A Major Project (Stage I) Report

on

Human Action Imitation using Gait Classifier

Submitted in partial fulfillment of the requirements

for the award of degree of

BACHELOR OF TECHNOLOGY

in

Information Technology

by

Akshitha Reddy N(18WH1A1234)

Srivalli CH (18WH1A1238)

Mahalakshmi T (18WH1A1244)

Under the esteemed guidance of

Dr. P S Latha Kalyampudi

Associate Professor

Department of Information Technology

BVRIT HYDERABAD College of Engineering for Women

Rajiv Gandhi Nagar, Nizampet Road, Bachupally, Hyderabad – 500090 (Affiliated to Jawaharlal Nehru Technological University Hyderabad)

(NAAC 'A' Grade & NBA Accredited- ECE, EEE, CSE IT)

January, 2022

DECLARATION

We hereby declare that the work presented in this project entitled “Human Action Imitation using Gait Classifier” submitted towards completion of the project in IV year I sem of B.Tech IT at “BVRIT HYDERABAD College of Engineering for Women”, Hyderabad is an authentic record of our original work carried out under the esteem guidance of Dr. P S Latha Kalyampudi, Associate Professor, Department of IT.

Akshitha Reddy N (18WH1A1234)

Srivalli CH (18WH1A1238)

Mahalakshmi T (18WH1A1244)

VISHNU
UNIVERSAL LEARNING

BVRIT HYDERABAD

College of Engineering for Women

Rajiv Gandhi Nagar, Nizampet Road, Bachupally, Hyderabad – 500090

(Affiliated to Jawaharlal Nehru Technological University Hyderabad)

(NAAC 'A' Grade & NBA Accredited- ECE, EEE, CSE IT)

CERTIFICATE

This is to certify that the major project (Stage I) report on **"Human Action Imitation using Gait Classifier"** is a bonafide work carried out by **Akshitha Reddy N (18WH1A1234), Srivalli CH (18WH1A1238)** and **Mahalakshmi T (18WH1A1244)** in the partial fulfillment for the award of B.Tech degree in **Information Technology, BVRIT HYDERABAD College of Engineering for Women, Bachupally, Hyderabad** affiliated to Jawaharlal Nehru Technological University, Hyderabad under my guidance and supervision.The results embodied in the major project (Stage I) work have not been submitted to any other university or institute for the award of any degree or diploma.

Internal Guide
Dr. P S Latha Kalyampudi
Associate Professor
Department of IT

Head of the Department
Dr. Aruna Rao S L
Professor & HoD
Department of IT

ACKNOWLEDGEMENT

We would like to express our profound gartitude and thanks to **Dr. K. V. N. Sunitha, Principal, BVRIT HYDERABAD college of Engineering for Women** for providing the working facilities in the college.

Our sincere thanks and gratitude to **Dr. Aruna Rao S L, Professor & Head, Department of IT, BVRIT HYDERABAD college of Engineering for Women** for all the timely support, constant guidance and valuable suggestions during the period of our project.

We are extremely thankful and indebted to our internal guide, **Dr. P S Latha Klayampudi, Associate Professor, Department of IT, BVRIT HYDERABAD college of Engineering for Women** for her constant guidance, encouragement, and moral support throughout the project.

Finally, we would also like to thank our Project Coordinators **Dr. P Kayal, Associate Professor, Dr. P S Latha Kalyampudi, Associate Professor**, all the faculty and staff of the Department of IT who helped us directly or indirectly, parents and friends for their cooperation in completing the major project(Stage I) work.

Akshitha Reddy N (18WH1A1234)

Srivalli CH (18WH1A1238)

Mahalakshmi T (18WH1A1244)

ABSTRACT

Human action imitation is a computer vision problem that tries to estimate human body joints location and decide how they are connected to each other. Human action imitation is a way of retrieving videos emerged from Content Based Video Retrieval (CBVR). Human action imitation has gained popularity because of its wide applicability in automatic retrieval of videos of particular action using visual features. The most common Stages for action recognition includes: object and human segmentation, feature extraction, activity detection and classification. Human action imitation has a strong theoretical significance and wide application prospect in the field of video surveillance, human-computer interaction, virtual reality, etc. To imitate the actions of the person Gait Classifier mechansim and DBSCAN(Density based spatial clustering application with Noise) is used. Gait analysis is widely used in clinical practice to help in understanding the gait actions where as DBSCAN is perfomed to improve the performance of clustering. This work present an approach of Human Action Imitaion using Gait Calssifier, where the input video is given to 3D CNN and VGG16 to train the video. 3D CNN gave more accuracy than VGG16, where 3D CNN gave 99.3 % accuracy and VGG16 gave 58.1 % accuracy in the Stage I. So, the video which is trained with CNN is classified with gait mechanism to get the correct imitation of the action of the given person.

LIST OF FIGURES

Figure No.	Figure Name	Page No.
1.1	Human Pose	3
2.1	Output of Real-Time 3D Human Pose	4
2.2	Output of Detection and Tracking for Human Pose	5
2.3	Output of Real-Time Gait Analysis	6
2.4	Comparison Survey on Human action recognition	8
3.1	Use-Case Diagram	9
3.2	Sequence Diagram	10
4.1	Block diagram of proposed system	12
4.2	Action Imiation of the Person	14
4.3	The tendency of clustering according to Eps.	15
4.4	The tendency of clustering according to minPts.	16
4.5	VGG16 Network Architecture	16
4.6	3D covolution	17
4.7	A 3D CNN architecture for human action imitation	18
5.1	Key Points Recognition	20
5.2	Human action Imitation 1	21
5.3	Human action Imitation 2	21
5.4	Accuracy of comparisons	22
6.1	Imitation on the person	23
6.2	Imitation beside the person	23

CONTENTS

TOPIC	PAGE NO.
ABSTRACT	**V**
LIST OF FIGURES	**VI**
1. Introduction	1
1.1 Objective	2
1.2 Problem Statement	3
2. Literature Survey	
2.1 Real-Time 3D Human Pose Estimation and Action Recognition	4
2.2 Combining Detection and Tracking for Human Pose Estimation	5
2.3 Real-Time Gait Analysis Using Convolutional Neural Networks	6
2.4 Multi-modality sensor fusion for gait classifiaction	7
2.5 DBSCAN Clustering Algorithm Based on Density	7
2.6 A survey on Human action recognition from videos	7
3. System Design	
3.1 UML Diagrams	9
3.1.1 Use-Case Diagrams	9
3.1.2 Sequence Diagrams	10
3.2 S/W & H/W Requirements	
3.2.1 H/W Requirements	11
3.2.2 S/W Requirements	11
4. Methodology	
4.1 Architecture	12
4.2 Modules	
4.2.1 Pre-Processing	13
4.2.2 Feature Extraction	13

4.2.3 Pose Detection 14
4.3 Algorithms
4.3.1 VGG16 (Visual Geometry Group 16) 16
4.3.2 CNN & 3D CNN 17
5. Parital Implementation & Results 20
6. Conclusion & Extension Plan for Stage II 23
References 24

1. Introduction & Background

Human action imitation 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. 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.

When attempting to recognize human activities, one must determine the kinetic states of a person, so that the computer can efficiently recognize the activity. Human activities, such as "walking" and "running," arise very naturally in daily life and are relatively easy to recognize. On the other hand, more complex activities, such as "peeling an apple," are more difficult to identify. Complex activities may be decomposed into other simpler activities, which are generally easier to recognize. Usually, the detection of objects in a scene may help to better understand human activities as it may provide useful information about the ongoing event

Human activities have an inherent hierarchical structure that indicates the different levels of it, which can be considered as a three-level categorization. First, for the bottom level, there is an atomic element and these action primitives constitute more complex human activities. After the action primitive level, the action/activity comes as the second level. Finally, the complex interactions form the top level, which refers to the human activities that involve more than two persons and objects. In this paper, we follow this three-level categorization namely action primitives, actions/activities, and interactions. This three-level categorization varies a little from previous surveys and maintains a consistent theme. Action primitives are those atomic actions at the limb level, such as "stretching the left arm," and "raising the right leg." Atomic actions are performed by a specific part of the human body, such as the hands, arms, or upper body part. Actions and activities are used interchangeably in this review, referring to the whole-body movements composed of several action primitives in temporal sequential order and performed by a single person with no more person or additional objects. Specifically, we refer the terminology human activities as all movements of the three layers and the activities/actions as the middle level of human activities. Human activities like walking, running, and waving hands are categorized in the actions/activities level. Interactions are human activities that involve two or more

persons and objects. The additional person or object is an important characteristic of interaction. Typical examples of interactions are cooking which involves one person and various pots and pans and kissing that is performed by two persons.

Gestures are considered as primitive movements of the body parts of a person that may correspond to a particular action of this person. Atomic actions are movements of a person describing a certain motion that may be part of more complex activities. Human-to-object or human-to-human interactions are human activities that involve two or more persons or objects. Group actions are activities performed by a group or persons. Human behaviors refer to physical actions that are associated with the emotions, personality, and psychological state of the individual. Finally, events are high-level activities that describe social actions between individuals and indicate the intention or the social role of a person.

Most of the work in human activity recognition assumes a figure-centric scene of uncluttered background, where the actor is free to perform an activity. The development of a fully automated human activity recognition system, capable of classifying a person's activities with low error, is a challenging task due to problems, such as background clutter, partial occlusion, changes in scale, viewpoint, lighting and appearance, and frame resolution. In addition, annotating behavioral roles is time consuming and requires knowledge of the specific event. Moreover, intra- and interclass similarities make the problem amply challenging. That is, actions within the same class may be expressed by different people with different body movements, and actions between different classes may be difficult to distinguish as they may be represented by similar information. The way that humans perform an activity depends on their habits, and this makes the problem of identifying the underlying activity quite difficult to determine.

1.1 Objective

Human action imitation 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. Depending on their complexity, human activities are categorized into:

1. gestures
2. atomic actions
3. human-to-object or human-to-human interactions
4. group actions
5. behaviors
6. events

1.2 Problem Statement

Human action imitation is one of the key problems in computer vision that has been studied for well over 15 years. The reason for its importance is the abundance of applications that can benefit from such a technology. Human action imitation or HAI for short, is the problem of predicting what a person is doing based on a trace of their movement by imitating their actions using Gait Classifier.

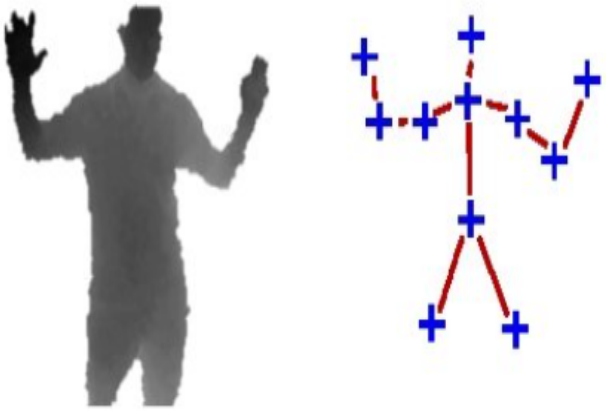

Figure 1.1: Human Pose

2. Literature Survey

2.1 Real-Time 3D Human Pose Estimation and Action Recognition

In Diogo C. Luvizon[1], In Multi-Task Deep Learning for Real-Time 3D Human Pose Estimation and Action Recognition, they have presented a new approach for human pose estimation and action recognition using multi-task deep learning. The proposed method for 3D pose provides highly precise estimations with low resolution feature maps and departs from requiring the expensive volumetric heat maps by predicting specialized depth maps per body joints. The proposed CNN architecture, along with the pose regres- sion method, allows multi-scale pose and action supervision and re-injection, resulting in a highly efficient densely su- pervised approach. The method can be trained with mixed 2D and 3D data, benefiting from precise indoor 3D data, as well as “in-the-wild” images manually annotated with 2D poses. This has demonstrated significant improvementsfor 3D pose estimation. The proposed method can also be trained with single frames and video clips simultaneouslyand in a seamless way.

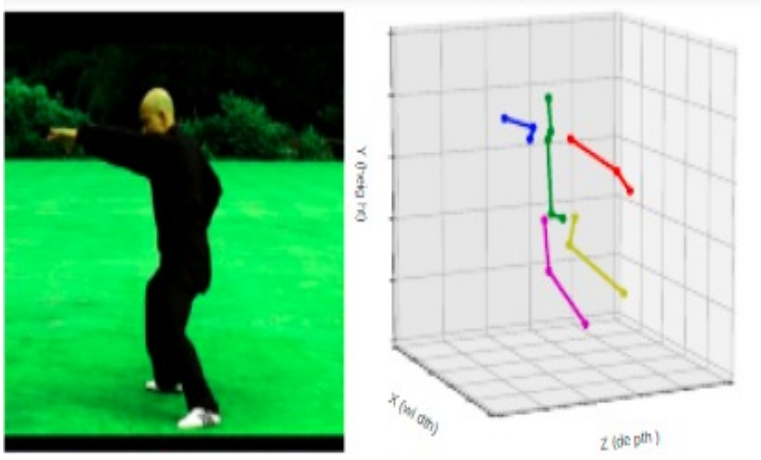

Figure 2.1: Output of Real-Time 3D Human Pose

Predicted 3D poses from RGB images for both 2D and 3D datasets as shown in Figure 2.1 using CNN.

2.2 Combining Detection and Tracking for Human Pose Estimation

In J. Tighe, D. Modolo[2], Combining Detection and Tracking for Human Pose Estimation in Videos, they have presented a novel top-down approach for multiperson pose estimation and tracking in videos. The approach can recover from failures of its person detector by propagating known person locations through time and by searching for poses in them. The approach consists of three components. Clip Tracking Network was used to jointly perform joint pose estimation and tracking on small video clips. Then, Video Tracking Pipeline was used to merge tracklets predicted by Clip Tracking Network, when these belonged to the same person. Finally, Spatial-Temporal Merging was used to refine the joint locations based on a spatial-temporal consensus procedure over multiple detections for the same person. They showed that this approach is capable of correctly predicting people poses, even on very hard scenes containing severe occlusion and entanglements. Finally, they showed the straight of our approach by achieving state-of-the-art results on both joint detection and tracking, on both the PoseTrack 2017 and 2018 datasets, and against all top-down and bottom-down approaches.

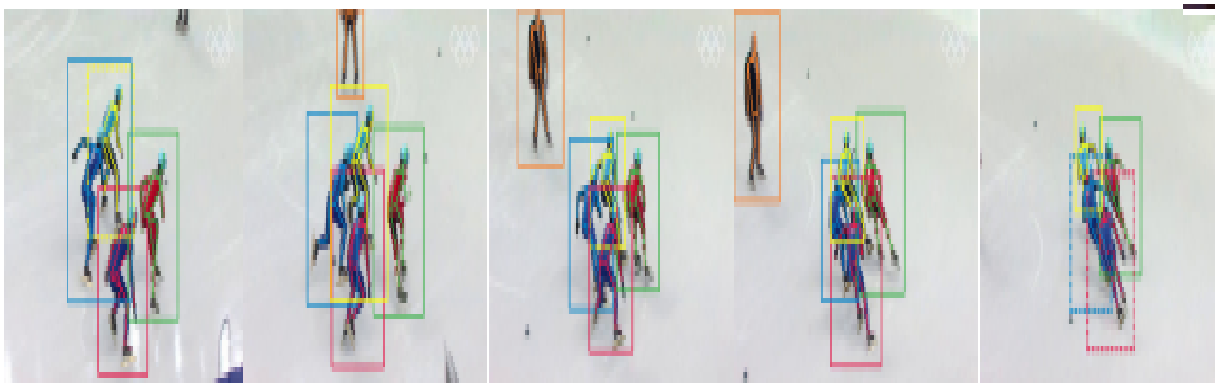

Figure 2.2: Output of Detection and Tracking for Human Pose

Visualization of the output of approach on videos from the PoseTrack dataset. Bounding boxes and poses are color coded using the track id predicted by the model. Solid bounding boxes indicate that the instance was localized by the person detector, while dotted bounding boxes were originally missed by the detector as shown in Figure 2.2, but recovered by this approach.

2.3 Real-Time Gait Analysis Using Convolutional Neural Networks

In Ali Rohan[3], In Human Pose Estimation-Based Real-Time Gait Analysis Using Convolutional Neural Networks, an approach to develop an efficient gait analysis mechanism using deep learning tools such as Convolutional Neural Network (CNN)-based classifier is presented. The proposed approach uses the human pose estimation method to classify the abnormalities found in a specific person's gait. By introducing a proper data collection and training scheme for CNN, the proposed approach addresses the problems related to gait analysis methods used previously. The experimental results are promising in solving some of the typical problem constraints involved in the gait analysis system. The accuracy achieved for classifying normal and abnormal gait is 97.3 % that proves the applicability of the proposed mechanism. Furthermore, in the future, the system can be used to distinguish different people's gait by adding more data in the training process of the classifier.

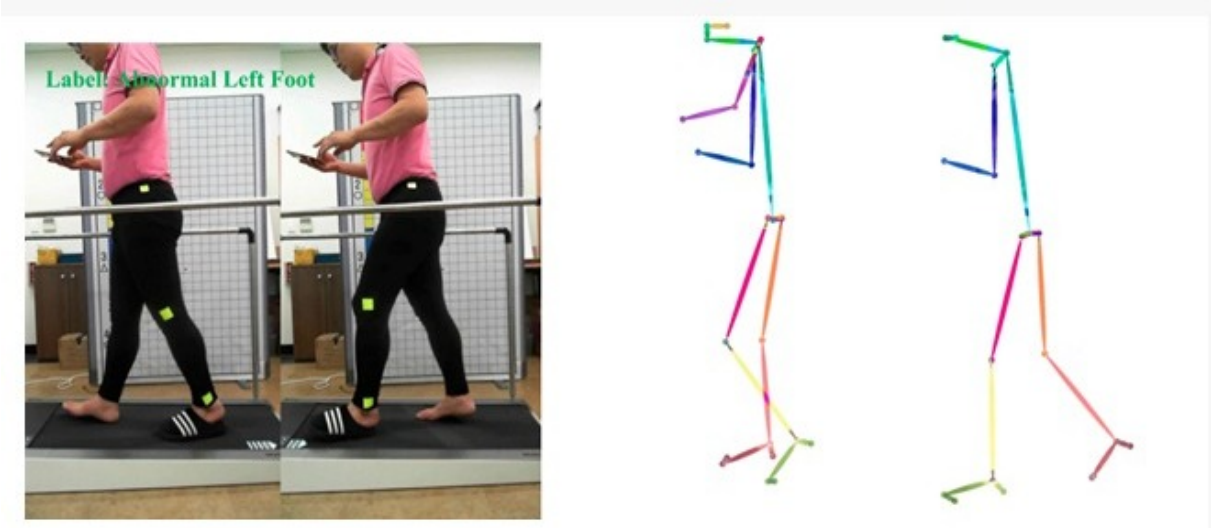

Figure 2.3: Output of Real-Time Gait Analysis

Example of the recorded data for category Abnormal Left Foot and its extracted skeletal image using pose estimation as shown in the Figure 2.3.

2.4 Multi-modality sensor fusion for gait classifiaction

In Yunas, Syed Usama[4], Multi-modality sensor fusion for gait classifiaction is evident that CNN outperforms ANN in terms of lesser classification time and higher accuracies. Feeding input data as 2-D arrays to CNN architecture and fully connected merging strategy is able to extract gait features in a robust manner. Therefore our multi-modality fusion yields the best results positioning itself as a candidate to set a new state-of-the-art in the case under study. This can be taken up further by expanding the dataset with a larger number of subjects, varying the number/position of the AIS sensors, as well as defining additional manners of walking.

2.5 DBSCAN Clustering Algorithm Based on Density

In Dingsheng Deng[5], DBSCAN Clustering Algorithm Based on Density, the paper discussed about DBSCAN. In the era of big data, human electric interaction is transformed into a series of data, which contains great value. Machine Learning has shown excellent results in data mining, and has gradually become the main technology of data mining. The lack of e data in actual production makes unsupervised learning more adaptable. It is widely used in many scenarios, such as commodity recommendation and numerical prediction. However, in this scenes, the numerical range of data is very wide, and sum of them have customized personalized services, which not only requires clustering algorithm to be suitable for the non-uniform density dataset with numerical vast density gradually sparse, but also needs to have diversified results and high homogeneity. Cluster analysis place an extremely important role in data mining and can make a very important contribution in large number of data analysis business. Now a days, the data volume is increasing rapidly, so it is urgent to improve efficiency and reliability of clustering algorithm in the Stage of clustering analysis.

2.6 A survey on Human action recognition from videos

In Chandni J. Dhamsania[6], A survey on Human action recognition from videos, this paper comprehensively discusses the methods and limitations in the field of human action recognition as shown in Figure 2.6. Trajectory based approach, hierarchical approach, semantic descriptor based approach, spatio-temporal interest point based approaches are used widely for human action recognition. Thus the human action recognition methods conclude that the progress in the field of action

recognition is encouraging.

Type of Action	Methodology	Results	Dataset	Limitations
Two person or person-object interaction	In the proposed approach [7] Multiple Instance Learning (MIL) is used for subsequence action classification and then features such as trajectory, HOG, HOF, MBH is computed and bag of feature encoding technique and SVM for classification	57.42% for full sequence classifier and 59.80% for combined sequence classifier	Hollywood2	Action recognition fails when body motion is not clearly visible.
Two person or person-object interaction	This framework [8] uses three layer convolution Independent Subspace Analysis and PCA for human interaction recognition from segmented and unsegmented videos.	For segmented videos accuracy obtained is 93% and 90% on set1 and set2 respectively. Similarly, for unsegmented videos it is 90% and 85% for set1 and set2 respectively.	UT-interaction	Spatial and temporal localization of activities are not possible using proposed approach
Two person or person-object interaction	In this approach [9] support vector machine is trained using motion interchange pattern for shot detection, optical flow field of motion salient pixels from region of interest is computed, self similarity matrix for view invariant feature and Fisher encoding technique is used.	50.1%	TV Human Interaction	Very low accuracy is achieved
Two person or person-object interaction	In this framework [10] human interaction is carried out using body pose features. For this skeleton of person are created, Joint features of skeleton are extracted and MIL is used.	87.3%	RGBD Videos captured using Microsoft Kinect Sensors.	The experiment is carried out for the videos from specific view point.
Two person or person-object interaction	In the proposed approach [23] human action recognition is carried using dense trajectories, the feature points are tracked using optical flow. Trajectory, histogram of oriented gradient, histogram of optical flow and motion boundary histogram features are computed, codebook is generated using Bag of features. Support vector machine is used for classification.	58.3%	Hollywood-2	The accuracy obtained is low.

Figure 2.4: Comparison Survey on Human action recognition

In the future, there are some performance issues that need to be solved for real time deployment. Many challenges like high computation cost, change in appearance, change in illumination, changing camera view point and low recognition rate need to be solved.

3. System Design

3.1 UML Diagrams

The Unified Modeling Language (UML) is used to specify, visualize, modify, construct, and document the artifacts of an object-oriented software-intensive system under development. UML offers a standard way to visualize a system's architectural blackprints.

3.1.1 Use Case Diagram

Use-case diagrams model the behavior of a system and help to capture the requirements of the system. Use-case diagrams describe the high-level functions and scope of a system.

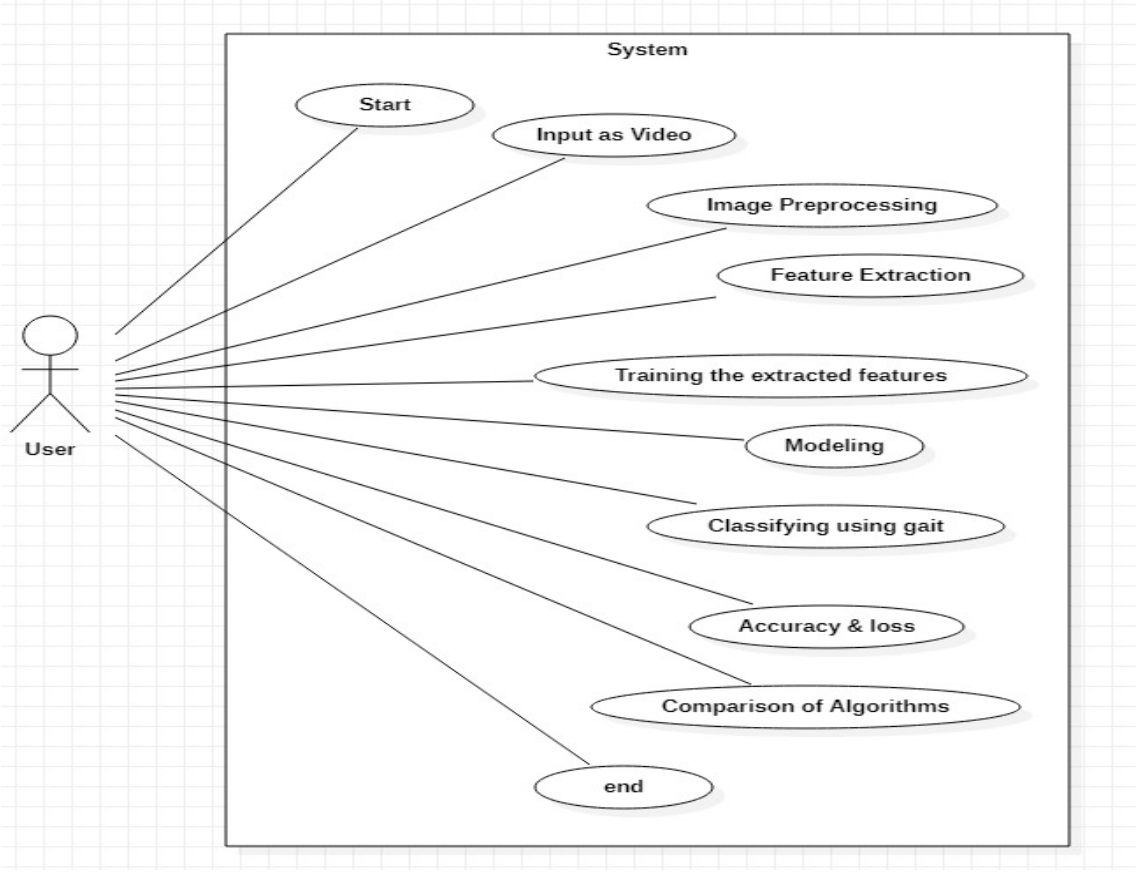

Figure 3.1: Use-Case Diagram

These diagrams also identify the interactions between the system and its actors as shown in Figure 3.1. The use cases and actors in use-case diagrams describe what the system does and how the actors use it, but not how the system operates internally. Use-case diagrams illustrate and define the context and requirements of either an entire system or the important parts of the system. A complex system can be modeled with a single use-case diagram, or create many use-case diagrams to model the components of the system. You would typically develop use-case diagrams in the early phases of a project and refer to them throughout the development process.

3.1.2 Sequence Diagram

Sequence Diagrams Represent the objects participating in the interaction horizontally and time vertically. Sequence Diagrams are interaction diagrams that detail how operations are carried out. They capture the interaction between objects in the context of a collaboration.

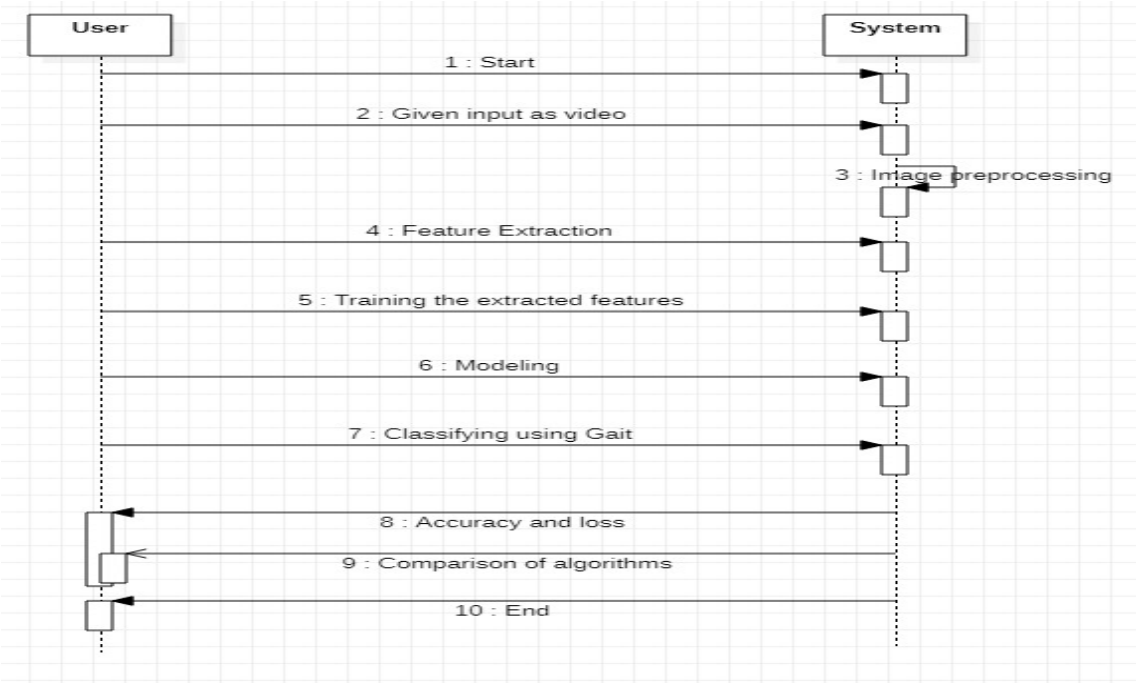

Figure 3.2: Sequence Diagram

Sequence Diagrams are time focus, and they show the order of the interaction visually by using the vertical axis of the diagram to represent time, what messages are sent, and when. A sequence diagram shows, as parallel vertical lines (lifelines), different processes or objects that live simultaneously, and, as horizontal arrows, the messages exchanged between them, in the order in which they occur as shown Figure 3.2. This allows the specification of simple runtime scenarios in a graphical manner. The sequence diagram represents the flow of messages in the system and is also termed as an event diagram. It helps in envisioning several dynamic scenarios. It portrays the communication between any two lifelines as a time-ordered sequence of events, such that these lifelines took part at the run time. In UML, the lifeline is represented by a vertical bar, whereas the message flow is represented by a vertical dotted line that extends across the bottom of the page. It incorporates the iterations as well as branching.

3.2 S/W & H/W Requirements

3.2.1 Hardware Requirement

1. Processor : 2.16 GHz(Minimum)
2. RAM : 4 GB(Minimum)
3. Application architecture : 64-bit

3.2.2 Software Requirement

1. Operating System : Ubuntu/Windows.
2. Programming Platform : Python IDE 3.7, Sublime.
3. User Interface : Tkinter

4. Methodology

4.1 Architecture

The Working flow of the Project consists of five steps as shown in Figure 4.1.

1. Input Video Image
2. Image Preprocessing
3. Feature Extraction
4. Human Action Imitation
5. Output Imitation Results

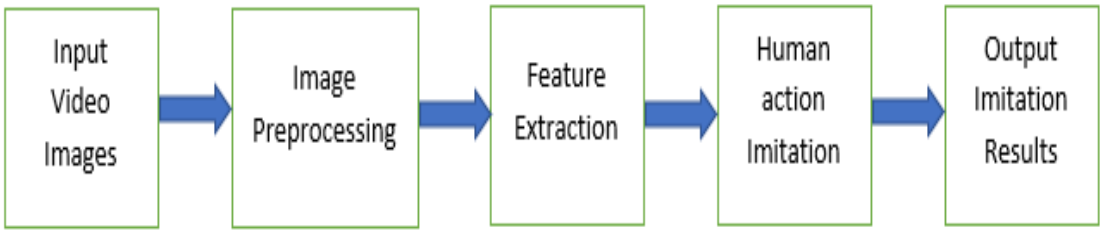

Figure 4.1: Block diagram of the proposed system

Initially give video as input to the image pre-processing. In image pre-processing it will clean, integrate and reduce the frames in the video which are not required. The output of data pre-processing acts as input to the feature extraction. In the feature extraction step it will extract the key points from the frames. And then it will enter into the next step called Human action recognition. In human action recognition, used gait classifier mechanism using CNN and VGG16. When trained the video with CNN and VGG16, CNN gave more accuracy with 99.33%.

4.2 Modules

The modules included in Human Action Imitation Using Gait Classifier are

1. Pre-Processing
2. Feature Extraction
3. Pose Detection

4.2.1 Pre-Processing

The nature of 3D pose data, an identical pose can be expressed differently depending on the viewing angle and the height or shape of a person. This characteristic becomes a problem when learning because even in the same pose, it can be recognized as a different pose in the clustering process which leads to performance degradation of the model. Therefore, proceed to normalize and align pose so that pose can be recognized under the same conditions before clustering. The formula expressing normalize is as follows (1) .

$$norm = (x_i - min_x)/(max_x - min_x) \tag{1}$$

The pose can be normalize by applying the normalization ratio to all the data. Also, rotate the data based on the left and right pelvic joints of pose data to process all data to have the same angle. Through these methods, a problem in which a similar pose is recognized as a different pose due to an angle can be solved.

4.2.2 Feature Extraction

The output of the pre-processing module is given as the input to the feature extraction module. Raw video sequence consists of massive spatio-temporal pixel intensity variations that contribute nothing to the action itself, such as pixels related to the color of clothes and cluttered background. Feature extraction is a process that detects and extracts most representative information from raw data as features. Any video sequence will generate a specific number of features, and different video sequences will have distinctive number of features. Feature representation is a process to give a unique representation for every video sequence based on the extracted features. The final representation should be

of the same dimension among different videos. The video sequences is divided in 255 frames and the output of this module is given to pose detection module, to train the data.

4.2.3 Pose Detection

The output of the feature extraction module is given as the input to pose detection module and the input data is be trained with Gait Classifier and DBSCAN techniques and the imitating the actions of the person is the output of this module as shown in Figure 4.2.

Figure 4.2: Action Imiation of the Person

Gait Classifier:

Gait(limb movement) from videos by using this classification output peoples can easily understand whether their rehabilitation gait movement are going accurately[4]. To implement this project, the module used different limb movements such as Knee movement, toe, ankle etc.

- **Gait Pattern Extraction:**

  Using this module various limb movements will be extracted during walking from videos and this features will be used to perform gait (limb movement) phase classification.

- **Gait pattern clustering:**

  Extracted above module features will be clustered using DBSCAN clustering algorithms and the similar movement will automatically goes to same cluster and for each movement such as ankle, knee and toe will have one cluster and their movement will go to its appropriate cluster.

- **Gait Phase Reconstruction:**

  Using this module, the model can analyse cluster to get similar movements features and then perform classification to predict limb movement.

Using above 3 modules algorithm are building gait classification application and this application will accept input video from user and then start classification of limb movement.

DBSCAN(Density based spatial clustering application with Noise):

Cluster analysis is an unsupervised learning method that classifies data according to their similarity to understand the characteristics of a large database. For most cluster analysis, the k-means clustering method that determines the number of clusters in advance is adopted frequently due to its simplicity[5]. DBSCAN has the advantage of creating an appropriate cluster on its own.he process of tuning parameters to fit dataset has continued in past because it is important to properly tune two parameters(minPts, Eps) to make DBSCAN perform well as shown in the Figure 4.3 and Figure 4.4.

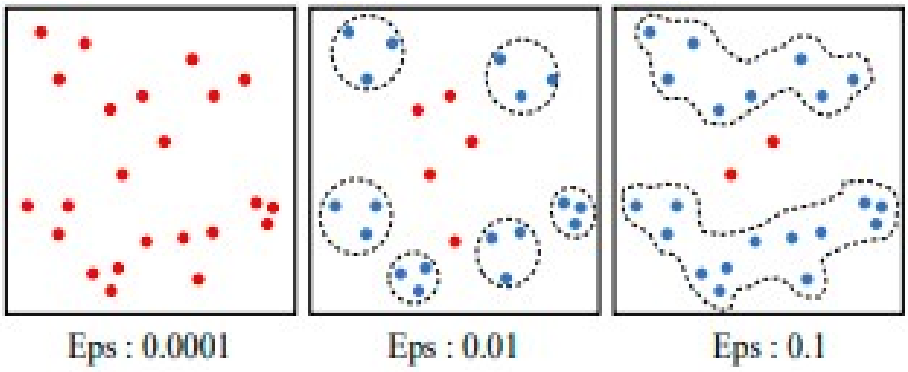

Figure 4.3: The tendency of clustering according to Eps.

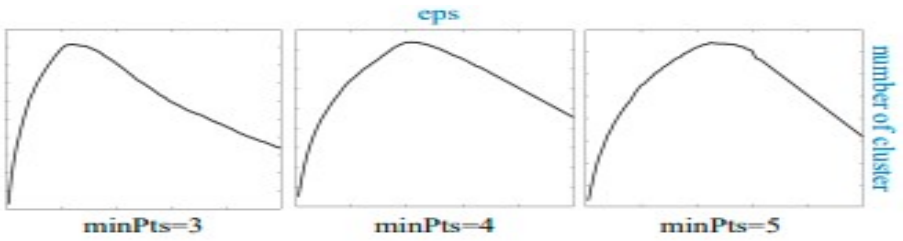

Figure 4.4: The tendency of clustering according to minPts.

4.3 Algorithms

The input data is trained with two algorithms like CNN and VGG16, to know which algorithm has given more accuracy for the given input data.

4.3.1 VGG16(Visual Geometry Group 16)

The input of the VGG16 is a fixed size image of 224×224×3 as shown in Figure 4.7. This input image is passed through a stack of convolutional layers (boxes that use ReLus as activation functions). The convolutional layers are usually accompanied by max-pooling layers (max pooling boxes) then two dense layers (boxes fully connected and ReLu as activation function) of 4096 nodes each[2]. Finally, a dense layer (boxes with Softmax activation function) of 1000 nodes, yields the output of this CNN.

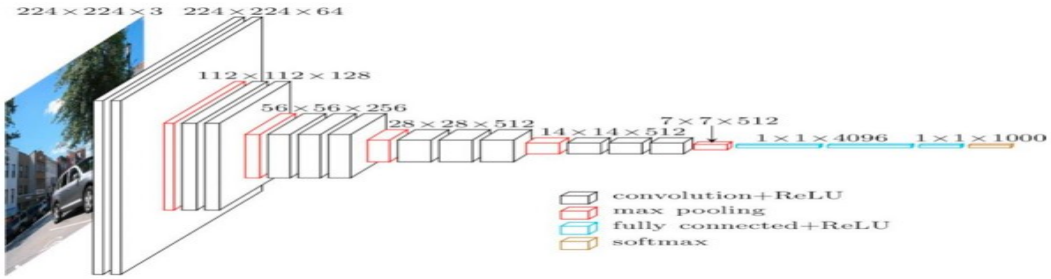

Figure 4.5: VGG16 Network Architecture

4.3.2 CNN & 3D CNN

CNN(Convolutional Neural Networks):

Convolutional Neural Networks (CNN) are typically used to analyze image data and map images to output variables. In this model, perform 3D convolutions in the convolution Stages of CNNs to compute features from both spatial and temporal dimensions. The 3D convolution is achieved by convolving a 3D kernel to the cube formed by stacking multiple contiguous frames together[1]. By this construction, the feature maps in the convolution layer is connected to multiple contiguous frames in the previous layer, thereby capturing motion information as shown in Figure 4.5. Formally, the value at position (x, y, z) on the jth feature map in the ith layer is given by equation (2).

$$v_{ij}^{xyz} = tanh\left(b_{ij} + \sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1} w_{ijm}^{pqr} v_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \quad (2)$$

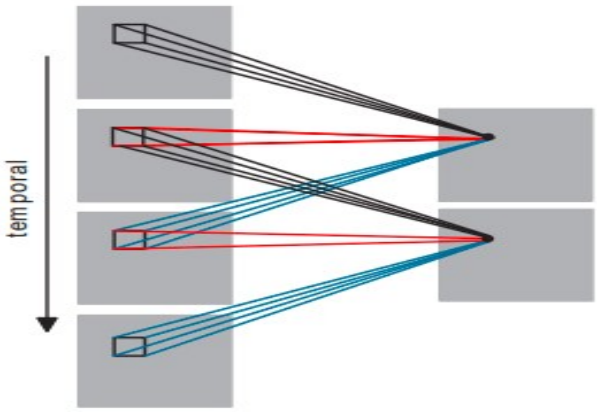

Figure 4.6: 3D convolution

A 3D convolutional kernel can only extract one type of features from the frame cube, since the kernel weights are replicated across the entire cube. A general design principle of CNNs is that the number of feature maps should be increased in late layers by generating multiple types of features from the same set of lower-level feature maps.

3D CNN(3D Convolutional Neural Networks):

In 3D CNN architecture shown in Figure 4.6, consider 7 frames of size 60×40 centered on the current frame as inputs to the 3D CNN model. In this model, first apply a set of hardwired kernels to generate multiple channels of information from the input frames. This results in 33 feature maps in the second layer in 5 different channels known as gray, gradient-x, gradient-y, optflow-x, and optflow-y[3]. The gray channel contains the gray pixel values of the 7 input frames. The feature maps in the gradient-x and gradient-y channels are obtained by computing gradients along the horizontal and vertical directions, respectively, on each of the 7 input frames, and the optflow-x and optflow-y channels contain the optical flow fields, along the horizontal and vertical directions, respectively, computed from adjacent input frames. This hardwired layer is used to encode our prior knowledge on features, and this scheme usually leads to better performance as compared to random initialization.

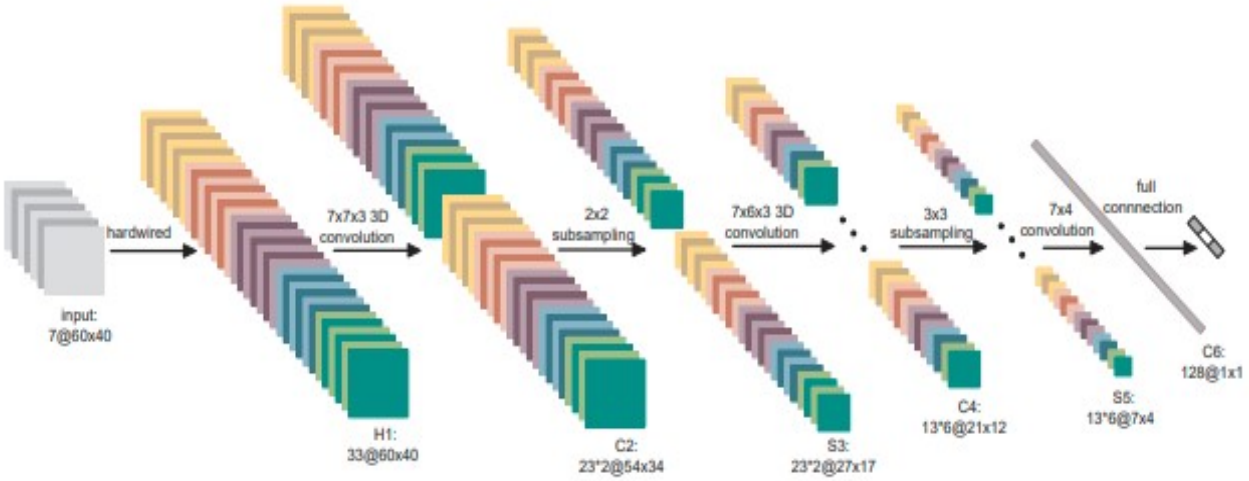

Figure 4.7: A 3D CNN architecture for human action imitation

After, that apply 3D convolutions with a kernel size of 7 × 7 × 3 (7 × 7 in the spatial dimension and 3 in the temporal dimension) on each of the 5 channels separately. To increase the number of feature maps, two sets of different convolutions are applied at each location, resulting in 2 sets of

feature maps in the C2 layer each consisting of 23 feature maps. This layer contains 1,480 trainable parameters. In the subsequent subsampling layer S3, apply 2×2 subsampling on each of the feature maps in the C2 layer, which leads to the same number of feature maps with reduced spatial resolution. The number of trainable parameters in this layer is 92. The next convolution layer C4 is obtained by applying 3D convolution with a kernel size of $7 \times 6 \times 3$ on each of the 5 channels in the two sets of feature maps separately. To increase the number of feature maps, apply 3 convolutions with different kernels at each location, leading to 6 distinct sets of feature maps in the C4 layer each containing 13 feature maps[3]. This layer contains 3,810 trainable parameters. The next layer S5 is obtained by applying 3×3 subsampling on each feature maps in the C4 layer, which leads to the same number of feature maps with reduced spatial resolution. The number of trainable parameters in this layer is 156. At this Stage, the size of the temporal dimension is already relatively small (3 for gray, gradient-x, gradient-y and 2 for optflow-x and optflow-y), so perform convolution only in the spatial dimension at this layer. The size of the convolution kernel used is 7×4 so that the sizes of the output feature maps are reduced to 1×1. The C6 layer consists of 128 feature maps of size 1×1, and each of them is connected to all the 78 feature maps in the S5 layer, leading to 289,536 trainable parameters.

5. Partial Implementation & Results

The approach to develop an efficient gait classifier mechanism using deep learning tools such as Convolutional Neural Network (CNN) and VGG16 based classifier is presented. When compared both the algorithms CNN got more accuracy, so we proceeded with CNN as shown in Figure 5.1. The proposed approach uses the human action imiation method to classify the actions found in a specific person's gait. By introducing a proper data collection and training scheme for CNN.

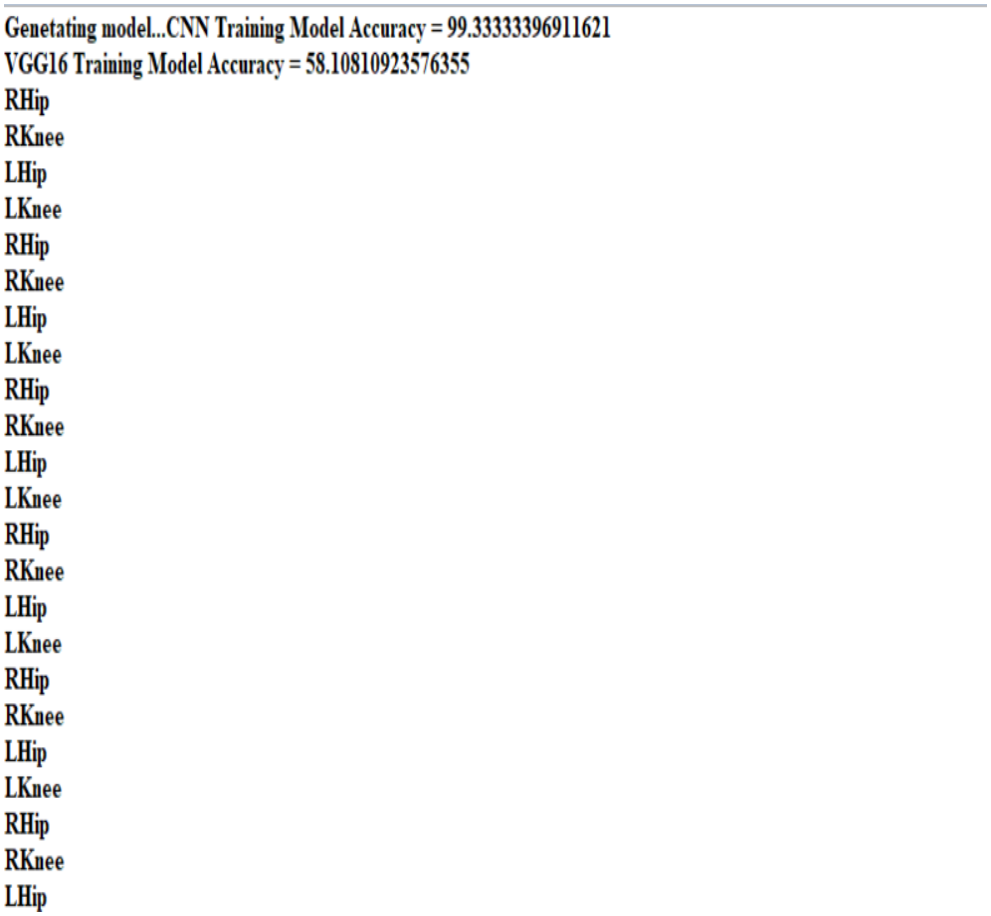

Figure 5.1: Key Points Recognition

Figure 5.2: Human action Imitation 1

The experimental results are promising in solving some of the typical problem constraints involved in the gait analysis system as shown in Figure 5.2 and Figure 5.3. The accuracy achieved for classifying action of a person is 99.33 % that proves the applicability of the proposed mechanism.

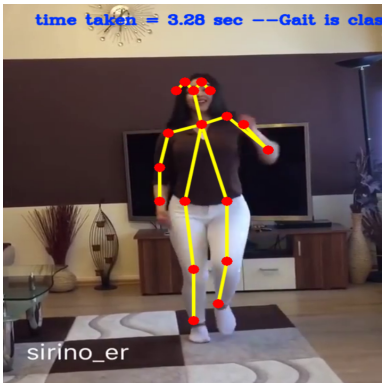

Figure 5.3: Human action Imitation 2

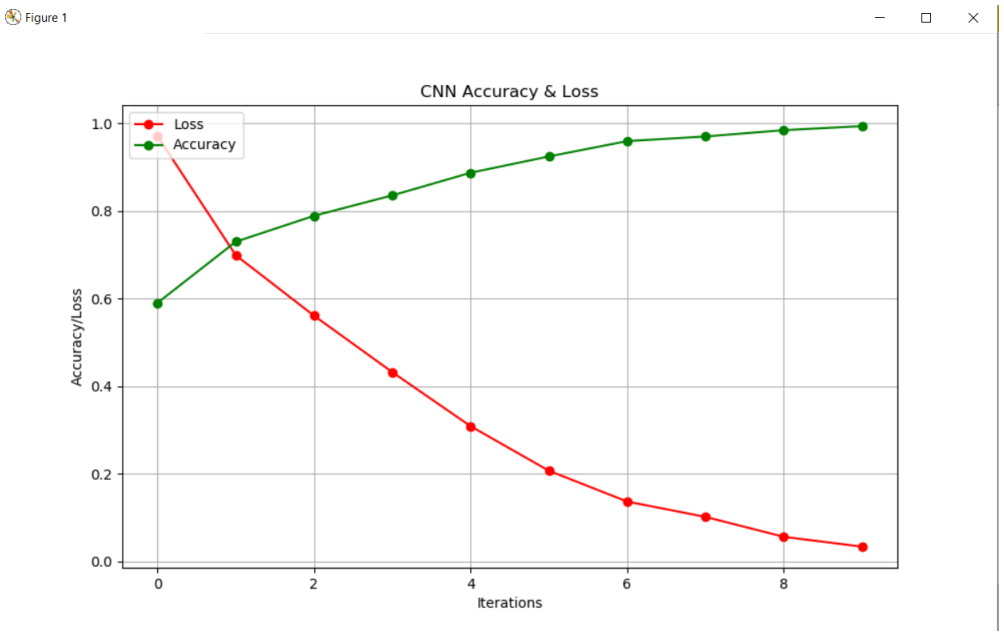

Figure 5.4: Accuarcy of comparisons

The above Figure 5.4 shows about the variation of loss and accuracy of CNN when it is trained with the input data.In CNN Accuracy and Loss,in which CNN algorithm we used 2d CNN in which CNN have different layers and from those layers we get the accuracy and loss.In this algorithm we have activation layers like relu and softmax and we have CNN loss like categorical crossentropy and we used optimizer like adam for training deep learning models.

6. Conclusion & Extension Plan for Stage II

The main purpose of the model is to imitate the actions of the person beside him or her. In Stage I, the partial execution of the project, was just implemented upto imitating the actions on the person, as shown in the Figure 6.1.

Figure 6.1: Imitation on the person

While imitating the actions of the human, this model will get the imitated features on the human. So, in extension plan for Stage II, the model can implement the actions of the person beside him or her as shown in the Figure 6.2. This model will be demonstrating on the live videos and want to test the optimized videos on different deep learning algorithms.

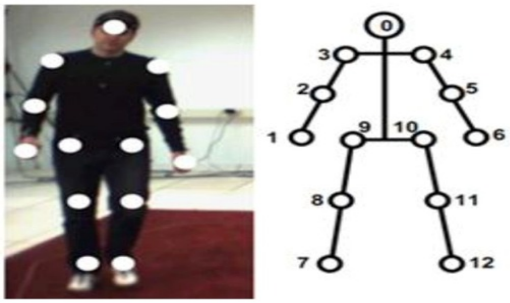

Figure 6.2: Imitation beside the person

REFERENCES

[1] D. C. Luvizon, D. Picard and H. Tabia, ”Multi-Task Deep Learning for Real-Time 3D Human Pose Estimation and Action Recognition,ïn IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 43, no. 8, pp. 2752-2764, 1 Aug. 2021, doi: 10.1109/TPAMI.2020.2976014.

[2] M. Wang, J. Tighe and D. Modolo, Çombining Detection and Tracking for Human Pose Estimation in Videos,”2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 11085-11093, doi: 10.1109/CVPR42600.2020.01110.

[3] A. Rohan, M. Rabah, T. Hosny and S. H. Kim, ”Human Pose Estimation-Based Real-Time Gait Analysis Using Convolutional Neural Network,ïn IEEE Access, vol. 8, pp. 191542-191550, 2020, doi: 10.1109/ACCESS.2020.3030086.

[4] S. U. Yunas, A. Alharthi and K. B. Ozanyan, ”Multi-modality sensor fusion for gait classification using deep learning,”2020 IEEE Sensors Applications Symposium (SAS), 2020, pp. 1-6, doi: 10.1109/SAS48726.2020.9220037.

[5] D. Deng, ”DBSCAN Clustering Algorithm Based on Density,”2020 7th International Forum on Electrical Engineering and Automation (IFEEA), 2020, pp. 949-953, doi: 2020.

[6] C. J. Dhamsania and T. V. Ratanpara, .A survey on Human action recognition from videos,”2016 Online International Conference on Green Engineering and Technologies (IC-GET), 2016, pp. 1-5, doi: 10.1109/GET.2016.7916717.

[7] B. Antic, T. Milbich and B. Ommer, ”Less is more: video trimming for action recognition”, International Conference on Computer Vision Workshops (ICCVW), IEEE, pp. 515-521, 2013.

[8] N. Nguyen and A. Yoshitaka, ”Human interaction recognition using independent subspace analysis algorithm”, International Symposium on Multimedia (ISM), IEEE, pp. 40-46, 2014.

[9] B. Zhang, Y. Yan, N. Conci and N. Sebe, ”You talkin’ to me?: recognizing complex human interactions in unconstrained videos”, International Conference on Multimedia, ACM, pp. 821-824, 2014.

[10] H. Wang, A. Kläser, C. Schmid and C. Liu. .Action recognition by dense trajectories”, Computer Vision and Pattern Recognition, IEEE, pp. 3169-3176, 2011.

[11] K. Yun, J. Honorio, D. Chattopadhyay, T. Berg and D. Samaras, ”Two-person interaction detection using body-pose features and multiple instance learning”, Computer Society Conference on Computer Vision and Pattern Recognition Workshops, IEEE, pp. 28-35, 2012.