Gestures Imitation and Modeling using Concept Hierarchies and Hidden Markov Model

Patrick M. Emaase, and Byoung-Tak Zhang
School of Computer Science and Engineering
Seoul National University, Seoul 151-744, Korea
{pemaase, btzhang}@bi.snu.ac.kr

Abstract

Imitation poses a unique problem in determining the pattern of motor activation that will make actions look like a connection of hidden states. Often any movement requires complex set of mechanisms that map an observed movement of a teacher onto one's own movement apparatus. Imitation is based on the automatic activation of motor representation by movement observation. This imitation capacity depends on learned perceptual-motor correspondence. We propose Gesture Imitation Learning through hypernetworks approach for human-machine interaction. Hypernetwork model provides dynamic and versatile connectivity ideal for robust multimodal learning.

1. Introduction

Gestures play an important role by providing key information about location, method and timing of movements, and about spatial relationship among objects. Development of intelligent multi-modal interface for natural interaction. Natural interaction is cognitively transparent, effortless multi-modal way of information exchange that happens between people. This makes it possible for human-robot interaction – robot being able to understand what the user is say and doing and user. The user can simply behave. On the other hand, robots have been developed to perform a specific set of tasks in highly constrained and deterministic environment. This require to embed specific controllers. The desired approach has to be scalable, with ability to control multiple degrees of freedom, work in highly variable environment. The main reason for humanoids to interact with humans in their daily environment. Many symbols or motions often are used in natural environments. The quality of gestural interface for any HCI is directly related to the proper modeling of hand gestures. How to model hand gestures depends, primarily on the intended application within the HCI context.

Gesture play an important role in multimodal interaction especially for conveying spatial information. The hypernetworks have a fluidic, reconfigurable molecular structure which are essential for gesture learning in dynamic sitations [1]. We propose Gestures Imitation and Modeling using Concept Hierarchies and Hidden Markov Model (HMM). This will enable us to learn gestures and predict user intentions. Key elements of this approach is the use of deep hypernetwork framework.

Gestures definitions still are grey and therefore raises the need to candidly clarify ones perspective. The basic taxonomy of gestures is illustrated in figure 1. Often gestures generate a mental concept which can be expressed through the motion of arms and hands. The mental concepts represents the intentions of the gesturer.

There is a need to learn gestures and intension for valid communication between teacher's gestures and the student's body. This will improve learning performance. To solve the ambiguity problem, gesture taxonomy is necessary. They can be intentional (gesture) or unintentional hand movement.

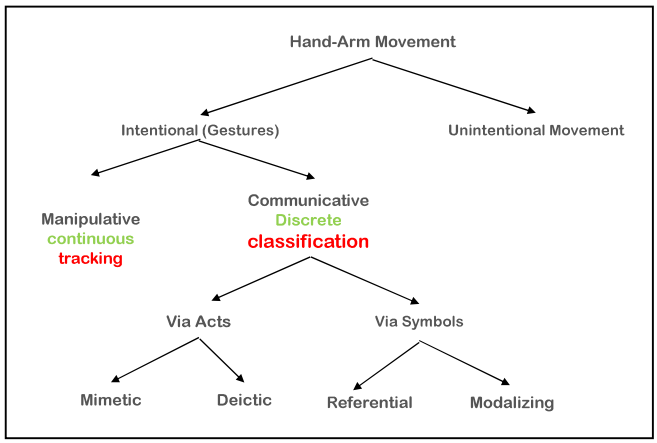

Figure 1: Hierarchy of Gestures

2. Backgrounds & Related Works

2.1 Physiological Response

Human emotions and emotions in general are complex and intuitive based on the nature and extent of stimuli in direct with. Their evocation cause discrete physiological response that is predictive. Emotion response is a cognitive function that can be largely unconscious but can be quantified dynamically. A response to gestural stimuli provides an interesting academic

pursuit in Machine learning since it can be characterized along dimensions of intention and response. The ability to accurately predict the quality and quantity of response is subtle in Machine learning. Associating specific physiological pattern to emotion response pattern requires clearly distinguishing and classifying latent responses.

Skin segmentation

We use off-the-shelf sensors to detect skin color method to compute a binary skin mask at time t, M_t^S, based on the RGB image. We also find the user mask, M_t^U obtained from the Kinect SDK based on the depth image. We align the two to find their intersection $Mt^{S \wedge U}$, which indicates the user's skin region.

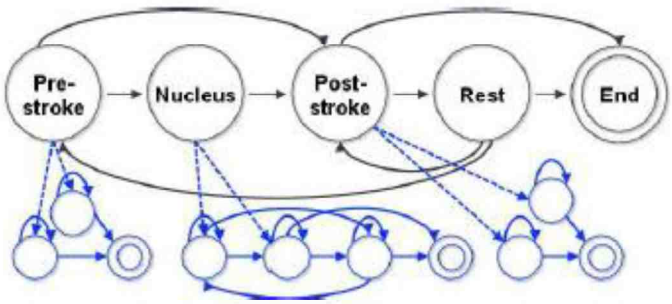

Figure 2: Temporal Gestures model with different faces. Each face can be modeled as an HMM

3. Dynamical Models for Physiological Response

Dynamic model provide appropriate solutions to sequential learning problem to learn physiological response to multimedia stimuli. The common ones are sliding window, recurrent sliding windows, input-output Markov models, conditional fields and graph transformer network as shown in figure 2.

3.1 Sliding Window Method

The sliding window algorithm converts the sequential multisensory input into the classical supervised learning problem. It uses a window classier that maps an input window of width w into an individual output value y. For our case, input w are features collected from gestures stimuli exposure. Sliding window is "half-width" of the window. For our case the width of the window is five seconds with an overlap of 2.5 seconds. A sliding window is therefore defined by a fixed number of recently generated data elements which is the target of data mining. The window classifier is trained by converting each training examples into windows and then applying a standard supervised learning algorithm through regression models.

Recurrent Sliding Windows

This is an improvement of sliding window method where the predicted value is fed as an input to predict subsequent value. It recursively improves the predictive quality of the systems. This has been applied in various dynamic modeling environments to model dynamic responses. It has the ability to improve the quality of results.

4. Experimental Results

Seventeen neurologically healthy participants (8 males and 7 females), aged between 20 and 29 (mean 24.04, standard deviation 2.29), undergraduate or graduate students, all Korean, participated in the experiment. All participants could clearly understand the stimuli mode of communication. This was supported by their respective response to stimuli.

Experimental Setup

We conducted an experiment with people teaching the Restaurant Intelligent Service (RISE) robot, and compare this to the robot learning on its own in the same environment. The experimental scenario is a RESTAURANT, in which RISE can is taught basic movements. To show experimentally which aspects of the learning process are influenced by a non-expert human partner's interaction we collected data from two types of learning sessions: SELF—the robot learning on its own; and GUIDED—the robot learning with a human teacher. We collect several independent measures during the Gestures learning sessions with human subjects. We then compare means between the two groups with t-tests in order to understand some specific details of how guidance and self-learning differ. Experimental subjects were shown their goal to help RISE learn about the actions. The subjects are told they can help RISE by making action suggestions, by naming objects and by testing the named aspects learned. The desired goals are benchmarked to determine variability. Results are collected and analyzed.

Hierarchical Gestural Analysis

Intentional gestures have distinct meaning that can be inferred from gestural movement. Unintentional movements often are fast, repetitive, and close to the body; manipulative gestures tend to be slower, and have definite hand poses.

Hypernetwork Concept Hierarchy (HCH)

Gestures form a hierarchy of intentions that are probabilistic in nature. They can be used to model and explain the interaction between behaviors at different levels of abstraction. These can be mapped into different levels in the gesture taxonomy into the levels of abstraction in Hypernetwork Concept Hierarchy (HCH). The higher levels represent the mental intention of the gesturer. We need to determine the optimal number of levels to use in HCH. The bottom level in the HCH consists of observations and states in a typical HMM. An HMM is suitable for modeling sequential data such as time series, and has been

used widely for dynamic gesture recognition with reasonable success.

Gesture Spotting and Recognition

Previous research posits that a gesture consist of three phases: pre-stroke, nucleus, and post-stroke. The pre-stroke phase consists of a preparatory movement that sets the hand in motion from some resting position. The nucleus of a gesture has some "definite form and enhanced dynamic qualities". Finally, the hand either returns to the resting position or repositions for the new gesture phase. Each gesture phase includes a sequence of hand/arm movement that can be modeled using HMMs. Therefore we can train an HMM for each phase since we have ground truth. Each phase has a variable length, we can model the termination probability for each hidden state *s* as *t*(END|*s*). This can be summarized as follows:

$$p(X_1^T, S_1^T; \theta) = t(s_1)t(END|_{ST}) \prod_{t=2}^{T} t(st|st-1) \prod_{t=1}^{T} e(xt|st)$$

Where θ represents the model parameter vector which includes the initial state probabilities t(s), the state transition probability to model user variations.

Given N training sequences, the expectation maximization (EM) algorithm can be used to estimate the model parameters. In particular, the update for the termination probability during the ith iteration.

Because there are 3 rest positions, we use 3 hidden states for both the pre-stroke and post-stroke phases. Each hidden state can be the start state and can only remain in its own state or go to the end state as shown in figure 3.

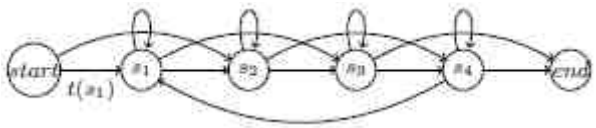

Figure 3: A state transition diagram of a modified 4 state Bakis Model for the nucleus phase

Experimental Tools used

Weka used to preprocess and analyze results. The Base learner was set to be Linear Regresssion with Test predicted at steps. This was possible after considering Training of future predictors and test future predictors to maintain the integrity of dataset and future results.

Algorithms: Basic regression models were used and compared for their accuracy and appropriateness. The evaluation were based on statistical validation parameters Correlation Coefficients (CC), Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and measured by cross-validation.

Table 1 summarizes the findings which show that prediction of gestural state using Kinect sensor is possible and feasible. This is confirmed by the generally low values of MAE and RMSE. The variation of Training and Testing data was used to show the benefits of setting aside least amount of data for testing. This gave a DAC (Direction Accuracy) of over 60%.

Table 1. Summary of Validation results

Tr:Te	MAE	RMSE	DAC
7:3	**0.0363**	**0.0424**	**67.43**
5:5	**0.0263**	**0.0344**	**53.57**
3:7	**0.0385**	**0.0477**	**56.56**

5. Discussion and recommendation

Results from figure 1 (a) and (b) show that prediction is possible and feasible. This is confirmed by the respective values of CC, MAE and RMSE. There is an increasing body of evidence supporting the application of gestural imitation and modeling using Hypernetwork Concept Hierarchies responses. The results of the experiments presented above indicate that gestural responses are predictive hence can be generalized and learned autonomously. Additional empirical work is needed, in order to establish dynamic modeling of cognitive response to. Further studies to elaborate dynamic gestural responses in a dynamic environment need to be explored. Wearables are on the rise, hence there is a need to understand spatio-temporal responses through further research. Overall, we found that different gestures can be modeled and that it is predictive in nature.

Acknowledgement

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2010-0017734-Videome), supported in part by ICT R&D program funded by the Korea government (MSIP/IITP) (10035348-mLife, 14-824-09-014, 10044009-HRI.MESSI).

References

Ahn, H.I. and Picard, R.W. (2014). Measuring Affective-Cognitive Experience & Predicting Market Success. Vol. 5, No. 2. Pp. 173 – 186.

Ekman, P. (1992). Facial Expressions of Emotions: New Findings, New Questions. American Psychological Society. Vol. 3 (1). Pp 384 - 392

Gudjonson, G.H. (1983). Suggestibility, Intelligence, Memory Recall and Personality: Experiment Study. British Journal of Psychiatry. Pp 35- 38

Healey, C.G. & Enns, J.T. (2012). Attention and Visual Memory in Visualization and Computer Graphics. IEEE Transactions on Vis. & Computer Graphics. Vol. 18 (7) pp. 1170 - 1788

Jiang, X,G., Xu, Baohan & Xue Xiangyang (2014). Predicting Emotions in User-Generated Videos. Association for the Advancement of Artificial Intelligence. Pp. 1-13