Application of Deep Learning in Music Therapy

Hsiang-Jan Hung[1], Firdaus Golam[1], Chieh-Ming Yang[1], Chi-Hui Chen[2], Jen-Yeu Chen[1], and Jui-Ling Hsiao[3], Han-Chieh Chao[4] and Wei-Che Chien[5]

[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
[4]Department of Artificial Intelligence, Tamkang University, New Taipei City, Taiwan
[5]Department of Computer Science and Information Engineering, National Dong Hwa University, Hualien, Taiwan
masonh3008@gmail.com

Abstract. This project utilizes image recognition and hand-tracking technology to create a program that aids music therapists in conducting piano-playing activities during music therapy. Recording the effectiveness of activities can reduce the burden on therapists during the therapy process and maximize the therapy's effectiveness. Piano playing is a music therapy activity where the individual presses the corresponding piano keys to enhance concentration, fine motor skills, and coordination, among other benefits. This project aims to determine if the correct fingers are used and if the piano keys are pressed in the correct order. When pressed incorrectly, the system will set a reminder for the therapists, reducing their burden, and record the student's playing situation for long-term effect evaluation simultaneously.

Keywords: Music Therapy, Deep Learning, Object Detection, Classification, Piano, MediaPipe.

1 Introduction

1.1 A Subsection Sample

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. Numerous research papers have highlighted the effectiveness of music therapy. A comprehensive research by H. Kamioka summarized evidence for the effectiveness of music therapy [1]. Furthermore, G. Bharathi *et al.* research on music therapy (MT) found that it had a positive impact on individuals with autism spectrum disorder (ASD) compared to a placebo treatment [2]. 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, attention deficit hyperactivity disorder, 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 Fig. 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 Fig. 2.

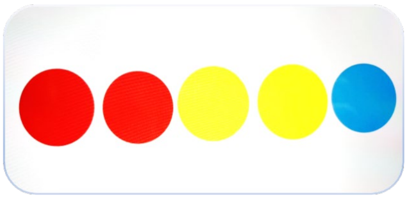


Fig. 1. Color score.

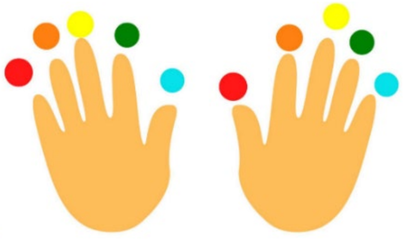


Fig. 2. Color for all fingers.

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.

There have been several studies in recent years about detecting pressed piano keys. For instance, A. OZKAYA and S. I. Tuncer [3] established a piano frequency detection system using the Goertzel algorithm, which uses the frequency of sound to detect the keys. Another study by J. Lee *et al.* [4] employed a two-stream convolutional neural network and multi-task learning to accurately detect piano keys, a novel and accurate approach. Another study by P. Suteparuk [5] he use only brightness to detect piano keys pressed in video. But it can’t use in real time camera which cannot be used in this situation.

In this project, A self-trained YOLOv7 [6] model was used to identify the positions of the piano keys on the screen. YOLOv7 is a deep learning-based object detection algorithm. Then, a self-trained convolutional neural network (CNN) [7] model was used to identify whether a key was pressed. CNN model is a type of deep learning algorithm designed for analyzing visual imagery. Additionally, MediaPipe [8], a cross-platform, customizable machine learning solutions framework, was used to identify the position of the fingertips. This step confirmed whether the correct finger was used for pressing the key.

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.

2 Methods

The system flowchart of this project is shown in Fig. 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 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 model: The system uses a self-trained CNN model to identify pressed key. Through this model, the system can determine whether each key is pressed.
5. MediaPipe 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.

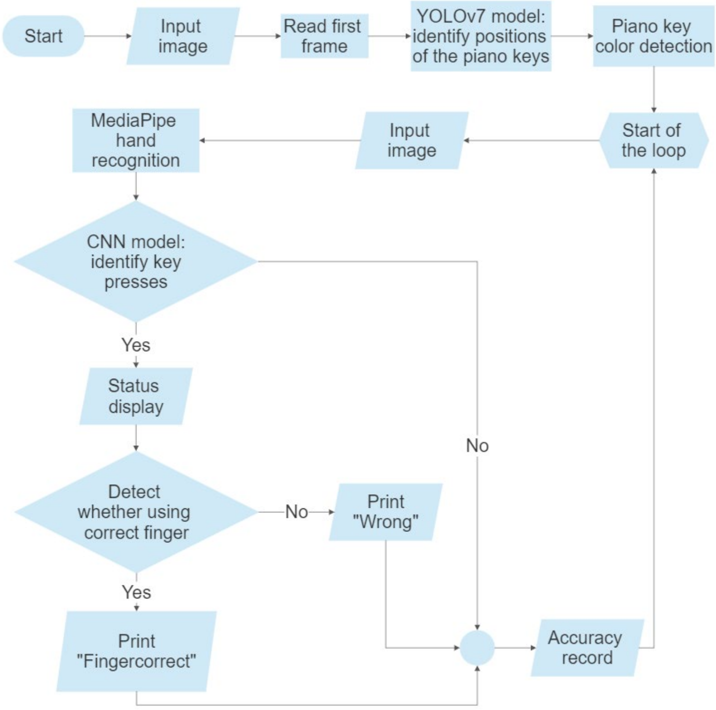

Start
Input image
Read first frame
YOLOv7 model: identify positions of the piano keys
Piano key color detection
Start of the loop
Input image
MediaPipe hand recognition
CNN model: identify key presses
Yes
No
Status display
Detect whether using correct finger
No
Print "Wrong"
Yes
Print "Fingercorrect"
Accuracy record

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.

To evaluate the performance of piano key defect detection, various metrics are used, including specificity, sensitivity, precision, and F_1 score. Specificity, sensitivity, and precision are defined in terms of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN) as shown in equations (1) to (3). The F_1 score is defined as the harmonic mean of precision and recall (sensitivity), calculated using Equation (4). The object detection model can be evaluated by incorporating the 101-point interpolation Average Precision (AP) metric, defined as calculated in equation (5), where $\rho(\tilde{r})$ is the measured precision at recall $\tilde{r}$.

$$Specificity = \frac{TN}{TN + FP}, \tag{1}$$

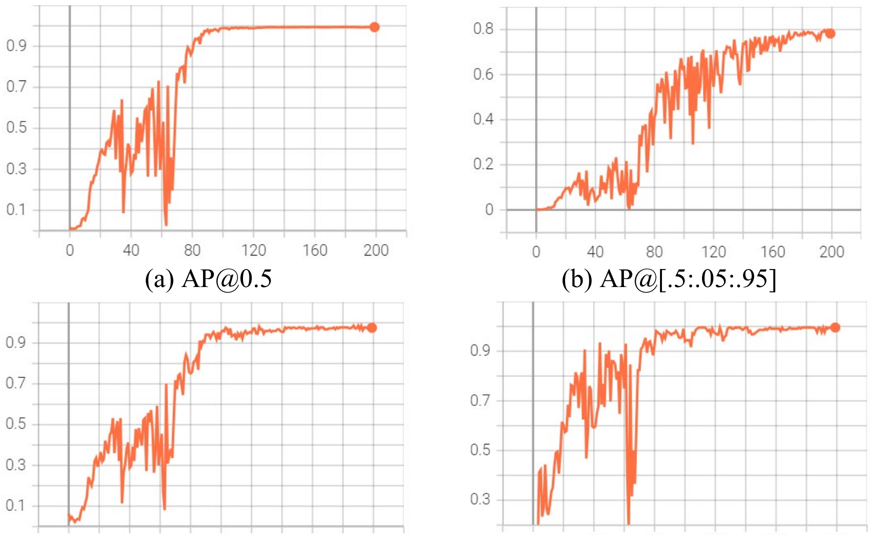
$$Sensitivity\ (Recall) = \frac{TP}{TP + FN}, \tag{2}$$

$$Precision\ = \frac{TP}{TP + FP}, \tag{3}$$

$$F_1\ score = \frac{2 \times\ Precision\ \times\ Recall}{Precision + Recall}, \tag{4}$$

$$AP\ = \frac{1}{101} \sum_{r \in \{0, 0.01, \ldots, 1\}} \max_{\tilde{r}: \tilde{r} \geq r} \rho(\tilde{r})\,. \tag{5}$$

The training results are illustrated in Fig. 4, which displays the learning curve. The final average precision (AP) reached 99.42%. Fig. 4 consists of four subplots: (a) AP@0.5, (b) AP@[.5:.05:.95], (c) precision, and (d) recall.



(a) AP@0.5

(b) AP@[.5:.05:.95]

(c) Precision

(d) Recall

Fig. 4. Learning curves for YOLOv7 model.

2.2 CNN

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

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.

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 Fig. 5 and the training model architecture is shown in Fig. 6. For the training process, Fig. 7 illustrates the Loss and Accuracy curves of the training set, while Fig. 8 depicts those of the testing set, both showing complete convergence. The accuracy of the testing set reaches 99.99%, with a loss of 0.0009. Accuracy is defined in terms of TP, TN, FP, and FN as shown in equations (6).

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}, \tag{6}$$

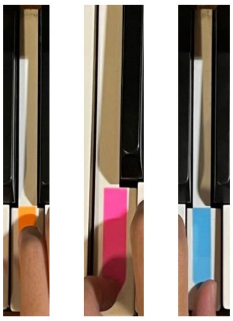


Fig. 5. Examples of the dataset.

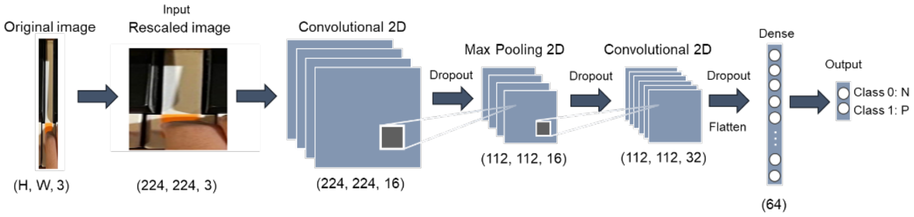


Fig. 6. CNN model architecture.

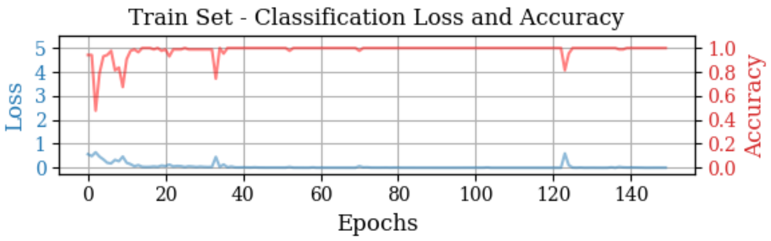


Fig. 7. Training set loss and accuracy curves for CNN.

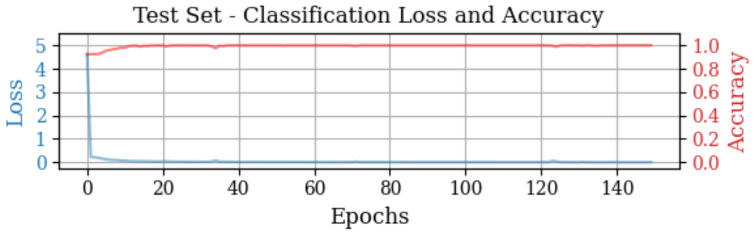


Fig. 8. Test set loss and accuracy curves for CNN.

3 Results Demo

The user sets up a camera to open the program. After pressing a piano key, the system displays the key that has been pressed in the lower left corner of the screen using numbers 1~5 while displaying "Fingercorrect" indicating the correct finger or "Wrong" indicating the incorrect finger on the upper left corner of the screen, as shown in Fig. 9.

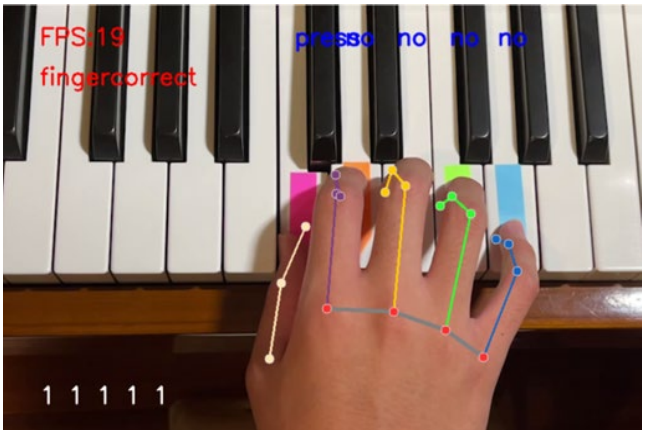


Fig. 9. Results demo.

4 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 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 helpful it may be considered to be ported into mobile platforms for convenience in the future.

5 Future work

In future research, we aim to explore several avenues. Firstly, we need to train models capable of recognizing various angles and piano configurations, enhancing their robustness to accurately identify keys from different perspectives. Secondly, refining the models to improve their accuracy and responsiveness in detecting key presses is crucial, particularly for real-time applications. Additionally, there is potential in creating tailored mini-games for individuals to practice piano skills at home, leveraging trained models to provide interactive platforms. Lastly, we see promise in developing beginner-specific piano teaching software, incorporating research insights to offer personalized learning experiences and assist novices in mastering fundamentals and techniques.

References

1. Kamioka, H., Tsutani, K., Yamada, M., Park, H., Okuizumi, H., Tsuruoka, K., Honda, T., Okada, S., Park, S. J., Kitayuguchi, J., Abe T., Handa, S., Oshio, T., Mutoh, T.: Effectiveness of music therapy: a summary of systematic reviews based on randomized controlled trials of music interventions. Patient Preference and Adherence, **8**, 727–754 (2014)
2. Bharathi, G., Venugopal, A., Vellingiri, B.: Music therapy as a therapeutic tool in improving the social skills of autistic children. The Egyptian Journal of Neurology, Psychiatry and Neurosurgery, **55**(44), (2019)
3. OZKAYA, A., Tuncer, S. I.: Pressed piano key detection and transcription by visual motion analysis. Politesi, (2020)
4. Lee, J., Doosti, B., Gu, Y., Cartledge, D., Crandall, D., Raphael, C.: Observing pianist accuracy and form with computer vision. In: IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA (2019)
5. Suteparuk, P.: Detection of piano keys pressed in video. Dept. of Comput. Sci., Stanford Univ., Stanford, CA, USA, Tech. Rep. (2014)
6. Wang, C. Y., Bochkovskiy, A., Liao, H. Y. M.: YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, Canada (2023)
7. Lecun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. In: IEEE **86**(11), 2278–2324 (1998)
8. Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.L., Grundmann, M.: MediaPipe hands: On-device real-time hand tracking. Conference on Computer

Vision and Pattern Recognition (CVPR) Workshop on Computer Vision for Augmented and Virtual Reality, (2020)