Marker-Less Piano Fingering Recognition using Sequential Depth Images

Akiya Oka, Manabu Hashimoto
Graduate School of Information Science and Technology
Chukyo University
101, Tokodachi, Kaizu-cho, Toyota-shi, Aichi, Japan 470-0393
Email: {oka, mana}@isl.sist.chukyo-u.ac.jp

***Abstract*—Piano fingering is one of the important skills for piano performance, especially for beginners. Consequently, technology for recognizing a player's fingering is required in order to develop an automated piano lesson system. The term "piano fingering" refers to which fingers are used for pressing piano keys. In this paper, we propose a method for recognizing piano fingering by analyzing motion of multiple fingers of a piano player through the use of depth images continuously acquired with a depth sensor. Our method makes it possible to develop a practical system that does not require the use of any special markers such as color labels on fingers. First, a dictionary data set for various fingering patterns is registered. Each data element consists of a depth image, the name of a pressed key, correct information for its fingering, and the wrist positions of the player in an image. Next, a fingering pattern for unknown depth images is identified by matching acquired images to those in the dictionary data set. To reduce the search space size, the wrist position detected from an input image and a note name signal obtained from a MIDI keyboard are effectively used. The Nearest Neighbor search algorithm is utilized to search for solutions. Experimental results obtained using actual piano pieces for beginners demonstrate that the system achieves an 91.6% recognition rate and that its processing time is less than 120 msec per note.**

I. Introduction

In this paper, we define "piano fingering" as a combination of a finger name and the name of a note played by pressing a piano key. Since playing the piano with correct fingering is very important for beginners[1], composers often denote fingering information on a musical score. However, when beginners play the piano, they often find it difficult to discern whether their fingering is correct or not by themselves. For them it would be helpful to have an assistance system that could automatically recognize their piano fingering and indicate their mistakes.

With respect to related work on recognition methods involving the human hand, a number of studies have been conducted in the field of gesture recognition [2]. However, there are few examples of piano fingering recognition because it requires very precise ways of measuring complicated finger movements. One previous research work about piano fingering recognition is based on region extraction using the background-subtraction technique with a standard CCD camera[3]. However, it is difficult to extract hand regions stably because the contrast between the hand and the piano's white keys is not very high. Another method[4] utilizes a motion capture system for measuring finger positions three-dimensionally and is used for generating high quality computer graphics. Still another method[5] makes use of a monocular camera, where the objective was to achieve practical performance. The important thing is that all of these methods require the use of specified markers such as optical reflective material or color labels on the palms or fingers. These markers surely must be useful for achieving reliable recognition, but they will disturb beginners' natural piano practice.

The purpose of this research is to propose a method to recognize piano fingering without the need for any markers. In this paper, we propose a method for recognizing fingering by matching input depth images to images in a dictionary data set. Depth image is more suitable for extracting than CCD camera image, even in various skin colors or illumination conditions. The data consists of depth images acquired when players press the keys, note names, and finger names. In the recognition process, wrist positions detected from input images and note name signals obtained from a MIDI keyboard are utilized to reduce the number of candidates to be compared with the input data. Real-time recognition is achieved through efficient prescreening of the dictionary.

II. Proposed method

A. Representations of fingering information

As stated previously, we define "fingering information" as a combination of a finger name and the name of a note played by pressing a piano key. Hereafter we refer to these names as "identification codes" or simply "codes". We assign identification codes to each of the 88 piano keys (notes) and to each of the ten fingers of the hands. The finger codes are assigned in order from the thumb to the little finger as shown in figure 1(a). Each of the key codes is a combination of the relative note name and its absolute pitch as shown in figure 1(b). For example, if the fingering information is "R1C4", it means that the key C4(center "C") is pressed by the finger R1 (right thumb).

B. Overview of proposed method

The proposed method consists of a learning module and a recognition module. Figure 2 shows a schematic block diagram of the method.

978-1-4673-5621-3/13/$31.00

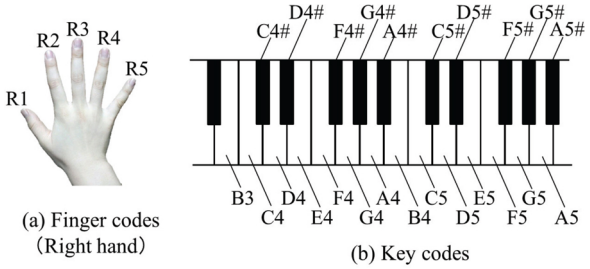

Fig. 1. Definition of code-numbers of fingers and piano keys.

The learning module generates a dictionary data set for the recognition module. Basically, the dictionary data set comprises a lot of depth images for various key pressing patterns. Each image has certain attributes that function as supervising information, i.e., a finger code, a key code, and a wrist position. The recognition module uses Nearest Neighbor search algorithm to compare input depth images with all possible dictionary data elements. The final output, i.e., the fingering recognition result, is determined as the dictionary data set that is most similar to and consistent with the input image. Incidentally, we should note that using an electronic piano makes it easy to get signals for the names of notes played in pressing piano keys by using a standard MIDI interface. The nearest neighbor algorithm is based on simple distance comparison between an input data and large amount of dictionary data. So, we can easily screen the dictionary data in advance, using attribute of the input data such as note name or wrist position. We therefore use these signals for screening the dictionary data set. The dictionary data set space size is initially very large but can be reduced by extracting data that has the key code (note name) attributes. This data reduction makes high-speed searching for solutions possible. In this research, we carried out the screening process by using wrist positions as well as note names.

C. Learning: Dictionary created using MIDI signal

Figure 3 shows flow of getting learning images using MIDI signal. In this research, depth images were obtained through the use of a versatile depth sensor, i.e., the Kinect sensor manufactured by Microsoft Corp. Acquired depth images are pre-processed by a spatial median filter to reduce random noise. Hand regions are then extracted. The depth sensor is fixed tightly and firmly above the piano keyboard to ensure that the distance between the sensor and the keyboard remains constant. This enables us to use the background subtraction technique to extract hand regions from images. Here, the background depth images were captured in advance, i.e., before the piano was played. The threshold value for binarizing the subtracted depth images was determined experimentally.

Next, We explain how to determine the wrist position. In general, the direction of the wrist's motion in piano playing is parallel to the keyboard line. Therefore, we can prepare a ROI (region of interest) in an input image for wrist detection. The left and right edges of the wrist are detected by scanning

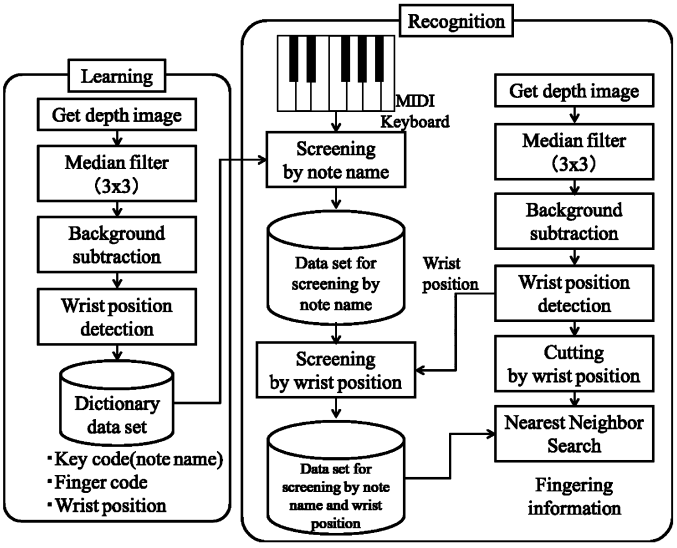

Fig. 2. Block diagram of proposed learning and recognition.

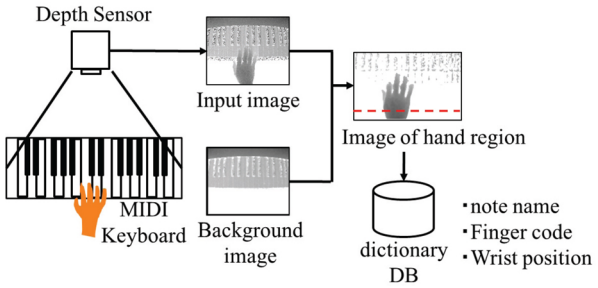

Fig. 3. A flow of generating a dictionary database consists of depth images and MIDI signals.

the ROI and the wrist position is determined as the center of both side edges.

The MIDI standard uses the terms "note-ON" and "note-OFF" signals to respectively mean the timing for the finger to begin and finish pressing a piano key. Therefore, using these signals makes it possible to determine the appropriate timing for transferring a trigger signal to the depth sensor for image capturing. Learning images are provided with codes identifying the finger name, note name and wrist position. Figure 4 shows an example of images which have been used for learning.

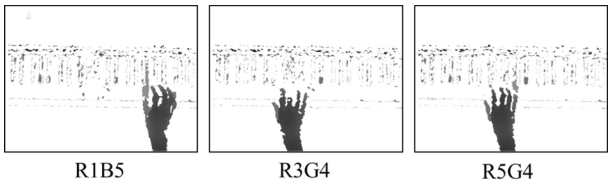

Fig. 4. Example of depth images stored in the dictionary DB. 780 images of hand region with correct fingering data are registered.

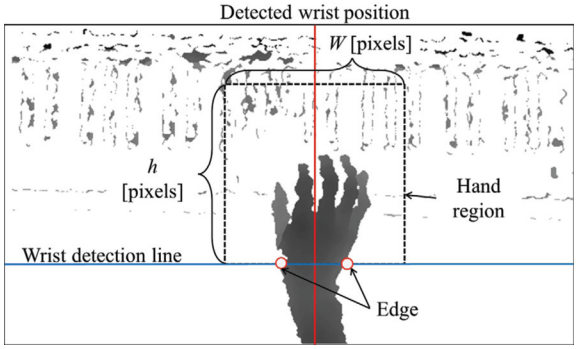

Fig. 5. Detection of the wrist position from a depth image.

D. Recognition: Fingering recognition using Nearest Neighbor method

Our method's basic recognition process scheme is based on the Nearest-Neighbor search technique.

The size of the data set generated as described in Subsection is reduced by screening using note name information obtained as MIDI signals. We calculate the wrist position from the input image in a similar way in the learning module. This information can be also used for screening the dictionary data set. Differences between the wrist position in the input image and those in the dictionary are then calculated. If the distance value is larger than d pixels, the learning data is discarded from the dictionary data set. Before executing the recognition process, sub-regions are extracted from an input image and images in the dictionary data set in advance. The sub-regions include the hand region. Actually, a $w \times h$ pixel sub-region is automatically extracted using wrist position data as the center of the sub-region. This region size was determined so as to include the hand region through preliminary experiments. Figure 5 shows the process for detecting the wrist position from the image. The horizontal line in the figure is for detecting the edges of the right wrist and the wrist's center position.

In the Nearest Neighbor search scheme, images are represented as vectors of 40,000 dimensions. The similarity between two images is calculated as the distance between two vectors in a feature space. A vector corresponding to an input image is compared with all vectors in the dictionary data set. Here, the Euclidean distance is used to calculate the distance between the vectors of an input image and learning images. Our method achieves high speed and high reliability by reconstructing the reduced dictionary data set dynamically using note name signal and wrist position information. Figure 6 shows an example of two-step reduction of learning data set by screening. In this example, five classes of learning data in the original dictionary are reduced to three classes by screening using the note name signal "C4" and further reduced to two classes by using wrist position information. We can execute efficient searches by using such a reconstructed dictionary instead of the original dictionary.

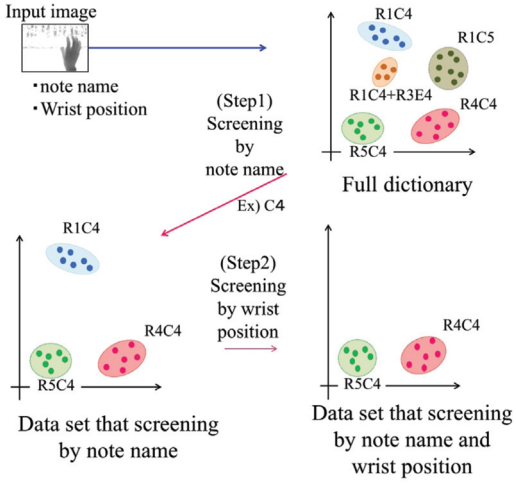

Fig. 6. Reduction of search space by screening using note name signal and wrist position.

TABLE I
RECOGNITION RATES USING IMAGE OF MONOTONE AND TRIAD.

Input image	Recognition rate[%]
Image of Monotone	93.2(110/118)
Image of Triad	95.7(112/117)

III. EXPERIMENTS AND DISCUSSION

A. Experimental set up

In this experiment, we used the Kinect sensor mentioned in Subsection II-D to obtain depth images. The sensor was fixed to a height of 70 cm above the keyboard surface. This setup enables measurement of the approximately four-octave depth of a piano keyboard. A Database for the recognition of the right hand consists of D4 to B5 of the note name. A Database of the left hand consists of G2 to E4 of the note name. Six learning data elements per class are stored. When screening database using wrist position, threshold d is set to 25. The size of cut-out of hand region w,h are set to 200 pixels.

B. Performance for keying a chord

We measured the recognition rate using images of monotone and triad. Figure 7 shows examples of monotone and triad images. In this way, there is difference the shape of a hand due to the number of keying note. We used 118 images of monotone and 117 images of triad as test images. Table I shows the result of recognition by using these images. Recognition rate of triad is higher than monotone. The shape of keying triad has stronger constraint than that of keying monotone. So, variation of input image is small.

C. Estimation of recognition rate in music for beginners

The well-known songs "Mary had a Little Lamb" and "Lightly Row" and J.S. Bach's "Menuet in G Minor" were used to experimentally estimate the recognition success rate. Figure 8 shows the music for J.S. Bach's "Menuet"[6]. "Mary

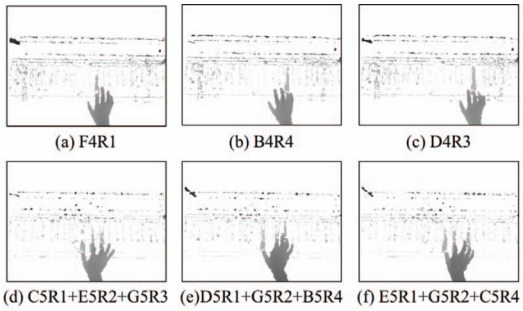

Fig. 7. Example of image of monotone and triad.(a)∼(c) are images of keying monotone.(d)∼(f) are that of keying triad.

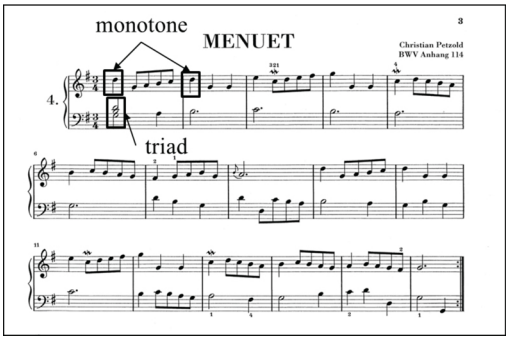

Fig. 8. Example of sheet music for J.S. Bach's "Menuet in G Minor"[6].(Opening 16 bars are illustrated.)

had a Little Lamb" and "Lightly Row", only the right hand part was performed and its recognition success rate estimated in the experiment. "Menuet" is played with both hands. The dictionary dataset comprised 780 data elements in 130 classes. The finger name assigned in the musical score was used for the correct fingering for learning data. If no finger name was assigned, we added it to reflect the natural performance for a beginner pianist.

As Table II shows, the recognition rates were 100 for "Mary had a Little Lamb" and "Lightly Row". The main reason such rates were achieved is the relatively slow tempo of these songs. In comparison, the recognition rates for "Menuet" were low. The reason for the comparatively low rates may be that the thumb was hidden by the other fingers as shown in Figure 9(a). Figure 9(b) shows recognition result, when we used Figure 9(b) as input image. Occlusion of this type presents a difficult problem for our method.

TABLE II
RECOGNITION RATES USING MUSIC FOR BEGINNERS.

Song title	Recognition rate[%]
"Mary had a Little Lamb"	100.0
"Lightly Row"	100.0
J.S.Bach "Menuet"	91.6

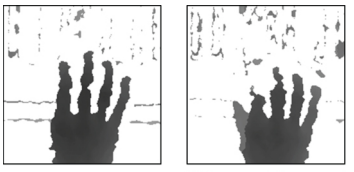

Fig. 9. Example of recognition success and failure.

D. Processing time of the proposed method

An ordinary desktop PC was used to measure the processing time of our method's modules, which consists of preprocessing and recognition. We used the dictionary data set (780 data elements in 130 classes) to measure the processing time each module needed to recognize a single input image. The results are shown in Table III. The recognition time was 120 msec per image. Limitation of proposed method is 8 keying per sec. For example, this is corresponding to an eighth note in allegro of tempo marks. This means that our method is applicable to slow-tempo songs for piano beginners.

TABLE III
PROCESSING TIME FOR EACH MODULE.(CPU: INTEL PENTIUM DUAL-CORE2.3GHZ, SYSTEM MEMORY: 4 GB)

Module	Processing time[msec]
Preprocessing	9
Recognition	120

IV. CONCLUSION

We have proposed a method for piano fingering recognition that uses depth images and does not require the use of any particular markers. It achieves high speed and high reliability through efficient screening using note name and wrist position information from the standard MIDI interface of an electronic piano keyboard. Experiments using short piano pieces have confirmed that our method achieves practical piano fingering recognition performance.

ACKNOWLEDGMENT

This work was partially supported by Grant-in-Aid for Scientific Research (B) 24300088.

REFERENCES

[1] E. Clarke, R. Parncutt , M. Raekallio and J. Slobaba, "Talk-Fingers: an interview study of pianists' views on fingering", Musicae Scientiae, Vol.1, No.1, pp.87-107, 1997.
[2] Yale Song, David Demirdjian and Randall Davis "Tracking Body and Hands For Gesture Recognition: NATOPS Aircraft Handling Signals Database", In Proceedings of the 9th IEEE International Conference on Automatic Face and Gesture Recognition (FG), pp500-506, 2011.
[3] D.O. Gorodnichy and A. Yogeswaran "Detection and tracking of pianist hands and fingers", In Proc. the Canadian conference Computer & Robot Vision, 2006.
[4] Noriko Nagata, Nozomi Kugimoto, Rui Miyazono, Kosuke Omori, Takeshi Fujimura, S.Furuya, Haruhisa Katayose, Hiroyoshi Miwa, "CG Animation for Piano Performance", In Proc. 15th Japan-Korea Joint Workshop on Frontiers of Computer Vision, pp.302-307, 2009.
[5] Takegawa, Terada, Nishio, "Design and Implementation of a Real-Time Fingering Detection System for Piano Performances", Proceeding of International Computer Music Conference, pp.67-74, 2006.
[6] J. S. Bach, "Notenbuchein fur Anna Magdalena Bach", G. Henle Verlag.