One-Shot Gesture Recognition: One Step Towards Adaptive Learning

Cabrera Maria E[1], Sanchez-Tamayo Natalia[2], Voyles Richard[1] and Wachs Juan P[1]
[1] Purdue University, West Lafayette, IN. USA
[2] Los Andes University, Bogota, Colombia

Abstract— **User's intentions may be expressed through spontaneous gesturing, which have been seen only a few times or never before. Recognizing such gestures involves one shot gesture learning. While most research has focused on the recognition of the gestures themselves, recently new approaches were proposed to deal with gesture perception and production as part of the recognition problem. The framework presented in this work focuses on learning the process that leads to gesture generation, rather than treating the gestures as the outcomes of a stochastic process only. This is achieved by leveraging 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 state-of-the-art classifiers. Classification performance is evaluated in terms of recognition accuracy and coherency; the latter being a novel metric that determines the level of agreement between humans and machines. Specifically, the referred machines are robots which perform artificially generated examples. Coherency in recognition was determined at 93.8%, corresponding to a recognition accuracy of 89.2% for the classifiers and 92.5% for human participants. A proof of concept was performed towards the expansion of the proposed one shot learning approach to adaptive learning, and the results are presented and the implications discussed.**

I. INTRODUCTION

The problem of recognizing gestures from a single observation is called One-Shot Gesture Recognition [1]. In the considered scenario a single training observation is available for each gesture to be recognized. The limited amount of information provided by that single observation makes this problem ill-posed [2]; achieving generalization in recognition becomes even more challenging without enough resources to mine information from. That is why pure machine learning approaches have not yet offered a significant result compared to state-of-the-art recognition based on multiple training examples. For example, HMM, SVM, PCA based techniques were applied to this challenge obtaining suboptimal performances as reported at the ChaLearn 2011 competition [3]. Therefore, one can conclude that currently those algorithms are not the most suitable to tackle the one-shot recognition problem. To overcome this limitation, we resort to learn from context. In the case of gesture production, context is given by the entity generating the observation (e.g. human or robot). By understanding the gist of the kinematic and psychophysical processes leading to the generation of that single observation [4], a series of new artificial observations can be generated. These artificial observations can later be used as training examples for classical machine learning classifiers.

Yet, determining the underlying principles governing gesture production is particularly challenging due to human variability in gesture performance, perception and interpretation. In the approach presented in this paper, artificial observations are generated based on distinctive elements (placeholders) in the 3D gesture trajectory which are highly correlated to neurological signatures, as was recently reported by Cabrera et al. [5]. These placeholders are further used to propagate and augment a dataset of gestures which all preserve the salient characteristics of each gesture type while embedding the variabilities in human motion within a compact representation [4]. Once the datasets are created, classifiers are trained with the artificial observations and subsequently tested on real gestures datasets. Two metrics are used to assess performance: recognition accuracy and recognition coherency. The first is about the percentage of observations correctly classified over the total number of observations. The second metric refers to the level of agreement that displayed both the human observers and the machines. Whether a machine and an observer agree on which gestures are recognized or not, would indicate the level of success with which our algorithm can mimic human performance.

We expect that once machines are trained using the one-shot learning paradigm they would be able to interact naturally with their users even when encountering gestures seen only once before. During real-time interaction, gestures will be observed again and again which suggests adopting an adaptive learning approach (from 1-shot to N-shot). Thus, in this paper we also present an incremental progression of the developed framework for one-shot gesture recognition to adaptive learning. The idea of adaptive learning is implemented by applying the same methodology of dataset augmentation for the training examples but with different proportions. As more real-life gestures become available when interacting with a robot, the ratio of real over artificial data would increase, thus acting as a form of learning from demonstration [6] paradigm.

II. BACKGROUND

A. Relevance of gestures in communication/interaction

Gestures are a form of engaging our body into expressions with the objective of conveying a message, completing an action, or as a reflection of a bodily state. Humans are quite adept at communicating effectively with gestures even when

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.

978-1-5090-4023-0/17 $31.00 © 2017 IEEE
DOI 10.1109/FG.2017.98

some of the gestures are spontaneously evoked during interaction [7]. Communication grounding and context allows the observers to infer the meaning of the gesture even when that specific expression form was not seen before.

It would be beneficial to enable machines to understand these forms of spontaneous physical expressions that have been only seen once before. To achieve that goal, one should consider existing mechanism of communication that include not only the outcome and meaning of a particular gesture, but the process involved during gesture production that are common to different human beings. Such process involves both cognitive and kinematic aspects.

The cognitive aspects referred are those events that occur during the production of human gestures. Such events have been related to improvement of memory and problem solving [8]–[10]. Research has been conducted to relate gestures to speech on the neurological level [11], [12], yet the cognitive processes related to gesture production and perception have not been considered as a source of valuable information representative of gestures. These events (fluctuations in EEG signals related to mu rhythms oscillations) have been linked recently to gesture comprehension [5]. These cognitive signatures related to observed gestures may be used to compress a gesture in memory while retaining its intrinsic characteristics. When a gesture is recalled, these key points associated with the cognitive signatures are used to unfold the gesture into a physical expression. We plan to use these key points as a global form of gesture representation.

B. One-Shot learning in gesture recognition

One-Shot learning in gesture recognition has gained much traction since initial works proposed for the ChaLearn gesture challenges in 2011 and 2012 [13]. Results of the challenge were reported by Guyon et al. using the Levenshtein Distance (LD) metric, where LD=0 is considered the perfect edit distance [3], [14] and LD=1 a complete error. One approach by Wan et al. was based in the extension of invariant feature transform (SIFT) to spatio-temporal interest points. In that work the training examples were clustered with the K-Nearest Neighbors algorithm to learn a visual codebook. Performance was reported by an LD=0.18 [15].

Histogram Oriented Gradients (HOG) have being used to describe image based representation of gestures. DTW was implemented as the classification method obtaining LD=0.17 [16]. Another method relied on extended motion History Image as the gesture descriptor. The features were classified using Maximum Correlation Coefficient leading to an LD=0.26 [17]. In that work dual modality inputs from RGB and depth data were used from Kinect sensor.

Fanello et al. relied on descriptors based on 3D Histograms of Scene Flow (3DHOF) and Global Histograms of Oriented Gradient (GHOG) to capture high level patterns from the gestures. Classification was performed using a Support Vector Machine (SVM) using sliding window with LD=0.25 [18].

C. Adaptive learning in gesture recognition

Adaptive shot learning involves switching from 1-shot to N-shot learning paradigms as more data becomes available. An example of this is given by a robot that was trained using a single observation in a fixed environment, and sudden is relocated to a new place and needs to immediately interact with its users. As new real data becomes available, the robot adopts these new observations to re-train itself, to better recognize future instances of the gestures. Thus, gradually moving from a 1- shot learning paradigm to 2-shot, and eventually N-shot. It is expected that as real-data becomes available, performance would improve. This concept is relatively new, and few examples exist in the literature. Hasanuzzaman et al., [19] update the training set of a robot interacting with different humans, to account for changes in lighting and faces using multi-clustering approaches to incorporate new visual features to the ones already available. Pfister et. al [19] one-shot algorithm was used to detect sections of a given gesture class within a video reservoir. Video sections were then used as training samples for a new classifier, with higher performance. This approach included adaptation to human pose and hand shape that enabled generalization to videos of different size and resolutions; with different people and gestures performed at different prosody and speed.

This paper will demonstrate our one-shot learning approach through a variety of experiments and also present preliminary results on adaptive shot learning to gesture recognition.

III. METHODOLOGY

In this section implementation details are provided. First, we describe the gesture data set used and present the framework for one-shot learning from one example of each gesture class. We do this by extracting a set of salient points within the gesture trajectory and finding a compact representation of each gesture class. This representation is then used to augment the number of examples of each gesture class artificially, maintaining intrinsic characteristics of the gestures within that class. Then the selection of classification algorithms and the training/testing methodology is presented. The used performance metrics are described, and finally an extension to adaptive learning approach is presented with preliminary results.

A. Implementation details

The data used to test the approach presented consists of the Microsoft Research Cambridge-12 Kinect gesture data set. It includes 6,244 instances of 12 different gestures related to gaming commands and media player interaction. Gesture observations for the one-shot learning problem presented in this paper relies on a reduction of Microsoft Research MSRC-12 dataset, that includes a lexicon of 8 gesture classes. This reduction was due to the nature of some of the gestures, like taking a bow or kicking, that are gestures not related to motions of the upper limbs. The data set comprises tracking information of 20 joints collected using Kinect pose estimation pipeline from 30 people performing the gestures. The subset of gestures in the lexicon, was selected to include only gestures involving upper-limbs motion (Fig. 1).

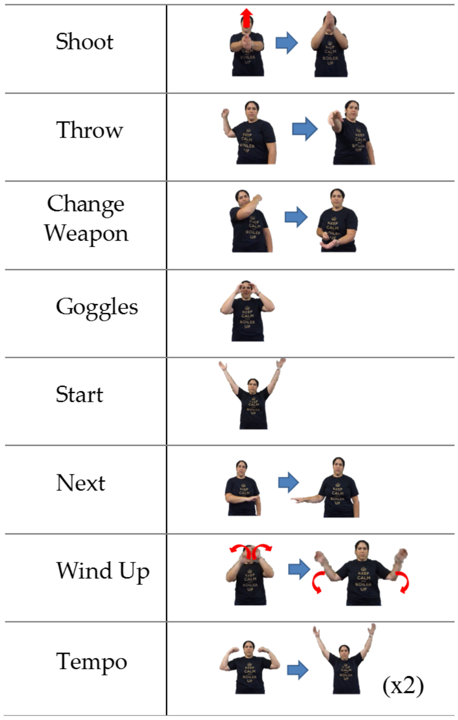

Fig. 1. Gesture lexicon from MSRC-12

B. Formal definition of the One-Shot Learning problem

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 $g^i_k \in \mathcal{G}_i$, where $k = 1, ..., M_i$. Where Mi is the number of observations of gesture class i. Gesture observations are a concatenated trajectory of 3D points with h as the total number of sample points within an observation.

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

The N-shot classification paradigm involves gathering multiple observations of each gesture class to obtain adequate classifying solutions. Instead, in the case of one-shot learning, we rely on one observation as the basis to generate several others. Thus, g^i_k 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 Mi is the desired number of instances of that class required for training. Equation (2) is applied to a single observation g^i_1, and is used to extract a set of inflection points labeled as x^i_q, where $q = 1, ..., l$ and $l < h$.

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

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. These set of inflection points, $\tilde{G}_i$ (3), will serve as the basis to create artificial gesture instances, $\hat{g}^i_k$, $\in \mathbb{R}^h$ 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 (Fig. 2). Function $\mathcal{A}$ is described further in [4].

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

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

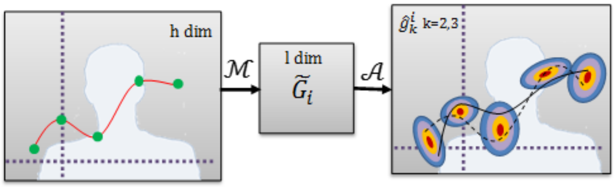

Fig. 2. Changes in gesture dimensionality through functions M and A

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

$$\Psi : \hat{g}^i_k \to \mathcal{G}_i \tag{5}$$

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

$$\begin{gathered}\text{Max Z} = \mathcal{W}\{\Psi(g^u), \mathcal{G}_i\} \\ \text{s.t. } i \leq N;\ i \in \mathbb{Z}^+;\ \mathcal{G}_i = \Psi(g^i_1);\ \Psi(g^u) \in \mathcal{L}\end{gathered} \tag{6}$$

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.

The motivation behind this form of gesture encoding is to replicate the way that humans perceive gestures in order to later decode them to generate human-like arm/hand trajectories.

The main form of encoding relies on keeping only the inflections points within a trajectory together with a variance associated to that point. It can be argued that encoding using the inflections points may not be the most effective form of compact representation of a gesture. Yet, in a preliminary experiment, it was found a relationship between the timing of mu oscillations and kinematic inflection points, such that inflection points were followed by interruptions in mu suppression approximately 300 ms later. This lag is consistent with the notion that inflection points may be utilized as place holders involved in conscious gesture categorization. The fact that positive correlations have been observed between abrupt changes in motion and spikes in electroencephalographic (EEG) signals associated with the motor cortex supports the hypothesis of a link between inflection points in motion and cognitive processes [5]. Therefore, these points can be used to capture large variability within each gesture while keeping the main traits of the gesture class.

C. Classification algorithms

Four different classification algorithms were trained using 200 artificial observations per gesture class. The performances were evaluated using 100 testing gesture examples from the public dataset mentioned earlier. The algorithms selected were DTW, CRF, HMM and SVM, renowned for their use in state-of-the-art gesture recognition approaches. In the case of HMM and SVM, a one-vs-all scheme was used, while CRF and DTW provide a metric of likelihood to the predicted result after training is completed.

The DTW classification algorithm was implemented using the Gesture Recognition Toolkit (GRT) [20], which is a C++ machine learning library specifically designed for real-time gesture recognition.

Each HMM is comprised by five states in a left-to-right configuration and trained using the Baum-Welch algorithm, which has been previously shown to generate promising results in hand gesture recognition [21].

For the SVM, each classifier in the one-vs-all scheme was trained using the Radial Basis Function (RBF) kernel. The library available in MATLAB was used to implement SVM.

In the case of CRF, the training examples were encoded using the BIO scheme to determine the beginning (B), inside (I), and outside (O) of a gesture. The CRF++ toolkit was used to train and test this classification algorithm [22].

D. Performance metrics

The recognition accuracy metric, $A_{cc}\%$, is used to evaluate the percentage of correct classification over total number of observations. It is defined in (7) as the ratio of the number of true estimations, E_{true}, to the total number of testing examples, $E_{samples}$. Accordingly, recognition accuracy is equivalent to the sum of diagonal elements of a confusion matrix divided by the sum of all elements of matrix.

$$A_{cc}\% = \frac{E_{true}}{E_{samples}} \times 100\% \tag{7}$$

Results of overall accuracy are calculated as the average of gesture accuracy per class.

A second metric is applied to measure the level of coherence between the performance of the classifiers and human observers. Both the classifiers and the human assess gestures performed artificially by a robot. High coherence found between human and machine classification indicates that the method used to generate artificial examples encompasses variability that humans understand as being part of the same gesture class.

The second proposed metric, coherency $\gamma(\cdot)$ indicates the resemblance between classifiers and human observers' recognition of gestures. The goal with this metric is to evaluate how well the method presented can mimic human production, perception and recognition. Coherency (8) is defined as the intersection between the sets of agreement indices (*AIx*) for both humans (*MH*) and machines (*MM*). The intersection is measured when both machine and humans either identify correctly a gesture or misidentify it regardless of the class where the agents classified them.

$$\gamma = \frac{AIx_{machine} \cap AIx_{human}}{\|AIx_{human}\|} \times 100\% \tag{8}$$

The *AIx* is measured as the median of a set of Boolean values of gesture recognition, which indicate whether the gesture was being correctly classified (1) or not (0). The value $\|AIx_{human}\|$ indicates the count of elements in each set, which is identical for humans and machines.

E. Progression of One-Shot framework to Adaptive Learning

The developed methodology for artificial generation of gesture examples is applied to achieve adaptive learning. Assuming the number of samples per class M_i is kept constant throughout the process of adaptive learning, the proportion between original samples $g^i_{k_O} \in \mathbb{R}^{3\times h}$ and artificially generated examples $\hat{g}^i_{k_A} \in \mathbb{R}^{3\times h}$ changes. $\alpha\%$ is the percentage of artificial data that needs to be generated for each real observation. The number of samples for real and artificial observations is given by, k_O and k_A, respectively. These parameters are related through (9):

$$\mathcal{M}_i = k_O + k_A, \qquad \alpha = \frac{(M_i - k_O)}{k_O M_i} \\ 1 \leq k_O \leq M_i \tag{9}$$

For example, for $M_i = 200$ and assuming the one-shot learning case, we have $k_O = 1$, $k_A = 199$. Then $\alpha\% = 99.5\%$ of the data needs to be generated from the single observation. If instead $k_O = 2$ then $k_A = 198$, but only 49.5% of the data needs to be generated for each. As the value of k_O grows, the number of artificial examples created from each original example decreases. This notion is shown in Fig. 3. When $k_O = M_i$, (N-shot learning) there is no need to generate artificial examples at all.

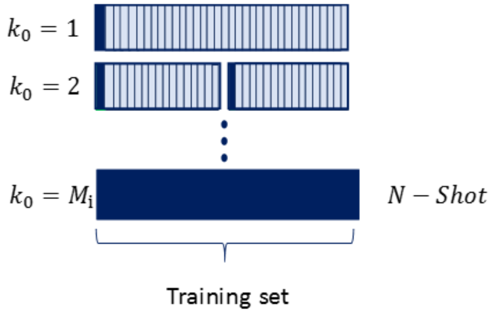

Fig. 3. Visual representation of the proportion of original and artificial examples in the training set

Conversely, with fewer number of observations available, the true observations have higher representation in the artificially generated data set. This could result in two opposing scenarios: either increasing the performance of the trained classifier by providing variability associated to the gesture class, or worsen the performance of the trained classifier if the example was an outlier within the gesture class.

IV. Results

A. Recognition accuracy

In order to test recognition accuracy, a testing set comprised of 100 examples per class was used. The average recognition per gesture class, as well as $A_{cc}\%$ are shown in table I. All the results are comparable, with highest recognition for the SVM and lowest for CRF. It is considered a positive result to have comparable accuracies for different classifiers, since the method developed for one-shot learning is agnostic of the classification method used.

TABLE I
Recognition Accuracy (%) for trained classifiers

Gesture	**HMM**	**SVM**	**CRF**	**DTW**
Start	92	92	85	93
Next	95	96	89	95
Goggles	93	93	87	90
WindUp	89	91	86	90
Shoot	89	91	91	91
Throw	91	92	92	90
ChangeWeapon	89	91	86	90
Tempo	88	89	93	92
Overall	90.9	91.9	89.1	91.4

B. Coherency

To test the coherency of the gestures generated using our approach, ten participants were recruited. Each was asked to watch a video of a person performing one labeled example of each gesture class in the selected subset of the MSRC-12 dataset. Next, each participant observed the Rethink Robotics' Baxter robot performs a total of 16 artificially generated gesture instances, two of each gesture class, in random order. They were then asked to assign a label to each gesture instance performed by Baxter.

The same artificial trajectories were performed by Baxter and detected using a Kinect. Using blob segmentation and tracking the trajectories of Baxter's end-effectors were determined and used as testing examples for the trained classifiers. These classifiers, in turn, predicted the labels of all the artificial gesture instances mentioned previously.

Table II shows the recognition accuracies found for each interaction combination: robot-human (*MH*) and robot-classifiers (*MM*). The last two columns represent the overall recognition for *MM* and *MH* respectively.

Using the recognition results from the previous two scenarios, the metric of coherency was calculated using (8). The AIx among users and machines were calculated for each gesture type and instance. The recognition coherence found across the 20 lexicon sets tested was $\gamma = 93.8\%$. Given that 100% is perfect coherency, the obtained value indicates a very good coherency between humans and machines.

TABLE II
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

C. One-Shot to Adaptive Learning

The developed methodology for adaptive learning is implemented using two different classification algorithms: DTW and HMM. Ten different values of original samples are assessed and their overall accuracy reported. The Number of artificial samples used for training was 200.

Fig. 4 is the graphical representation of the overall accuracy as a function of the number of original examples used to generate artificial examples to augment the training data set. A reference value for the one-shot learning was included based on the recognition accuracy of each classification algorithm.

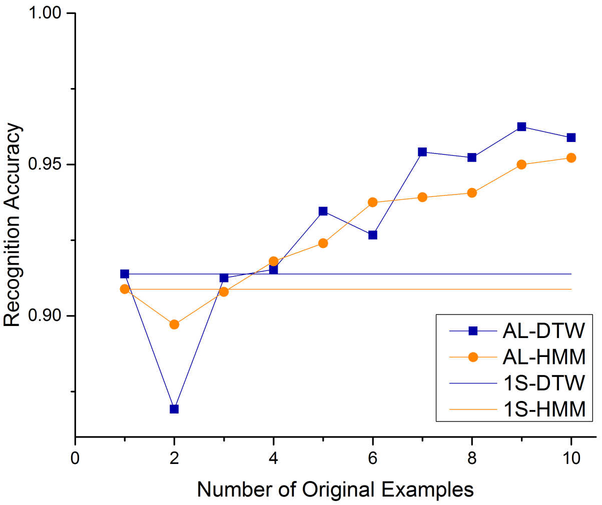

Fig. 4. Accuracy Recognition for adaptive learning framework as number of original examples varies

As the number of real observations grow, there is an increasing trend for the recognition accuracy in both classifiers used. Interestingly, when only two real samples were provided, the recognition accuracies were below the one-shot baseline. A possible explanation is that the observation used was an outlier among the typical observations of that class. This occurrence is found in the results of both classifiers, though the HMM result for k = 2 was influenced less lowering the baseline performance. Using an outlier that was not well recognized, will lead to augmenting the dataset of observations with hard to classify gesture instances. Nevertheless, as more observations are added, the recognition accuracy picks-up again and reaches 95.89% for DTW and 95.22% for HMM.

V. CONCLUSION

This paper presents the methodology and metrics associated with a framework to achieve one-shot gesture recognition. This framework is based on the extraction of salient points in gesture trajectories which are used as a compact representation of each gesture class. Using this compact representation, an augmented data set of realistic artificial samples is generated and used to train classifiers. Four classifiers commonly used in state-of-the-art for gesture recognition were used to evaluate the effect of classifier on the overall performance. Recognition accuracy was measured for each classifier using a subset of the publicly available data set of Microsoft Research MSRC-12. The recognition accuracy of all classifiers showed comparable results with an average recognition of 90%. A different metric was used to measure the artificially generated examples in terms of recognition coherence between humans and machines. For this, a robotic platform was used to perform a set of these artificial examples. The resulting coherence was 93.8%, indicating a high level of agreement in recognition when the recognizing agent in the interaction is changed.

Results regarding adaptive learning implementation were shown in terms of recognition accuracy as a function of the number of original examples included in the artificial generation process. Two classification algorithms were used: DTW and HMM. Recognition accuracy was computed when only one observation was provided to generate the artificial data sets, then two observations and so on until reaching 10 observations. It was observed that, as expected, the performance grows with a stronger representation of real-data within the mixed dataset.

VI. 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] E. Rodner and J. Denzler, "One-shot learning of object categories using dependent gaussian processes," in *Joint Pattern Recognition Symposium*, pp. 232–241, Springer, 2010.

[3] I. Guyon, V. Athitsos, P. Jangyodsuk, H. J. Escalante, and B. Hamner, "Results and analysis of the chalearn gesture challenge 2012," in *Advances in Depth Image Analysis and Applications*, pp. 186–204, Springer, 2013.

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

[5] 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 (in press)*, Jan. 2017. arXiv: 1701.05921.

[6] B. D. Argall, S. Chernova, M. Veloso, and B. Browning, "A survey of robot learning from demonstration," *Robotics and autonomous systems*, vol. 57, no. 5, pp. 469–483, 2009.

[7] D. McNeill, *Language and gesture*, vol. 2. Cambridge University Press, 2000.

[8] M. Chu and S. Kita, "The nature of gestures' beneficial role in spatial problem solving.," *Journal of Experimental Psychology: General*, vol. 140, no. 1, p. 102, 2011.

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

[10] K. Muser, "Representational Gestures Reflect Conceptualization in Problem Solving," *Campbell Prize*, 2011.

[11] R. M. Willems, A. zyrek, and P. Hagoort, "When Language Meets Action: The Neural Integration of Gesture and Speech," *Cerebral Cortex*, vol. 17, pp. 2322–2333, Oct. 2007.

[12] A. zyrek, R. M. Willems, S. Kita, and P. Hagoort, "On-line Integration of Semantic Information from Speech and Gesture: Insights from Event-related Brain Potentials," *Journal of Cognitive Neuroscience*, vol. 19, pp. 605–616, Apr. 2007.

[13] "ChaLearn Looking at People."

[14] 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.

[15] 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.

[16] J. Konen and M. Hagara, "One-shot-learning gesture recognition using hog-hof features," *The Journal of Machine Learning Research*, vol. 15, no. 1, pp. 2513–2532, 2014.

[17] 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.

[18] 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.

[19] M. Hasanuzzaman, T. Zhang, V. Ampornaramveth, H. Gotoda, Y. Shirai, and H. Ueno, "Adaptive visual gesture recognition for humanrobot interaction using a knowledge-based software platform," *Robotics and Autonomous Systems*, vol. 55, pp. 643–657, Aug. 2007.

[20] N. E. Gillian, *Gesture recognition for musician computer interaction*. PhD thesis, Queens University Belfast, 2011.

[21] M. G. Jacob and J. P. Wachs, "Context-based hand gesture recognition for the operating room," *Pattern Recognition Letters*, vol. 36, pp. 196–203, 2014.

[22] "CRF++: Yet Another CRF toolkit."