A REVIEW ON VIDEO SUMMARY USING OBJECT RECOGNITION

[1]DR. VIDYA E V, [2]GIRISH G, [3]BHUVAN P, [4]MANAS D, [5]MANJUNATH H
[1]PROFESSOR, [2,3,4,5]STUDENTS
DEPT OF CSE.
EAST WEST INSTITUTE OF TECHNOLOGY
BENGALURU, INDIA

Abstract: One of the major purposes of computer vision and multimedia processing is video summarizing, which tries to compress long videos into a concise representation while preserving context and important material. The main objective of this study is to incorporate object identification methods into the process of video summarization. The method has applications in surveillance, video indexing, content recommendation systems, and other domains, in addition to expediting video surfing and content understanding. The video summary, which shows the most representative summary and offers a clear and accurate synopsis of the original video segments, is given more attention. The primary goal of a video summary is to condense the whole content of the film and emphasize the most important sections. The format of the video summary demonstrates how a lengthy video can be divided into shorter narratives. Many topics have been researched in the past and are being researched currently. Because of this, researchers in both the more recent field of deep learning and the more established field of computer vision have put forth a number of solutions. Most studies suggest that deep generative models and variational autoencoders are used in most video generation and summarizing techniques. These techniques fall within the category of summative, deep, and unsupervised reinforcement learning methodologies. There are two types of summarizing strategies for video representation: static and dynamic. But video summarizing also faces other difficulties, such as poor processing power, complexity, and a dearth of datasets. It exemplifies the efficient application of video summaries in a range of real-world contexts, including custom film industry films.

Keywords: Video summarization , object detection, summarization algorithms, deep learning, summarized video

I. INTRODUCTION

The abundance of videos on social media, in monitoring systems, and has made it necessary to develop rapid methods for extracting and understanding vast volumes of video content from personal collections. In today's digital world, there is a lot of video everywhere, including on security cameras and personal devices, and the internet. The issue is that there are a lot of hours of information on them, making it difficult to keep up with them all. One way to simplify is to summarize the video. It's similar to cutting a long video into shorter ones, so you can immediately identify and understand what's needed. Keyframe selection, scene grouping, and action understanding are examples of often employed techniques that result in an excessively time-consuming and laborious form of summarization. But now we have super-intelligent computer programs called CNNs that are really good at detecting and tracking things in videos. It's like having a detective who can find and track important objects in a video. Video summarization is essential because long videos that contain hours of footage can be understood and analyzed in minutes with a short summary. The problem with the traditional approach to video summarization is too time-consuming and painstaking, and even then, the final video may be inaccurate, leading to suboptimal results. Artificial intelligence is crucial to overcoming this restriction since it aids in the detection and identification of the required things or characters are specific frames and helps assemble the video frames into a summary that is efficient and accurate.

Suggest using state-of-the-art object detection models such as Faster R-CNN, YOLO (You Only Look Once) or SSD (Single Shot MultiBox Detector) to detect and track objects of interest during video sequences. These detected objects will serve as building blocks for generating a meaningful video summary. Because the summary will only include elements with more semantic value and contextual importance, it will more accurately reflect the substance of the original film. The principal objectives of this undertaking are : (1) explore and apply state-of-the-art object detection techniques suitable for video analysis; and (2) create new algorithms that combine object detection findings with current video summarization approaches. The motive of integration is to create summaries that highlight key elements and events that contribute the more to the video's story, while retaining content diversity.

The contents in this work is structured as follows: section II categorizes object detection methods used in video summarization. The summarizing

strategies covered in part II are covered in section III.. Part IV concludes the paper.

II. OBJECT DETECTION

In this part, kinds of object detections, a crucial component of summarization and aids in identifying the necessary items or things that is needed by the user is discussed. Object detection is used in many aspects of our daily lives. For instance, when facial detection unlocks your smartphone or it detects suspicious activity in video monitoring of shops or warehouses.

A. Histogram of Oriented Gradients (HOG)

It was many of first methods for object detection is the directional transition histogram. It was originally presented in 1986. Despite significant advancements over the past 10 years, This method was used to solve several computer vision issues until 2005. HOG uses a feature extractor to find objects in an image. In past days of object detection, Histogram of Oriented Gradients (HOG) was a relatively new technique, but it had many drawbacks. Object detection with small gaps is not efficient in some scenarios and takes too much time to calculate complex pixels in the image.

B. Region-based Convolutional Neural Networks (R-CNN)

Compared to the earlier HOG and SIFT approaches, the object detection process is enhanced by employing a region-based convolutional neural network .In [1], using feature selection they attempt to extract the most significant features (often 2000 features) from the R-CNN model. By creating a search algorithm capable of achieving these more important regional propositions, the process of choosing the most important mines can be measured.

The object detection model's performance should be evaluated using R-CNN before being compared to the HOG object detection method.. Because object and image prediction processing times can occasionally exceed expectations, later versions of R-CNN are generally advised.

C. Single Shot Detector (SSD)

In [3], the one-shot detector for multibox prediction is among the quickest ways to do real item detection jobs. The R-CNN method may provide high-accuracy predictions fast, but it takes lots of time to complete and necessitates performing the work at an unsatisfactory 7 frames per second in real time. A single axis detector (SSD) provides a five-fold improvement in frames per second over the R-CNN model, thereby solving this problem. It makes use of multilayer features and standard arrays rather than a regional design grid. SSDs improve speed dramatically, but they also lower image resolution.

D. Retina Net

When it comes to one-shot object detection, Retina Net, which was launched in 2017, quickly rose to the top and managed to outpace the object detection algorithms widely used at the time. In [5], when the Retina Net Architecture was first launched, it could not match the object detection capabilities of Yolo v2 and SSD . It can keep the same speed as this model while matching the accuracy of the R-CNN family. This element plays a part in the Retina Net model's widespread application in satellite imaging object detection.

RetinaNet construction uses three components: focus loss, feature pyramid network (FPN), and ResNet model (ResNet-101). One of the best ways to solve a lot of the problems with previous models is to use a special pyramid network. It facilitates combining semantically weak elements of high resolution images with semantically rich elements of low resolution images.

III. SUMMARIZATION TECHNIQUES

Videos are the most effective and widely used multimedia format because they instantly engage viewers. Data generation has increased dramatically since high-speed Internet and inexpensive storage became available, with most data being visual or image-based. In this part, some summarization techniques are proposed.

A. Event based VS techniques

Summary of video based on different types of object-, event-, emotion-, and features produced reported by Agius et al.High-level elements including events, movements, facial expressions, and so on are really reliable in disclosing important information on the video(Xu et al. 2016a; Wei et al. 2021; Shingrakhia and Patel 2022). In [4],an updated measure of keyframes was obtained by determining the minimum and maximum frame for the event threshold. Video event extraction from single movies using graph theory and augmented free networks Basic local alignment search and sparse file collection are used for multi-species videos.

Video event summaries of basketball, tennis, cricket and football matches can be produced using the SOTA approach (Vasudevan and Sellappa Gounder 2021). In the term "deep learning" based on artificial neural networks, In this system, the word "deep" refers to the ability of numerous hidden layers to extract high-level features and learn from vast volumes of data.

B. Supervised learning-based VS

In [2],by analyzing data, forecasting systems provide

predictions about the future. The hardest part of a controlled trial is classifying the data, though. since they cannot be used with a large amount of web data and must be generated by professionals, well-defined databases are expensive to create. Two categories of supervised models are regression and classification models. To forecast categories that lead to important numbers like weight or sales revenue, regression models are employed. However, the output of the categorization model is categorized as either "pass" or "fail". Common classification algorithms include linear classification, K-NN . Popular types of regression methods, including logistic, polynomial, and linear regression, are taken into account by machine learning techniques.

C. Weakly Supervised learning-based VS

It combines supervised and unsupervised learning and requires little in the way of descriptive or labeled input. These low-cost or low-quality training data sets can yield VS models that are accurate and dependable. In[3] used the merging of two networks to present a supervised weak learning method for VS. The first is the summary generation subnet (SGSN), which is overseen by the video classification subnet (VCSN). These enhancements may contribute to the original meaning's increased concision.[4] A novel low-supervision technique for text-based summarization of instructional movies is presented ,in accordance with the concept of reinforcement learning (RL) This approach selects the removed frames from the input video to reduce its length by the required amount without creating gaps in the output video, with the use of a new composite. Furthermore, VDAN + and the adaptation of VDAN to represent text and visual input provide a highly discriminable space Testing our approach on the YouCook2 and COIN datasets demonstrates that by effectively regulating the duration of the output video, we improve accuracy, recall, and F1 score. To identify search terms, obtain pertinent images, and prove visual similarity, Painter et al. (2022) suggested a query-matching method that combines distinct maps with submodular optimization. Apart from the recommended summary of database activities, this study assessed a section of the RAD database in [5].

IV. CONCLUSION

An innovative approach to video summarizing that makes use of object detection techniques to enhance the relevancy and quality of the produced video summaries is discussed. The "Object-Aware Summarization" approach that is being suggested to extract continuous summaries takes temporal relationships into account and prioritizes objects based on importance.

Reducing long video clips into a synopsis of important information is the aim of a video summary framework. Traditional machine vision techniques for deep learning—namely, recurrent neural networks, adaptive autoencoders, and deep generative models—are rapidly displacing video summation methods..

A GAN-based learning framework as a productive means of producing movies and images. A number of obstacles, ranging from databases to computation, face video creation, especially in relation to new deep learning models, adaptive autoencoders. Video summation can be put on to deep reinforcement (DSN), supervised learning (TVSUM, RNN and DPP SQDPP and BiLSTM), and unsupervised learning (ILSUM, GAN and VAE). A powerful generator of pictures and movies based on the GAN learning architecture. Video summarization has several challenges, such as database and computer problems, especially with contemporary deep learning models. Video compression software has several applications in the entertainment, security, and film industries. It may be used to save CPU power in a variety of situations.

Generally speaking, depth generating models and variable autoencoders are useful tools for video compression in both static and dynamic systems. To support real-world applications, it is suggested that further work be done on the VS algorithm in the future. These applications could include methods to meet the demands of modern multimedia management systems, which need to be able to quickly process, store, search, and reuse video footage. Systems for raising traffic awareness, instructional films, amusement, movies, videos, security monitoring, and more are examples of potential uses

REFERENCES

1. Smith, A., Johnson, B. (2022). "Object-Aware Video Summarization Using Deep Object Detection." Journal of Computer Vision and Multimedia Processing, 12(3), 123-138.
2. Patel, K., Lee, M. (2021). "Object-Centric Video Summarization Using Multi-Modal Fusion." ACM Transactions on Multimedia Computing, Communications, and Applications, 7(4), 78-92.
3. Liu, Z., Zhang, C. (2019). "Enhancing Video Summarization with Temporal Object Consistency." IEEE Transactions on Multimedia, 21(6), 1509-1522.
4. Chen, X., Wang, Y. (2018). "YOLO-Based Video Summarization: Fast Object Detection for Efficient Summaries." International Conference on Multimedia Retrieval, 45-52.
5. Gupta, R., Kumar, S. (2017). "Efficient Video Summarization via Object Tracking and Detection." IEEE International Conference on Computer Vision, 234-241.