2019 International Conference on Robotics and Automation (ICRA)
Palais des congres de Montreal, Montreal, Canada, May 20-24, 2019

Video-based Prediction of Hand-grasp Preshaping with Application to Prosthesis Control

Luke T. Taverne[1,2*], Matteo Cognolato[2,3*], Tobias Bützer[2], Roger Gassert[2], Otmar Hilliges[1]

***Abstract*— Among the currently available grasp-type selection techniques for hand prostheses, there is a distinct lack of intuitive, robust, low-latency solutions. In this paper we investigate the use of a portable, forearm-mounted, video-based technique for the prediction of hand-grasp preshaping for arbitrary objects. The purpose of this system is to automatically select the grasp-type for the user of the prosthesis, potentially increasing ease-of-use and functionality. This system can be used to supplement and improve existing control strategies, such as surface electromyography (sEMG) pattern recognition, for prosthetic and orthotic devices. We designed and created a suitable dataset consisting of RGB-D video data for 2212 grasp examples split evenly across 7 classes; 6 grasps commonly used in activities of daily living, and an additional *no-grasp* category. We processed and analyzed the dataset using several state-of-the-art deep learning architectures. Our selected model shows promising results for realistic, intuitive, real-world use, reaching per-frame accuracies on video sequences of up to 95.90% on the validation set. Such a system could be integrated into the palm of a hand prosthesis, allowing an automatic prediction of the grasp-type without requiring any special movements or aiming by the user.**

I. Introduction

Although modern hand prostheses are becoming increasingly more advanced and capable of a great number of grasp types [1–3], there is a distinct lack of an intuitive, low-latency and robust method for the selection of grasp-type. We address this using a framework capable of choosing a grasp-type without any additional actions required; the user reaches toward an object, the system selects the appropriate grasp, and the user can open and close the hand using a standard proportional sEMG control. We created a custom dataset for this purpose and present a model that generates predictions for each frame in the video. To the best of our knowledge, this is the first approach to select grasp-type based purely on video data, requiring no additional input or special actions by the user.

Currently, the grasp-type selection in dexterous hand prostheses is often performed by eliciting specific sEMG signal patterns via muscle contractions, such as co-contractions of antagonist muscles [4]. The process for a grasp can be divided into: *a)* grasp initiation, *b)* grasp-type selection and *c)* grasp execution. With sEMG sequences, *a)* and *b)* require explicit, specific and non-intuitive [5] actions from the user; something humans do not have to do to operate their biological hands.

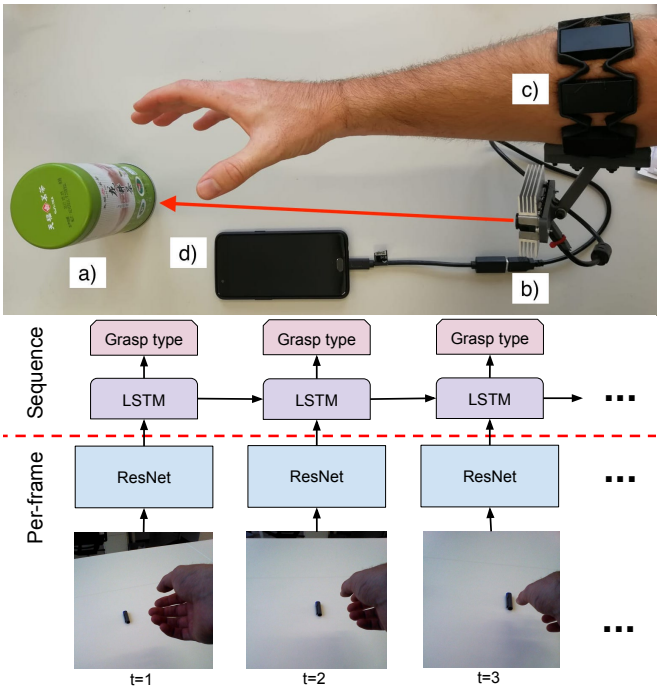

Fig. 1. Top: System schematic. As the user reaches for the desired object (a), the forearm-mounted RGB-D video camera (b) and Myo armband (c) stream data to a smartphone application (d). The system uses this data to generate the required grasp-type (in this example, *medium wrap*). Bottom: Model diagram for the grasp-type prediction over a grasping sequence with video input.

Our method seeks to eliminate the need for special, non-intuitive actions from the user for steps *a)* and *b)*; instead, our system will detect the onset of *a)*, also known as *gesture spotting*, and predict the grasp-type needed for *b)*. One commercial device, the CoApt Complete Control (CoApt Engineering, USA), also seeks to eliminate these non-intuitive actions, and is able to control many commercially available prostheses using sEMG-based pattern recognition. This does, however, rely on a minimum of 8 electrodes to be attached to the user[1]. Other methods, such as via a smartphone application or proximity sensors, have been developed for grasp-type selection (iLimb, Touch Bionics, UK)[2].

For human grasping, there is a relationship between the distance between subject and object, the object's size and shape, and grasp prehension [6]. There is also information concerning the size and fragility of the object embedded in the velocity and acceleration profiles of the hand [7], which also affects the choice of grasp. We take this as motivation

*Luke T. Taverne and Matteo Cognolato are co-first authors.
[1]Department of Computer Science, Advanced Interactive Technologies Lab (AIT), ETH Zurich, 8092 Zurich, Switzerland ltaverne@ethz.ch / otmar.hilliges@inf.ethz.ch
[2]Department of Health Sciences and Technology, Rehabilitation Engineering Lab (RELab), ETH Zurich, 8092 Zurich, Switzerland first.last@hest.ethz.ch
[3]Institute for Business Information Systems, University of Applied Sciences Western Switzerland (HES-SO), 3960, Sierre, Switzerland

[1]http://www.coaptengineering.com/
[2]http://www.touchbionics.com/

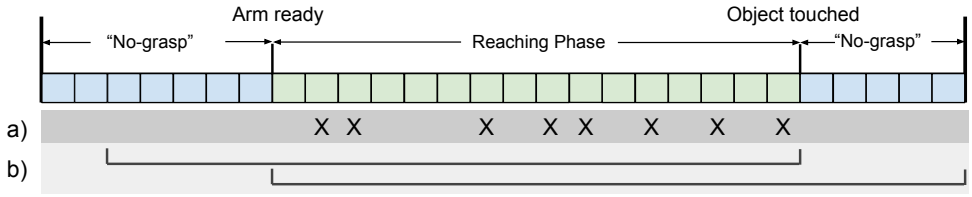

Fig. 2. A visualization of the input to the two models. The left and right of the series of boxes represent the start and end of the recorded video sample, with each box representing a single frame of the video. The timestamps for "arm ready" and "object touched" are shown, with the reaching phase between them. The frames outside the *reaching phase* are labeled with *no-grasp*, and the frames within the *reaching phase* are labeled with the grasp label selected by the user for the given sample (i.e. one of the grasps show in Figure 3). a) shows the selected *single frames*, and b) shows two examples of the *full sequence* input.

to use the visual properties of the desired object, along with the motion of the forearm during grasping, in the selection of the grasp-type.

The use of video data for choosing a grasp-type is non-trivial; it is essentially two nested tasks. First, it must be determined *if* the user is performing a grasp or not (referred to as *gesture spotting* in gesture recognition literature), and if they are performing a grasp, determine which type of grasp is being used. The system must also deal with a moving camera, mounted on the user's body, and target objects that may be occluded by the user's body or by parts of the environment. Differing lighting conditions also affect the video representation of the objects. As deep learning architectures have had success in solving similar problems, such as various gesture recognition [8–10] and image recognition [11] tasks, we will focus our efforts on these types of models.

Our goal is to use this information embedded in the objects and the motion data embedded in the video to increase the intuitiveness and reduce the latency of the grasp-type selection process, thereby reducing the cognitive burden placed on the user of a hand prosthesis. Reliability and consistency are two essential requirements for assistive and restorative solutions, so we focus on a subset of common grasps and design a dataset to investigate the maximum performance that can be achieved using such a system. The dataset consists of a *baseline* dataset (used for training), where we limit ourselves to a few indoor locations, and a *hard* dataset (used only for testing), where we test our model on unseen locations and novel objects. We also design the system to be wearable and portable, so that it could be easily adapted for real-life use in the future. The ideal system would have the RGB-D camera embedded into the palm of the hand prosthesis, respectively a wrist support for an orthosis. However, in order to validate the concept the system was designed to be effective for the data acquisition with a healthy subject and was subsequently adapted for this purpose, as depicted in Figure 1.

II. Related Work

One of the most commonly researched methods to achieve more intuitive prosthetic control is sEMG pattern recognition [12]. High classification accuracies (around 90%) have been achieved in laboratory settings [13]. However, limitations such as lack of robustness and reliability have limited the translation of pattern-recognition-based myoelectric control systems into clinical practice and commercial devices. Such limitations have prompted the investigation of alternative methods to supplement or replace the sEMG-based control [14]. In particular, the use of additional sources of information such as inertial data or computer vision, have been shown to be a promising approach [5].

Focusing on implementations exploiting computer vision, the first work using a camera for the purpose of grasp-type selection for dexterous hand prostheses comes, to the best of our knowledge, from [15]. This project uses a rule-based system and combines an ultrasound distance sensor with an RGB webcam. It applies traditional computer vision techniques to estimate the size of the object in order to identify the appropriate grip aperture and one of four grasp types. This system requires a so-called "aiming-phase", which is an explicit movement to target the desired object. Ghazaei et al. [16] included a more flexible deep learning based approach and also mounted an RGB webcam to the back of the hand and relied on the same "aiming-phase" requirement as [15]. Other systems use depth information to improve grasp-type selection. Markovic et al. [17] mounted a pair of stereo cameras onto glasses and combined this with an augmented reality feedback system to select the required grasp-type and aperture for a given object. Štrbac et al. [18] collected RGB-D video data directly via a tripod-mounted Kinect to select the aperture and grasp-type.

These efforts toward automating the grasp-type selection for hand prostheses often add some additional burden to the user, whether it be additional non-intuitive movements like an explicit "aiming phase", or a restricted workspace. In the case of the "aiming phase," this constraint also adds latency to the system, limiting the efficiency of the user in interacting with the environment. The aim of this work is to provide a new portable, low-latency and intuitive system able to identify the intended grasp without adding additional burden to the user.

A human-grasping dataset for activities of daily living, containing various modalities such as RGB-D video, full-body IMU, and egocentric RGB video was recently released and is presented in [19]. Like our system, the RGB-D camera is mounted on the forearm and captures a view of the object being approached by the hand. With 3826 grasp-samples spread across 33 different types and 13 subjects, this dataset has too much variation for our current purposes. If we consider restricting ourselves to a subset of this dataset consisting of a single subject and fewer grasps, the resulting subset would not have enough samples for training a deep learning model. This dataset would also require a large amount of manual effort in terms of relabeling, cropping videos and timestamping to match our dataset. However, the similarities between this recent dataset and our own work suggest the feasibility and applicability of this type of approach.

III. Methods

A. System Overview

We designed our system with the hardware shown in Figure 1; an RGB-D camera collects video data as the hand approaches an object and is attached to an armband via a 3D-printed mount, while a smartphone application handles the user interaction and recording of data. Using the collected samples from the Handcam *baseline* dataset, we train and evaluate several deep learning models for predicting the required grasp-type. We also present a *hard* dataset, which contains new locations and novel objects, and use it to test our best model's performance. In the interest of future work, the armband we selected is the Myo armband (Thalmic Labs, Canada), which we also use to collect IMU data.

B. Grasp Recognition Pipeline

Fig. 3. The six grasps selected for use in this system, along with some examples of the types of objects used for each category. Grasp diagrams from [20].

1) Grasp Types: We chose six grasp types from [20] that are commonly used in activities of daily living. The selected grasps are: power sphere, medium wrap, tip pinch, precision disk, lateral pinch, and writing tripod; depicted in Figure 3. We add a seventh grasp category, which we call the *no-grasp*. The *no-grasp* class was introduced in an effort to both (1) help the system identifying when there is no valid object for grasping, and (2) to discourage the system from learning the grasp type based on the movements of the hand. Regarding (1), these are obtained by including two *no-grasp* phases in each acquisition: one before the starting of the reaching phase, before there is a target object in view, and a second one after the object was touched/grasped, since the configuration of the hand should not change automatically while the object is being held. Regarding (2), since the data acquisition was performed by an able-bodied subject and his hand was visible in the video, the system could in principle identify the grasp-type based on the hand movement instead of using information about the object and the motion. If this were the case, the grasp-selection may not work in a real application; if the hand motion is used for selection and the hand is not visible because the device is embedded into the prosthetic hand's palm, then it may never select a grasp-type. We rather want the system to identify the grasp-type based on the object and scene that the hand approaches; therefore, *no-grasp* movements are recorded by performing one of the other six grasps toward an area which contains no target object.

2) Models: Figure 2 represents our usage of each sample in the dataset. Since each frame in the sequence has a label (one of the grasp types in Figure 3 or *no-grasp*), we can train models of two types: single frames and sequences (more on frame labeling in Section III-C). For single frames we train a ResNet [11] with cross-entropy loss:

$$J_{CE}(\theta) = -\sum_{i}^{M} y_i \log(P(\hat{y}_i)) \quad (1)$$

Where θ are the model parameters, M is the number of classes, y_i is the true label and $\hat{y}_i$ is the predicted label. The ResNet is applied to the individual frames, as shown in Figure 1 (bottom), with a final fully-connected layer in place of the LSTM.

For our sequence models, we remove the final fully-connected layer from our trained single frame ResNets and use the exposed feature vector as an input to an LSTM [21] and also use cross-entropy loss. A visualization of the sequence models is shown in Figure 1 (bottom). For each of the two model types (single frames, sequences), we trained a separate network on RGB, depth-only, and RGB-D input modalities. We used 10-fold cross-validation with class-balanced 90% train / 10% validation splits. All models were subject to early stopping based on the validation accuracy. The ResNet was trained from scratch.

C. Data Collection - Handcam Baseline Dataset

The proposed system collects RGB-D video for each sample. The RGB-D video was recorded with a structured light camera (Orbbec Astra Mini S, Orbbec 3D, USA) at 30FPS with a resolution of 320x240px and a depth resolution of 100 μm. We also collect IMU data, sampled at 50Hz via the Myo armband. The IMU data is unused in this paper, and is collected for future use. Although the Myo armband is also designed to collect sEMG data, the official MyoSDK for Android does not yet support streaming raw sEMG data. To support the goal of performing gesture spotting (the decision between *grasp* and *no-grasp*), we also record timestamps corresponding to the start of the reaching phase of the grasp (*arm-ready*), and when the object is first touched (*object-touched*). For a visualization of how these timestamps are used in a sample, see Figure 2. The data collection was handled by a custom application on an Android smartphone for portability. To begin a data collection session, the user connects the camera and Myo armband to the smartphone. After the application confirms proper configuration and connections, the grasp-type icons are enabled and recording can begin. The procedure for collecting a grasp sample is as follows:

1) Place object on table.
2) Arm at side, object is on table.
3) Select the grasp-type on app (recording begins).
4) Prepare arm to reach for object. Touch screen to record *arm ready* timestamp when arm first points toward object.
5) Reach forward and grasp object using chosen grasp-type. Upon first contact with object, touch screen to record *object touched* timestamp.
6) Finish grasp, hold object. Touch screen to stop recording.

Data collection was performed using this process on around 200 different objects (recorded in approximately 8h over several days). Each object was grasp approximately 5 times in each of two data collection locations, with the object placed in a random orientation and starting position for each sample. The two data collection locations were both indoors, one with a low white table and the other a medium height dark wooden table. One healthy subject (24 years-old, male) collected all 2212 samples in these two locations, with balanced classes of 316 samples for each of the 7 grasp-types (including the *no-grasp*). An example image from a video sample for a "medium wrap" grasp is shown in Figure 4.

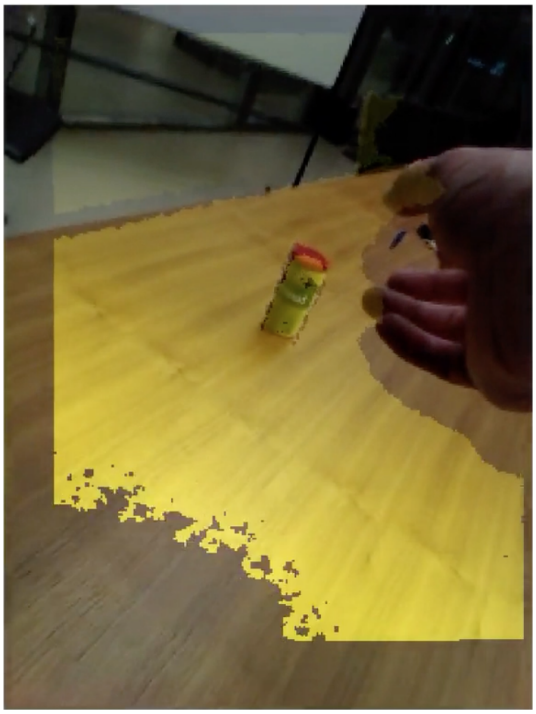

Fig. 4. An image from the *reaching phase* of a "medium wrap" grasp. The calibrated depth is shown as a yellow overlay onto the RGB, where brighter corresponds to a closer depth value. The pixelization on the bottom left is due to the minimum usable distance of the depth sensor; this part of the table is within 0.3m of the camera.

1) Preprocessing: The collected samples are first checked for consistency in order to verify that all files are present and non-empty. Timestamps across the different data sources are then synchronized. For the IMU data from the Myo armband we subtract the first timestamp from all future timestamps. To synchronize the grasp-event timestamps *arm-ready* and *object-touched* with the camera frames, a special sample video was recorded; with the camera pointed at the screen of the smartphone, the screen was touched to record the *arm ready* and *object touched* timestamps. The timestamps of these video frames were used as an offset to synchronize the video and grasp-event data for all samples.

D. Data Collection - Handcam "Hard" Dataset

We designed several test sets to evaluate the limitations of our models. Our training set is limited to two locations, so we investigate the performance in an additional, unseen location. The new location is also indoors, but is a table of different color, texture, height, and lighting conditions than the other two. Since we did not record which unique object is used in each sample, the validation set contains novel *samples* but there is no guarantee for how many novel *objects* it contains. This because each sample is only tagged with a grasp-type and each unique object was grasped approximately 5 times in each location. We then randomly sampled 284 grasp-samples from each of the 7 grasp-types for each validation split, maintaining the class balance. So it is also possible, although improbable, that all the samples of one unique object ended up in the training set with none in the validation set, or vice-versa. Therefore we recorded samples of several objects that were not included in either the training or validation sets. We recorded one test set for each combination of the above variations, yielding a total of 3 test sets. For each test set, two objects for each grasp-type were used, and they were grasped two times each, yielding a total of 24 grasp samples in each of the 3 test sets. We emphasize that this *hard* dataset was not used for training purposes and was only used as a test of the model's ability to generalize. The test sets were recorded by the same subject as the main training and validation sets, and followed the same procedure as in Section III-C.

IV. Results

For *accuracy*, a prediction is considered correct if the model chooses exactly the correct grasp-type for the given frame, including the *no-grasp* label. For sequences this is applied to each frame, so to achieve 100% accuracy for a given sample, the network must choose the correct label for every frame in the sequence. For sequence data, we also present precision and recall. They are reported as average and standard deviation over the 10 splits for the validation set, and over the results obtained with the 10 models on the hard set.

A. Handcam - Baseline Set

1) Single frames: We extracted 20 single frames at random, without replacement, from each sample in the Handcam *baseline* dataset and used them to train a wide ResNet [11] to predict the grasp-type (from the 7 grasp-types, including the *no-grasp*). We examined two model variations: a ResNet-18 with image size of 112x112px, and a ResNet-50 bottleneck with input size of 224x224px. The ResNet-18 is small enough to be suitable for later use in end-to-end training with an LSTM [21], while the ResNet-50 is too large for efficient end-to-end training but has a higher model capacity. For each of the two model variations, we trained a separate network on RGB, depth, and RGB-D input modalities. For data augmentation we applied random crops. We train these two ResNets and compare their accuracy in Table I.

TABLE I
VALIDATION SET - SINGLE FRAMES - ACCURACY (%)

Type	*RGB*	*Depth*	*RGB-D*
ResNet-50	**94.62 ± 1.89**	**76.90 ± 3.97**	**95.01 ± 1.80**
ResNet-18	92.98 ± 1.65	73.24 ± 2.45	93.27 ± 2.00

As expected, the ResNet-50 outperforms the ResNet-18 in all modalities, likely due to its larger model capacity.

We note that the depth significantly underperforms both RGB and RGB-D for the two models. Due to the minimum distance of the depth sensor (~0.3m), some depth frames will contain no object data when the object is closer than this minimum distance. The models therefore have a more difficult time with single depth frames.

2) Sequences: Here we use a model that can predict the required grasp-type for each frame in a video sequence. We apply a ResNet, reusing the single frame models trained in Section IV-A.1, to each frame in the sequence. The fully-connected layer is removed from the ResNet, and the underlying feature vector is used as an input for a 1 layer LSTM with 1024 hidden units. In this section, we have two types of sequence models: *frozen*, where we freeze the ResNet weights during training, and *end-to-end*, where we allow the gradient to flow through the LSTM into the ResNet weights. In order to allow model convergence in the latter, the learning rate we used for updating the ResNet weights was 10x lower than the learning rate for the LSTM. For training we chose a sequence length of 60 frames and randomly sampled these subsequences from each grasp-sample, as a means of data augmentation. Validation samples are always evaluated on their full sequence length. We applied a central crop to all image sequences. The results for the sequence models are presented in Tables II and III.

TABLE II
VALIDATION SET - SEQUENCES - ACCURACY (%)

ResNet Size	*RGB*	*Depth*	*RGB-D*
50 (frozen)	95.12 ± 0.62	**93.72 ± 1.13**	94.33 ± 2.02
18 (frozen)	94.64 ± 0.76	92.65 ± 1.21	95.29 ± 0.59
18 (end-to-end)	**95.50 ± 0.31**	93.27 ± 0.64	**95.90 ± 0.61**

TABLE III
VALIDATION SET - SEQUENCES - RGB-D PRECISION & RECALL

Grasp	*Precision*	*Recall*
Power Sphere	0.91 ± 0.03	0.91 ± 0.05
Medium Wrap	0.92 ± 0.03	0.96 ± 0.01
Tip Pinch	0.91 ± 0.05	0.93 ± 0.03
Precision Disc	0.91 ± 0.04	0.91 ± 0.05
Lateral Pinch	0.92 ± 0.03	0.97 ± 0.02
Writing Tripod	0.92 ± 0.02	0.96 ± 0.01
No-Grasp	0.98 ± 0.00	0.97 ± 0.01

The sequence ResNet-18 models performed nearly as well as, and sometimes better than, the frozen ResNet-50/LSTM. This shows that although the ResNet-18 had a lower accuracy on the single frames, the LSTM was able to leverage the temporal information to overcome the more limited model capacity of the ResNet-18, emphasizing the importance of the temporal component in solving this task. We also note that the accuracy of the depth modality drastically improved compared to the single frame model, suggesting that the LSTM is able to ignore frames that contain no object depth information, as described in Section IV-A.1. The *end-to-end* training of the ResNet-18/LSTM allowed for a slight further increase of the accuracy.

B. Handcam - Hard Sets

We test our best sequence model, the end-to-end ResNet-18, on the Handcam *hard* test sets described in Section III-D and present the results in Tables IV and V.

TABLE IV
HARD SET - SEQUENCES - ACCURACY (%)

		End-to-end ResNet-18/LSTM		
New location	*New objects*	*RGB*	*Depth*	*RGB-D*
-	-	95.50 ± 0.31	93.27 ± 0.64	95.90 ± 0.61
-	✓	86.96 ± 2.00	**88.65 ± 1.27**	**86.64 ± 2.01**
✓	-	**75.97 ± 2.52**	68.01 ± 1.39	73.67 ± 2.82
✓	✓	64.86 ± 1.47	64.21 ± 0.55	64.35 ± 0.66

As expected, the models perform worse on the hard sets than on the baseline set. However, the networks were able to generalize to new objects with a limited reduction in accuracy (lower than 10% for all models), suggesting that the training set was diverse enough in the number and types of objects. The model struggled with new environments, reaching a classification accuracy of up to 75.97%, but with low precision and recall. This result is to be expected, as the training and validation sets were recorded in only two different environments. The relatively poor generalization to new locations would likely be improved by recording and training on additional samples in new locations. This applies also to the unseen objects and locations condition, where the value of the accuracy was mostly coming from the classification of the no-grasp, as indicated by the precision and recall.

We also evaluated our best model, the ResNet-18/LSTM RGB-D, on a cluttered environment with new/old objects, and new/old environments. The system was generally unreliable in all cases, which is to be expected as it was trained on single objects. For more information, please see the supplemental material at http://ait.ethz.ch/projects/2019/handcam.

V. DISCUSSION

We presented a video-based approach to automatically predict grasp-types, requiring no explicit action by the user. The proposed system achieves per-frame accuracies of up to 95.90% on the video sequence data, and was able to generalize to completely novel objects with accuracy of up to 88.65%. Despite the low performance on new locations, the system was able to reliably spot the difference between *no-grasp* and *grasp*. Furthermore, there is no need for specific actions from the user, making the system capable of recognizing the grasp type with low-latency.

In order to evaluate the relative offline performance of the proposed system, we look to the most closely-related learning-based approach, from [16]. It is important to emphasize that the following considerations involve models trained on different datasets and therefore they are not directly comparable, although they have the same goal of supporting grasp-type classification. The following comparisons will refer to the end-to-end ResNet-18/LSTM RGB-D, as we consider it to be our best and most promising model. In the offline experiments in [16], the system is evaluated using two

TABLE V
HARD SET - SEQUENCES - RGB-D PRECISION & RECALL

Grasp	*New Objects*		*New Location*		*New Objects & Location*	
	Precision	*Recall*	*Precision*	*Recall*	*Precision*	*Recall*
Power Sphere	0.91 ± 0.08	0.84 ± 0.08	0.95 ± 0.05	0.53 ± 0.22	0.23 ± 0.36	0.04 ± 0.07
Medium Wrap	0.87 ± 0.08	0.77 ± 0.16	0.40 ± 0.42	0.07 ± 0.09	0.26 ± 0.40	0.03 ± 0.05
Tip Pinch	0.84 ± 0.10	0.54 ± 0.17	0.88 ± 0.30	0.57 ± 0.22	0.00 ± 0.00	0.00 ± 0.00
Precision Disc	0.70 ± 0.12	0.71 ± 0.26	0.50 ± 0.50	0.04 ± 0.05	0.00 ± 0.00	0.00 ± 0.00
Lateral Pinch	0.61 ± 0.43	0.23 ± 0.26	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00	0.00 ± 0.00
Writing Tripod	0.94 ± 0.03	0.93 ± 0.08	0.97 ± 0.04	0.34 ± 0.23	0.19 ± 0.39	0.02 ± 0.03
No-Grasp	0.90 ± 0.02	0.98 ± 0.01	0.72 ± 0.03	1.00 ± 0.00	0.64 ± 0.00	1.00 ±0.00

different cross-validation techniques: within-object cross-validation (WOC), where the model is evaluated on different views of seen objects, and between-object cross-validation (BOC), where the model is tested on objects not seen during training. The WOC method, which achieved 85.29% accuracy for the 4 grasp-types, is most comparable to the results achieved on the Handcam *baseline* set, where the model achieved 95.90% accuracy on the 7 grasp types. The BOC method, which achieved 74.74%, corresponds well to the unseen objects portion of the Handcam *hard* dataset presented in this work, where the model achieved 86.64%. The results obtained with the proposed approach are therefore promising, supporting the feasibility of the system presented in this work.

In general, the proposed system also allows the elimination of both the aiming phase and image pre-processing step, which contributes to increased latency in such a system. Furthermore, the use of video sequences, the inclusion of depth data and the exploitation of the temporal information via the LSTM can substantially improve the ability of the system to distinguish between large and small objects of the same shape on-the-fly during the reaching phase. In a practical implementation, the system could also identify the moment in which a grasp-type must be chosen in order to start the pre-shaping of the hand. This can be estimated on-the-fly by evaluating the approaching speed and using the mechanical proprieties of the prosthetic/effector, such as the time needed to fully close the fingers. Finally, the results indicate that the system is able to automatically identify the intention to grasp, with no need for a trigger such as a predefined voluntary contraction from the user, providing a low-latency and intuitive interface.

A. Limitations

Considering the results on the baseline and hard sets, a clear limitation is the weak generalization to new locations. This is however to be expected, as the number of locations in the training set was very low compared to the variety of objects (2 locations vs ~200 objects). This can be addressed by increasing the number of locations used in future iterations of the data collection. The other major limitation of the system is the type and placement of the video camera. The type of depth camera we used (structured light) are generally quite affordable, although this type of camera has the drawback of a minimum distance to the object, below which the camera is unable to provide depth information. To improve on this, we can use a more expensive time-of-flight camera, which have a much lower minimum range. In addition, time-of-flight cameras can be quite small, so the camera placement could be moved to the wrist or palm of the hand, which would be embedded in the prosthesis or orthosis in a real-world application, making wearing our system quite seamless for the user. Our system was also not designed to handle cluttered scenes, which are very common in the real-world.

B. Future Work

The system should be tested with intact and prosthesis users to determine if the offline results will transfer to a real application scenario, and to learn if real users would find such a device useful in their everyday life. For use with a hand prosthesis, the camera could be moved to the palm of the prosthetic hand, reducing the footprint of the device. The device can also be adapted for use with the RELab tenoexo [22], a hand exoskeleton to assist subjects with loss of hand function due to neurological disease or trauma. To handle cluttered scenes where the user grasps one object from a group of many, we may be able to implement a model which uses attention mechanisms to focus on different areas of the image [23], or use a fast object detector like YOLO v2 [24]. For the purposes of building an especially robust system for everyday use, we must add additional adversarial examples. Other deep architectures may provide better performance and smaller model size at a higher training cost, such as DenseNet [25], and some size-efficient networks like SqueezeNet [26] have been implemented for inference on FPGAs [27]. Furthermore, we could use a model like C3D [28] to convolve the spatial and temporal information.

VI. CONCLUSION

We have presented a novel, fully-wearable system and framework for automatically choosing the required grasp-type for powered hand prostheses and orthoses. The presented methods systematically outline the contributions of each input-modality and model architecture, and we examine the limitations and constraints of our system through several additional hard datasets. Our system places no additional cognitive or physical burden on the user for the operation of their assistive device, and has realistic limitations for basic real-world use. The results show that the system and framework are a promising approach to automatically selecting the grasp-type for powered hand prostheses and orthoses, while providing a road-map for which areas require improvement in future iterations of this system.

VII. ACKNOWLEDGMENT

This work was supported by the ETH Zurich Foundation in collaboration with Hocoma AG. The authors would also like to thank NVIDIA for providing the GPUs for this project.

REFERENCES

[1] P. Heo, G. M. Gu, S.-j. Lee, K. Rhee, and J. Kim, "Current hand exoskeleton technologies for rehabilitation and assistive engineering," *International Journal of Precision Engineering and Manufacturing*, vol. 13, no. 5, pp. 807–824, may 2012. [Online]. Available: http://link.springer.com/10.1007/s12541-012-0107-2

[2] H. K. Yap, Jeong Hoon Lim, F. Nasrallah, J. C. H. Goh, and R. C. H. Yeow, "A soft exoskeleton for hand assistive and rehabilitation application using pneumatic actuators with variable stiffness," in *2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, may 2015, pp. 4967–4972. [Online]. Available: http://ieeexplore.ieee.org/document/7139889/

[3] M. Atzori and H. Müller, "Control Capabilities of Myoelectric Robotic Prostheses by Hand Amputees : A Scientific Research and Market Overview," *Frontiers in systems neuroscience*, pp. 1–13, 2015.

[4] I. Kyranou, A. Krasoulis, M. S. Erden, K. Nazarpour, and S. Vijayakumar, "Real-time classification of multi-modal sensory data for prosthetic hand control," in *2016 6th IEEE International Conference on Biomedical Robotics and Biomechatronics (BioRob)*. IEEE, jun 2016, pp. 536–541. [Online]. Available: http://ieeexplore.ieee.org/document/7523681/

[5] D. Farina and S. Amsüss, "Reflections on the present and future of upper limb prostheses," *Expert Review of Medical Devices*, no. 4, pp. 321–324.

[6] M. Santello and J. F. Soechting, "Gradual Molding of the Hand to Object Contours," *Journal of Neurophysiology*, vol. 79, no. 3, pp. 1307–1320, mar 1998. [Online]. Available: http://www.physiology.org/doi/10.1152/jn.1998.79.3.1307

[7] R. G. Marteniuk, J. L. Leavitt, C. L. MacKenzie, and S. Athenes, "Functional relationships between grasp and transport components in a prehension task," *Human Movement Science*, vol. 9, no. 2, pp. 149–176, apr 1990. [Online]. Available: https://www.sciencedirect.com/science/article/pii/0167945790900259

[8] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *Workshop at the European conference on computer vision*. Springer, 2014, pp. 474–490.

[9] P. Molchanov, S. Gupta, K. Kim, and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," in *2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, jun 2015, pp. 1–7. [Online]. Available: http://ieeexplore.ieee.org/document/7301342/

[10] D. Wu, L. Pigou, P.-J. Kindermans, N. D.-H. Le, L. Shao, J. Dambre, and J.-M. Odobez, "Deep Dynamic Neural Networks for Multimodal Gesture Segmentation and Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1583–1597, aug 2016. [Online]. Available: https://ieeexplore.ieee.org/document/7423804/

[11] S. Zagoruyko and N. Komodakis, "Wide Residual Networks," may 2016. [Online]. Available: http://arxiv.org/abs/1605.07146

[12] M. Asghari Oskoei and H. Hu, "Myoelectric control systems – A survey," *Biomedical Signal Processing and Control*, no. 4, pp. 275–294, oct.

[13] A. H. Al-Timemy, G. Bugmann, J. Escudero, and N. Outram, "Classification of Finger Movements for the Dexterous Hand Prosthesis Control With Surface Electromyography," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 3, pp. 608–618, may 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6471724/

[14] C. Castellini, P. Artemiadis, M. Wininger, A. Ajoudani, M. Alimusaj, A. Bicchi, B. Caputo, W. Craelius, S. Dosen, K. Englehart, D. Farina, A. Gijsberts, S. B. Godfrey, L. Hargrove, M. Ison, T. Kuiken, M. Marković, P. M. Pilarski, R. Rupp, and E. Scheme, "Proceedings of the first workshop on peripheral machine interfaces: Going beyond traditional surface electromyography," *Frontiers in Neurorobotics*, vol. 8, no. AUG, pp. 1–17, 2014.

[15] S. Došen and D. B. Popović, "Transradial Prosthesis: Artificial Vision for Control of Prehension," *Artificial Organs*, no. 1, pp. 37–48.

[16] G. Ghazaei, A. Alameer, P. Degenaar, G. Morgan, and K. Nazarpour, "Deep learning-based artificial vision for grasp classification in myoelectric hands," *Journal of Neural Engineering*, no. 3, p. 036025, jun.

[17] M. Markovic, S. Dosen, C. Cipriani, D. Popovic, and D. Farina, "Stereovision and augmented reality for closed-loop control of grasping in hand prostheses," *Journal of Neural Engineering*, vol. 11, no. 4, p. 046001, aug 2014. [Online]. Available: http://stacks.iop.org/1741-2552/11/i=4/a=046001?key=crossref.920cc30034714dfe1da6df3291ab5200

[18] M. Štrbac, S. Kočović, M. Marković, and D. B. Popović, "Microsoft kinect-based artificial perception system for control of functional electrical stimulation assisted grasping." *BioMed research international*, vol. 2014, p. 740469, aug 2014. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/25202707http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4151575

[19] A. Saudabayev, Z. Rysbek, R. Khassenova, and H. A. Varol, "Human grasping database for activities of daily living with depth, color and kinematic data streams," *Scientific Data*, vol. 5, p. 180101, may 2018. [Online]. Available: http://www.nature.com/articles/sdata2018101

[20] I. M. Bullock, J. Z. Zheng, S. D. L. Rosa, C. Guertler, and A. M. Dollar, "Grasp Frequency and Usage in Daily Household and Machine Shop Tasks," *IEEE Transactions on Haptics*, vol. 6, no. 3, pp. 296–308, jul 2013. [Online]. Available: http://ieeexplore.ieee.org/document/6469076/

[21] F. Gers, "Learning to forget: continual prediction with LSTM," in *9th International Conference on Artificial Neural Networks: ICANN '99*, vol. 1999. IEE, 1999, pp. 850–855. [Online]. Available: http://digital-library.theiet.org/content/conferences/10.1049/cp{_}19991218

[22] T. Bützer and R. Gassert, "RELab tenoexo."

[Online]. Available: http://www.relab.ethz.ch/content/specialinterest/hest/rehabilitation-engineering-lab/en/research/current-research-projects/robotic-hand-orthosis-for-therapy-and-assistance-in-activities-of-daily-living.html

[23] K. Xu, J. L. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," pp. 2048–2057, 2015. [Online]. Available: https://nyu-staging.pure.elsevier.com/en/publications/show-attend-and-tell-neural-image-caption-generation-with-visual-

[24] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jul 2017, pp. 6517–6525. [Online]. Available: http://ieeexplore.ieee.org/document/8100173/

[25] G. Huang, Z. Liu, K. Q. Weinberger, and L. van der Maaten, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, vol. 1, no. 2, 2017, p. 3.

[26] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally, and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and," feb 2016. [Online]. Available: http://arxiv.org/abs/1602.07360

[27] D. Gschwend, "ZynqNet: An FPGA-Accelerated Embedded Convolutional Neural Network," Master Thesis, ETH Zurich, 2016. [Online]. Available: https://github.com/dgschwend/zynqnet

[28] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning Spatiotemporal Features with 3D Convolutional Networks," in *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015, pp. 4489–4497. [Online]. Available: http://ieeexplore.ieee.org/document/7410867/