Learning Gestures for the First Time

Cabrera Maria E[1]. Major Professor: Juan Wachs[1].
[1] Purdue University, West Lafayette, IN. USA

***Abstract*— Humans are able to understand meaning intuitively and generalize from a single observation, as opposed to machines which require several examples to learn and recognize a new physical expression. This trait is one of the main roadblocks in natural human-machine interaction. Particularly, in the area of gestures which are an intrinsic part of human communication. In the aim of natural interaction with machines, a framework must be developed to include the adaptability humans portray to understand gestures from a single observation. Most approaches used previously for One-Shot Learning rely heavily on purely numerical solutions, and leave aside the mechanisms humans use to perceive and execute gestures. This gap leads to suboptimal solutions. A framework is proposed to incorporate the processes of cognition, perception and execution related to gesturing to the paradigm of one-shot gesture recognition. By extracting the "Gist of a Gesture" and considering biomechanical features of the human arm and embodied cognition, context in recognition can be achieved. The performance of the method is evaluated in terms of independence from the classifying method, efficiency in terms of comparing to traditional N-shot learning approaches, and coherence in recognition among machines and humans.**

I. INTRODUCTION

Gestures are a key component in human to human interaction. As such, we expect machines and service robots to understand this form of interaction, as intuitively, as humans do. We can see a gesture only once, and be able to recognize it the next time that is presented because of our capability to learn from few examples and ability to associate between concepts. Starting from a very young age, children are able to understand meaning intuitively and generalize from a single observation, as opposed to machines which require several examples to learn and recognize a new physical expression. Because of it, learning sessions must be spent before machines can be used in a natural and straightforward setting, limiting the scope of natural human-robot interaction.

The problem of recognizing gestures from a single observation is called One-Shot Gesture Recognition. In the aim of natural interaction with machines, a framework must be developed to include the adaptability that current approaches lack. The limited amount of information provided by a single observation makes this an ill-posed problem to apply approaches based exclusively on data mining or statistics; some form of context is required.

This work proposes that such context can be extracted by observing, understanding and modeling the mechanisms involved in gesture generation rather than the single example provided of a determined gesture class. Such mechanisms involve cognition, perception and motor execution. The proposed approach integrates in a coherent form aspects of human perception and cognition through salient features; physical action through a biomechanical model of the human body, in particular the upper limbs; and the actual generation of motion and its spatio-temporal traces. Such an approach offers a holistic framework to deal with the problem of one shot gesture recognition and generation. This approach is validated through experiments studying human and machine recognition in a variety of scenarios, whereas the success of this framework is determined by mainly the ability to mimic human gesture production and recognition, rather than machine recognition accuracy alone.

Solving the problem of One-Shot gesture recognition in the scope of Human-Machine Interaction, paves the way for other learning paradigms such as Zero-Shot learning or unsupervised One-Shot learning. Including cognitive processes and modeling human perception and execution of gestures can enrich context and semantics, which can then be used to infer the meaning of an unseen gesture. Potential new applications exist in the field of medicine, assistive technologies and interactive platforms. For example, in HCI one observation could be enough to customize a user interface to a specific user. This reduces training sessions times, and reduces muscular fatigue associated to repeated muscular performance. In diagnostics, one shot learning can be used to understand the neural processes that are involved in gesture perception and generation. This could potentially be used for early detection of degenerative diseases or disorders like apraxia.

II. BACKGROUND

A. Relevance of gestures in communication/interaction

A large body of research has been published and surveyed around gesture recognition systems, mostly in the context of human-computer interaction. Ibraheem and Khan [1] focused on the use of gestures with instrumented gloves or colored markers, additionally to vision-based approaches. Other sensor-based gesture recognition is covered by Corera and Krishnarajah [2] including accelerometers, electromyographic (EMG), and inertial measurement units (IMU) sensors. Some of the vision-based techniques reviewed by Mitra and Archaya [3] include particle filtering, optical flow, skin color models, with various combinations for both hand and head gestures. A typical pipeline of such operations include thresholding in different color spaces, edge and corner detection, morphological operations, blob detection, optical flow, gradients of orientation, difference of Gaussians, motion history, and motion energy. Some of these techniques can be highly dependent on external illumination, skin color and/or occlusion [4].

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 some of the gestures are spontaneously evoked during interaction [5]. 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 [6]–[8]. Research has been conducted to relate gestures to speech on the neurological level [9], [10], 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 [11]. 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. The framework proposed in this work uses 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 [12]. 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 [12], [13] 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 [14].

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 [15]. 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 [16]. In that work dual modality inputs from RGB and depth data were used from a 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 [17].

III. METHODOLOGY

Fig. 1 shows the general pipeline for the proposed implementation. The user's motions associated with one gesture example are detected using a Kinect sensor and the skeleton information is processed to acquire the trajectories for the joints in the upper limbs. This is followed by a process of extracting a compact representation of the gesture class, referred to as the "gist of the gesture". This representation is then used to generate an enlarged data set of artificial gesture examples through two different methods, a forward and a backward approach, which consider the kinematics of the human arm. Once the enlarged data set has been completed for all the gesture classes in a lexicon, different state-of-the-art classifiers are trained, namely HMM, SVM, CRF, DTW. The proposed method is agnostic to classification method used. Different performance metrics are used to evaluate performance, among which are accuracy, efficiency and coherency.

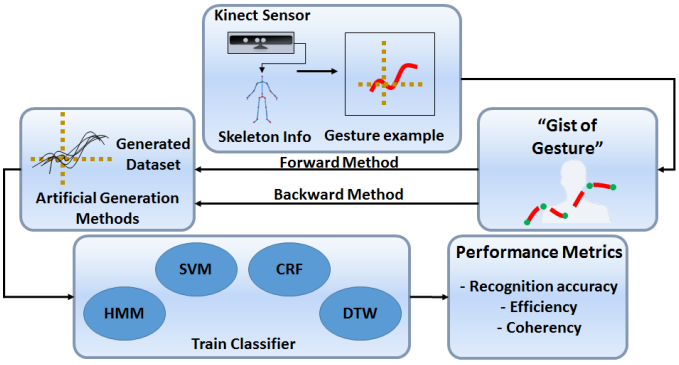

Fig. 1. Overview of methodology

A. Extracting a compact gesture representation

The placeholders of a gesture were obtained in the following way: the number and location of inflexion points for each hand's trajectory (an array of three-dimensional points), the type of curvature present between a pair of inflexion points (e.g. convex, straight, and concave), and the sequence of the movement described by the quadrant where each inflexion point is located with respect to the gesturer's shoulder, considering the YZ plane as above and below the shoulder, and closer to or further away from the body centroid. The specific magnitude of this coordinate space is based on anthropometry. The use of inflection points has been found in previous literature and associated with depicting the gesturer's intentions [18], [19]. 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 [11]. Therefore, these points can be used to capture large variability within each gesture while keeping the main traits of the gesture class.

B. Artificial dataset enlargement

1) Forward Method: This method uses the compact representation of the gesture class mentioned previously and uses the determined inflexion points as the mean for a Gaussian Mixture Model (GMM). The associated variance for each mixture of Gaussian is calculated using the different points within the gesture trajectory. This method was further explained in [20].

2) Backward Method: This process of generating artificial observations also relies on the information extracted, or the "gist of the gesture".

However, this method focuses on finding a set of V possible inverse kinematic solutions S_j for the joint angles in the human arm at each inflexion point of the gesture example. Given a target position for the arm's end effector, with 7 degrees of freedom (DOF), several solutions can be obtained. Choosing an optimal solution S_j^* at each point is accomplished through a cost function $\varphi(\cdot)$ that considers physical effort and kinematic optimization options like force-torque considerations or energy expenditure. To connect the optimal joint solutions at each inflexion point, a minimum jerk function is followed while keeping within absolute constraints of each joint.

This process becomes iterative, when one joint angle value is altered using a weighted average of other possible solutions and propagated forward in the gesture motion producing a new solution $\hat{s}_j^*$.

A different approach might also be considered, where one perturbation in one joint at a given point is propagated to simulate the human ability to compensate overreaching a target. This process also becomes iterative as different joints at different points in the trajectory are altered.

C. 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 (1) 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. Results of overall accuracy are calculated as the average of gesture accuracy per class.

$$A_{cc}\% = \frac{E_{true}}{E_{samples}} \times 100\% \tag{1}$$

The proposed method has to display generalization capabilities, which will be compared to N-Shot Learning approaches to empirically determine the number of samples required for each classifier to reach the same recognition accuracy obtained by training them with artificially generated samples. Thereby we also propose a metric of efficiency $\eta(\cdot)$ (2) that expresses by how much the presented approach is more efficient than the standard N-Shot learning approach, given the number of samples $samples_{cutoff}$ required to reach the same baseline for accuracy recognition.

$$\eta = \frac{samples_{cutoff} - 1}{samples_{cutoff}} \tag{2}$$

Another metric is applied to measure the level of coherence $\gamma(\cdot)$ 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 goal with this metric is to evaluate how well the method presented can mimic human production, perception and recognition. Coherency (3) 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.

$$\gamma = \frac{AIx_{machine} \cap AIx_{human}}{\|AIx_{human}\|} \times 100\% \tag{3}$$

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.

IV. RESULTS

In order to test the proposed method for One-Shot gesture learning, subsets of two publicly available datasets were used as lexicons: 10 gestures from the ChaLearn Dataset 2013 (CGD13), and 8 gestures from the Microsoft Research Dataset (MSRC-12). The decision to use subsets of the original datasets relates to the scope of the proposed method, limited to gestures performed with the upper limbs and avoiding gestures that are distinguishable by detecting hand configuration.

A. Recognition accuracy

The average recognition per gesture lexicon, 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. Further details on these results are reported in [20], [21].

TABLE I
RECOGNITION ACCURACY (%) FOR TRAINED CLASSIFIERS AND DIFFERENT DATASETS

Data set	HMM	SVM	CRF	DTW
CGD13	91.0%	92.6%	86.4%	89.5%
MSRC-12	90.6%	93.3%	91.4%	91.1%

B. Efficiency compared to N-Shot Learning

The efficiency of the approach was compared with that obtained with N-shot learning. This comparison used the recognition accuracy obtained in the previous section as a baseline to determine the number of samples required to achieve similar recognition results in a traditional N-shot learning approach. The cutoff values for the number of samples where the recognition accuracy for each classifier reached the baseline were used to determine the efficiency metric η. Results are shown in Table II. These results are fully presented in [21].

TABLE II
EFFICIENCY METRIC (η) FOR TRAINED CLASSIFIERS PER DATASET

η	HMM	SVM	CRF	DTW
CGD13	0.982	0.979	0.981	0.983
MSRC-12	0.979	0.976	0.985	0.985

This metric is related to the ability to save data acquisition time. This means reducing the time needed to acquire and process numerous training samples. This perspective on classification performance is an advance on current views of the one-shot learning problem.

V. FUTURE PLAN AND CHALLENGES

It is considered within the future work of this dissertation to develop the backward method related to inverse kinematics of the human arm and the selection of solutions based on ergonomics and minimum effort needs to be implemented. Additional experiments have to be conducted to further assess the independence of the system of the classification methods chosen, the impact of additional variability towards accuracy, and coherence of the method when the roles between humans and machines performing and recognizing gestures are interchanged. Another expectation is to increase the number of datasets tested, and the number of gesture classes evaluated.

REFERENCES

[1] N. A. Ibraheem and R. Z. Khan, "Survey on various gesture recognition technologies and techniques," *International journal of computer applications*, vol. 50, no. 7, 2012.

[2] S. Corera and N. Krishnarajah, "Capturing hand gesture movement: a survey on tools techniques and logical considerations," *Proceedings of chi sparks*, 2011.

[3] S. Mitra and T. Acharya, "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, 2007.

[4] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer vision and image understanding*, vol. 115, no. 2, pp. 224–241, 2011.

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

[6] 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.

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

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

[9] 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.

[10] 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.

[11] 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.

[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] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] F. Despinoy, D. Bouget, G. Forestier, C. Penet, N. Zemiti, P. Poignet, and P. Jannin, "Unsupervised Trajectory Segmentation for Surgical Gesture Recognition in Robotic Training," *IEEE Transactions on Biomedical Engineering*, vol. 63, pp. 1280–1291, June 2016.

[19] K. Buchin, M. Buchin, J. Gudmundsson, M. Lffler, and J. Luo, "Detecting commuting patterns by clustering subtrajectories," *International Journal of Computational Geometry & Applications*, vol. 21, no. 03, pp. 253–282, 2011.

[20] 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.

[21] M. E. Cabrera and J. Wachs, "A Human-Centered approach to One-Shot Gesture Learning," *Frontiers in Robotics and AI*, vol. 4, 2017.