Hand motion estimation in the wild via IMUs and gaze-tracking: Can an efficient linear model suffice?

1st Eitan Gerber
Faculty of Mechanical Engineering
Technion - Israel Institute of Technology
Haifa, Israel
eitangerber@campus.technion.ac.il

2nd Nili E. Krausz
Faculty of Mechanical Engineering
Technion - Israel Institute of Technology
Haifa, Israel
nili.krausz@me.technion.ac.il

***Abstract*—One of the major current challenges in robotics is the simulation-to-reality (sim2real) gap for training robot controllers particularly due to the limited ability of simulations to recreate real physics. This limitation could be partially mitigated by training controllers on human recordings. Transhumeral prostheses, meant to accurately recreate the motion of the missing part of the limb, could also benefit from such controllers. In this work, we propose a controller for transhumeral prostheses that utilizes a network of sensors worn by the user to enable naturalistic and intuitive control of the prosthesis, and demonstrate proof-of-concept hand motion predictors trained on an open-access dataset, based on a) linear regression and b) a UNet. The training data included 8 subjects, and we tested using a 9th subject. Our results demonstrate that the upper arm position is an important input for hand motion prediction, and adding additional sensor modalities can add benefit. Both methods generalized well to the test subject, yet surprisingly, the UNet's prediction error was only $\sim 24\%$ lower than the efficient linear model's while requiring substantially more computation. Future work will extend this towards a sensor network to help address dynamics and hand kinematics during grasping and manipulation for a proposed trasnhumeral prosthesis controller.**

***Index Terms*—robotics, prostheses, ML, AI, shared control**

I. Introduction

The field of robotics has undergone many changes on its path from science-fiction to semi-ubiquity in diverse fields such as industry, medicine, and driving. A major current challenge in robotics is the transfer from simulations to physical reality (sim2real) [1], as the resulting controllers tend to suffer mainly due to limitations in the ability of existing simulations to recreate the dynamics of the real world. One solution to the sim2real problem is the use of human data for demonstration, particularly in cases where robots are expected to replace humans or to work alongside each-other [2], [3].

A field of robotics that may particularly benefit from training controllers on human data is robotic upper limb prosthetics [4]. Millions of people world-wide have a limb loss or difference [5]. Currently, functionally useful prostheses are most limited for individuals with amputations at the transhumeral level (above the elbow) or higher. Typically, powered (robotic) upper-limb prostheses are controlled using surface electromyography (sEMG) [4], by performing learned voluntary (often unnatural) muscle contractions, or via pattern recognition methods [6]. Clinically available prostheses are presently limited to sequential control, meaning the user can control only a single degree of freedom (DOF) or discrete action at a time - potentially leading to unnatural motions of the prosthesis and compensatory motions.

While several groups have done work towards simultaneous control of a prosthetic hand [7], [8], we believe that true naturalistic motions and intuitive control require the use of additional sensors such as vision/gaze [9], and/or knowledge of proximal kinematics such as shoulder angles [10].

II. A Proposed Controller

We propose a controller that uses a multi-modal sensor network to semi-autonomously control a transhumeral prosthesis during reach-to-grasp tasks and other activities of daily living (ADL). Such a controller would address many of the gaps identified by Marinelli et al. [4] and Gou et al. [11] by reducing the cognitive burden on the user and enabling simultaneous control of multiple DOFs of the prosthesis. As it stands, most work on shared autonomy in upper limb prostheses is focused on grasping [12], though some research has looked at shared control of more proximal joints, such as the elbow [13]. Our proposed controller would be focused on endpoint prediction and naturalistic motion of the entire prosthesis.

The proposed controller would receive inputs from a body-mounted sensor network including sEMG sensors measuring the activity of selected muscles, inertial measurement units (IMU) mounted on the torso and residual limb, a gaze-tracking system with integrated IMU, and possibly one or more RGB cameras, see Fig. 1. In this work we show a proof-of-concept (POC) for the proposed controller, using a reduced set of proximal sensors to create an endpoint predictor and assess the generalizability and performance of two predictor types.

III. Methods

We first searched for a dataset for our POC. While Krausz et al.'s [9] research into multi-modal intent prediction showed promising results, it featured highly constrained and unnatural gaze behaviour and object use. Many other datasets that were surveyed including [2], [3], [14]–[16] had similar constraints such as limited sensor modalities, highly artificial tasks, or missing data. Therefore, for our POC we selected a dataset

This research is funded by the Israel Ministry of Innovation, Science and Technology Grant #0007458.



Fig. 1: Illustration of a transhumeral amputee whose prosthesis uses our proposed controller. The sensor modalities featured are integrated sEMG+IMU units on the arm and shoulder, gaze tracking glasses with an outward-facing RGB camera (with each eye's gaze vector highlighted in red), and an RGB camera on the back of the prosthetic hand.

that includes a multi-modal sensor network and naturalistic ADL tasks, Ma et al.'s Nymeria [17] dataset, included in Gen 1 of Meta's Project ARIA.

The Nymeria dataset includes a wealth of data related to egocentric motion, gaze tracking, video and sound, and highly-detailed segmentation and labeling of the subjects' actions. We limited ourselves to the dataset's full-body IMU-based motion capture (MoCap) and gaze tracking data, with particular focus on data available to subjects with transhumeral amputations. We arbitrarily selected 9 subjects from the Nymeria dataset for analysis, due to storage limitations (9 subjects required $\sim 1.6[TB]$). The datastreams we picked were:

- **Hand**: Left & right hand coordinates. These were the "targets" for our endpoint predictor.
- **Shoulder**: Left & right upper-arm coordinates.
- **Elbow**: Left & right forearm coordinates.
- **SA**: Left & right 3D shoulder joint angles.
- **GV**: The central gaze vector.
- **GPY**: Gaze pitch and yaw angles.

From these we also computed:

- **AP**: Arm Position, mean shoulder and elbow coordinates.
- **AV** and **AA**: Arm Velocity and Arm Acceleration.
- **SAV** and **SAA**: Shoulder Angular Velocity and Shoulder Angular Acceleration.

The body kinematics datastreams were sampled at $240[Hz]$ by the XSENS MVNX IMU-based MoCap suits used, and the gaze-related datastreams were upsampled from $\sim 10[Hz]$ to $240[Hz]$ using nearest-neighbor resampling. The hand (Hand) and arm coordinates (Shoulder, Elbow) and gaze vector (GV) are all 3D vectors with the gaze vector specifically being a unit vector, and the gaze pitch/yaw (GPY) angles are scalar values measured in radians. As provided, Hand, Shoulder, and Elbow are in a static "world" coordinate system, yet we wanted to use these in a dynamic coordinate system to track the subject. Thus, prior to use we subtracted from them the (provided) pelvis coordinates, given by Nymeria as the "root" of the person. The recorded datafiles were all synchronized and resampled with a Python script that used Project Nymeria's native tools, and exported as .csv and .h5 files. An important stage of the preprocessing was finding the mean of the upper-arm and forearm coordinates at every timestamp, as that mean describes a point near the middle of the upper arm - we refer to this point as the arm position (AP). We also compute the velocity and acceleration of the AP and SA, by taking the 1^{st} or 2^{nd} order differentials of the joint angles and coordinates, as relevant. We refer to each item from the above list (except Hand, Shoulder and Elbow) as a 'predictor group'.

We chose to test the various datastreams' predictiveness of the hand coordinates using two approaches: a) linear regression (LR), and b) a modified UNet [18], to assess whether a simple linear approach performs as well as a commonly used timeseries network. For the LR we wrote a MATLAB script using the *fitlm()* function (from the Statistics and Machine Learning Toolbox), to select the desired hand (left/right/both) and which predictors. As the model returned by the function only predicts one variable, our script runs three times for each target defined - once per X,Y,Z axis. We learned and tested $255 \cdot 6 = 1530$ linear regressors, utilizing different predictor group combinations to predict the XYZ coordinates of both hands (6 linear regressors per predictor group combination).

For the UNet, we replaced the choke-point at the bottom of the "U" with a 2-layer Gated Recurrent Unit (GRU) to "inject" a temporal dependency into the model. Additionally, the model ensures causality by employing left-side zero-padding in the convolutional layers and by replacing the standard pooling layers with strided causal convolutions. During the upsampling stage of the model, we make use of nearest-neighbor interpolation (instead of convolutions) to maintain causality. As this is a UNet based model, we trained it sequence-by-sequence on 180-timestep windows with a stride of 9, for a 95% overlap between windows. While training the model we used a loss function formulated to promote natural kinematics, using a weighted sum of the hand position Euclidean error, the velocity loss (stepwise velocity of the prediction VS the ground-truth) to enforce smooth natural motion, and the jerk loss to comply with the principle of minimum jerk [19].

In terms of the data split for learning our models, we used the recordings from 8 subjects for the train/val dataset and the recordings from the 9^{th} subject as the test dataset. Between epochs of the UNet's training we iterated which of the 8 subjects is used for validation, to avoid overfitting the model and maintain generalizability. The 9^{th} (test) subject was left out of the training process, used only to evaluate the final model's inference accuracy and speed. Each subject has 8 recording of $\sim$800 to $\sim$1000 seconds, for a per-subject total of roughly two hours of recorded data, or about 18 hours of motion data used in this work. This work included 4 male / 5 female subjects, 4 aged 31-35, 4 aged 25-30, and one subject aged 18-24, with a mean height of $175.11 \pm 8.97[cm]$ and a mean bodyweight of $71.44 \pm 11.92[kg]$.

IV. Results

We assessed the performance of the linear regressors trained on a single predictor group, see Table I, and note that most individual predictor groups other than AP lead to an MAE

TABLE I: MAE $\pm$ SD overall and per-axis, comparison between different LR predictor sets. AP=Arm Position, AV=Arm Velocity, AA=Arm Acceleration, GV=Gaze Vector, GPY=Gaze Pitch&Yaw, SA=Shoulder Angle, SAV=Shoulder Angular Velocity, SAA=Shoulder Angular Acceleration. The gaze vector is a unit-vector, all other columns have units as listed.

Axis	**AP** $[m]$	**AV** $[m/s]$	**AA** $[m/s^2]$	**GV**	**GPY** $[rad]$	**SA** $[deg]$	**SAV** $[deg/s]$	**SAA** $[deg/s^2]$
Euc.	0.203 ± 0.025	0.346 ± 0.050	0.345 ± 0.050	0.346 ± 0.051	0.346 ± 0.051	0.332 ± 0.046	0.346 ± 0.050	0.346 ± 0.050
X	0.090 ± 0.009	0.192 ± 0.023	0.191 ± 0.023	0.192 ± 0.022	0.192 ± 0.022	0.192 ± 0.024	0.192 ± 0.023	0.192 ± 0.023
Y	0.087 ± 0.006	0.160 ± 0.038	0.159 ± 0.038	0.160 ± 0.039	0.160 ± 0.039	0.161 ± 0.035	0.160 ± 0.038	0.160 ± 0.038
Z	0.130 ± 0.027	0.178 ± 0.050	0.178 ± 0.050	0.177 ± 0.050	0.176 ± 0.051	0.150 ± 0.042	0.178 ± 0.050	0.178 ± 0.050

in the $0.346 \pm 0.051[m]$ range. Moreover, as seen in Fig. 2 when combining predictor groups there was no meaningful change to the prediction accuracy when neglecting the AP group. The mean 3D MAE without the AP predictor group ranged from $0.344 \pm 0.055[m]$ to $0.331 \pm 0.047[m]$ with 1-7 predictor groups. In contrast, when using the AP group alone we obtain a mean MAE of $0.203 \pm 0.025[m]$, and using all 8 groups together produced a mean MAE of $0.186 \pm 0.019[m]$. Clearly, AP is the single most predictive group, yet a $\sim 8\%$ improvement in accuracy can be obtained by appending other predictor groups to AP.

We evaluated the performance of the modified UNet predictor after training for 152 epochs, of which the final 25 showed no increase in validation accuracy, leaving us with the model weight saved after the 127^{th} epoch. We obtained an overall MAE of $0.142 \pm 0.080[m]$, a $\sim \%24$ improvement in hand motion prediction VS the best linear regressor.

To better understand the source of hand prediction error, we consider an example recording from the test dataset as seen in Fig. 3. We note that the MAE for the linear regressor for the two hands appears to be somewhat correlated, such that when the error for the left hand increases often so does the error for the right hand. In Fig. 3a we identify two regions with different error levels. We see in Fig. 3b that when the subject was using their hands to hang a streamer while directly looking at it (at the recording's $289.2[sec]$ mark) the error was significantly lower than while the person was using their hands to bunch-up a cloth while focusing on a different person's actions (at the recording's $330[sec]$ mark), as seen in Fig. 3c.

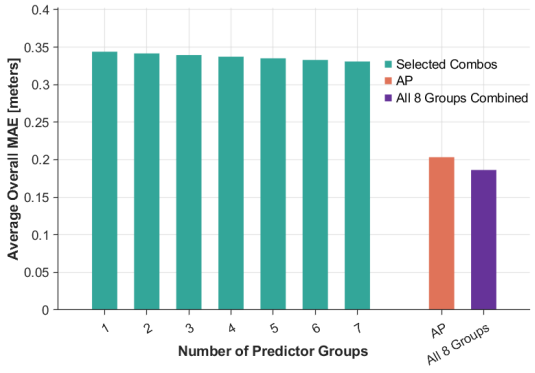
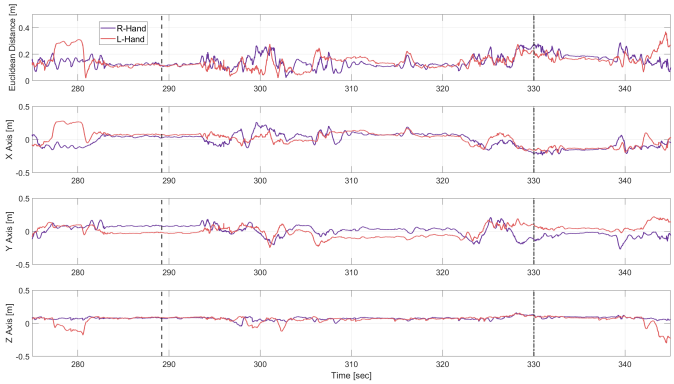
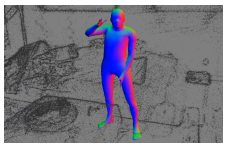


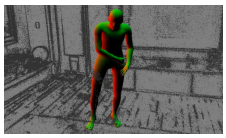
Fig. 2: Overall average 3D MAE produced by LR with all possible combinations of predictor groups other than the arm position (AP) vs. the number of predictor groups used together. On the right-hand side of the graph we show mean 3D MAE using AP group alone, vs. all 8 predictor groups.



(a) Euclidean and per-axis distance between LR predicted and recorded hand positions. The dashed line (left) is at t = $289.2[sec]$ and the dot-dashed line (right) is at t = $330[sec]$.



(b) Estimated 3rd-person pose at $289.2[sec]$.



(c) Estimated 3rd-person pose at $330[sec]$.

Fig. 3: Distance graphs for a representative portion of the test data, specifically 275-345s of the "*20231208_s0_ronald_guerra_act5_y2gscm*" recording [17] showing segments with higher and lower LR prediction accuracy, corresponding to the hands being used to hang a streamer and the hands being used to bunch-up a cloth.

V. Discussion

In this work we present results for hand motion prediction during naturalistic behavior based on motion of the upper arm and gaze tracking, for POC of our proposed prosthesis controller. We trained on recorded data from 8 subjects, and tested on data from a single novel subject. Our results show good prediction of hand position based on upper arm position, with errors of $\sim 19[cm]$ across multiple hours of data with a linear regressor. While not ideal, the predictor generalized well from the training subjects to the test subject and between unstructured ADLs, and we show that by adding sensor modalities we can marginally improve performance, as compared to using arm position alone. When using the UNet-based predictor our results had $\sim \%24$ lower errors, at the cost of significantly increased training time and computation - over 24 hours to train the UNet VS minutes to train the LR, both on the same PC. The per-frame prediction times

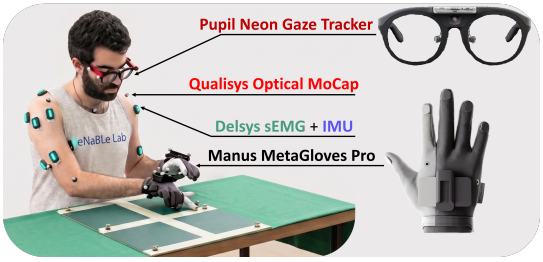


Fig. 4: New multi-modal sensor network, seen during a gaze tracking task. Not shown, we use an OT Bioelettronica (Italy) HD-sEMG system to measure the activity of forearm muscles.

followed a similar trend, with the UNet needing a mean of $1.49[ms]$ while the LR needed a mean of $0.54 \cdot 10^{-3}[ms]$. This presents a potential trade-off whereby selection of the right sensor modalities may offset the need for a complex model and reduce the computational costs.

Our results suggest that while Project Nymeria's available data forms a strong basis for such a predictor, additional sensor modalities are indeed needed to improve prediction accuracy. For example, Krausz et al. [9] collected multi-modal (gaze tracking, sEMG, MoCap) recordings of subjects performing highly "artificial" tasks, while in Project Nymeria [17] the relevant data is limited to gaze tracking and MoCap but the subjects are performing ADL in highly realistic scenarios. However, this comes with challenges in understanding where/how our predictor struggles or fails, as the realistic scenarios are highly unstructured. An additional limitation may be posed by the specifics of the different datasets, such as Project Nymeria's gaze-tracking being limited to a sample rate of $\sim 10[Hz]$. Moreover, the lack of force/torque sensing is a critical limitation in most human datasets for robotics, where closing the sim2real gap relies on knowledge of joint and task dynamics. Additionally, while current prostheses leverage biosignals such as sEMG that can provide user intent [20], Project Nymeria does not include such data.

To bridge the gap between these datasets we are collecting a multi-modal dataset, consisting of sEMG, IMUs, optical MoCap, MoCap gloves, and gaze tracking (Pupil Neon, Pupil Labs GmbH, Germany) as demonstrated in Fig. 4. Subjects will perform a number of ADLs along with tasks from standardized assessments [21], which we believe will strike a good balance between overly structured "artificial" tasks and tasks so unstructured that generalized conclusions can't be drawn from the recordings.

Importantly, our dataset will include both sEMG and ground truth hand kinematics based on MoCap gloves (Metagloves PRO, Manus, The Netherlands), which can be leveraged for both grasping and dexterous manipulation, even during occluded tasks where optical mocap solutions might fail. We do not directly capture force data, but amplitude of sEMG is highly correlated with muscle effort [20] which could be a useful proxy for the force/torque exerted (for solving sim2real) as well as providing the *intent* for *intent prediction*.

Our results demonstrate that computationally efficient LR for hand motion prediction, if trained with the right dataset, performs almost to the level of a deep neural network, suggesting the notion that simplicity may at times compensate for reduced accuracy. In the context of an amputation with no distal signal remaining, the ability to predict hand position from upper-arm kinematics holds potential for our ultimate goal of naturalistic hand motion prediction for dexterous manipulation and naturalistic control of a prosthetic arm, based on intelligent use of body-based sensing.

References

[1] E. Aljalbout *et al.*, "The reality gap in robotics: Challenges, solutions, and best practices," *Annu. Rev. Control. Robot. Auton. Syst.*, vol. 9, pp. 403–432, May 2026.
[2] P. Kratzer *et al.*, "Mogaze: A dataset of full-body motions that includes workspace geometry and eye-gaze," *IEEE Robot. Autom. Lett.*, vol. 6, pp. 367–373, Apr. 2021.
[3] R. Ren *et al.*, "Enhancing robotic skill acquisition with multimodal sensory data: A novel dataset for kitchen tasks," *Sci. Data*, vol. 12, p. 476, Mar. 2025.
[4] A. Marinelli *et al.*, "Active upper limb prostheses: a review on current state and upcoming breakthroughs," *Prog. Biomed. Eng.*, vol. 5, p. 012001, Jan. 2023.
[5] C. L. McDonald *et al.*, "Global prevalence of traumatic non-fatal limb amputation," *Prosthet. Orthot. Int.*, vol. 45, p. 105, Apr. 2021.
[6] I. Kyranou, S. Vijayakumar, and M. S. Erden, "Causes of performance degradation in non-invasive electromyographic pattern recognition in upper limb prostheses," *Front. Neurorobot.*, vol. 12, Sep. 2018.
[7] M. Nowak *et al.*, "Simultaneous and proportional real-time myocontrol of up to three degrees of freedom of the wrist and hand," *IEEE Trans. Biomed. Eng.*, vol. 70, pp. 459–469, Feb. 2023.
[8] J. M. Hahne *et al.*, "Simultaneous control of multiple functions of bionic hand prostheses: Performance and robustness in end users," *Sci. Robot.*, vol. 3, p. eaat3630, Jun. 2018.
[9] N. E. Krausz *et al.*, "Intent prediction based on biomechanical coordination of EMG and vision-filtered gaze for end-point control of an arm prosthesis," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 28, pp. 1471–1480, Jun. 2020.
[10] A. Akhtar, L. J. Hargrove, and T. Bretl, "Prediction of distal arm joint angles from EMG and shoulder orientation for prosthesis control," in *Proc. IEEE EMBC*, Aug. 2012, pp. 4160–4163.
[11] W. Guo *et al.*, "Toward human-in-the-loop shared control for upper-limb prostheses: A systematic analysis of state-of-the-art technologies," *IEEE Trans. Med. Robot. Bionics*, vol. 5, pp. 563–579, Aug. 2023.
[12] J. Mouchoux *et al.*, "Impact of shared control modalities on performance and usability of semi-autonomous prostheses," *Front. Neurorobot.*, vol. 15, Dec. 2021.
[13] M. Merad, E. . de Montalivet, and N. Jarrasse, "Can we achieve intuitive prosthetic elbow control based on healthy upper limb motor strategies?" *Front. Neurorobot.*, vol. 12, Feb. 2018.
[14] D. Di Domenico *et al.*, "Reach&grasp: a multimodal dataset of the whole upper-limb during simple and complex movements," *Sci. Data*, vol. 12, p. 233, Feb. 2025.
[15] M. V. Liarokapis *et al.*, "Learning human reach-to-grasp strategies: Towards EMG-based control of robotic arm-hand systems," in *Proc. IEEE ICRA*, May 2012, pp. 2287–2292.
[16] B. Lento *et al.*, "3d-arm-gaze: a public dataset of 3d arm reaching movements with gaze information in virtual reality," *Sci. Data*, vol. 11, p. 951, Aug. 2024.
[17] L. Ma *et al.*, "Nymeria: A massive collection of multimodal egocentric daily motion in the wild," in *Proc. ECCV*, 2025, pp. 445–465.
[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. MICCAI*, 2015, pp. 234–241.
[19] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," *J. Neurosci.*, vol. 5, pp. 1688–1703, Jul. 1985.
[20] P. Konrad, *The ABC of EMG*. Noraxon, 2006.
[21] L. Resnik *et al.*, "Development and evaluation of the activities measure for upper limb amputees," *Arch. Phys. Med. Rehabil.*, vol. 94, pp. 488–494.e4, Mar. 2013.