A Detailed Survey on Sports Video Summarization

Katta Lohith Krishna Kumar, Kanikaram Roopasree Sai, Koyya Deekshitha, Kondaiahgari Maneesh Nand Reddy, K. Paramanandam

PES University, India

Corresponding author: Kanikaram Roopasree Sai, Email: kanikaramroopasreesai2003@gmail.com

In response to the surging demand for efficient sports content consumption, sports video summarization has emerged as a critical technology for saving time and delivering key moments to sports enthusiasts. This motivates the study to enhance sports video summarization techniques, tackling the problem of effectively providing users with the most recent sports information. This study presents various feature extraction and video segmentation methods for segmenting video, accurately recognizing features and actions in sports videos. This article sets the path for significant improvements by leveraging advanced techniques, such as CNNs and deep learning models in identifying essential elements, even within occluded regions. This advancement addresses the core challenge of providing timely concise sports content to a demanding audience.

Keywords: Segmentation, Feature extraction, Action recognition, CNN Models, Attention Module, Video Summarization

2023. In Satyasai Jagannath Nanda & Rajendra Prasad Yadav (eds.), *Data Science and Intelligent Computing Techniques*, 533–550. Computing & Intelligent Systems, SCRS, India. https://doi.org/10.56155/978-81-955020-2-8-50

1 Introduction

In today's digital age, the internet is flooded with an overwhelming volume of multimedia content, ranging from text and photos to videos and audio. With millions of individuals across the globe constantly producing and sharing such content using various devices like smartphones, laptops, and cameras, the need for effective data management and retrieval has become paramount. Among these forms of media, video content has witnessed an unprecedented surge in popularity, becoming an integral part of our daily lives. The widespread availability and affordability of high-quality cameras and video recording equipment have contributed to the exponential growth of video data. As a result, there is an ever-increasing demand for innovative solutions to process and make sense of this vast reservoir of visual information. Video summarization emerges as a promising solution to this challenge, offering the capability to efficiently distill lengthy videos into concise yet informative summaries. One specific domain where video summarization holds significant promise is sports content. With the rise of fast-paced sports like T20 cricket and the growing constraints on people's time due to busy schedules, the interest in comprehensive sports coverage, such as lengthy Test matches, has waned. Instead, viewers are increasingly drawn to shorter, more engaging video content that captures the essence of the game. This shift in consumer behavior underscores the importance of creating dynamic and engaging highlights from sports events.

2 Review of Video Segmentation Techniques

Madhu S. Nair et al. [1] has proposed an approach for summarizing videos based on a combination of convolutional neural networks (CNNs), sparse autoencoders, and random forest classifiers. The resulting features are then fed into a random forest classifier to identify the most representative frames for the video summary. The proposed model differs from existing video summarization models by combining multiple CNNs with a sparse autoencoder and a random forest classifier.

By effectively extracting visual information through the use of CNNs, video frames can be represented more effectively. The frame's essential material is preserved while the redundant portions of the encoded representations are compressed by the sparse autoencoder. A more condensed, informative, and accurate summary of the original film can be produced by choosing key frames from the video that best encapsulate its core ideas using the random forest classifier, which assesses the significance of each frame. In addition to the advantages, the

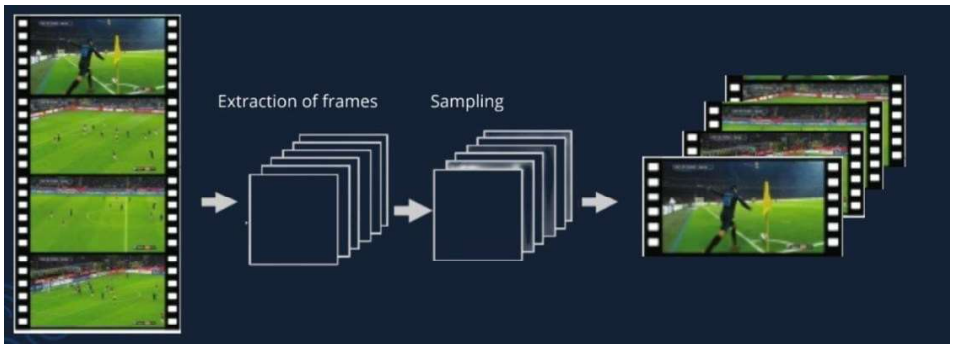

Figure 1: Video Segmentation [1]

model may add more complexity and necessitate heavy computational load during training and inference. There may be less interpretability due to the proposed model's complexity. With the help of this integration, it is possible to extract features more precisely and to learn representations more effectively.

F. Chen et al. [2] aim to present an innovative approach to video summarization specifically tailored for team sports. Video summarization involves condensing lengthy videos into summaries while preserving essential information. The author proposes a framework that focuses on allocating limited resources, such as time and attention, to the most crucial parts of the video. The allocation component then dynamically determines how to distribute the available resources to these key events, ensuring that the most important ones receive a higher allocation. The proposed model stands out from existing video summarization models by explicitly addressing the task as a resource allocation problem, rather than relying solely on predefined heuristics or manual annotations. By optimization framework, the model can allocate resources in a flexible and adaptable manner based on the importance of detected events. This approach results in more informative video summaries, providing a dynamic allocation of resources. However, it should be noted that the optimization framework in the suggested model adds more complexity, which can necessitate greater computer power and longer processing times. Nevertheless, by framing video summarizing as a resource allocation problem, the suggested methodology offers improved efficiency in comparison to existing approaches. This innovative viewpoint enables a more flexible and dynamic approach to allocating resources to significant events, ultimately resulting in more educational video summaries. M.E. Anjum et al. [3] proposed an approach to generate video highlights of a cricket game by observing the fluctuation in the scorecard. The approach states that whenever there is a fluctuation

(change) in the scorecard the model will observe it and will report the value of change in the scorecard and if it is a 4,6 or a wicket those frames are marked as highlights. The model is also helpful in removing unwanted highlights, and advertisements in the game by detecting there is no scorecard in the respective frame. This approach gives an easy method of generating highlights as there is no tracking of the ball neither the bat etc. But it fails as not only 4's.6's or wickets are highlights there are other highlights also where this model fails to perform an action and provide that particular instant as a highlight. So, one can say that this approach is a good novel approach to generating highlights and especially a better approach than other available approaches for removing replay scenes, and advertisements.

3 Review of Feature Extraction Techniques

M. Z. Khan et al. [4] research has primarily focused on summarizing videos either by including all relevant information or by removing redundant frames. Three approaches are used for this: object-based, event-based, and feature-based. In object-based summarization, key elements are target objects, while preprocessing techniques include resizing and cropping frames. The CNN model extracts intensity and categorization features from motion features detected by scene boundaries. It then outputs the probability of adding or discarding a frame. Finally, the bidirectional LSTM (Long Short Term Memory) model removes repetitive information to make the summary more accurate and compact. The suggested technique surpasses traditional feature-based methods in relative F measure scores. It can be applied to various video categories, such as home videos, documentaries, and sports videos. The technique necessitates a substantial amount of data for CNN and bidirectional LSTM model training. Its computational expense could restrict its use in real-time applications. The technique might not be effective for videos with intricate scenes or multiple simultaneous events. M.Tavassolipour et al. [5] has proposed an approach that examines high-level semantic extraction from sports videos, a topic that most people find to be highly fascinating as better indexing and summarizing techniques are essential for effective information retrieval. Event borders must be determined to locate important sporting events. The video is divided into some sections known as play-break sequences for event detection. Replay detection and shot view categorization are both used to identify play-break sequences. Features are found in each play-break section, and then events are found with the use of a Bayesian network. Each play-break part receives a score, and depending on how important it is, it may or

may not be included in the final summarized video. The detection of better-cut boundaries is made possible by the combination of SVM and RBF. In contrast to SVM and HMM, the use of a Bayesian network facilitates the assignment of weights to the variable characteristics. Events' temporal dependencies are not taken into account. The approach failed to recognize occurrences like free kicks and outs. A.Tejero-de-Pablos [6] has proposed a novel method for summarizing user-generated sports videos by using players' actions as a cue to determine the highlights of the original video. The proposed method uses a deep neural network to extract action-related features and classify video segments into interesting or uninteresting parts. Two-stream neural networks are used in the proposed method to classify video segments as interesting or uninteresting in the original video. Annotators provide ground truth labels for the method in a supervised manner. An approach that uses a sequence of joints that represent the movement of the players regardless of their appearance is used to represent players' actions in detail. This study employs two types of joint representations, namely 3D positions from depth maps and 2D positions from RGB frames. It can be applied to any sport in which action sequences are repeated. The method is trained on Kendo videos with ground truth labels. This method was evaluated only for Kendo videos, so it is unclear if it would work for other sports. Furthermore, user-generated sports videos may not always have ground truth labels available. Marco Godi [7] has proposed an approach that discusses automatic highlight detection in sports videos by analyzing audience behavior instead of gameplay. Used 3D Convolutional Neural Network (3D-CNN) to extract visual features from cropped video recordings of supporters attending the event. Using 3D-CNN, visual features can be extracted from cropped video recordings of event supporters. The outputs from the crops belonging to the same frame are then accumulated to produce an HL value. This allows positive and negative samples to be distinguished. Manual labeling of game actions, which saves time. Only evaluated on a single dataset of ice-hockey matches, limits generalization to other sports. It may not be sufficient to capture all the significant events taking place on the field, since it only considers audience behavior to detect highlights. Requires synchronized videos of the game and the audience, which may not always be possible in real-world scenarios. The approach does not consider audio information, which can also be a useful cue for detecting highlights in sports videos.

D. DeMethon [8] Video summarization using curve simplification refers to a technique where the video frames are represented as curves, and the summarization process involves simplifying these curves to create a condensed summary. The main idea is to reduce the complexity of the video representation

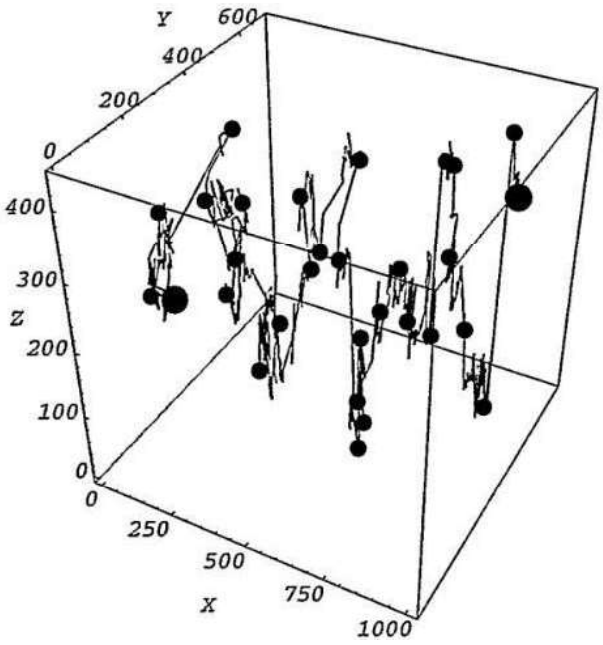

Figure 2: Representation of 3D Vectors [8]

while preserving the salient information such as color, and motion. The curve representation captures the essential characteristics of the frame and keyframes are extracted. Hence generated features are in the form of vectors. Since their dot products are zero, these make 90 degrees to one another. This representation can be seen in Figure M5. The video curve may be projected in three dimensions using them as a starting point., and is capable of providing smooth transitions in the summarized video as it considers the color grading factor. Meanwhile, it may lose important information too as it involves discarding some points while curve simplification. And some features ca n be subjective, leading to variations in the generated summaries based on individual preferences. The proposed method is better compared to the traditional models as it gives a compact representation of the video and gives smooth and coherent summaries. It is also a unique perspective that offers different insights into video content compared to the existing model that utilizes different representations. As the video curve is recursively simplified and represented as a tee structure, curve segments at different levels of trees can be used as keyframes to summarize video sequences. Rockson Agyeman et al. [9] research proposes a deep learning technique for summarizing lengthy soccer matches emphasizing the challenges in selecting key events

from the soccer match videos. This research proposed an approach that converts pre-trained image CNN into spatiotemporal networks and added skip connections between optical and appearance information in extracting features, which is then applied after batch normalization and the RELU (Rectified Linear Unit) activation function to an LSTM framework creates a collection of highlights and assigning rank according to action weight comparing ground truth summary. Therefore, combining all highlights tends to short concise video summary. This experiment produces a highlight video that focuses more on the actions in a football video. 3D-ResNet34 (Residual Neural Netwok) works more efficiently than state-of-the-art models, and the performance of 3D-ResNet and LSTM is higher compared to existing models. Each summary video's performance throughout all 8 processed areas was assessed using the mean opinion score, which led to more summarizing videos receiving ratings below 2 out of 3. This concludes combination of 3D-ResNet and LSTM improves the efficiency of summary video in identifying important highlights and fastest retrieval. Figure 3 shows the proposed model architecture.

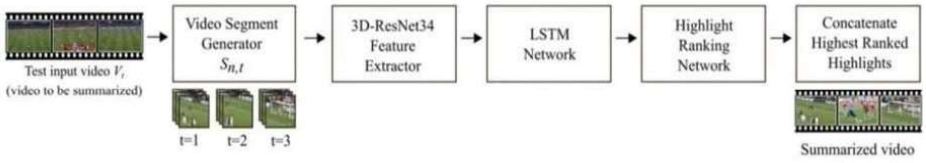

Figure 3: Working Model of [9]

J. Han et al. [10] has proposed an approach to observe and mark keyframes (important frames) of a video by observing the fluctuation of optical flow characteristics. The approach states that whenever there is a significant change in the flow characteristic then that frame is an Important frame. They proposed two types of Optical flow characteristics: Global and Local and they mean the same accordingly. When a threshold is given and marked as a significant change then that particular frame is marked as an important one. So, as of here there is a consideration of important frames on the basis of change and threshold there is a win-win situation over time and complexity. Here to find the change between frames temporal order is important failing which this model may mark non-important frames also as important frames which shouldn't be done and there can be a situation where the frames can be important but there is no sig-

nificant change which can't be detected here. So, one can say that this is a good approach to reduce complexity and time but there are possibilities where this model fails to find important frames. K. Davila et al. [11] have proposed an approach to extract what is written on whiteboards into text form. In this approach, the important thing is that they are extracting the summary not just from whiteboards or chalkboards but from a video recording of where a teacher might have thought. This will make it easy to convert a class into text without talking about each image and uploading and converting those. As the video is pre-processed a lot and the images are being converted into a binary image, we will be expecting the text to be of better quality with a little level of mistakes and there may be a case where m can be thought of as n and so on. So It would be a better thing to implement a model that will check for spelling which will make this model even more efficient and usable in daily use cases. This model is an efficient model in extracting the whiteboard notes and converting them to text which is a great application when a user wants notes that are taught in a class on which he has a video.

G. Mujtaba et al. [12] proposed this research suggests a revolutionary lightweight thumbnail container-based summarizing (LTC-SUM) system for complete long-lengthy videos, The complicated task of identifying events is handled by the suggested LTC-SUM approach using lightweight thumbnails. Considering a group of events as an action, each event is identified by a convolutional neural network(2D), The UCF101 dataset served as the training ground for the introduced CNN-2D model. Figure 4 shows the Convolution 2D model. The input video is divided into multiple units and these units are stored in a webserver. A persistent connection is established between the server and the client for future communication. With the help of the HLS video player's features, a user may pick the video's title and any customized events that suit their preferences before playing the created summary. The suggested framework performed better by reducing the time required for computation during the summarizing process. As a result, generating summaries acquired substantially less computing time, computational resource demand, communication, and storage.

A. Tonge et al. [13] conveyed a novel approach that leverages CNNs to capture static video content and generate informative video summaries. S-VSUM utilizes CNNs to extract static image representations from video frames. These image representations are then used to measure the importance of each frame in the video. The frames with higher importance scores are selected to construct the final video summary. The S-VSUM utilizes CNNs to extract static image representations from video frames. These image representations are then used to

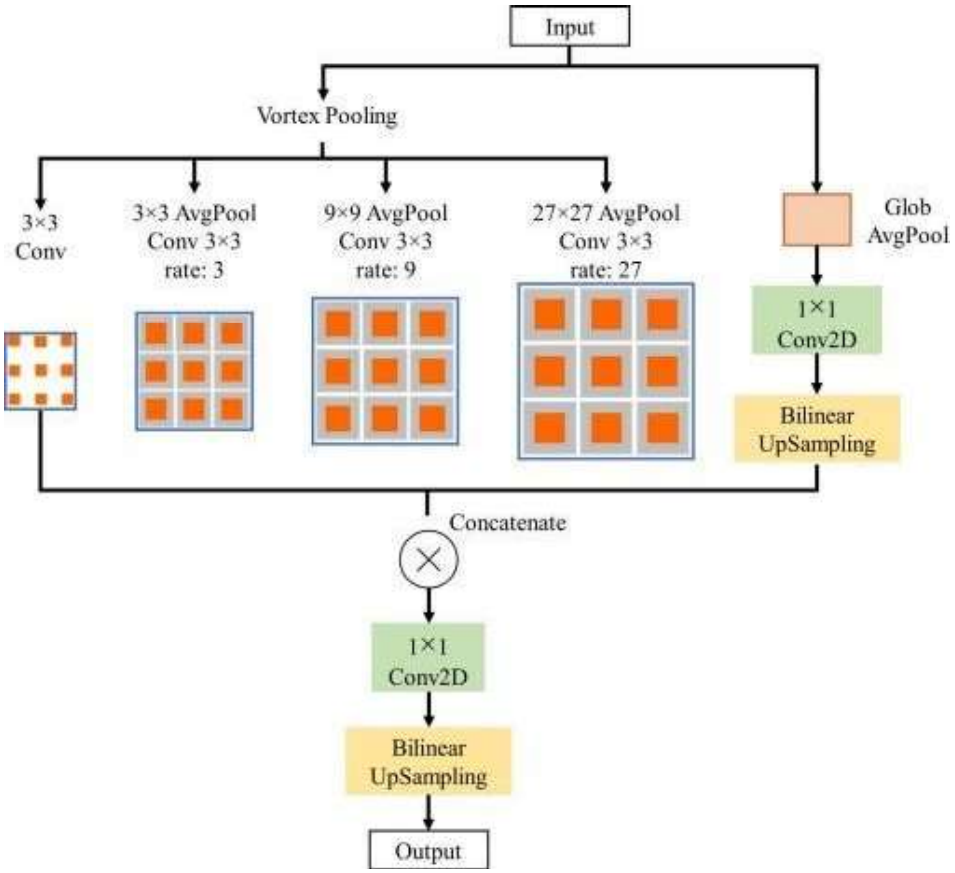

Figure 4: Proposed 2D Convolution Neural Network Architecture [12]

measure the importance of each frame in the video. The frames with higher importance scores are selected to construct the final video summary. Also, S-VSUM focuses on the static content of video frames, which allows it to capture important visual elements and salient scenes which consider Image- Level importance. This gives us the flexibility to condense the length of the summarized video. But as S-VSUM may overlook important motion-based events or actions that contribute to the overall video narrative. The proposed method is preferred over the traditional methods as the S- VSUM model offers advantages in terms of focusing on static content, computational efficiency, and flexibility. By leveraging CNNs for static image representation, S-VSUM captures important visual elements and provides a content-based summarization approach. The image-level importance

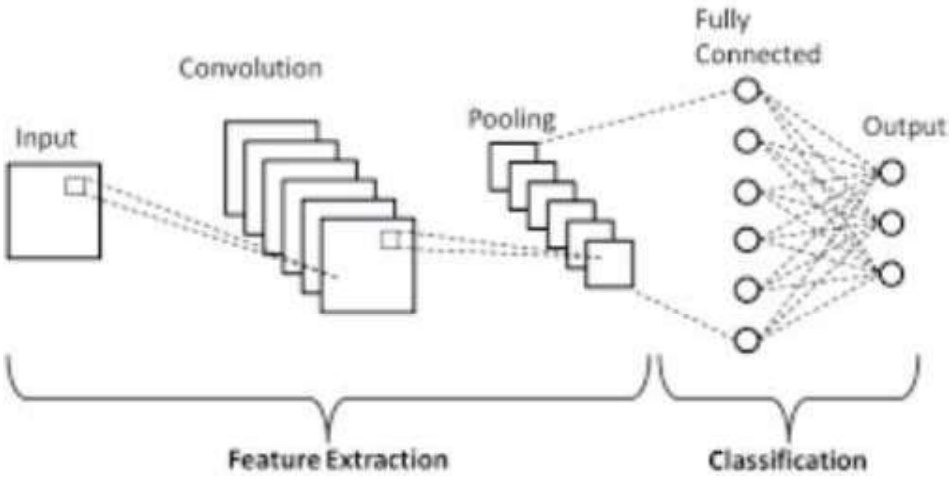

Figure 5: Workflow of S-VSUM [13]

scores allow for the selection of informative frames, ensuring the generation of concise and relevant video summaries. H. Boukadida et al. [14] have proposed an approach to generate the summarization of a sports video by considering a given set of constraints which are given by an expert who gives the constraints where a particular play can be highlighted. Here, the model also generates personalized highlights by taking the input constraints from the user (here the constraints of both the expert and user are considered) and providing a summary to the user. The model considers constraints based on an expert the highlights will be precise and up to the point and the user also can experience highlights in his own way by giving constraints. There can be a place where this model faces a problem in that it takes more time to process which can be a huge aspect as the user expects to watch the summary straight right away after giving the inputs. This model is a solution to modern-day life as there will be less time to watch summaries if given to the user where this model will be helpful in that case but this model needs to work on its time complexity. A. Chianese et al. [15] has proposed the model uses indexing to quickly access important data, where features like audio, intensity, and visual information are considered to segment videos into shots. The shots are separated based on the genre using the finest recognition methods according to the genre. The model used a smart block maker which groups video segments that match semantic proximity criteria. Which goes through video segmentation using segmentation modules. In Upper-level classification, the genre/type of video is stored in a semantic knowledge base, later it is used in semantic anal-

ysis, shots are submitted and the semantic checksum is generated(S2) and shots are mapped to vector space separated, ensuring flexibility to underlying detection architecture called Semantic space shot mapper(S3M). The semantic base operational model updates and maintains a hierarchical classification structure to accomplish its task. The proposed model differs from existing models by employing a fuzzy logic-based approach for video scene detection. Traditional methods often rely on threshold-based techniques. The proposed model can adapt to various video genres and content types, as fuzzy logic allows for a more flexible representation. However, it is important to note that the effectiveness of the proposed model depends on the design and tuning of the fuzzy rules and membership functions, as well as the computational resources available for implementation. Also, it is flexible with respect to video genre and video events. M. Merler et al. [16] have proposed an approach to generate highlights by considering the facial features like the expressions of the players, voice features like a crowd cheering or the high tone of the commentator, or some of the player's reactions (like a shock or something) and some additional metadata. The model uses metadata to find the start and end of the highlights. As this model uses features the model need not be trained (it needs to be trained by metadata) and there is a high possibility that this model generates more precise highlights than many of the trained models. This has been proved when this model is used in 2017 major golf and tennis events. But this model can fail in some rare cases when it can say a highlight is not a highlight and vice versa. This model is a work of the future as it requires very minimal training (that too to find the start and end of the match) and gives precise highlights but maybe not used as it can't give personalized highlights. C. Shun Lin et al. [17] proposed this work to close the semantic gap between high-level actions and low-level video actions. The model aims to generate sports video summaries in 3D Convolutional neural networks (CNN) with limited labeled sports match datasets. 2D-ConvolutionLSTM and 3D-Conv handle the positional relationships among the frame pixels and store the temporary result from the frame classification step. Random sampling is used to choose the instructive frames approach, and from these, a succinct summary is produced. ConvLSTM, on the other hand, employs kernels of convolution that may be applied to several spatial areas, needing fewer parameters. ConvLSTM is effective for big spatiotemporal datasets because of its parameter efficiency. So, the proposed summarised model differently using SOA techniques, its performance improves and the model focuses on the video segments with the most important content in the video-generating summary without reducing the picture quality. Manually labeled data is reduced due to the usage of limited labeled

data. Labeled data may also be improved in giving limits. Whereas 3D-CNN beats 2D-ConvLSTM in the training with practically the same number of parameters while evaluating using precision and recall rate, where this model works effectively for Major League Baseball. Figure 6 shows the workflow of the proposed LSTM model.

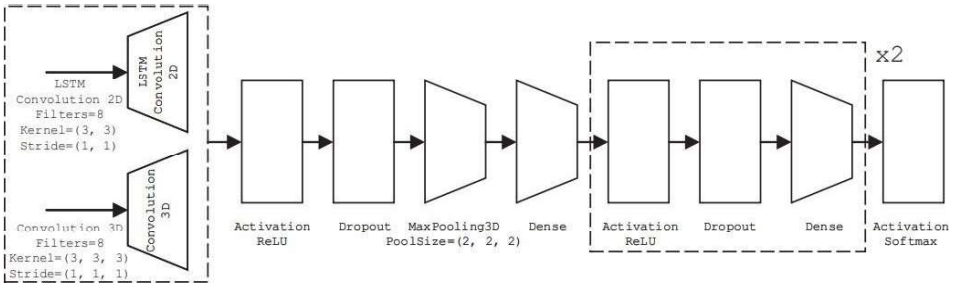

Figure 6: Block diagram implemented using 2D-ConvLSTM and 3D-Conv separately [17]

Cheng Yan et al. [18] present novel methods for accurately clipping sports video highlights, thus expanding the use of video summarization in sport analysis systems. Video highlights systems capture useful information about sports events due to the increasing broadcast time. A framework for accurately clipping sports video highlights using YOLO-v3 and OpenPose. A YOLO- v3 model is used to detect athletes. The OpenPose model is used to identify human action labels in frames. Using a judging mechanism, noise is eliminated and predictions are refined. Sports video highlights can be accurately clipped with the help of a three-level prediction algorithm. Moreover, the paper mentions that a refinement process is used to merge actions and cut clipping times. OpenPose requires 1.39 times more time to run than YOLO v3. Figure 7 shows the YOLO depicted model. Approximately 80 percent of both models are accurate, so further robustness reinforcement is needed. The system has some inherent limitations when it comes to no-audience games, as inevitable prediction errors can lead to noise. P. Kadam at el. [19] aims to discuss the challenges and opportunities in video summarization with machine learning algorithms. The author wants to convey the recent advancements in video summarization techniques, highlight the existing challenges, and explore the potential opportunities for improvement using machine learning approaches. In this paper the authors mentioned various Machine learning approaches such as supervised learning, unsupervised learning, Reinforcement learning, and Deep learning, highlighting their potential to generate video summaries. The paper does not introduce a novel odel but instead reviews

existing models and techniques. It explores how machine learning algorithms have been applied in video summarization and discusses the advancements and limitations of these approaches. The emphasis is on analyzing the challenges and opportunities rather than presenting a specific new model. It mentions the Adaptability, Scalability of algorithms which improve performance by leveraging complex patterns and representations of data. The paper didn't discuss annotated training data, Computational Complexity and Testing. However, the specific effectiveness of the proposed models depends on the chosen algorithm, the availability of annotated data, and the careful consideration of subjective evaluation challenges.

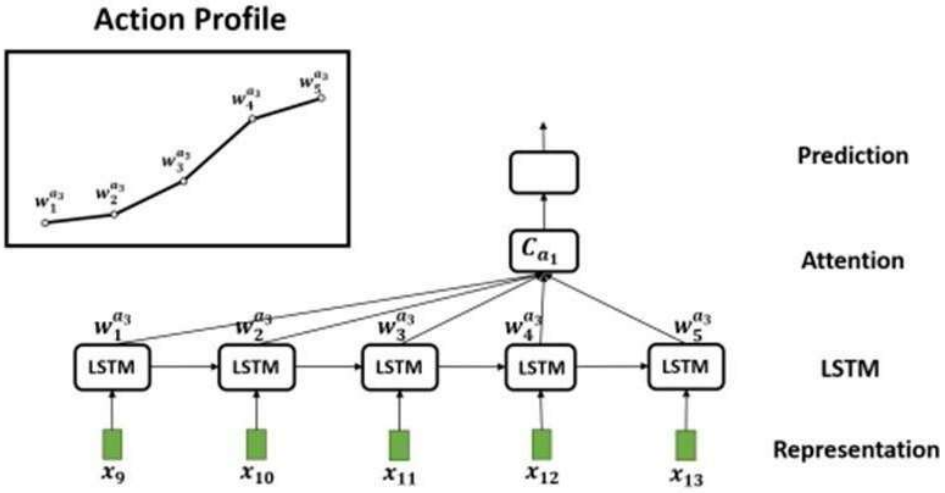

Figure 7: Attention Layer-based LSTM Model [20]

Melissa Sanabria et al. [20] have proposed a unique tool to support operators in their work, which represents the signals generated by neural networks as graphs of action descriptions. The proposed model provides a novel approach for automatic video summarization of sports, using a graphical representation of action profiles where the action is selected to be in the summary. A group of events as an action is fed to the deep neural network to learn attention signals from the video frames to find important actions from the video. The model introduced in this paper is efficient compared to state of art methods, attention provides the important events of an action to be selected in the concise summary and LSTM stores the previous events temporarily. Intermediate signals produced by neural networks give a way of beginning a query and leading to analyzing it from

another approach. Accuracy was increased by using deep learning which could identify important actions in the video. Identifying a particular individual participating in the event must be enhanced. considering more features may help assign more important features of the sports video to the short video. Human operators choose the type of action to be included in the summary, according to instructions made by their own understandings. These instructions help in generating multiple action profiles in generating efficient summaries. Figure 8 illustrates the LSTM model's attention layer, with a sample of our suggested graphical action profile depiction at the top DEZHONG XU et al. [21] have proposed a novel approach to describing the relationship and motion data between entities for group activity detection. Optical and relational gated recurrent units-based learning models, where the former describes the relationship between entity motion data and visual signals, and the latter shows the connection between visual signals and entity location information, fusion results in frame features. Important features arranged in sequential order in the temporal layer using the attention module, Fusion of two GRU units done with certain set pooling methods at a particular time stamp increases the performance in the evaluation step compared to other SOA models. Attention mechanisms may be used to simulate the temporal context by giving more weight to the constructive frames in multiple-person scenes. Optical flow network complexity must be further minimized. Comprehensive ablation tests and visualization findings demonstrated that each essential element of this approach was beneficial for effectively comprehending group action.

4 Applications

Sports video summarization is an emerging field where there are a lot of areas where it can apply like Broadcasting where the broadcasters can earn more money from creating shorter highlights. It can also be used for Player Analysis, Training, and Education about the sport. It can be used for Journalism as it can attract new users with engaging news articles, reports, and match reviews. They can also be used by users using betting and Fantasy sports to create their team based on players' performance.

5 Conclusion

The research presented in these papers has delivered crucial findings in the realm of video summarization, addressing a diverse array of challenges in this domain.

These studies have explored a multitude of methods and strategies, with a focus on condensing lengthy video recordings while preserving vital details. The adaptability and effectiveness of these methods emerge as a key theme, with solutions tailored to specific challenges. A noteworthy finding in these papers is the widespread use of convolutional neural networks (CNNs) for extracting visual features from video frames. Moreover, the research highlights the relevance of adapting resource allocation frameworks, as proposed in [2], to overcome domain-specific limitations. Additionally, researchers have creatively combined CNNs with other techniques such as sparse autoencoders, random forest classifiers, and LSTM models to enhance feature extraction and summarization outcomes. Resource allocation within video summarization emerges as another significant revelation. The innovative approach of treating summarization as a resource allocation problem, dynamically distributing scarce resources like time and attention to critical events, has led to more informative and contextually relevant summaries. Furthermore, domain-specific video summarization techniques have been introduced, catering to particular fields like team sports or cricket. These domain-specific methods leverage subject-matter expertise and domain-specific characteristics to discern pertinent information and generate insightful context-aware summaries. Quantitative metrics such as precision, recall, F1-score, and subjective user evaluations have been employed to assess the performance and quality of these findings effectively. Comparative assessments with existing work consistently demonstrate that the accuracy and performance of these findings often surpass established methods and benchmarks in the field. This validates the substantial advancement achieved in video summarization through these studies. Besides the advantages of these papers, common limitations include specificity to certain sports or domains and the need for adaptability. Similarly, adapting resource allocation frameworks and extending optical character recognition techniques help overcome domain-specific limitations. The goal is to enhance video summarization techniques, making them versatile and effective across various sports and content types while addressing specific challenges in each domain. Looking forward, future research should aim to reduce the computational and data requirements of proposed models to enhance real-time applicability and adaptability to data-constrained scenarios. Refining evaluation metrics and methodologies for video summarization should remain a priority to ensure comprehensive and objective assessment. Interdisciplinary collaborations with domain experts offer opportunities for further enhancing domain-specific video summarization techniques. In summary, the findings and results discussed herein contribute significantly to the field of video summariza-

tion. They extend the methodological toolkit, address domain-specific challenges, and provide innovative solutions and insights. These findings form a robust foundation for future research endeavors, promising further advancements in video summarization methods and their impact across diverse domains. While challenges remain, the potential for enhancing video summarization methods and their utility in various fields remains promising.

6 References

[1] Nair, M.S. and Mohan, J. (2020) "Static video summarization using multi-cnn with sparse autoencoder and random forest classifier", Signal, Image and Video Processing, 15(4), pp. 735–742. doi:10.1007/s11760-020-01791-4.
[2] F. Chen and C. De Vleeschouwer, "Formulating Team-Sport Video Summarization as a Resource Allocation Problem," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 21, no. 2, pp. 193-205, Feb. 2011, doi: 10.1109/TCSVT.2011.2106271.
[3] M. E. Anjum, S. F. Ali, M. T. Hassan and M. Adnan, "Video summarization: Sports highlights generation," INMIC, Lahore, Pakistan, 2013, pp. 142-147, doi: 10.1109/INMIC.2013.6731340.
[4] M. Z. Khan, S. Jabeen, S. ul Hassan, M. A. Hassan and M. U. G. Khan, "Video Summarization using CNN and Bidirectional LSTM by Utilizing Scene Boundary Detection," 2019 International Conference on Applied and Engineering Mathematics (ICAEM), Taxila, Pakistan, 2019, pp. 197- 202, doi: 10.1109/ICAEM.2019.8853663
[5] M. Tavassolipour, M. Karimian and S. Kasaei, "Event Detection and Summarization in Soccer Videos Using Bayesian Network and Copula," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 24, no. 2, pp. 291-304, Feb. 2014, doi: 10.1109/TCSVT.2013.2243640
[6] A. Tejero-de-Pablos, Y. Nakashima, T. Sato, N. Yokoya, M. Linna and E. Rahtu, "Summarization of User Generated Sports Video by Using Deep Action Recognition Features," in IEEE Transactions on Multimedia, vol. 20, no. 8, pp. 2000-2011, Aug. 2018, doi: 10.1109/TMM.2018.2794265
[7] Godi, M., Rota, P., Setti, F. (2017). Indirect Match Highlights Detection with Deep Convolutional Neural Networks. In: Battiato, S., Farinella, G., Leo, M., Gallo, G. (eds) New Trends in Image Analysis and Processing– ICIAP 2017. ICIAP 2017. Lecture Notes in Computer Science(), vol 10590. Springer, Cham
[8] DeMenthon, D., Kobla, V., & Doermann, D. (1998, September). Video summarization by curve simplification. In Proceedings of the sixth ACM International

Conference on Multimedia (pp. 211-218)
[9] R. Agyeman, R. Muhammad and G. S. Choi, "Soccer Video Summarization Using Deep Learning," 2019 IEEE Conference on Multimedia Information Processing and Retrieval (MIPR), San Jose, CA, USA, 2019, pp. 270-273, doi: 10.1109/MIPR.2019.00055
[10] J. Han, Y. Zhang and Z. Sun, "Key Frame Extraction Based on Sports Statistical Characteristics," 2022 IEEE Asia-Pacific Conference on Image Processing, Electronics and Computers (IPEC), Dalian, China, 2022, pp. 908-911, doi: 10.1109/IPEC54454.2022.9777583.
[11] K. Davila, F. Xu, S. Setlur and V. Govindaraju, "FCN-LectureNet: Extractive Summarization of Whiteboard and Chalkboard Lecture Videos," in IEEE Access, vol. 9, pp. 104469-104484, 2021, doi: 10.1109/ACCESS.2021.3099427.
[12] G. Mujtaba, A. Malik and E. -S. Ryu, "LTC-SUM: Lightweight Client-Driven Personalized Video Summarization Framework Using 2D CNN," in IEEE Access, vol. 10, pp. 103041-103055, 2022, doi: 10.1109/ACCESS.2022.3209275.
[13] A. Tonge and S. D. Thepade, "S-VSUM: Static Video Content SUMmarization using CNN," 2022 International Conference on Signal and Information Processing (IConSIP), Pune, India, 2022, pp. 1-5, doi: 10.1109/ICoNSIP49665.2022.10007516.
[14] H. Boukadida, S. -A. Berrani and P. Gros, "Automatically Creating Adaptive Video Summaries Using Constraint Satisfaction Programming: Application to Sport Content," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 27, no. 4, pp. 920-934, April 2017, doi: 10.1109/TCSVT.2015.2513678.
[15] Chianese,A.,Miscioscia,R.,Moscato,V.,Parlato, S.,&Picariello, A.(2004, August). A fuzzy approach to video scene detection and its application for soccer matches. In Proceedings of the 4th International Conference on Intelligent Systems Design and Application
[16] M. Merler et al., "Automatic Curation of Sports Highlights Using Multimodal Excitement Features," in IEEE Transactions on Multimedia, vol. 21, no. 5, pp. 1147-1160, May 2019, doi: 10.1109/TMM.2018.2876046.
[17] C. Lin and Y. Chen, "Sports Video Summarization with Limited Labeling Datasets Based on 3D Neural Networks," 2019 16th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Taipei, Taiwan, 2019, pp. 1-6, doi: 10.1109/AVSS.2019.8909872.
[18] C. Yan, X. Li and G. Li, "A New Action Recognition Framework for Video Highlights Summarization in Sporting Events," 2021 16th International Conference on Computer Science & Education (ICCSE), Lancaster, United Kingdom, 2021, pp. 653-666, doi: 10.1109/ICCSE51940.2021.9569708
[19] P. Kadam et al., "Recent Challenges and Opportunities in Video Summariza-

tion With Machine Learning Algorithms," in IEEE Access, vol. 10, pp. 122762-122785, 2022, doi: 10.1109/ACCESS.2022.3223379.
[20] M. Sanabria, F. Precioso and T. Menguy, "Profiling Actionsfor Sport Video Summarization: An attention signal analysis," 2020 IEEE 22nd International Workshop on Multimedia Signal Processing (MMSP), Tampere, Finland, 2020, pp. 1-6, doi: 10.1109/MMSP48831.2020.9287062.
[21] D. Xu, H. Fu, L. Wu, M. Jian, D. Wang and X. Liu, "Group Activity Recognition by Using Effective Multiple Modality Relation Representation With Temporal-Spatial Attention," in IEEE Access, vol. 8, pp. 65689-65698,2020, doi: 10.1109/ACCESS.2020.297974