Recognizing Unfamiliar Gestures for Human-Robot Interaction through Zero-Shot Learning

Wil Thomason
Department of Computer Science
Cornell University
wbthomason@cs.cornell.edu

Ross Knepper
Department of Computer Science
Cornell University
rak@cs.cornell.edu

***Abstract*—Human communication is highly multimodal, including speech, gesture, gaze, facial expressions, and body language. Robots serving as human teammates must act on such multimodal communicative inputs from humans, even when the message may not be clear from any one modality alone. In this paper, we explore a method for achieving increased understanding of complex, situated communications by leveraging coordinated natural language, gesture, and context. These three problems have largely been treated separately, but unified consideration of them can yield gains in comprehension [11, 1]. We develop and present a novel model for incorporating context and coincident speech to better understand a wider class of gestures. We show preliminary results demonstrating the capability of our system to provide reasonable descriptions of gestural examples from classes both known and unknown to the system, and conclude with a discussion of the future directions of the project.**

I. Introduction

Human communication is highly multimodal, including speech, gesture, gaze, facial expressions, and body language. Robots serving as human teammates must act on such multimodal communicative inputs from humans, even when the message may not be clear from any one modality alone. In this paper, we explore a method for achieving increased understanding of complex, situated communications by leveraging coordinated natural language, gesture, and context. These three problems have largely been treated separately, but unified consideration of them can yield gains in comprehension [11, 1].

Gesture recognition has been an area of investigation from the early days of computer vision, but modern gesture recognition systems remain fragile. Most approaches focus on speed and accuracy of recognition, but they remain restricted to a fixed gestural lexicon [6, 22, 16, 2, 7, 12] and cannot recognize gestures outside of a small pre-trained set with any accuracy [13, 2].

Our work departs from this traditional model in that the set of gestures it can recognize is not limited to the gestural lexicon used for its training. Even in simplified domains, naive classifiers can fail to recognize instances of trained gestures due to human gestural variability. Humans resort to gesture when speech is insufficient, such as due to inability to recall a word, inability to be heard, or inadequate time to formulate speech. For these reasons, gesture is prevalent in human discourse. Yet gestures defy attempts at canonical classification both due to variations within and among individuals and due to their subjective interpretations. We define the **unfamiliar gesture understanding problem**: given an observation of a previously unseen gesture (i.e. a gesture of a class not present in any training data given to the system), we seek to output a contextually reasonable description in natural language of the gesture's intended meaning.

This problem is an instance of the machine learning problem of zero-shot learning, a burgeoning area of machine learning that seeks to classify data without having seen examples of its class in the training stage. Most prior work in the area [17, 9, 15] makes use of a multimodal dataset to perform the zero-shot task. However, the zero-shot task has not yet been demonstrated for gestural data. In the related one-shot learning task, gesture understanding has been shown from only one example of a given class in the training stage [20, 19, 18]. The primary drawback of such approaches is their reliance on a fixed lexicon of gestures. We remove this drawback by creating a novel multimodal embedding space using techniques from convolutional neural nets to handle variable length gestures and allow for the description of arbitrary unfamiliar gestural data.

The ChaLearn 2013 multi-modal gesture recognition challenge explored techniques for increasing the robustness of understanding by combining gesture and text [5]. However, the entries still only recognize a small fixed set of gestures.

Other work in situated multimodal understanding systems has been limited to combining simple diectic (pointing) gestures with speech, to differentiate among a small set of referent objects [3]. These pointing gestures represent a small and relatively simple subset of human gestures. In this paper, we contribute a novel approach to understanding unfamiliar language, gesture, and context in order to be able to understand diverse and varied gestures.

II. Approach

Two key insights of our approach to derive meaning from unfamiliar gestures are to recognize physical similarities among gestures by commonalities in their constituent "sub-gestures," and to leverage redundant information contained in simultaneous, situated speech and gesture. We begin with some intuition for these two insights.

First, whereas gestures with similar high-level physical form do not always have similar meanings, many gestures

with related meanings share common "sub-gestural" motion components. For instance, pushing and pointing gestures both involve an outward motion, indicating a semantically-related position away from the gesturer.

Second, a common mode of gestural use in conversation is to add redundancy to spoken information to increase the chance of the speaker's meaning being correctly inferred. For example, when giving instructions, a speaker may make gestures that represent physically the actions their words describe. By sampling coincident speech and gesture in a variety of contexts, we can therefore construct from experience an approximate partial map between the meanings of the two modes of communication.

These two insights combined allow us to understand unfamiliar gestures. First, we can exploit the structural similarity of gestures with related meanings to map an unfamiliar gesture to a location in an embedding space of gestures that reflects its relation to other gestures we have previously seen. We can then use this placement and the partial map between gestures and speech that we have established during training to determine a reasonable meaning for the unfamiliar gesture.

A. Details

Our approach is built around a multi-stage pipeline which takes individual gestures formatted as RGB-D data as its input and outputs a natural-language description of the gesture. The stages of the pipeline are as follows, in order:

1) Gesture Embedding: The first step of our approach is to create an embedding space mapping gestures to the corresponding words. For a gesture g encoded as a series of RGB-D frames, we first compute a discrete wavelet transform vector $\vec{\psi_g}$ of g, by creating windows of 120 ms, each overlapping by 20 ms, and taking the discrete wavelet transform of each window.

We then use $\vec{\psi_g}$ as the input to a neural network composed of two 1-D convolutional layers separated by a max pooling layer to allow for variable-length inputs, and followed by three fully-connected layers. We assume that there exists a bag of words $W = \{\vec{w_1}, \ldots, \vec{w_k}\}$ associated with each g, where each $\vec{w_i}$ is encoded as a vector in a pre-trained word embedding (in particular, we use Word2Vec [14]). At training time this is given; in practical usage we aim to recover this bag of words. As such, we train the network to minimize the following loss function, where f is the function computed by the network:

$$\mathcal{L}(\vec{\psi_g}, W) = \left\| \frac{\sum_{\vec{w_i} \in W} \vec{w_i}}{k} - f(\vec{\psi_g}) \right\| \quad (1)$$

This loss function is simply the norm of the difference between the centroid in the pretrained word embedding space of the words corresponding to g and where in this space f places g. In other words, we learn a mapping which places gestures closest to those words most strongly associated with them.

In usage, we compute $f(\vec{\psi_g})$ and examine its k nearest neighbors in the word embedding space to approximate the set of words most strongly associated with g.

2) Salience Heuristic: Although the above multimodal embedding produces a set of candidate words to describe a gesture, it does not take into account any notion of dynamic context, i.e. context from specific, recent interactions. We propose a simple salience heuristic to filter down the set of possible descriptor words as the final stage in our pipeline. This heuristic, which is inspired by Eldon et al. [3], imposes an ordering on the candidate descriptors by computing a variant on the common tf-idf metric [10] for each. This variant is a direct analogue of tf-idf for the gestural context, and computes:

$$\mathcal{S}(w) = \left(1 + \log\left(\sum_{i=1}^{m} \tfrac{1}{i} \mathcal{I}_w(\mathcal{O}_i)\right)\right) \cdot \left(\log\left(1 + \frac{N}{\sum_{i=1}^{N} \mathcal{I}_w(\mathcal{C}_i)}\right)\right) \quad (2)$$

where the $\mathcal{O}_i$ are the m most recent bags of words recorded by the system (in the order of recording), the $\mathcal{C}_i$ are bags of words associated with known (training) gestures, $\mathcal{I}_w(x)$ is an indicator function that is 1 if word w is present in bag of words x, and 0 otherwise, and N is the total number of known gestures. This heuristic therefore favors words which have recently been relevant to gestures used in the current conversation (i.e. favoring topic continuity) while avoiding words which are relevant to a large number of gestures and are therefore unlikely to be very specific descriptors of a given gesture. If the embedding in Section II-A1 returns k possible descriptors, the top $\ell < k$ according to their ranking by $\mathcal{S}$ are chosen for the final output of the system.

III. Results

We have conducted preliminary experiments assessing the performance of both the zero-shot learning model and the salience heuristic.

A. Zero-Shot Model

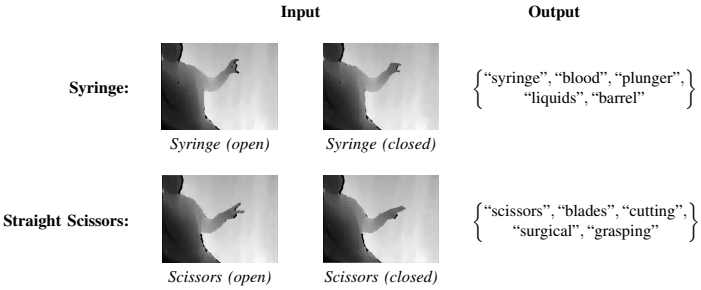

Fig. 1. The output of our zero-shot learning system for both known (syringe) and unknown (straight scissors) classes of gesture.

We trained our zero-shot model on a subset of the data from Guyon et al. [8] consisting of surgical hand signals. We refer to this subset as a "meta-class", as it contains several classes of gestures and is thus a class of classes. As these data did not include the language accompanying the gestures, we created a set of plausible accompanying words for each gesture, constructed by randomly sampling salient words from a textual description of the object or action indicated by each class of gesture.

Figure 1 shows an example of our results using this subset. We withheld all examples of the "straight scissors" class from the training process. After training, we evaluated the performance of the model at generating reasonable descriptions

for gestures from both the known and unknown classes. As shown in Figure 1, we are able to successfully generate sets of words describing each gesture, regardless of whether or not the gesture's class was present in the training data.

Given that the goal of our system is to produce a "reasonable" (as judged by humans) description for an unfamiliar gesture, we believe that the best means of evaluation for our system is direct human assessment. However, these are preliminary results, and we have not yet conducted a human evaluation study (we discuss plans for such a study in Section IV). The results in the below table show two means of automatic assessment. First, we compute the average Hausdorff distance between the bag of words output by our system for an instance of the held-out gesture class and a bag of words generated by a human for the same gesture. The properties of the Hausdorff metric ensure that outputs are penalized both for including irrelevant words and excluding relevant words. This method provides insight into the performance of our system, but is ultimately insufficient. We measure the distance between words as distance between their corresponding word vectors in the word2vec embedding space. While this space has been shown to possess certain algebraic properties and place some similar words close together in the embedding space (e.g. Mikolov et al. [14]), it does not necessarily position words in such a way that their relative distance is meaningful. The second automatic means of assessment that we present in the below table is qualitative; we list a subset of the words output for the held-out gesture, for human judgment.

Meta-Class (Held-out Class)	**Avg. Dist.**	**Salient Descriptors**
Surgical (Scissors)	0.2828	"cutting", "pliers", "blades"

Although we only show results for a single held-out class here, we note that holding out several classes from the training process produced lower-quality results. Given that our training dataset was very small (approx. 100 gesture examples, total), we attribute this drop in performance to this change causing the system to have insufficient training data. We propose a data collection experiment to remedy this problem in Section IV-B.

B. Salience Heuristic

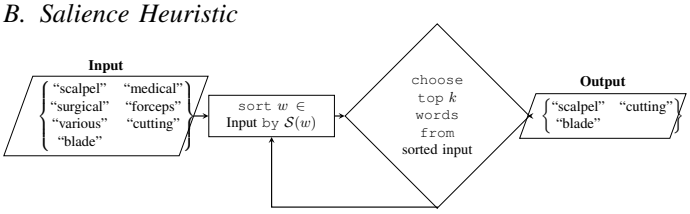

Fig. 2. The output of our salience heuristic on an example "conversation".

To test the performance of our salience heuristic, we constructed a set of "conversations" composed of a sequence of simulated past outputs of our system and a simulated output of our zero-shot model (as the next element in the sequence). We then applied our salience heuristic to these data, and qualitatively assessed the results in terms of the salience of the words selected. We show an example of these results in Figure 2. The result shown is for a shortened conversational sequence due to space constraints; we assessed the system on longer sequences. As shown, we succeed in selecting descriptors which are more recently relevant and more relevant to the conversation overall.

These preliminary results establish the viability of our approach. We are able to generate a set of reasonable descriptors for unfamiliar gestures without losing the capability to do so for gestures in training classes. Further, we are able to remove contextually irrelevant words from the generated set of descriptors to improve the overall accuracy of the final set of descriptors. This set is useful for understanding the meaning of gestures.

IV. Planned Future Work

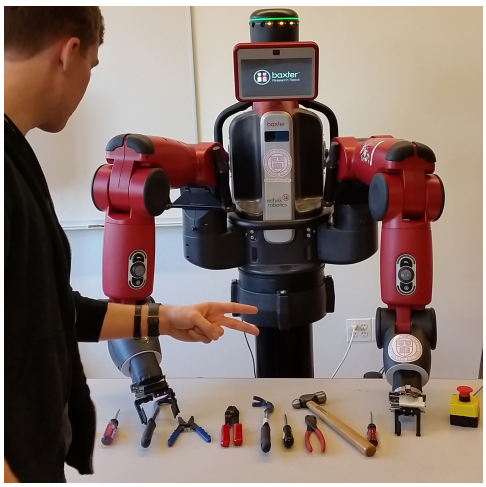

Fig. 3. An example experimental scenario. If the user makes the dynamic gesture shown and issues the request "Can you hand it to me?", our system should be able to disambiguate both the referent object and desired action.

As the above results are preliminary, we have several experiments scheduled to be completed, as follows:

A. End-to-End Object Identification

The most important experiment for our system is a full-scale end-to-end verification of its functionalities. We intend to do so by training our proposed zero-shot model on more subsets of the ChaLearn 2011 dataset and running the entire system (as detailed in Section II), on a Rethink Robotics Baxter robot (pictured in Figure 3). We will be able to capture the empirical performance of our system in a realistic scenario by using Baxter to perform an object identification task. The human user will indicate to Baxter through various blends of speech, speech and gesture, and independent gesture the object which they wish to obtain (e.g. with an ambiguous phrase such as "the red one" and an accompanying gesture to indicate that, of the available red objects, they mean a hammer). We will assess Baxter's performance at identifying the correct object both in the presence and absence of gesture to better quantify

the contribution of our system's abilities. As we are aware of no direct baselines (i.e. no other systems capable of performing zero-shot learning on gestures), we will compare our system to the current state of the art in gesture recognition and natural language understanding (e.g. [11, 21, 23, 22]), trained on the same data as we use to train our system.

This experiment should provide useful data on the viability of our system for use in real-world robotic applications. Further, by tweaking the parameters of our model during separate repetitions of this experiment, we can gain greater insight into the performance characteristics and areas of strength or weakness of our system. For example, by removing the salience heuristic, we can determine how well the embedding performs on its own. By changing the dataset used for training, we can establish the generality of our approach. Finally, by substituting the features used to describe gestures in the input to our embedding space and retraining the embedding, we can further tailor our selected features to provide optimal performance.

As mentioned in Section III, the best means of assessing our zero-shot model is direct human rating of the quality of its output. Although participants in the real-robot study will be asked to provide ratings of the quality of Baxter's overall performance, we believe that it is important to evaluate each component of the system in isolation. We will use a large-scale Amazon Mechanical Turk study to obtain ratings of the quality of the output of our zero-shot model in terms of its applicability to the relevant gestural input and its comprehensibility. We will seek ratings of the quality of our model's outputs both in isolation and relative to the output of several baselines: state of the art gesture recognition systems (e.g. [21, 23, 22]) and randomly selected bags of words drawn from the overall corpus of speech associated with each gesture.

B. Multimodal Corpus Collection

A dearth of multimodal data limits the development of algorithms for situated gesture and language understanding. Guyon et al. [8] and Escalera et al. [4] have provided a good starting point, but we see possible improvement in areas such as the artificial nature of the gestures contained (i.e., the performers were instructed to gesture) and the dataset's focus on beat and emblematic gestures. We plan to complete an experiment to collect a new gestural dataset for use in training our model and eventually for public release.

Participants in the experiment will be placed in pairs in a room. The room will contain two tables and two small blinds concealing papers in front of each subject. One participant, the builder, will be given origami paper, and the other, the instructor, will receive a set of intentionally vague instructions for folding origami. The participants will be told that the instructions have been algorithmically generated, and that we wish to test their correctness and interpretability. By concealing the true purpose, this pretense ensures that the gestures produced are natural. The instructor will be asked to convey the directions for constructing the origami to the builder, using any speech or gestures desired, but without showing the other participant their instructions. The participants' speech and gestures will be recorded by microphones and Kinect sensors.

Based on a pilot trial of this study, we believe that it will result in a large quantity of high-quality iconic and metaphoric gestures and their accompanying speech. These recordings will be manually segmented and annotated by users on Amazon Mechanical Turk and used for further training and analysis. This annotated corpus will be publicly released to accompany the final paper.

We believe that this gestural data collection experiment has high potential for both immediate direct impact and longer-term indirect impact. The obvious benefit of the experiment is that it provides us with more data to use in training. By increasing both the quantity and quality of our training data, we hope to be able to attain better performance at the unfamiliar gestures task. More broadly, however, the collected dataset will enable further studies to be conducted by both our lab and other researchers. The dataset is intentionally general — nothing in its framing or collection is inherently robotics-specific. This generality makes the dataset potentially interesting to researchers across the fields of psychology, computer vision, machine learning, HCI, HRI, and general robotics. The data collected will be realistic, as participants will be kept oblivious of the true purpose of the study, and no special effort will be made to elicit or force gestures. While the task is artificial, it still represents a realistic example of a collaborative problem-solving task. We intend to release the collected dataset to the public, enabling these and further uses.

References

[1] Yoav Artzi and Luke Zettlemoyer. *UW SPF: The University of Washington Semantic Parsing Framework*.

[2] Qing Chen, Nicolas D. Georganas, and E.M. Petriu. Real-time vision-based hand gesture recognition using haar-like features. In *Instrumentation and Measurement Technology Conference Proceedings, 2007. IMTC 2007. IEEE*, pages 1–6. doi: 10.1109/IMTC.2007.379068.

[3] Miles Eldon, David Whitney, and Stefanie Tellex. Interpreting multimodal referring expressions in real time. URL https://edge.edx.org/asset-v1:Brown+CSCI2951-K+2015_T2+type@asset+block@eldon15.pdf.

[4] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Isabelle Guyon, Vassilis Athitsos, Hugo Escalante, Leonid Sigal, Antonis Argyros, Cristian Sminchisescu, and others. Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 365–368. ACM, .

[5] Sergio Escalera, Jordi Gonzàlez, Xavier Baró, Miguel Reyes, Oscar Lopes, Isabelle Guyon, Vassilis Athitsos, and Hugo Escalante. Multi-modal gesture recognition challenge 2013: Dataset and results. In *Proceedings of the 15th ACM on International conference on multimodal interaction*, pages 445–452. ACM, .

[6] Piotr Gawron, Przemysław Głomb, Jarosław Adam Miszczak, and Zbigniew Puchała. Eigengestures for natural human computer interface. 103:49–56. doi: 10.1007/978-3-642-23169-8_6. URL http://arxiv.org/abs/1105.1293.

[7] S. S. Ge, Y. Yang, and T. H. Lee. Hand gesture recognition and tracking based on distributed locally linear embedding. 26(12):1607–1620. ISSN 0262-8856. doi: 10.1016/j.imavis.2008.03.004. URL http://www.sciencedirect.com/science/article/pii/S0262885608000693.

[8] Isabelle Guyon, Vassilis Athitsos, Pat Jangyodsuk, and Hugo Jair Escalante. The ChaLearn gesture dataset (CGD 2011). 25(8):1929–1951. ISSN 0932-8092, 1432-1769. doi: 10.1007/s00138-014-0596-3. URL http://link.springer.com/article/10.1007/s00138-014-0596-3.

[9] Saumya Jetley, Bernardino Romera-Paredes, Sadeep Jayasumana, and Philip Torr. Prototypical priors: From improving classification to zero-shot learning. URL http://arxiv.org/abs/1512.01192.

[10] Karen Spärck Jones. A statistical interpretation of term specificity and its application in retrieval. 28:11–21.

[11] Thomas Kollar, Stefanie Tellex, Matthew R Walter, Albert Huang, Abraham Bachrach, Sachi Hemachandra, Emma Brunskill, Ashis Banerjee, Deb Roy, Seth Teller, and others. Generalized grounding graphs: A probabilistic framework for understanding grounded language. URL https://people.csail.mit.edu/sachih/home/wp-content/uploads/2014/04/G3_JAIR.pdf.

[12] Y. Kondo, K. Takemura, J. Takamatsu, and T. Ogasawara. Body gesture classification based on bag-of-features in frequency domain of motion. In *2012 IEEE RO-MAN*, pages 386–391. doi: 10.1109/ROMAN.2012.6343783.

[13] Dan Luo and Jun Ohya. Study on human gesture recognition from moving camera images. In *2010 IEEE International Conference on Multimedia and Expo (ICME)*, pages 274–279. doi: 10.1109/ICME.2010.5582998.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. URL http://arxiv.org/abs/1301.3781.

[15] Mark Palatucci, Dean Pomerleau, Geoffrey Hinton, and Tom Mitchell. Zero-shot learning with semantic output codes. In *Neural Information Processing Systems (NIPS)*.

[16] Vaughn Segers and James Connan. Real-time gesture recognition using eigenvectors. URL http://www.cs.uwc.ac.za/~jconnan/publications/Paper%2056%20-%20Segers.pdf.

[17] Richard Socher, Milind Ganjoo, Hamsa Sridhar, Osbert Bastani, Christopher D. Manning, and Andrew Y. Ng. Zero-shot learning through cross-modal transfer. URL http://arxiv.org/abs/1301.3666.

[18] Hafiz Imtiaz Upal Mahbub. One-shot-learning gesture recognition using motion history based gesture silhouettes. doi: 10.12792/iciae2013.037.

[19] Jun Wan, Qiuqi Ruan, Wei Li, and Shuang Deng. One-shot learning gesture recognition from RGB-d data using bag of features. 14(1):2549–2582. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2567709.2567743.

[20] Di Wu, Fan Zhu, and Ling Shao. One shot learning gesture recognition from RGBD images. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 7–12, . doi: 10.1109/CVPRW.2012.6239179.

[21] Jiaxiang Wu, Jian Cheng, Chaoyang Zhao, and Hanqing Lu. Fusing multi-modal features for gesture recognition. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 453–460. ACM, . ISBN 978-1-4503-2129-7. doi: 10.1145/2522848.2532589. URL http://doi.acm.org/10.1145/2522848.2532589.

[22] Ying Yin and Randall Davis. Gesture spotting and recognition using salience detection and concatenated hidden markov models. In *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*, ICMI '13, pages 489–494. ACM. ISBN 978-1-4503-2129-7. doi: 10.1145/2522848.2532588. URL http://doi.acm.org/10.1145/2522848.2532588.

[23] Yin Zhou, Kai Liu, Rafael E. Carrillo, Kenneth E. Barner, and Fouad Kiamilev. Kernel-based sparse representation for gesture recognition. 46(12):3208–3222. ISSN 0031-3203. doi: 10.1016/j.patcog.2013.06.007. URL http://dx.doi.org/10.1016/j.patcog.2013.06.007.