Unraveling Human Actions: From Video to Understanding

Mohit Gupta
Dept. of CSE
IIIT, Delhi
mohit22112@iiitd.ac.in

Nidhi Verma
Dept. of CSE
IIIT, Delhi
nidhi22044@iiitd.ac.in

Pratik Chauhan
Dept. of CSE
IIIT, Delhi
pratik221128@iiitd.ac.in

Abstract

Human action recognition (HAR) is a subfield of computer vision that focuses on the analysis and categorization of human activities from visual data. This field is of significant importance because it has numerous practical applications, such as healthcare monitoring, sports analytics, security surveillance, and entertainment. In this project, we aim to build a robust HAR classifier that can accurately identify and categorize various human actions from video frames. To achieve this, we have experimented with several popular models such as MobileNet, MobileNetV2, Xception, BiLSTM, etc., which are known for their excellent performance in image classification tasks. We have utilized the UCF-50 dataset, which is a standard benchmark dataset for HAR, to train and evaluate our models. The key objective of our project is to develop a comprehensive and accurate classifier that can accurately distinguish between different types of human actions with a high degree of accuracy. To achieve this, we have implemented various techniques such as feature extraction, data augmentation, and hyperparameter tuning. The final outcome of this project is a robust and accurate classifier that can be trained on any dataset of human actions and applied to real-world scenarios.

Keywords: *Image Classification, Deep Learning, Recurrent Networks, Sequence Learning, Action Recognition.*

1 Introduction

HUMAN Action Recognition and human behavior understanding are becoming increasingly important and have attracted many research efforts over the past two decades Shany et al. (2012), Cheng et al. (2013). Moreover, human activity analysis is a key enabler in the development of many applications, such as smart rooms, interactive virtual reality systems, people monitoring, and environment modeling Baek and Yun (2010), San-Segundo et al. (2016), Yu et al. (2016). Two essential categories of approaches to human action recognition can be distinguished: wearable sensors-based and vision-based methods Zerrouki et al. (2018). Nowadays, small sensors are available and embedded in many daily devices such as smartphones and smartwatches. On the other hand, vision-based approaches focus on using information extracted from video sequences to recognize human actions Zerrouki et al. (2016). Approaches in this category have become much more important than before due to the advancements in computer vision, pattern recognition, and image processing fields Yu et al. (2012).

In this project, we are using the vision-based approach for classifying human action. There are numerous state-of-the-art models available for HAR classification problems like ResNet50, AlexNet, Inception, etc. In this project we are using 4 models in total, two models named MobileNetV2 and Xception on UCF50 image dataset and rest two are fusion based models named MobileNetV2 + BiLSTM and Xception + BiLSTM on UCF50 video dataset.

2 Literature Review

Wu et al. (2014) proposes a method that uses features extracted from sensor data and uses SVM for classification. The output of SVM is then fed into HMM to learn temporal dependencies between different activity classes. The method achieves high accuracy rates of 98.22% on their own dataset of 60 classes, surpassing other state-of-the-art methods. Guo and Lai (2014) reviews techniques for recognizing human actions in still images. Emphasizes the importance of benchmark datasets like Weizmann, KTH, and HMDB for evaluating these methods. Jain and Kanhangad (2017) presents a method for classifying human activities using data collected from smartphone sensors. The method involves classification using three machine learning algorithms like k-nearest neighbors, support vector machines, and decision trees to classify six activities. The proposed method achieves an average accuracy of over 95.0%. Zerrouki et al. (2018) proposes a vision-based method for classifying human actions in video using the adaptive boosting algorithm. The method achieves high accuracy on the KTH and Weizmann datasets, outperforming other state-of-the-art methods. Features are extracted from video frames and used as input to the adaptive boosting algorithm, which learns weak classifiers that are combined into a strong classifier. Jegham et al. (2020) provides an overview of vision-based human action recognition, discussing challenges such as robust feature extraction, inter-class similarities and intra-class variations, and the need for real-time processing and scalability in practical applications.

3 Proposed Methodology

3.1 Dataset[1]

For this problem, we have researched many datasets related to this problem, the details are shown in Table 1. In our baseline, we have experimented on UCF 50[2] images only dataset. Now we are working on videos as well. As described earlier the UCF50 dataset contains 50 classes, each class videos are grouped into 25 groups, where each group consists of more than 4 action clips. Due to lack of compute resources, we have reduced our training dataset to 25 video clips per class. We have removed extra redundant videos of each group. We have splitted our whole dataset into train, test, and validation sets with the split ratio of 80:10:10 and stratified with class label. In our baseline dataset, there is no association between images as when we split the dataset, the video frames were randomly distributed among different sets. But in our final approach, we are splitting the dataset based on the videos, then dividing the images into frames.

3.2 Model Architecture

In our baseline, we have performed experiments on the images only dataset of UCF50 with MobileNet, and MobileNetV2 dataset. Having the very large dataset, these are light-weight high accuracy deep learning models for image classification which gives the accuracies given in table 2. We have done our baseline experiments on two different setups of same dataset. First experiment is on original dataset, and second is on augmented dataset.

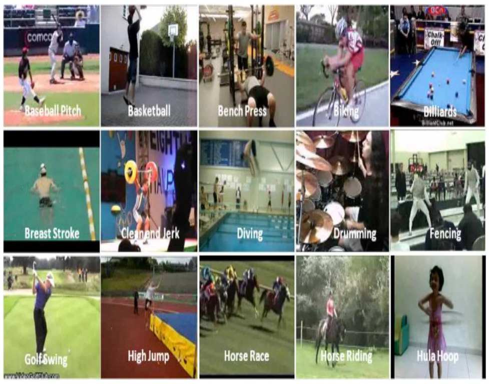


Figure 1: Sample Video Frames

Now comes the final proposal, as our title suggests, our aim is to recognize human activity from **VIDEOS**. Videos are a set of frames, and the current frame of a video has some association with previous, and future frames. So to recognize actions from the videos, we need some kind of model which will learn the sequential dependencies. We successfully classified images in baseline, now with same models we are fusing BiLSTM to get the dependencies between video frames, and classify actions from them. The detailed architecture visualization can be seen from the fig. 2. We have also added one more Deep Architecture to improve the accuracy of the model which is **Xception**. We have used pre-trained **Xception** with input shape (224*224*3) for extracting features from the video frames, then these extracted features will be passed to multi-layered **Bi-LSTM** with first GRU layer having 64 neurons, and then the second GRU layer having 32 neurons to learn the sequential dependencies, with dropout of 0.4. then this vector sequence is passed to a dense layer of 8 neurons with reLU activation, and the final output dense layer of 50 neurons having softmax activation function. is passed to **softmax** layer which will return the probability of belonging of each class.

4 Experimental Setup & Results

In our baselines, we had worked on UCF50 images only, and trained two models MobileNet, and MobileNetV2 and among those two, MobileNet outperforms with 96.0% accuracy on test set. Now we are directly working on videos, in which we first extracts the frames from the video, then extract their features, and after that using BiLSTM, with a classification head, we are classifying the

[1] https://tinyurl.com/3eyvjj9v

[2] https://www.crcv.ucf.edu/data/UCF50.php

Feature	UCF11	UCF50	UCF101	HAR
Number of action classes	11	50	101	15
Number of Video Clips	1600	6612	13320	18000
Video Duration (seconds)	10	10	10	images
Resolution	320*240	320*240	320*240	320*240

Table 1: Feature Details of each Dataset

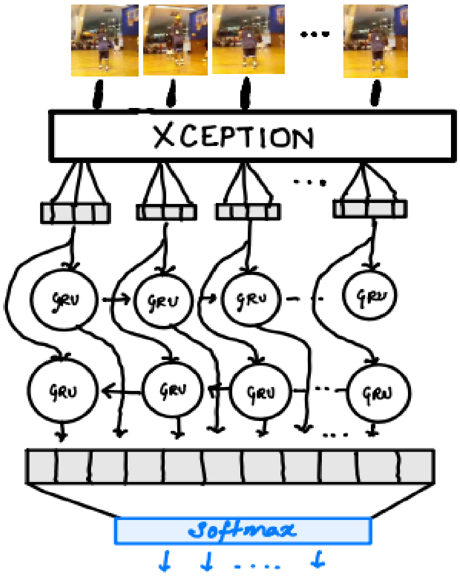
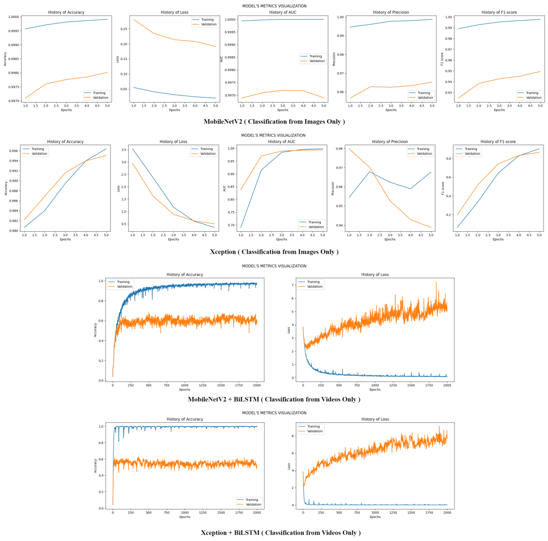


Figure 2: Xception + BiLSTM Architecture



MobileNetV2 (Classification from Images Only)

Xception (Classification from Images Only)

MobileNetV2 + BiLSTM (Classification from Videos Only)

Xception + BiLSTM (Classification from Videos Only)

Figure 3: Different Visualizations for Model Training

different classes.

We have trained two fusion based architectures to measure the performance of our classification model. In our first fusion setup, we first used the pre-trained MobileNetV2[3] model, which is trained on Imagenet[4] dataset with BiLSTM on our UCF50 videos dataset with input image shape of (224, 224, 3) as all the images are in RGB format. We have added some extra layers to the architecture, such as a Dense layer with Relu activation, dropout of 0.5, and softmax activation at the output layer. We have used Adam as the optimizer function and sparse categorical cross entropy as the loss function. The configuration used for this model is epochs 2000, batch_size 32, train-validation split 80-20. The dimension of extracted feature vector is 1*2048

Then on the same splits, we have trained Xception[5] with Bi-LSTM, and here the LSTM unit is GRU with the same configuration. The results obtained are very good with this combination as compared to the previous combination results. The reason may be that the GRU is good at handling long sequences, and somehow tackling vanishing gradient problem. That is why, Xception + Bi-LSTM (GRU) gives the best results for our project. The metric visualizations can be seen from the fig. 3.

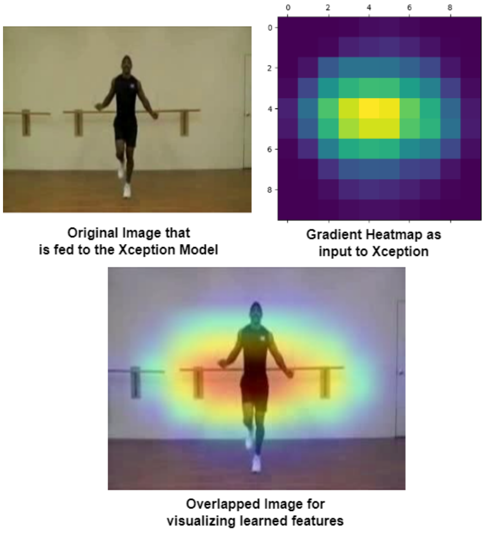
4.1 Explainability

Explainability refers to the ability to understand and interpret the decision-making processes of an algorithm or model. GRAD-CAM as a tool for explainability - GRAD-CAM (Gradient-weighted Class Activation Mapping) is a technique used for visualizing and interpreting the decisions made by deep neural networks in image classification tasks. We have also used gradcam to visualize the part of video frame which is used to classify the image. The gradcam results can be seen from the fig. 4

4.2 Results

To concluding the results and observation of our whole project. We have experimented on two broad setups to explore the vision domain. First we worked on image classification only with MobileNet, and MobileNetV2 in which MobileNet outperforms with the test accuracy of 96.0%. Then

[3] https://arxiv.org/abs/1704.04861
[4] https://www.image-net.org/
[5] https://arxiv.org/abs/1610.02357



Original Image that is fed to the Xception Model

Gradient Heatmap as input to Xception

Overlapped Image for visualizing learned features

Figure 4: Gradient Heat-Map for Original Image as Input to Xception

after, to classify the action from the videos, we need to learn the sequential dependencies between different video frames. For this we have used two combinations.

1. **MobileNetV2 + BiLSTM**
2. **Xception + BiLSTM (GRU)**

Out of these two combinations, the Xception + BiLSTM combination works really good, and gives the test accuracy of 54.16 %. The detailed metric results can be seen from table 2.

NOTE: * meaning the baseline model, and ** meaning the fine-tuned model [2]

4.3 Hardware and Software Specifications

The project is designed and developed using Python and its libraries as Numpy, Pandas, Sci-kit, matplotlib, and DL libraries. Google Colaboratory is used to develop and run the code on the Web Browser itself. The specification for the platform are as for GPU is 1xTesla K80, having 2496 CUDA cores with 12GB GDDR5 VRAM. The CPU specifications are as 1xsingle Core Hyper-Threaded Xeon Processors @2.3Ghz, i.e. (1 core, 2 threads), and the available RAM is approximately 12.6 GB with a disk space of almost 33 GB.

5 Error Analysis & Observations

The dataset we are using is balanced, so we don't have a problem with the unbalanced dataset, which is a major issue in classification problems. After using mobileNetV2, and Xception on the UCF50 dataset, we have observed that, as the dataset is very large and clean, both the models are giving very good training and validation accuracy for a very less number of epochs.

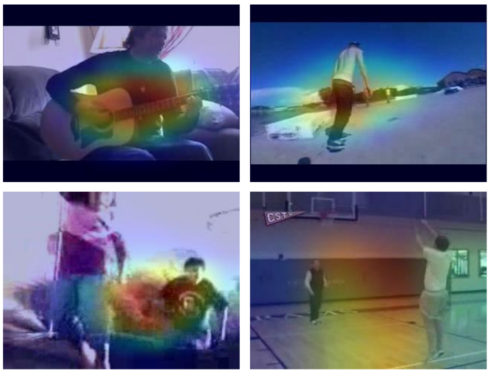


Figure 5: Error Analysis using GradCam

Now If we see the fig. 5, the first row shows the gradcam images for still frames in which the object does not moving much from its position, but in the second row, the main object is moving or not still.

In first case, the model clearly able to focus on the object, which we can see from the gradcam heatmap. Our model is working good for still video frames, but on the other hand, if we see the second case, the objects are not in center or moving, the model is not able to catch the object properly as gradcam shows.

From this we can conclude that, our model works good or correctly able to classify actions in which most of the portion of the video is still and only some part is moving like **PlayingGuitar**, or **Skiingg**, and on the other hand, for second case it is not completely fails, but also not correctly able to classify actions in which majority of the section is moving in the video like **BasketBall**, or **TrampolineJumping**.

6 Future Work

We have already outlined the future scope of our project in the mid-semester report, and we are pleased to report that our initial results have been quite promising. However, we have identified another area for improvement in our work that we believe has the potential to significantly enhance the accuracy of our fusion-based model. Specifically, our model incorporates both Deep CNN and BiLSTM architecture to handle long sequential dependencies. While this approach has yielded positive results so far, we believe that adding an attention mechanism to the BiLSTM architecture could

Dataset	Model	Type of Input	Precision	Recall	Accuracy
UCF - 50	MobileNet*	Images	**0.96**	**0.96**	**0.9613**
	MobileNetV2*	Images	0.95	0.94	0.9467
UCF - 50	MobileNetV2**	Images	**0.97**	**0.97**	0.97
	Xception	Images	0.95	0.86	**0.99**
UCF - 50	MobileNetV2 + BiLSTM	Video Frames	–	–	0.45
	Xception + BiLSTM (GRU)	Video Frames	–	–	**0.54**

Table 2: Result Evaluations

further enhance our model's performance.

By introducing an attention mechanism, we can enable the BiLSTM to learn dependencies in a more targeted and efficient manner, thereby boosting the overall accuracy and precision of the model. We are confident that this approach will prove to be a valuable addition to our project and look forward to exploring it further. Multimodal data fusion techniques are being explored to address the challenge of zero-shot learning, where the system needs to recognize actions it hasn't seen before. To improve real-time processing of recognition data, 5G and edge computing are being leveraged. Additionally, integrating with augmented reality and virtual reality opens up new use cases, expanding the potential applications of this technology.

7 Contributions

Although all the team members contributed equally to this project. But if we broadly define what each of our team member has done.

- MobileNetV2 - on image dataset - Nidhi Verma

- Xception - on image dataset - Pratik Chauhan

- MobileNEtV2 + BiLSTM - on video dataset - Mohit Gupta

- Xception + BiLSTM - on video dataset - All members

References

Jonghun Baek and Byoung-Ju Yun. 2010. Posture monitoring system for context awareness in mobile computing. *IEEE Transactions on Instrumentation and Measurement*, 59(6):1589–1599.

Jingyuan Cheng, Oliver Amft, Gernot Bahle, and Paul Lukowicz. 2013. Designing sensitive wearable capacitive sensors for activity recognition. *IEEE Sensors Journal*, 13(10):3935–3947.

Guodong Guo and Alice Lai. 2014. A survey on still image based human action recognition. *Pattern Recognition*, 47(10):3343–3361.

Ankita Jain and Vivek Kanhangad. 2017. Human activity classification in smartphones using accelerometer and gyroscope sensors. *IEEE Sensors Journal*, PP:1–1.

Imen Jegham, Anouar Ben Khalifa, Ihsen Alouani, and Mohamed Ali Mahjoub. 2020. Vision-based human action recognition: An overview and real world challenges. *Forensic Science International: Digital Investigation*, 32:200901.

Rubén San-Segundo, Julián David Echeverry-Correa, Christian Salamea, and José Manuel Pardo. 2016. Human activity monitoring based on hidden markov models using a smartphone. *IEEE Instrumentation Measurement Magazine*, 19(6):27–31.

Tal Shany, Stephen J. Redmond, Michael R. Narayanan, and Nigel H. Lovell. 2012. Sensors-based wearable systems for monitoring of human movement and falls. *IEEE Sensors Journal*, 12(3):658–670.

Haitao Wu, Wei Pan, Xingyu Xiong, and Suxia Xu. 2014. Human activity recognition based on the combined svmhmm. In *2014 IEEE International Conference on Information and Automation (ICIA)*, pages 219–224.

Liu Yu, Haibin Li, Xiaowei Feng, and Jizhong Duan. 2016. Nonintrusive appliance load monitoring for smart homes: recent advances and future issues. *IEEE Instrumentation Measurement Magazine*, 19(3):56–62.

Miao Yu, Adel Rhuma, Syed Mohsen Naqvi, Liang Wang, and Jonathon Chambers. 2012. A posture recognition-based fall detection system for monitoring an elderly person in a smart home environment. *IEEE Transactions on Information Technology in Biomedicine*, 16(6):1274–1286.

Nabil Zerrouki, Fouzi Harrou, Ying Sun, and Amrane Houacine. 2016. Accelerometer and camera-based strategy for improved human fall detection. *Journal of Medical Systems*, 40.

Nabil Zerrouki, Fouzi Harrou, Ying Sun, and Amrane Houacine. 2018. Vision-based human action classification using adaptive boosting algorithm. *IEEE Sensors Journal*, 18(12):5115–5121.