OpenArm 2.0: Automated Segmentation of 3D Tissue Structures for Multi-Subject Study of Muscle Deformation Dynamics

Yonatan Nozik*, Laura A. Hallock*, Daniel Ho, Sai Mandava, Chris Mitchell, Thomas Hui Li, and Ruzena Bajcsy

***Abstract*— We present a novel neural-network-based pipeline for segmentation of 3D muscle and bone structures from localized 2D ultrasound data of the human arm. Building from the U-Net [1] neural network framework, we examine various data augmentation techniques and training data sets to both optimize the network's performance on our data set and hypothesize strategies to better select training data, minimizing manual annotation time while maximizing performance. We then employ this pipeline to generate the OpenArm 2.0 data set, the first factorial set of multi-subject, multi-angle, multi-force scans of the arm with full volumetric annotation of the biceps and humerus. This data set has been made available on SimTK (`https://simtk.org/projects/openarm`) to enable future exploration of muscle force modeling, improved musculoskeletal graphics, and assistive device control.**

I. Introduction

A deep understanding of human motion dynamics requires a deep understanding of our (muscle) actuators. Although a number of systems exist that attempt to model multi-muscle structures [2, 3], these frameworks make strong optimization-based assumptions about the distribution of force across synergistic muscles that preclude modeling of many motions of interest. In particular, these systems often assume that humans are exerting the least muscle force possible to execute a particular dynamic trajectory, an assumption that prevents modeling of stiffness, balance, and dexterity in both healthy individuals and those who exhibit musculoskeletal pathology (e.g., stroke-induced plegia, antagonistic co-contraction due to Parkinsons' disease). In addition, these high-dimensional systems largely rely on sparse and aggregate data to fit models to individuals, resulting in substantial modeling error.

These limitations — inaccurate force attribution between muscles and inadequate consideration of musculoskeletal geometry variation across individuals — could be mitigated by improved measurement of *individual muscles*. At the same time, single-muscle force and stiffness properties cannot be readily disambiguated from joint torque values (the result of multiple muscles' efforts) or surface electromyography (sEMG) data (an aggregate and noisy activation signal of all muscles near the sensor's electrode). In this paper, we argue that *muscle deformation* represents a promising signal to gain insight into both passive tissue properties and patterns of muscle activity, as deformation occurs under both passive changes in kinematic configuration and active loading [4]. In particular, deformation provides a *mechanical* signal that is more tightly coupled to muscle dynamics than standard *neurological* biosignals like sEMG, and it can be much more precisely localized to specific structures of interest. Improved understanding and modeling of this deformation and how it relates to body dynamics could therefore revolutionize a number of fields, including biosignal-driven prosthesis control, studies of musculoskeletal pathology and motor control, and graphical rendering of human motion.

Nevertheless, muscle deformation remains an under-utilized and under-explored signal in the biomechanics community. A primary reason for this lack of usage is that the observed signals are both poorly characterized (due to complex internal muscle and external contact dynamics) and prohibitively time-consuming to annotate. In this work, we seek to address both of these limitations by developing an *automated pipeline to generate 3D annotated muscle data from localized 2D ultrasound scans via convolutional neural network (CNN) segmentation*.

The contributions of this paper are as follows:

- a novel CNN-based pipeline for generating 3D annotated scans of bone and muscle structures, trained in a proof-of-concept setting to annotate the humerus and biceps brachii in a scan of the full arm of multiple subjects under multiple conditions;
- a preliminary quantitative evaluation of neural network architectures, and their possible modifications, for use in the ultrasound tissue segmentation domain;
- all annotation code, including both trained networks and training infrastructure for application to new tissues; and
- all generated arm ultrasound scans and their tissue annotations, which encompass multiple subjects under multiple loading conditions and joint configurations and thus form a compelling data set with which to study muscle deformation.

The latter two contributions have been made available for general research use as part of the OpenArm project on SimTK (`https://simtk.org/projects/openarm`).

II. Design Decisions & Related Work

Although there has been some exploration of muscle motion in the graphics community [5–7], as well as preliminary use of related imaging signals for device control [8, 9] and measurement of musculoskeletal geometry [10], studies of muscle deformation are sporadic and distributed across a number of fields. As a result, there is little consensus on which specific deformation signals are most informative of musculoskeletal dynamics, and studies that use the signals

*These authors contributed equally to this work.

This work was supported by the NSF National Robotics Initiative (award no. 81774), Siemens Healthcare (85993), the NVIDIA Corporation GPU Grant Program, eZono AG, and the NSF Graduate Research Fellowship Program.

The authors are with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, Berkeley, CA 94720, USA {nozik, lhallock, ho, smandava, chrismitchell, thomashli, bajcsy}@berkeley.edu

are largely restricted to black-box models that yield little scientific insight.

In an attempt to disambiguate these informative signals, a preliminary study on a single individual — which generated the first OpenArm data set — showed the existence of quantifiable muscle deformation with changes in both muscle force and kinematic configuration, as well as the substantial degree to which measured variation changes based on the region of the muscle observed [4]. These findings support the necessity of studying deformation throughout the full muscle volume, but cannot be generalized without increasing the subject cohort; because manual annotation of muscle volumes is prohibitively time-intensive, such a study has not yet been possible.

This paper directly addresses this segmentation challenge via an automated framework to generate 3D annotated scans of bone and muscle structures from localized 2D ultrasound data. While the pipeline we present could easily be re-trained for various tissue segmentation tasks, we here present a proof-of-concept application to upper-limb muscle deformation with the immediate goal of generating the next OpenArm data set. We argue for both the importance of this data set and our specific framework design decisions in the subsections below.

A. The Case For Full-3D Deformation Imaging

Although muscle deformation can be observed in one or two dimensions (e.g., muscle thickness or cross-sectional area changes, respectively), there do not exist models that can translate these observations into clinically-relevant quantities like stiffness, activation, and force output. A few models exist in the graphics community [5–7], but none (to our knowledge) have been validated against in vivo human data. Internal muscle dynamics are complex and largely unobservable; extending micro- or mezzo-scale biological models of muscle fibers [11, 12] to macro-scale deformation phenomena would require an immeasurably precise knowledge of both motor unit distribution and contact dynamics of the surrounding tissues. Musculoskeletal geometry also varies substantially across individuals in ways that significantly impact system dynamics [13, 14], and the dominance of tissue contact effects means that deformation study is largely restricted to in vivo — rather than ex vivo — study.

This complexity of muscle dynamics has meant that the most successful macro-scale dynamics models are *phenomenological* — i.e., models constructed from data using system identification techniques rather than grounded in a known biological process. (The most widely-adopted example is perhaps the Hill model [15].) We argue that muscle deformation could benefit from similar phenomenological modeling, which requires careful and principled observation of the muscle in all three dimensions across all domains we hope to describe (namely, the variation in shape during both kinematic and activity changes, generalizable across individuals).

B. Segmentation Targets & Conditions: Biceps Brachii & Humerus under Varied Loading & Joint Angle

We focus on a limited number of tissue structures — the biceps brachii and the humerus — as a concession to the time-intensive nature of 3D data collection. (Generating scans of an individual under multiple configurations is a time-consuming process, as is generation of — largely manual — ground truth segmentation validation data.) All scans are collected under static, though loaded, conditions; we hope to expand the work to encompass full dynamic models in the future with the insights gained during static study.

We target these structures for two primary reasons. First, the two structures are representative of the tissue segmentation challenge in that they have significantly different appearance and properties (e.g., different elasticity properties, tissue-specific artifacts like bone shadow); successful segmentation of these structures thus constitutes strong evidence that segmentation of other muscles and bones would be successful using similar methodologies. Second, limiting the number of structures examined allows us to tractably perform segmentation across subjects and under changes in loading condition and kinematic configuration. The generated scans — of both the rigid humerus, which can be used for alignment, and the deforming biceps — therefore constitute a data set that will allow for unprecedented exploration of muscle deformation models across multiple individuals under multiple conditions.

C. Imaging Modality: Localized 2D Ultrasound

To generate 3D tissue scans, MRI is perhaps the most obvious imaging modality and has been used in prior studies of muscle motion [16]. However, limited bore size makes scanning under natural arm configurations difficult or impossible, and long scan times make collecting data under multiple conditions prohibitively expensive and time-consuming. This is especially true when scanning under loaded configurations, as muscle fatigue may significantly influence the observed deformation signal.

Instead, we collect 3D scans via sweeps of a 2D brightness mode (B-mode) ultrasound probe — a technology that is cheaper, better studied, and more readily available than 3D ultrasound — whose position is spatially registered via motion capture. This technique has been widely used in both human- and robot-guided ultrasound imaging for surgical [17] and exploratory [18, 19] applications, and comprehensive tools for scan generation are widely available [20]. While the system is vulnerable to significant ultrasound-specific artifacts, including bone shadow and imaging of the surface gel, structures of interest (bone surfaces, muscle fascia, etc.) are readily visible. Moreover, the comparatively short scan time and unrestricted workspace permit numerous scans of a single subject.

D. Segmentation Approach: Convolutional Neural Networks

Convolutional Neural Networks (CNNs) are a promising tool in many medical imaging domains [21], including echocardiogram annotation [22], brain lesion segmentation [23], and localization of organ structures [24]. In this work, we adapt existing CNN-based methods to segment our tissue structures of interest, the humerus and the biceps brachii.

On a conceptual level, this problem is well suited to a deep learning approach. Rules for segmentation are difficult to enumerate a priori (due to both the modality-specific artifacts mentioned above and the varied geometric behaviors of different tissue structures, which make classical registration

TABLE I

VOLUMETRIC DATA COLLECTION CONDITIONS

Manipulated Factors		Levels‡
θ	elbow flexion angle	0°, 30°, 60°, 90°
LC	elbow load condition	fully supported (FS) 0% MVC (P0), 10% MVC (P1), 30% MVC (P3), 50% MVC (P5)

‡*Data were collected from 10 subjects at each of the 4 flexion angles and 5 loading conditions listed above in a factorial manner, for a total of 20 scans per subject. Additional data were collected from a single elderly subject at 30° and 60° flexion angles (all loading conditions) and 0° and 90° (FS only), for a total of 12 scans. Note that fully supported (FS) trials occurred while the arm was held in position by the experimental jig (i.e., to measure "pure" kinematic deformation), while the latter 4 loading conditions denote percentages of maximum voluntary contraction (MVC) level.*

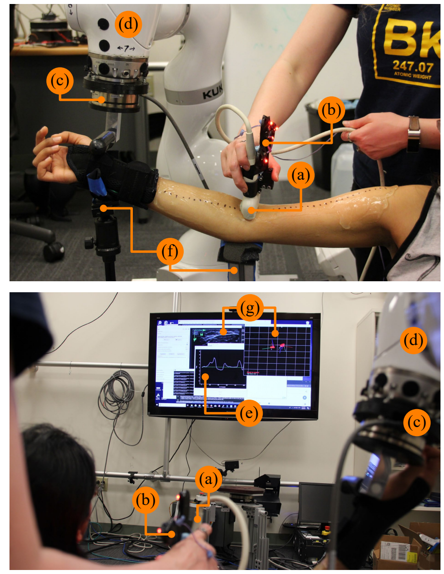

Fig. 1. Experimental setup for the collection of full-arm upper-limb morphology data under multiple elbow angles and loading conditions (shown here at a 30° angle of elbow flexion under FS (*top*) and P1 (*bottom*) loading conditions). Setup includes ultrasound probe (a) (with attached active motion capture markers (b) used for spatial tracking); force/torque sensor (c), held statically in place by KUKA LBR iiwa 14 R820 robot (d) and used by the subject via real-time visual feedback (e) to maintain constant force output during loaded trials; mechanical jig (f) used to support the elbow (during all trials) and the forearm (during FS trials, *top*); and real-time ultrasound and motion capture data (g) for continuous system status monitoring.

approaches difficult); at the same time, the domain is well-specified and highly constrained, such that generating domain-spanning training data is a tractable problem.

As discussed in Section IV, we find that CNN-based segmentation significantly outperforms classical registration-based approaches in both quantitative segmentation accuracy and qualitative manual cleanup time.

III. METHODS

In the following section, we outline our methods for both data set collection and CNN-based segmentation.

A. Data Set Collection

Volumetric data of the full anterior surface of the arm were collected from 10 individuals under 4 kinematic configurations (i.e., elbow angles) and 5 subject-specific loading conditions in a full factorial manner, with factors and levels as listed in Table I. Collection methods are largely those used in the generation of the first OpenArm data set [4], aside from updates to subject demographics, posture, and loading conditions, as noted below.

1) Subject Biometric Data & Consent: Data were collected from the right arm of 10 subjects (6 male, 4 female, all right-handed, age 21.4±2.46, mass 66.9±10.1 kg, height 1.72±0.0876 m, body mass index 22.5±2.63).* All subjects were healthy, with a wide variety of exercise regimes and body types. An incomplete set of scans, as noted in Table I, was collected from an additional elderly subject (female, right-handed, age 85, mass 61.2 kg, height 1.52 m, body mass index 26.4) for preliminary evaluation of generalizability across age groups. The study protocol was approved by the University of California Institutional Review Board for human protection and privacy, under Protocol ID 2016-01-8261. Each subject was first informed of the experimental procedure and written informed consent was obtained.

2) Data Collection: During data collection, the test subject sat erect in a low stadium chair, with legs comfortably extended and right arm extended laterally from the body at a 90 degree shoulder abduction angle. The forearm was fully supinated, with the upper arm supported at the distal end of the humerus, as shown in Figure 1. Scans were then collected with the subject's elbow held statically at each of the 4 angle values listed in Table I (as measured from full extension) under 5 loading conditions (fully supported by a jig at the wrist and unsupported while pressing upward on a force-torque sensorized handle with 4 prescribed levels of force), for a total of 20 trials. Subjects wore a brace to limit wrist flexion force and more completely isolate the elbow.

Loading conditions were selected for each participant based on the subject's maximum voluntary contraction (MVC) value. To measure this MVC value, subjects were asked to press upward on the handle with maximum possible force, then hold for several seconds, then release. Subjects performed this sequence twice, both at full elbow extension, and the maximum overall force value was recorded as MVC. (Empirically, this value varied substantially across subjects, with mean and standard deviation 66.1±28.8 N for the 10 primary subjects tested.)† To allow subjects to maintain a given force for the several minutes required to generate a full 3D scan, force conditions were chosen as 0, 10, 30, and 50 percent of the MVC value. Note that the same MVC value, and thus the same force conditions, were used for all angle conditions for the same subject, to allow for development of models explicitly relating muscle deformation to force. Collecting this value at full extension — the angle at which muscles are weakest [25] — ensured that subjects could maintain the forces required under all tested conditions. Subjects maintained the prescribed force during each trial by matching visual feedback from the attached force/torque sensor (ATI Axia80, ATI Industrial Automation, Apex, NC, USA) to a marked goal value on a real-time series plot, as

*Statistics are reported as *mean* ± *standard deviation*.

†This sequence was performed at 30° of extension for the single elderly subject to avoid hyperextension injury, at which the measured MVC was 42.5 N.

shown in Figure 1. All subjects were able to consistently maintain this value within several newtons.

During each trial, ultrasound images were collected using a portable commercial ultrasound scanner (eZono 4000, eZono AG, Jena, Germany) equipped with a 3-12MHz linear transducer (L3-12 NGS, eZono AG, Jena, Germany) by an experienced operator in the same manner as those of the first OpenArm data set [4]. As in previous work, the machine was configured to collect B-mode data at a depth of 4 cm, with a 3.8 cm transducer footprint, and the 2D ultrasound data was spatially localized using a PhaseSpace active motion capture system (PhaseSpace Inc., San Leandro, CA, USA) and calibrated both spatially and temporally using the open-source PLUS calibration toolkit [20]. Scans were again streamed to an external computer at a rate of 30 fps through an OpenIGTLink server [26] and reconstructed using the volume reconstruction application provided by the PLUS toolkit.

The full experimental setup is shown in Figure 1, and representative volumetric data can be seen in Figure 2 as the spatial intensity map from which volumes are segmented.

B. Candidate Segmentation Architectures & Modifications

To segment the humerus and biceps brachii structures from the generated scans, we explored the following neural net architectures and data augmentation techniques. All networks were trained in Tensorflow [27] using the Adam optimizer [28] and a cross-entropy loss function on a custom-built desktop machine with an INTEL Core i7-5820K six-core CPU and an NVIDIA Titan Xp GPU. Hyperparameter values and full architecture details have been made available with code release.

1) Baseline Architecture: 2D U-Net: Although our objective was to generate 3D scans, we chose the 2D U-Net [1] as our baseline segmentation architecture, which we applied to axial slices of the full upper-arm volume; each slice of each test scan was then predicted individually to generate full volumetric predictions. The U-Net was designed to perform well when trained on relatively small data sets and has been widely applied to various biomedical image segmentation tasks [22, 29], and is more computationally efficient than many of its 3D counterparts. All examined networks built upon the original U-Net architecture, with one additional concatenation block (corresponding to four additional 3x3 convolutions, one additional 2x2 max pooling operation, and one additional "up-convolution").

2) Data Augmentation: To generate additional training data without prohibitively time-consuming manual annotation, we artificially increased the size of our training data set via both rotational and elastic deformation, a useful and common practice in neural network training when comprehensive data is not available [30]. Specifically, we generated additional scans through arbitrary rotation and random elastic deformation [31] and trained networks both with and without this augmented data, as noted in Section IV and Table II.

3) Baselines/Controls for Comparison: We evaluate the neural network approaches against a classical registration-based approach in which the tissue structures from one scan are mapped to another by finding the optimal transformation between the two spatial intensity maps. Specifically, we compare the methods above with both pure rigid registration (as a simple baseline) and a set of sequentially higher-degree-of-freedom transformations (rigid, affine, and B-spline-parameterized) in which each transformation is used to initialize the next registration (as a more complex baseline that better represents the upper limits of classical registration's efficacy). Optimal transformations were calculated using the SimpleElastix image registration library [32]; transformation quality was evaluated at each optimization step using the mutual information criterion. Manual translation was performed prior to automated registration to better align the scans, and sensitive hyperparameters were tuned via grid search to further optimize the results. Baseline registration code and associated hyperparameter values have been released with the rest of the OpenArm code base.

C. Generation of a Ground-Truth Data Set

To train and evaluate the segmentation pipeline across kinematic configurations, loading conditions, and subjects, we manually segmented full-scan volumetric data for the following subjects:

- *Sub1* – all angles and loading conditions (20 scans)
- *Sub2* – all loading conditions at 30° (5 scans)
- *Sub3* & *Sub4* – single scans at 30°, FS (1 scan each)

Scans were selected to allow for comparison across variables of interest while remaining tractable. Each scan required approximately 10–12 hours of expert annotation time.

IV. Experiments & Results

The objectives of our investigations into neural network architectures were twofold: first, we sought a network that could reliably generate annotations across a variety of subjects and configurations; second, we aimed to derive principles to guide future architecture design in similar domains.

Two major factors influenced neural network performance: use of data augmentation and selection of training data. While a full factorial analysis under all possible augmentations and training data sets was intractable, we tested both the augmentation types above and various promising sets of training data (incorporating single or multiple angle conditions, loading conditions, and subjects) in a principled manner, in both isolation and combination, as outlined in Table II. Networks were restricted to using at most 500 slices of manually segmented data to allow fair comparison across training strategies. For networks in which multiple scans of training data were used, these 500 slices were distributed uniformly at random across scans. Up to 1000 slices of total augmented data were used, again distributed uniformly at random across augmented scans, to maintain tractable training time. Of these slices, 5% were reserved for validation at each epoch and 15% for final testing.[‡] Networks were trained for 40 epochs — after which, empirically, they had reliably converged, as shown in Figure 3 — and the epoch with best performance was selected for comparison.

The performance of these models under various conditions is shown in Table II, and example qualitative segmentation data from select architectures are shown in Figure 2.

[‡]Segmentation accuracy on this test data was used internally to confirm that networks were not overfitting; in this paper, we instead report accuracies across new 3D scans not used in training, as generalization across conditions, rather than across slices of the same 3D image, is of primary interest.

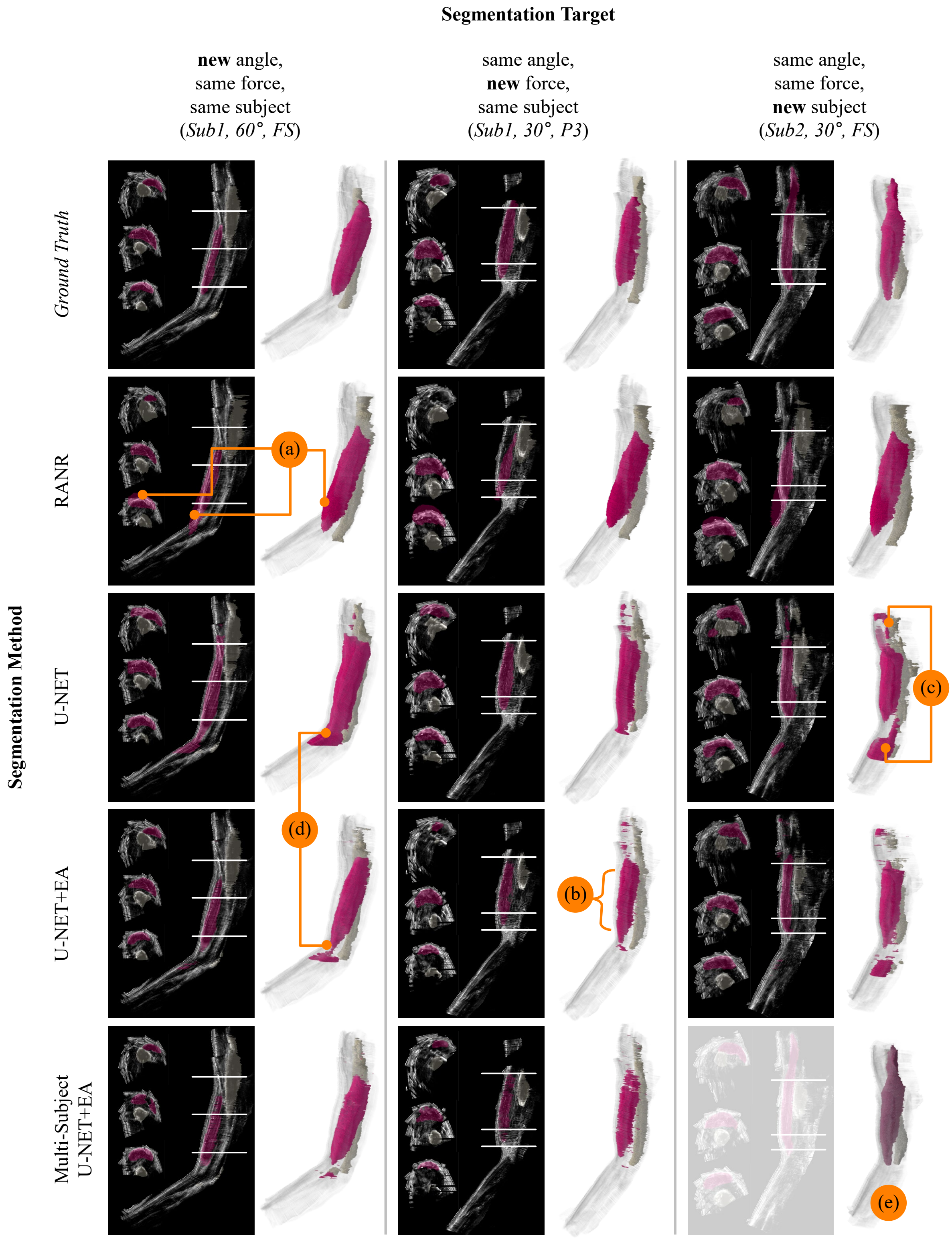

Fig. 2. Exemplar volumetric data, as segmented manually ("Ground Truth"), via optimized classical image registration (RANR), and via neural network, unmodified (U-NET), using elastic deformation data augmentation (U-NET+EA), and using an augmented multi-subject data set (Multi-Subject U-NET+EA). Data used for RANR were ground truth values of (Sub1, 30°, FS); data for training the optimized neural networks are those described in Section IV and Table II. Data were trained, tested, and predicted only on the upper part of the arm, above the elbow; raw lower-arm intensity maps are provided for context. Although superficially smooth and well-formed, RANR segmentation poorly localizes biceps and humerus (a), resulting in poor segmentation accuracy; in contrast, neural network methods perform reliably along the middle section of the biceps (b) but segment more poorly near the ends of the structure (c). Adding elastic deformation augmentation data generally helps smooth the data and improve accuracy (d), though many artifacts remain. *Note that Sub2 Multi-Subject U-NET prediction (e) shows a scan used in network training; its high level of accuracy thus represents network memorization, and the scan is presented for completeness only.*

TABLE II
SEGMENTATION ACCURACY

ARCHITECTURE / STRATEGY	ACCURACY (IoU, *Pixel Accuracy*)			
	same angle, same force, same subject[1]	**new** angle, same force, same subject[2]	same angle, **new** force, same subject[3]	same angle, same force, **new** subject[4]
RR (*rigid registration*)	0.431, *0.980*	0.320, *0.970*	0.537, *0.985*	0.244, *0.953*
RANR (*rigid-affine-nonlinear hierarchical registration*)	0.722, *0.991*	0.545, *0.980*	0.450, *0.980*	0.386, *0.960*
U-NET (*unmodified U-Net [1]*)	0.875, *0.996*	0.593, *0.982*	0.604, *0.988*	0.422, *0.961*
U-NET+RA (*U-Net + rotational augmentation*)	0.929, *0.998*	0.560, *0.984*	0.464, *0.986*	0.393, *0.971*
U-NET+EA (*U-Net + elastic deformation augmentation*)	0.950, *0.999*	0.677, *0.988*	0.573, *0.989*	0.533, *0.978*
U-NET+RA+EA	0.936, *0.998*	0.577, *0.984*	0.544, *0.988*	0.499, *0.978*
Multi-Angle U-NET	0.886, *0.997*	0.691, *0.989*	0.614, *0.989*	0.470, *0.972*
Multi-Angle U-NET+EA	0.906, *0.997*	0.717, *0.989*	0.651, *0.990*	0.523, *0.975*
Multi-Force U-NET	0.885, *0.997*	0.617, *0.985*	0.770, *0.994*	0.452, *0.972*
Multi-Force U-NET+EA	0.902, *0.997*	0.682, *0.988*	0.793, *0.994*	0.519, *0.977*
Multi-Subject U-NET	0.884, *0.997*	0.657, *0.987*	0.536, *0.988*	0.885, *0.995*
Multi-Subject U-NET+EA	0.908, *0.998*	0.687, *0.989*	0.565, *0.989*	0.909, *0.996*

RR, RANR, U-NET, U-NET+RA, U-NET+EA, and U-NET+EA were all trained on (or mapped from) the single Sub1 scan at 30° and FS conditions, with and without augmented data from the same scan(s) as noted. Multi-Angle U-NETs were trained on Sub1 scans at all angle conditions and FS loading; similarly, Multi-Force U-NETs were trained on Sub1 scans at 30° and all force conditions, and Multi-Subject U-NETs were trained on all subjects at 30° and FS conditions. Note that grayed values constitute network "memorization" — i.e., predictions are calculated over data included in network training. These values are presented as baselines for the maximum performance we expect to achieve from a given strategy.
[1] *Accuracy on Sub1 scan at 30° angle and FS loading conditions.*
[2] *Mean accuracy on Sub1 scans at 0°, 60°, and 90° angle and FS loading conditions.*
[3] *Mean accuracy on Sub1 scans at 30° angle and P0, P1, P3, and P5 loading conditions.*
[4] *Mean accuracy on Sub2, Sub3, and Sub4 scans at 30° angle and FS loading conditions.*

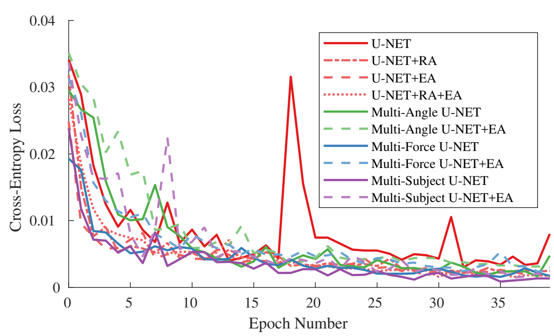

Fig. 3. Training error for all reported network architectures over 40 epochs, over which networks reliably — if sometimes messily — converged. Accuracy values reported in Table II and predictions shown in Figure 2 are computed from the minimum-loss epoch network for each respective architecture.

A. Registration vs. CNN-Based Methods

Quantitatively, as shown in Table II, neural networks perform almost uniformly better than registration-based methods in terms of both intersection-over-union (IoU) values and overall pixel accuracy (calculated as the total fraction of correctly classified voxels). Registration is particularly inadequate when segmenting new subjects. In addition, as shown in Figure 2, the sources of this error are fundamentally different. Registration-based segmentation generates well-formed tissue structures, but these structures are significantly misaligned, so errors are distributed along the length of the scan; in contrast, neural network methods perform well in the central belly of the muscle, but struggle with endpoints, such that errors are concentrated at the top and bottom of the muscle. Manual error correction is thus much more time-efficient on neural-network-segmented scans, as a much smaller subset of slices must be corrected. Alternatively, analysis can be restricted to the center of the muscle with a reasonable expectation of data reliability. This difference in error type also suggests future work in ensemble approaches that incorporate the successful aspects of each strategy.

B. Elastic Deformation Augmentation

The augmentation of the training data set with elastically deformed scans substantially improved performance in most tested cases, suggesting that this method of data augmentation is useful in tissue annotation. In fact, as shown in Table II, a U-Net trained on a single scan along with the elastically-augmented version of that scan (U-NET+EA) performed almost as well on scans at new angles as a network trained on multiple angles (Multi-Angle U-NET). Gains across force conditions and subjects were more modest, but elastic augmentation still showed improved performance as compared with unaugmented counterparts. Qualitatively, elastic augmentation seems to sharpen scans and and remove some artifacts, as shown in Figure 2.

The addition of rotationally augmented data did not improve performance; in fact, performance was diminished across all tested categories, most likely because both test and training scans were collected at similar poses.

C. A Note on Multi-Subject Training and Performance

To effectively scale up muscle deformation analysis requires generalization to new subjects. While neural network segmentation performed reasonably well in this respect, considering the small number of scans seen — even an unmodified U-Net trained on a single scan successfully segmented the middle portion of most subjects' biceps and humerus — performance remains significantly lower than when generalizing across angle or force conditions for the same subject. More explicitly, a network trained on multiple subjects performs better on new angles and force conditions than a network trained on multiple angles or force conditions performs on new subjects;

thus, our objective of multi-subject data sets will be best served in the future by focusing on generation of training data across multiple subjects rather than multiple angle or force conditions. In the future, we will continue to experiment with networks trained on combinations of these variables to further enhance performance.

V. Conclusions & Future Work

The optimized U-Net-based segmentation pipeline described in this work was used to generate the OpenArm 2.0 data set, which has been released to SimTK along with our optimized networks and generating code. Future releases of the data set will be manually cleaned to maintain accuracy along the full length of the tissue structures.

OpenArm 2.0 represents the first multi-subject multi-condition annotated muscle deformation data set, which we will employ in future work to validate existing muscle deformation models — both data-driven and biology-motivated — and to develop new ones. Automating the segmentation process will also allow us to greatly expand the data we examine, which will in the future include multiple muscles under both static and dynamic conditions, as well as greater variety in subjects' age and pathology. We emphasize that this data set is our primary contribution, and we hope others will use it to expand the muscle modeling field.

Acknowledgment

The authors acknowledge the aid and advice of Dr. Gregorij Kurillo, Akira Kato, Jaeyun Stella Seo, David Wang, Sachiko Matsumoto, and Nandita Iyer in both hardware development and data collection; of Jeffrey Zhang and Kireet Agrawal in development of initial U-Net and image processing code; and of Ian McDonald in development of image registration code.

References

[1] O. Ronneberger, P. Fischer, and T. Brox. "U-Net: Convolutional networks for biomedical image segmentation". In: *MICCAI*. Springer. 2015, pp. 234–241.

[2] S. L. Delp et al. "OpenSim: Open-source software to create and analyze dynamic simulations of movement". In: *IEEE Trans. on Biomed. Eng.* 54.11 (2007), pp. 1940–1950.

[3] M. Damsgaard et al. "Analysis of musculoskeletal systems in the AnyBody Modeling System". In: *Simulation Modelling Practice and Theory* 14.8 (2006), pp. 1100–1111.

[4] L. A. Hallock, A. Kato, and R. Bajcsy. "Empirical Quantification and Modeling of Muscle Deformation: Toward Ultrasound-Driven Assistive Device Control". In: *ICRA*. IEEE. 2018, pp. 1825–1832.

[5] S. I. Park and J. K. Hodgins. "Data-driven modeling of skin and muscle deformation". In: *TOG*. Vol. 27. 3. ACM. 2008, p. 96.

[6] R. R. Lemos et al. "Modeling and simulating the deformation of human skeletal muscle based on anatomy and physiology". In: *Computer Animation and Virtual Worlds* 16.3-4 (2005), pp. 319–330.

[7] Q.-H. Zhu, Y. Chen, and A. Kaufman. "Real-time biomechanically-based muscle volume deformation using FEM". In: *Computer Graphics Forum*. Vol. 17. 3. Wiley Online Library. 1998, pp. 275–284.

[8] M. Sagardia et al. "Ultrapiano: A novel human-machine interface applied to virtual reality". In: *ICRA*. IEEE. 2014, pp. 2089–2089.

[9] D. J. Farris, B. D. Robertson, and G. S. Sawicki. "Elastic ankle exoskeletons reduce soleus muscle force but not work in human hopping". In: *Journal of Applied Physiology* 115.5 (2013), pp. 579–585.

[10] E. Passmore et al. "Application of ultrasound imaging to subject-specific modelling of the human musculoskeletal system". In: *Meccanica* 52.3 (2017), pp. 665–676.

[11] L. S. Krivickas et al. "Relationship between force and size in human single muscle fibres". In: *Experimental Physiology* 96.5 (2011), pp. 539–547.

[12] J. Gulati and A. Babu. "Tonicity effects on intact single muscle fibers: Relation between force and cell volume". In: *Science* 215.4536 (1982), pp. 1109–1112.

[13] K. R. Holzbaur et al. "Upper limb muscle volumes in adult subjects". In: *Journal of Biomechanics* 40.4 (2007), pp. 742–749.

[14] G. Ettema, G. Styles, and V. Kippers. "The moment arms of 23 muscle segments of the upper limb with varying elbow and forearm positions: Implications for motor control". In: *Human Movement Science* 17.2 (1998), pp. 201–220.

[15] A. Hill. "The heat of shortening and the dynamic constants of muscle". In: *Proceedings of the Royal Society of London B: Biological Sciences* 126.843 (1938), pp. 136–195.

[16] X. Zhong et al. "Imaging two-dimensional displacements and strains in skeletal muscle during joint motion by cine DENSE MR". In: *Journal of Biomechanics* 41.3 (2008), pp. 532–540.

[17] T. Ungi et al. "Navigated breast tumor excision using electromagnetically tracked ultrasound and surgical instruments". In: *IEEE Trans. on Biomed. Eng.* 63.3 (2016), pp. 600–606.

[18] M.-A. Janvier et al. "A 3-D ultrasound imaging robotic system to detect and quantify lower limb arterial stenoses: In vivo feasibility". In: *Ultrasound in Medicine & Biology* 40.1 (2014), pp. 232–243.

[19] Z. Baum et al. "Visual aid for identifying vertebral landmarks in ultrasound". In: *Medical Imaging 2018: Image-Guided Procedures, Robotic Interventions, and Modeling*. Vol. 10576. International Society for Optics and Photonics. 2018, 105760Z.

[20] A. Lasso et al. "PLUS: Open-source toolkit for ultrasound-guided intervention systems". In: *IEEE Trans. on Biomed. Eng.* 10 (Oct. 2014), pp. 2527–2537.

[21] G. Litjens et al. "A survey on deep learning in medical image analysis". In: *Medical Image Analysis* 42 (2017), pp. 60–88.

[22] J. Zhang et al. "Fully automated echocardiogram interpretation in clinical practice: Feasibility and diagnostic accuracy". In: *Circulation* 138.16 (2018), pp. 1623–1635.

[23] K. Kamnitsas et al. "Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation". In: *Medical Image Analysis* 36 (2017), pp. 61–78.

[24] P. Moeskops et al. "Deep learning for multi-task medical image segmentation in multiple modalities". In: *MICCAI*. Springer. 2016, pp. 478–486.

[25] L. A. Frey-Law et al. "Knee and elbow 3D strength surfaces: Peak torque-angle-velocity relationships". In: *Journal of Applied Biomechanics* 28.6 (2012), pp. 726–737.

[26] A. L. Tamas Ungi and G. Fichtinger. "Open-source platforms for navigated image-guided interventions". In: *Medical Image Analysis* 33 (Oct. 2016), pp. 181–186.

[27] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015.

[28] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: *arXiv:1412.6980* (2014).

[29] M. Z. Alom et al. "Recurrent residual convolutional neural network based on U-Net (r2u-net) for medical image segmentation". In: *arXiv:1802.06955* (2018).

[30] P. Y. Simard, D. Steinkraus, and J. C. Platt. "Best practices for convolutional neural networks applied to visual document analysis". In: *ICDAR*. IEEE. 2003, p. 958.

[31] G. van Tulder. *Elastic deformations for N-dimensional images (Python, SciPy, NumPy, TensorFlow)*. https://github.com/gvtulder/elasticdeform. 2018.

[32] K. Marstal et al. "SimpleElastix: A user-friendly, multi-lingual library for medical image registration". In: *CVPR*. 2016, pp. 134–142.