Vision for Prosthesis Control Using Unsupervised Labeling of Training Data

Vijeth Rai [1] *Member, IEEE,*, David Boe [1] *Member, IEEE,* and Eric Rombokas [1,2] *Member, IEEE*

***Abstract*— Transitioning from one activity to another is one of the key challenges of prosthetic control. Vision sensors provide a glance into the environment's desired and future movements, unlike body sensors (EMG, mechanical). This could be employed to anticipate and trigger transitions in prosthesis to provide a smooth user experience.**

A significant bottleneck in using vision sensors has been the acquisition of large labeled training data. Labeling the terrain in thousands of images is labor-intensive; it would be ideal to simply collect visual data for long periods without needing to label each frame. Toward that goal, we apply an unsupervised learning method to generate mode labels for kinematic gait cycles in training data. We use these labels with images from the same training data to train a vision classifier. The classifier predicts the target mode an average of 2.2 seconds before the kinematic changes. We report 96.6% overall and 99.5% steady-state mode classification accuracy. These results are comparable to studies using manually labeled data. This method, however, has the potential to dramatically scale without requiring additional labeling.

I. Introduction

Successful prosthesis control has two important requirements: 1) Understanding user intent, 2) Understanding the dynamic demands of the environment and movement activity. Neuromuscular-mechanical data gathered from electromyogram (EMG) and mechanical sensors are ideally suited for the first requirement as they provide a direct window into the user state. To explicitly sense the environment, vision as a sensing modality has recently garnered interest [1, 2, 3, 4].

Current powered prosthesis control is generally based on the categorization of the environment into finite classes or "modes", such as flat ground or stair ascent. Sensor data is used as inputs to a trained classifier to predict the right mode to match the environmental demands at any given time.

EMG implicitly senses the environment by relying on user actions, which are modulated by the user in response to the environment. EMG is also a user-dependent sensor requiring subject-specific calibration and training. Its ability to provide a window into the future actions of the user and the upcoming environment changes is limited by the timing of the user's actions. This has posed a challenge for anticipating the right locomotion mode, especially during transitions [5, 6]. However, [7] have shown that augmenting neuromuscular-mechanical data with direct environment information improves performance.

A. Vision for Explicitly Sensing the Environment

Humans use vision to sense the environment and fluidly adapt to upcoming changes. Even with a wide variability of terrains, humans consistently look about 1.5 seconds ahead of their current location [8]. This is similar to look ahead timing seen in research on other motor actions, such as stair climbing [9]. This means that the environmental information present in vision data will generally precede kinematics data. This look-ahead window, known as lead-time, is the main benefit of using vision to anticipate transitions.

Computer Vision is a user-independent sensor that could improve classification accuracy by directly sensing the environment [10]. It can anticipate transitions in advance to trigger a change in controls at the right moment. It would also improve the overall robustness of the system by not relying purely on user-dependent sensing modalities [11, 12].

In the neuromuscular-mechanical sensing-based approach, machine learning is employed for pattern recognition of data for mode classification. Likewise, machine learning is also used in state-of-the-art image classification tasks, as well as vision-based locomotion mode research. Support Vector Machines [13] and Finite State Machines [2] have been used for mode classification from depth camera data. These methods have the advantage of being computationally efficient. At the expense of greater computation, more accurate results have been obtained using variants of deep learning models. This approach e.g. [1, 3], has yielded an overall classification accuracy (OCA) of 95% but transitions were relatively worse with about 89% CA. The training data for all studies were manually labeled which has presumably led to a low sample size of subjects and training data.

B. Issues with Using Vision for Control

These promising successes suggest that vision can play a valuable role in prosthetic control. One of the key bottlenecks for realizing this promise is the difficulty of manually labeling large quantities of training data. Here, we explore the use of unsupervised techniques to automatically create pseudo-labels of modes, solely learned from the kinematic regularities that appear simultaneously with the captured images.

To train a vision-based mode classifier requires images with corresponding prosthetic activity mode labels as targets. Studies currently employ manual methods of labeling, based on physical location (kinematics) or visible features in image data. One challenge with this approach is that there is human labeling subjectivity, especially about when a transition actually occurs[4]. Consider an image as the user approaches the stairs. Exactly when does the transition occur? These choices can yield variability and inter-rater discrepancies in the labels. Overall, this process is time and resource-consuming, even for a modest number of modes. For ex-

[1] V. Rai and E Rombokas are with the Department of Electrical and Computer Engineering, University of Washington Seattle,WA, 98102 USA e-mail: (raiv@uw.edu).

ample, Laschowski et al. [4] manually labeled 37000 vision frames for data from a *single* subject with 3 mode classes. This is a titanic effort and the challenge for proceeding with this strategy is clear.

Practical industrial application of deep learning requires a large training dataset. Comparing other domains that use vision for robotic control [14], the size of training data in the recent prosthetic control studies have been far fewer. This can limit generalization as the deep learning models learn only the features of the limited training data used in that study. This also prevents comparisons between classification algorithms from different researchers [15]. Exonet [16] is currently the only publicly available data pertaining to the field of prostheses and exoskeletons, where approximately 923,000 images were manually labeled and organized into 12 locomotion mode classes.

We apply two techniques from the field of machine learning to mitigate these issues. We use an unsupervised learning strategy to automatically acquire labels for the training images. To improve generalization and accuracy, we transfer knowledge learned from bigger, publicly available image dataset by a technique known as Transfer Learning.

C. Unsupervised Labelling and Transfer Learning

An unsupervised machine learning algorithm is used to draw inferences from datasets consisting of data without labeled responses. We employ a technique known as *cluster analysis*. Cluster analysis in general is a longstanding and mature field. For analysis of human movement, it has been successfully applied to identify patterns of gait deviations in children with cerebral palsy [17], and to distinguish the walking parameters of young from elderly subjects [18].

We use time-normalized knee gait cycles as input features. Knee gait cycles are clustered, or grouped into similar classes, on a similarity-based distance metric. These cluster labels constitute ”modes” by virtue of their similarity in the data. The regularities present in natural movements dictate the grouping of images of the environment. This allows for modes to arise naturally and with minimal bias from human interpretation. The auto-generated labels are then used to train a vision classifier to detect upcoming modes and transitions.

To demonstrate the benefits of using vision, we calculate the lead time of vision-based predictions relative to those based only on mechanical sensors. Since the visual features of new terrain are visible before the body mechanically responds to it, we hypothesize that we can achieve lead times similar to typical human look-ahead duration [8]. We report the overall and steady-state classification accuracy of the vision classifier for a test set with manually labeled terrain labels.

To summarize, the contributions described in this manuscript are:

1) Demonstration of an unsupervised model to automatically label gait data.
2) Application of those labels as targets to train a vision classifier.
3) Quantification of accuracy and lead time using the vision classifier compared to mechanical sensor (kinematic) data.

II. Methods

A. Data Collection and Experiments

Ambulation data was collected for a total of 10 healthy participants with no amputation or other mobility impairments. Recruitment and human subject protocols were performed in accordance with the local University of Washington Institutional Review Board approval and each subject provided informed consent.

The subjects' anthropometric details, such as height, body segment lengths, etc.) were recorded and 17 wearable motion capture sensors (see Instrumentation below) were placed on their body. This was followed by a calibration procedure and a brief test to see the quality of data being recorded. A head-mounted ego-centric camera from Pupil Labs [19] was calibrated and used to collect visual data.

Subjects performed 10-15 minute trials of ambulation tasks, including transitions between them. The subjects were instructed to walk naturally at a self-selected pace. The data for these activities was collected in public spaces during active business hours, with the intent that normal gait dynamics and corrections would appear in the example data.

Each trial started with participants performing flatground walking in a cluttered classroom environment with chairs, followed by sections of flatground walking in open corridors. The participants were then instructed to ascend a flight of stairs to reach the next level, which involved flatground walking in corridors. Participants returned to the original level by descending the flight of stairs. There were brief sections of atypical movements such as opening the classroom door to enter the corridor section, short sections (2-3 steps) of flatground in between flight of stairs. There were also instances in which the participants needed to navigate around other people walking in the spaces.

B. Instrumentation

We collected locomotion data using the Xsens Awinda suit [20], consisting of 17 body-worn sensors placed at key locations. Each sensor has a tri-axial gyroscope, accelerometer, magnetometer, and barometer. Xsens Analyse software integrates these individual sensors and renders a full-body avatar. After a system specified calibration, the software provides position and joint kinematics in a 3D environment. Although other data such as limb segment position, orientation, acceleration are available, we used only joint angles for this study. All angles are in 1x3 Euler representation of the joint angle vector (x, y, z) in degrees, calculated using the Euler sequence ZXY using the International Society of Biomechanics standard joint angle coordinate system [21]. Data, sampled at 60 Hz, from a total of 22 joints in 3 anatomical planes (sagittal, frontal, transverse) were captured for each trial, which results in 66 total possible features for our machine learning methods.

A head-mounted ego-centric camera from Pupil Labs [19] was used to collect visual data at 30fps and a resolution of 640x480 pixels.

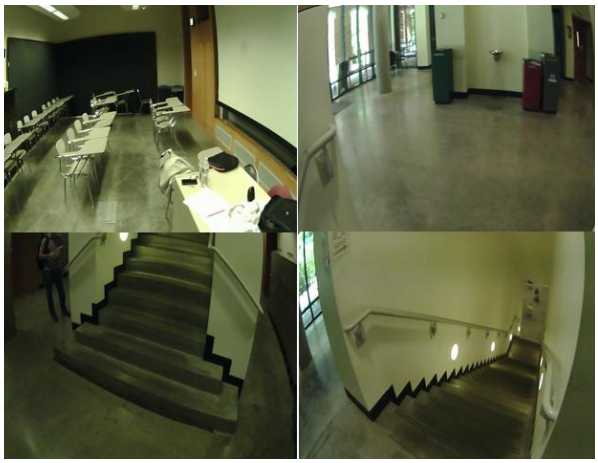

Fig. 1: Examples of images captured by the head-mounted pupil camera. 4 types of terrain were part of the environment - flatground walking with and without obstacles, stair ascent and descent.

C. Clustering for Unsupervised Labeling of Training Data

We apply Hierarchical Agglomerative Clustering (HAC) to cluster knee gait cycles based on relative similarity measured using euclidean distance.

Lack of 'correct' labels or responses makes objective validation of cluster results challenging. However, quantitative and qualitative methods, including visual analysis of resulting clusters have been used to successful ends in prior studies [17, 18, 22]. We verified the optimal number of clusters using several metrics such as R-ratio, Dunn Index, Silhouette score.The number of resulting clusters dictate the locomotion mode labels available to the vision classifier (See Section II-D). For brevity, we report only R-ratio results.

1) Data-processing: Lower-limb joints as input features have consistently shown the most predictive power in gait-related applications of clustering algorithms[17]. The input to our clustering models was the right knee joint kinematics. Segmented gait cycles were temporally normalized to 100 percent. Each gait cycle was considered a single sample, in total yielding 6766 gait cycles in the training data from 9 subjects. Relative euclidean distance between each sample was used to measure similarity for the clustering process.

Due to the lack of insole data, knee gait cycles were segmented by using the contralateral knee peak flexion as cutoffs (Fig.2). Similar methods have been used in other studies to segment gait data without using force plate data [23]. Although this method lacks the moment-to-moment precision afforded by force-plate data, it is sufficient for clustering entire gait cycles. MATLAB command $findpeaks$ was used for determining the peaks. A threshold value of 15 degrees was used for this data.

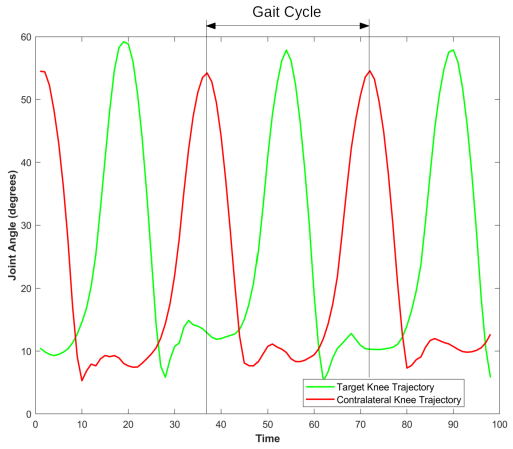

Fig. 2: Knee gait cycles were segmented using the peak flexion angle of the contralateral knee joint as beginning and end of the cycle. Although an imperfect method, it yields qualitatively consistency results amenable for automatic gait categorization via clustering.

2) Unsupervised Machine Learning: Clustering algorithms can be broadly classified as hierarchical methods or non-hierarchical methods such as K-Means. The HAC algorithm we use is a type of hierarchical procedure in cluster analysis which are characterized by the rule-based development of a tree-like structure known as dendrogram. The dendrogram aids in visualizing and interpreting the resulting cluster groups. Going from the top, the dendrogram allows informed decisions about the similarity of the occurring patterns or cluster groups in the data. Leaves represent the final clusters. Leaves in the same branch are similar in terms of linkage distance between the clusters.

3) Analysis: We use the R Ratio and within-cluster sum of squared error as a means to select the optimal number of cluster groups in the training data.

R-ratio is used by prior studies as a computational metric to determine the optimal number of clusters [24, 22]. This is a measure of the reduction of the within-cluster variability.

$$R = \left[\frac{e(N,K)}{e(N,K+1)} - 1\right](N-K+1)$$

where e(N, K) is defined as the summation of within-cluster square distances for N patterns and K clusters. The algorithm is stopped when the R ratio peaks, corresponding to a large reduction in the within-cluster variability. This indicates that the clusters are quite homogeneous.

Along with computation methods like R-ratio, a final visual inspection is generally considered crucial [17]. We use the mean kinematic pattern of each of the resulting clusters for visual analysis. These represent the different environments evidenced by differing knee kinematics. In the next section, we show these mean kinematic patterns and the dendrogram of the clustering process to elucidate relative similarity between groups.

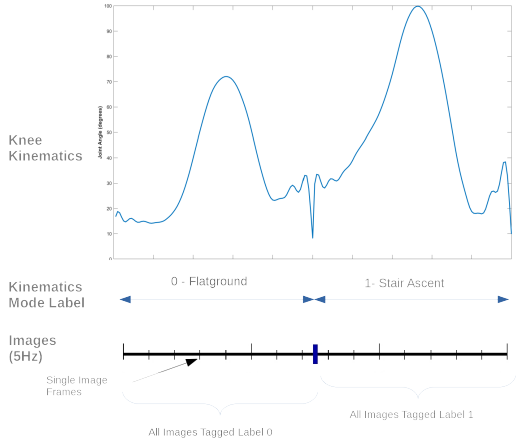

Fig. 3: Each gait cycle was assigned a label using clustering. All images corresponding to the duration of the gait cycle were tagged with the same label for the purposes of training the vision classifier.

D. Vision Classification

An 18-layer Residual Convolutional Network pre-trained on ImageNet data [25] is fine-tuned to predict 3 locomotion mode classes based on image data.

1) Data Processing:

Image labeling: Image data were collected using the head-mounted camera and down-sampled to 5Hz. An unsupervised cluster model was used to generate mode labels for every gait cycle (See Section II-C). An average human gait cycle duration is about 1-second [26]. At a sample rate of 5Hz for image data, a single gait cycle contains about 5 images. All images corresponding to the duration of the gait cycle are tagged with the same label as shown in Fig 3.

2) Data: A multi-modal sensor data brings in challenges in training and evaluation.

a) Kinematic and Vision Labels: Current research on vision for prosthesis control distinguishes between 2 types of labels or ground truth for the data. Terrain labels can be based on

i Current kinematic behavior using foot position.
ii Visible features of the terrain in image data.

Since visible features of the new terrain can be in the image from a long distance, vision-based labeling is prone to subjective bias [13, 4]. For example, while approaching a flight of stairs in the distance, the exact vision frame to switch the mode from flatground to stair ascent can be unclear. More importantly, for our purpose, this also means that vision-based labels will generally precede kinematics-based labels. This lead-time is the main benefit of using vision to anticipate transitions.

b) Steady-State and Transition Data: For training and performance evaluation a common trend in prosthetic control research is to separate steady-state and transition sections [27, 13]. Steady-state comprises sections of data without any terrain or activity changes, where a periodic gait cycle is repeated. Transitions consist of sections with locomotion mode changes. For e.g [3] consider 5 seconds of data before terrain change as transition data. Other studies use gait events such as mid-swing for stair ascent or heel strike for flat ground transitions as deadlines to assess performance[2]. We use a shorter section of 3 seconds before mode label change as the transition section. The performance evaluation of transition, however, is less clearly established, which we discuss below.

3) Machine Learning: We train a ResNet-18 a convolutional neural network that is 18 layers deep[28] with image data to classify terrains into 3 classes(flatground, stair-ascent, and stair descent). However, the initial weights of this Base Model is randomly initialized as is common practice.

To improve generalization and to offset a relatively small training data size, we apply a technique called *Transfer Learning*. This is achieved by training a model on a large dataset (eg. ImageNet) and then transferring the learned knowledge (features, weights) by fine-tuning the model to a different dataset. The features learned from the large dataset usually carries over to the new dataset and improves accuracy compared to a randomly initialized model[29].

The pre-trained model is adapted or 'fine-tuned' to the new dataset by training only the last layer, known as the head, responsible for generating the class labels. The rest of the model, known as the body or backbone, is frozen to retain the learning from the large dataset.

The network is pre-trained on more than a million images from the ImageNet database [25] with images from 1000 object categories, such as a keyboard, mouse, pencil, and many animals. As a result, the network has learned rich feature representations for a wide range of images. We fine-tune this network by further retraining on our data to classify terrains into 3 classes(flatground, stair-ascent, and stair descent). The network has an image input size of 224-by-224.

4) Training and Inference Process:

a) Training: The training process uses labels generated by the unsupervised model to train the vision classifier.

1) Knee gait cycles in the training dataset are clustered into an optimal number of clusters using an unsupervised clustering algorithm. Each group is assigned a kinematic mode label.
2) Each gait cycle has several corresponding images (Fig 3). The label associated with a gait cycle is used to tag all the corresponding images.
3) Vision Classifier is trained with images as inputs and corresponding labels.

Training with Steady-State Data Only: The approach described above relies on the transference of kinematic-based labels to vision data. Kinematic sensor data are inherently delayed and hence the labels acquired by the clustering method represent the mode class *after* transitioning to a new terrain type. As such, this labeling scheme is only valid for steady-state data.

The vision classifier predicts labels based on the visible features of current or upcoming terrain in the image data. To induce clear binary distinctions between terrains during

training, only steady-state data is used. About 3 seconds of data before transitions are considered 'transition' data and omitted from training.

For an unseen test subject approaching a stair ascent terrain, the image data will be classified as stair ascent mode based on visible features of stairs. The predicted label will lead the kinematic-based actions and is one of the key benefits of vision sensors.

b) Inference:: Since a new terrain will be visible before body kinematics adapt, the predicted vision label (V_t) will lead the kinematic based $\hat{K_t}$. We define the lead time as

$$\Delta t = \mid \hat{V_t} - \hat{K_t} \mid$$

1) During inference, images are forwarded to the CNN classifier to predict vision labels$\hat{V_t}$
2) For steady-state gait, the predicted $\hat{V_t}$ will match the $\hat{K_t}$ mode label.
3) For an upcoming transition, new terrain will be visible before the kinematics change to adapt. Hence the mode labels predicted by the CNN classifier precede kinematic-based labels by a lead-time of Δt. That is $\hat{V_t}$ will match the mode class of $\widehat{K_{t+\Delta t}}$

5) Performance Evaluation and Analysis: Vision-based prosthesis studies employ 3 evaluation metrics to evaluate various characteristics of resulting performance. Overall classification accuracy (CA) is the percentage of accurately labeled images relative to all images in the test data. Steady-state classification accuracy is similar but omits transition sections of the data.

Evaluation of performance for transitions is not clearly defined, and several methods are observed in studies In EMG studies [30], specific gait events such as toe-off from the level ground to upstairs and heel contact from downstairs to the level ground are used to estimate the deadline of locomotion transition. Prediction lead time with respect to this deadline is used to quantify the anticipatory response of the systems. [2] use a similar approach by segmenting gait and use the mid-swing as the critical deadline. However, they report the percentage of transitions detected before the foot lift of the leading limb. [1] report lead time for a CNN-based vision classifier with respect to kinematic based labels. [3] defined the transition period as the 5 second period before a transition and report terrain classification accuracy percentage during these periods. [13] report percentage of total transitions detected.

We report overall CA, steady-state CA, and the lead time of the vision-based label with respect to the kinematic labels for every terrain type.

III. Results

The system noted 99% steady-state and 96% overall accuracy of mode classification using the vision sensor. Transitions were detected with a best-case average of 2.2 secs before kinematics changed to adapt to the new terrain.

A. Unsupervised Labeling using Clustering

R ratio peaked at K=3 number of clusters indicating 3 dominant kinematic patterns.

The parameter K, the number of desired clusters decides the truncation of the clustering process shown as the tree-like dendrogram (Fig4). This decision can be based on prior knowledge of the terrain types in the training data. Analyzing the dendrogram and the resultant cluster groups could also lend insight.

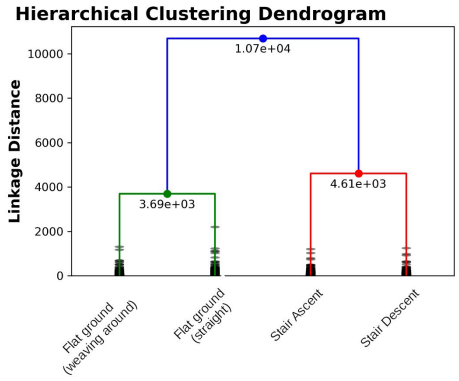

Fig. 4: The dendrogram is a graphical representation of the cluster groups and their relationships. E.g., with group labels starting with 0 from the left, cluster groups 0 and 1 belong to the same branch. Cluster groups 2 and 3 are distinct from groups 0 and 1. Visual analysis of the patterns in each of the groups reveals that groups 0 and 1 correspond to flatground in two different scenarios. Group 2 cluster contains all the stair ascent samples and group 3 was determined to contain stair descent samples.

In our case, the desired level of resolution of the environment was constrained to flatground, stair-ascent, and stair descent activities. This matched the peak R-ratio. Hence, the clustering process was halted when 3 cluster groups remained. The average kinematic patterns extracted in the 3 groups are shown in Fig 5. These patterns correspond to flatground, stair-ascent, and stair descent respectively.

If the dendrogram is truncated at a higher level, the process is halted earlier, to result in 4 clusters as shown in Fig 4. The fourth cluster, in this case, corresponding to a variant of flatground walking that included avoiding obstacles (Fig 7).

B. Vision Classification

The performance of the vision classifier was evaluated on a test set with manually labeled ground truth. Fig 6 shows the predicted and actual mode labels for unseen test subject data. Transitions and lead times are also shown.

The transfer learning classifier pre-trained on the ImageNet data achieved an overall CA of 96.6% and steady-state CA of 99.5%. In comparison, the classifier initialized with random weights had an overall CA of 95.7% and a steady-state accuracy of 99.14%.

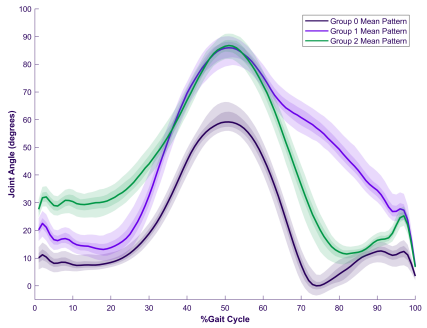

Fig. 5: Mean patterns and 25% percentile of samples when the number of clusters is chosen to be K=3

	Transition	Base Model	Tranfer-Learning Model
a	FG - SA	0.8s	1.1s
b	SA - FG	2.5s	2.6s
c	FG -SD	1.5s	1.0s
d	SD -FG	4.5s	2.5s
e	FG -SA	1.1s	1.0s
f	SA -FG	3.1s	2.8s
g	FG - SD	1.8s	1.8s

TABLE I: Transition lead-times for the base model (with randomly initialized weights) and the transfer-learning model (with pre-trained weights). Base model had a better average lead time of 2.2 seconds compared to 1.8 seconds achieved by the transfer-learning model. Certain particular transitions such as stair-ascent (SA) to flat ground (FG) and stair-descent (SD) to flatground had relatively much lead time with the base model.

The base model (randomly initialized) had a better average lead time of 2.2 seconds than the transfer learning classifier with an average lead time of 1.8 seconds. The lead-times for some of the transitions such as transitioning to flat-ground from stair ascent (f in Fig 6) and stair descent (d) was markedly better for the base model (Table I). These results were contrary to our hypothesis. We discuss the small variability in the test set as a limitation below.

IV. Discussion

Vision sensors are a great way to explicitly sense the environment. It allows looking ahead at upcoming transitions to anticipate change in control and improve prosthetic performance. However, for robust deployment, training requires labeling examples which can be a laborious process. We show here that a vision classifier can be trained using machine-generated labels and achieve similar performance compared to manually labeled data.

While training, neural networks learn general features of images to classify data. We train a CNN classifier on pure examples of the terrain types by eliminating transition sections from training data. This helps build a clearer understanding of the corresponding terrain type e.g. stairs in the network. While approaching a flight of stairs, the vision data shows features of stairs (albeit mingled with other segments of flatground in the image) almost 5 - 7 seconds before the user takes the first step. The visual features of the new terrain get gradually more distinct as the distance to the stairs reduces. This eventually triggers a label change in the vision classifier predictions, well before the bio-mechanical changes in the user. We noted best-case lead-time of +4.5 secs before a terrain transition and worst case lead-time of +0.8 seconds (Table I). In all cases the vision classifier detected terrain change before the actual transition.

EMG activity precedes a change in physical behavior with less than 100ms lead time [31]. This precludes its usage to detect the environment and predict locomotion modes of amputees in advance. EMG is also limited in its ability to recover from errors as the system is relying solely on the user state to indirectly gauge the demands of the current environment [32]. Explicit sensing of the environment to anticipate changes will improve robustness and overall performance [11]. This is achievable using vision sensors. It is also a user-independent sensor and would allow off-the-shelf usage in locomotion mode recognition systems [13].

The unsupervised clustering of gait cycles was shown to be a viable method to derive terrain labels for images with minimal human intervention. The categorization of data is dictated by the uniqueness of the movements thereby eliminating subjective bias.

The dendrogram (Fig 4) shows the last 4 clusters of gait patterns present in the data. The clustering algorithm differentiates between flat ground walking in long corridor sections versus flat ground walking inside classrooms (See top row of Fig 1). The latter which involved avoiding chairs (shown as Group 3 in Fig 7) could be considered a different 'mode' for prosthetic control. Most manual labeling schemes would generalize these into a single flat ground mode despite their slight differences. With unsupervised labeling, the choice rests upon the researcher/clinician. When this level of distinction is not desired, the group will be merged into a single 'flatground' group, still distinct from stair ascent and descent.

We hypothesized that transfer learning will be able to improve performance and generalization. The transferred model was pre-trained on a million images of the ImageNet dataset. Our results showed that transfer learning improved classification accuracy from 95.6% to 96.7% for 3 modes. This OCA is comparable to the best results from studies with manually labeled training data (95% [1] and 98% [3]). But as authors of [4] noted individual mode accuracy matters. Most datasets, including ours, are unbalanced consisting mostly (80%+) of flatground.

The average lead time for transitions was better for the randomly initialized model (2.2 secs vs 1.8 secs). Transition to flatground from both SA and SD had much larger lead times which contributed to a larger average lead time (Table I). This could be due to the classifier over-fitting to the training set. The size and terrain classes of the test set were limited and were very similar to the training set, especially for flatground examples. Data with a wider range of environments will be

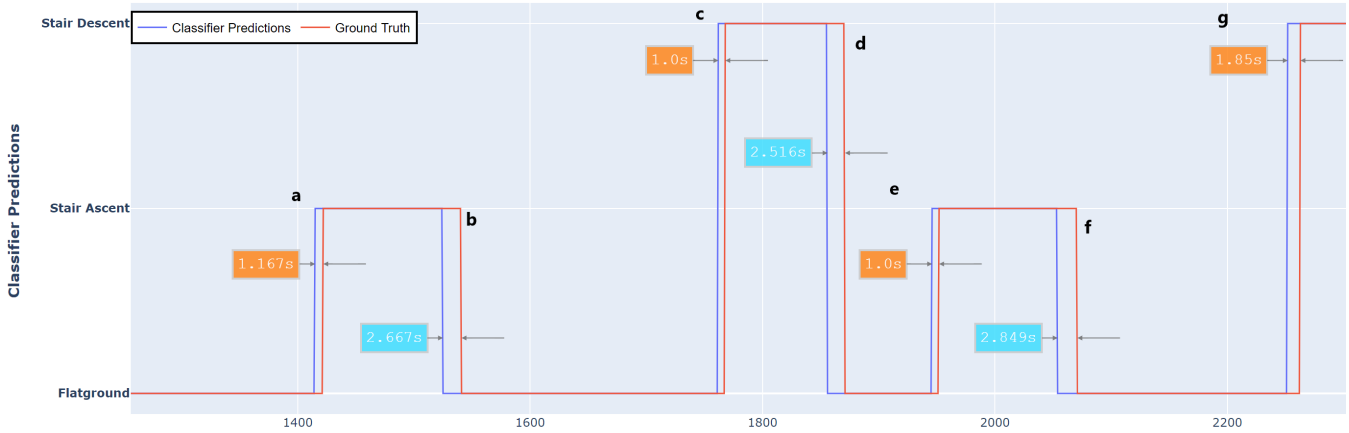

Fig. 6: Predicted and actual mode labels for an unseen test subject data. Transfer learning based model showed overall classification accuracy of 96% and an average transition lead-time of 1.8 seconds. In contrast, the base model with randomly initialized weights had an OCA of 95% and 2.2 seconds. Lead-time is defined as the time elapsed between the vision system detecting a change in terrain and the body kinematics of the user changing to adapt to the new terrain.

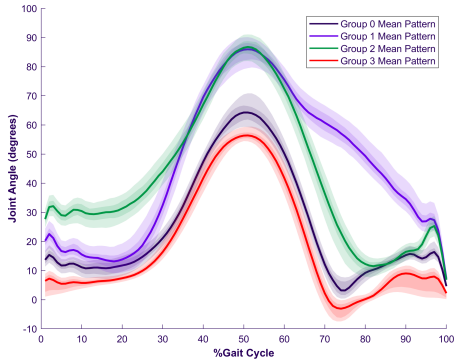

Fig. 7: Mean patterns and 25% percentile of samples when the number of clusters is chosen to be K=4

better to evaluate the transfer learning benefits. As future work, the classifiers will be evaluated on ExoNet [15] with 12 classes.

This study improves current methods for obtaining environment information using vision. However, relying solely on vision data for control is not a recommended option since the user gaze data can shift unexpectedly to glance and inspect the environment. Vision information would supplement, rather than replace, the control decisions from neuromuscular-mechanical data. When integrated with control hardware of lower-limb prostheses, it could be useful to minimize the search space of possible movements well ahead of the actual transition. This lead-time would allow controllers to provide a smoother experience for the end-user.

V. Conclusions

Advances in computer vision and artificial intelligence are allowing prostheses to adapt more intuitively to dynamic environments. However, small-scale training datasets have hindered the widespread development and deployment. We show here that kinematic data can be used to acquire machine-generated labels to train vision classifiers with minimal human intervention. Performance can be improved by transferring learned features from publicly available large-scale datasets.

References

[1] Kuangen Zhang et al. “Environmental features recognition for lower limb prostheses toward predictive walking”. In: *IEEE transactions on neural systems and rehabilitation engineering* 27.3 (2019), pp. 465–476.

[2] Tingfang Yan et al. “A locomotion recognition system using depth images”. In: *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2018, pp. 6766–6772.

[3] Boxuan Zhong et al. “Environmental Context Prediction for Lower Limb Prostheses With Uncertainty Quantification”. In: *IEEE Transactions on Automation Science and Engineering* (2020).

[4] Brock Laschowski et al. “Preliminary design of an environment recognition system for controlling robotic lower-limb prostheses and exoskeletons”. In: *2019 IEEE 16th International Conference on Rehabilitation Robotics (ICORR)*. IEEE. 2019, pp. 868–873.

[5] Levi J Hargrove et al. “Robotic leg control with EMG decoding in an amputee with nerve transfers”. In: *New England Journal of Medicine* 369.13 (2013), pp. 1237–1242.

[6] Huseyin Atakan Varol, Frank Sup, and Michael Goldfarb. “Multiclass real-time intent recognition of a powered lower limb prosthesis”. In: *Biomedical Engineering, IEEE Transactions on* 57.3 (2010), pp. 542–551.

[7] Fan Zhang et al. “Preliminary design of a terrain recognition system”. In: *Engineering in medicine and biology society, EMBC, 2011 annual international conference of the IEEE*. IEEE. 2011, pp. 5452–5455.

[8] Jonathan Samir Matthis, Jacob L Yates, and Mary M Hayhoe. "Gaze and the control of foot placement when walking in natural terrain". In: *Current Biology* 28.8 (2018), pp. 1224–1233.

[9] Aftab E Patla and Joan N Vickers. "How far ahead do we look when required to step on specific locations in the travel path during locomotion?" In: *Experimental brain research* 148.1 (2003), pp. 133–138.

[10] A. J. Young and L. J. Hargrove. "A Classification Method for User-Independent Intent Recognition for Transfemoral Amputees Using Powered Lower Limb Prostheses". In: *IEEE Transactions on Neural Systems and Rehabilitation Engineering* 24.2 (2016), pp. 217–225.

[11] Michael R Tucker et al. "Control strategies for active lower extremity prosthetics and orthotics: a review". In: *Journal of neuroengineering and rehabilitation* 12.1 (2015), p. 1.

[12] Michael Tschiedel, Michael Friedrich Russold, and Eugenijus Kaniusas. "Relying on more sense for enhancing lower limb prostheses control: a review". In: *Journal of NeuroEngineering and Rehabilitation* 17.1 (2020), pp. 1–13.

[13] Yerzhan Massalin, Madina Abdrakhmanova, and Huseyin Atakan Varol. "User-independent intent recognition for lower limb prostheses using depth sensing". In: *IEEE Transactions on Biomedical Engineering* 65.8 (2017), pp. 1759–1770.

[14] Sergey Levine et al. "End-to-end training of deep visuomotor policies". In: *The Journal of Machine Learning Research* 17.1 (2016), pp. 1334–1373.

[15] B. Laschowski et al. "Comparative Analysis of Environment Recognition Systems for Control of Lower-Limb Exoskeletons and Prostheses". In: *2020 8th IEEE RAS/EMBS International Conference for Biomedical Robotics and Biomechatronics (BioRob)*. 2020, pp. 581–586. DOI: 10.1109/BioRob49111.2020.9224364.

[16] Brock Laschowski et al. "ExoNet Database: Wearable Camera Images of Human Locomotion Environments". In: *Frontiers in Robotics and AI* 7 (2020), p. 188.

[17] Brigitte Toro, Christopher J Nester, and Pauline C Farren. "Cluster analysis for the extraction of sagittal gait patterns in children with cerebral palsy". In: *Gait & posture* 25.2 (2007), pp. 157–165.

[18] Eric Watelain et al. "Gait pattern classification of healthy elderly men based on biomechanical data". In: *Archives of physical medicine and rehabilitation* 81.5 (2000), pp. 579–586.

[19] Moritz Kassner, William Patera, and Andreas Bulling. "Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction". In: *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*. UbiComp '14 Adjunct. Seattle, Washington: ACM, 2014, pp. 1151–1160. ISBN: 978-1-4503-3047-3. DOI: 10.1145/2638728.2641695. URL: http://doi.acm.org/10.1145/2638728.2641695.

[20] *MTw Awinda - Products - Xsens 3D motion tracking*. https://www.xsens.com/products/mtw-awinda/. (Accessed on 10/29/2018).

[21] Ge Wu et al. "ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—part I: ankle, hip, and spine". In: *Journal of biomechanics* 35.4 (2002), pp. 543–548.

[22] Vassilios G Vardaxis et al. "Classification of able-bodied gait using 3-D muscle powers". In: *Human Movement Science* 17.1 (1998), pp. 121–136.

[23] .. Meinard Haji Ghassemi, Jochen Klucken, and Björn M. Eskofier. "Segmentation of Gait Sequences in Sensor-Based Movement Analysis: A Comparison of Methods in Parkinson's Disease". In: *Sensors* 18.1 (2018). ISSN: 1424-8220. URL: https://www.mdpi.com/1424-8220/18/1/145.

[24] J-JJ Chen and Richard Shiavi. "Temporal feature extraction and clustering analysis of electromyographic linear envelopes in gait studies". In: *IEEE Transactions on Biomedical Engineering* 37.3 (1990), pp. 295–302.

[25] Olga Russakovsky et al. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252. DOI: 10.1007/s11263-015-0816-y.

[26] M Pat Murray. "Gait as a total pattern of movement: Including a bibliography on gait". In: *American Journal of Physical Medicine & Rehabilitation* 46.1 (1967), pp. 290–333.

[27] He Huang et al. "Continuous locomotion-mode identification for prosthetic legs based on neuromuscular–mechanical fusion". In: *Biomedical Engineering, IEEE Transactions on* 58.10 (2011), pp. 2867–2875.

[28] Kaiming He et al. "Deep Residual Learning for Image Recognition". In: *CoRR* abs/1512.03385 (2015). arXiv: 1512.03385. URL: http://arxiv.org/abs/1512.03385.

[29] Simon Kornblith, Jonathon Shlens, and Quoc V Le. "Do better imagenet models transfer better?" In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2019, pp. 2661–2671.

[30] Aaron J Young et al. "Classifying the intent of novel users during human locomotion using powered lower limb prostheses". In: *Neural engineering (NER), 2013 6th international IEEE/EMBS conference on*. IEEE. 2013, pp. 311–314.

[31] P Artemiadis. "EMG-based robot control interfaces: past, present and future". In: *Advances in Robotics & Automation* 1.2 (2012), pp. 1–3.

[32] Levi J Hargrove et al. "Intuitive control of a powered prosthetic leg during ambulation: a randomized clinical trial". In: *JAMA* 313.22 (2015), pp. 2244–2252.