Recognition Coherency for Human-Robot Interaction in One-Shot Gesture Learning

Maria E. Cabrera
School of Industrial Engineering
Purdue University
West Lafayette, IN 47906
Email: cabrerm@purdue.edu

Richard M. Voyles
School of Engineering Technology
Purdue University
West Lafayette, IN 47906
Email: rvoyles@purdue.edu

Juan P. Wachs
School of Industrial Engineering
Purdue University
West Lafayette, IN 47906
Email: jpwachs@purdue.edu

***Abstract*—With the aim of achieving natural interactions between humans and robots, user's intentions may be expressed through spontaneous gesturing, which may have been seen only a few times or never before. Recognizing such gestures involves one shot gesture learning. The framework presented in this work focuses on learning the process that leads to gesture generation, rather than mining the gesture's associated features. This is achieved using kinematic and cognitive aspects of human interaction. These factors enable the artificial production of realistic gesture samples originated from a single observation, which in turn are used as training sets for four different state-of-the-art classifiers. Classification is evaluated in terms of coherency of gesture recognition between humans and robots. Two scenarios are considered to address the metric of coherency, one where a robotic platform performs gestures and the classifiers recognize them; the other, where the recognition is done by human participants. Performance is obtained through recognition percentages for all the agents. Then the proposed metric of coherency determines the level of agreement between these two conditions. Experimental results showed 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%. While this new metric is not directly comparable to raw accuracy or other pure performance-based standard metrics, it provides a quantifier for validating how realistic the robotic generated gestures are.**

I. Introduction

In the gesture recognition community, interest has grown recently in the problem of One-Shot Gesture Recognition [1, 2], which equates to recognizing gestures from a single observation. Recognizing gestures is difficult for two reasons: gestures are intrinsically imprecise, encompassing a great deal of variability, and the humans that perform them are also imprecise, injecting characteristics of their own preferences and biomechanics. When only one example is provided this task becomes even more challenging, increasing the risk of low generalization capabilities [3].

By including the human aspect within the framework, the kinematic and psycho-physical attributes of the gesture production process are used to support recognition. This approach presents a strategy to generate a dataset of realistic samples based on a single example. Using a single labeled example, multiple instances of the same class are generated synthetically, augmenting the dataset and enabling one-shot learning.

The recognition problem is framed by the idea of using the generation process for gestural instances rather than the instance itself. The proposed method captures significant variability while maintaining a model of the fundamental structure of the gesture to account for the stochastic process involved in the gesture production, associated with inherent non-linearity of human motor control [4]. We rely on global salient characteristics in a given gesture example that transcend variability due to human nature and are present in all examples of the same gesture class. These characteristics are referred as the gist of a gesture, and are used towards an artificial and realistic gesture generation process [5].

The main focus of this paper is determining just how realistic the produced synthetic gestures are in the scope of human-robot interaction. Literature regarding brain activity shows that there are similarities in the motor cortex responses when a human observes other humans, and robots alike, perform gestures [6]. A robotic platform is used to perform these synthetic gestures in two different scenarios and determine the coherency between them.

II. Background

A. Gesture Communication

Gestures are a basic form of communication between human beings. Young children use gestures to communicate before they learn how to talk [7]. Not only are the outcome and meaning of a gesture important, but also what gestures can tell us about the cognitive processes involved during gesture generation [8, 9]. Recent studies showed that gesturing plays a causal role in learning and that gesturing can promote learning [10].

Gestures offer a potential interface modality that includes control through symbolic commands, as for keyboards, and pointing attributes similar to those of the mouse, but in a more flexible, natural, and expressive form. Promoting forms of gesture recognition that are similar to the mechanisms existing in humans will allow a more natural communication than the existing ones. Thomason and Knepper [11] present a relevant example to the field of HRI and related to using spontaneous gestures as mean of communication with a robotic platform.

B. One-Shot Learning in Gesture Recognition

An important landmark in one-shot learning applied to gestures was the Microsoft initiative to start ChaLearn Looking at People Challenge in 2011. The yearly challenge employs computer vision and machine learning techniques to recognize human actions and features through multi-modal interaction. For the One-Shot Learning challenge, a vast data set of both development and validation batches, was used worldwide as training and testing data in the competition; the results were reported in Guyon et.al. with partial success [12].

A common theme in the proposed methods emphasized gesture representation as strictly machine learning and classification of observations regardless of the process involved in their generation [13, 14]. No relevance was mentioned towards the shape or characteristics of the human body performing the gestures. More recent methods are described by Escalante et al. [1], where a 2D map of motion energy is obtained per each pair of consecutive frames in a video and then used for recognition after applying Principal Component Analysis.

III. Methodology

This section presents details of the implementation of a method to achieve one-shot gesture recognition through the "gist of the gesture". An upper-limb gesture performed by a user is detected and recorded using a Kinect sensor. Using the skeleton data provided by the sensor, salient characteristics are extracted, which we refer to as the gist of the gesture, and are used to recreate new realistic observations that resemble the one provided by the user [5]. This process is repeated until a large data set of observations is generated. The motivation behind the gist of the gesture is to understand how humans gesture and what determines the forms of the gestures produced. Furthermore, multiple instances of the same gesture will share key common motion components among all instances, regardless of the variability associated with human performance [15].

A. One-Shot Learning Problem Definition

Let $\mathcal{L}$ describe a set or lexicon formed by N gesture classes, $\mathcal{G}_i$ where $\mathcal{L} = \{\mathcal{G}_1, \mathcal{G}_2, ..., \mathcal{G}_i, ..., \mathcal{G}_N\}$. Each gesture class is formed by the set of gestures instances $g_k^i \in \mathcal{G}_i$, where $k = 1, ..., M_i$. Where M_i is the number of observations of gesture class *i*. Each gesture observation is a concatenation of trajectory points in 3D (1) where *h* is the total number of points within that gesture observation.

$$g_k^i = \{(x_1, y_2, z_1), ..., (x_h, y_h, z_h)\} \tag{1}$$

In the case of one-shot learning, we rely on one observation as the basis to generate several others. Thus, g_k^i exists only for $k = 1$. We resort to create additional instances which could be used to later train and test a variety of classification algorithms. The new artificial instances result in an increase in dataset size from $k = 1$ to $k = M_i$. This parameter M_i is the desired number of instances of that class required for training. Equation (2) is applied to a single observation g_1^i, and is used to extract a set of inflection points labeled as x_q^i, where $q = 1, ..., l$ and $l < h$.

$$\begin{gathered}\tilde{G}_i = \left\{\mathbf{x_q^i} = (x_q, y_q, z_q) : \mathbf{x_q^i} \in g_k^i, q = 1, ..., l, l < h\right\} \\ \tilde{G}_i \in \tilde{G}_{\mathcal{L}}, i = 1, \ldots, N\end{gathered} \tag{2}$$

The set of inflection points, is a compact representation obtained using the function $\mathcal{M}$ (3) that maps the gesture dimension *h* to a reduced dimension *l* by extracting the salient points of a given gesture instance. The methods considered for salient point extraction range from manual annotation to automatic selection based on differentiation of the gesture trajectory.

The sets of inflection points, $\tilde{G}_i$ (3), will serve as the basis to create artificial gesture instances, $\hat{g}_k^i$, $\in \mathbb{R}^h$ for each $\mathcal{G}_i$. Then, artificial gesture examples for $\tilde{G}_i$ are generated through the function $\mathcal{A}$ (4), which maps from dimension *l* to gesture dimension *h*. Functions $\mathcal{M}$ and $\mathcal{A}$ are described further in [5].

$$\begin{gathered}\tilde{G}_i = \mathcal{M}(g_k^i), k = 1, i = 1, \ldots, N; \\ g_k^i \in \mathbb{R}^{3\times h}; \tilde{G}_i \in \mathbb{R}^{3\times l}; l < h\end{gathered} \tag{3}$$

$$\hat{g}_k^i = \mathcal{A}(\tilde{G}_i), k = 1, \ldots, M_i; i = 1, \ldots, N \tag{4}$$

The function Ψ (5) maps gesture instances to each gesture class using the artificial examples.

$$\Psi : \hat{g}_k^i \rightarrow \mathcal{G}_i \tag{5}$$

For future instances g^u the problem of one-shot learning gesture recognition is defined in (6) as:

$$\begin{gathered}\text{Max } \mathrm{Z} = \mathcal{W}\{\Psi(g^u), \mathcal{G}_i\} \\ \text{s.t. } i \leq N;\ i \in \mathbb{Z}^+;\ \mathcal{G}_i = \Psi(g_1^i);\ \Psi(g^u) \in \mathcal{L}\end{gathered} \tag{6}$$

Where, g^u are the unseen instances of an unknown class and $\mathcal{W}$ is the selected metric function, for instance accuracy or F-Score.

B. Implementation Details

The approach proposed in this paper is independent to a specific form of classification, Ψ. Furthermore, it is not conceived with a specific classification approach in mind. This idiosyncratic approach is tested by training four different classification methods, currently used in state-of-the-art N-shot gesture recognition approaches, namely Hidden Markov Models (HMM), Support Vector Machines (SVM), Conditional Random Fields (CRF) and Dynamic Time Warping (DTW), and adapt them to be used in one-shot gesture recognition.

The selected data set to test the proposed framework is the Microsoft Research MSRC-12 [16]. This data set consists of sequences of human movements, representing 12 different iconic and metaphoric gestures related to gaming commands and interacting with a media player.

A subset of this data set was selected. The number of gesture classes in the lexicon was reduced to 8. This reduction is to avoid gesture classes performing whole-body motions (like kicking or taking a bow) since the focus of this paper is on gestures performed with the upper limbs.

C. Generation of Artificial Trajectories

To generate artificial observations $\hat{g}_k^i$, the recorded location for each inflection point $\mathbf{x_q^i}$ where $q = 1, ..., l$ is used as the mean value μ for a mixture of Gaussians, while the quadrant information for each point of the hand's trajectory relative to the user's shoulder is used to estimate the variance. Each point in the sample trajectory is assigned to a quadrant with respect to the user's shoulder reference frame, as shown in Fig. 1.

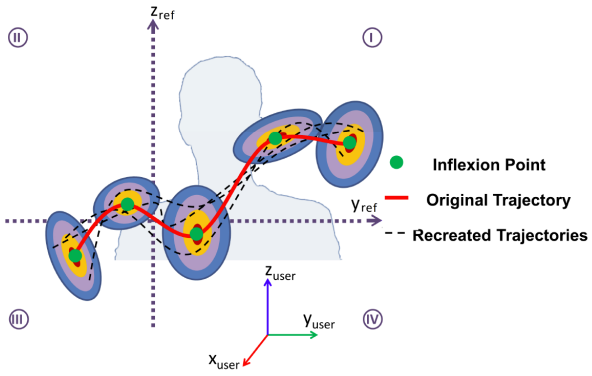

Fig. 1. Visual representation of the Artificial Trajectory Generation

Equation (3) is used as the function to draw samples from these GMM centered at each inflection point, and generate smooth interpolated trajectories in between them, using the curvature information extracted from the original example.

D. Coherency Metric

A metric is proposed to measure the level of coherence $\gamma(\cdot)$ (7) between the recognition accuracy obtained by the proposed method, and that found when humans observe the gestures. The higher the coherence, the better the mimicry of human perception, recognition and gesture execution. This metric of coherence is related to agreement indices for recognition of gestures by machines $AIx_{machine}$ and humans AIx_{human}.

$$\gamma = \frac{AIx_{machine} \cap AIx_{human}}{\|AIx_{human}\|} \times 100\% \qquad (7)$$

Coherency is defined 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 correctly recognized each gesture or not. The value $\|AIx_{human}\|$ refers to the count of all elements in the set. Given that both sets have the same number of elements it is indifferent to select the set from human or machines.

It is important to note that when computing the coherency, the intersection of incorrect identifications is said to exist when both the machine and human make any incorrect identification. It is not necessary for the machine and human to misidentify a particular gesture instance as the same class. In a case of confusion, both machine and human misidentify the gesture.

IV. Results

The results for two different scenarios used to recognize gestures performed by a robotic platform are presented in this section. The gestures are the result of an artificial generation process based on salient characteristics extracted from a single example of each gesture class as explained earlier in Section III. In the first scenario (MM), the gestures are recognized using four different classification algorithms. In the second scenario (MH), ten participants recognized the gestures performed by Baxter. Recognition accuracies were found for each scenario and then used to determine coherency. The recognition accuracies found in both scenarios are summarized in Table I.

TABLE I
Recognition Accuracy (%) for different interaction combinations: Robot-Human (MH) and Robot-Machine (MM)

Gesture	Robot-Machine (MM)					Robot-Human (MH)
	HMM	SVM	CRF	DTW	**ALL**	
Start	90	90	90	85	**88.8**	**90**
Next	95	95	90	95	**93.8**	**100**
Goggles	90	90	90	90	**90**	**90**
WindUp	85	85	80	90	**87.5**	**100**
Shoot	90	90	85	90	**88.8**	**70**
Throw	95	95	90	95	**93.8**	**95**
ChangeWeapon	85	90	85	90	**85**	**100**
Tempo	85	90	85	85	**86.3**	**100**
OVERALL	**89.4**	**90.6**	**86.9**	**90.0**	**89.2**	**92.5**

The results obtained with the classification algorithms are comparable with state-of-the-art results reported by Ellis et al. [17] reaching 88.7% accuracy, or by Ramirez-Corona et al. [18] with 91.82%. However, the limited number of samples used for testing (20 per gesture class), may have had a detrimental effect on the performance.

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 do with the hand position that comes natural for humans to mimic a shooting gesture, which is very difficult to be reproduced smoothly with the robotic platform. This supports the need for including the hand configuration during the salient point extraction process.

A. Coherency

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 = 93.8%.

Despite the "Shoot" gesture being the lowest accuracy (70%) for human recognition in the MH scenario, most of the participants could understand the performed gesture as indicated by the AIx_{human} resulting in high values (e.g. AIx_{human} = 1) for the different lexicon sets tested.

V. Conclusion

This paper explores gesture recognition in a general way relevant to Human-Robot Interaction; we also introduce a new metric of coherency to the problem of one-shot gesture recognition in HRI. An existing framework was used which focuses on the gesture generation process, using kinematic and cognitive characteristics of human interaction, to extract salient features of a gesture class from a single example, and use those to generate an enlarged artificial data set of realistic gesture samples. These artificial samples were validated using two different scenarios where a dual-arm robotic platform is used to execute the gesture trajectories: the first scenario involved the use of state-of-the-art classification methods to recognize the performed gestures, and the second relied on human recognition.

The calculated coherency metric is our main indicator that the generated gestures capture human-like variations of gesture classes. This metric may also enhance natural interactions between humans and robots, by providing some insight on the way mistakes are made. Even when it is related to recognition accuracy, it shows a different perspective of how the messages are conveyed between agents.

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 coherency in the context of other approaches for artificial gesture generation, with different dual-arm robotic platforms.

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 author and are not necessarily endorsed by the Department of Defense.

References

[1] H. J. Escalante, I. Guyon, V. Athitsos, P. Jangyodsuk, and J. Wan, "Principal motion components for one-shot gesture recognition," *Pattern Analysis and Applications*, pp. 1–16, May 2015.

[2] S. R. Fanello, I. Gori, G. Metta, and F. Odone, "Keep it simple and sparse: Real-time action recognition," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2617–2640, 2013.

[3] L. Fe-Fei, R. Fergus, and P. Perona, "A Bayesian approach to unsupervised one-shot learning of object categories," in *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pp. 1134–1141, IEEE, 2003.

[4] D. M. Wolpert, "Computational approaches to motor control," *Trends in cognitive sciences*, vol. 1, no. 6, pp. 209–216, 1997.

[5] M. E. Cabrera and J. P. Wachs, "Embodied gesture learning from one-shot," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*, pp. 1092–1097, IEEE, 2016.

[6] B. A. Urgen, M. Plank, H. Ishiguro, H. Poizner, and A. P. Saygin, "EEG theta and Mu oscillations during perception of human and robot actions," *Frontiers in neurorobotics*, vol. 7, p. 19, 2013.

[7] L. P. Acredolo and S. Goodwyn, *Baby signs: How to talk with your baby before your baby can talk*. Random House, 2000.

[8] S. Goldin-Meadow, "The role of gesture in communication and thinking," *Trends in cognitive sciences*, vol. 3, no. 11, pp. 419–429, 1999.

[9] D. McNeill, *Hand and mind: What gestures reveal about thought*. University of Chicago press, 1992.

[10] A. Segal, *Do Gestural Interfaces Promote Thinking? Embodied Interaction: Congruent Gestures and Direct Touch Promote Performance in Math.* ERIC, 2011.

[11] W. Thomason and R. Knepper, "Recognizing Unfamiliar Gestures for Human-Robot Interaction Through Zero-Shot Learning," in *International Symposium on Experimental Robotics*, pp. 841–852, Springer, Cham, 2016.

[12] I. Guyon, V. Athitsos, P. Jangyodsuk, B. Hamner, and H. J. Escalante, "Chalearn gesture challenge: Design and first results," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pp. 1–6, IEEE, 2012.

[13] J. Wan, Q. Ruan, W. Li, and S. Deng, "One-shot learning gesture recognition from RGB-D data using bag of features," *The Journal of Machine Learning Research*, vol. 14, no. 1, pp. 2549–2582, 2013.

[14] D. Wu, F. Zhu, and L. Shao, "One shot learning gesture recognition from RGBD images," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 7–12, June 2012.

[15] M. Cabrera, K. Novak, D. Foti, R. Voyles, and J. Wachs, "What makes a gesture a gesture? Neural signatures involved in gesture recognition," *12th IEEE International Conference on Automatic Face and Gesture Recognition*, May 2017.

[16] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 1737–1746, ACM, 2012.

[17] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. Laviola Jr, and R. Sukthankar, "Exploring the trade-off between accuracy and observational latency in action recognition," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 420–436, 2013.

[18] M. Ramrez-Corona, M. Osorio-Ramos, and E. F. Morales, "A non-temporal approach for gesture recognition using microsoft kinect," in *Iberoamerican Congress on Pattern Recognition*, pp. 318–325, Springer, 2013.