electronics

MDPI

Article

Static Video Compression's Influence on Neural Network Performance

Vishnu Sai Sankeerth Gowrisetty * and Anil Fernando

Department of Computer & Information Sciences, University of Strathclyde, Glasgow G1 1XQ, UK
* Correspondence: vishnu.gowrisetty@strath.ac.uk

Abstract: The concept of action recognition in smart security heavily relies on deep learning and artificial intelligence to make predictions about actions of humans. To draw appropriate conclusions from these hypotheses, a large amount of information is required. The data in question are often a video feed, and there is a direct relationship between increased data volume and more-precise decision-making. We seek to determine how far a static video can be compressed before the neural network's capacity to predict the action in the video is lost. To find this, videos are compressed by lowering the bitrate using FFMPEG. In parallel, a convolutional neural network model is trained to recognise action in the videos and is tested on the compressed videos until the neural network fails to predict the action observed in the videos. The results reveal that bitrate compression has no linear relationship with neural network performance.

Keywords: bitrate compression; convolutional neural network; action recognition

Citation: Gowrisetty, V.S.S.; Fernando, A. Static Video Compression's Influence on Neural Network Performance. *Electronics* **2023**, *12*, 8. https://doi.org/10.3390/electronics12010008

Academic Editors: Hüseyin Kusetogullari, Turgay Celik, Chafik Samir and Amir Yavariabdi

Received: 1 August 2022
Revised: 1 December 2022
Accepted: 10 December 2022
Published: 20 December 2022

1. Introduction

With the rapid advancement of technology, envisioning a future filled with smart gadgets is becoming more feasible. Smart security cameras are wire-free cameras that can do more than just record video and take pictures. These gadgets include added features that allow you to check on the condition of your home even if you are hundreds of miles away. These devices provide you with the assurance that your home is secure and safe from intruders. The detection of human actions from real-time CCTV video data streams is a major subject. It is very useful for video surveillance and anomaly detection.

1.1. Motivation

In the domain of security, there are two conflicting criteria for smart monitoring. The first thing that is necessary is high-quality video. Having visual representations of a problem is useless if you cannot get valuable information from it. The capacity to recognize persons or interpret descriptions of objects can be decisive in determining whether a video includes important information. The overall graphical fidelity of videos is affected by their resolution. The better the resolution, the clearer the video. Video file size is also affected by resolution. As a result, a high-definition video has a larger file size than a standard-definition video of the same runtime. The availability of data storage is the second necessity for smart security monitoring. Following the completion of the recording, the video must be saved, which requires both memory and processing power if it is to be broadcast. This memory and processing capability necessitates a substantial financial commitment. As a result, the size of a video file accounts for a significant portion of the entire cost of any security system.

Video compression could be beneficial in fixing this issue. Parts of the data that are deemed repetitious are removed using a video codec, reducing the size of the video. However, because video compression results in data loss, some of the lost data may be critical, as a deep learning neural network may rely on the missing data. We define critical data as data that have a statistical link with poor deep learning neural network prediction

performance if it they are omitted or altered. As a result, it is critical to pick a video compression format that strikes the optimal balance between removing unnecessary data and retaining crucial data. In addition, standard video compression standards are intended for human video consumption. Most videos are currently processed by machines, and human consumption is linked to validating machine analysis results [1]. A new video codec permits lower bitrates without compromising machine analysis performance.

1.2. Research Purpose

Humans watch and evaluate video in a very different way than machines. Disparities in colour, shape, and size of things are influenced by the objects' distance, lighting, familiarity, and image resolution. The impact of compression on human perception of digital information can be measured using metrics such as Structural Similarity Index Metric (SSIM) [2,3] and Peak Signal-to-Noise Ratio (PSNR) [3]. The goal of this research is to identify the maximum bitrate compression that a video may go through for a neural network to still be able to predict the action in the film.

This is accomplished by examining the outputs of a Convolutional Dense Neural Network (5C3DNN) tested on several different bitrate versions of video sequences from the i3DPost Multi-view Human Action Dataset [4]. The 5C3DNN is trained on the HMDB51 Dataset [5].

2. Related Work

From previous research, it can be assumed that the major adverse effects on neural network prediction probability appear to be caused by several types of blur and noise. To test the performance of a new capsule-based architecture [6] on data that include intrinsic noise, an image classification architecture based on the clustering of neurons into capsules and a new dynamic routing protocol was explored. In order to accomplish this, the efficiencies of six widely used Convolutional Neural Networks were reviewed on two distinct datasets with varying image-quality distortions. It was demonstrated that high classification accuracy is not achievable when a dataset comprises a large number of classes, and the new capsule architecture is resistant to specific image loss.

The effects of $\times$264 video encoding optimized for human vision on neural network performance in the realm of autonomous vehicles was analysed to potentially demonstrate the relevance of $\times$264 video encoding as a method for decreasing the volume of data required for the sufficient performance of machine learning algorithms used for autonomous driving while preserving human perception levels of the data [7].

A range of network architectures and application domains, including end-to-end convolution, encoder–decoder, region-based CNN (R-CNN), dual-stream, and generative adversarial networks (GAN), have been examined to demonstrate a nonlinear and nonuniform link between network performance and the amount of lossy compression [8]. Notably, performance degrades drastically below a JPEG quality (quantization) level of 15% and an H.264 Constant Rate Factor (CRF) of 40%. In certain instances, however, retraining these architectures on pre-compressed imagery recovers network performance. In addition, there is a correlation between architectures employing an encoder–decoder pipeline and lossy image compression resistance.

Training the models [9] on uncompressed images increases their robustness when evaluated on compressed data; moderate image compression has no effect on the classification performance of deep learning-based models. On average, an image can be compressed by a quality factor of 10 for a JPEG encoder and 20 for a MozJPEG encoder while preserving its classification accuracy [10].

3. Methodology

3.1. Bitrate-Compression

The FFmpeg [11] tool is used to compress our dataset in terms of bitrate. We pick FFmpeg to compress our datasets since it is a high-performance, industry-leading, open-

source multimedia framework for encoding, decoding, transcoding, and other tasks. The rate at which video data are conveyed in bits per second is referred to as video bitrate. When a video is expected to be compressed, a maximum bitrate that must not be exceeded must be sent to the video codec. For example, if the video codec is designed to compress the video at 1 mbps, the encoder will compress each second of the video so that the decoder only receives 1 mb of data per second.

There are several algorithms for finding and applying a bitrate value for video compression, such as:

Constant Bitrate (CBR) [12]: video quality is sacrificed in order to keep a constant bitrate.

Variable Bitrate (VBR) [12]: maintains video quality while allowing the bitrate to fluctuate.

To understand how bitrate affects video quality, it is necessary to know how video compression works. When video is compressed, the compression technique uses Discrete Cosine Transform (DCT) [**?**] to transfer video from the pixel domain to the frequency domain, and Quantization [14] to reject a significant number of frequencies that the human eye cannot distinguish. When compressing a video, information is compromised in order to maintain the video's quality. A substantial amount of data are lost when compressing a video considerably, and the loss of data is obvious. If the video is not heavily compressed, the file size is huge, but the video quality is excellent. This is known as the rate–distortion trade-off in video compression. Assuming that the resolution is fixed, the lower the bitrate, the worse the video quality.

The i3DPost Multi-view Human Action Dataset is made up of consecutive png frames, and because FFmpeg can only encode video files, we first utilize it to compile this dataset into an mp4 file. As depicted in Figure 1, this video is fed into an H.264/HEVC [15] video encoder, which performs prediction, transformation, and encoding operations to produce a compressed bitstream. An H.264/HEVC video decoder completes the procedures of decoding, inverse transforming, and reconstruction to produce a decoded video sequence. This decoded video sequence is bitrate-altered before being evaluated with the 5C3DNN model.

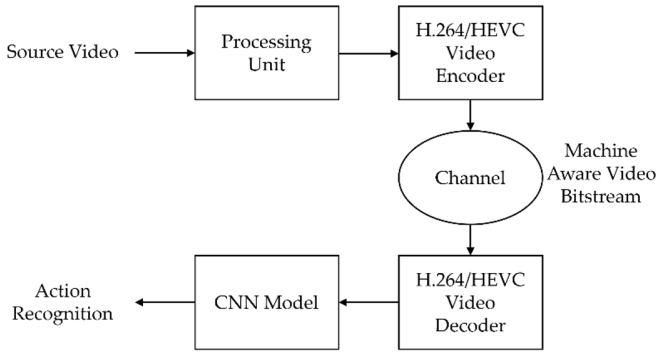

Figure 1. The framework.

3.2. Neural Network

A neural network classification model with five convolution layers is built. This is a basic classifier. A moving average technique [16] is employed. This method involves creating a function that generates a unique forecast for the entire video. This function takes 'n' frames from a video and predicts what will happen in each frame. Averaging the predictions of these 'n' frames yields the total score for the entire video. This is referred to as the Single-Frame CNN method.

The 5C3DNN model (Figure 2) is first trained on the HMDB51 dataset. The HMDB51 dataset includes videos from many sources, such as movies and web videos. The dataset includes 6849 video clips from 51 activity categories, each of which has at least 101 clips. This 5C3DNN model is an image classification model that operates on each and every frame of the video before averaging the individual probabilities to produce a final probabilities vector.

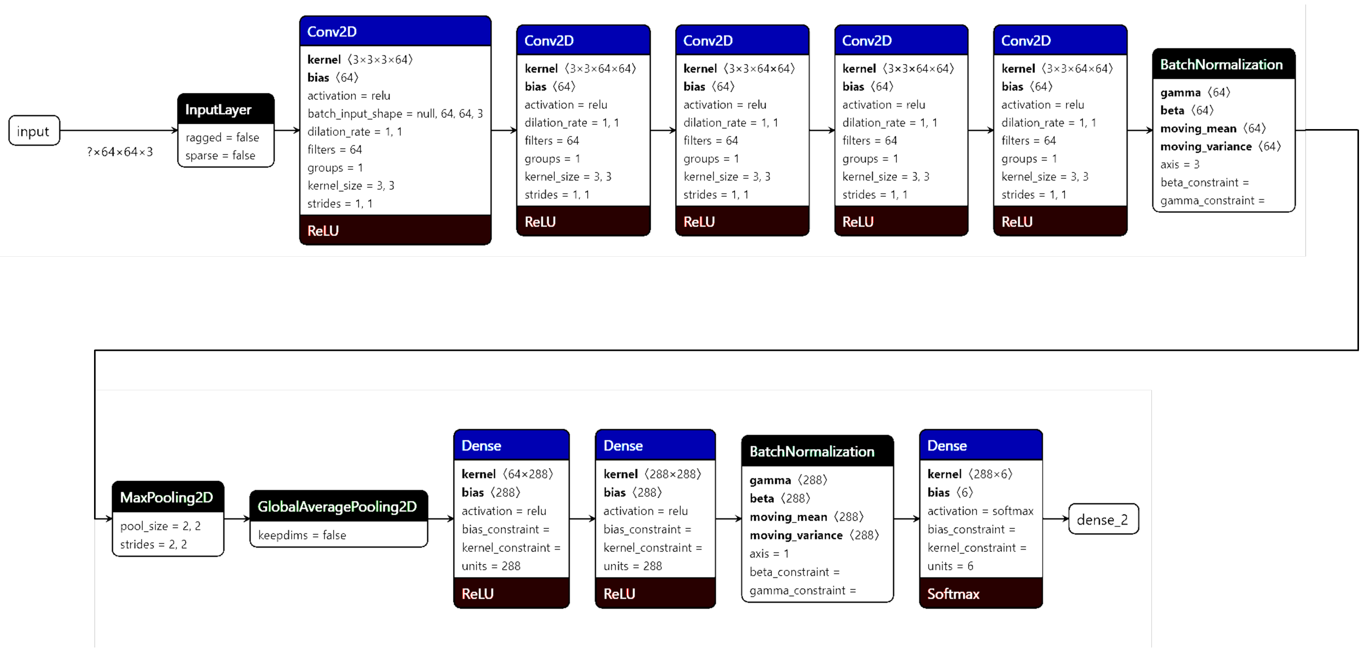

Figure 2. The 5C3DNN architecture.

Several constants are defined during the preparation of the dataset, including the frame's height and width (64 × 64) and the maximum number of frames collected from each action to train the model. The sizes of all of these frames are standardised, and this is now a feature. Each feature is assigned an action label. There are now two lists: features and labels. Labels are converted to hot-encoded vectors. Following the division of these lists into training and testing sets, the training data are fed into the 5C3DNN model. On test data, the model predicts with 96% accuracy after training.

3.3. *YUV*

YUV is a pixel format employed by video applications. YUV is the real identity of the colour space shared by all "YUV" pixel formats. In contrast to RGB (Red–Green–Blue) formats, YUV colours are represented with one "luminance" component, called Y (similar to grey scale), and two "chrominance" components, called U (blue projection) and V (red projection).

The YUV 4:2:0 format comprises luminance plane Y, followed by the U and V chrominance planes. The two chrominance planes (blue and red projections) are sub-sampled by a factor of two in both the horizontal and vertical dimensions. For a 2 × 2 square of pixels, there are four Y samples but only one U sample and one V sample. This format uses 48 bits per 4 pixels, or 12 bits per pixel, giving it a depth of 12 bits per pixel.

The sequential png frames from the i3DPost Multi-view Human Action Dataset are grouped and converted into YUV 4:2:0 format video, as it considered to be a raw dataform and is helpful for calculating the PNSR with respect to the YUV 4:2:0 format of bitrate-altered videos.

4. Results and Analysis

The i3DPost Multi-view Human Action Dataset has 125 sequential uncompressed frames with an aspect ratio of 1920 × 1080. The bitrate of the uncompressed video formed using these frames is calculated based on one uncompressed frame:

pixels per frame = 1920 * 1080 = 2,073,600 pixels.

In YUV420, there is one U and one V value per 2 × 2 group of Y (which means the two chroma components are sampled at half the sample rate of luma both horizontally and vertically.

Hence, a 2 × 2 block of an uncompressed frame has 4*4-bit Y values + one 4-bit U + 4-bit V. Therefore, for 2 × 2 pixel: 16 + 4 + 4 = 24 bits and 1 pixel: 6 bits.

Hence, bits per frame = 6 * 2,073,600 = 12,441,600 bits.

The uncompressed video runs at 30 frames per second (FPS). Therefore,

bit rate = 30 * size of one uncompressed frame = 30 * 12,441,600 = 373,248,000 bits/s.

The bitrate of each uncompressed video sequence that is formed from the uncompressed images in the i3DPost Multi-view Human Action Dataset is 373,248,000 bits/s.

The trained 5C3DNN model is further tested on bitrate-altered videos to find out the extent to which we can decrease the bitrate of the video so that the neural network model fails to predict the action. For this, eight video sequences with bitrates alerted at variable intervals are fed into the model and accuracies are noted. The following graph shows the results obtained for each video sequence.

The results are consistent and maintain a non-linear relationship with respect to bitrate compression. The neural network appears to perform really well on video sequences that are too compressed for human consumption. From Figure 3, we can observe that in the range of 25 Kbps–35 Kbps, the neural network is losing its ability to predict the action in the video.

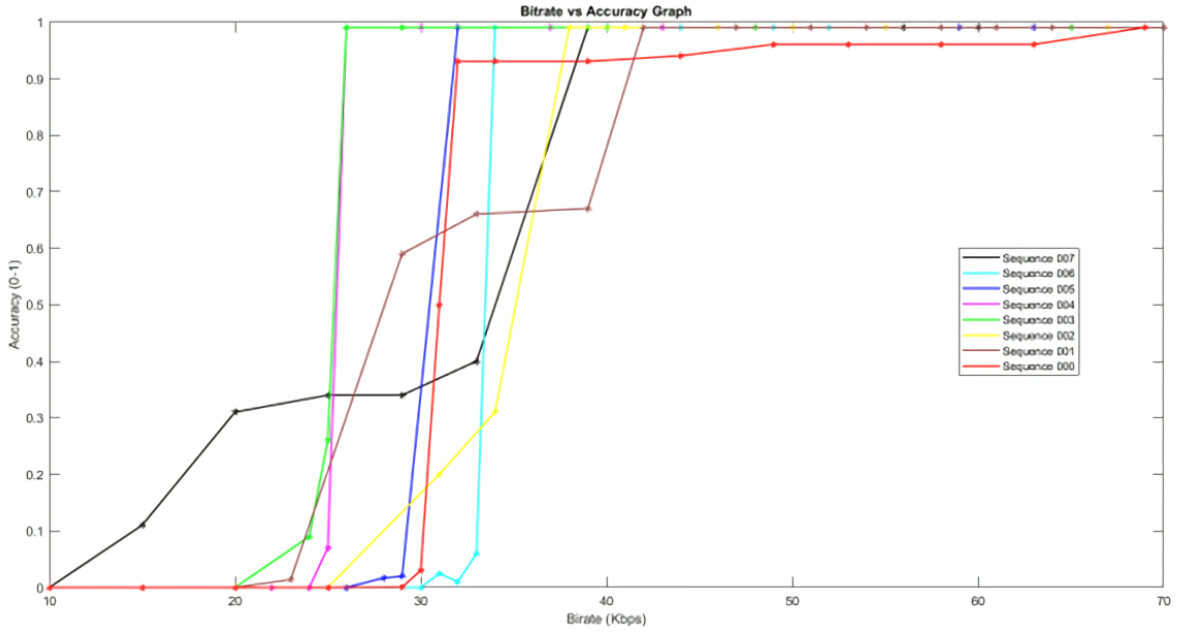

Figure 3. Bitrate vs. accuracy of video sequences.

To calculate the PSNR value between the uncompressed video and the bitrate-altered videos, a MATLAB function 'yuvpsnr' is used. The bitrate-altered videos are initially converted from mp4 to YUV420 format, as the function yields solutions only between two YUV format videos.

Both of the Bitrate vs. PSNR graphs (Figures 4 and 5) portray the same fall of the quality with bitrate reduction that was depicted in the Bitrate vs. Accuracy graph (Figure 3) earlier.

The Bitrate vs SSIM graphs (Figure 6) prove the fall of quality with respect to the bitrate shown in Figure 3.

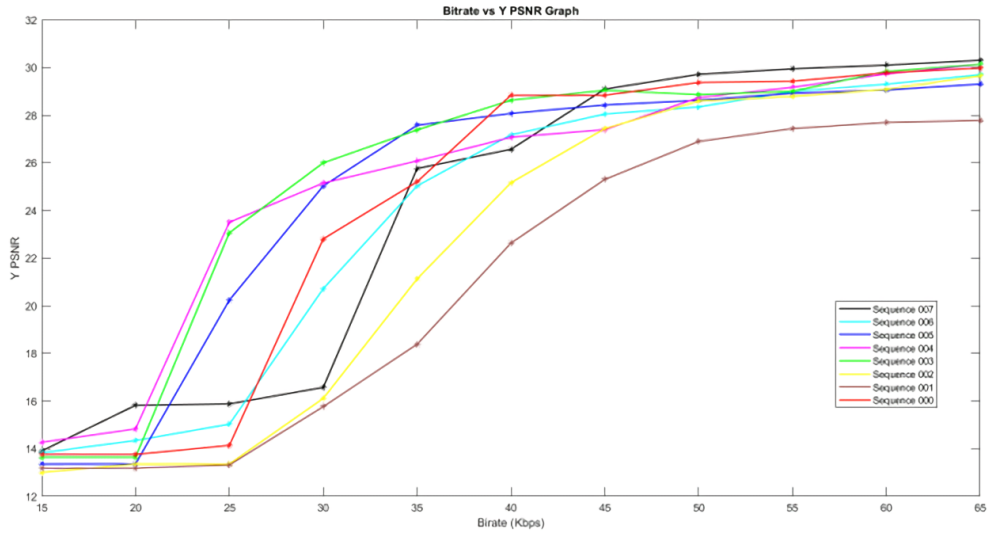

Figure 4. Bitrate vs. Y PSNR of video sequences.

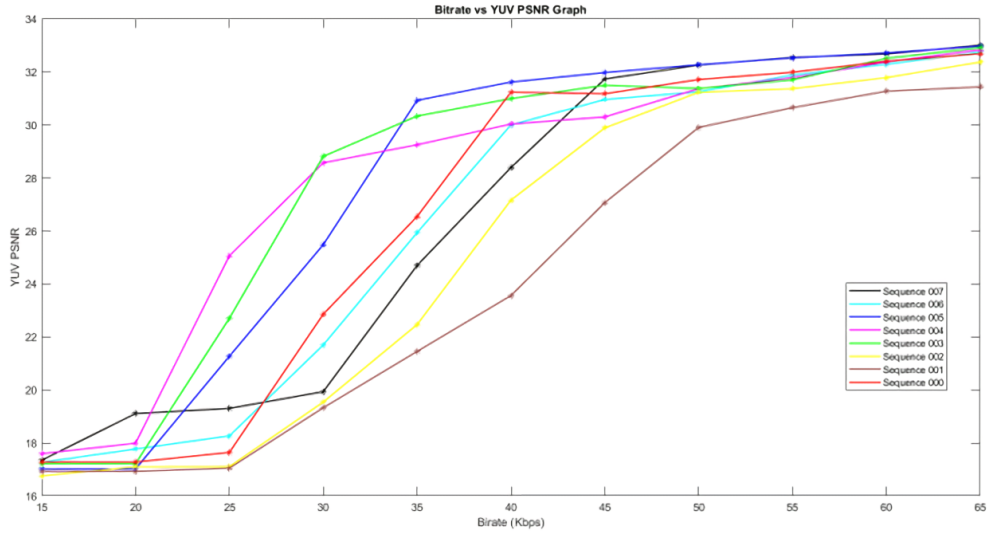

Figure 5. Bitrate vs. YUV PSNR of video sequences.

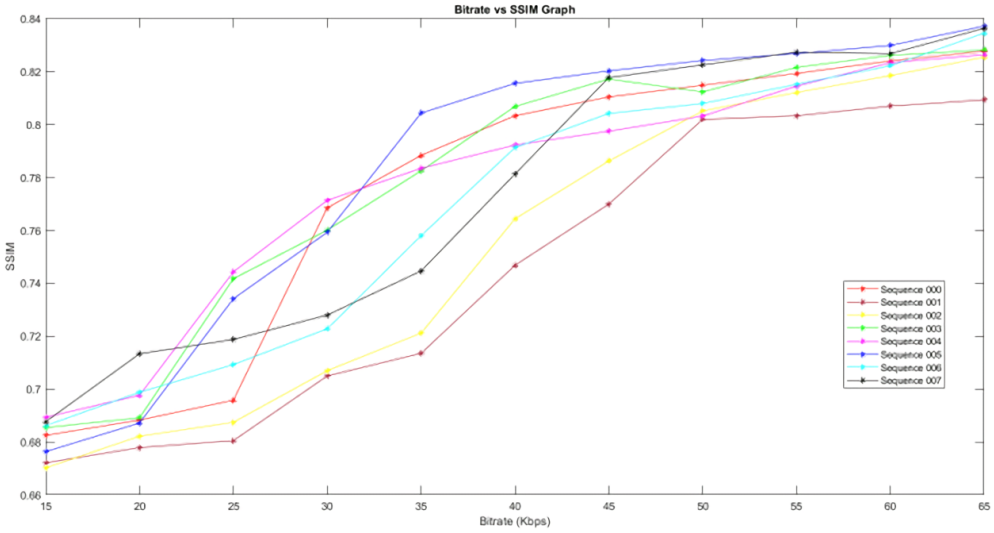

Figure 6. Bitrate vs. SSIM of video sequences.

The follwing quartet of Figures 7–10 portray the dissipation of the data with respect to the reduction of the bitrate and thus losing the accuracy in predicting the activity at 35 Kbps. We can see that model is performing well even at 40 Kbps bitrate, where human consumption of the video is not possible.

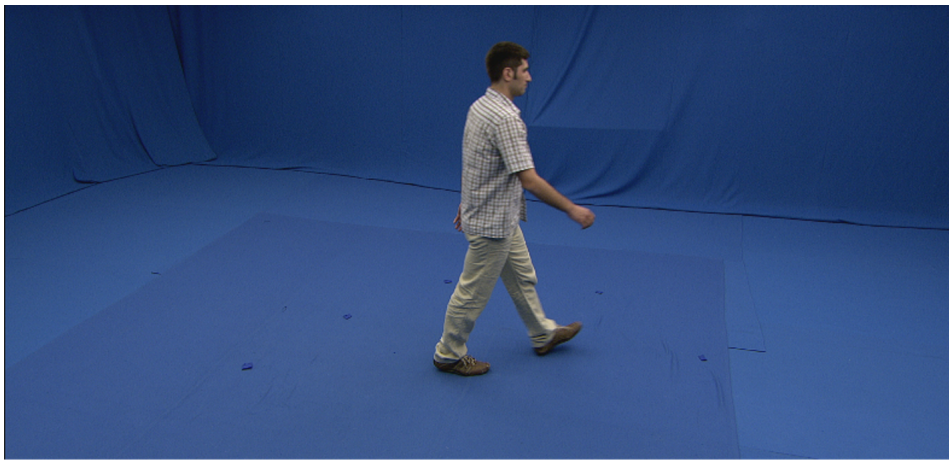

Figure 7. Uncompressed frame: Accuracy = 0.99.

Figure 8. Frame of a video with a bitrate of 40 Kbps: Accuracy = 0.99.

Figure 9. Frame of a video with a bitrate of 70 Kbps: Accuracy = 0.99.

Figure 10. Frame of a video with a bitrate of 35 Kbps: Accuracy = 0.40.

The Table 1 gives the approximate compression rates of the video sequences at the breakeven points.

Table 1. Compression ratio at breakeven point of each video sequence.

Sequence No.	Bitrate at Breakeven Point (Kbps)	Compression Ratio at Breakeven Point
000	31	12,040.25
001	28	13,330.28
002	34	10,977.88
003	25	14,929.92
004	25	14,929.92
005	29	12,870.62
006	33	11,310.54
007	33	11,310.54

The neural network model presented in this paper is trained and tested on only six classes of the HMBD51 Dataset ("climb_stairs", "fall_floor", "run", "sit","stand", and "walk"). To evaluate the performance of this model, it is compared with proven neural network architectures in the Table 2.

Table 2. Comparison between architectures.

Model	Accuracy on HMDB51 Test Data (5 Classes)	Average Bitrate at Breakeven Point (Kbps) (i3DPost)
Rank pooling [17]	82.74%	39.82
SP-CNN [18]	85.27%	35.61
BERT, 3D CNN [19]	95.33%	31.93
5C3DNN [Ours]	96.96%	29.75

5. Conclusions

In this research work, we empirically assessed the effect of bitrate reduction on the performance of neural networks for action recognition. In doing so, we hoped to demonstrate the applicability of video compression as a means of decreasing the quantity of data required for the proper performance of deep learning algorithms used for action detection while preserving human perception levels of said data.

We generated a dataset from raw uncompressed images and evaluated the bitrate-altered movies inferred from eight video sequences in that dataset on a 5C3DNN model trained on a different activity dataset. Following this, we drew two comparisons. First, we examined the accuracy of bitrate-modified videos and determined the range of bitrates

at which the neural network loses its capacity to predict activity. Next, we compared the PSNR values of the YUV version of bitrate-altered videos to those of the uncompressed raw videos to see if they complemented the prior findings.

Important findings from our studies include that the range of 25 Kbps to 35 Kbps contains the breakeven bitrate points at which the neural network could not predict. This range of videos has a compression ratio between 10,977.88 and 14,929.92. It was discovered that there is no linear relationship between bitrate compression and 5C3DNN performance. Later, the model was compared to different architectures and proved its capability to perform better than them at lower bitrates.

The datasets reported in this work are static; hence, additional work on dynamic videos is necessary to strengthen the outcomes of this experiment. Future research on this subject would benefit from the inclusion of even more-diverse datasets and ones that include other luminance settings. Future research on this topic would also benefit from testing the effectiveness of another neural network model created and trained to meet a different application domain other than Action Recognition.

Author Contributions: Methodology, V.S.S.G.; Software, V.S.S.G.; Formal analysis, V.S.S.G.; Investigation, V.S.S.G. and A.F.; Resources, V.S.S.G.; Data curation, V.S.S.G.; Writing—original draft, V.S.S.G.; Writing—review and editing, V.S.S.G.; Visualization, V.S.S.G.; Supervision, A.F.; Project administration, A.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: HMDB51: https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/ (accessed on 1 August 2022). i3DPost Multi-view Human Action Dataset: https://kahlan.eps.surrey.ac.uk/i3dpostaction/ (accessed on 1 August 2022) (Available on request).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Duan, L.; Liu, J.; Yang, W.; Huang, T.; Gao, W. Video Coding for Machines: A Paradigm of Collaborative Compression and Intelligent Analytics. *arXiv* **2020**, arXiv:2001.03569.
2. Wang, Z.; Bovik, A.; Sheikh, H.; Simoncelli, E. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, **13**, 600–612. [CrossRef] [PubMed]
3. Horé, A.; Ziou, D. Image quality metrics: PSNR vs. SSIM. In Proceedings of the 20th International Conference on Pattern Recognition, ICPR 2010, Istanbul, Turkey, 23–26 August 2010; pp. 2366—2369. [CrossRef]
4. Gkalelis, N.; Kim, H.; Hilton, A.; Nikolaidis, N.; Pitas, I. The i3DPost Multi-View and 3D Human Action/Interaction Database. In Proceedings of the 2009 Conference For Visual Media Production, London, UK, 12–13 November 2009; pp. 159–168.
5. Kuehne, H.; Jhuang, H.; Garrote, E.; Poggio, T.; Serre, T. HMDB: A large video database for human motion recognition. In Proceedings of the International Conference On Computer Vision (ICCV), Barcelona, Spain, 6–13 November 2011.
6. Roy, P.; Ghosh, S.; Bhattacharya, S.; Pal, U. Effects of Degradations on Deep Neural Network Architectures. *arXiv* **2018**, arXiv:1807.10108.
7. Stanoevich, M.; Partain, J. Effects of Video Compression formats on Neural Network Performance. 2019. Available online: http://hdl.handle.net/2077/62552 (accessed on 1 August 2022).
8. Poyser, M.; Abarghouei, A.; Breckon, T. On the Impact of Lossy Image and Video Compression on the Performance of Deep Convolutional Neural Network Architectures. *arXiv* **2020**, arXiv:2007.14314
9. Jo, Y.; Choi, Y.; Park, H.; Lee, J.; Jung, H.; Kim, H.; Ko, K.; Lee, C.; Cha, H.; Hwangbo, Y. Impact of image compression on deep learning-based mammogram classification. *Sci. Rep.* **2021**, *11*, 7924. [CrossRef] [PubMed]
10. Benbarrad, T.; Salhaoui, M.; Anas, H.; Arioua, M. Impact of Standard Image Compression on the Performance of Image Classification with Deep Learning. *Innov. Smart Cities Appl.* **2022**, *5*, 901–911.
11. Tomar, S. Converting video formats with FFmpeg. *Linux J.* **2006**, *146*, 10.
12. Mohsenian, N.; Rajagopalan, R.; Gonzales, C. Single-pass constant- and variable-bit-rate MPEG-2 video compression. *IBM J. Res. Dev.* **1999**, **43**, 489–509. [CrossRef]
13. Strang, G. The Discrete Cosine Transform. *SIAM Rev.* **1999**, **41**, 135–147. [CrossRef]
14. Ropert, M.; Le Tanou, J.; Blestel, M. Mastering Quantization is key for Video Compression. *Smpte Motion Imaging J.* **2022**, *131*, 45–53. [CrossRef]
15. Richardson, I. *The H.264 Advanced Video Compression Standard*; Wiley Publishing: Hoboken, NJ, USA, 2010.

16. Alzubaidi, L.; Zhang, J.; Humaidi, A.; Al-Dujaili, A.; Duan, Y.; Al-Shamma, O.; Santamaría, J.; Fadhel, M.; Al-Amidie, M.; Farhan, L. Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J. Big Data* **2021**, *8*, 53. [CrossRef] [PubMed]
17. Fern, O.B.; Gavves, E.; Oramas, M.J.; Ghodrati, A.; Tuytelaars, T. Rank Pooling for Action Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 773–787. [CrossRef] [PubMed]
18. Yu, S.; Cheng, Y.; Su, S.; Cai, G.; Li, S. Stratified pooling based deep convolutional neural networks for human action recognition. *Multimed. Tools Appl.* **2017**, *76*, 13367–13382. [CrossRef]
19. Kalfaoglu, M.; Kalkan, S.; Alatan, A. Late Temporal Modeling in 3D CNN Architectures with BERT for Action Recognition. In *Computer Vision—ECCV 2020 Workshops*; Springer: Cham, Switzerland, 2020; pp. 731–747.