Turkish Journal of Computer and Mathematics Education (TURCOMAT) Vol.11 No.1(2020),1717-1724

Research Article

A Robust Music Note Recognition System Using Convolutional Neural Network

Mr. Singaraiah[1], Dr. Rakesh Mutukuru[2]

[1]Research Scholar, Department of Electronics and Communication Engineering, Shri Venkateshwara University, Gajraula, Uttar Pradesh, India
[2]Research Supervisor & Professor, Department of Electronics and Communication Engineering, Shri Venkateshwara University, Gajraula, Uttar Pradesh, India

Abstract: The task of automatically recognizing musical instruments poses significant challenges within the domain of music information retrieval. Learning to play the piano, on the other hand, demands expert instruction and substantial practice. Due to the hectic nature of modern life, many individuals find it difficult to commit to systematic training. Additionally, the scarcity of qualified piano teachers and the high costs associated with lessons further discourage potential students. If a computer could recognize and assess a learner's piano performance in real time, it would enable learners to identify and correct their mistakes promptly. Although there are existing music recognition technologies, most suffer from several limitations. Currently, music processing systems that incorporate models for chord progressions achieve high accuracy in tasks such as music structure analysis, multi pitch analysis, and automatic composition or accompaniment. pitch patterns are treated as observations derived from the hidden states within the chord progression model. Convolutional Neural Networks (CNN) have been successfully applied to chord recognition. The CNN approch will give high accuracy, precision and F1-Score.

Keywords: Automatically Recognizing Musical Instruments, Chord Progressions, Convolutional Neural Networks (CNNs).

I. INTRODUCTION

Music significantly enhances individuals' lives by serving as a key source of entertainment for experts and listeners, and at times offering good benefits. With advancements in multimedia and technology, various music players have emerged featuring capabilities such as fast-forward, reverse, variable playback speeds, local and streaming playback, and volume modulation, among others. Although these features meet basic user needs, they still require users to manually browse through playlists and select songs that match their current mood and behavior. This manual process can be labor-intensive and may result in difficulty finding an appropriate playlist.

In recent years, the field of deep learning has made substantial strides, particularly in music classification. Convolutional Neural Networks (CNNs), a prevalent deep learning architecture, have proven highly effective for music classification, recognition and segmentation. These networks imitate the human brain with artificial neurons organized in multiple hierarchical layers. By processing inputs through convolutions, applying weights, biases and activation functions, CNNs can perform complex music analyses. More data can improve the accuracy of these deep learning models.

A Music Recognition System utilizing Convolutional Neural Networks (CNNs) represents a cutting-edge approach to enhancing user experience by automating the selection of music based on emotional state. Leveraging advancements in deep learning, this system employs CNNs to accurately analyze facial expressions and identify emotions such as happiness, sadness, or anger. By integrating this technology, the system can dynamically curate and play music that aligns with the user's current mood, eliminating the need for manual playlist browsing and creating a more personalized and seamless listening experience.

II. LITERATURE SURVEY

R. Su, L. Wang and X. Liu et. al. Recently, various audio-visual speech recognition (AVSR) systems have been developed by using multimodal learning techniques. One key issue is that most of them are based on 2D audio-visual (AV) corpus with the lower video sampling rate. To address this issue, a 3D AV data set with the higher video sampling rate (up to 100 Hz) is introduced to be used in this paper. Another issue is the requirement of both auditory and visual modalities during the system testing. To address this issue, a visual feature generation based bimodal convolutional neural network (CNN) framework is proposed to build an AVSR system with wider application. In this framework, long short-term memory recurrent neural network (LSTM-RNN) is used to generate the visual modality from the auditory modality, while CNNs are used to integrate these two modalities. On a Mandarin Chinese far-field speech recognition task, when visual modality is provided, significant average character error rate (CER) reduction of about

27% relative was obtained over the audio-only CNN baseline. When visual modality is not available, the proposed AVSR system using the visual feature generation technique outperformed the audio-only CNN baseline by 18.52% relative CER [1].

J. Calvo-Zaragoza, A. -J. Gallego and A. Pertusa et. al. There are large collections of music manuscripts preserved over the centuries. In order to analyze these documents it is necessary to transcribe them into a machine-readable format. This process can be done automatically using Optical Music Recognition (OMR) systems, which typically consider segmentation plus classification workflows. This work is focused on the latter stage, presenting a comprehensive study for classification of handwritten musical symbols using Convolutional Neural Networks (CNN). The power of these models lies in their ability to transform the input into a meaningful representation for the task at hand, and that is why we study the use of these models to extract features (Neural Codes) for other classifiers. For the evaluation we consider four datasets containing different configurations and notation styles, along with a number of network models, different image preprocessing techniques and several supervised learning classifiers. Our results show that a remarkable accuracy can be achieved using the proposed framework, which significantly outperforms the state of the art in all datasets considered.

K. -Y. Choi, B. Coüasnon, Y. Ricquebourg and R. Zanibbi et. al. State-of-the-art Optical Music Recognition system often fails to process dense and damaged music scores, where many symbols can present complex segmentation problems. We propose to resolve these segmentation problems by using a CNN-based detector trained with few manually annotated data. A data augmentation bootstrapping method is used to accurately train a deep learning model to do the localization and classification of an accidental symbol associated with a note head, or the note head if there is no accidental. Using 5-fold cross-validation, we obtain an average of 98.5% localization with an IoU score over 0.5 and a classification accuracy of 99.2% [3].

Y. Liu and Y. Chen et. al. Automatical recognition of facial expression is an interesting and challenging problem, which has so many applications such as expression synthesis, human-robot interaction, metal state identification, intelligent tutoring systems, operator fatigue, music for mood, and clinical medicine. The vital step of a successful approach is deriving features from raw facial image. The existed methods of features extraction are the hand-crafted features based on geometric features or appearance features, and the auto-learned features. To utilize the benefit of low computation of hand-crafted features and the high-representation of auto-learned features, we firstly proposed the combined features CNN-CBP with putting together Centralized Binary Patterns (CBP) features and Convolutional Neural Network (CNN) features. And then, we classified the features using Support Vector Machine (SVM). With the help of the CNN-CBP features, we achieved average recognition accuracy of 97.6% on the Extended Cohn-Kanade datasets and 88.7% on the Japanese Femal Facial Expression datasets based on 10-cross validation [4].

S. Deebika, K. A. Indira and Jesline et. al. The paper constitutes the implementation of Convolutional neural network for the emotion detection and thereby playing a song accordingly. Segregating the songs and playing them in accordance to one's mood could facilitate the music lover. Although there exist a lot of algorithms designed for it, the computation is not as expected. This paper eradicates such an issue by using CNN. In order to obtain minimal processing, multilayer perceptron are implemented by CNNs. In comparison to various algorithms for image classification, CNNs observed to have little-processing. This implies that the filters used in CNNs are advantageous when compared to traditional algorithm. The visualization of features directly can be less informative. Hence, we use the training procedure of back-propagation to activate the filters for better visualization. The multiple actions such as capturing, detecting the emotion and classifying the same can all be confined as one step through the use of CNN. The slow performances of the real-time approaches could be enhanced by regularizing the methods and by visualizing the hidden features. Hence the proposed approach could enhance the accuracy and the computation speed [5].

III. METHODOLOGY

Music recognition systems, such as those used in Music Information Retrieval (MIR) and Automatic Music Transcription (AMT), involve several stages of pre-processing to prepare the audio data for further analysis. These stages aim to convert raw audio signals into a more manageable and informative representation. Here's an overview of typical pre-processing steps:

Audio Acquisition: Convert the audio signal to a standard sampling rate (e.g., 44.1 kHz or 16 kHz) if it's not already in this format. Mono Conversion: If the audio is in stereo, it is often converted to mono by averaging the two channels. This simplifies the analysis and reduces computational load.

Noise Reduction: Detect and remove sections of the audio that contain only silence or background noise to focus on the parts with musical content. Denoising: Apply noise reduction techniques such as spectral subtraction or Wiener filtering to enhance the quality of the audio signal.

Normalization: Adjust the volume of the audio to a standard level to ensure consistent amplitude across different recordings. Peak Normalization: Scale the audio signal so that the highest amplitude is at a specific level, usually just below the maximum possible value.

Segmentation: Frame Blocking: Divide the audio signal into short, overlapping frames (e.g., 20-50 ms with 50% overlap) to facilitate short-time analysis. Windowing: Apply a window function (e.g., Hamming, Hanning, or Blackman) to each frame to reduce spectral leakage during Fourier Transform.

Feature Extraction: Compute the Short-Time Fourier Transform (STFT) to convert the time-domain signal into a time-frequency representation. Spectrogram:

Create a spectrogram by plotting the magnitude of the STFT over time, which visually represents the energy distribution across frequencies. Mel-Frequency Cepstral Coefficients (MFCCs):

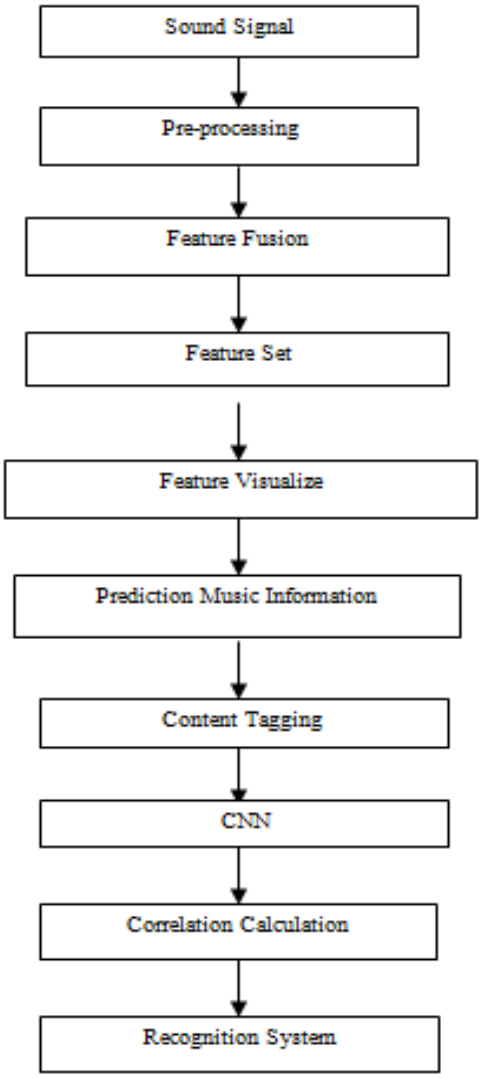

Fig. 1: Block Diagram

Create a spectrogram by plotting the magnitude of the STFT over time, which visually represents the energy distribution across frequencies. Mel-Frequency Cepstral Coefficients (MFCCs): Extract MFCCs, which provide a compact representation of the spectral characteristics of the audio. Chromagram: Compute a chromagram to represent the harmonic and melodic content by mapping the spectrum to 12 pitch classes (chromas).

Onset Detection: Apply functions to detect the beginning of musical events (e.g., notes, beats) in the audio signal. Thresholding: Use thresholding techniques to determine significant onsets.

Pitch Detection: Use autocorrelation methods to detect periodicities in the signal, which correspond to the fundamental frequencies of musical notes. Harmonic Product Spectrum: Apply harmonic product spectrum methods to enhance pitch detection accuracy.

Tempo and Beat Tracking: Analyze the periodicity of onsets or other rhythmic features to estimate the tempo of the music. Beat Detection: Detect the regular beat intervals, which can be used for aligning the music in time.

Data Augmentation: Pitch Shifting: Shift the pitch of the audio without altering the tempo to create augmented data for training recognition models. Time Stretching: Alter the tempo of the audio without changing the pitch for further data augmentation.

Format Conversion: Combine the extracted features into a format suitable for input to machine learning models, such as feature vectors or matrices.

These pre-processing steps transform raw audio into a set of features that are informative and manageable for subsequent stages of music recognition, such as classification, transcription, or similarity analysis. The specific choice of pre-processing techniques can vary depending on the exact application and the characteristics of the audio data.

Feature fusion in music recognition refers to the integration of multiple types of features extracted from music signals to enhance the performance of recognition tasks. These tasks can include music genre classification, mood detection, instrument recognition, and more. By combining different features, systems can capture a more comprehensive representation of the audio, leading to improved accuracy and robustness.

Content tagging in music recognition involves identifying and labeling various elements within a piece of music to enhance its discoverability and analysis. This process uses algorithms to detect features such as genre, mood, tempo, key, instruments, and vocal presence. By tagging these attributes, music recognition systems can categorize and recommend songs more accurately, facilitate personalized playlists, and enable detailed music searches. This advanced tagging also aids in copyright management, licensing, and providing deeper insights into musical trends and listener preferences.

Convolutional Neural Networks (CNNs) have become instrumental in music recognition due to their ability to efficiently process and analyze audio data. By converting audio signals into spectrograms, CNNs can identify patterns and features within the frequency domain, making them particularly effective for tasks such as genre classification, instrument recognition, and even music recommendation. The hierarchical structure of CNNs allows for the extraction of both low-level features like pitch and rhythm, as well as high-level features such as melody and harmony, enabling precise and nuanced music analysis.

Correlation calculation in music recognition involves measuring the statistical relationship between different musical features to identify patterns or similarities. For instance, by analyzing the correlation between the spectral components of audio signals, algorithms can match a recorded clip to a song in a database. This technique helps in recognizing melodies, rhythms, or harmonies by comparing how closely the features of a sample align with known pieces, thus enabling accurate music identification and classification.

IV. RESULTS

In this section a robust music note recognition system using convolutional neural network is observed.

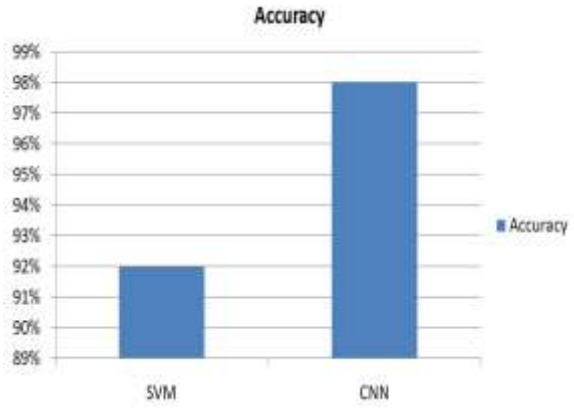

Fig. 2: Accuracy Comparison Graph

In Fig.2 shows the accuracy comparison graph is observed between SVM and CNN. The CNN shows the high comfort.

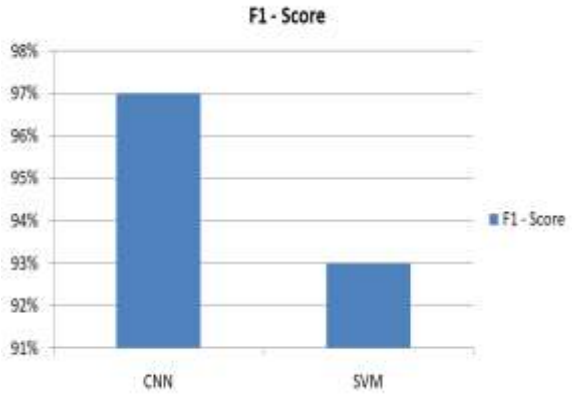

Fig. 3: F1 - Score Comparison Graph

In Fig.3 shows the f1-score comparison graph is observed between CNN and SVM. The CNN shows the high f1-score.

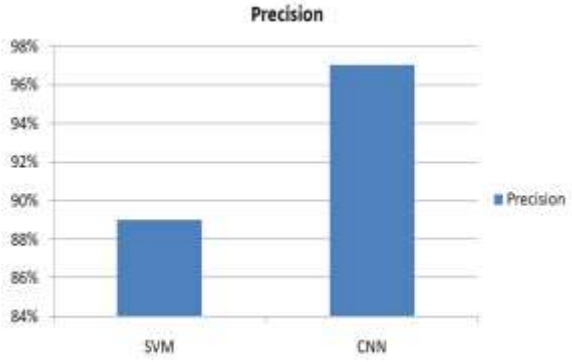

Fig. 4: Precision Comparison Graph

In Fig.4 shows the precision comparison graph is observed between SVM and CNN. The CNN shows the high precision.

V. CONCLUSION

The integration of advanced music recognition technologies, particularly those utilizing convolutional neural networks (CNNs), holds promising potential for enhancing both music information retrieval and music generation. These systems not only improve the accuracy of tasks such as music structure analysis and multi-pitch detection but also offer new possibilities for real-time performance assessment. This could revolutionize music education by providing accessible, cost-effective, and efficient tools for learners. Convolutional Neural Networks (CNN) have been successfully applied to chord recognition. Hence, this analysis achieves better results in terms of accuracy, f1-score and precision.

VI. REFERENCES

[1] R. Su, L. Wang and X. Liu, "Multimodal learning using 3D audio-visual data for audio-visual speech recognition," 2017 International Conference on Asian Language Processing (IALP), Singapore, 2017, pp. 40-43, doi: 10.1109/IALP.2017.8300541.

[2] J. Calvo-Zaragoza, A. -J. Gallego and A. Pertusa, "Recognition of Handwritten Music Symbols with Convolutional Neural Codes," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 691-696, doi: 10.1109/ICDAR.2017.118.

[3] K. -Y. Choi, B. Coüasnon, Y. Ricquebourg and R. Zanibbi, "Bootstrapping Samples of Accidentals in Dense Piano Scores for CNN-Based Detection," 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR), Kyoto, Japan, 2017, pp. 19-20, doi: 10.1109/ICDAR.2017.257.

[4] Y. Liu and Y. Chen, "Recognition of facial expression based on CNN-CBP features," 2017 29th Chinese Control And Decision Conference (CCDC), Chongqing, China, 2017, pp. 2139-2145, doi: 10.1109/CCDC.2017.7978869.

[5] S. Deebika, K. A. Indira and Jesline, "A Machine Learning Based Music Player by Detecting Emotions," 2019 Fifth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), Chennai, India, 2019, pp. 196-200, doi: 10.1109/ICONSTEM.2019. 8918890.

[6] Apurva A. Mehta and Malay S. Bhatt. Optical music notes recognition for printed piano music score sheet. In International Conference on Computer Communication and Informatics, Coimbatore, India, 2015.

[7] Kia Ng, Alex McLean, and Alan Marsden. Big data optical music recognition with multi-images and multi recognisers. In EVA London 2014 on Electronic Visualisation and the Arts, pages 215–218. BCS, 2014.

[8] David Bainbridge and Tim Bell. A music notation construction engine for optical music recognition. Software: Practice and Experience, 33(2):173–200, 2003.

[9] Jorge Calvo-Zaragoza and Jose Oncina. Recognition of pen-based music notation: The HOMUS dataset. In 22nd International Conference on Pattern Recognition, pages 3038–3043. Institute of Electrical & Electronics Engineers (IEEE), 2014.

[10] Jorge Calvo-Zaragoza and David Rizo. Camera-primus: Neural end-to-end optical music recognition on realistic monophonic scores. In 19th International Society for Music Information Retrieval Conference, pages 248–255, Paris, France, 2018.

[11] Jorge Calvo-Zaragoza and David Rizo. End-to-end neural optical music recognition of monophonic scores. Applied Sciences, 8(4), 2018.

[12] Jorge Calvo-Zaragoza, Alejandro Toselli, and Enrique Vidal. Handwritten music recognition for mensural notation: Formulation, data and baseline results. In 14th International Conference on Document Analysis and Recognition, pages 1081– 1086, Kyoto, Japan, 2017.

[13] I. Fujinaga, A. Hankinson and J. E. Cumming, "Introduction to SIMSSA (single interface for music score searching and analysis)", Proceedings of the 1st International Workshop on Digital Libraries for Musicology DLfM@JCDL 2014, pp. 1-3, September 12, 2014.

[14] A. Rebelo, G. Capela and J. S. Cardoso, "Optical recognition of music symbols: A comparative study", International Journal on Document Analysis and Recognition, vol. 13, no. 1, pp. 19-31, Mar. 2010.

[15] A. Rebelo, I. Fujinaga, F. Paszkiewicz, A. Marcal, C. Guedes and J. Cardoso, "Optical music recognition: state-of-the-art and open issues", International Journal of Multimedia Information Retrieval, vol. 1, no. 3, pp. 173-190, 2012.

[16] A. Sharif Razavian, H. Azizpour, J. Sullivan and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition", The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2014.

[17] J. Calvo-Zaragoza and J. Oncina, "Recognition of pen-based music notation: The HOMUS dataset", 22nd International Conference on Pattern Recognition ICPR 2014, pp. 3038-3043, August 24–28, 2014.

[18] R. M. Pinheiro Pereira, C. E. Matos, G. Braz, J. a. D. De Almeida and A. C. De Paiva, "A deep approach for handwritten musical symbols recognition", Proceedings of the 22nd Brazilian Symposium on Multimedia and the Web ser. Webmedia 16, pp. 191-194, 2016.

[19] S. Lee, S. J. Son, J. Oh and N. Kwak, "Handwritten music symbol classification using deep convolutional neural networks", Proceedings of the 3rd International Conference on Information Science and Security, 2016.

[20] L. Bottou, "Large-scale machine learning with stochastic gradient descent" in Proceedings of COMPSTAT2010, Springer, pp. 177-186, 2010.