

Video Compression by Eliminating Irrelevant Frames

Miss. Kadambari Ramesh Kowale¹, Prof. Mr. Hirendra R. Hajare²

M-Tech Final Year Student, Computer Science And Engineering, Ballarpur Institute of Technology, Ballarpur, Maharashtra, India¹

Assistant professor and HOD, Computer Science And Engineering, Ballarpur Institute of Technology, Ballarpur, Maharashtra, India²

Abstract-

Regarding the security of the ATM systems situated at various locations throughout the world, constant monitoring via surveillance at the ATM machine is required. Even though ATMs are equipped with manual protection, several theft events have made headlines over the past few decades. Since continuous surveillance provides an alternative to manual security, it is not feasible for the banks to do so at every one of these ATMs. With so much data generated by the monitoring system inside and outside the ATM, it becomes expensive for the banks to manage in terms of memory. As a result, the goal of this job is to reduce the number of frames in the video that was collected by removing all of the pointless frames. The technique employs transfer learning on a Mask RCNN Network to identify frames containing people before removing any such frames. Mask RCNN network, which is typically trained on the COCO dataset, will be used to detect humans in the so-called relevant frames and subsequently to filter out the frames with no humans. The pertinent frames are after that turned into video and kept in the backup. The plan decreases the amount of storage needed for continuous monitoring.

Keywords- Region-Based Convolutional Neural Network (RCNN), Convolutional Neural Networks (CNN), Closed Circuit Television(CCTV), Discrete cosine transform (DCT),

GPU (Graphics Processing Unit), Deep Learning techniques, area classification,

INTRODUCTION

Since the last several decades, there has been a sharp rise in the usage of CCTV (closed circuit television), sometimes known as video surveillance. Numerous situations, both indoors and outdoors, might benefit from video monitoring. In order to provide safety and security, surveillance cameras are becoming an essential component of daily life. In the Indian government's development programs like Digital India, one of the main priorities is e-surveillance. It will still have video surveillance. Video surveillance has several benefits, including effective monitoring, reduced labor needs, affordable auditing capabilities, and the capacity to adapt new security measures, among others.

The use of artificial intelligence, machine learning, and deep learning in video surveillance is a new field. The use of artificial intelligence enables computers to think like people. Making predictions based on future data and learning from training data are crucial aspects of machine learning. Deep learning is employed since there are now GPU (Graphics Processing Unit)

processors and large datasets accessible. One of the most effective designs for tackling challenging learning challenges is deep neural networks. Deep Learning models automatically extract characteristics from picture input and provide high-level representations. This is more general since the feature extraction procedure is fully automated. Directly from the pixels of an image, convolutional neural networks (CNNs) may learn visual patterns. Therefore, we suggest a fresh strategy based on Deep Learning techniques.

Installation of CCTV is necessary in a number of industries, including banks, ATMs, stores and multiplexes, city roads and highways, government offices and buildings, classrooms, universities, etc. With increasing spatial resolutions, frame rates, and dynamic ranges being used as a result, video service providers have been expanding the video parameter space, which has significantly increased the amount of storage space needed. Video surveillance storage is a problem that many industries are dealing with as the quality of surveillance footage improves. Consequently, there is an increasing need for video storage. It is foolish to believe that simply increasing memory size will fix the issue. Thus, this surveillance film was wiped after a certain amount of time, such as six months for ATM security footage, one month for classroom surveillance, and three months for office surveillance footage, which resulted in the loss of pertinent data as well. So, to save pertinent data with the least amount of space, a domain-specific compression algorithm is required.

Discrete cosine transform (DCT) and block-based motion estimation are two hand-crafted modules that have been popular in recent years in video compression techniques to cut down on redundant information in video sequences. The fact that they are hard-coded prevents them from keeping up with the diversity and increasing demand for video use cases, such as object identification and VR broadcasting, despite the fact that they have been painstakingly built and fine-tuned. Numerous DNN-

based techniques have been developed for intra-prediction, residual coding, mode choice, entropy coding, and post processing for the video compression challenge. In typical video compression algorithms, these methods are employed to increase the effectiveness of a particular module.

A relevance-based compression strategy with two modules was proposed in our study effort. 1) Using a neural network for object detection, relevance-based object detection 2) Content adaptive video compression, which eliminates pointless frames. ATM surveillance video with brief periods of motion is the application case taken into consideration for this study endeavor. The ATM room's CCTV system records continuous video monitoring. However, it has been noted that the security footage only shows the stable object (an ATM machine) for the remainder hours and just a small portion of the transaction activity that took place in an ATM room. The majority of storage capacity is spent on non-active portions of surveillance film, which creates the crucial data storage problem in surveillance systems.

STATE OF ART

In [1] paper, compared to the preceding VOC 2012, R-CNN has a 30% increase in mean average precision (mAP). There are three modules in R-CNN: i) To create proposals, use selective search to produce sparse set proposals. ii) area classification—use a series of SVM classifiers; iii) feature extraction—extract fixed length feature vectors from each proposal. R-CNN, however, has two significant drawbacks linked to how time-consuming training and testing are, and how it relies on selective search to provide high quality recommendations.

The paper [2] R-CNN and SPP net have constraints in terms of training and testing time as well as detection accuracy. Fast R-CNN is an innovative solution that addresses these issues. Fast R-CNN is also known as a single-stage training method since it optimizes all training-related processes, including feature extraction, classification, and regression, from beginning to end. When it comes to detection accuracy, training efficiency, and inference speed, Fast R-CNN surpassed R-CNN and SPP-net. The generation of suggestions continued to be done using traditional techniques like Selective Search and Edge Boxes despite improvements in object identification.

In [3] paper, By eliminating the cropping and warping components, the SPP-net proposal improved the R-CNN approach. With the aid of a deep convolutional network, this model computes the features from the full picture and uses spatial pyramid pooling to extract fixed-length feature vectors onto a feature map. In SVM classifier and bounding box regression, these retrieved features are used as input. Although SPP-net outperforms R-CNN in terms of performance, it was unable to undergo full optimization owing to multistage training. Furthermore, because SPP did not do back propagation, all the parameters were locked.

In [4] paper Feature Pyramid Networks (FPN) have been developed to enable object recognition in feature maps at various sizes by mixing shallow layer features with deep layer features. In addition to being widely employed in video detection, FPN has made significant strides in the identification of multi-scale objects. The multiscale problem was fixed by

FPN. From a single-scale input, FPN employs a top-down architecture with lateral connections to construct an in-network feature pyramid.

In paper [5], established the obvious and natural notion of Mask-RCNN. Aligning pixels one to one is the main component of Mask R-CNN. In addition to the current branches for classification and bounding box regression, Mask R-CNN builds on the principles of Faster R-CNN by adding a branch for predicting segmentation masks for each Region of Interest (RoI). Faster R-CNN operates at 5 frames per second, while Mask R-CNN only incurs a little cost. Mask R-CNN is also simple to generalize to other applications, such as assessing human postures within the same environment. The score of a detection (or instance segmentation) hypothesis is based on the classification score component with the highest weight.

In paper [6], For "fully convolutional instance segmentation" (FCIS), a multi-stage cascade that combines the segment proposal system and object identification system has been suggested. This cascade predicts segment suggestions from bounding-box proposals in the first stage. A collection of position-sensitive output channels are to be totally convolutional predicted.

In paper [7], the system is quick thanks to these channels since they handle object classes, boxes, and masks simultaneously. Contrarily, FCIS exhibits systematic errors on overlapping instances and creates fictitious edges, demonstrating that it has trouble overcoming the basic challenges of segmenting instances.

PROPOSED WORK

The primary components of the suggested methodology are two modules: -

1. Content-adaptive video compression
2. Relevance-based object recognition

Content-adaptive video compression: As seen in fig. 1, the output of the network for relevance-based object recognition serves as the input for the network for content-adaptive video compression. The material is compressed using various quantization parameters (QP) allotted to various code tree units (CTU) by a compression module. both the CTU of the pertinent Frame (QPr) and the CTU of the irrelevant Frame (QPi). Relevant content is compressed using default QP, whereas unrelated content is either eliminated or compressed with a lesser quality using a higher QP.

Relevance-based object recognition: Video frames are classified as being inactive or being in action using relevance object detection. Two types of frames are recognized. 1) Idle frame (irrelevant frame) - refer to the no-action frame, i.e., the only stable object is an ATM machine 2) Action frame (relevant frame): This frame is used to categorize the pixels of an activity into ROI and Non-ROI. The suggested model will select the appropriate classification network to divide video surveillance frames into idle and active frames.

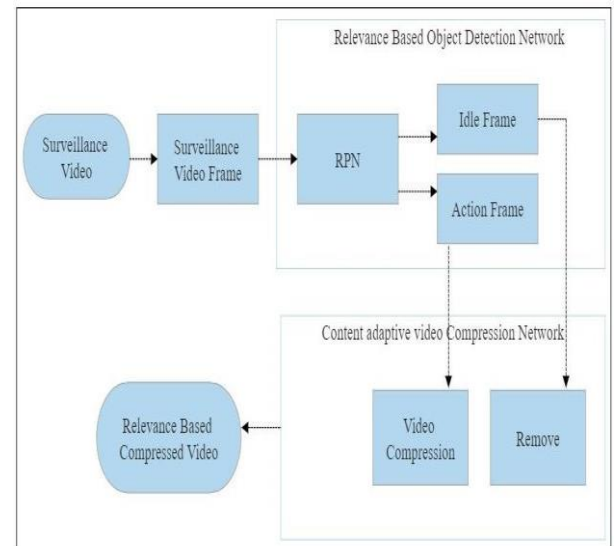


Fig. 1 Proposed System Architecture

METHODOLOGY

Mask RCNN:

On the foundation of Faster Rcn, Mask Rcn is an enhanced network. The RPN network is primarily fed by Faster RCN's entry of the final layer of the feature map that CNN has retrieved. The target feature is less visible in the original input image because it was too tiny, even if it is more noticeable in this feature map. Despite the fact that some basic characteristics can be recovered when they are fed into the RPN network, the convolution layer is likely to filter out tiny targets as noise in the midst of the hidden layer. In the subsequent creation of 256-dimensional vectors, a few Anchor targets are probably going to be seen as background. The issue of tiny targets detection must thus be resolved. Mask Rcn used the Feature Pyramid Network (FPN). In order to prevent filtering out tiny target features, FPN applies the same feature map as that derived by CNN to the RPN network. By expanding the residual structure, the "top-down"

feature map is combined with the initial convolution to improve the target's feature.

$$f(B_1) \approx \frac{x_2 - x}{x_2 - x_1} * f(A_{11}) + \frac{x - x_1}{x_2 - x_1} * f(A_{21})$$

$$f(B_2) \approx \frac{x_2 - x}{x_2 - x_1} * f(A_{12}) + \frac{x - x_1}{x_2 - x_1} * f(A_{22})$$

$$f(P) \approx \frac{y_2 - y}{y_2 - y_1} * f(B_1) + \frac{y - y_1}{y_2 - y_1} * f(B_2)$$

$$\begin{aligned} f(x, y) \approx & \frac{f(A_{11})}{(x_2 - x_1)(y_2 - y_1)} * (x_2 - x)(y_2 - y) \\ & + \frac{f(A_{21})}{(x_2 - x_1)(y_2 - y_1)} * (x - x_1)(y_2 - y) \\ & + \frac{f(A_{12})}{(x_2 - x_1)(y_2 - y_1)} * (x_2 - x)(y - y_1) \\ & + \frac{f(A_{22})}{(x_2 - x_1)(y_2 - y_1)} * (x - x_1)(y - y_1) \end{aligned}$$

RESULT



Screenshot 1: - Sample screenshot of idle frame



Screenshot 2: - Sample screenshot of action frame

The action frame in the accompanying screenshot is where the ATM machine and the person are both visible. These frames are referred to be action frames since they feature an item other than the ATM machine. In order to facilitate further compression, these frames are sometimes known as key frames or I-frames. Following input from the preceding module, the content

adaptive video compression module is in charge of removing the unnecessary frames and compressing the necessary frames using the default QP.

Deployment and comparison of Model: In order to demonstrate that the suggested work gets the best results when compared to the existing system, Relevance based Object identification model and Content adaptive video Compression model are merged by looking for appropriate tools.

Sr. No.		Parameters	Value
1	Original Video	Total number of frames in the video	18000
2		Number of frames considered (15:1)	1200
3		Size of the video	351 MB
4	Mask RCNN	Number of RELEVANT frames detected	403
5		Number of non-RELEVANT frames	797
6		Accuracy	84%
7		Size of compressed video	52 MB

Table 1 – Performance of Mask RCNN and YOLOV5s Networks

CONCLUSION

In this paper, we proposed a RCNN methodology to compress the video. By allowing Mask RCNN to recognize just those video frames when the ATM is utilized, the Deep Mask RCNN network, which can identify 80 items, may be used to identify the video frames with human presence. In order to simply remove the empty (human-free) frames, the weights controlled by the Mask RCNN utilizing the COCO dataset are able to recognize such pertinent frames. After stacking all the pertinent frames in the proper order, a compressed video may be created for storage and later use in applications involving risks and security. The Mask RCNN network may be used with transfer learning, and it also has the same ability as any other pre-trained network to be utilized with custom learning.

REFERENCES

- [1] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," *2014 IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, OH, USA, 2014, pp. 580-587, doi: 10.1109/CVPR.2014.81.
- [2] R. Girshick, "c," *2015 IEEE International Conference on Computer Vision (ICCV)*, Santiago, Chile, 2015, pp. 1440-1448, doi: 10.1109/ICCV.2015.169.
- [3] K. He, X. Zhang, S. Ren and J. Sun, "Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015, doi: 10.1109/TPAMI.2015.2389824.
- [4] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pp. 91–99, 2015.
- [5] K.-H. Kim, S. Hong, B. Roh, Y. Cheon, M. Park, "PVANET: deep but lightweight neuralnetworks for real-time object detection", in: arXiv preprint arXiv:1608.08021, 2016.
- [6] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 779-788, doi: 10.1109/CVPR.2016.91.

- [7] Liu W. et al. (2016) SSD: Single Shot MultiBox Detector. In: Leibe B., Matas J., Sebe N., Welling M. (eds) *Computer Vision – ECCV 2016*. ECCV 2016. Lecture Notes in Computer Science, vol 9905. Springer, Cham. https://doi.org/10.1007/978-3-319-46448-0_2
- [8] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, Stronger," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 6517-6525, doi: 10.1109/CVPR.2017.690.
- [9] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan and S. Belongie, "Feature Pyramid Networks for Object Detection," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, 2017, pp. 936-944, doi: 10.1109/CVPR.2017.106.
- [10] K. He, G. Gkioxari, P. Dollár and R. Girshick, "Mask R-CNN," *2017 IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017, pp. 2980-2988, doi: 10.1109/ICCV.2017.322.
- [11] J. Dai, K. He and J. Sun; *Procc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 3150-3158.
- [12] L. Zonglei and X. Xianhong, "Deep Compression: A Compression Technology for Apron Surveillance Video," in *IEEE Access*, vol. 7, pp. 129966-129974, 2019, doi: 10.1109/ACCESS.2019.2940252.
- [13] Negin Ghamsarian, "Enabling Relevance-Based Exploration of Cataract Videos", In *Proc. ICMR.Association for Computing Machinery*, NewYork, USA, pp. 378–382, June 2020, doi/10.1145/3372278.3391937.
- [14] S. Kim et al., "Adversarial Video Compression Guided by Soft Edge Detection," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Barcelona, Spain, 2020, pp. 2193-2197, doi:10.1109/ICASSP40776.2020.9054165
- [15] J. Lin, D. Liu, H. Li and F. Wu, "M-LVC: Multiple Frames Prediction for Learned Video Compression," *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, pp. 3543-3551, doi: 10.1109/CVPR42600.2020.00360
- [16] S. Hao, Y. Zhou, Y. Guo, "A Brief Survey on Semantic Segmentation with Deep Learning," *Neurocomputing*, Volume 406, 2020, Pages 302-321 doi: 10.1016/j.neucom.2019.11.118.
- [17] X. Wu, D. Sahoo and S.C.H. Hoi, "Recent advances in deep learning for object detection", *Neurocomputing*, Volume 396, 2020, Pages 39-64, ISSN 0925-2312, <https://doi.org/10.1016/j.neucom.2020.01.085>.
- [18] Xu, J. "A deep learning approach to building an intelligent video surveillance system". *Multimed Tools Appl* 80, 5495–5515 (2021). <https://doi.org/10.1007/s11042-020-09964-6>.
- [19] S. Ma, X. Zhang, C. Jia, Z. Zhao, S. Wang and S. Wang, "Image and Video Compression With Neural

Networks: A Review," in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1683-1698, June 2020, doi: 10.1109/TCSVT.2019.2910119.

[20] Cloud-Computing-server-utilisation <https://aws.amazon.com/blogs/aws/cloud-computing-server-utilization-the-environment>, 2020. [Online; accessed 11-November-2020].