Coherence in One-Shot Gesture Recognition for Human-Robot Interaction

Maria E. Cabrera
Purdue University
West Lafayette, IN
cabrerm@purdue.edu

Richard M. Voyles
Purdue University
West Lafayette, IN
rvoyles@purdue.edu

Juan P. Wachs*
Purdue University
West Lafayette, IN
jpwachs@purdue.edu

ABSTRACT

An experiment was conducted where a robotic platform performs artificially generated gestures and both trained classifiers and human participants recognize. Classification accuracy is evaluated through a new metric of coherence in gesture recognition between humans and robots. Experimental results showed an average recognition performance of 89.2% for the trained classifiers and 92.5% for the participants. Coherence in one-shot gesture recognition was determined to be $\gamma = 93.8\%$. This new metric provides a quantifier for validating how realistic the robotic generated gestures are.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; **Gestural input**; *Interaction paradigms*;

KEYWORDS

Robotics, Gesture Recognition, One-Shot Learning

ACM Reference Format:
Maria E. Cabrera, Richard M. Voyles, and Juan P. Wachs. 2018. Coherence in One-Shot Gesture Recognition for Human-Robot Interaction. In *HRI '18 Companion: 2018 ACM/IEEE International Conference on Human-Robot Interaction Companion, March 5–8, 2018, Chicago, IL, USA.* ACM/IEEE, Chicago, IL, USA, 2 pages. https://doi.org/10.1145/3173386.3176977

1 INTRODUCTION

The topic of mutual grounding is highly relevant to communication between two humans, between a human and a machine or between two machines, but humans have the unique ability to quickly adjust their context, often from only one example. This is called *one-shot learning* [3, 4] and it involves developing machines capable of recognizing gestures from a single observation. This research topic is difficult, not only because of the lack of training data, but also because the bulk of machine learning algorithms are focused on N-shot problems, where N is often very large.

Most existing research has tried to maximize the accuracy of recognition of one-shot learning. We propose a new metric that chooses not to maximize the raw accuracy of the recognition of a gesture, but to maximize coherence with the recognition characteristics of other humans. In other words, we choose to maximize agreement with humans, both in when humans classify correctly and when humans mis-classify.

By including the human aspect within the framework – through virtual generation of $N-1$ additional examples – the human kinematic and psycho-physical attributes of the gesture production process are used to support recognition. Motion features within gestures were found to be correlated with neural signals associated to activation of motor and visual cortices [1].

The main focus of this paper is determining just how "realistic" the produced synthetic gestures are in the scope of Human-Robot Interaction (HRI). A robotic platform is used to perform these synthetic gestures in two different scenarios and determine the coherency between them.

2 METHODOLOGY

The overview of the method to achieve one-shot gesture recognition from a single example of each gesture class is shown in Figure 1. Using the hands' trajectories from Kinect's skeleton data, the "gist of the gesture" is extracted; it contains salient motion points where abrupt changes in speed and orientation occur. A human-centered approach leveraging spatial variability allows to artificially enlarge the data set to train different state-of-the-art classifiers capturing significant variability while maintaining a model of the fundamental structure of the gesture to account for its stochastic process [2].

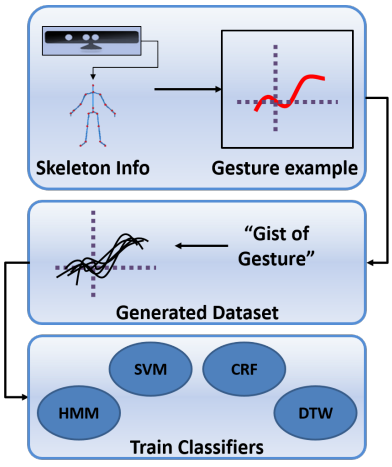

Figure 1: Overview of one-shot gesture recognition framework.

*Corresponding author

HRI '18 Companion, March 5–8, 2018, Chicago, IL, USA

ACM ISBN 978-1-4503-5615-2/18/03.
https://doi.org/10.1145/3173386.3176977

This idiosyncratic approach is tested by training four different classification methods, namely Hidden Markov Models (HMM), Support Vector Machines (SVM), Conditional Random Fields (CRF) and Dynamic Time Warping (DTW).

The dual-arm robotic platform Baxter, from Rethink Robotics, was used to perform artificially generated gestures. To determine coherence in one-shot gesture recognition, two scenarios (shown in Figure 2) were considered: in the first scenario (MH), humans recognize the gestures. In the second scenario (MH), the gestures are recognized using four different classification algorithms.

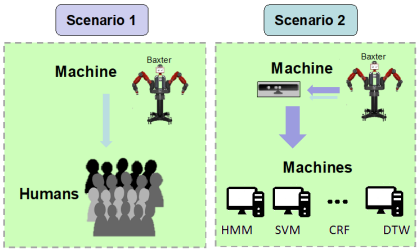

Figure 2: Scenarios used to determine recognition coherence between humans and machines.

The selected data set to test the proposed framework is the Microsoft Research MSRC-12 [5]. The number of gesture classes in the lexicon was reduced to 8 to avoid gesture classes that are not performed with the upper limbs.

A metric is proposed to measure the level of coherence between the recognition accuracy obtained by trained classifiers, and the accuracy found when humans observe the robot perform artificial gestures.

Coherency (γ) is defined in Eq. 1 as the intersection between the sets of AIx for both humans and machines. The Agreement Index (AIx) is the median in the set of all Boolean values for recognition, whether each agent ($AIx_{machine}$ and AIx_{human}) correctly recognized each gesture or not. The value $\|AIx_{human}\|$ counts all elements in the set. The higher the coherence, the better the mimicry of human perception and gesture execution and recognition.

$$\gamma = \frac{AIx_{machine} \cap AIx_{human}}{\|AIx_{human}\|} \times 100\% \quad (1)$$

Ten participants were asked to watch a video of a person performing one example of each gesture class in the MSRC-12 lexicon with the gesture's respective label. Next, each participant observed Baxter perform two instances of each gesture class (16 total) in random order; finally, they were asked to assign a label to each. Once the experiment was concluded, participants filled out a questionnaire inquiring whether the characteristics of each gesture were maintained when the robot performed the gestures.

3 RESULTS

Recognition accuracies were found for each scenario using 20 lexicon sets and then used to determine coherence γ. The recognition accuracies found in both scenarios are summarized in Figure 3.

The recognition accuracy of the participants on the testing dataset was 92.5%. The gesture 'Shoot' showed the lowest recognition rate among the participants. One possible explanation has to

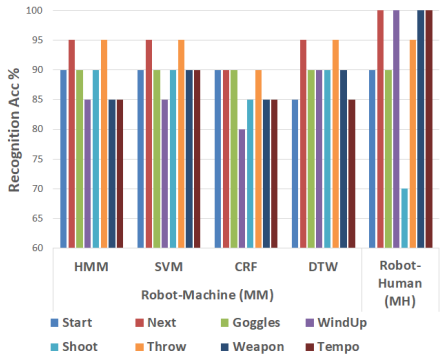

Figure 3: Recognition Accuracy (%) for different interaction scenarios: Robot-Human (MH) and Robot-Machine (MM)

do with the hand configuration that comes natural for humans to mimic a shooting gesture, which is very difficult to be reproduced with the robotic platform. Using the recognition results from the previous two scenarios, the metric of coherency was calculated. The AIx among users and machines were calculated for each gesture type and instance. The coherency found was $\gamma = 93.8\%$.

4 CONCLUSION

This paper explores gesture recognition and introduces a new metric of coherency to the problem of one-shot gesture recognition in HRI. The calculated coherency metric is our main indicator that the generated gestures capture human-like variations of gesture classes, affirming the desired mutual grounding. Experimental results provide an average recognition performance of 89.2% for the trained classifiers and 92.5% for the participants. Coherency in recognition was determined at 93.8% in average for all 20 lexicon sets performed by Baxter and recognized by classifiers (MM) and humans (MH).

Future work includes computing coherence in the context of other approaches for artificial gesture generation and its inherent use for gesture imitation.

ACKNOWLEDGMENTS

This work was supported by the Office of the Assistant Secretary of Defense for Health Affairs under Award No. W81XWH-14-1-0042. Opinions, interpretations, conclusions and recommendations are those of the authors and are not necessarily endorsed by the Department of Defense

REFERENCES

[1] Maria Cabrera, Keisha Novak, Daniel Foti, Richard Voyles, and Juan Wachs. 2017. What makes a gesture a gesture? Neural signatures involved in gesture recognition. *12th IEEE Int Conf on Automatic Face and Gesture Recognition* (2017).

[2] Maria Eugenia Cabrera and Juan Wachs. 2017. A Human-Centered approach to One-Shot Gesture Learning. *Frontiers in Robotics and AI* 4 (2017).

[3] Hugo Jair Escalante, Isabelle Guyon, Vassilis Athitsos, Pat Jangyodsuk, and Jun Wan. 2015. Principal motion components for one-shot gesture recognition. *Pattern Analysis and Applications* (2015), 1–16.

[4] Sean Ryan Fanello, Ilaria Gori, Giorgio Metta, and Francesca Odone. 2013. Keep it simple and sparse: Real-time action recognition. *The Journal of Machine Learning Research* 14, 1 (2013), 2617–2640.

[5] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. 2012. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1737–1746.