Vision Based Human Activity Recognition

PROJECT SUPERVISOR

Ms. Farah Sadia

PROJECT CO SUPERVISOR

Dr. Jawwad Shamsi

PROJECT TEAM

Muhammad Owais Mushtaq	[18k-1177]
Muhammad Usman Umar	[18k-1069]
Shaharyar Amjad	[18k-1371]

Submitted in partial fulfilment of the requirements for the degree of Bachelor of Science
in
Computer Science.

FAST SCHOOL OF COMPUTING
NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES

KARACHI CAMPUS
June 2022

Project Supervisor	Ms. Farah Sadia
Project Team	Muhammad Owais Mushtaq [18k-1177] Muhammad Usman Umar [18k-1069] Shaharyar Amjad [18k-1371]
Submission Date	June 06, 2022

Ms. Farah Sadia _

Supervisor

Dr. Jawwad Shamsi _

Co-Supervisor

Dr. Zulfiqar Ali Memon _

Head of Department

FAST SCHOOL OF COMPUTING
NATIONAL UNIVERSITY OF COMPUTER AND EMERGING SCIENCES
KARACHI CAMPUS

ACKNOWLEDGEMENT

It is indeed with a great pleasure and immense sense of gratitude that we acknowledge the help of these individuals. We are highly indebted to our Director **Dr. Muhammad Atif Tahir,** FAST-National University of Computer and Emerging Sciences, for the facilities provided to accomplish this main project.

We would like to thank Dr. **Zulfiqar Ali Memon**, Head of the Department of Computer Science, FAST-National University of Computer and Emerging Sciences, for this constructive criticism throughout our project.

We feel elated in manifesting our sense of gratitude for our internal project guide **Miss. Farah Sadia, Lecturer, Department of Computer Science**, FAST-National University of Computer and Emerging Sciences. She has been a constant source of inspiration for us and we are very deeply thankful to her for his support and valuable advice. Plus, we are also very thankful to **Dr. Jawwad Shamsi, Dean, Department of Computer Science,** FAST-National University of Computer and Emerging Sciences, for his continuous guidance throughout our project timeline.

We are extremely grateful to our Departmental staff members, Lab technicians and Non-teaching staff members for their extreme help throughout our project.

Finally we express our heart-full thanks to all of our friends who helped us in successful completion of this project.

PROJECT ASSOCIATES

Muhammad Owais Mushtaq [18k-1177]

Muhammad Usman Umar [18k-1069]

Shaharyar Amjad [18k-1371]

ABSTRACT

Automated systems (particularly surveillance systems) are in demand for the classification of various human actions as the number of cameras grows day by day. Furthermore, human action recognition (HAR) has emerged as one of the most appealing study subjects across a wide range of computer vision applications. The precise detection of human activity on an uncertain basis, on the other hand, remains a mystery. There has been limited study on human action recognition processes in real situations, which drives us to pursue research in this application space. In this paper, we examined several approaches, including CNN+LSTM, Inception-V3+CNN, MobileNet, and Inception-V3. Experiments are carried out on the UCF-101 dataset to demonstrate the effectiveness of new models.

CONTENTS

Page

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

ANN	Artificial Neural Network
CV	Computer Vision
CNN	Convolutional Neural Network
DL	Deep Learning
DNN	Deep Neural Network
GAN	Generative Adversarial Network
GRU	Gated Recurrent Unit
HAR	Human Action/Activity Recognition
HCI	Human-Computer Interface
KDA	Kernel Discriminant Analysis
LDA	Linear Discriminant Analysis
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Neural Language Processing
PCA	Principal Component Analysis
RBD	Reduced Basis Decomposition
RNN	Recurrent Neural Network

This page intentionally left blank

INTRODUCTION

A HAR's purpose is to identify actions or activities carried out by a single person or a group of people. HAR has piqued the interest of several research groups in recent years, and it is now considered an active study area due to applications such as visual surveillance, HCI, education, medical, and abnormal activity recognition, and many others [26]. Activity recognition is critical to mankind because it records individuals' behaviour with data that computing systems may use to monitor, evaluate, and aid them in their daily lives [27]. Video-based systems and sensor-based systems are the two main types of HAR systems. Sensor-based systems use on-body or ambient sensors to recognize people's motions and log their activity traces. While Video-based systems use cameras to take images or videos to identify people's motion [1]. Automated solutions for the classification of such actions using computationally intelligent techniques are in demand due to an increase in the use of cameras.

Action Detection and Action Classification are two approaches used by the HAR framework (which is also divided into two sub methods: Action Representation and Interaction Representation). HAR is further subdivided into the following sections: A general stage to finish the procedure in video-based HAR is: a) Action representation – feature extraction and encoding, b) Dimensionality reduction techniques – original features are transformed by removing redundant information using different models such as PCA, RBD, LDA, and KDA, and c) action analysis-based HAR – action classification techniques are used, which primarily include traditional ML and DL techniques such as CNN, RNN, LSTM, GRU, and GAN [26].

RELATED WORK

TRADITIONAL ML-BASED HAR

ML approaches have been used in HAR in a number of earlier studies [25]. Feature extraction techniques such as time-frequency transformation [3], statistical approaches [4], and symbolic representation [5] are heavily used. The traits that are extracted, on the other hand, are carefully developed and heuristic. To successfully capture distinguishing traits for human actions, there were no uniform or systematic feature extraction methodologies.

DL-BASED HAR

In recent years, DL has shown remarkable success in modelling high-level abstractions from complex data [6] in a variety of fields, including CV, NLP, and speech processing. Along with the effectiveness [7, 8, 9] and inevitable evolution of DL in HAR, latest research is being done to solve specific challenges. However, Researchers are still sceptical about DL because of its spontaneous success, frantic innovation, and lack of theoretical underpinning.

The appeal of deep learning is its layer-by-layer structure, which allows it to scale learning from simple to abstract data. Because of DL's exceptional learning capabilities, the activity recognition system can evaluate multimodal data in depth for accurate recognition [28]. We've also seen how DNN has made significant improvements in video comprehension [10]. As a result, there has been a surge in interest in enhancing the efficiency of video models [11].

FEATURE EXTRACTION

A step-by-step strategy that includes feature representation using feature extraction techniques and action classification techniques can be applied for HAR. In image classification, feature extraction is a crucial step. It enables us to portray visual material as accurately as possible. Deep neural networks have a variety of architectures that encode features from various angles. CNNs, for example, are good at capturing the local connections of multimodal sensory input, and the translational invariance introduced by localization results in accurate recognition [12]. CNN architecture contains three steps for feature extraction: convolution, activation, and pooling. For image-based recognition tasks, CNN-based models are effective [2]. As a result, it can produce good results and identify spatial relationships from RGB data, as well as efficiently extract complex human movements with their temporal movements using various filters and pooling operations [13]. However, CNN has the disadvantage of requiring large training data and high parameter tuning [14], and it is also ineffective for modelling sequence data or time series problems. Recurrent Neural Networks (RNN) are useful for modelling data with temporal fluctuations [14], such as labelling sequences or time series, but they can suffer from the problem of vanishing gradient [14]. In the temporal domain, Long Short Term Memory (LSTM) can represent long-term contextual information, but it cannot extract spatial information [15]. However, CNN can be integrated with LSTM to incorporate both spatial and temporal characteristics [15]. Many researchers have worked on hybrid models that incorporate deep structures, such as Inception Network [29], Yolo, X3D [17], and I3D [19], among others. These hybrid models have been developed for use in a variety of applications, not just HAR.

CLASSIFICATION

The most recent trend in HAR hybrid models has been to focus on computational efficiency in order to expand to increasingly bigger datasets and be used in real-world applications. Hidden TSM [16], X3D [17], and AssembleNet [18], for example. There were two other tendencies previously; one of the initial attempts was to apply CNN to optical flow streams. Second, initiatives like I3D [19], ShowFast [11], and others used 3D convolutional kernels to model video temporal information. GAN is also an effective method for building semi-supervised classifiers, however these networks are difficult to train [20].

DATASET

When the amount of training data increases, deep learning algorithms usually improve in accuracy. This suggests that in order to develop successful models for video action recognition, we'll require large-scale annotated datasets [32]. Datasets are important for evaluating multiple algorithms for a certain goal, and task-specific algorithm evaluation is based on factors relevant to each dataset. [31]. Few datasets that we have reviewed for this research includes: HMDB51 [21], a collection of 7K footage organised into 51 action categories, each with a minimum of 101 clips, was released in 2011. There had already been a lot of study done on this dataset, with positive results. UCF101 [22], which was introduced in 2012, is larger and superior than HMDB51 since it has 13.32K YouTube videos spanning 101 categories of human behaviour. Because of its size and classes, it will demand more computing power than the previous one, but it will also produce more important results. Charades [23], a 2016 public dataset with continuous action films featuring 9848 videos (each video is 30 seconds) of 157 multi-label daily indoor activities performed by 267 different persons, is one of the largest public datasets with continuous action videos. Because of the length and variety of the activities, this is a difficult dataset to work with. HVU [24], a dataset for multi-label multi-task video comprehension, was released in 2020. There are 572K videos and 3,142 labels in this collection. For train, validation, and test, the official split has 481K, 31K, and 65K videos, respectively. Scene, object, action, event, attribute, and concept are the six task categories in this dataset. For each label, there are around 2; 112 samples. The length of the videos varies, with the longest being 10 seconds. Despite the fact that this dataset is the most recent and contains more features, processing it demands a lot of computing effort due to its vast size.

Table 1: Dataset Details

Dataset	Year	#Action Classes	#Samples	Average Length
HMDB51	2011	51	7K	~5s
UCF101	2012	101	13.32K	~6s
Charades	2016	157	9.84K	30.1s
HVU	2020	739	572K	10s

REQUIREMENTS

Table 2: Hardware Requirements

Requirements	Minimum	Preferred
GPU for training model	NVIDIA 2070	NVIDIA 3090
RAM	16 GB	64GB
HARD Drive	1 TB	Depends on the length of Dataset

Table 2: Software Requirements

Tools	Version
Python	3.10
Tensor Flow	2.7
Cuda	11.2
Anaconda (Jupyter Notebook)	4.10.3
Operating System	Windows 10 (Preferred)

DESIGN AND IMPLEMENTATION

DATASET SELECTION

To illustrate the feasibility of models, experiments are performed on the UCF-101 [22] dataset. UCF101, which was launched in 2012, is a benchmark dataset consisting of 101 different types of human actions and 13.32K YouTube videos (spanning 27 hours) (Figure. 1). Brushing Teeth, Cliff Diving, Floor Gymnastics, and Playing Cello are among the 51 new classes included in this extension of UCF50. Human Object Interaction, Body Motion Only, Human–Human Interaction, Playing Musical Instruments, and Sports are the categories. There are three different train–test splits available. But for most of our experiments, we have used the split ratio of Training, Validation, and Testing as 80:10:10.

Figure 1: A few example classes from UCF101 dataset

CNN+LSTM IMPLEMENTATION

We began our research with a basic hybrid deep learning model, CNN with LSTM. On the front end, CNN layers are added, followed by LSTM layers with a Dense layer on the output to create a CNN+LSTM. In Keras, we have created a CNN+LSTM model by first specifying the CNN levels, then wrapping them in a TimeDistributed layer, and last defining the LSTM and output layers. For brief, CNN+LSTM is an LSTM architecture intended primarily for sequence prediction issues including spatial inputs such as photos or videos. CNN layers for feature extraction on input data are paired with LSTMs to facilitate sequence prediction in the CNN+LSTM architecture. Convolution, activation, and pooling are the three processes in the CNN architecture for feature extraction. For image-based recognition tasks, CNN-based models are effective [2]. As a result, it can produce good results and identify spatial relationships from RGB data, as well as efficiently extract complex human movements with their temporal movements using various filters and pooling operations [13]. However, CNN has the

disadvantage of requiring large training data and high parameter tuning [14], and it is also ineffective for modelling sequence data or time series problems. LSTM, on either hand, can describe long-term contextual information in the temporal domain but not in the spatial domain [15]. However, CNN can be integrated with LSTM to incorporate both spatial and temporal characteristics [15]. As a result, we used UCF101 to train and test CNN+LSTM with the split ratio of 70: 30. All the parameters selected for training have been disclosed in table 3. Moreover, for feature extraction, we have used four fully connected ConvLSTM2D layers followed by Max-Pooling and Time-Distributed layers and a flatten layer at the end followed by Dense layer for classification (Figure 2)

Table 3: Details for parameters of CNN + LSTM

Technique	Value
Image size	64,64
Loss Function	Categorical Cross Entropy
Optimizer	Adam
Evolution Matrices	Accuracy
Number of Epochs	1000
Batch Size	32
Class Mode	Categorical
Validation Split	0.2
Shuffle	True

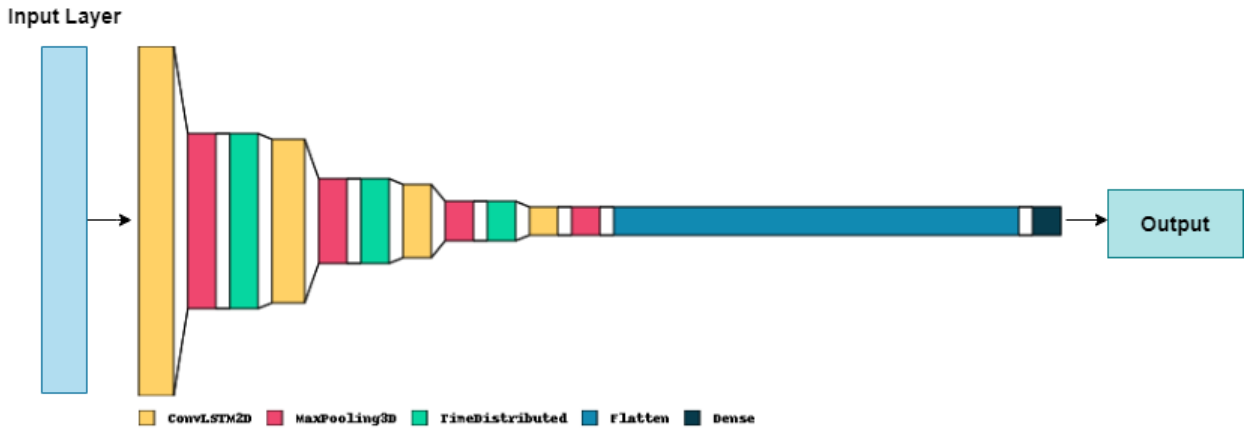

Figure 2: CNN+LSTM architecture diagram

INCEPTION-V3 IMPLEMENTATION

We chose Inception-V3 among the several DL models. The use of Inception-v3 for feature extraction was motivated by its high accuracy (70%) on the ImageNet dataset [29]. For greater model adaption, the Inception V3 model uses a number of approaches to optimise the network. It has a more extensive network than the Inception V1 and V2 models, but its speed is unaffected. It is less expensive in terms of computing. As regularizers, it employs auxiliary Classifiers. Auxiliary classifiers are used to improve the convergence of very deep neural networks. In very deep networks, the auxiliary classifier is primarily employed to tackle the vanishing gradient problem. In the early stages of the training, the addition of auxiliary classifiers had no effect. However, when compared to the network without auxiliary classifiers, the network with auxiliary classifiers had better accuracy in the end. As a result, in Inception V3 model design, the auxiliary classifiers operate as a regularizer. In addition, to reduce the number of model parameters and increase efficiency, the Inception-v3 has used multiple optimization techniques such as factoring large convolutions into a network of smaller convolutions, using batch normalisation with auxiliary classifiers, and reducing grid size by concatenating parallel pooling and convolutional layers [30]. Convolutions, average pooling, max pooling, dropouts, and fully linked layers are among the minor modules that make up the model. (Figure. 3) Other parameters such as Image size, Loss Function, Data Pre-processing steps, and others are shown in table 4. Furthermore, it had never been implemented on UCF101 previously.

Table 4: Parameters details for Inception V3

Technique	Value
Image size	224, 224
Loss Function	Categorical Cross Entropy
Optimizer	Adam
Evolution Matrices	Accuracy
Number of Epochs	100
Batch Size	32
Class Mode	Categorical
Rescale	1.0/255
Share Range	0.2
Zoom Range	0.2
Horizontal Flip	True

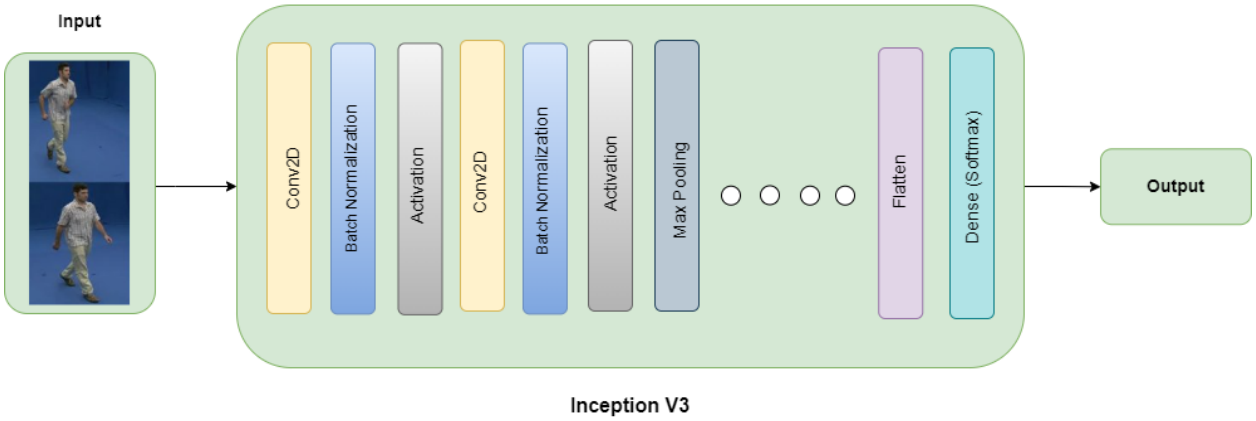

Figure 3: Inception-V3 architectural diagram

INCEPTION-V3 + CNN IMPLEMENTATION

When it comes to human-machine interfaces, video classification is critical since it aids in the analysis of diverse actions. Transfer learning approaches can assist us in producing accurate predictions. Transfer Learning strategies are crucial for increasing learning outcomes. Transfer learning is based on the idea of taking labelled data from a certain area of interest and using it to improve the performance of a machine learning algorithm in that domain. We expanded our research in order to better our testing findings. Using weights from "imagenet," we extracted features using the InceptionV3 pre-trained model and transferred them to the fully connected layer. We upgraded our previously implemented Inception-V3 (Figure. 3) after deploying it on UCF101 by freezing the first 172 levels and adding three CNN layers towards the end. (Figure. 4) Convolutional layers, batch normalisation layers, activation layers, pooling layers, and concatenation layers are used to connect many inception levels in each inception layer. However, the training parameters for this model were the same as those for Inception-V3, as shown in Table 4.

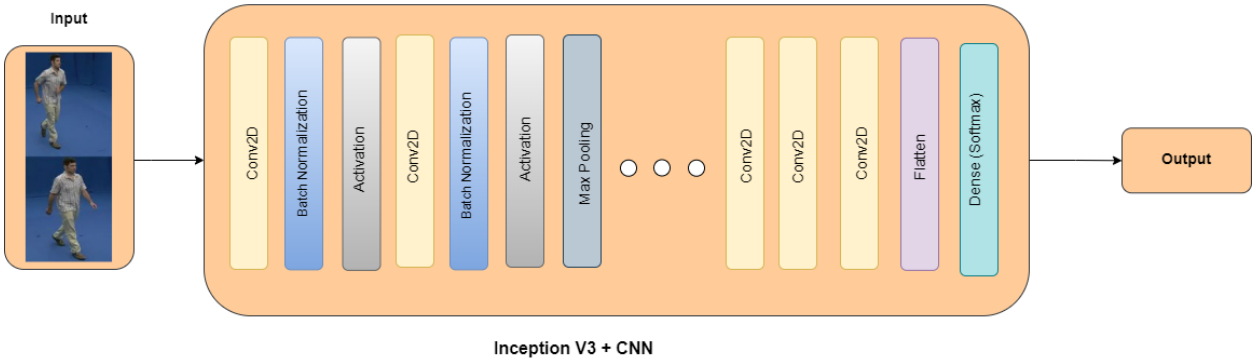

Figure 4: Inception-V3 + CNN architectural diagram

MOBILENET IMPLEMENTATION

To broaden our research, we also applied a streamlined architecture model (Figure. 5), Mobile Net, for better results. The goal of MobileNet is to develop lightweight deep neural networks by using depthwise separable convolutions. The convolution kernel or filter is applied to all of the channels of the input image in a normal convolutional layer by doing a weighted sum of the input pixels with the filter, then sliding to the next input pixels across the pictures. Only the first layer of MobileNet uses standard convolution. The depth wise separable convolutions, which are a combination of depthwise and pointwise convolutions, are the following layers. The depthwise convolution performs the convolution independently for each channel. The input channels are filtered using depthwise convolution. The pointwise convolution is the following phase, which is similar to conventional convolution but with a 1x1 filter. The goal of pointwise convolution is to combine the depthwise convolution's output channels to create additional features. As a result, the computing labour required is less than with traditional convolutional networks. However, we were unable to match Inception-V3 + CNN in terms of testing accuracy. However, we noticed that Mobile Net was faster than Inception-V3 in terms of training and testing during the training and testing phase. The training settings for this model were identical to those for Inception-V3, as shown in Table 4.

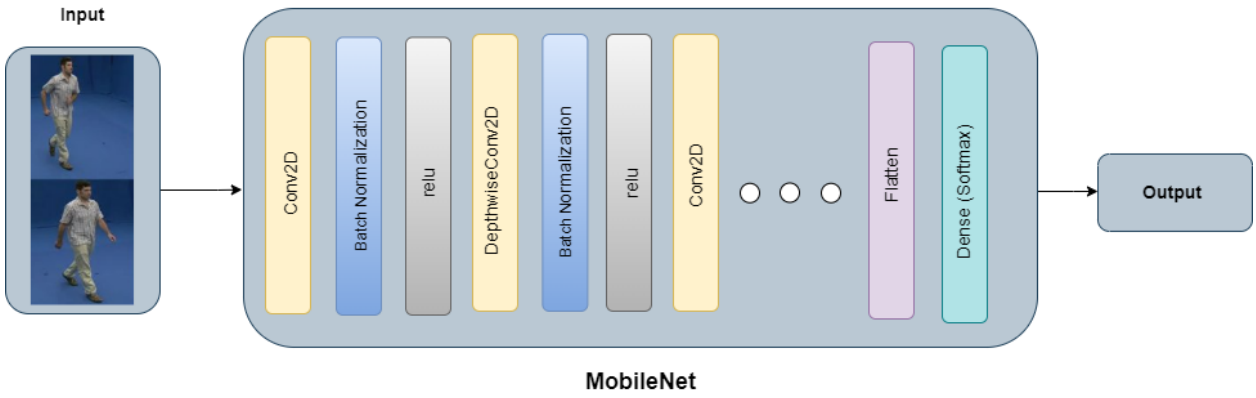

Figure 5: MobileNet architectural diagram

TESTING AND EVALUATION

Initially, we trained Inception-V3 over UCF101. Although our feature extraction accuracy was good throughout training and validation, we didn't get decent results during testing. As a result, we increased our implementation and trained MobileNet, which was faster during training and produced better results in both training and testing, but the testing results were still below expectations. After further investigation, we discovered that Inception-V3's low testing accuracy was related to its vast number of layers. As a result, we froze the first 172 layers before adding three fully connected layers at the end. And our efforts resulted in better results. Which was superior to both of the prior methods. As a result, we tried CNN+LSTM with fewer layers to see what the outcomes were, and while the validation accuracy was lower, we got the best testing accuracy out of all the models. Table.5 shows the evaluation accuracies of each of the four models, with a comparison provided in Figure 6.

Table 5: Models Summary

Model	Training Accuracy	Validation Accuracy	Testing Accuracy
Inception-V3	91.12	93.75	56.34
MobileNet	96.16	96.25	60.97
Inception V3 + CNN	85.56	88.34	74.4
CNN + LSTM	99.88	80.57	81.57

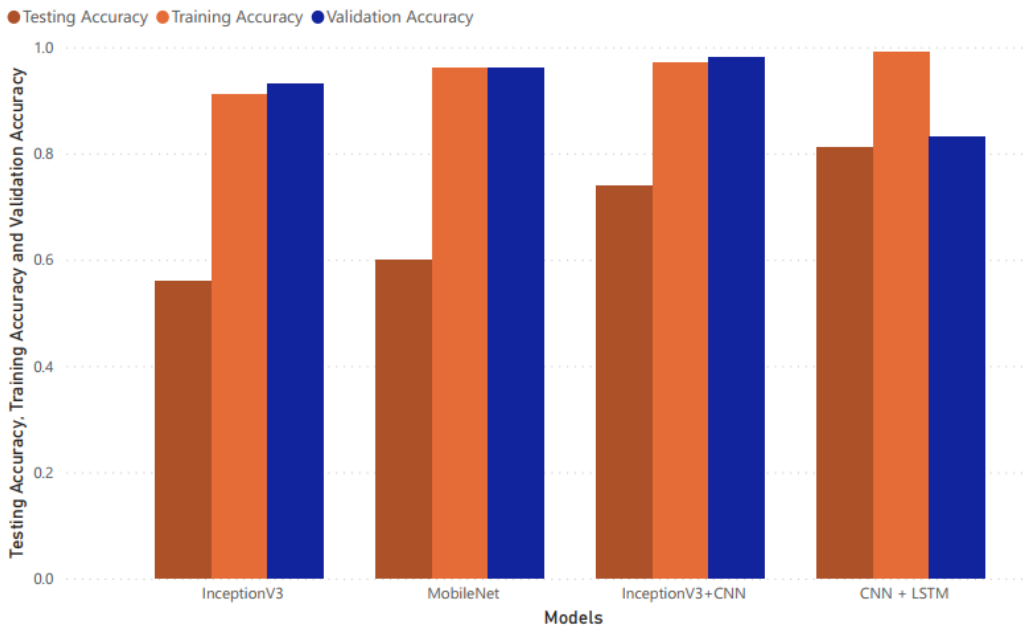

Figure 6: Comparison of Models

CONCLUSION

Computer vision and deep learning technology enable extensive research opportunities in the realm of solving Human Activity Recognition. After successfully implementing different feature extraction algorithms on UCF101, including CNN+LSTM, Inception-V3, Inception-V3+CNN, and MobileNet, we discovered that CNN+LSTM produced better results. With our research, we have also demonstrated the comparative analysis of these models when used individually. Furthermore, we have shown that a high-performing classifier can be trained using a purely synthetic dataset like UCF101. These findings (presented in Table 5) serve as proof of concept, implying that synthetic data can be a useful tool in recognising human activities.

However, HAR is a computationally intensive topic that necessitates the development and testing of more efficient models using new fine-grained and large-scale labelled datasets. Our contribution is significant because it provides a novel perspective on training and assessing various models on the labelled HAR dataset, and it is undeniably a step toward a highly accurate and optimised model in the near future.

The following are some possible future work directions:

- More good results could be achieved in future, by updating implemented models or using new hybrid models.
- This study has applications in a variety of fields, including tracking several people's behaviours and activating notifications if necessary, assisting visually impaired persons in identifying various activities, receiving early notification during emergency circumstances regardless of human presence, and so on.
- Success of good results in this domain will lead us to cover more kinds of different activities and will also open the new doors for research in Video Capturing.

REFERENCES

[1] Sina Mokhtarzadeh Azar, Mina Ghadimi Atigh, Ahmad Nickabadi, and Alexandre Alahi. 2019. Convolutional relational machine for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7892–7901.

[2] Akilan T, Wu QJ, Safaei A, Jiang W 2017. A late fusion approach for harnessing multi-CNN model high level features. In: 2017 IEEE international conference on systems, man, and cybernetics (SMC).
IEEE, pp 566–571.

[3] Tâm Huynh and Bernt Schiele. 2005. Analysing features for activity recognition. In *Proceedings of the Joint Conference on Smart Objects and Ambient Intelligence: Innovative Context-aware Services: Usages and Technologies*. ACM, 159–163.

[4] Eoin Brophy, José Juan Dominguez Veiga, Zhengwei Wang, Alan F. Smeaton, and Tomas E. Ward. 2018. An interpretable machine vision approach to human activity recognition using photoplethysmography sensor data. *arXiv preprint arXiv:1812.00668*.

[5] Jessica Lin, Eamonn Keogh, Stefano Lonardi, and Bill Chiu. 2003. A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*. ACM, 2–11.

[6] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. 2018. A survey on deep learning: Algorithms, techniques, and applications. *Comput. Surv.*51, 5 (2018), 92.

[7] Sojeong Ha, Jeong-Min Yun, and Seungjin Choi. 2015. Multi-modal convolutional neural networks for activity recognition. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 3017–3022.

[8] Nicholas D. Lane and Petko Georgiev. 2015. Can deep learning revolutionise mobile sensing? In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications*. ACM, 117–122.

[9] Ianbo Yang, Minh Nhut Nguyen, Phyo Phyo San, Xiao Li Li, and Shonali Krishnaswamy. 2015. Deep convolutional neural networks on multichannel time series for human activity recognition. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence.*

[10] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. 2018. Non-local neural networks. *In Proceedings of the IEEE conference on computer vision and pattern recognition, pages* 7794–7803.

[11] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. *In Proceedings of the IEEE international conference on computer vision, pages* 6202–6211.

[12] Nils Y. Hammerla, Shane Halloran, and Thomas Plötz. 2016. Deep, convolutional, and recurrent models for human activity recognition using wearables. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence*. 1533–1540.

[13] Moya Rueda F, Grzeszick R, Fink G, Feldhorst S, ten Hompel M 2018. Convolutional neural networks for human activity recognition using body-worn sensors. *In: Informatics, vol 5. Multidisciplinary Digital Publishing Institute, p* 26

[14] Razzak MI, Naz S, Zaib A. 2018. Deep learning for medical image processing: overview, challenges and the future. *In: Classification in BioApps. Springer, pp* 323–350

[15] Li C, Wang P, Wang S, Hou Y, Li W. 2017b. Skeleton-based action recognition using LSTM and CNN. *In: 2017 IEEE international conference on multimedia and expo workshops (ICMEW). IEEE, pp* 585–590

[16] Ji Lin, Chuang Gan, and Song Han. 2019. TSM: Temporal Shift Module for Efficient Video Understanding. In *The IEEE International Conference on Computer Vision (ICCV)*.

[17] Christoph Feichtenhofer. 2020. X3D: Expanding Architectures for Efficient Video Recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[18] Michael S. Ryoo, AJ Piergiovanni, Mingxing Tan, and Anelia Angelova. 2020. AssembleNet: Searching for MultiStream Neural Connectivity in Video Architectures. *In The International Conference on Learning Representations (ICLR).*

[19] Joao Carreira and Andrew Zisserman. Quo Vadis. 2017. Action Recognition? A New Model and the Kinetics Dataset. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] A. Ahsan U, Sun C, Essa I. 2018. Discrimnet: Semi-supervised action recognition from videos using generative adversarial networks. *ArXiv preprint arXiv:*1801.07230

[21] Hildegard Kuehne, Hueihan Jhuang, Est´ıbaliz Garrote, Tomaso Poggio, and Thomas Serre. 2011. HMDB: A Large Video Database for Human Motion Recognition. In *The IEEE International Conference on Computer Vision (ICCV)*.

[22] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *arXiv preprint arXiv:1212.0402*.

[23] Gunnar A. Sigurdsson, Gul Varol, Xiaolong Wang, Ali ¨Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding. In *The European Conference on Computer Vision (ECCV)*.

[24] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jurgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large Scale Holistic Video Understanding. *European Conference on Computer Vision*, pages 593–610. Springer, 2020

[25] Oscar D. Lara and Miguel A. Labrador. 2013. A survey on human activity recognition using wearable sensors. *IEEE Commun. Surv. Tutor.* 1192–1209.

[26] Preksha Pareek, Ankit Thakkar. 2021. A survey on video-based Human Action Recognition: recent updates, datasets, challenges, and applications, *Artificial Intelligence Review* 54:2259–2322

[27] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2021. Deep Learning for Sensor Based Human Activity Recognition: Overview, Challenges, and Opportunities. *ACM Comput. Surv.* 54, 4, Article 77 (May 2021), 40 pages

[28] Kaixuan Chen, Dalin Zhang, Lina Yao, Bin Guo, Zhiwen Yu, and Yunhao Liu. 2021. Deep Learning for Sensor Based Human Activity Recognition: Overview, Challenges, and Opportunities. *ACM Comput. Surv.* 54, 4, Article 77 (May 2021), 40 pages.

[29] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. 2016. Rethinking the inception architecture for computer vision, *in Proc. CVPR, pp.* 2818 – 2826.

[30] Y. Zahid, M. A. Tahir and M. N. Durrani. 2020. Ensemble Learning Using Bagging And Inception-V3 For Anomaly Detection In Surveillance Videos, *in Proc. ICIP, pp.* 588 - 592.