Q2)

Computer vision is commonly used in robotics for object identification and dexterous object manipulations, however, one translation of this technology into the tangential neighboring field of prosthetics was only recently introduced.

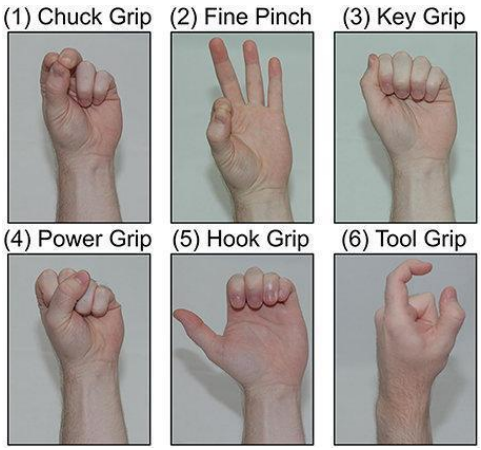

Image 0: Earley's Grasps

For preliminary context, most high dexterity forms of grasps can be categorized into 6 grasps - varying between the age of the user and research. Prosthetics in the past were triggered by point of leverage or EMG sensors, however for many patients with neurological damage (notably stroke victims) or amputees without intra-muscular insertion of nerve endings (another way of saying extra nerve endings are looped back up to futureproof for more advanced EMG prosthetics and reduce phantom pain), rendering myoelectric sensors are not an option. While in many cases, amputees can resort to a 2-finger binary opening ability, five-finger dexterous manipulation with different grasps can potentially enable a better quality of life. However, with the difficulty of higher numbers of nodes to control, the issue of limited control ability remains. As such, much recent research over the past decades has begun exploring multi-modal or alternative control methods to EMG sensor input as a manner to control grasps. Modes of input span from integration of camera for edge detection (Casley 2014, DeGol 2016),

to AR glasses for eye tracking (Makovic 2014), to voice control (Gruppioni 2008), to depth sensor (Lenz 2014). Of these modes of control, a culmination of visual and depth perception has shown the greatest promise

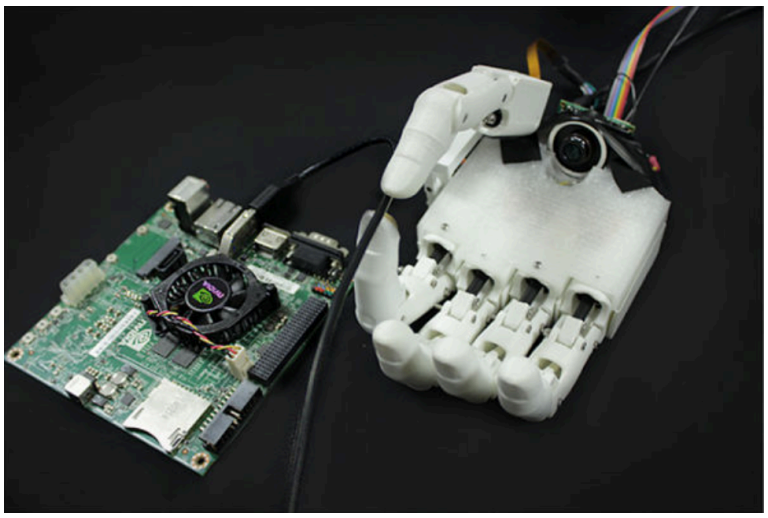

Image 1: DeGol 2016 Camhand

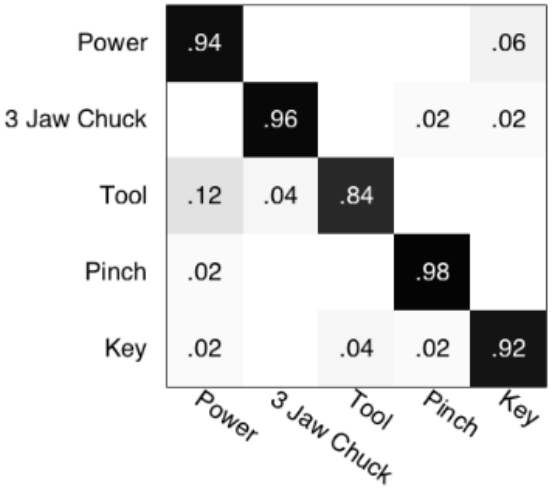

Image 2: DeGol's grasp correlation results.

The most notable research in this niche is DeGol's Camhand (Image 1) research, where he runs a Convolutional Neural Network through a set of images through manually labeled grasp assignment through imageNet. DeGol found great accuracy through his work (up to 98%),

however, he found bias in the cross-section that limited the accuracy of two grasps to 84% (Image 2). This displays great potential for computer vision's implementation into grasp toggling. I believe that this can be further pursued by implementing mocap or natural human grasp input for objects for grasp assignments as opposed to assigning grasps to static images via a screen, potentially eliminating biases.

Since DeGol's research, the multi-modal diffusion policy (inclusive of computer vision) (Chi 2023) has proven even greater potential for greater accuracy in different conditions beyond imageNet sample with computer vision, displaying even greater viability to implement automated grasp toggling for prosthetics. Lenz's 2014 is aggregate research exploring multiple prior research on different modes of input and explains greater potential to implementing depth-sensing ability. This idea of implementing depth sensing was further perpetuated by Chi's 2024 UMI research with robotic learning from mimicking human input with RGB and depth as input allowing for greater positional data accuracy and dexterity.

I believe an updated recreation of the DeGol camera hand with the latest policy learning capabilities can greatly aid the process of creating an automated grasp-toggling ability for rehab and prosthetic users.

Reference:

- Chi 2023 Diffusion policy: https://diffusion-policy.cs.columbia.edu/
- Chi 2024 UMI: https://umi-gripper.github.io/
- Lenz 2014: http://arxiv.org/abs/1301.3592
- DeGol 2016: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5325038/
- Markovic 2014: https://pubmed.ncbi.nlm.nih.gov/24891493/
- Gruppioni 2008: https://www.academia.edu/24769901/A_Voice_Controlled_Prosthesis_Test_of_a_Vocabulary_and_Development_of_the_Prototype
- Casley 2014 IRIS hand: https://digital.wpi.edu/concern/student_works/z603r0024?locale=en
- Early 2016: https://ieeexplore.ieee.org/document/7523682

Q3)

Data obtained to train an ML model can be split into a few different formats. I believe data can come in a few formats, it can either come in a depth map, regular RGB video with post-depth or post-gaussian processing, or static images.

In terms of a static dataset, many existing data banks such as imageNet, can be something like an existing bank of images such as imageNet, deepGrasp, or this random object assorted huggingFace, with subsequent.

In terms of depth map, while existing depth maps, such as this Kaggle bank have the potential to a Gazebo hand forming around objects, it fundamentally doesn't work with the objective of the experiment of assigning grasps to objects' shape and size without the physical object in front of the assignee. In other words, a depth map would only be as helpful as correlating the size and shape of an object with a corresponding grasp in the physical space, which requires live collection of data.

In terms of RGB video, the same issue as the depth map persists, where the accuracy of correlating grasps with objects in the video - even with post-processing - wouldn't yield a high accuracy, where the same issue as DeGol's experiment would persist.

In other words, the machine learning pipeline of this project would come from correlating the capture of a human hand positioning to one of the few predefined grasps and identifying grasps that work best of the general shape and size through a diffusion policy.

In summary, if limited to using existing data to train models, we would use existing image databank and manually assigning one of six grasps and running a CNN or transformers with a layer of diffusion for highest correlation of grasps.