Development of a Musical-based Interaction System for the Waseda Flutist Robot – Implementation of a Real-time Vision Interface Based on the Particle Filter Algorithm

Jorge Solis, Atsuo Takanishi
(*Faculty of Science and Engineering*, *Waseda University*, *Tokyo* 169-8555, *Japan*)

Abstract – The aim of this paper is to create an interface for human-robot interaction. Specifically, musical performance parameters (i.e. vibrato expression) of the Waseda Flutist Robot No. 4 Refined IV (WF-4RIV) are to be manipulated. This research focused on enabling the WF-4RIV to interact with human players (musicians) in a natural way. In this paper, as the first approach, a vision processing algorithm, which is able to track the 3D-orientation and position of a musical instrument, was developed. In particular, the robot acquires image data through two cameras attached to its head. Using color histogram matching and a particle filter, the position of the musician's hands on the instrument are tracked. Analysis of this data determines orientation and location of the instrument. These parameters are mapped to manipulate the musical expression of the WF-4RIV, more specifically sound vibrato and volume values. The authors present preliminary experiments to determine if the robot may dynamically change musical parameters while interacting with a human player (i.e. vibrato etc.). From the experimental results, they may confirm the feasibility of the interaction during the performance, although further research must be carried out to consider the physical constraints of the flutist robot.

Key words – humanoids; human-robot interaction; vision; music

Manuscript Number: 1674-8042(2010)02-0133-05

dio: 10.3969/j.issn.1674-8042.2010.02.08

1 Introduction

The relation between art and robots has a long history dating since the golden area of automata. As a result from the great efforts from researchers from both musical engineering and biomechanical engineering fields, nowadays we may distinguish two basic research approaches: developing human-like robots and developing robotic musical instruments[1-2].

The first approach, formally named Musical Performance Robots, is based on the idea of developing anthropomorphic robots capable displaying musical skills similar to human (from the point of view of intelligence and dexterity). The first attempt of developing an anthropomorphic musical robot was done by Waseda University in 1984. In particular, the WABOT-2 was capable of playing a concert organ. Then, in 1985, the WASUBOT built also by Waseda, could read a musical score and play a repertoire of 16 tunes on a keyboard instrument[3]. Prof. Kato argued that the artistic activity such as playing a keyboard instrument would require human-like intelligence and dexterity. Other examples can be found in Ref.[4-7].

From the second research approach, a robotic musical instrument is a sound-making device that automatically creates music with the use of mechanical parts, such as motors, solenoids and gears. By implanting algorithms of Musical Information Retrieval (MIR), the robotic musical instruments are simple mechanisms designed to embed sensors to analyze the human behavior and to provide physical responses on the actuated musical instrument. In other words, this approach may facilitate the introduction of novel ways of musical expression that cannot be conceived through conventional methodologies. A number of engineers and artist have made headway in this area. The art of building musical robots has been explored and developed by musicians and scientists such as Ref.[8-11].

More recently, few researchers have been focused on integrating basic perceptual modules to the musical performance robots in order to interact with human musicians. In particular, Singer et al.[12] developed the GuitarBot which has been designed to create new of musical expression. In particular, their approach is based in developing robotic instruments that can play in way that humans can't or generally don't play. The instruments provide composers with an immediacy of feedback, similar to composing on synthesizers. However, as opposed to synthesizers, physical instruments resonate, project and interact with sound spaces in richer, more complex ways. All robotic instruments are controlled by custom developed Musical Instrument Digital Interface (MIDI) hardware and software, based around PIC microcontrollers. Another example is the Haile developed by Weinberg et al.[13], which is a robot designed to utilize autonomous behaviors that support expressive collaboration with human musicians. Haile is composed by a robotic arm that can hit the drumhead in different locations, speeds and strengths. The mechanism of the arm is reproduced by a sliding mechanism con-

* **Received:** 2010-03-06
Corresponding author: Jorge Solis(solis@ieee.org)

trolled by a solenoid. From the musical perceptual level, different Musical Information Retrieval algorithms have been implemented to modify the performance of Haile.

Even though GuitarBot and Haile are able to interact with musicians using conventional MIR algorithms, their physical mechanisms are too simple. However, if we want to understand in more detail the human while interacting in a musical way, we may require to increase the complexity of the mechanism of the musical performance robots as well as enhancing the perceptual capabilities of the robot (not only to process aural information but also visual, etc.) while considering physical constrains (such as breathing points, etc.).

Since 1990 at Waseda University we have been performing the research on musical performance robots[4]. As a result, we have been developing an anthropomorphic robot that is capable of producing the flute sound similar to an intermediate player. In order to add expressiveness to the flute performance, we have implemented musical performance rules based on neural networks so that the robot can extract the musical content of the human player before doing the interaction. However, when we tried to perform experiment where the robot interacting in real-time with a human musician (band context), still the robot lacks of cognitive capabilities to process the coming musical information from the performance of the partner.

Up to now, several researchers have been providing advanced techniques for the analysis of human musical performance. However, in our case, we are talking about not just analyzing the human performance, but also we are required to map those musical parameters into control parameters of the robot. This means that we are also required to take into account the physical constraints of the robot. Due to the complexity of doing this task, we have proposed to continue our research based on two approaches: enhancing the cognitive capabilities of the Waseda Flutist Robot to process visual/aural information and developing a new musical performance robot such as a Waseda Saxophonist Robot. Therefore, as a long-term goal, we would like to enable the interaction between two human-like performance robots that are able to interact at the same level of perception as humans. From this, we may understand more in detail, from a scientific point of view, how humans can interact in musical terms. This may also contribute in finding new ways of musical expression that have been hidden behind the rubric of musical intuition.

In this paper we provide an overview of the current research achievements on the WF-4RIV towards the implementation of a Musical-Based Interaction System (MbIS). A set of experiments were carried out to verify the effectiveness of the proposed vision processing to enable the flutist robot to actively interact with musicians.

2 Waseda Flutist Robot

The research on the Waseda Flutist Robot, since 1990, has been carried out as an approach to understand the human motor control from an engineering point of view as well as introducing novel ways of musical teaching. Thanks to the improvements on the technical skills of Waseda Flutist Robot, we have enabled the robot to enhance its musical expressiveness[4]. In particular, we have been focused on improving the mechanical design of the lungs, vocal cord, mouth, etc. as well as the implementation of advanced control strategies to control the implement auditory feedback system[14] and enhancing some of the perceptual capabilities such as automatic melody recognition[15], human face tracking[16], etc. As a result of our research, the latest version of the flutist robot, the WF-4RIV is able of playing the flute nearly similar to the performance of an intermediate flutist[17]. Moreover, towards proposing novel applications for humanoid robots, a General Transfer Skills System was implemented on the flutist robot to emulate the presence of a musical tutor[18].

However, the WF-4RIV is able of playing a human-like performance in terms of timing and variation of tone modulation (i. e. vibrato effect), but it does not actively react to what other instrumentalists are playing (Fig. 1). In order to integrate the robot into a realistic environment like a human band, it needs to be able to react to visual and acoustic cues from its partner musicians. When looking at how human instrumentalists interact with each other in a Jazz band, we see that the musicians give signals to each other by moving their instruments. When the members of the band take turns in playing solos a saxophonist player might direct the lead to the next musician by bending forward with his instrument in the direction of that person. To detect movements of an instrument we use two cameras that are integrated into the head of the flutist robot. We experimented with different vision processing algorithms to find out, which method is most suitable to track movements of a musician's instrument in a realistic environment. Such a realistic environment might be a concert stage or a studio room. In both cases various lighting conditions, from very bright illumination to a quite dark environment might occur.

2.1 Particle tracking algorithm

We concentrate on creating a visual interface that enables a human to control the robot through gestures with his musical instrument as freely as possible. A further emphasis here is to create an interface, that allows on the one hand controlling the musical expression of the robot accurately, but also robustly and in a computationally efficient way. In the past we have exhibited the robot on several occasions and we found that each place had very different lighting and background conditions. Our gesture detection method should not only work in an optimized laboratory environment, but also under realistic circumstances; thus, we need to find a way to cope with these varying ambiances. There might be a constantly changing background (i.e. people passing by, stopping to watch the

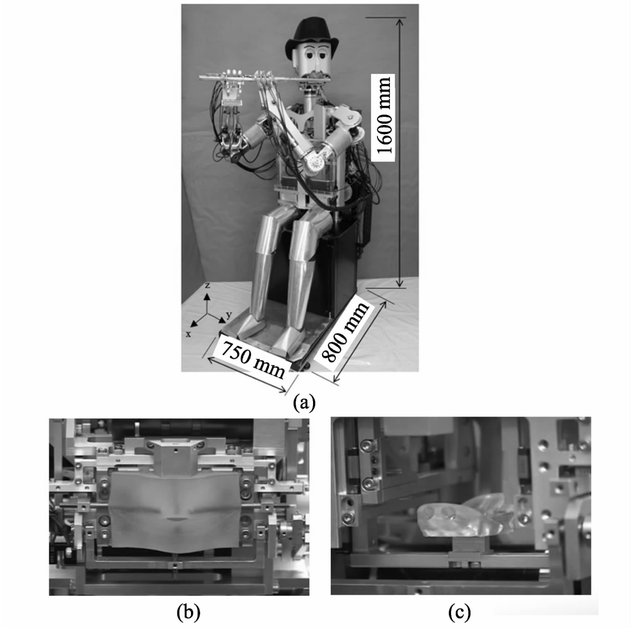

Fig. 1 Waseda Flutist Robot No. 4 Refined Ⅳ (WF-4RIV) has been designed to play a classical flute: (a) The WF-4RIV is composed by 41-DOFs, (b) artificial lips, (c) artificial tongue

robot perform) and below-optimum lighting sources (i. e. stage lighting facing into the cameras of the robot). Another difficulty is that the humanoid head of the robot can move during a performance. This requires our image processing algorithm to adapt rapidly to fundamental changes of background (fundamental in contrast to a small object being moved in the background).

In order to satisfy these requirements, we propose to implement color histogram matching[19] and particle tracking[20] to follow the movement of a musical instrument, while satisfying the previously introduced requirements. The combination of the two methods is an established way to follow an object with a certain color profile. The system is initialized manually by defining the starting positions of the player's hands. For the computation of 3D data, the algorithm makes use of a stereometry mapping technique.

We do not calculate a complete depth map of the scene, but limit our matching to the four patches found by the particle tracker. One might argue that we could also use 2D techniques to compute a depth image: as we try to emulate human perception, we prefer using two cameras and stereo pair matching. The technique also saves resources due to the limited number of points being calculated. To begin with, the user marks the initial location of the image areas to be tracked. Four areas are selected: one patch for each hand in the left eye camera and one patch in the right eye camera. A color histogram is generated from each of the selected image areas. Once a new video frame is acquired we find a random distribution of locations around the previous location of each of the patches. For each of these particles a histogram is generated and compared with the initialization histogram. The particle with the most similar color profile is the new position of the tracked patch. Particles with less likelihood are also saved, each weighted according to its probability of occurrence. The more particles are being used the more accurate the method becomes (Fig. 2).

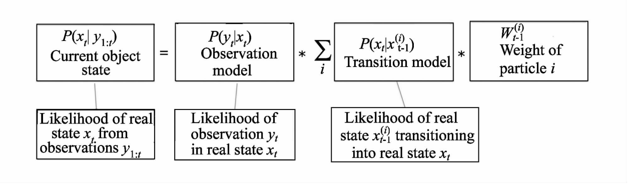
$P(x_t|y_{1:t})$
Current object state
=
$P(y_t|x_t)$
Observation model
$* \sum_i$
$P(x_t|x^{(i)}_{t-1})$
Transition model
*
$W^{(i)}_{t-1}$
Weight of particle i
Likelihood of real state x_t from observations $y_{1:t}$
Likelihood of observation y_t in real state x_t
Likelihood of real state $x^{(i)}_{t-1}$ transitioning into real state x_t

Fig. 2 Principle of the particle filter tracking

After we found the x-y -axis coordinates of the image patches, we now calculate the distance of a patch from the camera. To achieve this we examine the difference between the x -position of the patch in the left camera image to the x -position position in the right camera image. The larger the difference, the closer the patch is located to the camera. Without an exhaustive calibration, we cannot determine the absolute distance of the patch from the camera. However, if we work with relative walues to compute only changes in orientation and position, this may not be necessary.

For both vision processing methods, we model the shape of the saxophone as a line. The hands of the player are located on two spots along a line. The average of the position of the hands is recorded as the center position of the saxophone. Similarly we deal with the orientation: We consider a line drawn from the center of one hand to the center of the other. The inclination of the line is the orientation of the instrument. There is no ambiguity about the position, as normally a player would not hold the instrument upside down. From the 2D coordinates of the four hand particles we calculate the relative position, inclination and rotational angle of the instrument. To compute the depth values of both hands we use a z - transformation Eq. (1).

$$z = \frac{1}{\Delta x}\alpha, \tag{1}$$

where x_p is the distance between the x -coordinate of the patch in the left camera image (xpl) and the right camera image (xpr), as shown in Eq. (2).

$$\Delta x_p = | x_{pl} - x_{pr} |. \tag{2}$$

Accordingly z denominates the z -coordinate of the patch. We use as a constant to adjust the value of Δz for further calculations. Inclination and rotational angle are obtained by transforming the Cartesian coordinates resulting from Eq. (3)~(5) the object tracking into a cylindrical system.

$$\Delta x = x_1 - x_0, \tag{3}$$

$$\Delta y = y_1 - y_0, \tag{4}$$

$$\Delta z = z_1 - z_0, \tag{5}$$

Δx ,Δy , and Δz denote the vector between the saxophone player's hands (3-5). For Δx and $\Delta y > 0$; we use a Cartesian coordinate systemparallel to the view plane of the robot to calculate roll (φ) , pitch (θ) and yaw (ρ) inclination of the instrument.

$$\varphi = \arctan(\Delta y / \Delta x), \tag{6}$$

$$\theta = \arctan(\Delta y / \Delta z), \tag{7}$$

$$\rho = \arctan(\Delta x / \Delta z), \tag{8}$$

Although we adapt new object coordinates only from the particle with the highest likelihood, as the particle filter method works recursively, the information about the other particles is not lost. A particle with an initially lower than maximum likelihood is not discarded, but it can still propagate to gain more likelihood later. However, research on particle filters has shown that in case all particles are kept for the whole tracking run, all but one particle tend to be degraded to probabilities close to 0. There are several ways to counteract this behavior[1]. We have chosen the method of re-sampling. After each new predict-update cycle, particles with a probability lower than a certain threshold are exchanged for newly initialized particles. This threshold, as well as the optimum number of particles to be used, has been determined manually.

3 Experiments & results

The purpose of our experiments is to show how well a user can express his musical intention using the provided interaction setup. The interaction system itself resembles a closed control loop with the robot on the one side and the human musician on the other side. The fact that decides about the quality of this closed loop is how responsive the robot is to the actions of its interaction partner. Movements of the instrument in front of the robot camera were recorded as orientation values that change the vibrato amplitude of the flute sound.

We used a FFT spectral analysis algorithm to find pitch and amplitude of the music data. This enabled us to qualitatively extract pitch/melody pattern, song tempo and vibrato amplitude from the music data. To keep the analysis as simple as possible only one tone is played. The average volume of the sound output becomes higher with less vibrato effect as the amount of air streaming through the glottis mechanism of the robot that controls the vibrato, is at these times higher and produced louder volume. However, the vibrato oscillates over this average value and changes its amplitude according to the orientation value calculated by the particle tracking algorithm.

The experimental results are shown in Fig. 3. At 16 s, we observe a maximum of approximately 100 controller tics that relates to vibrato amplitude of around 10 dB in the volume plot. A minimum at 25 s accordingly produces very low vibrato amplitude. When we look at the volume plot we can identify areas, where the level suddenly drops for certain duration. These moments are called Breathing Points (BP) and relate to the time when the robot's lung system is deflated and needs to pull air in order to be able to produce the air-beam necessary to generate the flute sound. In the graph, we find these BPs at 9.5 s, 19 s and 29 s.

In regards to timing the distribution is equidistant and of equal duration. The re-filling period of the lung is the only time that the tone pitch is not reported as constant. At that time, when the robot does not produce any sound, the analysis of the environment noise resulted in analysis values that bear no significance for our experiment. Concluding from this we note, that there are certain systematical restrictions that need to be considered when approaching interaction with the Flutist Robot. The foremost limitation here is the breathing-in, breathing-out rhythm of the lung. During the time the robot is breathing out, a tone is generated, that can be manipulated by the user utilizing the controls provided by the interaction system. However, when the robot is breathing in, naturally, no tone is produced, which means that the user has to take a forced break in his control scheme. To be prepared for these interruptions in the flutist robot's play, the musician has to adapt his musical material in a way that is similar to creating musical material for a human, as a human also has certain physical limitations.

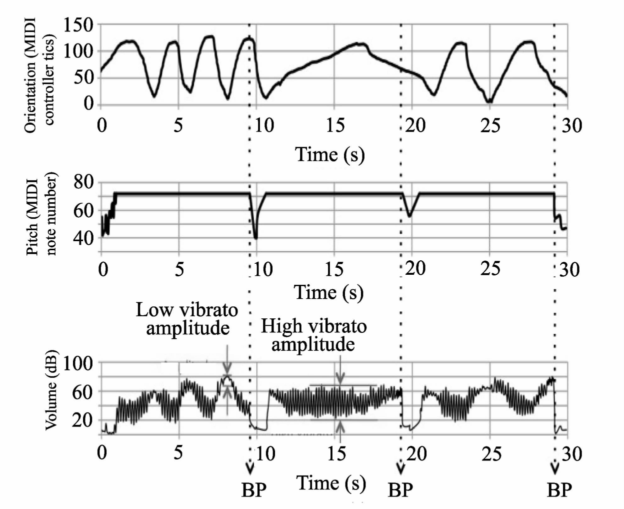

Fig.3 Experimental results for the Particle Filter-based object tracker controlling the WF-4RIV. The movements recorded in the upper graph manipulate the sound output of the Flutist Robot, shown in the lower two plots. Breathing Points are marked as BP

As a further application to control the performance parameters of the robot we decided to modulate the frequency of the vibrato output of the robot. The mode of experimentation here was the same as in the previous evaluations. However we changed our signal mapping scheme to relate instrument orientation data with the vibrato frequency instead of the vibrato amplitude of the robot's output. We set the target frequency range for the robot's vibrato output from 0 Hz (a constant note, without vibrato) to 10 Hz. Vertical orientation of the saxophone (0°) was assigned to the lowest possible vibrato frequency. The saxophone being moved forward by 90° resulted in the maximum vibrato frequency. We mapped the angle space between 0° and 90° linearly to the vibrato range. The angle data we retrieved from the vision processor had a resolution of 1°.

In Fig.3, we observe that also in this case, as far as the physical restraints of the robot allow it, the instrumentalist can freely influence the performance output. Slower as well as faster movements are cleanly detected and propagated by the Particle Filter-based tracker. Finally, we may consider about the qualitative applicability of the Particle Filter-based tracker, that the method bears the potential of being used to control more performance critical parameters of the Flutist Robot. The algorithm not only satisfies in respect to reproducing the timing of instrument movements during an improvised Jazz performance, but also sensitively records amplitudes and complies with available computational resources.

4 Conclusions & future work

In this paper, we have presented the implementation of a real-time vision tracking system to enable the interaction between the WV-4RIV with musicians. For this purpose, the particle filter algorithm was proposed to track the orientation of the instrument, in order to control the musical performance parameters like vibrato amplitude of the flute tone during a musical interaction.

As a long term goal our system is to be used by musicians without technical knowledge, we would to try to reduce the effort necessary to initialize the tracking system. On the other hand, we will focus our research as well to the implementation of an aural processing algorithm to track in real-time different kinds of musical parameters (i.e. pitch, tempo, etc.) implementation of an aural processing algorithm to track in real-time different kinds of musical parameters (i.e. pitch, tempo, etc.)

References

[1] A. Kapur, 2005. A History of Robotic Musical Instruments. Proc. of the International Computer Music Conference, p. 21-28.

[2] J. Solis, A. Takanishi, 2007. An Overview of the Research Approaches on Musical Performance Robots. Proc of the International Computer Music Conference, p. 356-359.

[3] S. Sugano, I. Kato, WABOT-2: Autonomous Robot with Dex-terous Finger-arm Coordination Control in Keyboard Performance. Proc. of the Int. Conference on Robotics and Automation, p. 90-97, 198.

[4] J. Solis, K. Chida, K. Suefuji, A. Takanishi, 2006. The development of the anthropomorphic flutist robot at Waseda University. *International Journal of Humanoid Robots*, 3(2): 127-151.

[5] S. Takashima, T. Miyawaki, 2006. Control of an Automatic Performance Robot of Saxophone: Performance Control Using Standard MIDI Files. Proc. of the IROS Workshop on Musical Performance Robots and Its Applications, p. 30-35.

[6] K. Shibuya, 2007. Toward Developing a Violin Playing Robot: Bowing by Anthropomorphic Robot Arm and Sound Analysis. Proc. of ROMAN, p. 763-768.

[7] H. Kuwabara, M. Shimojo, 2006. The Development of a Violin Musician Robot. IROS06 Workshop on Musical Performance Robots and Its Applications, p. 18-23.

[8] A. Kapur, E. Singer, M. Benning, G. Tzanetakis, 2007. Integrating Hyper Instruments, Musical Robots & Machine Musicianship for North Indian Classical Music. Proc. of the New Interfaces for Musical Expression, p. 238-241.

[9] R. B. Dannenberg, B. Brown, G. Zeglin, R. Lupish, 2005. McBlare: A Robotic Bagpipe Player. Proc. of the NIME2005, Vancouver.

[10] K. Beilharz, 2004. Interactively Determined Generative Sound Design for Sensate Environments: Extending Cyborg Control. Y. Pisan (eds), Interactive Entertainment, p. 11-18.

[11] E. Hayashi, 2006. Development of an Automatic Piano that Produce Appropriate: Touch for the Accurate Expression of a Soft Tone. Proc. of IROS Workshop on Musical Performance Robots and Its App., p. 7-12.

[12] E. Singer, 2004. LEMURs Musical Robots. Proc. of the Conference on New Interfaces for Musical Expression, p. 183-184.

[13] G. Weinberg, S. Driscoll, 2007. The Design of a Perceptual and Improvisational Robotic Marimba Player. Proc. of the 16th IEEE International Conference on Robot & Human Interactive Communication, p. 769-774.

[14] J. Solis, K. Taniguchi, T. Ninomiya, T. Yamamoto, A. Takanishi, 2008. Development of Waseda Flutist Robot WF-4RIV: Implementation of Auditory Feedback System. Proc. of International Conference on Robotics and Automation, p. 3654-3659.

[15] J. Solis, S. Isoda, K. Chida, A. Takanishi, K. Wakamatsu, 2004. Anthropomorphic Flutist Robot for Teaching Flute Playing to Beginner Students. Proc. of the IEEE International Conference on Robotics and Automation, p. 146-150.

[16] J. Solis, K. Suefuji, K. Chida, A. Takanishi, 2006. Imitating the Human Flute Playing by the WF-4RII: Mechanical, Perceptual and Performance Control Systems. Proceedings of the 1st IEEE/RAS- EMBS International Conference on Biomedical Robotics and Biomechatronics, p. 1024-1029.

[17] J. Solis, K. Taniguchi, T. Ninomiya, T. Yamamoto, A. Takanishi, 2008. The Mechanical Improvements of the Waseda Flutist Robot and the Implementation of an Auditory Feedback Control System. Proc. of the 17th CISM-IFToMM Symposium on Robot Design, Dynamics, and Control, p. 217-224.

[18] J. Solis, K. Suefuji, K. Taniguchi, T. Ninomiya, T. Yamamoto, A. Takanishi, 2006. Towards an Autonomous musical teaching system from the Waseda Flutist Robot to Flutist Beginners. Proc. of IEEE/RSJ International Conference on Intelligent Robots and Systems-Workshop: Musical Performance Robots and Its Applications, p. 24-29.

[19] R. F. D. Saxe, 1996. Toward Robust Skin Identification in Video Images. 2nd International Conference on Automatic Face and Gesture Recognition, p. 379.

[20] M. S. Arulampalam, S. Maskell, N. Gordon, T. Clapp, 2002. A tutorial on particle filters for online nonlinear/non-GaussianBayesian tracking. *IEEE Transactions on Signal Processing*, 50(2): 174-188.