*Proceedings of the 2nd International Conference on Multimedia and Human-Computer Interaction*
*Prague, Czech Republic, August 14-15, 2014*
*Paper No. 130*

# Musical Query by Movement

**Jay Clark, Mason Bretan, Gil Weinberg**
Georgia Institute of Technology
840 McMillan St., Atlanta, GA
jason.clark@gatech.edu; pbretan3@gatech.edu; gilw@gatech.edu

***Abstract***- As technology advances, new concepts of music consumption and retrieval continue to emerge. Many of these retrieval systems have moved beyond the traditional title and artist query methods and instead utilize higher level descriptors of music. We introduce Query By Movement (QBM), a music retrieval system which exploits the connection between music and body motion allowing users to query for music based on what they physically feel rather than what they may be able to articulate. The system analyzes a user's head and body movements using computer vision and plays an appropriate song suggested by the movement. We first explore the concept of QBM by investigating the analogous features that can be extracted from both video and audio signals. We select tempo and energy as candidate features and perform a user study which demonstrates a significant correlation between the extracted features in the movement and audio domains. We describe an architecture for leveraging the correlation through a QBM system and evaluate our system system with another user study. Our results indicate that our features and our system yield more favorable music retrieval than randomized queries.

***Keywords:*** Information retrieval, Music recommendation.

## 1. Introduction

Music retrieval systems have evolved greatly from the original title and artist query methods. Innovation in this field has allowed users to easily query and consume music in virtually any number of settings or circumstances, whether it be at home on a computer, at the gym, or driving in a car. Some systems utilize automatic playlist generation based on mood (Songza, Stereomood) or similarity (iTunes Genius, Spotify Radio, Pandora) allowing users to query general types of music rather than specific songs. Other systems, such as Query by Humming, allow the user to search for specific songs when the title or lyrics are unknown Ghias et al. (1995). Though several types of retrieval systems exist, each similarly leverages a specific characteristic or piece of information regarding the music in order to provide the user with an appropriate response to their query. Here we introduce a new music retrieval system which exploits the connection between music and dance and allows users to query for music based on what they physically feel rather than what they may be able to articulate through words or sonically express. We call this new system Query by Movement (QBM).

We believe that dancing and rhythmic body movements can be an effective and intuitive means of music retrieval. Studies suggest a strong cognitive link between music and body motion, likely due to their "logical association... as temporally organized art forms" Lewis (1988). Music is shown to promote motor responses in both children and adults that is often analogous to the music Mitchell & Gallaher (2001), Phillips-Silver & Trainor (2005). The connections between music and dance can be recognized by observers as well Krumhansl & Schenck (1997), even when the two are temporally separated Mitchell & Gallaher

(2001). Sievers' findings suggest that "music and movement share a common structure that affords equivalent and universal emotional expressions" Sievers et al. (2013). Gazzola observed that the neurons responsible for firing when performing an action also fire when hearing an audio representation of that action Gazzola et al. (2006). The connection is further supported by the historical interdependent development of music and dance Grosse (1897).

This connection is most often utilized to dance appropriately when listening to music. Some projects have explored the reverse relationship, where dancers generate analogous music with their movement. This paradigm has been investigated using worn hardware, held hardware, external sensors, and computer vision in a musical performance context Paradiso & Hu (1997), Paradiso & Sparacino (1997), Camurri et al. (2000), Aylward & Paradiso (2006), Winkler (1998) Samburg et. al's iClub furthered this investigation in a social music consumption context Samberg et al. (2002).

The aforementioned projects have been able to report "rich mappings that directly reflected the movement of the dancer"Paradiso & Hu (1997) because music and dance share a vocabulary of features which are descriptors for both art forms. These include ideas such as rhythm, tempo, energy, fluidity, dynamics, mood, and genre. We believe that by using the congruent features of dance to query for music, we can leverage the cognitive link between music and dance and bypass the deciphering and articulation necessary to query used by other methods.

In this paper we explore the idea of QBM by first examining the correlation between dance and music and the features that can be extracted from each. We then describe an architecture for leveraging the correlation through a QBM system. Finally, we evaluate this system with a user study.

## 2. Feature Extraction

In order to leverage the congruent features of dance to query for music, we must be able to extract and demonstrate a correlation between these features in both the movement and audio domains. A few candidate data acquisition systems were vetted from related works, including worn sensors, external sensors, and computer vision. We chose a computer vision implementation on the iPhone because of its ubiquity and relevance to our robotic musical companion robot, which uses a smart-phone for all of its sensing and higher order computation Bretan & Weinberg (2014).

### 2.1. Data Acquisition: Computer Vision

Motion tracking in computer vision can be performed in a number of ways. Most involve a trade-off between computational ease and degrees of freedom. For example, a powerful means of motion tracking involves object recognition, in which a classifier is trained to detect features, such as a face, hands, or facial features. This could be leveraged to track and consider different body parts individually. Conversely, a simple means of motion tracking is frame differencing, in which pixels in corresponding frames are subtracted from one another, and their difference is summed. This algorithm is inexpensive and robust, but can only supply a generalized view of the movement. We attempted to optimize this trade off by leveraging sparse optical flow, the tracking of a number of arbitrary pixels from frame to frame. This algorithm is cheaper than object detection and tracking, but allows more degrees of freedom than frame differencing, should we need them.

We use the Open Source Computer Vision (OpenCV) library's implementation of the sparse pyramidal version of the Lucas-Kanade optical flow algorithm, as detailed in Bouguet (2001). Because of the limitations of the iPhone's hardware, an optimization must be made between frame rate and the number of image features we can track with the optical flow. As a result, we initialize our image processing by running Shi and Tomasi's "Good Features to Track" corner detection algorithm Shi & Tomasi (1994) to identify the 100 strongest corners in the first frame.

We compute the optical flow for these 'good features' on each subsequent frame and generate candidate

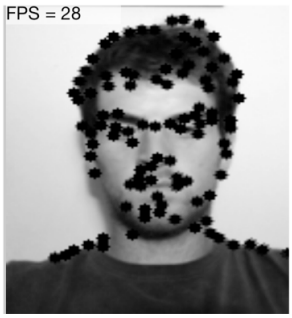


Fig. 1. Screenshot of the optical flow tracking.

feature vectors:

$$\sum_{i=1}^{N} |s[i]| \tag{1}$$

$$\sum_{i=1}^{N} |s_x[i]| \tag{2}$$

$$\sum_{i=1}^{N} |s_y[i]| \tag{3}$$

$$\sum_{i=1}^{N} s_x[i] \tag{4}$$

$$\sum_{i=1}^{N} s_y[i] \tag{5}$$

### 2.2. Feature Selection

The EchoNest Jehan et al. (2010) is a music intelligence platform whose API grants access to over a billion MIR data points about millions of songs. Furthermore, the platform allows an application to query for music by defining specific ranges of these data points. Many of these calculated features are analogous to dance, including genre, mood, tempo, and energy. Given our data acquisition system, we do not believe we can deterministically extract genre and mood from the optical flow of arbitrarily tracked pixels. Therefore, we selected tempo and energy as a 2-D feature space to map from movement to music query.

### 2.3. User Study

We designed a user study in order to validate our feature selection and optimize the algorithms used to compute them. The study aimed to demonstrate a correlation between calculated data and Echo Nest supplied labels.

#### 2.3.1. Method

We used the Echo Nest to query for nine songs at each of three tempos: 90, 120, and 150 bpm. The songs varied in energy. We attempted to control variables that may also contribute to higher-energy dancing by holding constant other Echo Nest-supplied proprietary features such as danceability, artist familiarity,

and song "hotttnesss." We allowed genre to be randomly selected. The research subjects were asked to bob their head to 9 short song segments in front of the iPhone's camera. Three songs of varying energy were randomly selected per tempo, and the three tempo classes were presented in a random order. For each dance, we computed the optical flow of their movement and generated the five candidate data vectors detailed in section 2.1.

### 2.3.2. Results

Twenty Georgia Tech students participated in the study. We analyzed results for both the tempo and energy features.

Tempo correlates strongly and explicitly between music and dance. Extracting the periodicity from a time-varying signal is a well-defined research problem Gouyon et al. (2006). We found that combining the data vectors (1) (2) and (3) minimized the error rate.

We perform an autocorrelation on these vectors and peak-pick the three strongest correlations per vector. This provides the lag indices which can be multiplied by the mean time elapsed between frames, $\delta$ T, to arrive at a candidate time per beat. We then convert this to beats per minute and scale the value to fall within the range of 77-153 bpm, and finally round it to the nearest 10 bpm. Each data vector yields 3 tempo candidates, and the mode of the candidates is selected as the tempo induction of a dance signal.

Using this method, we were able to detect the correct tempo in 88.33% of dances from the user study. Of the 11.77% error, half were off by just 10 bpm.

Energy is a more complicated feature in both the movement and audio domains. Energy is a proprietary Echo Nest feature (trained using hand-taggings) which quantifies the following: "Does [the song] make you want to bop all over the room, or fall into a coma?" Sundram (2013). We hypothesized that people tend to move in quicker, more jagged movements when dancing to high-energy music. We expected to see smoother accelerations and decelerations in lower - energy movement. Grunberg found a similar correlation between movement and emotion. When dancing to angrier music, gestures tend to "only take up about two-thirds of the beat" Grunberg et al. (n.d.).

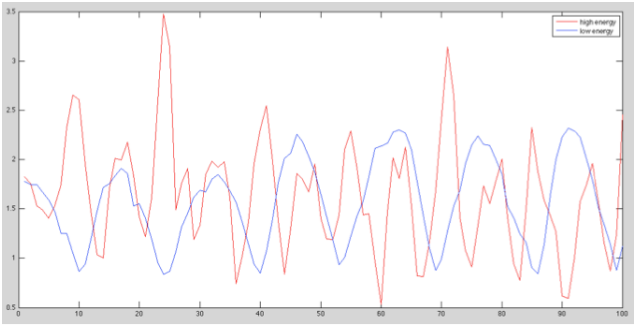


Fig. 2. Movement vector for a high and low energy song at the same tempo.

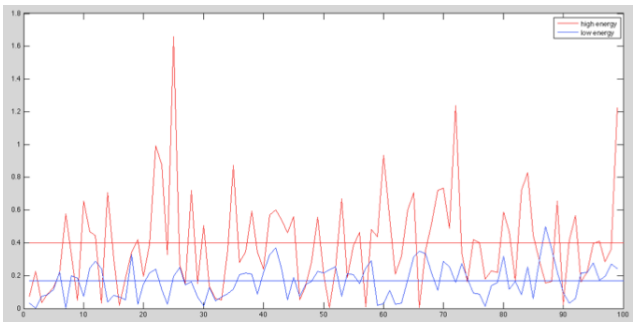


Fig. 3. Mean absolute value of the first-order difference of the data in figure 2.

We found our hypothesis to be represented in the data. Figure 2 depicts a subject's movement magnitude vector when dancing to a low energy song versus a high energy song at the same tempo. Notice the tall, thin spikes in velocity in the high-energy movement as compared to the wider, more gradual velocity fluctuations in the low energy song.

We found that energy is proportional to the mean absolute value of the first order difference of our movement vector. Figure 3 depicts the absolute value of the first-order difference of the above data. The

mean is superimposed on top and represents our final calculated energy value.

The study results indicate a significant correlation between calculated energy and Echo Nest labeled energy at each tempo, as illustrated in Figure 4 . When considering each individual subject's data separately, the correlations became more significant, as illustrated in Figure 5. Relatedly, we also found that the research subjects generally responded to an increase or decrease in song energy with a corresponding increase or decrease in their dance energy. This correlation was strong and significant, as illustrated in Figure 6.

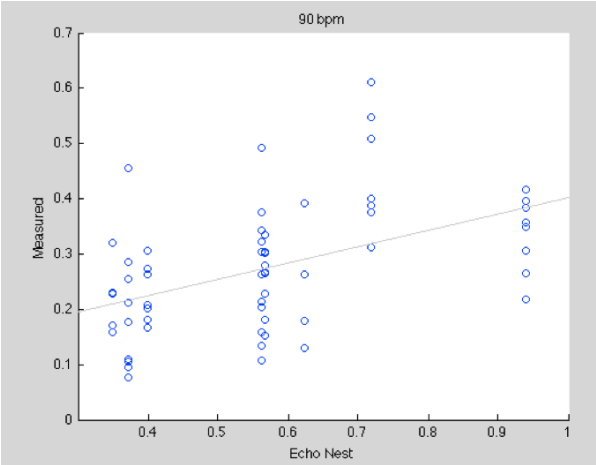


Fig. 4. Measured energy vs. Echo Nest supplied energy at 90 BPM. r = 0.4784, p<0.001

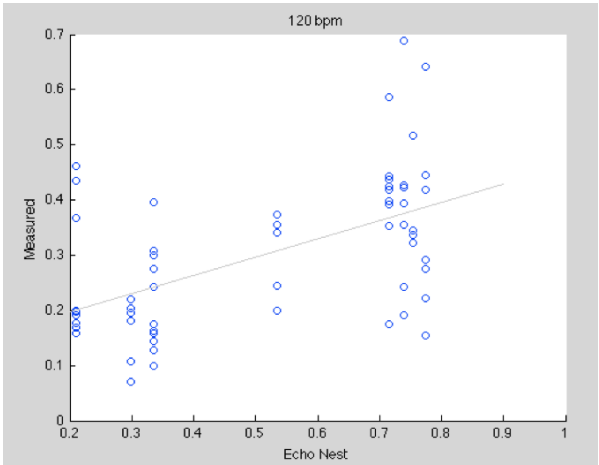


Fig. 5. Measured energy vs. Echo Nest supplied energy at 120 BPM. r = 0.4845, p<0.001

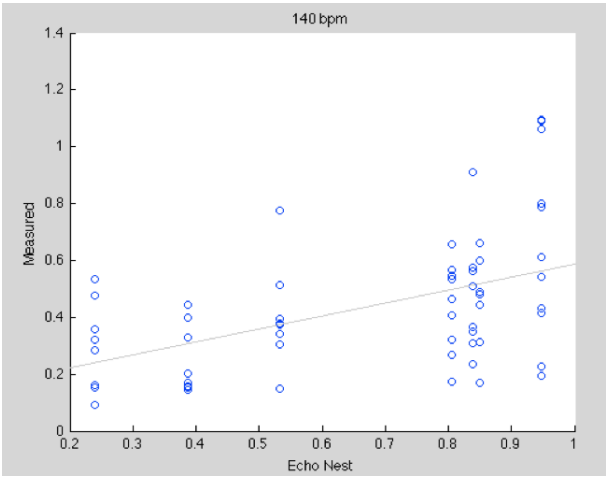


Fig. 6. Measured energy vs. Echo Nest supplied energy at 140 BPM. r = 0.5484, p<0.001

### 2.3.3. Discussion

The results of the study suggest an explicit correlation of tempo in the audio and movement domains. While the energy results suggest a correlation as well, the stronger individual correlations may suggest that scaling is an issue. For example, one user's medium energy movement may be similar to another user's high-energy movement.

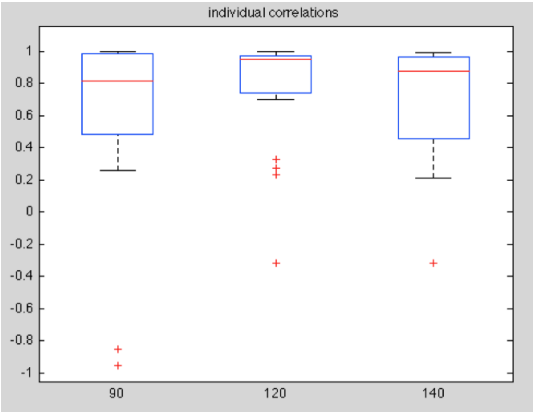


Fig. 7. Individual Correlations.

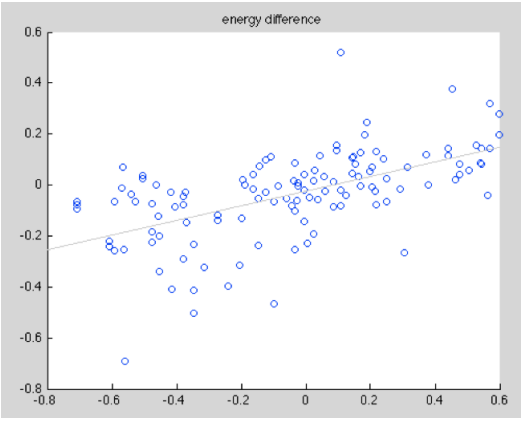


Fig. 8. Labeled vs. extracted changes in energy. r=0.57, $p<0.001$

## 3. Implementation

Query by Movement is implemented as an application for the iPhone 5. It uses the front-facing camera to capture frames at 30 fps and leverages the OpenCV implementation of the sparse pyramidal Lucas-Kanande optical flow algorithm [2] initialized with strong corner detection [8] in order to generate movement vectors. We extract energy and tempo information from these vectors using the algorithms optimized in Study 1 (Section 2.3). We use this two-dimensional feature space to query for music using libEchoNest, Echo Nest's iOS library for the API. Specifically, we query for songs within (+/-) 5 bpm of our extracted tempo and within (+/-) .1 of our extracted energy (out of 1). Our query also includes a minimum danceability – an Echo Nest feature which is closely related to beat strength and tempo stability – of 0.5 (out of 1). Finally, minimum artist familiarity and "song hotttnesss" were set at 0.5 (out of 1) under the assumption that the user would want to query for something familiar. The Echo Nest query returns an array of songs which match the query. The array contains information about each song, including a Spotify URL. We select a song at random and play it using the libSpotify API.

## 4. Evaluation

We validated our feature selection and our system by performing a subjective user study.

### 4.1. Method

Research subjects were asked to use the Query By Movement system to query for music. After extracting tempo and energy, the system performed each query in one of four ways:

1. Query using detected tempo and detected energy
2. Query using random tempo and detected energy
3. Query using detected tempo and random energy
4. Query using random tempos and energies

Each research subject queried 8 songs and the system responded in each of the four ways listed above in a random order. After hearing the system's response, the subject recorded a rating on a scale of 1-5, with 5 being the most appropriate. Each subject experienced all query conditions twice.

### 4.2. Results

10 graduate Georgia Tech students participated in the study. A one-way between subjects analysis of variance (ANOVA) was conducted to compare the effect of the different testing conditions. The standard query performed significantly better than the randomized queries, as depicted in Figure 7.

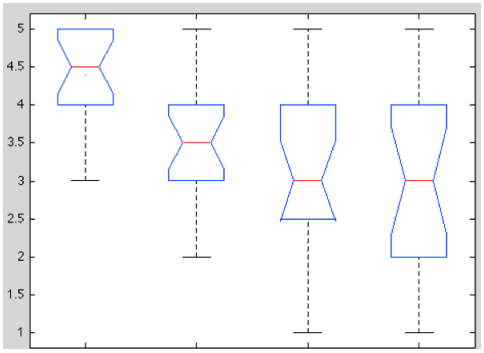


Fig. 9. ANOVA of the 1-4 testing conditions on x-axis and the user rating on y-axis.

### 4.3. Discussion

The results from this user study indicate that our calculated features contribute to an appropriate query of a Query By Movement system. Because appropriateness is a measure related to expectation, our results may suggest that our features are transparent across the audio and movement domains. We might also speculate that with more study participants, we may find that energy more strongly contributes to an appropriate query than tempo does

## 5. Future Work and Conclusions

A natural and logical continuation of investigating a Query By Movement system is the construction of a higher-dimensional feature space. In order to achieve this, new features must be able to be extracted from both the movement and audio domains. Features such as mood and genre may contribute to a more compelling system. The Microsoft Kinect is capable of sophisticated video feature tracking while remaining robust and somewhat ubiquitous. Tracking of individual body parts could possibly grant enough degrees of freedom to begin considering machine learning as a means to classify mood or genre in dance.

## References

Aylward, R. & Paradiso, J. A. (2006), Sensemble: a wireless, compact, multi-user sensor system for interactive dance, *in* 'Proceedings of the 2006 conference on New interfaces for musical expression', IRCAM-Centre Pompidou, pp. 134–139.

Bouguet, J.-Y. (2001), 'Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm', *Intel Corporation* **5**.

Bretan, M. & Weinberg, G. (2014), Chronicles of a robotic musical companion, *in* 'Proceedings of the 2014 conference on New interfaces for musical expression', University of London.

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

Gazzola, V., Aziz-Zadeh, L. & Keysers, C. (2006), 'Empathy and the somatotopic auditory mirror system in humans', *Current biology* **16**(18), 1824–1829.

Ghias, A., Logan, J., Chamberlin, D. & Smith, B. C. (1995), Query by humming: musical information retrieval in an audio database, *in* 'Proceedings of the third ACM international conference on Multimedia', ACM, pp. 231–236.

Gouyon, F., Klapuri, A., Dixon, S., Alonso, M., Tzanetakis, G., Uhle, C. & Cano, P. (2006), 'An experimental comparison of audio tempo induction algorithms', *Audio, Speech, and Language Processing, IEEE Transactions on* **14**(5), 1832–1844.

Grosse, E. (1897), *The beginnings of art*, Vol. 4, D. Appleton and Company.

Grunberg, D. K., Batula, A. M., Schmidt, E. M. & Kim, Y. E. (n.d.), 'Affective gesturing with music mood recognition'.

Jehan, T., Lamere, P. & Whitman, B. (2010), Music retrieval from everything, *in* 'Proceedings of the international conference on Multimedia information retrieval', ACM, pp. 245–246.

Krumhansl, C. L. & Schenck, D. L. (1997), 'Can dance reflect the structural and expressive qualities of music? a perceptual experiment on balanchine's choreography of mozart's divertimento no. 15', *Musicae Scientiae* **1**(1), 63–85.

Lewis, B. E. (1988), 'The effect of movement-based instruction on first-and third-graders' achievement in selected music listening skills', *Psychology of Music* **16**(2), 128–142.

Mitchell, R. W. & Gallaher, M. C. (2001), 'Embodying music: Matching music and dance in memory', *Music Perception* **19**(1), 65–85.

Paradiso, J. A. & Hu, E. (1997), Expressive footwear for computer-augmented dance performance, *in* 'Wearable Computers, 1997. Digest of Papers., First International Symposium on', IEEE, pp. 165–166.

Paradiso, J. & Sparacino, F. (1997), 'Optical tracking for music and dance performance', *Optical 3-D Measurement Techniques IV* pp. 11–18.

Phillips-Silver, J. & Trainor, L. J. (2005), 'Feeling the beat: Movement influences infant rhythm perception', *Science* **308**(5727), 1430–1430.

Samberg, J., Fox, A. & Stone, M. (2002), iclub, an interactive dance club, *in* 'ADJUNCT PROCEEDINGS', p. 73.

Shi, J. & Tomasi, C. (1994), Good features to track, *in* 'Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on', IEEE, pp. 593–600.

Sievers, B., Polansky, L., Casey, M. & Wheatley, T. (2013), 'Music and movement share a dynamic structure that supports universal expressions of emotion', *Proceedings of the National Academy of Sciences* **110**(1), 70–75.

Sundram, J. (2013), 'Danceability and energy: Introducing echo nest attributes'.

Winkler, T. (1998), Motion-sensing music: Artistic and technical challenges in two works for dance, *in* 'Proceedings of the International Computer Music Conference'.