2010 IEEE-RAS International Conference on Humanoid Robots
Nashville, TN, USA, December 6-8, 2010

Interactive Musical Participation with Humanoid Robots Through the use of Novel Musical Tempo and Beat Tracking Techniques in the Absence of Auditory Cues

Daniel M. Lofaro, Paul Oh, JunHo Oh, Youngmoo Kim

Abstract— **An interactive participant in a live musical performance requires a multitude of senses in order to perform. These senses include hearing, sight, and touch. Our long-term goal is to have our adult size humanoid robot Jaemi Hubo be an interactive participant in a live music ensemble. This work focuses on Jaemi Hubo's sense of sight as it pertains to a musical environment. Jaemi Hubo's musical awareness is increased through the use of novel musical tempo and beat tracking techniques in the absence of auditory cues through the use of computer vision and digital signal processing methods. Real time video frames of subjects moving to a regular beat is recorded using Jaemi Hubo's video capture device. For each successive video frame, an optic flow algorithm is implemented to discern the direction and magnitude of motion. A Fast Fourier Transform is then applied to obtain the spectral content of the motion data. Next, a Gaussian weight centered on the average musical tempo is applied to the normalized spectrum. The resulting maxima of the weighted spectrum is the tempo calculated from the video frames. A tempo based dynamic threshold of the first derivative of the motion data was used to find the beat location. Experiments using OpenCV, and Matlab produced accurate tracking of the true tempo and beat timing in the captured video.**

I. INTRODUCTION

"There seems to be an inherent disconnect between robotics and music" according to Johnathan Strickland, reporter for Discovery Communications. The Honda ASIMO performed with the Detroit Symphony Orchestra in May 2008 at the *Power of Dream Music Education Fund initiative* event. The piece played was *The Impossible Dream* from the musical *Man of La Mancha*. ASIMO's role was to conduct the entire orchestra. In preparation for the event, ASIMO studied the movements of Charles Burke, education director for the Detroit Symphony, while he conducted *The Impossible Dream* to a pianist. Six months later ASIMO successfully conducted the Detroit Symphony Orchestra. Though the event was impressive "some people might argue that ASIMO didn't really direct the Detroit Symphony Orchestra – rather, Burke did. After all, ASIMO was really recreating Burke's style and movements. The entire routine was simply a program running on ASIMO's operating system. The robot just ran through the series of motions from start to finish, and would have continued regardless of whether or not the orchestra followed the robot's lead." - Strickland.

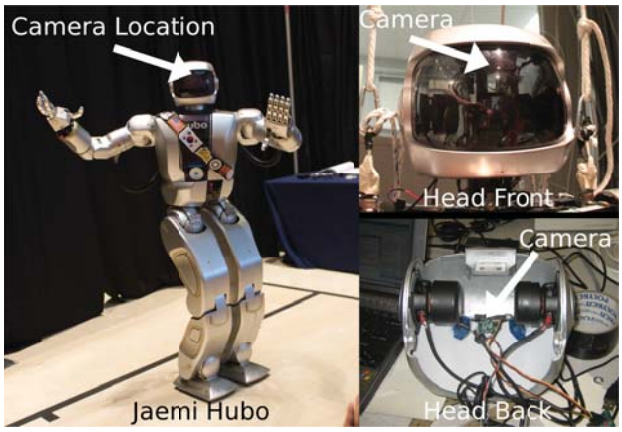

Fig. 1. (LEFT) Jaemi Hubo, 130cm tall 41 degree of freedom humanoid robot, full body view and camera location. (TOP RIGHT) Jaemi Hubo front head view and camera location. (BOTTOM RIGHT) Inside Jaemi Hubo's head and camera location.

This blind autonomy, i.e. lack of world feedback, is a cause of the inherent disconnect between robots and music. This is a common problem of robots when faced with creative human-robot interaction. Traditionally, *auditory feedback* is the mechanism of choice for interactive musical robots. The autonomous dancing humanoid (ADH) robot by DASL[1] and MET-lab[2] and the work done by Michalowski et al.[1] with the Keepon[3] are two examples of robotic platforms that intelligently interact with music [1], [2]. Both the ADH and Keepon listen to music and track the beat and tempo in real-time using auditory cues, then dance to the music.

Music that is best tracked from audio in real-time requires large magnitude changes and a constant beat present in the music [3]. Consequently pop and other similar musical styles track well, however, non-percussive melodic music, such as

This project was supported by the Drexel Autonomous Systems Lab (DASL) and the Music Entertainment Technology (MET) Lab.

D. Lofaro is a Ph.D. Candidate with the department of Electrical and Computer Engineering, Drexel University, Philadelphia, PA 19104, USA. dml46@drexel.edu

P. Oh is with the Department of Mechanical Engineering & Mechanics, Drexel University, Philadelphia, PA 19104, USA. paul@coe.drexel.edu

J. Oh director of Hubo Lab at the Korean Advanced Institute of Science and Technology, Daejeon, South Korea. jhoh@kaist.ac.kr

Y. Kim is with the Electrical and Computer Engineering Department, Drexel University, Philadelphia, PA 19104, USA. ykim@drexel.edu

[1]DASL: Drexel Autonomous Systems Lab, Drexel University, Philadelphia, PA

[2]MET-lab: Music Entertainment Technology Lab, Drexel University, Philadelphia, PA

[3]Created by Hideki Kozima: Creator of Keepon while at the National Institute of Information and Communications Technology (NICT) in Kyoto, Japan.

most orchestral pieces, does not. In orchestral performances a conductor can keep the beat during periods of absent auditory cues. One reason humans are able to accurately track the beat and tempo present in a multitude of different musical styles is because we can utilize our senses of hearing, sight and touch to do so. Our overarching goal is to make our adult size (130cm) humanoid robot Jaemi Hubo (see Fig. 1) musically aware and have it become an interactive participant in a live *human* ensemble. Bringing our goal to fruition means that all senses must be utilized.

This work focuses on Jaemi's sense of sight as it pertains to a musical environment. A novel approach to tempo and beat tracking in the absence of auditory cues through the use of computer vision and digital signal processing methods is used to give Jaemi Hubo an additional component (sight) to musical awareness. The approach is called visual beat tracking (VBT). Experiments showed that when equipped with the VBT Jaemi Hubo is able to accurately track the beat and tempo when watching a trained musician conduct to a steady tempo with its on board camera, see Fig. 1. Experiments were performed at multiple measure timings/meters. Fig. 2 shows the conducting pattern for the different timings/meters.

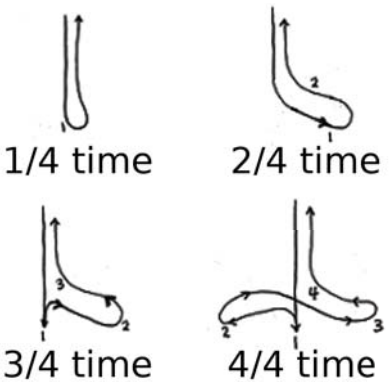

Fig. 2. Diagram of conducting timings/meters used. For each different timing the conductors hand follows the arrow. The conductors hand will be where the number is located at the beginning of each beat. The conductors hand starts at the beginning of the arrow at the beginning of each measure.

II. RELATED WORK

Using the sense of sight to perceive the ictus[4] (instant at which the beat occurs) has been shown to be effective. Park et al. [4] demonstrated vision system that is able to estimate the period of an unstructured beat gesture expressed in any part of the viewable scene. This system used the Lucas-Kanade algorithm [5] to calculate the optical flow. This locates the regions in the scene (between two consecutive frames) that contain relative movement. The trajectory of the center of gravity of the optical flow was used to influence the sound of a piano by actuating robotic fingers. Crick et al. [6] created a system that allowed the humanoid robot Nico to play a drum in concert with human drummers. Color target tracking was used to track the conductor's hand. Nico used both sight and sound to enable precise synchronization in performing its task.

The Lucas-Kanade algorithm used by Park et al. to calculate the optical flow requires a high contrast subject to be tracked limiting its effectiveness in a live *human* band. The Nico system is limited because it uses a simple threshold on only the versicle position of the conductor's hand. Crick et al. states that their system is only affective when the gestures are smooth and regular. The VBT presented in this work does not contain these limitations due to the robust and dynamic methods used to calculate the tempo and ictus.

[4]Ictus: "bottom" of the beat gesture where it changes vertical direction

III. METHODOLOGY

Our objective is to calculate the tempo and beat timing (ictus) directly from Jaemi Hubo's single camera live video feed. From musical conducting or rhythmic action the tempo and beat timing can be calculated by examining the spectral content of the mean magnitude and angle of motion in a given scene M_N (across two or more frames). To calculate M_N a real-time video feed (gray scale - intensity) must be captured. Each frame in the video is equalized to compensate for different lighting conditions. The size of the image is reduced to allow for faster computation time. The computer vision (CV) Horn-Schunck Optical Flow method is used to calculate the relative motion between the current frame and the previous frame [7][8]. The mean magnitude and angle moved between the two frames (M_N) is then calculated. The vector $\overline{V_N}$ containing the past N_{fft} values of M_N is created, where N_{fft} is the length of the forthcoming fast Fourier transform (FFT). The spectral content of $\overline{V_N}$ is calculated via the use of the FFT creating $\overline{S_{V_N}}$. To reduce the effects of harmonics a triangle piecewise weight is applied to $\overline{S_{V_N}}$ with the mean tempo for popular music as calculated by Moelants [9] receiving the greatest weight. The tempo corresponding to the location of the highest peak in the weighted $\overline{S_{V_N}}$ is the tempo contained in the captured video. The tempo is normally measured in beats per minute (BPM).

The beat timing (location of each beat or ictus) is calculated by examining the first derivative of the mean angle between the current and the previous frame. The time instance where the resulting value becomes less than the dynamic beat marking threshold (DBMT or B_N^c) marks the beginning of the beat. B_N^c is a function of the current calculated tempo that increases the probability that the beat will be detected at the next ictus by decreasing its magnitude as the likelihood of the next ictus increases.

The block diagram of the process to calculate the beat timing and the tempo contained in the captured video can be found in Fig. 3.

A. Movement Extraction

The Horn-Schunck optical flow algorithm is used to find the direction and magnitude of motion in each frame set I_N and I_{N-1}[7]. The Horn-Schunck algorithm calculates the flow velocity (u, v) of each point in the frame set.

$$\frac{dE}{dt} = E_x u + E_y v + E_t = 0 \tag{1}$$

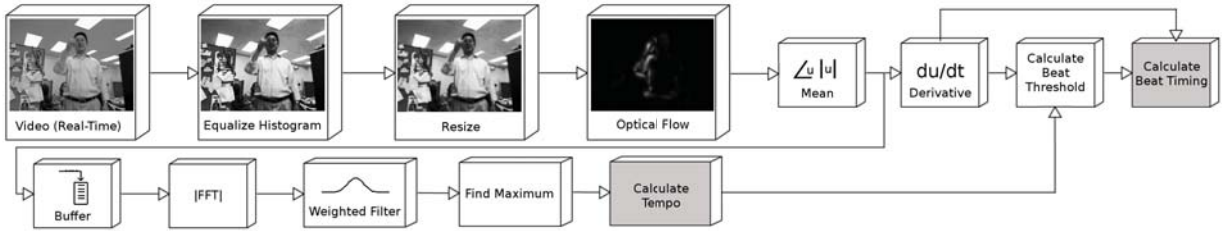

Fig. 3. Video taken from a real-time feed. Each frame is equalized to compensate for different lighting conditions. Current image I_N is resized (reduced in size) to allow for faster computation time. The optical flow is taken between the current frame I_N and the previous frame I_{N-1}. The resulting data consists of the magnitude and angle the each pixel moved from I_{N-1} to I_N. The magnitude and angle is averaged over the entire image resulting in a mean magnitude and angle moved for the image set I_N and I_{N-1}. This is referred to as the movement data. The FFT of the past M values of the movement data is calculated. A weighted filter is placed over the FFT spectrum to give higher weight to the most common tempos[9]. The maxima of the FFT spectrum correlates to the overall tempo present in the video.

where

$$u = \frac{dx}{dt} \quad and \quad v = \frac{dy}{dt} \tag{2}$$

$E(x, y, t)$ denotes the image brightness at point (x, y) in the image plane at time t. E_x, E_y and E_t are the partial derivatives for the image brightness in respect to x, y, and t respectively. The constraints on the flow velocity can now be defined as:

$$(E_x, E_y) \cdot (u, v) = -E_t \tag{3}$$

Running in discrete (frame by frame) time, our annotation and calculations must be modified. Let $E_{j,k,N}$ be the brightness of the pixel located on the j^{th} row, k^{th} column on the N^{th} image. j spans from 0 to the width of the image frame. k spans from 0 to the height of the image frame. The top left corner is the origin $(0, 0)$. The bottom right corner is the max width and height (j_{max}, k_{max}). Horn *et al.*[7] determined that E_x, E_y, and E_t can be estimated by:

$$\begin{aligned} E_x \approx \frac{1}{4}(&E_{j,k+1,N} - E_{j,k,N} + \\ &E_{j+1,k+1,N} - E_{j+1,k,N} + \\ &E_{j,k+1,N+1} - E_{j,k,N+1} + \\ &E_{j+1,k+1,N+1} - E_{j+1,k,N+1}) \end{aligned} \tag{4}$$

$$\begin{aligned} E_y \approx \frac{1}{4}(&E_{j+1,k,N} - E_{j,k,N} + \\ &E_{j+1,k+1,N} - E_{j,k+1,N} + \\ &E_{j+1,k,N+1} - E_{j,k,N+1} + \\ &E_{j+1,k+1,N+1} - E_{j,k+1,N+1}) \end{aligned} \tag{5}$$

$$\begin{aligned} E_t \approx \frac{1}{4}(&E_{j,k,N+1} - E_{j,k,N} + \\ &E_{j+1,k,N+1} - E_{j+1,k,N} + \\ &E_{j,k+1,N+1} - E_{j,k+1,N} + \\ &E_{j+1,k+1,N+1} - E_{j+1,k+1,N}) \end{aligned} \tag{6}$$

The non-thresholding methods used by the Horn-Schunck optical flow algorithm allows the system to be more tolerant of low contrast and poor lighting conditions when compared to other optical flow algorithms.

B. Tempo Extraction

The tempo can be extracted by finding the maximum peak in the FFT spectra vector $\overline{S_{V_N}}$ for the mean motion data vector $\overline{V_N}$ where:

$$\overline{S_{V_N}} = \text{FFT}(\overline{V_N}) \tag{7}$$

The peak will only be calculated for index values from 0 to $\frac{N_{fft}}{2}$ where N_{fft} is the FFT length. The tempo of the peak x_p is calculated using the index value of the peak i_p and Eq. 8.

The tempo x_i can be calculated in beats per minute (BPM) for any index point i from the FFT using:

$$x_i = \frac{60 \cdot i \cdot fps}{N_{fft}} \tag{8}$$

Where i = FFT index point and fps = Frames Per Second of the video feed. Fig. 4 shows the plot of $\overline{S_{V_N}}$ vs. x_i where $fps = 29.97$, $N_{fft} = 256$, and $N = 500$. The index point of the maximum value of Fig. 4 i_p is found by finding the maxima of the vector $\overline{S_{V_N}}$ and the corresponding index point. Eq. 8 and i_p are used to calculate the tempo x_T in the video feed:

$$tempo = x_T = \frac{60 \cdot i_p \cdot fps}{N_{fft}} \tag{9}$$

$\overline{S_{V_N}}$ can contain non-tempo related frequencies due to harmonics or other visual stimuli. Applying a piecewise tempo weight $\overline{T_W}$ to $\overline{S_{V_N}}$ reduces the effects of the non-tempo related frequencies. $\overline{S^W_{V_N}}$ is the weighted $\overline{S_{V_N}}$ vector and is described in Section III-C. $\overline{S^W_{V_N}}$ can be used as a direct replacement for $\overline{S_{V_N}}$ when the tempo weight is needed.

C. Tempo Weight

To reduce the effects of harmonics and to ensure the location of the highest peak in $\overline{S_{V_N}}$ depicts the correct tempo a weight is applied. The each value of the weight vector $\overline{T_w}$ is different for each discrete tempo value. T_w has the same length as $\overline{S_{V_N}}$. $\overline{T_w}$ is a triangular piecewise weight that

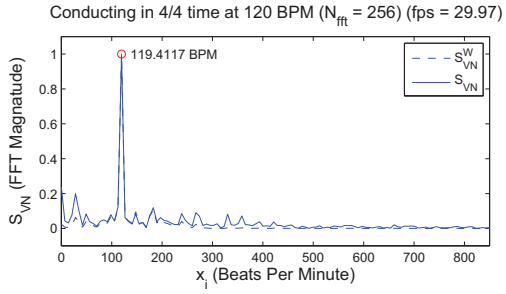

Fig. 4. Trained musician conducting in 4/4 time at 120 BPM. Plot of $\overline{S_{V_N}}$ vs. x_i and $\overline{S^W_{V_N}}$ vs. x_i where $fps = 29.97$, $N_{fft} = 256$, and $N = 500$. The index point of the maximum value of $i_p = 18$. The video feed is from a trained musician conducting in 4/4 time at 120 BPM. The tempo is calculated from Eq. 8 where $i = i_p$. The tempo calculated is $x_{i_p} = x_p = x_T = 119.4$ BPM, 0.5% diference from the true tempo 120 BPM.

peaks at the mean tempo T_m of popular music. D. Moelants calculated $T_m \approx 128BPM$ [9]. $\overline{T_w}$ is calculated by:

$$T_{w_i}(x_i) = \begin{cases} m \cdot x_i + b_1 & \text{if } x_i \leq T_m \\ -m \cdot (x_i - T_m) + b_2 & \text{if } T_m < x_i \leq 2T_m \\ b_1 & \text{if } 2T_m < x_i \leq x_{N_{fft}/2} \\ 0 & \text{else} \end{cases} \tag{10}$$

where i is the index corresponding to x_i defined in Eq. 8. i is all integer values from 0 to $(N_{fft} - 1)$.

$$\overline{T_w} = [T_{w_0}(x_0), T_{w_1}(x_1), \cdots, T_{w_{N_{fft}-1}}(x_{N_{fft}-1})] \tag{11}$$

and

$$m = K_{max} - K_{min} \tag{12}$$

$$b_1 = K_{min} \tag{13}$$

$$b_2 = (2 - T_m) \cdot K_{min} + (1 + T_m) \cdot K_{max} \tag{14}$$

Where K_{min} and K_{max} are the minimum and maximum desired gain values for T_{w_i}.

The plot of $\overline{T_w}$ for multiple tempos x_i is shown in Fig. 5.

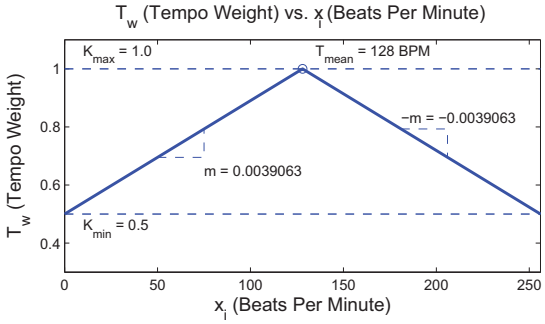

Fig. 5. Tempo weight $\overline{T_w}$ reduces the effects of non-tempo related spectral peaks in S_{V_N}. T_w is defined in Eq. 10.

The weighted spectrum $\overline{S^W_{V_N}}$ is calculated by:

$$\overline{S^W_{V_N}} = \overline{S_{V_N}} \circ \overline{T_w} \tag{15}$$

Where $\circ$ is the element by element product operator. Fig. 4 shows $\overline{S_{V_N}}$ and $\overline{S^W_{V_N}}$ plotted vs. x_i

D. Beat Timing Extraction

The beat timing (ictus) is extracted by examining the first derivative (discrete) between M^A_N and M^A_{N-1}, denoted by M^A_t:

$$M^A_t(N) = \frac{M^A_N}{fps} - \frac{M^A_{N-1}}{fps} \tag{16}$$

Where M^A_N and M^A_{N-1} is the mean angle moved at time point N and $N - 1$ respectively.

The beginning of the beat occurs when there is a sharp negative peak in M^A_t. The beat is marked according to Eq. 17.

$$B_N = \begin{cases} 1 & \text{if } M^A_t \leq B^c_N \\ 0 & \text{else} \end{cases} \tag{17}$$

where the rising edge of 1 represents the beginning of the beat (ictus) and 0 represents no beat. B^c_N is the dynamic beat marking threshold (DBMT) at discrete time N. B^c_N is a sliding window that averages over $\approx K_b \cdot 100\%$ the length of one beat.

$$B^c_N = \frac{-K_{b_w}}{N_b} \sum_{m=N-N_{b_w}}^{m=N} |M^A_t(m)| \tag{18}$$

where N_b is N_{b_w} less than the number of samples between two beats (rounded down to the nearest integer). K_{b_w} is a user defined static positive gain.

$$N_{b_w} = \text{floor}(K_b \cdot N_b) \tag{19}$$

K_b is the fraction of N_b that is used for averaging. To average over the time span of less than one beat $0 < K_b < 1$. N_b is the number of samples between two beats.

$$N_b = \frac{60 \cdot fps}{x_T} \tag{20}$$

where x_T is calculated in Eq. 9.

IV. EXPERIMENT & RESULTS

The capabilities of the vision beat tracker (VBT) were tested using two variables, the measure timing/metering and FFT length N_{fft}. The conducting tempo stayed constant at 120 BPM for all test. The FFT length was tested at $N_{fft} = 128$ and $N_{fft} = 256$. The measure timing/metering was tested at 4/4, 3/4, 2/4, and 1/4. Fig. 2 shows conducting pattern for the different timings. The accuracy of the calculated tempo x_p and the settling time for each permutation was calculated.

Note: for all tests a trained musician conducted in the specified meter to a metronome playing at 120 BPM. On each beat the metronome would play a *tick* sound. The audio

and video feed were saved to allow for timing comparison. The *tick* sound of the metronome is the benchmark tempo and beat timing. No other sounds are present in the bench mark videos.

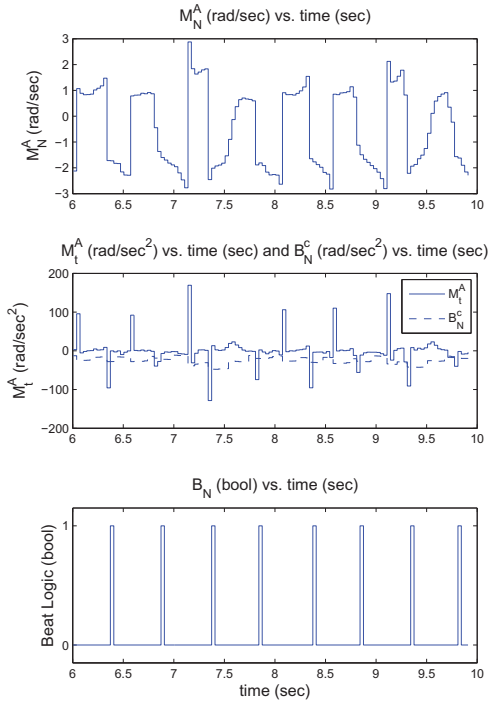

Fig. 7. The video feed is from a trained musician conducting in 4/4 time at 120 BPM. The tempo is calculated from Eq. 8 where $i = i_p$. The tempo calculated is $x_{i_p} = x_p = 119.4$ BPM. (TOP) M_N^A vs time: Mean direction objects in the frame moved from I_{N-1} to I_N. (MIDDLE) M_t^A and B_N^c vs. time: First derivative (discrete) of between M_N^A and M_{N-1}^A. B_N^c is the beat cut threshold. The start of the beat is when M_t^A falls below B_N^c. (BOTTOM) B_N vs time: The rising edge of B_N denotes the start of the next beat. B_N is calculated using Eq. 17. Note $fps = 29.97$, $N_{fft} = 256$, and $i = 18$.

A. Tempo Extraction Results

Table I and Fig. 8 contain the calculated x_T and settling times respectively for all metering and $N_{fft} = 256$. Table II and Fig. 9 contain the calculated x_T and settling times respectively for all metering and $N_{fft} = 128$.

It was found that the x_T calculated by the VBT had a smaller error with $N_{fft} = 256$ but a slower rise time when compared to $N_{fft} = 128$. The greater accuracy when $N_{fft} = 256$ is attributed to a smaller bin sizes for the calculated tempos. The faster rise time when $N_{fft} = 128$ is attributed to the system only reaching 5% acuracy as compaired to 0.5% acuracy.

B. Beat Timing Extraction Results

Fig. 7 shows M_N^A, M_t^A, B_N^c, and B_N vs. time with $fps = 29.97$, $N_{fft} = 256$, and $i = 18$. The video feed is from a trained musician conducting in 4/4 time at 120 BPM. The tempo is calculated from Eq. 8 where $i = i_p$. The tempo calculated is $x_{i_p} = x_p = 119.4$ BPM. It is important to note how B_N^c is more likely to be crossed the longer a beat has not been registered.

The VBT system placed a black circle in the upper left hand corner of the video feed on the frame that it detected the start of a beat, see Fig. 6. This black circle was used in the measurement of the accuracy of the system. The time between the beginning of the beat marked by the metronome's *tick* was measured and recorded.

Table III contains the time between the start of the beat and the marked beat for the different measure timings with $N_{fft} = 128$. Table IV contains the time between the start of the beat and the marked beat for the different measure timings with $N_{fft} = 256$.

There was little difference in the performance between the VBT with $N_{fft} = 128$ and $N_{fft} = 256$. On average there was a delay of 75ms between the start of a beat and the beat marker. This is equivalent to a 2-3 frame delay. Thus resulting in a delay of approximately 100ms at $fps = 29.97$, or a delay of approximately 50ms with $fps = 60fps$.

TABLE I

TEMPO ACCURACY AND SETTLING TIME FOR x_T VS. $time$, SEE FIG. 8. $N_{fft} = 256$, $fps = 29.97$, CONDUCTING TEMPO = 120 BPM

Meter	x_T (BPM)	Settling Time (sec)
4/4	119.41 BPM	1.60 sec
3/4	119.41 BPM	3.70 sec
2/4	119.41 BPM	4.04 sec
1/4	119.41 BPM	2.77 sec
-	average	3.03 sec

TABLE II

TEMPO ACCURACY AND SETTLING TIME FOR x_T VS. $time$, SEE FIG. 9. $N_{fft} = 128$, $fps = 29.97$, CONDUCTING TEMPO = 120 BPM

Meter	x_T (BPM)	Settling Time (sec)
4/4	112.39 BPM	0.87 sec
3/4	112.39 BPM	0.97 sec
2/4	126.44 BPM	1.44 sec
1/4	126.44 BPM	1.27 sec
-	average	1.14 sec

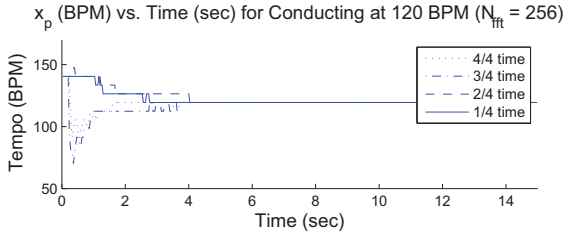

Fig. 8. Trained musician conducting at 120 BPM in multiple timings/meters. Plot of x_i vs. $time$ where $fps = 29.97$ and $N_{fft} = 256$. The average settling time for x_T is 3.03 sec.

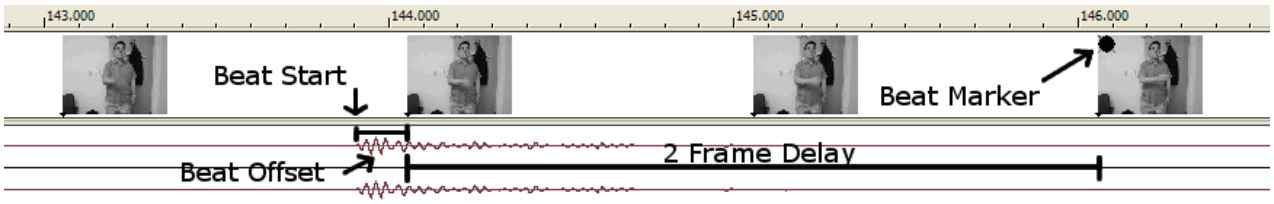

Fig. 6. Trained musician conducting at 120 BPM in 4/4 time. $N_{fft} = 128$, $fps = 29.97$. Displays the video feed (top) and the given audio (bottom). A metronome is playing a *tick* at 120 BPM and shows up in the audio display. No other audio is present in the video feed. A black dot shows up in the upper left hand corner of the frame when the VBT registers a beat. There is a 2-3 frame delay between beat activation and beat recognition.

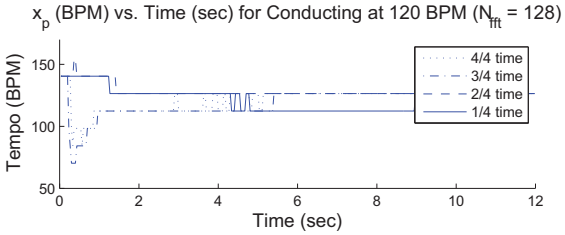

Fig. 9. Trained musician conducting at 120 BPM in multiple timings/meters. Plot of x_i vs. $time$ where $fps = 29.97$ and $N_{fft} = 128$. The average settling time for x_T is 1.14 sec.

TABLE III

BEAT TIMING WHERE THE OFFSET IS THE AMOUNT OF TIME BETWEEN WHEN THE BEAT OCCURRED AND THE VBT MARKED IT. $N_{fft} = 128$, $fps = 29.97$, CONDUCTING TEMPO = 120 BPM. NOTE: $\approx 0.033 \frac{sec}{frame}$

Meter	Frame Delay (frame)	Beat Offset (sec)	Total Delay (sec)
4/4	2	0.003 sec	0.069 sec
3/4	2	-0.007 sec	0.059 sec
2/4	2	0.017 sec	0.083 sec
1/4	2	0.023 sec	0.089 sec
-	-	average	0.075 sec

V. CONCLUSION & FUTURE WORK

With the use of the visual beat tracker (VBT), Jaemi Hubo has the ability to calculate the timing for a musical beat in 0.075 seconds and estimate the tempo in 4.0 seconds. The use of the Horn-Schunck optical flow algorithm allows the VBT to function in unstructured real-world environments. The dynamic beat marking threshold (DBMT) allowed the beat/ictus tracker to be unaffected by irregular conducting directions.

Jaemi Hubo now has one of the three senses (hearing, sight, and touch) required to achieve our overarching goal of creating a robot capable of being an interactive participant in a live ensemble. Jaemi's sense of sight gives it key information about a musical environment: tempo and ictus. The next step is to combine the audio beat tracker from Ellenberg's ADH system [2] with the VBT. As a result Jaemi will be able to follow the beat through visual and auditory cues. The addition of tactile and sensory multiplexing systems will close Jaemi Hubo's interactive participant gap and allow us to achieve our ultimate goal: having our adult size humanoid robot Jaemi Hubo be an interactive participant in a live music ensemble.

TABLE IV

BEAT TIMING WHERE THE OFFSET IS THE AMOUNT OF TIME BETWEEN WHEN THE BEAT OCCURRED AND THE VBT MARKED IT. $N_{fft} = 256$, $fps = 29.97$, CONDUCTING TEMPO = 120 BPM. NOTE: $\approx 0.033 \frac{sec}{frame}$

Meter	Frame Delay (frame)	Beat Offset (sec)	Total Delay (sec)
4/4	2	0.007 sec	0.073 sec
3/4	2	0.030 sec	0.096 sec
2/4	1	0.009 sec	0.042 sec
1/4	2	0.022 sec	0.088 sec
-	-	average	0.075 sec

VI. ACKNOWLEDGMENTS

Support for this work was provided by a National Science Foundation - Partnerships for International Research and Education grant (#0730206).

REFERENCES

[1] M. Michalowski, R. Simmons, and H. Kozima, "Rhythmic attention in child-robot dance play," in *Robot and Human Interactive Communication, 2009. RO-MAN 2009. The 18th IEEE International Symposium on*, sept. 2009, pp. 816 –821.

[2] R. Ellenberg, D. Grunberg, Y. Kim, and P. Oh, "Exploring creativity through humanoids and dance," in *5th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI) 2008, Proceedings of.* URAI, November 2008.

[3] E.D.Scheirer, "Tempo and beat analysis of acoustic musical signals," *Journal of the Acoustical Society of America*, vol. 103, pp. 588–601, 1998.

[4] K. Park, S. Jeong, C. Pelczar, and Z. Bien, "Beat gesture recognition and finger motion control of a piano playing robot for affective interaction of the elderly," *Intelligent Service Robotics*, vol. Volume 1, Number 3, pp. 1861–2776 (Print) 1861–2784 (Online), July, 2008.

[5] B. Lucas and T. Kanade, "An interactive image registration technique with application to stereo vision," *Proceedings of the DARPA image understanding workshop*, pp. 121–130, 1981.

[6] C. Crick, M. Munz, and B. Scassellati, "Synchronization in social tasks: Robotic drumming," in *Robot and Human Interactive Communication, 2006. ROMAN 2006. The 15th IEEE International Symposium on*, 6-8 2006, pp. 97 –102.

[7] B. Horn and B. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–204, 1981.

[8] T. Kim, S. Park, and S. Shin, "Rhythmic-motion synthesis based on motion-beat analysis," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 392–401, 2003.

[9] D. Moelants, "Dance music, movement and tempo preferences," in *Proc. 5th Triennal ESCOM Conference*, Ghent University, Belgium, 2003.