Get a Grip: Multimodal Visual and Simulated Tendon Activations for Grounded Semantics of Hand-related Descriptions

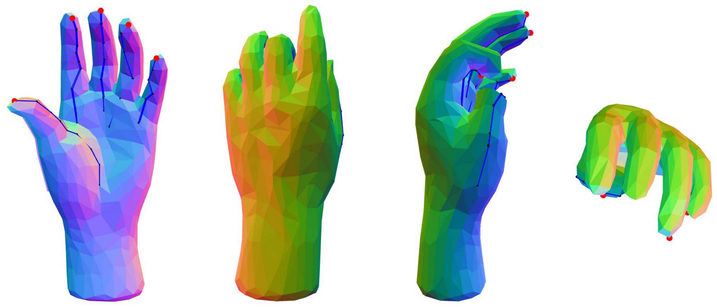
Daniele Moro
danielemoro@u.boisestate.edu
Dr. Casey Kennington

Can a machine learn to understand the meaning of hand poses?

- Semantic meaning of words is grounded in perceptual modalities (Harnad, 1990)
- Meaning of hand poses is grounded through visual features and embodied tendon activations

Generated 972 Simulated Hands

- *Forward Simulation Model* (Bern et al., 2017)
- 5 tendons that contract the fingers (0.0 to 1.0)
- 243 hand configurations and 4 perspectives



Data is collected and aggregated

- Tendon data: how much each finger is activated
- Perspective data
- Visual data from pre-trained CNN VGG19
- Descriptions from Amazon Mechanical Turk

Words as Classifiers learns fit between hand and description

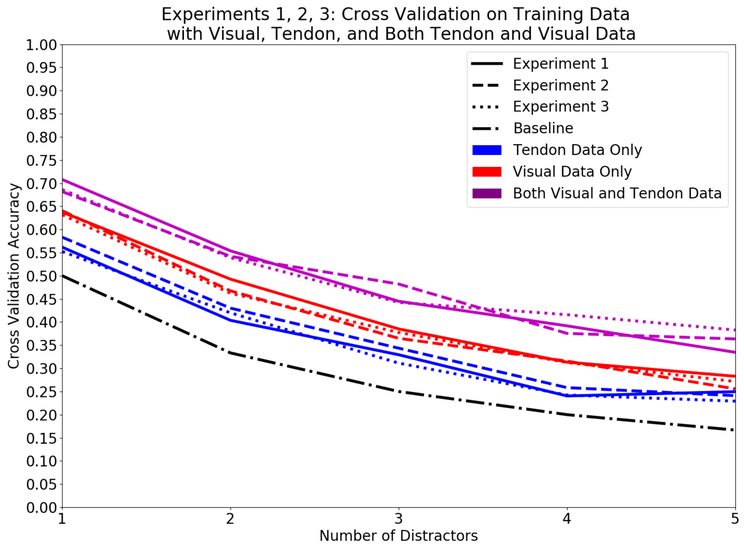
- Learns mapping between features from modalities and corresponding hand descriptions
- Builds classifier for every word in vocabulary
- Leverages logistic regression classifier

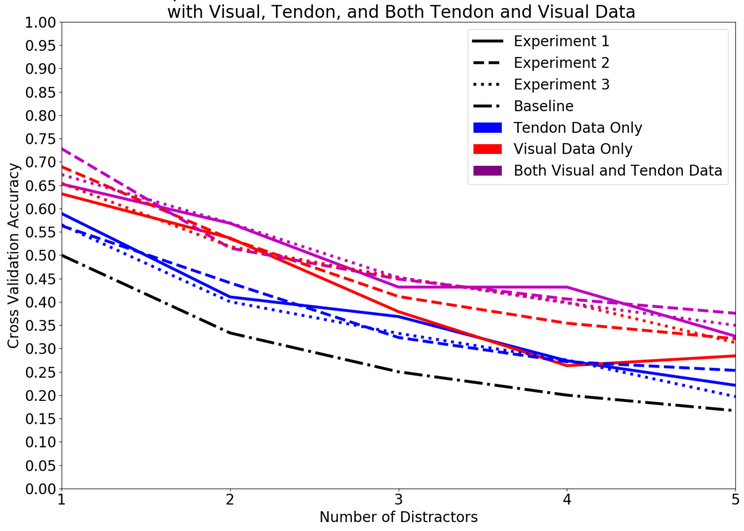
Experiments to Test Model

1. Image Retrieval: Given a description, model must retrieve correct image among distractors

2. Description Retrieval: Given an image, model must retrieve correct description among distractors

3. Mirror Neurons: Model generates tendon data from visual data





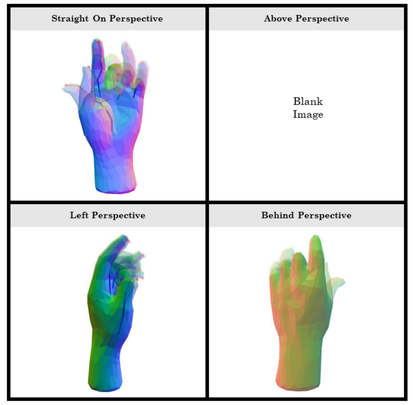
Task is Challenging

- Perspective of hand image plays a large role
- Distinct versus long descriptions challenge model
- Many distractors are similar

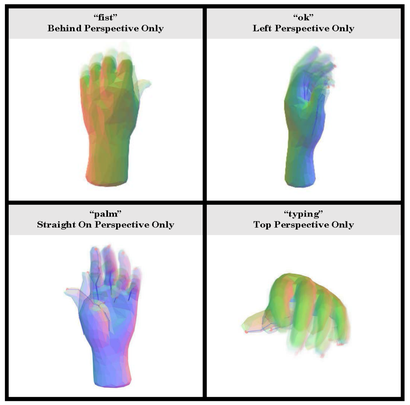
Analysis: Model Generates Hands

- Calculate the fit of each image to a certain word
- Blend top 100 images into the figures below

Interpretation of *pointing*



Interpretation of many words



Conclusion

- Our model can to some extent understand the **meaning** of **hand poses.**
- **Combined** modalities **increase** the **accuracy** of the model in recognizing semantics of hand poses
- **Mirror neurons** are good performers
- Future work: more data, make WAC more complex

References

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1-3):335– 346.

James M Bern, Grace Kumagai, and Stelian Coros. 2017. Fabrication, Modeling, and Control of Plush Robots. In *Proceedings of the International Conference on Intelligent Robots and Systems*.

Casey Kennington and David Schlangen. 2015. Simple Learning and Compositional Application of Perceptually Grounded Word Meanings for Incremental Reference Resolution. *In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 292–301, Beijing, China. Association for Computational Linguistics.

SLIM SPEECH, LANGUAGE & INTERACTIVE MACHINES

BOISE STATE UNIVERSITY