IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.105108

SUMMARISATION OF VISUAL CONTENT IN INSTRUCTIONAL VIDEOS

Dr. Sharath Kumar Y H[1], Rumana Saifi[2], Sahana H C[3], Sushmitha Murthy[4], Varshini K P[5]

Associate Professor, Department of Information Science and Engineering, MITM, Mysuru[1]

Student, Department of Information Science and Engineering, MITM, Mysuru[2-5]

Abstract: Instructional videos have become increasingly popular for sharing knowledge and providing step-by-step guidance in various domains, ranging from cooking and crafts to academic subjects and technical tutorials. However, these videos often contain a significant amount of visual content, making it challenging for users to quickly grasp the key information and instructions. To address this issue, the field of video summarization has emerged, aiming to automatically extract concise and informative summaries from instructional videos. This paper presents a comprehensive review and analysis of existing techniques for the summarization of visual content in instructional videos. We categorize the approaches into two main groups: frame-based and object-based methods. Frame-based methods focus on selecting key frames that represent the essential information in the video, while object-based methods aim to identify and summarize relevant objects or regions of interest within the video. We discuss various strategies employed by these methods, including visual saliency analysis, motion analysis, and semantic understanding. Furthermore, we explore the challenges associated with instructional video summarization, such as handling complex scenes, dealing with occlusions, and understanding temporal dependencies. We also highlight the evaluation metrics commonly used to assess the quality of video summaries, including content coverage, representativeness, and coherence. Additionally, we present existing benchmark datasets and discuss their limitations in capturing the diverse range of instructional videos. Finally, we provide insights into potential future research directions in this field, such as incorporating multimodal information, leveraging deep learning techniques, and exploring user preferences to personalize video summaries. By summarizing visual content in instructional videos effectively, we can enhance the accessibility and usability of these videos, allowing users to quickly grasp the key concepts and instructions. This survey serves as a valuable resource for researchers and practitioners interested in video summarization and lays the groundwork for further advancements in this area.

Keywords: Summarization, Visual Content, Instructional Videos, Video Summarization, Key Frames, Object-based Methods

I. INTRODUCTION

With the proliferation of online platforms and the increasing popularity of video-based content, instructional videos have emerged as a powerful medium for sharing knowledge and providing step-by-step guidance in various domains. These videos cover a wide range of subjects, including cooking, crafts, academic tutorials, DIY projects, and technical demonstrations. While instructional videos are valuable resources for learning and acquiring new skills, they often contain a large amount of visual content, which can make it challenging for users to quickly extract the key information and instructions. Video summarization, a subfield of multimedia analysis and retrieval, focuses on automatically extracting concise and informative summaries from videos. Traditional video summarization techniques primarily targeted broadcast news and surveillance videos, where textual information played a prominent role. However, the unique characteristics of instructional videos, with their emphasis on visual demonstrations and hands-on instructions, require tailored approaches for summarizing their visual content effectively.

The summarization of visual content in instructional videos poses several challenges. Firstly, these videos often comprise multiple frames with varying levels of relevance and importance, making it crucial to identify and select key frames that capture the essential information. Secondly, instructional videos frequently involve the demonstration of specific objects or regions of interest, requiring methods to identify and summarize relevant visual elements accurately. Additionally, instructional videos may exhibit complex scenes, occlusions, and temporal dependencies, necessitating sophisticated techniques to handle these challenges. This paper presents a comprehensive survey and analysis of existing techniques for the summarization of visual content in instructional videos. We categorize these approaches into two main groups: frame-based and object-based methods. Frame-based methods focus on selecting key frames that represent the most significant aspects of the video, enabling users to grasp the essence of the instructions quickly. Object-based methods, on the other hand, aim to identify and summarize relevant objects or regions of interest within the video, providing a more granular understanding of the visual content.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.105108

We discuss various strategies employed by these methods, including visual saliency analysis, motion analysis, and semantic understanding. Visual saliency analysis techniques aim to identify visually salient regions or objects that attract the viewer's attention. Motion analysis methods focus on detecting and summarizing important temporal changes, such as object movements or actions performed in the video. Semantic understanding techniques leverage high-level semantic information to extract meaningful visual content and summarize it effectively. Furthermore, we explore the evaluation metrics commonly used to assess the quality of video summaries in the context of instructional videos. These metrics include content coverage, which measures the extent to which the summary represents the essential content, representativeness, which evaluates the ability of the summary to capture the diversity of visual content, and coherence, which assesses the logical flow and organization of the summary.

Finally, we provide insights into potential future research directions in the field of summarization of visual content in instructional videos. These directions include incorporating multimodal information, such as audio and textual cues, to enhance the summarization process. Additionally, leveraging deep learning techniques, such as LSTM which may enable more accurate and context-aware video summarization. Furthermore, exploring user preferences and feedback to personalize video summaries according to individual needs and learning styles holds promise for enhancing the user experience.

II. LITERATURE SURVEY

In this paper [1], it addresses the problem of supervised video summarization by formulating it as a sequence-to-sequence learning problem, where the input is a sequence of original video frames, the output is a keyshot sequence. Their key idea is to learn a deep summarization network with an attention mechanism to mimic the way of selecting the keyshots of humans. To this end, they proposed a novel video summarization framework named Attentive encoder-decoder networks for Video Summarization (AVS), in which the encoder uses a Bidirectional Long Short-Term Memory (BiLSTM) to encode the contextual information among the input video frames. As for the decoder, two attention-based LSTM networks are explored by using additive and multiplicative objective functions, respectively. Extensive experiments are conducted on two video summarization benchmark datasets, i.e. SumMe, and TVSum. The results demonstrate the superiority of the proposed AVS-based approaches against the state-of-the-art approaches, with remarkable improvements from 0.8% to 3% on two datasets, respectively. They proposed a deep attentive framework for supervised video summarization. Specifically, two attention-based deep models named A-AVS and M-AVS are developed, respectively. To the best of our knowledge, our work is the first attempt to apply attention mechanisms in deep models for video summarization. The proposed models outperform the competing methods on two benchmark datasets by 0.8% - 3%. They also provided qualitative analysis and parameter sensitive analysis. In addition, the augmentation experiments also verify the effectiveness and superiority of AVS framework when applied augmented data.

They proposed [2] an innovative joint end-to-end solution, Abstractive Summarization of Video Sequences, which uses the deep neural network to generate the natural language description and abstractive text summarization of an input video. This provides a text-based video description and abstractive summary, enabling users to discriminate between relevant and irrelevant information according to their needs. Furthermore, our experiments show that the joint model can attain better results than the baseline methods in separate tasks with informative, concise, and readable multi-line video description and summary in a human evaluation. ASoVS which uses deep neural networks to generate natural language description and abstract text summary of an input video sequence. This automatic understanding of video semantics and then presenting this information in textual form can help users to understand huge volumes of data. For this purpose, the multitask features learning deep model mines rich information such as human attributes, objects, actions, interactions and scene information from a video to produce video description. This information is taken out through a CNN, which is fine-tuned for these specific tasks. This model produced comprehensive, concise and readable descriptions with good results on the three datasets. They can create a short and a fluent summary of a longer text document, which mines appropriate information from the input text to utilize the relevant information faster. This textual conversion not only reduces the size of video data but also enables users to index and navigate information through it.

In this paper [3], the convolutional neural network VGG19 model is used to extract the depth features of video frames on the TVSum50 dataset. The training set is trained by the sorting learning ListNet algorithm to obtain a model that can mark the frames of the test set. According to the scoring result, the top 25 frames in the test set are extracted as key frames, and the video summarization is synthesized. The video frames in digest which came into being by automatically generated and manual work can use pHash algorithm as assimilation judgment, and get the F1-score of the model. Compared with the current video summarization method, this method is to mark the frame, not the frame set, so the calculation efficiency is reduced, but it can improve accuracy. Through the test on the TVSum50 dataset, the method of this paper is significantly better than other video summarization methods.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.105108

In this paper [4], they proposed a novel deep summarization framework named Bi-Directional Self-Attention with Relative Positional Encoding for Video Summarization (BiDAVS) that can be highly parallelized. Our proposed BiDAVS considers position information of input sequence and effectively captures long-range temporal dependencies of sequential frames by computing bi-directional attention. Extensive experiments on two popular benchmark datasets, i.e., SumMe and TVSum, show that the proposed model outperforms state-of-the-art approaches. In this paper, they proposed a novel deep summarization framework, Bi-Directional Self-Attention with Relative Positional Encoding for Video Summarization (BiDAVS). To the best of our knowledge, they are the first to apply multi-head self-attention with relative positional encoding in video summarization tasks. Our model can effectively capture long-range temporal dependencies of sequential frames by computing bi- directional attention. Experimental results show that our approach outperforms other state-of-the-art supervised methods on two benchmark datasets. Particularly, another evaluation using Kendall's τ and Spearman's ρ correlation coefficients obviously shows that our proposed BiDAVS framework is superior to others.

In this paper [5], they proposed a Detect-to-Summarize network (DSNet) framework for supervised video summarization. The DSNet contains anchor-based and anchor-free counterparts. The anchor-based method generates temporal interest proposals to determine and localize the representative contents of video sequences, while the anchor-free method eliminates the pre- defined temporal proposals and directly predicts the importance scores and segment locations. Different from existing supervised video summarization methods which formulate video summarization as a regression problem without temporal consistency and integrity constraints, our interest detection framework is the first attempt to leverage temporal consistency via the temporal interest detection formulation. Specifically, in the anchor-based approach, they first provide a dense sampling of temporal interest proposals with multi-scale intervals that accommodate interest variations in length, and then extract their long-range temporal features for interest proposal location regression and importance prediction. Notably, positive and negative segments are both assigned for the correctness and completeness information of the generated summaries. In the anchor-free approach, they alleviate drawbacks of temporal proposals by directly predicting importance scores of video frames and segment locations. Particularly, the interest detection framework can be flexibly plugged into off-the-shelf supervised video summarization methods. They evaluate the anchor-based and anchor-free approaches on the SumMe and TVSum datasets. Experimental results clearly validate the effectiveness of the anchor-based and anchor-free approaches.

III. METHODOLOGY

Step 1: Data Collection

- Obtain the YouTube video link for which the text summarization is required.
- Retrieve the video transcript either by using YouTube's API or by using transcription services.

Step 2: Preprocessing

- Convert the transcript text into a suitable format for analysis (e.g., lowercase, remove punctuation).
- Perform tokenization to split the text into individual words or sentences.
- Remove stop words (commonly used words with little semantic value) to reduce noise in the data.

Step 3: Text Summarization Technique Selection

- Choose a suitable text summarization technique based on the requirements and available resources. Some common techniques include:
- Extractive Summarization: Selecting important sentences or phrases from the original text.
- Abstractive Summarization: Generating new sentences that capture the essence of the original text.
- Graph-based Summarization: Analyzing the relationships between sentences to identify important information.

Step 4: Implementation of Text Summarization

- Develop or utilize an existing text summarization algorithm based on the chosen technique.
- Implement the algorithm to process the preprocessed text data and generate a summary

Step 5: Evaluation and Refinement

- Evaluate the generated summaries based on predefined metrics such as ROUGE (Recall-Oriented Understudy for Gisting Evaluation) or BLEU (Bilingual Evaluation Understudy).
- Refine the summarization algorithm based on the evaluation results and user feedback.

Step 6: Integration and Deployment

- Create an interface to accept the YouTube video link as input.
- Integrate the summarization system with the YouTube API or other necessary components.
- Deploy the system, making it accessible for users to provide video links and receive text summaries.

IARJSET ISSN (O) 2393-8021, ISSN (P) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified ⁜ Impact Factor 8.066 ⁜ Peer-reviewed / Refereed journal ⁜ Vol. 10, Issue 5, May 2023

DOI: 10.17148/IARJSET.2023.105108

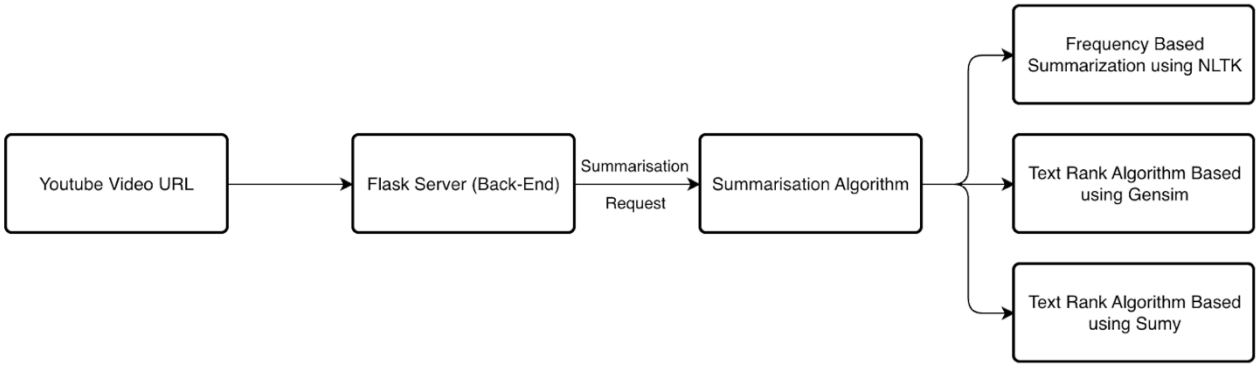

Fig. 1 Methodology

IV. CONCLUSION

Text summarization of YouTube video links is a valuable and challenging task that involves condensing the content of a video into a concise and informative summary. By leveraging the video's transcript and employing various text summarization techniques, it becomes possible to extract the most important information and present it in a condensed form. The design activity for text summarization of YouTube video links encompasses several key steps. It begins with data collection, where the video link is obtained and the transcript is retrieved. Preprocessing follows, involving text formatting, tokenization, and the removal of stop words to prepare the data for analysis. The choice of text summarization technique is critical and can include extractive, abstractive, or graph-based approaches.

Implementation involves developing or utilizing an existing text summarization algorithm based on the chosen technique. The algorithm processes the preprocessed text data and generates a summary that captures the essential information from the video. Evaluation and refinement are necessary steps to ensure the quality and effectiveness of the generated summaries, employing metrics such as ROUGE or BLEU. Finally, the summarization system is integrated with the YouTube API or other relevant components, creating an interface that accepts video links as input and provides text summaries as output. Deployment of the system allows users to easily access and benefit from the text summarization capabilities for YouTube videos.

REFERENCES

[1]. C. Ordonez, Y. Zhang, and S. L. Johnsson, "Scalable machine learning computing a data summarization matrix with a parallel array DBMS," Distrib. Parallel Databases, vol. 37, no. 3, pp. 329–350, 2019, doi: 10.1007/s10619-018-7229-1.

[2]. M. Mauro, L. Canini, S. Benini, N. Adami, A. Signoroni, and R. Leonardi, "A freeWeb API for single and multi-document summarization," ACM Int. Conf. Proceeding Ser., vol. Part F1301, 2017, doi: 10.1145/3095713.3095738.

[3]. A. T. Sarda and M. Kulkarni, "Text Summarization using Neural Networks and Rhetorical Structure Theory," Int. J. Adv. Res. Comput. Commun. Eng., vol. 4, no. 6, pp. 49–52, 2015, doi: 10.17148/IJARCCE.2015.4612.

[4]. G. Silva, R. Ferreira, S. J. Simske, L. Rafael Lins, M. Riss, and H. O. Cabral, "Automatic text document summarization based on machine learning," DocEng 2015 - Proc. 2015 ACM Symp. Doc. Eng., pp. 191–194, 2015, doi: 10.1145/2682571.2797099.

[5]. T. Jo, "K nearest neighbor for text summarization using feature similarity," Proc. - 2017 Int. Conf. Commun. Control. Comput. Electron. Eng. ICCCCEE 2017, pp. 1–5, 2017, doi: 10.1109/ICCCCEE.2017.7866705.

[6]. B. Mutlu, E. A. Sezer, and M. A. Akcayol, "Multi-document extractive text summarization: A comparative assessment on features," Knowledge-Based Syst., vol. 183, p. 104848, 2019, doi: 10.1016/j.knosys.2019.07.019.

[7]. M. Afsharizadeh, H. Ebrahimpour-Komleh, and A. Bagheri, "Query-oriented text summarization using sentence extraction technique," 2018 4th Int. Conf. Web Res. ICWR 2018, pp. 128–132, 2018, doi: 10.1109/ICWR.2018.8387248.

[8]. J. N. Madhuri and R. Ganesh Kumar, "Extractive Text Summarization Using Sentence Ranking," 2019 Int. Conf. Data Sci. Commun. IconDSC 2019, pp. 1–3, 2019, doi: 10.1109/IconDSC.2019.8817040.