Using One-Shot Machine Learning to Implement Real-Time Multimodal Learning Analytics

Michael J Junokas[1], Greg Kohlburn[1], Sahil Kumar[1], Benjamin Lane[1], Wai-Tat Fu[1], and Robb Lindgren[1]

[1]ELASTIC[3]S, University of Illinois, 1310 S. Sixth St. Rm. 388, Champaign, IL 61820 NJ, USA
junokas@illinois.edu

Abstract. Educational research has demonstrated the importance of embodiment in the design of student learning environments, connecting bodily actions to critical concepts. Gestural recognition algorithms have become important tools in leveraging this connection but are limited in their development, focusing primarily on traditional machine-learning paradigms.

We describe our approach to real-time learning analytics, using a gesture-recognition system to interpret movement in an educational context. We train a hybrid parametric, hierarchical hidden-Markov model using a one-shot construct, learning from singular, user-defined gestures. This model gives us access to three different modes of data streams: *skeleton positions*, *kinematics features*, and *internal model parameters*. Such a structure presents many challenges including anticipating the optimal feature sets to analyze and creating effective mapping schemas. Despite these challenges, our method allows users to facilitate productive simulation interactions, fusing of these streams into embodied semiotic structures defined by the individual. This work has important implications for the future of multimodal learning analytics and educational technology.

Keywords: educational technology, gesture recognition, one-shot machine learning, cognitive embodiment

1 Introduction

The connection between embodiment and human cognition has become increasingly established within an array of academic domains [13][16][17][15][5][9]. Specifically, educational research has shown the importance of embodiment in designing effective student learning environments [10][19][11]. The idea that human cognition is embedded in our bodily interactions with our physical environment has provoked the need for technological tools that can explore, recognize, and apply users' movement in educational interventions. The ability to reinforce the embodied nature of cognition and to leverage it for effective learning has gained traction as the theoretical understanding and technical capabilities of motion-capture systems expand. With limiting factors eroding, we can start constructing models that fully take advantage of this developing

research and multimodal learning analytics (MMLA), creating applications that test the bounds of embodied learning with interactive visualization technologies.

While the opportunity to create more advanced models that engage MMLA for embodied learning exists, the majority of interactive learning systems employ direct-manipulation paradigms (i.e. pushing virtual buttons, grasping virtual objects) rather than empowering learners to use expressive gestures based on their own embodied intuitions [6][12].By using machine-learning algorithms, we are able to recognize and analyze symbolic gestures, mining parameters from users' movement, providing them with a rich and minimally constraining palate to interact and develop within systems.

Systems which utilize gesture-based input seem like a natural way to strengthen the connection between body and mind in digital interaction yet many challenges remain for the user and designer of such systems. One of the most significant inhibitors to effective interaction is the cognitive load it takes to memorize and perform precise, pre-defined inputs. Users are often forced to fit their movement into templates that have generalized any relatable physical nuance of their gesture. This prevents users from developing any kind of meaningful semiotic structure that gives them access to higher level relationships and more abstract concepts, often resorting to simple and direct connections between movement and ideas. In a fluid and dynamic interaction setting, especially where the development of a personalized language of movement would promote more established semiotic structures, this is not optimal. Users would be forced to learn a collective gesture library onto which they would have to project their expressions, remaining contained within impersonal constructs that are removed from their own conceptions.

2 Proposed Model

To empower users to define their own gestural relationships with symbols, we have implemented a novel system that nurtures an effective learning environment through real-time analytics, fostering embodied cognition. We train a hybrid hierarchical hidden-Markov model (HHMM) [8] with one-shot training [7] creating a system that is defined by and specific to the user. By specifying interaction to the individual, users are able to form stronger semiotic frameworks much quicker and with a greater level of satisfaction using our one-shot model in comparison to the same model that is traditional.

In order to collect motion data, we use the Microsoft Kinect V2 [2] due to it's relative robustness-cost ratio and its portability. The Kinect captures movement through the generation of depth maps, utilizing a camera and infrared sensor. From these depth maps a skeleton frame representing the spatial position of a subject can be extracted. Using the Kinect's application programming interface, our data collection software uses Open Sound Control [20] to send a `skeleton' of joint-positions to our feature analysis and gesture recognition components. This results in our first data stream, *skeleton position.*

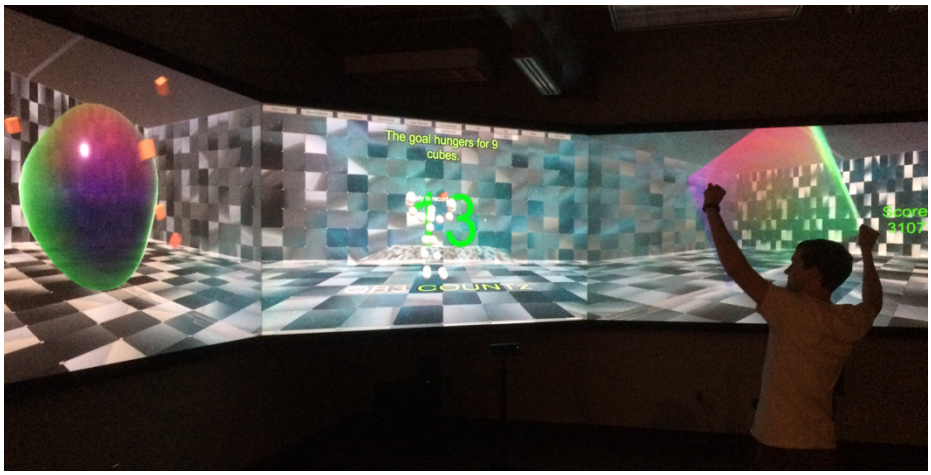

Fig. 1. User interacting with our model through an educational simulation

Our feature generation and data processing software is written in Max 6/Jitter [18], extracting an array of features including positional derivatives (i.e. velocity, acceleration, jerk), comparative features (i.e. relative positions and their derivatives, dot product between features), and statistical metrics (i.e. summation of speed across joints, analog for total body `energy', mean of feature windows) all at a variety of timeframes concurrently. This results in our second data stream, *kinematic features*.

From these features, we train a HHMM to recognize and extract abstract model features from user gestures. The user is able to define as many gestures as they wish, training each class in succession, using only one example for each class. Once the user has recorded their complete gesture space, an HHMM adapted from the IRCAM `MUBU' package [1] is trained. This model creates a machine-learned representation of the user's interaction, generating temporal, probabilistic parameters. This results in our third data stream, *internal model parameters.*

After training, we combine these data streams create a hybrid model that classifies gestures, extracts abstract model parameters, defines kinematic features, and measures summary statistics. This model can then be used to recognize and analyze users' gestures, allowing them to begin developing semiotic structure between abstract concepts and their movement. While these streams all ultimately rely on a singular source, they provide unique analytic opportunities and are thus considered different modes.

3 Model Challenges

While we have shown that our model can be successful in experimental tasks, immersive simulation theatres, and several other applications [14][3], we continue to address a variety of challenges including discovering the optimal set of features and creating the most effective mapping schemas to benefit the learners.

The biggest challenge we face is determining the most relevant features that engage the user and provide analytic insight. While we can gather and analyze a multitude of features, determining the most prescient features and how those features interact is an opportunity for realizing a more effective model in performance and accuracy. For our initial work, we chose to create a robust model that incorporates as many features as we could collect in order to empirically determine the best feature set. We've experimented with variety of approaches to focusing our features including setting empirically determined thresholds, selecting only statistically significant features, and providing users with the agency to select features of their choosing. While we've found some success with these methods, more investigation needs to be done.

While we are able to extract a variety of features, connecting them to the users beyond intuitive responses has proven elusive. The ability to adapt our analytics in real-time relies heavily on an efficient mapping schema that takes full advantage of all of the features we are extracting, physical and abstract. Additionally, this is essential to creating a more intuitive and empowering interaction for the user. In order to address this, we've incorporated real-time, visual feedback into our training and simulations, reinforcing user interaction through performance.

4 Conclusion

The ability to ground abstract concepts in tangible, intuitive, embodied metaphors is a vital foray into advancing effective education and other application interventions. By creating a model in which the user can specifically and quickly define their interactions through the fusion of different data streams, we offer a novel tool that begins to explore one-shot learning as a means to real-time multimodal learning analytics.

References

1. MUBU for Max. http://forumnet.ircam.fr/product/mubu-en/. (2015).
2. Developing with Kinect for Windows. https://developer.microsoft.com/en-us/windows/kinect/develop. (2017).
3. ELASTIC^{3}S. http://elastics.education.illinois.edu/home/about/. (2017)
4. Morphew, J., Mathayas, N., Alameh, S., and Lindren, R. Exploring the relationship between gesture and student reasoning regarding linear and exponential growth. *Proceedings of the International Conference of the Learning Sciences.* (2016).
5. Barsalou, L. Grounded cognition. *Annu. Rev. Psychol.* 59 (2008), 617-645.
6. Biskupski, A., Fender, A., Feuchtner, T., Karsten, M., and Willaredt, J. Drunken ed: a balance game for public large screen displays. *CHI '14 Extended Abstracts on Human Factors in Computing Systems*. ACM, (2014), 289-292.
7. Fei-Fei, L., Fergus, R., and Peronna, P. One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence.* 28, 4, (2006), 594-611.
8. Fine, S., Singer, Y., and Tishby, N. The hierarchical hidden Markov model: Analysis and applications. *Machine learning* 32, 1, (1998), 41-62.
9. Gallagher, S. *How the body shapes the mind.* Cambridge Univ Press.
10. Glenberg, A. Embodiment for education. *Handbook of cognitive science: An embodied approach.* (2008), 355-372

11. Goldin-Meadow, Susan. Learning through gesture. *Wiley Interdisciplinary Reviews: Cognitive Science.* 2, 6, (2011), 595-607.
12. Ibister, K., Karlesky, M., Frye, J., and Rao, R. Scoop!: a movement-based math game designed to reduce math anxiety. *CHI' 12 Extended Abstracts on Human Factors in Computing Systems.* ACM, (2012), 1075-1078.
13. Johnson, M. *The body in the mind: The bodily basis of meaning, imagination, and reason.* University of Chicago Press.
14. Junokas, M., Linares, N., and Lindgren, R. Developing gesture recognition capabilities for interactive learning systems: Personalizing the learning experience with advanced algorithms. *Proceedings of the International Conference of the Learning Sciences.* (2016).
15. Lakoff, G. and Johnson, M. *Metaphors we live by.* University of Chicago Press.
16. O'Loughlin, M. *Embodiment and education.* 15, (2006), Springer.
17. Peters, M. Education and the Philosophy of the Body: Bodies Knowledge and Knowledges of the Body. *Knowing bodies, moving minds.* (2004), 13-27, Springer.
18. Puckette, M. Max/MSP (Version 6): Cycling'74. (2014).
19. Roth, M. Gestures: Their role in teaching and learning. *Review of Educational Research.* 71, 3 (2001), 365-392.
20. Wright, M. Freed, A., and others. Open sound control: A new protocol for communicating with sound synthesizers. *Proceedings of the 1997 International Computer Music Conference.* Vol. 2013. 10.