Hand Cricket Game using CNN and MediaPipe

Sai Surya Teja Gontumukkala, Yogeshwara Sai Varun Godavarthi, Bhanu Rama Ravi Teja Gonugunta, Suja Palaniswamy*
Department of Computer Science & Engineering
Amrita School of Computing, Bengaluru
Amrita Vishwa Vidyapeetham, India
saisurya029@gmail.com, gysaivarun31@gmail.com, ravitejagonugunta150@gmail.com, p_suja@blr.amrita*

***Abstract*— In the era of computer-based games, every other game is being computerized. The work involves implementing a Hand Gesture Based Hand Cricket Game, i.e., Once the player starts playing by keeping his/her video, the hand gestures of the player are recognized and the score of the player is calculated until he gets out of the game or video. It is implemented in two Approaches i.e., first one using own dataset and custom-developed Convolutional Neural Network (CNN) and another using MediaPipe inbuilt model in Python Using OpenCV, Keras, and MediaPipe Packages. The trained CNN gave an accuracy of 0.9767 after ten epochs trained.**

Keywords*— *Hand Gesture, CNN, MediaPipe, OpenCV, Keras

I. Introduction

Hand Cricket game is one of the popular games among people across ages and especially children. It is one of the best games to play when there is no equipment available. It requires two people's involvement, where players play the game based on some set of rules. If the batsman wants to score 2 runs then the batsman show two’s gesture in hand like showing pointing index finger followed by middle finger. Similarly, players should show gestures of 1(pointing finger), 3(pointing, middle and ring fingers), 4(pointing, middle, ring, and pinky fingers), and 5(all fingers) to score 1, 3, 4, and 5 runs, respectively. A batsman can score a whopping 6 runs by showing only Thumb Finger. The game begins with choosing odd or even, and the winner gets an opportunity to select bowling or batting. The batsman keeps playing until the bowler and batsman shows same hand signs i.e., score the same runs, then the batsman gets out. If the batsman gets out then the game terminates at a particular score for the batsman, then the next player gets a chance to bat and should score more than the previous batsman, to win the other player. The total score of a player is the sum of runs he scored every time until he gets out.

Convolutional Neural Network (CNN) is a type of neural network, most commonly used for recognition, analysis, and classification of images. Keras is a package for deep learning algorithms in python. The MediaPipe framework is a cross-platform framework for developing multimodal applied machine learning pipelines. i.e., it holds trained ML Algorithms which can be used in many applications.

Hand Cricket is one of the simple and fun filled games. Till now, it has been played as a traditional physical game and console based computerized game. In this work it is implemented using image processing and deep learning techniques. This computerized implementation of hand cricket will be extremely helpful in the current scenario, because of the covid 19 pandemic physical games are not possible, people can play this type of games when they are alone. Differently abled people face difficulties in playing physical games such as cricket, hockey, etc. This game will be helpful for them as well.

The remaining sections of the paper is outlined as follows:
In Section II, the works related to this work are explained briefly. In Section III, the concepts of CNN architecture and grey level thresholding are elaborated. In Section IV, the complete implementation details along with the output of this work are shown. In Section V, the conclusion of this work is discussed.

II. Related Works

Image classification is the process of identifying and labelling groups of pixels or vectors within an image based on specific rules. There are various algorithms for image classification such as SVM, CNN, K-NN, and Decision Tree etc. CNNs have defied expectations and ascended to the throne as the most advanced computer vision technique. CNNs are by far the most successful among the several types of neural networks such as LSTM, ANN, RNN etc., and CNNs are widely used in image data applications [1]. CNN is a framework which has an ability to learn features of the specific domain on its own. Variations of the CNN in different applications are explored [2]. The feature vectors are sent to SVM, From the CNN. It is showed that using Pre-Trained CNN as feature extractor is an exceptionally reliable approach or the recognition of hand written digits [3]. A new method is proposed that combines SQI (Self Quotient Image) Algorithm with CNN to produce a bypass feature maps input. With the help of the basic image processing this structure allowed the appending of features with prior information to the CNN [4].

Breast Cancer Classification was done using Conventional Methods as well as Automatic Methods of Machine Learning Approaches. While both gave good accuracy, Extra Tree Classifier outperformed other algorithms [5]. An emotion recognition system named DPIIER was proposed which can work in different illumination conditions and for different head poses. The Model is evaluated using Type III-fold cross validation and it is trained and tested on GPU System [6]. NDVI's ability is used as the stand-alone parameter to change existing four bands CNN to two band with fewer filters for satellite image classification. The results showed that SAT-4 image classification achieved a 98.01% accuracy by using this method [7]. CNN based on thermal image is proposed to recognize people on an ultra-low power-resource constrained system. It is designed to fit in memory less than 500kB and will detect heads and count the number of people in a respective classroom with a 99 percent accuracy [8].

The analysis of CNN's performance on classification of sports video is done and proved that it is exactly accurate on a single CPU [9]. A CNN with few parameters than normal was proposed which solves problems like overfitting, gradient vanishing to classify images. An accuracy of 99.467 percent was achieved for MNIST Dataset using this model [10]. The sign language detection framework using RNN along with MediaPipe hand tracking module. The model is trained with Common words used in Vietnamese and produced accurate results [11].

In previous literature, image classification was used in various domains such as Colors Identification, Animals Classification, Handwritten Digits, Breast Cancer Detection, Emotion Recognition, Image and video content classification etc. The proposed model focused on the usage of image classification in the domain of gaming i.e., implementation of hand cricket game using CNN and MediaPipe, where the CNN is used to classify images in the first approach and MediaPipe's hand tracking module is used to implement the game in the second approach.

III. CONCEPTS

In this section, the architecture of the CNN that includes Convolution, ReLu, Pooling, SoftMax function and the concept of Grey level Thresholding is included.
A typical ANN is sensitive to image location and takes more computation. So, it is not recommended for image classification [12].

CNN is efficient to solve both. It uses filters which is used to extract the features so that it is not sensitive to image location i.e., recognizes the regions of interest irrespective of its location and pooling for reducing the dimensions to decrease computation. The architecture of a typical CNN is shown in Fig. 1.

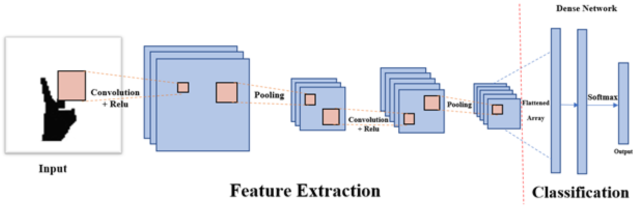


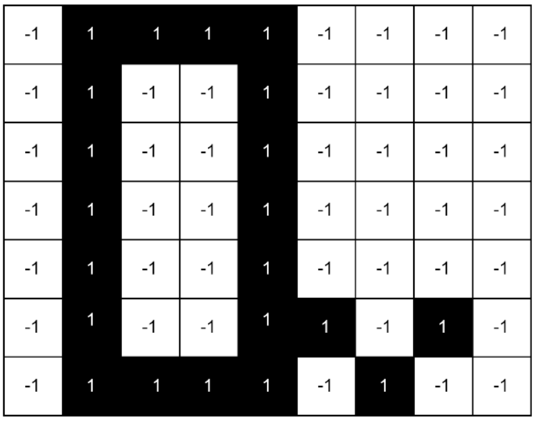
Fig. 1. Diagrammatic Representation of a CNN

A. Convolution:

Consider a hand written digit 'a' as shown in the Fig. 2.

Fig.2. Hand written alphabet

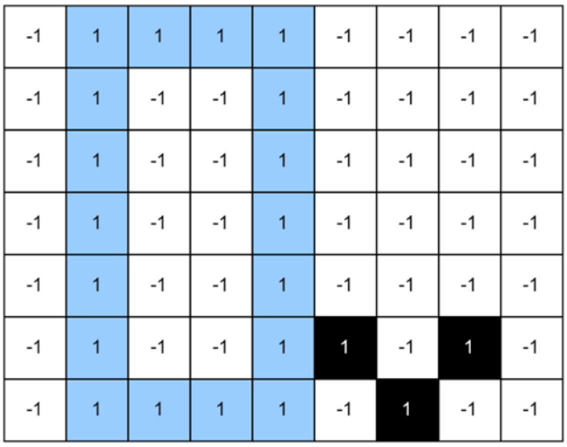
Consider, representing the image having the alphabet "a" with a 7 x 9 grid with intensity values in it, as shown in Fig. 3.



-1	1	1	1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	1	-1	1	-1
-1	1	1	1	1	-1	1	-1	-1

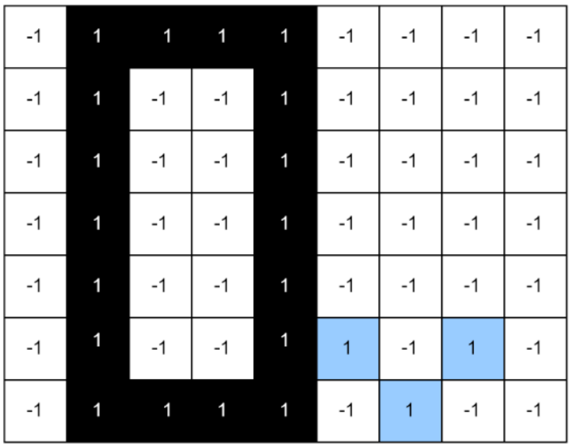
Fig.3. Letter "a" in pixel format

Consider three features namely loopy pattern filter as shown in Fig.3.a, curved pattern filter as shown in Fig.3.b, Diagonal pattern filter as shown in Fig.3.c to identify the hand written alphabet 'a' [3].



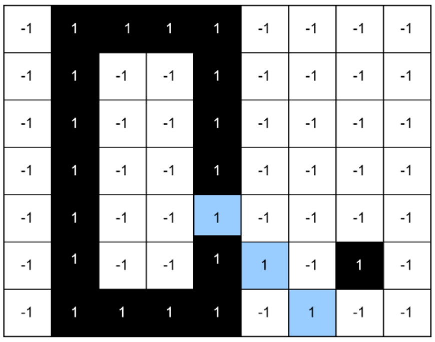
-1	1	1	1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	1	-1	1	-1
-1	1	1	1	1	-1	1	-1	-1

Fig.3.a: Loopy Pattern Filter



-1	1	1	1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	1	-1	1	-1
-1	1	1	1	1	-1	1	-1	-1

Fig.3.b: Curved Pattern Filter



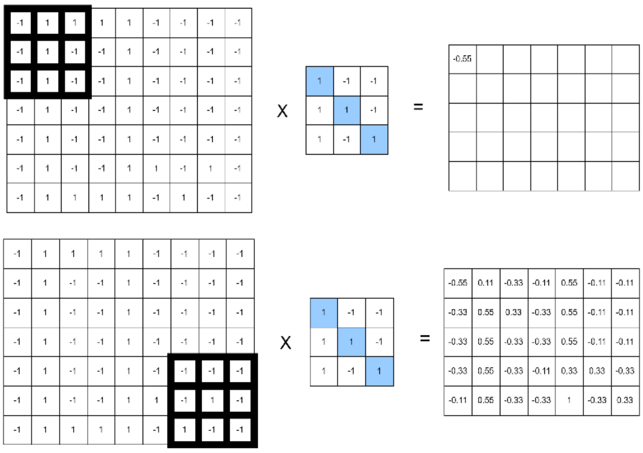
-1	1	1	1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	1	-1	1	-1
-1	1	1	1	1	-1	1	-1	-1

Fig.3.c: Diagonal Pattern filter

If the filter is n x n, a n x n matrix is considered and individual elements will be multiplied followed by averaging and it is entered in a new matrix which is called as the feature map since it maps all the features.

Example:

For Loopy Pattern Filter:



-1	1	1	1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	1	-1	1	-1
-1	1	1	1	1	-1	1	-1	-1

X

1	-1	-1
1	1	-1
1	-1	1

=

-0.55						

-1	1	1	1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	-1	-1	-1	-1
-1	1	-1	-1	1	1	-1	1	-1
-1	1	1	1	1	-1	1	-1	-1

X

1	-1	-1
1	1	-1
1	-1	1

=

-0.55	0.11	-0.33	-0.11	0.55	-0.11	-0.11
-0.33	0.55	0.33	-0.33	0.55	-0.11	-0.11
-0.33	0.55	-0.33	-0.33	0.55	-0.11	-0.11
-0.33	0.55	-0.33	-0.11	0.33	0.33	-0.33
-0.11	0.55	-0.33	-0.33	1	-0.33	0.33

Fig.4.a&b: Working of Convolution with filter with an example

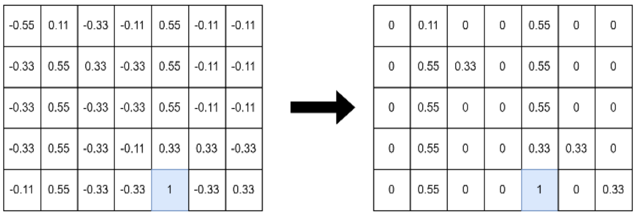
Here Fig.4.a&b depicts the process of convolution of image with filter for an example, first element of the 3 x 3 grid of the entire matrix is multiplied with the first element of the filter, second element of the grid with is multiplied second element of the filter and so on and all the obtained values will be summed and it is divided with no. of elements present in the filter.
And so, on up to the last possible 3 x 3 grid.

Benefits of convolution:

- Not every node relates to other nodes (sparse networks), which reduces over fitting.
- Convolution and Pooling gives a Location invariant Feature Detection.
- Parameter sharing i.e., a parameter learned by a filter can be shared to another filter.

B. *ReLu:*

It is an activation function that will convert the negative numbers to '0' and positive numbers remains same. The working of a ReLu activation function is clearly conveyed through Fig. 5.



-0.55	0.11	-0.33	-0.11	0.55	-0.11	-0.11
-0.33	0.55	0.33	-0.33	0.55	-0.11	-0.11
-0.33	0.55	-0.33	-0.33	0.55	-0.11	-0.11
-0.33	0.55	-0.33	-0.11	0.33	0.33	-0.33
-0.11	0.55	-0.33	-0.33	1	-0.33	0.33

0	0.11	0	0	0.55	0	0
0	0.55	0.33	0	0.55	0	0
0	0.55	0	0	0.55	0	0
0	0.55	0	0	0.33	0.33	0
0	0.55	0	0	1	0	0.33

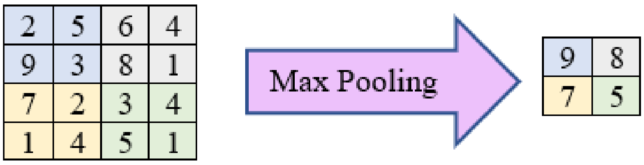
Fig.5. Working of ReLu Activation Function with an example

Benefits of ReLu:

- It makes computation faster.
- It introduces non-Linearity.
- Nonlinearity is required in activation functions because their goal in a neural network is to produce a nonlinear decision boundary using nonlinear weight and input combinations.

C. *Pooling:*

Pooling is used to reduce the dimensions of the feature map. The following example in Fig. 6 is a Maximum Pooling with 2 x 2 filter and stride '2'. In Maximum Pooling, the maximum number will be taken out of all numbers.



2	5	6	4
9	3	8	1
7	2	3	4
1	4	5	1

9	8
7	5

Fig.6. Working of Max Pooling with an example

D. *Thresholding:*

- Thresholding is among the most important methods of image segmentation, in which pixels with equivalent grey scale (or any other feature) are clustered or grouped together.
- An image histogram is commonly used to determine the best threshold setting (s).
- A single threshold is appropriate, since some images (such as scanned text) are bimodal [2,4,7-9,11].

Grey-level Thresholding:

Let, 'T' be the Threshold Value

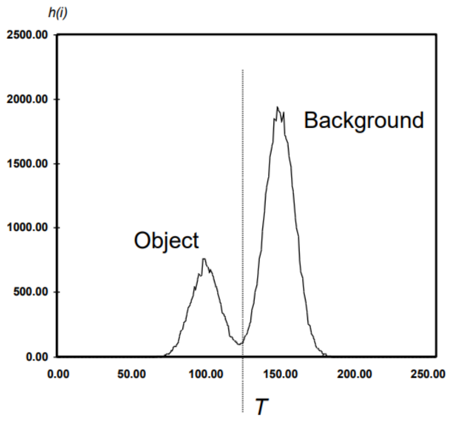


Fig.7. Gray Level Thresholding

Grey level thresholding algorithm works as follows: –

```
If (grey level of pixel p <=T):
        "Pixel p is an object pixel"
Else:
        "Pixel p is a background pixel"
```

Any threshold divides the histogram into two groups as displayed in Fig.7; each with its own set of statistics (mean, variance) Each group's homogeneity is assessed using the variation within groups. The best threshold is the one that satisfies all the above criteria [1,6].

E. SoftMax Function:

SoftMax is a mathematical function that transforms a vector of integers into a vector of probabilities, with each value's probability proportional to its relative scale in the vector [5].

Example:

$$\begin{bmatrix}3\\1\\0.2\\4\end{bmatrix} \quad \rightarrow \quad \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \quad \rightarrow \quad \begin{bmatrix}0.255\\0.034\\0.015\\0.694\end{bmatrix}$$

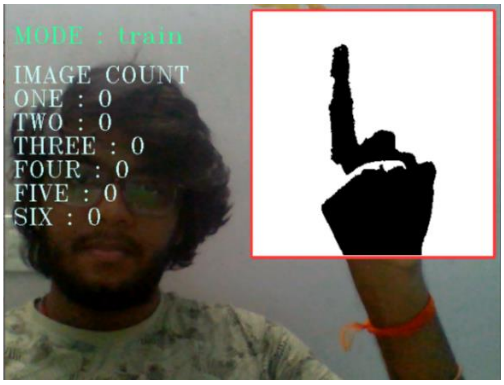
Output Layer   SoftMax activation Function   Probabilities

IV. IMPLEMENTATION

In this work, we have followed two approaches: first one using CNN and second one using hand tracking module in MediaPipe.

A. Using CNN:

An in-house developed custom data set is used in this approach. It is created using Open-CV. The coordinates of region of interest (ROI) are given and the hand region which is the ROI is extracted. "The hand gestures dataset" size is 6300, it has six categories i.e., 1, 2, 3, 4, 5, 6 where each category has 1000 images for training 50 images for testing.

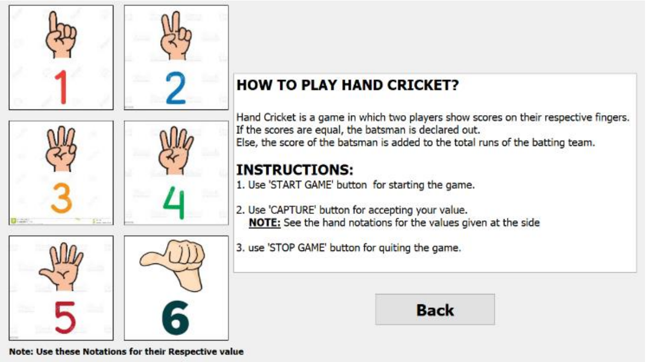


MODE : train
IMAGE COUNT
ONE : 0
TWO : 0
THREE : 0
FOUR : 0
FIVE : 0
SIX : 0

Fig.8. Dataset Creation for hand gesture "one"

The dataset can be created as shown in Fig.8, in that the image is corresponding to the 1's gesture. The gesture's number can be given through keyboard input so that the image will get stored in that directory.

Keras Module is used to build the neural network. CNN consists of two convolution layers of 32 3 x 3 Filters (no. of parameters are restricted, so that more no. of un related features can be restricted) with ReLu Activation (it makes computation faster and makes the neural network non-linear), each Layer Followed by a Max Pooling Layer (it down scales the image and it helps in extracting the most key features of the image) and the Dense Layer Containing a Hidden Layer, SoftMax function at the end. The CNN classifier used Adam optimizer (it is the best performing optimizer and works good with sparse data), categorical cross entropy loss function and accuracy as the evaluation metrics. The images are converted to black and white using gray level thresholding which helps in the dividing the image into two parts i.e., hand will be converted to black and the rest to white before training the model. The model is trained with 10 epochs and testing data set is given as the validation set. The GUI is implemented using PyQt5. The instructions of the game will appear when the instructions button is clicked in the home menu as shown in Fig.9 below. After reading the instructions the player can go back and start the game through home menu. A sample game play is shown in Fig. 10.



HOW TO PLAY HAND CRICKET?

Hand Cricket is a game in which two players show scores on their respective fingers.
If the scores are equal, the batsman is declared out.
Else, the score of the batsman is added to the total runs of the batting team.

INSTRUCTIONS:

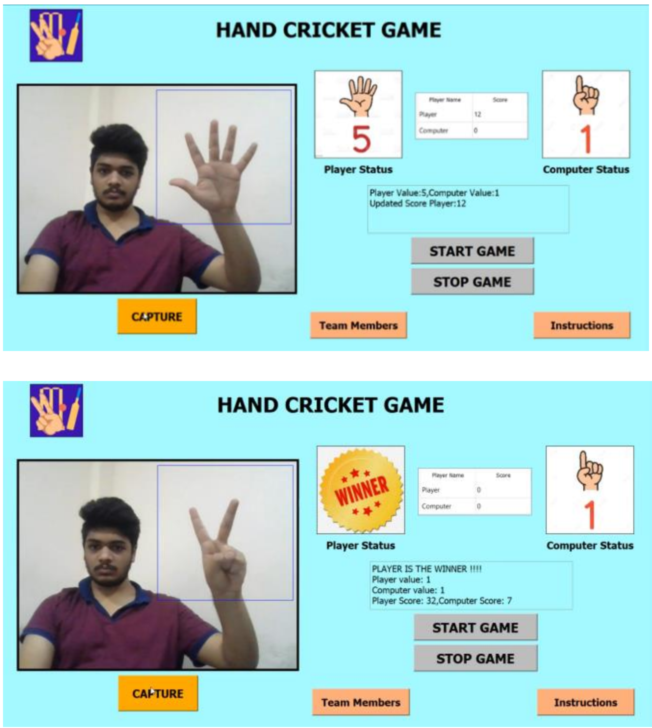
1. Use 'START GAME' button for starting the game.

2. Use 'CAPTURE' button for accepting your value.
   NOTE: See the hand notations for the values given at the side

3. use 'STOP GAME' button for quiting the game.

Note: Use these Notations for their Respective value

Fig.9. Instructions of the game



Player Name	Score
Player	12
Computer	0

Player Value:5,Computer Value:1
Updated Score Player:12

Player Name	Score
Player	0
Computer	0

PLAYER IS THE WINNER !!!!
Player value: 1
Computer value: 1
Player Score: 32,Computer Score: 7

Fig.10. a&b: Game Play using CNN model

B. *Using MediaPipe:*

MediaPipe is an open source frame work developed by google which contains customizable ML solutions. It is platform friendly since it can run in any environment. Hand-Tracking module of MediaPipe is used to recognize the number of fingers opened. It can be identified by the notation given in Fig.11.

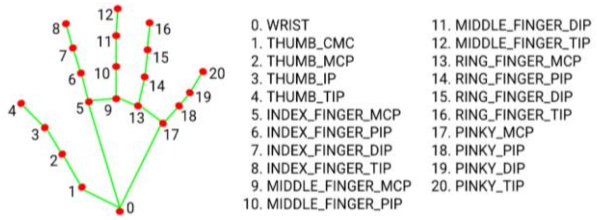
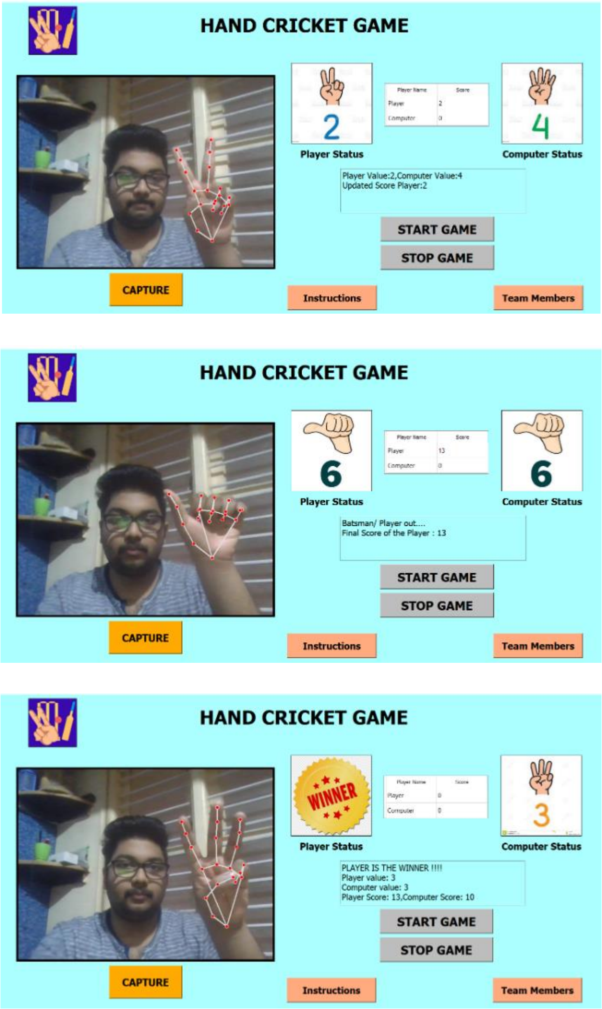


Fig.11. Hand landmarks and their indices

As shown in the above image, each joint will have a number associated to it. Hand tracking module will return an array of size 21 each value representing whether joint of left hand is visible or not. If all the joints of a certain finger are visible, then the finger is considered to be open and its count will be added to the count of total number of fingers opened. The total number of fingers open is the runs scored by the batsman. The game play of hand cricket using MediaPipe is exhibited in Fig.12. a,b&c.



Player Name	Score
Player	2
Computer	0

Player Value:2,Computer Value:4
Updated Score Player:2

Player Name	Score
Player	13
Computer	0

Batsman/ Player out....
Final Score of the Player : 13

Player Name	Score
Player	0
Computer	0

PLAYER IS THE WINNER !!!!
Player value: 3
Computer value: 3
Player Score: 13,Computer Score: 10

Fig.12. a,b&c: Game Play using MediaPipe

The trained convolution neural network gave an accuracy of 0.9767 after ten epochs trained. As MediaPipe is an in-built framework the accuracy will be perfect.

V. CONCLUSION AND FUTURE SCOPE

In this work, we implemented a CNN for hand cricket game using Python and another implementation using MediaPipe. The CNN model was trained using the custom home-made data set and achieved good accuracy. The custom-built dataset was created using left hand images, so the player should play only with his left hand. In future even the right-hand images can be added to the dataset so that the game works for both the hands and it can be developed as two player game instead of computer being second player.

REFERENCES

[1] S. S. Teja Gontumukkala, Y. S. Varun Godavarthi, B. R. Ravi Teja Gonugunta, R. Subramani and K. Murali, "Analysis of Image Classification using SVM," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 01-06, doi: 10.1109/ICCCNT51525.2021.9579803.

[2] N. Aloysius and M. Geetha, "A review on deep convolutional neural networks," 2017 International Conference on Communication and Signal Processing (ICCSP), 2017, pp. 0588-0592, doi: 10.1109/ICCSP.2017.8286426.

[3] Yoshihiro Shima, Yumi Nakashima, and Michio Yasuda. 2018. Handwritten Digits Recognition by Using CNN Alex-Net Pre-trained for Large-scale Object Image Dataset. In Proceedings of the 3rd International Conference on Multimedia Systems and Signal Processing (ICMSSP '18). Association for Computing Machinery, New York,NY,USA, 36–40. DOI:https://doi.org/10.1145/3220162.3220163

[4] Xingrun Xing, Min Dong, Cheng Bi, and Lin Yang. 2019. Self-Quotient Image based CNN: A Basic Image Processing assisting Convolutional Neural Network. In Proceedings of the 2019 3rd International Conference on Digital Signal Processing (ICDSP 2019). Association for Computing Machinery, New York, NY, USA, 17–21. DOI:https://doi.org/10.1145/3316551.3316567

[5] A. J. B and S. Palaniswamy, "Comparison of Conventional and Automated Machine Learning approaches for Breast Cancer Prediction," 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA), 2021, pp. 1533-1537, doi: 10.1109/ICIRCA51532.2021.9544863.

[6] S. Palaniswamy and Suchitra, "A Robust Pose & Illumination Invariant Emotion Recognition from Facial Images using Deep Learning for Human-Machine Interface," 2019 4th International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), 2019, pp. 1-6, doi: 10.1109/CSITSS47250.2019.9031055.

[7] Unnikrishnan, Anju & Vishvanathan, Sowmya & Kp, Soman. (2019). A Two-Band Convolutional Neural Network for Satellite Image Classification. 10.1007/978-981-13-0212-1_17.

[8] Andres Gomez, Francesco Conti, and Luca Benini. 2018. Thermal image-based CNN's for ultra-low power people recognition. In Proceedings of the 15th ACM International Conference on Computing Frontiers (CF '18). Association for Computing Machinery, New York, NY, USA, 326–331. DOI:https://doi.org/10.1145/3203217.3204465

[9] Rani, N Shobha & Rao, Pramod & Clinton, Paul. (2018). Visual recognition and classification of videos using deep convolutional neural networks. International Journal of Engineering and Technology(UAE). 7. 85-88. 10.14419/ijet.v7i2.31.13403.

[10] Huang Shuo and Hoon Kang. 2021. Deep CNN for Classification of Image Contents. In 2021 3rd International Conference on Image Processing and Machine Vision (IPMV) (IPMV 2021). Association for Computing Machinery, New York, NY, USA, 60–65. DOI:https://doi.org/10.1145/3469951.3469962

[11] Bach Duy Khuat, Duong Thai Phung, Ha Thi Thu Pham, Anh Ngoc Bui, and Son Tung Ngo. 2021. Vietnamese sign language detection using MediaPipe. In 2021 10th International Conference on Software and Computer Applications (ICSCA 2021). Association for Computing Machinery, New York, NY, USA, 162–165. DOI:https://doi.org/10.1145/3457784.3457810

[12] Dhaval Patel, Image classification using CNN (CIFAR10 dataset), Jul. 2021. Available: https://www.skillbasics.com/courses/deep-learning-with-tensorflow-keras-and-python/lecture/186