Visual Guitar/Ukulele Chord Recognition Using *MediaPipe*

Sunny Kumar Tuladhar

School of Engineering and Technology, Master's in Data Science and Artificial Intelligence,
Asian Institute of Technology, Klong Luang, Pathum Thani, Thailand
Sunny.Tuladhar@ait.ac.th

Abstract **- This paper aims to describe a process to visually recognise guitar chords from video in real time. Recognising chords has always been an important and difficult task in music. Generally this task is done using audio but this has its limitations such as being able to detect the correct voicing that is being played by the musician. Here we use *MediaPipe* to detect hand landmarks and crop the landmark image. Then we use simple 2DCNN to recognise the chord shapes. This was trained using a local dataset and then fine tuned to other similar instruments such as ukulele. We see certain chord shapes generalise to other instruments while some pairs of chords are confused even in the same instrument. After training the data we see average results in guitar chord recognition but poor results when fine tuned to the ukulele instrument.**

Keywords**— Music, Computer Vision, Chords classification, *MediaPipe*, Hand Landmarks, Deep Learning**

1. Introduction

A chord is basically combinations of more than two notes that form a unique sound. The interval between each note is what determines what type of chord it is and its key is the main root of the chord. The process of automatic chord recognition [2] is one of assigning a chord label to a section of audio.

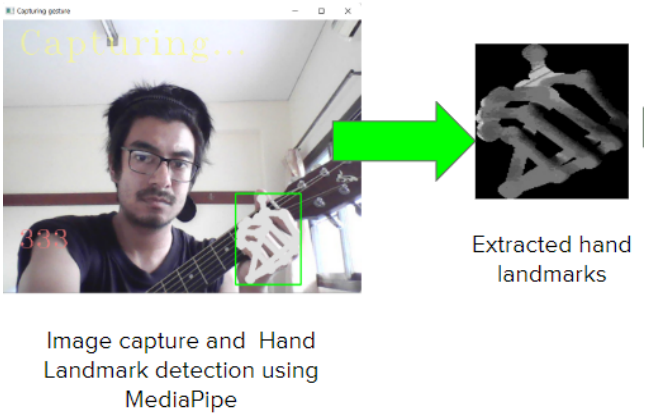

Fig1: Image capture using OpenCV and hand landmark detection using MediaPipe. The cropped images of the extracted landmarks are shown on the right.

Recognition of chords is always a key part in learning to play a piece of music, be it a professional or a beginner. Traditionally this is done by training the ears of musicians to recognise notes in the music. This is a tedious task and requires years of practice and experience. Nowadays automatic chord recognition techniques which use the audio and tell the chord label of the section of audio are also used.

Audio chord recognition has many applications in the area of music information retrieval, such as annotating the harmonic content of audio files in a database or for use in music transcription systems. [1]

In this paper we drift from audio recognition into visual chord recognition. The aim is to identify the chords based on video frames and in real time. This is done so that a user might be able to play the chords being played in a video and in the right voicing as played by the musician.

Here, we try to identify 8 chords each for the guitar and then 8 chords on the ukulele and see how well they fare for practical uses when used in real time. We use *MediaPipe* to detect hand landmarks and then crop them to 200 x 200 images to be fed into a 2DCNN and train the model to see the results.

Ukulele is an instrument very similar to guitar but instead of 6 strings like the latter, only has 4 strings as shown in Fig 2. The soprano ukulele that we are using, is also tuned 5 semitones above the guitar. Bearing this in mind, we try a generalisation test where we simply try a live demo on the ukulele using the model trained only on the guitar. We also fine tune our guitar model to see how it works with the ukulele.

Fig 2: Guitar below and ukulele on top. Ukulele is a similar instrument as the guitar but with 4 strings and smaller fretboard and frets.

2. Related work

2.1 Audio based recognition

Audio based chord recognition has been a very common topic for music based research. Many types of models have been developed to detect chords aurally including Hidden Markov Model [3] and another model where a unique pitch profile was created for each chord and classification was attempted via a nearest-neighbors approach. [4]

3.1 Visual based recognition

Visual approaches have been used to track a guitarists' fingertips and analyse guitar playing [5]. Wang et al.(2018) utilized CNNs to segment hand images and used pose estimation to predict the quality of a guitarist's performance. [6]. Although the aim is not to classify guitar chords the method is similar and requires fingertip and its position detection.

An approach similar to ours using *MediaPipe* was also done by Bahadur, Akshay [8] where he used *MediaPipe* to detect chords in the electric guitar. He has detected famous similar chords and not accounted for chords with similar shape. His work is the major inspiration for this project.

3. Datasets

For this experiment, the data was created using the webcam and is only suitable for a local experiment in a similar context. 800 images were taken from the webcam while each chord was played live. We use Python's OpenCV to capture each frame every 0.05 seconds. We then detect the hand landmarks using MediaPipe and then crop the images to only include the hand landmark as shown in Fig 1. The cropped images were all 200 x 200 pixels as shown in Figure 3. The chords for the guitar were *'A', 'Bm', 'C', 'Cadd9', 'D', 'F', 'G'* and *'C/G'* . The chord paris *C* and *C/G, F* and *Bm*, *Cadd9* and *G* have similar shapes and were deliberately selected to see the model against such similarities. So in total for the guitar we have 6400 images.

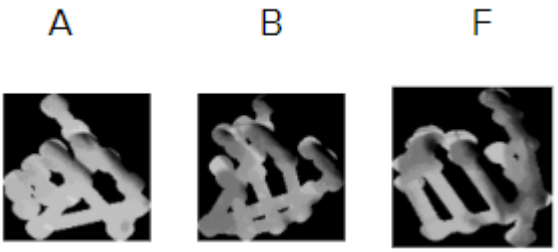

Fig 3: Sample of 200 x 200 images cropped from the video to be used for training in the dataset. The images shows the hand shape for the A , B and the F chord on the Guitar

For the ukulele dataset we only took 200 images and used the same landmark detection and cropping to create the dataset for fine tuning. Since the ukulele has a different string count and chord shape I selected different chords for this dataset. The 8 ukulele chords were *'A'*, *'Am'*, *'C'*, *'D'*, *'Dm'*, *'Em'*, *'F'* and *'G'*. In total for the ukulele we have 1600 images.

4. Method

4.1 *MediaPipe Hands*

MediaPipe is Google's machine learning algorithm that detects human activities including Faces, Iris, Hands, Pose etc. We are interested in the *MediaPipe*'s hand tracking. This offers 21 landmarks in 3d with multi-hand support, from a single frame as shown in Fig 4. *MediaPipe Hands* utilizes an ML pipeline consisting of multiple models working together: A palm detection model that operates on the full image and returns an oriented hand bounding box. A hand landmark model that operates on the cropped image region defined by the palm detector and returns high-fidelity 3D hand keypoints.[7]

Here we use MediaPipe Hands to detect the hand that is playing the chords in the guitar and also detect the landmarks. We then crop the selected section of the players hands landmarks and feed it into the Convolutional Neural Network (CNN)

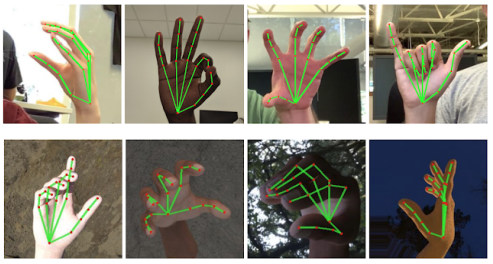

Fig 4: 21 Landmarks for the hand detected by the MediaPipe algorithm. It is also able to display a 3D map of the hands.

4.2 Classification Model

We trained a self created 2D CNN model with the above 200 x 200 images. The model is made up of 3 CNN blocks

and 2 Dense layers as shown in Fig 5. The first block is a Convolution 2D with 32 channels and a kernel size of (5,5) and Relu activation followed by Maxpool2D (2,2)with no padding.

The second block is a Convolution 2D with 64 channels and a kernel size of (5,5) and Relu activation followed by Maxpool2D (5,5) with a stride of (2,2) with no padding.

The third block is a Convolution 2D with 128 channels and a kernel size of (5,5) and Relu activation followed by Maxpool2D (5,5) with a stride of (2,2) with no padding.

The final block is a fully connected layer with 1024 nodes and Relu activation with a Dropout of 0.6, followed by nodes equal to the number of classes i.e 8 with the softmax activation for classification.

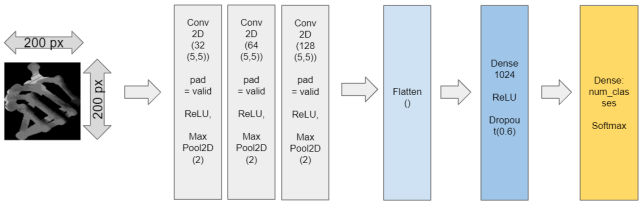

Fig 5: The Convolutional Neural network that was trained on the hand landmarks

5. Training and results

For the guitar training the 800 cropped hand landmark images of 200 x 200 pixels each of the 8 chords , totaling 6400 images, were split 80-20 percent into training and test. The images were trained using the 2DCNN model for 10 epochs using Adam optimizer with a batch size of 64.

The results for these guitar shapes gave a validation accuracy of 83.4% but since the image data is very localized and restricted to a single instrument and a single person at almost the same angle it does not generalise well.

The performance of the model in a live video demonstration shows good performance while detecting the chords ‘D’ and ‘A’ . As expected the model gets confused between the chords ‘G’ and ‘Cadd9’, ‘F’ and ‘Bm’ and ‘C’ and ‘C/G’. The hand landmarks of these chord pairs are shown in Fig 6.

A generalisation test of this guitar trained model on the ukulele to detect the ‘G’ chord on the ukulele seemed to perform well. The ‘G’ on the ukulele is the exact same shape as the ‘D’ in the guitar.

Then I used this guitar trained model and fine-tuned it with the ukulele hand landmarks for the ukulele chord shapes. Here I used 200 images for 8 ukulele chords different from the guitar. The result gave a good accuracy of 84% but similar to above does not generalise well and when tried on the live video the results are not convincing and useful enough with only fine tuning.

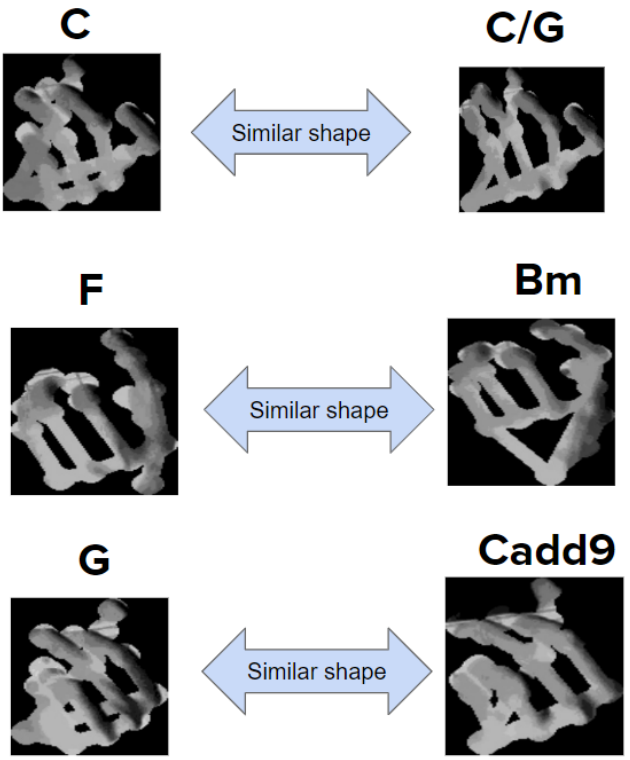

Fig 6: Hand Landmarks of different chords. The chords at the left and right have similar shapes but sound different.

7. Discussion/Future Works

The good performance on the guitar could be attributed to the localized dataset and the similar environment it was tested on. But it is still a good sign that indicates that increasing the image variation on the dataset with different angles, different hands and more variations could give a more general and accurate model that could be used to detect chords from live video recordings of famous musicians given a big enough chord catalog.

The similar chord pair confusion was caused due to the similarity in the finger shapes while playing although they sound very different when listened to. So this model could be combined with audio to get a combined better result to predict the chord sound and the chord shape to be played.

The poor performance on the ukulele compared to the guitar, could be due to the less data used for fine-tuning the model. If more data could be used, a better model could be formed and the same generalisation criteria as the guitar could be implemented in future datasets. Also the ukulele has smaller frets so the fingers are more crammed up so detection could be more difficult in this case.

If some commercial dataset in this domain could be made available with images from a practical case, such as music videos of guitarists with correct labels it would greatly help improve this domain.

Some recommendations for future work could be

1. Use fretboard information for classification of Bar chords in the correct key
2. Combine audio signals with visual recognition to better distinguish similar looking and similar sounding chords.
3. MediaPipe Hands can create 3D shape of the hand landmarks which could be used to train the model instead of the images
4. A better commercial dataset in this domain of cropped guitar chord hands from actual music videos by famous artists with correct chord labels could help this domain of research
5. Train the models using better and more efficient CNN models such as ResNet or ResNext.

8. Conclusion

Here we took MediaPipe hand landmarks of the hand playing chord shapes in the guitar and ukulele and trained them in a simple CNN and got good results for the guitar and below average for the ukulele.

Given good enough data and a good model this approach could be useful for beginner and intermediate guitar players to copy the greats with ease without having to rely on years of theory knowledge or audio training. This opens new possibilities for learning as it will be able to showcase *how* a chord is being played instead of just *what* chord. How a chord is being played sometimes changes the true feeling of playing music and is done very well by professionals.

This will also help transcribers to give more accurate tabular representation of the guitar musical piece instead of having to rely only on their ears and eyes.

Thus further research in this visual recognition of the chords will definitely benefit the musician community for the beginners and the professionals.

Acknowledgment

This research was possible due to the project work we had to do as a part of the Master's in Data Science and Artificial Intelligence course in the subject Computer Vision taught by Prof. Matthew Dailey. We would also like to thank our Teaching Assistant ,Alisa Kunapinun , for her continuous support throughout the course.

References

[1] A. Stark and M. Plumbley, 'Real-Time Chord Recognition for Live Performance', 08 2009.

[2] C. A. Harte and M. B. Sandler, "Automatic chord identification using a quantised chromagram," in Proc AES 118th Convention, no. 6412, Barcelona, 2005.

[3] Sheh, A. and Ellis, D.P.W. "Chord Segmentation and Recognition Using EM-Trained Hidden Markov Models", Proc. ISMIR, 2003.

[4] Fujishima, T. "Realtime Chord Recognition of Musical Sound: a System Using Common Lisp Music", Proc. ICMC, pp.464–467, 1999.

[5] Kerdvibulvech, Chutisant, and Hideo Saito. "Vision-based detection of guitar players' fingertips without markers." Computer Graphics, Imaging and Visualisation (CGIV 2007). IEEE, 2007.

[6] Wang, Z, et al. "A 3D Guitar Fingering Assessing System Based on CNN-Hand Pose Estimation and SVR-Assessment." Society for Imaging Science and Technology. 2018.

[7] Hands. (2022). Retrieved 24 July 2022, from https://google.github.io/mediapipe/solutions/hands

[8] "GitHub - akshaybahadur21/Guitar-Learner: Guitar chord detection and classifier for humans 🎸", GitHub, 2022. [Online]. Available: https://github.com/akshaybahadur21/Guitar-Learner. [Accessed: 24- Jul- 2022]