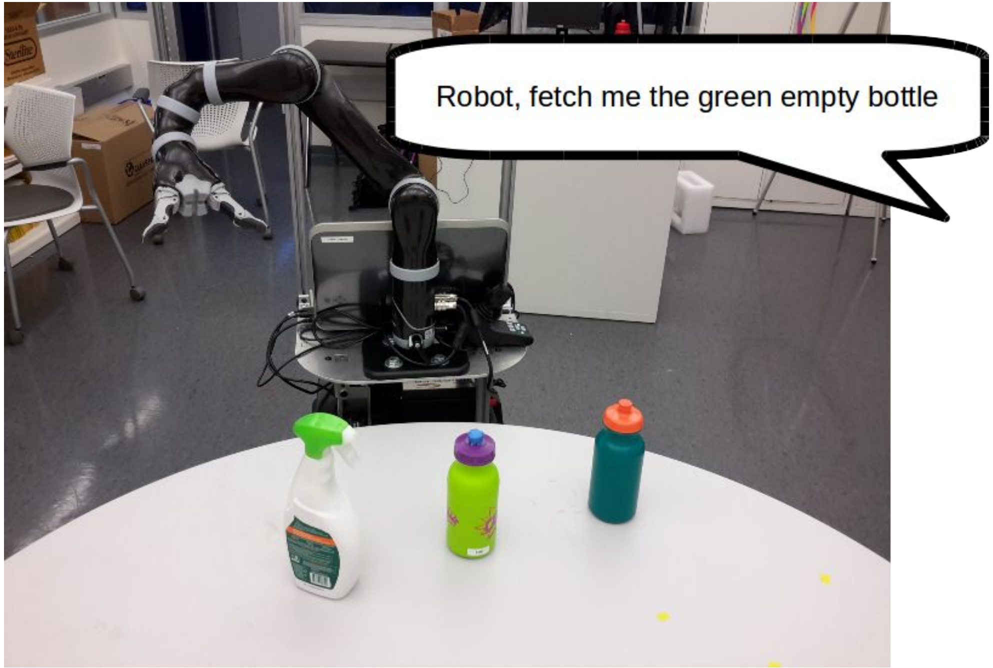
Grounding Language in Exploratory Behaviors and Multi-Modal Perception

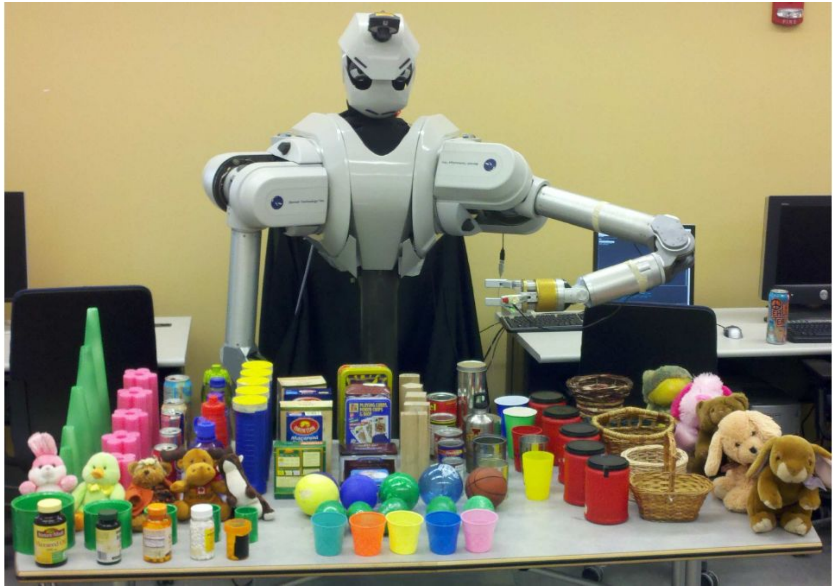
Zach Osman, Jivko Sinapov

Motivation

- Grounded Language Learning
- Multimodal vs Vision only

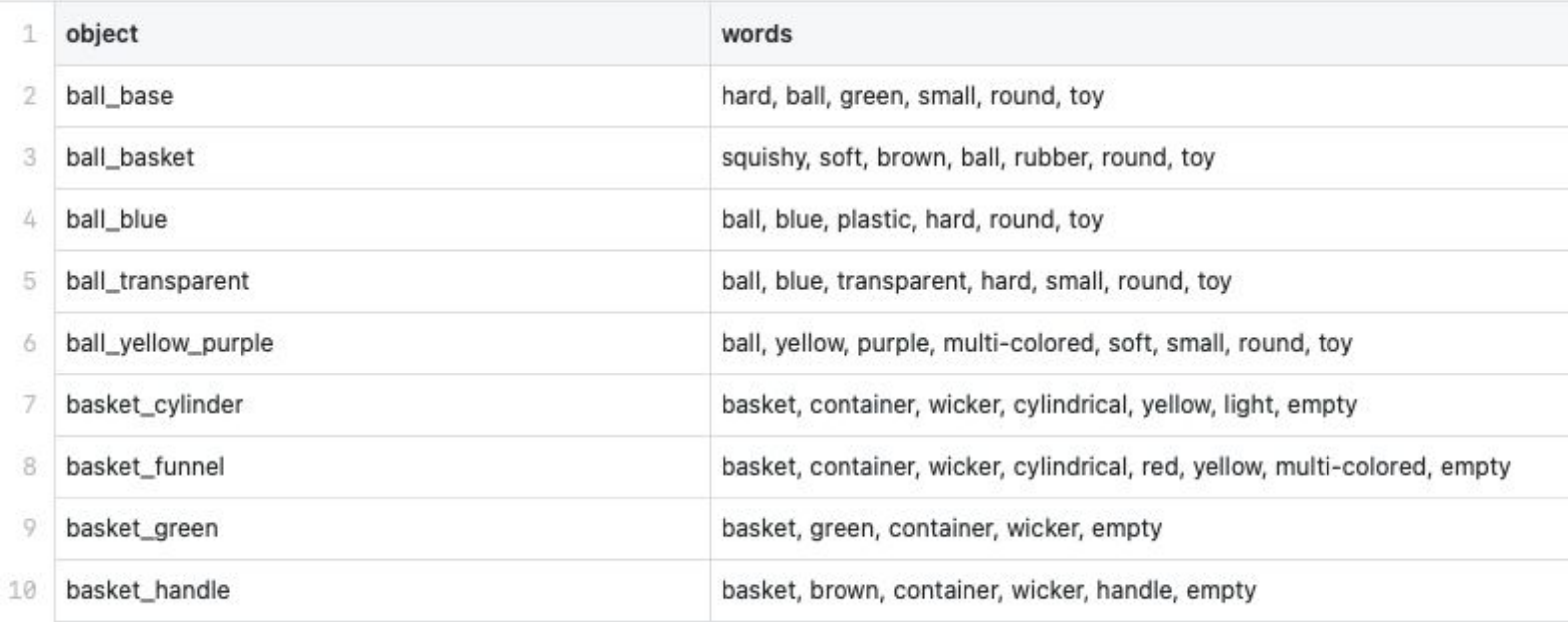


Objects



Tatiya, G., and Sinapov, J. (2019) **Deep Multi-Sensory Object Category Recognition Using Interactive Behavioral Exploration** In proceedings of the IEEE International Conference on Robotics and Automation (ICRA)

Data - Object-Word Labels



object	words
ball_base	hard, ball, green, small, round, toy
ball_basket	squishy, soft, brown, ball, rubber, round, toy
ball_blue	ball, blue, plastic, hard, round, toy
ball_transparent	ball, blue, transparent, hard, small, round, toy
ball_yellow_purple	ball, yellow, purple, multi-colored, soft, small, round, toy
basket_cylinder	basket, container, wicker, cylindrical, yellow, light, empty
basket_funnel	basket, container, wicker, cylindrical, red, yellow, multi-colored, empty
basket_green	basket, green, container, wicker, empty
basket_handle	basket, brown, container, wicker, handle, empty

Data - Sensorimotor Features

- 48 different behavior modality combinations
- Modalities used: audio, vibration, flow, haptics, SURF, finger position (only for grasp), color (only for look)

grasp

lift

hold

shake

drop

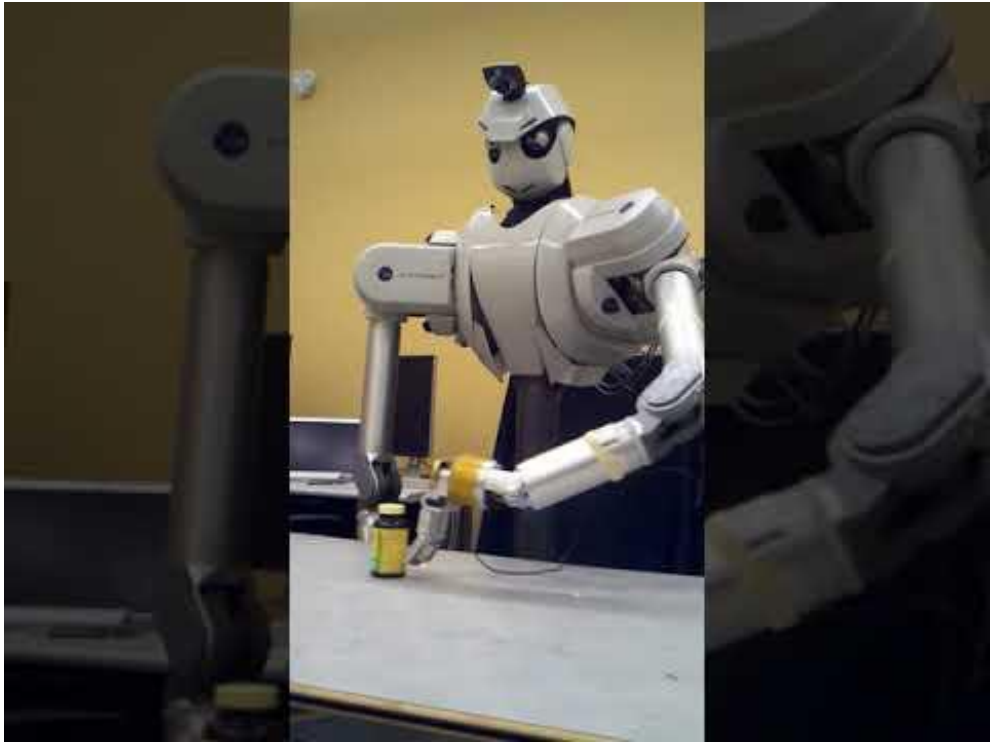
tap

poke

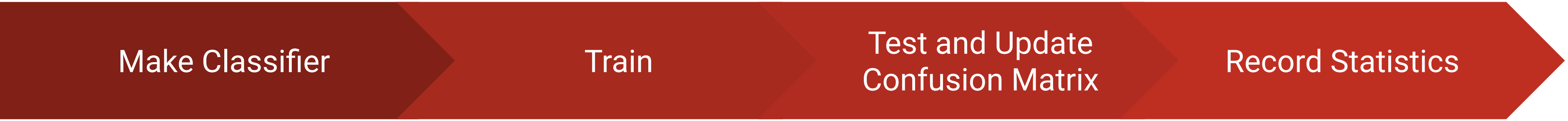
push

press

Object Exploration by Robot



Train Classifiers for each Word-Sensorimotor Combination



Make Classifier

Train

Test and Update Confusion Matrix

Record Statistics

For the word ‘hard’, the behavior ‘tap’, and the modality ‘audio’, make classifier ‘hard_tap_audio_0’ for the first train-test split

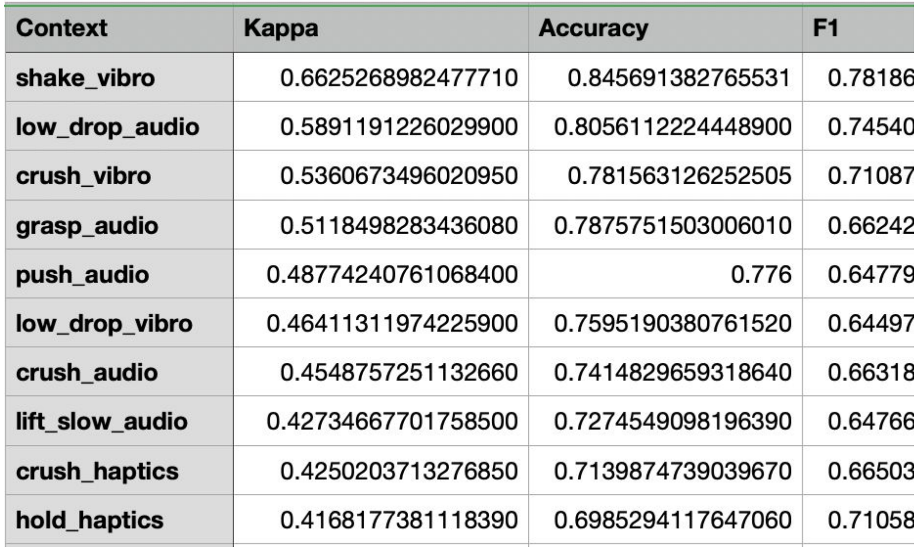
Train classifier using data from ‘tap_audio’ sensorimotor data file

Test classifier on test set of objects using ‘tap_audio’ data and update confusion matrix accordingly

Record Kappa, Accuracy, Recall, Precision, and F1 from confusion matrix

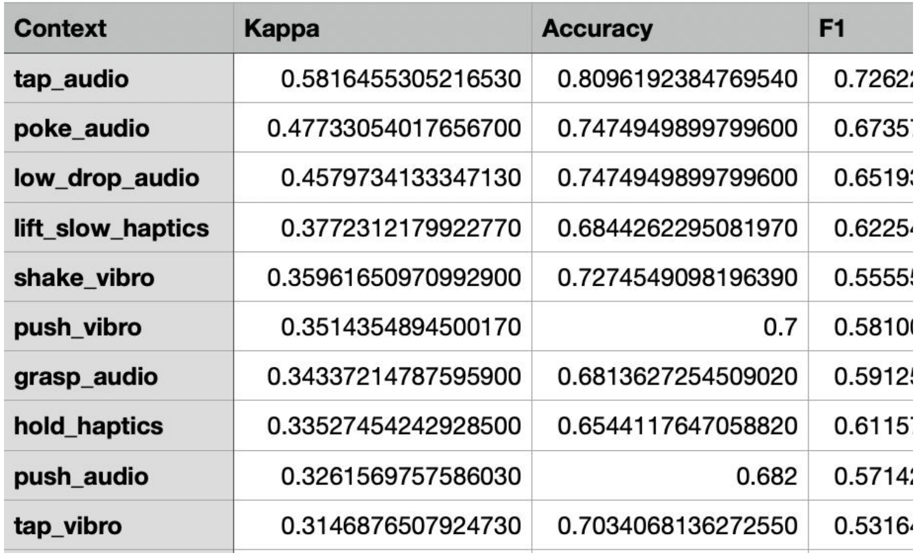
Results

empty_stats



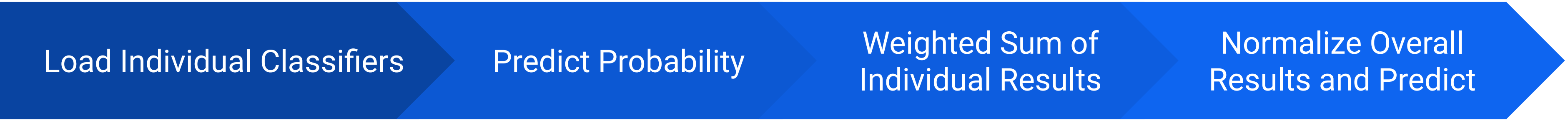
Context	Kappa	Accuracy	F1
shake_vibro	0.6625268982477710	0.845691382765531	0.78186
low_drop_audio	0.5891191226029900	0.8056112224448900	0.74540
crush_vibro	0.5360673496020950	0.781563126252505	0.71087
grasp_audio	0.5118498283436080	0.7875751503006010	0.66242
push_audio	0.48774240761068400	0.776	0.64779
low_drop_vibro	0.46411311974225900	0.7595190380761520	0.64497
crush_audio	0.4548757251132660	0.7414829659318640	0.66318
lift_slow_audio	0.42734667701758500	0.7274549098196390	0.64766
crush_haptics	0.4250203713276850	0.7139874739039670	0.66503
hold_haptics	0.4168177381118390	0.6985294117647060	0.71058

hard_stats



Context	Kappa	Accuracy	F1
tap_audio	0.5816455305216530	0.8096192384769540	0.7262
poke_audio	0.47733054017656700	0.7474949899799600	0.6735
low_drop_audio	0.4579734133347130	0.7474949899799600	0.6519
lift_slow_haptics	0.3772312179922770	0.6844262295081970	0.6225
shake_vibro	0.35961650970992900	0.7274549098196390	0.5555
push_vibro	0.3514354894500170	0.7	0.5810
grasp_audio	0.34337214787595900	0.6813627254509020	0.5912
hold_haptics	0.33527454242928500	0.6544117647058820	0.6115
push_audio	0.3261569757586030	0.682	0.5714
tap_vibro	0.3146876507924730	0.7034068136272550	0.5316

Combining all Sensorimotor Features



Load Individual Classifiers

Predict Probability

Weighted Sum of Individual Results

Normalize Overall Results and Predict

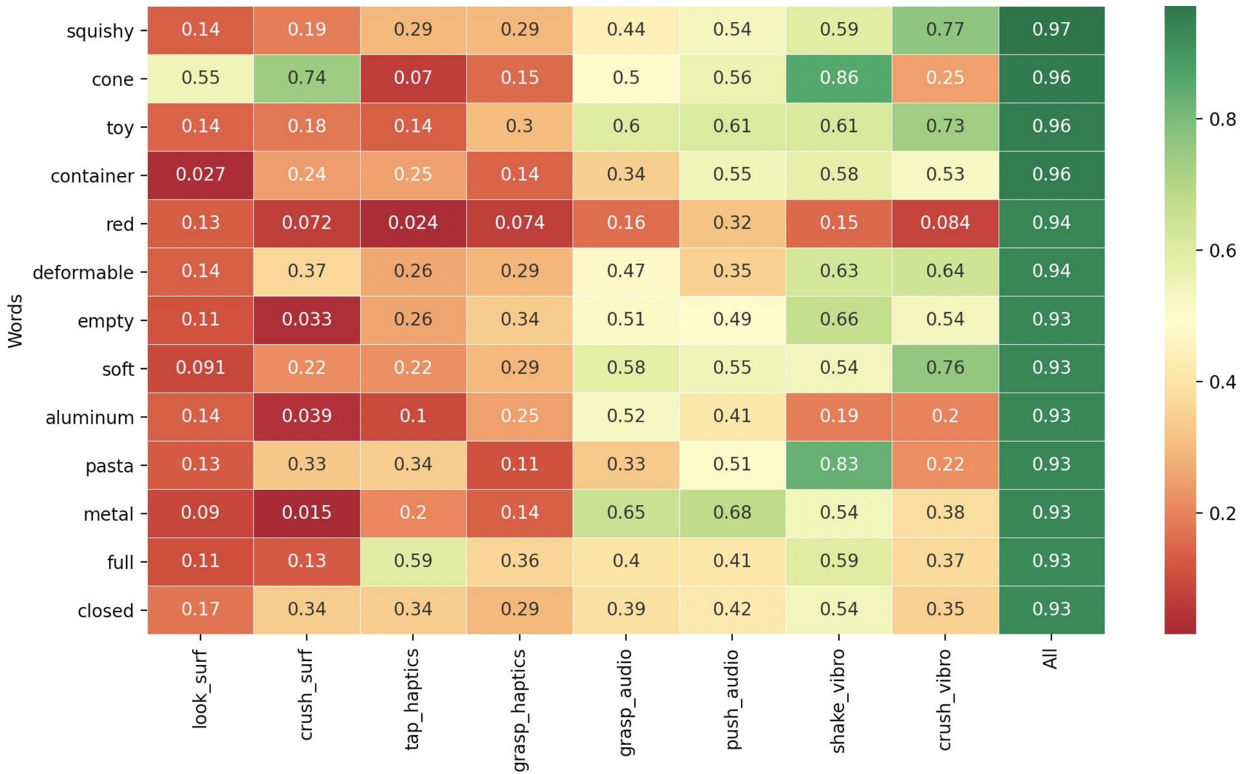
Load all classifiers for a particular word, such as ‘hard’

For each individual classifier, predict probability of word applying to an instance of an object with that context’s data

Calculate weighted sum based on Kappa values for all individual classifiers to get overall class distribution probabilities

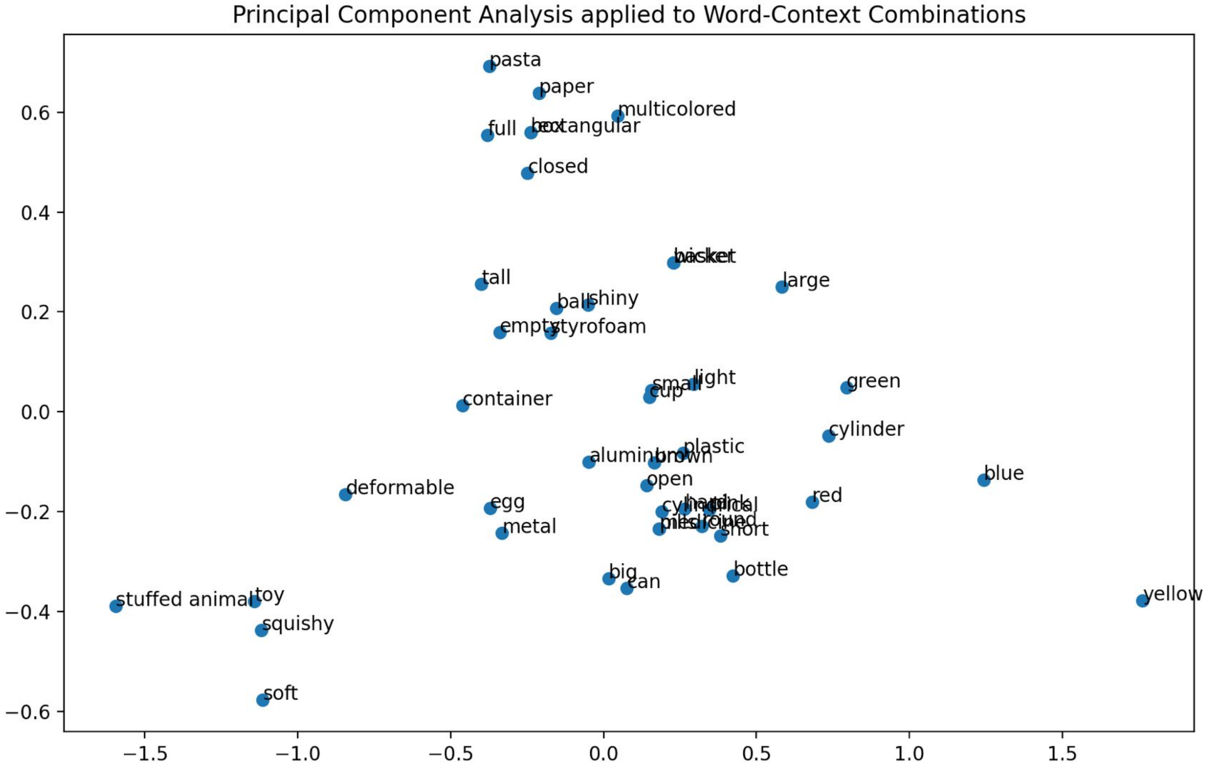
Normalize results and make a prediction if the word applies to the object or not and compare to the ground truth

Results - Heat Map

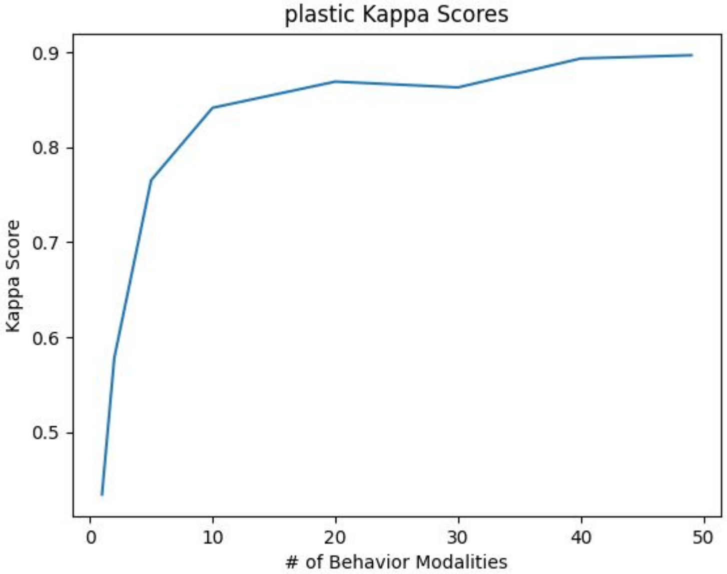


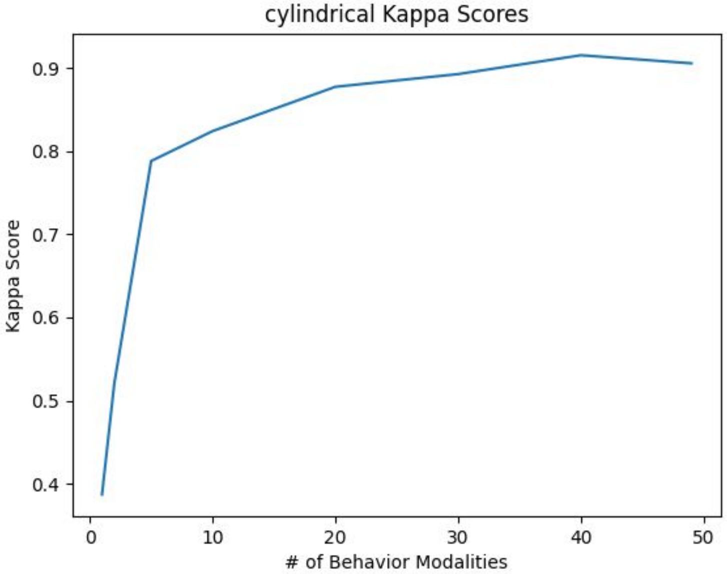
Words	look_surf	crush_surf	tap_haptics	grasp_haptics	grasp_audio	push_audio	shake_vibro	crush_vibro	All
squishy	0.14	0.19	0.29	0.29	0.44	0.54	0.59	0.77	0.97
cone	0.55	0.74	0.07	0.15	0.5	0.56	0.86	0.25	0.96
toy	0.14	0.18	0.14	0.3	0.6	0.61	0.61	0.73	0.96
container	0.027	0.24	0.25	0.14	0.34	0.55	0.58	0.53	0.96
red	0.13	0.072	0.024	0.074	0.16	0.32	0.15	0.084	0.94
deformable	0.14	0.37	0.26	0.29	0.47	0.35	0.63	0.64	0.94
empty	0.11	0.033	0.26	0.34	0.51	0.49	0.66	0.54	0.93
soft	0.091	0.22	0.22	0.29	0.58	0.55	0.54	0.76	0.93
aluminum	0.14	0.039	0.1	0.25	0.52	0.41	0.19	0.2	0.93
pasta	0.13	0.33	0.34	0.11	0.33	0.51	0.83	0.22	0.93
metal	0.09	0.015	0.2	0.14	0.65	0.68	0.54	0.38	0.93
full	0.11	0.13	0.59	0.36	0.4	0.41	0.59	0.37	0.93
closed	0.17	0.34	0.34	0.29	0.39	0.42	0.54	0.35	0.93

Results - PCA


Principal Component Analysis applied to Word-Context Combinations
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
pasta
paper
multicolored
full
box
rectangular
closed
basket
large
tall
ball
shiny
empty
styrofoam
small
light
cup
green
container
cylinder
plastic
brown
blue
open
deformable
egg
cylindrical
hard
pink
red
metal
round
short
big
can
bottle
stuffed animal
toy
yellow
squishy
soft

Results - Multi-Context Kappas





Challenges

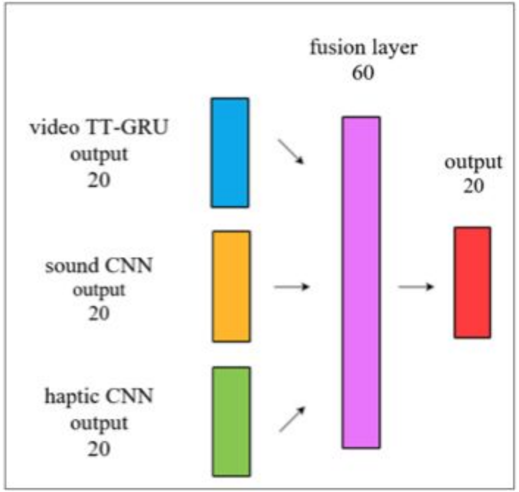
- Problem: Unbalanced Data Set
- Solution: upsampling

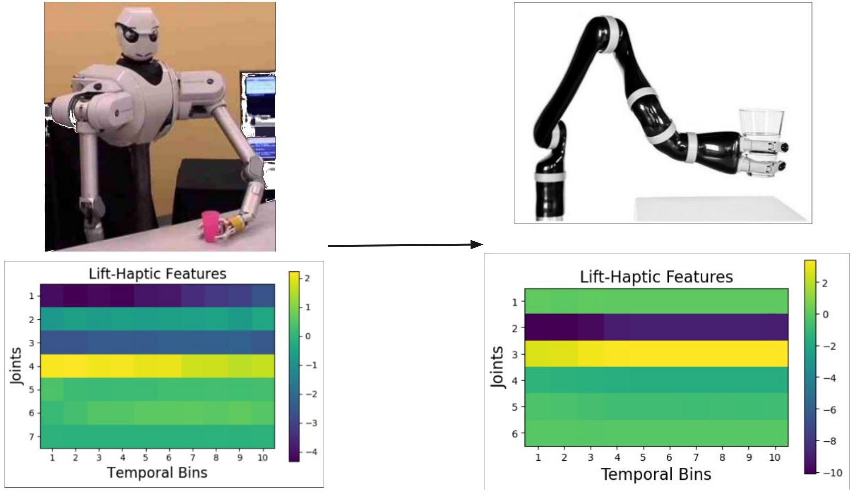
Conclusion

- Multimodal approach allows us to better identify properties of an object that can not be picked up through visual data alone
- Using a combination of behavior-modalities produces more accurate results

Future Work

- Deep Neural Network
- Knowledge Transfer between robots (different joints, different sensors)





Acknowledgements

- Professor Sinapov

T-TRIPODS

DIAMONDS

NSF

Thank You!

Any questions?