Capturing and Recognizing Expressive Performance Gesture

Michael J Junokas, Kyungho Lee, Mohammad Amanzadeh, Guy E Garnett
Illinois Informatics Institute at the University of Illinois, Urbana-Champaign
junokas@illinois.edu, klee141@illinois.edu, amanzad2@illinois.edu, garnett@illinois.edu

Abstract

A better understanding and control of expressive performance gesture potentially could have a large and disruptive impact on electronic media and movement performance practice. We use digitally captured positional data, features extracted from this positional data, and a variety of machine-learning algorithms, to improve the accuracy of recognizing expressive qualities of performance gestures, using concepts derived from Laban Movement Analysis (LMA). Through these methods, we seek to develop better human-computer interfaces, to expand expressive movement vocabularies, and to shift movement aesthetics, by empowering users to exploit their full performance capabilities.

Keywords

Machine Learning, Expressive Performance Gesture, Expressive Movement Recognition

Introduction

Human-computer interfaces rely on merging user abilities and technological tools to form an enhanced performance environment. While advances in technology have allowed users to perform more complex tasks with greater ease, the added technological challenges can hinder the movement and expressive capabilities of the user. This interference limits functionality and compromises the diverse pallet of expressive movement qualities. We aim to create a more cognitively transparent computer interaction system that maximizes embodied knowledge based on movement, quantifiably capturing and recognizing expressiveness present in performance gesture.

Gesture in performance couples the functionality of achieving a task with the expression of aesthetic qualities, creating a dynamic and complex dual role. Our belief is that the expressiveness in performance gesture comes from subtleties of movement, extending beyond its practical function (i.e., "how a gesture is performed"(Caramiaux, Donnarumma, and Tanaka 2015). Musical conducting provides an excellent example of a gesture system that gleans expression from nuanced gesture, expanding the range of a conductor's performance from micro cues to a grandiose breadth. The work done in (Kolesnik and Wanderley 2004), (Maes et al. 2013), and (Morita, Hashimoto, and Ohteru 1991) all point toward taking advantage of this broad vocabulary, driving different technologies through movement. Our work similarly attempts to dissect a performer's gestural range, discovering the subtleties and range of their movement through feature extraction, statistical measures, and machine learning. Through this analysis, gestural nuance can be applied to technological interactions with greater control and more expressive means.

We describe the current state of our research in capturing, analyzing, and applying expression in performance gesture. To apply machine-learning algorithms to performance gesture, we needed to develop a vocabulary to clarify the movement qualities we wished to quantify. For this purpose, we chose Laban Movement Analysis (LMA)(Laban and Ullmann 1966) and used it to inform and guide the creation of our dataset and our methods of feature extraction. We describe these features and our use of several machine-learning algorithms (specifically k-means clustering, Hidden-Markov Models (HMM), and autoencoders), followed by a discussion of the results and implications of these approaches. Finally, we will offer conclusions and future directions we wish to pursue based on this research.

Classifying Movement through Laban Movement Analysis

Laban Movement Analysis is a method and language created by Rudolf Laban (1879-1958) analyzing, describing, and explaining movement in terms of functionality, tendencies, intention, and expression. The method is principally used by dancers and choreographers as a way to gain insight into movement from an expressive and intentional realm, but has expanded as a descriptive vocabulary for movement itself (Maletic 1987). We are centrally concerned with the goal that Laban states:

> Basically one has to start with the description of movement .... Our aim is thus the mastery of movement through explanation. (Laban and Ullmann 1966)

The 'mastery of movement through explanation is executed in our work by utilizing LMA as a framework from which we can build a classifiable dataset of labels that encompasses our movement capabilities. From this data set, we can perform digital analysis using machine-learning techniques. Using concepts from LMA, we develop a structure to represent movement, intention, and expression.

LMA has five principal components that collectively create a comprehensive symbolism for movement: Body, Space, Effort, Shape, and Relationship. In our work, we focus on Effort, which relates most directly to expressive characteristics we are seeking.

There are four distinct components of Effort: Space, Weight, Time, and Flow. In our work, we focus on the first three components, omitting Flow, which tends to be based on the interconnection of other movement qualities. Each Laban Effort Component represents a continuum between an *indulging* and a *fighting* Basic Effort Factor (BEF). Space can be represented on an axis from direct (fighting: focused, channeled) to indirect (indulging: multi-focused, all around awareness). Weight can be represented on an axis from strong (fighting: forceful, firm) or light (indulging: fine touch, buoyant). Time can be represented on an axis from sudden (fighting: urgent, instantaneous) or sustained (indulging: lingering, gradual).

Through the combinations of these limits, eight Basic Effort Actions (BEA) can be created: Float (sustained, indirect, light), Flick (sudden, indirect, light), Wring (sustained, indirect, strong), Slash (sudden, indirect, strong), Glide (sustained, direct, light), Dab (sudden, direct, light), Punch (sudden, direct, strong), and Press (sustained, direct, strong) (Hackney 2003)(Laban and Lawrence 1947). These BEAs are the vocabulary upon which our model is built and establish a framework from which we can build a quantifiable, labeled dataset of expressive movement.

Data Capture System and Feature Extraction

In order to analyze BEAs, we turned to digital movement analysis. For our corpus, we asked 8 performers (6 non-experts and 2 with LMA training) to perform each of BEAs in isolation and used the Microsoft Kinect to capture the gesture data. From the captured data, we derived a skeleton using the x, y, and z position of 21 distinct joints of the performer, of which we focused on the right wrist.

The positional data was transformed into higher-level features such as velocity and acceleration. We also extracted comparative features to measure movement curvature, such as the dot product between successive positional or successive velocity vectors. Additionally, we applied Fourier transform to each of these features. All of these features were used singularly or in combination to create different views of the dataset through the features. These computations allowed us to view performance gestures as a series of feature segments, giving us a means to explore the expressiveness of each performance gesture.

In addition to using these features in their raw form, we turned to dimensionality reduction using Principle Component Analysis (PCA) (Wold, Esbensen, and Geladi 1987) and Independent Component Analysis (ICA) (Hyvärinen and Oja 2000) in tandem to isolate the features that contained the most representative aspects of the gesture. We also used unsupervised learning, specifically autoencoding (Ng 2011), to find machine-derived feature combinations that could optimally represent our dataset.

Recognition of Gestures

We applied a variety of machine learning algorithms to our extracted features in an attempt to best classify and recognize performance gestures. We experimented with k-means clustering, hidden Markov Models, and using autoencoding features with logistic regression models and support vector machines . Within each of these models, we attempted classification at varying time windows and with varying feature sets

K-Means Clustering

The k-means clustering algorithm allows us to categorize data with similar characteristics into discrete clusters. After generating a number of cluster centroids using the algorithm, we are able to characterize the data by its nearest centroid (Arthur and Vassilvitskii 2007). For a given BEF, a normalized motion profile histogram is created by tallying the nearest centroids of all the training BEFs data points, creating an average histogram over the number of samples. These motion profiles are then compared to input data histograms and classified as whichever BEF motion profile it is closest.

Through empirical evaluation, we found the best classification generating 32 distinct clusters using the combination of the velocity and normalized dot product of changes in positional data.

	Weight	Space	Time
Indulging	0.60	0.80	0.83
Fighting	0.72	0.45	0.55

Table 1: F1 scores of BEF classification with K-means clustering with 32 clusters using 8 frame feature windows of velocity and normalized dot product data

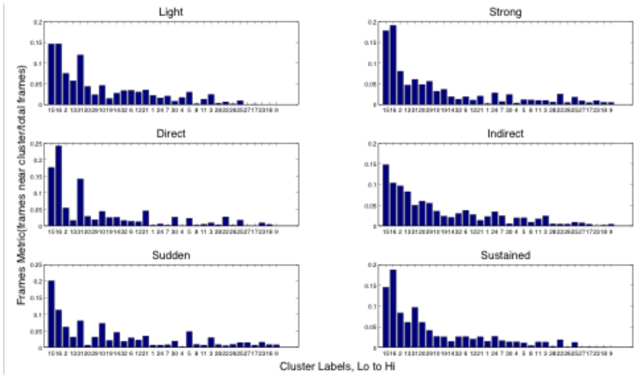

Figure 1: Motion profile histograms generated from k-means algorithm and velocity data

While using k-means clustering did not produce the highest F1 scores, it was an important step towards real-time movement analysis and provided direction for using more sophisticated algorithms.

Hidden Markov Models

Hidden Markov Models (HMM) have been widely used in modeling sequential data such as movement, and particularly, speech (Rabiner 1989). The model assumes that a hidden process with a finite number of states controls the observed data and assumes that the probability of being in a state at each time only depends on the state of the model at the previous time. Using these stipulations, the algorithm models a latent distribution of the data in each state of the model. Given a trained HMM, we can identify the likelihood of observing the input data within that model.

Like the template-matching model used with the k-means clustering, Laban Effort Component classifiers were made with two HMMs, each modeling the indulging or fighting BEF of the Laban Effort Component. We compared the likelihood of the input gesture to both HMMs and chose the model that resulted in the higher likelihood. Through empirical evaluation, we determined that using a moving average filter (Smith and others 1997) with 12 frames width on velocity resulted in the most representative feature of the data. We sliced the sequences into non-overlapping segments of 16 frames and trained HMMs with 8 hidden states. In order to predict the BEF of each frame of the test data, we sliced them into segments of 16 frames with 15 frames overlap.

	Weight	Space	Time
Indulging	0.71	0.81	0.87
Fighting	0.70	0.79	0.67

Table 2: F1 scores of BEF classification with HMM using 15 frame feature windows of velocity data

When testing HMM classification with different features, we noticed that the Fourier transform of features performed worse than non-transformed features. We decided to investigate how the Kinect was smoothing or filtering out information from our gestures.

To test this, we performed a hand movement moving periodically up and down at a rate of one beat per second (1 Hz) and gradually increased the rate to our maximum capability. We simultaneously recorded the movement with the Kinect and the accelerometer within a Nintendo Wiimote and compared the two. The Wiimote's accelerometer recorded the increased frequency throughout the recording. However, the Kinect stopped showing the increased frequency around 6 Hz and began to filter out a majority of the information (see Figure 2), potentially removing several important features of our data. The smoothing of the joint data can be altered within the Kinect Software Development Kit (Jana 2012) and will have to be further investigated if we wish to use FFT analysis to its fullest capabilities.

Autoencoder Features

In addition to utilizing algorithms for classifying, we were also able to obtain a machine-based representation of the data through unsupervised learning, specifically autoencoder activations. Similar to principal component analysis, autoencoders generalize datasets to a representative collection of features, allowing us to reduce dimensionality and present a non-linear representation of our dataset (Hinton and Salakhutdinov 2006) (Ng 2011).

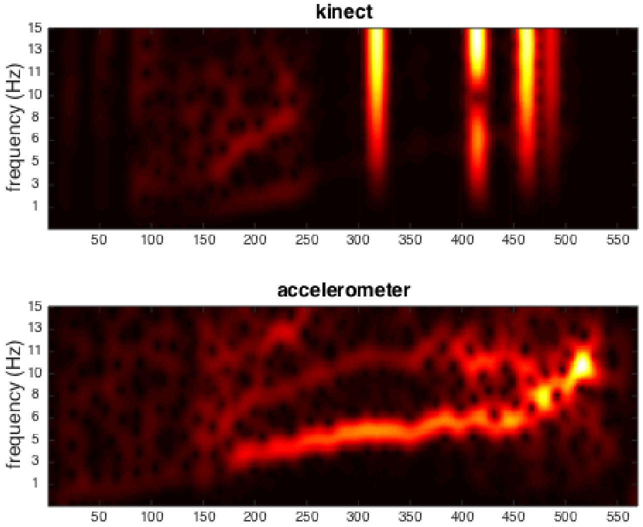

Figure 2: Comparing the Fourier spectrogram of kinect and Wiimote accelerometer as we increase the frequency of hand movement

	SVM	SVM w/AE	LR	LR w/AE
Sudden	0.78	0.81	0.49	0.57
Sustained	0.72	0.71	0.61	0.75
Strong	0.51	0.57	0.54	0.58
Light	0.68	0.69	0.46	0.55
Direct	0.52	0.56	0.52	0.60
Indirect	0.67	0.67	0.49	0.51

Table 3: F1 scores of BEF classification with support vector machine (SVM) and logistic regression (LR) classifiers using 16 frame feature windows of velocity data and activations from autoencoder (w/AE)

Using windows of 16 frames on velocity data as source data, we generalized our movement using 48 activation states. Using these activation states, we created training and testing feature sets from our LMA recordings. These feature sets were then fed to logistic regression and support vector machine classifiers. (see Table 3).

Discussion of Results

While the overall accuracy of our tests did show some level distinction between LMA gestures, the levels were not as high as we anticipated. This could be due to several reasons.

The windowed segments of the data could be less distinctive than the labels imply. In preliminary tests, classification was attempted on the entirety of gestures rather than windowed segments. This resulted in much higher accuracies, but was not done in real-time, our ideal system setting. This suggests that there are distinctive elements within the gestures, but our windowing was unable to isolate those elements from the rest of the data. Steps must be taken to consider the balance between considering the whole gesture versus a real-time computing system. Using infinite impulse response filters or segmenting by specific points of interest in the gesture may provide a more ideal compromise between optimizing our data usage and working in real-time.

The features from a singular joint may not contain enough information to provide clear distinction between gestures. In our tests, we used only a fraction of the total features and only derived those from the right wrist. Testing with more features from additional joints in combination could lead to higher classification scores.

Autoencoding activations were used in replacement of velocity features for our tests. Alternatively, these activations could be concatenated with the original features, creating combinatorial feature representation of both sets, potentially improving accuracy.

Conclusions and Future Work

Our research in recognizing nuanced expression in gestures is beginning to show promising results and has directed our plans for future investigation. The combinatorial nature of feature representation still needs to be further explored in order to find the best representation of the data. A more generalized movement vocabulary that is simpler and more basic than LMA could allow for a more fundamental approach to movement, reducing intentional context. Locating and focusing on the most essential data segments rather than all segments could direct our research to the most representative gesture elements. While autoencoding has generated better feature representations of our data, that representation has only been used with select algorithms and should be used with others including HMMs or deep-belief networks. Kinect filtering will lead us into exploring the internal configuration of the latest release of Microsofts sensor in hope of manipulating the filtering that affects Fourier transforms of the data.

While there is still much work to be done, we have found promise in our current research. Through understanding the basic building blocks of performance gesture through machine learning, we can begin to more effectively understand and generalize that gesture. Through this understanding, we can start taking full advantage of our full physical capacities in human-computer interactions.

A deeper understanding of the expressiveness within performance gesture could lead to more efficient, liberated, and expressive human-computer interactions, which would foster user-driven innovation, providing more refined and robust methods of information control and exploration. This expansion of expression could redefine the very fundamentals of movement performance practice, disrupting the current paradigm and forcing a new approach to technology and movement aesthetics. Using digital feature capture, data analysis, and machine-learning algorithms, we seek definable expression and intention in performance gesture to realize this paradigm disruption.

References

Arthur, D., and Vassilvitskii, S. 2007. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, 1027–1035. Society for Industrial and Applied Mathematics.

Caramiaux, B.; Donnarumma, M.; and Tanaka, A. 2015. Understanding gesture expressivity through muscle sensing. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21(6):31.

Hackney, P. 2003. *Making connections: Total body integration through Bartenieff fundamentals*. Routledge.

Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *Science* 313(5786):504–507.

Hyvärinen, A., and Oja, E. 2000. Independent component analysis: algorithms and applications. *Neural networks* 13(4):411–430.

Jana, A. 2012. *Kinect for Windows SDK Programming Guide*. Packt Publishing Ltd.

Kolesnik, P., and Wanderley, M. 2004. Recognition, analysis and performance with expressive conducting gestures. In *Proceedings of the International Computer Music Conference*, 572–575.

Laban, R., and Lawrence, F. C. 1947. *Effort*. Macdonald & Evans.

Laban, R., and Ullmann, L. 1966. *Choreutics*. Macdonald and Evans.

Maes, P.-J.; Amelynck, D.; Lesaffre, M.; Leman, M.; and Arvind, D. 2013. The conducting master: an interactive, real-time gesture monitoring system based on spatiotemporal motion templates. *International Journal of Human-Computer Interaction* 29(7):471–487.

Maletic, V. 1987. *Body-space-expression: The development of Rudolf Laban's movement and dance concepts*, volume 75. Walter de Gruyter.

Morita, H.; Hashimoto, S.; and Ohteru, S. 1991. A computer music system that follows a human conductor. *Computer* 24(7):44–53.

Ng, A. 2011. Sparse autoencoder. *CS294A Lecture notes* 72.

Oblong Industries. 2009. Oblong G-speak. `http://www.oblong.com/g-speak/`.

Rabiner, L. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2):257–286.

Smith, S. W., et al. 1997. *The scientist and engineer's guide to digital signal processing*. California Technical Pub. San Diego.

Wold, S.; Esbensen, K.; and Geladi, P. 1987. Principal component analysis. *Chemometrics and intelligent laboratory systems* 2(1):37–52.