The Seventeenth International Symposium on Artificial Life and Robotics 2012 (AROB 17th '12),
B-Con Plaza, Beppu, Oita, Japan, January 19-21, 2012

Development of a singing robotic system from music scores in real time

Shou Imamoto[1], Yukiharu Yamauchi[1], Sei-ichiro Kamata[2]
Naoto Ohata[2], and Yusuke Murayama[2]

[1] Kitakyushu National College of Technology, Kitakyushu, Fukuoka 802-0985, Japan
(Tel: 81-93-964-7297, Fax: 81-93-964-7298)
(yamauchi@kct.ac.jp)
[2] Waseda University Graduate School, Kitakyushu, Fukuoka 808-0135, Japan

Abstract: In this paper, we propose a novel robotic entertainment system that can read a musical score and sing a song in real time. The objective of this system is to provide on-demand entertainment to onlookers. Previously, we developed a robotic system in collaboration with the Graduate School of Information, Production and Systems, Waseda University, and Shanghai Jiao Tong University. This system, named Ninomiya-kun, can read a book. It is equipped with a camera to read printed material placed on a book stand. In this study, we attempt to realize a robotic system that can sing a song by producing human-like sounds on the basis of notes and words extracted from a music score. Experimental results show that the proposed technique consistently outperforms well-established procedures.

Keywords: Robot vision, entertainment robots, music score recognition, singing robotic system

1 INTRODUCTION

Recently, robotic systems have been enhanced considerably, and they have become an integral part of the manufacturing industry. In addition, robots play an active role in medical care, living assistance, and security. The demand for robots for living assistance is expected to increase with the ongoing development of robotic systems.

Fig. 1. Robotic system Ninomiya-kun

Previously, we developed a robotic system in collaboration with the Graduate School of Information, Production and Systems, Waseda University and Shanghai Jiao Tong University [1, 2]. This system, named Ninomiya-kun (Fig. 1), can read a book. It is a child-sized humanoid that does not have legs; hence, it cannot move around. Further, it is equipped with two cameras for processing visual information and it can communicate with humans via sounds and gestures. It reads text by focusing its camera on printed material, such as a newspaper, magazine, or catalog, which is placed on a special book stand. Using a built-in computer with character recognition software, the robot translates the text into spoken words, which are produced by a voice synthesizer.

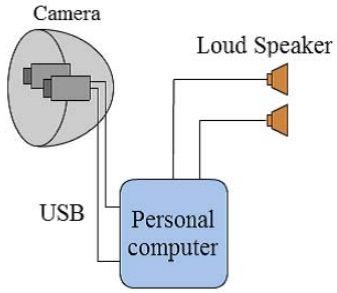


Fig. 2. System configuration of Ninomiya-kun

We have started enhancing Ninomiya-kun with new capabilities, and it must be trained to provide entertainment. It is important for an entertainment system to provide on-demand entertainment to onlookers. Our objective is to develop a robotic system that can read a musical score and sing a song in real time. We attempt to realize a system that can capture a printed music score using the camera installed in its the head, and sing a song by producing human-like sounds on the basis of the notes and words extracted from the captured music score[3, 4]. The input image should be properly processed for the successful recognition of the music score. In the first part of this paper, we discuss an image preprocessing technique for detecting musical notes, and we evaluate its performance. In the second part of this paper, we propose an adaptive technique for recognizing musical notes in order to deal with notes that belong to the same pitch class. Ex-

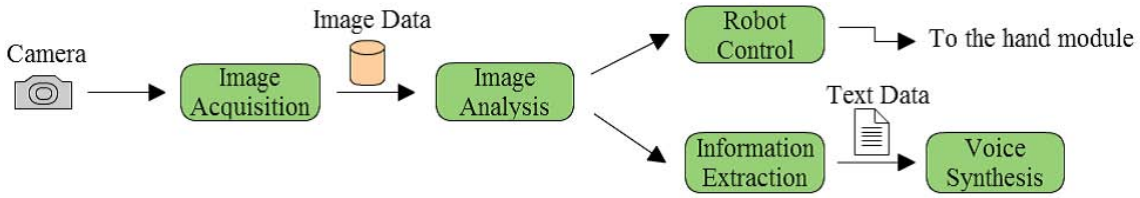


Fig. 3. Structure of the software system

perimental results indicate excellent music score recognition performance.

2 SYSTEM CONFIGURATION

The proposed robotic system has a backpack that consists of three units: a personal computer unit, an image acquisition unit, and a loudspeaker unit, i.e., an electro-acoustic transducer that produces sounds on the basis of the text that is translated into spoken words[11]. It has two hands that are driven by servomotors, enabling it to open a book and turn its pages; the rotational speed and angles of the hands are determined by the personal computer unit. Its head is equipped with two cameras as visual sensors. The video signals recorded by the cameras are transferred to the personal computer unit through the image acquisition unit, which provides high-resolution digitized images. The system configuration is shown in Fig. 2 [12, 13].

3 SOFTWARE CONFIGURATION

In this section, we describe an image processing technique for extracting notes and lyrics from printed music scores. A staff, or stave, is a set of five horizontal lines and four spaces, each of which represents a different musical pitch; it provides important information for musicians. As shown in Fig. 3, the software for the proposed robotic system consists of five modules: the image acquisition module, the image analysis module, the information extraction module, the voice synthesis module, and the robot control module. The image acquisition module captures a score using the camera and the image acquisition unit connected to the personal computer; then, this module performs image strain compensation and image binarization. Next, the image analysis module receives the binarized image and extracts the position of the staff. In the information extraction module, the pitch class of each note (C, C# , D etc.) is determined by the staff position information; in addition, it is determined whether the relative duration of each note is based on the ordering of sounds on a frequency-related scale. Simultaneously, the lyrics that are printed below the staff are extracted in this module. Accordingly, the voice synthesizer in the voice synthesis module produces human-like sounds; when required, requests for a rest period, i.e., is an interval of silence, are sent to the voice synthesis module by the information extraction module [5, 6].

4 EXTRACTION OF HEAD AND STEM

In an automatic music score recognition system, it is very important to extract the heads and stems of notes because these are the most ubiquitous and musically relevant symbols in a score [7, 8]. The objective of the system is to extract note heads quickly and accurately. It acquires the music score continuously in order to extract the heads and stems. In this study, we emphasize on the developmental aspect of the robotic system; hence, we employ a simple score with no chords. Initially, the camera is focused on the beginning of a score, and the score is continuously captured until a rest period is requested. Figure 4 shows the score that is used as a target. The image area that includes the candidates for heads and stems is extracted [9].

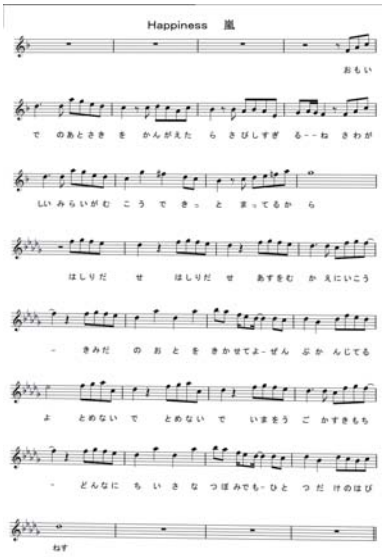


Fig. 4. Simple music score with no chords

The image area that includes the candidates for heads and stems is extracted. In order to extract the heads and stems, the stave is extracted by calculating the horizontal histogram of the music score, and the candidates for the areas with the lyrics are detected (Fig. 5). The black elliptical parts of the notes are the note heads; they indicate the note value (i.e.,

rhythmic duration), and they are extracted according to the relative position in the staff. The height of the head is equal to the interval between the lines in the staff, and the width of the head is 1.2 times its height. A template image is generated for heads whose height is less than the interval between the lines [10]. This template is compared with the overlapped image areas, while the detected stave is erased from the image (Fig. 6). The pitches of the notes are determined according to the position of the extracted heads. In the next step, stems, flags, and accidental notations are detected by template matching.

Fig. 5. Histogram of the music score

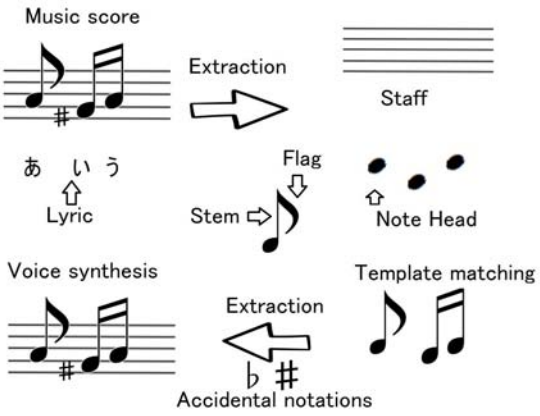


Fig. 6. Procedure for detecting notes

5 EXPERIMENTS

An experiment was conducted using the robotic system Ninomiya-kun. A captured music score is shown in Fig. 7. The robot extracted lyrics and music notes by focusing its camera on a printed music score placed on a special music stand. Fig. 8 (a) shows the recognition process; its side view is shown in Fig. 8 (b). The experiment was conducted as follows. Two music scores were used. One was "Kimigayo,"the national anthem of Japan; it is a short composition. The other was a 30-s extract from "Happiness," a song by the Japanese group Arashi. We used two scores in order to check whether the system is able to sing all the lyrics of different music scores. Figure 9 shows the staff extracted by calculating the histogram of "Kimigayo"; it was extracted with high accuracy. The accuracy of staff extraction from other music scores is more or less the same. The notes in the music scores "Kimigayo" and "Happiness" were detected by the template matching procedure, as shown in Fig. 10. The detected note heads are denoted by red boxes. However, the half notes were not detected, as shown in Fig. 10. A half note is a note having a white elliptical head. The template matching procedure fails to recognize a white note head that should be recognized because the pixel size of the elliptical part is different from that of the template image. We plan to devise ways to improve the accuracy off half notes detection.

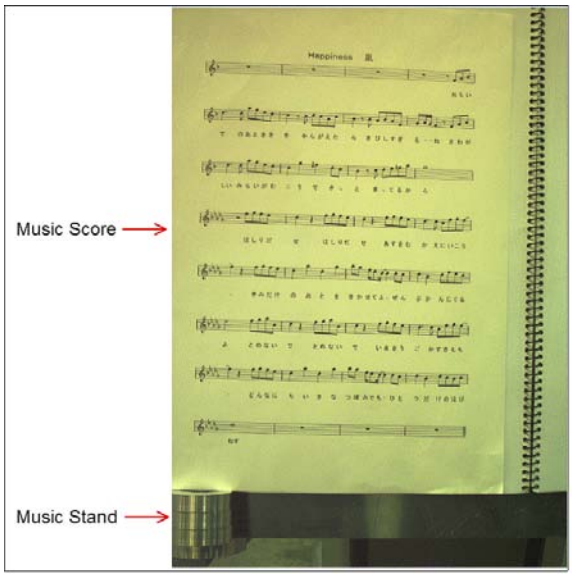
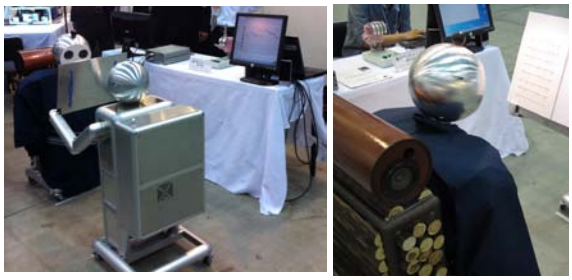


Fig. 7. Music score captured by Ninomiya-kun



(a) (b)

Fig. 8. Ninomiya-kun extracting lyrics and music notes

6 CONCLUSION

We proposed a novel robotic entertainment system that can extract notes from printed music scores using solely visual information. To realize the recognition of simple music scores, we devised a strategy based on image processing. This strategy was evaluated experimentally using the robotic system Ninomiya-kun, and satisfactory results were obtained. Further development of the software is required in order to realize the recognition of more complex music

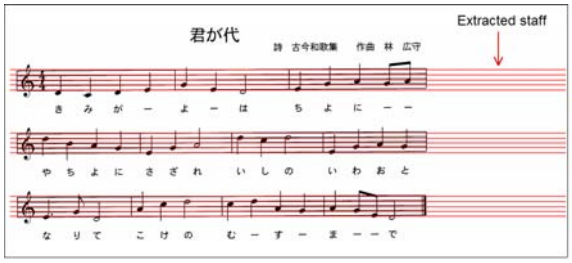


Fig. 9. Extracted staff in "Kimigayo"

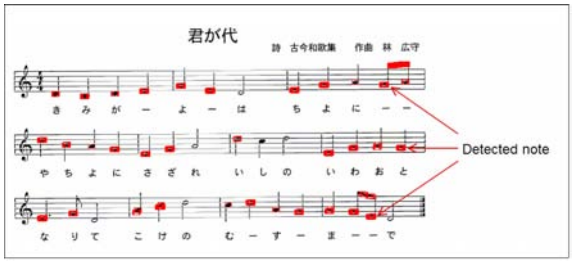


Fig. 10. Detected notes in "Kimigayo"

scores; investigation to this end is currently underway. In addition, a comprehensive evaluation of the recognition procedure remains to be established [14].

REFERENCES

[1] Y. Komatsubara, Y. Yamauchi, S. Kamata (2008), Recognizing and reading for characters on books (in Japanese). ITE Annual conference 2008

[2] Y. Komatsubara, Y. Yamauchi, S. Kamata (2009), Character segmentation and reading from Japanese character (in Japanese). ITE Annual conference 2009

[3] Jaime S. Cardoso, Artur Capela, Ana Rebelo, Carlos Guedes (2008), A connected path approach for staff detection on a music score. ICIP 2008:1005-1008

[4] Arshia Cont (2010), A coupled duration-focused architecture for real-time music-to-score alignment. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.32, No.6:974-987

[5] J. Keshet, S. Shalev-Shwartz, Y. Singer, and D. Chazan (2007), A large margin algorithm for speech-to-phoneme and music-to-Score alignment. IEEE Trans. Audio, Speech, and Language Processing, Vol.15, No.8:2373-2382

[6] M. Johnson (2005), Capacity and complexity of HMM duration modeling techniques. IEEE Signal Processing Letters, Vol.12, No.5:407-410

[7] L. Rabiner (1989), A tutorial on hidden Markov models and selected applications in speech recognition. Proc. IEEE, Vol.77, No. 2:257-285

[8] C. Raphael (1999), Automatic segmentation of acoustic musical signals using hidden Markov models. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol.21, No.4:360-370

[9] A. Cont (2006), Realtime audio to score alignment for polyphonic music instruments using sparse non-negative constraints and hierarchical HMMS. Proc. IEEE Acoustics, Speech, and Signal Processing

[10] C. Raphael (2006), Aligning music audio with symbolic scores using a hybrid graphical model. Machine Learning, vol. 65:389-409

[11] T. Kawabuchi, Y. Yamauchi, S. Kamata (2010), Technique for Reading documents with multiple better type (in Japanese), SICE Kyushu Annual conference 2010, 203A-4

[12] S. Imamoto, Y.Yamauchi (2011), Head and stem extraction from printed music scores, International Workshop on Target Recognition and Tracking 2011, pp.72

[13] N. Habu, Y.Yamauchi (2011), Printed Japanese characters recognition system, International Workshop on Target Recognition and Tracking 2011, pp.73

[14] T. Hakui, Y. Yamauchi (2011), Two-finger robotic hand for a touch panel system, International Workshop on Target Recognition and Tracking 2011, pp.74