Two-Streams: Dark and Light Networks with Graph Convolution for Action Recognition from Dark Videos (Student Abstract)

Saurabh Suman, Nilay Naharas, Badri Narayan Subudhi, and Vinit Jakhetiya

Indian Institute of Technology Jammu
NH44, Nagrota, Jagti, Jammu and Kashmir, India-181221
2019uee0132@iitjammu.ac.in, 2019uee0112@iitjammu.ac.in, subudhi.badri@iitjammu.ac.in, vinit.jakhetiya@iitjammu.ac.in

Abstract

In this article, we propose a two-stream action recognition technique for recognizing human actions from dark videos. The proposed action recognition network consists of an image enhancement network with Self-Calibrated Illumination (SCI) module, followed by a two-stream action recognition network. We have used $R(2+1)D$ as a feature extractor for both streams with shared weights. Graph Convolutional Network (GCN), a temporal graph encoder is utilized to enhance the obtained features which are then further fed to a classification head to recognize the actions in a video. The experimental results are presented on the recent benchmark "ARID" dark-video database.

Introduction

Action recognition in dark-light environments is an important task in computer vision. It has a wide range of applications, including military surveillance systems, night security, autonomous robots, mobile robots, etc. Despite an increase in research interest in video processing tasks in the dark environment, an adequate study in this area is still lacking. The lack of research for action recognition in low-light situations could be attributed to a lack of low-light datasets and appropriate image enhancement approaches for enhancing low-light data. Although it is difficult to recognize actions in dark video data, it would be very advantageous to use effective models for dark video action identification. Several pioneer works have been developed to solve the action recognition problem in dark-light conditions. A dark video dataset, action recognition in dark (ARID) dataset has been developed by (Xu et al. 2021).

A delta sampling approach has been introduced by (Hira et al. 2021), which makes the use of a Zero-DCE module to enhance the low-light video data and a combination of $R(2+1)D$ and BERT networks are deployed to classify the actions. An action recognition called the DarkLight Networks is proposed by (Chen et al. 2021) where the Gamma intensity correction (GIC) is used to enhance the dark-light frames sampled from the video and a self-attention mechanism is used for the action classification.

In this abstract, we propose a novel human action recognition architecture for dark videos. The proposed scheme used a two-stream network which consists of a dark stream and a light stream. The dark video frames are directly fed to the dark stream while the light stream uses Self-Calibrated Illumination (SCI) (Ma et al. 2022) to enhance the low-light frames. The dark and low light frames are fed to the weight-shared $R(2+1)D$ features extractor. The extracted features from the dark and low-light networks are complementary in nature. Further, these complementary features are concatenated and then fed into a three-layer Graph Convolutional Network (GCN) to enhance and smooth the obtained features. After that, the classification head is used to recognize the actions in the video.

We test the suggested method using the "ARID" dark video dataset. Top-1 and Top-5 accuracy have been employed as assessment metrics to demonstrate the proposed method's performance. A total of fifteen state-of-the-art (SOTA) techniques are used to evaluate the performance of the proposed scheme, and the results corroborate our findings.

Proposed Methodology

It may be observed that many SOTA algorithms use CNN architectures for action recognition in dark videos. However, most of the SOTA techniques are ineffective in characterizing the spatio-temporal information in low-light videos for action recognition. As a result, in this work, we have utilized the capabilities of the $R(2+1)D$ and a temporal graph encoder for recognizing actions in low-light videos. In this paper, we have proposed a novel two-stream action recognition network for dark videos. The two streams network consists of a dark stream and a light stream. The dark stream uses dark input video frames. Similarly, the light stream uses the enhanced version of the dark dark frame using the Self-Calibrated Illumination (SCI) (Ma et al. 2022) approach. Further, we have used the pre-trained $R(2+1)D$ network to extract the deep features from the low-light video frames to enhance them. The deep extracted features are complimentary in nature and characterize the video frame with a deeper level. Both the $R(2+1)D$ networks operate in weight sharing mechanism. The reason behind choosing $R(2+1)D$ as a feature extractor is that, $R(2+1)D$ factors the 3D convolutional filters into discrete spatial and temporal components which provides considerable accuracy gains.

The dark stream in the proposed two-stream model extracts deep textural features which may be disrupted by an

image enhancement module or light network. The image enhancement technique SCI improves the brightness of dark-light images. Self-Calibrated Illumination (SCI) is a learning framework for robustly illuminating dark frames in low-light environments. It sets up a cascaded illumination learning mechanism with weight-sharing connections to address the objectives. It builds a self-calibrated module to avoid the computing weight of the cascaded pattern, which concedes the convergence between results at each level and produces benefits that only need a single basic block for inference, significantly reducing computation costs. It then improved the model's ability to adapt to generic scenarios using unsupervised training loss.

Further, the three-layer GCN network (a feature smoother) fuses and selects enhanced spatio-temporal information from the two-streams: Dark and Light. The classification head is utilized to recognize the actions in the video.

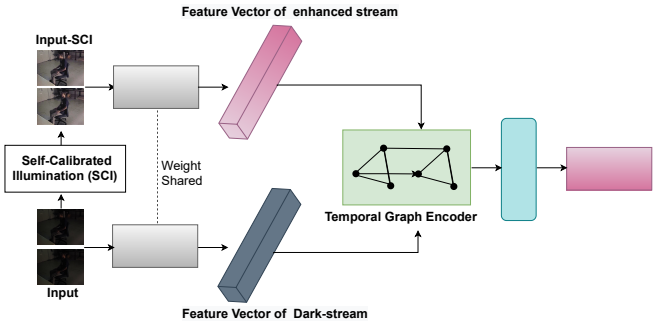

Figure 1: Proposed two-stream action recognition network

Implementation details The input frames of dimension $3 \times 64 \times 112 \times 112$ are fed directly into the feature extractor $R(2+1)D$ used as a dark-stream as well as to SCI and $R(2+1)D$, which forms the light-stream. Both streams provides the output of dimension $512 \times 8 \times 7 \times 7$. After the average pooling is applied on both streams, which gives the features of dimension 512×8 and then transposed to get the features of size 8×512 in each stream. Then two stream features are concatenated to a dimension of 16×512, which is further fed to the temporal graph encoder. The three-layer GCN provides the output of dimension 16×256. After applying the adaptive average pooling, which outputs the features of dimension 256. This is fed to a classification head to classify the action. The ADAMW optimizer with a learning rate of 10^{-5} is utilized for training in the proposed scheme. We have used the same action class setting as the other fifteen SOTA techniques to evaluate the result.

Experimental Results

We have tested the proposed technique on the ARID database. The effectiveness of the proposed technique is verified by comparing the results obtained with those of the fifteen SOTA techniques. The effectiveness of the proposed action recognition network is verified using two evaluation measures: Top-1 and Top-5 accuracies. Table 1 shows the performance comparison of the proposed method using Top-1 and Top-5 accuracy without replicating the results of considered SOTA techniques. It may be observed that the proposed technique provides better results than all fifteen SOTA techniques. In the proposed method, the use of GCN has reduced the number of parameters by more than one million in comparison to (Chen et al. 2021). The references for all the considered techniques used for comparison are available at (SupplementaryMaterials 2022).

Models	Top-1 Accuracy	Top-5 Accuracy
VGG-TS	32.08%	90.76%
TSN	57.96%	94.17%
C3D	40.34%	94.17%
Separable-3D	42.16%	93.44%
3D-ShuffleNet	44.35%	93.44%
3D-SqueezeNet	50.18%	94.17%
3D-ResNet-18	54.68%	96.60%
I3D-RGB	68.29%	97.69%
3D-ResNet-50	71.08%	99.39%
3D-ResNet-101	71.57%	99.03%
Pseudo-3D-199	71.93%	98.66%
I3D Two-stream	72.78%	99.39%
3D-ResNext-101	74.73%	98.54%
DarkLight-ResNeXt-101	87.27%	99.47%
DarkLight-R(2+1)D-34	94.04%	99.87%
Proposed Method	**95.86%**	**99.87%**

Table 1: The Top-1 and Top-5 accuracy results on ARID V1.0 of a few competitive models and our.

Conclusions

We proposed a two-stream action recognition network model for recognizing the actions from low light or dark videos. The use of complementary deep dark and light features in a graph convolutional networks improves the efficiency of the network. This is verified by comparing it against fifteen SOTA techniques. To improve its accuracy, further tuning of the parameters is in progress.

References

Chen, R.; Chen, J.; Liang, Z.; Gao, H.; and Lin, S. 2021. DarkLight Networks for Action Recognition in the Dark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 846–852.

Hira, S.; Das, R.; Modi, A.; and Pakhomov, D. 2021. Delta Sampling R-BERT for limited data and low-light action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 853–862.

Ma, L.; Ma, T.; Liu, R.; Fan, X.; and Luo, Z. 2022. Toward Fast, Flexible, and Robust Low-Light Image Enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5637–5646.

SupplementaryMaterials. 2022. :[Online] Available. https://github.com/cdshmnsh/Supplementary-material-for-AAAI-23. Accessed: 16/09/2022.

Xu, Y.; Yang, J.; Cao, H.; Mao, K.; Yin, J.; and See, S. 2021. Arid: A new dataset for recognizing action in the dark. In *International Workshop on Deep Learning for Human Activity Recognition*, 70–84. Springer.