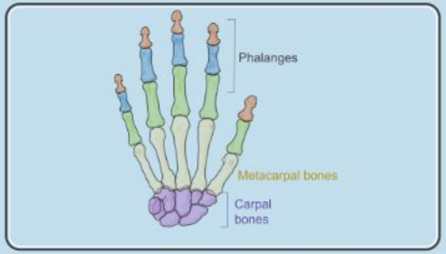
Devi's SSG Journey Talk

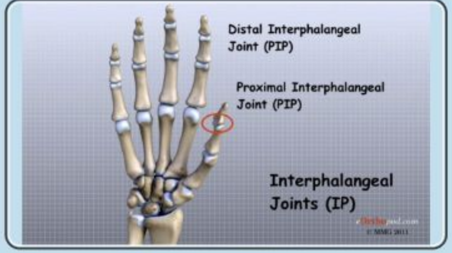
Exploring Computer Vision for Robotic Manipulation

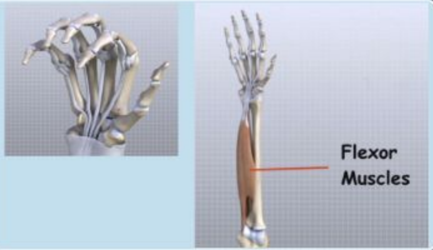
Past Research

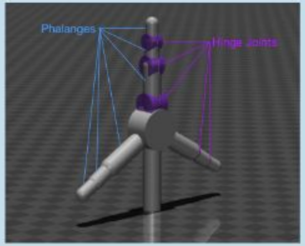
Human hands are able to pick up and move a wide variety of objects

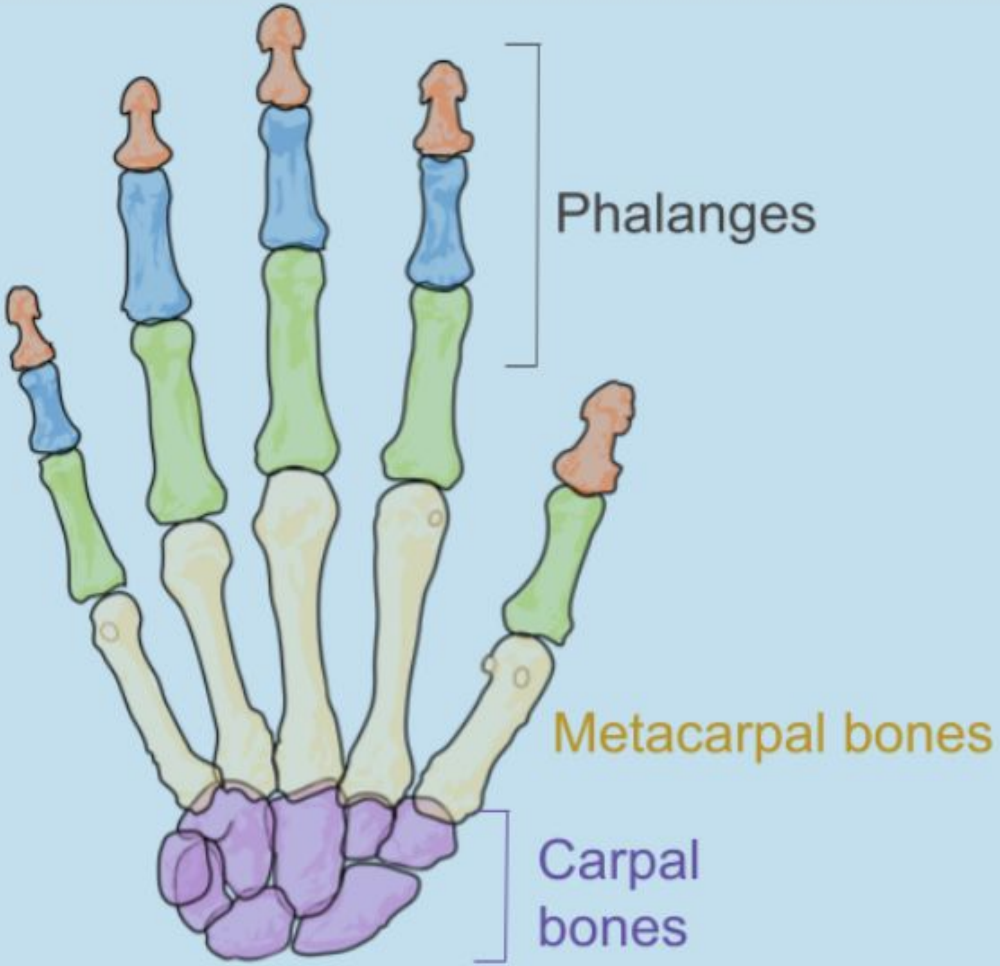
How can I design a Robotic Hand that is simpler but still effective?

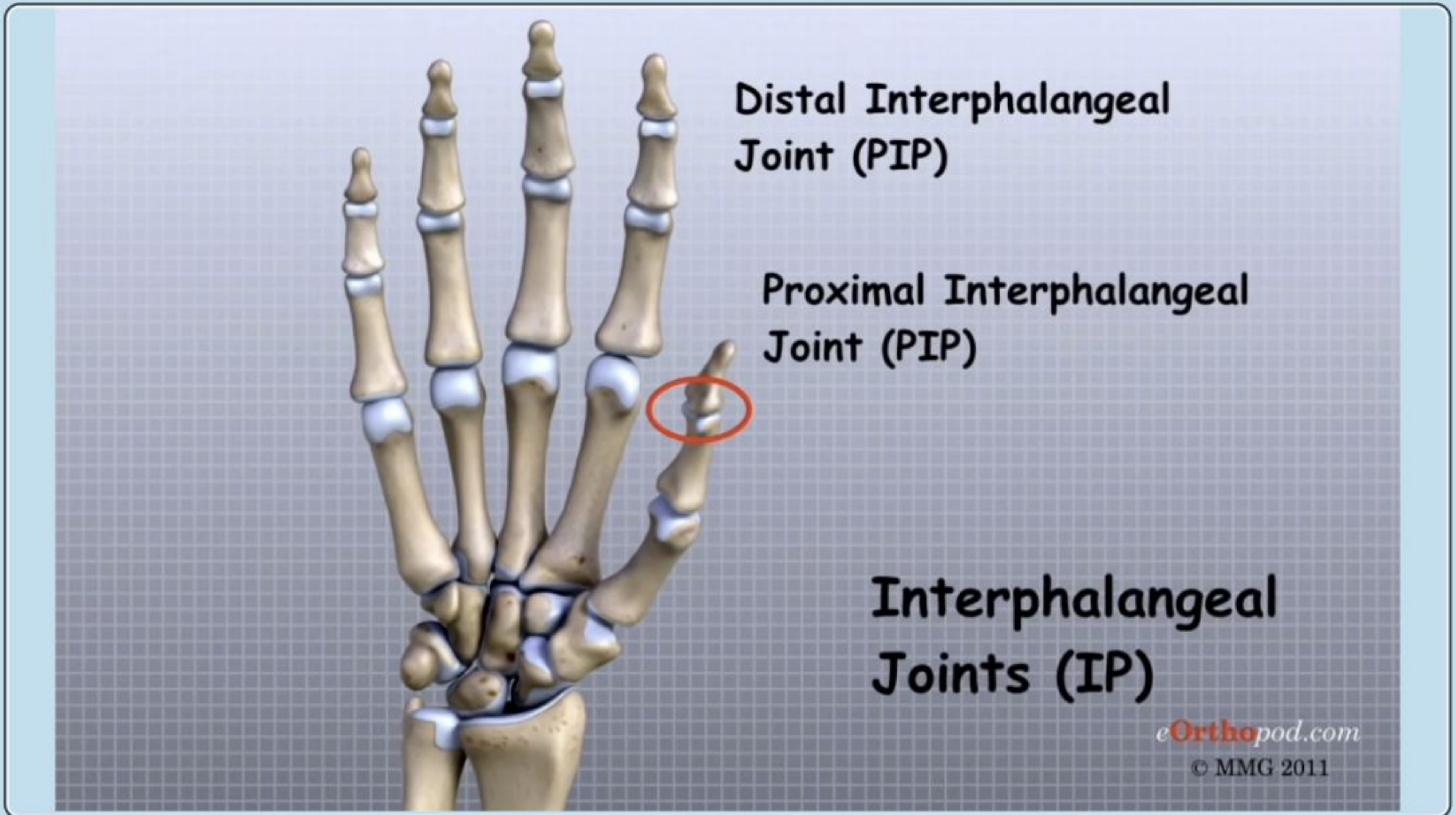


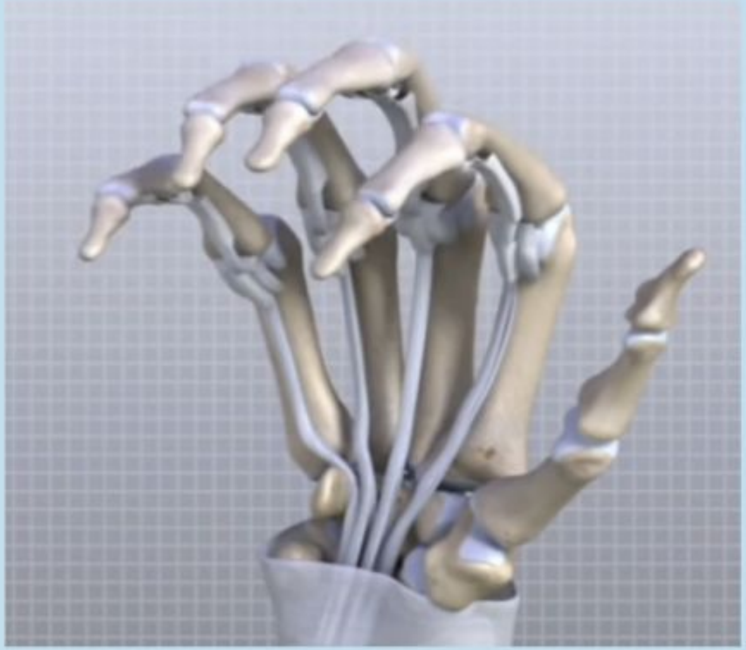


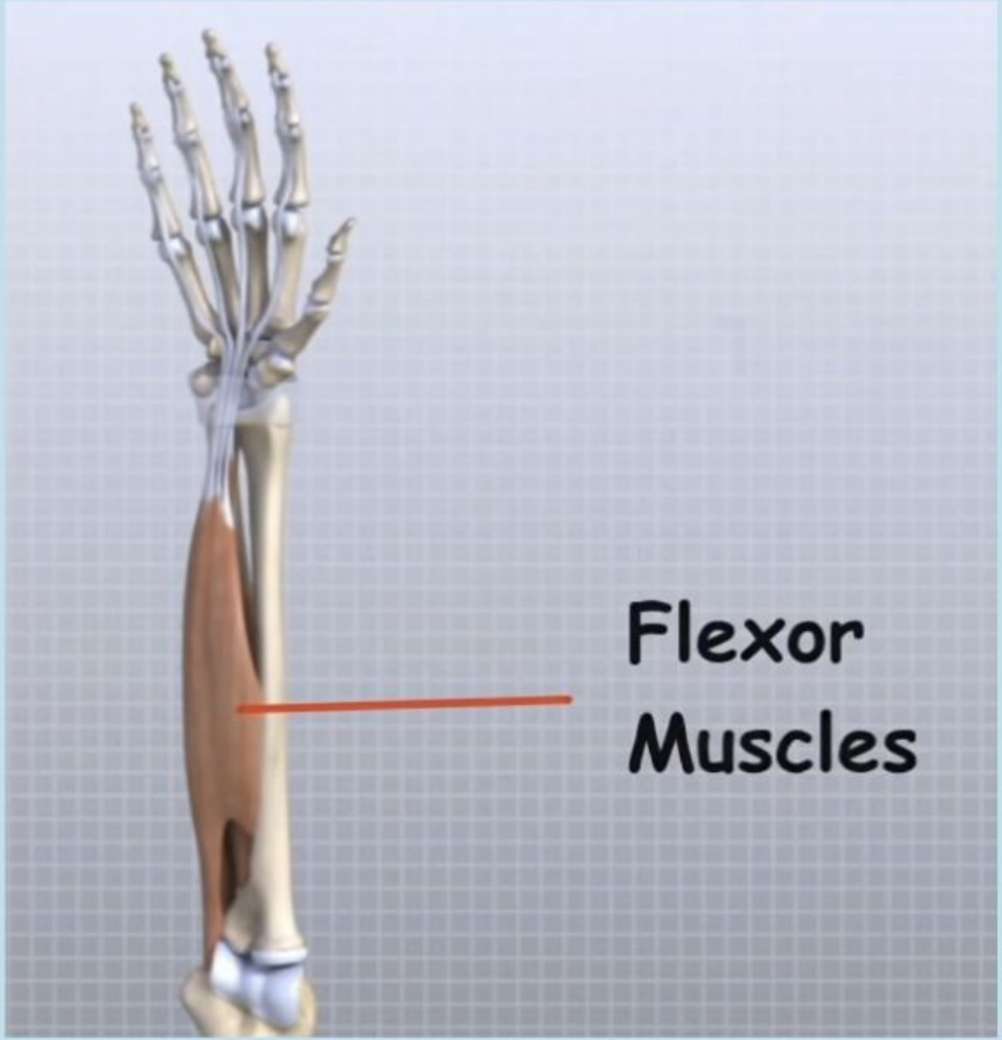


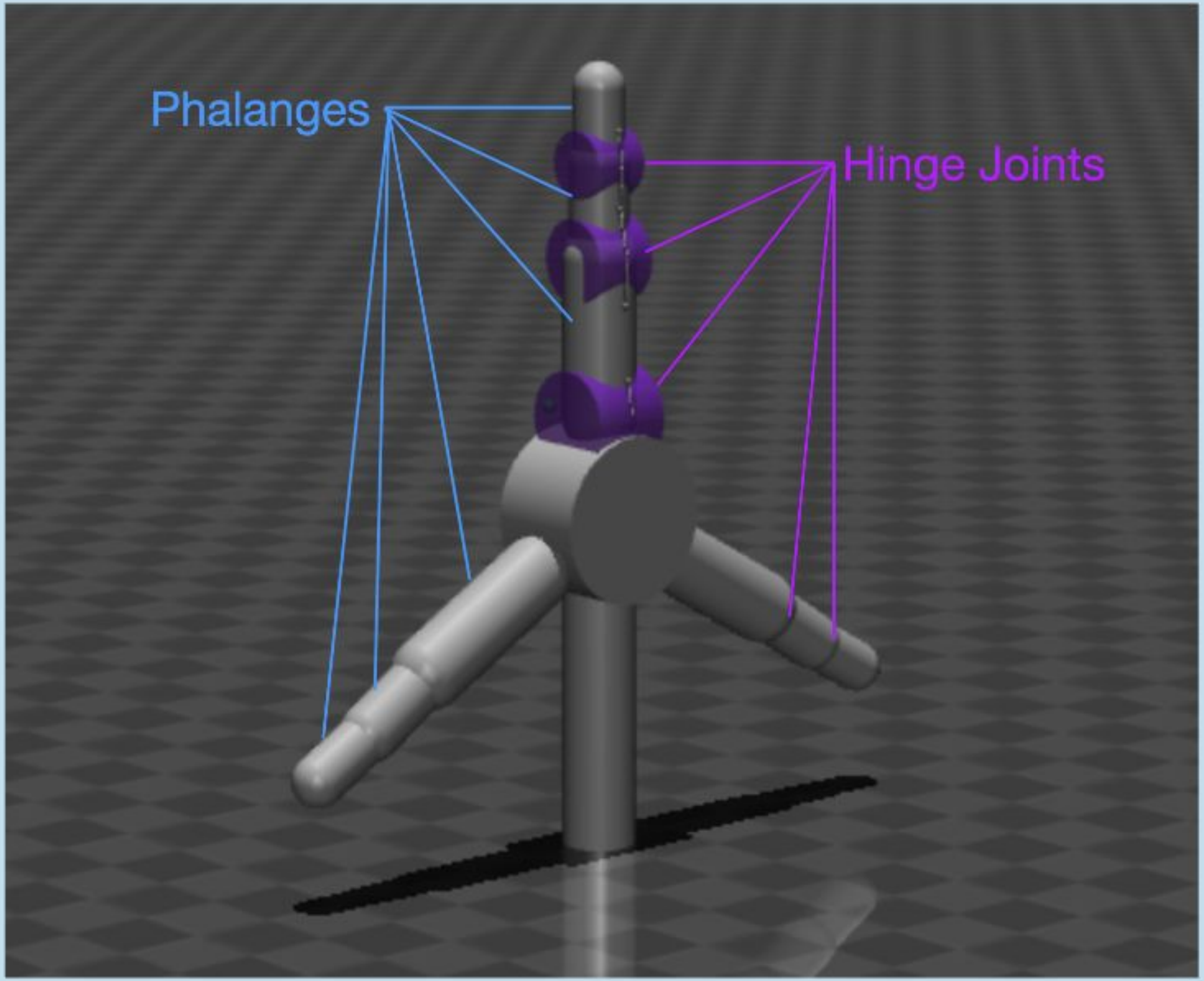



Phalanges
Metacarpal bones
Carpal
bones

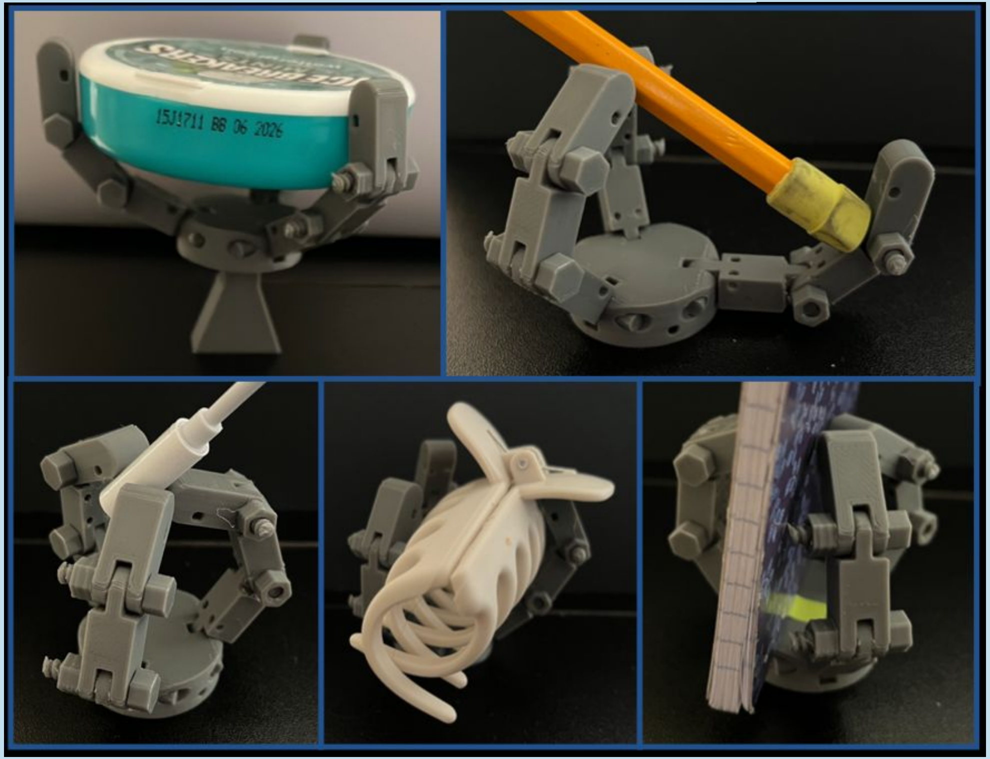

Distal Interphalangeal
Joint (PIP)
Proximal Interphalangeal
Joint (PIP)
Interphalangeal
Joints (IP)
eOrthopod.com
© MMG 2011

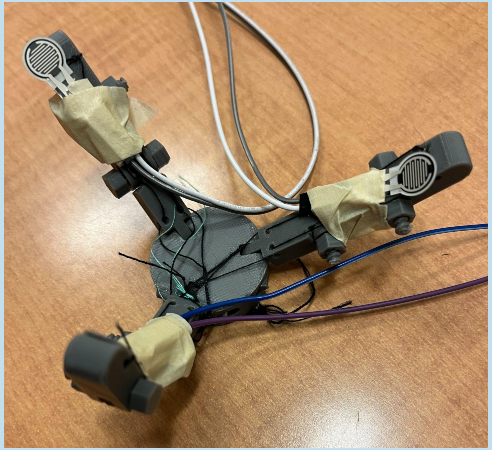


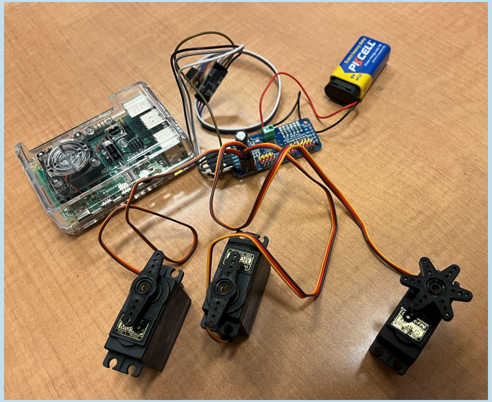

Flexor
Muscles


Phalanges
Hinge Joints

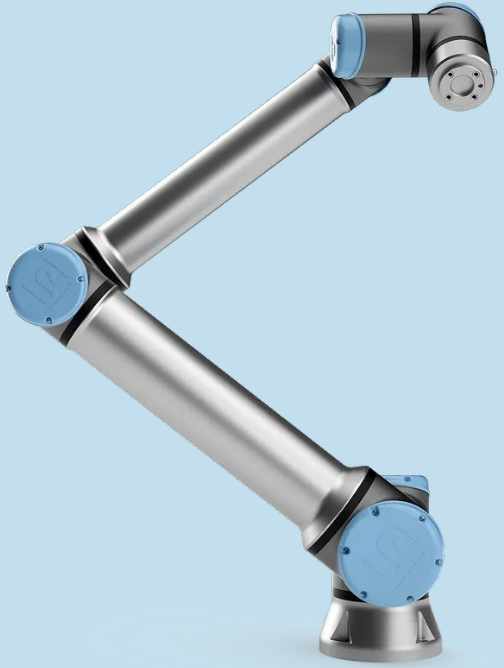
Hand Design

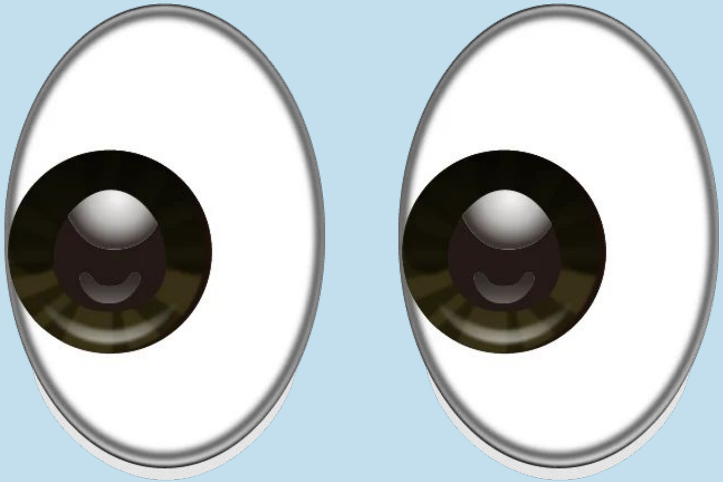






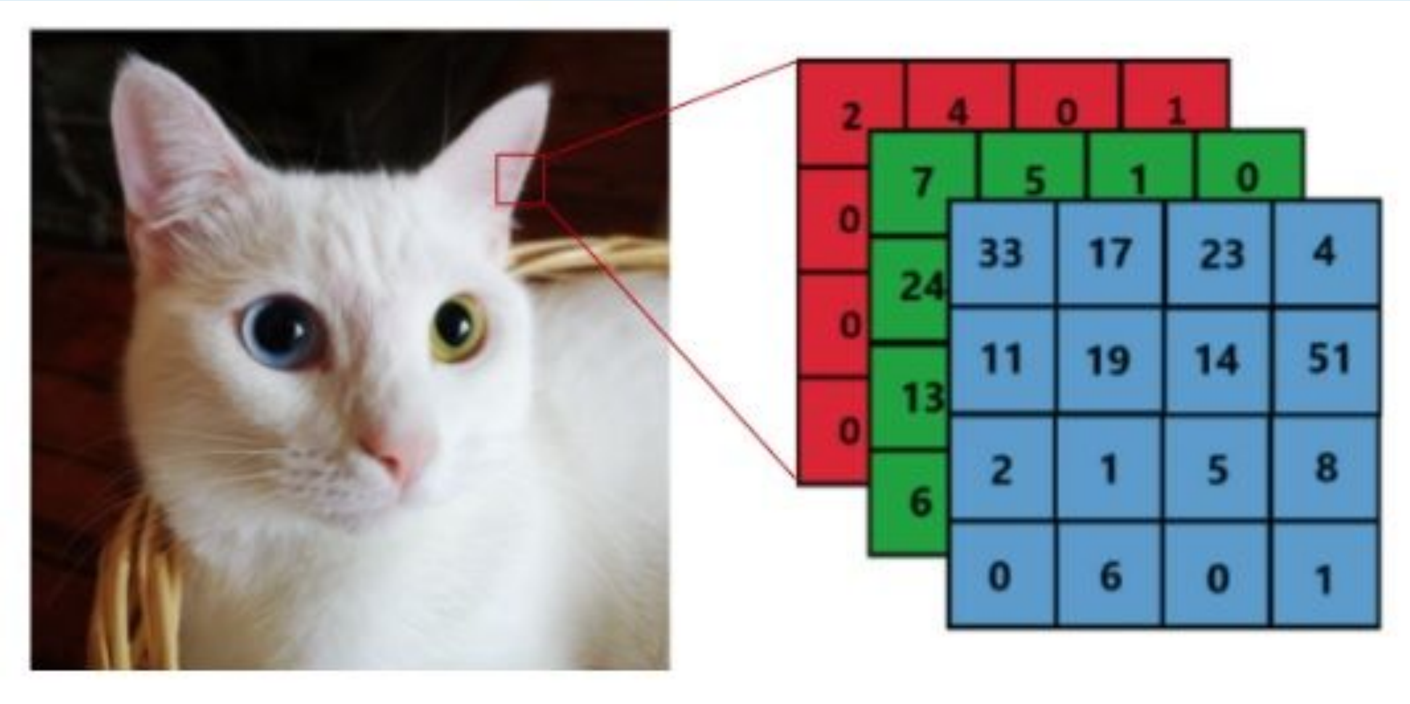
How to use the Robotic Hand?





https://emojiisland.com/products/eyes-emoji-icon

Background Information



https://www.midokura.com/unveiling-the-world-of-computer-vision-a-comprehensive-overview/

Background Information

https://indiaai.gov.in/article/image-segmentation-the-deep-learning-approach

Image Segmentation

Object Detection

CURIOSITY QUESTION

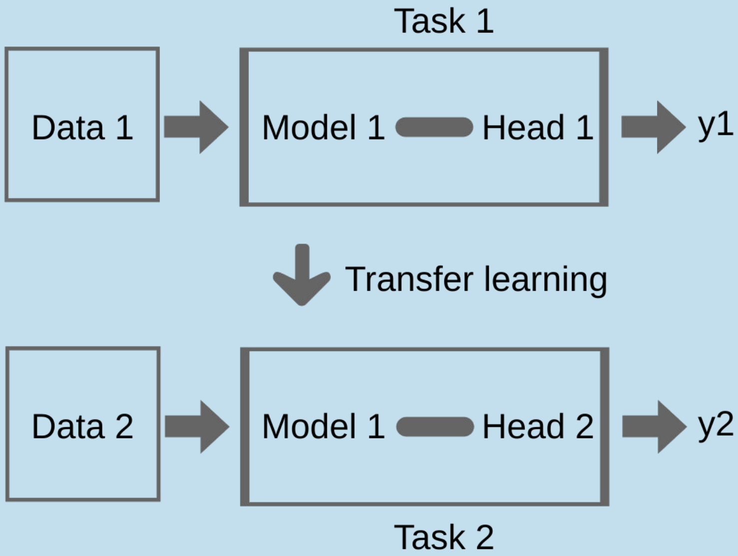
How to train a computer vision model to quickly and reliably determine the 3 dimensional position and shape of an object from visual data?

HYPOTHESIS

Using images from different points of view along with transfer learning from models trained on pose estimation tasks can help reliably determine the 3D position and shape of an object.

HYPOTHESIS

https://unblast.com/free-mug-on-table-mockup-psd/



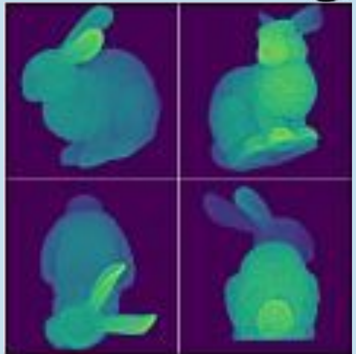
https://en.wikipedia.org/wiki/Transfer_learning

IMAGINED EXPERIMENT

Train 3 different types of models to reconstruct 3D position and shape of household objects from images

Single-image

Multi-image



Transfer Learning

https://viso.ai/computer-vision/point-cloud-processing/

IMAGINED EXPERIMENT

3D position accuracy

Euclidean distance error in mm between predicted and ground truth object center.

3D shape similarity

Overlap in terms of surface area of actual and predicted object shapes.

IMAGINED EXPERIMENT

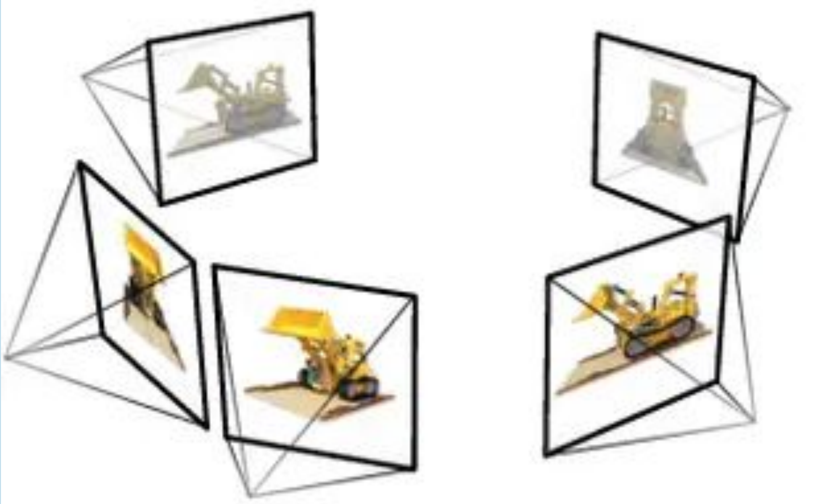
Positive Control

Location that is accurately measured and 3D pose captured by LiDAR.

Negative Control

Model trained on mismatched labels. Should fail to learn meaningful 3D representations.

Neural Radiance Fields (NeRFs)



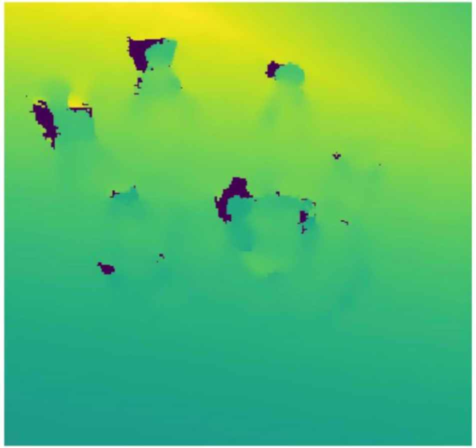
Images + accurate camera poses

3D scene representation

Dex-NeRF for Transparent Objects

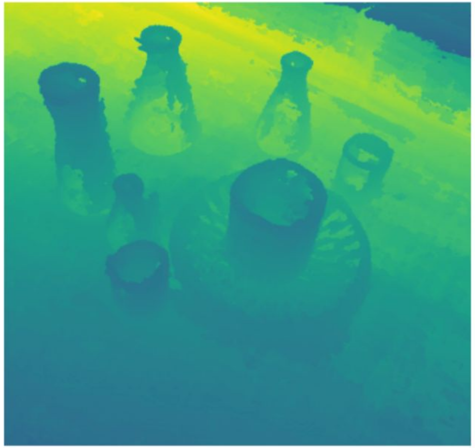
Real-world Scene

Real-world scenes in labs, homes, workplaces and more, have transparent objects that existing depth sense have difficulty isolating.



RealSense D410 Depth Image

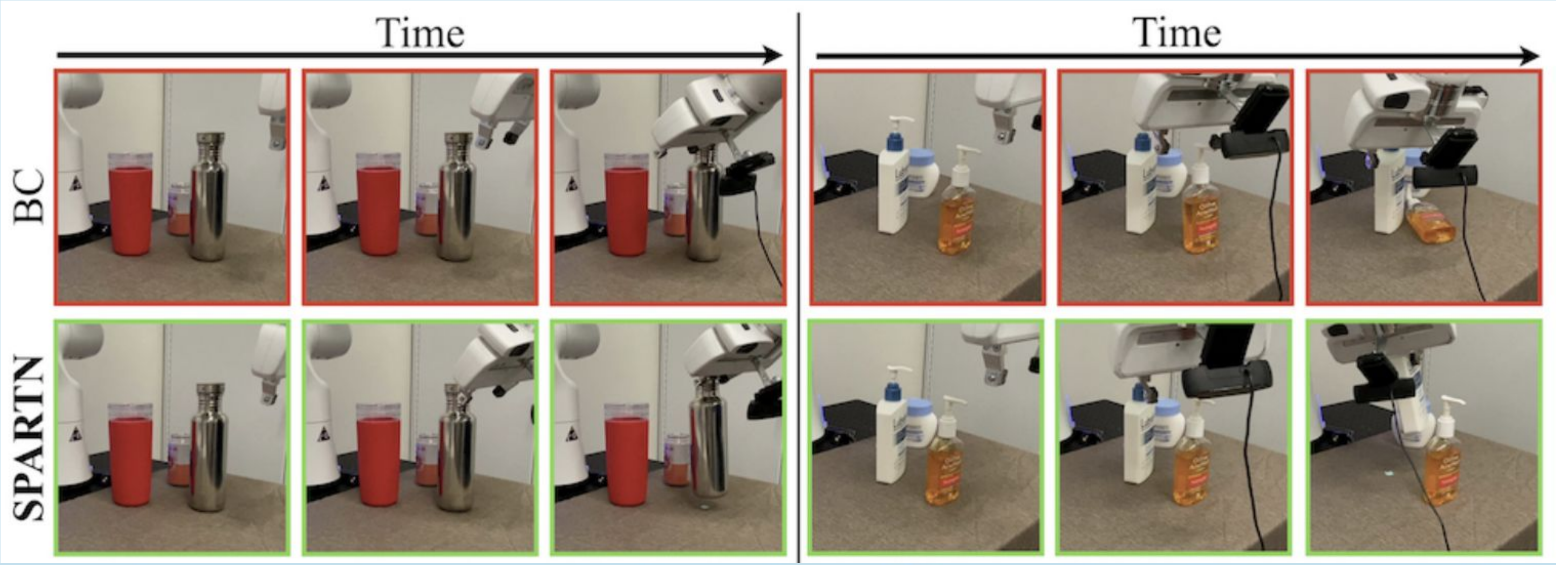
Here, the depth output from the Intel RealSense D410 camera has missing objects and pixels, and smudged and hard-to-distinguish outlines.



Depth Map (Ours)

Using NeRF, on the same scene, we're able to recover a depth map with all objects and pixels, crisp object outlines.

SPARTN for Robustness



On 8 complex grasping scenarios (e.g., shiny objects, clutter), it improved grasp success by 22.5% on average.

H-InDex Pre-trained Model

Pretraining

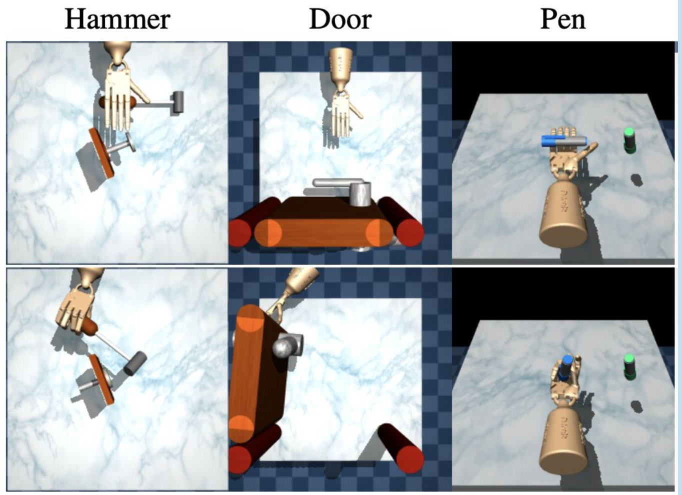
Recognize 3D human hand poses.

Adapting

Fine-tunes using a self-supervised method. Learns to find key points like corners or edges without needing labeled data.

Reinforcement Learning

Train through RL with knowledge from original hand-pose model.



CONCLUSION

Multi-view inputs enable detailed 3D reconstruction from RGB images.

Transfer learning from human hand pose estimation leads to better performance in robotic manipulation tasks.

Approaches like **SPARTN** and **H-InDex** demonstrate that offline augmentation and pretrained representations boost learning efficiency and robustness.

FURTHER RESEARCH

Evaluate RGB vs. depth-sensing cameras

Mount both a depth camera and Raspberry Pi camera

Test vision methods on a real robot

Explore ways to accelerate the process


Thank you