Automatic Grasp Selection using a Camera in a Hand Prosthesis

Joseph DeGol, Aadeel Akhtar, Bhargava Manja, and Timothy Bretl

***Abstract*— In this paper, we demonstrate how automatic grasp selection can be achieved by placing a camera in the palm of a prosthetic hand and training a convolutional neural network on images of objects with corresponding grasp labels. Our labeled dataset is built from common graspable objects curated from the ImageNet dataset and from images captured from our own camera that is placed in the hand. We achieve a grasp classification accuracy of 93.2% and show through real-time grasp selection that using a camera to augment current electromyography controlled prosthetic hands may be useful.**

I. Introduction

This paper presents a prosthetic hand system augmented with a camera for use in automatic grasp selection (Figure 1). Initial work with grasp selection for prosthetic hands has used electromyography (EMG) [1]. The grasps most often classified using EMG include the power grasp, pinch grasp, tool grasp, 3-jaw chuck, and key grasp, and classification accuracies between 90-95% have been reported in laboratory conditions by Kuiken et. al. [2]. However, Castelini et. al. [3] showed that EMG control is still imperfect because EMG signals are stochastic and have issues with robustness. These issues manifest themselves in classification errors when selecting between multiple grasps in real-world settings.

An alternative to EMG is to use RFID chips as was done with the Infinite Biomedical Technologies Morph system [4]. However, Morph only works for objects that have been pre-tagged with RFID chips. Cameras offer another alternative. Work by Markovic et al. [5] used camera data fused with EMG for grasp selection, but the user had to wear the camera on his or her head and the camera was not used to directly classify appropriate grasps for objects. Our system investigates how a camera can be embdedded in a prosthetic hand for grasp selection and does not require object tagging (e.g. RFID systems) or external sensors (e.g. head camera).

We investigate how to use a camera embedded in a prosthetic hand for automatic grasp selection. Images captured by the camera in the hand are sent to a portable embedded processor that classifies each image as one of the five grasps (Kuiken et. al. [2]: power, pinch, tool, 3-jaw chuck, and key). We also contribute a dataset of annotated image data that maps close range images of objects to the appropriate grasp type and demonstrate that our system achieves 93.2% accuracy in grasp classification. Lastly, we provide snapshots of grasp classification experiments done in real time and provide the videos of these experiments on our website[1].

J. DeGol, A. Akhtar, B. Manja, and T. Bretl are with the University of Illinois, Urbana, IL 61801, USA {degol2, aakhta3, manja2, tbretl}@illinois.edu, bretl.csl.illinois.edu

[1]bretl.csl.illinois.edu/prosthetics

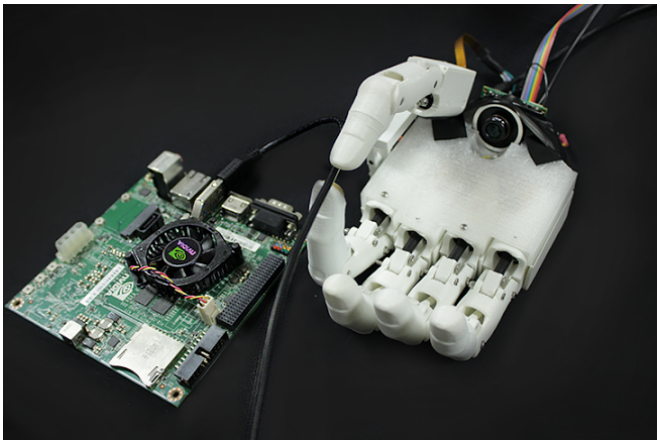


Fig. 1: In this paper, we augment the prosthetic hand of Slade et. al. [6] by adding a camera in the palm. We then use a subset of the ImageNet [7] dataset to train a convolutional neural network [8] to learn grasps. Finally, we validate our system by achieving 93.2% grasp classification accuracy on a test set of images captured from our hand camera.

These results are promising for the inclusion of cameras in future prosthetics and open the door to further investigation on how a hand camera can be used in conjunction with electromyography to improve grasp selection.

II. Dataset

In this section, we detail the data we used for testing automatic grasp selection with our prosthetic hand.

A. DeepGrasping

The DeepGrasping dataset [9] consists of images of objects on a table. Each image has a resoluton of 640 pixels x 480 pixels. In total, there are 1035 images of 280 objects. We augment this dataset by providing annotations for one of our five possible grasps. We labeled each object based on which grasp we felt was most natural for that object. If an object had more than one reasonable grasp choice, we chose the grasp that would cause a more uniform representation of grasps. The percentage of each grasp is shown in Table I. Note that most objects are labeled for power grasp, three jaw chuck, and pinch grasp. Tool grasp and key grasp are minimally represented. Because of the bias in this dataset, we were motivated to create a new dataset that more uniformly represents each grasp. Figure 2a shows some example images from the DeepGrasping dataset.

978-1-4577-0220-4/16/$31.00 ©2016 IEEE

DeepGrasping ImageNet HandCam

Fig. 2: We use three datasets to evaluate our system: Deep Grasping, ImageNet, and HandCam. Because of bias towards power, pinch, and three jaw chuck in Deep Grasping, we chose to create a dataset from ImageNet that more uniformly represents all five grasps. We then created the HandCam dataset from images captured by our hand camera to test the validity of our system.

Grasp	DeepGrasping	ImageNet	HandCam
Key	0.0 %	11.8 %	20.0 %
Pinch	21.8 %	10.6 %	20.0 %
Power	47.0 %	47.5 %	20.0 %
Three Jaw Chuck	28.0 %	19.2 %	20.0 %
Tool	3.2 %	10.9 %	20.0 %

Table I: This table shows the % of each grasp that is represented in the DeepGrasping, ImageNet, and HandCam datasets using our annotations. Key grasp and tool grasp are largely underrepresented in the DeepGrasping dataset so we created new datasets from ImageNet and our own hand camera that more uniformly represent each grasp.

B. ImageNet

Because of the lack of representation of key grasp and tool grasp in the DeepGrasping dataset. We created a new dataset. First, we downloaded images of common graspable objects from ImageNet [7] (a common dataset for object recognition). This data was curated down to 5180 images. The resolution of the images varies. These images can be grouped into 25 object categories: Ball, Basket, Blowdryer, Bowl, Calculator, Camera, Can, Cup, Deodorant, Flashlight, Glassware, Keys, Lotion, Medicine, Miscellaneous, Mugs, Paper, Pen, Remote, Scissors, Shears, Shoes, Stapler, Tongs, and Utensils. The Miscellaneous category contains various tools such as drills, knives, and hammers. These images were then annotated with what we felt was the most natural grasp. If an object had more than one reasonable grasp choice, we chose the grasp that would cause a more uniform representation of grasps. The percentage of each grasp is shown in Table I. Figure 2b shows some example images from the ImageNet dataset.

C. HandCam Testing Set

The goal is to evaluate how accurately a camera in a prosthetic hand can identify the appropriate grasp for a given object. Therefore, we create a third testing dataset consisting entirely of images captured using the camera in the hand. Each image has a resolution of 640 pixels x 480 pixels. For each grasp, ten objects were chosen and were photographed from five different perspectives. This gives us a total of 50 new objects and 250 test images. Examples are shown in Figure 2c.

It is worth clarifying that the HandCam dataset is used purely for testing; this means that the convolutional neural network (CNN) never trains using any images from the HandCam dataset. This is interesting because it means that the CNN does not train with any images captured by the hand camera. Lastly, note that because we chose ten objects for each grasp type, we have a uniformly represented test set; which is desirable for evaluation because a biased dataset can achieve high classification accuracy by simply weighting guesses towards the biased category.

III. Method

In this section, we describe adding a camera to a prosthetic hand and how automatic grasp selection is performed.

A. Adding a Camera to the Prosthetic Hand

We modified the design of [6] to fit our camera. Based on the dimensions of our camera (a PointGray FireflyMV USB2.0 [10]) and the space restrictions of the prosthetic hand, we found it most appropriate to place the camera in the palm. Our modifications allow the camera to sit in the palm and face outward towards an intended graspable object (Shown in Figure 2c). Images captured by the camera are then sent to an NVIDIA Tegra [11] for processing. The Tegra is a mobile processor equipped with a GPU which allows for complex image processing. As stated, the Tegra is a mobile processor and can, in principle, be equipped with the hand; making the entire system portable.

B. Automatic Grasp Selection

Convolutional Neural Networks (CNNs) have become the state-of-the-art method for object recognition in the computer vision community [8], [12]. These networks operate by taking an image as input and performing successive operations on the image such as filtering, max pooling, and rectification. Each of these operations is referred to as a layer. The network

of [12] (often referred to as VGG-VeryDeep-16 because of its 16 layers) has become one standard architecture for CNNs for object recognition. This network has achieved exemplary results in classifying 1000 objects in the ImageNet dataset [7].

We use the VGG-VeryDeep-16 architecture for automatic grasp selection. Specifically, we input images and corresponding grasp labels to the network for training. To get the network to classify five grasps, we edit the second to last layer of the network to consider only five possible classes. We decrease the learning rate of the architecture to properly tune the network to our new data and altered architecture. This tuning procedure is standard for training new CNNs.

IV. Results and Discussion

In this section, we demonstrate the utlity of our automatic grasp selection method and discuss the features the network learned for choosing grasps.

A. Grasp Classification Accuracy

Table II shows the classification accuracy of grasps on the HandCam test set when our CNN was trained using either the DeepGrasping or ImageNet data described in Section II. It is not surprising that training with DeepGrasping results in poor classification results in comparison to ImageNet because ImageNet consists of thousands of additional images at more varied perspectives and also contains objects of all five grasp classes. However, we include the results for training with DeepGrasping and testing with HandCam for completeness.

Focusing on the results when training with ImageNet and testing on HandCam, we see that the classification accuracy is 93.2%. This is a promising result that is on par with state-of-the-art EMG systems [2].

Table II also shows accuracy for each class. We see that the pinch grasp achieves the highest classification accuracy and the tool grasp achieves the lowest classification accuracy. We suspect that pinch performed the best because the objects labeled as pinch were often small and numerous (e.g. a pile of pills) or long and thin (e.g. a pen); which is easily differentiable from larger, singular objects often associated with the other grasps (e.g. bowls, bottles, balls, shoes, etc...).

We can see from the confusion matrix in Figure 3 that the reason tool grasp achieves the lowest accuracy is because it is often confused with power grasp. This is unsurprising because one of the key differentiators between tool and power grasps is the presence of a trigger; which might be occluded by the body of the object from certain camera views. For example, spray bottle (which has a trigger) is an object we labeled as a tool grasp; however, if the camera views it from certain perspectives, the trigger may be hidden, making the object look more like an object the network learned to associate with power grasps.

B. Grasping Objects in Real-Time

Figure 4 provides snapshots of our camera hand system successfully identifying the correct grasp type for five objects. In each case, an object was placed in front of the hand and the camera captured an image to classify the correct grasp (shown in the first column). Then, the hand was actuated into the correct grasp (columns two through five). All five grasps are represented, one per row. The videos that these snapshots were taken from, along with five additional real-time experiments can be found at bretl.csl.illinois.edu/prosthetics.

Testing with HandCam (%)	Training with Deep Grasping	Training with ImageNet
Mean Accuracy	19.6	93.2
Key	0.02	92.0
Pinch	0.0	98.0
Power	52.0	94.0
Three Jaw Chuck	44.0	96.0
Tool	0.0	86.0

Table II: Classification accuracies when testing on the HandCam dataset. The training data is either DeepGrasping (column 2) or ImageNet (column 3). The per-grasp testing accuracy is also shown. Mean classification accuracy on the HandCam testing set is 93.2% when training with ImageNet.

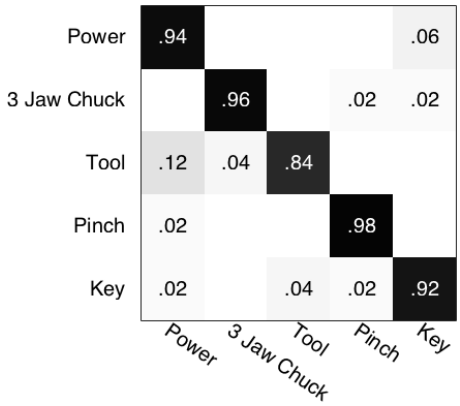
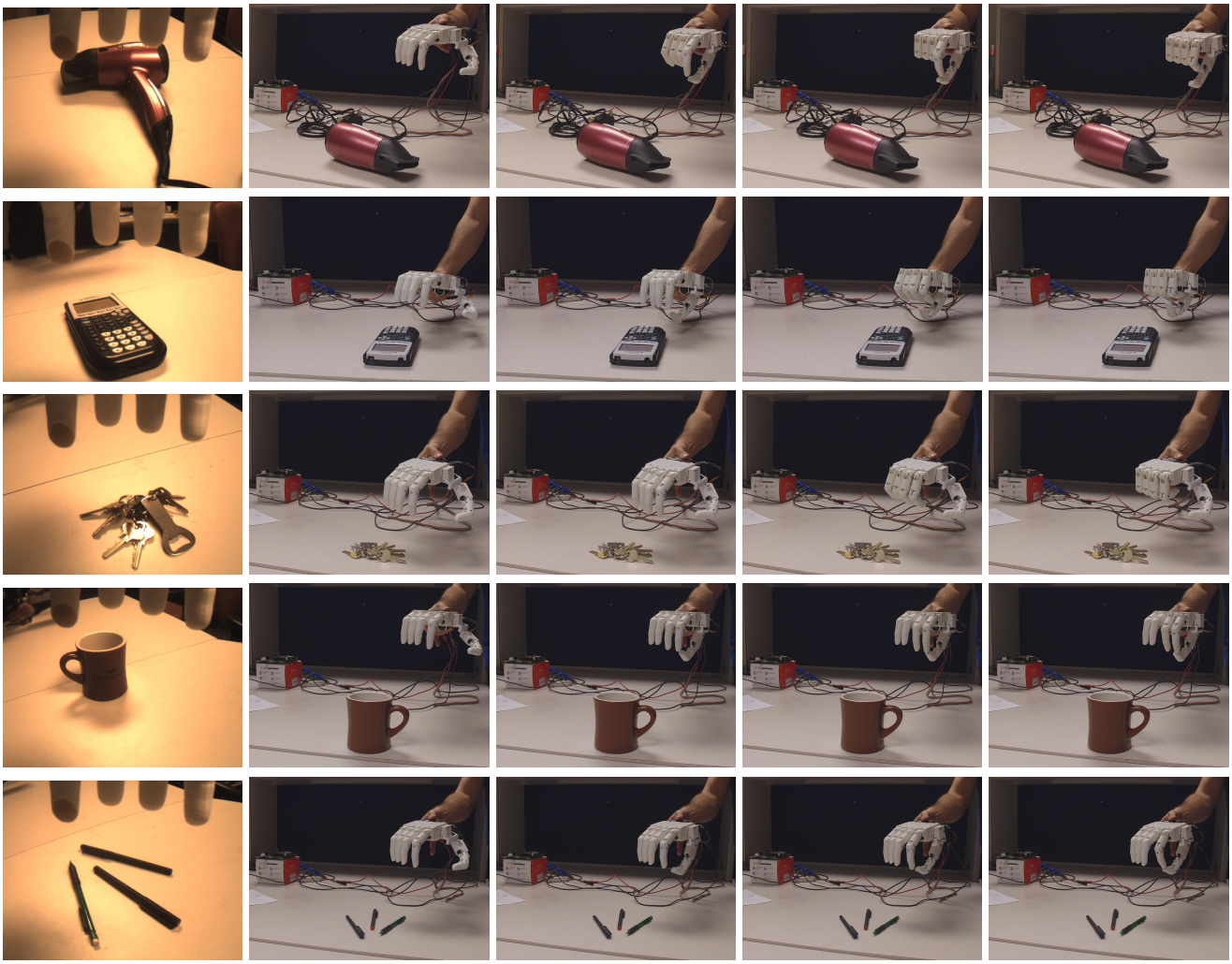


Fig. 3: The confusion matrix shows that the overall accuracy when training with our ImageNet data and testing on our HandCam data is 93.2%, which is a promising result that is on-par with current EMG systems. Note also that the most confused grasp is tool graps being labeled as power. We suspect that this confusion arrises because triggers (a defining feature of tool grasp objects) are occluded by the body of the object in the confused instances.

V. Conclusions

In this paper we augment a prosthetic hand with a camera and show that our system is a viable option for automatic grasp selection. Given the stochastic nature of EMG signals, a hand camera system can be especially useful in resolving large ambiguity in EMG grasp classification. Moreover, CNNs have been shown to perform well for classifying 1000 objects, so we expect to achieve high accuracies with the addition of more grasps (e.g. [13] identifies 24 unique grasps). With the addition of new grasps however, comes the ambiguity in choosing between several reasonable grasps for a given object; perhaps EMG signals can disambiguate these options. We leave determining the best way to combine EMG and camera data for grasp selection for future work.



Classified Image Grasp Snapshots

Fig. 4: Snapshots (columns 2 to 5) show automatic grasp selection in real-time using our prosthetic hand with an embedded camera. Column 1 shows the image that was processed for automatic grasp selection. Each row is a different grasp being correctly selected for the given object. From top to bottom, the grasps are Key, Pinch, Power, Three Jaw Chuck, and Tool.

ACKNOWLEDGMENT

This work is supported by the DoD National Defense Science and Engineering Graduate Fellowship (NDSEG), NSF grant IIS-1320519, and NIH grant F30HD084201. We gratefully acknowledge the support of NVIDIA Corporation for the donation of the Tegra used for this research. We also thank Patrick Slade for helping install the camera in the hand.

REFERENCES

[1] B. Hudgins, P. Parker, and R. N. Scott, "A new strategy for multifunction myoelectric control," *IEEE Trans. Biomed. Eng.*, 1993.

[2] T. Kuiken, G. Li, B. Lock, R. Lipschutz, L. Miller, K. Stubblefield, and K. Englehart, "Targeted muscle reinnervation for real-time myoelectric control of multifunction artificial arms," *JAMA*, 2009.

[3] 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. A. Kuiken, M. Markovic, P. M. Pilarski, R. Rupp, and E. Scheme, "Proc. of the first workshop on peripheral machine interfaces: Going beyond traditional surface electromyography," *Frontiers in Neurorobotics*, 2014.

[4] "Morph. infinite biomedical technologies. http://www.i-biomed.com."

[5] 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," *J. Neural Eng.*, 2015.

[6] P. Slade, A. Akhtar, M. Nguyen, and T. Bretl, "Tact: Design and performance of an open-source, affordable, myoelectric prosthetic hand," in *IEEE 2015 ICRA*.

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

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems 25*, 2012.

[9] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *IJRR*, 2014.

[10] "Point grey firefly mv usb2.0 camera. https://www.ptgrey.com/.."

[11] "Nvidia tegra mobile processor. http://www.nvidia.com/object/tegra.html.."

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[13] T. Feix, I. Bullock, and A. Dollar, "Analysis of human grasping behavior: Object characteristics and grasp type," *IEEE Trans. Haptics*, 2014.