SE367 INTRODUCTION TO COGNITIVE SCIENCE

A Grounded Framework for Gestures and its Applications

Debidatta Dwibedi

April 18, 2013

ABSTRACT

If we are to imagine a world in the near future where humans and robots work in tandem to solve everyday problems, we need to prepare for the use of gestures as an effective mode of communication. Till now the use of gestures has been limited to very specific domains like direction giving embodied cognition agents(ECA). Furthermore, such systems are trained specifically for their domain of use. Both its knowledge and the gestures it produces are fed to the ECA beforehand. Due to recent developments in grounding language in perception it is possible to visualize a system that not only learns gestures on the go but also the words associated with it. Doing so for an extended period of time would lead to the system learning a vocabulary of gestures and words/phrases associated with them. This knowledge gained can be used to then produce gestures to improve communication between humans and robots.

INTRODUCTION

Social robots are on the rise. A lot of them have human-like appearances but not one is free from the curse of the uncanny valley. Apart from the candour of emotions presented by these robots, multi-modal interaction with humans can go a long way in overcoming the aforementioned barrier. In the recent years, major progress has been made in the interaction between humans and artificial agents using both natural language and speech as input. It is but inevitable that the next mode of communication between humans and robots would be through the use of gestures and human-like movements.

A theme that seems to be gaining momentum is that of lifelong learning[Thrun '95]. In this paradigm, systems are armed with powerful machine learning algorithms but they are not trained with any data. Hence, any data to these systems is unseen data. They, much like humans, keep learning throughout their lifespan. I want to keep true to this theme as much as possible throughout this project as I believe that systems that are able to make associations on their own are the closest we can come to simulate the human mind[Searle '68]. So, in this project I aim to develop a lifelong gesture learning framework.

Humans use gestures and diagrams inadvertently and excessively in two situations - while giving out route directions and while explaining to someone how to assemble or use an object. Words, by themselves, are meaningless symbols and it is these gestures and drawings that help transfer complicated thoughts about objects and actions because of visual similarity. This redundancy of information is necessary to both the person who needs the explanation as well as the person who is explaining. The other advantage that gestures provide is that they are a more universal means of communication.

This project aims at associating gestures with language. This framework can then be used to generate gestures helpful in route descriptions and assembly instructions. These domains are chosen for the sake of demonstration. The framework, however, is domain independent and can be trained to produce gestures in any context.

RELATED WORK

Researchers are looking at ways to integrate speech and gesture production in humanoid robots and study how multimodality affects a robot's interaction with humans[Kopp '08]. Experiments have also been carried out regarding the design aspects of a robot's utterance, gesture and timing[Okuno '09].

There are three kind of gestures: 1) deictic gestures that point to or indicate things in the environment; 2) iconic gestures that resemble what they are meant to convey and 3) beat gestures that are used in conjunction with speech and emphasize certain words or phrases[Tversky '09]. While the importance of deictic gestures has been well established and modelled in a direction giving situation[Striegnitz '05], the importance of iconic gestures needs to be studied more. They play a minor role in the route learning task because one needs iconic representation only for landmarks that are unknown to the asker. Many a time it is possible for the direction giver to provide the correct path without referring to any landmarks. However, in case of the parts assembly scenario iconic gestures are key. People usually represent the size, the shape and the function of various parts using iconic gestures.

Gesture recogntion has been a classic problem in computer vision. One of the most common approaches to it has been that of HMM[Chen '03]. However, the process of training an HMM requires determining states and transitions and probabilities associated with them. Recent approaches that use dynamic time warping point us towards a one-shot gesture learning paradigm using which requires no dataset of seen gestures and no training on that dataset.

Robots should be able to pick up new gestures from new interactions and be able to use this newly learnt gesture later on. Matuszek et al[Matuszek '12] have successfully associated physical attributes with language. Similarly, this project aims at grounding words associated with actions and objects with their related gestures. Nayak et al[Nayak '12] propose method to ground language by using association measures which minimalizes the amount of training data required.

Integrating some of these ideas, would automate the process of an agent learning gestures and their associated words without any further human supervision.

Lifelong Gesture Learning Framework

The framework for a lifelong gesture learning system can be outlined as follows:

1. Human describes and performs gestures.
2. If gesture is in database, increase vocabulary by discovering synonyms or associted words of that gesture.
3. If gestures is not present, record new gesture and try to discover the word associated with that gesture.

Each aspect of the learning system will now be described in detail.

Recording and Recognizing Gestures

I used a Kinect to record the gestures being performed. Microsoft Kinect SDK is very good at tracking the human skeleton. It was used to keep track of the coordinates of the hands of the human relative to his hip center. These coordinates at different times were used to represent each gesture. Hence, for each gesture we have a $6 \times N$ where N is the number of frames the human took to perform the gesture.

Now another human performs a query gesture. It is represented in a similar manner in a $6 \times M$ where M is the number of frames the second human took to perform the gesture. The problem reduces to finding the similarity between these two curves in time both of which may represent the same gesture but be of different durations in time.

We used the Frechet distance between these two vectors as a similarity metric to recognize gestures. It provides a one-shot gesture learning technique that employs dynamic programming two find similarity between two time series. Intuitively, the Frechet distance between two curves is the minimum length of a leash required to connect a dog and its owner, constrained on two separate paths, as they walk without backtracking along their respective curves from one endpoint to the other. Query gestures with their Frechet distance beyond a certain threshold were treated as new gestures and added in the database.

Associating Words with Gestures

In the proposed framework, I have assumed the availability of a good speech to text converter. Using temporal coherence, the system should be able to match gesture produced with the utterances at that time. However as most of my test subjects were local residents and there is no reliable speech to text converter for Indian accents, I decided to collect my test data as input text. This is a major bottleneck in this framework.

Two datasets were collected. One is a set of route descriptions given by 36 different IIT Kanpur students. It also contained one description from Google Maps. The students were aksed to describe in their natural language how to get to particular locations on campus. The second dataset was a set of instructions on how to asemble a TV stand provided by Ikea online.

The natural language system does stemming after removing stopwords. Every sentence has an associated gesture that was performed as that sentence was spoken. The system searches for non-trivial frequent words and then non-trivial frequent bigrams in the dataset it has learned. These words become candidates for words which might refer to the gestures. The problem here is to find which words might actually be referring to the gesture. For doing this the following association measure was used

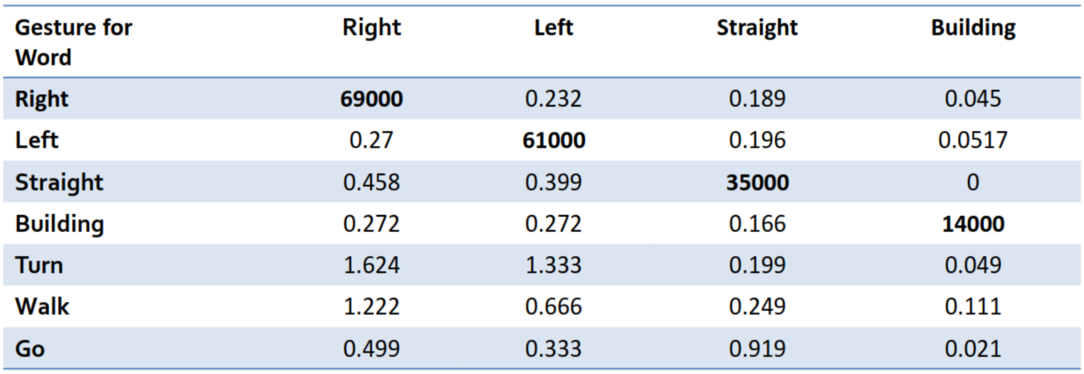
$$A(G, W) = \frac{P(G|W)}{P(\overline{G}|W)}$$

where G represents a recorded gesture, W represents a candidate word and $\overline{G}$ represents all other gestures. Each gesture in the database is assigned the word for which their association measure is the maximum.

This association measure though effective is still very simple and to an extent naive. Denominator may become equal to 0. Hence, a very small constant needs to be added. But doing so makes the measure depend on frequency of word which makes it flawed. Chi-square distributions give a good measure of co-occurrence and can be used to find these associations.

Words and phrases discovered in route descrption dataset: right, left , turn, straight, hall, road, walk, building ,('take', 'right'), ('take', 'left'), ('go', 'straight'), ('turn', 'left'),('turn', 'right'), ('right', 'turn')

Words and phrases discovered in assembly descrption dataset: shelf, frame, glass, place, top, bottom, bolts , ('allen', 'wrench'), ('shelf', 'frame'), ('bottom', 'shelf') , ('glass', 'shelf'), ('top', 'shelf')



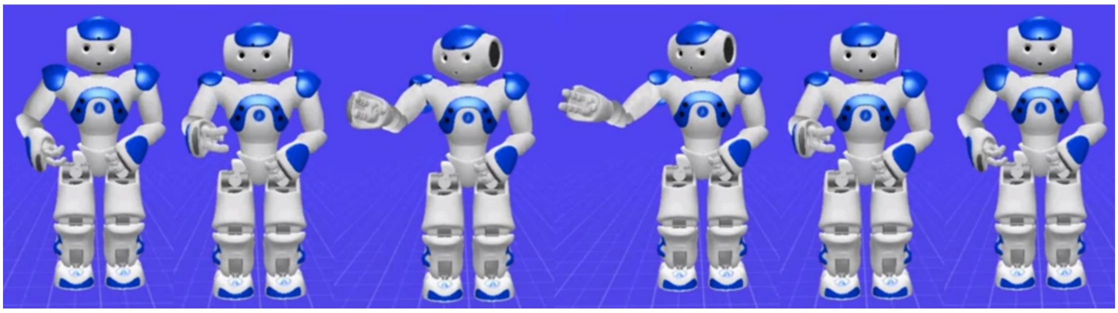
Gesture for Word	Right	Left	Straight	Building
Right	**69000**	0.232	0.189	0.045
Left	0.27	**61000**	0.196	0.0517
Straight	0.458	0.399	**35000**	0
Building	0.272	0.272	0.166	**14000**
Turn	1.624	1.333	0.199	0.049
Walk	1.222	0.666	0.249	0.111
Go	0.499	0.333	0.919	0.021

Association Measures for Gestures and Words

The extremely high association measures between the correct gesture and word is mostly because we had almost perfect information because of the description being present with us in form of text. This might not be the case when the gestures are associated automatically from speech. However, because the measure of association between the correct pair of gesture and word is so very high we can expect that the system should be robust to those impending errors.

TRANSFERRING GESTURES AND WORDS TO ECA

A simulation of a Nao robot was chosen as the ECA that will finally perform the gestures. The Nao robot can easily take a series of joint angles as input and perform the given motion. However, the recorded gestures are in cartesian coordinates. Hence, those cartesian coordinates were converted into joint angles of the robot using inverse kinematic relations. The simulation was then run on Choregraph. This software also allows one to directly upload the gesture to an actual Nao robot. Nao also has the ability to convert text to speech. Hence, the system can now read instructions from Google Maps while performing the corresponding gestures.



Nao performing "Take right turn"

Another advantage of simple transferring these gestures to the robot is that there is no need for a gesture giving system to now have hand-coded gestures which are usually found by employing a hit and trial process on the joint angles of the ECA to get a natual looking motion. Since the gesture is transferred from a human it will look natural and can be transferred to the robot without hassle. Currently, the movement is slightly jerky and needs to be smoothed.

Future Work

- An important part of the system is the module that parses knowledge from systems like Google Maps to find what are the gestures that can be performed.
- Integrate all the modules into one lifelong gesture learning system.
- Use of gestures in a collaborative setting. Utterances and gestures can become very complicated when a robot is actually helping out a human in action. It poses a very interesting problem.
- Use speech input to find association between gestures and words. Speech brings a certain sense of ambiguity and proabability to the system. However since it is also a datastream in time it can easilty be correlated temporally with the gestures being performed.
- Carry out usability studies
  - No human-robot interaction system is complete without carrying out usability studies to show how comfortable humans are interacting with the system and how effective it is in communicating with humans.
  - Impact of orientation of robot, timing of utterances and gestures etc. on comprehension
  - Provide gestures and descriptions to humans and see effectiveness of such a system

CONCLUSION

This project tried to explore the lifelong learning paradigm with respect to learning gestures. Although, there have been certain assumptions taken and certain drawbacks in the system, most of them revolve around integrating different modules which given time can be overcome.

ACKNOWLEDGEMENTS

I would like to thank Shashank Sonkar for helping me out with one-shot gesture learning and Nehchal Jindal for providing me with the code for converting cartesian coordinates into joint angles of the robot. I would also like to acknowledge Prof. Mukerjee for all the guidance and support he has provided throughout this semester.

REFERENCES

[1] Tversky, Barbara, et al. *"Explanations in gesture, diagram, and word."* Spatial Language and Dialogue, Coventry, KR, Tenbrink, T., and Bateman, J.(Eds.), Oxford University Press, Oxford (2009).

[2] Kopp, Stefan, Kirsten Bergmann, and Ipke Wachsmuth. *"Multimodal communication from multimodal thinking - towards an integrated model of speech and gesture production."* International Journal of Semantic Computing 2.01 (2008): 115-136.

[3] Okuno, Yusuke, et al. *"Providing route directions: design of robot's utterance, gesture, and timing."* Human-Robot Interaction (HRI), 2009 4th ACM/IEEE International Conference on. IEEE, 2009.

[4] Striegnitz, Kristina, et al. *"Knowledge representation for generating locating gestures in route directions."* Proceedings of Workshop on Spatial Language and Dialogue (5th Workshop on Language and Space). 2005.

[5] Matuszek, Cynthia, et al. *"A Joint Model of Language and Perception for Grounded Attribute Learning."* arXiv preprint arXiv:1206.6423 (2012).

[6]Sushobhan Nayak and Amitabha Mukerjee,*" Grounded Language Acquisition: A Minimal Commitment Approach"* COLING 2012: 24th International Conference on Computational Linguistics Mumbai, India, December 8-15, 2012.

[7]Thrun, Sebastian, and Tom M. Mitchell. *"Lifelong robot learning."* Robotics and autonomous systems 15.1 (1995): 25-46.

[8]Chen, Feng-Sheng, Chih-Ming Fu, and Chung-Lin Huang. *"Hand gesture recognition using a real-time tracking method and hidden Markov models."* Image and Vision Computing 21.8 (2003): 745-758.