Action Recognition in the Dark

Yijn Yang
School of Electrical and Electronic Engineering
Nanyang Technological University
yang0748@e.ntu.edu.sg

***Abstract*—The rapid increase in the volume of video data as a result of the prevalence of video devices and applications has led to a boom in automated video analysis, which has attracted the interest of an increasing number of researchers. And human action recognition (HAR) is one of the fundamental tasks in video analytics. The task of human action recognition in video has advanced significantly with the release of multiple video datasets, but is typically based on videos in normal environments. When this type of model is migrated to video datasets in less-than-ideal environments, such as dark environments with inadequate lighting, human action recognition generally fails to perform well. The purpose of this report is to develop and evaluate a project for human action recognition in dark datasets from scratch using the deep learning (CNN) method. The results indicate that applying frame enhancement methods to the dark dataset can improve the accuracy of HAR to some degree, but the overall performance is still unsatisfactory and does not allow for effective action category recognition.**

***Index Terms*—video analytics, dark environments, human action recognition, convolutional neural networks**

I. Introduction

Due to the proliferation of video data, action recognition techniques in the field of video analytics are utilised in a variety of applications, including intelligent surveillance to track human trajectories [1], smart vehicles to detect pedestrians and recognise their body movements [2], and factory material and safety monitoring [3]. Human action recognition has been a popular topic in recent years, with numerous new approaches being proposed. For instance, [4] utilised the CNN-LSTM model to process and store frame information and temporal information. The authors of [5] propose using RF signals in conjunction with representations of the human skeleton to recognise human movements in dim light and across walls. The paper [6] proposed a 2D-CNN extraction of skeleton information combined with 3D-CNN using infrared data for human action recognition. In addition, there is also the use of a thermal imaging camera combined with deep learning approaches [7]. However, the data for these methods are derived from video datasets of humans in normal environments or from special signals captured by specialised cameras and other devices. The robustness of projects in normal environments is insufficient, and accuracy decreases when migrating to harsh environments. In addition, the cost of additional equipment in addition to the video itself is an issue for dark environment data projects.

In deep learning, the mainstream approaches for video-based motion detection are to use 3D-CNN jointly to extract spatial and temporal information [1], [8], or to use dual-flow networks that provide motion information in the temporal dimension by means of optical flow [4]. A traditional 2D-CNN can also be used to process the video as multi-frame images, taking into account only the spatial domain.

In addition, there are synthetic dark ambient videos, which are created by transforming normal lighting video samples. However, there is an imperceptible difference between the pseudo-dark environment and the real dark environment. For instance, the real dark environment has both low luminance and contrast, whereas the synthetic dark environment video has low luminance but higher contrast, which does not match the real environment [9].

Therefore, this project will be trained and evaluated using a video dataset captured in a real dark environment. Action Recognition in the Dark (ARID) dataset is currently the only dataset dedicated to human action recognition captured in a real dark environment using a regular camera [9].

The rest of the report is structured as follows: Section 2 will discuss some methods used in the Human Action Recognition domain with general environment datasets, followed by some of the methods applied to dark datasets. In Section 3, the datasets and CNN methods used for this project as well as the precise experimental procedures and adjustments will be described. Section 4 will focus mainly on the accuracy result of the validation dataset and its interpretation. Section 5 will conclude the project and discuss prospective outcomes.

II. Related Works

Larger video datasets, such as HMDB51 [10], UCF101 [11], Activity Net [12], and Kinetics [13], have pushed the boundaries of video analysis, including human action recognition. However, as the vast majority of these mainstream video datasets are captured in conventional environments with sufficient light intensity, this limitation has led to research models that cannot easily be applied to the evaluation of human action recognition in non-conventional environments, i.e. dark environments. In recent years, with the release of the ARID dataset and the collection of private datasets by researchers themselves, there has been increased research into the development of models using data obtained in dark environments. This section will begin with a review of model studies based on conventional datasets, followed by a discussion of models developed using dark datasets.

A. Human Action Recognition

3D-CNNs, dual-stream networks, temporal networks, and human skeleton representations are the predominant approaches to HAR. For example, the authors of [4] propose two approaches, the first of which uses multiple convolutional layers for feature extraction and a pooling layer to combine the information from each frame in order to reveal their relationship. The second is a temporal network CNN + LSTM approach, which also employs multiple convolutional layers for feature extraction and then feeds the extracted features into an LSTM network, gradually identifying the long-term temporal relationships between each frame as the cell state is updated. Additionally, the author [4] investigates the use of optical flow images as input to improve the model. In the UCF-101 dataset, CNN, CNN+LSTM models, and the optical flow input method all achieved very high accuracy rates, greater than 87%.

The authors of [14] used 3D-CNN, RestNeXt-101 [15], to evaluate on all three large datasets and obtained a good performance of 94.5% in UCF-101, 70.2% in HMDB-51, and 78.4% on Kinetics' training set, proving that 3D-CNN is more effective.

B. Human Action Recognition in the Dark

In [5], the authors take advantage of the fact that RF signals can bypass and reflect off objects and the human body to generate and combine a representation of the human skeleton, and then use a neural network with an attention feature learning network to extract spatial and temporal features from the skeleton and pass them into a multi-proposal network for action recognition, thus achieving recognition of human actions in poor conditions of dark light, and even through walls. The dataset for this project is the author team's collection of human action photographs taken with RF signals and multi-view cameras.

The author of [16] proposed a Delta-sampling strategy and a Resnet BERT with Resnet-34 as the backbone, and achieved an accuracy of 90.46% on a subset of the ARID dataset, demonstrating the effectiveness of sampling methods, image enhancement techniques, and models.

III. Methodology

This section first introduces the dataset utilised for the project. The enhanced technology, gamma intensity correction, that was applied to the data is then described. Following this, the CNN networks used are presented. Finally, the training and validation operational principle is described.

A. Dataset

The dataset used for this project was derived from the ARID dataset with ten classes of human actions, including Drink, Jump, Pick, Pour, Push, Run, Sit, Stand, Turn and Walk. Each video in the dataset had a frame rate of 30 fps and a resolution of 320 by 240 pixels. 70% of this was the training set (730 videos) and 30% was the test set (320 videos).

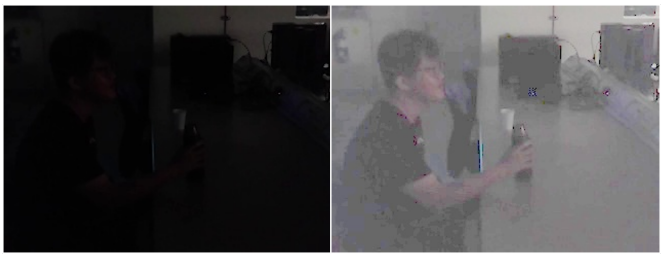

Fig. 1. Comparison of the same drinking picture before and after Gamma Correction

B. Gamma Intensity Correction

Gamma correction is a non-linear transformation which can compensate for the loss of luminance. Eq.1 is the formula of gamma correction. In this project, its value is set to $\gamma = 5$.

$$V_{out} = AV_{in}^{\gamma} \tag{1}$$

Figure 1 intuitively depicts luminance compensation via gamma correction. Before correction, it is more difficult to recognise the action as a drinking action with the naked eye; after correction, the action becomes evident.

C. Networks

The training network for this project is the pre-trained convolutional neural network from PyTorch, allowing the network to converge more quickly. 2D-CNN, ResNet-34, was used in the first stage of training, while 3D-CNN, Resnet's variant ResNeXt-50 (32×4d), was used in the later stages.

D. Implementation

This section will detail the implementation and operation principles of the code.

1) Training: The training dataset is sampled every 8 frames to ensure that no useful information is missed. Gamma correction is applied to the sampled images to increase their luminance. After initialising the pre-training weights, the images are fed to the 2D-CNN or 3D-CNN for training in human action recognition.

In the initial stage, each input into the network consisted of a sequence of 16 enhanced frames, whereas in subsequent stages, each input consisted of a sequence of 32 enhanced frames. After each training epoch on the dataset, the model's parameters are automatically saved. During training, the learning rate begins at $1e^{-4}$ and is divided by 10 when the accuracy of the validation set reaches a plateau.

2) Recognition: The validation set employs a slightly different but comparable methodology to the training set. Based on the total number of frames, each video in the validation set is sampled over 7 frames uniformly. The sampled images are then also gamma corrected and fed into the model in batches of 16/32 for validation. Similar to CNN classification of single images, each image has its label.

Creating a list with the same dimensions as the validation set. When the output of the validated image matches the label, one is added to the value of the corresponding position in the

TABLE I
COMPARISON OF THE ACCURACY(%) OF RESNET-34 WITH DIFFERENT LEARNING RATES UNDER BATHCHSIZE 16

ResNet-34	acc($lr = 1e^{-4}$)	acc($lr = 1e^{-5}$)	acc($lr = 1e^{-6}$)
1	27.19	22.81	14.69
2	26.88	24.69	18.44
3	24.38	23.44	21.56
4	21.25	23.13	23.13
5	22.50	20.63	21.88
6	25.31	24.38	23.13
7	20.63	25.31	22.50
8	22.81	22.50	23.13
9	25.00	24.38	25.31
10	27.50	24.06	24.69
11	25.31	23.44	24.38
12	28.13	23.75	25.00
13	30.00	24.69	26.25
14	27.81	25.31	24.06
15	28.75	25.00	25.31
16	27.81	25.63	23.13
17	26.25	25.31	25.31
18	21.56	25.94	25.63
19	23.13	25.31	25.00
20	21.56	23.13	24.38

TABLE II
COMPARISON OF THE ACCURACY(%) OF RESNET-34 WITH DIFFERENT LEARNING RATES UNDER BATHCHSIZE 32

ResNet-34	acc($lr = 1e^{-4}$)	acc($lr = 1e^{-5}$)	acc($lr = 1e^{-6}$)
1	24.06	19.06	12.19
2	25.31	24.38	13.13
3	24.69	23.75	14.69
4	20.00	24.38	16.88
5	21.56	20.38	17.50
6	23.44	24.69	20.00
7	19.69	25.94	20.00
8	19.06	24.69	20.31
9	20.00	23.75	20.31
10	21.56	23.44	20.94
11	22.81	25.00	20.94
12	20.94	23.44	22.81
13	23.75	24.06	21.56
14	21.88	23.13	22.50
15	23.44	22.81	22.81
16	24.06	24.38	21.88
17	23.44	23.13	22.81
18	24.06	24.06	22.50
19	24.69	20.94	22.19
20	25.00	25.31	22.81

list; otherwise, one is subtracted. Finally, the overall accuracy of the validation set can be computed based on the number of positive values of all positions in the list (only the odd number 7 of frames are sampled). When the correctness rate has stabilised, training and validation are completed.

IV. EXPERIMENT

This section will show all the validation results from the training process, and the final validation dataset results, with two decimal places only.

The images before gamma correction were first passed into ResNet-34 for training and evaluation, but the results of the evaluation were rather surprising, with accuracy rates essentially around 10.3% at learning rates of 1e-4 and 1e-5. when the learning rate was 1e-6, the accuracy rate improved but could only plateau at a maximum of 14.68%. This indicates that it is difficult for ResNet-34 to learn features in dark images. And due to time and memory constraints, ResNeXt-50 32x4d was not trained to evaluate images prior to gamma correction.

For images after Gamma Correction, Table I displays the percentage accuracy of ResNet-34 for different learning rates (ranging from 1e-4 to 1e-6) at a batch size of 16. It can be seen that at the learning rate of 1e-4, the accuracy fluctuates significantly, and a single training epoch may have an accuracy fluctuation of approximately 5%. This indicates that the model cannot converge because the learning rate is set too high. At a learning rate of 1e-5, we can observe that the model still oscillates to some extent in the early stages of training, but the magnitude is smaller than before, and the results of the last few epochs are relatively stable. When the learning rate is 1e-6, the gradient decreases more slowly, so the loss in early epochs of 1e-6 is high and the accuracy rate is low, but it fluctuates the least and rises steadily until it is stable.

TABLE III
COMPARISON OF THE ACCURACY(%) OF RESNEXT-50 32×4D WITH DIFFERENT BATHCHSIZE UNDER SAME LEARNING RATE 1E-5

ResNeXt-50 32×4d	batch size (16)	batch size (32)
1	25.31	26.88
2	26.88	30.00
3	29.06	27.50
4	29.69	27.50
5	30.31	28.44
6	27.19	28.44
7	28.75	28.13
8	26.88	29.06
9	27.50	29.06
10	27.81	28.44
11	27.50	29.06
12	29.38	30.00
13	28.13	28.75
14	27.19	30.00
15	26.88	29.69
16	27.50	31.25
17	27.19	27.50
18	26.88	27.50
19	26.88	28.13
20	28.13	28.44

Table II shows the accuracy of the validation dataset after increasing the batch size. The increase in batch size results in a more precise gradient descent direction and less oscillation compared to Table I.

ResNext-50 32x4d was only trained and evaluated at learning rates of 1e-5 due to time and memory limitations; the results are shown in Table III. As can be seen, the average accuracy improves as the batch size grows and reduces some fluctuations, but they are all extremely small.

As for test dataset, I chose the model parameters with the highest validation accuracy for Resnet-34 and ResNext-50 32×4d based on the validation accuracy of these two networks (Resnet-34: lr=1e-4, epoch=13, batch size=16; ResNext-50

32×4d: lr=1e-5 epoch=16, batch size=32). They are separately loaded into their respective models and the accuracy of the test dataset is assessed. The test dataset was evaluated identically to the validation dataset, with sampling and gamma correction, prior to being fed into the model with loaded training parameters. The accuracy of the Resnet-34 model was 36.875%, while the accuracy of the ResNext-50 32x4d model was 32.50%. This is primarily due to the fact that the ResNext-50 32x4d model's training is still relatively lacking.

The codes and pretrained models used in this study are publicly available in https://github.com/Lucky-Robin/Human-Action-Recognition-in-the-Dark.

V. Conclusion

This project accomplished human action recognition in the dark dataset ARID using 2D-CNN and 3D-CNN, with a best performance of 36.875% on the final test set. In addition, the completion of this project from scratch led to an increase in experiential learning.

Despite the relatively satisfactory results achieved by this project, there are still many directions for continued improvement in both the CNN model and the pre-processing process. The following are some of the future recommendations for development and improvement of this research:

1) The 3D-CNN model was not sufficiently trained due to time and memory limitations, and the number of training sessions and batch size for the 3D-CNN should be increased next time.
2) In the future, histogram equalisation and post-image enhancement denoising techniques could be added to the project's image pre-processing techniques.

References

[1] A. Ullah, K. Muhammad, W. Ding, V. Palade, I. U. Haq, and S. W. Baik, 'Efficient activity recognition using lightweight CNN and DS-GRU network for surveillance applications', Appl. Soft Comput., vol. 103, p. 107102, May 2021.

[2] L. Chen et al., 'Survey of pedestrian action recognition techniques for autonomous driving', Tsinghua Sci. Technol., vol. 25, no. 4, pp. 458–470, Aug. 2020.

[3] C. Xingyu and L. Hao, 'Application Analysis of Infrared thermal imaging Technology in Intelligent Manufacturing Field', J. Phys. Conf. Ser., vol. 1693, no. 1, p. 012129, Dec. 2020.

[4] J. Y.-H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, 'Beyond Short Snippets: Deep Networks for Video Classification'. arXiv, Apr. 13, 2015.

[5] T. Li, L. Fan, M. Zhao, Y. Liu, and D. Katabi, 'Making the Invisible Visible: Action Recognition Through Walls and Occlusions', arXiv, Feb. 2020.

[6] A. M. De Boissiere and R. Noumeir, 'Infrared and 3D Skeleton Feature Fusion for RGB-D Action Recognition', IEEE Access, vol. 8, pp. 168297–168308, 2020.

[7] G. Batchuluun, D. T. Nguyen, T. D. Pham, C. Park, and K. R. Park, 'Action Recognition From Thermal Videos', IEEE Access, vol. 7, pp. 103893–103917, 2019.

[8] S. Ji, W. Xu, M. Yang, and K. Yu, '3D Convolutional Neural Networks for Human Action Recognition', IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 1, pp. 221–231, Jan. 2013.

[9] Y. Xu, J. Yang, H. Cao, K. Mao, J. Yin, and S. See, 'ARID: A New Dataset for Recognizing Action in the Dark'. arXiv, Aug. 19, 2022.

[10] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, 'HMDB: A large video database for human motion recognition', in 2011 International Conference on Computer Vision, Nov. 2011, pp. 2556–2563.

[11] K. Soomro, A. R. Zamir, and M. Shah, 'UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild'. arXiv, Dec. 03, 2012.

[12] F. C. Heilbron, V. Escorcia, B. Ghanem, and J. C. Niebles, 'ActivityNet: A large-scale video benchmark for human activity understanding', in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2015, pp. 961–970.

[13] Carreira and A. Zisserman, 'Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset', in 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Jul. 2017, pp. 4724–4733.

[14] K. Hara, H. Kataoka, and Y. Satoh, 'Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?' arXiv, Apr. 01, 2018.

[15] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, 'Aggregated Residual Transformations for Deep Neural Networks'. arXiv, Apr. 10, 2017.

[16] S. Hira, R. Das, A. Modi, and D. Pakhomov, 'Delta Sampling R-BERT for limited data and low-light action recognition'. arXiv, Jul. 12, 2021.