Available online at www.ijpe-online.com

International Journal of Performability Engineering

vol. 19, no. 12, December 2023, pp. 817-823
DOI: 10.23940/ijpe.23.12.p6.817823

YouTube Video Summarizer using NLP: A Review

Yogendra Singh, Rishu Kumar, Soumya Kabdal, and Prashant Upadhyay*

Department of Computer Science and Engineering, Sharda University, Greater Noida, India

Abstract

This review paper delves into the emerging realm of YouTube video summarization utilizing Natural Language Processing (NLP) techniques, a critical area of research with increasing prominence in our multimedia-rich digital age. The paper commences with a broad overview of the field, elaborating on the need for automated video summarization tools to navigate and condense the massive, ever-growing sea of YouTube content. Further, we systematically scrutinize the role and implementation of NLP methods in extracting meaningful textual data from videos, focusing on video transcripts, closed captions, user comments, and associated metadata. Subsequent sections dissect seminal and recent works, studying various NLP techniques such as text summarization, sentiment analysis, topic modeling, and deep learning architectures employed in this context. The paper also focuses on the various metrics used for evaluation and shows datasets generally used to assess the performance of these summarization systems. Finally, we identify current challenges and potential future directions for research in the area, acknowledging the evolving landscape of online video platforms and AI technologies. This review aims to provide researchers and practitioners with an encompassing perspective on the pivotal role of NLP in enabling more efficient, accurate, and intuitive navigation of YouTube content ultimately shaping our digital consumption experiences.

Keywords: Natural Language Processing (NLP); YouTube video summarization; text summarization; video content analysis; artificial intelligence

1. Introduction

As one of the most popular video-sharing platforms worldwide, YouTube has become a vast ocean of information, covering an expansive range of subjects. This widespread proliferation of video content underscores the necessity for efficient and automated summarization tools that enable users to quickly understand the essence of a video without having to watch it in its entirety. This is where Natural Language Processing (NLP) steps in, offering a promising avenue for summarizing video content by processing and condensing associated textual information. This review paper aims to illuminate the intersection of NLP and YouTube video summarization, a rapidly growing field at the nexus of artificial intelligence and multimedia information retrieval.

The use of NLP to summarize YouTube videos has garnered increasing interest from both academia and industry, as it holds the potential to transform the way we navigate and consume online video content. By systematically exploring various aspects of this field, this paper endeavors to provide a comprehensive understanding of the state-of-the-art techniques, the challenges being faced, and the future trajectories that research in this area might take. Ultimately, this review serves as a repository of knowledge for researchers, engineers, and enthusiasts alike, embarking on a journey to delve deeper into the exciting possibilities that arise when the power of NLP is harnessed to summarize the endless stream of YouTube videos.

YouTube is an expansive global platform that provides a space for sharing and accessing a rich array of information. However, with over 500 hours of video uploaded every minute, the volume of content could be higher, leading to an overwhelming deluge of data that can be challenging for users to navigate. To make sense of this vast repository and efficiently find relevant content, there is an increasing need for automated tools to summarize video content, distilling it to its essential elements. This forms the backdrop for exploring the realm of YouTube video summarization using Natural Language Processing (NLP).
NLP a branch of artificial intelligence (AI) concerned with the interaction between computers and human language. By applying NLP techniques, we can extract meaningful information from the textual elements associated with a video - such as closed captions, transcripts, user comments, and metadata - and generate concise summaries that accurately capture the video's main ideas and points of interest.

* Corresponding author.
E-mail address: prashanttheace@gmail.com

Videos have become omnipresent across various domains such as first-person perspective, surveillance, sports, education, and entertainment. Technological progress and widespread access to video recording devices have led to an exponential surge in available video data. This influx poses network traffic, bandwidth, and browsing issues, as noted in [1, 2]. An enormous amount of data is produced every day due to the rise in video popularity. Finding meaningful information while sifting through countless video collections might be a difficult undertaking. The ability to extract semantic information from low-level audio or visual input is necessary to find relevant video material. As proposed in [3], automatic video summarization methods can solve these concerns, facilitating easier navigation and browsing of extensive video collections.

Internet browsing for information has become integral to various sectors such as education, recreation, and business. Efficient content indexing and access can expedite browsing and facilitate the retrieval of critical materials in a reduced timeframe, as outlined in [4]. Video summarization offers the capability to navigate through vast volumes of video data. According to [5], this procedure, sometimes described as the development of a reduced video version without sacrificing important information, helps offer a summary of longer movies. Personalized video summarization has become a cutting-edge method in multi-model summarising, as mentioned in [6]. As described in [7], and [8], attention approaches have also been applied to artificial video summary to facilitate online video data browsing and navigation.

This review paper provides a comprehensive overview of the utilization of NLP in YouTube video summarization, an emergent field of study that sits at the intersection of AI and multimedia information retrieval. While the techniques and principles of NLP have seen wide application in traditional text-based domains, their utilization in video summarization presents novel challenges and opportunities. This is due to the multidimensional nature of video content, which incorporates visual and auditory elements in addition to a range of textual data. The implementation of NLP techniques in YouTube video summarization is rapidly gaining traction, driven by the potential to revolutionize how users interact with video content. Such applications could make video content more accessible, enhance content discovery, and provide more efficient ways for individuals to glean relevant information from videos without needing to watch them in their entirety.

However, the path to implementing NLP in video summarization has its challenges. Challenges range from extracting and processing textual data from multimedia sources to the accurate representation and summarization of video content that might cover a wide variety of topics, languages, and contexts. Furthermore, the fast-paced evolution of both NLP technologies and the nature of YouTube content calls for an ongoing adaptation and refinement of techniques. With these challenges in mind, this review offers a detailed overview of the state of the art in YouTube video summarization using NLP.

The review will encompass the key techniques employed, the practical applications and results achieved, and the major challenges faced by researchers in this field. As one navigates through this review, one will witness the confluence of artificial intelligence, NLP, and multimedia information processing and explore their potential to shape the way digital content is consumed in the future. Figure 1 illustrates the foundational design of text-oriented video summarization. With the aid of widely used deep neural networks, visual attributes are gleaned from the video. Concurrently, textual attributes are sourced from the accompanying text in the context of video summarization. A combined embedding process is performed to uphold the semantic associations within the video, utilizing both visual and textual feature vectors. Consequently, a summary is produced, employing the most relevant frames that are arranged in a chronological sequence.

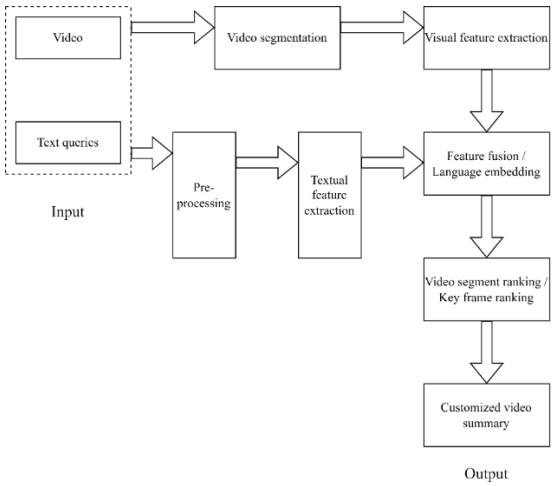

Figure 1. Fundamental framework for text-based video summarization

2. Literature Review

Recent years have seen a surge in research activity related to video summarization, with various techniques garnering significant results. Many of these techniques are based on methods that employ keyframes or are driven by structure. Keyframe-based techniques generate a summary using a selection of semantically relevant keyframes from the video. Their respective attention can influence the selection of keyframes. Their respective attention scores.
A model was created that included three basic algorithms, according to Hritvik Gupta et al [9]. We used TF-IDF, Bidirectional Encoder Representations from Transformers (BERT), and Latent Semantic Analysis (LSA). Using shortened Singular Value Decomposition (SVD), LSA makes extracting all relevant themes from a text easier. Contrarily, TF-IDF is used to pick important words from each phrase in a text. In this process, BERT is responsible for encoding the sentences and obtaining the positional embedding of topics. This work by Hritvik Gupta et al. [9] allows for a thorough comprehension of the TF-IDF algorithm. A separate approach for text summarization, presented by Swaranjali Jugran et al. [10], utilizes the tools NLTK and Spacy. NLTK is understood as a Python-based toolkit for processing natural/human languages, and Spacy is a well-known open-source package explicitly made for NLP tasks. Their study indicates that Spacy does text summarising more effectively than NLTK. Therefore, as demonstrated by their work, Spacy is selected as the preferred tool in our project. The work of Surabhi Adhikari et al. [11] provides comprehensive insights into various text summarization methods. These approaches, widely used in Natural Language Processing (NLP), include the BERT model, TF-IDF, Sentence Ranking, K-means clustering, and K-nearest neighbours. A model was presented by J. N. Madhuri et al. [12] to turn extractive text summaries of a document into audio. To summarise, the model used NLTK and sentence rating. However, there were shortcomings in their approach: it struggled with large documents and needed a mechanism to convert the summarized text into audio.

Kaiz Merchant et al. [13], on the other hand, provided a model for automatic text summarization based on latent semantic analysis and natural language processing (NLP). Their work specifically focused on legal judgments, where they attempted to transform lengthy judgments into shorter, more useful summaries. However, the fact that the model's evaluation was mostly dependent on word similarity rather than the complete notion was a fundamental flaw in their method. According to the study in [14], a video is treated as a comprehensive undirected graph. This video is partitioned into clusters by employing a cut detection algorithm, forming a temporal graph. Further scene detection in the video is achieved using a shortest-path algorithm. The video summary is then generated from the temporal graph by leveraging structural and attention information. In contrast, the study outlined in [15] introduces an approach based on super-frame segmentation for keyframe extraction. Here, each super-frame's visual appeal or 'interestingness' is assessed based on low, mid, and high-level features. The boundaries of these frames, which align with a cut, form an optimal summary based on their interestingness scores.

The research outlined in [2] leverages deep learning algorithms for video summarization, implementing techniques such as pipelining to dissect critical components of the video. The input video and the modified version are also compared throughout this process, and the correctness of the summarised text is assessed. [16] explores abstractive summarization using deep neural networks for video sequence summarization. The suggested joint model outperforms existing approaches in this scenario by enabling users to distinguish between pertinent and irrelevant information.

The method introduced in [17] sets itself apart from other research by initially categorizing videos into static and dynamic types before performing the required transcript translation. This process employs an algorithm known as short boundary Detection. The research in [18] is centered on creating a prototype for video indexing, tailored explicitly for lectures, utilizing Syntactic Similarity Measures. An auto-caption generator tool adds captions to the video by utilising dynamic programming approaches. A review of current video summarising approaches is also included in the paper. A real-time video summary approach for mobile platforms is given in the article referenced in [19]. This approach analyses the footage when a camera records it in real-time and simultaneously produces a summary. The method looks at both linked intrinsic video data, such as the video stream's content and related external video metadata, like the external camera information. According to the study in [20], a standard summarising system with basic capabilities may not fulfil user requests, necessitating the development of a bespoke system. The proposed method generates a video summary based on the user's choices and the top-scoring, semantically relevant films. In [21], and the summarization is performed using supervised learning for videos of a similar category. The summaries of the other films in the same subset are created once a collection of movies has been analysed and analyzed. Loss function cross-validation scores are employed, and every transition between frames in a movie is regarded as a state.

Figure 2 summarizes the core steps in a YouTube video summarizer using NLP, from input to user interaction. Diagramming tools or software can create a visual diagram based on this structure.

1) Video Input: Represents the input video content.
2) Video Processing (NLP): Denotes the application of NLP techniques to process the video content.

3) Feature Extraction: This involves extracting key visual and audio features from the video.
4) Text Generation: This involves generating textual summaries based on the extracted features.
5) Summarized Text Output: This represents the final output of the video summarization process.

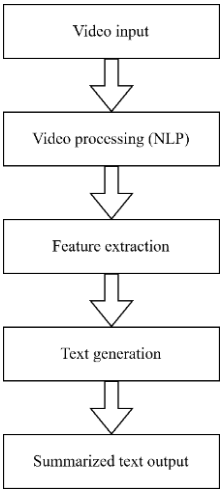

Figure 2. The flowchart for a YouTube Video summarizer using NLP

In this section, the authors contrast both supervised and unsupervised text-based video summarising techniques. Table 1 presents supervised approaches, and Table 2 describes unsupervised alternatives.

Table 1. Performance evaluation using guided techniques

Author	Dataset	Technique	Limitations
Huang et al. [22]	Dataset based on query-video pairs	Specialized attention network and GPT-2 (Static)-based contextualized word representations	Embedding dimension influences model training speed and effectiveness.
Narasimhan et al. [23]	Query-focused dataset for video summarizing	Bi-Modal Transformer for the creation of dense video captions and CLIP-It, a language-guided multimodal transformer (Static)	Unsuitable prejudices are embedded. However, these biases can be enhanced.
Huang et al. [24]	Summary of many models of video	Feature fusion and Dictionary-based BOW (Static)	Semantic understanding can be lacking in the BOW model due to its isolated word treatment.
Xiao et al. [25]	Query-focused dataset for video summarizing	QSAN, or Query-Biased Self-Attentive Network: A dynamic caption generator with reinforcement	Substantial pre-processing, such as curating and sanitizing caption data, is necessary.
Nalla et al. [26]	Dataset for query-focused video summarization	Local and global focus Dynamic feature fusion	In a feature fusion model, there's a possibility of some data getting lost.
Xiao et al. [27]	Video summarization dataset with a query focus	CNN, local media, and worldwide exposure	The complexity of computation is significant.
Jiang et al. [28]	Query-focused video summarization dataset	A variable autoencoder with a module for multilayer self-attention. Utilize user-oriented diversity and a stochastic (random) latent variable (Dynamic) for the diversity factor.	Queries are expressed as words, which can result in increased computational time.
Vasudevan et al. [29, 30]	Dataset with relevance and variety	Submodular combination of goals (LSTM) (Static)	Long videos can lead to a substantial increase in inference cost.
Sharghi et al. [31]	Query focused video summarization dataset	A parameterized sequential determinant point process in a memory network. (Dynamic)	Sequential processing leads to longer computational time.
Sharghi et al. [32]	TV episodes, UT egocentric	SH-DPP, short for Sequential and Hierarchical Determinant Point Process.	SH-DPP continues to demand considerable computational resources.

Table 2. Performance analysis based on unsupervised methods

Author	Dataset	Technique	Limitations
Zhang et al. [33]	Query focused video summarization dataset	Three-player Generative Adversarial Network	It's highly dependent on parameter decisions and is computationally demanding.
Zhang et al. [34]	Query-focused video summarization dataset	Network-based on deep reinforcement learning for summarizing	Too many states can negatively impact the results.
Sreeja et al. [35]	Films demonstrating vehicles and academic inspections	DL-based object detection and ontologies from the semantic web for query inferences	Focusing solely on key objects can hinder the proper handling of semantic relationships.

Sekh Arif et al. [36]	CCTV datasets from Sherbrook Street and VIRAT	Based on the clustering of tubes	Consideration is given solely to significant objects and their temporal motion.
Junyu Gao et al. [37]	Highlights from SumMe, TVSum, and YouTube	Relationship-aware hard assignments based on graph neural networks for choosing key clips and assignment-learning graph	Video graphs lack the representation of object and scene semantics.

Different models in qualitative research provide varied frameworks or lenses through which data is collected and interpreted. The choice of a model can shape the direction of the study, the type of questions asked, the way data is collected, and the ensuing interpretation. Table 3 shows the qualitative analysis done by various researchers using different models.

Table 3. Qualitative analysis of models

Author	Approach	Advantages	Drawbacks
Sandra Eliza Fontes de Avila et al. [38]	Using hue colour histograms as feature descriptors and K Mean clustering, this method chooses aesthetically different groups.	1. Colour histograms are low-level feature descriptors that are resilient to even minor changes in camera position. 2. This method uses fewer resources to get good results since clustering takes a lot less time and computing power than neural net structures.	1. The Clustering Algorithm disregards the chronological order 2. Colour histograms do not account for the orientation or the colour dispersion in the frame. As a result, a picture with identical colours scattered differently is compared to an image with same colours spread differently.
Ke Zhang et al. [39]	The model is built on the BiLSTM architecture. At each temporal step, A multi-layer perceptron receives the output from these LSTM layers together with a visual frame feature to produce either a binary frame label or a frame level significance score.	Because of its ability to retain prior information, LSTMs successfully simulate the changeable temporal dependencies.	1. Because recurring models computations cannot be parallelized, computing resources are well-spent. 2. Poor performance when simulating scenarios with dynamic changes because the structural or- der needs to be upheld.
Behrooz Mahas-seni et al. [40]	By utilizing a variable autoencoder to produce a corresponding summary and feeding it to the discriminator in an adversarial situation, the architecture trains a keyframe selector LSTM.	1. Unsupervised Approach eliminates the necessity for difficult to get annotated user data. 2. Acquires knowledge of intermediate representations, which may be valuable in other contexts. 3. In some circumstances, negative criticism may be more helpful than user-reported facts. 4. Regularisation might be applied in different places, focussing on different meanings.	1. It's well known that GANs may be trained in both time and memory. 2. Because the goal is to achieve global representation, subtle changes that can be significant to consumers are difficult to record. 3. User semantics need to be recorded while creating a video summary.
Mrigank Rochan et al. [41]	For video summarization, a fully convolutional model has been modified.	1. Since convolution models are independent of prior outcomes, they allow parallel computing. Comparing this to LSTM methods, training is made more efficient. 2. CNNs can represent the entire range at a considerably lower depth than LSTMs, enabling high level of network to context aggregation sooner. In LSTMs, the last node is only considered once.	1. Loss of resolution and low-level semantics induced by the convolutional layer stack and repeateddownsampling of the inputs biases the output towards contextual information while ignoring local knowledge. 2. Excessive upsampling spreads a limited number of values across a vast area, reducing uniqueness and making many nodes seem same.
Kaiyang Zhou et al. [42]	Encoder-Decoder Architecture training using reinforcement learning framework. Rewards are determined by the resulting summary's variety and representativeness.	Encoder-Decoder Architecture training using reinforcement learning framework. Rewards are determined by the resulting summary's variety and representativeness.[40]	1. A focus on diversity creates issues when there are slow or subtle changes. 2. LSTMs result in ineffective training. 3. Includes no compensation for long-term dependency.

Figure 3 [43] aims to display the daily metrics captured from video surveillance cameras. These cameras are strategically placed (often in public or private areas) to monitor activities, ensure security, or gather evidence. The data from these cameras can provide invaluable insights into various aspects, depending on the specific metrics presented in the figure, like data source, nature of data, purpose and interpretation. The data or videos captured by these cameras need to be summarized so that it can be possible to gather all the required information quickly.

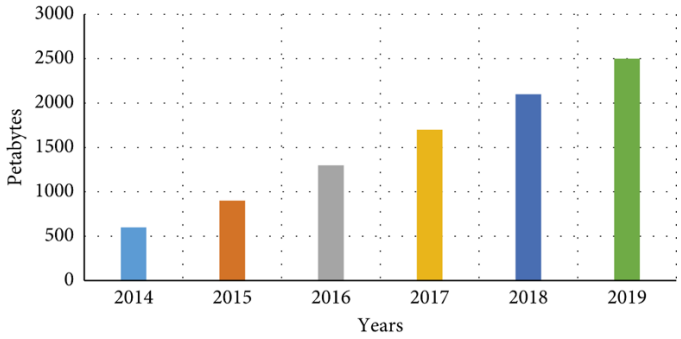

Figure 3. Data daily generated by surveillance cameras

3. Conclusion

This work provides insight into the development of text-based video summarization. We scrutinize various methods in this review and offer a detailed comparison to support ongoing research in this domain. The challenges and potential enhancements to existing strategies for the future are also touched upon. The critical insight derived from this study is the superior performance of supervised learning methods compared to unsupervised and weakly supervised techniques, which is contingent on the training effectiveness and dataset size. Past research is primarily constrained by limited support for broad-based queries, and there's significant scope for enhancement, particularly in multi-modal video summarization. This modest contribution aims to aid individuals in selecting an appropriate technique for personalized video summarization.

References

[1] Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., and Yokoya, N. Video Summarization using Deep Semantic Features. In *Computer Vision–ACCV 2016: 13th Asian Conference on Computer Vision, Taipei, Taiwan, November 20-24, 2016, Revised Selected Papers, Part V 13*, Springer International Publishing, pp. 361-377, 2017.

[2] Apostolidis, E., Adamantidou, E., Metsai, A.I., Mezaris, V., and Patras, I. Video Summarization using Deep Neural Networks: A Survey. *Proceedings of the IEEE*, vol. 109, no. 11, pp. 1838-1863, 2021.

[3] Zhang, S., Zhu, Y., and Roy-Chowdhury, A.K. Context-Aware Surveillance Video Summarization. *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5469-5478, 2016.

[4] Kwon, J. and Lee, K.M. A Unified Framework for Event Summarization and Rare Event Detection from Multiple Views. *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 9, pp. 1737-1750, 2014.

[5] Sridevi, M. and Kharde, M. Video Summarization using Highlight Detection and Pairwise Deep Ranking Model. *Procedia Computer Science*, vol. 167, pp. 1839-1848, 2020.

[6] Varini, P., Serra, G., and Cucchiara, R. Personalized Egocentric Video Summarization of Cultural Tour on User Preferences Input. *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2832-2845, 2017.

[7] Ji, Z., Xiong, K., Pang, Y., and Li, X. Video Summarization with Attention-Based Encoder–Decoder Networks. *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1709-1717, 2019.

[8] Fajtl, J., Sokeh, H.S., Argyriou, V., Monekosso, D., and Remagnino, P. Summarizing Videos with Attention. In *Computer Vision–ACCV 2018 Workshops: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers 14*, Springer International Publishing, pp. 39-54, 2019.

[9] Gupta, H. and Patel, M. Method of Text Summarization using LSA and Sentence Based Topic Modelling with Bert. In *2021 international conference on artificial intelligence and smart systems (ICAIS)*, IEEE, pp. 511-517, 2021.

[10] Jugran, S., Kumar, A., Tyagi, B.S., and Anand, V. Extractive Automatic Text Summarization using SpaCy in Python & NLP. In *2021 International conference on advance computing and innovative technologies in engineering (ICACITE)* IEEE, pp. 582-585, 2021.

[11] Adhikari, S. Nlp Based Machine Learning Approaches for Text Summarization. In *2020 Fourth International Conference on Computing Methodologies and Communication (ICCMC)*, IEEE, pp. 535-538, 2020.

[12] Madhuri, J.N. and Kumar, R.G. Extractive Text Summarization using Sentence Ranking. In *2019 international conference on data science and communication (IconDSC)*, IEEE, pp. 1-3, 2019.

[13] Merchant, K. and Pande, Y. Nlp Based Latent Semantic Analysis for Legal Text Summarization. In *2018 international conference on advances in computing, communications and informatics (ICACCI)*, IEEE, pp. 1803-1807, 2018.

[14] Ngo, C.W., Ma, Y.F., and Zhang, H.J. Video Summarization and Scene Detection by Graph Modeling. *IEEE Transactions on circuits and systems for video technology*, vol. 15, no. 2, pp. 296-305, 2005.

[15] Gygli, M., Grabner, H., Riemenschneider, H., and Van Gool, L. Creating Summaries from User Videos. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part VII 13*, Springer International Publishing, pp. 505-520, 2014.

[16] Dilawari, A. and Khan, M.U.G. ASoVS: Abstractive Summarization of Video Sequences. *IEEE Access*, vol. 7, pp. 29253-29263, 2019.

[17] Smaïli, K., Fohr, D., González-Gallardo, C.E., Grega, M., Janowski, L., Jouvet, D., Komorowski, A., Koźbiał, A., Langlois, D., Leszczuk, M. and Mella, O. A First Summarization System of a Video in a Target Language. In *Multimedia and Network Information Systems: Proceedings of the 11th International Conference MISSI 2018 11*, Springer International Publishing, pp. 77-88, 2019.

[18] Jaiswal, S. and Misra, M. Automatic Indexing of Lecture Videos using Syntactic Similarity Measures. In *2018 5th International Conference on Signal Processing and Integrated Networks (SPIN)*, IEEE, pp. 164-169, 2018.
[19] Choudhary, P., Munukutla, S.P., Rajesh, K.S., and Shukla, A.S. Real Time Video Summarization on Mobile Platform. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, pp. 1045-1050, 2017.
[20] Kannan, R., Ghinea, G., Swaminathan, S., and Kannaiyan, S. Improving Video Summarization Based on User Preferences. In *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG)*, IEEE, pp. 1-4, 2013.
[21] Basak, J., Luthra, V., and Chaudhury, S. Video Summarization with Supervised Learning. In *2008 19th International Conference on Pattern Recognition*, IEEE, pp. 1-4, 2008.
[22] Huang, J.H., Murn, L., Mrak, M., and Worring, M. Gpt2mvs: Generative Pre-Trained Transformer-2 for Multi-Modal Video Summarization. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pp. 580-589, 2021.
[23] Narasimhan, M., Rohrbach, A., and Darrell, T. Clip-It! Language-Guided Video Summarization. *Advances in Neural Information Processing Systems*, vol. 34, pp. 13988-14000, 2021.
[24] Huang, J.H. and Worring, M. Query-Controllable Video Summarization. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pp. 242-250, 2020.
[25] Xiao, S., Zhao, Z., Zhang, Z., Guan, Z., and Cai, D. Query-Biased Self-Attentive Network for Query-Focused Video Summarization. *IEEE Transactions on Image Processing*, vol. 29, pp. 5889-5899, 2020.
[26] Nalla, S., Agrawal, M., Kaushal, V., Ramakrishnan, G., and Iyer, R. Watch Hours in Minutes: Summarizing Videos with User Intent. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, Springer International Publishing, pp. 714-730, 2020.
[27] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., and Polosukhin, I. Attention Is All You Need. *Advances in neural information processing systems*, vol. 30, 2017.
[28] Jiang, P. and Han, Y. Hierarchical Variational Network for User-Diversified & Query-Focused Video Summarization. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pp. 202-206, 2019.
[29] Vasudevan, A.B., Gygli, M., Volokitin, A., and Van Gool, L. Query-Adaptive Video Summarization via Quality-Aware Relevance Estimation. In *Proceedings of the 25th ACM international conference on Multimedia*, pp. 582-590, 2017.
[30] Gygli, M., Grabner, H., and Van Gool, L. Video Summarization by Learning Submodular Mixtures of Objectives. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3090-3098, 2015.
[31] Sharghi, A., Laurel, J.S., and Gong, B. Query-Focused Video Summarization: Dataset, Evaluation, and a Memory Network Based Approach. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4788-4797, 2017.
[32] Sharghi, A., Gong, B. and Shah, M. Query-Focused Extractive Video Summarization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*, Springer International Publishing, pp. 3-19, 2016.
[33] Zhang, Y., Kampffmeyer, M., Liang, X., Tan, M., and Xing, E.P. Query-Conditioned Three-Player Adversarial Network for Video Summarization. *arXiv preprint arXiv:1807.06677*, 2018.
[34] Zhang, Y., Kampffmeyer, M., Zhao, X., and Tan, M. Deep Reinforcement Learning for Query-Conditioned Video Summarization. *Applied Sciences*, vol. 9, no. 4, pp. 750, 2019.
[35] Sreeja, M.U. and Kovoor, B.C. A Unified Model for Egocentric Video Summarization: An Instance-Based Approach. *Computers & Electrical Engineering*, vol. 92, pp. 107161, 2021.
[36] Ahmed, S.A., Dogra, D.P., Kar, S., Patnaik, R., Lee, S.C., Choi, H., Nam, G.P., and Kim, I.J. Query-Based Video Synopsis for Intelligent Traffic Monitoring Applications. *IEEE Transactions on Intelligent Transportation Systems*, vol. 21, no. 8, pp. 3457-3468, 2019.
[37] Gao, J., Yang, X., Zhang, Y., and Xu, C. Unsupervised Video Summarization via Relation-Aware Assignment Learning. *IEEE Transactions on Multimedia*, vol. 23, pp. 3203-3214, 2020.
[38] De Avila, S.E.F., Lopes, A.P.B., da Luz Jr, A., and de Albuquerque Araújo, A. VSUMM: A Mechanism Designed to Produce Static Video Summaries and a Novel Evaluation Method. *Pattern recognition letters*, vol. 32, no. 1, pp. 56-68, 2011.
[39] Zhang, K., Chao, W.L., Sha, F., and Grauman, K. Video Summarization with Long Short-Term Memory. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VII 14*, Springer International Publishing, pp. 766-782, 2016.
[40] Mahasseni, B., Lam, M., and Todorovic, S. Unsupervised Video Summarization with Adversarial LSTM Networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pp. 202-211, 2017.
[41] Rochan, M., Ye, L., and Wang, Y. Video Summarization using Fully Convolutional Sequence Networks. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 347-363, 2018.
[42] Zhou, K., Qiao, Y., and Xiang, T. Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[43] Ul Haq, H.B., Asif, M., Ahmad, M.B., Ashraf, R., and Mahmood, T. An Effective Video Summarization Framework Based on the Object of Interest using Deep Learning. *Mathematical Problems in Engineering*, vol. 2022, 2022.