RV Educational Institutions ®

RV College of Engineering ®

Autonomous
Institution Affiliated
to Visvesvaraya
Technological
University, Belagavi

Approved by AICTE,
New Delhi, Accredited
By NAAC, Bengaluru
And NBA, New Delhi

Go, change the world

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

MINOR PROJECT REPORT

ON

Human Action Recognition in Sports

Submitted by,

Navnith Bharadwaj	**1RV19CS098**
Kumar Prakhar	**1RV19CS078**
Mohamed Moin Irfan	**1RV19CS089**

Under the guidance of

Dr. Praveena T
Assistant Professor
Department of CSE
RV College of Engineering

In partial fulfillment for the award of degree
of
Bachelor of Engineering
in
Computer Science and Engineering
2020-2021

RV COLLEGE OF ENGINEERING®, BENGALURU-59

(Autonomous Institution Affiliated to VTU, Belagavi)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

CERTIFICATE

Certified that the minor project work titled ***'Human Action Recognition in sports'*** is carried out by **Navnith Bharadwaj(1RV19CS098), Kumar Prakhar(1RV19CS078), Mohamed Moin Irfan(1RV19CS089),** who are bonafide students of RV College of Engineering, Bengaluru, in partial fulfillment for the award of degree of **Bachelor of Engineering in Computer Science and Engineering** of the Visvesvaraya Technological University, Belagavi during the year 2020-2021. It is certified that all corrections/suggestions indicated for the Internal Assessment have been incorporated in the minor project report deposited in the departmental library. The Minor Project report has been approved as it satisfies the academic requirements in respect of minor project work prescribed by the institution for the said degree.

Signature of Guide	**Signature of Head of the Department**	**Signature of Principal**
Dr.Praveena T	**Dr. Ramakanth Kumar P**	**Dr.K.N.Subramanya**

External Viva

Name of Examiners **Signature with Date**

1

2

RV COLLEGE OF ENGINEERING®, BENGALURU-59

(Autonomous Institution Affiliated to VTU, Belagavi)

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING

DECLARATION

We **Navnith Bharadwaj, Mohamed Moin Irfan, Kumar Prakhar,** students of sixth semester B.E., department of CSE, RV College of Engineering, Bengaluru, hereby declare that the minor project titled **'Human Action Recognition in Sports'** has been carried out by us and submitted in partial fulfillment for the award of **Bachelor of Engineering** in **Computer Science and Engineering** during the year 2020-21.

Further we declare that the content of the report has not been submitted previously by anybody for the award of any degree or diploma to any other university.

We also declare that any Intellectual Property Rights generated out of this project carried out at RVCE will be the property of RV College of Engineering, Bengaluru and we will be one of the authors of the same.

Place: Bengaluru

Date:

Name	Signature
1. **Navnith Bharadwaj**(1RV19CS098)	
2. **Kumar Prakhar** (1RV19CS078)	
3. **Mohamed Moin Irfan**(1RV19CS089)	

Acknowledgement

We are indebted to our guide**, Dr. Praveena T, Assistant professor, Department of CSE** for her wholehearted support, suggestions and invaluable advice throughout our project work and also helped in the preparation of this thesis.

We also express gratitude to our Minor Project lab faculty **Prof**.**Revathi S A, Assistant Professor,** Department of Computer Science and Engineering for her valuable comments and suggestions.

Our sincere thanks to **Dr. Ramakanth Kumar P.**, Professor and Head, Department of Computer Science and Engineering, RVCE for his support and encouragement.

We express sincere gratitude to our beloved Principal, **Dr. K. N. Subramanya** for his appreciation towards this project work.

We thank all the **teaching staff and technical staff** of the Computer Science and Engineering department, RVCE for their help.

Lastly, we take this opportunity to thank our **family** members and **friends** who provided all the backup support throughout the project work.

Abstract

With the rapid development of computer vision technology, human action recognition technology has occupied an important position in this field. It has important practical value and research value in security protection, advanced human-computer interaction, video search analysis and sports analysis. Due to the non rigid body characteristics of the human body, the change of illumination, and the influence of the changeable surrounding environment, human action recognition is more challenging. Virtual reality technology is an important subject in the computer field. It simulates real scenes by means of computer software and hardware technology. Observe students' learning situation dynamically according to the scene. At present, this technology is still in the research and development stage, and there are still many problems in the application process. However, the exchange of students' learning information and simulated scenes is of great help to teaching activities. Moreover, the data obtained by virtual technology also provides guidance for teaching and research.

Our study includes the LSTM model and CNN, LSTM model is an advanced RNN, a sequential network, that allows information to persist. It is capable of handling the vanishing gradient problem faced by RNN. A recurrent neural network also known as RNN is used for persistent memory.The obtained dataset is fed into the OPENPOSE CNN model which is the first real-time multi-person system to jointly detect human body, hand, facial, and foot key-points (in total 135 key-points) on single images. It gives 17 joints as numerical quantities.

The obtained dataset is fed into the LSTM model which gives the action as output. These actions are numbers which are used to determine the action class. The model performs with an accuracy of 95.44% after performing 300 epochs.

Availability of proper dataset or making of huge dataset which covers all the categories of various actions is responsible for the efficiency of the current model. Human Action Recognition (HAR) aims to understand human behavior and assign a label to each action. It has a wide range of applications, and therefore has been attracting increasing attention in the field of computer vision. Human actions can be represented using various data modalities, such as rgb, **skeleton**, depth, infrared, point cloud, event stream, audio, acceleration, radar, and WiFi signal, which encode different sources of useful yet distinct information and have various advantages depending on the application scenarios

Table of Contents

Chapter 1
Introduction

1.1. State of Art Developments

With the rapid development of computer vision technology, human action recognition technology has occupied an important position in this field. It has important practical value and research value in security protection, advanced human-computer interaction, video search analysis and sports analysis. Due to the non rigid body characteristics of the human body, the change of illumination, and the influence of the changeable surrounding environment, human action recognition is more challenging**.** Technology powered by Artificial intelligence is used in various HAR(Human Action recognition) systems. Our project focuses on building a model that recognises human actions accurately from a given video. This can be used in entertainment, athlete performance analysis, etc.

1.2 Motivation

Human activity recognition plays a significant role in human-to-human interaction and interpersonal relations. Because it provides information about the identity of a person, their personality, and psychological state, it is difficult to extract. The human ability to recognize another person's activities is one of the main subjects of study of the scientific areas of computer vision and machine learning.

The goal of human activity recognition is to examine activities from video sequences or still images. Motivated by this fact, human activity recognition systems aim to correctly classify input data into its underlying activity category.

1.3 Problem Statement

In today's world, sports has become an important part of society, culture and gaming.

- Many systems have been developed to help athletes improve their performance and have better engagement with the audience.

- Particularly, sports players and coaches often use action scenes to improve their skill and performance. However, they have to manually analyze action scenes which is expensive and time consuming. This makes Human Action Recognition (HAR) a very important area in sports as it makes the above process automatic and inexpensive.
- The task of HAR, which is the focus of this project, is to recognize which human actions are in a particular video sequence, at what time they occur, and where they are located.

1.4 Objectives

Based on the problem statement and current scenario, the objectives have been identified and formulated:

1. Since intent classification for kannada is not widely explored, we have to work on all parts of machine learning lifecycle(dataset preparation, modeling)
2. Collection of kannada questions and labeling its intent manually, to feed it to a model.
3. Building a model to identify the intent of the provided text.
4. To convert the data entries in the created dataset to suitable format to feed it to the defined model
5. Fine tuning the hyper parameters to increase the efficiency of the model.

1.5 Methodology

- Split each video into frames and do feature extraction.
- Split the data set into a training and testing set.
- Use Openpose for feature extraction which gives 17 key points.
- After preprocessing data, convert it into an tuple and pass it to RNN.
- Perform Training using LSTM-RNN on the output of the Openpose model.
- Then analyze the results of the above model using a testing set and evaluate its performance.

Chapter 2
Literature Survey

2.1 Introduction

Human Action Recognition (HAR) aims to understand human behavior and assign a label to each action. It has a wide range of applications, and therefore has been attracting increasing attention in the field of computer vision. Human actions can be represented using various data modalities, such as RGB, skeleton, depth, infrared, point cloud, event stream, audio, acceleration, radar, and WiFi signal, which encode different sources of useful yet distinct information and have various advantages depending on the application scenarios. Consequently, lots of existing works have attempted to investigate different types of approaches for HAR using various modalities.

2.2 Related work

Kun Luo wrote "Elements and construction of sports visual image action recognition system based on visual attention analysis" in 2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications.It had broad system diagram given, Common feature classification table.The sports visual image system constructed by various complex sports images contains rich visual content, highlights the charm of sports, and plays an important role in the dissemination and promotion of sports culture.Although it was not implemented only theoretical.[1]

Cheng Yan, Xin Li and Guoqiang Li wrote a paper A New Action Recognition Framework for Video Highlights Summarization in Sporting Events in 2021,The 16th International Conference on Computer Science & Education. They propose to utilize the three-level prediction model based on two open-source structures to automatically recognize players' action, and then efficiently summarize the sports highlights. They recognize player actions using both YOLO v3 and OpenPose, respectively, and compare their performance.Although the training is slow. Cannot distinguish between match relevant and match irrelevant players and errors occur when too many people in the video. [2]

Haoran Wei and Nasser Kehtarnavaz wrote a paper Simultaneous Utilization of Inertial and Video Sensing for Action Detection and Recognition in Continuous Action Streams,IEEE Sensors Journal (Volume: 20, Issue: 11, June1, 1 2020).The inertial and video data are captured simultaneously via a wearable inertial sensor and a video camera, which are turned into 2D and 3D images. These images are then fed into a 2D and a 3D CNN with their decisions fused in order to detect and recognize a specified set of actions of interest from continuous action streams.Although wearable inertial sensors need to be worn on the body.[3]

Nihat İnanç , Murat Kayri and Ömer Faruk Ertuğrul wrote "Recognition of Daily and Sports Activities",2018 IEEE International Conference on Big Data (Big Data). In this paper, the daily and sports activities dataset was employed in order to recognize action type, action, gender and also to investigate the effect of the position of the sensor node and the sensor type on the overall accuracy.ELM classification is used. Results showed that the proposed approach can be successfully employed in order to distinguish the action types. But the accuracy in detecting the actions, and gender is low. Also it was found that near similar accuracies were achieved by using the logs of a single sensor, which is ACC in Z-dimension on the right arm. Therefore, it can be noted that the performance of an action recognition system is highly dependent on the sensor type and sensor position.[4]

Youlin Song wrote "Research on Sports Image Recognition and Tracking Based on Computer Vision Technology",2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT).It is about a hybrid algorithm (S-D) is proposed by combining semi-naive Bayes classification algorithm (SNBC) and dense trajectory algorithm (DT). SNBC has a high efficiency in image recognition, but it is not accurate enough to recognize the movement of people in sports images. The research combines DT and SNBC, proposes the S-D algorithm, and builds a S-D model based on it. Results show that the S-D algorithm can effectively improve the recognition accuracy of sports images and has high practicability. However, the study doesn’t analyze the recognition effect of the S-D algorithm in a complex environment. [5]

Jianwei Li , Hainan Cui , Tianxiao Guo , Qingrui Hu and Yanfei Shen wrote a paper "Efficient fitness action analysis based on spatio-temporal feature encoding",2020 IEEE International Conference on Multimedia & amp; Expo Workshops (ICMEW). This paper proposes a simple yet efficient spatio-temporal skeleton encoding method, and designs a novel human action analysis method to recognize and evaluate fitness actions.This paper focuses on spatio-temporal skeleton encoding. First the Skeleton extraction and simplification is done then Skeleton feature encoding is performed to get the spatio-temporal skeleton image and later action analysis is performed. Further increasing the recognition accuracy of complex dynamic movements is required.[6]

Doan Yen Nhi Le wrote a paper "Analyzing the Trend of Stock Market and Evaluate the performance of Market Prediction using Machine Learning Approach" in 2022,Computational Intelligence and Neuroscience his paper gave a description about the model based on deep learning (DL) and clustering extraction algorithm. The neural network (NN) is applied to the sample set containing images of nonathletes, and the negative training sample set is iteratively enhanced according to the generated false positives, and the results are optimized by clustering method. Although the non-DL method relies too much on artificial prior knowledge, which requires a higher sports action video library.[7]

Mohammad Ashraf Russo, Laksono Kurnianggoro, Kang-Hyun Jo wrote "Classification of sports videos with combination of deep learning models and transfer learning",2019 International Conference on Electrical, Computer and Communication Engineering.In the project CNN extracted features are combined with temporal information from RNN to formulate the general model to solve the problem. Initially, they make a small dataset consisting of only 5 sports classes-Basketball, Cricket, Football, Ice Hockey, Tennis. Each class contains 60 sequences in total and each sequence contains 64 sequential frames. Later dataset is scaled to 10 classes. In the recurrent part, the gated recurrent units (GRU) were used. After CNN and RNN, transfer learning model VGG-16 is used. When scaling up the dataset for larger classification tasks, such as 15 classes which have ambiguity in themselves, it becomes quite difficult.There is still much room for improvement, other deep learning based techniques that could be applied for such task are data augmentation, fine tuning to adopt weights more to the specific dataset.[8]

Tsuyoshi Masuda , Ren Togo , Takahiro Ogawa and Miki Haseyama wrote "Sports Action Detection Based on Self-Supervised Feature Learning and Object Detection",2021 IEEE 10th Global Conference on Consumer Electronics (GCCE). The proposed method realizes action detection without a fine grained annotation based on self-supervised feature learning and object detection. The self-supervised feature learning works well when a single person is on a video. Thus, we introduce object detection into our method to achieve action detection for multiple persons by tracking each person.[9]

Kamal Kant Verma, Brij Mohan Singh wrote "Vision based Human Activity Recognition using Deep Transfer Learning and Support Vector Machine", 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON).The paper explains feature extraction and representation has been performed using a fine-tuned VGG-19 model. Then, a non-linear multi-class support vector machine is used to classify the extracted features into the activity classes.Dataset: UCF Sports Action dataset.Although the number of frames in each sports video sequence varies from 62 to 387 for UCF Sports Action dataset therefore, some loss of information exists.The total false positive rate (error) of the proposed approach is 0.291. The main cause of this error is due to the three sports activities namely Running, Skateboarding and Walking. The reason being is that the activities Running, Skate-boarding and Walking in the confusion matrix have higher confusion as compared to the other activities of the dataset.[10]

2.3 Summary

Everyone has implemented the model using different techniques, CNN, transfer learning techniques, VGG-19 architectures, deep learning (DL) and clustering extraction algorithms, efficient spatio-temporal skeleton encoding method and many more. Most efficient method is using CNN and RNN (CNN extracted features are combined with temporal information from RNN to formulate the general model to solve the problem).

Chapter 3
Software Requirements

3.1 Functional Requirements

- Collection of data from different sources
- Removing unwanted data and formatting it according to the needs of the model.
- Feature extraction from openpose
- Pre-process data into an array and pass it to RNN

3.2. Non-Functional Requirements

- Reliability: The system is highly reliable.
- Accessibility: It can be easily accessible i.e click & run.
- Efficiency: Resource consumption for a given load is quite low.
- Robustness: Our system is not capable of coping with errors during execution.
- Scalability: Our project is scalable i.e. we can add more resources to our project without disturbing the current scenario

3.3. Hardware Requirements

- Central Processing Unit (CPU) — Intel Core i5 6th Generation processor or higher. An AMD equivalent processor will also be optimal.
- RAM — 8 GB minimum, 16 GB or higher is recommended.
- Most machine learning techniques that require more than 16GB of RAM now leverage cloud computing to speed up processing.
- Graphics Processing Unit (GPU) — NVIDIA GeForce GTX 960 or higher. AMD GPUs are not able to perform deep learning regardless.
- Operating System — Ubuntu or Microsoft Windows 10. It is recommended to update Windows 10 to the latest version before proceeding forward

3.4. Software Requirements

- Operating system- Windows 7 or above is used as the operating system as it is stable and supports more features and is more user friendly
- Environment - Jupyter Notebook and Google Colab are used to implement the project.
- Libraries : Pandas, Numpy, Matplotlib, Seaborn, Tensorflow, Keras.
- Language: Python(3.6 version and higher)
- IDE: Google colab for online development, Jupyter Notebook for offline development.
- Jupyter Notebook and google colab for executing the machine learning models flexible interface allows users to configure and arrange workflows in data science, scientific computing, computational journalism, and machine learning.

3.5. Summary

This project requires an i5 6th generation processor with operating system of windows 7 and above. It uses various python libraries like pandas and numpy.It collects data from various sources and processes the data.

Chapter 4
Design of Artificial Intelligence in motion

4.1. High Level design

In the Fig 4.1, the video is split into frames passed to the openpose platform, where feature extraction takes place.Further the data is divided into two parts: the training set and testing set. In the training set the preprocessed data is passed to the LSTM-RNN where accuracy is increased by repeated iterations.finally the testing set is used to evaluate the performance of the model.

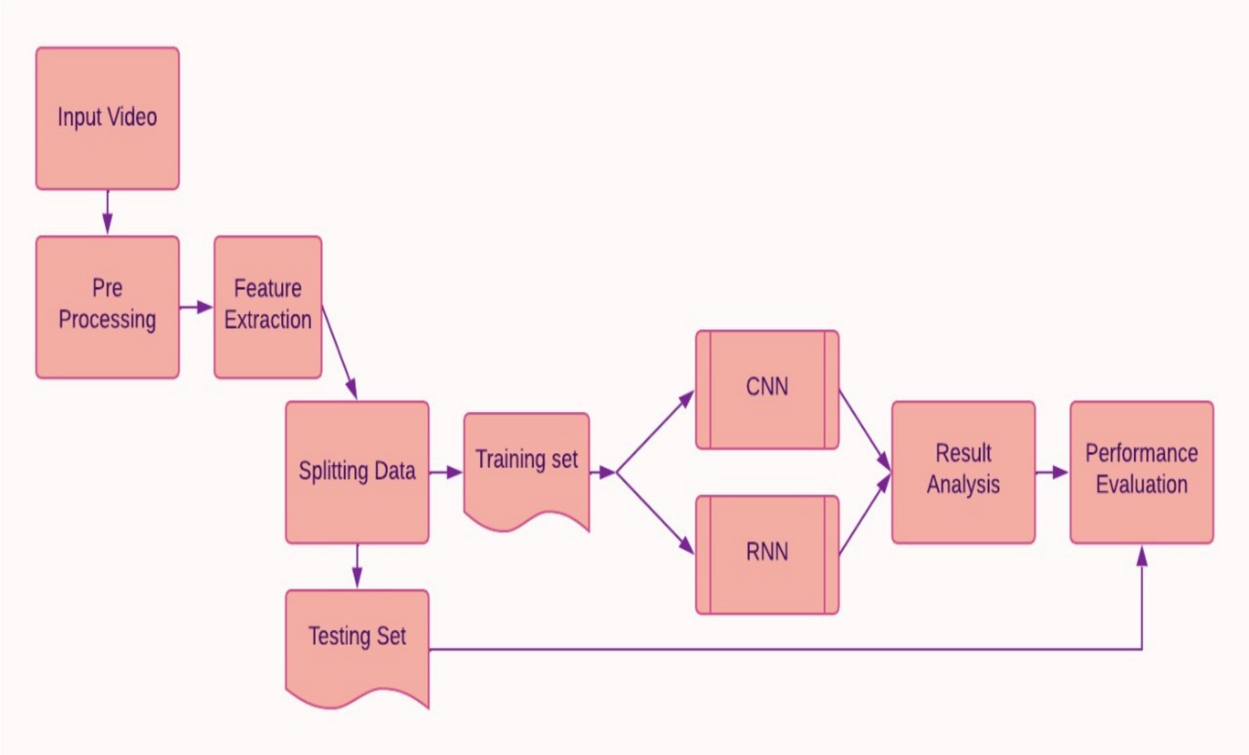


Figure 4.1: Showing the high level structural design of the model

4.1.1. System Architecture

In Fig 4.2, the OpenPose has represented the first real-time multi-person system to jointly detect human body, hand, facial, and foot keypoints (in total 135 keypoints) on single images.

Input: Image, video, webcam, Flir/Point Gray, IP camera, and support to add your own custom input source (e.g., depth camera).

Output: Basic image + keypoint display/saving (PNG, JPG, AVI, ...), keypoint saving (JSON, XML, YML, ...), key points as array class, and support to add your own custom output code (e.g., some fancy UI).

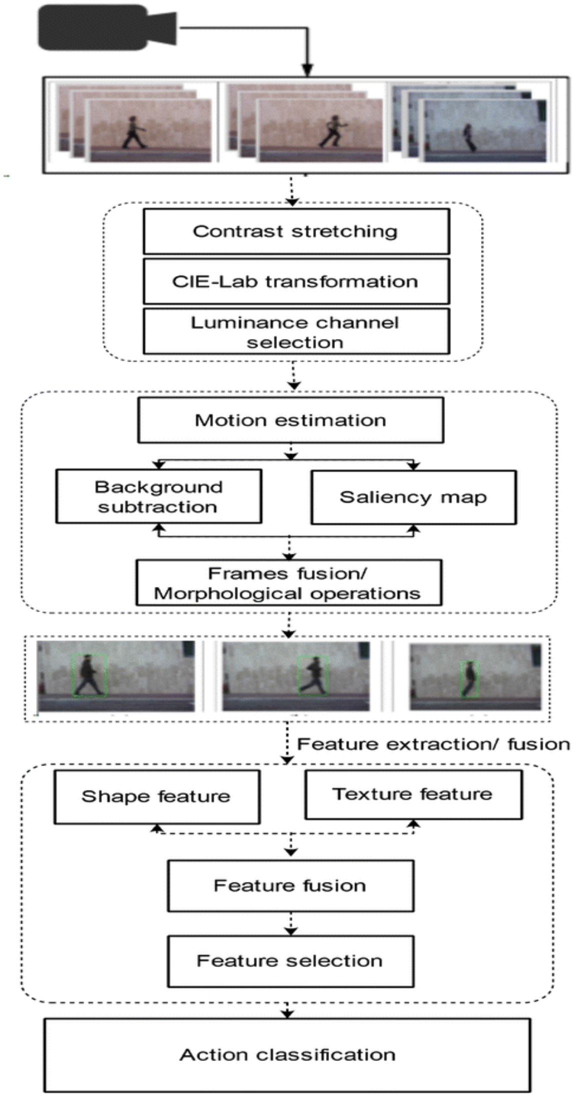


Figure 4.2: Shows the detailed data flow through the model

4.3. Summary

In short the dataset is processed by CNN to extract features and the preprocessed data is converted into an array.further the data is passed through LSTM-RNN where we train the model and increase our accuracy based on the number of epochs.Then we test the model by remaining part of our data and evaluate the accuracy. We also use the optimization function to slightly increase the final accuracy

Chapter 5
Implementation of Artificial Intelligence in HAR

5.1. Programming Language Selection

Python is an interpreted, object-oriented, high-level programming language with dynamic semantics. It’s easy to learn syntax emphasizes readability and therefore reduces the cost of program maintenance. Python supports modules and packages, which encourages program modularity and code reuse.To reduce development time, programmers turn to a number of Python frameworks and libraries. Python, with its rich technology stack, has an extensive set of libraries for artificial intelligence and machine learning. Here are some of them:

- Keras, Tensorflow and SciKit Learn for machine learning.
- Numpy for data analytics.
- SciPy for advanced computing.
- Pandas for general purpose data analysis.

Python’s popularity is that it’s a platform independent language. Python is supported by many platforms including Linux, Windows, and macOS. The numpy package is used to do the computation faster, for example, matrix multiplication. The pandas are used for data analysis and also for importing csv files and other file formats to perform machine learning techniques. The scikit learn is the most popular library used for ML, which makes the coding simpler and has a wide variety of functions.

5.2. Platform Selection

For this Project, Since Python is our only language to run Machine learning modules, we are using google colab, which allows developers to write and execute Python code through their browser. Google Colab is an excellent tool for deep learning tasks. It is a hosted Jupyter notebook that requires no setup and has an excellent free version, which gives free access to Google computing resources such as GPUs.

There are several reasons to opt to use Google Colab instead of a plain Jupyter Notebook instance:

- Pre-Installed Libraries
- Saved on the Cloud
- Collaboration
- Free GPU Use

5.3. Summary

We have used Google colab platform because of its excellent performance for deep learning tasks. It supports numpy, pandas and various other libraries/packages and also provides use of GPU to some extent which is not available in jupyter notebook. Python is the most popular language for implementing ML and deep learning models and we will be using python for implementation.

Chapter 6
Experiment results and testing of Artificial Intelligence in HAR

6.1. Evaluation Metrics

The evaluation metric used to evaluate our model for classification is accuracy. For calculating the efficiency of the model, our job is to check, for how many sentences in the test data the model has given the correct intent label. This can be achieved with the help of already defined metric **Accuracy, Precision, Recall, F-score.**

Accuracy is an evaluation metric that allows you to measure the total number of predictions a model gets right. A confusion matrix displays counts of the True Positives, False Positives, True Negatives, and False Negatives produced by a model. Accuracy can be calculated with the help of confusion matrix as shown in the eq(1):

$$Accuracy = \frac{TN+TP}{TP+FP+TN+FN} \qquad \ldots .(1)$$

Accuracy can also be directly calculated as shown in eq(2):

$$Accuracy = \frac{Total\ no.\ of\ corrected\ predictions}{Total\ no.\ of\ predictions} \qquad \ldots .(2)$$

To compute the efficiency of our model we are using the later equation.
The obtained dataset is fed into the LSTM model which gives the word embeddings as output. These word embeddings are the feature vectors, used to determine the intent of a sentence. This word embedding is sent to a multi-class classification model which outputs the category of intent it belongs to.

Precision evaluates how precise a model is in predicting positive labels. The top of the formula is the number of positive observations that a model predicted correctly. The denominator is the number of times the model predicted a positive label in total. Precision is a good evaluation metric to use when the cost of a false positive is very high and the cost of a false negative is low.

The formula is below as shown in eq(3):

$$Precision = \frac{TP}{TP+FP} \qquad \ldots(3)$$

Recall calculates the percentage of actual positives a model correctly identifies (True Positive). When the cost of a false negative is high, you should use recall. The numerator is the number of true positives or the number of positives the model correctly identified. The denominator is the number of actual positives predicted by the model and the number of positives incorrectly predicted as negative by the model. The formula for recall is below as shown in eq(4):

$$Recall = \frac{TP}{TP+FN} \qquad \ldots(4)$$

The **F-score**, also called the F1-score, is a measure of a model's accuracy on a dataset. It is used to evaluate binary classification systems, which classify examples into 'positive' or 'negative'. The F-score is a way of combining the precision and recall of the model, and it is defined as the harmonic mean of the model's precision and recall. The F-score is commonly used for evaluating information retrieval systems such as search engines, and also for many kinds of machine learning models, in particular in natural language processing. The formula as shown in eq(5):

$$F1 = \frac{TP}{TP+\frac{1}{2}(FP+FN)} \qquad \ldots(5)$$

6.2. Experimental Dataset

The dataset that has been used is the Berkley MHAD. The Berkeley Multimodal Human Action Database (MHAD) contains 11 actions performed by 7 male and 5 female subjects in the range 23-30 years of age except for one elderly subject. All the subjects performed 5 repetitions of each action, yielding about 660 action sequences which correspond to about 82 minutes of total recording time.

6.3. Performance Analysis

The performance of the dataset is evaluated on the basis of how it predicts the label.The dataset is divided into 80% for training data and 20% for testing data. After training the model on a training dataset , the model is evaluated against the testing data. The model predicts the labels based on given values. In order to calculate the accuracy of the model on how it is performing, the predicted values are compared with the test values and categorized into True positive, True negative, False positive and False Negative. Using the formulae given in above section, the Accuracy, precision, recall and F1 score calculated are shown in the Table 5.1:

Table 5.1: Table shows how the model performs on different metrics

Metrics	**Efficiency(%)**
Precision	95.500332335475%
Recall	95.44427056164146%
Accuracy	95.44426798820496%
F1-score	95.45884537525906%

In the RNN deep learning model, for each iteration, the data is given to the model and the model tries to improve by reducing the loss every iteration by optimizing the values based on the given loss function and optimization function. In the Fig 5.1, the accuracy , loss after epoch is calculated and depicted as follows

A confusion matrix is a table that is used to define the performance of a classification algorithm. Confusion matrix visualizes and summarizes the performance of a classification algorithm. The confusion matrix for the model is shown in Fig 5.2

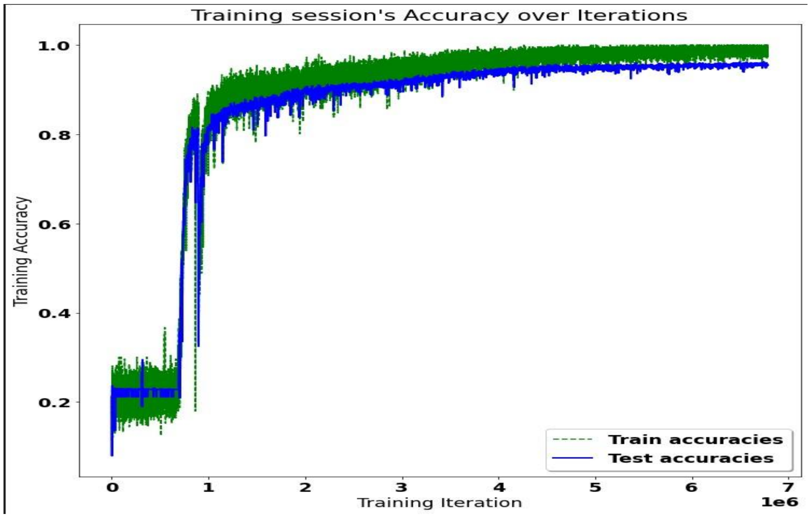
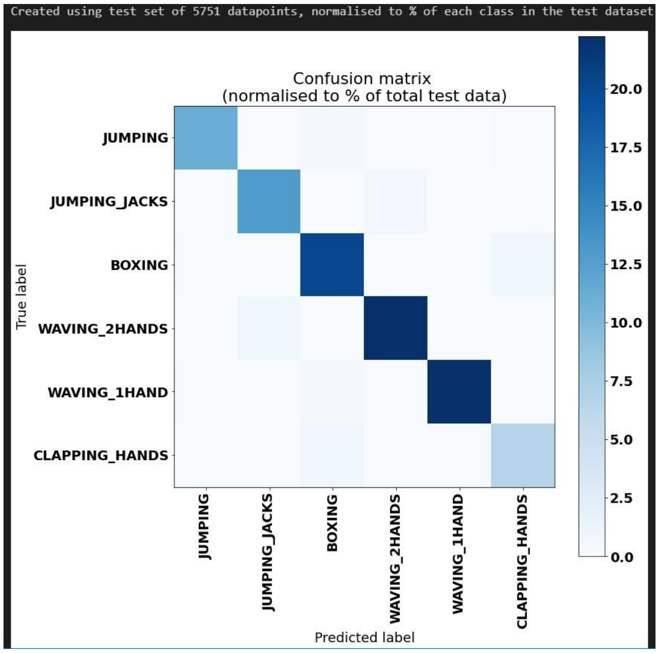


Figure 5.1: Accuracy vs iteration graph



Created using test set of 5751 datapoints, normalised to % of each class in the test dataset

Figure 5.2: Confusion Matrix

6.4 Summary

The accuracy of the model on training data is 98% and on the test set it is 95%. This accuracy is possible due to the availability of good datasets and the research that has been done in this field.

Chapter 7
Conclusion and Future Enhancement

7.1. Limitations of the Project

The dataset has not been normalized and data augmentation has not been done in this project. Furthermore, it is evident that there exists a great need for efficiently manipulating training data that may come from heterogeneous sources. The number and type of different modalities that can be used for analyzing human activities is an important question. The combination of multimodal features, such as body motion features, facial expressions, and the intensity level of voice, may produce superior results, when compared to unimodal approaches, On the other hand, such a combination may constitute over-complete examples that can be confusing and misleading. The proposed multimodal feature fusion techniques do not incorporate the special characteristics of each modality and the level of abstraction for fusing. Therefore, a comprehensive evaluation of feature fusion methods that retain the feature coupling is an issue that needs to be assessed.

7.2. Conclusion

Recognizing human activities from video sequences or still images is a challenging task due to problems, such as background clutter, partial occlusion, changes in scale, viewpoint, lighting, and appearance. Many applications, including video surveillance systems, human-computer interaction, and robotics for human behavior characterization, require a multiple activity recognition system. In this work, we provide a detailed review of recent and state-of-the-art research advances in the field of human activity classification. We propose a categorization of human activity methodologies and discuss their advantages and limitations. In particular, we divide human activity classification methods into two large categories according to whether they use data from different modalities or not. Then, each of these categories is further analyzed into sub-categories, which reflect how they model human activities and what type of activities they are interested in. Moreover, we provide a comprehensive analysis of the existing, publicly available human activity classification datasets and examine the requirements for an ideal human activity recognition dataset. Finally, we report the characteristics of future research directions and present some open issues on human activity recognition.

7.3. Future Enhancements

The dataset has not been normalized and data augmentation has not been done in this project. Furthermore, it is evident that there exists a great need for efficiently manipulating training data that may come from heterogeneous sources. The number and type of different modalities that can be used for analyzing human activities is an important question. The combination of multimodal features, such as body motion features, facial expressions, and the intensity level of voice, may produce superior results, when compared to unimodal approaches, On the other hand, such a combination may constitute over-complete examples that can be confusing and misleading.

7.4. Summary

A significant gap exists in our knowledge of how domain-specific feature extraction compares to unsupervised feature learning in the latent space of a deep neural network for a range of temporal applications including human activity recognition (HAR).

References

[1] Kun Luo, Elements and construction of sports visual image action recognition system based on visual attention analysis,2021 IEEE International Conference on Advances in Electrical Engineering and Computer Applications.

[2] Cheng Yan; Xin Li; Guoqiang Li, 2022,Computational Intelligence and Neuroscience,2021,The 16th International Conference on Computer Science & Education

[3] Haoran Wei; Nasser Kehtarnavaz,IEEE Sensors Journal (Volume: 20, Issue: 11, June1, 1 2020),

[4] Nihat İnanç , Murat Kayri, Ömer Faruk Ertuğrul, Recognition of Daily and Sports Activities,2018 IEEE International Conference on Big Data (Big Data),.

[5] Youlin Song, Research on Sports Image Recognition and Tracking Based on Computer Vision Technology,2021 5th Asian Conference on Artificial Intelligence Technology (ACAIT)

[6] Jianwei Li , Hainan Cui , Tianxiao Guo , Qingrui Hu and Yanfei Shen , Efficient fitness action analysis based on spatio-temporal feature encoding ,2020 IEEE International Conference on Multimedia & amp; Expo Workshops (ICMEW).

[7]Analyzing the Trend of Stock Market and Evaluate the performance of Market Prediction using Machine Learning Approach,2022,Computational Intelligence and Neuroscience

[8] Mohammad Ashraf Russo; Laksono Kurnianggoro; Kang-Hyun Jo, Classification of sports videos with combination of deep learning models and transfer learning, 2019 International Conference on Electrical, Computer and Communication Engineering.

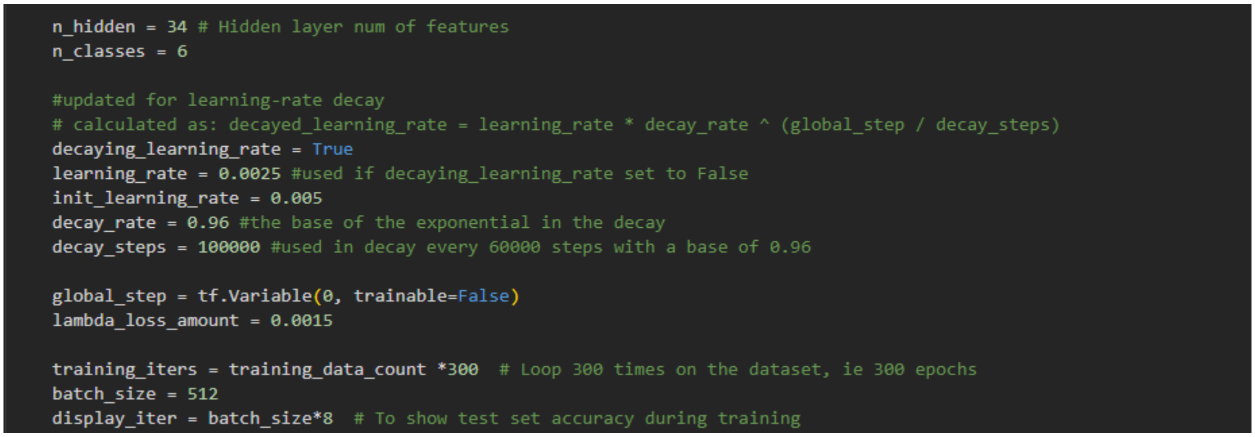
[9] Tsuyoshi Masuda , Ren Togo , Takahiro Ogawa† and Miki Haseyama, Sports Action Detection Based on Self-Supervised Feature Learning and Object Detection,2021 IEEE 10th Global Conference on Consumer Electronics (GCCE).

[10]Kamal Kant Verma, Brij Mohan Singh ,"Vision based Human Activity Recognition using Deep Transfer Learning and Support Vector Machine", 2021 IEEE 8th Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON).

Appendices

Appendix 1: Screenshots

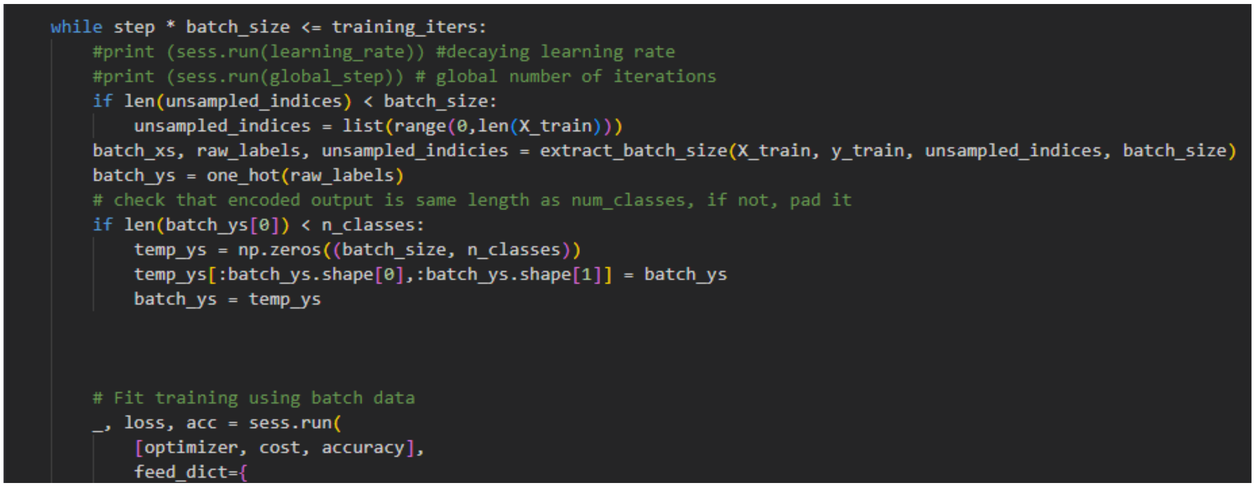
Hyperparameters are parameters whose values control the learning process and determine the values of model parameters that a learning algorithm ends up learning. In the Fig A1.1 below shows the hyperparameters passes to the LSTM model



```
n_hidden = 34 # Hidden layer num of features
n_classes = 6

#updated for learning-rate decay
# calculated as: decayed_learning_rate = learning_rate * decay_rate ^ (global_step / decay_steps)
decaying_learning_rate = True
learning_rate = 0.0025 #used if decaying_learning_rate set to False
init_learning_rate = 0.005
decay_rate = 0.96 #the base of the exponential in the decay
decay_steps = 100000 #used in decay every 60000 steps with a base of 0.96

global_step = tf.Variable(0, trainable=False)
lambda_loss_amount = 0.0015

training_iters = training_data_count *300  # Loop 300 times on the dataset, ie 300 epochs
batch_size = 512
display_iter = batch_size*8  # To show test set accuracy during training
```

Figure A1.1.Shows the code snippet of LSTM hyperparameters

It is a variety of recurrent neural networks (RNNs) that are capable of learning long-term dependencies, especially in sequence prediction problems. LSTM has feedback connections, i.e., it is capable of processing the entire sequence of data, apart from single data points such as images.In the Fig A1.2 shows the LSTM training on the training set for 300 epochs.



```
while step * batch_size <= training_iters:
    #print (sess.run(learning_rate)) #decaying learning rate
    #print (sess.run(global_step)) # global number of iterations
    if len(unsampled_indices) < batch_size:
        unsampled_indices = list(range(0,len(X_train)))
    batch_xs, raw_labels, unsampled_indicies = extract_batch_size(X_train, y_train, unsampled_indices, batch_size)
    batch_ys = one_hot(raw_labels)
    # check that encoded output is same length as num_classes, if not, pad it
    if len(batch_ys[0]) < n_classes:
        temp_ys = np.zeros((batch_size, n_classes))
        temp_ys[:batch_ys.shape[0],:batch_ys.shape[1]] = batch_ys
        batch_ys = temp_ys



    # Fit training using batch data
    _, loss, acc = sess.run(
        [optimizer, cost, accuracy],
        feed_dict={
```

Figure A1.2.Shows the code snippet of the model LSTM