the human soleus muscle. *Journal of Physiology*, *587*(17). https://doi.org/10.1113/jphysiol.2009.174912

Cronin, N.J., & Lichtwark, G. (2013). The use of ultrasound to study muscle-tendon function in human posture and locomotion. *Gait and Posture*, *37*(3).

https://doi.org/10.1016/j.gaitpost.2012.07.024

Cronin, N J, Carty, C. P., Barrett, R. S., & Lichtwark, G. (2011). Automatic tracking of medial gastrocnemius fascicle length during human locomotion. *Journal of Applied Physiology (Bethesda, Md.: 1985)*, *111*(5), 1491–1496. https://doi.org/10.1152/japplphysiol.00530.2011

Cronin, N J, Peltonen, J., Ishikawa, M., Komi, P. V, Avela, J., Sinkjaer, T., & Voigt, M. (2008). Effects of contraction intensity on muscle fascicle and stretch reflex behavior in the human triceps surae. *Journal of Applied Physiology (Bethesda, Md.: 1985)*, *105*(1), 226–232. https://doi.org/10.1152/japplphysiol.90432.2008

Cronin, N J, Ishikawa, M., Grey, M. J., af Klint, R., Komi, P. V, Avela, J., … Voigt, M. (2009). Mechanical and neural stretch responses of the human soleus muscle at different walking speeds. *The Journal of Physiology*, *587*(Pt 13), 3375–3382. https://doi.org/10.1113/jphysiol.2008.162610

Cunningham, R., Sánchez, M., May, G., & Loram, I. (2018). Estimating Full Regional Skeletal Muscle Fibre Orientation from B-Mode Ultrasound Images Using Convolutional, Residual, and Deconvolutional Neural Networks. *Journal of Imaging*, *4*(2), 29. https://doi.org/10.3390/jimaging4020029

Darby, J., Li, B., Costen, N., Loram, I., & Hodson-Tole, E. (2013). Estimating skeletal muscle fascicle curvature from B-mode ultrasound image sequences. *IEEE Transactions on Biomedical Engineering*, *60*(7), 1935–1945. https://doi.org/10.1109/TBME.2013.2245328

Drazan, J. F., Hullfish, T. J., & Baxter, J. R. (2019). An automatic fascicle tracking algorithm quantifying gastrocnemius architecture during maximal effort contractions. *PeerJ*, *2019*(7). https://doi.org/10.7717/peerj.7120

Dumoulin, V., & Visin, F. (2016). *A guide to convolution arithmetic for deep learning*. Retrieved from http://arxiv.org/abs/1603.07285

Farris, D. J., & Lichtwark, G. A. (2016). UltraTrack: Software for semi-automated tracking of muscle fascicles in sequences of B-mode ultrasound images. *Computer Methods and Programs in Biomedicine*, *128*, 111–118. https://doi.org/10.1016/j.cmpb.2016.02.016

Franz, J. R., & Thelen, D. G. (2016). Imaging and simulation of Achilles tendon dynamics: Implications for walking performance in the elderly. *Journal of Biomechanics*, *49*(9), 1403–1410. https://doi.org/10.1016/j.jbiomech.2016.04.032

Fukashiro, S., Itoh, M., Ichinose, Y., Kawakami, Y., & Fukunaga, T. (1995).

Ultrasonography gives directly but noninvasively elastic characteristic of human tendon in vivo. *European Journal of Applied Physiology and Occupational Physiology*, *71*(6), 555–557.

Fukunaga, T., Ichinose, Y., & Ito, M. (1997). Determination of fascicle length and pennation in a contracting human muscle in vivo. *Journal of Applied …*, 354–358. Retrieved from http://www.jappl.org/content/82/1/354.short

Fukunaga, T., Kubo, K., Kawakami, Y., Fukashiro, S., Kanehisa, H., & Maganaris, C. N. (2001). In vivo behaviour of human muscle tendon during walking. *Proceedings. Biological Sciences / The Royal Society*, *268*(1464), 229–233. https://doi.org/10.1098/rspb.2000.1361

Garcia-Garcia, A., Orts-Escolano, S., Oprea, S., Villena-Martinez, V., & Garcia-Rodriguez, J. (2017). *A Review on Deep Learning Techniques Applied to Semantic Segmentation*. Retrieved from http://arxiv.org/abs/1704.06857

Ghesu, F. C., Krubasik, E., Georgescu, B., Singh, V., Zheng, Y., Member, S., … Comaniciu, D. (2016). Marginal Space Deep Learning: Efficient Architecture for Volumetric Image Parsing. *IEEE TRANSACTIONS ON MEDICAL IMAGING*, *35*(5). https://doi.org/10.1109/TMI.2016.2538802

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). *Rich feature hierarchies for accurate object detection and semantic segmentation*. Retrieved from http://arxiv.

Greenspan, H., Van Ginneken, B., & Summers, R. M. (2016, May 1). Guest Editorial Deep Learning in Medical Imaging: Overview and Future Promise of an Exciting New Technique. *IEEE Transactions on Medical Imaging*, Vol. 35, pp. 1153–1159. https://doi.org/10.1109/TMI.2016.2553401

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *ArXiv*. Retrieved from http://image-net.org/challenges/LSVRC/2015/

Herbert, R. D., & Gandevia, S. C. (1995). Changes in pennation with joint angle and muscle torque: in vivo measurements in human brachialis muscle. *The Journal of Physiology*, *484*(2), 523–532. https://doi.org/10.1113/jphysiol.1995.sp020683

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507. https://doi.org/10.1126/science.1127647

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, *1*, 448–456. International Machine Learning

Society (IMLS).

Ito, M., Kawakami, Y., Ichinose, Y., Fukashiro, S., & Fukunaga, T. (1998). Nonisometric behavior of fascicles during isometric contractions of a human muscle. *Journal of Applied Physiology (Bethesda, Md. : 1985)*, *85*(4), 1230–1235. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22190314

Jahanandish, M. H., Fey, N. P., & Hoyt, K. (2019). Lower Limb Motion Estimation Using Ultrasound Imaging: A Framework for Assistive Device Control. *IEEE Journal of Biomedical and Health Informatics*, *23*(6), 2505–2514. https://doi.org/10.1109/JBHI.2019.2891997

Kamnitsas, K., Ledig, C., Newcombe, V. F. J., Simpson, J. P., Kane, A. D., Menon, D. K., … Glocker, B. (2016). *Efficient Multi-Scale 3D CNN with Fully Connected CRF for Accurate Brain Lesion Segmentation*. https://doi.org/10.1016/j.media.2016.10.004

Karamanidis, K., Travlou, A., Krauss, P., & Jaekel, U. (2016). Use of a Lucas-Kanade-based template tracking algorithm to examine in vivo tendon excursion during voluntary contraction using ultrasonography. *Ultrasound in Medicine and Biology*, *42*(7), 1689–1700. https://doi.org/10.1016/j.ultrasmedbio.2016.02.019

Kawakami, Y., Abe, T., & Fukunaga, T. (1993). Muscle-fiber pennation angles are greater in hypertrophied than in normal muscles. *Journal of Applied Physiology (Bethesda, Md. : 1985)*, *74*(6), 2740–2744. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8365975

Kawakami, Y., Ichinose, Y., & Fukunaga, T. (1998). Architectural and functional features of human triceps surae muscles during contraction. *Journal of Applied Physiology (Bethesda, Md. : 1985)*, *85*(2), 398–404. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9688711

Körting, C., Schlippe, M., Petersson, S., Pennati, G. V., Tarassova, O., Arndt, A., … Wang, R. (2019). In vivo muscle morphology comparison in post-stroke survivors using ultrasonography and diffusion tensor imaging. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-47968-x

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *International Conference on Neural Information Processing Systems*, 1097–1105. https://doi.org/10.1145/3065386

Kubo, K., Kanehisa, H., Takeshita, D., Kawakami, Y., Fukashiro, S., & Fukunaga, T. (2000). In vivo dynamics of human medial gastrocnemius muscle-tendon complex during stretch-shortening cycle exercise. *Acta Physiologica Scandinavica*, *170*(2), 127–135. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/11114950

Kurokawa, S., Fukunaga, T., & Fukashiro, S. (2001). Behavior of fascicles and tendinous structures of human gastrocnemius during vertical jumping. *Journal of Applied Physiology (Bethesda, Md.: 1985)*, *90*(4), 1349–1358. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11247934

Kwah, L. K., Pinto, R. Z., Diong, J., & Herbert, R. D. (2013, March 15). Reliability and validity of ultrasound measurements of muscle fascicle length and pennation in humans: A systematic review. *Journal of Applied Physiology*, Vol. 114, pp. 761–769. https://doi.org/10.1152/japplphysiol.01430.2011

Lai, A., Lichtwark, G. A., Schache, A. G., Lin, Y.-C., Brown, N. A. T., & Pandy, M. G. (2015). In vivo behavior of the human soleus muscle with increasing walking and running speeds. *Journal of Applied Physiology (Bethesda, Md.: 1985)*, *118*(10), 1266–1275. https://doi.org/10.1152/japplphysiol.00128.2015

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building Machines That Learn and Think Like People. *Behavioral and Brain Sciences*, *40*. Retrieved from http://arxiv.org/abs/1604.00289

Lecun, Y., Bengio, Y., & Hinton, G. (2015, May 27). Deep learning. *Nature*, Vol. 521, pp. 436–444. https://doi.org/10.1038/nature14539

Leitner, C., Hager, P. A., Penasso, H., Tilp, M., Benini, L., Peham, C., & Baumgartner, C. (2019, October 1). Ultrasound as a tool to study muscle–tendon functions during locomotion: A systematic review of applications. *Sensors (Switzerland)*, Vol. 19. https://doi.org/10.3390/s19194316

Lichtwark, G. A., & Wilson, A. M. (2006). Interactions between the human gastrocnemius muscle and the Achilles tendon during incline, level and decline locomotion. *Journal of Experimental Biology*, *209*(21), 4379–4388. https://doi.org/10.1242/jeb.02434

Lichtwark, G. a, Bougoulias, K., & Wilson, a M. (2007). Muscle fascicle and series elastic element length changes along the length of the human gastrocnemius during walking and running. *Journal of Biomechanics*, *40*(1), 157–164. https://doi.org/10.1016/j.jbiomech.2005.10.035

Litjens, G., Kooi, T., Bejnordi, B. E., Setio, A. A. A., Ciompi, F., Ghafoorian, M., … Sánchez, C. I. (2017, December 1). A survey on deep learning in medical image analysis. *Medical Image Analysis*, Vol. 42, pp. 60–88. https://doi.org/10.1016/j.media.2017.07.005

Liu, S., Wang, Y., Yang, X., Lei, B., Liu, L., Li, S. X., … Wang, T. (2019, April 1). Deep Learning in Medical Ultrasound Analysis: A Review. *Engineering*, Vol. 5, pp. 261–275. https://doi.org/10.1016/j.eng.2018.11.020

Long, J., Shelhamer, E., & Darrell, T. (2015). *Fully Convolutional Networks for Semantic Segmentation*.

Loram, I. D., Maganaris, C. N., & Lakie, M. (2004). Paradoxical muscle movement in human standing. *The Journal of Physiology*, *556*(Pt 3), 683–689. https://doi.org/10.1113/jphysiol.2004.062398

Loram, I. D., Maganaris, C. N., & Lakie, M. (2006). Use of ultrasound to make noninvasive in vivo measurement of continuous changes in human muscle contractile length. *Journal of Applied Physiology*, *100*(4), 1311–1323. https://doi.org/10.1152/japplphysiol.01229.2005

Ma, J., Wu, F., Jiang, T., Zhao, Q., & Kong, D. (2017). Ultrasound image-based thyroid nodule automatic segmentation using convolutional neural networks. *International Journal of Computer Assisted Radiology and Surgery*, *12*(11), 1895–1910. https://doi.org/10.1007/s11548-017-1649-7

Magnusson, S. P., Hansen, P., Aagaard, P., Brønd, J., Dyhre-Poulsen, P., Bojsen-Moller, J., & Kjaer, M. (2003). Differential strain patterns of the human gastrocnemius aponeurosis and free tendon, in vivo. *Acta Physiologica Scandinavica*, *177*(2), 185–195. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12558555

Marzilger, R., Legerlotz, K., Panteli, C., Bohm, S., & Arampatzis, A. (2018). Reliability of a semi-automated algorithm for the vastus lateralis muscle architecture measurement based on ultrasound images. *European Journal of Applied Physiology*, *118*(2), 291–301. https://doi.org/10.1007/s00421-017-3769-8

Milletari, F., Ahmadi, S.-A., Kroll, C., Plate, A., Rozanski, V., Maiostre, J., … Navab, N. (2016). Hough-CNN: Deep Learning for Segmentation of Deep Brain Regions in MRI and Ultrasound. *Computer Vision and Image Understanding*, *164*, 92–102. Retrieved from http://arxiv.org/abs/1601.07014

Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). *V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation*. Retrieved from http://arxiv.org/abs/1606.04797

Mitchell Waldrop, M. (2019). What are the limits of deep learning? *Proceedings of the National Academy of Sciences of the United States of America*, *116*(4), 1074–1077. https://doi.org/10.1073/pnas.1821594116

Narici, M. V., Maganaris, C. N., Reeves, N. D., & Capodaglio, P. (2003). Effect of aging on human muscle architecture. *Journal of Applied Physiology*, *95*(6), 2229–2234. https://doi.org/10.1152/japplphysiol.00433.2003

Narici, M. V, Binzoni, T., Hiltbrand, E., Fasel, J., Terrier, F., & Cerretelli, P. (1996). In vivo human gastrocnemius architecture with changing joint angle at rest and during graded isometric contraction. *The Journal of Physiology*, *496 (Pt 1*, 287–297. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1160844&tool=pmcentrez&rendertype=abstract

Nicolson, M., & Fleming, J. E. E. (John E. E. (2013). *Imaging and imagining the fetus : the development of obstetric ultrasound*. Johns Hopkins University Press.

Rana, M., Hamarneh, G., & Wakeling, J. M. (2009). Automated tracking of muscle fascicle orientation in B-mode ultrasound images. *Journal of Biomechanics*, *42*(13), 2068–2073. https://doi.org/10.1016/j.jbiomech.2009.06.003

Ravishankar, H., Venkataramani, R. B., Thiruvenkadam, S., & Sudhakar, P. (2017). Learning and Incorporating Shape Models. *Miccai 2017*, *10433*(2), 203–211. https://doi.org/10.1007/978-3-319-66182-7

Reeves, N. D., Narici, M. V., & Maganaris, C. N. (2004). In vivo human muscle structure and function: Adaptations to resistance training in old age. *Experimental Physiology*, *89*(6), 675–689. https://doi.org/10.1113/expphysiol.2004.027797

Ronneberger, O., Fischer, P., & Brox, T. (2015). 2015-U-Net. *ArXiv*, 1–8. Retrieved from arxiv:1505.04597v1

Rutherford, O. M., & Jones, D. A. (1992). Measurement of fibre pennation using ultrasound in the human quadriceps in vivo. *European Journal of Applied Physiology and Occupational Physiology*, *65*(5), 433–437. https://doi.org/10.1007/bf00243510

Schindelin, J., Arganda-Carreras, I., Frise, E., Kaynig, V., Longair, M., Pietzsch, T., … Cardona, A. (2012, July). Fiji: An open-source platform for biological-image analysis. *Nature Methods*, Vol. 9, pp. 676–682. https://doi.org/10.1038/nmeth.2019

Schmidhuber, J. (2015, January 1). Deep Learning in neural networks: An overview. *Neural Networks*, Vol. 61, pp. 85–117. https://doi.org/10.1016/j.neunet.2014.09.003

Seynnes, O. R., De Boer, M., & Narici, M. V. (2007). Early skeletal muscle hypertrophy and architectural changes in response to high-intensity resistance training. *Journal of Applied Physiology*, *102*(1), 368–373. https://doi.org/10.1152/japplphysiol.00789.2006

Seynnes, O.R., Bojsen-Møller, J., Albracht, K., Arndt, A., Cronin, N. J., Finni, T.,

& Magnusson, S. P. (2015). Ultrasound-based testing of tendon mechanical properties: A critical evaluation. *Journal of Applied Physiology*, *118*(2). https://doi.org/10.1152/japplphysiol.00849.2014

Seynnes, Olivier R., & Cronin, N. J. (2019). *Simple Muscle Architecture Analysis (SMA): an ImageJ macro tool to automate measurements in B-mode ultrasound scans*. Retrieved from http://arxiv.org/abs/1905.09490

Shelhamer, E., Long, J., & Darrell, T. (2017). Fully Convolutional Networks for Semantic Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(4), 640–651. https://doi.org/10.1109/TPAMI.2016.2572683

Shen, D., Wu, G., & Suk, H.-I. (2017). Deep Learning in Medical Image Analysis. *Annual Review of Biomedical Engineering*, *19*(1), 221–248. https://doi.org/10.1146/annurev-bioeng-071516-044442

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*. International Conference on Learning Representations, ICLR.

Sipila, S., & Suominen, H. (1991). Ultrasound imaging of the quadriceps muscle in elderly athletes and untrained men. *Muscle & Nerve*, *14*(6), 527–533. https://doi.org/10.1002/mus.880140607

Sousa, F., Ishikawa, M., Vilas-Boas, J. P., & Komi, P. V. (2007). Intensity- and muscle-specific fascicle behavior during human drop jumps. *Journal of Applied Physiology (Bethesda, Md.: 1985)*, *102*(1), 382–389. https://doi.org/10.1152/japplphysiol.00274.2006

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). *On the importance of initialization and momentum in deep learning*.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). *Going Deeper with Convolutions*.

Tajbakhsh, N., Shin, J. Y., Gurudu, S. R., Hurst, R. T., Kendall, C. B., Gotway, M. B., & Liang, J. (2016). Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning? *IEEE Transactions on Medical Imaging*, *35*(5), 1299–1312. https://doi.org/10.1109/TMI.2016.2535302

Topol, E. J. (2019, January 1). High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, Vol. 25, pp. 44–56. https://doi.org/10.1038/s41591-018-0300-7

Wan, L., Zeiler, M., Zhang, S., Lecun, Y., & Fergus, R. (2013). *Regularization of Neural Networks using DropConnect*.

Wu, H., Zhang, J., Huang, K., Liang, K., & Yu, Y. (2019). *FastFCN: Rethinking Dilated Convolution in the Backbone for Semantic Segmentation*. Retrieved from http://arxiv.org/abs/1903.11816

Wu, L., Cheng, J. Z., Li, S., Lei, B., Wang, T., & Ni, D. (2017). FUIQA: Fetal ultrasound image quality assessment with deep convolutional networks. *IEEE Transactions on Cybernetics*, *47*(5), 1336–1349. https://doi.org/10.1109/TCYB.2017.2671898

Zhang, Y., Ying, M. T. C., Yang, L., Ahuja, A. T., & Chen, D. Z. (2017). Coarse-to-Fine Stacked Fully Convolutional Nets for lymph node segmentation in ultrasound images. *Proceedings - 2016 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2016*, 443–448. https://doi.org/10.1109/BIBM.2016.7822557

Zhou, G. Q., Chan, P., & Zheng, Y. P. (2015). Automatic measurement of pennation angle and fascicle length of gastrocnemius muscles using real-time ultrasound imaging. *Ultrasonics*, *57*(C), 72–83. https://doi.org/10.1016/j.ultras.2014.10.020

Zhou, G. Q., Zhang, Y., Wang, R. L., Zhou, P., Zheng, Y. P., Tarassova, O., … Chen, Q. (2018). Automatic myotendinous junction tracking in ultrasound images with phase-based segmentation. *BioMed Research International*, *2018*. https://doi.org/10.1155/2018/3697835