IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

Customized Video Summarization with Thumbnail Containers and 2D CNN

Vaishnavi MN[1], Shashank Reddy[2], Kiran YC[3]
Students, Department of Information Science and Engineering[1,2]
Professor, Department of Information Science and Engineering[3]
Global Academy of Technology Bangalore, India

Abstract: *This project focuses on acquiring customized video summaries using thumbnail container-based summarization framework and 2D CNN model to select and extract specific features from thumbnails. This framework creates a custom keyshot summary for two or more concurrent users by leveraging the computing power of the user's device. The most advanced methods that collect and analyze complete video data to create video summaries consume a lot of computing power. In this context, we use the thumbnail containers framework which implements light thumbnails to manage the complex detection of events. This minimizes computational complexity and increases communication and storage performance by overcoming computational and privacy issues in resource constrained end-user environments. We aim on developing a user interactive customized video summarization tool which will be trained utilizing diverse datasets leading to the generation of personalized video summaries for feature-length videos.*

Keywords: Video summarization, 2D CNN, thumbnail containers, keyframes

I. INTRODUCTION

Video summarization has become increasingly crucial in recent years due to the explosive growth of multimedia content, particularly videos. As more users engage with smart devices capable of recording high-quality videos, and as social media platforms continue to serve as primary communication channels for billions of users, the volume of video content being generated and shared has reached unprecedented levels. This exponential increase in video content poses a significant challenge for users who seek to efficiently navigate through vast libraries of videos to find content relevant to their interests.

Video summarization addresses this challenge by providing condensed versions of full-length videos, extracting and presenting only the most meaningful segments. By creating succinct summaries, viewers can quickly grasp the essence of the video without needing to watch the entire duration.For instance, an extended cricket match video can be condensed to spotlight pivotal instances such as free hits and sixes. Traditionally, two methods have been employed: keyframes, which are static representations, and keyshots, which are dynamic summaries capturing continuous video segments. While keyframe-based summaries are lighter, they often sacrifice contextual information and original sound, making keyshot-based methods more preferred.

These summarization techniques cater to various video types, from short-form content like Tik-Toks to long-form content such as movies and sports matches. However, processing long-form videos poses computational challenges due to the vast amount of data involved. Deep learning approaches exacerbate this by requiring segmented processing, making them unsuitable for resource-constrained devices. Additionally, video summarization is inherently subjective, leading to the need for personalized summaries tailored to individual preferences. Yet, generating such summaries in real-time demands significant computational resources and raises privacy concerns when centralized servers handle user data as we embark on this data-driven journey, the goal is to transcend traditional boundaries of research and foster a holistic comprehension of the intricate web connecting population dynamics, public health outcomes, and economic landscapes. In doing so, we aim to provide a valuable resource for policymakers, researchers, and stakeholders striving to navigate the complexities of our evolving world

IJARSCT
ISSN (Online) 2581-9429
International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
Impact Factor: 7.53
Volume 4, Issue 1, February 2024

II. LITERATURE SURVEY

In [1],The keyframe generation method employs the Local Binary Pattern (LBP) operator to characterize image texture features, ensuring gray invariance and rotation invariance. The LBP operator is improved and integrated with color features for a comprehensive representation of image information. The techniques used in the study include Hierarchical Clustering, K-Means, and Silhouette Coefficient. Datasets utilized are the Open Video Project (OVP) Dataset and YouTube Dataset. The proposed method demonstrates superior performance in keyframe generation for video summarization, especially when compared to existing algorithms, across different datasets. The fusion of color and texture features enhances the representation of image information, leading to high-quality keyframes. The use of silhouette coefficients and hierarchical clustering contributes to improved clustering results.

In [2],Video summarization methodologies encompass frame selection, segmentation, clustering, feature extraction, temporal analysis, user perspective considerations, and hybrid approaches, collectively aiming to distill key content, enhance understanding, and cater to user preferences in a condensed form. Video summarization leverages machine learning algorithms, computer vision techniques, deep learning models, data compression technologies, and a range of computational tools to extract, analyze, and compress visual data for effective content condensation and efficient storage. Datasets utilized include Office, Lobby, VISIOCITY, Campus, TVSum, SumMe, Tour20, UT Ego, and YouTube 8M Segments. The integration of video summarization technology addresses the challenge of efficiently processing large volumes of visual data, providing a user-centric solution with diverse applications, particularly in domains such as surveillance, sports, and medical fields, while acknowledging ongoing challenges and opportunities for improvement.

In [3],FCN-LectureNet: Extractive Summarization of Whiteboard and Chalkboard Lecture Videos. Scene text detection involves adapting general object detection architectures and utilizing component and pixel-level approaches. Handwritten content extraction employs FCN-LectureNet, which integrates binarization, text mask estimation, and background estimation. Temporal segmentation introduces a novel method based on detecting major content deletion events, and key-frame selection relies on a spatial-temporal index of connected components (CCs). Deep learning, including architectures like SSD and U-NET, feature prominently. Scene text detection methods leverage fully convolutional networks (FCN) and instance segmentation, with a focus on transfer learning. Datasets used include Access Math and Lecture Math. Binarization methods are evaluated using DIBCO metrics over the AccessMath dataset, and an ablation study compares different methods over the LectureMath dataset. Specific experiment details or observations are not provided in the text.

In[4],ASoVS: Abstractive Summarization of Video Sequences. Methodologies involve the extraction of visual features using CNN, such as VGG-16, for multi-line video description. LSTM is utilized for language modeling, and a two-tier process is implemented for Subject-Verb-Object (SVO) trios. Recent adoption of deep learning methods combines CNN for visual information and RNN (LSTM) for text interpretation in video description. Abstractive text summarization methodologies include sequence-to-sequence models, attention mechanisms, pointer-generator networks, and the application of generative adversarial networks (GAN). Coverage mechanisms are implemented to prevent repetition. Technologies encompass Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM). Bidirectional LSTM and sequence-to-sequence learning models are applied for video description tasks, and attention mechanisms, Pointer Generator Networks, and Generative Adversarial Networks (GAN) are employed for abstractive text summarization. Evaluation measures consist of METEOR for video description, ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L) for abstractive summarization, and human evaluation for both tasks. Baseline models for comparison in video description include MP-LSTM, Semantic Compositional Networks (SCN-LSTM), Task-Specific Feature Encoding, and Multimodal Stochastic (MS) RNN. For abstractive summarization, baseline models include Pointer Generator Networks and a Generative Adversarial Network (GAN)

In [5],CNN and HEVC Video Coding Features for Static Video Summarization. Methodologies employed encompass feature extraction with CNNs and HEVC, dimensionality reduction using Sparse Autoencoder (SAE) and Stepwise Regression, frame elimination techniques based on low-level HEVC features and motion estimation/compensation, and video summarization using a random forest classifier. Technologies used in the study include Convolutional Neural Networks (CNNs) such as GoogleNet, AlexNet, Inception-ResNet-v2, and VGG16 for visual feature extraction, as well as the High-Efficiency Video Coding (HEVC) codec for video coding and low-level feature extraction. Datasets utilized are the Open Video Project (OVP) Dataset and VSUMM Dataset. Experiments involved testing the performance of feature extraction, dimensionality reduction, frame elimination, and video summarization techniques on benchmark datasets, comparing results with baseline models in video description and summarization tasks, such as MP-LSTM, SCN-LSTM, Task-Specific Feature Encoding, and Multimodal Stochastic RNN

In [6],The approach for dynamic video summarization, based on descriptors in space-time video volumes and Sparse Autoencoder, entails several methodologies. These methodologies comprise a thumbnail-based strategy aimed at enhancing computational efficiency, the utilization of a 2D CNN model for personalized event detection, transfer learning for model training, superframe segmentation to facilitate shot division, and the application of LPQ-TOP (Local Phase Quantization - Temporal Extension) and SAE (Sparse Autoencoder) for spatiotemporal and high-level feature extraction. In terms of technology, the proposed method makes use of HTTP 2.0 persistent connections, HLS (HTTP Live Streaming), FFmpeg, EfficientNet-B0, and an HTML5 HLS video player. The experimentation phase involves the UCF101 action recognition dataset, a set of 18 video titles, and the SumMe dataset. Experiments include evaluating an action recognition model on the UCF101 dataset, comparing the proposed LTC-SUM method with baseline approaches regarding computation time on devices with varying computational resources, and showcasing the efficacy of the proposed method in generating summaries for different video genres, catering to user preferences.

In [7],Spatiotemporal Modeling for Video Summarization Using Convolutional Recurrent Neural Network presents CRSum, an innovative approach integrating Convolutional Recurrent Neural Networks (CRNN) and 3D Convolutional Neural Networks (CNNs). CRSum employs Sobolev loss, a novel gradient-based content loss, to capture video temporal intricacies. It harmoniously fuses 3D deep and shallow features, predicting importance scores for video summarization. In ATS, CRSum leverages CRNN for robust spatiotemporal modeling, with 3D CNNs extracting features. Evaluations on SumMe, TVSum50, and VTW show CRSum's superior SumMe performance and competitiveness on TVSum50, with mini-batch Adam optimization and Sobolev loss highlighting its effectiveness. The experiments underscore Sobolev loss's pivotal role and nuanced contributions of feature types in CRSum's methodology.

In [8],Multi-Sensor Integration for Key-Frame Extraction from First-Person Videos addresses Key-Frame Extraction by selecting representative frames in diverse scenes. Sparse Representation, utilizing the l1-norm, reduces noise in First-Person Videos (FPV). Graph Modeling expresses conditional dependence structures, while Multi-Sensor Integration enhances key-frame selection using data from both video and motion sensors. For FPV video classification, summarization, and key-frame extraction, Sparse Modeling (SMRS), Graph Models, Pre-trained Deep Neural Network (DNN), Alternating Direction Method of Multipliers (ADMM), and Local Submodular Approximations-Trust Region (LSA-TR) were employed. The datasets CMU-MMAC (Carnegie Mellon University Multimodal Activity Database) and Daily Activities Dataset were utilized. Parameter tuning via cross-validation optimized parameters, including the regularization parameter (α) in sparse modeling. Algorithm comparison assessed key-frame extraction algorithms on datasets like CMU-MMAC, with specific experiments targeting datasets such as CMU-MMAC's brownie dataset and the daily activities dataset to analyze algorithm performance and optimize parameters

In [9],Summarization of Wireless Capsule Endoscopy Video Using Deep Feature Matching and Motion Analysis presents a method for summarizing Wireless Capsule Endoscopy (WCE) videos. The methodology involves shot segmentation based on feature matching and motion analysis. The Convolutional Autoencoder Neural Network (CANN) is utilized for unsupervised feature extraction, followed by shot segmentation through the analysis of dissimilarity measures and motion characteristics between consecutive frames. Keyframe extraction is then conducted based on motion analysis to remove redundant frames.

In [10],Video Summarization via Nonlinear Sparse Dictionary Selection introduces a methodology involving several key steps. Shot boundary detection identifies changes between frames, clustering groups similar frames, motion analysis focuses on motion information, sparse representation selects a subset of frames, and deep learning leverages neural networks for summarization. The video summarization process employs various keyframe extraction techniques, including shot boundary-based, clustering-based, motion-based, sparse representation-based, and deep learning-based methods.

In [11],A Survey of Content-Aware Video Analysis for Sports" provides insights into methodologies employed in sports video analysis. These methodologies encompass object extraction, tracking, naming, and action recognition.

Object extraction techniques utilize probabilistic models, Dynamic Bayesian Networks (DBNs), and trajectory-based approaches. Tracking involves region-based detection, template matching, particle filters, and camera calibration algorithms. Naming methods include face recognition, text recognition on athletes' clothing, and player number recognition. Action recognition encompasses feature representation, classifier learning, and deep learning approaches.

In [12],The Real-Time Event-Driven Road Traffic Monitoring System Using CCTV Video Analytics employs a methodology that encompasses various steps. These include the utilization of synthetic data for training the DCNN model, image augmentation for data preprocessing, and an event-driven approach for efficient video summarization. Moreover, the system integrates security and privacy measures to adhere to EU-GDPR regulations.

In [13],Discovery of Shared Semantic Spaces for Multiscene Video Query and Summarization," the proposed framework utilizes latent Dirichlet allocation (LDA) to learn local scene activities, spectral clustering for multilayer scene clustering, and k-center summarization for multiscene video summarization. It introduces shared activity topic bases (STBs) to represent activities across scenes and employs cross-scene query-by-example and classification techniques.

In [14],The Intermedia-Based Video Adaptation System employs a methodology involving various steps. These include the extraction of motion data through Multiple-QP Motion Estimation, the utilization of H.264/AVC and DXTC for texture compression, the application of the ρ-domain model for accurate rate control, the capture of structural characteristics through shot detection and key frame extraction, and the definition of ROI areas based on Attention Objects (AOs). The proposed video adaptation framework leverages technologies such as Multiple-QP Motion Estimation, H.264/AVC-encoded bit streams, DirectX Texture Compression (DXTC), the ρ-domain model for rate control, and a flexible Group of Pictures (GOP) structure for structural description.

Table 1: Table Analysis

Sl no	Author/ year	Research /Work Paper	Methodology	Technique	Dataset/ Input	Experiment/ Observation
1	FengsuiWang,JingangChen,Furong Liu (2021)	Keyframe Generation Method via Improved Clustering and Silhouette Coefficient for Video Summarization	The keyframe generation method employs the Local Binary Pattern (LBP) operator to characterize image texture features, ensuring gray invariance and rotation invariance. The LBP operator is improved and integrated with color features for a comprehensive representation of image information.	The techniques used in the study include Hierarchical Clustering, K-Means and Silhouette Coefficient	Open Video Project (OVP) Dataset and YouTube Dataset	The proposed method demonstrates superior performance in keyframe generation for video summarization, especially when compared to existing algorithms, across different datasets. The fusion of color and texture features enhances the representation of image information, leading to high-quality keyframes. The use of silhouette coefficients and hierarchical clustering contributes to improved clustering results.

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 1, February 2024

2	Payal kadam, Deepalivora, Sashikala mishraet al. (2022)	Recent Challenges and Opportunities in Video Summarization With Machine Learning Algorithms	Video summarization methodologies encompass frame selection, segmentation, clustering, feature extraction, temporal analysis, user perspective considerations, and hybrid approaches, collectively aiming to distill key content, enhance understanding, and cater to user preferences in a condensed form.	Video summarization leverages machine learning algorithms, computer vision techniques, deep learning models, data compression technologies, and a range of computational tools to extract, analyze, and compress visual data for effective content condensation and efficient storage.	Office [30] Lobby [28] VISIOCITY [77] Campus [26] TVSum [64] SumMe [63] Tour20 [29] UT Ego [27] YouTube 8M Segments [64]	The integration of video summarization technology addresses the challenge of efficiently processing large volumes of visual data, providing a user-centric solution with diverse applications, particularly in domains such as surveillance, sports, and medical fields, while acknowledging ongoing challenges and opportunities for improvement.
3	Kenny Davila, Fei Xu,SrirangarajSetlur, Venu Govindaraju (2021)	FCN-LectureNet: Extractive Summarization of Whiteboard and Chalkboard Lecture Videos	Scene text detection involves adapting general object detection architectures and utilizing component and pixel-level approaches. Handwritten content extraction employs FCN-LectureNet, which integrates binarization, text mask estimation, and background estimation. Temporal segmentation introduces a novel method based on detecting major content deletion events, and key-frame selection relies on a spatial-temporal index of connected components (CCs).	Deep learning, including architectures like SSD and U-NET, feature prominently. Scene text detection methods fully leverage convolutional networks (FCN) and instance segmentation, with a focus on transfer learning.	AccessMath, LectureMath	Binarization methods are evaluated using DIBCO metrics over the AccessMath dataset, and an ablation study compares different methods over the LectureMath dataset. Specific experiment details or observations are not provided in the text.

4	ANIQA DILAWARI,MUHAMMAD USMAN GHANI KHAN (2019)	ASoVS: Abstractive Summarization of Video Sequences	Methodologies include the extraction of visual features using CNN, such as VGG-16, for multi-line video description. LSTM is utilized for language modeling, and a two-tier process is implemented for Subject-Verb-Object (SVO) trios. Recent adoption of deep learning methods combines CNN for visual information and RNN (LSTM) for text interpretation in video description. Abstractive text summarization methodologies involve sequence-to-sequence models, attention mechanisms, pointer-generator networks, and the application of generative adversarial networks (GAN). Coverage mechanisms are implemented to prevent repetition.	The technologies involved in the research encompass Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory (LSTM). Bidirectional LSTM and sequence-to-sequence learning models are applied for video description tasks, and attention mechanisms, Pointer Generator Networks, and Generative Adversarial Networks (GAN) are employed for abstractive text summarization.	Video Description, Abstractive Text Summarization	Evaluation measures consist of METEOR for video description, ROUGE metrics (ROUGE-1, ROUGE-2, ROUGE-L) for abstractive summarization, and human evaluation for both tasks. Baseline models for comparison in video description include MP-LSTM, Semantic Compositional Networks (SCN-LSTM), Task-Specific Feature Encoding, and Multimodal Stochastic (MS) RNN. For abstractive summarization, baseline models include Pointer Generator Networks and a Generative Adversarial Network (GAN).
5	OBADA ISSA,TAMER SHANABLEH (2022)	CNN and HEVC Video Coding Features for Static Video Summarization	Methodologies employed encompassed feature extraction with CNNs and HEVC, dimensionality reduction using Sparse Autoencoder (SAE) and Stepwise Regression, frame elimination techniques based on	Technologies used in the study include Convolutional Neural Networks (CNNs) such as GoogleNet, AlexNet, Inception-ResNet-v2, and VGG16 for	OVP (Open Video Project) Dataset,VSUMM Dataset	Experiments involved testing the performance of feature extraction, dimensionality reduction, frame elimination, and video summarization techniques on benchmark datasets, comparing results with baseline models

			low-level HEVC features and motion estimation/compensation, and video summarization using a random forest classifier.	visual feature extraction, as well as the High-Efficiency Video Coding (HEVC) codec for video coding and low-level feature extraction.		in video description and summarization tasks, such as MP-LSTM, SCN-LSTM, Task-Specific Feature Encoding, and Multimodal Stochastic RNN.
6	JESNA MOHAN,(Student member, IEEE), MADHU S. NAIR (2018)	Dynamic Summarization of Videos Based on Descriptors in Space-Time Video Volumes and Sparse Autoencoder	Methodologies include a thumbnail-based approach for computational efficiency, a 2D CNN model for personalized event detection, transfer learning for model training, superframe segmentation for shot division, and the use of LPQ-TOP (Local Phase Quantization - Temporal Extension) and SAE (Sparse Autoencoder) for spatiotemporal and high-level feature extraction.	The proposed method leverages technologies such as HTTP 2.0 persistent connections, HLS (HTTP Live Streaming), FFmpeg, EfficientNet-B0, and an HTML5 HLS video player.	UCF101 action recognition dataset,Set of 18 video titles,SumMe dataset	Experiments involve evaluating an action recognition model on the UCF101 dataset, comparing the proposed LTC-SUM method with baseline approaches in terms of computation time on high and low computational resource devices, and demonstrating the efficiency of the proposed method in generating summaries for various video genres based on user preferences.
7	YUAN YUAN, (Senior Member, IEEE), HAOPENG LI,QI WANG, (Senior Member, IEEE) (2019)	Spatiotemporal Modeling for Video Summarization Using Convolutional Recurrent Neural Network	CRSum introduces a distinctive architecture that utilizes CRNN and 3D CNNs. It employs Sobolev loss, a novel gradient-based content loss function, to better model the temporal structure of video during training. The network fuses learnable 3D deep features and shallow features, allowing it to learn and predict importance scores for video summarization	The proposed method, CRSum, leverages Convolutional Recurrent Neural Networks (CRNN) and 3D Convolutional Neural Networks (CNNs) for video summarization. CRNN combines the	SumMe,TVSum50,VTW	The experiments involve 5-fold cross-validation, and videos are preprocessed by sub-sampling and resizing. The proposed model is trained using mini-batch Adam optimization, and the Sobolev loss is compared to Mean Squared Error (MSE) loss. CRSum achieves the best performance on SumMe and competitive results

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 1, February 2024

			in an end-to-end manner the ATS task	strengths of CNN and Recurrent Neural Networks (RNN) for spatiotemporal modeling, while 3D CNNs are used to extract spatiotemporal features directly from videos.		on TVSum50, demonstrating its effectiveness. The experiments also highlight the importance of Sobolev loss and the role of different types of features in the proposed method.
8	YUJIE LI (Member, IEEE), ATSUNORI KANEMURA (Member, IEEE), HIDEKI ASOH (Member, IEEE), TAIKI MIYANISHI,MOTOAKI KAWANABE(2020)	Multi-Sensor Integration for Key-Frame Extraction From First-Person Videos	Key-Frame Extraction involved selecting representative frames from various scenes, Sparse Representation used the l1-norm for noise reduction in FPV videos, Graph Modeling expressed conditional dependence structures, and Multi-Sensor Integration enhanced key-frame selection with data from both video and motion sensors.	Sparse Modeling (SMRS), Graph Models, Pre-trained Deep Neural Network (DNN), Alternating Direction Method of Multipliers (ADMM), and Local Submodular Approximations -Trust Region (LSA-TR) were harnessed for FPV video classification, summarization, and key-frame extraction.	CMU-MMAC (Carnegie Mellon University Multimodal Activity Database),Daily Activities Dataset	Parameter Tuning through cross-validation optimized parameters like the regularization parameter (α) in sparse modeling, algorithm comparison assessed key-frame extraction algorithms on datasets like CMU-MMAC, and multi-sensor integration impact was explored for quality improvement in FPV videos. Specific experiments targeted datasets such as CMU-MMAC's brownie dataset and the daily activities dataset, analyzing algorithm performance and optimizing parameters.
9	SUSHMA AND P. APARNA, (Senior Member, IEEE) (2020)	Summarization of Wireless Capsule Endoscopy Video Using Deep Feature Matching and Motion	The method involves shot segmentation based on feature matching and motion analysis in WCE videos. The Convolutional Autoencoder Neural	The proposed method utilizes Convolutional Autoencoder Neural Network (CANN) for feature extraction in	KID Dataset, Dataset-2 from the Department of Gastroenterology,	The experiments involve training the Convolutional Autoencoder Neural Network (CANN) using around 50,000 WCE frames from the datasets.

		Analysis	Network (CANN) is employed for unsupervised feature extraction. Shot segmentation is performed by analyzing the dissimilarity measure and motion characteristics between consecutive frames. Keyframe extraction is then carried out based on motion analysis to eliminate redundant frames.	Wireless Capsule Endoscopy (WCE) videos. It leverages technologies such as Sparse Modeling, Graph Models, and Alternating Direction Method of Multipliers (ADMM) for video summarization. Additionally, the method employs motion analysis using features like motion score, motion direction, and motion energy.	Manipal Hospital, Bangalore, India	Batchwise training is performed, and the network is trained to minimize the mean squared error (MSE) between input and output over all image samples. Evaluation is conducted using keyframes located by an expert gastroenterologist as ground truth. Performance metrics include F-score and compression ratio, and comparisons are made with other WCE video summarization methods. Motion analysis experiments include testing different similarity thresholds and motion direction thresholds to optimize the summarization performance.
10	MINGYANG MA, SHAOHUI MEI, SHUAI WAN, ZHIYONG WANG,DAGAN FENG , (Fellow, IEEE) (2019)	Video Summarization via Nonlinear Sparse Dictionary Selection	Shot boundary detection identifies changes between frames, clustering groups similar frames, motion analysis focuses on motion information, sparse representation selects a subset of frames, and deep learning leverages neural networks for summarization.	Video summarization utilizes keyframe extraction techniques, including shot boundary-based, clustering-based, motion-based, sparse representation-based, and deep learning-based methods.	VSUMM Dataset,TVSum Dataset	Evaluation metrics for the VSUMM dataset include Precision, Recall, and F-score, while for the TVSum dataset, importance scores of uniform two-second shots are converted to keyframe scores, and the mean score of selected keyframes is proposed as a metric.
11	Huang-Chia Shih,	A Survey of Content-Aware	The methodologies involve object	The survey discusses	Object Extraction	Various experiments are reported, such as

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 1, February 2024

	(Member, IEEE) (2018)	Video Analysis for Sports	extraction, tracking, naming, and action recognition. Object extraction techniques use probabilistic models, DBNs, and trajectory-based approaches. Tracking involves region-based detection, template matching, particle filters, and camera calibration algorithms. Naming methods include face recognition, text recognition on athletes' clothes, and player number recognition. Action recognition encompasses feature representation, classifier learning, and deep learning approaches.	various technologies for sports video analysis, including object extraction techniques using probabilistic multimedia objects (Multijects), dynamic Bayesian networks (DBN), and object segmentation algorithms.	and Tracking,Naming Objects,Action Recognition	shot boundary detection using color-based methods, object tracking using particle filters, trajectory-based ball detection for basketball videos, and action recognition using deep learning models. Observations include the detection of critical events, play-and-break analysis, and the use of keyframes and highlights for summarizing sports videos. Benchmark datasets are used for evaluating the performance of algorithms in action recognition and event detection. Accompanied by error analysis to identify common error patterns and challenges.
12	MEHWISH TAHIR, (Graduate Student Member, IEEE), YUANSONG QIAO, (Member, IEEE),NADIA KANWAL, (Senior Member, IEEE), BRIAN LEE,	Real-Time Event-Driven Road Traffic Monitoring System Using CCTV Video Analytics	The methodology involves the utilization of synthetic data for training the DCNN model, image augmentation for data preprocessing, and an event-driven approach for efficient video summarization. The system also integrates security and privacy measures to comply with EU-GDPR.	The proposed system employs technologies such as Deep Convolutional Neural Networks (DCNN), CCTV cameras, and smart city infrastructure for real-time event-driven road traffic monitoring, accident detection, and	VSUMM Dataset,SumMe	In one experiment, the system achieved an overall accuracy of 82.3% on real-time CCTV videos. Video summarization reduced the duration of five test videos from 56.3 seconds to 43.3 seconds (23.1%), demonstrating the efficiency of the proposed approach in capturing relevant events.

ISSN 2581-9429 IJARSCT

	MAMOONA N. ASGHAR, (Senior Member, IEEE) (2023)			video summarization.		
13	Xun Xu, Timothy M. Hospedales,Shaogang Gong (2017)	Discovery of Shared Semantic Spaces for Multiscene Video Query and Summarization	The proposed framework leverages latent Dirichlet allocation (LDA) to learn local scene activities, spectral clustering for multilayer scene clustering, and k-center summarization for multiscene video summarization. It introduces shared activity topic bases (STBs) to represent activities across scenes and employs cross-scene query-by-example and classification techniques.	The paper discusses the use of technologies such as latent Dirichlet allocation (LDA) for learning local scene activities, spectral clustering for multilayer scene clustering, and k-center summarization for multiscene video summarization in the context of surveillance scene understanding.	Multiscene Surveillance Dataset,Real Traffic Surveillance Videos,Junction and Roundabout Dataset	The experiments involve learning local activities using LDA, clustering scenes based on semantic similarity, and discovering shared activity topic bases (STBs) within scene clusters. The proposed framework is evaluated for cross-scene query-by-example, classification, and multiscene summarization using the collected dataset, demonstrating its effectiveness in surveillance scene understanding. The experiments also include annotation of behaviors in the dataset and assessing annotation consistency among multiple annotators.
14	Dong Zhang, Bin Li, Houqiang Li(2012)	Intermedia-Based Video Adaptation System: Design and Implementation	The methodology involves extracting motion data through Multiple-QP Motion Estimation, employing H.264/AVC and DXTC for texture compression, utilizing the ρ-domain model for accurate rate control,	The proposed video adaptation framework leverages technologies such as Multiple-QP Motion Estimation, H.264/AVC-encoded bit		The study reveals CE-BERT is an efficient Twitter rumor detection model with reduced computational requirements, outperforming state-of-the-art models in source text scenarios.

			capturing structural characteristics through shot detection and key frame extraction, and defining ROI areas based on Attention Objects (AOs).	streams, DirectX Texture Compression (DXTC), the ρ-domain model for rate control, and a flexible Group Of Pictures (GOP) structure for structural description.		

III. CONCLUSION

According to the literature survey on customized video summarization, there is a growing need for personalized video summaries amidst the exponential growth of multimedia content. By employing a thumbnail-based approach and a 2D CNN model, as per the survey the aim is to efficiently condense lengthy videos into keyshot summaries customized to individual user preferences. This approach leverages the computing power of end-user devices, minimizing computational complexity while enhancing communication and storage efficiencies. By offering a user-driven, interactive video summarization tool, it not only enhances the user experience but also addresses privacy concerns associated with centralized data handling. Overall, this initiative represents a significant step towards facilitating efficient and personalized video content consumption in today's multimedia-rich landscape.In addition to catering to the escalating demand for personalized video summaries, according to the survey it also tackles the challenges posed by the sheer volume of video content generated daily across various platforms. With the proliferation of smart devices capable of high-quality video recording and the widespread use of social media channels for communication, users are inundated with an overwhelming amount of video material. This surge in content consumption underscores the critical importance of effective video summarization techniques.

REFERENCES

[1] FengsuiWang,JingangChen,Furong Liu. (2021) "Keyframe Generation Method via Improved Clustering and Silhouette Coefficient for Video Summarization"
[2] Payal kadam, Deepali vora, Sashikalamishraetal.(2022) "Recent Challenges and Opportunities in Video Summarization With Machine Learning Algorithms"
[3] Kenny Davila, Fei Xu,SrirangarajSetlur,Venu Govindaraju (2021)"FCN-LectureNet: Extractive Summarization of Whiteboard and Chalkboard Lecture Videos"
[4] ANIQA DILAWARI,MUHAMMAD USMAN GHANI KHAN.(2019) "ASoVS: Abstractive Summarization of Video Sequences"
[5] OBADA ISSA,TAMER SHANABLEH (2022) "CNN and HEVC Video Coding Features for Static Video Summarization"
[6] JESNA MOHAN,(Student member, IEEE), MADHU S. NAIR (2018) "Dynamic Summarization of Videos Based on Descriptors in Space-Time Video Volumes and Sparse Autoencoder"
[7] YUAN YUAN, (Senior Member, IEEE), HAOPENG LI,QI WANG, (Senior Member, IEEE)(2019) "Spatiotemporal Modeling for Video Summarization Using Convolutional Recurrent Neural Network"
[8] YUJIE LI (Member, IEEE), ATSUNORI KANEMURA (Member, IEEE), HIDEKI ASOH (Member, IEEE), TAIKI MIYANISHI,MOTOAKI KAWANABE(2020) "Multi-Sensor Integration for Key-Frame Extraction From First-Person Videos"
[9] B. SUSHMA AND P. APARNA, (Senior Member, IEEE) "Summarization of Wireless Capsule Endoscopy Video Using Deep Feature Matching and Motion Analysis"

IJARSCT

ISSN (Online) 2581-9429

International Journal of Advanced Research in Science, Communication and Technology (IJARSCT)

International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal

Impact Factor: 7.53

Volume 4, Issue 1, February 2024

[10] MINGYANG MA, SHAOHUI MEI, SHUAI WAN, ZHIYONG WANG,DAGAN FENG , (Fellow, IEEE)(2019) "Video Summarization via Nonlinear Sparse Dictionary Selection"

[11] Huang-Chia Shih, (Member, IEEE) (2018) "A Survey of Content-Aware Video Analysis for Sports"

[12] MEHWISH TAHIR, (Graduate Student Member, IEEE), YUANSONG QIAO, (Member, IEEE),NADIA KANWAL, (Senior Member, IEEE), BRIAN LEE, MAMOONA N. ASGHAR, (Senior Member, IEEE)(2023) "Real-Time Event-Driven Road Traffic Monitoring System Using CCTV Video Analytics"

[13] Xun Xu, Timothy M. Hospedales,Shaogang Gong (2017) "Discovery of Shared Semantic Spaces for Multiscene Video Query and Summarization"

[14] Dong Zhang, Bin Li, HouqiangLi(2012) "Intermedia-Based Video Adaptation System: Design and Implementation"