Human Action Recognition using Deep Learning

CS4099 Project Final Report

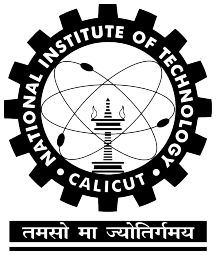
Submitted by

Lakavath Akshay Kumar **Reg No: B210531CS**

Shreya **Reg No: B211267CS**

Under the Guidance of

Dr. Pranesh Das



Department of Computer Science and Engineering

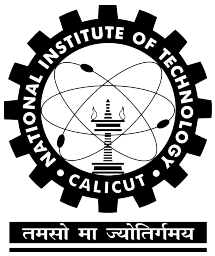
National Institute of Technology Calicut

Calicut, Kerala, India - 673 601

April 15, 2025

NATIONAL INSTITUTE OF TECHNOLOGY CALICUT, KERALA, INDIA - 673 601

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING



2025

CERTIFICATE

Certified that this is a bonafide record of the project work titled

Human Action Recognition using Deep Learning

done by

Lakavath Akshay Kumar

Shreya

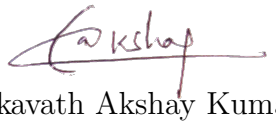
of eighth semester B. Tech in partial fulfillment of the requirements for the award of the degree of Bachelor of Technology in Computer Science and Engineering of the National Institute of Technology Calicut

Project Guide	**Head of Department**
Dr. Pranesh Das	Dr. Subashini R
Assistant Professor	Associate Professor

DECLARATION

We hereby declare that the project titled, **Human Action Recognition using Deep Learning)**, is our own work and that, to the best of our knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of the university or any other institute of higher learning, except where due acknowledgement and reference has been made in the text.

Place : NIT Calicut
Date : 10-04-2025

Signature : 
Name : Lakavath Akshay Kumar
Reg. No. : B210531CS

Signature: 
Name: Shreya
Reg. No. : B211267CS

Abstract

Traditional human action recognition (HAR) models typically perform well in normal lighting conditions but exhibit significant limitations in low-light or no-light environments. A major problem in this area is the availability of limited datasets that represent such dark conditions. In this work, we address this challenge by creating a new dataset with the help of an existing low-light dataset. We have also enhanced it to reflect degraded conditions by applying transformations like increased brightness and darkness, rotation, shifting, gamma correction, and noise. In addition to that, to check the usefulness of the dataset, a basic model is implemented. Furthermore, a comparative study is conducted using the same model between **NoctAct-HAR(Nocturnal Action Dataset for Human Action Recognition)** that is our dataset and the widely used dataset **ARID** for consistency.

ACKNOWLEDGEMENT

We wish to extend our heartfelt appreciation to Dr. Subashini R, the Head of the Computer Science Engineering Department at NIT Calicut, for granting us the opportunity to undertake this project. Our gratitude also extends to Dr. S Sheerazuddin and Dr. Anil Pinapati, our project coordinators, whose meticulous organization of project milestones was instrumental in our success. We are deeply grateful to Dr. Pranesh Das, our invaluable guide and mentor, whose unwavering support, guidance, and encouragement propelled us throughout the lifecycle of the project. Our appreciation also goes to our parents and the esteemed faculty members for their constant motivation and support. We also acknowledge the invaluable assistance provided by the dedicated staff of the CSE Department at NIT Calicut. Lastly, we extend our thanks to our friends whose cooperation and assistance were indispensable throughout the course of the project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Human Action Recognition (HAR) has become a vital component of real-world applications. It plays an important role in areas such as healthcare monitoring, smart homes, surveillance, human interaction with computer and other systems. With advancements in deep learning, HAR models have significantly evolved and shown remarkable results in recognizing actions in normal or rather well-lit environments. However, it remains a challenge in a poorly lit environment, which is required for nighttime surveillance or low-light indoor areas. Although researchers are trying to bridge this gap, a major problem is lack of proper datasets that reflect the same condition.

The Action Recognition in the Dark (ARID) dataset [2] was introduced to address this issue in 2020, it still has been the only proper dataset available for this work. Until Extremely Low-Light Action Recognition (ELLAR) [3] was introduced in 2024 for the same purpose. This problem of data availability has restricted the development of HAR in dark.

To face this challenge, our paper focuses on creating a new, and enhanced dataset. The dataset is a mix of videos which are collected by us and ARID dataset. We have also applied a range of transformations to simulate extreme conditions. These include adjustments in brightness and darkness levels, rotation and shifting, gamma correction and synthetic noise. The result is a

diverse dataset which contains over **6613 videos** and **12 classes**.

To evaluate the performance of NoctAct-HAR, we implemented a basic model and compared its performance across NoctAct-HAR and widely used ARID dataset. By using the same model for evaluation, we have ensured a consistent comparison. We have used two lightweight yet effective image enhancement techniques. We also experimented using different number of classes.

Our results highlight that the enhanced dataset introduces a higher degree of complexity and provides a more rigorous benchmark for HAR systems. In particular we observed the improvement in accuracy. NoctAct-HAR achieved a 4% higher accuracy than ARID dataset.

This work contributes toward building more robust and challenging dataset that can help the research community design systems capable of functioning reliably in adverse lighting conditions.

Chapter 2

Literature Survey

2.1 Human Action Recognition

Human Action Recognition (HAR) is an important area in computer vision. It focuses on identification and classification of human actions from videos and images. Traditional methods were very rigid as they relied on handcrafted features like Space-Time Interest Points (STIP), Histograms of Oriented Gradients (HOG), and Improved Dense Trajectories (iDT). It was combined with other machine learning classifiers. However they struggled with challenges like occlusion and lighting conditions.

With the rapid growth of deep learning, HAR has experienced shift. Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) have enabled in learning spatio-temporal features. These models effectively capture both spatial and temporal feature which is important for correct understanding. Although these advancements are significant but there are still limitations, such as lack of large datasets.

These development and challenges are thoroughly discussed in the survey by Herath et al. (2017) [4] which can be considered as comprehensive guide for future development in HAR.

2.2 Human Action Recognition in Dark

Human Action Recognition (HAR) in Dark remains one of the greatest challenge in this field. Limited visibility hampers the detection of motion, scenes and objects. Models trained on well-lit datasets perform poorly under such conditions. Low light leads to the loss of both spatial and temporal features, resulting in inaccurate interpretations of human actions. To address this, researchers have started exploring pre-proceesing methods such as image enhancement and contrast enhancement to make features visible properly.

Singh et al. (2022) [5] proposed a method which uses Bidirectional Encoder Representations from Transformers (BERT) for HAR in dark. Their approach leverages transformer's ability to capture spatial features in low-light frames and model long-range temporal dependencies. Their model demonstrated superior performance over CNN and RNN-based methods under dark conditions.

2.3 Deep Learning Techniques

2.3.1 3D CNN

3D CNNs are a powerful approach for Human Action Recognition (HAR). While 2D CNNs process only spatial dimensions whereas 3D CNNs incorporate the time dimension, enabling the model to capture motion dynamics.

The comprehensive review done about video-based HAR methods by Pham et al. (2022) [6] highlighted the effectiveness of 3D CNNs in learning temporal-patterns from raw videos.

Experimental comparison done between CNN+LSTM architectures and 3D CNNs with Two-Stream CNNs by Yu and Yan [7] reveals that 3D CNNs performed well by learning spatial and temporal features together. However, the absence of large-scale dataset and need for significant computational resources increase the model's complexity.

2.3.2 Two-Stream CNN

Because of their ability to handle spatial and temporal features effectively Two-Stream CNNs have become one of the popular choices in HAR. Yu and Yan (2020) [7] discussed in their paper that the architecture of Two-Stream CNN consists of two separate streams. One processes optical flow to get temporal features (motion-related) and other processes RGB frames to capture spatial features. This helps the model to understand how scene changes over time.

A discussed in the paper Two-Stream CNNs achieve promising results on HAR datasets but it also pointed that producing optical flow is computationally expensive, making it difficult for real-world utilization. Despite this the architecture remains strong baseline for HAR.

2.3.3 Convolutional Neural Network + Long Short-Term Memory

Convolutional Neural Network + Long Short-Term Memory (CNN+LSTM) is a hybrid model which has become an effective approach for HAR. In this model, CNNs first extract spatial features from individual frames and captures important patterns along with reducing dimension using pooling operations. These are then passed into LSTM networks, which model temporal dependencies by using gating mechanisms and memory cells.

Yu and Yan (2020) [7] discussed how this approach improved the performance by learning not only appearance in single frame but also the temporal flow of actions across various frames. Similarly, Pham et al. (2022) [6] also highlighted the strength of CNN+LSTM in capturing motion patterns across time. However, both papers also noted that this model can be computationally intensive, especially for longer video sequences and require careful tuning for good performance.Despite these problems it remains widely used model in HAR tasks.

2.3.4 Long-term Recurrent Convolutional Network

Long-term Recurrent Convolutional Network (LRCNs) are a hybrid deep learning model which was introduced by Donahue et al. [8] in 2015. This model uses CNNs for spatial and LSTM networks for temporal feature extraction. In this approach, rich spatial features are extracted first using CNNs and then passed to LSTMs, which capture temporal features overs time, allowing the network to learn both motion patterns and sequential dependencies across frames.

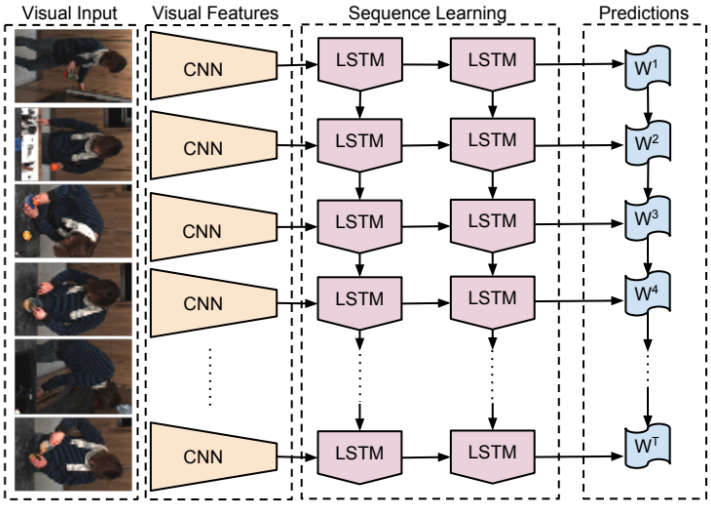


Figure 2.1: LRCN model

The framework was evaluated across multiple vision tasks such as image captioning, activity recognition, and video description. It demonstrated improved performance for action recognition using both context provided by frame sequences and content of individual frames. The Authors also emphasized the importance of motion information and end-to-end learning in combining appearances. Even after high computationality it still remains strong base for sequence modeling and inspires other architectures in the HAR domain.

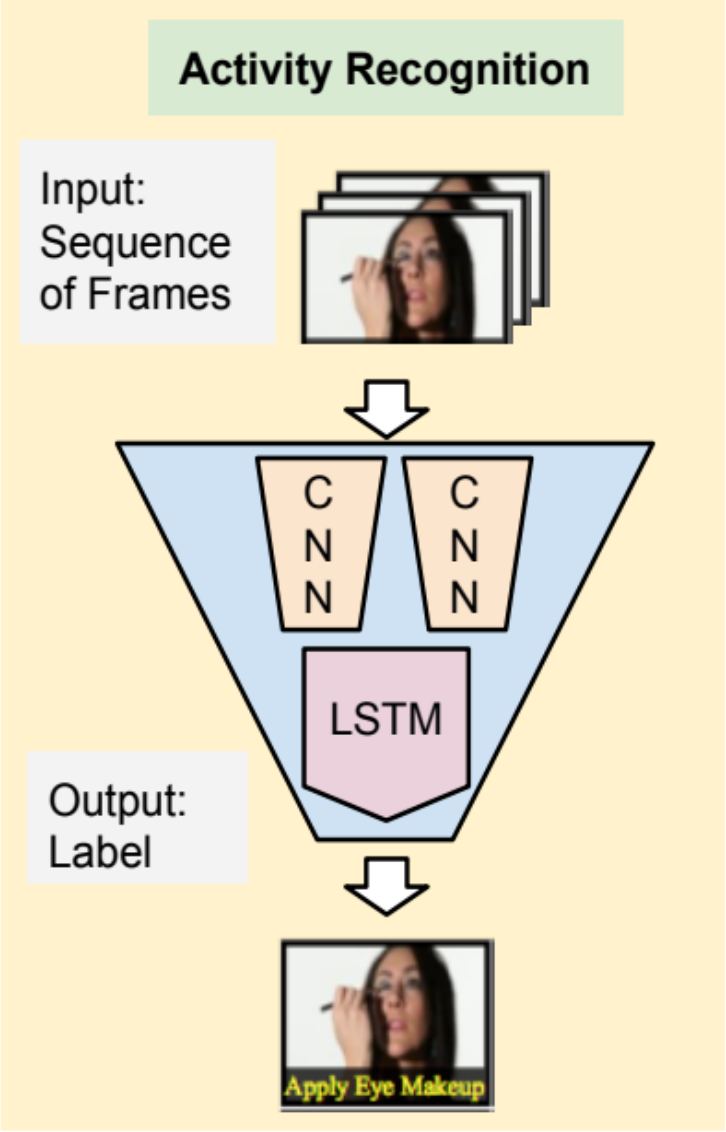


Figure 2.2: Activity Recognition using LRCN.

2.3.5 Self-attention LRCN

Although LRCN is a well known model, it has a limitation: the uniform treatment of all frames which reduces effectiveness of the model in complex settings. To overcome this, Athilakshmi et al. [9] proposed an enhanced model in 2024 which uses self-attention mechanism to focus on informative frames.

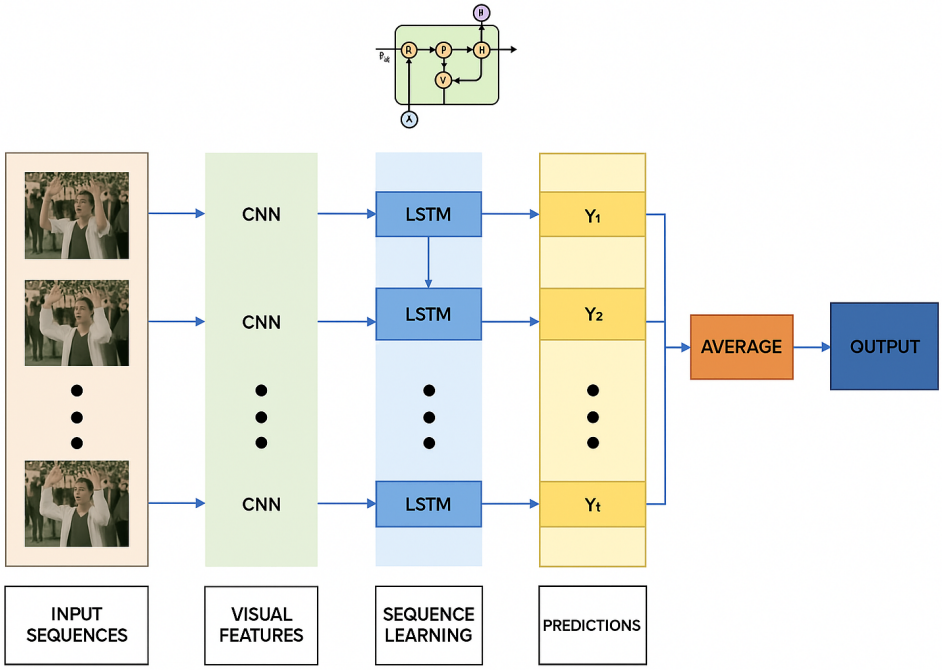


Figure 2.3: Self-attention LRCN model

This model improved the interpretability and performance of HAR. It allowed the model to assign higher weights to frames which are more relevant to action performed which is very useful in surveillance, where cluttered scenes, occlusions or repeated frames hinder performance. The results highlighted that the proposed model outperformed standard LRCNs model on benchmark datasets, which made ut one of the promising architecture for real-world surveillance systems.

2.3.6 Gated Recurrent Units

Gated Recurrent Units(GRU) are a lightweight yet effective alternative to LSTMs. GRUs have a simpler architecture by combining forget and input gate into one update gate, which reduces the number of parameters speeding the training.

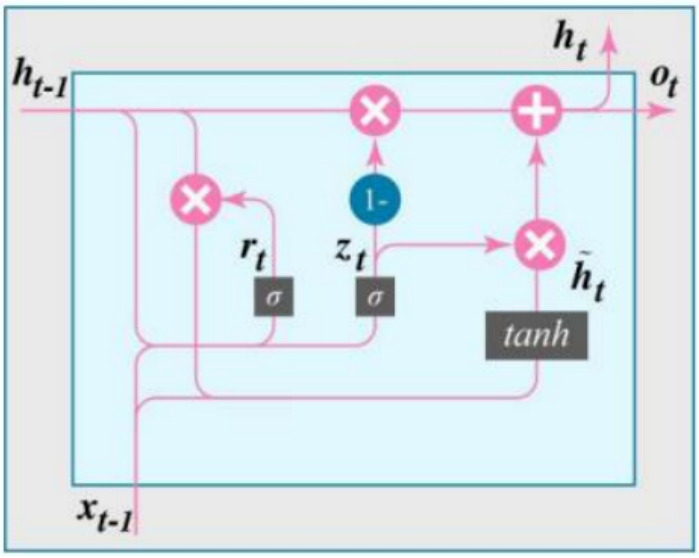


Figure 2.4: Structure of GRU unit [1]

A comparative study was conducted by Muhamad and Mohammed [10] in 2022, where they proposed enhanced GRU and LSTM models for HAR. The improved GRU model consisted of two layers with modified gating mechanisms and 128 units each, allowing it to better adapt motion sequences. It was evaluated on UCF-101 dataset and achieved accuracy of 92.9%. The model outperformed many traditional models.

This showed that GRUs offer a strong balance between computational efficiency and accuracy, making it well-suited for real-time HAR applications.

2.4 Image Enhancement Techniques

2.4.1 CLAHE

Quality of frames has great impact on HAR and object detection models, however in low-light environment it deteriorates leading to a negative impact on these systems. Zhikang Yuan et al. [11] proposed an image enhancement technique to address this problem called CLAHE. To enhance the brightness of low-light images, CLAHE was applied to improve visibility.

In this technique contrast is adjusted adaptively along with preserving chrominance features to maintain natural color tones. According to the results, this method improved object detection in low-light environment. It is not only valuable for preprocessing step but also for industrial use.

2.4.2 Multi Scale Retinex

The contrast and visibility of video frames are degraded in low-light environments. Liu et al. (2016) [12] proposed an image enhancement technique based on MSR algorithm. This technique is inspired by Retinex theory, which mimics human visual perception by decomposing a frame into reflectance components and illumination.

In this approach visibility is enhanced by adjusting illumination along with amplifying reflectance. Thereby improving contrast and brightness without introducing noise. This method enhanced frame quality and improved the motion in videos and clarity of objects captured under low-light conditions.

2.5 Available Dataset

2.5.1 ARID

HAR in dark environment poses a significant challenge. Most of the datasets are for well-lit environment which fails for low-light scenarios. To address this Yuecong Xu et al.[2] introduced a new dataset specifically designed for dark conditions: ARID (Action Recognition In Dark).

The dataset initially had 3,780 video clips spanning 11 classes, all taken in poor lighting conditions. In its updated version ARID v1.5, dataset expanded to have 5,572 video clips having same classes.

The distribution of clips of ARID dataset is shown in fig. 2.5:

2.5.2 ELLAR

Introduction of ARID dataset addressed the problems to some extent, however having one dedicated dataset was still not enough. Ha et al. introduced the ELLAR (Extremely Low-Light condition Action Recognition) dataset [3].

ELLAR consists of over 12,000 real-world videos, captured in dark environment not synthetically generated. The authors also proposed a technique called DGAM to enhance video frames by adaptively adjusting gamma values.

Recent developments in HAR in dark have shed light on the important role of unique datasets that capture the challenges posed in low-light environments. ARID and ELLAR are two of the notable works in this field. Both are designed to judge the performance of HAR models under low-light conditions. A detailed comparison between these datasets is provide in Table 2.1.

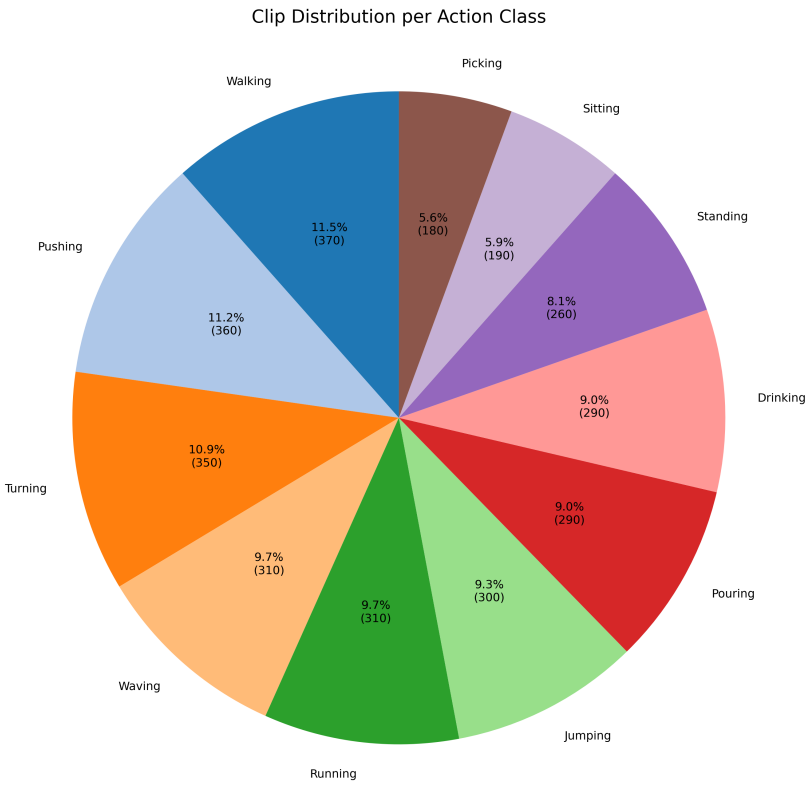


Figure 2.5: Distribution of video clips

Some of the contributions in given HAR for dark scenarios are given in Table 2.2

Table 2.1: Comparison between ARID and ELLAR datasets.

Aspect	**ARID**	**ELLAR**
Total Clips	5,572	12,078
Lighting Conditions	Low-light (not extreme)	Extremely low-light (severe noise, blur)
Action Classes	11	12
Best Model	DarkLight (94.04%)	DGAM (38.42%)
Best Generalization	Fails on ELLAR (13.21%)	Trains well on ELLAR and generalizes to others

Table 2.2: Literature Study for HAR

Paper Title	**Year**	**Publication**	**Dataset Used**	**Key Contribution**
Going deeper into action recognition: A survey (Samitha Herath et al.) [4]	2017	Image and Vision Computing	KTH, UCF101, HMDB51	Comprehensive survey between traditional and deep-learning based approaches, also emphasized importance of temporal dependencies and real-world challenges.

Continued on next page

Table 2.2 – continued from previous page

Paper Title	**Year**	**Publication**	**Dataset Used**	**Key Contribution**
Action recognition in dark videos using spatio-temporal features and bidirectional encoder representations from transformers (Himanshu Singh et al.) [5]	2022	IEEE Transactions on Artificial Intelligence	Custom dark video dataset	A transformer based framework using spatio-temporal features and BERT, showing robustness in dark environments.
Video-based human action recognition using deep learning: a review (Pham et al.) [6]	2022	arXiv preprint arXiv:2208.03775	UCF101, HMDB51, Kinetics	Review of deep learning methods for HAR, emphasized the role of 3D CNNs snd RNNs in extracting spatial-temporal features.
Human action recognition using deep learning methods (Zeqi Yu and Wei Qi Yan) [7]	2020	IVCNZ Conference	UCF101, HMDB51	Compared 3D CNNs, CNN+LSTM, and Two-Stream CNNs, also 3D CNNs can learn spatial and temporal features but are computationally expensive.

Continued on next page

Table 2.2 – continued from previous page

Paper Title	**Year**	**Publication**	**Dataset Used**	**Key Contribution**
Long-term recurrent convolutional networks for visual recognition and description (Donahue et al.) [8]	2015	CVPR	UCF101, HMDB51	Proposed LRCN model.
Action recognition for intelligent surveillance system using LRCN with attention mechanisms (Athilakshmi et al.) [9]	2024	ICAECT (IEEE)	UCF50, UR Fall Detection Dataset	LRCN model was enhanced with self-attention mechanism.
A comparative study using improved LSTM/GRU for human action recognition (Azhee Wria Muhamad and Aree Ali Mohammed) [10]	2022	–	UCF101	Comparison of GRU and improved LSTM models, GRU achieved high accuracy (92.9%) with fewer parameters.

Continued on next page

Table 2.2 – continued from previous page

Paper Title	**Year**	**Publication**	**Dataset Used**	**Key Contribution**
CLAHE-based low-light image enhancement for robust object detection in overhead power transmission system (Zhikang Yuan et al.) [11]	2023	IEEE Transactions on Power Delivery	Custom overhead transmission dataset	Introduction of CLAHE-based image enhancement.
ARID: A new dataset for recognizing action in the dark (Yuecong Xu et al.) [2]	2020	arXiv preprint arXiv:2006.03876	ARID	Introduction of a dataset ARID of low-light videos for HAR in Dark.
ELLAR: An Action Recognition Dataset for Extremely Low-Light Conditions with Dual Gamma Adaptive Modulation (Minse Ha et al.) [3]	2024	Asian Conference on Computer Vision	ELLAR	Intoduction of new dataset ELLAR for extremely low-light, DGAM method to enhance frames, improved recognition by 3.39%.

Continued on next page

Table 2.2 – continued from previous page

Paper Title	**Year**	**Publication**	**Dataset Used**	**Key Contribution**
Low-light video image enhancement based on multiscale retinex-like algorithm (H. Liu et al.) [12]	2016	Chinese Control and Decision Conference (CCDC)	–	Introduction of new image enhancement technique MSR(Multiscale Retinex-like enhancement method).

2.6 Observation from the papers

While the available dataset has given promising results but the limited number of dataset is evident. We need a more diverse dataset which has different classes from different fields and in varied lighting conditions.

2.7 Contribution

Our primary contribution is to create a new dataset which has different classes from different fields and which can help and support the research in this field.

Chapter 3

Problem Definition

Developing a Human Action Recognition (HAR) system for low-light environment is essential for application smart homes, nighttime surveillance. There are a few datasets like ARID and ELLAR which has enabled progress in this area, but still there are limitations. This project addresses the gap by creating a challenging dataset for HAR in dark. NoctAct-HAR combines self-recorded videos captured in dark environment and videos taken from ARID dataset. We also introduced variations like brightness reduction, gamma correction, synthetic noise to simulate extreme scenarios. This dataset aims to support future research on HAR system. Comparative evaluations with existing dataset is also conducted to benchmark its effectiveness.

Chapter 4

Methodology

We have used our proposed dataset for this entire process.

4.1 Preprocessing

Video sequences are loaded, converted into frames of 32x32 and normalized. It may be converted into grayscale depending on the image enhancement technique used.

4.2 Image Enhancement Technique

In the CLAHE[11] technique, the input frame is first converted to HSV then back to BGR, and finally to grayscale, after that CLAHE is applied and brightness and contrast are adjusted using OpenCV function and gamma correction.

In the MSR[12] approach the frame is converted to RGB from BGR and then MSR algorithm is applied which uses Gaussian blurs at multiple scales. Each scale computes log difference between blurred and original image and then results are averaged and normalized. Image enhancement technique is used to enhance frame visibility under dark conditions.

4.3 Model

We use LRCN[8] as our model which combines CNN and GRU for spatial and temporal feature extraction respectively. After each frames are enhanced they are then passed to CNN. CNN has Conv2D, ReLU, BatchNorm and MaxPooling layers and dropout for regularization which is optional. The model captures spatial features in this layer. After which the the CNN features are then wrapped in TimeDistributed and passed to GRU[10]. In this layer the model extracts temporal dependencies between the frames, which are important for understanding the motion in HAR. Model architecture is shown in the fig. 4.2.

4.4 Output and Evaluation

The output that we get after GRU layer is passed to Dense layers followed by softmax activation function which returns probability distribution of classes as output. The model is evaluated using Accuracy, Precision, Recall, and confusion matrix.

Initially, we used CLAHE (Contrast Limited Adaptive Histogram Equalization) but later switched to MSR (Multi-Scale Retinex) because MSR offers better enhancement as it mimics human visual perception. CLAHE improves local contrast but it may lose small details because of different lighting conditions, whereas MSR uses multiple scales and works on color channels. It preserves both color fidelity and fine textures under low-light which makes it more effective for enhancing frames in dark video frames.

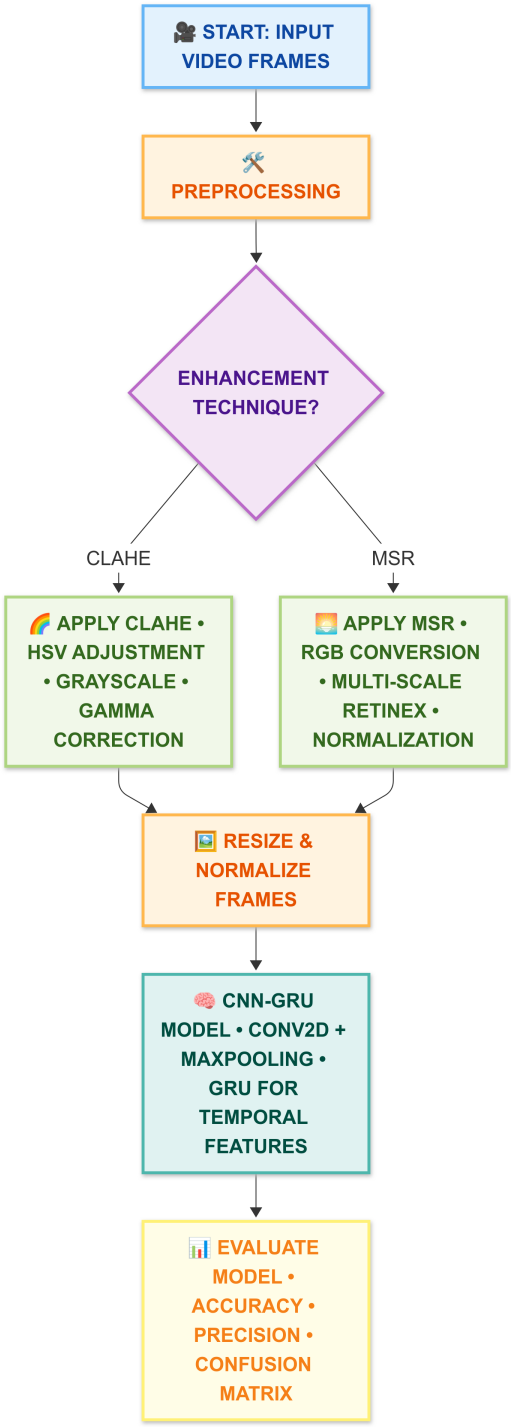


Figure 4.1: Flowchart

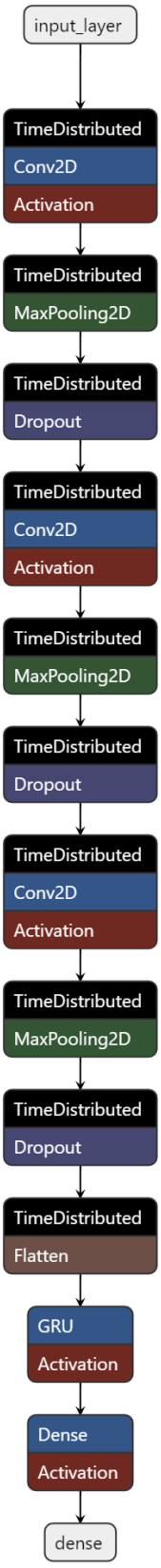


Figure 4.2: Model Architecture

Chapter 5

Results and Analysis

5.1 Dataset

The limitation of available datasets in dark low-illumination continues to be a challenge, to overcome this we built a robust dataset for HAR in dark. The dataset has a total of 6613 videos spanning 12 distinct classes. The videos are in .mp4 format and each videos focuses on single action. The videos are divided into different folders each folder consisting of one action.

The dataset is collected by us on our smartphones and for some classes the videos are taken from ARID dataset. We have also applied transformation on some videos like brightness and darkness adjustments, rotation, gamma correction. The videos are a mix of some containing original videos and some containing transformed videos in each class. The actions are normal human activity. They classes are as follows: **Boxing, Drink, Lifting Weights, Picking, Push, Receiving the Phone, Run, Stand, Throwing objects, Walking on Stairs, Walking with Flashlight, Waving.** Each class consists of 583 videos with an exception in lifting weights which has 423 videos and walking on stairs which has 360 videos.

The total duration of the videos is 17969.42 seconds and the videos are taken outside in dark except those taken from the ARID dataset. To get

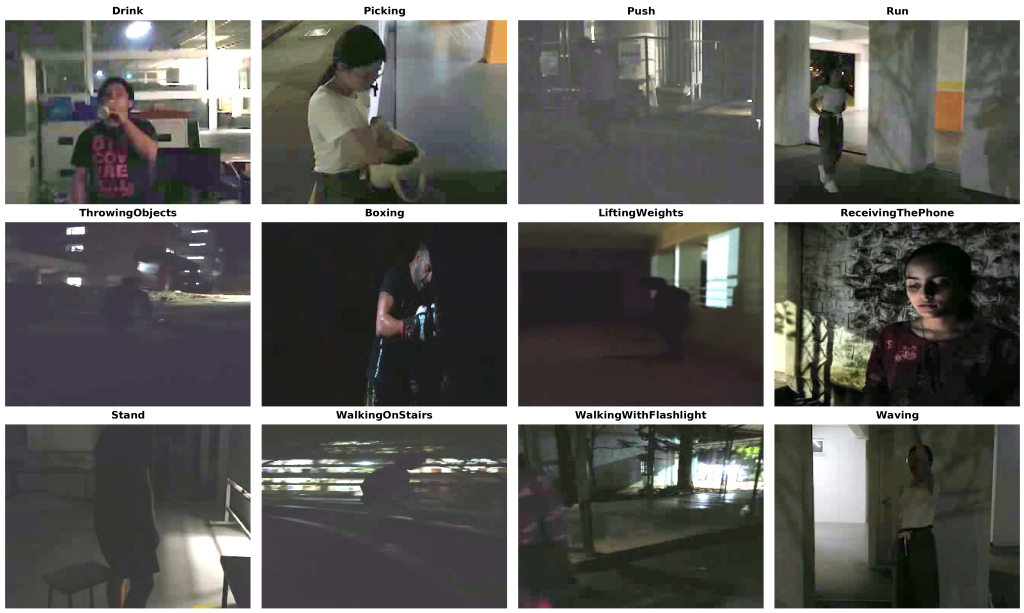


Figure 5.1: Sample frames for each of the 12 action classes of the NoctAct-HAR dataset. All samples are tuned properly for better display.

a more clear idea about the total duration video count and other statistics here is a visual summary:

To get in-depth understanding of this dataset we have also illustrated a pie chart. Fig 5.2 shows the percentage of videos per class and total duration(in seconds) of videos for each class. The dataset contains some of the original videos and some of the transformed videos which is valuable resource for training and testing the models for dark environments.

Table 5.1: Per-Class Statistics for Video Dataset

Class Name	**Number of Videos**	**Total Duration (seconds)**
boxing	583	1102.04
Drink	583	1975.23
lifting weights	423	833.10
Picking	583	1736.60
Push	583	1634.83
receiving the phone	583	1416.19
Run	583	1759.57
stand	583	1849.89
Throwing objects	583	1577.13
walking on stairs	360	725.53
walking with flashlight	583	1642.78
waving	583	1716.51
Total	**6613**	**17969.42**

5.2 Evaluation Matrix

These are the evaluation matrix we have used where :True Positive(TP),True Negative(TN),False Positive(FP),False Negative(FP)

5.2.1 Accuracy

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FP}$$

5.2.2 Precision

$$\text{Precision} = \frac{TP}{TP + FP}$$

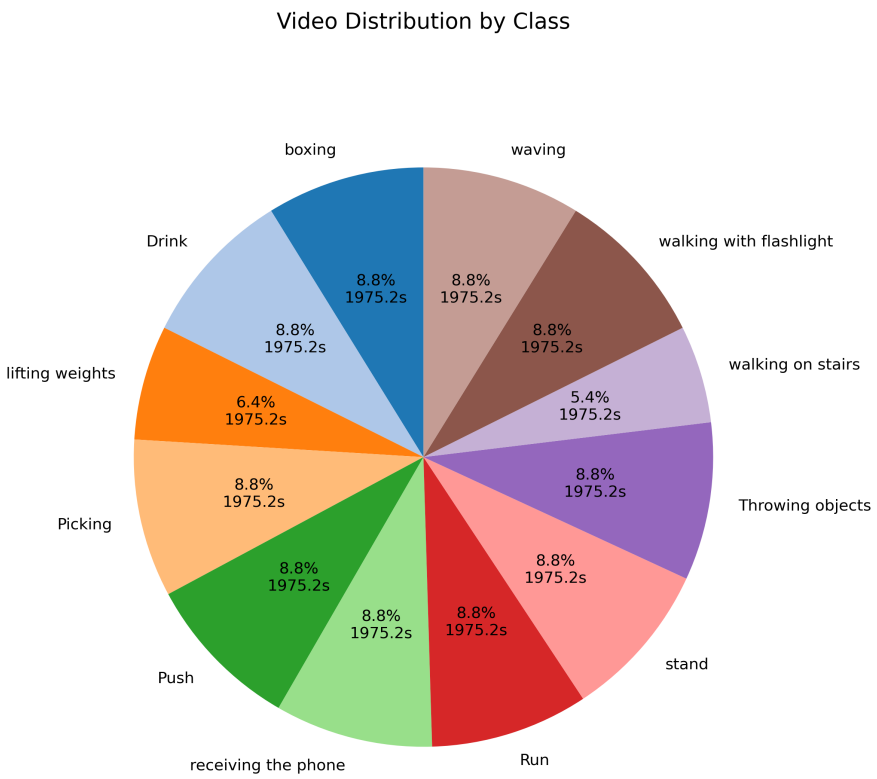


Figure 5.2: Percentage of Videos and duration of videos per class in the NoctAct-HAR dataset.

5.2.3 Recall

$$\text{Recall} = \frac{TP}{TP + FN}$$

5.3 Results

We propose the performance of NoctAct-HAR when used CLAHE and MSR in this section. We have used confusion matrix, accuracy recall and precision

to access the effectiveness.

A quick summary about the model's performance on CLAHE-based approach is provided in table 5.2.

- Accuracy: 78.87%
- Precision: 0.80
- Recall: 0.79

Table 5.2: Confusion Matrix for CLAHE

148	**10**	**0**	**1**	**1**	**2**	**2**	**0**	**2**	**0**
0	**98**	**0**	**9**	**4**	**1**	**0**	**6**	**0**	**8**
0	**0**	**120**	**0**	**1**	**0**	**2**	**1**	**0**	**0**
0	**9**	**0**	**99**	**1**	**0**	**2**	**0**	**0**	**13**
0	**12**	**8**	**59**	**0**	**34**	**6**	**7**	**0**	**15**
1	**7**	**0**	**4**	**10**	**1**	**84**	**5**	**2**	**14**
1	**13**	**0**	**10**	**1**	**5**	**78**	**9**	**0**	**8**
0	**1**	**0**	**1**	**1**	**0**	**0**	**136**	**1**	**6**
0	**0**	**2**	**0**	**0**	**2**	**2**	**0**	**88**	**3**
0	**2**	**0**	**6**	**3**	**0**	**1**	**7**	**1**	**125**

The results after switching to MSR-based approach where model's overall performance improved is shown in table. 5.3.

- Accuracy: 83.73%
- Precision: 0.85
- Recall: 0.84

Table 5.3: Confusion Matrix for MSR

161	**1**	**0**	**0**	**0**	**0**	**0**	**3**	**1**	**0**
2	**90**	**0**	**15**	**8**	**3**	**3**	**9**	**1**	**3**
0	**0**	**123**	**0**	**1**	**0**	**0**	**2**	**0**	**1**
3	**0**	**0**	**112**	**2**	**0**	**9**	**2**	**0**	**2**
0	**3**	**2**	**81**	**0**	**38**	**12**	**1**	**0**	**4**
0	**1**	**0**	**1**	**1**	**125**	**5**	**0**	**6**	**2**
1	**3**	**0**	**7**	**7**	**2**	**118**	**5**	**1**	**6**
0	**1**	**0**	**0**	**0**	**2**	**1**	**141**	**0**	**0**
0	**0**	**0**	**2**	**1**	**4**	**0**	**0**	**138**	**1**
0	**0**	**0**	**4**	**0**	**12**	**5**	**0**	**0**	**123**

From the above tables, it evident that MSR-based approach outperformed CLAHE-based approach in terms of accuracy, precision, and recall. The confusion matrix also showed improvements acroos multiple classes. Overall, MSR is better for preprocessing as it enhances model's visibility under low-light conditions.

5.4 Comparison

The comparison between NoctAct-HAR and ARID dataset is given in the Table. 5.4.

Table 5.4: Comparison of ARID and NoctAct-HAR dataset

Feature	**ARID [2]**	**NoctAct-HAR (proposed)**
Number of Classes	11	12
Classes	Wave, Walk, Turn, Stand, Sit, Run, Pick, Push, Pour, Jump, Drink	**Boxing**, Drink, **Lifting Weights**, Picking, Push, **Receiving the Phone**, Run, Stand, **Throwing objects**, **Walking on Stairs**, **Walking with Flashlight**, Waving
Total Videos	5,572	6,613
Video Format	AVI, MP4	MP4
Lighting Environment	Low-light indoor (controlled)	Low-light outdoor (uncontrolled) and low-light indoor (controlled)
Camera Setup	Static (fixed camera)	Mobile and fixed camera
Environmental Control	Controlled	Uncontrolled and Controlled

We compared the model's performance on both the dataset. The table 5.5 shows the difference.

Table 5.5: Performance Comparison on ARID and NoctAct-HAR Dataset using CLAHE and MSR

Dataset	**Enhancement**	**Accuracy (%)**	**Precision**	**Recall**
ARID[2]	CLAHE	79.83	0.80	0.80
NoctAct-HAR(Ours)	CLAHE	78.87	0.80	0.79
ARID[2]	MSR	79.19	0.80	0.79
NoctAct-HAR(Ours)	MSR	**83.73**	**0.85**	**0.84**

Table 5.6 shows the top-1 accuracy of ARID dataset along with its enhanced version using different techniques across some of the models[2].

Table 5.6: Comparison of Top-1 Accuracy for Different Existing Models with ARID Dataset

Dataset	Top-1 Accuracy			
	C3D	I3D-RGB	3D-ResNet	3D-ResNext
ARID-GIC[2]	44.09	69.14	75.15	78.06
ARID-HE[2]	39.49	63.67	65.49	75.82
ARID-LIME[2]	39.61	73.02	75.45	77.40
ARID-BIMEF[2]	45.23	68.89	68.28	73.39
ARID-KinD[2]	46.64	67.55	70.59	69.62
ARID without any enhancement	40.34	68.29	71.57	74.73

From the above tables it is evident that our model and enhancement technique performed better on ARID as well as NoctAct-HAR. MSR enhancement technique when used on our proposed dataset outperformed ARID dataset achieving higher accuracy on our model. NoctAct-HAR's diversity

may have helped in better generalization, which makes it suitable for practical use.

For better understanding we have provided the bar chart of mean and standard deviation values in Fig. 5.3. Our dataset has high mean which indicates somewhat bright pixels because of the transformations and high deviation which shows variation in pixel intensity. It is good for generalization.

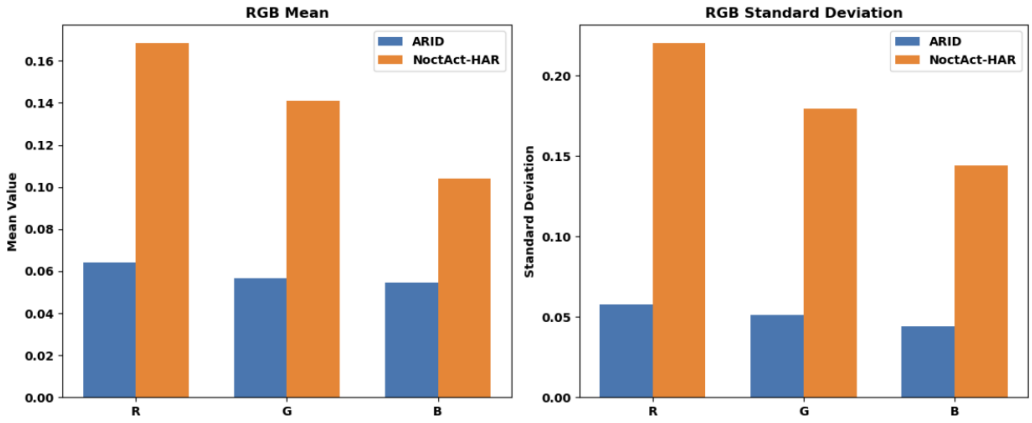


Figure 5.3: Bar charts of RGB Mean (Left) and RGB Standard Deviation (Right) values for ARID and NoctAct-HAR.

We have also shown the RGB values and Y value histograms of NoctAct-HAR and ARID in Fig. 5.4 .

The histograms of ARID and NoctAct-HAR in Fig. 5.4 depicted the characteristics of both the dataset. ARID dataset is concentrated towards lower region values whereas there is slight deviattion in NoctAct-HAR probably because of transformations applied. This is justified further by sample frames and their RGB and Y values between NoctAct-HAR and ARID dataset, shown in Fig. 5.5.

NoctAct-HAR has slightly bright pixels even after taking the videos in darkness because of the transformations applied on some of the videos how-

ever it still is a good choice for further improvement in HAR in dark environment.

To analyze the effectiveness of CLAHE and MSR, we compared the original frames and the enhanced frames using both the techniques. The fig. 5.6 demonstrates the same.

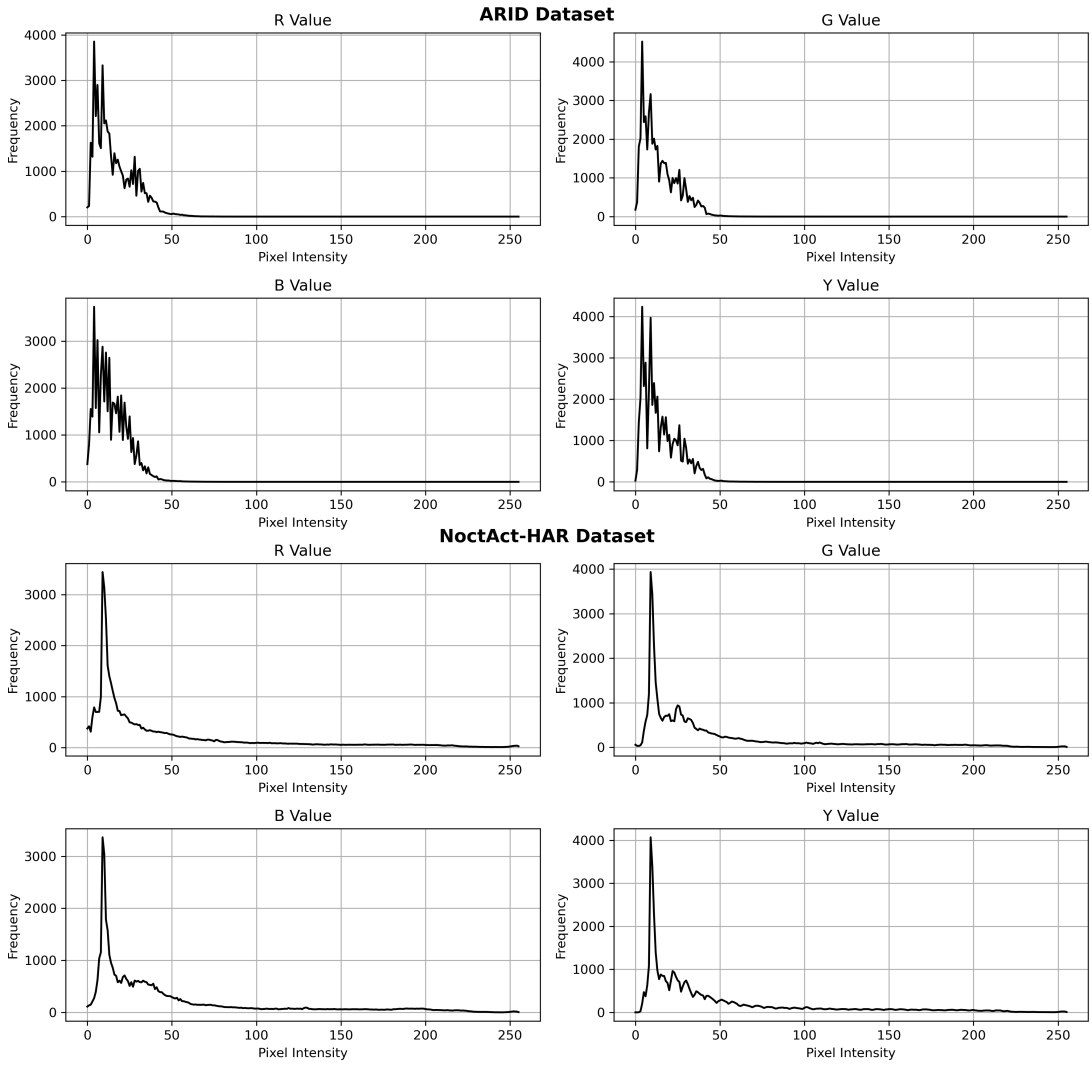


Figure 5.4: Histograms for RGB and Y values of ARID (Top) and NoctAct-HAR(Bottom).

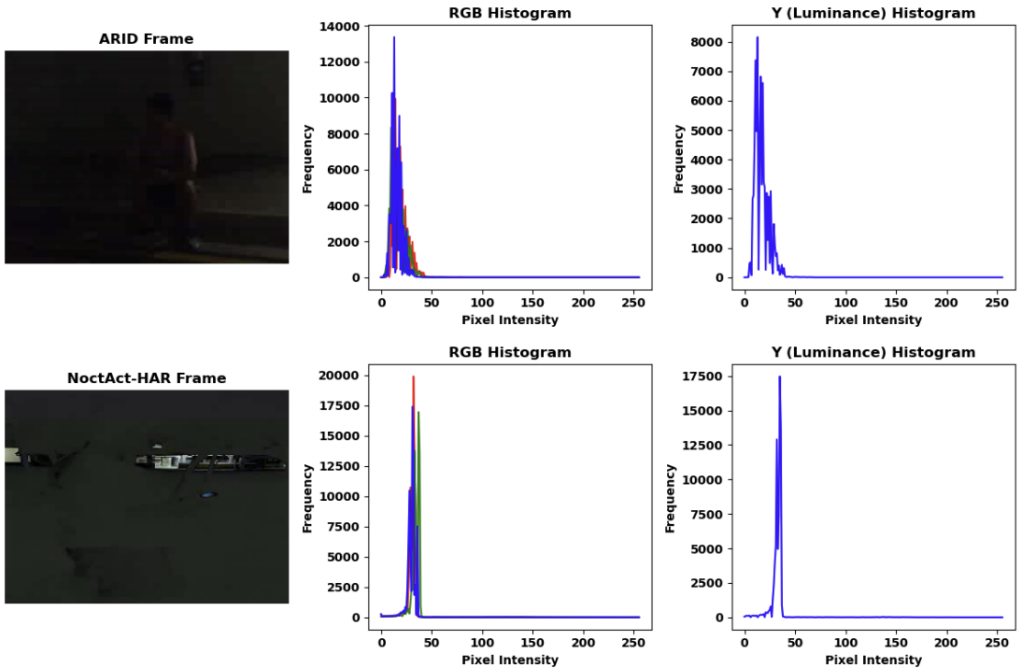


Figure 5.5: Comparison of sampled frames and the RGB and Y value histograms of their corresponding videos from ARID dataset and NoctAct-HAR.

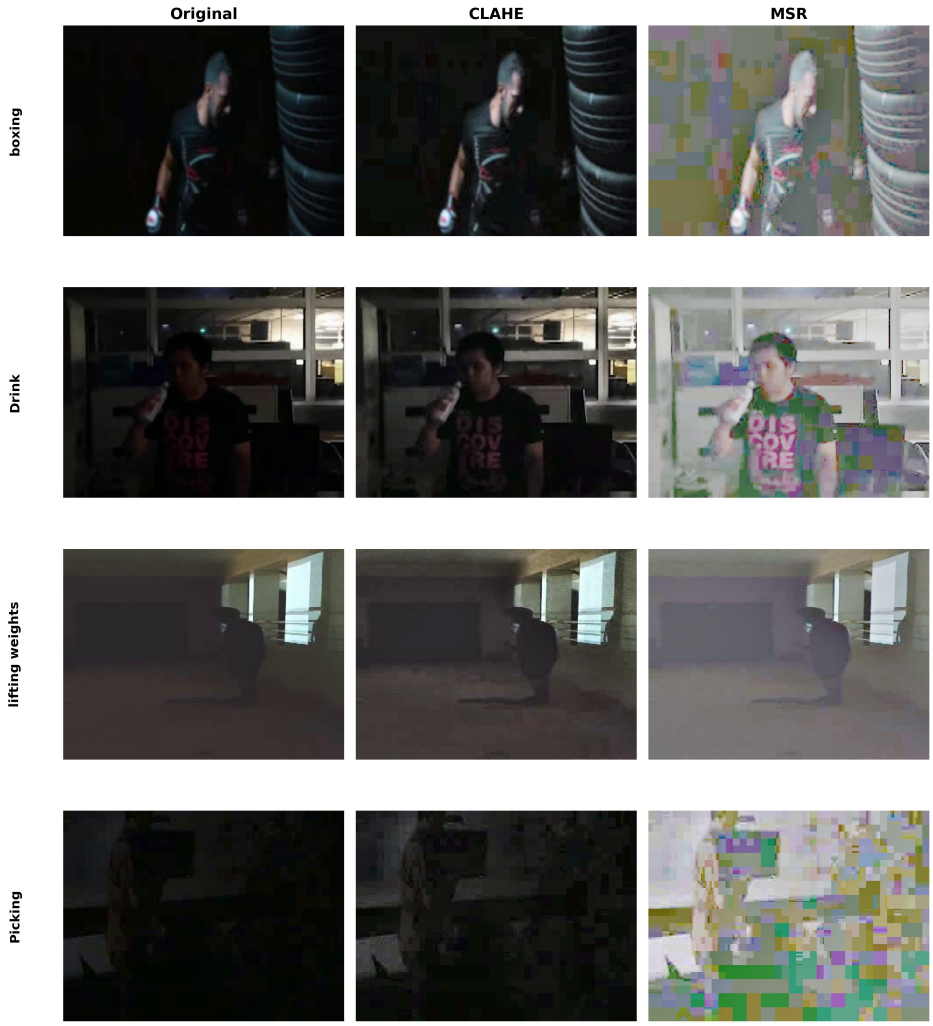


Figure 5.6: **Original, CLAHE and MSR**-enhanced frames across few classes.

Chapter 6

Conclusion and Future work

In this paper, we proposed a new dataset for action recognition in dark environment which is designed to overcome the lack of datasets in this domain. NoctAct-HAR includes more number of classes and hence more number of videos, in which some are even transformed for more challenging scenarios. We also saw the effectiveness of image enhancement techniques- CLAHE and MSR on model's overall performance.

In the future, the dataset can be further expanded to support extensive research in this field. Diverse actions can be included along with the varied lighting conditions. Multi-person action can also be added. Finally, releasing the dataset publicly would contribute in the advancement of research in HAR in Dark.

References

[1] F. M. Shiri, T. Perumal, N. Mustapha, and R. Mohamed, "A comprehensive overview and comparative analysis on deep learning models: Cnn, rnn, lstm, gru," *arXiv preprint arXiv:2305.17473*, 2023.

[2] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, "Arid: A new dataset for recognizing action in the dark," in *Deep Learning for Human Activity Recognition* (X. Li, M. Wu, Z. Chen, and L. Zhang, eds.), (Singapore), pp. 70–84, Springer Singapore, 2021.

[3] M. Ha, W.-G. Bae, G. Bae, and J. T. Lee, "Ellar: An action recognition dataset for extremely low-light conditions with dual gamma adaptive modulation," in *Proceedings of the Asian Conference on Computer Vision*, pp. 800–817, 2024.

[4] S. Herath, M. Harandi, and F. Porikli, "Going deeper into action recognition: A survey," *Image and vision computing*, vol. 60, pp. 4–21, 2017.

[5] H. Singh, S. Suman, B. N. Subudhi, V. Jakhetiya, and A. Ghosh, "Action recognition in dark videos using spatio-temporal features and bidirectional encoder representations from transformers," *IEEE Transactions on Artificial Intelligence*, vol. 4, no. 6, pp. 1461–1471, 2022.

[6] H. H. Pham, L. Khoudour, A. Crouzil, P. Zegers, and S. A. Velastin, "Video-based human action recognition using deep learning: a review," *arXiv preprint arXiv:2208.03775*, 2022.

[7] Z. Yu and W. Q. Yan, "Human action recognition using deep learning methods," in *2020 35th International Conference on Image and Vision Computing New Zealand (IVCNZ)*, pp. 1–6, 2020.

[8] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2625–2634, 2015.

[9] R. Athilakshmi, C. K. T. Muddangula, V. Tummala, P. S. C. SaiNagaKrishna, S. C. Kota, and V. J. Prasad, "Action recognition for intelligent surveillance system using lrcn with attention mechanisms," in *2024 Fourth International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT)*, pp. 1–6, IEEE, 2024.

[10] A. W. Muhamad and A. A. Mohammed, "A comparative study using improved lstm/gru for human action recognition," 2022.

[11] Z. Yuan, J. Zeng, Z. Wei, L. Jin, S. Zhao, X. Liu, Y. Zhang, and G. Zhou, "Clahe-based low-light image enhancement for robust object detection in overhead power transmission system," *IEEE Transactions on Power Delivery*, vol. 38, no. 3, pp. 2240–2243, 2023.

[12] H. Liu, X. Sun, H. Han, and W. Cao, "Low-light video image enhancement based on multiscale retinex-like algorithm," in *2016 Chinese Control and Decision Conference (CCDC)*, pp. 3712–3715, 2016.