Grasp Pre-shape Selection by Synthetic Training: Eye-in-hand Shared Control on the Hannes Prosthesis

Federico Vasile[1,3] Elisa Maiettini[1] Giulia Pasquale[1] Astrid Florio[2] Nicolò Boccardo[2] Lorenzo Natale[1]

Abstract— We consider the task of object grasping with a prosthetic hand capable of multiple grasp types. In this setting, communicating the intended grasp type often requires a high user cognitive load which can be reduced adopting shared autonomy frameworks. Among these, so-called *eye-in-hand* systems automatically control the hand pre-shaping before the grasp, based on visual input coming from a camera on the wrist. In this paper, we present an *eye-in-hand* learning-based approach for hand pre-shape classification from RGB sequences. Differently from previous work, we design the system to support the possibility to grasp each considered object part with a different grasp type. In order to overcome the lack of data of this kind and reduce the need for tedious data collection sessions for training the system, we devise a pipeline for rendering synthetic visual sequences of hand trajectories. We develop a sensorized setup to acquire real human grasping sequences for benchmarking and show that, compared on practical use cases, models trained with our synthetic dataset achieve better generalization performance than models trained on real data. We finally integrate our model on the Hannes prosthetic hand and show its practical effectiveness. We make publicly available the code and dataset to reproduce the presented results[†].

I. Introduction

Latest advancements in the development of prosthetic arms [1], [2] and specifically of myoelectric devices [3], [4] have led to the design of novel control systems based on electromyography (EMG) or mechanomyography (MMG) inputs. While effective, these systems need a significant cognitive effort for their control [5]. In fact, they generally require the user to execute specific muscle activation in order to drive the device. This increases the effort and the cognitive load required to get acquaintance with the prosthesis, contributing to the rejection of the device [6]. Therefore, while representing useful information on user motor intentions, these sensors are not sufficient to achieve a simple and intuitive control of devices with several degrees of freedom. In this perspective, a *shared autonomy* (or *shared control*) of the device has been introduced [5]. The main idea is to split the task between the user and an automatic controller which relies on external contextual measurements as input. It has been shown in [7] that a semi-autonomous control of a prosthesis allows a user to accomplish tasks in shorter time and providing less explicit inputs with respect to volitional
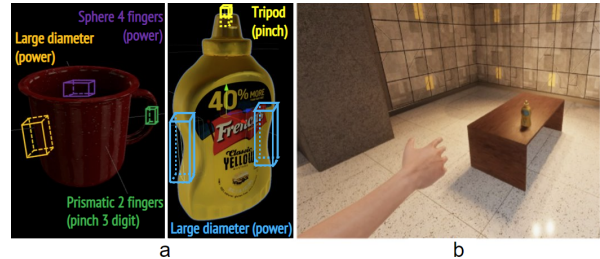

Fig. 1. Examples of the proposed multi grasp type per object annotation for the Mug and the Mustard (each object part is labeled with a different grasp type and pre-shape) (a) and of the table-top setup rendering (b).

control only. In particular, for tasks like object grasping, in order to be useful for controlling the action, the sensing might happen remotely, with no physical contact (*teleceptive sensing* [8]), before the actual action. A recently explored solution is the integration of vision sensors (e.g., RGB and Depth cameras) to exploit visual information of the environment and of the other interacting elements. Specifically, for the grasping task in a *shared autonomy*, the prosthesis control system could infer fundamental information from contextual measurements, such as (i) the user's target object, (ii) the part of the object that is intended to be grabbed and (iii) the intended use of the object. All these aspects concur to identify the grasp planned by the user (i.e., the *grasp type*), which determines the correct hand pre-shaping (i.e., the *pre-shape* phase), wrist orientation and hand closure, that would be otherwise unknown to the control system. Interestingly, it has been shown that the inclusion of visual information significantly increases the average grasp type classification accuracy [9]. Specifically, in *shared autonomy*, the hand pre-shaping and wrist orientation can be controlled automatically, based on visual input, while the closure of the fingers can be left to the user.

In this work, we focus on the problem of the automatic control, with vision, of the hand pre-shaping before the grasp, hence addressing a so-called *hand pre-shape classification* problem. Machine learning has been adopted in the recent literature for this task [12], [13], [5]. While effective, these methods require large and varied training sets. In the referenced literature on vision-based prosthetic control, the data is gathered for the purpose, by setting up data acquisition sessions that are typically long and tedious. In fact, while training deep learning systems on synthetic datasets is a quite established practice in computer vision and robotics, this approach is seldom adopted in prosthetics. We present a methodology and related tool for the synthetic generation of human-like grasping RGB sequences, which allows to

[1] Humanoid Sensing and Perception, Istituto Italiano di Tecnologia, Genoa, Italy (email: `name.surname@iit.it`)
[2] Rehab Technologies Lab, Istituto Italiano di Tecnologia, Genoa, Italy (email: `name.surname@iit.it`)
[3] DIBRIS, University of Genoa, Genoa, Italy
[†] `https://github.com/hsp-iit/prosthetic-grasping-simulation`

Pre-shape	Lateral	Power					Pinch	Pinch 3 Digit	
Definition	Adducted thumb	Central thumb					Abducted thumb	Abducted thumb	
Grasp type	Adducted thumb	Large diameter	Small diameter	Medium wrap	Sphere 4 fingers	Power sphere	Prismatic 4 fingers	Tripod	Prismatic 2 fingers
Objects	Pitcher[1], Plate[2], Spatula[2], Scissors[1]	Chips can[2], Mug[1], Mustard[2]	Hammer[1]	Meat can[2]	Chips can[1], Meat can[2], Mug[2]	Plum[1], Baseball ball[1]	Spoon[1], Large marker[1], Scissors[1], Spatula[1], Banana[1], Pitcher[1]	Red wood block[1], Mustard[1], Banana[1]	Mug[1]

TABLE I

ASSOCIATION OF PRE-SHAPES WITH DIFFERENT GRASP TYPES AND OBJECTS ACCORDING TO [10], [11], DEFINING THEM WITH THE THUMB POSE. THE SUPERSCRIPT DIGIT ON THE OBJECT NAME REFERS TO THE NUMBER OF OBJECT PARTS HAVING THAT GRASP TYPE.

produce low-cost training sets for learning-based pre-shape classification. Remarkably, we show that models trained with the proposed synthetic dataset achieve comparable or higher performance on real use cases, than model trained on real data. We then present a prosthesis control pipeline, trained on the presented synthetic dataset, that allows to predict the user's intended hand pre-shape and consequently move the prosthesis fingers accordingly. Note that, differently from similar work [13], we consider the case in which different object parts can be associated to different grasp types (and thus, pre-shapes). This is a critical aspect since it has been shown that one reason for amputees dissatisfaction with the prosthesis is the lack of the device adaptability to different object properties [14]. Thus, the target is to enable the user to grasp an object from each side and for all possible usages. Moreover, we develop a sensorized setup. We adopt it to empirically validate our design choices (e.g., the sensor placement). Then, we use it both (i) for collecting real human grasping sequences for benchmarking and (ii) for testing online the proposed approach. Finally, we integrate it with the Hannes prosthesis [4] to test the presented pipeline, showing its effectiveness.

This paper is organized as follows. In Sec. II, we review the related literature. Then, in Sec. III and IV, we describe the proposed approach for pre-shape classification and for human-like trajectories generation. In Sec. V, we illustrate the experimental analysis carried out on datasets and with the Hannes prosthesis. Finally, in Sec. VI we draw conclusions.

II. RELATED WORK

Vision-aided prosthetic grasping. Recent methods that exploit visual cues for prosthetic grasping differ mainly for sensor placements. For instance, [15] uses an external static camera to acquire RGB and Depth images and to segment the target object via background subtraction. This information is then used by a learning-based pipeline to predict the corresponding grasp type. This is not suitable for real-world applications since it limits the work space to a static camera field of view and requires a preliminary sensor calibration. Moreover, it has been shown in [16] that static external cameras allow for a less accurate control with respect to other placements, like, on the prosthesis (*eye-in-hand* [16]) or on a user's headset (*egocentric* point of view).
In approaches with an *egocentric* placement, the sensor has a wide field of view and additional information, as the user's gaze, is available. For instance, [17] and [18] use a stereo vision camera mounted on augmented reality glasses. The

target object is identified when the user directs their gaze, and therefore the glasses, towards it. Then the geometrical properties of the object are retrieved by analyzing the Depth information from the camera and the system estimates the grasp type. In [19], instead, a pre-trained object detector is fine-tuned for the purpose of grasp detection. The resulting grasp detector provides bounding boxes of possible objects to be grasped and the closest box to the user's gaze is selected. This approach requires the user to direct gaze to the object of interest, and to wear a headset or glasses. Conversely, the *eye-in-hand* configuration can be completely transparent to the user and allows to gather closer views of the objects to grasp. This makes it easier to identify the target with visual and motion cues [20]. For instance, in [21] the geometrical information (centroid and major axes) of the target object is inferred by using an RGB-D camera placed on the prosthesis and it is used to control the wrist orientation through visual servoing. In [22], instead, the same information is used to select the correct grasp type according to an *IF-ELSE* set of rules. More recent approaches rely on deep learning techniques to extract useful information for the grasp like, target object segmentation [23], [24], [25], object parts affordance [26] and grasp type prediction [12], [13].

In this work, we compare the performance of learning-based methods in both configurations (*egocentric* and *eye-in-hand*) in the same setting. We find that the *eye-in-hand* provides better visual cues for the considered task. Moreover, we develop a learning-based prosthesis control pipeline that allows to predict the user's intended hand pre-shape and consequently execute it on the device.

Object grasping datasets. To achieve good performance and generalization capabilities, supervised learning algorithms require quality annotated and abundant training data. For instance, recent work [27], [28], [29] presents novel multi-modal datasets of humans grasping or manipulating objects which are remarkably valuable for benchmarking and performance analysis purposes. However, they are acquired with highly sensorized, extremely controlled setups. This approach does not scale to the need of a high variability in the data for training purposes. A recent trend for object grasping is to synthetically generate the data for model training. For instance, [30] generates a dataset under the hypothesis that common household objects can be decomposed into one or more primitive shapes each of those having a family of associated grasps. Other approaches in robotics, like [31], identify contact maps on objects of interest, specifying the dexterous hand configuration for grasping it. However these
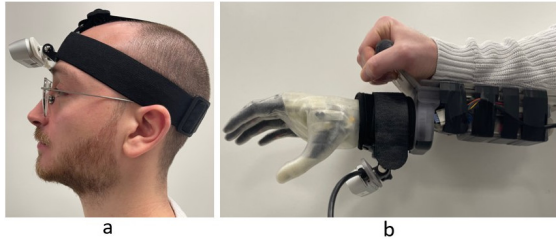
Fig. 2. Pictures of the developed wearable sensorized experimental setup and of the used Hannes prosthesis [4].

approaches limit their contribution to generating the correct fingers position on the object, while in the task of pre-shape classification the arm trajectory towards the target is also important to discriminate between different pre-shapes for the same object. In prosthetics, the HandCam dataset [13] has been recently presented. The proposed setting is similar to the one used in this work. However, the dataset has a great limitation, i.e., it does not consider the realistic case in which different grasps can be associated to different parts of the same object [32] and this reduces the problem to an object categorization. Finally, regarding synthetic visual data generation, this technique is rarely applied to prosthetics. To the best of our knowledge, only one work [33] has been proposed that generates a synthetic dataset for the task of target object identification.

In this work, we generate a synthetic dataset to address the task of hand pre-shape classification. Specifically, we consider the possibility to associate different grasp types to different object parts (see Fig. 1a). This makes the task more difficult since identifying the object is not enough for correctly predicting the pre-shape. The arm trajectory towards the object part, during the approach, has to be considered as well. Finally, we apply domain randomization [34] to improve generalization capabilities of the model.

III. METHODS

In the *shared autonomy* framework, the prosthesis user and the control system need to cooperate in order to achieve the target grasping task. Ideally, this cooperation should be transparent to the user to allow a natural prosthesis utilization with the lightest cognitive load possible. At the same time, the user should feel to be in full control of the prosthesis and not vice-versa. To this aim, the start of a grasp movement, the timing for the fingers closure around the object and the force to apply could be left to the user (e.g., by means of muscle activation read through EMG sensors). On the contrary, the selection of the hand pre-shape can be delegated to the prosthesis control system.

Thus, differently from previous work [13], we focus on pre-shape (instead of grasp type) classification. We propose a system that can predict the best prosthesis pre-shape to grasp a given object, based on visual input. In order to exploit motion cues, we consider sequences of RGB frames taken in the time interval between the start of the arm movement towards the object and the moment before the hand touches the target. Note that, in this setting, the sequence starts when

the user has their hand pointing to the object (i.e., in view from the *eye-in-hand* camera) and finishes when the hand touches it for grasping it. Hence, we frame the task as a pre-shape classification of visual sequences. We consider four pre-shape classes, namely, *Power*, *Lateral*, *Pinch* and *Pinch 3 Digit* (refer to Tab. I for the thumb configuration for each pre-shape). Then, we add the class *No grasp*, to discriminate when no pre-shape needs to be executed. Finally, note that we hypothesize to have only one object in the scene. The extension to multiple objects would include a preliminary step of target object identification such as the ones proposed in [33], [20] which is out of the scope of this paper but can be considered as future work.

In this paper we evaluate the advantages of using synthetic data for pre-shape classification learning. To this aim, we rely on two well-established learning-based models that share the same high-level architecture, i.e. (i) Convolutional Neural Network (CNN) based feature extraction (CNN backbone), followed by (ii) a pre-shape classifier. The first block encodes each image into a convolutional feature vector which is then taken as input by the classifier to predict a pre-shape for each image. In this work, we use *Mobilnet V2* [35] as CNN backbone, pre-trained* on the ImageNet dataset [36]. For pre-shape classification, we adopt either *Fully-connected layer* (*CNN + FC*) or *Long short-term memory* [37] (*CNN + LSTM*). For the former, we use one FC layer of dimension C, where C is the number of classes. For the latter, we use an LSTM with 256 hidden units, followed by a FC layer with dimension C to obtain the class scores. The input image size is 224x224px. The batch size is 32 when training on real dataset and 256 when training on synthetic dataset. In both cases, we fix the weights of the CNN backbone, while we train from scratch the classifier on top with a cross-entropy loss. In order to choose the number of training epochs, we evaluate performance on a validation set. Specifically, we reduce the learning rate (i.e. learning rate reduction on plateau) by a factor of 0.1 if the validation loss does not decrease after 6 epochs, and we stop the training (i.e. early stopping) if the validation loss does not decrease after 10 epochs (we initialize the learning rate to 0.0005). Moreover, in order to counteract any unbalancing in the dataset, before starting each epoch we randomly downsample every class to the cardinality of the minority one.

Finally, the models are used to predict a pre-shape for each sequence. Specifically, they predict one of the four pre-shapes (or the class *No grasp*) for each frame. The final prediction for the sequence is given by a majority voting, i.e., the pre-shape class that has been predicted for most of the frames (excluding the class *No grasp*).

IV. SYNTHETIC DATA GENERATION

We simulate the work space as a room with a table in the center (refer to Fig. 1b, for an example of the simulated setting). We render a human arm and we rigidly attach a

---

*https://download.pytorch.org/models/mobilenet_v2-b0353104.pth

Fig. 3. Sample frames of the proposed synthetic dataset and of the collected real sequences for the different training and test sets.

RGB sensor below the wrist, simulating the view a camera would have if placed as in Fig. 2b. Then, we place the 3D model of an object on the table and we simulate a human arm trajectory that makes the hand reach the object to grasp.

Different grasp types for different object parts. For this work, we chose a subset of 15 objects from the YCB-Video [38] dataset (refer to the last row of Tab. I for the list of all the objects) and the corresponding 3D models[†]. We choose objects from this dataset in order to ease experiments reproducibility and comparison with the literature. However, the same approach can be applied to any object with a 3D model available. Firstly, we assign one grasp type to each graspable object part. Specifically, we take inspiration by [10] for the objects grasp types association but we exclude or modify those that are not possible to execute with our prosthesis. Then, since in the *shared autonomy* framework we aim to address a pre-shape classification task, we associate a pre-shape to each considered grasp type, following [11]. Tab. I overviews the resulting associations of the chosen grasp types, pre-shapes and objects. Note that, in this way, we create a link between each object part and the corresponding grasp type and pre-shape. In order to implement this in the rendering pipeline, we overlay a parallelepiped to each object part of interest and we assign it the correct grasp type (refer to Fig. 1a for grasp type annotation examples). If, while executing an arm trajectory, the hand clashes with one parallelepiped, the frame sequence is annotated as the corresponding grasp-type. Note that, differently from previous work [13], we consider the case in which different object parts can be associated to different grasp types. This is a fundamental aspect because, in practice, one object can be grasped differently from different parts, depending on its use [32]. In this case, simply recognizing the object is insufficient, we show in Sec. V that such a naïve approach brings to a grasping performance degradation even in the ideal case of 100% object classification accuracy.

Human-like arm trajectory generation. The arm trajectory towards the object part, together with the object type, conveys useful information to predict the grasp type. Hence, simulating realistic trajectories is key in this setting. To this purpose, we model the arm and the hand as a mesh with a RGB camera rigidly attached below the wrist. In this setting, the trajectory to model consists of the movement

of a 6D point in space. Since our input is only RGB frames, we aim to model a camera trajectory resembling the one that the camera would have if mounted on a real person approaching the target object with their hand, to grasp it. In our setup (both real and synthetic), we placed the camera such that no finger occlusions occur. However, since we adopt a photorealistic arm model[‡], simulating different camera placements —including such occlusions— would be possible if needed. To make a realistic hand approach, referring to the literature on human arm movements [39], we implement a minimum jerk trajectory, i.e., on a straight line, with a bell-shaped velocity profile. Moreover, to simulate the pronation-supination and flexion-extension of the wrist at the end of the reaching, we rotate the camera. In particular, while approaching the object, the hand rotates such that at the end of the movement the normal to the palm becomes parallel to the normal to the external face of the parallelepiped linked to the target object part.

Domain randomization. In order to cover the well-known sim-to-real gap [34] and to generalize to novel conditions, it is necessary to introduce variability in the synthetically generated training set. To this end, by following a well-established practice in computer vision [34], we randomize the textures of the room walls, floor and table and we randomly vary the light conditions. Moreover, we randomize the object pose on the table (refer to Fig. 3, for example of generated frames). Finally, we addressed the problem of introducing sources of randomization in the process of arm trajectory generation. Differently from the above aspects, this one is peculiar of the considered application and hence we devised a novel randomization strategy. Specifically, for each simulated approaching trajectory we fix the final point as the center of the chosen parallelepiped, whereas we randomly select the initial one. Precisely, this is randomly sampled from a plane perpendicular to the table-top and placed at a distance from the table border such that both the object and part of the background are in view. Before a sequence execution, given an initial and final points couple, the following procedure is executed. Firstly, a straight line trajectory is generated to check for collisions. Then, if the collision with the considered parallelepiped happens, the trajectory is executed and the sequence is labeled with the grasp type represented by the parallelepiped. Otherwise, e.g.,

---

[†]http://ycb-benchmarks.s3-website-us-east-1.amazonaws.com

[‡]https://www.cgtrader.com/free-3d-models/character/man/fps-arms-pack

if the trajectory collides with the object mesh, it is discarded. This ensures that given a *grasp type - object part* pair, only reasonable sequences are generated (e.g., it is impossible to grab a mug by the handle if this is not visible). Thus, the final dataset is composed of video sequences taken from the RGB camera on the wrist of the simulated arm reaching the object. Each sequence is annotated with the chosen object, grasp type and pre-shape.

We developed the described data generation approach using the Unity[§] simulation engine to obtain photo-realistic frame sequences and integrating the Perception package[¶] to ease domain randomization and sequence labeling.

V. EXPERIMENTS

A. Experimental setup

In order to evaluate the proposed synthetic data generation technique, we generate a dataset of 47 sequences for each *grasp type - object part* pair (as defined in Tab. I), resulting into 1457 videos. Notice that the superscripts on the object names in Tab. I indicate the number of object parts having the same grasp type (e.g., the mustard in Fig. 1a has the *large diameter* grasp type on both bottle sides). Where not differently specified, we use this dataset to train the considered learning models.

With the aim to test all the components of the proposed work, we develop a wearable sensorized setup composed of two Intel RealSense D435 cameras: one is worn on the head and the other on the wrist, with two elastic bands. We use this setup both to collect sequences for benchmarking purposes and to test the control pipeline directly on the Hannes prosthesis. Specifically, three healthy subjects collected several sequences. The first subject stands in front of the object to grab, placed on a table-top and collected 311 sequences. For each *grasp type - object part* pair, he was asked to vary the approaching direction and object pose on the table. Each sequence is 3 seconds long, where the first 2.5 seconds are labeled as the grasp type executed by the subject and the last 0.5 seconds as *No Grasp* (since the corresponding frames show the moment of hand-object contact). We split this set of sequences into three sub-sets, balancing the different objects and grasp types presence. This results in sub-sets of 46, 58 and 207 sequences. We use the first as a validation set for model training to choose the number of optimization epochs as described in Sec. III. The second one, referred to as *Same person*, is used as a test set to evaluate the performance of the proposed approach for this subject in the following sections. Finally, the third one is used as a real training set, for comparing the performance of models trained on it with the proposed synthetic dataset. Refer to Fig. 3 (*Same person*) for example frames of this dataset. Then, other two healthy subjects collected four further sets, performing two trials for each *grasp type - object part* in each set. Refer to Fig. 3 for example frames of each of them. The sets are designed as follows:

- *Different velocity*. The grasp is completed in 1.5 seconds instead of 3 seconds. The hand movement starts from the same distance to the object as for the first subject, thus resulting in a faster approach.
- *From ground*. Rather than starting with the object in view, (i) the arm is initially extended along the side, (ii) the subject raises their arm, (iii) approaches the object and (iv) grasps it. All steps are performed smoothly. Starting with the arm along the side produces different trajectories resulting in different object views. For our experiments, the first part of the sequence is trimmed out at inference time as the camera points at the ground.
- *Seated*. The subject performs the grasp while seating in front of the table. The different body posture influences the initial hand pose with respect to the object and thus the approach. The grasp is completed in 1.5 seconds.
- *Different background*. We vary the tablecloth underneath the object to be grasped. We do this to evaluate the system performance on a different background.

Since the considered learning models output a prediction for each frame but the final prediction for the entire sequence is given by majority voting (see Sec. III), the performance is evaluated by considering a per-video Accuracy. Results are reported as mean and standard deviation over 3 trials of the same experiments, by varying the random seed.

B. System design motivation

In this section, we empirically motivate our main design choices.

One object, multiple grasps. In order to demonstrate the importance of having different grasps for different object parts, we show the performance loss that one would achieve on our dataset with a model predicting only one grasp type for each object. For doing this, we consider an ideal model that can discriminate with 100% accuracy among the 15 objects chosen for our experiments. Then, an object-grasp type mapping is required. Hence, we consider the grasp type-object association of Tab. I. Note that, for objects with multiple grasp possibilities, the most frequent one is chosen. This theoretical model is evaluated on the *Same person* test set (see the first row of Tab. II) that we labeled with multiple grasps per object. By using the theoretical single grasp model, a user could correctly grasp the object only 79.3% of the time. Precisely, single grasp objects are predicted with 100% accuracy, while multi grasp objects obtain only 56% accuracy even considering perfect object classification. This motivates the need for considering different grasp types for different object parts when learning the pre-shape classification model and confirms that our dataset well reflects this challenge. Moreover, the higher accuracy obtained by our model (see second row in Tab. II) confirms its disambiguation capability between different pre-shapes for the same object.

Egocentric vs eye-in-hand. The camera positioning is another crucial aspect of the system design. It affects the quality of the collected RGB sequences and therefore the prediction performance. Moreover, some placements might be invasive and affect the prosthesis usability. To compare the two

configurations, we consider data acquired by the first subject with our setup (see Sec. V-A) from both cameras (on head and hand) and we use them to train and test the two different models presented in Sec. III. Results are presented in Tab. II. In general, eye-in-hand models obtain higher performance and less variability than the egocentric ones. Additionally, in Fig. 4 we report the per-video accuracy achieved at each frame by the two CNN + FC models throughout the approaching sequence. We compare performance obtained for single grasp (dashed line) and multi grasp (solid line) objects. As it can be observed, regarding single grasp objects, in both setups the model identifies the correct pre-shape since the beginning of the sequence (because in this case this is uniquely identified by the recognized object). Differently, for multi grasp objects, the eye-in-hand data clearly brings more information to discriminate between different pre-shapes for the same object between second 0.7 and 1.7. This may be due to the fact that as the camera gets closer to the object, the target part to grasp becomes the most visible one. This is shown by the accuracy peak at time 1.5s. Note that, there is a drop for all models around time 2.5s. This is due to the label switch from pre-shape to *No grasp*. To further support our design choice, we notice that past literature [40] shows that the maximum hand aperture, during grasping, happens at 70% of the reaching phase (i.e., ~2.1s in our case). Since for our setting and for the considered sequences length, the performance peak for the eye-in-hand configuration happens at ~1.5s, we believe that this is a promising result to obtain a smooth reach-and-grasp movement since the prediction seems to come in useful time to actuate the hand aperture. Considering also the minor invasiveness of the eye-in-hand configuration, we consider this configuration for our work.

FC vs LSTM. Previous work [13] suggests that recurrent models (CNN + LSTM) should outperform single frames ones (CNN + FC) for the task of hand pre-shape classification. However, our problem is different from [13] since we consider the case in which the same object can be associated to different grasp types. Therefore, we compare the two different learning architectures described in Sec. III in our multi grasp setting. To this aim, we consider the data acquired by the first subject and we use it to train and test the two architectures. Results are presented in Tab. II. As it can be observed, in our case, the single frames model (CNN + FC) clearly outperforms the recurrent one (CNN + LSTM). Interestingly, all the errors for the recurrent models come from the multi grasp objects. This is due to the fact that, as previously discussed, the first frames of the grasping sequences are ambiguous for those objects. This ambiguity is an issue especially for the recurrent models since for them it is more difficult to recover from an early error than for the single frames one. Note that, considering only single grasp objects as in [13] might hide this issue. Given these results, we consider the CNN + FC architecture for our work.

C. Learning pre-shape classification from synthetic data

One major contribution of this work is the development of a synthetic visual data generation tool for prosthetic grasping.

		Egocentric Video acc (%)	Eye-in-hand Video acc (%)
Single-grasp Training	Perfect classification	79.3 ± 0.0	79.3 ± 0.0
Multi-grasp Training	CNN + FC	96.6 ± 1.4	98.9 ± 0.8
	CNN + LSTM	87.4 ± 3.5	92.0 ± 2.1

TABLE II

COMPARISON BETWEEN MODELS TRAINED WITH THE EGOCENTRIC AND EYE-IN-HAND CONFIGURATIONS FOR BOTH SINGLE GRASP (CASE OF PERFECT CLASSIFICATION) AND MULTI GRASPS PER OBJECT SETTINGS.
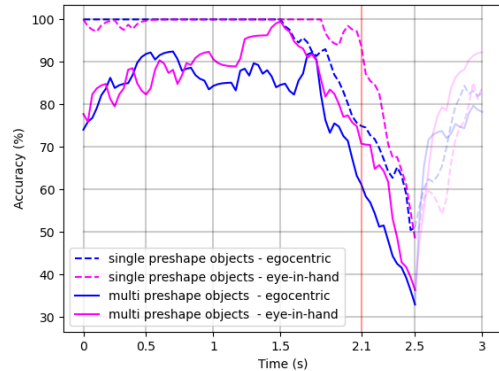

Fig. 4. Accuracy trends comparison over time during a grasping sequence of models trained with egocentric (blue) and eye-in-hand (pink) data. We report performance for single (dashed) and multi grasp (solid) objects.

In this section, we evaluate the effectiveness of the proposed approach by using the generated data to train a learning-based model for pre-shape classification. We compare the obtained performance against a model trained on real data, i.e., on the 207 sequences collected with the first subject, by testing on different real use case conditions, namely, over the test sets collected with our sensorized setup as described in Sec. V-A. We report results in Tab. III.

Firstly, we evaluate the models on the *Same person* test set (first row in Tab. III). In this case, the real training and test sets are drawn from the same distribution, being acquired by the same subject under the same visual conditions, while varying only the approach direction. Conversely, the synthetic training set has been generated with no knowledge of the subject posture and background. As expected, in this case a model trained on the real dataset fits extremely well the user and visual conditions while the one trained on synthetic data performs worse. Interestingly, while the overall accuracy of 80.2% is close to the 79.3% obtained by the theoretical model trained on single-grasp labels discussed previously, in this case, $38.1\% \pm 0.02$ of the wrong predictions come from single grasp objects while $61.9\% \pm 0.02$ come from multi grasp ones.

Then, we analyze the performance of the same models in case of trajectory variations of the arm during the approaching phase (rows from 2 to 4 in Tab. II). Notably, in the first two test sets the synthetic model achieves performance comparable to the real one, and in the third test set the synthetic model performs clearly better. Specifically, the *Seated* set is the most challenging one since its sequences present a more curved trajectory and consequently different viewpoints

	Real tr. set Video acc (%))	Synthetic tr. set Video acc (%))
Same person	98.9 ± 0.8	80.2 ± 0.9
Different velocity	81.7 ± 0.9	79.7 ± 0.8
From ground	76.2 ± 1.0	76.0 ± 0.9
Seated	63.9 ± 1.0	68.1 ± 1.0
Different background	56.2 ± 1.7	76.4 ± 2.0

TABLE III

VIDEO ACCURACY OF THE MODELS TRAINED ON A REAL AND ON THE PROPOSED SYNTHETIC DATASETS, COMPARED ON DIFFERENT USE CASES

while approaching the object. While the real model exhibits a more pronounced drop in this case, the synthetic model has a smaller decrease. This confirms the robustness of the trajectories generated in the proposed synthetic dataset.

Finally, we compare the accuracy obtained by the learned models on the *Different background* set (last row in Tab. III). The different pattern in the background produces a performance gap for both the real and synthetic models with respect to the accuracy on *Same person*. However, this gap is remarkably smaller ($\sim 4\%$) for the synthetic model than for the real model ($\sim 42\%$). This confirms results of previous work on domain randomization [34] and demonstrates the effectiveness of the proposed approach for pre-shape classification under different visual conditions.

As a final remark, these experiments show that except from tests on the same subject used for training, models trained on the synthetic dataset have comparable performance to models trained on real data. Moreover, in case of more challenging conditions in terms of arm movements and background variations, the synthetic dataset resulted to be crucial, while the real one led to a significant performance drop. Therefore, a real training set can be considered the best solution only for scenarios in which a dataset is available and the conditions are known, as shown in a similar work [13]. However this is practically unfeasible for prosthetic applications. Moreover, we show that the proposed approach to model eye-in-hand camera trajectories is not only effective to fill the sim2real gap but also achieve higher trajectory generalization.

D. Application on the Hannes prosthesis

As a further contribution, we test the presented eye-in-hand shared control system on an improved version of the Hannes prosthetic hand [4]. Hannes is a myoelectric poly-articulated prosthesis equipped with electromyographic sensors, a battery pack designed to last up to 1 day, and control electronics, all placed inside the wearable socket (refer to [4] for further details). Remarkably, for object grasping, the version of Hannes used in this work comprises an active abduction/adduction joint for the thumb and a three-digital modality can be enabled. This allows to execute the four different pre-shapes considered in this work (namely, *Power*, *Lateral*, *Pinch* and *Pinch 3 Digit*).

A more advanced eye-in-hand system prototype would have the RGB camera embedded into the palm of the hand prosthesis. However, for the first testing of our pipeline on Hannes, we integrate it with our sensorized setup and we use an external able-bodied adapter to allow a healthy user

to drive it. The final setup is shown in Fig. 2. All the computation for pre-shape classification is performed on an external GPU-enabled laptop which is connected to the on board control electronics of Hannes via serial connection. The experiment is designed as follows. A healthy user, handling the prosthesis, stands in front of the object to grasp placed on a table. After a starting trigger (for this analysis we considered a keyboard signal), the user approaches the object, moving the prosthesis towards it as to grasp it. The approach lasts ~ 2 seconds and the system collects RGB frames from the wrist-mounted camera. These frames are fed to the classifier (the CNN + FC model trained on synthetic data from Sec. V) and a pre-shape is predicted. The pre-shape class is converted into the corresponding thumb and three-digital configuration and this is sent to the on board control electronics which translates it into control signals for the joints. Next, the pre-shape is executed and then the fingers close around the object. Note that, the aim of this experiment is to integrate the proposed approach with a prosthesis and evaluate performance, therefore in order to automatise the starting trigger and the fingers closure we just temporised them, but we plan to implement more refined approaches also for these steps (e.g., with EMG signals). A video demonstration of the resulting system is attached as supplementary material and it shows the usability of the proposed approach for prosthetic grasping.

VI. CONCLUSIONS

We present an eye-in-hand shared control pipeline for prosthetic grasping. Specifically, we tackle the problem of hand pre-shape classification of RGB sequences from a wrist mounted camera with a learning-based approach. Generalization capabilities to different conditions (e.g., backgrounds and arm trajectories) are critical for prosthetic applications. To address this, we develop a synthetic human-like trajectories generation tool and we demonstrate that models trained with the proposed synthetic dataset achieve comparable or higher performance than models trained on real data on practical use cases. To the best of our knowledge, this is the first attempt to use synthetic visual data for the task of pre-shape classification for prosthetic grasping.

Differently to previous literature, we considered the case where different grasp types can be associated to different object parts. While increasing the task complexity, this aspect is fundamental in order to devise an effective prosthesis control system, and thus increase the user satisfaction. We show that, by using well established predictive models, the eye-in-hand configuration allows to get better results in this setting than the egocentric one. Classification performance might be improved by exploiting more sophisticated techniques, such as the ranking labels [41]. Furthermore, the current model and data generation technique can be extended to deal with a more fine-grained supervision. For instance, each object part can be segmented and associated to the corresponding pre-shape, leading to the affordace segmentation task [26].

REFERENCES

[1] M. G. Catalano, G. Grioli, E. Farnioli, A. Serio, C. Piazza, and A. Bicchi, "Adaptive synergies for the design and control of the pisa/iit softhand," *The International Journal of Robotics Research*, vol. 33, no. 5, pp. 768–782, 2014.

[2] P. Weiner, J. Starke, F. Hundhausen, J. Beil, and T. Asfour, "The kit prosthetic hand: Design and control," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2018, pp. 3328–3334.

[3] G. K. Patel, C. Castellini, J. M. Hahne, D. Farina, and S. Dosen, "A classification method for myoelectric control of hand prostheses inspired by muscle coordination," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 26, no. 9, 2018.

[4] M. Laffranchi, N. Boccardo, S. Traverso, L. Lombardi, M. Canepa, A. Lince, M. Semprini, J. Saglia, A. Naceri, R. Sacchetti, *et al.*, "The hannes hand prosthesis replicates the key biological properties of the human hand," *Science Robotics*, vol. 5, no. 46, p. eabb0467, 2020.

[5] M. Gardner, C. S. Mancero Castillo, S. Wilson, D. Farina, E. Burdet, B. C. Khoo, S. F. Atashzar, and R. Vaidyanathan, "A multimodal intention detection sensor suite for shared autonomy of upper-limb robotic prostheses," *Sensors*, vol. 20, no. 21, p. 6097, 2020.

[6] M. Simao, N. Mendes, O. Gibaru, and P. Neto, "A review on electromyography decoding and pattern recognition for human-machine interaction," *Ieee Access*, vol. 7, pp. 39 564–39 582, 2019.

[7] J. Mouchoux, M. A. Bravo-Cabrera, S. Dosen, A. F. Schilling, and M. Markovic, "Impact of shared control modalities on performance and usability of semi-autonomous prostheses," *Frontiers in Neurorobotics*, vol. 15, 2021.

[8] N. E. Krausz and L. J. Hargrove, "A survey of teleceptive sensing for wearable assistive robotic devices," *Sensors*, vol. 19, no. 23, 2019.

[9] M. Cognolato, M. Atzori, R. Gassert, and H. Müller, "Improving robotic hand prosthesis control with eye tracking and computer vision: A multimodal approach based on the visuomotor behavior of grasping," *Frontiers in Artificial Intelligence*, p. 199, 2022.

[10] N. Wake, D. Saito, K. Sasabuchi, H. Koike, and K. Ikeuchi, "Object affordance as a guide for grasp-type recognition," *arXiv preprint arXiv:2103.00268*, 2021.

[11] T. Feix, J. Romero, H.-B. Schmiedmayer, A. M. Dollar, and D. Kragic, "The grasp taxonomy of human grasp types," *IEEE Transactions on human-machine systems*, vol. 46, no. 1, pp. 66–77, 2015.

[12] 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*, vol. 14, no. 3, 2017.

[13] L. T. Taverne, M. Cognolato, T. Bützer, R. Gassert, and O. Hilliges, "Video-based prediction of hand-grasp preshaping with application to prosthesis control," in *2019 International Conference on Robotics and Automation (ICRA)*, 2019, pp. 4975–4982.

[14] P. J. Kyberd, C. Wartenberg, L. Sandsjö, S. Jönsson, D. Gow, J. Frid, C. Almström, and L. Sperling, "Survey of upper-extremity prosthesis users in sweden and the united kingdom," *JPO: Journal of Prosthetics and Orthotics*, vol. 19, no. 2, pp. 55–62, 2007.

[15] C. Shi, D. Yang, J. Zhao, and H. Liu, "Computer vision-based grasp pattern recognition with application to myoelectric control of dexterous hand prosthesis," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 28, no. 9, pp. 2090–2099, 2020.

[16] T. Lampe and M. Riedmiller, "Acquiring visual servoing reaching and grasping skills using neural reinforcement learning," in *The 2013 international joint conference on neural networks*. IEEE, 2013.

[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, 2014.

[18] M. Markovic, S. Dosen, D. Popovic, B. Graimann, and D. Farina, "Sensor fusion and computer vision for context-aware control of a multi degree-of-freedom prosthesis," *Journal of neural engineering*, vol. 12, no. 6, p. 066022, 2015.

[19] M. Zandigohar, M. Han, M. Sharif, S. Y. Gunay, M. P. Furmanek, M. Yarossi, P. Bonato, C. Onal, T. Padir, D. Erdogmus, *et al.*, "Multimodal fusion of emg and vision for human grasp intent inference in prosthetic hand control," *arXiv preprint arXiv:2104.03893*, 2021.

[20] Y. He, R. Kubozono, O. Fukuda, N. Yamaguchi, and H. Okumura, "Vision-based assistance for myoelectric hand control," *IEEE Access*, vol. 8, pp. 201 956–201 965, 2020.

[21] D. K. Madusanka, R. Gopura, Y. Amarasinghe, and G. K. Mann, "Hybrid vision based reach-to-grasp task planning method for transhumeral prostheses," *IEEE Access*, vol. 5, pp. 16 149–16 161, 2017.

[22] S. Došen, C. Cipriani, M. Kostić, M. Controzzi, M. C. Carrozza, and D. B. Popović, "Cognitive vision system for control of dexterous prosthetic hands: experimental evaluation," *Journal of neuroengineering and rehabilitation*, vol. 7, no. 1, pp. 1–14, 2010.

[23] F. Hundhausen, D. Megerle, and T. Asfour, "Resource-aware object classification and segmentation for semi-autonomous grasping with prosthetic hands," in *2019 IEEE-RAS 19th International Conference on Humanoid Robots (Humanoids)*. IEEE, 2019, pp. 215–221.

[24] F. Hundhausen, J. Starke, and T. Asfour, "A soft humanoid hand with in-finger visual perception," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2020.

[25] F. Hundhausen, R. Grimm, L. Stieber, and T. Asfour, "Fast reactive grasping with in-finger vision and in-hand fpga-accelerated cnns," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, pp. 6825–6832.

[26] E. Ragusa, C. Gianoglio, S. Dosen, and P. Gastaldo, "Hardware-aware affordance detection for application in portable embedded systems," *IEEE Access*, vol. 9, pp. 123 178–123 193, 2021.

[27] 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, no. 1, pp. 1–13, 2018.

[28] Y.-W. Chao, W. Yang, Y. Xiang, P. Molchanov, A. Handa, J. Tremblay, Y. S. Narang, K. Van Wyk, U. Iqbal, S. Birchfield, *et al.*, "Dexycb: A benchmark for capturing hand grasping of objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9044–9053.

[29] F. Krebs, A. Meixner, I. Patzer, and T. Asfour, "The kit bimanual manipulation dataset," in *2020 IEEE-RAS 20th International Conference on Humanoid Robots (Humanoids)*, 2021, pp. 499–506.

[30] Y. Lin, C. Tang, F.-J. Chu, and P. A. Vela, "Using synthetic data and deep networks to recognize primitive shapes for object grasping," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 10 494–10 501.

[31] T. Zhu, R. Wu, X. Lin, and Y. Sun, "Toward human-like grasp: Dexterous grasping via semantic representation of object-hand," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 741–15 751.

[32] T. Feix, I. M. Bullock, and A. M. Dollar, "Analysis of human grasping behavior: Correlating tasks, objects and grasps," *IEEE transactions on haptics*, vol. 7, no. 4, pp. 430–441, 2014.

[33] B. Zhong, H. Huang, and E. Lobaton, "Reliable vision-based grasping target recognition for upper limb prostheses," *IEEE Transactions on Cybernetics*, 2020.

[34] J. Tobin, R. Fong, A. Ray, J. Schneider, W. Zaremba, and P. Abbeel, "Domain randomization for transferring deep neural networks from simulation to the real world," in *2017 IEEE/RSJ international conference on intelligent robots and systems*. IEEE, 2017, pp. 23–30.

[35] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 4510–4520.

[36] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[37] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[38] Y. Xiang, T. Schmidt, V. Narayanan, and D. Fox, "PoseCNN: A convolutional neural network for 6d object pose estimation in cluttered scenes," 2018.

[39] T. Flash and N. Hogan, "The coordination of arm movements: an experimentally confirmed mathematical model," *Journal of neuroscience*, vol. 5, no. 7, pp. 1688–1703, 1985.

[40] M. Jeannerod, "The timing of natural prehension movements," *Journal of motor behavior*, vol. 16, no. 3, pp. 235–254, 1984.

[41] M. Zandigohar, M. Han, D. Erdoğmuş, and G. Schirner, "Towards creating a deployable grasp type probability estimator for a prosthetic hand," in *Cyber Physical Systems. Model-Based Design*. Springer, 2019, pp. 44–58.