Toward Contextual Grounding of Unfamiliar Gestures for Human-Robot Interaction

Wil Thomason and Ross Knepper
Department of Computer Science, Cornell University, Ithaca, USA

I. RESEARCH VISION

Interaction between humans and robots is inherently limited by understanding. Robots must understand all aspects of human communication to collaborate with humans naturally. While recent work has advanced the state of the art in language understanding, there remains a way to go. This gap is particularly prominent in *non-verbal communication*. When speech fails to adequately communicate a concept (*e.g.* due to noise, ambiguity, *etc.*), humans rely on other channels of communication. Of these channels, we are most interested in gestures — specifically, in *unfamiliar gestures*.

An unfamiliar gesture is a gesture not before seen by an observer. Humans use unfamiliar gestures in conversation when we improvise a gesture for a concept. As no gesture recognition system can train on the entire set of possible gestures a human could use in discourse, many more gestures are unfamiliar to robots. We hypothesize that we can understand unfamiliar gestures by leveraging situated gestures and speech in training.

II. PRIOR WORK

We recently presented an initial approach to understanding unfamiliar gestures [3] that utilizes *zero-shot learning*: training a system to understand examples from a class of inputs that were unseen at training time. By observing that humans often use gestures to "support" or emphasize the meaning of their speech, we can use examples of situated gesture and speech to train a multi-modal neural network for this task. We start with a pre-trained word embedding [2] and use pairs of gestures and words describing the same concept to train a network to compute an embedding function for gestures aligned with the word embedding space. We primarily rely on data from [1] augmented with verbal data and compute a novel feature on the gesture input. When training is complete, we input a gesture and get out the bag of words most closely related to the gesture in the embedding space. We then use a novel heuristic for contextual salience to filter down this bag of words, resulting in a set of contextually relevant descriptors for the unfamiliar gesture.

III. CURRENT WORK

Although we have achieved some success [3], we were limited by the scale and suitability of the available data to our task. To mitigate this limitation, we have conducted a data collection experiment to build our own large scale, high quality dataset of situated gestures and speech.

We elicit gestures by asking participants to use a set of instructions to teach the study operator how to fold a piece of origami. These instructions are designed to be frustrating to convey using only speech. In a pilot study, we found that participants made substantial use of simultaneous gesture and speech to give the instructions. Thus, by recording the speech and skeleton movements of the participants, we can record gestures and speech that have the same meaning.

We have conducted about 35 trials of 20 minutes each of this experiment. We are currently cleaning the data, which we will release in both raw and gesture-segmented form.

IV. FUTURE WORK

We hope that our new dataset will allow us to improve the model we designed [3]. In particular, we believe that adding depth to our embedding network will allow us to more accurately learn the desired embedding[1]. Additionally, we wish to investigate sequence learning models to better capture gestures which gain new meaning when combined.

We also wish to improve the precision of our process. We currently "ground" gestures to a set of possible descriptors. To make our approach more useful to robot applications, we need to be able to ground gestures to specific, contextually relevant meanings. We have considered grounding gestures directly to actions — inferring from an unfamiliar gesture what the human wants the robot to do. This approach is interesting in domains such as collaborative construction, but lacks generality. We believe that we may be able to come closer to our goal by constructing a new kind of embedding space mapping gestures and words to grounded meanings. This space will have an additional dimension representing time to allow the grounded meaning of a gesture to vary as the context changes.

REFERENCES

[1] I. Guyon, V. Athitsos, P. Jangyodsuk, and H. J. Escalante. The ChaLearn gesture dataset (CGD 2011). 25(8):1929–1951.
[2] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space.
[3] W. Thomason and R. A. Knepper. *Recognizing Unfamiliar Gestures for Human-Robot Interaction Through Zero-Shot Learning*, pages 841–852. Springer International Publishing, Cham, 2017.

This material is based upon research supported by the Office of Naval Research under Award Number N00014-16-1-2080. We are grateful for this support. We will be able to cover any additional travel costs not covered by the award.

[1] We restricted the size of the original network to avoid overfitting.