Robust Multi-User In-Hand Object Recognition in Human-Robot Collaboration Using a Wearable Force-Myography Device

Eran Bamani, Nadav D. Kahanowich, Inbar Ben-David and Avishai Sintov

***Abstract*— Applicable human-robot collaboration requires intuitive recognition of human intention during shared work. A grasped object such as a tool held by the human provides vital information about the upcoming task. In this paper, we explore the use of a wearable device to non-visually recognize objects within the human hand in various possible grasps. The device is based on Force-Myography (FMG) where simple and affordable force sensors measure perturbations of forearm muscles. We propose a novel Deep Neural-Network architecture termed *Flip-U-Net* inspired by the familiar U-Net architecture used for image segmentation. The Flip-U-Net is trained over data collected from several human participants and with multiple objects of each class. Data is collected while manipulating the objects between different grasps and arm postures. The data is also pre-processed with data augmentation and used to train a Variational Autoencoder for dimensionality reduction mapping. While prior work did not provide a transferable FMG-based model, we show that the proposed network can classify objects grasped by multiple new users without additional training efforts. Experiment with 12 test participants show classification accuracy of approximately 95% over multiple grasps and objects. Correlations between accuracy and various anthropometric measures are also presented. Furthermore, we show that the model can be fine-tuned to a particular user based on an anthropometric measure.**

I. Introduction

To develop a natural Human-Robot Collaboration (HRC) system, it is necessary that the robot unambiguously perceive the task carried out by a human. For this purpose, an object within a human hand provides significant information about the upcoming task and enables a substantial reduction in the set of possible actions that the human might perform. With such information, the robot can deliver a complementary object or plan an assistive trajectory [1]. For instance, a robotic arm can handover objects or assist an upper-limb amputee in completing dual-arm tasks. Examples for such scenarios are seen in Figure 1.

Signaling a robot of a desired assistive task in HRC has been widely researched. Nonetheless, common HRC approaches bear unnatural control methods including, for example, sensing brain activities [2] or human gestures [3]. These require human pre-training and may hinder the work flow. Vision has also been employed to identify objects in hand [4]. However, such approach requires a direct line-of-sight with the working area while the grasped object and its usage may be occluded. A different approach is

This work was supported by the Israel Science Foundation (grant No. 1565/20).

E. Bamani, N. D. Kahanowich, I. Ben-David and A. Sintov are with the School of Mechanical Engineering, Tel-Aviv University, Israel. e-mail: eranbamani@mail.tau.ac.il.

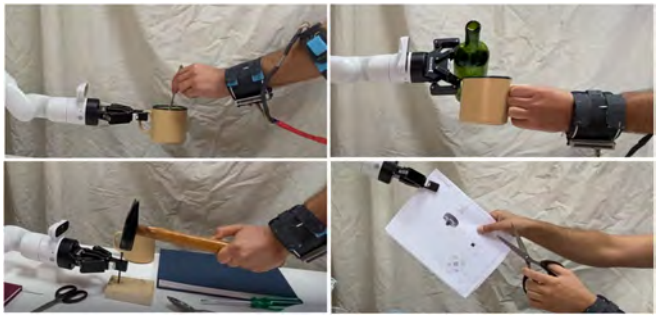

Fig. 1. A robotic arm assisting in dual-arm tasks or handing-over complementary objects to a human user based on observations from an FMG wearable device.

Electro-myography (EMG) where electrical muscle signals are sensed and mapped to limb movements [5]. The accuracy of the EMG technology, however, is commonly compromised by sweat, electrode placement and crosstalk [6]. An alternative solution is Force-myography (FMG) which non-invasively measures perturbations of the musculoskeletal system and has been proposed for user intention recognition in HRC [7]. FMG utilize low-cost force sensors and a simple acquisition device. A recent comparative work has shown that FMG outperforms EMG both in gesture recognition and regression-based control [8].

Early work by Amtf et al. [9] has introduced pattern identification of forearm muscle activities using body-worn force sensors. Since then, FMG signals were shown to be simple to acquire with a relatively high-accuracy. Consequently, FMG was used in data-based classification of hand gestures [10]–[12]. However, a classifier was retrained using new collected data when replacing the sensors on the arm. Hence, once the sensors have been removed and replaced, the trained classifier cannot be reused. Naturally, the classifier would also not be transferable to a different user. Furthermore, another comparison between FMG and EMG has shown that FMG is less sensitive to positioning variations and does not require direct contact with the skin [13].

A work by Gigli et al. [14] uses visual perception of objects to augment surface EMG classification of future grasps for a prosthetic hand. A different work has shown the ability to classify types of grasps rather than specific objects using FMG [15]. Recent work by Kahanowich and Sintov [16] on FMG-based object recognition has proposed a data-based iterative algorithm for robust recognition of grasped objects. The method used a low-cost wearable device with 15 force-sensitive resistors (FSR). The device was used to collect data and train a classifier to recognize in-hand objects.

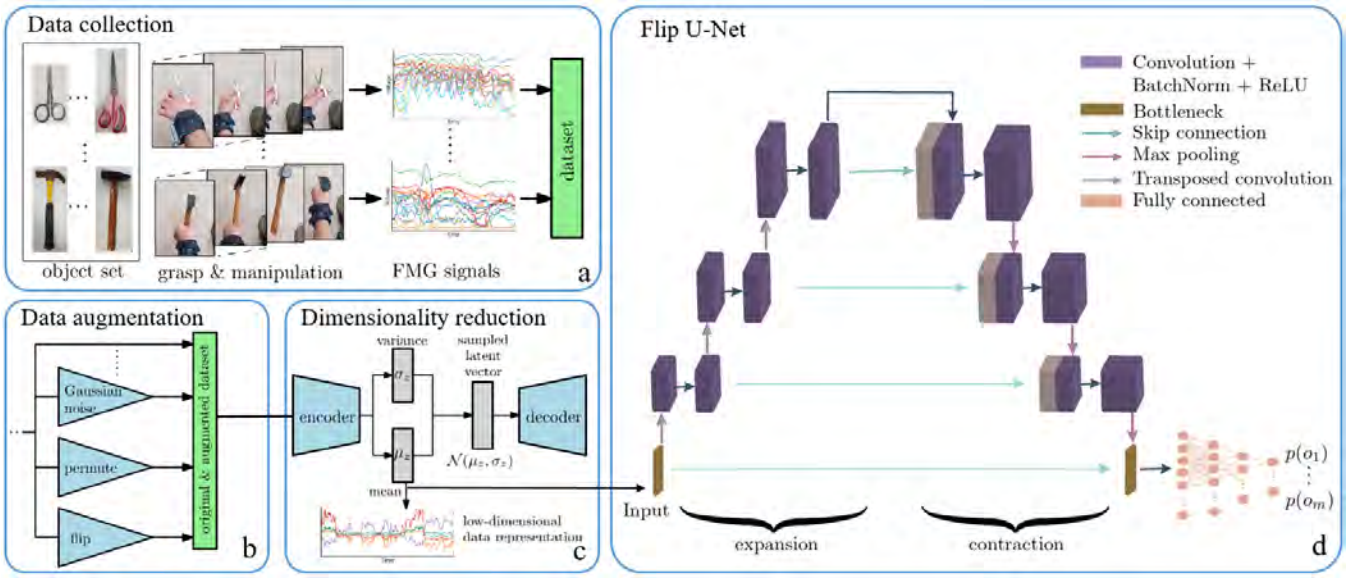

Fig. 2. The proposed method for data treatment and model training for object recognition using FMG measurements: (a) Data collection process in which multiple human subjects record various grasps of different everyday objects; (b) data augmentation; (c) dimensionality reduction using a Variational Autoencoder (VAE). Reduced representation of the data is taken from the mean of the latent space; and, (d) the proposed Flip-U-Net architecture to classify processed FMG signals to grasped objects.

Using an iterative algorithm, the approach was shown to be accurate while robust to re-positioning of the device, i.e., once the classifier has been trained over collected data, it maintains its accuracy even if the device has been taken off previously. Nevertheless, the classifier was trained solely on a single user and did not exhibit a transferable property where different humans can use it accurately. In addition, a classifier was trained on one specific task-based grasp of each object while different grasps of the object would most likely lead to classification failure.

In this work, we explore the use of FMG measurements in order to recognize grasped objects across a wide number of users while also robust to positioning variations. We utilize the same wearable FMG device introduced in [16]. Unlike the previous work by Kahanowich and Sintov which relied on single-user and single-task-based grasp for each object, we propose a multi-user classification approach where an object can be recognized according to various possible grasps. Furthermore, an object class does not include a single object whereas the classifier is able to identify untrained objects from the same class, e.g., identify drinking cups or scissors of different shapes and sizes. In the proposed approach, labeled data collected from multiple human participants is augmented and used to train a Variational Autoencoder (VAE) [17] for dimensionality reduction mapping. Furthermore, the processed data is used to train a novel Deep Neural-Network (DNN) architecture, termed *Flip-U-Net*. The proposed approach for training a robust multi-user classifier is illustrated in Figure 2.

Flip-U-Net is based on the U-Net originally proposed for segmentation of bio-medical images [18]. U-Net evolved from the traditional Convolutional Neural Network (CNN) and focuses on semantic segmentation of images. The U-Net architecture consists of two symmetric paths which give it the U-shaped architecture: contraction and expansion paths. These two paths enable context recognition and precise localization, respectively. We propose a flipped version of the U-Net where signals are expanded and then contracted. Such approach enables the estimation of the signal structure and extracts vital information for object recognition. Hence, Flip-U-Net is shown to exploit, along with a VAE, dominant information in a low-level non-visual application and provide accurate classification.

The proposed approach is analyzed based on the anthropometric measures of the users. We explore the effect of various anthropometric measures of a user on accuracy. While the overall accuracy of the model is shown to be high, we show the benefit of fine-tuning the model with data from training data associated to a subject with a similar anthropometric measure. Hence, the model can reach better accuracy on a new user when tuned with specific part of the training data.

The wearable FMG device and proposed method could enable intuitive and non-verbal communication with a robotic assistant. The robot is first informed of the object in the human hand which, in turn, implies future tasks. Hence, the robot can infer about future actions of the human beforehand and plan a trajectory accordingly. Since the method is non-visual, there is no dependency on a direct line-of-sight or lighting. This work leads the way for a robot to be able to observe human motion, recognize the task and promptly act to assist. By doing so in real-time during motion, a robotic arm will be able to efficiently interact with a human to complete a shared task.

II. Methods

A. FMG data acquisition device

We base our work on a wearable FMG device presented in previous work [16]. The system is based on 15 low-cost

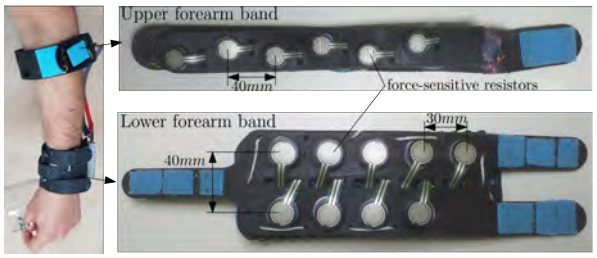

Fig. 3. FMG wearable device made of two parts for the upper and lower forearm. Each part includes a set of force-sensitive resistors (FSR) designed to sense perturbations of the musculoskeletal system.

Force-Sensitive Resistors sensors (FSR), model FSR-402 by Interlink Electronics. FSR sensors are made of polymer films that vary their electrical resistance with change in surface pressure. It has been shown that having a wide coverage along the forearm yields high accuracy. Hence, the device consists of an upper forearm band with six FSR sensors and a lower forearm (wrist) band with nine sensors organized in two rows as seen in Figure 3. The bands were fabricated by 3D printing with an elastic polymer (Thermoplastic polyurethane) and include a flexible bulge for each sensor to ensure proper attachment to the skin. This design provides flexibility during arm motion.

The system also includes a data acquisition system based on an Arduino Mega 2560 board. The FSR sensors are connected to the analog pins of the Arduino through a voltage divider of 4.7 kΩ resistor. The system provides real-time data stream of all the given sensors in a frequency of up to 300 Hz. The described system is composed of low-cost (the prototype costs approximately \$150) and light-weight (0.23 Kg) hardware which is appealing and suited for easy arm movements.

B. Data collection

We aim to identify an object grasped by any adult human user solely by measuring FMG signals by the device. Given a set of m object classes $\{\mathcal{O}_1, \dots \mathcal{O}_m\}$, we require to identify an object class from the set. That is, we require robust real-time classification based on pattern recognition of the input signals. This is achieved through supervised learning trained on diverse data collected from k human subjects, from a variety of objects of the same class (e.g., scissors of various sizes and shapes) and from multiple re-positioning of the device on the arm. A large amount of participants adds variance to the data due to differences in arm thickness, length and body fat.

Let $\mathbf{x}_t \in \mathbb{R}^{15}$ be the observable state of the musculoskeletal system measured by the FMG system with 15 FSR sensors. For each object class $\mathcal{O}_i$, training data is collected by holding several items of the same class. Figure 4 shows an example of classes of scissors and drinking cups with several items sampled during the experiments. This exposes the model to a wider number of variants from each object class. The data collection process includes a large number of grips and manipulations on each object as seen in Figure 2a. For each human subject, the collection process is composed of M episodes where, in each episode and between objects swapping, the FMG device is taken-off and re-positioned. To cope with different tightening forces at each episode, we consider episode values relative to the initial forces after strapping in. During each episode, N_e samples are recorded while the human manipulates the object and regrasps it. Ultimately, the resulting training data is a set of N labeled FMG signals $\Phi = \{(\mathbf{x}_1, l_1), \dots, (\mathbf{x}_N, l_N)\}$ where label l_i corresponds to object $\mathcal{O}_{l_i}$.

C. Data augmentation

We employ data augmentation to increase diversity in the collected training data and improve the generalization ability of an Artificial Neural-Network (ANN) model, without collecting new data. Data augmentation techniques such as scaling, padding and flipping are commonly used for generating new realistic samples from the true data distribution prior to training a DNN [19]. Augmentation is performed without altering the class label. Furthermore, these techniques have been shown to be effective for reducing overfitting [20], [21] and can help networks overcome small datasets [22] or datasets with imbalanced classes [23], [24].

Data augmentation techniques are usually employed on image datasets while our FMG signals are unidimensional. Hence, one dimensional data augmentation is yet to be well defined. In this paper, we formulate seven data augmentation operations on the recorded data where four are signal-wise. *Jittering* is the addition of noise of some distribution to the signal. We include two jittering variants: Exponential and Gaussian noise. *Scaling* is the multiplication of the signal by a scalar $q > 0$. Furthermore, *Flipping* is vertical flip of the signal vector (for a vertical vector). The remaining data augmentation operations are applied to an entire recorded episode. With *Permutation*, we rearrange segments of an episode in order to produce a new pattern. Within a window sliding along the episode, we slice the data into equal length slices and randomly permute the slices to generate a new window. Next, *Rotation* rotates the $15 \times N$ data array of the episode by a specified degree. Lastly, in *Magnitude-Warping*, we change the magnitude of each signal by convolving the data episode with a smooth curve varying with a normal distribution around one. For each signal $\mathbf{x}_t$, we include in the training dataset all augmented variants and the original signal $\mathbf{x}_t$, as seen in Figure 2b.

D. Dimensionality Reduction

Dimensionality Reduction (DR) is the process of reducing the number of input features in a dataset. The reduction removes redundant features and noise from the training data which, in turn, leads to improvement in model accuracy. An Autoencoder (AE) is a feed-forward ANN that learns efficient encoding of data in an unsupervised manner [25], [26]. AE consists of an encoder and a decoder. The encoder is used to capture key information from the data. It takes the input data and compresses it to produce a latent representation $\mathbf{z} \in \mathbb{R}^d$. The decoder takes the compressed data and decompresses it to reconstruct the original data. The latent

representation must be of lower-dimension than the input so that the AE cannot simply learn the identity function. The encoder and decoder can learn more complex operations than projection and linear combinations, and capture non-linear features of the data. AE is normally trained to reconstruct the input $\mathbf{x}$ by minimizing the objective function $\|\mathbf{x} - \hat{\mathbf{x}}\|^2$ where $\hat{\mathbf{x}}$ is the output of the decoder. Variational Autoencoder (VAE) is a generative variant of AE [17]. The training of VAE is regularised to prevent overfitting and to ensure the ability to generate new data in the latent space. While AE encodes the input as a single low-dimensional vector, VAE encodes it as a distribution over the latent space. The input $\mathbf{x}$ is mapped onto a distribution $Q(\mathbf{z}|\mathbf{x})$ and the latent space is sampled from it, i.e., $\mathbf{z} \sim Q(\mathbf{z}|\mathbf{x})$. The latent layer $\mathbf{z}$ of the AE is replaced with a multivariate Gaussian distribution including two sets of layers: one representing the mean μ_z in each of the dimensions of the latent space and one representing the variance σ_z such that $\mathbf{z} \sim \mathcal{N}(\mu_z, \sigma_z)$. Therefore, $\mathbf{z}$ is sampled from the distribution during the training and passed to the decoder to generate $\hat{\mathbf{x}}$. VAE is trained to minimize the evidence lower bound function which is the sum of the AE reconstruction loss and the Kullback-Leibler (KL) divergence between the latent distribution of the input $Q(\mathbf{z}|\mathbf{x})$ and the prior $P(\mathbf{z}) \sim \mathcal{N}(0, \mathbf{I})$. Hence, the objective function is given by

$$loss = \|\mathbf{x} - \hat{\mathbf{x}}\|^2 + D_{KL}(Q(\mathbf{z}|\mathbf{x})\|P(\mathbf{z})). \tag{1}$$

The KL divergence between two probability distributions P and Q is a measure of the difference between the two distributions and is defined as $D_{KL}(Q\|P) = \sum Q \log \frac{Q}{P}$.

We use the VAE to reduce the dimension of the training data, as illustrated in Figure 2c, prior to the training of a classifier network. The VAE is trained with the original and augmented data while minimizing (1). While we sample from the latent distribution $\mathbf{z} \sim \mathcal{N}(\mu_z, \sigma_z)$ for training the VAE, we use only the mean μ_z for generating lower-dimensional data. Hence, the trained encoder is used to map labeled data into a lower-dimensional space prior to using the Flip-U-Net.

It is important to note that, in AE, we have no control on the resulted distribution of the data in the latent space. AE focuses on the reconstruction of the input while compressing it at the bottleneck to a latent space in some underlying and unknown distribution. Thus, AE provides no guarantees on the organization of the latent space and is highly dependent on the distribution of the input data and on network architecture. Consequently, the latent space can lack regularity and an exploitable structure. On the other hand, VAE enforces the distribution of the latent space to be close to a standard normal distribution. Hence, this ensures regularization of the latent space and leads to better performance in the classification task. During the training of the VAE, the dimension of the latent space is forced to be $d < 15$. Preliminary experiments have shown the ability of VAE to properly embed the dominant features of the FMG signals and ease the training of a classifier, compared to other DR methods including AE. We include further analysis in the experimental section.

E. Flip-U-Net

The augmented and dimension reduced data is now ready for training a classifier. As presented in Section I, we propose the use of a modified U-Net network termed Flip-U-Net. The architecture of U-Net is symmetric and consists of contraction and then expansion parts. The contraction part consists of a repeated series of convolutions, rectified linear units (ReLU), max-pooling and down-sampling operations. Spatial information is lost in both convolutional and down-sampling operations. Hence, the expansion is an inverted path with convolutions and up-sampling operations. These operations help in inverting and compensating for the loss of spatial resolution. Furthermore, U-Net introduced skip-connections between the contraction and expansion which enable both parts to share information. Moreover, the benefit of using convolutional layers in the U-Net is in the ability to learn from sequential raw data directly without the need for manual feature engineering by an expert.

Unlike U-Net, Flip-U-Net consists of expansion of the data followed by contraction taking the shape of a flipped *U* as seen in Figure 2d. The architecture starts with a bottleneck which is the output vector of the VAE. Next, the expansion consists of three blocks. Each block passes the input to two 1×3 convolution layers, batch normalization and a ReLU. Each block doubles the number of filters and is followed by a 1×2 up-sampling layer implemented as transposed convolution. The contraction part is made of three blocks. Each block passes the input through two 1×3 convolution layers, batch normalization and ReLU, followed by down-sampling. The down-sampling is implemented by a 1×2 max-pooling operation with kernel size and stride 2 assisting in more accurate extraction of key features. Similar to the U-Net, skip-connections connect between the two parts such that each block in the expansion part is concatenated with the corresponding one in the contraction part. Skip-connections assist in recovering spatial information lost during the down-sampling. In addition, they ensure that features learned in the expansion are used in the reconstruction. Furthermore, skip-connections provide additional paths for the gradient and are often beneficial for model convergence. The above architecture is symmetric such that the number of expansion blocks is the same as of the contraction.

During training, sequential array batches of size $d_s \times d$ are fed into the Flip-U-Net, where d_s is the number of FMG samples included in the batch. When testing with a single query sample, d_s is equal to 1. Each signal at the output of the contraction part is flattened into the bottleneck and also concatenated with the input bottleneck through a skip-connection summing up to an output of size m_s. The last section of the Flip-U-Net is a fully-connected neural network. It consists of several hidden layers with ReLU activations. Dropout is added solely at the dense layer before the output layer. The width of the input is of size m_s while the output layer is of width m yielding class probabilities. Next, we evaluate the performance of the above approach for in-hand object recognition using an FMG device.

Fig. 4. Example of two object classes used in the data collection: (left) drinking cups and (right) scissors.

III. Experiments and Analysis

In this section, we test and analyse the proposed FMG device and classification approach with multiple human subjects and over a set of object classes. Videos of the data collection, experiments and demonstrations can be seen in the supplementary material.

A. Database and Descriptors

We have picked $m = 11$ everyday object classes including: books (or notebooks), bottles, cellphones, forks, hammers, drinking cups, plates, rulers, scissors, screwdrivers and spoons. For each object class, five objects of the same class were used in training. Figure 4 shows examples for the scissors and hammer classes. Each object of the class has different dimensions, topology and weight. The variety in physical properties of the objects affect the FMG signals when holding them and, therefore, provide data variance during training. For testing, we have used 11 additional objects (each from a different class) not used in the training. Further description on all objects along with collected data and code are available in a dedicated Git repository[1].

We have included multiple human participants in the data collection for training and testing. For training, we have recorded a dataset based on $k = 9$ subjects including three females and six males. For each subject, we measured the forearm length L, wrist circumference D_1, upper forearm circumference D_2, weight and height. Weight and height were used to compute the subjects' Body Mass Index (BMI). Table I presents the mean and standard deviation of these anthropometric measures for the subjects in training. Only data collected from these nine subjects was used to train various classifiers. On the other hand, data from 12 subjects (five females and seven males) was used for testing. The 12 test subjects were not included in training in any way. The test subjects were picked out such that they represent a large anthropometric variance. A detailed list of anthropometric measures sorted by BMI is also included in Table I.

In a training dataset collection session, the participants were instructed to hold an object both in a task-based grasp (e.g., grab the hammer or screwdriver by the handle) and in any intuitive way he or she decides. The participants freely manipulated their hands during the recording and switched between various grasps of the object. Examples of manipulating scissors and an hammer in different configurations are shown in Figure 2a. It is estimated that each object class was grasped with 2-3 main grasp taxonomies as defined in [27]. For example, the bottles were held by the body with a Medium wrap (22) grasp or by the cap with a Tripod (14). The participants were also asked to perform other manipulations such as hitting a polyethylene foam cube with the hammer and cut paper with scissors. Hence, the data collection process included a large number of different grasps and manipulations for each object. In between episodes, the FMG device was taken off and repositioned in slightly different variations. In conclusion, data was recorded by each participant on all objects of the 11 classes, yielding $M = 55$ episodes. Each episode includes $N_e = 10,000$ samples recorded in 200 Hz. Consequently, the non-augmented training set Φ is comprised of $N = 4,950,000$ labeled data points from various human subjects, objects, grasps and manipulations.

[1] https://github.com/eranbTAU/Robust-MUO

TABLE I

Details of participant demographics and anthropometry

	Gender	Age	D_1 (mm)	D_2 (mm)	L (mm)	BMI (kg/m^2)
	Training dataset - 9 human subjects					
Mean	3 × F	30.2	171.44	257.22	170.89	24.81
Std.	6 × M	2.82	22.36	33.74	27.38	3.35
Subj.	Testing dataset - 12 human subjects					
1	F	24	155	223	159	19.95
2	F	28	165	250	170	21.30
3	F	26	150	215	155	21.34
4	M	32	180	270	180	21.46
5	M	38	170	250	180	21.55
6	F	50	145	215	165	22.65
7	F	35	153	240	163	23.31
8	M	30	180	280	180	25.31
9	M	32	165	275	175	26.00
10	M	29	175	265	182	28.09
11	M	25	199	319	209	29.21
12	M	27	170	268	178	32.30

B. Data processing and model training

The collected training set is augmented as described in Section II-C. Hence, for each recorded data point we generate augmented variants and store them all together. All original and augmented training data are then used to train the VAE. Hyper-parameters optimization of the Flip-U-Net and VAE has produced the best loss value with a latent space of dimension $d = 5$. The reduced data and its corresponding labels are used to train the Flip-U-Net. The hyper-parameters optimization yielded a network of 15 convolutional layers and five fully-connected layers (Figure 2d). Thus, the total number of trainable parameters is 163,751. We used the ADAM optimizer along with a cross-entropy loss function with adaptive learning rate initialized at 0.000142, L2 regularization of 0.03, batch size of $d_s = 64$ and 30 epochs.

C. Model Evaluation

Using the training data, we have trained the VAE and Flip-U-Net as described above. Furthermore, we have conducted a comparison to other common classifiers, including: Nearest-Neighbors, Naive Bayes, Decision Trees, Random Forests, Support Vector Machines (SVM) with a linear kernel, Linear Discriminant Analysis (LDA) [28]. In addition, we include results for a simple ANN (115,715 trainable parameters) and an ANN classifier (130,499 trainable parameters) trained similarly to [16]. In the latter case, a signal is formed by the concatenation of the mean and standard deviation of a sensor

TABLE II

CLASSIFICATION SUCCESS COMPARISON FOR DIFFERENT CLASSIFIERS

Classifier	Success rate (%) w/o Aug. w/o VAE	w/o Aug. w/ VAE	w/ Aug. w/ VAE
Nearest Neighbors	70.12%	78.51%	80.36%
Naive Bayes	74.57%	80.99%	81.12%
Decision Tree	52.55%	63.56%	69.59%
Random forest	86.63%	75.31%	75.82%
SVM	65.17%	79.83%	82.72%
ANN	84.33%	77.91%	78.18%
LDA	41.15%	68.28%	71.39%
Mean-Filter + ANN [16]	77.82%	-	-
Flip-U-Net	88.91%	92.45%	94.83%

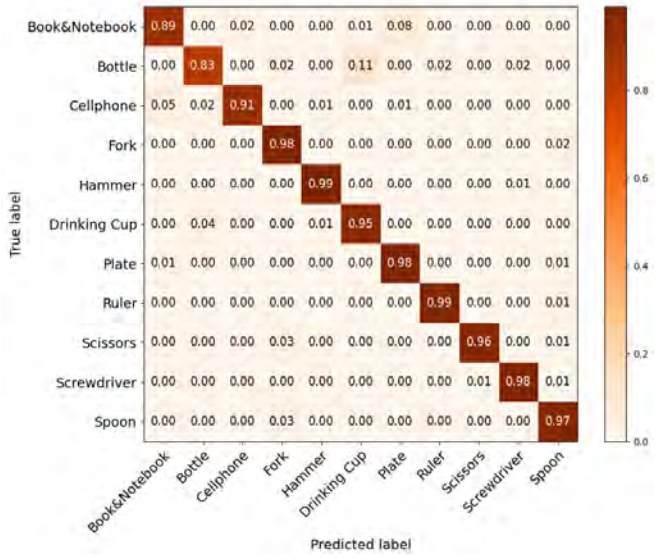

Fig. 5. Confusion matrix for the Flip-U-Net with data augmentation and dimensionality reduction using VAE, with a total success rate of 94.83%.

measurement along a sliding window of width $w = 100$ prior to training a fully-connected ANN classifier.

All classifiers were trained with the same data. Table II reports the classification success rate for these classifiers over the test data with and without VAE and data augmentation. Results show the importance of the data augmentation and VAE. These steps improve accuracy in most cases. Overall, it is clear that the Flip-U-Net approach along with augmentation and VAE outperforms standard methods. Figure 5 presents the confusion matrix for the Flip-U-Net classifier with data augmentation and VAE. Each different data augmentation technique provides some improvement to robustness. Individual analysis for each technique has shown relative improvement as low as 0.43% for the Magnitude-Warping and as high as 1.49% for the Rotation, this with respect to the 92.45% success rate without any augmentation.

TABLE III

CLASSIFICATION SUCCESS COMPARISON FOR DIFFERENT DR METHODS

DR method	Success rate (%) w/o Aug.	w/ Aug.
PCA	61.45%	65.24%
T-SNE	80.77%	84.27%
AE	88.89%	91.01%
VAE	92.45%	94.83%
Encoder	89.07%	91.63%

Grasps of several objects during experiments by two test subjects can be seen in Figures 6-8 along with model certainties for the predicted objects. Note that some grasps of different object classes may look similar. For example, the grasps of the hammer and screwdriver in Figures 6a and 8b, respectively, are both Adducted thumb grasps [27]. However, weight and center-of-mass location have a significant effect on the FMG signals along with the object's geometry. Nonetheless, other object classes, such as the bottle and drinking cup or the notebook and plate, share some grasp taxonomies (Medium wrap and Parallel extension, respectively) that are harder to distinguish. Hence, they have larger classification errors as seen in Figure 5. Figure 1 (and the supplementary video) presents demonstrations of complementary assistance of a collaborative robotic arm acting based on FMG observations of the objects held by the user. Real-time recognition of the object is performed at a frequency of 50 Hz. Hence, an object can be identified almost instantly when grasped. Overall, the experiments demonstrate the ability of the classifier to accurately recognize objects based on different grasps.

We analyze and compare the use of other DR methods prior to training the Flip-U-Net, including the following methods: Principle Component Analysis (PCA), t-distributed Stochastic Neighbor Embedding (T-SNE), AE and VAE. We also include a baseline in which the *Encoder* of the VAE is directly connected to the Flip-U-Net and trained with it to minimize classification error. Table III presents results for comparison with and without data augmentation. Evidently, VAE along with augmentation provides best accuracy.

Fig. 6. Grasps of an hammer by subjects 4 (a-c) and 9 (d-f). The classification model certainties $p(hammer)$ about the object are (a) 86.47%, (b) 97.29%, (c) 82.42%, (d) 93.1%, (e) 93.29% and (f) 93.36%.

Fig. 7. Grasps of a drinking cup by subject 4. The classification model certainties $p(cup)$ about the object are (a) 98.65%, (b) 95.71%, (c) 85.43%, (d) 96.77% and (e) 92.96%.

Fig. 8. Grasps of a screwdriver by subject 9. The classification model certainties $p(screwdriver)$ about the object are (a) 94.92%, (b) 93.05%, (c) 90.66%, (d) 94.06% and (e) 93.78%.

In the next analysis, we observe the multi-user robustness property with regards to data size. Recall that data was acquired by recording nine participants for all objects yielding $N = 4,950,000$ data points. Hence, we arranged all data sequentially as recorded and without any shuffling. In order to observe classification success rate with regards to data size, the entire training procedure (i.e., augmentation, VAE and Flip-U-Net) was repeatedly trained for a varying

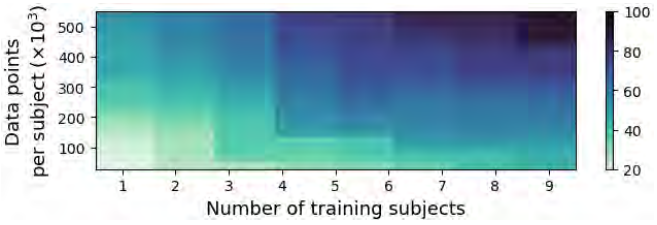

Fig. 9. A heat-map of the classification success rate (%) with regards to the number subjects and the number of recorded data points per subject.

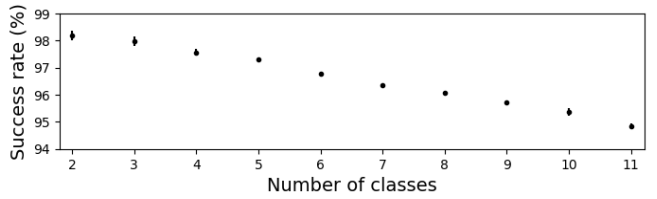

Fig. 10. Mean success rate with regards to the number of classes.

number of subjects and data points per subject. A heat-map of the success rate over the test data can be seen in Figure 9. The classification success rate improves with the increase of human subjects and data points per subject, reaching 94.83% success rate with N points. As expected, additional participants in the training data yields a more robust classifier. Additionally, Figure 10 presents the mean success rate with regards to the number of object classes. Results show a moderate decline in accuracy when increasing the number of classes leading to 94.83% with 11 classes.

D. Feature Importance

We now explore the importance of the FSR sensors in the device on prediction accuracy. Permutation feature importance is a common method to evaluate the impact of each feature in an ANN [29]. We measure the increase in the prediction error after permuting the values of each single feature in the validation data separately. The score is the accuracy reduction resulting from the permutation of a sensor's values and is computed according to $e_i = \frac{q-q_i}{q} \times 100\%$, where q is the success rate of the non-permuted model. q_i is the success rate when feature i is permuted. The results of feature importance evaluation are illustrated in Figure 11 along with sensor placements. The relative accuracies indicate a relatively strong dependence on the lower forearm sensors. All sensors along the arm are significant to a robust and accurate object recognition.

E. Correlation Analysis

We wish to analyze the accuracy of different test subjects with regards to anthropometric measures. Table IV presents the classification success rates for each individual test subject and for each gender. We have also included two metrics of signal strength: mean of all signals over all sensor measurements made by the subject and mean of the maximal signal measurements over all sensors. Signal measurements are voltage values normalized to be within $[0, 1]$.

Table V presents the Pearson correlation (r) coefficients between demography and anthropometric measures, and classification success rate. Results show that forearm measures

i	e_i (%)
1	10.98
2	3.61
3	4.05
4	6.49
5	19.51
6	8.37
7	3.26
8	12.92
9	16.50
10	21.83
11	11.01
12	8.49
13	20.66
14	13.78
15	12.14

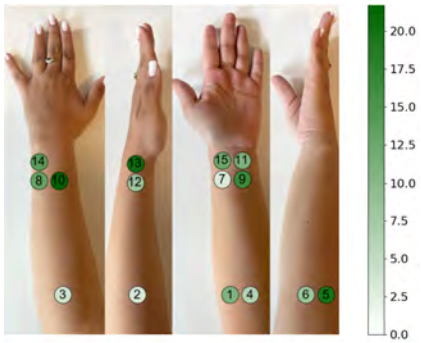

Fig. 11. Illustration of the sensor locations and importance score computed with the permutation feature importance method.

TABLE IV
CLASSIFICATION SUCCESS RATE OF THE HUMAN TEST SUBJECTS

Subject	Success rate (%)	Signal Strength	
		Mean	Mean-Max.
1	92	0.311	0.61
2	95	0.231	0.66
3	95	0.264	0.73
4	94	0.376	0.75
5	95	0.318	0.76
6	92	0.231	0.55
7	93	0.288	0.59
8	97	0.410	0.85
9	94	0.360	0.66
10	96	0.273	0.71
11	98	0.405	0.79
12	97	0.45	0.79
Females	96	0.265	0.63
Males	93	0.370	0.74

(i.e., D_1, D_2 and L) and BMI have relatively high impact on the accuracy of the model. This is assumed to be because of better attachment of the sensors to the skin for subjects with larger forearms. Gender and age have low to medium correlation to model accuracy since they both have some bias with regards to anthropometric measures. The male subjects, on average, have a much higher BMI and are younger. Furthermore, subjects with higher signal strength are more likely to receive accurate predictions.

F. Anthropometric-based Model Tuning

The above results show accuracy correlation to anthropometric measures. Hence, we now test the ability to fine-tune the model based on one measure. We take the BMI measure as a test case. Given a test subject, we retrain the model with data from a train participant of similar BMI. The retraining is done with a learning rate of 10^{-5} and 30 epochs. Figure 12 shows the accuracy gain achieved for four test participants with regards to the BMI difference ΔBMI. ΔBMI is the difference between the BMI of the test participant and of the train subject whose data was used for retraining. Accuracy gain was evaluated using the test data of the corresponding

TABLE V
CORRELATION (r) BETWEEN ANTHROPOMETRY AND ACCURACY

Measure	r	Measure	r
Gender	0.65	D_1	0.78
Age	-0.49	D_2	0.75
Mean-Max. signal strength	0.87	L	0.72
Mean signal strength	0.59	BMI	0.70

test participant and was averaged over 10 training trials. Results show strong negative correlation between accuracy gain and BMI similarity. In other words, accuracy is better improved with data from a subject who has similar BMI. Therefore, the model can be tuned to a particular subject based on his or her BMI measure.

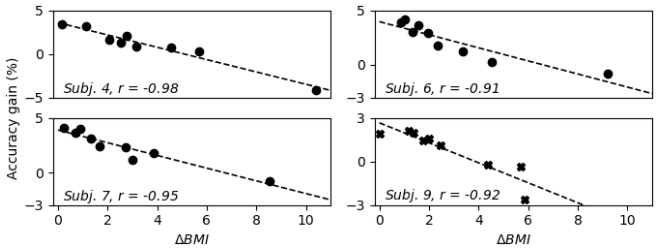

Fig. 12. Accuracy gain of a tuned model for subjects 4, 6, 7 and 9 with different train subjects and with regards to ΔBMI.

IV. CONCLUSIONS

In this paper, we have proposed an end-to-end approach for robust multi-user object recognition using a wearable FMG device with application in HRC. We introduced and evaluated a novel deep learning architecture called Flip-U-Net. Results show that Flip-U-Net along with data augmentation and VAE can achieve better performance than standard methods. Moreover, we examined the influence of each sensor of the FMG device on the accuracy. Correlation analysis have shown relations between accuracy and anthropometric measures of the test subjects. While the model by itself provides high accuracy, the BMI measure can be used as a selection criterion for fine-tuning the model to a new user.

While the approach provides fast and accurate predictions of grasped objects, it requires sufficient data from human subjects and long model training time. Furthermore, transition between objects in real-time may provide faulty predictions. Future work should address the identification of these transitions to increase certainty. This could also include information from other sensors. Future work may also include using network architectures that take sequential data. In addition, Conditional VAE could be used to fine-tune a model for a new user and generate synthetic training data with bias towards some anthropometric measure.

REFERENCES

[1] A. Ajoudani, A. M. Zanchettin, S. Ivaldi, A. Albu-Schäffer, K. Kosuge, and O. Khatib, "Progress and prospects of the human-robot collaboration," *Autonomous Robots*, 10 2017.

[2] L. F. Nicolas-Alonso and J. Gomez-Gil, "Brain computer interfaces, a review," *Sensors*, vol. 12, no. 2, pp. 1211–1279, 2012.

[3] B. Gleeson, K. MacLean, A. Haddadi, E. Croft, and J. Alcazar, "Gestures for industry: Intuitive human-robot communication from human observation," in *Proceedings of the ACM/IEEE International Conference on Human-robot Interaction*, 2013, pp. 349–356.

[4] W. Yang, C. Paxton, M. Cakmak, and D. Fox, "Human grasp classification for reactive human-to-robot handovers," *IEEE/RSJ Inter. Conf. on Intelligent Robots and Systems*, pp. 11 123–11 130, 2020.

[5] Z. Khokhar, Z. Xiao, and C. Menon, "Surface EMG pattern recognition for real-time control of a wrist exoskeleton," *Biomedical engineering online*, vol. 9, p. 41, Aug. 2010.

[6] E. Fujiwara, Y. T. Wu, C. K. Suzuki, D. T. G. de Andrade, A. R. Neto, and E. Rohmer, "Optical fiber force myography sensor for applications in prosthetic hand control," in *Proceedings of the IEEE International Workshop on Advanced Motion Control (AMC)*, 2018, pp. 342–347.

[7] X. Li, Q. Zhuo, X. Zhang, O. W. Samuel, Z. Xia, X. Zhang, P. Fang, and G. Li, "FMG-based body motion registration using piezoelectret sensors," in *Int. Conf. of the IEEE Eng. in Medicine and Biology Society*, 2016, pp. 4626–4629.

[8] A. Belyea, K. Englehart, and E. Scheme, "FMG versus EMG: A comparison of usability for real-time pattern recognition based control," *IEEE Trans. on Biomed. Eng.*, vol. 66, no. 11, pp. 3098–3104, 2019.

[9] O. Amft, H. Junker, P. Lukowicz, G. Troster, and C. Schuster, "Sensing muscle activities with body-worn sensors," in *Int Work Wearable Implant Body Sens Networks*, 05 2006, pp. 138–141.

[10] G. Ogris, M. Kreil, and P. Lukowicz, "Using FSR based muscule activity monitoring to recognize manipulative arm gestures," in *IEEE Int Symp Wearable Comput*, 2007, pp. 45 – 48.

[11] N. Li, D. Yang, L. Jiang, H. Liu, and H. Cai, "Combined use of FSR sensor array and SVM classifier for finger motion recognition based on pressure distribution map," *Journal of Bionic Engineering*, vol. 9, no. 1, pp. 39–47, 2012.

[12] H. K. Yap, A. Mao, J. C. H. Goh, and C. Yeow, "Design of a wearable FMG sensing system for user intent detection during hand rehabilitation with a soft robotic glove," in *IEEE Int. Conf. on Biomedical Rob. and Biomech.*, 2016, pp. 781–786.

[13] X. Jiang, L.-K. Merhi, Z. G. Xiao, and C. Menon, "Exploration of force myography and surface electromyography in hand gesture classification," *Medical Eng. & Physics*, vol. 41, pp. 63 – 73, 2017.

[14] A. Gigli, V. Gregori, M. Cognolato, M. Atzori, and A. Gijsberts, "Visual cues to improve myoelectric control of upper limb prostheses," in *IEEE Int. Conf. on Biomed. Rob. and Biomech.*, 2018, pp. 783–788.

[15] U. Zakia, X. Jiang, and C. Menon, "Deep learning technique in recognizing hand grasps using fmg signals," in *IEEE Annual Info. Tech., Elect. and Mobile Comm. Conf.*, 2020, pp. 0546–0552.

[16] N. D. Kahanowich and A. Sintov, "Robust classification of grasped objects in intuitive human-robot collaboration using a wearable force-myography device," *IEEE Robotics and Automation Letters*, vol. 6, no. 2, pp. 1192–1199, 2021.

[17] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[18] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[19] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *Inter. Interdisciplinary PhD workshop*, 2018, pp. 117–122.

[20] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, pp. 1–48, 2019.

[21] T. T. Um, F. M. Pfister, D. Pichler, S. Endo, M. Lang, S. Hirche, U. Fietzek, and D. Kulić, "Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks," in *Proc. ACM Inter. Conf. on Multimodal Interaction*, 2017.

[22] M. Olson, A. J. Wyner, and R. Berk, "Modern neural networks generalize on small data sets," in *Proc. International Conference on Neural Information Processing Systems*, 2018, pp. 3623–3632.

[23] L. Lusa *et al.*, "Improved shrunken centroid classifiers for high-dimensional class-imbalanced data," *BMC bioinformatics*, vol. 14, no. 1, pp. 1–13, 2013.

[24] R. Hasibi, M. Shokri, and M. Dehghan, "Augmentation scheme for dealing with imbalanced network traffic classification using deep learning," *arXiv preprint arXiv:1901.00204*, 2019.

[25] J. Zhai, S. Zhang, J. Chen, and Q. He, "Autoencoder and its various variants," in *IEEE International Conference on Systems, Man, and Cybernetics*, 2018, pp. 415–419.

[26] S. K. Portillo, J. K. Parejko, J. R. Vergara, and A. J. Connolly, "Dimensionality reduction of sdss spectra with variational autoencoders," *The Astronomical Journal*, vol. 160, no. 1, p. 45, 2020.

[27] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on Human-Machine Systems*, vol. 46, no. 1, pp. 66–77, 2016.

[28] A. Singh, N. Thakur, and A. Sharma, "A review of supervised machine learning algorithms," in *International Conference on Computing for Sustainable Global Development*, 2016, pp. 1310–1315.

[29] J. Yang, K. Shen, C. Ong, and X. Li, "Feature selection for mlp neural network: The use of random permutation of probabilistic outputs," *IEEE Trans. on Neural Net.*, vol. 20, no. 12, pp. 1911–1922, 2009.