# Privacy-aware Remote Monitoring System by Skeleton Recognition

Yoshihisa Nitta
Department of Computer Science
Faculty of Liberal Arts
Tsuda University
nitta@tsuda.ac.jp

Yuko Murayama
Department of Computer Science
Faculty of Liberal Arts
Tsuda University
murayama@tsuda.ac.jp

## Abstract

*As the number of elderly people living alone increases, the need for remote monitoring system is also increasing. The system automatically checks the safety of the elderly and notifies the state to remote areas in case of anomalies. However, how to protect the privacy of the watched person becomes a problem.*

*We propose that skeleton recognition technology is useful to monitor people with high accuracy while protecting the privacy. It can be used not only to investigate his/her posture and motion, but also to selectively analyze the voice emitted by himself/herself.*

*We developed a system that combines skeleton recognition and selective speech recognition by the audio direction. In this paper, we will explain the improvement of our system and report some experiment results.*

## 1. Introduction

Recently, the number of elderly people living alone in Japan continues to increase. In [1], the current state and trends on the elderly and their environment are stated as follows.

- Households with elderly people are about half of all households.
- "Single household" and "Couple only household" are the majority.
- The number of living with children is decreasing.
- The number of the elderly person living alone is increasing.

In particular, the increase of elderly people living alone for those 65 years of age or older is significant both for males and females.

The increase tendency of elderly people living alone is shown in Tab. 1 and Fig. 1. The data up to 2015 is based on the national census of the Ministry of Internal Affairs and Communications [2], and the data after 2016 is based on "Estimation of the number of Japanese households in the future (estimated in January 2013)" [3] of National Institute of Population and Social Security Research [4]. There were about 190,000 males and about 690,000 females living alone in 1980, and such men has increased to about 1.92 million and women to about 4 million in 2015.

As the number of elderly people living alone increases, the need for a system to watch the elderly from a remote place is also increasing. The system

**Table 1. Trends on elderly people living alone for aged 65 or over.**

| year | Number of Households (male) | (female) |
|---|---|---|
| 1980 | 193 | 688 |
| 1985 | 233 | 948 |
| 1990 | 310 | 1313 |
| 1995 | 460 | 1742 |
| 2000 | 742 | 2290 |
| 2005 | 1051 | 2814 |
| 2010 | 1386 | 3405 |
| 2015 | 1924 | 4003 |
| 2020 | 2173 | 4506 |
| 2025 | 2296 | 4710 |
| 2030 | 2433 | 4865 |
| 2035 | 2608 | 5014 |

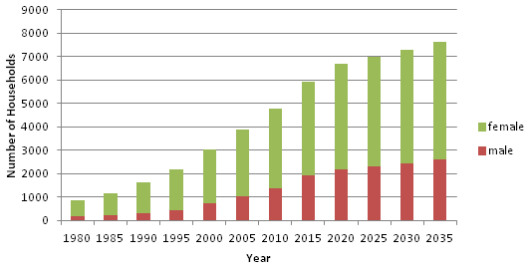



**Figure 1. Trends of elderly people living alone for those aged 65 or over.**

URI: https://hdl.handle.net/10125/59498
ISBN: 978-0-9981331-2-6

checks the state of the safety of the elderly and notifies the state to remote areas in case of anomalies.

## 2. Conventional system to watch the elderly living alone

Many systems have been proposed in the past to monitor the safely of elderly people living alone. Such watching services for the elderly are classified in [5] .

The Ramrock system [6] detects loitering and fall of an elderly alone for automatic warning, but sending images of a surveillance camera may cause privacy problems. In consideration of privacy, there are some methods using sensors other than camera to watch more loosely. In the system of Zojirushi [7], the usage records of a pot is sent twice a day, but such information is too indirect to detect emergency state. The system that watches the elderly by attaching sensors to furnitures such as a bed [8] has been studied, but it is inevitably a relatively large system. In the system of Philips [9] where a watched person wears a pendant type sensor for fall detection, there is a troublesome disadvantage that the sensor must be worn at all times. In the system of Fujitsu [10] which uses "sound analysis" of daily sound such as coughing and snoring to confirm the safety and estimates the health condition of elderly, the privacy is kept, but the accuracy of grasping the state of the elderly from the sound might become a problem.

The important points for watching the elderly living alone are as follows.

- Do not wear special devices such as sensors and markers.
- Use sensors other than surveillance camera, with which accurate state can be grasped with privacy.

From the above consideration, the method of recognizing the skeleton with non-attached sensor and directly grasping the human motion and posture is a superior method to watch people with privacy, and has a great advantage for other methods.

## 3. Skeleton recognition technology

It is getting popular to obtain skeleton information with the image cameras and depth sensors.

To grasp the safety and health condition without attaching a special device to the target person, skeleton recognition technology is useful. Furthermore, proper use of this technology will not cause privacy problems.

### 3.1. Conventional technology for skeleton recognition

The traditional way for skeleton recognition has been conducted by use of marker based systems [11][12][13]. A subject needs to put on markers around his/her body and the locations of markers are tracked and detected out of images and motion pictures taken by camera. Tracking would be done various ways such as optical motion picture and the use of an infrared camera. While marker-based systems have been used extensively, they are expensive.

Recently a more economical markerless systems, *Kinect for Windows V2* [14][15] (hereinafter called *"Kinect V2"*) is available.

Leap motion [16] is also a device that does not require a marker, but the aim is mainly to recognize the position and movement of the hands.

### 3.2. Kinect V2

The Kinect V2 developed by Microsoft is a device with many functions such as skeleton tracking, face tracking and voice direction acquisition. Human body data can be obtained with sufficient accuracy without contact with a special marker on the human body. Woolford [17] suggests that such a system, the Kinect V2, may be useable compare to a traditional clinical system, in a healthcare environment because the target person does not need to put on any marker device so that tracking would be done without making physical contact with that person. That is, the skeleton can be recognized cheaply with sufficient accuracy using Kinect V2.

*Kinect for Windows SDK 2.0* is the official SDK of Kinect V2. The original code API for C++ of the SDK consists of too many methods; 54 kinds of *Interface*, and the total number of methods is 277 [18].

### 3.3. NtKinect library

To improve the difficulty of using the official SDK, we have developed a class library *NtKinect* [19] [20] for C++ so that it is easier for programming to use the Kinect V2. The library has been released as an Open Source of MIT license.

With our library, one can easily perform skeleton tracking (Body Framework information) as well as face recognition [21][22]. Furthermore, we developed *NtKinectDLL* [23] which makes *NtKinect* a Dynamic Linking Library. This makes *NtKinect* available in many other programming languages and development environments like *Unity*. We distribute *NtKinectDLL* with wrappers for C# (in Unity) and Python programming language.

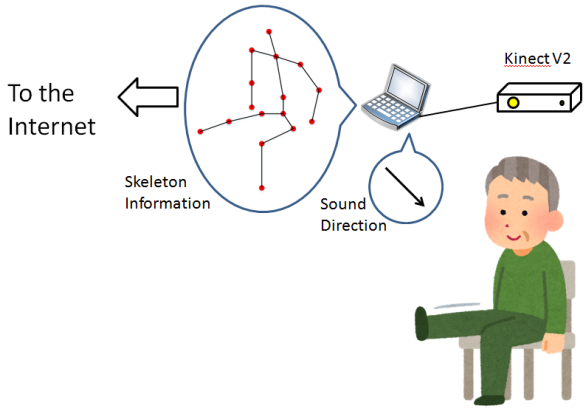



**Figure 2. Watching the elderly with Privacy**

**Figure 3. Skeleton Recognition and RGB Image.**

The tool has been distributed widely and used by several users already for their research work such as natural user interface and computer art systems [24][25][26][27][28][29][30].

## 4. Watching over by skeleton recognition

### 4.1. New improvement of our monitoring system

We have been facing a demographic problem of increasing population of elderly in Japan. A remote monitoring for their safety and security has been implemented to monitor those who live alone. As mentioned in Section 2, several attempts have been made with IT technology [6] [7] [8] [9] [10]. While these attempts are useful, there remain some issues of privacy and accurate grasp to be solved.

We proposed that skeleton recognition is useful for monitoring people with privacy and have developed such a monitoring system [31]. The outline of the system is shown in Fig. 2.

In this paper, we newly added the follwoing improvements to the system.

**Figure 4. Representaion of Skeleton using wireframe.**

**Figure 5. Representation of posture using Avatar.**

- To prevent identifying the individual from skeleton featuers and habits of actions.
- To recognize speech emitted by watching target selectively.

### 4.2. Preventing identifying the individual

Skeleton expression of human posture is useful for anonymity, but there is a possibility of knowing “who they are” and “what they are doing” from the features of the length between the joints or their habits of movement [32][33].

If the skeleton is recognized in Fig. 3 and the results are displayed as it is in Fig. 4, an individual may be identified from skeleton features like the length of arms and legs, shoulder width, etc.

For this reason, instead of directly using the position of the joint recognized, we improved to use the angle of the joints. and display the posture with avatars (3D Model) like Fig. 5. To do this, we use the Unity C# wrapper of *NtKinectDLL*.

Also, in order to prevent estimation of gestures and motions of the target from sequential skeleton information, we reduced the frame rate of posture

display. By dropping its rate to about 1 frame per second, the third person cannot estimate the action that the target is performing, but we can still detect the target's anomalies.

### 4.3. Use of sounds emitted by the target person

Speech recognition of the target is useful for detecting danger and anomalies, especially when the speaker needs help. However, everyday life is full of various sounds. Even if a human voice is detected, it has nothing to do with the state of the target person, if it is a voice emitted from television, radio, etc. In order to watch with sound, it is necessary to separate the sound emitted by the watching target from other sounds.

In our system, the position of the target is detected by the skeleton recognition. So, if the direction of the sound can be detected correctly, by comparing with the position of the skeleton, the target person's voice can be separated from other sounds.

However, to this purpose, it is necessary to verify the accuracy of the detected voice direction. We experimented the accuracy of the speech direction of Kinect V2. The result of the experiments is shown in chapter 5.

## 5. Experiment: Detection of speech direction and its accuracy

In our watching over system, Kinect V2 is used to perform skeleton recognition, voice direction detection, and voice data acquisition. Kinect V2 detects the sound direction by four array microphones lined up in its front. However, sound waves may be reflected by walls and floors. When sound reaches the microphones through multiple paths, it is difficult to detect the sound direction correctly.

Experiments measuring the accuracy of sound direction are greatly affected by the shape of the room and the arrangement of furnitures, so it is diffucult to obtain data that is acceptable everywhere. Therefore, in this experiment, we measured the sound direction in our laboratory sized near the assumed elderly room and examined the trend of the data.

In this experiment, a subject stood 2 meters apart from Kinect V2 device and spoke to the 4 different directions, (a), (b), (c) and (d).

- (a) facing the center of Kinect V2
- (b) the direction rotated by $90^\circ$ degree from (a)
- (c) the direction rotated by $180^\circ$ degree from (a)

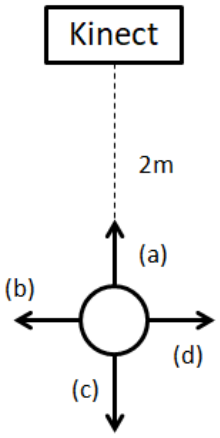



**Figure 6. Audio Direction from Front**

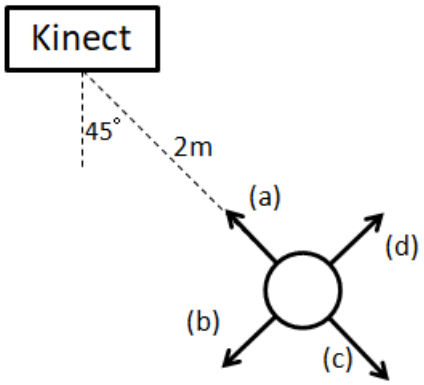



**Figure 7. Audio Direction from 45 degree**

- (d) the direction rotated by $270^\circ$ degree from (a)

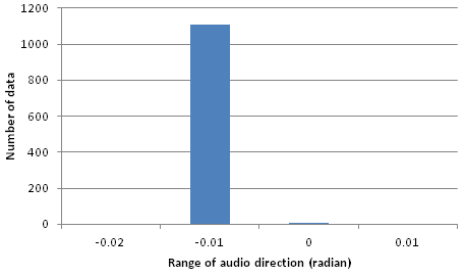
The voice direction detected by the Kinect V2 device is recorded. We performed this experiment in two kinds of positions, one from the front of Kinect V2 (Fig. 6), the other from the $45^\circ$ degree position (Fig. 7). The measurement results are shown in Fig. 8 and Fig. 9. The vertical axis of the graph is the number of data of the sound direction measured with the confidence factor 0.5 or more, and the horizontal axis is the detected direction in radians.

Table 2 shows the range of the fluctuation of the detected sound direction in each case.

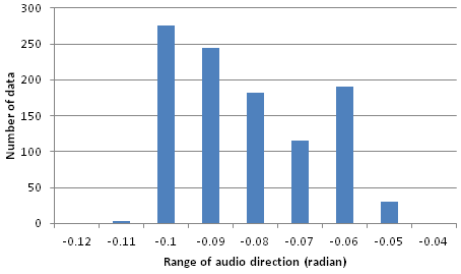
When the subject spoke from the front of the device as in Fig. 6, varying the face direction of (a), (b), (c) and (d), the measured sound direction is shown in Fig. 8. The range of the sound direction in this experiment is shown on the left side of Tab. 2. The varying

**Table 2. Experiment result of audio direction.**

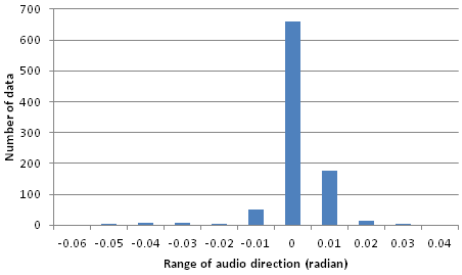
| position | Front | | | | $45°degree$ | | | |
|---|---|---|---|---|---|---|---|---|
| rotation | (a) | (b) | (c) | (d) | (a) | (b) | (c) | (d) |
| direction range (radian) | 0.01 | 0.06 | 0.03 | 0.03 | 0.05 | 0.03 | 0.21 | 0.10 |
| direction range (degree) | 0.57 | 3.44 | 1.72 | 1.72 | 2.87 | 1.72 | 12.0 | 5.73 |



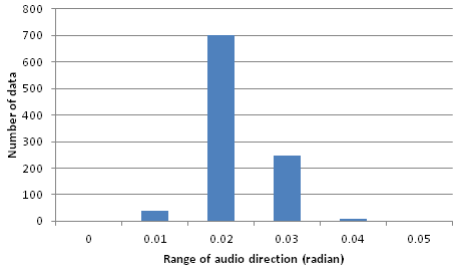

(a) $+0°$




(b) $+90°$

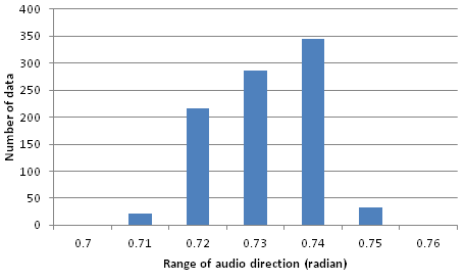



(c) $+180°$

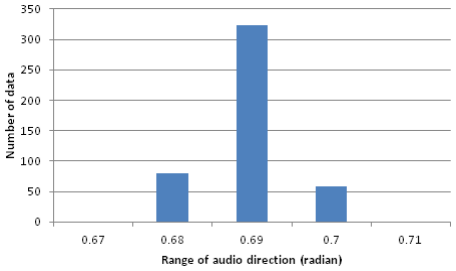



(d) $+270°$

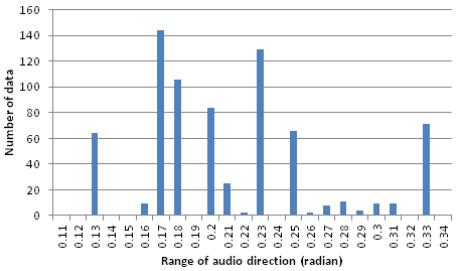
**Figure 8. Audio Direction from the Front.**



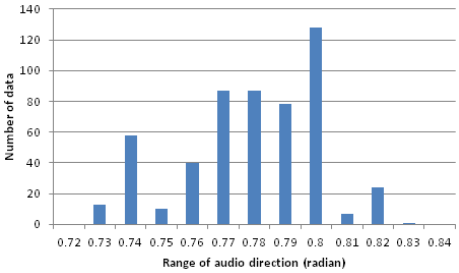

(a) $+0°$




(b) $+90°$




(c) $+180°$




(d) $+270°$

**Figure 9. Audio Direction from 45 degree**

```
<?xml version="1.0" encoding="utf-8" ?>
<grammar version="1.0" xml:lang="en-US"
  root="rootRule" tag-format="semantics/1.0-literals"
  xmlns="http://www.w3.org/2001/06/grammar">
  <rule id="rootRule">
    <one-of>
      <item>
        <tag>HELP</tag>
        <one-of>
          <item> Help </item>
          <item> Call </item>
        </one-of>
      </item>
      <item>
        <tag>AMBULANCE</tag>
        <one-of>
          <item> Amblulance </item>
        </one-of>
      </item>
    </one-of>
  </rule>
</grammar>
```

**Figure 10. Example of *MS Speech SDK*'s grammar file**

range in the voice direction is between $0.57°$, and $3.44°$. The minimum case is (a) and the the maximum case is (b). From the front position, we can detect the sound direction stably in all the case of (a), (b), (c) and (d).

When the subject spoke from the $45°$ degree position as in Fig. 7, varying the face direction of (a), (b), (c) and (d), the measured sound direction shown in Fig. 9. Since it is $45°$ degree, the value $\frac{\pi}{4} \cong 0.785$ should be detected in radians, but slightly smaller values tended to be obtained in the experiment. Since the measured values can be compensated to the correct values, it is important that the detected values are stable. The range in the voice direction in this experiment is shown on the right side of Tab.2. The varying range in the voice direction is between $1.72°$ and $12.0°$. The minimum case is (b), and the maximum case is (c). From the $45°$ position, the varying range in the sound direction fell within $2.85°$ only in cases (a) and (b). But, in case of (c) and (d), the obtained voice direction changes widely. We think this is because the sounds reached to the microphones through multiple paths by reflections.

From the above results, it can be said that the correct sound direction can be detected relatively stably when facing at least in the direction of (a) at both of the positions Fig. 6 and Fig. 7. That is, it is considered that the direction of the voice consciously issued toward the Kinect V2 device by the target can be correctly detected.

## 6. Recognition of speech content

In our system, we use the following two services for speech recognition.

- Microsoft, Speech Platform SDK v11 [34] (hereinafter called *"MS Speech SDK"*)
- Gooble Cloud Platform, Speech API [35] (hereinafter called *"Google Speech API"*)

```
200
{
  "results": [
    {
      "alternatives": [
        {
          "transcript": "help me call the ambulance",
          "confidence": 0.91519946
        }
      ]
    }
  ]
}
```

**Figure 11. Example of *Google Speech API*'s Recognition Result**

### 6.1. MS Speech SDK

In order to use *MS Speech SDK*, it is necessary to register words to be recognized in the grammar file. An example of grammar file is shown in Fig. 10. *MS Speech SDK* can recognize speech on the local computer, so it has advantages both in network bandwidth and privacy. Because some people can not put up with the situation being eavesdropped, even if it is mechanically processed.

### 6.2. Google Speech API

In order to use *Google Speech API*, there is no need to register words in advance. But it is necessary to send voice data to *Google Cloud Platform* through the network. This can be a drawback for the network bandwidth and privacy. An example of voice recognition result of *Google Speech API* is shown in Fig. 11.

The recognition accuracy of *Google Speech API* is higher than that of *MS Speech SDK* .

### 6.3. Speech Recognition and Privacy

What is important in speech recognition is that the watched people do not feel anxious about invasion of their privacy. Therefore, it is not appropriate to send voice data of everyday life constantly to the network. So, we adopted a combination of the good points of the above two services. In our system, voice data is processed mainly by *MS Speech SDK* on the local computer. When some specific keywords are recognized locally, a series of subsequent voice data is sent to the cloud of *Google Speech API* through the network and recognized.

In order to recognize the voice content accurately, it is necessary to acquire the voice sound without interruption. But using a Kinect V2 device, the audio data is acquired intermittently, because audio data cannot be acquired during performing skeleton recognition. This might deteriorate the precision of the speech recognition [36]. In our system, sound can be acquired as almost continuous data without interruption and the contents of the speech can be recognized smoothly. This is because the frequency of skeleton recognition is set to once per second to prevent identification of the individual and his/her action.

## 7. Conclusion

We propose that the following two points are useful for the remote monitoring system with privacy.

- Skeleton recognition
- Selective speech recognition of the target person

We developed such a remote monitoring system which has the following features.

- In order to eliminate the possibility of identifying an individual from the skeleton, the system expresses the skeleton with the angle of the joints and display it as a 3D avatar.
- Drawing frame rate is reduced in order to eliminate the possibility of the identifying the individual and his/her actions.
- The speech emitted by the watching target is recognized selectively, and speech is recognized with privacy.

We conducted experiments to make sure that sound direction detected by Kinect V2 can be used for selective speech recognition. We got the result that the voice direction can be used when the voice is emitted toward Kinect V2 device.

The advantage of monitoring people using the skeleton recognition is not only that privacy can be protected, but also that data is smaller than conventional method of using live streaming of images. Since the position or angle of one joint can be represented by three 32-bits floating point numbers, in the case of expressing skeleton information of a person with 25 points, it can be represented by $4 \times 3 = 300$ byte. This is considered to be very advantageous in terms of network traffic volume.

In our system, speech recognition does not generate network traffic so much, because the sound data is selected by its direction and the speech recognition is performed only on the local computer in most cases.

Consequently, our system is considered to be suitable not only for monitoring with high privacy, but also for monitoring through narrow bandwidth network.

## References


[1] Cabinet Office of Japan, "White Paper on Aging Society: 2017." http://www8.cao.go.jp/kourei/whitepaper/w-2017/html/zenbun/index.html, (2018/03/24 access) *(in Japanese)*.

[2] Ministry of Internal Affairs and Communication, JAPAN, "Information and Communications Policy Site." http://www.soumu.go.jp/main_sosiki/joho_tsusin/eng/index.html, (2018/06/14 access).

[3] Cabinet Office of Japan, "Number and Percentage of households with persons over 65 years old, White Paper on Aging Society: 2017 (in Japanese)." http://www8.cao.go.jp/kourei/whitepaper/w-2017/html/zenbun/csv/z1_2_1_01.csv, (2018/03/24 access).

[4] National Institute of Population and Social Security Research, "National Institute of Population and Social Security Research." http://www.ipss.go.jp/index-e.asp, (2018/06/14 access).

[5] T. Seiki, S. Sachio, M. Shinsuke, and N. Masahide, "A Classification Method of Remote Monitoring Service for Elderly Person," *IEICE technical report, Vol. 113, No.469*, pp. 169–174, 2014 *(in Japanese)*.

[6] Ramrock, "Care Support System 'Ramrock System'." http://www.ramrock.co.jp/, (2017/08/29 access) *(in Japanese)*.

[7] Zojirushi, "MIMAMORI Hot Line." http://www.mimamori.net/, (2017/08/29 access) *(in Japanese)*.

[8] D. Bradford, J. Freyne, and M. Karunanithi, "Sensors on My Bed: The Ups and Downs of In-Home Monitoring," *vol.7910, Lecture Notes in Computer ScienceSpringer Berlin Heidelberg*, 2013.

[9] Philips, "Medical Alert Service." http://www.lifeline.philips.com/, (2017/08/29 access).

[10] Fujitsu, ""The deciding factor is IoT and sound", Elderly monitoring service to the next step." http://www.fujitsu.com/jp/innovation/digital/life/dl-contents/2017/topics-05/, (2017/06/08 access) *(in Japanese)*.

[11] A. C. Sementille, L. E. Lourenço, J. R. F. Brega, and I. Rodello, "A motion capture system using passive markers," *Proc. of the 2004 ACM SIGGRAPH international conference on Virtual Reality continuum and its applications in industry (VRCAI '04)*, pp. 440–447, 2004.

[12] A. Barber, D. Cosker, O. James, T. Waine, and R. Patel, "Camera tracking in visual effects an industry perspective of structure from motion," *ACM Proc. of the 2016 Symposium on Digital Production (DigiPro '16)*, pp. 45–54, 2016.

[13] M. Schröder, J. Maycock, and M. Botsch, "Reduced marker layouts for optical motion capture of hands," *Proc of the 8th ACM SIGGRAPH Conference on Motion in Games (MIG '15)*, pp. 7–16, 2015.

[14] Microsoft, "Meet Kinect for Windows." https://developer.microsoft.com/en-us/windows/kinect, (2018/06/08 access).

[15] Microsoft, "Kinect for Windows SDK 2.0, Programming Guide, Body tracking." https://msdn.microsoft.com/en-us/library/dn799273.aspx, (2017/09/03 access).

[16] D. Avola, A. Petracca, G. Placidi, M. Spezialetti, L. Cinque, and S. Levialdi, "Markerless Hand Gesture Interface Based on LEAP Motion Controller," *Proc. of the 20th International Conference on Distributed Multimedia Systems: Research papers on distributed multimedia systems, distance education technologies and visual languages and computing*, pp. 27–29, 2014.

[17] K. Woolford, "Defining accuracy in the use of Kinect v2 for exercise monitoring," *ACM, Proc. of the 2nd International Workshop on Movement and Computing (MOCO '15)*, pp. 112–119, 2015.

[18] Microsoft, "Kinect for Windows SDK C++ Reference." https://msdn.microsoft.com/en-us/library/dn791993.aspx, (2018/06/08 access).

[19] Y. Nitta, "NtKinect: C++ Class Library for Kinect V2," *the 172-th conference SIG Human-Computer Interaction*, 2017 *(in Japanese)*.

[20] Y. Nitta, "NtKinect: Kinect V2 C++ Programming with OpenCV on Windows10." http://nw.tsuda.ac.jp/lec/kinect2/index-en.html, (2018/06/08 access).

[21] Y. Nitta, "NtKinect: How to recognize human skeleton with Kinect V2." http://nw.tsuda.ac.jp/lec/kinect2/KinectV2_skeleton/index-en.html, (2018/06/08 access).

[22] Y. Nitta, "NtKinect: How to recognize human face with Kinect V2 in ColorSpace coordinate system." http://nw.tsuda.ac.jp/lec/kinect2/KinectV2_face/index-en.html, (2018/06/08 access).

[23] Y. Nitta, "NtKinectDLL - DLL and Wrappers (Unity C#, Python) for NtKinect." http://nw.tsuda.ac.jp/lec/NtKinectDLL/index-en.html, (2018/03/31 access).

[24] H. Ichikawa, S. Iijima, and Y. Nitta, "Natural User Interface using Gesture on VR Space," *the 178-th conference SIG Human-Computer Interaction*, 2018 *(in Japanese)*.

[25] M. Tsuchiya, T. Itoh, and Y. Nitta, "Interactive Light Painting System Using Human Recognition," *NICOGRAPH 2016, P-4, The Society for Art and Science*, 2016 *(in Japanese)*.

[26] M. Tsuchiya, T. Itoh, and Y. Nitta, "An Interactive System for Light-Art-Like Representation of Human Silhouettes," *WISS 2016, P-213, JSSST*, 2016 *(in Japanese)*.

[27] M. Tsuchiya, T. Itoh, and Y. Nitta, "An Interactive System for Light-Art-Like Representation of Human Silhouettes," *Interaction 2017, IPSJ*, 2017 *(in Japanese)*.

[28] M. Tsuchiya, T. Itoh, Y. Nitta, M. Neff, and Y. Liu, "A System for Light-Art-Like Representation of Human Silhouettes," *ITE-AIT2018-71, Vol. 42, no. 12*, pp. 107–110, 2018 *(in Japanese)*.

[29] M. Tsuchiya, T. Itoh, Y. Nitta, M. Neff, and Y. Liu, "An Interactive System for Light-Art-Like Representation of Human Silhouettes," *Interaction 2018, IPSJ*, 2018 *(in Japanese)*.

[30] F. Kinugawa, Y. Hayashi, and K. Seta, "Posing Learning Environment Aiming at Improvement of Posture Control Ability," *JSiSE Student Research Presentation 2017, pp.97-98*, 2018 *(in Japanese)*.

[31] Y. Nitta and Murayama, "Software Support for Skeleton Recognition and Monitoring People with Privacy," *Proceedings of the 51st Hawaii International Conference on System Science (HICSS-51)*, pp. 200–206, 2018.

[32] F. Gossen and T. Margaria, "Comprehensive people recognition using the Kinect's face and skeleton model," *2016 IEEE International Conference on AQRT*, 2016.

[33] S. Yoshida, M. Izumi, and H. Tsuji, "A research on the ability of Kinect to discriminate people," *ITE Technical Report Vol.36, No.8, ME2012-32*, 2012 *(in Japanese)*.

[34] Y. Nitta, "NtKinect: How to recognize speech with Kinect V2." http://nw.tsuda.ac.jp/lec/kinect2/KinectV2_speech/index-en.html, (2018/03/02 access).

[35] Y. Nitta, "NtKinect: How to recognize Kinect V2 audio by Cloud Speech API of Google Cloud Platform." http://nw.tsuda.ac.jp/lec/kinect2/KinectV2_GoogleSpeech/index-en.html, (2018/03/02 access).

[36] Y. Nitta, "NtKinect: How to run Kinect V2 in a multi-thread environment." http://nw.tsuda.ac.jp/lec/kinect2/KinectV2_thread/index-en.html, (2018/03/25 access).