Interactive Sound Creation System with Depth Camera

Hiroyo Ishikawa* Takeki Ihara† Takuya Ujihara‡ Hideo Saito§

Graduate School of Science and Technology, Keio University, Yokohama, Japan

Abstract

In this paper, we propose an interactive sound creation system with a depth camera. When an object enters in the selected area of the depth camera, a sound is created interactively according to the acquired data (depth data). This system decides the three elements (pitch, tone and volume) of the sound by the statistical value of the depth data, position and motion of the objects. This system can create two kinds of sound. One is sound effects, like *Theremin*, which is created by fluctuating pitch frequency and harmonics structure, the other is creating a sound by mixing four kinds of tone selected from the musical source of a synthesizer. In the latter, when pitch is decided from the pitch included in a musical scale, melodies can be created automatically by the motion of objects. When the selected tone and the musical scale are changed, the system can create various sounds and melodies of another musical style. Additionally, an object do not necessarily have to be a human. Therefore this system is a new style virtual musical instrument to create sound effects and melodies unconsciously.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces—Interaction styles; H.5.5 [Information Interfaces and Presentation]: Sound and Music Computing—Systems; I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Depth cues; I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Motion; J.5 [Arts and Humanities]—Performing arts

1 Introduction

Recently, a new kind of musical instrument have been developed as a new entertainment systems. In particular, instruments using human motion have intensively been developed [1][2][3][4][5]. The Yamaha Miburi [1] is a musical instrument with body suit that senses a variety of body movements. Based on the movements, MIDI(Musical Instrument Digital Interface)[6] data is emitted, which can also be subsequently used to generate or modify sound or any other media. The GypsyMIDI [2] Motion Capture MIDI Controller from Sonalog is a musical instrument which consists of mechanical exoskeleton attached to the user's arms and shoulders. This instrument converts body movements to MIDI commands, so that the body movements can control the music. Even though such instruments can generate music with the movement of the human, the user needs to wear or attach the special sensors for making the instruments detect the motion of the user.

Theremin is a musical instrument that can be controlled by human motions without any attachment to the user. *Theremin* senses the change of electrostatic capacity, this means that the distance between each hand and the sensor antennas on the instruments can change tone of the musical scores. While *Theremin* only detects the distance of each hand from the antennas, we propose to use a ToF depth camera for capturing depth map of the body, which controls musical sound.

D-BEAM controller installed in musical instruments of Roland [7] has the similar musical control functionality based on the depth measurement from the system. However, it cannot measure the shape of the objects, but only measure the depth along with the fixed direction of an infra-red light beam. Our proposed system uses the depth camera, which enables to adapt with variety of shapes, poses, and positions in the space for variety of musical controlling functionalities.

The depth camera can directly capture the distance to the object by measuring the time of reflected infrared light for each pixel, so that the distance map (depth map) to the scene can be captured in video rate (i.e. 30FPS). Since the depth of each point of body surface can be real-timely detected using the depth camera, we can develop a new musical instrument with which player's body motion can precisely control various aspect of the sound. In addition to the body motion, we can also use the change of the surface shape of any object. For example, a folded paper can be used as a musical controlling material. In this way, interaction with any material can be a musical instrument by using the depth camera.

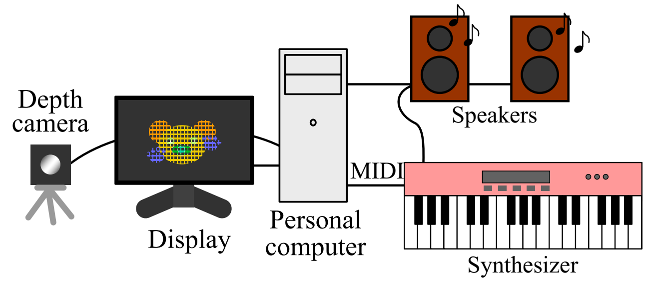


Figure 1: Sound creating system.

2 Interactive Sound Creation System

2.1 System over view

The system structure is shown in figure 1. The system consists of a depth camera, a personal computer, a computer display, a synthesizer and speakers. The user interface of the system is the depth camera which can acquire depth data at max of 30 fps. While an object exists in the selected range of the depth camera, the system sounds. Data processing and some device control are executed by a personal computer (Intel®Core™2 Duo CPU, 3.00GHz, 4.0GB RAM). The acquired depth data is displayed at a computer display as 3D mesh colored according to depth value in HSV color model.

This system can create two kinds of sound. In method 1, it needs a personal computer and speakers when the system creates sound effects like *Theremin*. In method 2, the system uses sound resource GM2 (General MIDI system level 2)[6]. In this system we use GM2 sound source of a synthesizer (Roland JUNO-Di). However, if a computer has GM2 sound source, the system can implement creating sound without a synthesizer.

*e-mail:hiroyo@hvrl.ics.keio.ac.jp
†e-mail:tihara@hvrl.ics.keio.ac.jp
‡e-mail:ujihara@hvrl.ics.keio.ac.jp
§e-mail:saito@hvrl.ics.keio.ac.jp

20th International Conference on Artificial Reality and Telexistence (ICAT2010), 1-3 December 2010, Adelaide, Australia
ISBN: 978-4-904490-03-7 C3450 ©2010 VRSJ

2.2 Depth Camera

The Depth camera which is used in this system is the SwissRanger SR4000. It's an infrared camera which measures distance based on the Time of Flight (ToF) principle[8]. It can directly capture the distance to an object by measureing the time of reflected infrared light for each pixel. Unlike a stereo camera system which relies on corresponding points in the images to determine the distance, the ToF system can measure distance of the objects with low or no textures like hands. According to the SR4000's manual, infrared light used in the SR4000 is in the 850nm range, and eye-safe under all operating conditions. So it is safe to use the SR4000 to detect body movements. The SR4000 acquires 176×144 pixel depth data at 30 fps. It also provides the "confidence map" that indicates quality of each pixel's measurement, based on distance, amplitude measurements and their temporal variations. It enables us to remove low quality pixels from statistic calculation.

2.3 Depth data and system flow chart

Depth data is represented by the right hand coordinate system, in which the optical axis of the depth camera is z axis, and the origin of the system is in the camera, as shown in figure 2. The depth data is acquired as a depth map, in which each pixel has information of the 3D position (x,y,z) of the surface on the depth map. This system uses depth data of the selected area which is set previously by the user. The selected area is defined by "near" and "far" as shown in figure 2. An example of the depth image which is acquired by the depth camera is shown in figure 3(1), and its 3D representation is shown in figure 3(2). Farther points from the camera are shown as brighter color in both images. Invalid pixels, which are out of the selected range, in the depth image are shown in black color.

The system flow chart is shown in figure 4. Method 1 is shown in figure 4(1). Noise is reduced from depth data, the nearest point and some statistical values are extracted from the data, and a sound is created. Method 2 is shown in figure 4(2). In method 2, the system can create more than two sounds simultaneously, so it includes process of segmentation of a depth image. When the system creates a sound, the nearest point and some statistical values are detected from the data after noise reduction in the same way as method 1. On the other hand, when the system creates more than two sounds simultaneously, the depth image is segmented into some subareas, the system selects nearer subareas from the camera. Then the nearest points are detected from each subarea. In the next step, the system checks that the data change between the previous frame and the current frame. When the change is at more than the threshold which is set previously by the user, the system create sounds.

An example of segmented areas is shown in figure 5. A depth image which is the left is segmented into subareas shown as figure 5(2). If each distance of four connected pixels is less than a threshold, these pixels are included in the same subarea (four connected pixels labeling). In this example, the depth image is segmented into five subareas, in which the same area is shown as the same color.

2.4 Method of Creating Sound

We have prepared two methods of creating sound, Method 1 and Method 2. Method 1 creates sound by addition of sinusoidal waveforms, and Method 2 by using a synthesizer of musical instrument. In the case of Method 1, since data from depth camera is supposed to be updated at a constant interval, it creates sound of constant length once the data is updated. On the other hand, Method 2 creates sound on the request of new sound from the system, and sustains it until the next request.

2.4.1 Method 1

When it creates sound, Method 1 determines loudness of sound by average distance between a detected object and the depth camera, and selects predetermined tone by the quadrant on X-Y plane where

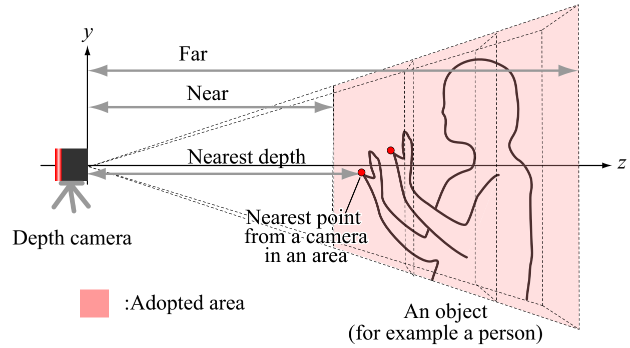


Figure 2: The adopted area which is acquired by a depth camera.

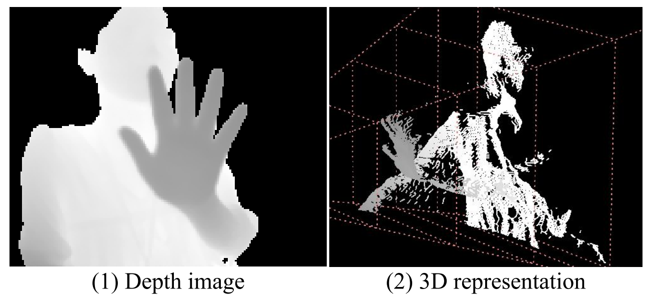


Figure 3: An example of a depth image (gray scale) and its 3D representation. Farther points from the camera are shown by brighter color. Invalid points are painted by black in the left image.

the nearest point lies. The method modifies tone based on variance of distance, but does not change pitch.

The detail of Method 1 creating sound $s(n)$ is as described below.

$$s(n) = A(Z_{avg}) \sum_{i=0}^{p-1} h_j(i) \sin\left\{ \frac{2\pi n(i+1)(F_f+\Delta f)}{F_s} + \phi(i) \right\}$$

Z_{avg} : average of effective depth
F_s : sampling frequency
F_f : pitch frequency
$\phi(i)$: initial phase of i th harmonics

Herein

$$A(Z_{avg}) = \frac{Z_{avg} - Near}{Far - Near} A_{\max} + \frac{Far - Z_{avg}}{Far - Near} A_{\min}$$

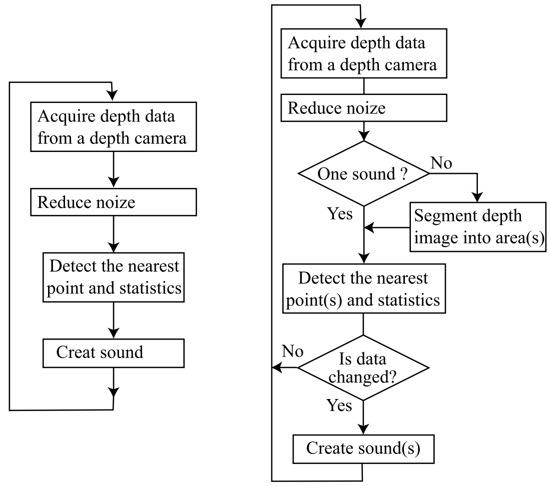
$A_{\max}$: maximum value of A
$A_{\min}$: minimum value of A

$$\Delta f = N(V_Z - M)$$

V_Z : variance of effective depth
M, N : determination coefficients of range of Δf

$A(Z_{avg})$ is a scaling factor which specifies amplitude of created sound. As average distance between detected object and depth camera becomes shorter, Method 1 creates louder sound. h_j used, which characterizes tone, is determined according to the quadrant where X-Y coordinate of the nearest point is located, as shown in figure 6. For example $h_j(i)$ is set to create triangular or rectangular waves as below.

$$\begin{aligned}
h_1(i) &= \{10000, 4000, 0, 3000, 0, 2000, 0, 1000\} \\
h_2(i) &= \{10000, 1000, 4000, 1000, 3000, 0, 2000, 0\} \\
h_3(i) &= \{10000, 2000, 0, 3000, 0, 3000, 0, 2000\} \\
h_4(i) &= \{10000, 1000, 6000, 0, 3000, 0, 2000, 0\}
\end{aligned}$$



(1) Method 1 (2) Method 2

Figure 4: The system flow chart.

(1) Depth image (2) Segmentation result

Figure 5: An example of a segmentation result. The pixels in the same area are shown at the same color.

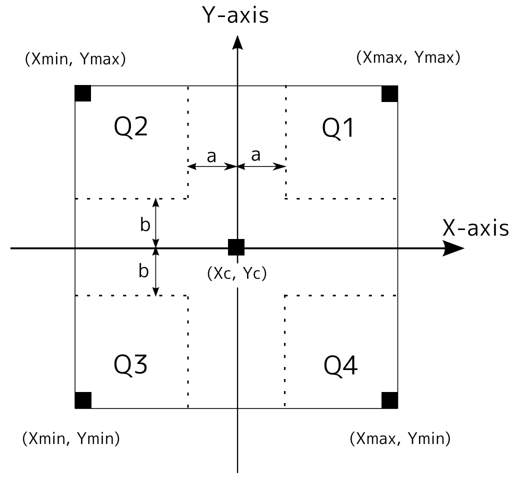


Figure 6: X-Y plane of depth data

Δf is used so that pitch frequency fluctuates and harmonics structure of created sound $s(n)$ changes slightly based on variance of effective depth value. The degree of the fluctuation depends on the coefficients M and N.

2.4.2 Method 2

Method 2 creates sound by using a synthesizer which is connected to a personal computer via MIDI(Musical Instrument Digital Interface). MIDI is a protocol which specifies how musical instruments are connected between themselves and computers and how performance information such as volume, pitch, tone, etc. is transmitted[6]. The method creates louder sound as average distance between a detected object and the depth camera comes shorter. Regarding tone, the method selects predetermined one according to the quadrant on X-Y plane where the nearest point lies, and modifies it. The method changes pitch so that higher Y coordinate of the nearest point position corresponds to higher pitch.

The detail of Method 2 is as follows. Now let $S(L,T,p)$ denote sound of loudness L, tone T and pitch p. Four sound $S_i(Li,T_i,p_i)$ $\{i=1,2,3,4\}$ are allocated to quadrants Q_i $\{i=1,2,3,4\}$ shown in figure 6. Loudness L is calculated as:

$$L_i = \frac{Z_{avg}-Near}{Far-Near}L_{i,\max} + \frac{Far-Z_{avg}}{Far-Near}L_{i,\min}$$

$L_{i,\max}$: maximum value of L
$L_{i,\min}$: minimum value of L

As for tone, in addition to the way the selected quadrant determines a basic tone, variance of detected distance modifies power spectrum of the tone by changing cutoff parameter of MIDI, which changes cutoff frequency of sound. Regarding pitch p_i, in the quadrant Q_1 or Q_2

$$p_i = round\left(\frac{Y+b-Y_c}{Y_{\max}+b-Y_c}p_{i,\max} + \frac{Y_{\max}-Y}{Y_{\max}+b-Y_c}p_{i,\min}\right)$$

$round(p)$ denotes integer nearest to p

in the quadrant Q_3 or Q_4

$$p_i = round\left(\frac{Y-Y_{\min}}{Y_c+b-Y_{\min}}p_{i,\max} + \frac{Y_c+b-Y}{Y_c+b-Y_{\min}}p_{i,\min}\right)$$

Since rapid sound change can happen around border of quadrants, interpolation of sound is processed to prevent it. For example interpolated sound S around center of X-Y plane, (X_c,Y_c) shown in figure 6 is:

$$\begin{aligned} S \quad = \quad & \frac{b+Y-Y_c}{2b}\left(\frac{a+X-X_c}{2a}S_1 + \frac{a-X+X_c}{2a}S_2\right) \\ & + \frac{b-Y+Y_c}{2b}\left(\frac{a+X-X_c}{2a}S_4 + \frac{a-X+X_c}{2a}S_3\right) \end{aligned}$$

We can select four sound for the system from sound set of General MIDI 2 known as GM2, which defines specific features of MIDI[6]. An example of allocated sound is shown in table1.

Table 1: allocated sound

Q2: violin	Q1: synth brass3
Q3: percussive organ2	Q4: timpani

3 Experiments

3.1 Experimental results

The proposed system is shown in figure 7. The depth data from the depth camera is shown on the computer display. We examined two methods. Here, “far” is 150cm and “near” is 50cm in figure 2.

3.1.1 Method 1

Sound creation with simulation data Input data of an experiment is simulation data which is generated by the following way. It is supposed that a sphere moves along the circle line on a

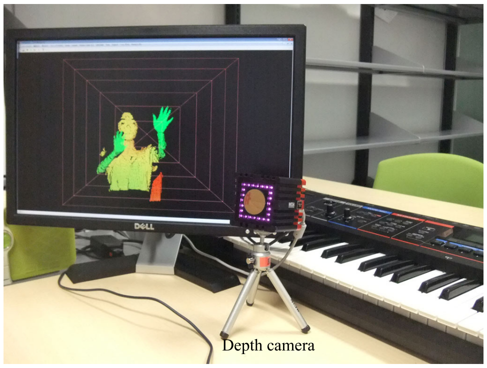


Figure 7: An experiment scene of the proposed system

plane which is parallel to the xy-plane and z equals 80cm. When the sphere is moving, its radius is changing from 2 cm to 10 cm and back to 2 cm twice. The variance of data is changed by changing the radius. The simulation data is generated from this model for 200 frames and is shown in figure 8.The quadrant numbers in which the nearest point exists are described at top of the graph.

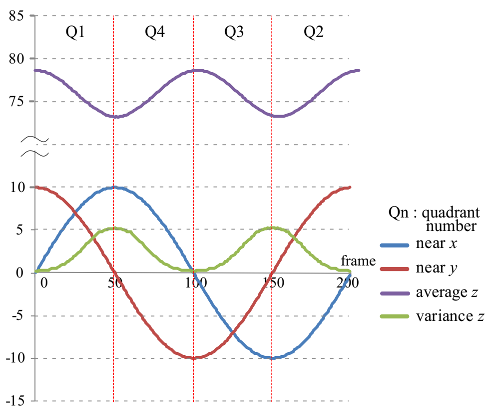


Figure 8: An example of simulation data in method 1. The nearest point position and average and variation of z are shown.

The result of creating sound with the simulation data is shown in figure 9(1)(2). From the graphs, it can be seen that frequency ($F_f + \triangle f$) is changed according to variance of z and the volume is changed according to average of z.

Sound creation with depth data An example of depth data with the depth camera for 50 frames is shown in figure 10. The quadrant numbers are described at top of the graph, and the result frequency ($F_f + \triangle f$) which is decided from variance of z is shown in figure 11 (1), the result volume which is decided from avarage of z is shown in figure 11 (2).

3.1.2 Method 2

Sound creation with simulation data Input data of this experiment is simulation data which is shown in figure 12. In the simulation data, the nearest point is moving along the circle line on a plane which is parallel to the xy-plane and z equals 60cm. The system creates a sound by mixture of four kinds of tone. And then, in this paper, we select "blue pentatonic scale" as a musical scale.

The result volume of four kinds of tone is shown in figure 13. Volume number means the quadrant number corresponding to four kinds of tone.

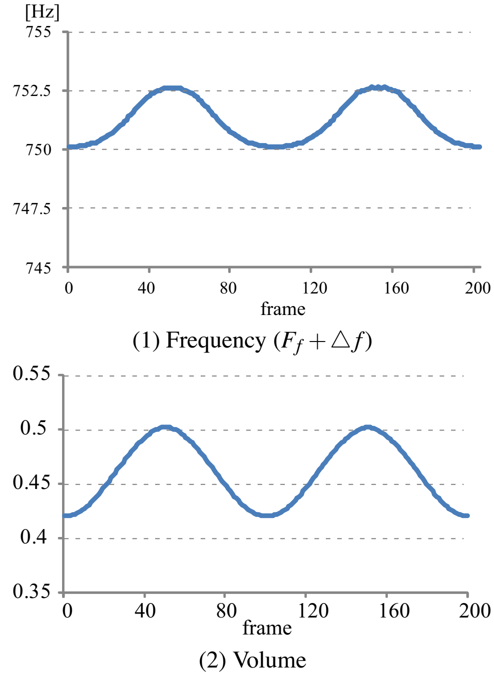


Figure 9: An example result of sound creation with simulation data in method 1

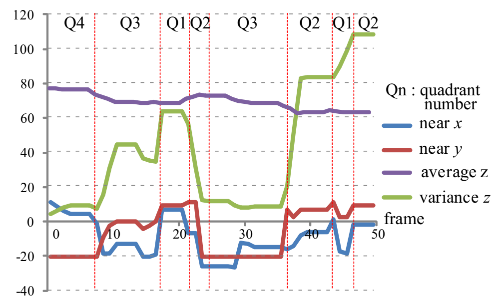
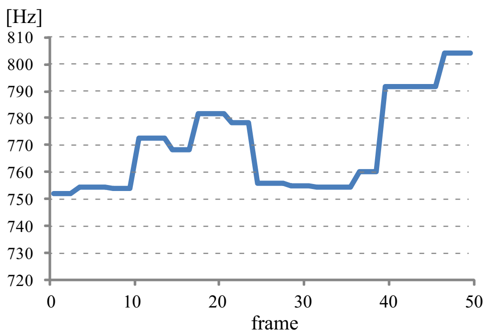


Figure 10: An example of depth data in method 1. The nearest point position and average and variation of z are shown.

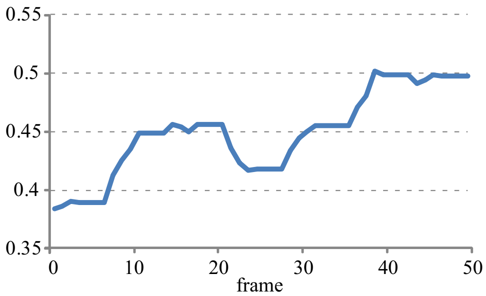
The result score of four kinds of pitch corresponding to four kinds of tone without the restriction of a musical scale is shown in figure 14(1), and the result score with the restriction of a musical scale is shown in figure 14(2). From figure 12 (1) and figure 14, It can be seen that pitch is changed according to the y position of the nearest point and volume of four tone is controlled by the nearest position. Additionally, in the case of using a restriction of a musical scale, pitch is included in the musical scale.

Sound creation with depth data In this experiment, input data is depth data from the depth camera and the system creates two sounds simultaneously by using a restriction of a musical scale (blue pentatonic scale). An example of the average and the nearest points of both areas which are selected from segmented areas for 50 frames is shown in figure 15. Area 1 is shown in figure 15(1), and area 2 is shown in figure 15(2). The volume of both sounds is shown in figure 16. Sound 1 created by area 1 is shown in figure 16(1), and sound 2 created by area 2 is shown in figure 16(2). Two sounds are created by these area data. Volume number means the quadrant for each area.

In figure 17, scores of created both sounds are shown. The result



(1) Frequency ($F_f + \triangle f$)



(2) Volume

Figure 11: An example of sound creation with depth data in method 1

score consists of four kinds of pitch corresponding to four kinds of tone.

3.2 Discussions

The system created sounds interactively by using depth data of objects including motion in two kinds of methods.

In method 1, the system created sound by fluctuating pitch frequency and harmonics structure. Although the sound is simple, it is interactive and various sound, because it is created by depth data of motion.

In method 2, the system created sound by mixing four kinds of tone which is controlled pitch and volume. In this experiment, the system created two sounds simultaneously. In this case, two sounds are created by two areas which are extracted from depth data, Because each sound is controlled by each area, the system can create two sounds asynchronously.

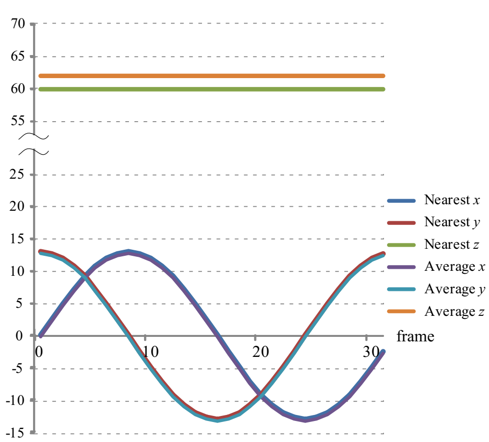


Figure 12: An example of statistical values of simulation data in method 2

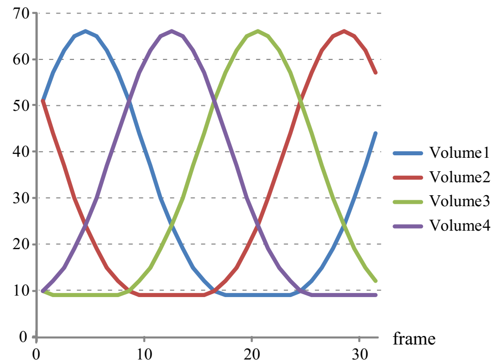
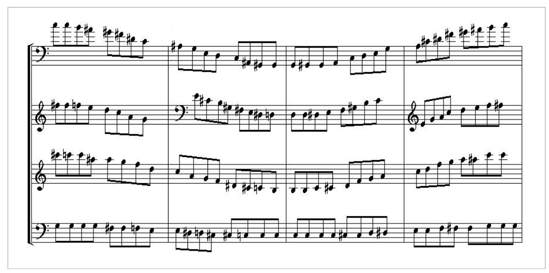
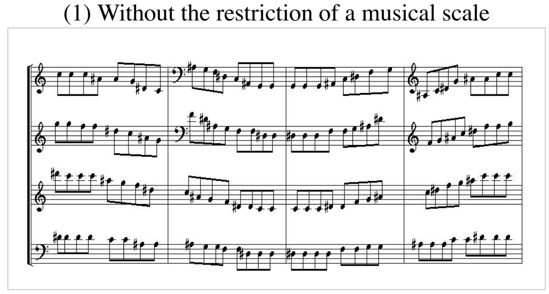


Figure 13: An example of volume of each tone with simulation data in method 2

(1) Without the restriction of a musical scale



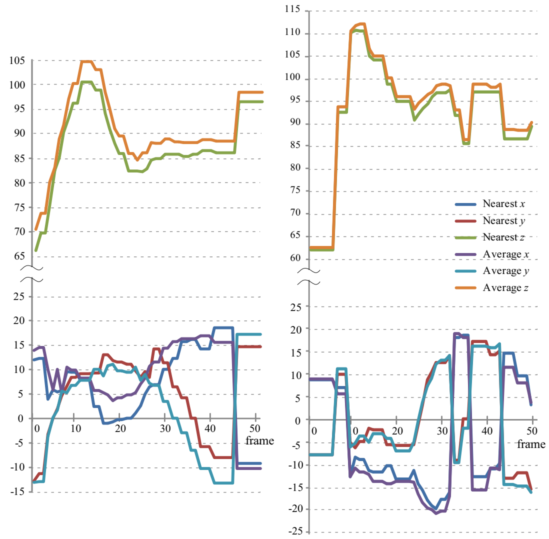
(2) With the restriction of a musical scale

Figure 14: The result score of four kinds of pitch

That is two sounds create two melodies which are made by different rhythms. It can be seen from scores in figure 17.

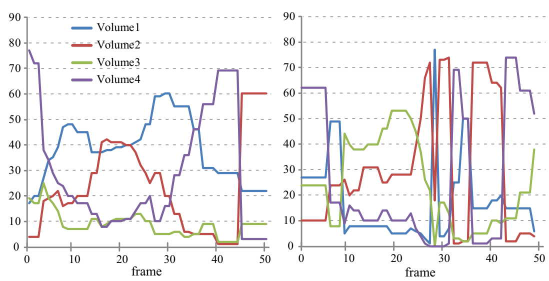
However, the positions of the nearest points change rapidly at several frames. It can be seen from the nearest point graph in figure 15. When an object moves, segmented areas change their positions. Therefore, a selected area in a previous frame may not be selected in a current frame, for example between frame 32 and frame 33, or the selected areas are exchanged each other, such as frame 42 and frame 43 in figure 15. In this case, sound pitch is changed unstably. It is necessary that the system keeps the continuity of the positions of the selected areas between frames.

Additionally, "cutoff frequency" was also changed by z variation. An example of the difference of z variation is shown in figure 18. The variation of figure 18(2) is more than the deviation (1). This parameter represents one characteristic of object shapes. Some tone has a good effect by changing "cutoff frequency",for example "brass", while some tone has less effect, for example "organ" and "timpani".



(1) Area 1 (2) Area 2

Figure 15: An example of statistic of depth data of two sounds in method 2



(1) Sound 1 (2) Sound 2

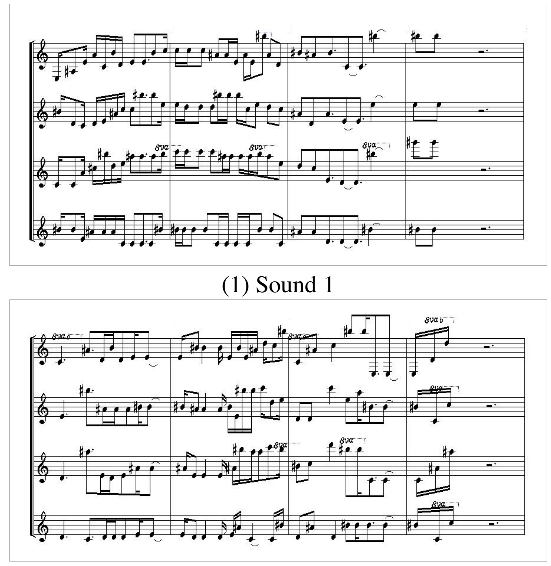
Figure 16: An example of result volume of each tone of two sounds in method 2

4 Conclusion

We proposed an interactive sound creation system with a depth camera. The proposed system can create sounds interactively with two kinds of methods from depth data of objects and its motion.

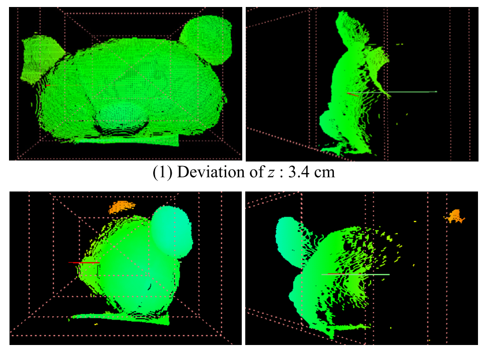
In method 1, sound effects, like *Theremin*, was able to be created by fluctuating pitch frequency and harmonics structure according to depth data. In this paper, pitch is fixed. If pitch can be changed by depth data, sound effects may be more interesting.

In method 2, a sound was able to be created by mixing four kinds of tone selected from the tone of a synthesizer according to depth data. When pitch is decided by using the restriction of a musical scale, melodies based on basic music theory are created automatically. Furthermore, when multi points are detected from depth data, we use two points in the case of this experiment, the system can create multi sound. Then in the experiments, we used "violin", "brass", "organ" and "timpani" as four tone and used "blue pentatonic scale" as a musical scale. When we choose some other four tones and some other musical tone, we can create various sounds and various melodies.

When we are creating sounds, we don't have to try to control pitch to play a musical instrument. We can play music unconsciously and, for example, cats also can create music by walking in front of the depth camera. However, for users which want to control pitch for themselves, it may be necessary that the system provides some way.

(1) Sound 1

(2) Sound 2

Figure 17: An example of a score of two sounds in method 2



(1) Deviation of z : 3.4 cm

(2) Deviation of z : 6.9cm

Figure 18: An example of displaying depth data representing deviation variation for changing "cutoff frequency". The left are front view images , and the right are side view images.

Acknowledgement

This work has been supported by "Foundation of Technology Supporting the Creation of Digital Media Contents" (CREST, JST), Japan.

References

[1] http://www.yamaha.co.jp/design/products/1990/miburi/.

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

[3] A. G. Mulder, S. S. Fels, and K. Mase. Design of virtual 3d instruments for musical interaction. In *Proceedings of the 1999 conference on Graphics interface*.

[4] F. Damiani, J. Manzolli, P. J. Tatsch, and A. M. Jr. A gestural control for a nonlinear sound synthesis method. In *Proceedings of the Third IEEE International Caracas Conference on Devices, Circuits and Systems, 2000*.

[5] C. Dobrian and F. Bevilacqua. Gestural control of music: using the vicon 8 motion capture system. In *Proceedings of the 2003 conference on New interfaces for musical expression*.

[6] http://www.midi.org.

[7] http://www.roland.com/products/en/exp/D_BEAM.html

[8] T. Oggier, M. Lehmann, R. Kaufmann, M. Schweizer, M. Richter, P. Metzler, G. Lang, F. Lustenberger, and N. Blanc. An all-solid-state optical range camera for 3D real-time imaging with sub-centimeter depth resolution (SwissRanger). In *Proc. SPIE*, volume 5249, pages 534–545, 2004.