Application of Image Recognition and Hand Tracking in Music Therapy

[1,*]*Hsiang-Jan Hung* (洪詳然), [1]*Firdaus Golam* (劉冠標), [1]*Chieh-Ming Yang* (楊哲旻),
[2]Chi-Hui Chen(陳綺慧), [1] *Jen-Yeu Chen* (陳震宇), and [3]Jui-Ling Hsiao(蕭瑞玲)

[1] Department of Electrical Engineering,
National Dong Hwa University, Hualien, Taiwan,
[2] Department of Music,
National Dong Hwa University, Hualien, Taiwan,
[3] Sunny Seeds ABA Learning Centre, Taipei, Taiwan
[*]E-mail: masonh3008@gmail.com

ABSTRACT

This project uses image recognition and hand tracking technology to develop a program that assists music therapists in conducting piano playing activities during music therapy. By recording the effectiveness of the activities, it can reduce the burden on therapists during the therapy process and maximizing the effectiveness of the therapy.

Piano playing activity is an activity in music therapy where the individual presses the corresponding piano keys to improve concentration, fine motor skills, and coordination, among other things. This project identifies whether the correct fingers are used and whether the piano keys are pressed in the correct order. When pressing incorrectly, the system will set a reminder for the therapists which reduce their burden, and record the student's playing situation for long-term effect evaluation at the same time.

Keywords: *Music therapy, object detection, object classification*

1. INTRODUCTION

Music therapy is a professional therapeutic method that uses music as a medium to improve physical and mental health and promote personal comprehensive development through the power of music. Combining theories and techniques from fields such as music, psychology, and medicine, music therapy uses rhythm, melody, sound, and expressiveness to influence an individual's emotions, cognition, behavior, and physical state. When designing music activities, music therapists choose appropriate music materials and activity methods based on the needs and goals of the individual. These activities may include listening to music, singing, playing instruments, composing music, musical games, and musical improvisation. By participating in music activities, individuals can express emotions, enhance self-awareness, increase creativity, improve language and motor coordination skills, promote social interaction and emotional regulation. Piano playing activity can improve concentration, fine motor skills, and coordination. The activity is suitable for individuals with ASD, Asperger's Syndrome, Cerebral palsy, ADHD, Down Syndrome, Developmental delay, emotional control, and other conditions. The individual presses the piano keys in sequence according to the color score provided by the therapist (as shown in Figure 1). By pasting five colors of stickers on the piano keys, the individual matches and presses the corresponding piano keys. The color used from thumb to the little finger are red, orange, yellow, green, and blue respectively (as shown in Figure 2).

In piano playing activities, therapists need to spend a lot of energy observing the physical and mental conditions of the individual and their fingers situation. In order to maximize therapeutic effects and provide long-term tracking, the motivation of this project is to develop a program to assist music therapists in recording the finger situation while pressing piano keys.

In this project, A self-trained YOLOv7 model was used to identify the positions of the piano keys on the screen. Then a self-trained CNN model was used to identify whether a key was pressed. At the same time, MediaPipe was used to identify the position of the fingertips to confirm whether the correct finger was used for pressing.

By recording the situation of fingers through the program, therapists can more accurately evaluate the progress and therapeutic effects of individuals. The program also provides real-time feedback and prompts to help therapists adjust the content and rhythm of music activities to more effectively achieve therapeutic goals.

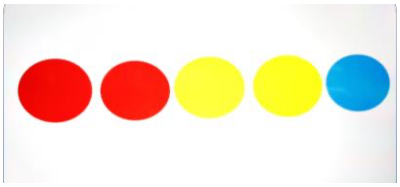

Fig. 1. Color score.

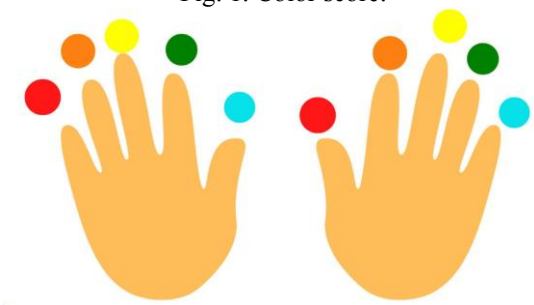

Fig. 2. Color for all fingers.

2. METHODS AND STEPS

The system flowchart of this project is shown in Figure 3. The following is a detailed description of the flowchart:

1. Input image: The system receives images of the music therapy scene as input.
2. YOLOv7[1] model: The system uses a self-trained YOLOv7 model to perform object detection on the image, mainly identifying the positions of the piano keys on the screen. These position coordinates will be stored for later use.
3. Piano key color detection: The system performs color detection on the identified piano keys, detecting red, orange, yellow, green, and blue. The five piano keys are continuously identified in an infinite loop.
4. CNN[2] model: The system uses a self-trained convolutional neural network (CNN) model to identify pressed key. Through this model, the system can determine whether each key is pressed.
5. MediaPipe[3] hand recognition: If a key is identified as being pressed, the system uses MediaPipe to identify the position of the fingertips to confirm whether the correct finger is used for pressing.
6. Status display: The system displays the keys that have been pressed in the lower left corner of the screen in numbers and marks them in blue font on the keys. At the same time, the system will display “Fingercorrect” in the upper left corner of the screen to indicate correct finger or “Wrong” to indicate incorrect finger.
7. Accuracy record: The system records the number of correct finger presses and the total number of presses to calculate accuracy. This data can be used for long-term observation of individual progress and therapeutic effects.

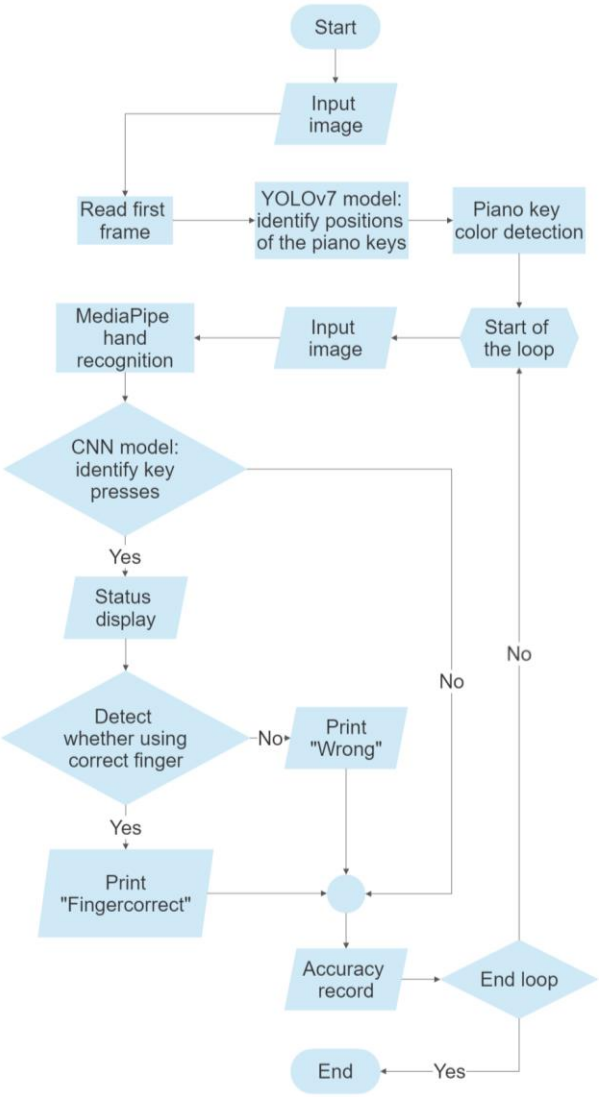

Fig. 3. System flowchart

2.1. YOLOv7

This project chose to use YOLOv7 to recognize piano key position coordinates instead of directly recognizing pressed piano keys because this project will be used for music therapy and requires very accurate results. Directly recognizing pressed piano keys has a greater possibility of misjudgment. Therefore, position coordinates and whether or not they are pressed are trained separately to ensure that the program judges accurately and reduces the possibility of misjudgment to a minimum. The result after training is shown in Figure 4. The final accuracy reached 99.42%, with its confusion matrix shown in Figure 5

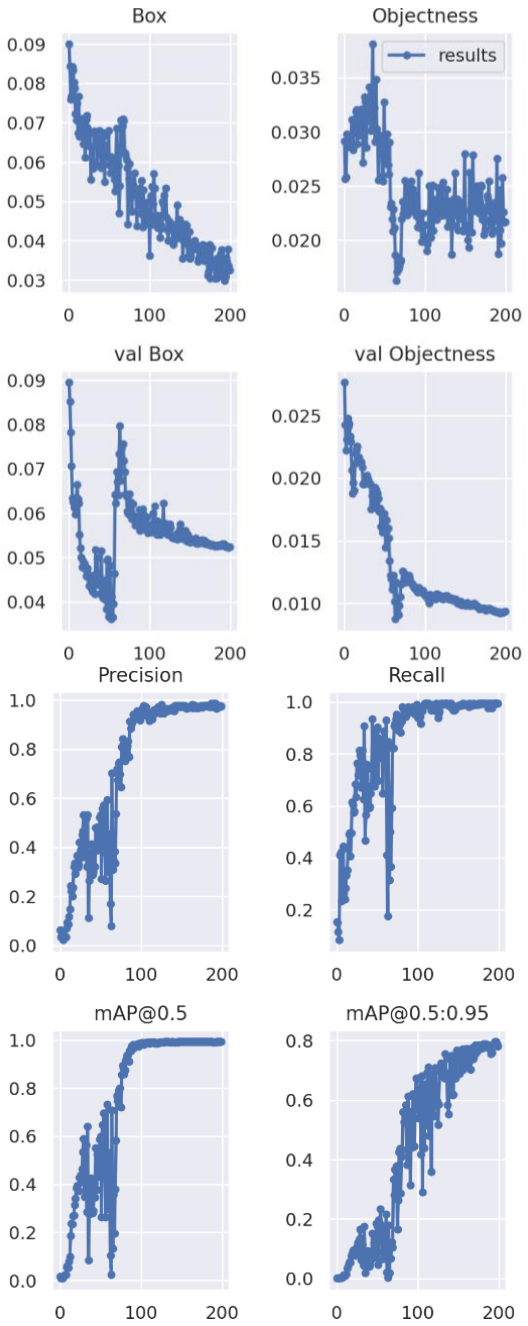

Fig. 4. Training result.

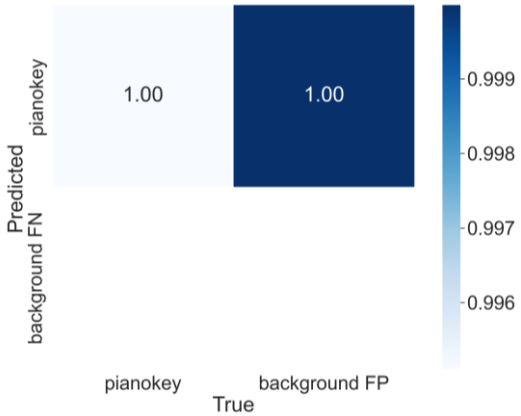

Fig. 5. Confusion matrix.

2.2. CNN (Convolutional Neural Network）

There are two reasons to use CNN model for identifying pressed key:

2.2.1. Spatial feature extraction

Whether or not a piano key is pressed is recognized through subtle changes in light and shadow. Therefore, identifying whether a piano key is pressed requires accurate extraction and classification of spatial features of piano keys. CNN has an advantage in processing image and visual data and can effectively capture local and global features in images. Therefore, it is more suitable for identifying spatial structures when piano keys are pressed.

2.2.2. Parameter sharing and weight sharing:

Identifying whether a piano key is pressed requires considering similarities between piano keys and shared features. Through parameter sharing and weight sharing design, CNN can effectively extract and utilize shared features between piano keys to improve model efficiency and accuracy.

Image dataset: 600 image data training sets are used in the training. Some examples of the dataset are shown in Figure 6 and the training model architecture is shown in Figure 7. For the training results, the accuracy is 99.99% as shown in Figure 8 and the loss value is 0.0009 as shown in Figure 9.

Fig. 6. Examples of the dataset.

Layer (type)	Output Shape	Param #
conv2d_2 (Conv2D)	(None, 224, 224, 16)	784
dropout_3 (Dropout)	(None, 224, 224, 16)	0
max_pooling2d_1 (MaxPooling2	(None, 112, 112, 16)	0
dropout_4 (Dropout)	(None, 112, 112, 16)	0
conv2d_3 (Conv2D)	(None, 112, 112, 32)	2080
dropout_5 (Dropout)	(None, 112, 112, 32)	0
flatten_1 (Flatten)	(None, 401408)	0
dense_2 (Dense)	(None, 64)	25690176
dense_3 (Dense)	(None, 2)	130

Fig. 7. Model architecture.

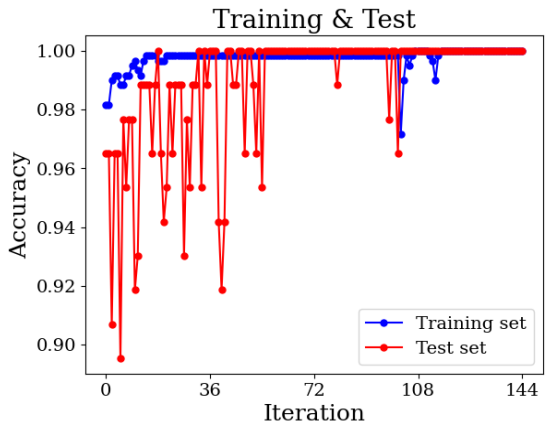

Fig. 8. Training result: accuracy.

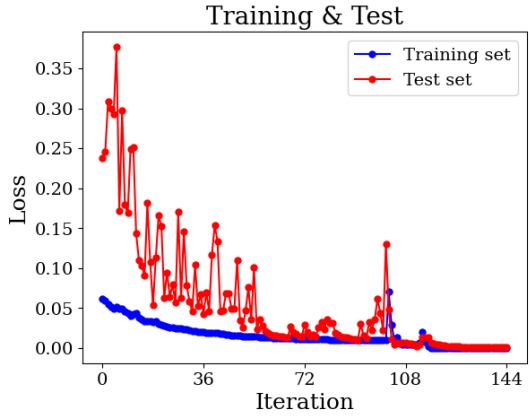

Fig. 9. Training result: loss.

3. RESULTS AND DISCUSSION

3.1. Results

The user sets up a camera to open the program. After pressing a piano key, the system displays key that have been pressed in the lower left corner of the screen using numbers 1~5 while displaying "Fingercorrect" indicating correct finger or "Wrong" indicating incorrect finger on the upper left corner of the screen.

Fig. 10. Results

3.2. Discussion

Music therapy is still in its infancy in Taiwan with few people applying digital technology to music therapy. Therefore, there are still too many unknowns about the effectiveness and and feasibility of combining digital technology with music therapy. Currently Ms. Chen Qihui is willing to apply this project's results to actual case treatment providing important practical application scenarios. If the system is actually helpful it may be considered to be ported into mobile platforms for convenience in the future.

3.3. Future Prospects

1. Train models that can recognize various angles and various pianos.
2. Train more accurate models with better responsiveness for recognizing whether or not they are pressed
3. Make mini games for individuals to practice at home
4. Develop piano teaching software for beginners learning piano

ACKNOWLEDGEMENT

This work was supported by the Ministry of Science and Technology, Taiwan, under Grant MOST 111-2622-8-259-001-TE2.

REFERENCES

[1] C.Y. Wang, A. Bochkovskiy, and H.Y.M. Liao, "YOLOv7: Trainable Bag-of-freebies Sets New State-of-the-art for Real-time Object Detectors," Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada, 2023.

[2] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based Learning Applied to Document Recognition," in Proceedings of the IEEE, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[3] F. Zhang, V. Bazarevsky, A. Vakunov, A. Tkachenka, G. Sung, C.L. Chang, and M. Grundmann, "MediaPipe Hands: On-device Real-time Hand Tracking," CVPR Workshop on Computer Vision for Augmented and Virtual Reality, 2020.