Multimodal Visual and Simulated Muscle Activations for Grounded Semantics of Hand-related Descriptions

Daniele Moro
Computer Science Department
Boise State University
danielemoro
@u.boisestate.edu

Casey Kennington
Computer Science Department
Boise State University
caseykennington
@boisestate.edu

Abstract

In this paper, we build on research which has applied visually-derived features for grounded semantics by leveraging an additional modality: simulated hand muscle activations. We apply the Words-as-Classifiers model of grounded semantics to learn a mapping between features from the two modalities and corresponding hand image descriptions. Our experimental results show that a multimodal fusion of both visual and muscle features yields improved results for the model than either of the modalities alone in image and description retrieval tasks. By simulating mirror neurons, we further show that the simulated muscle activations can be derived from the visual features and applied to our model.

1 Introduction

Part of the semantic representation and meaning of many words is *grounded* (Harnad, 1990) in how people perceive and experience the physical world. For example, the semantic meaning of the word *red* is grounded in a person's perception and experience in perceiving objects denoted by others as *red* through color vision. Though vision is an important and common modality for grounded semantics research, the semantic meaning of words can also be grounded in other perceptual modalities, such as auditory (Kiela and Clark, 2015) and olfactory (Grabski et al., 2012) perception. In this paper, we take inspiration from Roy (2005) which set forth a theoretical framework for language grounding from embodied, situated, sensorimotor primitives to words. We explore an embodied modality of simulated muscle activations in hands for grounded semantics. We hypothesize that the meaning of words and descriptions that are related to human hands, for example, *grip*, *point*, and *thumbs up*, are not only grounded in how those hand configurations are depicted visually, but also grounded in the muscle activations and muscle memory required to physically make those hand configurations. We explore this by simulating hand configurations using a virtual, soft-robotic inspired hand (Schlagenhauf et al., 2018; King et al., 2018) where the finger positions are defined by simulated muscle activations, which we use as features for the *Words-as-Classifiers* (WAC) model (Kennington and Schlangen, 2015; Schlangen et al., 2016) as well as a WAC-inspired neural network model and show that muscle activations coupled with visual features strengthen the grounded semantic meaning applied to image and description recall tasks. Our results could be used to augment human understanding in interactive robots by supporting a growing body of research around an embodied semantics, which postulates that grounding incorporates not only perceptual modalities, but also sensorimotor modalities (Johnson, 2008; Goertzel et al., 2010).

We further explore a potential approach to modeling mirror neurons in the brain–which discharge not only during action execution, but also during action observation (Kilner et al., 2009)–which allows our model to use muscle activations derived from a visual representation of the hands. If an embodied system (e.g., such as a robot) is to make use of both visual and muscle modalities, then both modalities must each come from some component that is part of the system where those features can be derived (i.e., a robot must have a camera for the visual and a soft-robotic hand for the muscles). It would be more common, and potentially more useful, for a system to make use of muscle activation information by simply observing someone else's hand configuration visually. For example, an individual's own neurons are activated for generating a *grip* in his own hand when that individual sees someone else making a

grip with her own hand. This is what mirror neurons afford humans, the existence of which has been fairly well supported (Kilner et al., 2009).[1] In summary, we make the following contributions: (1) We model a form of grounded semantics using muscle activations and visual/image representations of hand configurations, (2) we offer a set of data which includes images of hands with corresponding descriptions, simulated muscle activations, and visual/image features, (3) we further a notion of embodied semantics which leverages from perception (i.e., the outside world) as well as muscles (i.e., the inside, corporeal world) and offer a simple model for applying mirror neurons.

2 Related Work

Several areas of research play into this work including seminal (Roy and Reiter, 2005; Roy, 2005) and recent work in grounded semantic learning in various tasks and settings, notably learning descriptions of the immediate environment (Walter et al., 2014); navigation (Kollar et al., 2010); nouns, adjectives, and relational spatial descriptions (Kennington and Schlangen, 2015); attributes (Matuszek et al., 2012), verbs (She and Chai, 2016), and grounded distributional semantics (Bruni et al., 2014). We build on this previous work in that we represent the grounded semantics by linking meaning with visual features, yet we go beyond this work in that we consider representations of muscle activations as an additional modality of semantic meaning.

Other recent work has already gone beyond visual grounded semantics including olfactory perception (Kiela et al., 2015), auditory perception (Kiela and Clark, 2015), haptics (Alomari et al., 2017), and multimodal features including haptic, auditory, and proprioceptive (Thomason et al., 2016). Very similar to our goal of grounding into modalities beyond vision is Marocco et al. (2010) who grounded action words into sensorimotor actions of a simulated robot.[2] Our work is novel in that we are not solely focusing on a perceptual modality (e.g., such as vision); rather, we are building off of this line of research to explore corporeal modalities for an embodied semantics.

3 Model: Words-as-Classifiers

The WAC model follows Larsson (2015) as a simple approach to bridging grounded and formal semantics. It has recently been shown to yield state-of-the-art results in a reference resolution task using deep neural networks to represent photographs (Schlangen et al., 2016) as well as in real-time dialogue systems that can resolve references made to visual objects (Manuvinakurike et al., 2016). Following Zarrieß and Schlangen (2016), the WAC model is essentially a task-independent approach to predicting semantic appropriateness of words in physical contexts and can be flexibly combined with task-dependent decoding procedures. The WAC model pairs each word w in its vocabulary V with a classifier that maps the real-valued features x of an object obj to a semantic appropriateness (i.e., class membership) score:

$$[\![w]\!]_{obj} = \lambda \mathbf{x}.p_w(\mathbf{x}) \tag{1}$$

For example, to learn the connotative meaning of the word *grip*, the low-level features (i.e., visual, sensorimotor, etc.) of all objects described as *grip* in a corpus of referring expressions are given as positive instances to a supervised learning classifier. Negative instances are randomly sampled from the complementary set of utterances (i.e., not containing the word *grip*). This results in a trained $\lambda \mathbf{x}.p_{grip}(\mathbf{x})$, where x is a novel object (in our case, features representing a hand pose) that can be applied to $grip$ to determine class membership. Traditionally, the WAC model has been applied using independent linear classifiers, such as logistic regression. In this paper, we apply both this traditional approach to our task, and we also apply the WAC model using a neural network where the fitness score is applied to all words in the vocabulary (which makes up the top layer), thereby reducing the independence between the classifiers. We chose WAC because of its simplicity and interpretability, and neural networks have been shown to yield state-of-the-art performance in many tasks. Both approaches to WAC learn a mapping between non-linguistic features and words.

[1] Though the existence and function of mirror neurons is not without debate (Dinstein et al., 2008).

[2] Similar in some ways to Grabski et al. (2012), we also explore how mirror neurons can be used to derive muscle activations from visual features (originally published in French).

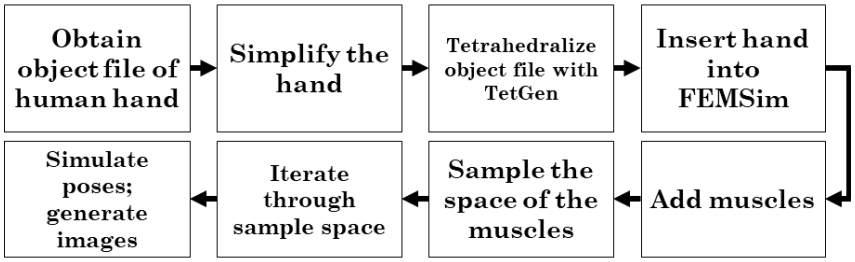

Figure 1: The process of gathering simulated hand pictures from a variety of poses

4 Data: The Multimodal Hand Corpus

In this section, we describe our approach to generating hand configurations with corresponding images, descriptions, and muscle activations which we used in our experiments.

4.1 Creating a Simulated Soft Hand

We generated simulated hand configurations by leveraging and expanding upon a soft body *Forward Simulation Model* (FEMSim) (Bern et al., 2017) which models a discretized two-dimensional soft object with simulated muscles that lie along the perimeter of the object. These muscles can be contracted, resulting in a state of the object (in our case, fingers of hands) where the potential energy is the lowest. For the purposes of this work, we expanded this model to support three dimensions, with muscles that run along the surface of the fingers. Because this model was more detailed than FEMSim simulation would reasonably support, we used the program MeshLab to perform a Quadric Edge Collapse Decimation and bring the model from over 56,000 faces to 1,000 faces. We then used a program called TetGen to tetrahedralize the mesh and generate the internal structure of the hand.

We added the muscles on the skin of the simulated hands with the constraint that each muscle must be able to naturally contract each finger. Each muscle is a collection of nodes of a mesh representing the simulated hand. The muscle imposes a soft constraint on the energy model of the soft object that allows the soft material to contract along the muscle nodes. A real number ranging from 0.0 to 1.0, which we denote as muscle activation values, is assigned to each muscle at every simulation step. Higher muscle activation values denote greater force that each muscle imposes in contracting the simulated nodes. Although the human hand contains dozens of muscles, we were constrained by the limits of our simulation to place 5 muscles on the hand to generate realistic motions. This resulted in a muscle from the tip of each finger to the palm. To allow for a greater range of motion and expression in the thumb, a 6th muscle connects the tip of the thumb and the wrist through the back of the hand. This approach led to the challenge of ensuring that the two thumb muscles worked together to produce natural thumb movements that mimicked human range of movements. As a result, we developed a coupling mechanism to abstract the two thumb muscles into a singular thumb muscle activation: higher thumb muscle activations result in the thumb approaching the palm of the hand.

4.2 Generating Hand Poses

After simulating a soft hand with acceptable movement fidelity to real hands, we captured images and recorded the corresponding muscle activations of different hand configurations. After placing the muscles on the simulated hands, we sampled the space of hand configurations by activating each muscle in the vector t in the activation space s and recording the resultant hand configuration. This resulted in $|s|^{|t|}$ total hand configurations. Because of this exponential nature, we constrained the muscle activation space s to be [0.0, 0.3, 0.7] as we found these activations to provide a meaningful range of distinguishable finger motions. In total, we generated 3^5, or 243 distinct poses.

We captured each hand pose through four different visual perspectives: *straight* (i.e., facing the palm), *above* (i.e., above the hand, facing downwards), *left* (i.e., with the thumb towards the camera), and *behind* (i.e., facing the back of the hand). See Figure 3 for one hand configuration from two of the four perspectives. This process resulted in a total of 972 images of hand configurations (243 hand poses * 4 camera angles; termed as *perspectives* below). Figure 1 depicts the entire process.

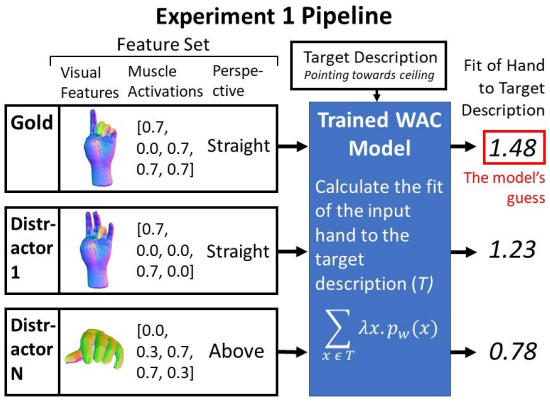

Figure 2: Evaluation strategy: select the highest scoring candidate from a set of N distractors and the gold image, given the description.

Figure 3: Two perspectives of the same muscle activation / hand configuration (left image: *left* perspective, right image: *straight* perspective).

4.3 Obtaining Image Descriptions

For the final part of our corpus, we used Amazon Mechanical Turk to obtain a description for each generated image. Each participant read and agreed to an informed consent, then they were taken to a web page that displayed 20 randomly selected hand images from our set, each with a text input box. They were given instructions to describe each hand pose as they would to a friend.

This collection resulted in two descriptions for each of the 972 images. After removing one description for inappropriate content, this resulted in 13,657 word tokens and a vocabulary size of 1,376. The average length of the descriptions was 7 words (std 4.72), where the most common number of words in a description was 2 (217 times). The most commonly used word was *hand* (685 occurrences) followed by *fingers* (525). 702 words occur once, and 185 words occur twice. Examples of words occurring once include *piano*, *scratch*, and *doornob*. The following are examples of descriptions from four images that were taken from the different perspectives (each perspective is denoted before each description) of the same hand configuration, which had a muscle activation of [0,0,0.3,0,0] (i.e., all fingers are straight except for the middle finger, which is slightly bent). Figure 3 corresponds to the *left* and *straight* descriptions:

1. *straight*: too little
2. *above*: the fingers and hand are curled as if holding a computer mouse but the thumb is outstretched
3. *left*: relaxed hand puppet
4. *behind*: fingers partially close thumb extended outward index finger slightly extended

We point out that these descriptions all described hands that had the *same configuration* (i.e., muscle activations), but since the task was a description of what they saw visually, and each depiction was from a different perspective, the descriptions can be quite varied. This tells us that our data captures something slightly more challenging than simply determining the name of an object: a configuration of a hand can be described in many different ways depending on the perspective.

5 Experiment 1: Hand Image Retrieval

In this section, we explain how we applied our model and data in an image retrieval task.

5.1 Task & Procedure

We follow Han et al. (2015) and Han and Schlangen (2017) and use a retrieval task to evaluate our model (we leave other informative evaluations, such as generating descriptions from features, as future work). That is, after our model has been trained, for each test instance we randomly select m distractor hand perspectives and our model is to pick out the correct hand perspective, given the description. We trained on all of the training data, and cross-validated the heldout data which comprised 10% of the data (i.e., 195 instances) with four folds, averaged over five runs, on three model variants which we describe below:

- **muscle** - only uses features related to 5 muscle activations and the orientation of the image

- **visual** - only uses visual features
- **muscle+visual** - use all muscle and all visual features

Representing the Features The *muscle* features are represented as real numbers between 0.0 and 1.0, where 0.0 represents no muscle activation and 1.0 represents full muscle activation (e.g., a hand where all 5 fingers are in a tight grip would have all five activations near 1.0; a relaxed hand would have all muscle activations near 0.0). For visual features, we apply a transfer learning approach (Pan and Yang, 2010) and use a pre-trained VGG19 convolutional neural network (CNN), which takes in an image at the bottom layer and outputs a softmax distribution over 1000 possible classes (Simonyan and Zisserman, 2014). The VGG19 was trained on the ILSVRC-2012 data set which contains 1.3 million images grouped into 1000 classes (i.e., the images depicted individual entities such as an animal or an object). We used the development data to empirically determine parameters, including which layer of the VGG19 model that we should use. We also included four binary features that represented the particular perspective (i.e., *straight, above, left, behind*) of the image. This resulted in 5 possible muscle features and 1004 possible visual features for our model.

Models We performed the experiment on two WAC model variants: logistic regression (WAC_{LR}) and a neural network (WAC_{NN}). For the WAC_{LR} variant, we used scikitlearn (Pedregosa et al., 2011) for each word w in the vocabulary by taking all descriptions where w was found and used the corresponding features for the hand configuration. This resulted in a separate classifier for each word in the vocabulary. For the WAC_{NN} variant, the input features were identical to that of WAC_{LR}, but in keeping in the spirit of WAC, the top layer was the full vocabulary. We used a dense input layer (activation=tanh) where the input shape was the number of features (which varied depending on the modalities being evaluated), an additional dense layer (activation=tanh) which had $|V| * 2$ neurons, and a top (activation=softmax) layer where the words in the vocabulary made up the class labels. We used the Adam (Kingma and Lei Ba, 2015) optimizer (learning rate=0.001) and categorical cross-entropy for gradient descent for 15 epochs (batch=256). We determined these parameters empirically by cross-validating on our training data.

Training For WAC_{LR} we train individual classifiers for each word, where each classifier can determine the probability of class "fit", and for WAC_{NN}, we train a single model which yields a softmax distribution for class "fit" for all words in the vocabulary. For the *muscle* variant, we only used the 5 muscle-related features, for the *visual* variant , we used the 1004 image features, and for the *muscle+visual* variant we used all of the features (i.e., muscle and visual concatenated) to give to each w as positive examples and randomly selected negative examples from descriptions that did not use w. For each positive example, we used three negative examples (the number of negative examples was also determined using the development set of our data). This means that, at a minimum, each word had at least 4 training instances (i.e., for words which only showed up once in our data). We removed several words from our vocabulary which were common in many of the descriptions (*hand, and, the, a, with, is, are, to,* and *of*) which would provide minimal semantic value. Note that for WAC_{NN}, we tested L1 and L2 regularization using a development set of data without any additional benefit.

Testing For each word w in each description, we apply the features of each of our distractor hand configurations as well as the true hand configuration as candidates to the WAC for that w (for the WAC_{NN} variant, we obtain the probability for w in the top layer's distribution) and compose a final probability over all of the candidates, adding together the results for each candidate. We take the *argmax* of the distribution as the model's guess and check if it is the true hand configuration which belongs to the description. This process is represented in Figure 2.

Metrics We use the accuracy of choosing the true hand configuration for all of the data in our cross-validation for each model variant using 1 to 5 distractors. The baseline for this model is random, or 1/(m+1) where m is the number of distractors. We hypothesize that the *muscle* variant will perform above baseline, but will not perform as well as *visual* because the descriptions were based on visual images, not on muscle activations. We further hypothesize that the multimodal *muscle+visual* variant will have the highest performance.

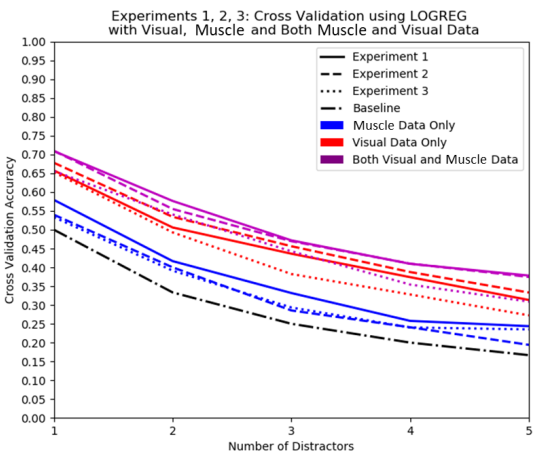

Figure 4: Results for Experiments 1 (image retrieval), 2 (description retrieval), and 3 (mirror neurons), where cross-validation was performed using the WAC$_{LR}$ variant of WAC. Experiment 1 results are solid, Experiment 2 results are dashed, Experiment 3 results are dotted.

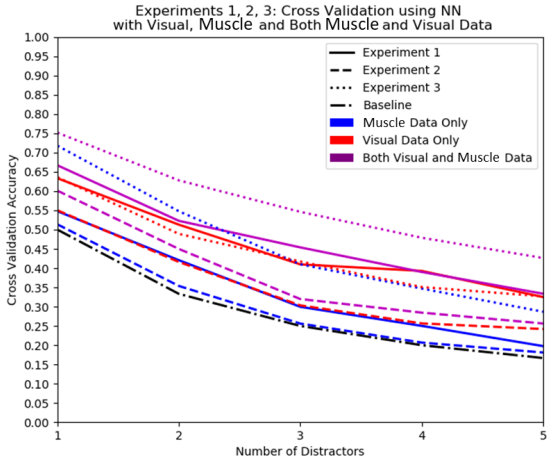

Figure 5: Results for Experiments 1 (image retrieval), 2 (description retrieval), and 3 (mirror neurons), where cross-validation was performed using the WAC$_{NN}$ variant of WAC. Experiment 1 results are solid, Experiment 2 results are dashed, Experiment 3 results are dotted.

Results For easier comparison across experiments, we grouped our results into two figures. The results for the cross-validation performed using the WAC$_{LR}$ is shown in Figure 4 as solid lines. The results for the cross-validation performed using the WAC$_{NN}$ is shown in Figure 5 as solid lines. As hypothesized, the multimodal *muscle+visual* model variant yields the highest performance. The muscle activations were above baseline, but do not perform very well on their own. Moreover, each hand configuration had multiple images associated with it (i.e., from different angles), each of which had distinct descriptions. This would cause confusion when WAC learned a mapping between muscle activations and words of a description. Unexpectedly, the differences in the modalities are more pronounced with the WAC$_{LR}$ model variant, despite is linearity assumption.

6 Experiment 2: Description Retrieval

In this section, we explain how we applied our model and data in a description retrieval task.

Task, Procedure & Metrics An equally important evaluation of our model reverses the retrieval task in Experiment 1. That is, for each test instance, we randomly select m distractor descriptions and task our model with picking out the correct description given the hand configuration features. We perform a 4-fold cross-validation using the heldout data on the three variants explained in Experiment 1. Moreover, instead of composing together the probability of each word in each description by summing, we average the probabilities of all words in each description so that longer descriptions are not favored by the model. The metrics for this experiment are similar to Experiment 1 except that we use the accuracy of the model choosing the true description for the corresponding hand configuration.

Results The results for the cross-validation performed using WAC$_{LR}$ is shown in Figure 4 and using WAC$_{NN}$ in Figure 5 as dashed lines. These results are comparable in trend to Experiment 1, with slightly lower scores overall. The visual+muscle modalities together perform better than visual or muscle alone. As in Experiment 1, WAC$_{NN}$ does not yield as high results as WAC$_{LR}$. This is possibly due to the sparsity of the data, but also potentially due to the way the two variants were approached: treating the WAC$_{LR}$ classifiers independently has some utility when the data are somewhat sparse.

7 Experiment 3: Simulation of Mirror Neurons

In this experiment, we repeat the task and procedure of Experiment 1 using muscle activations that are derived from the visual features.

Task, Procedure & Metrics The values that make up the *muscle* features are not directly observable like they were in Experiment 1. We train a Ridge Regression classifier that maps from the 1004 visual features to the 5 muscle features. We use the training data to learn this mapping, then pass each heldout image through the trained classifier to create a new set of muscle features that were derived from the visual features. This simulates, we claim, a very simplified function of mirror neurons; i.e., by observing a hand configuration visually, not only can a system use the visual features directly, the system can also derive muscle features from the visual features (RMSE on the development data is 0.0757). The training for WAC$_{LR}$ and WAC$_{NN}$ is the same as Experiment 1; we train using the original data (i.e., a robot that is making use of visual information to derive muscle information uses the model trained using its own muscle activations). The metrics for this experiment are the same as Experiment 1.

Results The results for the cross-validation performed using WAC$_{LR}$ is shown in Figure 4 as dotted lines; WAC$_{NN}$ is shown in Figure 5, also as dotted lines. The trend is largely the same as Experiments 1 and 2, with muscle working well above baseline, visual working well above muscle, and the visual+muscle performing the best for the WAC$_{LR}$ variant, though, as expected, lower overall when compared to Experiment 1, because the model is not using the true muscle values. For WAC$_{NN}$, the story is somewhat different from the first two experiments: not only is muscle above baseline, muscle alone performs better than visual, though visual+muscle perform the best. We explain this surprising, yet welcome, result by noting that the muscle features used in this experiment were derived from the visual features using Ridge Regression, resulting in muscle activation values that ranged more continuously between 0 and 1, whereas the muscle values in Experiments 1 and 2 were more discrete (i.e., values 0.0, 0.3, and 0.7). The WAC$_{NN}$ model can make use of finer distinctions in the features better than the WAC$_{LR}$ variant, and the wider variety in data may reduce overfitting in the WAC$_{NN}$ model .

8 Analysis

To understand the model's interpretation of the semantics of hand poses, we train the WAC$_{LR}$ and the WAC$_{NN}$ models as explained in Experiment 1 (i.e., *muscle+visual*) and isolate word w, then apply the model to all images, resulting in a fitness score p_w for that word. We then ranked the probabilities, resulting in the top x images for w. The x images are then grouped by perspective, and each of the four groups of perspectives are blended together to create four final images representing what a prototypical hand configuration would look like for w. For cases when a particular perspective was not represented in the top x images, then that perspective is labeled *Blank Image*. The more defined a region of the image, the more often this region of the image was represented in all of the x images. This allows us to analyze the overall "look" of a word by visualizing what configurations and perspectives in the image are more solid. Our chosen words are: *pointing*, *fist*, *ok*, *palm*, and *typing*.

pointing After applying all images to the trained WAC$_{LR}$ classifier for the word *pointing* (which occurred 103 times in our data), we took the 100 best fit images to produce Figure 6. We then repeated the above steps for the trained WAC$_{NN}$ classifier and generated Figure 7. All perspectives in both figures outline a pointing hand, with the index finger extended; other fingers mostly contracted. This shows that both models learned the prototypical grounded meaning of the word *pointing*. The results from the WAC$_{NN}$ are similar to the results from the WAC$_{LR}$, with the WAC$_{NN}$ variation showing a slightly more relaxed pointing hand than the WAC$_{LR}$ variation.

fist, ok, palm and **typing** Figure 8 shows the top 100 images that the WAC$_{LR}$ model learned to associate with the words *fist*, *ok*, *palm*, and *typing*. Each word occurs 62, 27, 184, and 23 times respectively in our data (we point out that WAC learned a reasonable semantics using only 23 examples for *typing*). Each word only has one perspective in the top 100 images (all four of the words' images are shown in one figure), showing that the perspective of hand pose may have a large impact on the image description. For

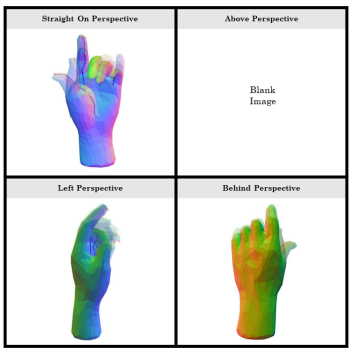

Figure 6: Blended fit for word *pointing* using WAC$_{LR}$ generated from the top 100 images. *Blank Image* means no images for that perspective.

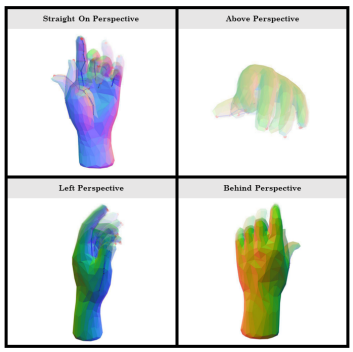

Figure 7: Blended fit of the word *pointing* using WAC$_{NN}$ generated from the top 100 images. In this case, all perspectives are represented.

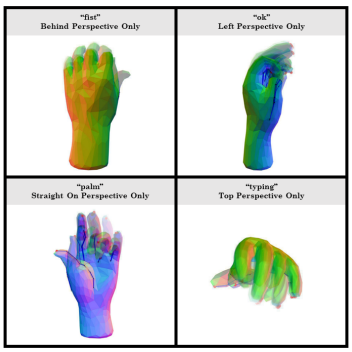

Figure 8: Blended fits for *fist*, *ok*, *palm*, and *typing* where only one perspective contained any images; generated from the top 100 images using WAC$_{LR}$.

fist, the *behind* perspective is shown, where all of the fingers are curled. For *ok*, the *left* perspective is shown, and the index finger is extended because human participants often used phrases such as *going to make an ok sign*. For *palm*, the *straight* perspective is shown, and the model learned that the position of the fingers plays little importance as long as the palm of the hand is showing. For *typing* the *above* perspective is shown. These demonstrate that the model learned the semantics of very specific words.

9 Discussion & Conclusion

In this paper, we presented novel, multimodal data which included simulated muscle activations with corresponding generated images and descriptions. We applied the multimodal data to the WAC model which learned a form of grounded semantics between words and two modalities of hand configurations: simulated muscle activations and visual representations. We showed that the model performed well above baseline using muscle activations alone, better with visual features derived from a VGG19 model, and the best when both modalities were present in a challenging image and description retrieval task. We also took inspiration from mirror neurons and applied a simplified approach to derive muscle features from the visual features, which yielded good results in an image retrieval task. We then analyzed our model and showed that WAC indeed learned to pick out prototypical hand configurations from the data.

A limitation of our solution is a lack of understanding of those external objects with which a hand might interact. For example, in the description *holding a small ball*, our model does not comprehend the physical meaning behind *ball*. We could augment our simulation with objects (such as balls) to provide our model information about external objects. Furthermore, instead of considering words independently, our model could be improved to develop a more contextual understanding.

Our work can potentially be applied to hand gesture recognition and sign language recognition through the use of mirror neurons, which we leave for future work. Moreover, our work furthers the notion that valuable information lies in embodied modalities. Roy (2005) stresses "the importance of binding symbols to sensorimotor representations, as evidenced by recent experiments that probe the embodied nature of cognitive processes." We envision a unified semantic theory that brings together distributional embeddings and grounded semantics (as posited in Thill et al. (2014)), which includes corporeal modalities. In the future, we hope to tackle a hand pose description generation task, as well as gathering hand descriptions from human participants experiencing the embodied sensation of particular hand poses. The code we used for the modeling and experiements and multimodal hand dataset are available.[3]

[3] https://github.com/bsu-slim/WAC-Hands

Acknowledgments

We would like to thank the anonymous reviewers for their useful comments and feedback.

References

Muhannad Alomari, Paul Duckworth, David C Hogg, and Anthony G Cohn. 2017. Natural Language Acquisition and Grounding for Embodied Robotic Systems AAAI17. In *In Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*.

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*.

Elia Bruni, Nam Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49:1–47.

Ilan Dinstein, Cibu Thomas, Marlene Behrmann, and David J Heeger. 2008. A mirror up to nature. *Current Biology*, 18(1):R13—-R18.

Ben Goertzel, Cassio Pennachin, Samir Araujo, Fabricio Silva, Murilo Queiroz, Ruiting Lian, Welter Silva, Michael Ross, Linas Vepstas, and Andre Senna. 2010. A general intelligence oriented architecture for embodied natural language processing. In *Artificial General Intelligence - Proceedings of the Third Conference on Artificial General Intelligence, AGI 2010*, pages 13–18.

Krystyna Grabski, Laurent Lamalle, and Marc Sato. 2012. Contrle prédictif et codage du but des actions orofaciales. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 1: JEP*, pages 289–296.

Ting Han and David Schlangen. 2017. Draw and Tell: Multimodal Descriptions Outperform Verbal- or Sketch-Only Descriptions in an Image Retrieval Task. In *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*.

Ting Han, Casey Kennington, and David Schlangen. 2015. Building and Applying Perceptually-Grounded Representations of Multimodal Scene Descriptions. In *Proceedings of SEMDial*, Gothenburg, Sweden.

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

Mark Johnson. 2008. *The meaning of the body: Aesthetics of human understanding*. University of Chicago Press.

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.

Douwe Kiela and Stephen Clark. 2015. Multi-and Cross-Modal Semantics Beyond Vision: Grounding in Auditory Perception. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2461–2470, Lisbon, Portugal.

Douwe Kiela, Luana Bulat, and Stephen Clark. 2015. Grounding Semantics in Olfactory Perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 231–236, Beijing, China, jul. Association for Computational Linguistics.

James M Kilner, Alice Neal, Nikolaus Weiskopf, Karl J Friston, and Chris D Frith. 2009. Evidence of Mirror Neurons in Human Inferior Frontal Gyrus. *Journal of Neuroscience*, 29(32):10153–10159.

Jonathan P King, Dominik Bauer, Cornelia Schlagenhauf, Kai-Hung Chang, Daniele Moro, Nancy Pollard, and Stelian Coros. 2018. Design, fabrication, and evaluation of tendon-driven foam manipulators. In *2018 IEEE RAS International Conference on Humanoid Robots*.

Diederik Kingma and Jimmy Lei Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Thomas Kollar, Stefanie Tellex, Deb Roy, and Nicholas Roy. 2010. Toward understanding natural language directions. *Human-Robot Interaction (HRI), 2010 5th ACM/IEEE International Conference on*, page 259.

Staffan Larsson. 2015. Formal semantics for perceptual classification. *Journal of Logic and Computation*, 25(2):335–369, dec.

Ramesh Manuvinakurike, Casey Kennington, David DeVault, and David Schlangen. 2016. Real-Time Understanding of Complex Discriminative Scene Descriptions. In *Proceedings of SigDial*.

Davide Marocco, Angelo Cangelosi, Kerstin Fischer, and Tony Belpaeme. 2010. Grounding action words in the sensorimotor interaction with the world: Experiments with a simulated icub humanoid robot. *Frontiers in Neurorobotics*, 4(MAY):7, may.

Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Bo Liefeng, and Dieter Fox. 2012. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proceedings of the 29th International Conference on Machine Learning*, pages 1671–1678.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, (10):1345–1359.

F Pedregosa, G Varoquaux, A Gramfort, V Michel, B Thirion, O Grisel, M Blondel, P Prettenhofer, R Weiss, V Dubourg, J Vanderplas, A Passos, D Cournapeau, M Brucher, M Perrot, and E Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1-2):1–12.

Deb Roy. 2005. Semiotic schemas: A framework for grounding language in action and perception. *Artificial Intelligence*, 167(1-2):170–205, sep.

Cornelia Schlagenhauf, Dominik Bauer, Kai-Hung Chang, Jonathan P King, Daniele Moro, Stelian Coros, and Nancy Pollard. 2018. Control of tendon-driven soft foam robot hands. In *2018 IEEE RAS International Conference on Humanoid Robots*.

David Schlangen, Sina Zarrieß, and Casey Kennington. 2016. Resolving References to Objects in Photographs using the Words-As-Classifiers Model. In *Acl*, pages 1213–1223.

Lanbo She and Joyce Y Chai. 2016. Incremental Acquisition of Verb Hypothesis Space towards Physical World Interaction. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 108–117.

Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. sep.

Serge Thill, Sebastian Padó, and Tom Ziemke. 2014. On the importance of a rich embodiment in the grounding of concepts: Perspectives from embodied cognitive science and computational linguistics. *Topics in Cognitive Science*, 6(3):545–558, jul.

Jesse Thomason, Jivko Sinapov, Maxwell Svetlik, Peter Stone, and Raymond J Mooney. 2016. Learning Multi-Modal Grounded Linguistic Semantics by Playing " I Spy ". In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*.

Matthew R Walter, Sachithra Hemachandra, Bianca Homberg, Stefanie Tellex, and Seth Teller. 2014. A framework for learning semantic maps from grounded natural language descriptions. *The International Journal of Robotics Research*, 33(9):1167–1190.

Sina Zarrieß and David Schlangen. 2016. Easy Things First: Installments Improve Referring Expression Generation for Objects in Photographs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.