IEEE Robotics and Automation Letters (RAL) paper presented at the
2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Image Transformation and CNNs: A Strategy for Encoding Human Locomotor Intent for Autonomous Wearable Robots

Ung Hee Lee[1], *Student Member, IEEE*, Justin Bi[2], Rishi Patel[2], David Fouhey[3], Elliott Rouse[1], *Member, IEEE*

Abstract— **Wearable robots have the potential to improve the lives of countless individuals; however, challenges associated with controlling these systems must be addressed before they can reach their full potential. Modern control strategies for wearable robots are predicated on activity-specific implementations, and testing is usually limited to a single, fixed activity within the laboratory (*e.g.* level ground walking). To accommodate various activities in real-world scenarios, control strategies must include the ability to safely and seamlessly transition between activity-specific controllers. One potential solution to this challenge is to the infer wearer's intent using pattern recognition of locomotion sensor data. To this end, we developed an intent recognition framework implementing convolutional neural networks with image encoding (*i.e.* spectrogram) that enables prediction of the upcoming locomotor activity of the wearer's next step. In this paper, we describe our intent recognition system, comprised of a mel-spectrogram and subsequent neural network architecture. In addition, we analyzed the effect of sensor locations and modalities on the recognition system, and compared our proposed system to state-of-the-art locomotor intent recognition strategies. We were able to attain high classification performance (error rate: 1.1%), which was comparable or better than previous systems.**

I. INTRODUCTION

Wearable robots, including powered prostheses and exoskeletons, have the potential to improve people's quality of life by enhancing their physical capabilities during locomotion [1], [2]. Despite the promise of these wearable technologies, challenges remain in the development of safe, intuitive, and versatile control systems. Recently, researchers have demonstrated exoskeletons that are able to apply substantial assistance, as well as reduce the metabolic expenditure during walking [1], [3]. To obtain these results, researchers typically develop control approaches that are intended for single activities, often tethered to a treadmill. For this approach to be applicable in daily life, these systems must be able to encompass multiple activities, including walking, running, and stair ascent or descent. To address the limitations associated with control systems meant for single activities, some researchers have developed methods for switching between multiple activity-specific controllers; however, often these transitions are initiated by commands such as visual, auditory, or touch (*e.g.* key-fob) cues which are non-intuitive and can increase cognitive burden [4]. Thus, for users to naturally perform the activities of daily life, it is imperative to develop control strategies that can infer the wearer's intended movement automatically without requiring external commands, and autonomously transition between different activity-specific controllers.

One approach to infer the wearer's intended activity is to use an intent recognition framework [4], [5], [6]. Intent recognition typically includes predicting the upcoming activities of the user each step using information from the wearer, robotic system, or environment prior to completing the movement (*e.g.* before heel contact or toe off of the current step) [7]. There have been several works that implemented intent recognition strategies employing sensor fusion for improving the performance [5], [8]. While these strategies demonstrated high performance on classifying users' locomotor activities (error rate $< 2\%$), they often rely on hand-crafted features, such as the mean, standard deviation, maximum and minimum of time-series data. This can be challenging because it may require domain specific knowledge and trial and error approaches to extract meaningful features [9].

Deep learning (DL) has been emerging as a tool to classify activities in human activity recognition (HAR) or intent recognition tasks [10], [11], [12]. Especially, convolutional neural networks (CNNs) have been used over other DL methods, due to their local dependency and scale invariance, which captures the invariant features of the same activities with variations (*e.g.* walking) [10]. Combined with recent advancement in processing capability and miniaturization of graphics processing units (GPUs), CNNs have been extensively employed for mobile and wearable sensors based tasks. To increase the performance of CNNs, several researchers have configured CNN architectures by adding additional layers and nodes or combined with other DL architectures (*e.g.* CNN + Long Short Term Memory) [13], [14]. These approaches can increase the computational complexity, which may not be ideal for low-power on-board sensors or microcontrollers [15]. In addition, due to the increased number of parameters in these architectures, it may be challenging to determine the optimal parameters from relatively small datasets [16].

To obtain better performance while minimizing the computational efforts, researchers in HAR have investigated various techniques for configuring input data, such as lin-

This work was supported by the the D. Dan and Betty Kahn Foundation.

[1]U. Lee and E. J. Rouse are with the Department of Mechanical Engineering and the Robotics Institute, University of Michigan, and also with the Neurobionics Lab, University of Michigan, Ann Arbor, MI 48109 USA; unghee@umich.edu, ejrouse@umich.edu

[2]J. Bi and R. Patel are with the Department of Electrical and Computer Engineering, University of Michigan.

[3]D. Fouhey is with the Department of Computer Science and Engineering and the Robotics Institute, University of Michigan.

ear interpolation, distance matrices, *etc.* [17]. Among these techniques, the use of image transformation (*i.e.* 2D representation) of time-series data as an input to CNNs, have been employed for classifying activities [11]. Especially, conversion to the spectrogram captures frequency features of the signals and is robust against variance of sampling rate [15]. While these image configuration techniques have achieved promising results, many researchers have focused on classifying the activity *after* the movement completion, rather than predicting the activity *before* the completion (*i.e.* intent). Specifically for HAR tasks of walking activities, researchers have focused on classifying the activity of the current or past step, rather than the activity of the subsequent step of the gait cycle [10], [15]. CNNs have been used to predict locomotor intent for use in powered prostheses; however, they either have directly applied time-series data or hand-crafted features as an input to the CNN, which resulted in similar or inferior performance compared to the feature-based classifiers (*e.g.* Linear Discriminant Analysis) [12], [18], [19]. Thus, the impact of these image configuration methods applied to CNNs for locomotor intent recognition tasks remains unknown.

The contributions from this paper include: (1) We propose a spectrogram-based CNN recognition framework for predicting the *intent* of the lower-limb locomotor activities. Inspired by [15], we modified this approach to be suitable for our tasks by developing an analysis pipeline composed of a lightweight neural network architecture and a mel-scaled spectrogram. (2) We compared the performance of our system to the state-of-the-art (SOTA) locomotor intent recognition strategies using bilateral neuromechanical signals. The proposed system achieved a classification error rate of 1.1%, which outperformed or was comparable to previous works [5], [19]. (3) We characterized the effect of sensor locations and modalities on the classifier performance; finally, (4) we qualitatively identified the region of the gait cycle responsible for the intention by visualizing the activation of the CNN. To our knowledge, this is the first work to use CNNs with image encoding of frequency content for lower-limb intent recognition with bilateral neuromechanical sensor fusion. The intent of our work is to enable future wearable robotic technologies to be used outside the laboratory, where a diverse range of activities is required.

II. System Design

A. Dataset

We used a publicly available dataset composed of kinematic and muscle activity signals to train our intent recognition framework. The dataset, named as the Encyclopedia of Able-bodied Bilateral Lower Limb Locomotor Signals (ENABL3S), was chosen over other datasets (*e.g.* UCI-HAR [21]), because it focuses on normal locomotion, includes rich biomechanical signals from multiple sensor modalities, and sampling rates are sufficient for online control purposes. The data were collected from wearable electrogoniometers (GONIO), surface electromyography (EMG) and internal measurement unit (IMU) sensors. The sampling rate of EMG, GONIO and IMU sensors were 1000, 500, and 500 Hz respectively and low-pass filtered at 350, 10 and 25 Hz respectively. All sensor data were processed to identify right and left heel contact and toe off (*i.e.* gait events). These sensors were placed on the lower limbs of 10 able-bodied human subjects (Fig. 1). Each subject performed 25 repetitions of a circuit consisting of walking on level ground (LW), ascending/descending a ramp with a 10 degree incline (RA/RD), and ascending/descending a four-step staircase (SA/SD). The odd-numbered trials had a sequence of these activities as follows: LW→ SA→ LW→ RD→ LW, while even-numbered trials had LW→ RA→ LW→ SD→ LW. The ground truth label was marked by the experimenter using a key fob. The preceding 500 ms of sensor data before each gait event was used as the input to our analysis pipeline and the activities after each gait event (*i.e.* upcoming activity) were used as the label for prediction [20].

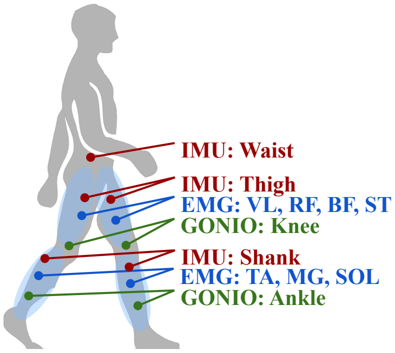

Fig. 1: Instrumentation setup for the dataset. The EMG electrodes were placed on seven muscle groups responsible for lower limb locomotion: tibialis anterior (TA), medial gastrocnemius (MG), soleus (SOL), vastus lateralis (VL), rectus femoris (RF), biceps femoris (BF), and semitendinosus (ST). The GONIOs were placed on knee and ankle joints and the IMUs were placed on the thigh and shank to measure angular position and velocity. [20]

B. CNN-based Intent Recognition

1) Image Encoding using Spectrograms: Due to the periodic nature of walking, we propose that the frequency domain information from the time-series data provides a more effective representation of lower-limb locomotor activities for CNN classification. To produce the frequency-domain representation, the Short-Time Fourier Transform (STFT) was performed on time-series data:

$$\mathrm{STFT}(x[n]; w, k) = \sum_{n=-\infty}^{\infty} x[n]w[n-k]e^{-j\omega n} \tag{1}$$

where the signal $x[n]$ was multiplied by a windowing function ω, shifted by an offset k. The squared magnitude of the STFT produced a spectrogram, and we further transformed the spectrogram using nonlinear scaling known as the *mel scale* (Eq. 2) which demonstrated its success as a preprocessing step in auditory classification tasks [22]. The mel scale originates from representing the human auditory system such that it has perceptually equal pitch (*i.e.* frequency-scale)

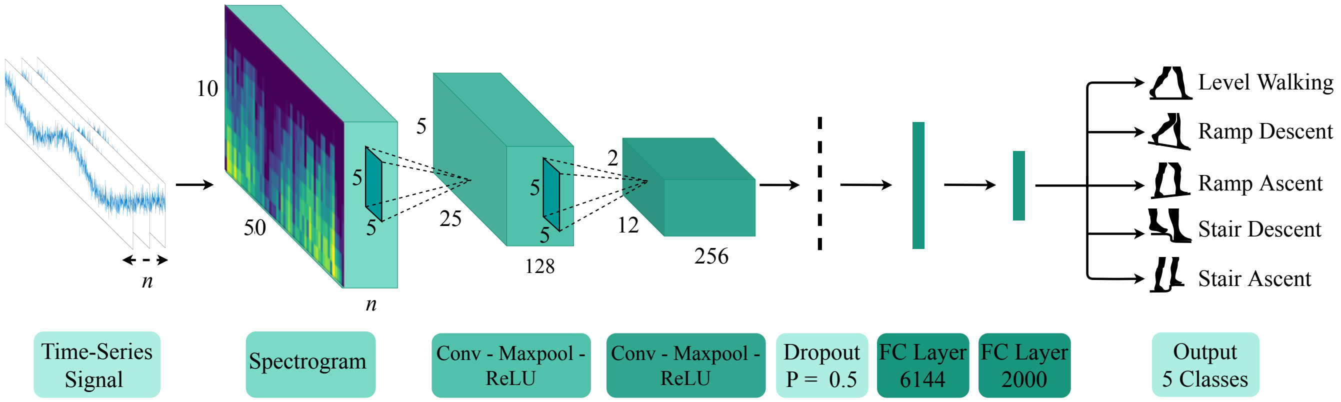

Fig. 2: Our proposed intent recognition pipeline from frequency domain representation, CNN architecture to the output activities. Convolutional layers consisted of kernels sized of 5x5, stride of 1, padding of 2, and two sequential linear layers had hidden units of 6144 and 2000, respectively. The dropout with 0.5 probability was added to improve the generalizability of the proposed system.

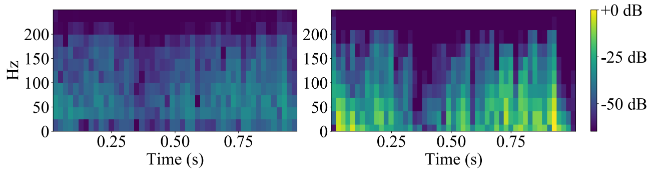

Fig. 3: A sample spectrogram (left) and mel-spectrogram (right) generated from the EMG signal of the right TA. Lower frequency signals were amplified from the mel scale conversion.

increments; in other words, as frequency (Hz) increases, larger intervals of frequency are required to produce the same magnitude of pitch increments. This scaling was chosen so that it amplifies the lower frequency content of the signal (Fig. 3), where much of the content of human locomotion is below 3.5 Hz [23]. The mel scale can be computed using

$$2595 \cdot log_{10}\left(1+\frac{f}{700}\right) \tag{2}$$

Lastly, the amplitudes were squared to further attenuate the higher frequency content and transformed to a decibel scale. For the windowing function ω, a Hann window of length 20, with an offset k of 10 was used. When converting to the mel scale, the Hz scale was partitioned into 10 bins (user-defined) prior to the transformation being applied. The selection of 10 bins was to balance classifier performance and processing overhead. All the steps outlined in this section were performed using the LibROSA package in Python [24].

2) LIR-Net Architecture: We designed the CNN architecture which consists of a series of 2D convolutional layers and pooling layers, followed by fully-connected layers (Fig 2). This proposed network is called *Locomotor Intent Recognition-Net* (LIR-Net), which is lightweight but provides high performance when classifying lower-limb neuromechanical spectrogram images. The spectrogram produced in the image encoding step (Section II-B.1) was provided as an input to the CNN and the softmax operation was applied to the output of the last linear layer, which represented probability distribution of the predicted class.

III. Experimental Protocol

The proposed system was evaluated against separate classifier configurations and compared with different classification strategies using the ENABL3S dataset. Furthermore, we investigated the effect of sensor modalities and laterality groups, where modality describes sensor type (*e.g.* IMU, EMG, GONIO) and laterality describes the side of the leg where a gait event was detected (*e.g.* ipsilateral, contralateral, and bilateral) [20]. Lastly, we implemented feature (*i.e.* unit) visualization of LIR-Net to identify the frequency region of the input spectrogram where the units were most activated.

A. Classifier Configuration

1) Generic: A *generic configuration* is defined as when only signal data or features were given as an input to a certain classification strategy without any information (*i.e.* ground truth) from the current activity provided. The *current activity* was defined as the activity before each gait event (*i.e.* before movement completion), whereas *upcoming activity* was defined as the future activity after each gait event.

2) Mode-Specific: The mode-specific strategy encodes the environment knowledge by providing the information of the current locomotor activity [4]. Specifically, the strategy employs separate classifiers depending on the current activity (*i.e.* mode), which has different number of outputs for each classifier (*e.g.* for the RA classifier, only transition to RA or LW is allowed). Combined with heuristic feature-based classifiers (*e.g.* Linear Discriminant Analysis (LDA), Support Vector Machine (SVM)), this configuration demonstrated low error rates ($< 2\%$) when classifying the locomotor intent [5].

B. Classification Strategies

1) Random Guesser: To understand the effect of distribution of the dataset on the classifier configurations, we created a baseline system that predicts activity based on distribution of samples, thereby always predicting the activity with the greatest likelihood.

Generic: Provided only the signal data without activity information, such a system can be represented as follows:

TABLE I: DATA DISTRIBUTION OF ENABL3S

Transition from	to	Number of Samples[#]
Level walking (LW)	LW	8886 (42.87%)
	RA	503 (2.43%)
	RD	474 (2.29%)
	SA	478 (2.31%)
	SD	477 (2.30%)
Ramp ascent (RA)	RA	2740 (13.22%)
	LW	481 (2.32%)
Ramp descent (RD)	RD	3416 (16.48%)
	LW	471 (2.27%)
Stair ascent (SA)	SA	934 (4.51%)
	LW	469 (2.26%)
Stair descent (SD)	SD	925 (4.46%)
	LW	476 (2.30%)

[#] The count of each activity transitions across all subjects.

$$\hat{i}_n = \arg\max_i P(i) \tag{3}$$

where $\hat{i}, i$ are the predicted class (*i.e.* upcoming activity) and true class label respectively, n represents the nth gait event, and P is the probability distribution of the class i. Therefore, the class with the largest representation (LW, Tab. I) was chosen every time.

Mode-Specific: Given the signal data and the current class of the signal data, we represented a similar system as follows:

$$\hat{i}_n = \arg\max_i P(i|i_{n-1}) \tag{4}$$

where P is the probability distribution of the class i given a current activity (i_{n-1}) of nth gait event. For example, given a RA, the classifier only outputs RA since the data distribution of RA-RA is (13%) larger than RA-LW (2%).

2) Heuristic Feature-based Classifiers: LDA and SVM have demonstrated their validity as classifiers for intent recognition, because they provide low classification error while it is computationally efficient [4], [5]. Especially, LDA combined with the mode-specific configuration achieved SOTA performance (1.43%) for intent recognition tasks [5]. To this end, we used LDA and SVM for our baseline classifiers to be compared with our proposed system. For an input to the classifiers, we used features previously known to be important for intent recognition when controlling powered prostheses. Features were extracted from the time-series data, including mean, standard deviation, maximum, minimum, initial, and final value, *etc* [5], [20].

Generic: The features extracted from bilateral sensor set with all sensor modalities of ENABL3S were provided as an input to the classifier, which were 332 features in total. To be consistent with the existing work, for the LDA classifer, features were normalized and principal component analysis was applied to maintain 95% variance, while the prior was set to be equally probable. For the SVM, a linear kernel was chosen with a regularization parameter of 10 [5]. The calculation was performed using the Sckit-learn software package in Python.

Mode-Specific: Separate LDA and SVM classifiers were trained to encompass all the gait events and locomotor activities. During the prediction, the mode-specific classifier was selected based on the current locomotion activity. The predictions (*i.e.* output) of the classifiers were limited by the number of transitions allowed on the previous activity.

3) LIR-Net: Generic: A generic configuration of LIR-Net followed the procedures of Section II-B.

Mode-Specific: To provide the current activity information to the network, we provided the mode information as a one-hot encoding vector and concatenated into the first linear layer of our intent recognition pipeline (Fig. 2). We chose this approach rather than explicitly following the conventional mode-specific scheme (*i.e.* training separate classifiers with differing number of output depending on the mode), since the performance of DL will likely suffer from the scarcity of the training samples due to the splitting.

C. Performance Evaluation

We compared the offline performance using both classifier configurations. We divided the dataset, including all sensor laterality groups and modalities, into testing and training sets, which were divided in two ways: 1. the division was randomized by 10-fold cross validation (90:10 split) within all subjects' data (*i.e.* user-dependent); 2. one out of ten subject's data were withheld as testing set, while the other nine subjects were grouped as the training set. This was repeated 10 times until all subjects were tested once (*i.e.* leave-one-out or user-independent cross validation) [4]. Each classifier was trained on the training set and evaluated on the testing set. For the user-dependent LIR-Net, the data were divided into training, validation, and test sets (80:10:10), and after finding the best hyperparameters, the validation set was added to the training set. Identical hyperparameters of the user-dependent LIR-Net were used for the user-independent classifier. The error rate of each classifier was determined by the number of correctly classified predictions divided by the number of each test set. Error rates were further categorized based on whether the misclassification occurred at the gait event where the previous and the following activities were identical (*i.e.* steady-state) or different (*i.e.* transitional) [5].

We conducted statistical analyses separately for each error types, and analyzed all classifiers on both classifier configurations and user-dependencies. We used three-way ANOVAs with error rate as a dependent variable, and classifier type, configuration, and user-dependency as independent variables, and subject as a random factor. We performed a *post hoc* comparison test using Tukey's Honestly Significant Difference Criterion (Tukey) to determine the statistical difference between the pairs of interest ($\alpha = 0.05$).

1) Training of LIR-Net: The network was trained to minimize the cross entropy loss which is described as:

$$Loss(q, p) = -\sum_i q(i) \log p(i) \tag{5}$$

where the $q(i)$ is the ground truth probability expressed as one-hot encoding and $p(i)$ is the predicted probability of class i. We used a stochastic gradient-based optimizer ADAM [25] with L2 regularization to prevent overfitting.

2) Hyperparameter Search of LIR-Net: We investigated different hyperparameters of LIR-Net to obtain the best performance. The parameters associated with spectrogram implementation were fixed (Section II-B.1). A grid search was then performed on hyperparameters to maximize validation accuracy. The hyperparameters found to give greatest accuracy were a batch size of 32, learning rate of 10^{-5}, L2 regularization strength of 10^{-3}, and 200 epochs. All calculations were performed using the PyTorch package [26].

3) Classification Latency of LIR-Net: The time required for the classifier to make a prediction is a critical factor in the real-time usability of intent recognition systems. To evaluate the latency of LIR-Net, we measured the elapsed time from spectrogram generation to prediction per activity using all sensors. We calculated the latency using a single-board computer (model: Jetson Nano, NVIDIA, Santa Clara, CA) with GPU acceleration. The calculation was repeated 10 times and was averaged to obtain the latency.

4) Comparison to ResNet: To validate the design choice of LIR-Net architecture, we compared the performance of ours with ResNet18 [27]. We used ResNet pre-trained on ImageNet to perform a fair comparison due to our scarcity of the number of data samples compared to the complexity of ResNet [28]. An identical training and evaluation procedure for LIR-Net was employed to quantify the performance of ResNet. A one-way ANOVA was used to measure the statistical difference between the two architectures.

D. Effect of Sensor Modalities and Locations on LIR-Net

To determine the effect of sensor types and locations, we trained LIR-Net classifiers on various subsets of the data and compared their performance. The subsets were determined by dividing the sensor data into four modality (EMG, GONIO, IMU, all) and three laterality groups (contralateral, ipsilateral, bilateral). The LIR-Nets were trained on each laterality group with a subset of the four modalities, and their error rates were recorded. The remaining sensor data were withheld during training. We divided all subsets of data into testing and training sets, where the division was randomized by user-dependent 10-fold cross validation.

We conducted statistical analyses of LIR-Nets by using a two-way ANOVA with the overall error rate as the dependent variable, modality and laterality as independent variables, and subject as a random factor. We performed a *post hoc* comparison test using Tukey's Criterion to determine the statistical difference between the pairs of interest ($\alpha = 0.05$).

E. Visualizing Activations of LIR-Net

Activation (*i.e.* output of the convolutional operations) visualization is a technique that can provide greater understanding of the internal operations of CNNs [29]. To this end, we visualized the activation of the trained LIR-Net after the first convolutional layer, given one sample of spectrogram. This was accomplished by first localizing the maximum value in each output channel (total 128 channels) and mapping each pixel from the first convolutional layer back to the input space (*i.e.* receptive field), where a highly activated pixel is likely important to the CNN. After identifying the mapping between the input and the activations, the receptive field of the input was weighted by the magnitude of its according maximal activation of each channel, and the max activations of all channels were summed and then normalized to create an averaged activation in the input space.

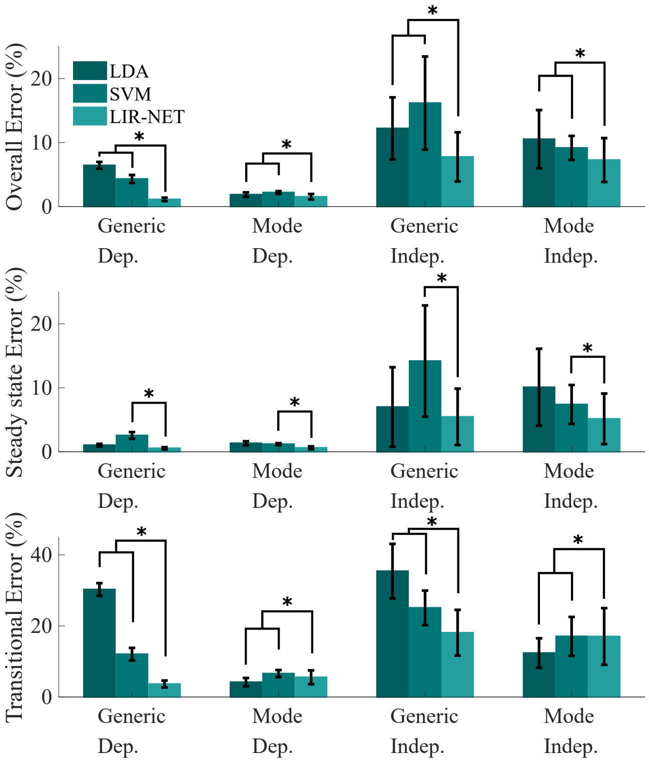

Fig. 4: Error rates of the classifiers on Generic and Mode-specific (Mode) configurations with user-independent (Indep.) and user-dependent (Dep.) conditions.

IV. Results

A. Performance Evaluation

We evaluated the performances of the classifiers on all classifier configurations and user-dependencies (Fig. 4). The interaction between all pairs of classifier types, configurations, and user-dependencies were all significant except the pair of configuration and user-dependency (p = 0.63).

1) Effect of Configuration: Error rates of the classifiers on both classifier configurations were compared (Tab. II). The generic and mode-specific configurations were statistically different across all classifiers and types of error rates. In general, the mode-specific configuration lowered the overall error rates of the random and heuristic-based classifiers, while the change in the error rate of LIR-Net was minimal.

2) Effect of Classifiers: Error rates of the random and heuristic classifiers were compared to that of LIR-Net (Tab. II) for each configuration and user-dependency. In general, the error rates were statistically different from LIR-Net across all error types, classifier configurations and user-dependencies, except the steady-state error of the generic

TABLE II: Error Rates of the Classifiers

	User-Dependent				User-Independent			
	Random	LDA	SVM	LIR-Net	Random	LDA	SVM	LIR-Net
Overall (%)								
Generic	47.99 [1.18]*	6.43 [0.53]*	4.30 [0.61]*	**1.11 [0.26]**	47.99 [2.65]*	12.21 [4.84]*	16.17 [7.27]*	**7.75 [3.84]**
Mode-specific	18.48 [1.03]*	1.85 [0.35]*	2.19 [0.20]*	**1.52 [0.42]**	18.52 [1.72]*	10.52 [4.55]*	9.16 [1.87]*	**7.26 [3.42]**
Steady-State (%)								
Generic	47.42 [1.55]*	1.03 [0.20]	2.54 [0.52]*	**0.54 [0.18]**	47.46 [3.18]*	6.99 [6.21]	14.18 [8.69]*	**5.45 [4.41]**
Mode-specific	**0.00 [0.00]***	1.32 [0.31]	1.19 [0.15]*	0.61 [0.23]	**0.00 [0.00]***	10.08 [6.01]	7.39 [3.03]*	5.14 [3.95]
Transitional (%)								
Generic	50.54 [2.00]*	30.26 [1.77]*	12.05 [1.76]*	**3.64 [0.98]**	50.46 [0.81]*	35.41 [7.67]*	25.08 [4.86]*	**18.08 [6.43]**
Mode-specific	100.00 [0.00]*	**4.18 [1.17]***	6.60 [0.99]*	5.54 [1.94]	100 [0.00]*	**12.37 [4.13]***	17.06 [5.46]*	17.03 [7.96]

Error rates (mean, [standard deviation]) of the generic and the mode-specific classifiers using the bilateral sensors with all modalities. Asterisks under random, LDA and SVM classifiers denotes statistically significant differences between the according classifiers and LIR-Net. The difference between the generic and the mode-specific configurations were all significant across all error types regardless of user-dependencies. Bold numbers represent the classifier with the lowest error rate for each type of error rate and configuration.

and mode-specific LDA classifiers. For overall errors, our proposed system achieved the lowest error rate (Dependent: [Generic: 1.1%, Mode-specific: 1.5%], Independent: [Generic: 7.7%, Mode-specific: 7.2%]) on both configurations and user-dependencies; whereas for the steady-state error, LIR-Net obtained the lowest error-rate on the generic configuration across all user-dependencies (Dependent: 0.5%, Independent: 5.4%) and the random classifier had the lowest error rate (0.0%) on the mode-specific configuration for both user-dependencies. In practice, the mode-specific random classifier simply predicted the current activity as the upcoming activity, which produced 0% error rates in steady-states. Since there were more steady-state than transitional cases in the dataset, the overall error rate of the mode-specific random classifier was lower than that of the generic classifier (Generic: 47.9 %, Mode-specific: 18.4%). The generic LIR-Net reached the lowest transitional error rates, while LDA had lowest transitional error within the mode-specific configurations across all user-dependencies.

3) Effect of User-Dependencies: We compared the error rates of the classifiers in the presence of different user-dependencies. The user-independent condition was statistically different from the user-dependent condition. For all classifiers, the error rates increased with the user-independent condition except the random classifier, the performance of which was governed by the distribution of the data (Tab. I).

4) Classification Latency of LIR-Net: The averaged latency was 136.07 ± 3.86 ms.

5) Comparison to ResNet: The overall error rate of the ResNet was compared with that of LIR-Net. The overall error rate of the pre-trained ResNet (1.29 ± 0.20%) was not statistically different (p = 0.13) to that of LIR-Net (1.11 ± 0.26%), which validated the choice of our network design.

B. Effect of Sensor Locations and Modalities on LIR-Net

The overall error rate of LIR-Net was statistically compared across all combinations of sensor laterality groups and modalities. The interaction between the modalities and laterality groups was significant. The effect of laterality groups was observed by comparing the classifier's performance with all sensor modalities combined (Tab. III). The error rate of the bilateral sensor set was statistically less (1.11%) than either ipsilateral or contralateral set. Similarly, the effect of sensor modalities was tested by comparing the performances of the classifier using the bilateral sensor set. The statistical significance was measured between all combined sensor modalities and individuals or combinations of two different sensor modalities. As a result, the error rate of the all combined sensor sets was significantly less than all individual sensor modalities and a pair of IMU and EMG sensors. For the single modalities with the bilateral sensor set, GONIO achieved the best performance, and for the two modalities, EMG and GONIO combination attained the lowest error rate.

TABLE III: LIR-Net performance on different sensor modalities and laterality groups

# of Sensor	Sensor Type	Ipsi (%)	Contra (%)	Bi (%)
1	I	4.68 [0.51]	5.99 [0.70]	2.58 [0.38]†
	G	3.85 [0.43]	4.00 [0.57]	**1.49 [0.29]**†
	E	7.66 [0.58]	8.00[0.51]	3.08 [0.46]†
2	I & G	3.05 [0.42]	3.33 [0.55]	1.29 [0.44]
	I & E	3.96 [0.55]	4.42 [0.52]	2.17 [0.71]†
	E & G	3.05 [0.55]	3.11 [0.32]	**1.15 [0.26]**
3	ALL	2.56 [0.50]*	2.89 [0.38]*	**1.11 [0.26]**

Overall error rates (mean, [standard deviation]) of LIR-Net using all possible combinations of laterality groups and sensor modalities: IMU (I), EMG (E), GONIO (G). The lowest error rates on each number of modalities are bolded.

* Asterisks under ipsilateral (Ipsi) and contralateral (Contra) classifiers denote the statistically significant differences between the according laterality and the bilateral (Bi) sensor set when all modalities are used.

† Daggers under sensor modalities denote the statistically significant differences between the according modality and all combined sensor (IMU & EMG & GONIO) with the bilateral sensors.

C. Visualizing Activations of LIR-Net

The network showed greater activations in the lower frequencies, where much of the information in the signals was localized (Fig. 5). Additionally, the greatest activations in the low frequency area occurred near to the gait events.

V. Discussion

In this paper, we proposed a CNN-based intent recognition system that utilized the spectrogram to represent the

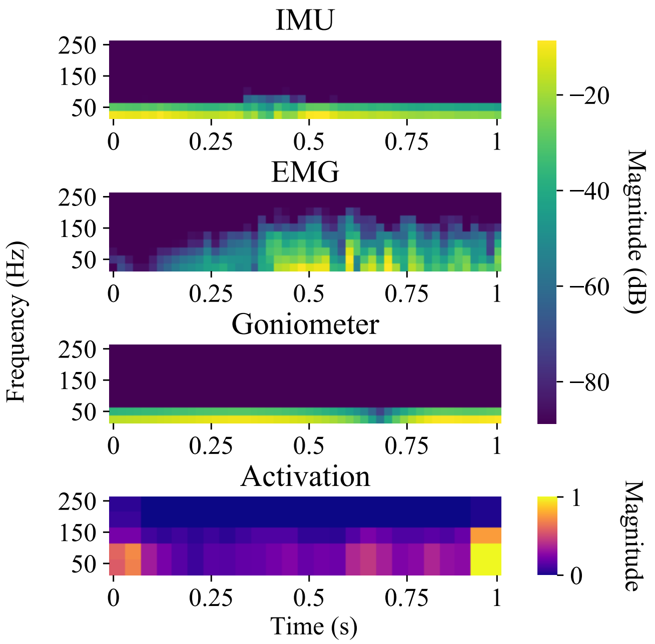

Fig. 5: A representative sample of the right shank IMU, right VL EMG, and right knee GONIO signals (top three), and activation visualization of LIR-Net (bottom).

frequency content of the input data. To this end, we studied the effect of sensor modalities and laterality groups on the proposed system, visualized the activation of CNNs, and compared our system to the state-of-the-art (SOTA) intent recognition classifiers [5], [19]. The overall error rate of our proposed system was 1.11% which exceeded the performance of the existing work in a generic classification scheme. The motivation of this paper is to improve the prediction capabilities of lower-limb locomotor intent recognition systems; ideally, providing a framework for autonomous wearable robots which can assist wearers with a diverse range of activities encountered in the real-world.

A. Limitations

The error rates of the user-independent classifiers were statistically higher than that of the user-dependent classifier. This result showed the classifiers, including our proposed system, were not able to generalize well to novel subjects. The performance reduction of our system was due to an increase in transitional error (LDA: +5%, SVM: +12%, LIR-Net: +14%) across user-dependencies, compared to the increase in steady-state error (LDA: +5%, SVM: +12%, LIR-Net: +5%). This relatively weak generalizability of LIR-Net to novel users in transitional states is likely due to the unbalanced number of activity samples in the ENABL3S dataset; specifically, the number of transitions were less than that of steady states (Tab. I). This stems from the fact that the data collection was conducted in a circuit that consisted of each activity, which is a convenient protocol, but may lead to sparseness in transition data [5]. Although our system obtained the lowest transitional error rates among all classifiers in the generic condition, the network had been trained and biased to lowering the overall error. In addition, deep learning (DL) generally performs better with more data, which may lead to greater improvements in performance when data are added, when compared to other classical machine learning algorithms. Thus, collecting more subject data with balanced number of samples, and techniques, such as data augmentation, can mitigate this limitation and improve the DL-based classifier [12].

B. Comparison to Past Works

Our proposed system was compared to the SOTA intent recognition systems using the ENABL3S dataset [5], [19]. For the combined mode-specific and user-dependent condition, our system performed comparably (1.52%) to the previous work of (1.43% [5]); but most importantly, in the generic classifier configuration, our work outperformed (1.11%) the heuristic-feature based classifiers (LDA: 6.43%, SVM: 4.30%) and CNNs with heuristic features as an input (3.7% [19]). Although mode-specific configurations could improve system performance, in real-world scenarios, relying on accurate knowledge of the previous step's activities (*i.e.* ground-truth) may be untenable. The performance of our intent recognition system in generic configuration demonstrates that our approach can be generalized across different environment conditions with various sequences of activities.

C. Effect of Sensor Locations and Modalities on LIR-Net

In general, as we fused more senor modalities and lateralites, the performance of the classifiers improved. This result was in accordance with the prior work using ENABL3S [5], [19]. Our findings showed GONIO had the best single modality performance with the bilateral sensor set; whereas IMU sensor data had the lowest error rate in previous works [5], [19]. For two modality sensors, EMG and GONIO combinations gave the best performance agreeing with the prior work [19]. Interestingly, the IMU and GONIO, EMG and GONIO combinations were not statistically different from all combined sensors, which suggests near optimal performance may be obtained from limited sensor selections.

D. Visualizing Activations of LIR-Net

Our technique of visualizing activations allowed a simple, but intuitive understanding of which features were learned by the network (Fig 5). The network had high activations nearby the gait events under the lower frequency region (<100 Hz). This shows that the information that dictates activity transitions are concentrated on the signals close to toe-off or heel-contact. To our knowledge, this is the first time that the activations of CNNs were qualitatively analyzed within a gait cycle, which is critical for identifying the intent using lower-limb neuromechanical signals. Although there was previous work visualizing the features of a CNN in lower limb sensor signals, the visualization was less intuitive and features were indistinguishable [18].

E. Application to Control of Wearable Robotics

Intent recognition is a control strategy which enables a wearable lower-limb robot to autonomously switch between controllers responsible for a specific task by inferring the wearer's locomotor intent. Typically, intent recognition is used as a high-level controller in a hierarchical control structure, where a mid-level controller encodes the activity-specific instructions for how to provide mechanical effort (*e.g.* via impedance or position control), and a low-level controller tracks the desired reference trajectories (*e.g.* feedback controller) [30]. A representative use case of the hierarchical controller is intent recognition in conjunction with mid-level finite-state controllers, where the gait cycle is divided by distinctive phases, and the transitions between these phases are based on heuristic rules. Since it is assumed that signals are stationary within each phase (*i.e.* identical activity), recognition-based classification strategies mitigate the time-varying characteristics of signals during the gait cycle [6].

The latency of our system was below what users may perceive (300 ms [31]) and within the time window required to ensure smooth transitions between activities following the gait events [32]. The latency can be further reduced by exploiting optimal sensor selection, and microcomputers with higher processing capabilities. Thus, this work demonstrates the usability of these techniques in real-time control of wearable robots.

ACKNOWLEDGMENT

The authors would like to thank to Dr. Gray Thomas of the Neurobionics Lab for his input on statistical analyses.

REFERENCES

[1] L. M. Mooney, E. J. Rouse, and H. M. Herr, "Autonomous exoskeleton reduces metabolic cost of human walking during load carriage," *Journal of neuroengineering and rehabilitation*, vol. 11, no. 80, 2014.

[2] A. F. Azocar, L. M. Mooney, L. J. Hargrove, and E. J. Rouse, "Design and characterization of an open-source robotic leg prosthesis," in *7th IEEE International Conference on Biomedical Robotics and Biomechatronics*, 2018, pp. 111–118.

[3] J. Zhang, *et al.*, "Human-in-the-loop optimization of exoskeleton assistance during walking," *Science*, vol. 356, no. 6344, pp. 1280–1284, 2017.

[4] A. J. Young and L. J. Hargrove, "A classification method for user-independent intent recognition for transfemoral amputees using powered lower limb prostheses," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 24, no. 2, pp. 217–225, 2015.

[5] B. Hu, E. Rouse, and L. Hargrove, "Fusion of bilateral lower-limb neuromechanical signals improves prediction of locomotor activities," *Frontiers in Robotics and AI*, vol. 5, no. 78, 2018.

[6] A. J. Young, A. M. Simon, N. P. Fey, and L. J. Hargrove, "Intent recognition in a powered lower limb prosthesis using time history information," *Annals of biomedical engineering*, vol. 42, no. 3, pp. 631–641, 2014.

[7] H. A. Varol, F. Sup, and M. Goldfarb, "Multiclass real-time intent recognition of a powered lower limb prosthesis," *IEEE Transactions on Biomedical Engineering*, vol. 57, no. 3, pp. 542–551, 2009.

[8] H. Huang, *et al.*, "Continuous locomotion-mode identification for prosthetic legs based on neuromuscular–mechanical fusion," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 10, pp. 2867–2875, 2011.

[9] T. Plötz, N. Y. Hammerla, and P. L. Olivier, "Feature learning for activity recognition in ubiquitous computing," in *22nd International Joint Conference on Artificial Intelligence*, 2011, pp. 1729–1734.

[10] M. Zeng, *et al.*, "Convolutional neural networks for human activity recognition using mobile sensors," in *6th IEEE International Conference on Mobile Computing, Applications and Services*, 2014, pp. 197–205.

[11] T. Hur, *et al.*, "Iss2image: A novel signal-encoding technique for cnn-based human activity recognition," *Sensors*, vol. 18, no. 3910, 2018.

[12] B. Hu, A. M. Simon, and L. Hargrove, "Deep generative models with data augmentation to learn robust representations of movement intention for powered leg prostheses," *IEEE Transactions on Medical Robotics and Bionics*, vol. 1, no. 4, pp. 267–278, Nov 2019.

[13] Y. Chen and Y. Xue, "A deep learning approach to human activity recognition based on single accelerometer," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 1488–1492.

[14] A. Zunino, *et al.*, "Predicting intentions from motion: The subject-adversarial adaptation approach," *International Journal of Computer Vision*, vol. 128, no. 1, pp. 220–239, 2020.

[15] D. Ravi, C. Wong, B. Lo, and G.-Z. Yang, "Deep learning for human activity recognition: A resource efficient implementation on low-power devices," in *2016 IEEE 13th international conference on wearable and implantable body sensor networks*, 2016, pp. 71–76.

[16] T. T. Um, *et al.*, "Data augmentation of wearable sensor data for parkinson's disease monitoring using convolutional neural networks," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 216–220.

[17] R. Zhang and C. Li, "Motion sequence recognition with multi-sensors using deep convolutional neural network," in *Intelligent Data Analysis and Applications*. Springer, 2015, pp. 13–23.

[18] B.-Y. Su, *et al.*, "A cnn-based method for intent recognition using inertial measurement units and intelligent lower limb prosthesis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 27, no. 5, pp. 1032–1042, 2019.

[19] K. Zhang, J. Wang, C. W. de Silva, and C. Fu, "Unsupervised cross-subject adaptation for predicting human locomotion intent," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 3, pp. 646–657, 2020.

[20] B. Hu, E. Rouse, and L. Hargrove, "Benchmark datasets for bilateral lower-limb neuromechanical signals from wearable sensors during unassisted locomotion in able-bodied individuals," *Frontiers in Robotics and AI*, vol. 5, no. 14, 2018.

[21] D. Anguita, *et al.*, "A public domain dataset for human activity recognition using smartphones." in *Esann*, vol. 3, 2013, pp. 437–442.

[22] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Sep. 2015, pp. 1–6.

[23] D. A. Winter, *Biomechanics and motor control of human movement*. John Wiley & Sons, 2009.

[24] B. McFee, *et al.*, "librosa: Audio and music signal analysis in python," in *Proceedings of the 14th python in science conference*, vol. 8, 2015, pp. 18–25.

[25] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *3rd International Conference on Learning Representations, ICLR 2015, Conference Track Proceedings*, 2015.

[26] A. Paszke, *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems*, 2019, pp. 8024–8035.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] J. Deng, *et al.*, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[29] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[30] M. R. Tucker, *et al.*, "Control strategies for active lower extremity prosthetics and orthotics: a review," *Journal of neuroengineering and rehabilitation*, vol. 12, no. 1, 2015.

[31] K. Englehart and B. Hudgins, "A robust, real-time control scheme for multifunction myoelectric control," *IEEE transactions on biomedical engineering*, vol. 50, no. 7, pp. 848–854, 2003.

[32] F. Zhang, M. Liu, and H. Huang, "Investigation of timing to switch control mode in powered knee prostheses during task transitions," *PLOS one*, vol. 10, no. 7, 2015.