

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

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

Volume 4, Issue 1, February 2024

Customized Video Summarization with Thumbnail Containers and 2D CNN

Vaishnavi MN¹, Shashank Reddy², Kiran YC³

Students, Department of Information Science and Engineering^{1,2} Professor, Department of Information Science and Engineering³ 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

Copyright to IJARSCT www.ijarsct.co.in





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

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

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 Networks (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

Copyright to IJARSCT www.ijarsct.co.in





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

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

Volume 4, Issue 1, February 2024

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 11-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, narrings and action recognition.

Copyright to IJARSCT www.ijarsct.co.in





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

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

Volume 4, Issue 1, February 2024

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.

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

 Table 1: Table Analysis

Copyright to IJARSCT www.ijarsct.co.in





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

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

tor: <mark>7.53</mark>

Volume 4, Issue 1, February 2024

2	D1	Decent	Vila	17:1	066 [20]	The inter the C
2	Payal	Recent	Video summarization	Video	Office [30]	The integration of
	kadam,	Challenges and	methodologies	summarization	Lobby [28]	video summarization
	Deepalivor	Opportunities	encompass frame	leverages	VISIOCITY	technology addresses
	a,	in Video	selection,	machine	[77]	the challenge of
	Sashikala	Summarization	segmentation,	learning	Campus [26]	efficiently processing
	mishraet	With Machine	clustering, feature	algorithms,	TVSum [64]	large volumes of
	al.	Learning	extraction, temporal	computer vision	SumMe [63]	visual data, providing
	(2022)	Algorithms	analysis, user	techniques,	Tour20 [29]	a user-centric
			perspective	deep learning	UT Ego [27]	solution with diverse
			considerations, and	models, data	YouTube	applications,
			hybrid approaches,	compression	8M	particularly in
			collectively aiming to	technologies,	Segments	domains such as
			distill key content,	and a range of	[64]	surveillance, sports,
			enhance	computational		and medical fields,
			understanding, and	tools to extract,		while acknowledging
			cater to user	analyze, and		ongoing challenges
			preferences in a	compress visual		and opportunities for
			condensed form.	data for		improvement.
				effective		_
				content		
				condensation		
				and efficient		
				storage.		
3	Kenny	FCN-	Scene text detection	Deep learning,	AccessMath,	Binarization methods
	Davila, Fei	LectureNet:	involves adapting	including	LectureMath	are evaluated using
	Xu,Srirang	Extractive	general object	architectures		DIBCO metrics over
	arajSetlur,	Summarization	detection	like SSD and		the AccessMath
	Venu	of Whiteboard	architectures and	U-NET, feature		dataset, and an
	Govindara	and Chalkboard	utilizing component	prominently.		ablation study
	ju	Lecture Videos	and pixel-level	Scene text		compares different
	(2021)		approaches.	detection		methods over the
			Handwritten content	methods		LectureMath dataset.
			extraction employs	leverage fully		Specific experiment
			FCN-LectureNet,	convolutional		details or
			which integrates	networks (FCN)		observations are not
			binarization, text	and instance		provided in the text.
			mask estimation, and	segmentation,		*
			background	with a focus on		
			estimation. Temporal	transfer		
			segmentation	learning.		
			introduces a novel	-		
			method based on			
			detecting major			
			content deletion			
			events, and key-			
			frame selection relies			
			on a spatial-temporal			
			index of connected		RESEARCH	11 50055
			components (CCs).		ISS	N
Сор	yright to IJAF	RSCT	DOI: 10.48175/IJA	RSCT-15397	2581- JJAR	9429
	w.ijarsct.co.				ALL AND	
					Marin •	<u>199</u>



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

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

Volume 4, Issue 1, February 2024

4	ANIQA	ASoVS:	Methodologies	The	Video	Evaluation measures
	DILAWA	Abstractive	include the extraction	technologies	Description,	consist of METEOR
	RI,MUHA	Summarization	of visual features	involved in the	Abstractive	for video description,
	MMAD	of Video	using CNN, such as	research	Text	ROUGE metrics
	USMAN	Sequences	VGG-16, for multi-	encompass	Summarizati	(ROUGE-1,
	GHANI	*	line video	Convolutional	on	ROUGE-2, ROUGE
	KHAN		description. LSTM is	Neural		L) for abstractive
	(2019)		utilized for language	Networks		summarization, and
	`		modeling, and a two-	(CNN),		human evaluation fo
			tier process is	Recurrent		both tasks.
			implemented for	Neural		Baseline models fo
			Subject-Verb-Object	Networks		comparison in video
			(SVO) trios. Recent	(RNN), and		description include
			adoption of deep	Long Short-		MP-LSTM, Semanti
			learning methods	Term Memory		Compositional
			combines CNN for	(LSTM).		Networks (SCN
			visual information	Bidirectional		LSTM), Task
			and RNN (LSTM) for	LSTM and		Specific Featur
			text interpretation in	sequence-to-		Encoding, and
			video description.	sequence		Multimodal
			Abstractive text	learning models		Stochastic (MS
			summarization	are applied for		RNN. For abstractiv
			methodologies	video		summarization,
			involve sequence-to-	description		baseline model
			sequence models,	tasks, and		include Pointe
			attention	attention		Generator Network
			mechanisms, pointer-	mechanisms,		and a Generative
			generator networks,	Pointer		Adversarial Network
			and the application of	Generator		(GAN).
			generative adversarial	Networks, and		
			networks (GAN).	Generative		
			Coverage	Adversarial		
			mechanisms are	Networks		
			implemented to	(GAN) are		
			prevent repetition.	employed for		
				abstractive text		
				summarization.		
5	OBADA	CNN and	Methodologies	Technologies	OVP (Open	Experiments involved
	ISSA,TA	HEVC Video	employed	used in the	Video	testing the
	MER	Coding	encompassed feature	study include	Project)	performance o
	SHANAB	Features for	extraction with CNNs	Convolutional	Dataset,VS	feature extraction
	LEH	Static Video	and HEVC,	Neural	UMM	dimensionality
	(2022)	Summarization	dimensionality	Networks	Dataset	reduction, frame
			reduction using	(CNNs) such as		elimination, and
			Sparse Autoencoder	GoogleNet,		video summarization
			(SAE) and Stepwise	AlexNet,		techniques of
			Regression, frame	Inception-		benchmark datasets
			elimination	ResNet-v2, and		comparing result
		1	techniques based on	VGG16 for	S ISS	Nwith baseline model
	yright to IJAF		DOI: 10.48175/IJA	DOOT 45007	2581-	9429 📓 73





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

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

Volume 4, Issue 1, February 2024

IJARSCT

			low-level HEVC features and motion estimation/compensat ion, and video summarization using a random forest classifier.	extraction, as well as the High-Efficiency		in video description and summarization tasks, such as MP- LSTM, SCN-LSTM, Task-Specific Feature Encoding, and Multimodal Stochastic RNN.
Me Sta me IE: M. S.	ESNA OHAN,(udent ember, EE), ADHU NAIR 018)	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	UCF101 action recognition dataset,Set of 18 video titles,SumM e 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.
YU (So Mo IE: HZ G WZ (So Mo IE:	UAN UAN, enior ember, EE), AOPEN LI,QI ANG, enior ember, EE) 019)	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	leverages Convolutional Recurrent Neural Networks (CRNN) and	SumMe,TV Sum50,VT W	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

Copyright to IJARSCT www.ijarsct.co.in





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

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

Volume 4, Issue 1, February 2024

	1		-			1
			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), ATSUNO RI KANEMU RA (Member, IEEE), HIDEKI ASOH (Member, IEEE), TAIKI MIYANIS HI,MOTO AKI KAWAN ABE(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 11-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),D aily Activities Dataset	ParameterTuning throughthroughcross- validationvalidationoptimizedparameterslikeregularizationparameter(α)insparsemodeling,algorithmcomparisonassessedkey-frameextractionalgorithmsondatasetslikeCMU-MMAC,cMU-MMAC,andmulti-sensorintegrationintegrationimpactwasexploredforqualityim FPVvideos.Specificexperimentstargeteddatasets suchasCMU-MMAC'sbrowniedatasetalgorithmperformanceperformanceandoptimizing
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 Gastroentero logy,	parameters.TheexperimentsinvolvetrainingtheConvolutionalAutoencoderNeuralNetwork(CANN)usingaround50,000WCEWCEframesthedatasets.

Copyright to IJARSCT www.ijarsct.co.in



expert

truth.

ratio.

other

video

Motion

testing

and

to

the

metrics

include

dataset.



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

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

IJARSCT

Network (CANN) is Wireless Manipal Batchwise training is Analysis performed, and the employed for Capsule Hospital, Bangalore, unsupervised feature Endoscopy network is trained to extraction. Shot (WCE) videos. India minimize the mean leverages squared error (MSE) segmentation is It performed by technologies between input and analyzing the such as Sparse output over all image dissimilarity measure samples. Evaluation Modeling, and motion Graph Models, is conducted using characteristics and Alternating keyframes located by Direction between consecutive an frames. Keyframe Method of gastroenterologist as extraction is then Multipliers ground (ADMM) carried out based on for Performance metrics video motion analysis to include F-score and eliminate redundant summarization. compression frames. Additionally, and comparisons are the method made with employs motion WCE analysis summarization using features like methods. motion score, analysis experiments motion include direction. different similarity and motion energy. thresholds motion direction thresholds optimize summarization performance. MINGYA Video Shot Video VSUMM 10 boundary Evaluation NG MA. Dataset, TVS Summarization detection identifies summarization for the VSUMM SHAOHU via Nonlinear utilizes um Dataset changes between dataset MEI. Sparse frames, clustering keyframe Precision, Recall, and I SHUAI Dictionary similar extraction F-score, while for the groups WAN, Selection frames, motion techniques, TVSum ZHIYON including shot analysis focuses on importance scores of G motion information, boundaryuniform two-second WANG,D sparse representation based, shots are converted to AGAN selects a subset of clusteringkeyframe scores, and FENG frames, and deep based, motionthe mean score of (Fellow, leverages based, selected keyframes is learning sparse representationproposed as a metric. IEEE) neural networks for (2019)summarization. based, and deep learning-based methods.

Volume 4, Issue 1, February 2024

Copyright to IJARSCT www.ijarsct.co.in

Chia Shih,

Huang-

А

Survey

Content-Aware

of

The

involve

11

DOI: 10.48175/IJARSCT-15397

object

The

discusses

survey

Object

methodologies



Extraction reported, such as

Various experiments





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

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

Volume 4, Issue 1, February 2024

IJARSCT

	f	Vila Australia			1	-1
· ·	Aember,	Video Analysis	extraction, tracking,	various	and	shot boundary
	EEE)	for Sports	naming, and action	technologies for	Tracking,Na	detection using color-
(20	018)		recognition. Object	sports video	ming	based methods,
			extraction techniques	analysis,	Objects,Acti	object tracking using
			use probabilistic	including object	on	particle filters,
			models, DBNs, and	extraction	Recognition	trajectory-based ball
			trajectory-based	techniques		detection for
			approaches. Tracking	using		basketball videos,
			involves region-based	probabilistic		and action
			detection, template	multimedia		recognition using
			matching, particle	objects		deep learning models.
			filters, and camera	(Multijects),		Observations include
			calibration	dynamic		the detection of
			algorithms. Naming	Bayesian		critical events, play-
			methods include face	networks		and-break analysis,
			recognition, text	(DBN), and		and the use of
			recognition on	object		keyframes and
			athletes' clothes, and	segmentation		highlights for
			player number	algorithms.		summarizing sports
			recognition. Action	-		videos. Benchmark
			recognition			datasets are used for
			encompasses feature			evaluating the
			representation,			performance of
			classifier learning,			algorithms in action
			and deep learning			recognition and event
			approaches.			detection.
			11			Accompanied by
						error analysis to
						identify common
						error
						patterns and challeng
						es.
2 M	EHWIS	Real-Time	The methodology	The proposed	VSUMM	In one experiment,
Н	TAHIR,	Event-Driven	involves the	system employs	Dataset,Sum	the system achieved
(G	Graduate	Road Traffic	utilization of	technologies	Me	an overall accuracy
Sti	udent	Monitoring	synthetic data for	-		of 82.3% on real-time
	ember,	System Using	training the DCNN	Convolutional		CCTV videos. Video
	EEE),	CCTV Video	model, image	Neural		summarization
	UANSO	Analytics	augmentation for data	Networks		reduced the duration
NO	G QIAO,	2	preprocessing, and an	(DCNN),		of five test videos
	Aember,		event-driven	CCTV cameras,		from 56.3 seconds to
ì	EE),NA		approach for efficient	and smart city		43.3 seconds
	IA		video summarization.	infrastructure		(23.1%),
	ANWAL		The system also	for real-time		demonstrating the
	(Senior		integrates security	event-driven		efficiency of the
M	lember,		and privacy measures	road traffic		proposed approach in
	EEE),		to comply with EU-	monitoring,		capturing relevant
	RIAN		GDPR.	accident	REBEARCH	exents.
	EE,			detection, and		N
	ght to IJAR	SCT	DOI: 10.48175/IJAI		2581-1 JARS	9429 1
	arsct.co.i				IJAR:	



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

,,					-	
13	MAMOO NA N. ASGHAR, (Senior Member, IEEE) (2023) Xun Xu, Timothy M. Hospedale	Discovery of Shared Semantic Spaces for	The proposed framework leverages latent Dirichlet allocation (LDA) to	video summarization.	Multiscene Surveillance Dataset,Real Traffic	The experiments involve learning local activities using LDA, clustering scenes
	s,Shaogan g Gong (2017)	Multiscene Video Query and Summarization	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.	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.	Surveillance Videos,Junct ion and Roundabout Dataset	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	ThemethodologyinvolvesextractingmotiondatadatathroughMultiple-QPMotionEstimation,employingH.264/AVCandDXTCfortexturecompression,utilizingthep-domainmodelforaccurateratecontrol,texture	The proposed video adaptation framework leverages technologies such as Multiple-QP Motion Estimation, H.264/AVC- encoded bit	All the second	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.

Copyright to IJARSCT www.ijarsct.co.in





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

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

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

Volume 4, Issue 1, February 2024

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, Srirangaraj Setlur, 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"

Copyright to IJARSCT www.ijarsct.co.in





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

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

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"

