Vision Based Multi-Modal Framework For Human Activity Recognition

Sarthak Jain - 191IT145
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: sarthak94511@gmail.com

Yash Gupta - 191IT158
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: guptayash1104@gmail.com

Rishit - 191IT141
Information Technology
National Institute of Technology Karnataka
Surathkal, India 575025
Email: rishit.191it141@nitk.edu.in

***Abstract*—Human activity recognition is important in many ways. Intelligent systems for video surveillance, public security, health care, and home monitoring are being developed.Humans' quality of life and security can be improved by recognising actions.In order to avoid unsafe situations,today automated, intuitive, and real-time technologies are typically required to recognise human actions and reliably identify odd behaviours.Human activity recognition (HAR) systems use data from many types of sensors to attempt to automatically recognise and analyse human actions.In this paper, we investigate how to create a fully-working and deployable system using a fusion of two modalities (RGB and skeletal data) to provide a comprehensive multi-modal framework for recognising human activities. Using graphical representations derived from a combination of dynamic RGB images and skeleton data representations, spatial information, body shape/posture, and time progression of actions are highlighted.As a result, each video is represented by two graphics that sum up the action. To extract significant features from these newly produced images, our method uses transfer learning from pre-trained models. Following that, we apply Principal Component Analysis and Linear discriminant analysis (LDA) to fuse retrieved features and train a neural network to categorise activities from visual descriptive photos. On the public UTD-MHAD datasets, experimental results confirmed the robustness of our feature-fusion architecture, which allows us to capture highly significant features and achieve the best performance and conlcusive results.**

***Keywords*—** HAR, RGB, Multi-Modal, UTD_MHAD, LDA

I. Introduction

The method of identifying and categorising a sequence of recorded data from ubiquitous or optical sensors into well-defined basic activities is known as human activity recognition (HAR). The process of identifying a particular action, also known as activity detection, entails temporally localising the movements of a person in the scene. While activity classification is the process of identifying the nature of a person's movements using some spatial and temporal cues or any other meaningful aspects that best describe the ongoing actions and assigning it to the appropriate class. Due to its extensive application domains, Vision-based HAR has become a very active research topic in computer vision and image processing. Automatic video surveillance, public security, virtual and augmented reality, and home monitoring are some of the use cases of Human Activity Recognition. Furthermore, advances in sensing technologies have inspired the creation of intelligent real-time systems that have the potential to influence the development of efficient human activity identification systems and improve people's quality of life and security.

A. Motivations

1) Dynamic image construction for RGB and retrieving skeleton images based on joint angle and location.
2) Transfer learning for feature extraction from above generated images.
3) Combining features from generated images.
4) Applying dimensionality reduction techniques on the data.
5) Training various Machine Learning and Deep Learning algorithms on the data for classifying the human action.

The majority of vision-based HAR research focuses on classifying activities using a single sensor modality. Due to environmental factors such as lighting, perspective changes, and cluttered background, degrades this results in certain limits when differentiating complicated operations . It is critical to leverage more than one modality in order to produce good results and enable robust HAR systems, and different fusion mechanisms are being investigated to this goal. In order to obtain high recognition accuracy, in this work we strive to merge two modalities from RGB and skeleton data for vision-based HAR. The proposed architecture combines the features from these two sets of data from the modalities. For RGB videos, an approximated rank pooling method is used to created dynamic images that will summarize the video. The skeleton data is utilised to construct pictures that encode the locations of the skeleton joints between video frames and so depict the action's temporal aspect. The process of feature extraction is then performed on these newly created images using the pre-trained state-of-the-art model. Then the features are fused together to get the final features. Due to large number of features, dimensionality reduction techniques are used to significantly reduce the dimensions of the data thus leading to better performance of the final model. Many Machine Learning and Deep Learning models are trained on the data and comparative study is done to choose the suitable model.

B. Innovations

- Fusion of feature extracted from dynamic RGB and skeltal data (euclidean based features extraction)
- Use of Linear Disciminant Analysis for dimensionality reduction after the fusion.
- Use of mobileNet as a main feature extractor from dynamic RGB images created from the RGB videos.

II. Related Work

Through [1] we learnt about what are multiple ways of human activity recognition. We learnt about the usage of RGB images, skeletal data, we also learnt about other forms as well. We also learnt about the usage of various models on these modes as mentioned above. Through [2] we studied about human activity identification using RGB data and also about human activity identification using skeletal data. Separate models were used and results were compared. Skeletal feature extraction in this model is also different then the one that we used. This uses pre-trained deep learning model for the feature extraction. Through [3] we learnt about 3d action recognition with respect to the view point, similar concept we have used in our project where feature extraction for skeltal data is done with respect to a fixed point. Through [4] we learnt about creation of dynamic images from the RGB videos as shown in Fig. 1 and than performing feature extraction on them and making classification on the extracted features. Through [5] we learnt about video modeling and how the data can effectively obtained from videos. We also learnt about methods for activity recognition from videos. Technique used in this paper is computationally expensive when compared with our multi modal implementation. Through [6] we learnt about the implementation of multimodal for human activity recognition, we saw comparison between various implementation of multi modals and also comparison was shown when different datasets are used. In [7] we saw one more method for implementation of multimodal. RGB videos data and depth data is used in this paper. Pretrained model is used for the feature extraction of both types of data than a model is used to make classification. Accuracy comparison is made on training model with different datasets. Through [8] we learnt about the evaluation of multimodal. We learnt about methods like canonical correlation analysis, linear discriminant analysis. We also saw comparisions between above mentioned fusion schemes. Final comparison is also made by training models on various datasets. In [9] we saw a comparison based study where multiple models are compared along with this comparisons are made by keeping model fixed and varying datasets. UTD-MHAD datasets that we have used for this project is also been showed here. In [10] we saw DMMS - based multiple features fusion for human action recognition and also comparison was shown when different datasets are used. Mobile-Net is used for feature extraction in this implementation. From [11],[12] and [13] we saw implementation of multi modal and also about the extraction of skeltal data from RGB videos and performing classification on them. From [14] and [15] we saw effective retrieval of videos for conversion into various forms without letting memory to overflow. Comparative study is shown in Table I.

III. Methodology

Dataset: Multi-Modal Human Activity Detection involves detection of activities like right arm swipe in both direction that is left and right, waving right hand, clapping, throw from right arm, cross arms in chest, shoot similar to that in basketball, drawing shapes like X, circle (both clockwise and counter-clockwise), triangle from right hand, bowling with right hand, sports activities like bowling, boxing, baseball swing, tennis swing by right hand, curling of arms, serve similar to game of tennis, pushing with both hands, action such as pickup and throw, knock on door, catching an object using right hand, jogging, walking, sit to stand, stand to sit, two arms stretch and left foot forward. The dataset used here is UTD-MHAD which have bot RGB and skeletal data. A Kinect camera is used to obtain the dataset. Videos are recorded at 30 frames per second. All above mentioned 27 actions are performed by 8 subjects, 4 males and 4 females. Each action is performed by subjects 4 times. Total data points in the dataset are 864.

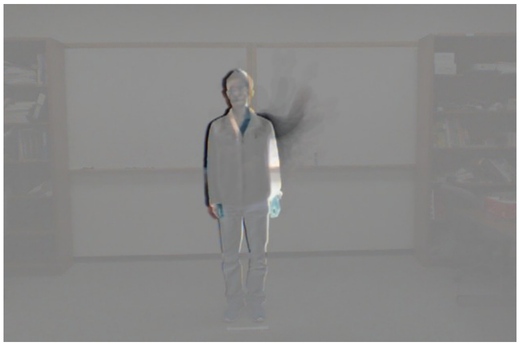
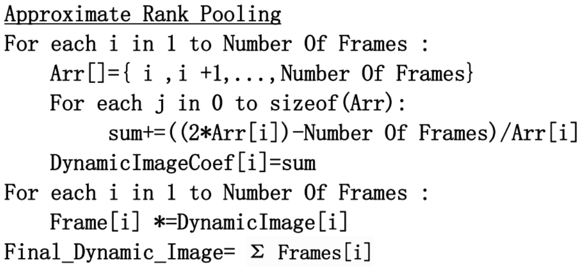


Fig. 1: Dynamic RGB



```
Approximate Rank Pooling
For each i in 1 to Number Of Frames :
    Arr[]={ i ,i +1,...,Number Of Frames}
    For each j in 0 to sizeof(Arr):
         sum+=((2*Arr[i])-Number Of Frames)/Arr[i]
    DynamicImageCoef[i]=sum
For each i in 1 to Number Of Frames :
    Frame[i] *=DynamicImage[i]
Final_Dynamic_Image= Σ Frames[i]
```

Fig. 2: Approximate Rank Pooling

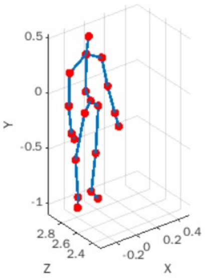


Fig. 3: Skeletal Image

Multi-Modal Human Activity Detection involves the merging of multiple modes by using some mathematical operations in order to obtain better results.Here the modes used are RGB videos and skeletal data, which after processing like feature extraction is merged by performing some mathematical operations and then classification is performed.

The first step involves preprocessing of RGB videos to form dynamic images.A technique called approximate rank pooling as shown in Fig. 2 is used. Example of Skeletal Image is shown in Fig. 3.

The overview of the methodology is shown in Fig. 4.

Once the dynamic images are obtained than feature extraction is performed using the Mobile Net model which is a pretrained model on ImageNet Dataset well capable of extracting features, detecting

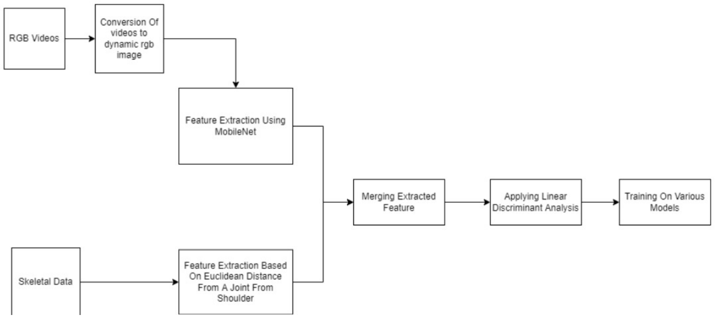

RGB Videos
Conversion Of videos to dynamic rgb image
Feature Extraction Using MobileNet
Merging Extracted Feature
Applying Linear Discriminant Analysis
Training On Various Models
Skeletal Data
Feature Extraction Based On Euclidean Distance From A Joint From Shoulder

Fig. 4: Methodology Diagram

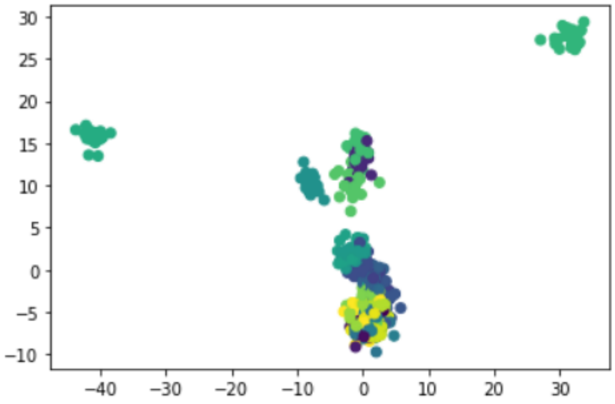


Fig. 5: LDA-2D

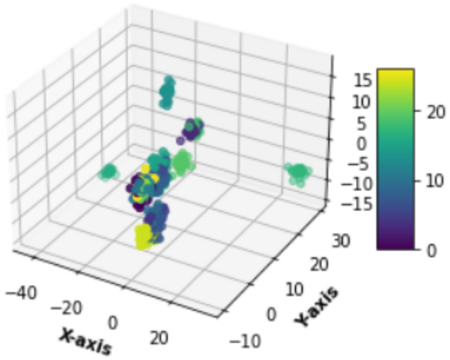
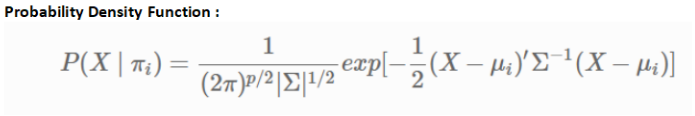


Fig. 6: LDA-3D



Probability Density Function :

$$P(X \mid \pi_i) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp[-\frac{1}{2}(X-\mu_i)'\Sigma^{-1}(X-\mu_i)]$$

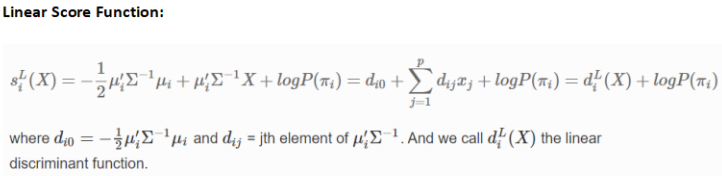
Fig. 7: Probability Density Function

edges and shapes. On the top of Extraction part of MobileNet model an average pooling layer is added. These Extracted Features will later be combined with the extracted features of skeletal data. For comparison purposes, in addition to above mentioned architecture, a dense layer with 27 neurons and a sigmoid function is used so that classification results can be obtained for the dynamic images obtained by performing above mentioned calculations. Learning rate is fixed to 0.0001 and binary cross-entropy as the loss function for the model, training is done till 240 epochs. The results obtained from this stage will be compared to the results from the other stages. Since the skeletal data consists of coordinates in X, Y, Z coordinate axes of all the joints, any movement will have a unique set of coordinates in euclidean space. So the shoulder center is considered here and distance of all the joints from the shoulder center is stored in the array. Euclidean distance is considered here in X,Y,Z coordinate axis. The obtained array can be considered as the array of extracted features of skeletal data. For comparison purposes, classification is made on obtained features. Algorithms used are XGBoost Classifier, ExtraTrees Classifier, Decision Tree, LightGBM Classifier results obtained will be later used to compare from multimodal that will be obtained on the fusion of extracted features obtained using above mentioned techniques. For feature fusion both the vectors containing the extracted features are concatenated and since the data is obtained from same subjects I.e. dynamic RGB and skeletal data merged is of same subject which implies data is highly correlated since dynamic images and skeletal data feature for same type of motion will result in highly correlated columns so dimensionality reduction can be performed by taking data to different feature space thus reducing the correlation and making the training faster and providing better classification. Techniques used here for dimensionality reduction are principal component analysis (PCA) and linear discriminant analysis (LDA).

Maths behind principal component analysis :

1) Mean of each dimension of whole dataset is calculated.
2) Calculating the covariance matrix for the entire dataset.
3) Calculating the eigenvectors and corresponding eigenvalues.
4) Sort the eigenvectors on basis of eigenvalues in decreasing order.
5) Use the final matrix to obtain data into the new vector space.

In Linear Discriminant Analysis decision is taken based on the linear score function mentioned below. Linear Score Function is function of the population means and pooled covariance-variance matrix. Logarithm of normal probability density function (Fig. 7) is also involved in the calculation of the linear score function (Fig. 8). Unknown parameters mew and sigma are estimated by using the values from the training dataset. The data visualized after applying LDA is shown in Fig. 5 and Fig. 6. New Data is obtained and trained on various machine learning algorithms mentioned above, and better results were obtained when compared to results from single modes. Results obtained are also compared with neural network with VGG-16 extractor and trained on imagenet but multi-modal outperformed it too. Algorithms used after feature extraction in this case are XGBoost Classifier, ExtraTrees Classifier, Decision Tree, LightGBM Classifier. A comparison-based study shows Use of multi modularity not only speeds up the training process but also out performs single modes of human activity recognition.



Linear Score Function:

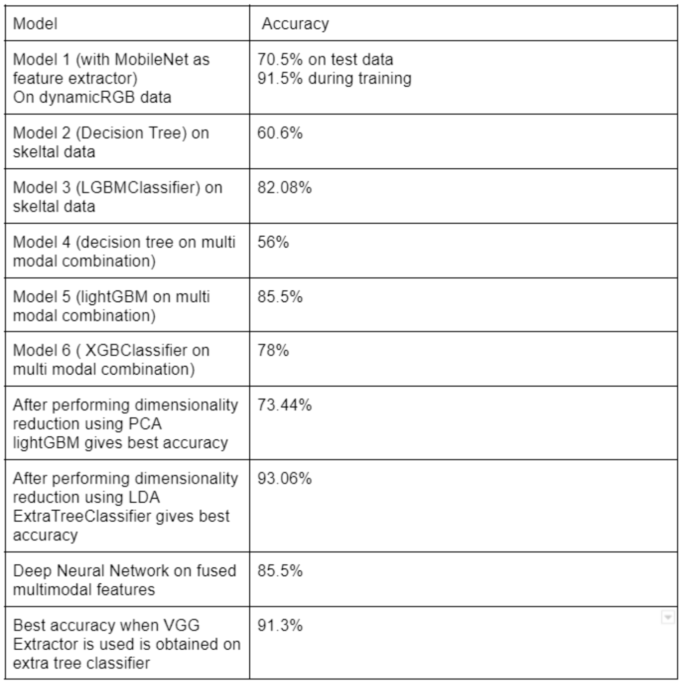
$$s_i^L(X) = -\frac{1}{2}\mu_i'\Sigma^{-1}\mu_i + \mu_i'\Sigma^{-1}X + logP(\pi_i) = d_{i0} + \sum_{j=1}^{p} d_{ij}x_j + logP(\pi_i) = d_i^L(X) + logP(\pi_i)$$

where $d_{i0} = -\frac{1}{2}\mu_i'\Sigma^{-1}\mu_i$ and d_{ij} = jth element of $\mu_i'\Sigma^{-1}$. And we call $d_i^L(X)$ the linear discriminant function.

Fig. 8: Linear Score Function

IV. Results and Analysis

Human Activity in Multiple Modes Detection includes activities such as right arm swipes in both left and right directions, waving



Model	Accuracy
Model 1 (with MobileNet as feature extractor) On dynamicRGB data	70.5% on test data 91.5% during training
Model 2 (Decision Tree) on skeltal data	60.6%
Model 3 (LGBMClassifier) on skeltal data	82.08%
Model 4 (decision tree on multi modal combination)	56%
Model 5 (lightGBM on multi modal combination)	85.5%
Model 6 (XGBClassifier on multi modal combination)	78%
After performing dimensionality reduction using PCA lightGBM gives best accuracy	73.44%
After performing dimensionality reduction using LDA ExtraTreeClassifier gives best accuracy	93.06%
Deep Neural Network on fused multimodal features	85.5%
Best accuracy when VGG Extractor is used is obtained on extra tree classifier	91.3%

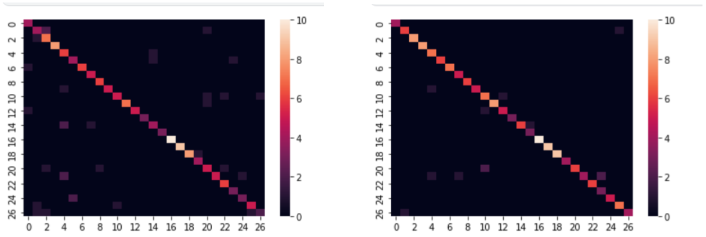
Fig. 9: Accuracy summary of various models

right hand, clapping, throw from right arm, cross arms in chest, shoot like in basketball, drawing shapes like X, circle (both clockwise and counter-clockwise), triangle from right hand,bowling with right hand, sports activities such as bowling, boxing,baseball swing,tennis swing by right hand,curling of arms,serve similar to game of tennis,pushing with both halves, UTD-MHAD is the dataset used here, which contains both RGB and skeletal data. The dataset was collected using a Kinect camera. Videos are captured at a frame rate of 30 frames per second. All of the 27 acts listed above are carried out by eight individuals. 4 males and 4 females.Each action is performed by subjects 4 times.Total data points in the dataset are 864.

Dynamic RGB (Red Green Blue) image which captures all the frame's actions into a single frame are created to assist action recognition.

The skeletal image contains the skeletal joints of the subject from which the joints co-ordinates are obtained in order to calculate the feature vectors.

Fig. 9 shows the accuracy scores of different models trained on the data.

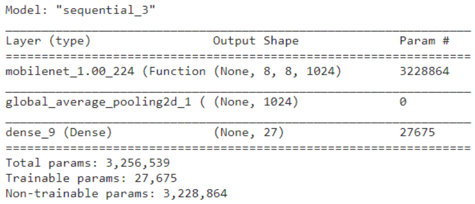


(a) LightGBM on seletal data

(b) LightGBM on multi-modal data after LDA

Fig. 10

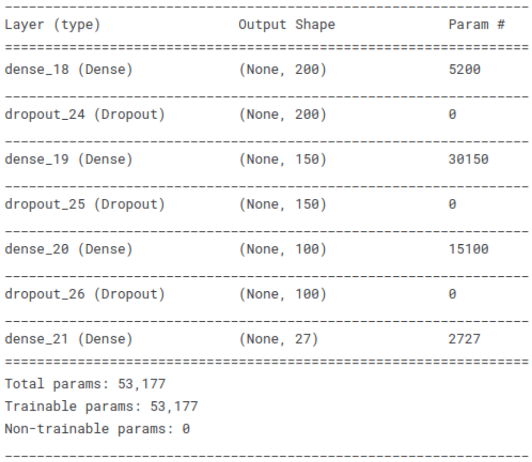
Mobilenet has been used to extract features from Dynamic RGB images which had a total of 3,256,539 parameters out of which 27,675 parameters were trainable and 3,228,864 were non trainable. As show



```
Model: "sequential_3"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
mobilenet_1.00_224 (Function (None, 8, 8, 1024)        3228864
_________________________________________________________________
global_average_pooling2d_1 ( (None, 1024)              0
_________________________________________________________________
dense_9 (Dense)              (None, 27)                27675
=================================================================
Total params: 3,256,539
Trainable params: 27,675
Non-trainable params: 3,228,864
```

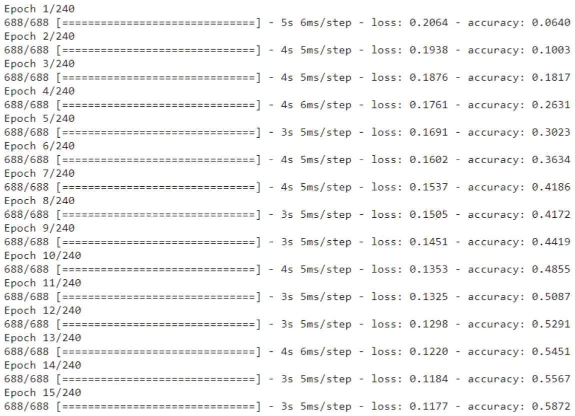
Fig. 11: Model 1 Deep Neural Network

in Fig. 11 and Fig. 12, Neural Network was created to train the model on the Dynamic RGB and skeletal data that was fused together in which the neural network had a total of 53,177 parameters out of which 53,177 parameters were trainable and 0 were non trainable.The accuracy score obtained on the test data was 85.54 percent on the data fused through LDA.



```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
dense_18 (Dense)             (None, 200)               5200
_________________________________________________________________
dropout_24 (Dropout)         (None, 200)               0
_________________________________________________________________
dense_19 (Dense)             (None, 150)               30150
_________________________________________________________________
dropout_25 (Dropout)         (None, 150)               0
_________________________________________________________________
dense_20 (Dense)             (None, 100)               15100
_________________________________________________________________
dropout_26 (Dropout)         (None, 100)               0
_________________________________________________________________
dense_21 (Dense)             (None, 27)                2727
=================================================================
Total params: 53,177
Trainable params: 53,177
Non-trainable params: 0
_________________________________________________________________
```

Fig. 12: Neural network

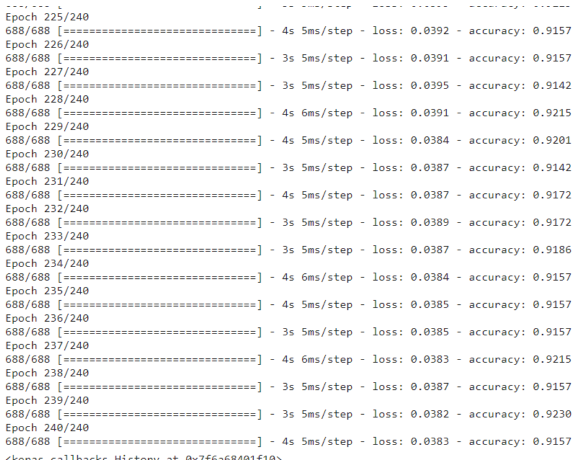


```
Epoch 1/240
688/688 [==============================] - 5s 6ms/step - loss: 0.2064 - accuracy: 0.0640
Epoch 2/240
688/688 [==============================] - 4s 5ms/step - loss: 0.1938 - accuracy: 0.1003
Epoch 3/240
688/688 [==============================] - 4s 5ms/step - loss: 0.1876 - accuracy: 0.1817
Epoch 4/240
688/688 [==============================] - 4s 6ms/step - loss: 0.1761 - accuracy: 0.2631
Epoch 5/240
688/688 [==============================] - 3s 5ms/step - loss: 0.1691 - accuracy: 0.3023
Epoch 6/240
688/688 [==============================] - 4s 5ms/step - loss: 0.1602 - accuracy: 0.3634
Epoch 7/240
688/688 [==============================] - 4s 5ms/step - loss: 0.1537 - accuracy: 0.4186
Epoch 8/240
688/688 [==============================] - 3s 5ms/step - loss: 0.1505 - accuracy: 0.4172
Epoch 9/240
688/688 [==============================] - 3s 5ms/step - loss: 0.1451 - accuracy: 0.4419
Epoch 10/240
688/688 [==============================] - 4s 5ms/step - loss: 0.1353 - accuracy: 0.4855
Epoch 11/240
688/688 [==============================] - 3s 5ms/step - loss: 0.1325 - accuracy: 0.5087
Epoch 12/240
688/688 [==============================] - 3s 5ms/step - loss: 0.1298 - accuracy: 0.5291
Epoch 13/240
688/688 [==============================] - 4s 6ms/step - loss: 0.1220 - accuracy: 0.5451
Epoch 14/240
688/688 [==============================] - 3s 5ms/step - loss: 0.1184 - accuracy: 0.5567
Epoch 15/240
688/688 [==============================] - 3s 5ms/step - loss: 0.1177 - accuracy: 0.5872
```

Fig. 13: Epoch data for first 15 epochs for model1

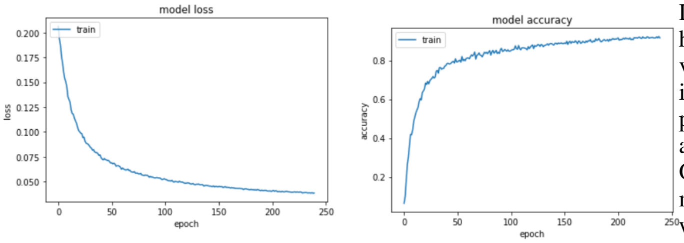
Fig. 12 and Fig. 13 represents the epoch data for the first 15 and last 15 epochs respectively in the neural network trained.

In order to fuse the features of Dynamic RGB data and skeletal data ,LDA was used and the data points were mapped onto a 2D plane as shown in Fig 5. As shown in Fig. 6, a 3D plot of these data points have been plotted onto a 3 dimensional plane using the Linear Discriminant Analysis.



```
Epoch 225/240
688/688 [==============================] - 4s 5ms/step - loss: 0.0392 - accuracy: 0.9157
Epoch 226/240
688/688 [==============================] - 3s 5ms/step - loss: 0.0391 - accuracy: 0.9157
Epoch 227/240
688/688 [==============================] - 3s 5ms/step - loss: 0.0395 - accuracy: 0.9142
Epoch 228/240
688/688 [==============================] - 4s 6ms/step - loss: 0.0391 - accuracy: 0.9215
Epoch 229/240
688/688 [==============================] - 4s 5ms/step - loss: 0.0384 - accuracy: 0.9201
Epoch 230/240
688/688 [==============================] - 3s 5ms/step - loss: 0.0387 - accuracy: 0.9142
Epoch 231/240
688/688 [==============================] - 4s 5ms/step - loss: 0.0387 - accuracy: 0.9172
Epoch 232/240
688/688 [==============================] - 3s 5ms/step - loss: 0.0389 - accuracy: 0.9172
Epoch 233/240
688/688 [==============================] - 3s 5ms/step - loss: 0.0387 - accuracy: 0.9186
Epoch 234/240
688/688 [==============================] - 4s 6ms/step - loss: 0.0384 - accuracy: 0.9157
Epoch 235/240
688/688 [==============================] - 4s 5ms/step - loss: 0.0385 - accuracy: 0.9157
Epoch 236/240
688/688 [==============================] - 3s 5ms/step - loss: 0.0385 - accuracy: 0.9157
Epoch 237/240
688/688 [==============================] - 4s 6ms/step - loss: 0.0383 - accuracy: 0.9215
Epoch 238/240
688/688 [==============================] - 3s 5ms/step - loss: 0.0387 - accuracy: 0.9157
Epoch 239/240
688/688 [==============================] - 3s 5ms/step - loss: 0.0382 - accuracy: 0.9230
Epoch 240/240
688/688 [==============================] - 4s 5ms/step - loss: 0.0383 - accuracy: 0.9157
<keras.callbacks.History at 0x7f6a68401f10>
```

Fig. 14: Epoch data for last 15 epochs for model1



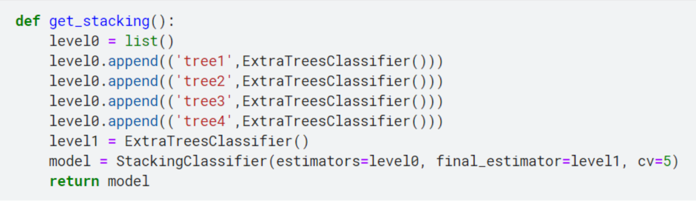
(a) Loss

(b) Accuracy

Fig. 15: Model

Fig. 15 shows the loss and accuracy of the finally trained model.

Finally Fig. 16 and Fig. 17 shows the stacking algorithm and result of stacking on multi modal fused data.



```
def get_stacking():
    level0 = list()
    level0.append(('tree1',ExtraTreesClassifier()))
    level0.append(('tree2',ExtraTreesClassifier()))
    level0.append(('tree3',ExtraTreesClassifier()))
    level0.append(('tree4',ExtraTreesClassifier()))
    level1 = ExtraTreesClassifier()
    model = StackingClassifier(estimators=level0, final_estimator=level1, cv=5)
    return model
```

Fig. 16: Stacking algorithm

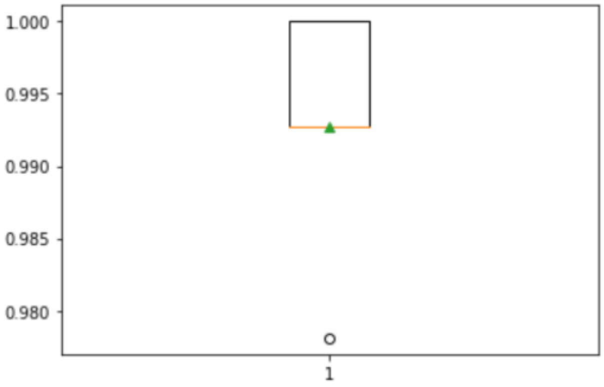


Fig. 17: Results on stacking on multi modal fused data

V. Conclusion

In this study, we describe a vision-based multi-modality fusion technique for recognising human activities. RGB dynamic images and skeleton images are created using RGB images and skeleton joint data respectively. Dynamic RGB (Red Green Blue) image captures all the frame's actions into a single frame. The pre-trained models that allow us to obtain relevant features from the picture sets are then used to generate features from these constructed visual images.For feature fusion, both the vectors containing the extracted features are concatenated, and because the data is from the same subjects (i.e. dynamic RGB and skeletal data merged), it is highly correlated. Because dynamic images and skeletal data feature for the same type of motion will result in highly correlated columns, dimensionality reduction was performed by moving the data to a different feature space, reducing the correlation and making the training faster and providing better classification. The feature fusion was done using Linear Discriminant Analysis and PCA.These were used to choose highly discriminative features from the two feature vectors for each video sequence. The feature fusion vectors that arise are then sent into a deep neural network to recognise and classify activities on the publicly accessible UTD-MHAD dataset. We then test our method and record the corresponding recognition accuracy score.

Our work in this project suggest that the outcomes of our proposed method are very accurate and can be used to recognise the postures with very high precision. For the UTD-MHAD dataset, we were able to achieve high recognition accuracy and beat the the accuracy scores of the current work in the same field thus establishing that combining two modes - Dynamic RGB and skleletal data can recognise human actions accurately.

Acknowledgement

We would like to express our gratitude to Dr. Dinesh Naik for his assistance in making our project successful, along with the IT department, NITK, for providing us the opportunity to work on this project.

References

[1] S. Jiang, X. Ding, F. Chen, E. Tan, and D. Zhang, "Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions," *Proceedings of the 4th conference on USENIX Conference*, vol. 4, 2005.

[2] B. Marshall and C. Welborn, "A multimodal approach for human activity recognition based on skeleton and rgb data." *Journal of Computing Sciences in Colleges*, vol. 27, no. 2, 2011.

[3] M. Abrams, C. Standridge, G. Abdulla, S. Williams, and F. EA., "3d action recognition from novel viewpoints," *Proceedings of the 4th WWW conference. Boston, MA*, 2000.

[4] P. Andersen and N. Petersen, "Action recognition with dynamic image networks," *Computer Science 1993*, vol. 18, no. 2, 1995.

[5] G. Kastaniotis, E. Maragos, V. Dimitsas, C. Douligeris, and D. Despotis, "Modeling video evolution for action recognition," *Proceedings of the 15th IEEE international symposium on modeling, analysis, and simulation of computer and telecommunication systems. Istanbul, Turkey*, pp. 132–137, 2007.

[6] C. Chang, T. McGregor, and G. Holmes, "Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor," *Proceedings of the Asia-Pacific web conference. Hong Kong, China*, 1999.

[7] S. Selvakumar, S.-K. Sahoo, and V. Venkatasubramani, "Human activity recognition by fusion of rgb and depth," *Elsevier, Computer Communications*, 2004.

[8] Y. Smirlis, E. Maragos, and D. Despotis, "Evaluating fusion of rgb-d and inertial sensors for multimodal human action recognition," *Elsevier, Applied Mathematics and Computation*, vol. 1, 2006.

TABLE I: Comparitive Study

Authors	Methodology	Merits	Limitations
S. Jiang, X. Ding, F. Chen, E. Tan, and D. Zhang	Data Fusion	Data fusion and multiple classifier systems for human activity detection	CCA method applied
P. Andersen and N. Petersen,	Multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor	Dynamic RGB images creation	Fusion of multiple modes not defined
C. Siriopoulos and P. Tziogkidis	Mms-based multiple features fusion for human action recognition	Multiple features fusion defined	Accuracy very low which is improved by PCA

[9] K. Psounis and B. Prabhakar, "Survey on human activity detection," *IEEE/ACM Transactions on Networking*, vol. 1, 2002.

[10] C. Siriopoulos and P. Tziogkidis, "Dmms-based multiple features fusion for human action recognition," *A DEA approach Elsevier*, 2010.

[11] Jalal and M. Mahmood, "Vision-based human activity recognition system using depth silhouettes: A smart home system for monitoring the residents. journal of electrical engineering technology," 2019.

[12] O. Popoola and K. Wang, "Video-based abnormal human behavior recognition," 2012.

[13] K. Sahu and Zamir, "An end-to-end spatio-temporal attention model for human action recognition from skeleton data," 2016.

[14] de Souza Alves and Szczerbicki, "From knowledgebased vision systems to cognitive vision systems: a review," 2016.

[15] S. Maybank and A. Farooq, "A survey on visual content-based video indexing and retrieval." 2010.