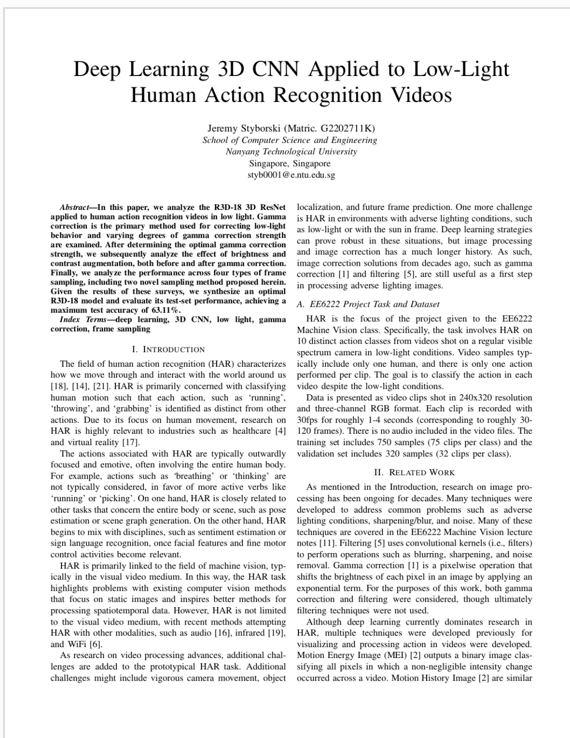
Digital Receipt

This receipt acknowledges that Turnitin received your paper. Below you will find the receipt information regarding your submission.

The first page of your submissions is displayed below.

Submission author:	SCSE STYBORSKI JEREMY ANDREW
Assignment title:	EE6222 Assignment: Action Recognition in the Dark
Submission title:	Styborski_DarkHAR_Report
File name:	Styborski_DarkHAR_Report.pdf
File size:	292.25K
Page count:	5
Word count:	3,836
Character count:	20,855
Submission date:	03-Nov-2022 11:57AM (UTC+0800)
Submission ID:	1943101923



Deep Learning 3D CNN Applied to Low-Light Human Action Recognition Videos

Jeremy Styborski (Matric. G2202711K)
School of Computer Science and Engineering
Nanyang Technological University
Singapore, Singapore
styb0001@e.ntu.edu.sg

***Abstract*—In this paper, we analyze the R3D-18 3D ResNet applied to human action recognition videos in low light. Gamma correction is the primary method used for correcting low-light behavior and varying degrees of gamma correction strength are examined. After determining the optimal gamma correction strength, we subsequently analyze the effect of brightness and contrast augmentation, both before and after gamma correction. Finally, we analyze the performance across four types of frame sampling, including two novel sampling method proposed herein. Given the results of these surveys, we synthesize an optimal R3D-18 model and evaluate its test-set performance, achieving a maximum test accuracy of 63.11%.**

***Index Terms*—deep learning, 3D CNN, low light, gamma correction, frame sampling**

I. Introduction

The field of human action recognition (HAR) characterizes how we move through and interact with the world around us [18], [14], [21]. HAR is primarily concerned with classifying human motion such that each action, such as ‘running’, ‘throwing’, and ‘grabbing’ is identified as distinct from other actions. Due to its focus on human movement, research on HAR is highly relevant to industries such as healthcare [4] and virtual reality [17].

The actions associated with HAR are typically outwardly focused and emotive, often involving the entire human body. For example, actions such as ‘breathing’ or ‘thinking’ are not typically considered, in favor of more active verbs like ‘running’ or ‘picking’. On one hand, HAR is closely related to other tasks that concern the entire body or scene, such as pose estimation or scene graph generation. On the other hand, HAR begins to mix with disciplines, such as sentiment estimation or sign language recognition, once facial features and fine motor control activities become relevant.

HAR is primarily linked to the field of machine vision, typically in the visual video medium. In this way, the HAR task highlights problems with existing computer vision methods that focus on static images and inspires better methods for processing spatiotemporal data. However, HAR is not limited to the visual video medium, with recent methods attempting HAR with other modalities, such as audio [16], infrared [19], and WiFi [6].

As research on video processing advances, additional challenges are added to the prototypical HAR task. Additional challenges might include vigorous camera movement, object localization, and future frame prediction. One more challenge is HAR in environments with adverse lighting conditions, such as low-light or with the sun in frame. Deep learning strategies can prove robust in these situations, but image processing and image correction has a much longer history. As such, image correction solutions from decades ago, such as gamma correction [1] and filtering [5], are still useful as a first step in processing adverse lighting images.

A. EE6222 Project Task and Dataset

HAR is the focus of the project given to the EE6222 Machine Vision class. Specifically, the task involves HAR on 10 distinct action classes from videos shot on a regular visible spectrum camera in low-light conditions. Video samples typically include only one human, and there is only one action performed per clip. The goal is to classify the action in each video despite the low-light conditions.

Data is presented as video clips shot in 240x320 resolution and three-channel RGB format. Each clip is recorded with 30fps for roughly 1-4 seconds (corresponding to roughly 30-120 frames). There is no audio included in the video files. The training set includes 750 samples (75 clips per class) and the validation set includes 320 samples (32 clips per class).

II. Related Work

As mentioned in the Introduction, research on image processing has been ongoing for decades. Many techniques were developed to address common problems such as adverse lighting conditions, sharpening/blur, and noise. Many of these techniques are covered in the EE6222 Machine Vision lecture notes [11]. Filtering [5] uses convolutional kernels (i.e., filters) to perform operations such as blurring, sharpening, and noise removal. Gamma correction [1] is a pixelwise operation that shifts the brightness of each pixel in an image by applying an exponential term. For the purposes of this work, both gamma correction and filtering were considered, though ultimately filtering techniques were not used.

Although deep learning currently dominates research in HAR, multiple techniques were developed previously for visualizing and processing action in videos were developed. Motion Energy Image (MEI) [2] outputs a binary image classifying all pixels in which a non-negligible intensity change occurred across a video. Motion History Image [2] are similar