Toward an Emotionally Intelligent Piano: Real-Time Emotion Detection and Performer Feedback via Kinesthetic Sensing in Piano Performance

Matan Ben-Asher
Music Engineering, University of Miami
1550 Brescia Avenue
Coral Gables, Florida
m.benasher@umiami.edu

Colby N. Leider
Music Engineering, University of Miami
1550 Brescia Avenue
Coral Gables, Florida
cleider@miami.edu

ABSTRACT

A system is presented for detecting common gestures, musical intentions and emotions of pianists in real-time using kinesthetic data retrieved by wireless motion sensors. The algorithm can detect six performer intended emotions such as *cheerful*, *mournful*, and *vigorous*, completely and solely based on low-sample-rate motion sensor data. The algorithm can be trained in real-time or can work based on previous training sets. Based on the classification, the system offers feedback in by mapping the emotions to a color set and presenting them as a flowing emotional spectrum on the background of a piano roll. It also presents a small circular object floating in the emotion space of Hevner's adjective circle. This allows a performer to get real-time feedback regarding the emotional content conveyed in the performance. The system was trained and tested using the standard paradigm on a group of pianists, detected and displayed structures and emotions, and it provided some insightful results and conclusions.

Keywords

Motion Sensors, IMUs, Expressive Piano Performance, Machine Learning, Computer Music, Music and Emotion

1. INTRODUCTION

In live performance, the dyad between performer and audience creates an intimate setting where extreme emotions manifest in music and gestural expressions. And it is the integration and interaction of senses – sound and sight – that the performer exploits in order to convey the verbally ineffable.

From the motivation of expressive performance in computer music, there have been ongoing attempts to quantify and objectify the way in which human expression can be embedded in an otherwise stale performance. The KTH rule-system developed by Bresin and Friberg [8], describes a set of rules applied to a musical score for it to sound lively and expressive upon computer playback. This has led to an enhanced understanding of how expression is conveyed in the audio of a music performance. More recent development and use of machine-learning algorithms has spawned research in human gesture recognition aimed at the control of audio effects via conducting gestures [5], and [14]. The motion of musicians in performance has also been approached to some extent, but mostly from a pedagogical point of view [12]. Friberg [7] implemented a fuzzy logic analyzer that uses audio data as well as video stream of a performer to map to specific expression. The algorithm calculates a parameter called *Quantity of Motion* (QoM) and along with audio analysis can detect *happiness*, *sadness*, or *anger.* Also recently, Gillian [11] developed a gesture recognition toolbox for the EyesWeb[1] environment that provides a variety of machine-learning algorithms that can operate in real-time. This environment was designed to explore and develop interactive multidimensional musical interfaces and displays [3].

It seems though, that the use of novel technologies such as motion sensors and machine intelligence has not yet been adopted by many musicians, which still mostly perform with instruments employing technology from decades ago. One justification to this phenomenon is that performers lack the bandwidth required to master additional controls [4], or in other words, they have their hands "tied". This implies that these technologies have still not been implemented in musical controllers in a way that is mature enough to create music that can be mastered, widely adopted and appreciated. It is possible that the inherent shortcoming these controllers exhibit is that they require the performer to display a new distinguishable gesture that the computer could detect. But why not rather have the controller learn to recognize the gestures that performers already create in their playing? Mastering a controller such as this would impose a minor requirement on musicians in terms of how much they need to alter their movements. Such an intelligent controller could be trained to detect various types of musical intentions, expressions, and musical emotions and use this information to interact with the performer almost like a fellow musician. This research presents a first step towards an emotionally intelligent controller that can detect musical expression and emotion in intuitive gestures based solely on the existing kinesthetic data of piano players in real time.

2. METHOD

Our system is comprised of wireless motion sensors, a computer, and a MIDI enabled piano. The system can work just as well with an acoustic piano. The MIDI data was only used to simplify data collection in the evaluation phase. Inertial measurment units (IMUs) are placed on each of the pianist's wrists. The pianist is instructed to play a spectrum of emotions. We use the emotion categories first described

NIME'13, May 27 – 30, 2013, KAIST, Daejeon, Korea.

[1] http://www.infomus.org/eyesweb_eng.php

in Hevner's *Adjective Circle* [13] (see figure 1). Another common measurement of emotions is using Russell's *Circumplex Model of Affect* where musical emotion is mapped in two dimensional space of *valence* vs. *arousal* [20]. However, the emotion categories in this model were not tailored for musical emotions, and some are difficult to convey in music. Moreover, it is possible to map Hevner's circle to a dimensional model similar to Russell [9].

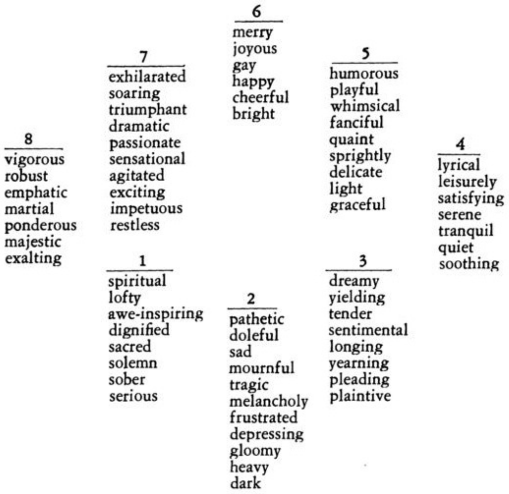


Figure 1: Hevner's Adjective Circle 1936 (from [13])

Based on preliminary testing we decided to omit category number 1 (*spiritual*) because most of the subjects had difficulty conveying such an emotion in music and could not perform it with enough repeatability. We also combined categories 7 and 8 (*exhilarated* and *vigorous*) because the difference between them was too subtle for the subjects to perform with distinction. The data from the sensors are recorded in synchronization with the MIDI signal. By doing this and prior knowledge of the musicians' intentions, we can train a classifier to learn these emotions as described in the following sections.

3. SYSTEM DESIGN

The system is comprised of *Opal* monitors from APDM[2]. These sensors transmit nine values wirelessly to an access point connected to a computer. The access point synchronizes the data coming from the different sensors and buffers it in to the operating system were the signal is received in MATLAB for initial processing. The data is then transmitted to the EyesWeb environment for real-time gesture recognition. The classification algorithm uses a Naïve Bayes approach employed in the SEC Machine Learning Toolbox in EyesWeb [10]. The algorithm detects the emotions and provides feedback to the user regarding the expressive character of the performance using two interactive displays.

3.1 Inertial Measurement Units

Opals are small wireless wrist worn motion sensors that measure and transmit nine values: acceleration x, y, z, angular velocity x, y, z, and magnetometer x, y, z. The magnetometer values are ignored because that would make the training data based on orientation relative to the north make the algorithm sensitive to the pianist's position. The data from the sensors is transmitted wirelessly to an access point for synchronization. The access point is read utilizing a set of functions in MATLAB in a dedicated SDK. The data is transmitted to the EyesWeb Environment via OSC messages at 64 samples per second.

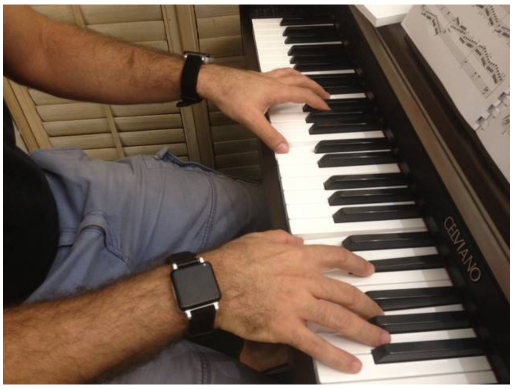


Figure 2: Opal monitors on hands.

3.2 Algorithm Description

Six dimensional data from two sensors are collected by the system at 64 samples per second. This data is buffered into three consecutive frames of 1 second each. This time frame corresponds to a one measure in *Andante* tempo (∼60 BPM) [19], and in faster playing even more, consistent with gestalt theory of music perception. Three windows are buffered to account for short-term memory based on the phonological loop buffer of 1–2 seconds for sound [1]. On each signal for each frame the following features are calculated: Mean, Standard Deviation,and RMS. In addition, the following features are continuously evaluated from the combination of the features: *Tempo*, *Dynamics*, and *Articulation*. These features have been reported as used by performers to convey emotions in the *lens model* paradigm [16]. The features are then fed to a Bayes classifier in the gesture recognition toolbox [11]. The intended emotion is also passed to the classifier. In the training phase, the trainer calculates the maximum likelihood parameters [6], μ and σ for each feature based on a Gaussian density function and stores it to a file. The predictor loads this file at start up or upon request (after retraining), compares the new input features to the trained data, and predicts the current class. For more information on the classifier see [10].

3.3 Visual Feedback

The data are collected, trained and classified in real time. The emotions are predicted as discrete integers 1–6 and displayed as text on the screen. To make the display more intuitive and to smooth the behavior of the algorithm, a moving average is computed over the previous three seconds for similar consideration of short term memory as above [1]. The detected emotions are then mapped to colors on an HSV color map and set the background of the piano roll based on the MIDI input being played. The mapping of emotional musical performances to colors has been studied in [2] and was used as a reference for our color mapping. Displaying the predictions this way allows for the feeling of a continuous flow in the music along with the musical emotion coloring that evolves in a natural rate similar to how we experience emotional responses in music.

A second feedback display is the adjective circle projected on an HSV color wheel. The detected emotion is shown as a circular object with a tail-like trace floating in two-dimensional emotion space. This creates a feeling of motion to the algorithm and presents the performers with the complete space while they adjust their playing to move from one

[2] http://www.apdm.com/.

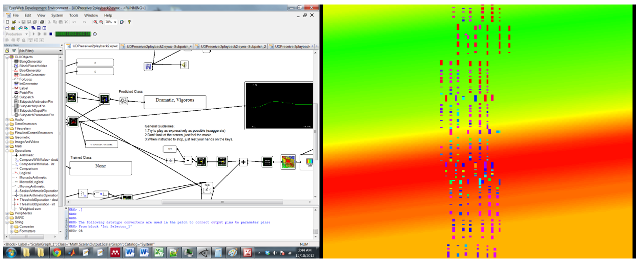


Figure 3: Display of piano roll with emotion colored background.

place to another, thus allowing the performer to “learn” the system after the system has been trained on the performer.

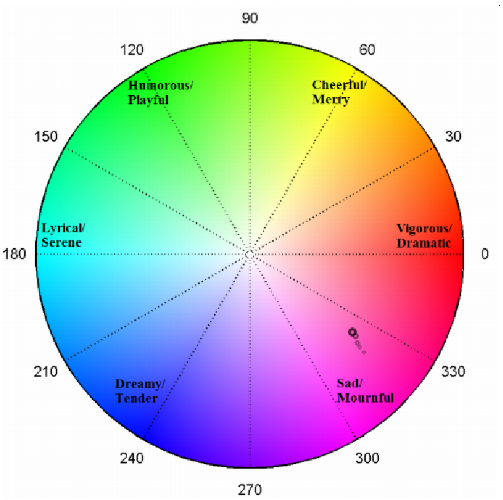


Figure 4: Adjective circle mapped to HSV Colorwheel with detected emotion.

4. EVALUATION

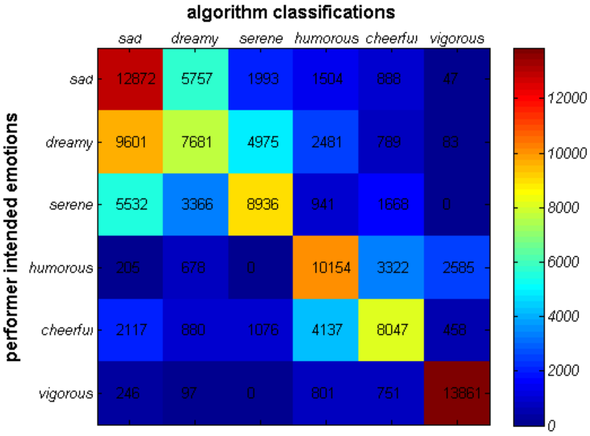
The system was evaluated on a homogenous group of 13 test subjects with an average playing experience of $\mu = 12.6$ ($\sigma = 4.8$) *yrs*. The age group statistic was $\mu = 21.8$ ($\sigma = 3.0$) *yrs*. The performers played the first few measures of Bach’s Minuet in G major, BWV 841 from the *Notebook of Anna Magdalena Bach*. The piece was played six times in the six different emotion categories to create the training data. This technique is the *standard paradigm* and has shown successful results in similar research [17]. The performers were only allowed to make variations in intensity, tempo, accents and slight pitch variations (major/minor and decorations). The latter was justified with the notion that the audio is not used for classification, so tonality is not an affecting factor. This was repeated twice while randomizing the order for training and testing. The algorithm-classified emotion categories were tested against the intended emotion categories. Each section was approximately equal in length and lasted 15–20 seconds.

4.1 Results

The classification performance is summarized in the total confusion matrix which is calculated by the sum of individual subject confusion matrices (Figure 5). The results show high performance in detecting most of the intended emotions. A clear diagonal is observed in the confusion matrix, with the main discrepancies occurring between the *sad*, *dreamy*, and *serene* categories.

In order to understand the confusion matrix more, one should observe it in varying levels of resolution. First, four major blocks are clearly seen, the two on the main diagonal are bright and the two on the remaining are dark. This is an indication of the algorithm’s strong ability in distinguishing between high and low arousal in emotions. The *sad*, *dreamy*, and *serene* are low arousal and the *humorous*, *cheerful*, and *vigorous* are high arousal categories.

Second, the inner diagonals are observed, this is an indication of the algorithm’s ability in distinguishing valence. The *sad* and *vigorous* categories are considered low in valence while the *lyrical*, *humorous*, and *cheerful* are in high valence. The *dreamy* category is generally assumed neutral in valence, a possible explanation as to why it gets confused the most. The relatively large number of confusions between adjacent categories such as *dreamy*, *sad*, and *serene* is also consistent with the previously mentioned dimensionality in the adjective circle [9].



algorithm classifications

performer intended emotions	sad	dreamy	serene	humorous	cheerfuı	vigorous
sad	12872	5757	1993	1504	888	47
dreamy	9601	7681	4975	2481	789	83
serene	5532	3366	8936	941	1668	0
humorous	205	678	0	10154	3322	2585
cheerfuı	2117	880	1076	4137	8047	458
vigorous	246	97	0	801	751	13861

Figure 5: Overall Confusion Matrix Stage 1. A clear diagonal is observed, the main discrepancies are between *dreamy*, *sad*, and *serene* .

Based on the total confusion matrix, the results for *Precision*, *Recall*, *Specificity*, and *Accuracy* [18] are be obtained per category (Table 1). The results show that the *vigorous* emotion scored highest in all categories. This was expected since the vigorous playing is very different and easily distinguishable from the other emotions. This is also consistent with the functionalist perspective [15], i.e. that we are programmed to be sensitive to emotions that can be life threatening and are imperative to our survival.

The *humorous* and *cheerful* categories had similar results, both were lower in precision and recall because they were confused between each other. Their accuracy however, is still relatively high because when detected, they were not confused with other categories.

The *sad* and *lyrical* also had similar results but lower than the other categories because they were often not only confused between each other but also with *dreamy*. The *dreamy* category had the lowest achievement in all categories. This too, matches our expectation regarding its neutral valence.

5. CONCLUSIONS AND FUTURE WORK

We have implemented and evaluated a system for detecting and displaying musical emotion and expression using kinesthetic data from motion sensors in real time. Our research shows that kinesthetic information from the pianists reveals information regarding the technical and emotional

Table 1: Precision, Recall, Specificity, and Accuracy as defined in [18]

Category	Precision	Recall	Specificity	Accuracy
sad/mournful	0.421	0.5582	0.8842	0.7647
dreamy/tender	0.4161	0.2999	0.8208	0.7578
lyrical/serene	0.5263	0.4371	0.8867	0.8351
humorous/playful	0.5072	0.5993	0.9311	0.8595
cheerful/merry	0.5203	0.4814	0.9159	0.8643
vigorous/dramatic	0.8137	0.8797	0.9813	0.9572

content of the music performed. This information can be displayed as feedback to allow interactive composition and performance. The results of our preliminary tests show that the algorithm predicts the intended emotion with accuracy in the range of 76% (*dreamy/tender*) up to 96% (*vigorous/dramatic*). Due to the low sample rate of this type of data and the availability of real-time machine learning techniques, it should be possible to achieve high accuracy in detection in real-time performance. Live information such as this could be used to facilitate emotionally augmented performances and a new experience of live music.

Further research now being performed with this system involves studying the evolution of emotion in time as the music progresses. Such evolution could be the motion path around the adjective circle during a section of the piece. These patterns could reveal helpful information regarding the structure and composition of the piece as well as the psycho-expressive dynamics during a musical performance. Furthermore, in this research the system was trained on performer intentions, future work could use listener evaluations to train the system. Then, a three-way comparison could be done comparing the performer intended emotion to the listener perceived emotions and the algorithm's predictions. This would allow real-time observation of the lens model suggested by Juslin [16].

6. ACKNOWLEDGMENTS

This research was partially funded by Ossür [3] as part of the Reload project and by the University of Miami [4].

7. REFERENCES

[1] A. D. Baddeley and G. Hitch. Working memory. *The psychology of learning and motivation*, 8:47–89, 1974.

[2] R. Bresin. What is the color of that music performance. In *Proceedings of the International Computer Music Conference*, pages 367–370, 2005.

[3] A. Camurri, S. Hashimoto, M. Ricchetti, A. Ricci, K. Suzuki, R. Trocca, and G. Volpe. Eyesweb: Toward gesture and affect recognition in interactive dance and music systems. *Computer Music Journal*, 24(1):57–69, 2000.

[4] P. Cook. Principles for designing computer music controllers. In *Proceedings of the 2001 conference on New interfaces for musical expression*, pages 1–4. National University of Singapore, 2001.

[5] R. Dillon, G. Wong, and R. Ang. Virtual orchestra: An immersive computer game for fun and education. In *Proceedings of the 2006 international conference on Game research and development*, pages 215–218. Murdoch University, 2006.

[6] R. Duda, P. Hart, and D. Stork. *Pattern Classification and Scene Analysis 2nd ed.*, chapter 2, pages 128–138. 1995.

[7] A. Friberg. A fuzzy analyzer of emotional expression in music performance and body motion. *Proceedings of Music and Music Science*, i:1–13, 2004.

[8] A. Friberg, R. Bresin, and J. Sundberg. Overview of the KTH rule system for musical performance. *Advances in Cognitive Psychology*, 2(2):145–161, Jan. 2006.

[9] A. Gabrielsson and E. Lindström. The influence of musical structure on emotional expression. In P. N. Juslin and J. A. Sloboda, editors, *In Music and Emotion: Theory and Research.*, pages 367–400. New York: Oxford University Press., 2010.

[10] N. Gillian, R. Knapp, and S. O'leModhrain. An adaptive classification algorithm for semiotic musical gestures. *the 8th Sound and Music Computing Conference*, 2011.

[11] N. Gillian, R. Knapp, and S. O'Modhrain. A machine learning toolbox for musician computer interaction. In *Proceedings of the 2011 International Conference on New Interfaces for Musical Expression (NIME11)*, 2011.

[12] A. Hadjakos, E. Aitenbichler, and M. Mühlhäuser. Potential use of inertial measurement sensors for piano teaching systems: Motion analysis of piano playing patterns. In *Proceedings of the 4th i-Maestro Workshop on Technology-Enhanced Music Education*, pages 61–68, 2008.

[13] K. Hevner. Experimental studies of the elements of expression in music. *American Journal of Psychology.*, 48(2):246–268, July 1936.

[14] A. Höfer, A. Hadjakos, and M. Mühlhäuser. Gyroscope-based conducting gesture recognition. *NIME09*, 2009.

[15] P. Juslin. Emotional communication in music performance: A functionalist perspective and some data. *Music perception*, pages 383–418, 1997.

[16] P. Juslin. Cue utilization in communication of emotion in music performance: Relating performance to perception. *Journal of Experimental Psychology: Human perception and performance*, 26(6):1797, 2000.

[17] P. Juslin and R. Timmers. Expression and communication of emotion in music performance. In P. N. Juslin and J. A. Sloboda, editors, *In Music and Emotion: Theory and Research.*, pages 454–489. New York: Oxford University Press., 2010.

[18] D. M. Powers. Evaluation: From precision, recall and f-factor to roc, informedness, markedness & correlation. *School of Informatics and Engineering, Flinders University, Adelaide, Australia, Tech. Rep. SIE-07-001*, 2007.

[19] G. Read. *Music notation: A manual of modern practice*, chapter 16. Taplinger Publishing Company, 1979.

[20] J. A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology.*, 39(6):1161–1178, November 2003.

[3] http://www.ossur.com/

[4] http://mue.music.miami.edu/