Thesis-AshishDhyani-2.docx

by Ashish Dhyani

Submission date: 05-Oct-2021 03:14PM (UTC+0100)
Submission ID: 160377315
File name: Thesis-AshishDhyani.docx (2.79M)
Word count: 17914
Character count: 94452

ACTION RECOGNITION IN DARK USING DEEP LEARNING

ASHISH DHYANI
Student ID - 975129

Under the supervision of
KARTHICK KALIANNAN NEELAMOHAN

Interim Report
Master of Science in Data Science
LIVERPOOL JOHN MOORES UNIVERSITY, UNITED KINGDOM

OCTOBER 2021

Acknowledgements

I like to take the chance to thank everyone who helped and supported in reaching this stage of the project and in producing this interim report document.

I would like to thank my supervisor, Karthick Kaliannan Neelamohan, for his continuous support and the valuable suggestions for this research which helped me to complete this project successfully.

I also thank Dr Manoj Jayabelan, the researcher at the Liverpool John Moores University, United Kingdom, who advised us throughout our tenure and gave many of the details on how to write effective thesis.

I cannot forget to thank my parents, my wife and son who encouraged me to pursue an MSc degree and upgrade my skill in Deep learning and artificial intelligence and supported me with all of the physiological and mentally available means.

I thank upGrad institute which introduced us with very good course, MSc. Degree with Liverpool John Moores University, on Deep Learning and also thank my student mentors who cleared my doubts time to time through-out my journey from post graduate diploma to masters program.

I thank all of my friends in Singapore and India who have always supported me with every encouraging words and motivation. Without all of their support and motivation, this report would have never been come out so nicely.

Proclamation

This document is prepared upon the request of the Liverpool John Moores University, United Kingdom as an interim report for the author's MSc project. A large portion of this document is expected to be included in the final report of the author's MSc Project after performing the appropriate modifications. The MSc project, which is titled as "Action Recognition in Dark using Deep Learning", is supervised by Karthick Kaliannan Neelamohan, Upgrad. The author proclaims that the entirety of this document is original and it obeys all valid documents rules and regulations of the Liverpool John Moores University, United Kingdom.

Ashish Dhyani

Abstract

Action recognition in the dark is closely related to real life scenarios, such as traffic management, animal breeding, human-computer interactions and crime preventions. Action recognition in the dark is the need of the hour where world is trying hard to build different algorithms which can augment our capabilities in this area. Videos captured in the low light conditions or in dark suffer from poor visibility and due to that extracting useful features from these videos becomes very hard which eventually makes the action recognition difficult. Most of the feature extraction methods are designed to work on high quality input videos. Some work is done on the low-light enhancement methods for the images which we can be utilized in the current paper. ARID Dataset, specially created for dark videos, is used in the current research and which explores different low light enhancement methods that are efficient in illumination of the each frame of videos and eventually help in feature extraction from videos. Furthermore, using the features extracted from video, current research proposes to build a deep neural network, using two-dimensional and three-dimensional convolutional neural networks, to extract the spatiotemporal features from the enhanced videos which will eventually help in recognizing the human actions in the videos. The challenges faced during the illumination phase of video sequences, features extraction from those enhanced images and some best practices in these areas are discussed.

Additionally, some of the state-of-the-art methods for low light illumination and human action recognitions on the ARID dataset are benchmarked and the method which is best in accuracy and low in computational time is proposed. It is important to find out whether the best state-of-the-art method for low light enhancement is really required for human action recognition because focus should be primarily on the human action recognition and not on the visual quality of the videos.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ARID – Action Recognition in Dark

CNN – Convolutional Neural Networks

RNN – Recurrent Neural Networks

2D/3D – two dimensional/three dimentional

GAN – Generative Adversarial Networks

LSTM – Long and Short Term Memory (RNN)

GRU – Gated Recurrent Unit (GRU)

Chapter 1. Introduction

1.1. Background of study / Problem Statement

In today's world, machines are helping us in various ways by identifying various situations and problems around us. Human action recognition is one of the areas where these machine learning algorithms can help us and augment our capabilities around it. It is important to have the capability to identify human actions not only in good lighting conditions but also in dark or low lighting conditions.

Action recognition on normal videos with good visibility has done a lot of progress in recent times and many machine learning algorithms are able to do it efficiently but action recognition in the low light situations is still pretty challenging and a hard nut to crack because the pictures or videos captured during the night time or low light situations suffer from poor visibility and contains a lot of noise and sometimes these videos are even difficult for human eyes to recognize accurately.

Innovations at hardware side also have done lot of progress where cameras are getting equipped with night vision capabilities but the night vision sensors degrade the quality of the images or videos by introducing lot of noise and it gets more difficult for present day feature extraction algorithms to deal with these noises. Another drawback with the night vision sensors is that they make the surveillance cameras more expensive and it's difficult for everyone to afford and adopt these technologies. So there is a need of software solutions which are capable of recognizing the actions in videos which are shot in dark without the help of any night vision capabilities.

Some use cases for dark videos action recognition are:

- Night surveillance at very sensitive areas around the cities for crime investigations
- Surveillance at the country borders to stop illegal infiltration
- Self driving cars at night time
- Outdoors home surveillance during night hours or in low light situations
- Studies for wild animals breeding and behaviour during night time without disturbing them.
- Traffic management during the night hours where there is no enough light.

- Robots which can facilitate human-computer interactions.

These all issues are very important which world would like to address. According to an article (TheSleepJudge Editorial Team, 2020) about crimes in United States, more than 50% of the major crimes happen during night time. And most of the crimes are never reported which happen in the dark because there are either no proofs or if it's get reported, video is not clear to identify the criminal. So it's a major issue not only for United States but also for India as well where crime rates are more and happen during the night time and to tackle these situations we need a model which can enhance the low light in the dark videos, eliminate the noise and identify the human action.

One of the articles (Guglielmi, 2018) tells about that how wild mammal animals turning nocturnal because of the human activities in the forests during day time. Also humans use cameras with flash lights to study the wild animals' behaviours at night which actually disturbs them a lot during night time as well. We need to learn to see in the dark without the use of flash lights.

1.2. Aim & Objectives

The aim of this research is to do the comparative study on the present capabilities of low light enhancement methods and build a model and recognize the human actions in ARID dataset. The goal of this study is to identify the best low light enhancement method for videos which are shot in dark and which works the best and gives the highest accuracy with an action recognition deep learning model.

The research objectives are formulated based on the aim of this study, which are as follows:

- To analyze the most popular low light enhancement methods and identify the best method among all of them which works the best for human action recognition.
- To propose a deep learning model which works the best in identifying the human actions with low light enhanced videos.
- To perform comparative study on 2D CNN and 3D CNN methods for human action recognition.

1.3. Research Questions

The below research questions are formulated based on Literature Review done in the field of human action recognition in dark:

- Are there any conclusions to use a particular low light enhancement method which actually helps in human action recognition?
- Is the best low light enhancement method really required to get high accuracy in human recognition or a simple low light enhancement method is good enough for the model to perform well in identifying the human actions? Here we are going to focus on the human recognition in dark, not on the visual quality of the videos.
- What will be the Human action recognition model's accuracy for different low light enhancement methods?

Can we use different algorithms for video enhancement based on business needs? User might be interested to see the actual enhanced video after human action recognition is done by the model.

1.4. Scope of Study

Due to lack of time frame, the scope of the study will be limited as below:

- The dataset is taken from ARID (Xu et al., 2021) dataset which is publicly available and data validation is not the part of this study. The scope of study is to analyze only 11 human actions mentioned in the data but this can be extended to any number of human actions in the future.
- This study contributes to the comparative study on different low light enhancement techniques, comparative study on human action recognition techniques and suggests the best combination of both which help in recognizing the human action in the videos which are shot in the dark.
- In case of low-light illumination, this study will limit to use some of the state-of-the-art techniques like KinD (Zhang et al., 2019), LIME (Guo et al., 2017), EnlightenGAN (Jiang et al., 2021) and Gamma Intensity Correction and plans to do the comparative analysis by comparing the different metrics.

- In case of human action recognition, this study will limit to use some of methods like action bank feature extraction (Ijjina and Mohan, 2014), utilize a rolling prediction average and utilize the 2D, 3D CNN, RNN methods for model building.

1.5. Significance of Study

The significance of this study is to explore various low light enhancement methods and find the best method which is efficient in recognizing the human action in the dark videos and will augment human capabilities during night time.

There are many areas where this research will help like night surveillance. To make our city and country safe, we need the capabilities to do surveillance at very crime sensitive areas. This model will help in recognizing the human actions during the night hours; we don't need to deploy a human at every place for surveillance. Most of the times it is very difficult for the humans as well to see in the dark without night vision capabilities. This action recognition method will help our police to protect our neighbourhood and people.

The second use case is Self Driving Cars; there are lot of researches going on for autonomous cars. As we know, self driving cars is still a big challenge and auto maker companies still facing a lot of challenges to make a perfect autonomous car where driving in the dark will be another hard nut to crack. I hope this study will be able to help this area to some extent because we need a model which can recognize things in dark.

The third use case is Human to Robot communication; there is lot of research going in the area of voice communication but it will great to see if robots can react based on the our actions instead of voice commands, then it will in true sense robot can help us in our day to day activities.

This model, with the comparatively analysis, will be able to help data scientists to find different methods to identify human actions in various scenarios. This study will be extended to give suggestions about different low-light enhancement methods which are best suited for different dark environment scenarios and also to the human action recognition related algorithm which do the best so far in these type of scenarios.

1.6. Structure of Study

This interim report aims to document what have already done in this project so far and putting the plan for what is expected to be done later on. Most of this report can be seen as a pre-final report for the analysis on the dataset and related research.

Chapter 2 presents a review of the literature related to this project. This chapter includes reviews on different illumination methods and human action recognition on the video sequences which have already been used in this project and/or which are expected to be used in the subsequent phases. This chapter mostly talks about different illumination methods based on deep learning with and without paired image supervision. This chapter also discusses about the pros and cons of the different methods and which all methods will be used in the comparative analysis. Although some of the other parts of the literature review haven't been touched yet in application, they are expected to be needed in the analysis on the dataset in the coming months.

Chapter 3 discusses the research methodology which was followed in the project so far. It presents most of the details of the proposed method and talks about different steps needs to be taken for model building, training and finally predicting. This chapter includes the details about different neural networks, different regularization techniques. Along with that there is information about "how the dataset was selected?", "what all hyper parameters will be used?", "what all pre-processing and data transformation technique will be used?" before we actually starts the model training and evaluation.

Appendix A contains the project plan in details and tells the current progress on the project and what all steps are remaining for a complete research. Appendix B contains the Research Proposal and Appendix C contains the ethics forms.

Chapter 2. Literature Review

2.1. Introduction

In the last decade, many low light enhancement methods were proposed with different capabilities and different methods to enhance the dark images. Some of the most popular methods are highlighted and discussed in this section which work the best with ARID dataset and eventually help in identifying the human action from the dataset.

Some methods are based on histogram equalization, some are based on Gamma Intensity Correction which amplify the low light in the images. Some methods are deep learning based which require images in pairs of dark and bright images. Some methods are based on Retinex Theory where it's suggested to split the image in two parts, Reflectance and illumination where sometimes reflectance is considered as the enhanced image.

2.2. Low illuminated videos and images

Sometimes when videos and images are shot in low lighting condition, they end up with grainy, under saturated, low-contrast, muddy video footage. The lack of light destroys the images or the videos. An expert (Hyman, 2010) suggests many ways to avoid low light images by using

- Micro Pro LED light to brighten up the scene.
- Choose the biggest aperture your video camera allows which helps more light to enter the camera.
- Slow down shutter speed which will allow to higher exposure to the light.
- Reduce the frame rate and video noise in post filters.
- Heat sensors in the cameras.

But sometimes, it's not possible to add the lighting to the scenes like surveillance cameras at country side or while shooting the videos and images in the forest during the night time, adding the light to camera can interfere in wild life. Choosing different apertures or shutter speed while shooting with surveillance cameras is not possible because it will make cameras very expensive and have these expensive cameras for surveillance purpose may not worth it, however we can use some good cameras at very sensitive areas.

2.3. Challenges to deal with low illuminated videos

Because of grainy, under saturated, low contrast images, videos suffer with lot of noise so it's hard to process those images and extract features; these noises interfere with feature extraction. There are lot of challenges to deal with these types of dark videos and images. An image captured with small aperture or fast shutter speed makes the image under exposed which leads to severe motion blur. Sometimes some cameras are not powerful enough to capture good quality of videos during night time.

Action recognition on normal videos with good visibility has done a lot of progress in recent times and many machine learning algorithms are able to do it efficiently but action recognition in the low light situations is still pretty challenging and a hard nut to crack because the pictures or videos captured during the night time or low light situations suffer from poor visibility and contains a lot of noise and sometimes these videos are even difficult for human eyes to recognize accurately. Images or videos of such type hard to process specially for neural networks, it will hard for algorithms to extract any useful features out of it and use it for prediction later.

Before starting action recognition on dark videos, it is important to illuminate and de-noise them properly and then start the feature extraction process which is going to take slightly more time in training the model and later predicting which makes it more challenging compare to properly illuminated videos or images.

2.4. Evolution of Low light illumination Techniques

Presently many algorithms are there in AI/ML world which are capable of illuminating the images and videos very well. If we look at the last decade history, there are a lot of inventions happened in this area. This area of research for low light enhancement techniques evolved from 'histogram equalization' to 'gamma intensity correction' to 'Retinex theory' to 'deep learning based neural networks'.

Some methods are based on colour correction using histogram equalization; some are based on Gamma Intensity Correction low light is amplified in the images and many methods are based on Retinex Therory where it is suggested to split the image in two parts, Reflectance and illumination where sometimes reflectance is considered as the enhanced image.

Recently, (Xu et al., 2021) proposed a new dataset (ARID) which contains 3784 video clips with 11 categories where each video is of around 1.2 seconds with 36 frames and all these videos are low contrast and low brightness videos. Figure1 shows snapshots from dark video sequence where we can clearly see, sometimes it's hard to recognize the action from naked eyes.

Together with dataset, a comparative study is done on the five different illumination methods which are based on Histogram Equalization, Gamma Intensity correction and Retinex Theory followed by human action classification on the videos. According to this study, authors advocate Gamma Intensity correction (GIC) which gives the highest accuracy (78.03%) among all the methods. Authors have also done comparison using 3D-ResNext-101(Hara et al., 2018) classification model with original dark videos and enhanced video with various methods and concludes that GIC method gives the best improvements (3.30%) in the accuracy of the model and KinD (Zhang et al., 2019) enhancement method actually lowers the accuracy of classification model by (5.11%) in ARID Dataset.

Figure 1 Dark Video frame in ARID dataset (Xu et al., 2021)

There are many algorithm which are based on Retinex Theory, for example: LIME (Guo et al., 2017) and Naturalness Preserved Enhancement Algorithm (Wang et al., 2013). One of the best Retinex based theories LIME was very successful illumination technique and had beaten most of the state-of-the-art techniques of its time.

LIME (Low-Light Image Enhancement via Illumination Map Estimation) method builds an illumination map and refines it by finding the maximum intensity of each pixel in R G B channels. For the refinement of illumination map, Augmented Lagrangian Multiplier (ALM) based algorithm is used which is also quite efficient algorithm and reduces the computational

cost. Illumination map estimation and refinement actually considers the neighbouring pixels which help in local consistency of illumination which helps in getting uniform illumination across the image. Gamma correction and denoising and recomposition are done to get the best results for illumination map. HDR dataset is used for this study and some low light images were chosen from the dataset and their LOE numbers are compared with all the other competitors. A comparative study is done on the 9 different image enhancement techniques with the proposed method and based on the visual comparison and LOE numbers, it can be clearly seen that the proposed method gives the clear advantage over all those state-of-the-art methods selected in the paper.

Naturalness Preserved Enhancement Algorithm (Wang et al., 2013), essentially focuses on preservation of naturalness of the image which is difficult to achieve in non-uniformly illuminated images. This paper as shown in Figure 2 proposes a new algorithm for non-uniform illumination images where it also proposes lightness-order-error measure for naturalness and bright-pass filter which helps in decomposing the image into reflectance and illumination and bi-log transformation is done to consider the balance between naturalness and details.

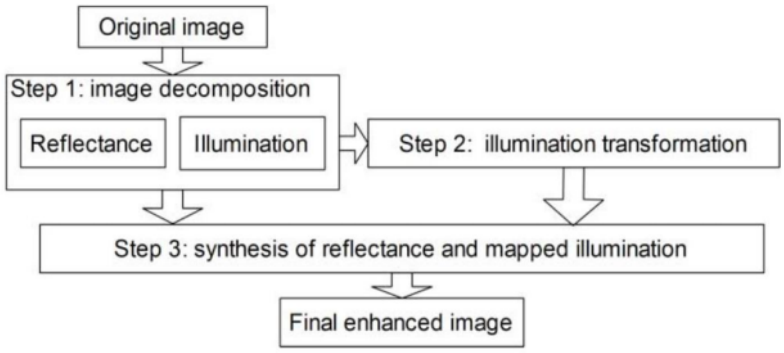


Figure 2 Flowchart for the proposed algorithm (Wang et al., 2013)

This paper tries to improve the local variation in the image and also doesn't harm the global trend of the intensity at the same time. So the paper focuses on the Reflectance extraction and relative order illumination compression. This paper introduces its own dataset which contains more than 150 images. A comparative study has been done on these images with 6 state-of-the-art methods together with the proposed algorithm and based on the visual comparison, quantitative

measurement results of discrete entropy and LOE numbers, it can be clearly seen that the proposed method gives the clear advantage over all those state of the art methods selected in the paper. According to the paper, their solution doesn't scale well with video files because it introduces slight flickering in the video files and authors plan to do further research in this area.

When cameras became smarter and we could leverage the hardware technology in building the training dataset by clicking the proper lighted images together with dark images either using some camera filters or using heat or infrared sensor technology. There are many dataset which offer the images or videos which come in pairs of low lighted and their proper lighted counter parts. Taking advantage of these datasets many papers proposed to have deep learning based low light enhancement methods where deep learning model were trained with both type of images and later use them for correcting the dark areas in the images. Some outstanding methods which are deep learning based are KinD (Zhang et al., 2019), Learning to See Moving Objects in the Dark (Jiang and Zheng, 2019).

KinD (Kindling the Darkness), is one of the major work which was done recently for image enhancement. It's a deep learning based method and is inspired by Retinex Theory. This paper mostly focuses on the poor visibility and different types of degradation like noise and colour distortion. This research is done on LOL Dataset which contains 500 low/normal light image pairs. 450 out of 500 images are used for training and 50 images are used for test.

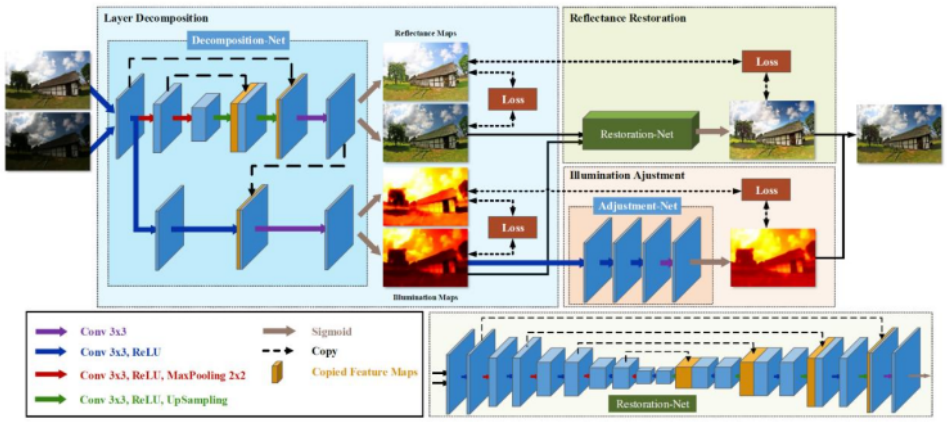


Figure 3 KinD Network Architecture (Zhang et al., 2019)

As shown in Figure3, KinD proposes layer decomposition and reflectance restoration neural networks, where restoration network adopts 5-layer UNet. It also proposes the Illumination Adjustment Net which can help in generating the ground-truth light level for images. In this research, batch size as 10 for layer decomposition net and batch size of 4 for reflectance and illumination adjustment net and stochastic gradient descent as optimization technique are used while building the models. A comparative study is done on the 10 different image enhancement methods with the method proposed in the paper. There are some visual comparison and some metrics (i.e. PSNR, SSIM) comparison between images (enhanced with different methods) are present in the paper and based on that, it concludes that the proposed method gives clear advantage over all the state-of-the-art methods selected in the paper and it outperforms all the competitors using LIME (Guo et al., 2017) and NPE (Wang et al., 2013) datasets as well. One thing to note here is, by looking at the visual comparison, it can be clearly seen that the images, enhanced with KinD**,** do not suffer from any noise or any sort of colour distortion.

Learning to See Moving Objects in the Dark (Jiang and Zheng, 2019), method proposes a new optical system which can capture both bright and dark videos of the same scene, to have both training and ground truth dataset. They discouraged the use of infrared sensors in the cameras because in forests it can disturb the animals and might trigger uncontrollable animal reactions. So they advocate enhancing the images which are captured by ordinary cameras with ND filter so the proposed camera can click both dark and bright images simultaneously. This paper introduces a new dataset which contains 179 pairs of videos consisting of 35800 extremely low-light images and their corresponding properly lighted images. This paper proposes a 3D U-net based network for low-light enhancement. A comparative study is done between state-of-the-art methods and the proposed method which concludes that the proposed method is the best among all the methods and also according to this paper, proposed method is able to tackle the flickering issues in the enhanced videos.

The only problem with deep learning methods is that they need the training data which has paired images or videos which we might not get very easily in every dataset. Presently we are analyzing the ARID dataset which videos don't have their properly lighted counter parts. So we will have to rule out the deep learning based methods for this research and work with non-deep learning based methods. Recently in 2021, (Jiang et al., 2021) proposed 'EnlightenGAN**:** Light Enhancement Without Paired Supervision' which is also deep learning method which can be

trained and evaluated without normal/low-light image pairs. More detail on Enlighten GAN is explained in the next sections.

2.5. Retinex Theory

Retinex Theory (Land, 1977) was introduced for full colour perception with colour consistency which was involved all level of visual processing. The term 'retinex' is made with retina and cortex.

In the experiment, three different source of narrow wave length illumination were used to shine on the colour display. Based on this experiment, if only one of those was shined on the display, no color variation could be seen, just various levels of light and dark. If somebody shines all 3 illuminators at once on the display, it is essentially shining white light on them because all 3 light primaries are shining on the display at once. Therefore, even though there are only 3 wavelengths present, and all other wavelengths are absent, these represent enough of the spectrum for colour to show up. Then, in the experiment, author tried to vary the quantity of wave length so that exactly the same amount of long, medium, and short wavelength light coming from any coloured area could be used. As shown in Figure 4, the light reflected for all the three wave lengths, together it is called spectral reflectance.

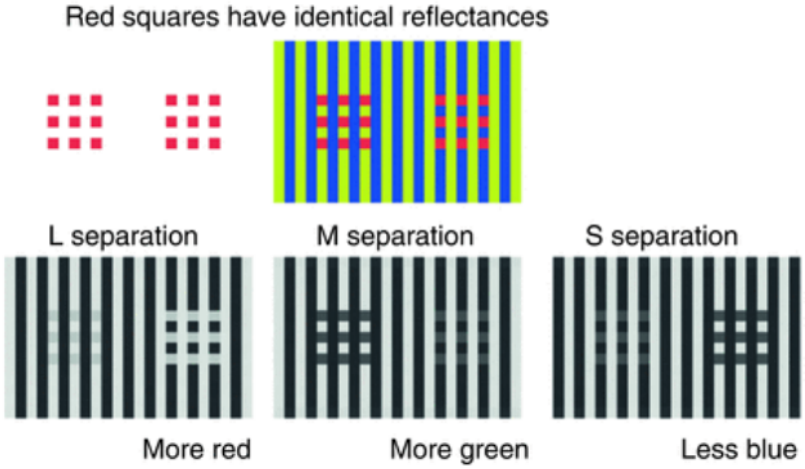


Figure 4 Retinex Theory (McCann, 2016)

Many people while experimenting with low light illumination on the dark images, considered reflectance for image classification and were quite successful in their experiments. Some algorithms like LIME, Naturalness Preserved Enhancement Algorithm are one of state-of-the-art algorithms which are based on Retinex Theory and KinD algorithm is also inspired by this theory.

2.6. Enlighten GAN

EnlightenGAN (Jiang et al., 2021) is deep learning based method which doesn't require any paired supervision unlike other deep learning based methods. This method, is another successful method which proposes a unsupervised dubbed EnlightenGAN (one path GAN), that can be trained and evaluated without normal/low-light image pairs unlike most of the deep learning based methods. Instead of ground truth data, information is extracted from input image and used for unpaired training.

In this paper, proposed method uses an attention U-Net neural network (as shown in Figure 5) as the generator and uses global and local discriminator to maintain the texture and details of the image. This paper advocates for self-regularized attention map instead of supervised learning. Also in the experiment, it uses data from multiple dataset where around 900 low light and 1K+ normal light images are collected from various datasets.

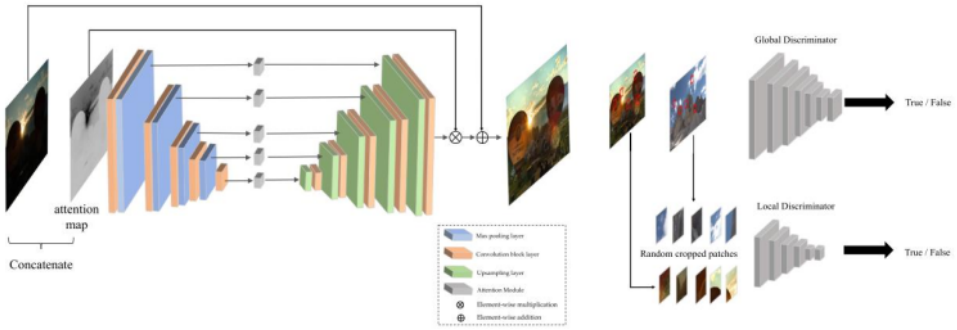


Figure 5 Enlighten GAN Architecture (Jiang et al., 2021)

EnlightenGAN is trained by 100 epochs and with learning rate of 1e-4 with Adam Optimizer and then another 100 epochs where learning rate was linearly decayed to zero. A comparative study is also done with 6 state-of-the-art methods together with EnlightenGAN and

based on the visual comparison and based on NIQE scores, it can be clearly seen that the proposed method gives the clear advantage over all the methods selected in the paper.

2.7. Human Action Recognition

Human activity recognition plays a significant role in human to human interaction and in interpersonal relations. Human ability to recognize other person's activity is a subject of study and mimicking it in any algorithm is quite hard. Human action Recognition from the video sequences is quite a challenging task because of multiple problems like movement in the background, changes in scale, lighting conditions and partial occlusion. Many applications like video surveillance and robotics for human behaviour characterization requires activity recognition algorithm which can help in classifying the human actions and take some action based on that. In addition to that, classifying behavioural actions is time consuming process and requires the knowledge of the event. Another problem arises because of the localization problem, which area we need to focus on, we need to focus on the dynamic and kinetic state of the person so that it's easier to recognize the human action.

A journal article 'A review of human activity recognition methods' (Vrigkas et al., 2015) explains very well about different types of human recognition methods for example: Rule based, Shape based, Behavioural, Stochastic, and Space Time methods. This article explains there are some activities which are day-to-day activities and it's easy to recognize but some actions like cutting vegetables and making tea difficult to recognize. These kind of complex activities we can further decompose in other simpler and regular activities.

Human action recognition methods evolved over time and every now and then people are able to invent a better algorithm. The evolution of human action recognition and different type of methods are explained in the next sections.

2.8. Challenges in action Recognitions

Human action recognitions can become challenging for various reasons like misjudgement of action because of intra-class and inner-class distance, complex environmental changes, or due to insufficient training data (Jiaxin et al., 2021). One example for intra-class and inner class distance is, for example, some people show their fist in order to say hello and while others may be trying

to hit someone. In this scenario, it is very important to consider the speed of the fist coming towards another person. If the speed of the fist is very fast that means it can be a fight between people or boxing game but if is slow, it can be a way to say hello to a person. Another example which is related to environment changes like rain or fog which is going to make the visual quality more worse and will put some noise in the video and may make the action recognition more harder for algorithms. Along with that, insufficient data also can cause issues during model building and prediction as deep learning methods are data hungry algorithms and it needs a lot of data to train the model.

2.9. Related Researches in Human Action Recognition

There are many researches happening around the action recognition. Some researchers suggest two-dimensional approach and some three-dimensional approach. (Vrigkas et al., 2015) talks about some researches propose tracking human actions with single view or multi view cameras and the paper gives a very detailed view on the evolution of activity recognition technique in the last two decades. The article also talks about different challenges we might face during action recognition which will be very helpful during the benchmark process in this research. In human action recognition, there are mostly two categories for analyzing the images or videos like top-down and bottom-up. 3D modelling is one of the best approaches which was proposed by (Chen et al., 2013) which promotes to create 3D representation of the human body which will help in recognizing the human action better than 2D representation. Another approach proposed by (Guo and Lai, 2014) which explains all the different methods for human activity recognition from the still images and then it puts then in two major categories based on the level of abstraction and types of the features.

There are different types of methods which can be used for human activity recognition, for example: Space time, stochastic, rule based and shape based which are called unimodal methods and multimodal methods which combine the features from different sources and use for activity recognition are: affective, behavioural and social networking methods.

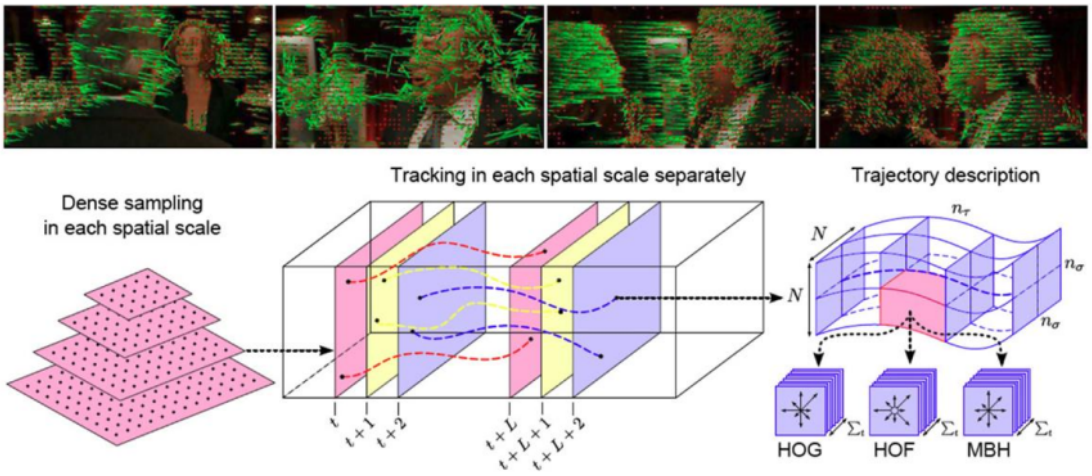


Figure 6 Human Space Time Representation (Vrigkas et al., 2015)

As shown in Figure 6, Space Time method is based on space-time features extracted from video sequence for recognizing the activities. There are lot of methods which are based on space time representation and most of them rely on the optical flow. Some methods are based on spatiotemporal features and facet model. (Gaidon et al., 2014) proposed an unsupervised method for learning human activities from short tracklets and used hierarchical clustering algorithm to represent videos with an unordered tree structure and compared all tree-clusters to identity the underlying activity.

Stochastic methods propose the entities to recognized as a stochastically predictable sequence of states. There are many methods which are based on stochastic methods. The most significant work is done by (Zhou and Zhang, 2014) which proposed a robust background clutter, camera motion and occlusions methods for recognizing the complex human activities. This paper proposed a new way to represent the human activities with a bag of Markov chains obtained from STIP and salient region feature selection.

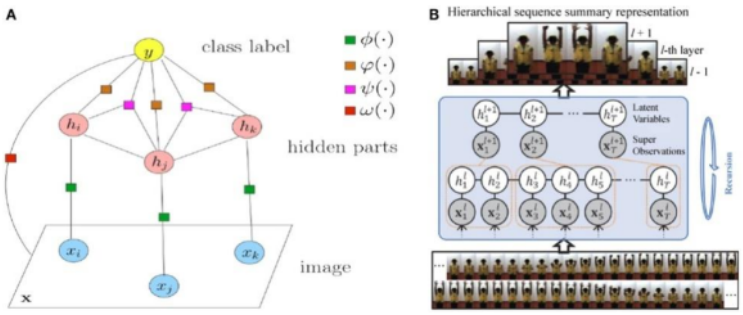


Figure 7 Representative Stochastic Methods (Vrigkas et al., 2015)

In another stochastic method, (Ni et al., 2014) decomposed the problem of complex activity recognition into two sequential sub-tasks with increasing granularity levels. First, the authors applied human-to-object interaction techniques to identify the area of interest, then used this context-based information to train a conditional random field (CRF) model and identify the underlying action. On the other hand, (Lan et al., 2012) proposed a hierarchical method for predicting future human actions, which may be considered as a reaction to a previous performed action. They introduced a new representation of human kinematic states, called "hierarchical movements," computed at different levels of coarse to fine-grained level granularity.

Rule-based methods (as shown in Figure 8) propose a way to modelling the human activity using some rules and set of attributes of the events where each activity is considered as a set of rules and attributes which enables the construction of a descriptive model for human activity recognition. For rule based method, (Kuehne et al., 2014) proposed a structured temporal approach for daily living human activity recognition. The author used HMMs to model human actions as action units and then used grammatical rules to form a sequence of complex actions by combining different action units. When temporal grammars are used for action classification, the main problem consists in treating long video sequences due to the complexity of the models. In order to solve the complexity of human activities in the video, (Donahue et al., 2017) suggests to decompose the video in smaller clips and generate of description from videos based on CNN model which can be used for activity recognition.

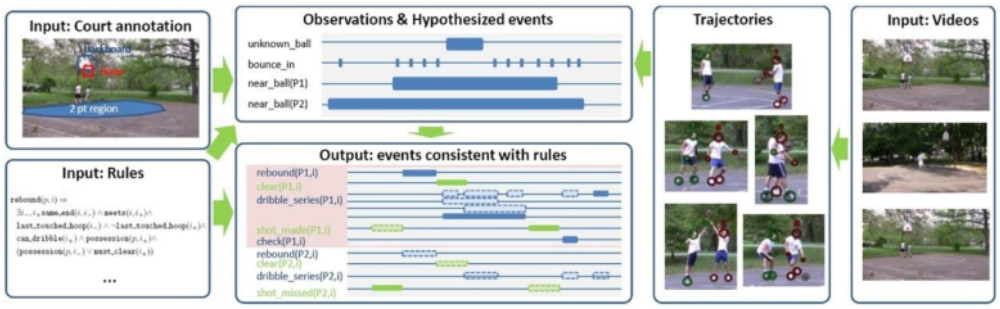


Figure 8 Rule based methods (Vrigkas et al., 2015)

Shape based methods mostly focuses on the body parts and their actions and we know that human activity based on human silhouette works very well. The human silhouette consists of limbs joints which are connecting to each other and so it is very important to extract the silhouette carefully from the video sequence to recognize the human activity efficiently. According to (Zhe Lin et al., 2010), identifying which body parts are most significant for recognizing complex human activities still remains a challenging task, paper proposes a out of the way method to overcome this situation.

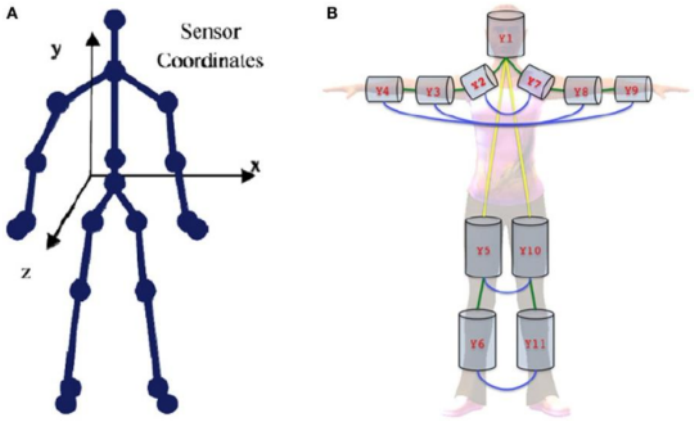


Figure 9 Human Body representation in 2D and 3D (Vrigkas et al., 2015)

Some methods are based on Action bank features. “What is an Action bank and how does it work?” is discussed in the next sections. Human action recognition based on recognition of linear patterns in action bank features using convolutional neural networks (Ijjina and Mohan, 2014), proposes a deep CNN architecture for recognizing human actions in videos using action

bank features. Action bank is nothing but a predefined set of videos converted to linear representation and saved as action bank features. This research proposes a convolutional network with linear mask which can capture the localized patterns for each action.

Generally, in case of action recognition in a video requires a 3D CNN to learn the spatiotemporal features from videos but this paper advocates using 2D CNN with action bank which is based on the concept of speech recognition using the spectrogram of audio data. They have designed a CNN which exploits this similarity in action bank features and this will drastically reduce the computational time for action recognition. UCF50 dataset containing 205 videos is used in this research, these input videos first get processed by the feature extraction module to extract the action bank, and then these action bank features are given as input to pattern recognition module for training. The CNN utilizes the similarity patterns to assign an action label to the videos. By looking at the results with this approach looks quite promising where the proposed method is able to achieve 93-94% accuracy in action recognition.

Human Action Invarianceness for Human Action Recognition (Sjarif and Shamsuddin, 2016), proposes to use human action shape or silhouette uniqueness to recognizing the human actions. Human action features can be extracted by using integration moment invariant. Action features are actually based on how silhouette moves in video frames. In human action invariance, the paper proposes three processes like extracting global features, similarity measurement between features and intra and interclass analysis.

Authors have used IXMAS dataset which contains 13 different actions performed by 10 people. The experiment is done with different video frames i.e. 30, 120 and 300. This research uses various techniques to achieve the higher accuracy like wavelet, PCA, Normalization, pre and post discretization. A comparative analysis on other methods is also done which other methods are not able to give the good accuracy with this dataset. But the method proposed by this paper is able to perform very well and is able to predict the human action with high accuracy up to 98-99%. This paper has given promising results and could be one of methods in my research in identifying the human action recognition in dark.

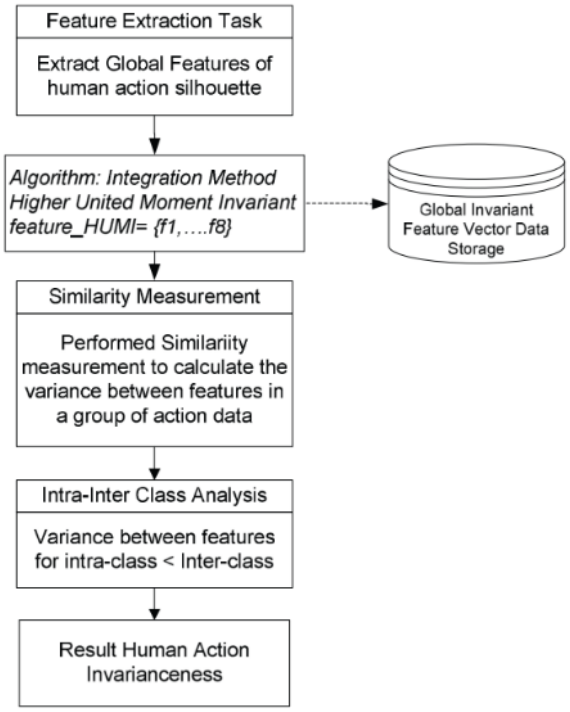


Figure 10 Human Action Invarianceness Procedure

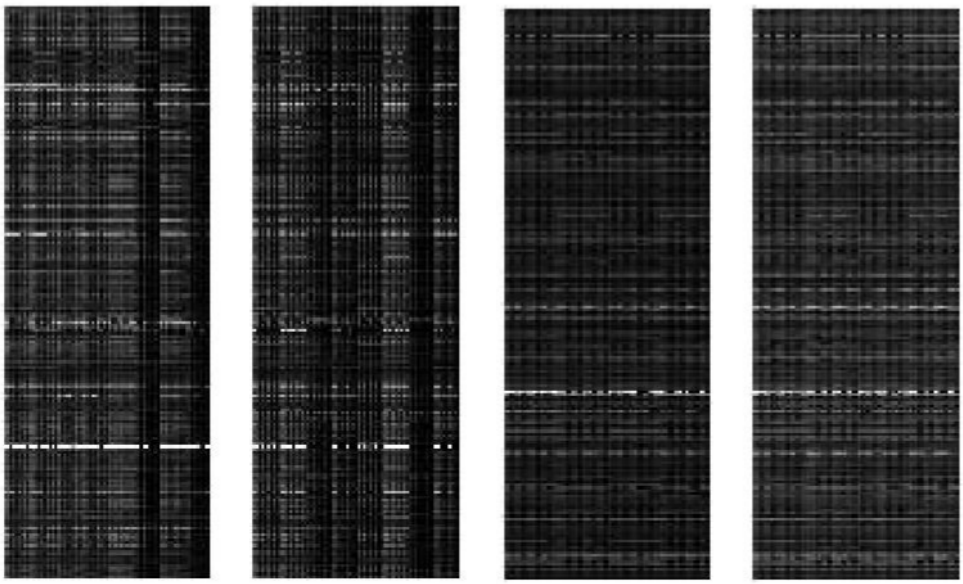
Learning Spatiotemporal Features with 3D Convolutional Networks (Tran et al., 2015), proposes deep 3D ConvNets which are trained on large dataset and suggest that these 3D ConvNets are better than 2D ConvNets for spatiotemporal feature extraction. This paper evaluates both the networks and concludes that 3D ConvNets gives much better accuracy (98%) compare to 2D ConvNets.

2.10. Action Bank for Video Sequences

Action Bank (Sadanand and Corso, 2012) is the two-dimensional representation of a video sequence. It is actually the collected output of many action detectors that each produces a correlation volume. A template based action detector is the primary element of the action bank. While building the action bank we need to take account into the scale and viewpoint and temp. To take care of scale we need multiple detector which can scan the video from left to right, right to left, top to bottom or bottom to top. Action bank is nothing but the feature vector of a video

sequence. Because of the 2D representation of video, we can use simply SVM classifier on the action bank feature vectors. Action bank gives a very good flexibility in choosing the detectors and multiple detectors can be used concurrently. In this paper, UCF50 data is used which contains around 50 classes and 6680 videos and generated the action bank for all the video sequences which were shot from different angles. Around 3-6 action templates were used to extract the features from each action.

Figure 11 shows the action bank features two-dimensional representation extracted from multiple action templates, where each horizontal line corresponds to an action bank feature generated by calculating the similarity of the video. These different pattern can uniquely identify an action and it can be clearly seen in the picture that Boxing#1 and Boxing#2 action have quite similar localized patterns and that is the same case for Running#1 and Running#2 action bank features. So using two-dimensional convolutional network might help in extracting the feature from it and classify according to these patterns and we no longer need to extract the feature in three-dimensional spatiotemporal space.



(a) Boxing #1 (b) Boxing #2 (d) Running #1 (e) Running #2

Figure 11 Action bank features 2D representation (Ijjina and Mohan, 2014)

Action bank helps in representing the video results in a fixed size irrespective of their length, illumination, or frame speed. Possibility of using 2D CNN can be explored for human action recognition by considering the action bank features as an input because model can be trained to classify the videos in one pass and we don't need to consider the temporal window on the video frames. In action bank we can clearly see the horizontal patterns which are associated with each action and can be utilized for classification using convolutional neural network.

2.11. 2D vs. 3D CNN for human action recognition

Convolutional neural network have received phenomenal success in computer vision tasks. CNN is a neural network which is based on convolutional operation. CNNs are used in research settings for the development of computational modeling, quantitative studies, and population-based analysis. Action recognition which is based on deep learning methods means to automatically analyze and recognize the motions of a single or multiple in a video sequence (Jiaxin et al., 2021). As shown in Figure12, deep learning CNN models can be used as the feature extraction and which is later can combined with classifier to recognize the actions.

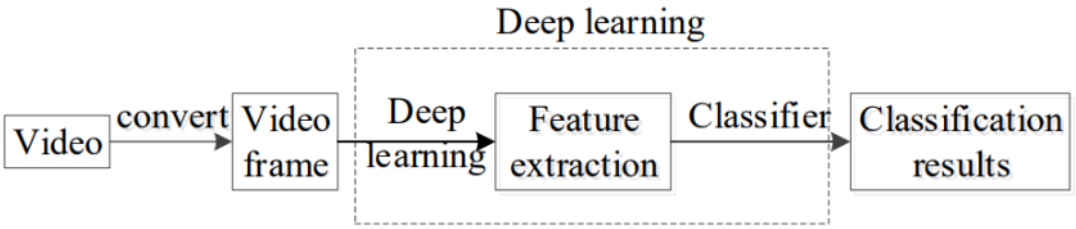


Figure 12 Action Recognition based on deep learning methods (Jiaxin et al., 2021)

There can be different type of convolutional methods which include 1D, 2D and 3D convolution. 2D and 3D convolution can be used for feature extraction in action recognition, however both methods are very different from each other in terms of extracting the features. The major difference between 2D CNNs and 3D CNNs is that 2D convolution extract the feature from two-dimensional feature map and time domain is compressed but the 3D convolution extract the feature in a three-dimensional feature map and includes the time domain information so it will capture each frame video frame individually and will also contain the information about the

progression as well which will ultimately help in action recognition of video sequences. Other difference is the number of parameters during the model training, 3D CNNs will have many more number of parameters compare to 2D CNNs. The more number of parameters leads to more computation time and resources so 2D CNNs are relatively faster to calculate but it will not contain the time domain information preserved in it.

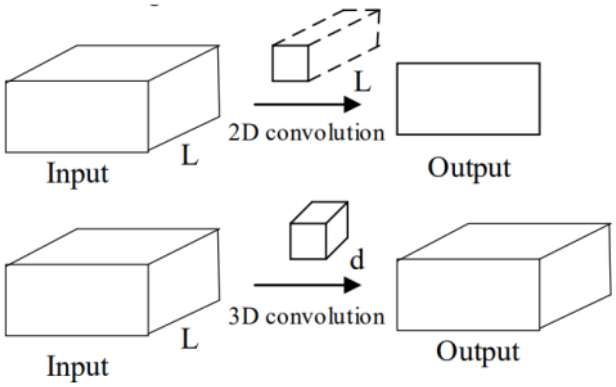


Figure 13 Feature extraction 2D vs. 3D CNNs

There are some methods propose the fusion of multiple methods which can perform the feature extraction together, which often leads to better recognition results.

However if there is a way to represent a video in two-dimensional manner like Action bank as explained in above section that will make the action recognition in much faster, however we need to evaluate which one performs better an Action bank approach or 3D CNN approach.

2.12. Discussion

From the analysis on different algorithm, we can state three major conclusions. First, action recognition cannot be done on the low light video sequences and synthetic dark videos cannot be used for model training in order to do action recognition in actual dark videos which were shot in the dark. Second, though there are many algorithms which are capable to make a video visually appealing but some frame enhancement methods actually reduces the action recognition accuracy. Some methods introduce the flickering in the videos after enhancing which might deteriorate the performance of the model, so we will have to careful in choosing the correct low light illumination method. Third, in dark video, many action recognition techniques fail to

identify the action because outlining of the action is not clear. For future work, we can focus on many more algorithms which can illuminate the videos well without introducing the noise in the videos and help action recognition algorithms to extract features from the videos.

2.13. Summary

In the above sections, we discussed about different algorithms for low light illumination and human action recognition. So based on these discussions we can summarize that there are many low light illumination techniques are there but we have to choose the best method which works well with the human action recognition because focus here is to recognize the human action not the visual quality of the videos. Action Bank for human action recognition and Enlighten GAN for low light illumination look quite promising, these will be the main focus of this research.

Chapter 3. Research Methodology

3.1. Introduction

Today, as our population is increasing, it is getting difficult to do the night surveillance manually by humans, we need some kind of solutions to solve the problems we face during night time; some scenarios are like night surveillance, illegal infiltrations at country borders, self driving cars during night time, analyze animal breeding and behaviours in the wild during night and human to robot interaction are different areas where computer vision will be needed. Humans need to augment their capabilities during the night. A lot of work in done on Human action recognition with the properly lighted videos and images but not enough work is done on action recognition in dark videos. However some of the state-of-the-art methods have done awesome job in enhancing the low lighted images and we are in the right direction.

The motive of this research is to analyze human actions in videos which are shot in the dark and with normal cameras. This research will definitely help fellow researchers who are doing the research in the similar areas where any sort of action recognition requirements are there. This research will also help camera manufacturers who are making cameras with built in human action recognition techniques. The model built with this research can be deployed at edge to any device and that can do the job of action in the dark while shooting the videos.

Since video shot at the dark suffer from poor visibility and all sort of noise and quality degradation, human action recognition directly on the dark videos will not be efficient because it will be difficult to extract features out of the videos. So this paper suggests the research in mostly two parts. Firstly, find the best method to pre-process and do the low light enhancement in the dark videos and remove the noise from the videos. Secondly, do the human action recognition on enhanced videos. There are some of the methods which are available for low light enhancement for videos and images. A comparative study needs to done on various methods and propose the best method which gives the highest accuracy and lowest computation time. Some methods enhance the images visually very well but a question to ask is “do we really need these kinds of refinements for action recognition in the dark?”

3.2. Research Methodology

This chapter addresses the different steps taken to perform the creation of an Action Recognition in dark system with focus on technical documentation. The following questions are answered over the course of this chapter:

- How can new knowledge be discovered?
- What goals and limitations influence the chosen method?
- What method provides the possibility to answer the research questions and how is it composed?
- What data attributes are required to apply the method?
- How can discovered knowledge be further applied?

In order to answer these questions, the structure of this chapter reflects an adaption of the earlier presented knowledge discovery process. The entire process is presented in Figure 14: The first step is to do the standardization on all the videos. In standardization process, video frames standardization will be done where speed of frames will be made same in all the videos. Video length and number of frames in the videos will also be standardized in order to make the process of feature extraction process simpler and effective. The second Step is to do Low light enhancement process where each frame of the video will be enhanced so that objects in the video clearly visible and extraction algorithm can extract the features out of it. There are many algorithms such as KinD (Zhang et al., 2019), LIME (Guo et al., 2017), EnlightenGAN (Jiang et al., 2021) methods will be used in low light enhancement.

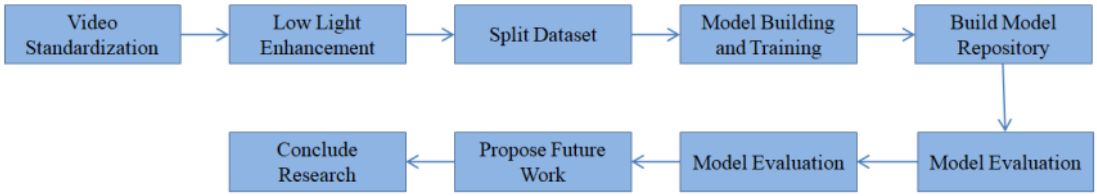


Figure 14 Research Steps

The third steps is to split the data in training and test sets where 70% of the videos will be kept for training the model and rest will be kept for testing and evaluation process. 70% videos are over 3700 videos files which are enough for the training process. It is important to split the dataset based on the different classes such that training and test dataset both should contain some video files from each class. The forth step is model building and training, this step is explained in details in the upcoming section where we will see model training with detailed instructions, outcomes and details on the hyper parameters used. The fifth step is to build the model repository where we will store all the model trained with different methods and algorithms and later all the models will be compared and benchmarked based on different matrices like accuracy, Precision, Recall and F1-Score. The training process involves lots of iteration based on different hyper parameters and number of epochs required for best accuracy.

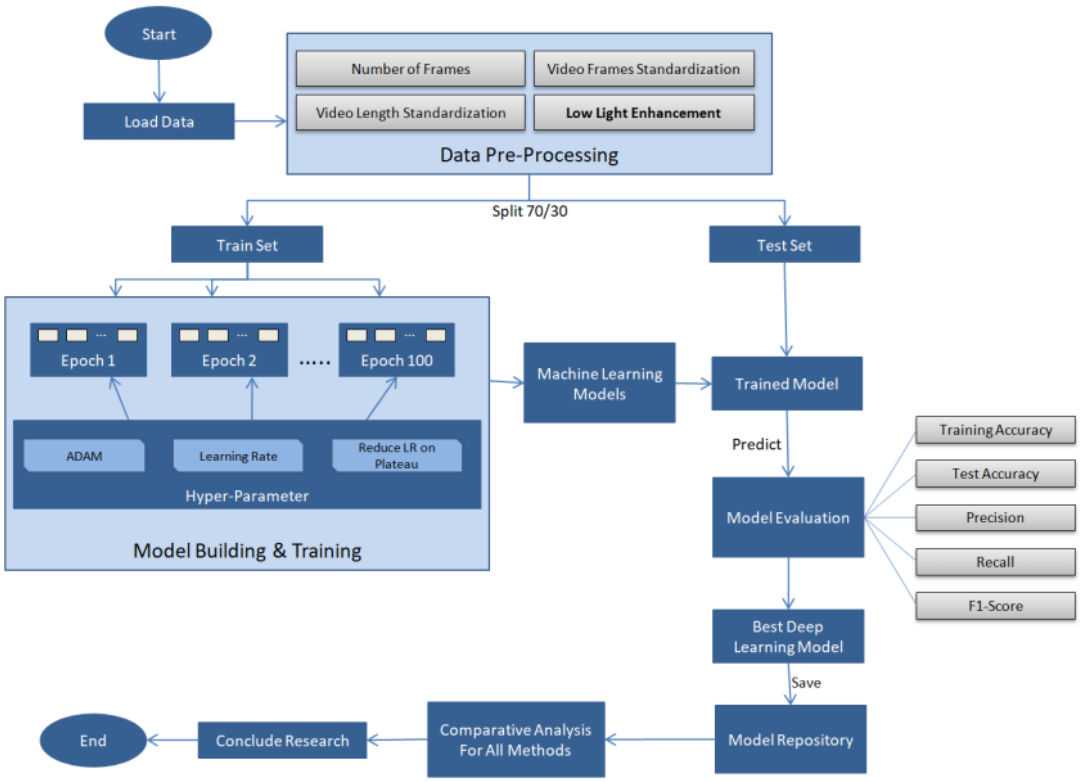


Figure 15 Research Methodology Detailed

The sixth step is to evaluate the trained model against the test set and check the accuracy, precision and recall values for that model. The seventh step, with the comparative analysis of different methods, decides which combination of low light enhancement method with action recognition method suits the best for ARID dataset. The eighth step is to propose the future improvements that can be done in this area or to give direction to explore certain methods which could not be used in this research due to time constraints and can be the good candidate for human action recognition in dark.

3.3. Data Selection

This section deals with action recognition in dark data selection process. Thereby, following questions will be answered:

- What is motivation behind choosing a particular dataset?
- What all other datasets available and why they can't be used for this research?
- What are the features of the chosen dataset and how it help this research?

There are lot datasets which can be used for this study, where we need the human action videos in dark, but many datasets contain video which were shot in the low light situation or contain synthetic dark videos. This research is planning to focus on the videos which are truly dark videos which were shot in dark without the presence of any kind of light source. There are dataset for Action Recognition like Kinetics (Carreira and Zisserman, 2017), HMDB51 (Kuehne et al., 2011) and UCF101 (Soomro et al., 2012). Out of these datasets, Kinetics dataset is much popular and contains around 400 action classes but these datasets were collected from web and shot with normal lighting conditions.

This research uses ARID (Action Recognition in Dark) dataset; this dataset contains more than 3780 video clips with 11 different kinds of human actions where each class contains around 110 clips For example: Drinking, Jumping, picking, pouring, running, sitting, standing, turning, walking and waiving. Each video clip has fixed frame rate of 30 frames/sec and 1.2 seconds long. All the videos are in .avi format and are compressed using DivX codec.

(Xu et al., 2021) have created this dataset specifically for analyzing dark videos. These dark videos are shot from the normal cameras without using any kind of night vision sensor. It's important to note that to create video in different illumination condition; different γ values were

used to synthesize different dark videos and γ values are taken from a normal distribution $N(\mu, \sigma^2)$ with the constraint of $\gamma \geq 0.1$. Some of videos are so dark that it is very difficult for even human eye to identify the actions in the videos. According to the author of ARID dataset, this is first ever dataset which is going to focus on human actions in dark.

The purpose of this dataset is to provide the dark videos which were really shot in the dark because we see most of the researches are done on synthetic dark videos or the dark videos which were created using some camera filters. Models built on synthetic dataset don't perform very well with the actual dark videos.

3.4. Data Pre Processing

This section deals with the pre-processing of datasets. Thereby following questions will be answered:

- Which all data standardization techniques are chosen before any other type pre-processing happens on the dataset?
- Which all illumination techniques are good for this scenario and why?
- What are the main challenges in this phase?

Before starting to build the model for action recognition, we need to pre-process all the videos so that videos are ready for the feature extraction. Pre-processing involves multiple stages like video standardization, frames per second, duration of videos and illumination of each video frame. During the video stream standardization process, we will have to make sure each video frame contains the same number of pixels and duration of video is exactly same, this standardization will eventually help in extraction process where algorithm will not have to face any kind of inconsistency.

For video standardization, this pre-processing stage loops through all the videos and apply the trimming on videos to keep the length to 2 seconds, frame size to 340x240 and frame speed to 30 frames/second and save it to a temporary directory. This temporary directory is the input to the next stage to apply more pre-processing steps such as low light enhancement to all the frames of the videos.

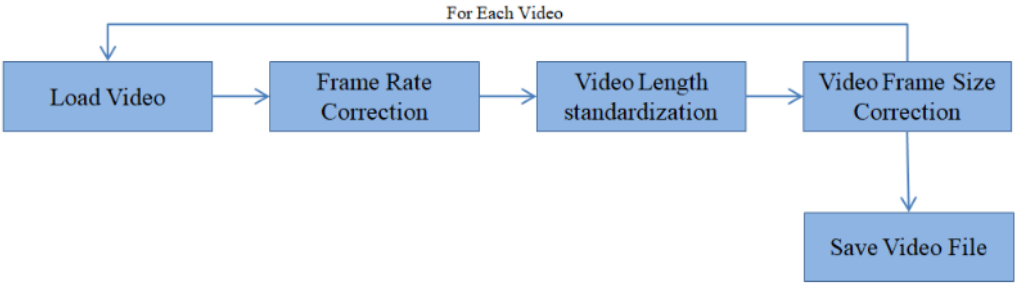


Figure 16 Video Files Pre-Processing

For the low light enhancement, this pre-processing stage loops through all the videos and then applies the low light enhancement techniques on the each of the video frames and saves again to a temporary directory. However, first stage of video standardization and second stage of low light enhancement techniques can be done together and save it once to the temporary directory. This will create multiple set of videos for each low light enhancement technique. For illumination of videos, this research uses some of the methods which make sense to use with ARID dataset. Some of methods which will work well with this dataset are: KinD (Zhang et al., 2019), LIME (Guo et al., 2017), EnlightenGAN (Jiang et al., 2021), Histogram Equalization (HE), Gamma Intensity Correction (GIC). EnlightenGAN seems to be the most impressive and promising methods as of now among all the low-light image enhancement methods.

Gamma Intensity Correction is chosen because ARID dataset has also done some video synthesis based on the gamma intensity and using gamma intensity will take less computational time and action recognition might perform well while extracting the feature because here main focus is action recognition in dark and not the visual quality of videos. LIME, KinD and Enlighten GAN methods are chosen because they directly enhance the low lighted images without the need of paired bright light images for training the model. These all the enhancement methods were invented for images, so these methods need to be applied on each of the video frames. During enhancement of the videos, a comparative study is planned where their computational time will be in great focus. Enlighten GAN described in great details in section 2.6 and is one of best methods which can illuminate the video files using deep learning methods.

The main challenge of this phase is that this phase is going to be compute intensive and time consuming. As there are multiple enhancement techniques so more storage space will

required for each type of enhancement technique that might be big issue at Kaggle environment where there is a limited storage and compute capacity.

3.5. Data Transformation

This section deals with transformation required on the dataset. Data transformation is mostly relevant to tabular dataset but in case of video data, data transformation can be done by creating the action bank for different videos, action bank is the 2D representation of videos. Action bank is explained in section 2.11 in details. For the same type of videos action bank features will be quite similar and once the action bank features are created neural network can be trained on the action bank dataset.

In this section, data transformation method will loop through all the videos and action bank will be created for each video sequence and will be saved to a file. Action bank representation will be very small is size so there is not much concern about the storage space.

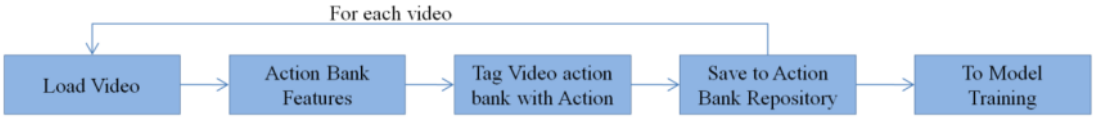


Figure 17 Data Transformation (Action Bank)

3.6. Hyper Parameters Selection

This section deals with hyper parameter selection which will be used during the model training phase. Thereby, the following questions will be answered:

- Which all parameters essentially required for model training process?
- What are the best practices while choosing the hyper parameters?
- Which all hyper parameters will be in action for this research?
- What is specific reason of using particular optimization technique?

The process for setting the hyper parameters required extensive expertise and trial and error. Based on how model is performing, we can change the parameters to neural network and see the results. Choosing the best learning rate is a critical task because choosing the wrong learning rate

might cause issues in model training and it might never find the global minima. If learning rate is too low it will take a lot of epochs to reach the global minima and if it is too high it might oscillate between global minima so learning rate has to be just right. This process of find the global minima are called gradient descent. Gradient descent is an optimization technique commonly used in training machine learning algorithms. The main aim of training ML algorithms is to adjust the weights w to minimize the loss or cost. By minimizing the cost function we can find the optimal parameters that yield the best model performance.

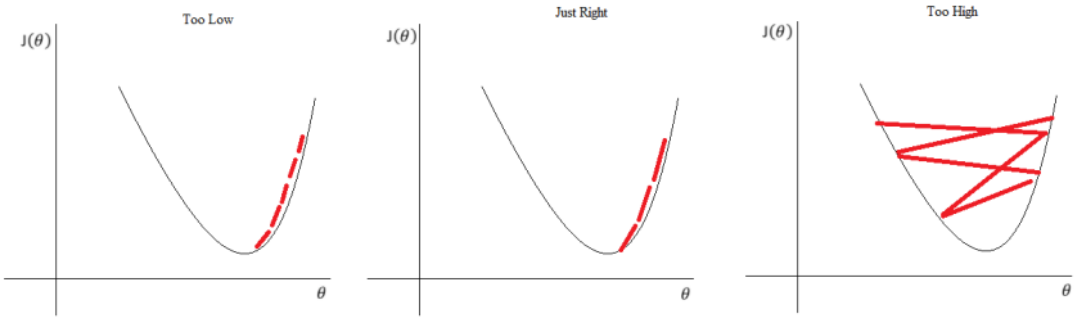


Figure 18 Learning Rate

Here in this research Adam is chosen as the optimization technique. Adam is the replacement optimization algorithm for SGD (Stochastic Gradient Descent) for training the deep learning models. Adam combines the best practices of AdaGrad and RMSProp algorithm to provide an optimization algorithm that can handle sparse gradients on noisy problems. Adam is relatively simple to configure because the default parameters generally do well in most of the situations. The default settings for Adam is alpha=0.001, beta=0.9, beta2=0.999 and epsilon=10E-8.

As shown in Table 1 and Figure 1, model will be trained with 100 Epochs with batch size of 30. Batch size 30 is decided based on the resource availability. For choosing the number of epochs will completely depend on model training accuracy, if required; this can be changed based on the results and accuracy obtained during the training process. These values are chosen based on the resource availability during this research. Other than that, together with ADAM optimization technique for finding the global minima, reduce the learning rate at the plateau with the factor of 0.1 and patience level of 5.

Parameter	Value
No. of Epochs	100
Batch Size	30
Optimization Technique	Adam
Learning Rate	Default for Adam
ReduceLROnPlateau	Patience=5, factor=0.1

Table 1 Hyper Parameters

For multiclass classification, it is always suggested to use categorical cross entropy for computing the loss value of the neural network. Categorical cross entropy requires data to be one-hot encoded and converted into categorical format. In Keras, this can be done with to_categorical which essentially applies one-hot encoding to the training and tests sets and now categorical cross entropy.

$$CCE(p,t) = -\sum_{c=1}^{C} t_{o,c} \log(p_{o,c})$$

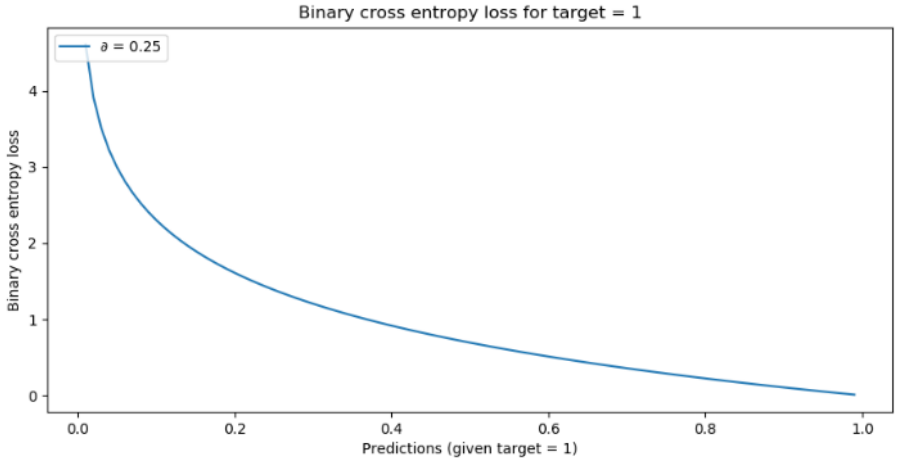


Figure 19 Categorical Cross Entropy

There is another method suggested by (Versloot, 2021) 'sparse categorical cross entropy' where one-hot encoding is not required. One-hot encoding might take a lot of time in case of huge dataset, using sparse categorical cross entropy we can avoid the one-hot encoding on the dataset.

3.7. Choice of Evaluation Metrics

This section deals with the selection, creation and pre-processing of the attributes in the data set. Thereby, the following questions will be answered:

- Which attributes were chosen for the given task and what is their main information contribution?
- How these evaluation metrics can deal with multiple methods for illumination and action recognition?
- What were the main problems encountered during this phase?

During model training, it will be important to measure the performance of the model. Without a proper approach to determine the quality of methods chosen, there is no point for further research. The most common approach is to check the categorical accuracy of the model but as we know accuracy is not everything, we need to look for other metrics like precision, recall and F1-score which are the true measurement of the model.

Table 2 metrics will be considered in this research which will tell us how the model is performing during model training and evaluation. Categorical accuracy calculates the percentage of predicted values (yPred) that matches with actual values (yTrue) for one hot labels and describes the model performance with the training data and validation categorical accuracy of the model tells the performance of model on the validation data. These accuracies should be in similar range in order to know that model is performing well in unseen data as well. Precision measures the percentage of true positives that were correctly classified and Recall measures the percentage of actual true positives that were correctly classified. All these metrics need to be be evaluated in order to find the best model.

Metrics	Description
categorical_accuracy	Accuracy of the model on the training data
val_categorical_accuracy	Accuracy of the model on the validation data
Precision	Proportion of positive identifications was actually correct
Recall	Proportion of actual positives was identified correctly
F1-Score	This a harmonic mean of Precision and Recall

Table 2 Model Evaluation Metrics

It is important to watch out for categorical accuracy and validation categorical accuracy especially so that we have balanced performance on train and test set. Both the accuracies should be in the same range and more than 80% for a good model. If categorical accuracy is too high and validation accuracy is too low then it will considered as over fitting which we have to avoid.
Only accuracy may not determine the performance of the model but other metrics like Precision and Recall are required to be analyzed which will help in determining the performance of the model.

Methodology described in Figure 1, will be repeated for every low light enhancement technique and for every action recognition algorithm. All the models will be saved to Model Repository. All the models will be evaluated and compare against various metrics and propose the best low light and human action recognition algorithm for ARID dataset.

3.8. Interactive Visual Analytics

This section deals with the exploratory data analysis on the dataset. Thereby, the following questions will be answered:

- What is ratio between videos present for different action category?
- Is there any data imbalance in different action categories?
- Is there any data augmentation required for the dataset?

This research plans to do the EDA on the dataset and check the imbalance present in the dataset. Based on the ratio for different categories, imbalance correction will be suggested. It is very important to do the imbalance correction in the dataset because neural network will not be able to

learn the features properly for the minority action category and neural network overall accuracy will suffer.

3.9. Data Augmentation

This section deals with data augmentation for videos, it is important to have enough data for neural network. Neural networks are data hungry and require a lot amount of data for training so data augmentation utmost important in this scenario. There are many technique which can be applied on the dataset.

There are many simple methods present like horizontal flipping of the each frame in the video, colour space augmentation and random cropping. A journal (Shorten and Khoshgoftaar, 2019) proposes many sophisticated techniques for data augmentation like geometric transformations, kernel filters, feature space augmentation, random erasing, GAN-based augmentation, adversarial training, neural style transfer and meta-learning techniques.

Figure 20 Data Augmentation (Shorten and Khoshgoftaar, 2019)

This research doesn't plan to implement all the data augmentation techniques but only few which can help generate some synthetic data which will be relevant for the video sequences. For this research due to the time constraint, some the data augmentation technique will be used like horizontal flipping, rotation by 10^{o}-20^{o} on the left of right axis and colour space transformation. These techniques need to be applied on the each frame of the videos.

Another advanced strategy for data augmentation is generative adversarial modelling and another variation is using GANs with auto-encoders which dramatically improve the quality of samples produced. Due to time constraints, this research will not focus on these advanced techniques based on GANs, but this can be excellent way to produce more data for the future work for this research.

3.10. Neural Network

Artificial neural networks are forecasting methods that are based on simple mathematical models of the brain. They allow complex non-linear relationships between the response variable and its predictors. A neural network can be thought of as a network of neurons which are organised in layers. The predictors (or inputs) form the bottom layer, and the forecasts form the top layer. There may also be intermediate layers containing hidden neurons.

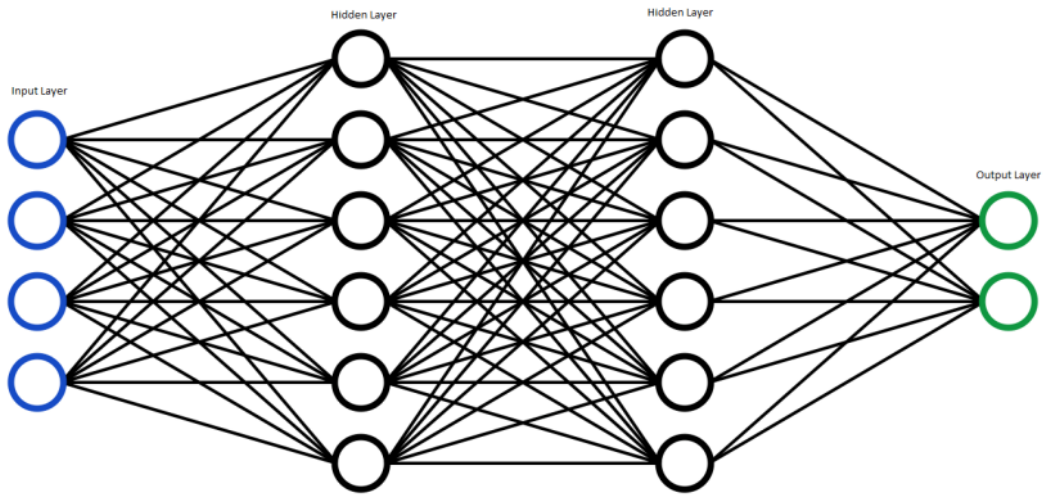


Figure 21 Neural Network

The simplest networks contain no hidden layers and are equivalent to linear regressions. Figure 21 shows the neural network version of a linear regression with four predictors. The coefficients attached to these predictors are called weights. The forecasts are obtained by a linear combination of the inputs. The weights are selected in the neural network framework using a learning algorithm that minimises a cost function such as the MSE. Of course, in this simple example, we can use linear regression which is a much more efficient method of training the model. Below formula shows the input to the hidden layer:

$$z_j = b_j + \sum_{i=1}^{n} w_{ij}\, x_i$$

Some advanced Neural Networks which this research plans to use are CNN and RNN. A convolutional neural network is a deep learning algorithm which can take in an image or video as an input and assign weights and biases to the neurons and based on these it will be able to differentiate one from the other. A CNN can be 2-dimentional or 3-dimentional, for normal black and white image feature extraction can be done using 2D CNN and while coloured images can be processed using 3D CNNs.

Another advanced neural network is RNN (Recurrent Neural Network) which is used to extract features from sequential data or time series data. A video is nothing but a sequence of images where RNN will be useful to extract features from it. RNN runs into two problems, known as exploding and vanishing gradients. In order to resolve these there are more advanced version of RNNs are LSTMs and GRUs. LSTMs and GRUs help to remember the context and propagates this context over time steps and solve the issue of exploding and vanishing gradients.

3.11. Regularization Techniques

Training a neural network which generalizes well to the unseen data is a challenging problem. A model with too little capacity cannot learn the problem and model with too much capacity can learn it too well that its start over fitting. Both these cases result in a model which doesn't generalize well. In order to make neural network work well, it needs to be regularized to avoid overfitting and underfitting. Regularization is a technique which makes slight modifications to

the learning algorithm such that the model generalizes better. This in turn improves the model's performance on the unseen data as well.

There are generally two approaches to reduce the overfitting, first is by training the network on more data and second is by changing the complexity of the network. A feature of very deep neural network is that their performance improves as they are fed larger datasets. Also if deep neural network has enough capacity to train on the data, there is a likelihood of overfitting. So we might have to trim down the capacity to regularize the model so that it doesn't overfit. There are different techniques for regularization, L1 and L2 regularization, dropouts, data augmentation and early stopping.

L1 and L2 regularization are the most common type of regularization techniques. These modify the cost function by adding another term know as the regularization term. Due to the addition of this regularization term, the values of weight matrices decrease because it assumes that a neural network with smaller weight matrices leads to simpler models. Therefore, it will also reduce overfitting to quite an extent. In L1 regularization technique, absolute value of the weights is penalized. Unlike L2, the weights may be reduced to zero here. Hence, it is very useful when we are trying to compress our model. Otherwise, usually prefer L2 over L1 regularization.

L2 Regularization:

$$\text{Cost function} = \text{Loss} + \frac{\lambda}{2m} * \sum ||w||^2$$

L1 Regularization:

$$\text{Cost function} = \text{Loss} + \frac{\lambda}{2m} * \sum ||w||$$

Another very popular regularization technique is dropouts. It also produces very good results and is consequently the most frequently used regularization technique in the field of deep learning. In this technique, some nodes are randomly selected and removed them along with along with all their incoming and outgoing connection.

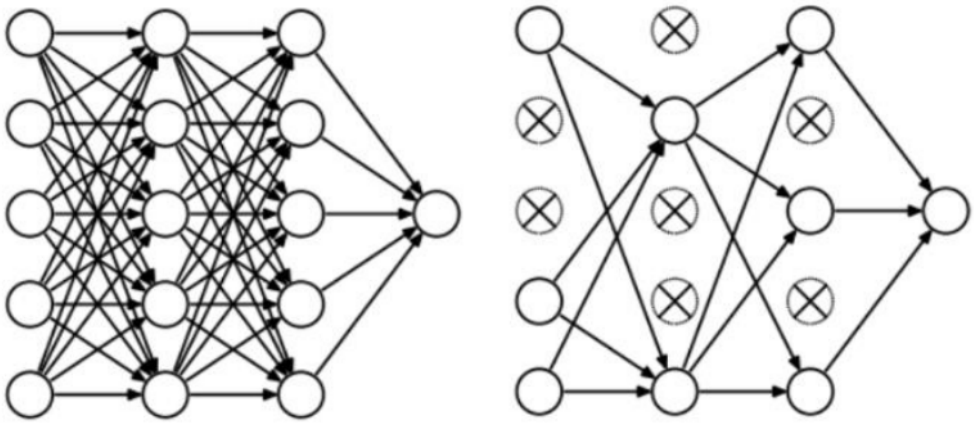


Figure 22 Neural Network with dropouts

So in each epoch, there will be different set of nodes and these results in a different set of outputs. This is just like an ensemble of multiple models which perform better than one single model. Due to these reasons, dropouts are usually preferred in deep learning regularization techniques. This research also plans to use the dropouts for its regularization technique. Other technique like data augmentation is already described in earlier section where synthetic data will be produced from the actual data and fed to that network for model training.

3.12. Required Resources

The research will need below hardware and software resources throughout the implementation

3.14.1 Software Requirements

The Predictive modelling will be implemented using python.

- Package Manager: Anaconda Navigator 1.9.12
- Presentation Layer: Jupyter lap 0.35.4
- Kaggle Environment to have access to GPUs and TPUs
- Language: Python 3.6.X
- Python Libraries:
    - Pandas and NumPy for data processing
    - Matplotlib and Seaborn for data visualization

- Keras framework and Tensor Flow as its backend for CNN and RNN model building and for model evaluation.

3.14.2 Hardware Requirement

A laptop with below configuration will be used:

- Operating System. Windows 10: 64-bit
- Processor: Intel(R) Core(TM) i7-7500U CPU @ 2.70GHz 2.90 GHz
- Memory: 16 GB

At Kaggle Environment below configuration will be used:

- Processor & Accelerator: NVIDIA TESLA P100 GPUs or TPU v3-8
- Duration: At least 20-30 Hours / Week
- Memory: 8GB per kernel

References

Carreira, J. and Zisserman, A., (2017) Quo Vadis, action recognition? A new model and the kinetics dataset. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017*, 2017-Janua, pp.4724–4733.

Chen, L., Wei, H. and Ferryman, J., (2013) A survey of human motion analysis using depth imagery. *Pattern Recognition Letters*, 34, pp.1995–2006.

Donahue, J., Hendricks, L.A., Rohrbach, M., Venugopalan, S., Guadarrama, S., Saenko, K. and Darrell, T., (2017) Long-Term Recurrent Convolutional Networks for Visual Recognition and Description. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 394, pp.677–691.

Gaidon, A., Harchaoui, Z., Schmid, C., Gaidon, A., Harchaoui, Z., Schmid, C., Gaidon, A., Harchaoui, Z. and Schmid, C., (2014) Activity representation with motion hierarchies To cite this version : 1073, pp.219–238.

Guglielmi, G., (2018) *Mammals turn to night life to avoid people*. [online] Nature (Nature). Available at: https://www.nature.com/articles/d41586-018-05430-4 [Accessed 3 Aug. 2021].

Guo, G. and Lai, A., (2014) A survey on still image based human action recognition. *Pattern Recognition*, [online] 4710, pp.3343–3361. Available at: https://www.sciencedirect.com/science/article/pii/S0031320314001642.

Guo, X., Li, Y. and Ling, H., (2017) LIME : Low-Light Image Enhancement via. 262, pp.982–993.

Hara, K., Kataoka, H. and Satoh, Y., (2018) Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.6546–6555.

Hyman, I., (2010) *7 Strategies to Shoot Video in Low Light*. [online] izzyvideo. Available at: https://www.izzyvideo.com/low-light-video/ [Accessed 7 Sep. 2021].

Ijjina, E.P. and Mohan, C.K., (2014) Human action recognition based on recognition of linear patterns in action bank features using convolutional neural networks. *Proceedings - 2014 13th International Conference on Machine Learning and Applications, ICMLA 2014*, pp.178–182.

Jiang, H. and Zheng, Y., (2019) Learning to see moving objects in the dark. *Proceedings of the IEEE International Conference on Computer Vision*, 2019-OctobIccv, pp.7323–7332.

Jiang, Y., Gong, X., Liu, D., Cheng, Y., Fang, C., Shen, X., Yang, J., Zhou, P. and Wang, Z., (2021) EnlightenGAN : Deep Light Enhancement. 30, pp.2340–2349.

Jiaxin, Y., Fang, W. and Jieru, Y., (2021) A review of action recognition based on Convolutional Neural Network. *Journal of Physics: Conference Series*, [online] 18271, p.12138. Available at: http://dx.doi.org/10.1088/1742-6596/1827/1/012138.

Kuehne, H., Arslan, A. and Serre, T., (2014) The language of actions: Recovering the syntax and semantics of goal-directed human activities. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.780–787.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T., (2011) HMDB: A large video database for human motion recognition. *Proceedings of the IEEE International Conference on Computer Vision*, pp.2556–2563.

Lan, T., Sigal, L. and Mori, G., (2012) Social roles in hierarchical models for human activity recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1354–1361.

Land, E.H., (1977) The Retinex Theory of Color Vision The Retinex Theory of Color Vision of radiant energy but correlated with. *Sci. Am.*, [online] 2376, pp.108–128. Available at: https://www.semanticscholar.org/paper/The-Retinex-Theory-of-Color-Vision-SCIENTIFIC-Land/2f3f8f151a52afa3c1e80505ddb09b8624162e35.

McCann, J., (2016) Retinex Theory BT - Encyclopedia of Color Science and Technology. In: M.R. Luo, ed. [online] New York, NY: Springer New York, pp.1118–1125. Available at: https://doi.org/10.1007/978-1-4419-8071-7_260.

Ni, B., Paramathayalan, V.R. and Moulin, P., (2014) Multiple granularity analysis for fine-grained action detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.756–763.

Sadanand, S. and Corso, J.J., (2012) Action bank: A high-level representation of activity in video. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.1234–1241.

Shorten, C. and Khoshgoftaar, T.M., (2019) A survey on Image Data Augmentation for Deep Learning. *Journal of Big Data*, [online] 61, p.60. Available at: https://doi.org/10.1186/s40537-019-0197-0.

Sjarif, N.N.A. and Shamsuddin, S.M., (2016) Human action invarianceness for human action recognition. *SKIMA 2015 - 9th International Conference on Software, Knowledge, Information Management and Applications*.

Soomro, K., Zamir, A.R. and Shah, M., (2012) UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. [online] November. Available at: http://arxiv.org/abs/1212.0402.

TheSleepJudge Editorial Team, (2020) *Crimes that Happen While You Sleep*. [online] Available at: https://www.thesleepjudge.com/crimes-that-happen-while-you-sleep [Accessed 3 Aug. 2021].

Tran, D., Bourdev, L., Fergus, R., Torresani, L. and Paluri, M., (2015) Learning spatiotemporal features with 3D convolutional networks. *Proceedings of the IEEE International Conference on Computer Vision*, 2015 Inter, pp.4489–4497.

Versloot, C., (2021) *How to use sparse categorical crossentropy with TensorFlow 2 and Keras?* [online] MachineCurve. Available at:

https://www.machinecurve.com/index.php/2019/10/06/how-to-use-sparse-categorical-crossentropy-in-keras/ [Accessed 30 Sep. 2021].

Vrigkas, M., Nikou, C. and Kakadiaris, I.A., (2015) A review of human activity recognition methods. *Frontiers Robotics AI*, 2NOV, pp.1–28.

Wang, S., Zheng, J., Hu, H.M. and Li, B., (2013) Naturalness preserved enhancement algorithm for non-uniform illumination images. *IEEE Transactions on Image Processing*, 229, pp.3538–3548.

Xu, Y., Yang, J., Cao, H., Mao, K., Yin, J. and See, S., (2021) ARID: A New Dataset for Recognizing Action in the Dark. *Communications in Computer and Information Science*, 1370, pp.70–84.

Zhang, Y., Zhang, J. and Guo, X., (2019) Kindling the darkness: A practical low-light image enhancer. *MM 2019 - Proceedings of the 27th ACM International Conference on Multimedia*, pp.1632–1640.

Zhe Lin, Zhuolin Jiang and Davis, L.S., (2010) Recognizing actions by shape-motion prototype trees. Iccv, pp.444–451.

Zhou, W. and Zhang, Z., (2014) Human Action Recognition With Multiple-Instance Markov Model. 910, pp.1581–1591.

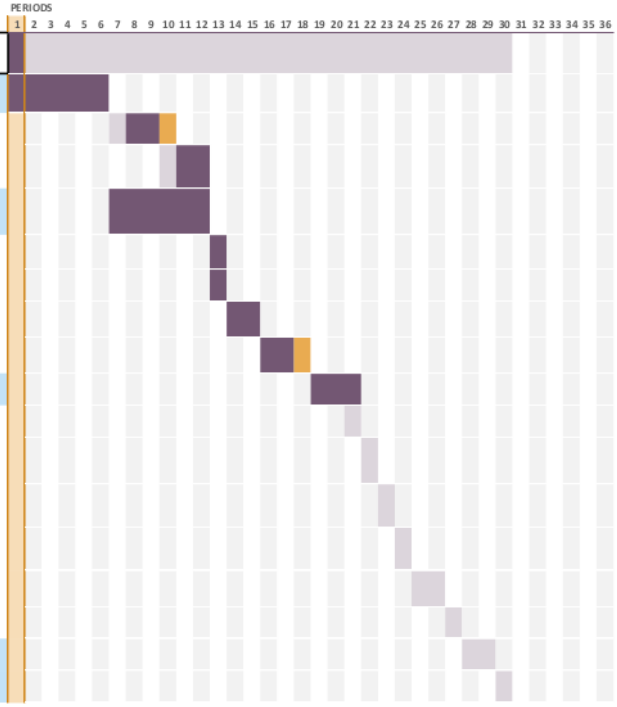
Appendix A: Research Plan

Project Planer - Action Recognition in Dark

Select a period to highlight at right. A legend describing the charting follows.

Period Highlight: 1 | Plan Duration | Actual Start | % Complete | Actual (beyond plan) | % Complete (beyond plan

ACTIVITY	PLAN START	PLAN DURATION	ACTUAL START	ACTUAL DURATION	PERCENT COMPLETE
Research Project: Action Recognition in Dark	1	30	1		60%
Milestone 1: Topic Submission with dataset	1	6	1	6	100%
Literature Review	7	3	8	3	100%
Document the Content for Research Proposal	10	3	11	2	100%
Milestone 2: Research Proposal Submission	7	6	7	6	100%
Get the data and prepare with hardware and software	13	0.5	13	0.5	100%
EDA and Data Preprocessing	13	1	13	1	100%
Document Low-Light Enhancement and action recognition techniques	14	2	14	2	100%
Document Research Methodology in details	16	2.5	16	3.5	100%
Milestone 3: Interim Report	19	3	19	3	100%
Pre-Process the video sequences	21	1			0%
Build Deep Learning Models for Action Recognition	22	1			0%
Comparitive Study between diffirent low light enhancement	23	1			0%
Evaluate the prediction by deep learning models	24	1			0%
Explaination for comparitive study between state-of-the-art methods	25	2			0%
Final Conclusion	27	1			0%
Milestone 4: Final Report	28	2			0%
Milestone 5: Video Presentation	30	1			0%



Appendix B: Research Proposal

1. Background

In today's world, machines are helping us in various ways by identifying various situations and problems around us. Human action recognition is one of the areas where these machine learning algorithms can help us and augment our capabilities around it. We need the capability to identify human actions not only in good lighting conditions but also in dark or low lighting conditions.

Action recognition on normal videos with good visibility has done a lot of progress in recent times and many machine learning algorithms are able to do it efficiently but action recognition in the low light situations is still pretty challenging and a hard nut to crack because the pictures or videos captured during the night time or low light situations suffer from poor visibility and contains a lot of noise and sometimes these videos are even difficult for human eyes to recognize accurately.

Unfortunately, there is not enough work done on this in machine learning space but we see lot of work going on at hardware side where cameras are getting equipped with night vision capabilities but as we all know and experience that night vision sensors degrade the quality of the images or videos by introducing lot of noise and it gets more difficult for present day feature extraction algorithms to deal with these noises. Another drawback with these sensors is that they make the surveillance cameras more expensive and it's difficult for everyone to afford and adopt these technologies. So there is a need of software solutions which are capable of recognizing the actions in videos which are shot in dark without the help of any night vision capabilities.

There are lot of research going on for

- Night surveillance at very sensitive areas around the cities for crime investigations
- Surveillance at the country borders to stop illegal infiltration
- Self driving cars at night time
- Outdoors home surveillance during night hours or in low light situations
- Studies for wild animals behaviour during night time without disturbing them.

These all issues are very important which world would like to address.

According to an article (TheSleepJudge Editorial Team, 2020) about crimes in United States, more than 50% of the major crimes happen during night time. And most of the crimes are never reported which happen in the dark because there are either no proofs or if it's get reported, video is not clear to identify the criminal. So it's a major issue for which we need the model which can enhance the dark videos and identify the human action.

One of the articles (Guglielmi, 2018) tells about that how wild mammal animals turning nocturnal because of the human activities in the forests during day time. Also humans use cameras with flash lights to study the wild animals' behaviours at night which actually disturbs then a lot during night time as well. We need to learn to see in the dark without the use of flash lights.

2. Problem Statement / Related Works

In last decade, many low light enhancement methods were proposed with different capabilities and different methods to enhance the dark images. Here, only few of the most popular methods are highlighted which work the best with ARID dataset and eventually help in identifying the human action from the dataset.

Some methods are based on histogram equalization, some are based on Gamma Intensity Correction where they try to amplify the low light in the images. Some methods are deep learning based where we need to have images in pairs of dark and bright images. Some methods are based on Retinex Theory where they suggest splitting the image in two parts, Reflectance and illumination where sometimes reflectance is considered as the enhanced image.

Recently, (Xu et al., 2021) proposed a new dataset (ARID) which contains 3784 video clips with 11 categories where each video is of around 1.2 seconds with 36 frames and all these videos are low contrast and low brightness videos. Together with dataset, a comparative study is done on the five different illumination methods which are based on Histogram Equalization, Gamma Intensity correction and Retinex Theory followed by human action classification on the videos.

According to this study, authors advocate Gamma Intensity correction (GIC) which gives the highest accuracy (78.03%) among all the methods. Authors have also done comparison using 3D-ResNext-101(Hara et al., 2018) classification model with original dark videos and enhanced video with various methods and concludes that GIC method gives the best improvements (3.30%) in the accuracy of the model and KinD(Zhang et al., 2019) enhancement method actually lowers the accuracy of classification model by (5.11%) in ARID Dataset.

One interesting thing to note here is, GIC doesn't necessarily do the best job in terms of image enhancement when we look at the image from the human lenses but its giving the highest accuracy (according to this paper) in action recognition from ML classification model point of view.

KinD (Zhang et al., 2019) (Kindling the Darkness), is one of the major work which was done recently for image enhancement. It's a deep learning based method which is inspired by Retinex Theory. This paper mostly focuses on the poor visibility and different types of degradation like noise and colour distortion. This research is done on LOL Dataset which

contains 500 low/normal light image pairs. 450 out of 500 images are used for training and 50 images are used for test.

KinD proposes layer decomposition and reflectance restoration neural networks, where restoration network adopts 5-layer UNet. It also proposes the Illumination Adjustment Net which can help in generating the ground-truth light level for images. In this research, batch size as 10 for layer decomposition net and batch size of 4 for reflectance and illumination adjustment net and stochastic gradient descent as optimization technique are used while building the models.

A comparative study is done on the 10 different image enhancement methods with the method proposed in the paper. There are some visual comparison and some metrics (i.e. PSNR, SSIM) comparison between images (enhanced with different methods) are present in the paper and based on that, it concludes that the proposed method gives clear advantage over all the state-of-the-art methods selected in the paper and it outperforms all the competitors using LIME (Guo et al., 2017) and NPE (Wang et al., 2013) datasets as well. One thing to note here is, by looking at the visual comparison, it can be clearly seen that the images, enhanced with KinD**,** do not suffer from any noise or any sort of colour distortion.

LIME (Guo et al., 2017) (Low-Light Image Enhancement via Illumination Map Estimation) is another very successful illumination technique and beats most of the state of art techniques of its time. This research warns that just recalling the visibility of the dark regions can create another problem of light saturation at the relatively bright regions so there will be a loss of details in the image.

Lime method is also based on Retinex theory but it mostly focuses on the illumination part. This method builds an illumination map and refines it by finding the maximum intensity of each pixel in R G B channels. For the refinement of illumination map, Augmented Lagrangian Multiplier (ALM) based algorithm is used which is also quite efficient algorithm and reduces the computational cost. Illumination map estimation and refinement actually considers the neighbouring pixels which help in local consistency of illumination which helps in getting uniform illumination across the image. Gamma correction and denoising and recomposition are done to get the best results for illumination map.

HDR dataset is used for this study and some low light images were chosen from the dataset and their LOE numbers are compared with all the other competitors. A comparative study is done on the 9 different image enhancement techniques with the proposed method and based on

the visual comparison and LOE numbers, it can be clearly seen that the proposed method gives the clear advantage over all those state-of-the-art methods selected in the paper.

Naturalness Preserved Enhancement Algorithm (Wang et al., 2013), is also based on Retinex Theory but it essentially focuses on preservation of naturalness of the image which is difficult to achieve in non-uniformly illuminated images.

This paper proposes a new algorithm for non-uniform illumination images where it also proposes lightness-order-error measure for naturalness and bright-pass filter which helps in decomposing the image into reflectance and illumination and bi-log transformation is done to consider the balance between naturalness and details. This paper tries to improve the local variation in the image and also doesn't harm the global trend of the intensity at the same time. So the paper focuses on the Reflectance extraction and relative order illumination compression.

This paper introduces its own dataset which contains more than 150 images. A comparative study has been done on these images with 6 state-of-the-art methods together with the proposed algorithm and based on the visual comparison, quantitative measurement results of discrete entropy and LOE numbers, it can be clearly seen that the proposed method gives the clear advantage over all those state of the art methods selected in the paper. According to the paper, their solution doesn't scale well with video files because it introduces slight flickering in the video files and authors plan to do further research in this area.

EnlightenGAN**:** Deep Light Enhancement Without Paired Supervision (Jiang et al., 2021), is another successful method which proposes a unsupervised dubbed EnlightenGAN (one path GAN), that can be trained and evaluated without normal/low-light image pairs unlike most of the deep learning based methods. Instead of ground truth data, information is extracted from input image and used for unpaired training.

In this paper, proposed method uses an attention U-Net neural network as the generator and uses global and local discriminator to maintain the texture and details of the image. This paper advocates for self-regularized attention map instead of supervised learning. Also in the experiment, it uses data from multiple dataset where around 900 low light and 1K+ normal light images are collected from various datasets.

EnlightenGAN is trained by 100 epochs and with learning rate of 1e-4 with Adam Optimizer and then another 100 epochs where learning rate was linearly decayed to zero. A comparative study is also done with 6 state-of-the-art methods together with EnlightenGAN and

based on the visual comparison and based on NIQE scores, it can be clearly seen that the proposed method gives the clear advantage over all the methods selected in the paper.

Human action recognition based on recognition of linear patterns in action bank features using convolutional neural networks (Ijjina and Mohan, 2014), proposes a deep CNN architecture for recognizing human actions in videos using action bank features. Action bank is nothing but a predefined set of videos converted to linear representation and saved as action bank features. This research proposes a convolutional network with linear mask which can capture the localized patterns for each action.

Generally, in case of action recognition in a video requires a 3D CNN to learn the spatiotemporal features from videos but this paper advocates using 2D CNN with action bank which is based on the concept of speech recognition using the spectrogram of audio data. They have designed a CNN which exploits this similarity in action bank features and this will drastically reduce the computational time for action recognition.

UCF50 dataset containing 205 videos is used in this research, these input videos first get processed by the feature extraction module to extract the action bank, and then these action bank features are given as input to pattern recognition module for training. The CNN utilizes the similarity patterns to assign an action label to the videos. By looking at the results with this approach looks quite promising where the proposed method is able to achieve 93-94% accuracy in action recognition.

Human Action Invarianceness for Human Action Recognition (Sjarif and Shamsuddin, 2016), proposes to use human action shape or silhouette uniqueness to recognizing the human actions. Human action features can be extracted by using integration moment invariant. Action features are actually based on how silhouette moves in video frames. In human action invariance, the paper proposes three processes like extracting global features, similarity measurement between features and intra and interclass analysis.

Authors have used IXMAS dataset which contains 13 different actions performed by 10 people. The experiment is done with different video frames i.e. 30, 120 and 300. This research uses various techniques to achieve the higher accuracy like wavelet, PCA, Normalization, pre and post discretization. A comparative analysis on other methods is also done which other methods are not able to give the good accuracy with this dataset. But the method proposed by this paper is able to perform very well and is able to predict the human action with high accuracy up to 98-

99%. This paper has given promising results and could be one of methods in my research in identifying the human action recognition in dark.

Learning Spatiotemporal Features with 3D Convolutional Networks (Tran et al., 2015), proposes deep 3D ConvNets which are trained on large dataset and suggest that these 3D ConvNets are better than 2D ConvNets for spatiotemporal feature extraction. This paper evaluates both the networks and concludes that 3D ConvNets gives much better accuracy (98%) compare to 2D ConvNets.

Learning to See Moving Objects in the Dark (Jiang and Zheng, 2019), method proposes a new optical system which can capture both bright and dark videos of the same scene, to have both training and ground truth dataset.

They discouraged the use of infrared sensors in the cameras because in forests it can disturb the animals and might trigger uncontrollable animal reactions. So they advocate enhancing the images which are captured by ordinary cameras with ND filter so the proposed camera can click both dark and bright images simultaneously. This paper introduces a new dataset which contains 179 pairs of videos consisting of 35800 extremely low-light images and their corresponding properly lighted images.

This paper proposes a 3D U-net based network for low-light enhancement. A comparative study is done between state-of-the-art methods and the proposed method which concludes that the proposed method is the best among all the methods and also according to this paper, proposed method is able to tackle the flickering issues in the enhanced videos.

Action Recognition in the Dark via Deep Representation Learning (Ulhaq, 2018), proposes action recognition from multiple video streams using deep multi-view representation learning. This paper talks about video captured in low lighting condition and multi-sensor scenario which is a novel method to fuse spatio-temporal deep correlations from multiple streams.

This paper proposes an approach to deal with very low quality dark night time videos. In the experiment, Night-Vision Video dataset (NV) is used. This dataset contains videos recorded by two different cameras Raytheon Thermal IR-2000B and Panasonic WV-CP470.

They were able to achieve average precision of 80% which is pretty impressive. But this method requires multiple videos captured via different sensors in cameras.

3. Research Questions

The below research questions are formulated based on Literature Review done in the field of human action recognition in dark:

- Are there any conclusions to use a particular low light enhancement method which actually helps in human action recognition?
- Is the best low light enhancement method really required to get high accuracy in human recognition or a simple low light enhancement method is good enough for the model to perform well in identifying the human actions?

  Here we are going to focus on the human recognition in dark, not on the visual quality of the videos.
- What will be the Human action recognition model's accuracy for different low light enhancement methods?
- Can we use different algorithms for video enhancement based on business needs? User might be interested to see the actual enhanced video after human action recognition is done by the model.

4. Aim and Objectives

The aim of this research is to do the comparative study on the present capabilities of low light enhancement methods and build a model and recognize the human actions in ARID dataset. The goal of this study is to identify the best low light enhancement method for videos which are shot in dark and which works the best and gives the highest accuracy with an action recognition deep learning model.

The research objectives are formulated based on the aim of this study, which are as follows:

- To analyze the most popular low light enhancement methods and identify the best method among all of them which works the best for human action recognition.
- To propose a deep learning model which works the best in identifying the human actions with low light enhanced videos.

- To perform comparative study on 2D CNN and 3D CNN methods for human action recognition.

5. Significance of the Study

The significance of this study is to explore various low light enhancement methods and find the best method which is efficient in recognizing the human action in the dark videos and will augment human capabilities during night time.

There are many areas where this research will help like night surveillance. To make our city and country safe, we need the capabilities to do surveillance at very crime sensitive areas. This model will help in recognizing the human actions during the night hours; we don't need to deploy a human at every place for surveillance. Most of the times it is very difficult for the humans as well to see in the dark without night vision capabilities. This action recognition method will help our police to protect our neighbourhood and people.

The second use case is Self Driving Cars; there are lot of researches going on for autonomous cars. As we know, self driving cars is still a big challenge and auto maker companies still facing a lot of challenges to make a perfect autonomous car where driving in the dark will be another hard nut to crack. I hope this study will be able to help this area to some extent because we need a model which can recognize things in dark.

This model, with the comparatively analysis, will be able to help data scientists to find different methods to identify human actions in various scenarios. This study will be extended to give suggestions about different low-light enhancement methods which are best suited for different dark environment scenarios.

6. Scope of the Study

Due to lack of time frame, the scope of the study will be limited as below:

- The dataset is taken from ARID (Xu et al., 2021) dataset which is publicly available and data validation is not the part of this study. The scope of study is to analyze only 11

human actions mentioned in the data but this can be extended to any number of human actions in the future.

- This study contributes to the comparative study on different low light enhancement techniques, comparative study on human action recognition techniques and suggests the best combination of both which help in recognizing the human action in the videos which are shot in the dark.
- In case of low-light illumination, this study will limit to use some of the state-of-the-art techniques like KinD (Zhang et al., 2019), LIME (Guo et al., 2017), EnlightenGAN (Jiang et al., 2021) and Gamma Intensity Correction and plans to do the comparative analysis by comparing the different metrics.
- In case of human action recognition, this study will limit to use some of methods like action bank feature extraction (Ijjina and Mohan, 2014), utilize a rolling prediction average and utilize the 2D, 3D CNN, RNN methods for model building.

7. Research Methodology

Today, as our population is increasing, it is getting difficult to do the night surveillance manually by humans, we need some kind of solutions to solve the problems we face during night time; some scenarios are like night surveillance, illegal infiltrations at country borders, self driving cars during night time and analyze animal behaviours in the wild during night. Humans need to augment their capabilities during the night. A lot of work in done on Human action recognition with the properly lighted videos and images but not enough work is done on action recognition in dark videos. However some of the state-of-the-art methods have done awesome job in enhancing the low lighted images and we are in the right direction.

The motive of this research is to analyze human actions in videos which are shot in the dark and with normal cameras. Human action recognition directly on the dark videos will not be efficient because it will be difficult to extract features out of the videos. So this paper suggests the research in two parts. Firstly, find the best method to pre-process and do the low light enhancement in the dark videos. Secondly, do the human action recognition on enhanced videos. There are some of the methods which are available for low light enhancement for videos and

images. A comparative study needs to done on various methods and propose the best method which gives the highest accuracy and lowest computation time. Some methods enhance the images visually very well but a question to ask is “do we really need these kind of refinements for action recognition in the dark?”

7.1 Dataset

This research uses ARID dataset; this dataset contains more than 3700 video clips with 11 different kinds of human actions where each class contains around 110 clips For example: Drinking, Jumping, picking, pouring, running, sitting, standing, turning, walking and waiving. Each video clip has fixed frame rate of 30 frames/sec and 1.2 seconds long. (Xu et al., 2021) have created this dataset specifically for analyzing dark videos. These dark videos are shot from the normal cameras without using any kind of night vision sensor. Some of videos are so dark that it is very difficult for even human eye to identify the actions in the videos.

The purpose of this dataset is to provide the dark videos which were really shot in the dark because we see most of the researches are done on synthetic dark videos or the dark videos which were created using some camera filters. Models built on synthetic dataset don’t perform very well with the actual dark videos.

7.2 Data Pre-Processing

Before starting to build the model for action recognition, we need to pre-process the data. Here this research plans to do the low light enhancement in the videos as part of the pre-processing. For low light enhancement, some of methods will be used which make sense to use with ARID [1] dataset. Some of methods which will work well this dataset are: KinD (Zhang et al., 2019), LIME (Guo et al., 2017), EnlightenGAN (Jiang et al., 2021), Histogram Equalization (HE), Gamma Intensity Correction (GIC). EnlightenGAN seems to be the most impressive and promising methods as of now among all the low-light image enhancement methods.

These methods are chosen because they directly enhance the low lighted images without the need of paired bright light images for training the model. These all the enhancement methods were invented for images, so these methods need to be applied on each of the video frames.

During enhancement of the videos, a comparative study is planned where their computational time will be on focus.

7.3 Model Building

After the dark videos enhancement, it is required to use multiple CNN based neural networks for feature extraction and find the best neural network which gives the highest accuracy and the lowest computational time on ARID dataset.

For model building, Keras framework is required for designing the 2D or 3D CNN models where Tensor Flow will act as a backend to Keras. There are multiple ways and methods to do the human action recognition, but this research plans to use three different methods for video classification

- Using feature extraction module to extract the action bank features (Ijjina and Mohan, 2014). We will create the action bank for all the 11 action categories in the dataset and train the model on the action bank and then build the 2D CNN model to do the video classification.

  This method is chosen because it will reduce the computational time drastically as we are using 2D CNN for video analysis instead of 3D CNN and this method is giving more than 90% accuracy in action recognition.

- Another way is to use 2D CNN and utilize a rolling prediction average (Rosebrock, 2019) where author suggested using a 2D CNN model and classifying each frame individually and keeping the list of predictions and taking the average of last K predictions and predicting the label with largest probability.

  Here this method is chosen because it's a 2D CNN model which will help in reducing the computational time and according to the results shown by the author its working nicely on the action recognition on the videos.

- Using neural network architectures like Long short-term memory (LSTMs) and Recurrent Neural Networks (RNN), because videos are nothing but sequence of images (frames). RNNs are well suited for sequential data classification.

  This method is chosen because a comparative study is required between 3D CNN model and 2D CNN model approaches to find the best method for ARID dataset.

7.4 Model Training and Evaluation

Dataset will be split into two parts (70% and 30%), 70% of the videos will be used for training the dataset and rest 30% will be used for validation of model.

While model training, below hyper parameters will be considered

Parameter	Value
No. of Epochs	100
Batch Size	30
Optimization Technique	Adam
Learning Rate	Default for Adam
ReduceLROnPlateau	Patience=5, factor=0.1

Table 3 Hyper Parameters

As shown in Table 1 and Figure 1, model will be trained with 100 Epochs with batch size of 30. These values are chosen based on the resource availability during this research. Other than that, ADAM will be used as the optimization technique while finding the global minima and reduce the learning rate at the plateau with the factor of 0.1 and patience level of 5.

Below metrics will be considered which will tell us how the model is performing during model training and evaluation:

Metrics	Description
categorical_accuracy	Accuracy of the model on the training data
val_categorical_accuracy	Accuracy of the model on the validation data
Precision	Proportion of positive identifications was actually correct
Recall	Proportion of actual positives was identified correctly
F1-Score	This a harmonic mean of Precision and Recall

Table 4 Model Evaluation Metrics

We have to watch out for categorical accuracy and validation categorical accuracy especially so that we have balanced performance on train and test set. Both the accuracies should be in the same range and more than 80% for a good model. If categorical accuracy is too high and validation accuracy is too low then it will considered as over fitting which we have to avoid.

Only accuracy may not determine the performance of the model but other metrics like Precision and Recall are required to be analyzed which will help in determining the performance of the model.

Methodology described in Figure 1, will be repeated for every low light enhancement technique and for every action recognition algorithm. All the models will be saved to Model Repository. All the models will be evaluated and compare against various metrics and propose the best low light and human action recognition algorithm for ARID dataset.

Thesis-AshishDhyani-2.docx

ORIGINALITY REPORT

8%	6%	3%	4%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	cs.uoi.gr Internet Source	1%
2	Submitted to University of Liverpool Student Paper	1%
3	Submitted to Liverpool John Moores University Student Paper	1%
4	Earnest Paul Ijjina, C. Krishna Mohan. "Human Action Recognition Based on Recognition of Linear Patterns in Action Bank Features Using Convolutional Neural Networks", 2014 13th International Conference on Machine Learning and Applications, 2014 Publication	1%
5	www.analyticsvidhya.com Internet Source	1%
6	mafiadoc.com Internet Source	1%
7	www.diva-portal.org Internet Source	1%

8	commons.trincoll.edu Internet Source	1%
9	otexts.com Internet Source	1%

Exclude quotes	On	Exclude matches	< 1%
Exclude bibliography	On		