IIAI Open Conference Publication Series
IIAI Letters on Informatics and Interdisciplinary Research
Vol.005, LIIR263
DOI: https://doi.org/10.52731/liir.v005.263

Virtual Drum System Development using Motion Detection

Soonmin Hong [*], Stephen Karungaru [*], Kenji Terada [*]

Abstract

This paper proposes an innovative approach to overcome the spatial, cost, and mobility constraints of traditional drum playing. By combining the latest deep learning technologies, YOLOv8 and the Pose Landmark model, we have developed a virtual drum system that precisely tracks the user's movements in real-time, allowing drumming anywhere. This technology significantly enhances musical creativity and accessibility by providing an experience similar to actual drumming without requiring expensive drum equipment. Furthermore, this system minimizes geographical and economic constraints in music education and practice, offering educators and learners a flexible and effective way to learn music. This study explores the technical details of the virtual drum system and its potential impact on music education and performance, making a significant contribution to the future of digital music performance tools.

Keywords: Deep Learning, Motion Recognition, Virtual Drumming, Real-time Tracking

1 Introduction

Music, especially drums, which form the foundation of rhythm, have always been a vital element in evoking human emotions. However, drum sets require high costs and large spaces, causing significant difficulties for many music enthusiasts in enjoying drumming. Economic or spatial constraints are particularly problematic for young music enthusiasts. To address these issues, research into developing new virtual drum playing systems using modern technology is essential.

Systems such as Aerodrums[1] and Pocket Drum[2] have been developed to tackle these challenges. Aerodrums provide an experience similar to playing actual drums by detecting infrared light reflected from markers attached to sticks and feet using a camera. However, in bright environments, the infrared light is not well detected, reducing the system's accuracy. For example, using Aerodrums in outdoor performances or brightly lit studios can make accurate drumming difficult. Additionally, Pocket Drum detects real-time foot movements

[*] Tokushima University, Tokushima, Japan

using gyro sensor-equipped sticks and dedicated hardware, reproducing drum playing actions. However, the battery life of the sensors is an issue, as the system may stop during performances if the battery runs out. For instance, during long performance sessions or outdoor events, the inconvenience of needing to charge the equipment mid-performance can arise.

The rapid development of deep learning technology has significantly improved computer vision performance [3], leading to notable advances in image segmentation and object detection fields [4, 5]. Based on this, Tolentino et al. [6] developed a computer vision-based virtual drum system that detects and tracks drum playing actions using a laptop camera and color-based markers. In a similar manner, this study uses the YOLOv8 model to track the position of drumsticks in real-time. Additionally, Yadid et al. [7] proposed a system that uses a smartphone and webcam to detect and play the tips of drumsticks in real-time with YOLOv5. However, these approaches are still far from actual drum playing, as they only recognize sticks.

This study combines the latest deep learning technologies, YOLOv8[8], and MediaPipe's pose landmark model [9] to recognize and track stick and foot movements in real-time, developing a virtual drum system similar to actual drum playing. Unlike previous studies that mainly focused on recognizing and tracking drumsticks, this system simultaneously tracks the movements of sticks and feet, providing a more accurate and realistic drumming experience. This system can contribute to enhancing musical creativity and expressiveness by providing an accessible, flexible, and exciting learning tool without economic and spatial constraints. Additionally, it offers new educational methodologies to music educators, providing opportunities to enhance learner motivation and engagement.

2 Methodology

This system integrates two distinct systems and is divided into three stages: (1) stick Tip Detection and Tracking, (2) Body Movement Detection and Tracking, and (3) Drum Sound Synthesis.

2.1 Stick Tip Detection and Tracking

Detecting and tracking the stick tip is a core element of the virtual drum playing system, requiring high precision and real-time responsiveness. For this, we use an approach that combines Roboflow[10] and YOLOv8[8].

Data Preparation and Labeling:
First, we recorded drum playing at 60fps using the iPhone 12 Pro front camera. We then divided the recorded video into individual images at 60 frames per second using Roboflow [10]. Each stick tip was labeled in the saved images, as shown in Figure 1. To clearly distinguish the right stick tip, we attached red tape and labeled it. Through this process, we labeled a total of 9,500 images, and during dataset generation with Roboflow, we added post-processing such as flipping and adjusting brightness by ±15%, creating a dataset of 25,000 images in total.

YOLOv8 Model Training:
YOLOv8[8] is the latest deep learning algorithm for object detection, capable of high-speed processing to identify and classify objects within images. This model detects and tracks the

position of the player's stick tip in real-time. Figure 2 shows the performance metrics obtained during the training and validation process of the YOLOv8 model. During both training and validation, box loss (Box Loss), classification loss (Cls Loss), and direction field loss (Dfl Loss) steadily decreased, indicating improvements in the model's prediction accuracy and object direction prediction capabilities. Precision reached approximately 0.92, and recall reached 0.85, demonstrating high accuracy and effective object detection. Mean Average Precision (mAP) reached 0.9 at 50% and 0.6 from 50% to 95%, indicating strong overall performance.

Model Performance Evaluation:
Figure 3 shows the confusion matrix results of the YOLOv8n model. The model demonstrated a high recognition accuracy of 93% for the right stick but had a high rate of misclassifying the background as the left stick. This indicates that the model struggles to distinguish between the background and the stick, rather than an issue with the data itself. To address this problem, the following approaches can be considered:

- Data Augmentation: Enhance the dataset by adding images with various backgrounds and more stick images to improve the model's discrimination ability.
- Additional Markers: Attach blue tape to the left stick to improve recognition accuracy.
- Improving the Model: Enhance the training methods to improve the model's ability to accurately recognize the background.

These improvements can further enhance the model's performance, and future research will reflect these to develop a more sophisticated virtual drum system. The trained model demonstrated the ability to accurately classify and track the tips of both sticks using a Logitech C920 webcam at 1080p resolution and 30fps, as shown in Figure 4. This technology provides a foundation for capturing the player's natural movements accurately, offering an experience similar to actual drum playing in a virtual environment.

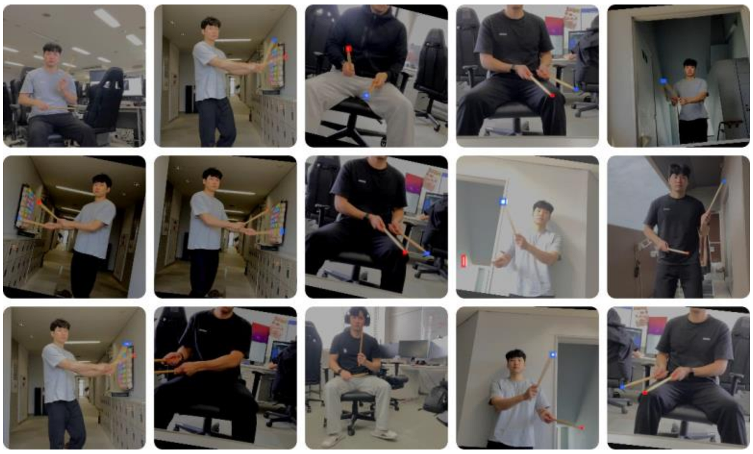

Figure 1: An example of bounding box work in Roboflow to train the position of stick tipsin various postures.

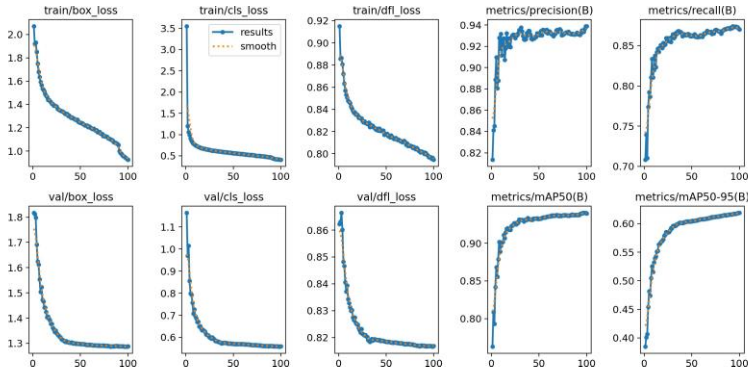

Figure 2: A graph showing the performance metrics obtained during the training and vali-dation process of the YOLOv8 model.

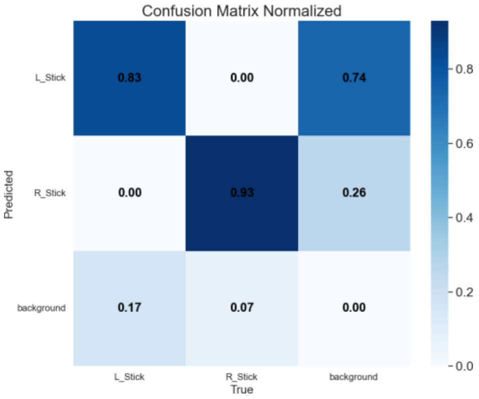

Figure 3: The normalized confusion matrix of the YOLOv8 model shows the model's performance in classifying the left stick (L-Stick), right stick (R-Stick), and background.

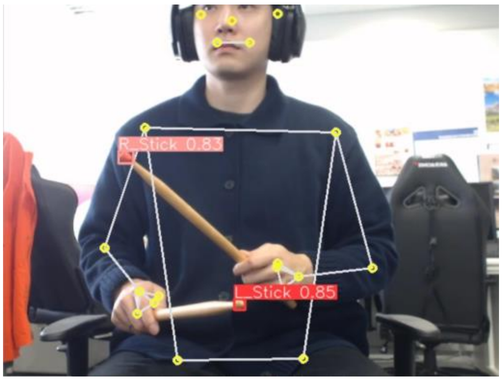

Figure 4: An example of real-time tracking of drumstick tips based on the dataset trained with YOLOv8. The scores at the top of the left (L-Stick) and right (R-Stick) drumsticks indicate the model's prediction probability.

2.2 Body Movement Detection and Tracking

Using a Logitech C920 webcam at 1080p resolution and 30fps, we demonstrated that it is possible to capture and analyze the user's natural drumming movements in real-time using MediaPipe's pose landmark model [9], as shown in Figure 5. This system precisely recognizes individual user postures and movements through MediaPipe's pose landmark model. The model uses machine learning algorithms to track the 3D coordinates of each part of the human body in real-time, allowing for accurate analysis of the user's overall body movements. Specifically, the system can detect various movements of the drummer, such as the movements of both arms and the complex movements of the front and heel of both feet. The system focuses on the user's right foot movements to enable the playing of the bass drum. Additionally, it tracks and analyzes the entire movement of the foot to accommodate various styles of pressing the drum pedal using the front and heel of the foot. This provides flexibility for users with different drumming styles to freely express their playing techniques. Furthermore, this system can be further developed to apply new technologies such as virtual reality (VR) [11] and augmented reality (AR) [12]. VR and AR allow users to play and experience music in a realistic and immersive environment. This enables users to have an experience similar to using an actual drum set, making it more effective for playing and creating music. The advancement of such technology can bring innovations in music education and production, and it is expected that more users will useit to enjoy and create music in the future. enjoy and create music.

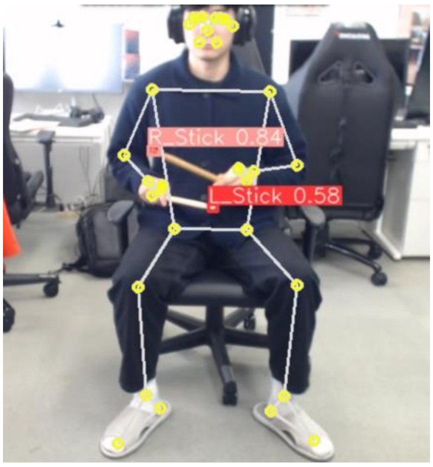

Figure 5: An example of real-time human posture tracking using the MediaPipe Pose Land-mark Model. The image shows the major landmark points of the entire body recognized bythe machine learning algorithm.

2.3 Drum Sound Synthesis

Using YOLOv8 Model

As shown in Figure 6, blue bounding boxes are created in positions similar to an actual drum, with the sounds of the hi-hat and snare assigned to each box. The YOLOv8 [8] model is used to track the exact position of the drumsticks in real-time. When the designated bounding box is hit, the color of the bounding box changes to red and the specified soundis played.

Using MediaPipe's Pose Landmark Model
The pose landmark model [9] is used to capture the entire body of the user in real-time. The position of the key points on the right foot is used to control various pedal techniques for bass drum sounds. This allows for the support of various drumming styles used by drum- mers on actual drums.

Combined System
This system combines YOLO [8] and the pose landmark model [9] to simultaneously detect the movements of the user's hands and feet and synthesize accurate drum sounds based onthese movements.

Minimizing sound delay is a critical task when creating a virtual drum system. The sound delay when hitting the drum should be very low so that users cannot perceive it. Generally, to ensure that users do not notice synchronization issues between audio and video, the delay should be within approximately 20-40 milliseconds [13]. Optimizing the delay within this range is essential for synchronizing the playing actions with the audio feedback. Therefore, the camera used should support a frame rate of at least 60 frames per second. A high frame rate improves motion detection accuracy and allows for more precise tracking of the player's movements. This technology provides users with an experience similar to using a physical drum set, significantly enhancing the practicality and immersion of digital drum systems. It can bring innovative changes not only in music playing but also in music education and creation.

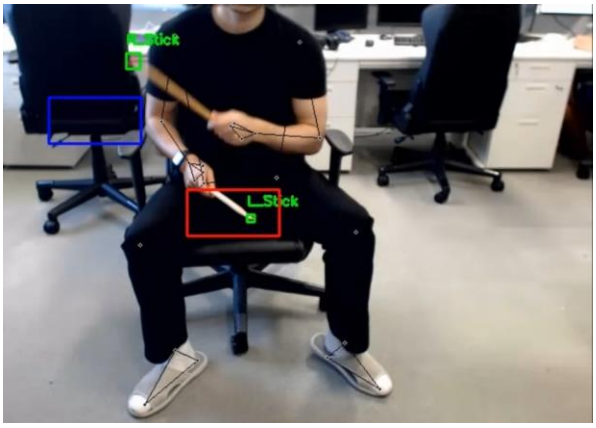

Figure 6: The image shows a system that tracks drumstick and body movements and generates drum sounds using the YOLOv8 model and MediaPipe's pose landmark model.

3 Conclusion and Future Research

The virtual drum playing system developed in this study serves as an important technological alternative that enhances accessibility and innovation in traditional drum playing through the use of motion recognition technology. By providing users with new musical experiences without the constraints of time and space, this system has the potential to bring innovative changes to the music industry and education.

Contributions of the Research:

- Technological Innovation: Proposes a motion recognition-based virtual drum system with high accuracy and real-time responsiveness by combining YOLOv8 and MediaPipe's pose landmark model.

- Economy and Accessibility: Provides an economical and portable system using common cameras and computer vision technology, making it accessible to more users.

- Offers new educational methodologies that enhance learner motivation and engagement, providing significant benefits to students who face economic and spatial constraints in accessing actual drum playing.

Impact of the Research:

- Enhancing Accessibility in Music Education: Can significantly enhance the accessibility of music education by providing an experience similar to actual drum playing without a physical drum set.

- Possibility of Real-time Collaboration: Can develop into an innovative platform that combines VR and AR technologies, allowing physically distant users to play and interact in a virtual space. This enables new forms of musical collaboration, such as remote ensemble performances, as well as advances in music education.

Future Research Directions: The current system's recognition rate is still in its early stages, primarily due to the lack of available YOLO datasets. Future research should focus on building extensive datasets, including many images of drumsticks and various backgrounds, to significantly improve recognition accuracy. This will enhance the user experience and increase the system's reliability.

Future research will include the following specific plans:

- Dataset Expansion: Expand the dataset and improve model recognition accuracy by collecting drumstick images under various backgrounds and lighting conditions.

- User Testing: Gather feedback from drummers and beginners to evaluate the system's accuracy and its differences from actual drumming, and use this feedback to improve the system.

- Application of VR and AR Technologies: Apply virtual reality (VR) and augmented reality (AR) technologies to provide users with an immersive playing experience and enable new forms of musical collaboration, such as remote ensemble performances.

- Evaluation of Educational Effectiveness: Assess how effectively the virtual drum system can be integrated into educational settings and music production studios.

These studies will help identify the practical advantages and limitations of the technology and ultimately optimize it to be widely used in music education and creation.

References

[1] Aerodrums, "Aerodrums: Air drums and virtual electronic drum kit," 2013. [Online]. Available: https://aerodrums.com/

[2] Aeroband, "Pocketdrum: Drum Anywhere, Anytime," 2016. [Online]. Available: https://www.aeroband.net/

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Commun. ACM, vol. 60, no. 6, pp. 84–90, 2017.

[4] J. Redmon and A. Angelova, "Real-time grasp detection using convolutional neural networks," in Proc. 2015 IEEE Int. Conf. Robot. Autom. (ICRA), 2015, pp. 1316–1322.

[5] M. Tan, R. Pang, and Q. V. Le, "Efficientdet: Scalable and efficient object detection," in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), 2020, pp. 10781–10790.

[6] C. T. Tolentino, A. Uy, and P. Naval, "Air drums: Playing drums using computer vision," in Proc. 2019 Int. Symp. Multimedia Commun. Technol. (ISMAC), 2019, pp. 1–6.

[7] H. Yadid, A. Algranti, M. Levin, and A. Taitler, "A2D: Anywhere Anytime Drumming," in Proc. 2023 IEEE Region 10 Symp. (TENSYMP), Canberra, Australia, 2023, pp. 1–6, doi: 10.1109/TENSYMP55890.2023.10223631.

[8] D. Reis, J. Kupec, J. Hong, and A. Daoudi, "Real-Time Flying Object Detection with YOLOv8," 2023. [Online]. Available: https://doi.org/10.48550/arXiv.2305.09972. arXiv:2305.09972 [cs.CV].

[9] I. Grishchenko et al., "BlazePose GHUM Holistic: Real-time 3D Human Landmarks and Pose Estimation," in Proc. CVPR Workshop Comput. Vis. Augmented Virtual Reality, New Orleans, LA, 2022, doi: 10.48550/arXiv.2206.11678. [Online]. Available: https://doi.org/10.48550/arXiv.2206.11678.

[10]F. Ciaglia et al., "Roboflow 100: A Rich, Multi-Domain Object Detec- tion Benchmark," arXiv:2211.13523 [cs.CV], Nov. 2022. [Online]. Available: https://doi.org/10.48550/arXiv.2211.13523.

[11] A. Bhardwaj, M. Bhardwaj, and A. Gaur, "Virtual Reality: An Overview," 2016.

[12]Y. Chen et al., "An overview of augmented reality technology," J. Phys. Conf. Ser., vol. 1237, no. 2, pp. 022082, 2019, doi: 10.1088/1742-6596/1237/2/022082.

[13]A. C. Younkin and P. J. Corriveau, "Determining the Amount of Audio-Video Synchronization Errors Perceptible to the Average End-User," IEEE Trans. Broadcast., vol. 54, no. 3, pp. 623–627, Sept. 2008, doi: 10.1109/TBC.2008.2002102.