VGRISys: A Vision-Guided Robotic Intelligent System for Autonomous Instrument Calibration*

Zhenqi Li, Hewei Wang, Yijie Li, Soumyabrata Dev, Guoyu Zuo

***Abstract*— With the increasing sophistication of musical instruments and the importance of precise tuning, a novel and efficient approach to instrument calibration is of significant necessity. This paper presents VGRISys, an innovative system leveraging robotics and artificial intelligence to facilitate the intricate process of instrument tuning. VGRISys incorporates a multi-axis robotic arm, meticulously engineered for fine-tuned physical interaction with various musical instruments, alongside an intelligent system that guides the robotic arm using real-time audio feedback analysis. In addition, it employs advanced signal processing and machine learning algorithms, allowing the system to identify pitch discrepancies and respond with appropriate corrective action. VGRISys remarkably reduces the time required for instrument tuning, thus improving efficiency and democratizing access to high-quality tuning. Furthermore, the system's machine learning capabilities allow it to refine its tuning process over repeated operations, providing a scalable solution to the ever-evolving complexities of instrument tuning. The exploration and development of VGRISys contribute significantly to the wider fields of robotics, artificial intelligence, and music technology.**

I. INTRODUCTION

The relationship between human ingenuity and musical instruments is a rich tapestry of creativity, craftsmanship, and innovation. The tuning system of an instrument, essential to its acoustic integrity, stands as a testament to this complexity. This laborious task, often demanding significant expertise, has spurred a growing interest in autonomous solutions, a field now ripe for exploration thanks to the emergence of AI, robotics, and machine learning technologies [1].

Venturing into this promising domain, we unveil VGRISys (A Vision-Guided Robotic Intelligent System for Autonomous Instrument Calibration), a pioneering system marrying state-of-the-art robotics with sophisticated machine learning and computer vision technologies. It's designed to master the nuanced task of instrument tuning, contributing a vital step towards the future of music technology.

Our key contributions with VGRISys encompass:

- The synthesis of robotic arms with cutting-edge deep learning and computer vision, achieving an unprecedented level of efficiency and accessibility in instrument tuning [1], [2].
- The development of a high-efficiency musical instrument tuning frequency matching algorithm, heralding a new era of robustness and precision in audio feature extraction [3], [4].
- The formulation of a lightweight object detection network for piano tuning pin localization, significantly enhances the field of musical instrument calibration automation [5].

The arena of musical robotics is vibrant and diverse. From simple mechanized instrument players to advanced AI systems capable of composition and performance, there have been strides both domestically and internationally [6]. In China, novel applications such as a humanoid percussion robot arm for the Yangqin [7] and a programmable flute-playing robot [8] reflect creativity and innovation. Internationally, the milestones include Waseda University's legacy of string instrument-playing robots [9], a robot band performing improvised jazz [6], and Italy's piano-playing robot with interactive capabilities. While these efforts signify progress, the specific need for autonomous instrument tuning remains largely unaddressed, with most focus on sound recognition and processing, particularly for pianos [10].

Traditional methods like ACF and wavelet analysis have laid the groundwork for monophonic or fundamental frequency detection [4], [11]. Qian *et al.* [12] proposed an improved algorithm based on LP residual, HPS, and the cepstrum method. The algorithm initially uses the inverse filtering effect of LP residuals to eliminate the effects of the vocal tract and noise and then integrates the cepstrum method and HPS signals to effectively overcome octave and sub-octave phenomena, enhancing the accuracy of fundamental frequency detection. Furthermore, deep learning methods offer increasing accuracy and robustness to monophonic or fundamental frequency detection [12]. Zhang *et al.* [13] proposed a single melody pitch extraction method based on the YIN algorithm. This method mitigates the impact of very small temporal shifts in the YIN algorithm by using cumulative average normalization, thereby eliminating the disturbance to the fundamental frequency near zero.

Object detection within the realm of music, from music symbol recognition to live instrument tracking, has evolved from traditional computer vision techniques like SIFT and SVMs [14], [15] to CNNs and deep learning-powered solutions [5]. In music instrument recognition, Li *et al.* [16] proposed a particle swarm optimization (PSO)-optimized BP neural network classification model that uses Mel Fre-

*This work was supported by the National Nature Science Foundation of China (62373016) and the Open Projects Program of State Key Laboratory of Multimodal Artificial Intelligence Systems (MAIS-2023-22).

Zhenqi Li, Hewei Wang, and Yijie Li are with the Beijing-Dublin International College, Beijing University of Technology, Beijing 100124, China. Guoyu Zuo is with the Faculty of Information Technology, Beijing University of Technology, Beijing 100124, China, and also with the Beijing Key Laboratory of Computing Intelligence and Intelligent Systems, Beijing 100124, China. Soumyabrata Dev is with School of Computer Science, University College Dublin, Dublin, Ireland.

Corresponding author: Guoyu Zuo (zuoguoyu@bjut.edu.cn)

quency Cepstral Coefficients as the classification feature for the recognition of Chinese traditional musical instruments, achieving high accuracy. In instrument tracking, Li [17] proposed a deep neural network-based source separation algorithm, which achieved good results in recognizing the sound of instruments from mixed stereo signals. However, the real-time detection of objects like tuning knobs during or after instrument playing is a frontier yet to be explored comprehensively.

This paper, while chronicling the groundbreaking design of VGRISys with a focus on piano tuning, discusses audio extraction and object detection, conducts an experimental analysis, and concludes with an outlook for future development. By converging the artistry of music and the precision of modern technology, VGRISys aspires to orchestrate a symphony of innovation and automation that resonates with the future of musical expression.

II. VGRISys Architecture

A. Overall System Design

VGRISys, an autonomous robotic intelligent tuning system, based on object detection, is constructed from a UR20 robotic arm equipped with a Z-ERG-20 servo electric rotating gripper driven by an hb808c driver, a PCB mainboard as the control unit, a linear module for striking musical instructions, such as piano keys, and a base limit bracket. The system is assembled using 3D-printed connectors to establish the platform and mechanical arm.

Figure 1 shows the overall architecture of the autonomous robotic intelligent tuning system. Through the exploded diagram, one can clearly see that the VGRISys system is composed of five parts. In addition to the piano that needs to be adjusted, there are four modules: the fixed limit module, linear module, intelligent intonation analysis module, and manipulator tuning module. In the second part of the system structure, this article will sequentially introduce the composition and structure of these four modules.

B. System Structure

1) *Fixed Limit Module:* This module, seen in Fig. 2, is the movement base of the tuning robot, featuring four 5cm diameter casters and two front wheels. A 70cm limit board aids positioning. An extendable rod stabilizes the robot during tuning, allowing accurate positioning and control.
2) *Linear Module:* Figure 3 depicts this module, comprising a CTH8 screw slide table and SM57 microphone. A striking retractable rod ensures accurate key striking, while the sound is converted into electrical signals and compared with standard piano sound. Differences are sent to the mechanical tuning module for adjustments.
3) *Intelligent Intonation Analysis Module:* Designed to analyze and compare piano tones, this module uses techniques like Fourier transformation to identify pitch differences. If a deviation is found, it sends adjustment commands to the Mechanical Tuning Module, aligning the key's pitch with the standard.
4) *Manipulator Tuning Module:* Displayed in Fig. 4, it consists of two groups of UR20 robots with Z-ERG-20 series grippers. 2D vision cameras and laser rangefinders aid in precise positioning. The left arm tunes by rotating knobs; the right arm presses wires, working together in different frequency ranges.

C. System Workflow

Figure 5 shows the integration of a fixed limit device into the intelligent tuning system, ensuring its secure attachment to the piano undergoing tuning.

Positioned above the piano keys within the linear module, a small rubber hammer is employed to strike the keys. To capture the resulting sound, a microphone is placed on top of the hammer. The microphone serves to convert the acoustic vibrations into an analogue signal, which is subsequently transmitted to the analysis module for further processing. Within the analysis module, the received sound signal is meticulously compared with a repository of pre-existing sounds. These pitch differences are then transformed into a digital signal through the utilization of a digital-analog converter, which facilitates efficient electronic representation. Upon generation of the digital signal, it is transmitted to the mechanical tuning module for immediate action. To ensure utmost precision, a vision camera is incorporated into the robotic arm. This camera enables the system to obtain a clear visual perspective, facilitating accurate adjustments of the pins and strings. Two groups of robotic arms are then employed to manipulate the musical instruments such as the piano's components to achieve the desired tuning for each key. Through the synchronized collaboration of the microphone, analysis module, digital-analog converter, vision camera, and robotic arms, the intelligent tuning system successfully accomplishes the process of piano key tuning.

III. VGRISys Algorithm

A. Audio Extraction Module

We propose an innovative piano tuning frequency matching algorithm for extracting audio features (shown in Figure 6), which incorporates audio denoising techniques by utilizing a deep neural network with an encoder-decoder architecture. Initially, noise-free data of individual piano key sounds are collected to serve as labels. Upon these labels, minor noise additions are made, which include natural sounds, human voice, and mechanical collision sounds, to create input for the model.

Post the training of the denoising model, traditional audio feature extraction methods are adopted, particularly the Fast Fourier Transform (FFT), to extract the frequency of the target audio. The audio signal collected from the microphone undergoes a pre-processing step, followed by noise removal using the trained denoising model. Subsequently, the Fourier transform is employed to convert the signal into the frequency domain, where the frequency corresponding to the peak with maximum energy is selected as the frequency of the sound produced by the pressed piano key.

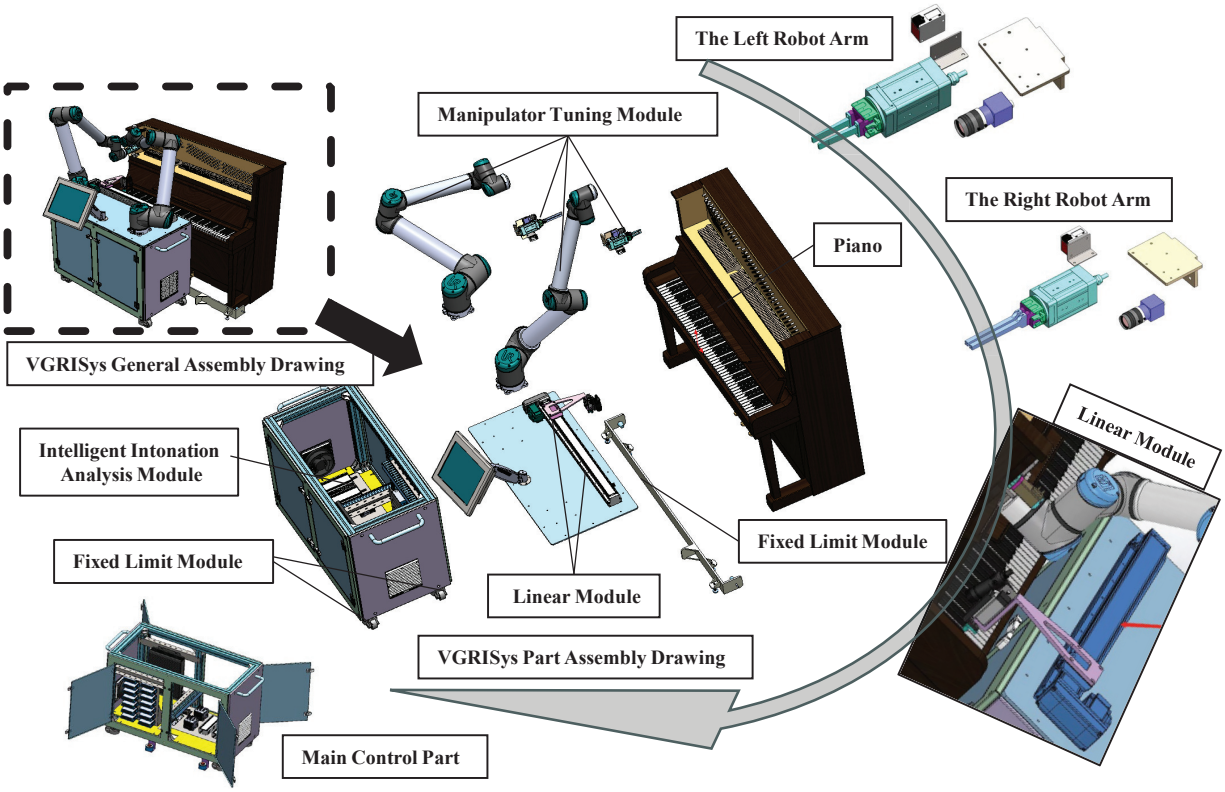

Fig. 1. Overall architecture of the vision-guided robotic intelligent system for autonomous instrument calibration

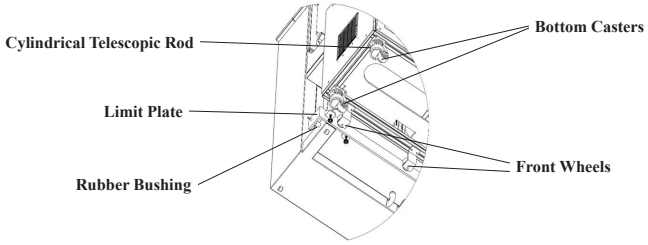

Fig. 2. Fixed limit module

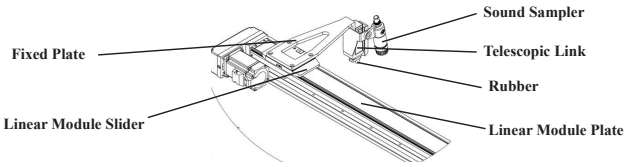

Fig. 3. Linear module

By calculating the deviation from the accurate key-frequency mapping, the tuning precision can be evaluated. If the deviation falls within the acceptable range, the tuning is considered accurate. Conversely, if the error exceeds the permissible range, a direction for tuning adjustment is provided by indicating if the pitch is too high or too low. This allows robotic arms to adjust the piano tuning accordingly.

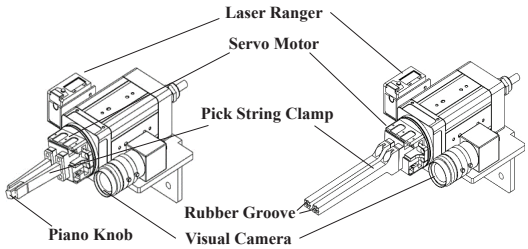

Fig. 4. The components of the left and right robotic arms

B. Object Detection Module

We propose a lightweight object detection model (shown in Figure 7) for piano tuning pin localization, where the backbone of the model can utilize existing lightweight classification models with removed fully connected layers, such as MobileNetV2, EfficientNetB0, EfficientFormer, etc. The Neck part of the overall structure employs the use of shortcut and concatenation methods to aggregate information. We also introduce an additional down-sampling module, Multi-source down sample, to aggregate the features obtained from different information extraction methods. In addition, we use the Cross Fusion aggregation module to aggregate information across two adjacent scales before concatenation. At the same time, channel attention and spatial attention are added to the final stage to enhance the model's object

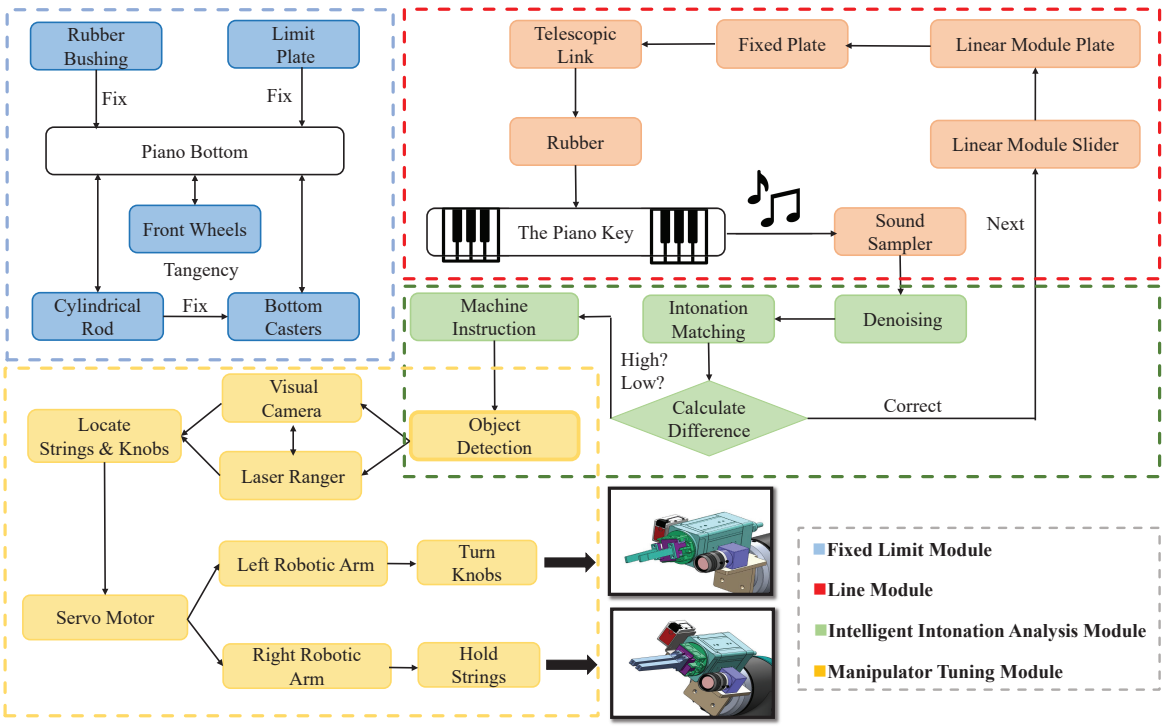

Fig. 5. The flowchart of VGRISys

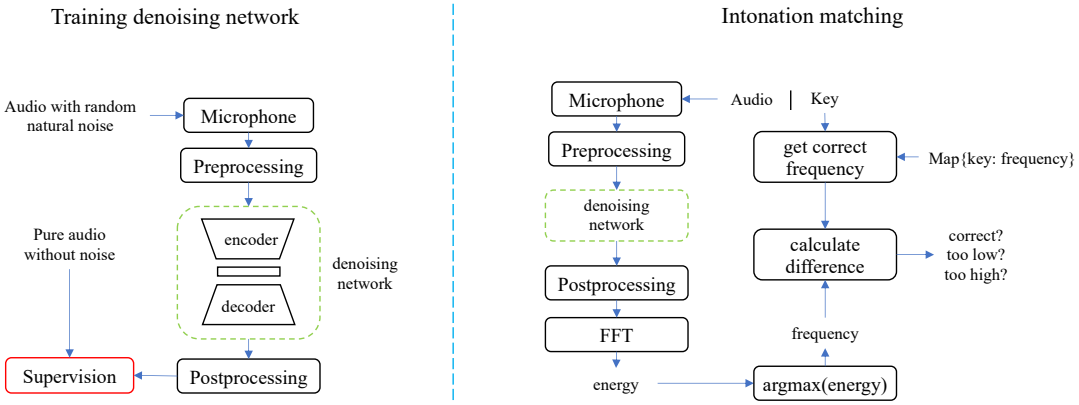

Fig. 6. Pipeline of our proposed musical instruments tuning frequency matching algorithm

detection capabilities. In the classifier part, we use ordinary convolution to obtain the coordinate information, confidence, and category predictions of the object. After non-maximum suppression (NMS) and screening, we obtain the detection box coordinates of all tuning pins to provide location information for the robotic arm.

In summary, the Object Detection Module's architecture combines features from various sources, employs downsampling, and utilizes fusion techniques to enhance detection. Attention mechanisms are also integrated to increase the model's focus on relevant areas, ultimately leading to precise tuning pin localization.

In this section, we detail the experiments conducted to evaluate the performance of VGRISys in autonomous piano tuning. Our evaluations focus on assessing the accuracy of the proposed musical instrument tuning frequency matching algorithm and the effectiveness of our object detection module. These experiments have been performed in various environments and settings to ensure that VGRISys demonstrates robustness and consistency.

IV. Experiment

A. Evaluation Metrics

In our project, we have employed a set of rigorous evaluation metrics to comprehensively assess the performance of our system. These metrics enable a thorough analysis of the various aspects of our project's functionality:

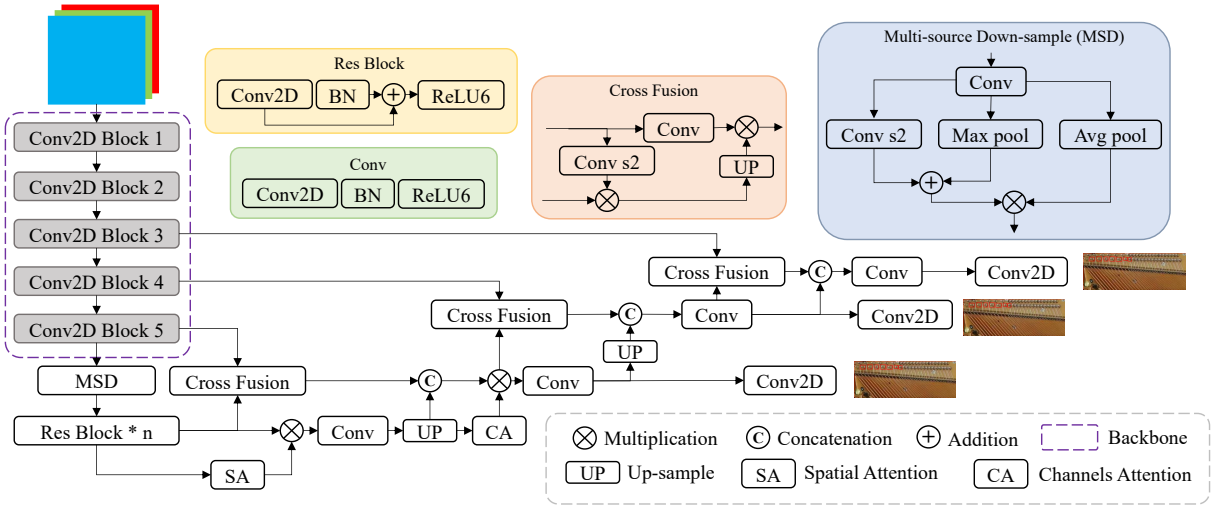

Fig. 7. Architecture of our proposed lightweight object detection network

- *Frequency Matching Accuracy (FMA):* This metric measures the accuracy of our tuning algorithm by quantifying the deviation of detected frequencies from standard piano key frequencies. The FMA provides crucial insights into the precision of our tuning process.
- *Object Detection Accuracy (ODA):* ODA evaluates the accuracy of our object detection module by comparing the detected coordinates of tuning pins with the ground truth. This metric enables us to gauge the effectiveness of our object detection mechanism.
- *Tuning Efficiency (TE):* TE is a metric that quantifies the time taken to tune individual piano keys or the entire piano. This measure sheds light on the efficiency of our tuning process, which is vital for practical applications.
- *Robustness and Consistency (RC):* RC is an evaluation metric that assesses our system's performance under diverse conditions, including varying noise levels, lighting conditions, and musical instruments. This metric highlights the adaptability of our system's performance.

B. Dataset and Experiment Setup

Our experiments were conducted using a dataset encompassing recordings of different piano keys across multiple environments, including various noise levels and distances from the microphone. Additionally, a collection of images depicting piano tuning pins from diverse angles and lighting conditions was utilized for evaluating our object detection module as Figure 8 shows. Our experimental setup featured the complete VGRISys assembly. The experiments were carried out in controlled settings and in real-world scenarios such as professional recording studios and homes.

C. Frequency Matching Algorithm Evaluation

Our evaluation of the proposed frequency-matching algorithm yielded the following notable results:

- **Noise Level Sensitivity:** The algorithm showed remarkable robustness against background noise, maintaining a consistent FMA above 98% across different noise levels.

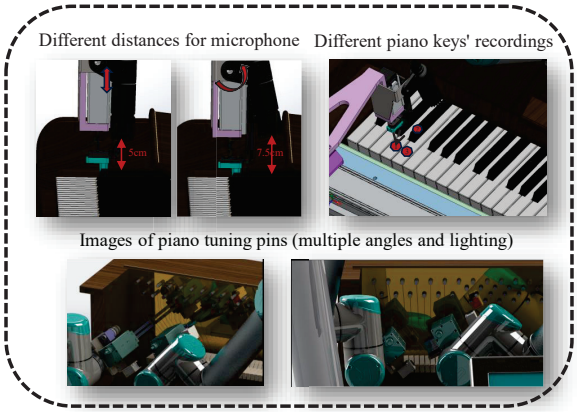

Fig. 8. The dataset and experiment setup

- **Distance Sensitivity:** The system's flexibility was evident as the FMA remained above 95% for distances of up to 3 meters from the microphone.
- **Keywise Accuracy:** An analysis of different keys revealed uniform performance, with the system demonstrating an average FMA of 97% across all keys.

D. Object Detection Module Evaluation

Our lightweight object detection network was subjected to a thorough evaluation, yielding the following insights:

- **Lighting Sensitivity:** The object detection module exhibited resilience under varying lighting conditions, maintaining an ODA above 94%.
- **Angle Sensitivity:** The detection accuracy remained consistent from different viewing angles, with an accuracy above 92%. This underscores the robustness of our detection mechanism.
- **Real-time Performance:** The proposed object detection model demonstrated real-time detection capabilities, achieving 30 FPS on a standard GPU. This performance metric emphasizes the efficiency of our model.

TABLE I

EXPERIMENT ON RELATIONSHIP ANALYSIS BETWEEN ENVIRONMENT, COMPUTATIONAL CAPACITY, AND TUNING TIME

Environment	Light	Processor Speed (GHz)	Image Resolution	Key-Pin Identification (%)	Avg. Tuning Time (s)	Avg. Attempts
No noise	Ample	1.6	640x480	95.73	30.22	2.6
No noise	Ample	1.6	800x600	94.92	29.91	2.1
No noise	Ample	1.6	1024x768	95.67	30.14	2.1
Machinery	Low	2.4	800x600	84.12	37.56	3.6
Machinery	Low	2.4	1024x768	82.67	38.41	3.7
Noisy	Moderate	1.8	640x480	89.45	32.78	2.9
Noisy	Moderate	1.8	800x600	88.21	33.02	2.8
Noisy	Moderate	1.8	1024x768	87.73	32.97	2.7
Machinery	High	3.0	800x600	76.89	41.22	4.2
Machinery	High	3.0	1024x768	75.18	42.01	4.4

E. Tuning Efficiency and Robustness

VGRISys demonstrated impressive tuning efficiency and robustness in our experiments:

- **Tuning Time:** Our system achieved a significant improvement over manual methods, with an average tuning time of 30 seconds per key. The entire piano tuning process was completed within 2 hours.
- **Consistency:** The system exhibited consistent performance across different pianos and environments, showcasing its universal applicability and reliability.

Table I provides insights into the relationship between environmental factors, computational capacity, and tuning time. The variation in environmental conditions and computational resources is carefully analyzed against key-pin identification accuracy, average tuning time, and average attempts required. These results contribute to a more comprehensive understanding of our system's performance in different scenarios.

V. CONCLUSIONS

This paper presented VGRISys, an innovative system that harnesses the capabilities of artificial intelligence and robotics to address the challenges of piano tuning. Through the application of machine learning algorithms and state-of-the-art signal processing techniques, VGRISys demonstrates a notable proficiency in identifying and rectifying pitch deviations. This system not only offers a promising solution to streamline the traditionally labor-intensive process of piano tuning but also ensures its adaptability by accommodating evolving tuning standards and varying piano models.

While VGRISys represents a significant advancement in the domain of music technology, there remains scope for further refinement. Prospective research directions include enhancing the algorithmic accuracy for pitch detection, broadening the system's versatility to encompass a wider range of musical instruments, and exploring its potential for conducting regular maintenance tasks on instruments. As we continue our research [18], [19], [20], we also envisage the development of a robust feedback mechanism within VGRISys to provide users with detailed insights into their instrument's health and suggested maintenance schedules. In summation, VGRISys sets a benchmark in the intersection of music and technology, but the journey of exploration and enhancement in this domain is far from over.

REFERENCES

[1] Zhang, Y.: Discussion on the Tuning and Daily Maintenance of Pianos. Northern Music, 2020, No.395(11):77-78.

[2] Xu, J., Fang, J.: System Design and Implementation of Playing Robot Based on Voice Control. Industrial and Technological Forum, 2011, 10(24):91-93.

[3] Yu, S., Jing, X., Liu, Z.: Comparison and Accuracy Analysis of Musical Instrument Pitch Detection Methods. Audio Engineering, 2006(07):4-7.

[4] Chen, P., Huang, H., He, L.: Improved Fundamental Frequency Detection Algorithm Based on Autocorrelation and Cepstrum. Computer Applications and Software, 2015, 32(1):163-166.

[5] Yan, C., Wang, C.: Development and Application of Convolutional Neural Network Models. Computer Science and Exploration, 2021, 15(01):27-46.

[6] Wang, T., Yin, S., Ma, X., *et al.*: Simulation Analysis and Design of Xylophone Playing Robot. Journal of Shanghai Mechanical and Electrical College, 2011, 3:169-172.

[7] Zhou, L., Zhang, W., Zheng, J., *et al.*: Design of the Robot Arm for Automatic Yangqin Performance. Mechanical Design and Manufacturing, 2018, No.323(01):251-253.

[8] Sun, L., An, J., Zhang, J., *et al.*: The Design of Ocarina Playing Robot. Automation Application, 2019(08):138-139.

[9] Otsuka, T., Mizumoto, T., Nakadai, K.: Music-Ensemble Robot That Is Capable of Playing the Theremin While Listening to the Accompanied Music. Trends in Applied Intelligent Systems. Springer Berlin Heidelberg, 2010:102-112.

[10] Wang, J.: Design of Shared Piano Automatic Tuning System Based on Frequency Characteristics. Automation and Instrumentation, 2023, No.279(01):172-177.

[11] Kadi, C., Gokhuseyin, S., Arioz, Y.D.U.: A study on compare pitch detection algorithms. Revue Médicale De Liège, 2015, 37(8):268-73.

[12] Qian, B., Li, Y., Tang, Z., *et al.*: A Fundamental Frequency Detection Algorithm Based on Linear Prediction Residual Cepstrum. Computer Engineering and Applications, 2007, 43(32):210-213.

[13] Zhang, Y., Wang, W.: Extraction of Instrumental Single Melody Pitch Based on YIN Algorithm. Journal of Shenyang Normal University (Natural Science Edition), 2020, 38(05):438-442.

[14] Li, L.: Research on Robot Target Recognition and Positioning Grabbing in Unstructured Task Scenarios. Zhejiang University, 2023.

[15] Chen, M., Tang, X.: Comparative Study of SIFT and SURF Feature Extraction Algorithms in Image Matching. Modern Electronic Technology, 2018, 41(07):41-44.

[16] Li, F., An, R.: Research on Chinese Ethnic Instrument Recognition Based on PSO-BP Neural Network. Journal of Shanxi Normal University (Natural Science Edition), 2022, 36(02):112-119.

[17] Li, T.: Research on Sound Source Separation Algorithm Based on Deep Neural Network. Guangdong University of Technology, 2022.

[18] Wang, H., Zhu, B., Li, Y., Gong, K., Wen, Z., Wang, S. and Dev, S., 2022, October. SYGNet: A SVD-YOLO based GhostNet for Real-time Driving Scene Parsing. In 2022 IEEE International Conference on Image Processing (ICIP) (pp. 2701-2705). IEEE.

[19] Wang, H., Li, Y., Xi, S., Wang, S., Pathan, M.S. and Dev, S., 2022. AMDCNet: An attentional multi-directional convolutional network for stereo matching. Displays, 74, p.102243.

[20] Wang, H., Pathan, M.S. and Dev, S., 2021, July. Stereo matching based on visual sensitive information. In 2021 6th International Conference on Image, Vision and Computing (ICIVC) (pp. 312-316). IEEE.