Proceedings of 12th IOE Graduate Conference
Peer Reviewed
ISSN: 2350-8914 (Online), 2350-8906 (Print)
Year: 2022 Month: October Volume: 12

Human Emotion Recognition from Gait Analysis using Graph Convolution Network

Umesh Kanta Ghimire [a], Saroj Shakya [b], Shashidhar Ram Joshi [c], Janardan Bhatta [d]

[a, c] *Pulchowk Campus, IOE, Tribhuvan University, Nepal*
[b, d] *Thapathali Campus, IOE, Tribhuvan University, Nepal*
[a] ukg@tcioe.edu.np, [b] sarojsh@tcioe.edu.np, [c] srjoshi@ioe.edu.np, [c] janardan.bhatta@ioe.edu.np

Abstract
Human walks refers to a constantly motion that represents not only flexibility, but it can too be used to identify the walker by either human spectators or computers. The mapping between differing emotions and walks patterns supports a new beginning for automated emotion recognition. The classification of perceived human emotion from gaits has been based on a Spatial Temporal Graph Convolutional Network (ST-GCN) architecture. From the RGB video of person's walking, the formulation implicitly used the gait features to classify the emotion of the human into one of the four emotions: Angry, Neutral, Happy and Sad. The annotated 1835 ELMD (Edinburgh Locomotion Mocap Database) dataset has been used both for 2D and 3D Model. The dataset has been increased up to 3841 after augmentation. The neural network model has been trained on annotated real-world gait videos and acquired an efficiency of 88.67% for 3D model and 96.77% for 2D model.

Keywords
Emotion Recognition, Gait, Gait Analysis, Gait Cycle, GCN, Human Gait, ST-GCN

1. Introduction

A person's gait is the way they walk. It shows a typical and significant daily activity through which onlookers can learn a great deal of interesting facts about the walker. The systematic study of human locomotion is known as gait analysis. Forensic walks study or forensic walks comparison is defined as the measurement and evaluation of the gait patterns and characteristics of the people/suspects and the equating of these characteristics with the scene of misbehavior evidence for criminal/private labeling [1]. In other words, since individualization of a person's walks of gait has not yet been entirely scientifically proven, forensic gait analysis can be thought of as a subscriber to the identification process rather than an individual's identification by labeling [2].

The purpose of this paper is to investigate the use of gait-related movement characteristics in a walking film for emotion perception. A gait is defined as an organized temporal series of body joint transformations (mostly translations and rotations) that occur during a single cycle of walking. The way someone walks is known as their gait, to put it simply.

There has been an increase in a number of different incidents and crime scenes, including homicides, kidnappings, sexual assaults, hit-and-runs, shoplifting, and HBT (House Break-in and Theft) etc. The world will be faced with a range of criminal actions if that keeps happening. We are inspired to work on this research because of the bad effects of crime, and We hope it will help the investigating police and crime scene investigators while they are dealing with the foot, footwear, and/or gait-related evidence at the crime scene. This information can further establish the suspect's physical or biological profile for individualization and identification, which helps partially offset these problems [3].

Various models have been already proposed where human emotion can be easily recognized from their facial expression as well as speech recognition as the distance between object and observers is very-very nearer. These kind of traditional human emotion recognition could not be remotely observable and there might be a high chance of active manipulation and imitation by object as there is a chance of less cooperation from the object as well. The use of hardware like different motion sensors in object's pocket is tedious and quite irrelevant for the emotion recognition as well. Hence We believe this work and

the model We used can be a good option for such needs as the model detects a human emotion from spatial and temporal features as traditional methods lacks temporal features like jerk, velocity and acceleration. The model predicts discrete perceived emotions from 3D pose sequences of human gaits from spatial temporal graph convolution network using posture and movement features.

Unlike CNN, Graph convolution network model learns the characteristics by looking at the nearby nodes. The primary distinction between CNNs and GNNs is that the convolution neural network was developed specifically to function on regular (Euclidean) organized data, whilst the GCN are generalized CNNs with varying numbers of connections and unordered nodes (irregular on non-Euclidean structured data) i.e. non regular structures. Since information is transferred towards neighboring unordered nodes within a graph and information is learned even before training, GCN were used to detect emotions in our model.

1.1 Main Contribution

The Spatial Temporal Graph Convolution Network (ST-GCN) implicitly extracts the person's stride from a walking video. To learn hybrid features, it blends thoroughly learned features with affective features. The Previous model [1] uses a variation auto-encoder and decoder techniques for emotion recognition. It uses both the labeled and unlabeled dataset so a semi-supervised learning. Another model [4] used a generative technique for dataset, which was not real dataset and hence generated. Our model uses only the labeled and real annotated dataset and hence supervised learning and does not depends upon the encoding and decoding technique. The model is trained on ELMD Dataset which was further processed for augmentation to increase the size of a dataset. Some joints are interpolated and processed. We develop a model that fits for input of both 2D and 3D coordinates and compares the result as well. The use of an Openpose for emotion recognition from gait analysis is a novel approach.

2. Literature Review

We thoroughly evaluate past work in action recognition and creation from gaits, which is a challenge linked to categorizing perceived emotions from gaits. A recent study [1] on auto encoder-based semi-supervised method for categorizing observed human emotions from walking patterns acquired in motion-captured data or videos and represented as sequences of 3D postures. Following the kinematic chains in the human body, they hierarchically pooled these joint movements in the encoder based on the motion on each joint in the pose at each time step retrieved from 3D pose sequences. Additionally, the latent embedding of the encoder was to only include the space of affective features that are psychologically driven and underlie the gaits. The classifier model was trained to link the latent embedding to emotion labels for the annotated data. On the Emotion-Gait (E-gait) benchmark dataset, which includes labeled and unlabeled gaits gathered from various sources, our semi-supervised technique obtains a mean average accuracy of 0.84.

The use of innovative generative network dubbed STEP-Gen by Uttaran Bhattacharya [1] , based on an ST-GCN dependent Conditional Variational Autoencoder created annotated synthetic gaits in addition to real-world gait videos for the training STEP (CVAE). The STEP classification accuracy is increased by pushpull regularization into the CVAE formulation of STEPGen. This paper used E-Gait, which includes thousands of artificial gaits in addition to 4,227 human gaits that have been annotated with perceived emotions. Compared to earlier techniques, STEP demonstrated classification accuracy of 88% on E-Gait.

In a study performed [5] , an architecture for a reliable and highly accurate Kinect-based gait identification system was proposed. It considers Joint relative cosine dissimilarity and joint relative triangular area which are two geometrical properties that were presented. Because both of the suggested features were view and pose invariant, recognition performance was improved. The feature vector of dynamic joint relative cosine dissimilarity and joint relative triangle area was used to train the neural network model. In this paper, the loss of the objective function is achieved by use of the Adam optimization approach. On two publicly accessible 3D skeleton-based gait datasets captured with the Microsoft Kinect sensor, the performance of the suggested deep learning neural network architecture was assessed. It had been demonstrated through experiments that the neural network architecture with dynamic geometric characteristics, outperformed previous techniques for Kinect skeleton-based gait

identification. After a 5-fold cross validation experiment, the suggested deep learning neural network, trained with the proposed geometric features, has an accuracy of 95.30% on the UPCV dataset and 98.08% on the GaitBiometry dataset.

Gait analysis was used in court in the case of Thomas Jackson in 1839 in London. Thomas Jackson was recognized by the witness due to his bowed left leg and limping when walking. [6] Although the method of forensic gait analysis is still debatable in terms of its accuracy and reliability, it was used for the first time as a form of admissible evidence/scientific evidence in the case of R v. Saunders at the Old Bailey Central Criminal Court in London, UK, by forensic podiatrist Dr. Haydn Kelly. The study explored forensic gait analysis techniques, human gait influences, forensic applications, and accuracy, reliability, and admissibility of the results. [7]

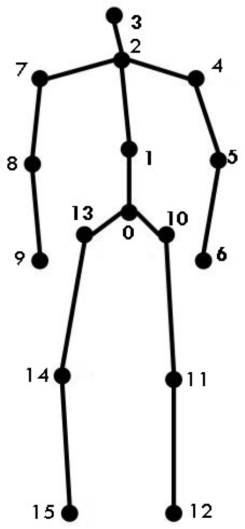
3. REQUIREMENT Analysis

3.1 Dataset Analysis

The source of dataset We have used consists of 1,835 real gaits. One data is comprised of 240 sets of 21 points in 3D coordinate system. These 1,835 data have been taken from the Edinburgh Locomotion MOCAP Database (ELMD) and annotated by Bhattacharya et. al to reduce the 21 points of joints to 16. The format for each data file is T x V, where T is the number of time steps and V is the number of coordinate locations or number of joints in the human skeleton system. The value of T in the data set is 240. The fps of the video to make this data is 60. So, the length of video in sec is 4. On the other hand, as mentioned earlier, value of V was 21 in the original dataset and the 5 trivial points has been dropped out by Bhattacharya et. al. fixed for all the files.

The points after removal of non-important points and re-arrangement the joint, 16 points (joints) are as shown in figure 1.

The 5 joints left toe, right toe, lower back, left hand index and right hand index has been removed. The labeled dataset consists of datasets of all 4 type of classes with different no. of sample. Since the number of data in each category is not evenly distributed, We have augmented the data with different transformation methods to make the number of data for all classes equal.



Joint Number	Joint Name
0	Root/Hip
1	Spine
2	Neck
3	Head
4	Left Shoulder
5	Left Elbow
6	Left Hand
7	Right Shoulder
8	Right Elbow
9	Right Hand
10	Left Hip
11	Left Knee
12	Left Foot
13	Right Hip
14	Right Knee
15	Right Foot

Figure 1: A Human Skeletal System showing 16 joints

3.2 Input Data

Openpose, an open source tool to extract human skeleton coordinates from the raw video of human, has been used to extract and feed the input data to our trained model. Openpose transforms raw gate sequence in to dynamic skeleton system. Primarily the skeletal system comprises of 25 joints. Later we preprocessed it to reduce the number of joints to 16 joints to 16 to match with our dataset. Each joint in a frame is a tuple of 2D pixel coordinates and a confidence score. The third value (confidence score) has not been used since the model expects coordinate system value. Since two models are trained for 2D and 3D data, the data from OpenPose is used as input data for 2D system.

4. SYSTEM ARCHITECTURE and Methodology

4.1 Theoretical Modeling

4.1.1 Graph Convolution Network

CNNs, graph convolution networks, in contrast to RNNs generate parameterized filters that are employed in multi-layer neural network models by utilizing problem-specific specialized designs that are known from spectral graphs. GCNs are a particular kind of convolutional neural network that can operate directly on graphs and benefit from their structural data. It resolves the issue of categorizing nodes in graphs (like citation networks) where labels are only

accessible for a small portion of nodes (such as documents) (semi-supervised learning).

The use of GCNs is rather straightforward. They function in layers; which we can combine to create any depth we desire. Three processes are taking place within each layer. The graph's structure is first standardized. Second, the node characteristics are multiplied by the normalized graph structure. The node attributes and weights are subjected to a non-linearity function as the final step inside each GCN layer. The new symmetric normalization equation becomes

$$D^{-\frac{1}{2}}.A.D^{-\frac{1}{2}}.X.$$

Finally a linear activation function ReLU is applied. The equation is modified in to

$$H^{[i+1]} = \sigma[\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{[i]}W^{[i]}]$$

4.2 System Architecture

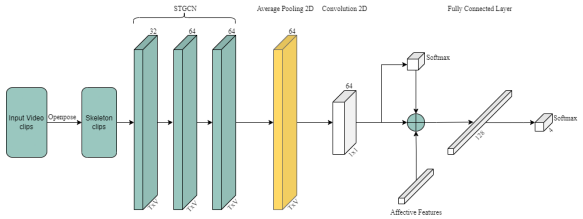


Figure 2: System architecture of proposed model

4.2.1 Pose extraction

Pose extraction is a process of extracting skeleton joints from a sequence of input videos. Openpose was used to extract the pose from the input video. The 3D coordinates of the skeleton joints (edges) and vertices consumed by Graph Convolution Network.

4.2.2 Feature extraction

The stride and step duration and cadence are the most often used temporal gait characteristics. In addition, the distance traveled between two successive ICs (step and stride length) can be used to define spatial gait characteristics. The gaits in our dataset were gathered from various angles and scales. Using the Umeyama approach, we convert all gaits to a common point of view in the world coordinates (1991). In our example, a gait is therefore an extracted per-frame view normalized skeletal graph that is temporally sequenced from a video/movie. We now give the term "gait" a proper definition. A graph G = (V, E) is used to represent a gait, where V stands for the collection of vertices and E stands for the set of edges. In addition to spatial and temporal features that is coordinates of vertices and edges of skeleton joints in time series, we added following two important features which might affect the result significantly.

4.2.3 Posture Extraction

The angles and separations between the joints, the areas of the various body parts (such as the area of the triangle formed by the neck, the right hand, and the left hand), and the bounding volume of the body are some of these.

4.2.4 Movement Features

These are the velocity and acceleration of individual joints in the gait. In our final network, we made use of the affective feature formulation created by Kleinsmith and Bianchi-Berthouze in 2013; Crenn et al. 2016b. In order to create hybrid feature vectors, we added the 29-dimensional affective feature to the last layer feature vector that was learned by our baseline network. In order to produce the final class labels, these hybrid feature vectors are passed through two fully connected layers with corresponding dimensions of 128 and 4.

4.3 ST-GCN

As seen in Figure 2, each input gait is processed by a group of three ST-GCN layers. There are 32 kernels in the first ST-GCN layer and next two have 64 kernels each.

4.4 Average Pooling

A 1×1 convolutional layer is applied after the output of the last ST-GCN layer has been average pooled in both the temporal and joint dimensions. Except for the fully connected layer, all ST-GCN layers are followed by a BatchNorm layer and the ReLU nonlinearity (not shown separately in Figure 2).

4.5 Convolution Layer

A softmax operation has been done to generate the class labels, the output of the convolutional layer is sent through a fully connected layer of dimension 4 (matching to the 4 emotion labels we have: angry, neutral, happy, and sad).

4.5.1 Fully Connected Layer

Finally, softmax activated fully connected layer is used to classify the emotion of human.

4.5.2 System Block Diagram

Our system is comprised of following different blocks, primarily input block to feed input video, post estimator with pre-trained model, feature extractor implemented with GCN and additional features and finally classifier to recognize the emotion.

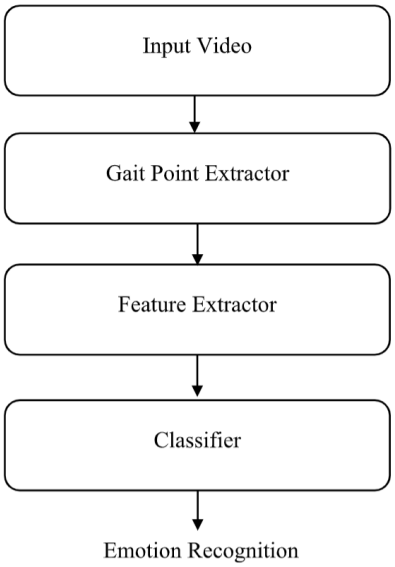


Figure 3: Spatial-Temporal Emotion Perception System Block Diagram

Input video can be the recorded video or video streamed from the camera. For the demonstration purpose a recorded video with person is taken as an input whose emotion is to be detected. This Gait point extractor block is a pre-trained model provided by OpenPose. Since primary focus of our research is on training the model to achieve maximum accuracy, the pose estimator block is thus borrowed from third party. This block helped us to extract features required to feed into GCN as input features. The input features are basically 3D space vectors of skeleton joints and connection between those joints. The connection between vertices of adjacent frames are called temporal features. So this features are times series data.The classifier comprised of ST-GCN layers, pooling layers, convolution layers and fully connected layers. These layers with appropriately chosen activation functions along with extra features like movement features and posture features that will help to classify the input gait data.

5. Implementation Details

5.1 Dataset Collection and Preparation

To train models using GCNs, data from ELMD has been taken. The dataset features have already been discussed in chapter 3. As discussed earlier, the ELMD dataset consists of altogether joint coordinates of 1835 clips. Each clip has 240 frames of 60 with 4 seconds duration (i.e. 60 fps). Those 1835 data have been annotated with 4 labels: Angry, Neutral, Happy and Sad. Out of 1835 data, 1048 samples belong to Angry, 454 belong to Neutral, 254 belong to Happy and 79 to Sad. The ELMD dataset contains 1835 different sequences of human motion captured with a MOCAP system at the School of Informatics, The University of Edinburgh. The data was captured on two separate occasions during which all actions were performed by the same male actor. The datasets contain mainly human locomotion data, including walking, jogging and running.

5.2 Dataset Preprocessing

For the training of 2D model, all data in 3D system needs to be converted into 2D system. The z-axis coordinate value has been eliminated to convert all the data in 2D coordinate system. To input the data in model trained using data in 2D coordinate system, OpenPose, an open-source pre-trained model has been used. Since, OpenPose gives the coordinates of human joints in a given video or live video from camera in 2D system with 25 joints, the 10 extra joints are removed and one additional joint is added using interpolation method. Hence the resultant joints match the joint in database which has been used to train the model. The elimination of 10 extra joints' value has not significant effect in the model performance because those joints are redundant and trivial. The joints right eye, left eye, right ear and left ear are eliminated to keep the nose as head joint. The neck is just centered and serves as shoulder center for our model. The right wrist serves as right hand and left wrist serves as left hand for our model. A spine coordinates was missing on a dataset and the coordinates were calculated by a technique called interpolation from the coordinates from left hip, right hip and neck.

The right heel, right big toe and right small toe in a 25-coordinate systems were removed by keeping only one joint, an foot joint in right and same goes to left foot joint part.

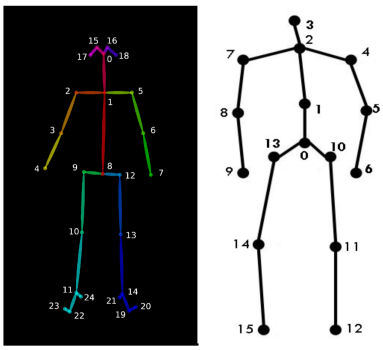


Figure 4: Openpose Standard of 25 joints (left) and 16 Joints model (right)

5.2.1 Augmentation

The mathematical expressions that have been used in augmentation are Translation and Scaling. Mathematically Translation operation can be formulated as follows:

$$(x', y', z') = f(x+a, y+a, z+a)$$

Where 'f' is the translation function and 'a' is a constant. The value of 'a' is negative for left-down and positive for right up Translation. The scalar value used for translation is in between 0-0.5 because our data is normalized in between 0-1.

Scaling is the process of increasing or decreasing magnitude of coordinate values.

$$(x', y', z') = f(nx, ny, nz)$$

The value of 'n' is a scalar number greater than 1 for scale up and less than 1 for scale down operation. Like translation, scaling has also been implemented to increase the data sample. The coordinate's values are scaled up and scaled down using value 0.5 to 1.5.

5.2.2 Point Reduction and Interpolation

For taking input for the 2D model, the coordinates from dataset that consists of 25 joints has been transformed in to 16 joints coordinates. For 3D model, input is given from test data. Figure 4 shows that the data obtained from Kinect sensor in fig. left is then interpolated and converted to few number of joints with 0 to 15; i.e. 16 in fig. right. The joints 15, 16, 17 and 18 don't have any effect in emotion recognition through gait so We ignore these dataset. The joints in right foot 11, 22, 23 and 24 in right figure is just converted to single coordinates as joint 15 in figure right. Similarly, the joints in left foot 14, 19, 20 and 21 in left figure is just converted to single coordinate as joint 12 in figure right.

5.3 Openpose

The OpenPose generally takes live video or recorded video or images as input and gives the coordinate of joints in 2D system. We took some self-videos captured from mobile camera and fed in to OpenPose and the human skeletal with 25-joints came with human skeletal form as shown in figure 5. Later it was processed to 16-joints.

Figure 5: An OpenPose output in 2-dimensional form showing 25 human index

5.3.1 Training

ELMD dataset as explained in chapter 3 has been used for the training purpose. It consists of 1048 Samples from Angry, 454 from Neutral, 254 from Happy and 79 from Sad before augmentation. The dataset after augmentation consists of 1048 Samples from Angry, 908 from Neutral, 1016 from Happy and 869 from Sad. Training is conducted for the total epoch of 200, 300, 500 and 1000 times with batch size 8 and an adaptive learning rate for both types of datasets. The computation of such matrices or tensor using CPU is inefficient as CPU only have a couple of cores to execute instructions. So, a GPU or TPU is used for computing. GPUs and TPUs have high numbers of cores which can be used in parallel for computation. For this work 16 GB RAM and 256 GB storage system and a single 4GB NVIDIA GTX 1660ti GPU has been used with 2.6 GHz Intel dual core processor.

6. Result and Analysis

6.1 Analysis Metrics

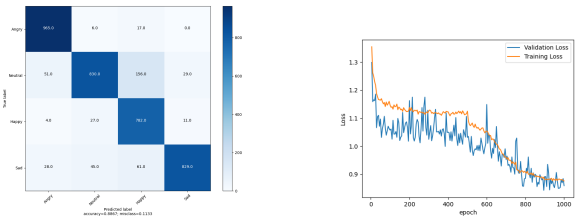
To calculate the performance of the proposed system, We used Precision, Recall, F1-Score, Sensitivity, ROC curve, Confusion matrix and Accuracy.

6.2 6.2 Analytics Metrics Summary of 3D Model

Table 1: Experiment Summary for Precision, Recall, F1 Score and Accuracy

Exp	Total Data	Class-wise Dataset Size	Epoch	Precision	Recall	F1 Score	Model Accuracy
Before Data Augmentation							
1.	679	Angry: 200 Neutral: 200 Happy: 200 Sad: 79	300	Angry: 0.99 Neutral: 0.61 Happy: 0.70 Sad: 0.0	Angry: 0.97 Neutral: 0.89 Happy: 0.70 Sad: 0.0	Angry: 0.99 Neutral: 0.72 Happy: 0.72 Sad: 0.0	75.56%
2.	679	Angry: 200 Neutral: 200 Happy: 200 Sad: 79	500	Angry: 0.99 Neutral: 0.59 Happy: 0.83 Sad: 0.0	Angry: 0.97 Neutral: 0.92 Happy: 0.70 Sad: 0.0	Angry: 0.98 Neutral: 0.72 Happy: 0.76 Sad: 0.0	76.29%
3.	1835	Angry: 1048 Neutral: 454 Happy: 254 Sad: 79	500	Angry: 0.98 Neutral: 0.61 Happy: 0.78 Sad: 0.0	Angry: 0.95 Neutral: 0.97 Happy: 0.29 Sad: 0.0	Angry: 0.96 Neutral: 0.75 Happy: 0.42 Sad: 0.0	82.13%
After Data Augmentation							
4.	3841	Angry: 1048 Neutral: 908 Happy: 1016 Sad: 869	500	Angry: 0.96 Neutral: 0.73 Happy: 0.90 Sad: 0.87	Angry: 0.90 Neutral: 0.87 Happy: 0.77 Sad: 0.91	Angry: 0.93 Neutral: 0.79 Happy: 0.83 Sad: 0.89	85.94%
5.	3841	Angry: 1048 Neutral: 908 Happy: 1016 Sad: 869	1000	Angry: 0.98 Neutral: 0.78 Happy: 0.95 Sad: 0.86	Angry: 0.92 Neutral: 0.91 Happy: 0.77 Sad: 0.95	Angry: 0.95 Neutral: 0.84 Happy: 0.85 Sad: 0.91	88.67%

In an experiment it shows that the sample has been 88.67% accurately classified and has been misclassified by 11.33%. The low accuracy is due to less number of datasets. Experiment-5 shows better accuracy all the experiment conducted previously due to training with more no. of dataset and epochs. The default value of threshold for confusion matrix We have used is 0.5.

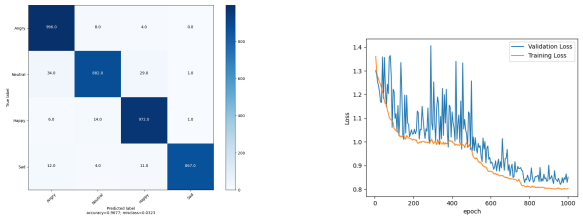


(a) confusion matrix **(b)** loss

Figure 6: Confusion Matrix (a) and loss (b) in an epoch of 1000 for 2D Model

All the Experiments were conducted by taking 90% Training and 10% testing dataset for an epoch of 500 and 1000. The result obtained were shown in figure 7 for every experiment conducted. From experiment-3, below shows that the sample has been 96.77% accurately classified and has been misclassified by 3.23%. It shows higher accuracy due to more number of data in the dataset and more number of epochs.

From graph, we have seen that the loss has been decreases with an increase in an epoch and increase in datasets as well. The Training and validation loss tends to converge towards zero. This is due to increase in number of epoch as the model learns itself

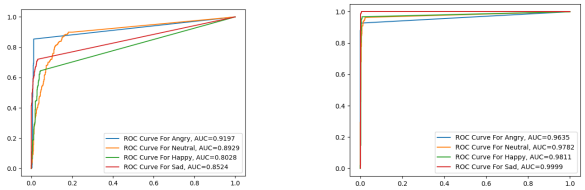


(a) confusion matrix **(b)** loss

Figure 7: Confusion Matrix (a) and loss (b) in an epoch of 1000 for 2D Model

from training samples Hence it can be seen that the training as well as testing loss has been gradually decreases and converges towards a zero along with the increase in number of epochs and size of dataset.

The ROC curve is plotted for an augmented data having Angry: 1048, Neutral: 908, Happy: 1016 and Sad: 869 Samples in an epoch of 500 and 1000 for both 2D and 3D Model and area under curve is analyzed. The x-axis of curve denotes False Positive rate and y-axis denotes True Positive Rate.



(a) roc 3d **(b)** roc 2d

Figure 8: ROC curve for 3D Model (a) and 2D Model (Left) in an epoch of 1000

From the figure 7 it was seen that 2D Model classifies True Positive and False Positive Data very well than 3D Model. The 2D Model classifies very well after increasing no. of epoch from 500 to 1000.

Table 2: Class-wise AUC for both 2D and 3D Model

Exp.	Model	Epoch	Class-wise Dataset	Area Under Curve	Comments
1.	3D	500	Angry: 1048 Neutral: 908 Happy: 1016 Sad: 869	Angry: 0.9209 Neutral: 0.8143 Happy: 0.7784 Sad: 0.8027	The Model Classifies Angry data very well
2.		1000	Angry: 1048 Neutral: 908 Happy: 1016 Sad: 869	Angry: 0.9197 Neutral: 0.8929 Happy: 0.8028 Sad: 0.8524	The Model classifies rest of 3 classes good on increasing epoch
3.	2D	500	Angry: 1048 Neutral: 908 Happy: 1016 Sad: 869	Angry: 0.9216 Neutral: 0.81 Happy: 0.8395 Sad: 0.5851	The Model Classifies Angry data very well
4.		1000	Angry: 1048 Neutral: 908 Happy: 1016 Sad: 869	Angry: 0.9635 Neutral: 0.9782 Happy: 0.9811 Sad: 0.999	The Model classifies All classes very good on increasing epoch.

7. LIMITATIONS and FUTURE ENHANCEMENT

Currently, the model can only produce a single persons' gait sequences. The quality of the video and the posture extraction technique both influence how accurate the categorization system is. We want to expand the strategy to handle crowd or multi-person videos as a future enhancement. This work can be enhanced by using more dataset and variety of dataset in order to increase the robustness for emotion recognition. We will enhance our research work with more dataset with more attributes for better accuracy. In future we are planning to detect emotions from other parameters in addition to gait like facial expression and body gesture. It would be much better if we are able to get 3 dimensional coordinates from by any means. At the end of the research it is expected to detect the human emotion of an individuals from his/her gait patterns.

8. Conclusion

Both 2D and 3D model is able to detect the human emotion of an individuals from his/her gait patterns. The perfect accuracy for 3D model is 88.67% and 96.77% for a 2D model. The both model was trained with same number of augmented data. The model accuracy can be further increased by taking more dataset, training the model with more number of epoch and hyper parameter tuning. The accuracy for a 3D model is slightly improving and almost same with Bhattacharya STEP Model (88%), but the accuracy for a 2D Model is very promising and novel method.

Acknowledgments

The authors are thankful to all the people who were directly or indirectly related during the research and experimentation.

References

[1] Uttaran Bhattacharya, Christian Roncal, Trisha Mittal, Rohan Chandra, Kyra Kapsaskis, Kurt Gray, Aniket Bera, and Dinesh Manocha. Take an emotion walk: Perceiving emotions from gaits using hierarchical attention pooling and affective mapping. In *Computer Vision – ECCV 2020*, pages 145–163. Springer International Publishing, 2020.

[2] Shihao Xu, Jing Fang, Xiping Hu, Edith Ngai, Wei Wang, Yi Guo, and Victor C. M. Leung. Emotion recognition from gait analyses: Current research and future directions, 2020.

[3] Kewal Krishan, Tanuj Kanchan, and John Dimaggio. Emergence of forensic podiatry–a novel sub-discipline of forensic sciences. *Forensic science international*, 06 2015.

[4] Uttaran Bhattacharya, Trisha Mittal, Rohan Chandra, Tanmay Randhavane, Aniket Bera, and Dinesh Manocha. STEP: Spatial temporal graph convolutional networks for emotion perception from gaits. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(02):1342–1350, apr 2020.

[5] A. S. M. Hossain Bari and Marina L. Gavrilova. Artificial neural network based gait recognition using kinect sensor. *IEEE Access*, 7:162708–162722, 2019.

[6] Birch Nirenberg M, Vernon W. A review of the historical use and criticisms of gait analysis evidence. 2018.

[7] Leslie Lamport. *Forensic Gait Analysis*. Treasure Island (FL): StatPearls Publishing USA, May, 2020.