SIMULATING IMITATIVE LEARNING IN A HUMANOID ROBOT FOR THE PURPOSE OF LANGUAGE ACQUISITION

By

Annlin Sheih

Senior Thesis in Computer Engineering

University of Illinois at Urbana-Champaign

Advisor: Stephen Levinson

May 2016

Abstract

Humans are born with an innate mechanism to recognize faces. Infants within weeks after birth are able to mimic facial gestures—an early but significant milestone towards adulthood. These skills are vital for interacting with their environment in order to develop their language from babbling to first word forms. Adapting the abilities from this period of infancy towards humanoid robots can produce a testable model of language acquisition. Learning a language requires more than just aural observation. By developing the visual component, the robot can observe features such as lips, teeth, and tongue in order to learn correspondence between speech and motor functions.

This thesis demonstrates progress toward the goal of developing a live lip-reading system. The first part describes experimentation with standard computer vision methods for human facial recognition, with emphasis on the mouth and jaw regions of the face. The next part shows methods to segment the mouth and discover salient patterns of lip configuration. The final challenge for this thesis is to implement this system live on the iCub, the humanoid robot of the Language Acquisition and Robotics Group, and have it respond in a way that demonstrates its interpretation of the lip movements.

Although the face and mouth can be consistently located within a given image frame, the challenge encountered was in segmenting the mouth into relevant features. Distinct attributes such as teeth and tongue were barely present or not visible in the image. This limited the analysis to mainly lip configurations. However, the analysis could still pinpoint distinct moments where teeth or tongue were showing.

Subject Keywords: imitation learning; computer vision; humanoid robot

Acknowledgments

I want to thank Professor Stephen Levinson, who has been a wonderful advisor and mentor throughout this research project. I also want to thank Ph.D. student Jacob Bryan for his guidance and help with the thesis experiments. Finally, I am grateful for the continuous support from the rest of the Language Acquisition and Robotics Group.

Contents

1. Introduction

The research for this thesis takes into account principles the Language Acquisition and Robotics Group hold to be true for the development of cognition. Firstly, a mind cannot be disembodied and must interact with the real world. Secondly, memory is associative, so learning occurs by finding correlation between different modes of input. This type of learning can occur with help from a complex sensory-motor system, which is necessary to simulate human cognition.

If a robot can develop a semantic and spatial model of the world based on these statements, the robot can share a similar view of the world and experience the world in a way similar to humans. Through interacting with humans, a robot can learn language and meaning through associating multiple sensory inputs with each other. By doing so, robots can eventually form a foundation for higher learning.

This thesis focuses on how a robot can acquire meaningful language through simulating the learning models of infants. Although the learning capabilities of infants are well-defined, the ways in which infants learn are still under investigation. However, imitation learning holds promise to be a key player in infant cognitive development. Modeling this approach for the purpose of language acquisition can potentially reduce the search space of language acquisition and speed up the learning process for robots.

2. Literature Review

Motivation for this thesis stems from work in the field of developmental psychology. This section evaluates existing psychological models of learning as an infant and computational models for language acquisition in humanoid robots.

2.1. Imitative Learning and Language Acquisition

Infants are not born with adult skills, so what allows them to develop these abilities as they grow? Sources of behavioral change that an infant experiences include maturation of the sensory, motor, and cognitive system, trial and error learning, independent invention and discovery, and imitative learning. This thesis focuses on the work done by Andrew N. Meltzoff and M. Keith Moore in imitative learning specifically. They conducted studies to show that neonates—infants that are just a few weeks old—are able to imitate facial gestures and other movements (Figure 1). Thus, imitation is not a skill that infants acquire over months of postnatal development; it is innate and thus the starting point for psychological development [1].

Babies are primed to learn from others. When they watch adults perform certain actions, babies' brains are activated as they try to understand those actions in the context of their own bodies. This mapping helps them make sense of the movements they witness and gives them the information needed to replicate the movement. Even if the action is not done perfectly, the infant can infer the correct action over time.

Figure 1 Meltzoff and Moore study, demonstrating tongue protrusion, mouth opening, and lip protrusion.

This concept can be applied towards language acquisition. As an infant watches an adult face and imitates their mouth shape, they start to understand the different movements a mouth can make. Gradually, they will make the connection of movement to sound coming out from the mouth. Relating the patterns of sounds to patterns of mouth movements allows babies to learn the first rudimentary aspects of language. This thesis will take these concepts into account.

2.2 Language Acquisition with the iCub

Although there has been significant growth in humanoid technology, there is still a gap in the robotic understanding of the cognitive needs for machine intelligence as well as understanding how the human brain functions and creates a cognitive being [2]. The RobotCub project is an ongoing initiative dedicated to solving these problems and creation of the iCub as a platform for studies in human and robotic cognition (Figure 2).

Imitative learning for robots poses complex problems such as how does the robot know when to imitate and what to imitate, and how to map the observed actions into a significant behavioral response. For the purpose of this experiment, we will have the iCub humanoid robot echo back what it observes from studying a human face. By echoing its observations of the mouth, the iCub can show that it can take in the pertinent information from the face and synthesize it with audio information to develop their language skills further.

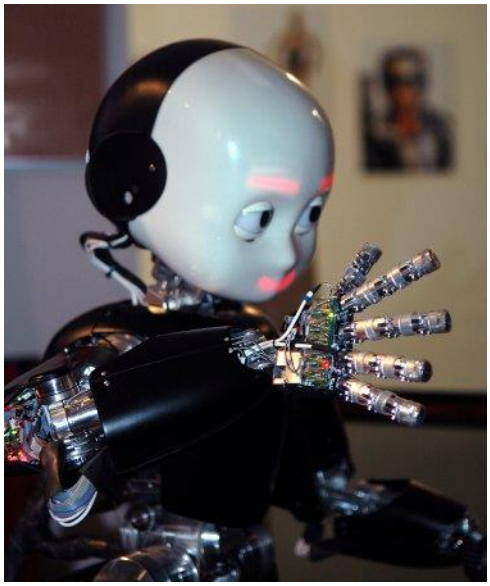

Figure 2 iCub expressing its ability to interact with its environment

3. Description of Research Results

3.1 Choice of Data

This thesis uses the Talking Face Data Set from Timothy Cootes, constructed as part of an experiment to model face behavior in natural conversation. This data set features 5000 frames of an adult male candidly expressing himself in different ways, ranging from total seriousness to engaging in laughter. The range of expressions the subject shows is a good sample of what an infant would experience and making sense of during its first few weeks. It also allows us to simulate Meltzoff and Moore's experiment with the iCub.

3.2 Feature Detection

The first step for imitative learning to occur is to find the face and its relevant features. Face and mouth detection was performed with OpenCV using Haar feature-based cascade classifiers originally proposed by Paul Viola and Michael Jones [3]. This method first trains the classifier with thousands of features found on a human face. Then, the features are grouped into different stages of classifiers and applied one by one to determine the face region. Thus, each frame approximates the face area with a bounding box. Each frame also approximates the mouth area with another bounding box.

After generating results for the Talking Face Data Set for each frame, several frames were hand-picked to illustrate different mouth configurations a subject can make. This search was based on the Meltzoff and Moore study, where they tested mouth opening, lip protrusion, and tongue protrusion (Figure 3).

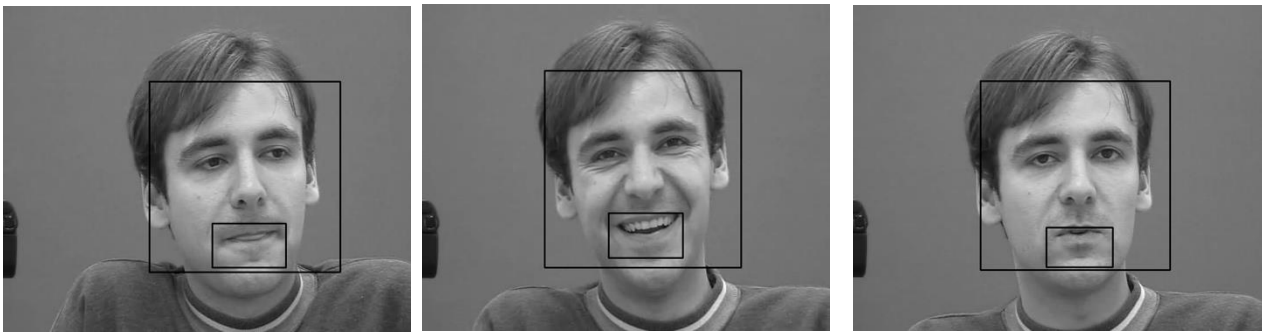

Figure 3 Tongue protrusion, mouth opening, and lip protrusion to match the configurations from Meltzoff and Moore's experiment.

At this point, different mouth configurations need to be tested to determine which one of them has the most distinct pattern. In order to do this, we must establish a mapping of mouth features. By doing so, we will have a way to maintain consistency in our bookkeeping and verify the correctness of our classification later on.

3.3 Mouth Segmentation

Lip tracking is an ongoing computer vision problem. Although there is a plethora of methods that have been implemented, the quality and quantity of information in each method is varied. The method used in this thesis is facial landmark detection, specifically using a coarse-to-fine search strategy on a Deformable Parts Model structured classifier of landmarks [4]. Landmark detection helps with face alignment, and is useful in scenes where face position cannot always be predicted or may be occluded.

The Talking Face data set includes annotations for face points generated with an Active Appearance Model (AAM) (Figure 4) [5]. Facial landmark detection generates similar feature points for the face, especially for the mouth and jawline regions. This method was implemented using the CLandmark library, which allows us to track face points either from static images or

from a live feed. The bounding boxes approximating face area earlier served as inputs to correctly align feature points to the face. This will be an important feature for the iCub to have. In order to perform lip reading we must maximize the amount of information so the iCub can develop its understanding of mouth shapes.

Figure 4 Top: Annotated points generated by AAM from given data set. Bottom: Annotated points produced from C2F-DPM method. Both methods produce similar points for mouth area and jawline region.

3.4 Verification

A calculation of coordinate variance was performed for each of the facial recognition models to compare their likeness. Shown below are the variances calculated for a single image frame. As shown, feature points M1 – M19 (mouth feature points) overall have less variance than feature points J1 – J13 (jaw feature points).

Table 1 Comparison of Feature Point Variances for a Single Image Frame

ID	AAM Coordinates	C2F-DPM Coordinates	Variance
M1	(338.418, 389.367)	(339.753, 391.975)	8.583889
M2	(360.296, 375.502)	(364.519, 380.216)	40.055525
M3	(387.264, 369.536)	(384.973, 371.673)	9.81545
M4	(399.558, 371.119)	(400.010, 373.891)	7.888288
M5	(408.819, 366.484)	(418.519, 368.144)	96.8456
M6	(433.487, 362.855)	(440.234, 366.723)	60.483433
M7	(457.765, 368.007)	(460.844, 372.845)	32.886485
M8	(444.035, 396.016)	(439.020, 401.219)	52.221434
M9	(426.237, 410.134)	(421.772, 414.087)	35.562434
M10	(403.353, 418.189)	(402.843, 417.460)	0.791541
M11	(380.699, 417.099)	(378.762, 419.303)	8.609585
M12	(359.151, 408.658)	(360.098, 410.386)	3.882793
M13	(364.842, 405.056)	(346.851, 390.710)	529.483797
M14	(402.535, 406.210)	(387.075, 383.542)	752.849824
M15	(434.540, 394.140)	(401.271, 381.012)	1279.170745
M16	(429.469, 373.462)	(419.780, 375.265)	97.12753
M17	(400.142, 381.207)	(451.379, 374.531)	2669.799145
M18	(369.036, 384.965)	(416.464, 397.893)	2416.548368
M19	(401.560, 394.740)	(402.267, 400.423)	32.796338
J1	(265.870, 332.048)	(274.296, 366.911)	1286.426245
J2	(273.543, 369.929)	(287.699, 401.251)	1181.46002
J3	(282.871, 400.888)	(307.779, 431.952)	1585.38056
J4	(301.982, 439.638)	(322.708, 461.123)	891.172301
J5	(326.057, 468.956)	(350.571, 480.643)	737.522165
J6	(366.260, 497.352)	(381.907, 492.200)	271.371713
J7	(403.871, 505.293)	(418.238, 490.622)	421.64893
J8	(441.439, 496.405)	(450.523, 479.972)	352.562545
J9	(478.562, 455.430)	(471.242, 459.140)	67.3465
J10	(499.617, 410.710)	(489.174, 436.356)	766.773565
J11	(508.862, 357.568)	(502.639, 402.125)	2024.051978
J12	(510.738, 326.282)	(505.799, 364.834)	1510.650425
J13	(509.163, 286.018)	(510.905, 324.747)	1502.970

3.5 Implementation on iCub

As the final step, this system will be ported onto the iCub to show that it can recognize and imitate the mouth movements it sees. The iCub's eyes contain cameras, which will enable it to perform face tracking on a given human subject. He will be able to process the visual information live through the code implemented in this thesis. The iCub's face is composed of arrays of LEDs for eyebrows and mouth, which it will use to express its understanding of the human subject. Mouth LEDs will light up according to the human subject's lips. Because the number of configurations of LEDs the iCub has is limited, basic mouth configurations will be tested on it to demonstrate proof of concept (Figure 5).

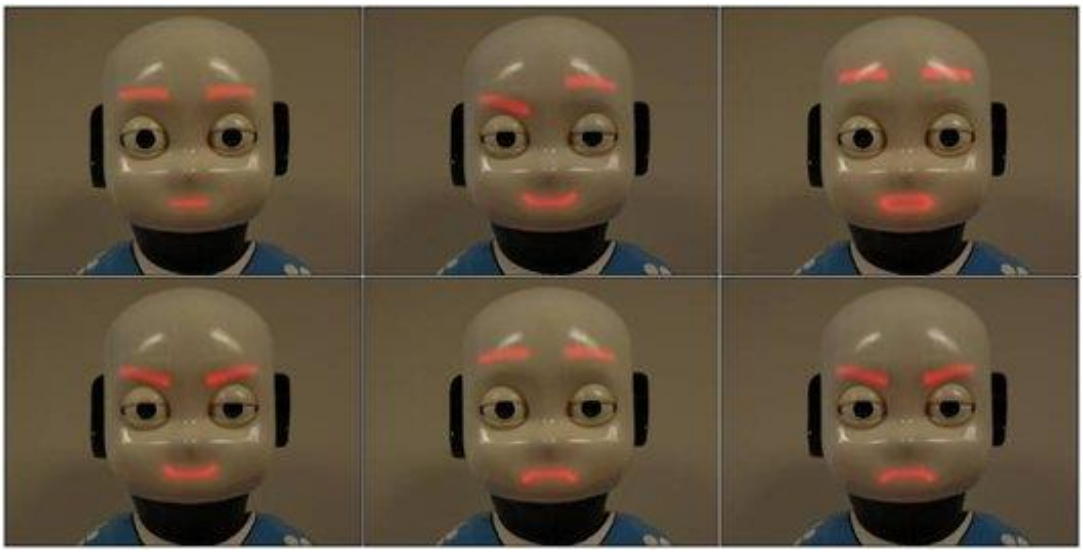

Figure 5 Sampling of mouth and eyebrow configurations iCub can make.

4. Conclusion

From my experiments, I learned how there are many problems in facial recognition that have yet to be solved, especially to understand visual patterns in language. The implemented system in this thesis can accurately detect the face and locate important points on the mouth and jawline. However, there is still room for improvement in constructing an accurate model of the mouth. Significant factors of speech come from the tongue and vocal tract for instance—places where vision is limited. This is the reason why the Language Acquisition and Robotics Group at UIUC believes that language acquisition is multi-modal [6].

Like human infants, the iCub learns through interacting with its environment. Sensory-motor function is essential; perceptual and cognitive functions are intertwined [7]. In the near future, the work from this thesis will be integrated with already-developed modules for the iCub's sensory input processing, speech recognition and generation, navigation, and associative learning, further improving on the iCub's ability to learn autonomously.

References

[1] A. N. Meltzoff and M. K. Moore, “Imitation of facial and manual gestures by human neonates,” *Science,* vol. 198, pp. 75-78, 1977.

[2] N.Tsagarakis, G. Metta, G. Sandini, D. Vernon, R. Beira, J. Santos-Victor, et al., “iCub: The Design and Realization of an Open Humanoid Platform for Cognitive and Neuroscience Research,” *International Journal of Advanced Robotics*, vol. 21, no. 10, pp. 1151-1175, 2007.

[3] P. Viola and M. Jones, “Rapid Object Detection Using a Boosted Cascade of Simple Features,” *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001.

[4] M. Uricár, V. Franc, and V. Hlavác, “Facial Landmark Tracking by Tree-based Deformable Part Model Based Detector,” Proceedings of the IEEE International Conference on Computer Vision Workshops, 2015.

[5] T.F. Cootes, *Talking Face,* http://personalpages.manchester.ac.uk/staff/timothy.f.cootes/data/talking_face/talking_face.html

[6] S. E. Levinson, L. Niehaus, L. Majure, A. Silver and L. Wendt, “Can a Robot Learn Language as a Child Does?” In 2012 AAAI Spring Symposium Series. 2012

[7] S.E. Levinson, K. Squire, R.S. Lin, and M. McClain, “Automatic Language Acquisition by an Autonomous Robot,” AAAI Spring Symposium on Developmental Robotics, March 21-23, 2005.