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

Wil Thomason and Ross A. Knepper

Department of Computer Science, Cornell University
{wbthomason,rak}@cs.cornell.edu

1 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 single modality. 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 [1, 12].

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, yet remain restricted to a fixed gestural lexicon [2, 6, 7, 13, 18, 25] and cannot recognize gestures outside of a small pre-trained set with any accuracy [2, 14].

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 wish 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 [10, 16, 19] 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 [21–23]. 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. Work in another direction has investigated the use of gestures by robots [9, 17]. Work in this area has focused on studying which gestures are most effective in robotic storytelling (e.g. Huang and Mutlu [9]), or on creating systems to make it easier for humans to encode gestures for robots to make. We aim to provide understanding of gestural meaning. Finally, the work of Takano, Hamano, and Nakamura [20] moves toward a general association between word labels and gestures through the use of correlated vector spaces. This work is focused on the retrieval of relevant motion data for a word query from a database, whereas our work seeks to construct a mapping from gestures to words. In general, the state of the art in recognition and gestural understanding, appears to be limited to pointing gestures as in [3], and other gestural recognition techniques which have been developed independently of robotic applications. 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.

2 Technical 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.

Intuitively, 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.

2.1 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:

Gesture Embedding: The first step of our approach is to create an embedding space mapping gestures to the corresponding words. We begin by splitting a gesture into its constituent sub-gestural motions. For a gesture g encoded as a series of RGB-D frames, we first partition the frames of g into windows of 120 ms, each overlapping by 20 ms. The purpose of these windows is to approximate sub-gestures. We rely on this approximation due to the recursive structure of the sub-gestural model: gestures are composed of sub-gestures, which may themselves be composed of sub-gestures, and so on. Thus, we use short overlapping windows to attempt to capture the "first level" of this structure, i.e. the sub-gestures which directly compose into gestures. The duration of these windows and their overlap was determined empirically. In future work, we hope to explore the possibility of dynamically-sized windows or other means of more accurately segmenting sub-gestures.

Next, we extract the human skeleton H of the user from each window, and compute the velocity $\boldsymbol{v}_{\boldsymbol{j}_i}$ of the joints $j_1, \ldots, j_n$ comprising H for each frame in the window. This process results in a time series $\boldsymbol{V}$ of joint velocities in the window. We complete feature computation by computing the discrete Fourier transform $\boldsymbol{\psi}_{\boldsymbol{g}}$ of $\boldsymbol{V}$. Specifically, we compute for each joint the 3-D Fourier transform of its velocity in $\boldsymbol{V}$. This feature is inspired by Kondo et al. [13] in its use of a transform of joint velocities as a means of describing gestures. However, we differ from Kondo et al. [13] in several ways. First, our feature representation is over sub-gestures rather than whole gestures. This difference is key to our model of gestures as a composition of smaller semantic units. Second, the features used in Kondo et al. [13] are histograms of frequency domain transforms of gestures, whereas this work uses the raw frequency domain representation of each sub-gesture.

After computing $\boldsymbol{\psi_g}$, we use it as the input to a neural network. This network is 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. This structure is simply a standard multi-layer perceptron placed atop a two-layer convolutional architecture often used in the field of object classification. The architecture of this network was chosen for its simplicity and ease of training; we hope to investigate the use of alternate architectures with our sub-gestural feature descriptor and zero-shot learning model in future work.

We assume that there exists a bag of words $W = \{\boldsymbol{w_1}, \dots, \boldsymbol{w_k}\}$ associated with each g, where each $\boldsymbol{w_i}$ is encoded as a vector in a pre-trained word embedding (in particular, we use Word2Vec [15]). 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:

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

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

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, Whitney, and Tellex [3], imposes an ordering on the candidate descriptors by computing a variant on the common tf-idf metric [11] for each. This variant is a direct analogue of tf-idf for the gestural context, and computes:

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

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 2.1 returns k possible descriptors, the top $\ell < k$ according to their ranking by $\mathcal{S}$ are chosen for the final output of the system.

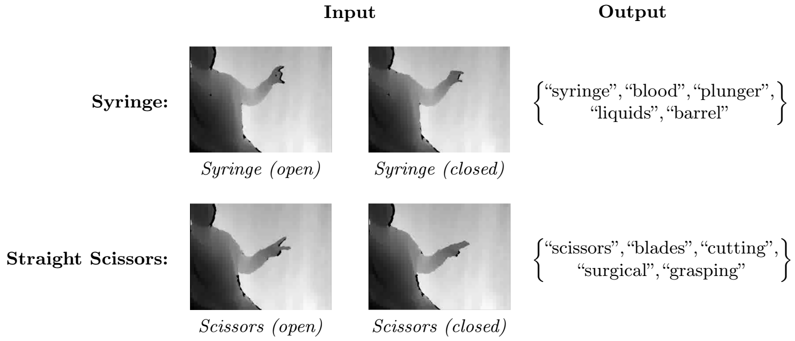
3 Experiments

We have conducted several experiments to validate the performance of our technique.

3.1 ChaLearn Dataset

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

3.2 Zero-Shot Model


Input
Output
Syringe:
{"syringe", "blood", "plunger", "liquids", "barrel"}
Straight Scissors:
{"scissors", "blades", "cutting", "surgical", "grasping"}

Syringe (open) *Syringe (closed)*

Scissors (open) *Scissors (closed)*

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. 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 surgical instruments indicated by each class of gesture. We withheld all examples of the straight scissors class from the training process as test data. 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. We note that holding out several classes produced lower-quality results; however, given that our training dataset was very small (100 gestures, total), we attribute this drop in performance to this change causing insufficient training data.

We have also performed an experiment in which we held out each class of surgical gesture in turn, and assessed the performance of our system. The goal of our unfamiliar gesture understanding system is to produce clusters of words for a gesture which a human would agree were reasonably associated with said gesture. As such, we have devised the following metric of performance: For each bag of words returned by our system, we label the result as "Not Relevant" if it contains fewer than four words deemed relevant to the input gesture by a human, "Relevant" if it contains between five and eight such words, and "Very Relevant" if it contains nine or ten such words (the size of the returned bag of words is ten). The results of our system's performance according to this metric are shown in Figure 2.

As may be seen, we achieve a majority of "Relevant" or "Very Relevant" results in a significant number of cases. However, there are notably some cases (such as when Army-Navy Retractor is the held-out class) for which our system performs very poorly. However, given the very low suitability of the ChaLearn data for our task, these results still demonstrate that our system is capable of providing reasonable descriptions of unfamiliar gestures.

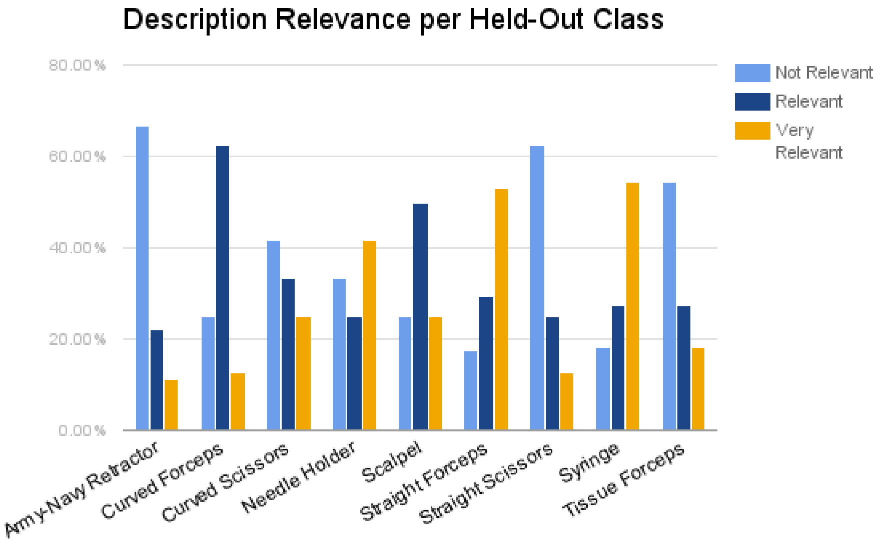


Fig. 2: The performance of our unfamiliar gesture understanding system for each held out class of surgical gesture

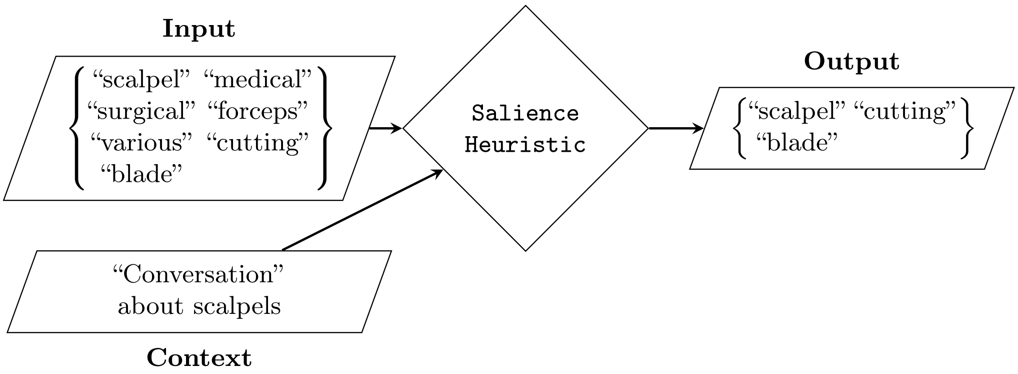

Input
"scalpel" "medical" "surgical" "forceps" "various" "cutting" "blade"
Salience Heuristic
Output
"scalpel" "cutting" "blade"
"Conversation" about scalpels
Context

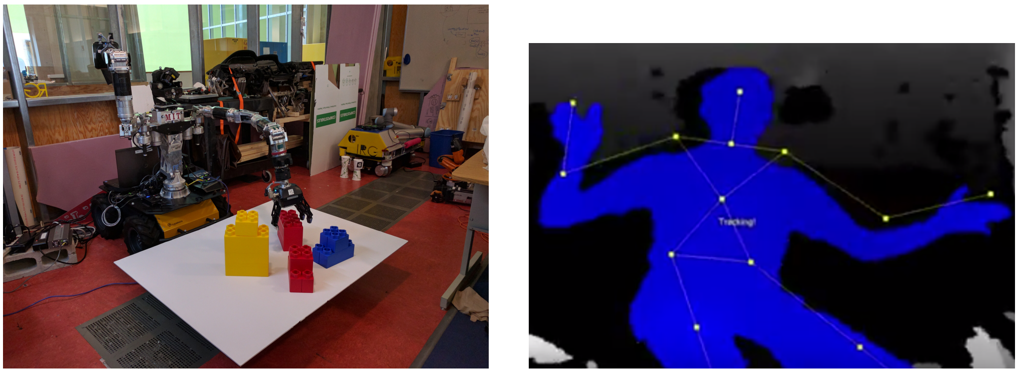
Fig. 3: The output of our salience heuristic on an example "conversation".

3.3 Salience Heuristic

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 3. 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. We ran trials on a large number of simulated conversations, injecting intentionally irrelevant terms into the input and testing if they were removed (without removing the relevant terms) after passing the conversation and input through the heuristic. In 61% of trials for these simulated data, we found that the heuristic scored the inserted irrelevant words as less relevant than the inserted relevant words, as desired.

These 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.

3.4 End-to-End Gesture Understanding



(a) The Optimus mobile manipulation platform in the experimental setup

(b) The skeleton of the gesturer, mid-gesture

Fig. 4: The unfamiliar gesture understanding system integrated with the Optimus mobile manipulation platform.

We have integrated our system into a real-world robotic platform to test its end-to-end functionality. The experimental setup (pictured in Figure 4a) was as follows: A human user and a mobile manipulation platform are positioned on opposite sides of a table, facing each other. A set of objects are placed onto the table. The human user makes a request for a particular object, and the manipulation platform must understand the request and grasp the correct object. Critically, the request made by the human contains both verbal and gestural elements, and is ambiguous without consideration of both components in conjunction. Specifically, the verbal component of the request identifies an object by color, but the table holds several objects of the specified color, making the referent object ambiguous. In this case, the gestural component is used to communicate the relative size of the referent object, disambiguating the request.

For this experiment, our system was made to run online and integrated with the verbal understanding and manipulation components of the overall platform. Thus, although this experiment is fairly simple in terms of the gestures, it serves to demonstrate the viability of our system for use in robotic applications.

In future work, we intend to measure the impact of our unfamiliar gesture understanding system on the overall understanding capability of a robot participating in a collaborative task with a human. We plan to run the entire system (as detailed above), on a Rethink Robotics Baxter robot. 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. We will run trials in which a human user will be asked to indicate to Baxter 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. [12, 24–26]), trained on the same data as we use to train our system. We will post the results of this experiment to our project site[1].

3.5 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 have begun to conduct an experiment to collect a new gestural dataset for use in training our model and eventually for public release.

Participants in the experiment are placed in a room with the study organizer. The room contains two tables, one for the participant and the other for the organizer. The table for the organizer holds a small blind, under which a piece of origami paper is placed. The participant is given a set of intentionally vague instructions for folding origami. They are 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 participant is asked to convey the directions for constructing the origami to the organizer, using any speech or gestures desired, but without showing the organizer their instructions. The participant's speech and gestures are recorded by microphones and Kinect sensors.

We have captured gestures from approximately 15 participants in this manner. Most sessions result in a large number of gestures describing the physical properties of the origami being folded: shapes, relative sizes, and fold structures (i.e. the direction and placement of a fold) are the concepts most commonly communicated through gesture. We are continuing to collect data, and hope to record a minimum of 50 participants before concluding the study.

We will be releasing the collected data on our project site[1]. The recordings from each trial will be transcribed and processed to extract the skeletal data of each participant. These transcriptions will be annotated with timing information. To ensure the anonymity of the study participants, we will release only the annotated transcripts and skeletal data for each trial to the public. We believe that this combination of data is sufficient to make the dataset useful for experiments in gestural understanding, linguistics, and other fields.

We hope that the completed dataset will have both immediate direct impact and longer-term indirect impact. The obvious benefit of the study is that it provides us with more data for 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 are realistic, as participants are kept oblivious of the true purpose of the study, and no special effort is made to elicit or force gestures. While the task is artificial, it still represents a realistic example of a collaborative problem-solving task. This means that it may be of interest to researchers in areas entirely separate from the topic of gesture, such as group dynamics and sociology.

[1] https://rpal.cs.cornell.edu/projects/unfamiliar-gestures

4 Conclusions

The largest weakness of our unfamiliar gesture understanding system is the lack of data suitable for use in training of the system. We are seeking to rectify this deficiency through our aforementioned data collection experiment; however, this experiment has not yet been concluded. This lack of data has limited evaluations of our system thus far to relatively simple applications. Even so, we are able to draw some conclusions about the performance and properties of our system.

First, it is apparent that the performance of the unfamiliar gesture understanding system is predicated on the quality of the word embedding space it uses. In the most basic sense, the word embedding must contain mappings for words which could reasonably be used to describe any gesture that the system hopes to be able to understand. We do not yet have a means of determining a threshold for suitability, which means that entirely unrelated words may be returned for a gesture in the absence of sufficiently many relevant words. Although our salience heuristic is designed to remove irrelevant words, it cannot determine if a word is relevant to a particular gesture, but only to a context. More subtly, we rely on the word embedding space placing similar words close to each other. While this property often holds, it is not universally true. Related to these issues is the tradeoff between coverage and comprehensibility. In other words, if the word embedding space contains more words and thus has better coverage, it may have lower comprehensibility, because it is more probable that an unrelated word will be closer to the point at which a gesture is embedded.

Second, we see room for improvement in experimentation with both the sub-gestural feature representation and the architecture of the neural network used to compute the aligned gesture embedding. In the latter case, we have experimented with the number of layers in the multi-layer perceptron component of the network, but the dearth of data available for training means that we quickly succumb to overfitting as more layers are added. In the former case, although our current approximation does a reasonable job of capturing small motions corresponding to sub-gestures, our intuition for sub-gestures suggests that they are not all of uniform or bounded duration, and thus that a more adaptive segmentation approach may have greater success.

Acknowledgements

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.

References

[1] Yoav Artzi and Luke Zettlemoyer. *UW SPF: The University of Washington Semantic Parsing Framework*. 2013.

[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*. May 2007, pp. 1–6. DOI: 10.1109/IMTC.2007.379068.

[3] Miles Eldon, David Whitney, and Stefanie Tellex. "Interpreting Multimodal Referring Expressions in Real Time". In: (2015). URL: https://edge.edx.org/asset-v1:Brown+CSCI2951-K+2015_T2+type@asset+block@eldon15.pdf.

[4] Sergio Escalera et al. "Chalearn multi-modal gesture recognition 2013: grand challenge and workshop summary". In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 365–368.

[5] Sergio Escalera et al. "Multi-modal gesture recognition challenge 2013: Dataset and results". In: *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 445–452.

[6] Piotr Gawron et al. "Eigengestures for natural human computer interface". In: *arXiv:1105.1293 [cs]* 103 (2011), pp. 49–56. DOI: 10.1007/978-3-642-23169-8_6. arXiv: 1105.1293. URL: http://arxiv.org/abs/1105.1293 (visited on 10/29/2015).

[7] S. S. Ge, Y. Yang, and T. H. Lee. "Hand gesture recognition and tracking based on distributed locally linear embedding". In: *Image and Vision Computing* 26.12 (Dec. 1, 2008), pp. 1607–1620. ISSN: 0262-8856. DOI: 10.1016/j.imavis.2008.03.004. URL: http://www.sciencedirect.com/science/article/pii/S0262885608000693 (visited on 11/18/2015).

[8] Isabelle Guyon et al. "The ChaLearn gesture dataset (CGD 2011)". In: *Machine Vision and Applications* 25.8 (Feb. 21, 2014), pp. 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 (visited on 02/03/2016).

[9] Chien-Ming Huang and Bilge Mutlu. "Modeling and Evaluating Narrative Gestures for Humanlike Robots." In: *Robotics: Science and Systems*. 2013.

[10] Saumya Jetley et al. "Prototypical Priors: From Improving Classification to Zero-Shot Learning". In: *arXiv:1512.01192 [cs]* (Dec. 3, 2015). arXiv: 1512.01192. URL: http://arxiv.org/abs/1512.01192 (visited on 01/29/2016).

[11] Karen Spärck Jones. "A statistical interpretation of term specificity and its application in retrieval". In: *Journal of Documentation* 28 (1972), pp. 11–21.

[12] Thomas Kollar et al. "Generalized grounding graphs: A probabilistic framework for understanding grounded language". In: *JAIR* (2013). URL: https://people.csail.mit.edu/sachih/home/wp-content/uploads/2014/04/G3_JAIR.pdf.

[13] Y. Kondo et al. "Body gesture classification based on Bag-of-features in frequency domain of motion". In: *2012 IEEE RO-MAN*. 2012 IEEE RO-MAN. Sept. 2012, pp. 386–391. DOI: 10.1109/ROMAN.2012.6343783.

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

[15] Tomas Mikolov et al. "Efficient Estimation of Word Representations in Vector Space". In: *arXiv:1301.3781 [cs]* (Jan. 16, 2013). arXiv: 1301.3781. URL: http://arxiv.org/abs/1301.3781 (visited on 03/30/2016).

[16] Mark Palatucci et al. "Zero-Shot Learning with Semantic Output Codes". In: *Neural Information Processing Systems (NIPS)*. Dec. 2009.

[17] Allison Sauppé and Bilge Mutlu. "Robot Deictics: How Gesture and Context Shape Referential Communication". In: *Proceedings of the 2014 ACM/IEEE International Conference on Human-robot Interaction*. HRI '14. New York, NY, USA: ACM, 2014, pp. 342–349. ISBN: 978-1-4503-2658-2. DOI: 10.1145/2559636.2559657. URL: http://doi.acm.org/10.1145/2559636.2559657 (visited on 11/19/2015).

[18] Vaughn Segers and James Connan. "Real-time gesture recognition using eigenvectors". In: (2009). URL: http://www.cs.uwc.ac.za/~jconnan/publications/Paper%2056%20-%20Segers.pdf.

[19] Richard Socher et al. "Zero-Shot Learning Through Cross-Modal Transfer". In: *arXiv:1301.3666 [cs]* (Jan. 16, 2013). arXiv: 1301.3666. URL: http://arxiv.org/abs/1301.3666 (visited on 01/25/2016).

[20] Wataru Takano, Seiya Hamano, and Yoshihiko Nakamura. "Correlated space formation for human whole-body motion primitives and descriptive word labels". In: *Robotics and Autonomous Systems* 66 (2015), pp. 35–43.

[21] Hafiz Imtiaz Upal Mahbub. "One-Shot-Learning Gesture Recognition Using Motion History Based Gesture Silhouettes". In: (2013). DOI: 10.12792/iciae2013.037.

[22] Jun Wan et al. "One-shot Learning Gesture Recognition from RGB-D Data Using Bag of Features". In: *J. Mach. Learn. Res.* 14.1 (Jan. 2013), pp. 2549–2582. ISSN: 1532-4435. URL: http://dl.acm.org/citation.cfm?id=2567709.2567743 (visited on 01/25/2016).

[23] 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)*. 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). June 2012, pp. 7–12. DOI: 10.1109/CVPRW.2012.6239179.

[24] Jiaxiang Wu et al. "Fusing Multi-modal Features for Gesture Recognition". In: *Proceedings of the 15th ACM on International Conference on Multimodal Interaction*. ICMI '13. New York, NY, USA: ACM, 2013, pp. 453–460. ISBN: 978-1-4503-2129-7. DOI: 10.1145/2522848.2532589. URL: http://doi.acm.org/10.1145/2522848.2532589 (visited on 03/31/2016).

[25] 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. New York, NY, USA: ACM, 2013, pp. 489–494. ISBN: 978-1-4503-2129-7. DOI: 10.1145/2522848.2532588. URL: http://doi.acm.org/10.1145/2522848.2532588 (visited on 01/22/2016).

[26] Yin Zhou et al. "Kernel-based Sparse Representation for Gesture Recognition". In: *Pattern Recogn.* 46.12 (Dec. 2013), pp. 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 (visited on 01/29/2016).