SN Computer Science (2024) 5:1010
https://doi.org/10.1007/s42979-024-03351-w

ORIGINAL RESEARCH

Enhancing Accessibility in Online Shopping: A Dataset and Summarization Method for Visually Impaired Individuals

Ratnabali Pal[1,2] · Samarjit Kar[2,3] · Arif Ahmed Sekh[4]

Received: 4 March 2024 / Accepted: 27 September 2024

Abstract
A visually impaired individual (VI) encounters numerous challenges in their daily activities, particularly in tasks reliant on visual systems such as navigation, educational pursuits, and shopping. Online shopping poses a heightened difficulty due to its reliance on visual representations of products in digital formats. The impact of visual impairment on product selection based on reviews remains inadequately investigated. This study endeavors to address two primary objectives. Firstly, we propose the creation of a dataset comprising product review videos (referred to as PVS10) tailored for visually impaired individuals. Secondly, we present a foundational summarization methodology designed to facilitate access to pertinent and informative content within extensive video collections for visually impaired individuals. Our dataset, gathered from YouTube, encompasses 10 distinct products, each associated with the top 10 review videos, totaling 100 videos of varying lengths. Utilizing the search term "review videos of PRODUCT NAME", we assembled the dataset to facilitate automated summarization processes aimed at maximizing salient information, minimizing redundant content, and preserving the overarching sentiment conveyed in the reviews. This research focuses on the challenges faced by visually impaired people in online shopping, particularly when selecting products based on customer reviews. Our study demonstrates that people with visual impairments may actively explore product reviews and only acquire the information they require.

Keywords Video summarization · Visually impaired · Review summarization · Visual assistance

Introduction

Advancements in artificial intelligence (AI), computer vision (CV), deep learning, and related fields have significantly broadened the possibilities for aiding individuals with visual impairments. These technologies have proven successful in practical applications such as question answering [1], navigation [2], and personal assistance [3].

In this domain, both images and videos serve as fundamental sources of information crucial for intelligent decision-making. Video analysis and information extraction play vital roles, particularly for visually impaired individuals. However, a significant challenge arises from the vast amount of video data recorded and shared across various platforms, making it time-consuming to watch such large volumes of content. To address this challenge, various video summarization methods have been proposed. These techniques aim to condense extensive video collections by extracting informative content, offering a concise and informative representation of visual content from a large set of videos [4–10]. A qualitative video summarization has two fundamental characteristics:

- Video summarization methods aim to ensure that the summaries generated are representative, meaning they include all the critical and relevant visual scenes from the original videos.

✉ Arif Ahmed Sekh
skarifahmed@gmail.com

Ratnabali Pal
rp.21ma1501@phd.nitdgp.ac.in

Samarjit Kar
skar@maths.nitdgp.ac.in

[1] Department of Mathematics, NIT Durgapur, Durgapur, WB 713209, India

[2] Department of Computational Sciences, Brainware University, Kolkata, WB 743248, India

[3] Department of Graphical Systems, Vilnius Gediminas Technical University, 10223 Vilnius, Lithuania

[4] Department of Compter Science, UiT The Arctic University of Norway, 9019 Tromsø, Norway

Published online: 02 November 2024

- The summary provides a concise synopsis of video sequences, minimizing redundancy while encapsulating the essential content.

Researchers have proposed numerous video summarization methods tailored for diverse application domains. These domains include episodes of web series [11], highlights of sports and games [12], music band performances, documentaries, personal videos, as well as medical videos [13] Additionally, video summarization techniques find applications in various other contexts such as surveillance, traffic management, and content browsing.

Videos can serve as valuable tools to aid buyers in online shopping, particularly through product review videos [14–16]. These videos provide substantial benefits for both businesses and customers alike. Customers have the opportunity to peruse various video-based reviews of products shared by fellow users, allowing for a more informed purchasing decision. In such scenarios, the core challenge persists: navigating and watching a large volume of videos is time-consuming. This challenge can be mitigated through the use of video summarization methods. However, the majority of existing video summarization techniques are designed for CCTV or scene-based applications. There is currently no available dataset specifically tailored for the video summarization of “product review videos”. The objective of such summarization should be to produce concise summaries comprising both visual and informative content. This task becomes even more challenging when considering visually impaired individuals, as they rely more on descriptive information available in audio or text formats rather than visual cues. Despite these challenges, research in the domain of video summarization for assisting visually impaired individuals remains limited.

To enhance the shopping experience for individuals with low vision, various technological solutions have been proposed, including barcode readers [17, 18], robotics [19, 20], computer vision-based systems [21], and braille product labels [22]. In this context, we introduce a system designed to assist visually impaired individuals in selecting suitable products from e-commerce platforms based on customer reviews available on YouTube. Here, we explore the potential of video summarization techniques tailored for “product review videos” to enhance the online shopping experiences of individuals with visual impairments.

This article focuses on addressing a specific challenge faced by visually impaired individuals and proposes a solution. We pose the question: “How can visually impaired individuals benefit from product review videos to enhance their online shopping experience?” While research on online product review video analysis has garnered significant attention, a publicly available dataset specifically catering to visually impaired individuals is yet to be created. To comprehensively evaluate our methodology, we compile a new large video dataset on product reviews from YouTube, taking into account various perspectives of visually impaired individuals and algorithmic challenges. The insights and opinions presented in these reviews can assist visually impaired individuals in making informed decisions and better understanding products.

Motivation and Contribution

Videos captured with various smart devices are considered a powerful assistive tool for visually impaired (VI) individuals, enhancing their independence and quality of life. The main motivation for developing such an assistive tool is to visualize the visual content and provide useful information in the form of speech. Low-sighted individuals can shop online independently by being provided with audio key review information from product videos. This information allows them to make choices that align with their preferences. We list the following summary of our contributions:

- For the first time, we introduce a product video summarization dataset specifically tailored for visually impaired individuals. Our dataset is designed to address the specific requirements of this demographic.
- We have proposed an audio-guided text-mining approach for video summarization, which is specifically tailored to meet the needs of visually impaired individuals. This approach relies on audio cues to guide the text-mining process, ensuring that the summarization output is accessible and informative for individuals with visual impairments.

The rest of this paper is structured as follows: The related work is presented in “Related Work” section. “Baseline Method” section outlines the data representation, feature analysis, fusion, and summarization. Descriptions of the datasets, evaluation metrics, and case study findings are found in “Dataset, Results and Discussion” section. “Ablation Study” section highlights the ablation study of the proposed work. “Conclusion” section concludes the findings and directions for further research.

Related Works

In this section, we conducted a systematic review to identify the current state of research in intelligent video analysis tailored for individuals with low vision. Social media platforms now serve as major sources of information, but visually impaired individuals face challenges accessing visual content. With widespread internet and smartphone usage, online activities like shopping have increased significantly, leading consumers to seek reassurance through online feedback and reviews. However, visually impaired individuals have limited access to such visual information. While various video summarization techniques exist for general users, their applicability for visually impaired individuals remains underexplored. Therefore, we provide an overview of these methods in this section.

Review Text Summarization

Numerous researchers in natural language processing (NLP) have explored text summarization techniques to condense lengthy reviews into shorter, representative sentences. Boorug's survey [23] evaluates various NLP methods for text summarization in online shopping contexts. Jiao et al. [24] proposed a relational extraction method linking product design features with user evaluations, identifying optimal feature combinations for review assistants. Fan et al. [25] conducted an in-depth analysis of product ranking based on online reviews, covering aspects such as customer satisfaction and market analysis. Shah [26] applied sentiment analysis and sequence-to-sequence (seq2seq) RNN with attention to summarization. Doğan et al. [27] used Word2vec and FastText LSTM for social media text summarization. Patel et al. [28] employed a rule-based fuzzy inference system for multi-document summarization, determining summarization outcomes based on sentence scores. Recently, in this article [29] authors proposed a novel approach to create synthetic datasets that combine information from reviews, product descriptions, and question-and-answers.

Video Summarization

Intelligent video summarization is a valuable solution to alleviate the tedious task of continuously watching multiple videos. Video summarization, as described in Basavarajaiah et al.'s survey [4], involves creating concise summaries by extracting key frames or clips from a large collection of original videos.

In Rochan et al.'s work [30], a fully convolutional sequence network (FCSN) was introduced for video summarization, adapted from semantic segmentation networks. Muhammad et al. [31] proposed a deep CNN framework with hierarchical weighted fusion to aggregate scores for each frame, resulting in intelligent feature fusion.

Hussain et al. [6] utilized deep bidirectional long short-term memory (BiLSTM) for multi-view video summarization. Ji et al. [32] discussed an attentive encoder-decoder framework employing Bidirectional Long Short-Term Memory (BiLSTM) in the encoder and attention-based LSTM networks in the decoder.

Zhang et al. [33] proposed a sequence-to-sequence learning model with metric learning loss for matching and targeted summaries, deriving semantic embeddings for both supervised and unsupervised videos. Rafiq et al. [34] used a pre-trained AlexNet Convolutional Neural Network (CNN) for cricket sports scene classification with high accuracy.

In Guntuboina et al.'s research [35], the YOLO object detection model was employed to identify scoreboards in sports videos, followed by OCR and classification to extract vital events. Additionally, Emon et al. [36] utilized an encoder-decoder architecture for deep cricket summarization, predicting the vital score of each frame in a test video. A Vision Transformer (ViT)-assisted deep pyramidal refinement network [37] has been proposed for video summarization, emphasizing the extraction and refinement of multi-scale features and the prediction of importance scores for each frame. In this study, Kawamura et al. [38] proposed a new video summarization methodology called "FastPerson" that considers visual and auditory information in lecture videos, enhancing the learning experience. This research method [39] focuses on a deep learning-based shot boundary detection pipeline as a preparatory phase and extracts keyframes using DBSCAN clustering algorithm. A genetic algorithm is used to optimize the hyper-parameters of DBSCAN [40] instead of requiring the user to pre-tune them, as the number of keyframes in a video might change depending on the content of the video. Recently, authors introduced a new Attention-Guided Multi-Granularity Fusion Model (AMFM) [41], enabling fusion and optimization modeling processes from context capturing.

Review Video Summarization

The primary focus of review video summarization is to extract and analyze positive or negative sentiments expressed in opinion sentences, extracting deep features to summarize customers' opinions during online shopping experiences. There are limited research on this topic.

In Otani et al.'s work [42], deep video features were extracted to identify essential segments, facilitating efficient video retrieval using the SumMe dataset [43]. Benkhelifa et al. [44] utilized an SVM classifier to extract objective texts from cooking recipe videos on YouTube, enabling classification and comparison between recipes. Im et al. [45] employed a pre-trained ResNet101 from ImageNet to extract multi-modal opinion summarization from Yelp and Amazon datasets.

In product review summarization for visually impaired individuals, researchers have addressed the challenges low-vision individuals encounter when shopping online. Stangl et al. [46] explored these challenges and implemented computer vision algorithms to develop intelligent online shopping assistants. Similarly, Kostyra et al. [47] investigated the food choices and eating experiences of visually impaired individuals, aiming to enhance the quality of the e-shopping experience. Yamaguchi et al. [48] proposed a method for timestamp analysis in review videos where dictionary-based product features are discussed to support online shopping (Table 1).

Multimodal Summarization for VI People

The article [55] proposed an audio summarization technique that combines script-based summarization with the classification of visual information from videos. Scientists developed a proof of concept and evaluated the summarization solution on eleven blind users to assess its effectiveness. Recently, Manojkumar et al. [56] presented a comparative analysis with Weighted TF_IDF [57], an efficient algorithm to automate text summarization which is taken from a text-speech. They stated that low-vision individuals would find great benefit from this proposed algorithm. A novel topic-guided summarization technique [53] has been proposed that allows VI people to navigate through the enormous video content of a product. Here, the researchers show how topic-based features can enhance video synopsis.

In Table 2 we discuss application-specific datasets used in Product review and video summarization and analysis.

Baseline Method

Here, we present a baseline method for summarizing product review videos. We start with a collection of N review videos denoted as $V_1, V_2, \dots, V_N$. The total duration of these videos is denoted as $T_{original}$. Our goal is to create a montage of videos with a significantly shorter duration, denoted as $T_{summary}$, where $T_{summary} << T_{original}$. Our approach considers both verbal demonstrations and visual information in the videos for summarization. The method is illustrated in Fig. 1, with the steps marked in the yellow circles discussed subsequently.

Table 2 Popular datasets used in various review and video summarization articles

Dataset	VS	TR	VR	VRVI
Product review (OPRA) [58]	✓	✗	✓	✗
Stanford Sentiment Treebank [59]	✗	✓	✓	✗
MVS [5]	✓	✗	✗	✗
Sports video [34]	✓	✗	✗	✗
Twitter dataset [60, 61]	✗	✓	✗	✗
BrowseWithMe [46]	✗	✓	✗	✓
Proposed PVS10	✓	✓	✓	✓

VS video summarization, *TR* text-based review, *VR* video-based review, *VRVI* video-based review for visually impaired

Data and Input: We aim to collect N number of videos for summarization. We conducted a search on YouTube$^{\text{TM}}$ for a specific product review. Subsequently, we obtained the top N videos for the search query as input to the proposed framework. Each video is then segmented into equal durations, with a duration of 1 min considered for each segment. Therefore, each video V_i is represented by a set of video segments as: $(V_i, S_1), (V_i, S_2), \dots, (V_i, S_M)$.

Audio-to-text: Subsequently, the verbal text is extracted using a state-of-the-art pipeline [62] and denoted as $(V_i, T_1), (V_i, T_2), \dots, (V_i, T_M)$. Following this, both the video and the text are input to the feature extraction pipeline.

Feature extraction: Our objective is to assign weights to the video segments for inclusion in the final summarization video. Figure 2 illustrates the products from which

Table 1 Characteristics and limitations of different summarization areas

Summarization (area)	Dataset	Applications	Limitations
Review text	(i) Amazon, Oposum and Flipkart [29] (ii) DUC 2004 [28]	(i) Pseudo-reviews with product description and question-answers	Not considering LLM [49] for multi-source opinion summarization
Video	(i) TVSum and SumMe [50] (ii) Open Video Project (OVP) and YouTube (YT) [39]	(i) Dynamic video generation enriched with audio (ii) Genetic-algorithm based summarization	Defective learning schemes in existing methods, contextual information with is still effective
Review video	(i) Yelp Dataset and Amazon Products review [51] (ii) SumMe [52]	(i) Multimodal opinion summarization (ii) Super frame segmentation	Unorganized information was not extracted effectively using advance image encoding
Visual assistance	(i) PRVDVI [53] (ii) CineAD [54]	(i) A novel topic-guided Summarization technique (ii) Visual Question Answering (iii) Automated audio description generation	Output was degraded due to the small size of the datsaet

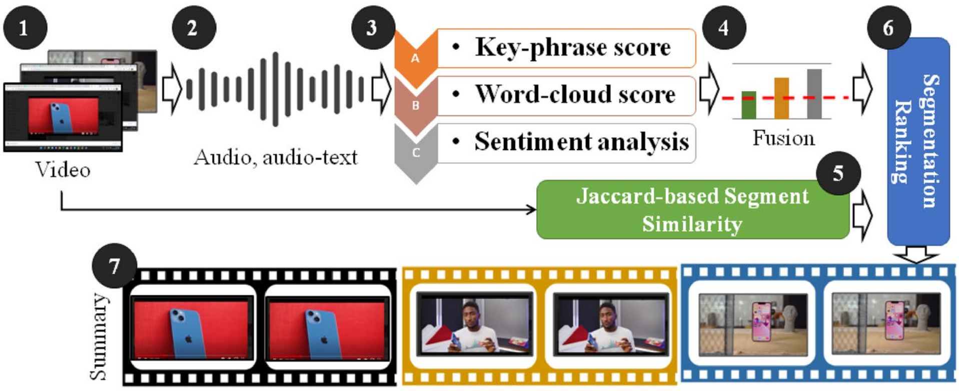

Fig. 1 The proposed baseline summarization method. The steps are marked using circles: 1. is a set of given "product review videos", 2. Audio and Audio-to-text extraction, 3. Feature extraction, 4. Informative segment extraction, 5. Visual similarity checking, 6. Segment ranking, and 7. Segment stitching and summarization

review videos have been collected. The weight of a segment is denoted by (W_i^j), which represents a combination of the verbal and visual importance of the segment. We extract the Keyword score (α), word-cloud score (β), and sentiment score (ψ) for the verbal feature, and hand activity (η) as the visual feature. The features are discussed subsequently.

Feature

In the proposed baseline, we have incorporated three features: keyword-based, word-cloud-based, and sentiment-based features.

Key-word score: Firstly, keywords or key phrases (up to bi-gram words) are extracted from all the verbal texts obtained from a particular product search. We utilize key-BERT [63, 64] for this purpose. Let these keywords be denoted by *key*. The keyword score of a video segment is defined as follows:

$$\alpha = \sum text \times frequency(key), \quad where\ text \in key,\ and\ text \in V_iT_j \tag{1}$$

Word-cloud score: Next, a word cloud is generated from all sets of video segments of a product review. We assign a weight to each word using WordCloud [65, 66]. The word-cloud score of a video segment is defined as:

$$\beta = \sum cloudValue(word),\ where\ word \in V_iT_j \tag{2}$$

Segment sentiment: Finally, a sentiment score is assigned to each segment based on the verbal text. The sentiment score (ψ) is extracted using the pre-trained model proposed in [67]. We utilize the sentiment probability distribution across multiple videos to obtain an unbiased summary.

$$\psi = sentimet(V_iT_j) \tag{3}$$

Segment similarity: To determine the uniqueness of a segment based on verbal demonstration, we employ a segment similarity approach using the Jaccard similarity algorithm [68, 69]. We calculate the similarity score (σ) to identify and remove redundant information in the summarized video. The Jaccard similarity is computed as the size of the intersection divided by the size of the union of two segments.

$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{4}$$

The Jaccard Similarity coefficient can be interpreted as the probability that an element picked at random from the universal set is present in both sets A and B. The weight of a segment, W_i^j is a combination of these features. While the exact formula for integrating these scores into a final weight isn't provided, it would likely look like:

$$W_i^j = [f(\alpha, \beta, \psi, \eta)] \tag{5}$$

where f is a function that combines these features, possibly using weighted summation or another method. By extracting and combining these features, each video segment can be assessed for inclusion in the final summary, ensuring that only the most informative and distinct segments are selected.

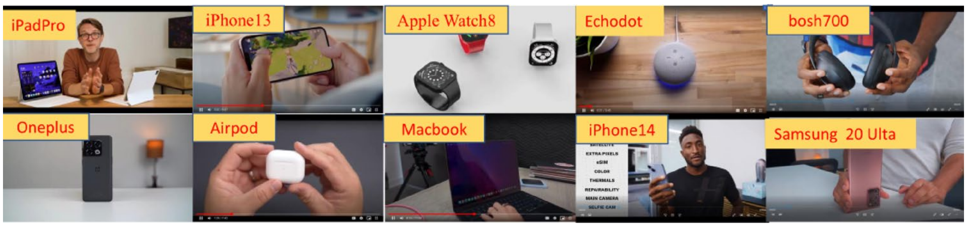

Fig. 2 Examples of products for which reviews have been collected

Fusion and Summarization

We approach this problem as a ranking problem, where our objective is to select the top-N segments (here, we consider $N = 10$) for inclusion in the final summary. Our goal is to maximize the presence of verbal and visual information in the summary, minimize redundant content, and preserve the overall sentiment of the review in the final video. To achieve this, we introduce three penalty terms, two of which are reward terms. The algorithm for this method analyses every input segment, determines and calculates its importance features, and assigns a score to each segment based on these features. It ranks the segments based on their scores, selects the top segments for the summary, and ensures diversity among the selected segments by applying a similarity threshold. The computational workflow for video summarization is presented in Algorithm 1.

Algorithm 1 Summary video generation

```
Input: $V = \{V_1, V_2, ..., V_N\}$ {V is a set of review videos}
Output: $S = \{(V_1, S_1), (V_1, S_2), ..., (V_N, S_M)\}$ {S is a the video segment}
for $i = 0$ to $N \times M$ do
    Calculate $\alpha, \beta, \psi$, and $\sigma$ for segment $S_i$
    Importance of $S_i = C1.\alpha + C2.\beta + C3.\psi + C4.\sigma$ {C1,...,C4 are predefined weights for each feature}
end for
$R_i$=rank of $S_i$ based on importance score
$\eta = \{S_1, ..., S_L\}$, where $S_i$ is taken based on rank and $\sigma(S_i, S_j) < th$ represents some similarity measure between segments $S_i$ and $S_j$ ensuring the selected segments are not too similar.
```

Dataset, Results and Discussion

Dataset

We presented a novel dataset of product review videos (referred to as PVS10) that offers an invaluable tool for researching and evaluating summarization algorithms, facilitating the more precise and efficient summarization of product review content. Our dataset consists of review videos for 10 distinct technical gadgets, which we have downloaded from YouTube. These videos offered in-depth evaluations and demonstrations of each one, providing information on their functionality and performance, features, and user experiences. This dataset can be incorporated into educational tools to explain product features and market trends to low-vision individuals and enhance their knowledge and skills. To ensure that the dataset collection process is inclusive and effective for visually impaired individuals, consider the following implementation steps:

Search on Youtube: Conduct a search on YouTube for a specific product review and select the top N videos.

Video Collection: Get the top N videos from the search query. Let's label these videos as $V_1, V_2, \ldots, V_N$.

Segmentation: Each video V_i is divided into one-minute equal segments. As a result, a video V_i with a total duration of T minutes will be divided into M segments, where M = T.

Representation: Each video V_i represented by a set of video segments. This can be written as: { (V_i, S_1),(V_i, S_2),...,(V_i, S_M)}, where S_j represents the j^{th} segment of video V_i.

Metric

We identified a gap in the metric for measuring the quality of the summarized video. To address this, we propose a novel performance measurement approach. Drawing from existing literature, we utilize ROUGE scores [70] for text, serving as a Unicode evaluation metric, and precision-recall-based metrics [71] for images and videos. Precision indicates how many selected frames are relevant, while recall determines how many relevant frames are selected. The video summarization method, along with the production of video segments and segmented bunches, is evaluated using objective criteria. The two criteria that encapsulate the entire framework of the video summarization scheme are Precision and Recall, calculated as follows in Eqs. (6) and (7):

$$Precision = \frac{matched\ frames}{Total\ frames\ in\ automaticsummary} \in [0, 1] \tag{6}$$

In our proposed video summarization method, a higher precision indicates that the majority of the segments included in the summary are relevant to the key content of the original video. It helps in determining how much of the summary is relevant while avoiding irrelevant or redundant information.

$$Recall = \frac{matched\ frames}{Total\ frames\ in\ user\ summary} \in [0, 1] \tag{7}$$

The automatic summary comprises detected key-frames obtained from different summarization methods. The Ground Truth user summary consists of various user summaries. The Image Euclidean distance is utilized for comparison.

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \in [0, 1] \tag{8}$$

Here, a higher recall indicates that the summary conveys the majority of the significant content from the original video.

F-measure [72] is the harmonic mean of both Precision and Recall, providing a measure of the accuracy of the experiment. It is especially useful when recall and precision need to be optimized simultaneously. To determine the accuracy of the Automatic summary compared to the Ground truth summary and to assess the similarity matching between segments, state-of-the-art methods use Fidelity, which is based on the semi-Hausdorff distance algorithm.

$$Fidelity = \frac{Sum\ Dist(F_i, AT_i)}{AT} \in [0, 1] \tag{9}$$

In our case, none of these metrics are accurate due to the repetitive information present in different videos. Our goal is to create a concise summary of the information. Therefore, we propose a straightforward question-answer F-measure metric for evaluation.

Question-answer F-measure: We implement straightforward accuracy metrics where questions and answers can be used to summarize the video synopsis. We have constructed a set of questions specific to product reviews for this task. For example, “what are the new features of that particular gadget?”

Table 3 Comparison of the summarization methods

Method	Video-based	Text-based	F-measure
Sentiment-based [26]	✗	✓	0.21
Text summarization [73]	✗	✓	0.31
Vision fusion [31]	✓	✗	0.25
Scene-based [34]	✓	✗	0.41
Proposed	✓	✓	0.75

Next, the F-measure is calculated based on how many answers are present in the summary. The F-measure is calculated using Eq. (8).

Results

Here, we present the results of different state-of-the-art methods for the task. Table 3 shows the F-measure obtained using different methods. It is observed that the proposed method achieves a significant improvement compared to the state-of-the-art.

Case Study

Here, we present a case study of a product from our proposed dataset. We selected the iPhone 13 for this study. We collected the top 5 videos of varying lengths. In this case study, human evaluators assessed the quality of a large set of summaries generated by different methods and classes.

Different parameters of the data are:

Product: iPhone 13

No. of Videos: 5

Total Duration: 68 min

Question: Good or bad? How is the sound quality? How is the camera quality? How is the battery life? What is new? What are the drawbacks? What are the smart features? How is the build quality? What is the price range? What is the specification?

Summary Duration: 10 min

F-measure: 0.69

Summary Text (May have grammatical errors, as it is the original output): I wanted you se nasht with headphones. Phone is well-balanced in easy to wear. Foster was in aqueous there were balance in love screen very good enough time for the here is exceptional, of course there’s no 3.5 mm jack. performances as well as 5 G connectivity in CPU test, iPhone 13th of the shot up to 15% higher the new iPhone 12 GPU. coronavirus started will be charged iPhone 13 from 0 to 54% in half an hour it also supports wireless charging which goes under 16. if you compare iPhone 13 photos to iPhone 12 there are basically indistinguishable in this is a barrel the continuous through our testing despite the new sensor. Phones barrel waver to upon your scoring want to have more as it used in 30 min and 30 Pro in and to have more hours. iPhone mini speaker is number one the battery life with, so many was the weak point now is not directly from bad to perfectly normal. iPhone 13 has a new camera strap. not at used both of them were pretty similar phones and hence the iPhone 12 was the best photo of the year but now after getting used to the 120 Hz refresh rate on my 13 Pro Max the 60 Hz display on the iPhone 13 is a big b and b initially. iPhone 13 is definitely not wanted@10awesome Max accessories which you can put on and take off in the second and a lot more other things that the overall experience of using this phone to a whole new Level.

Our observations are as below:

- The top four iPhone 13 keywords extracted by the key phrase score, as described in "Feature" section, are "iPhone," "camera," "accessories," and "Apple." The most popular phrase, "iPhone 13," is highlighted in the results.
- The word cloud cluster for the iPhone 13 comprises words such as "camera," "battery," "backup," "photos," "audio," etc.
- The overall sentiment of the product is positive, with a sentiment score of up to 70%.
- 69% of the questions are answered correctly using the proposed pipeline.

Figure 3 represents sentiment and word importance extracted from the audio text. Figure 4 shows the important word-based score of the dataset.

Ablation Study

In this section, we conduct ablation studies to assess the relative impact of each of our model's individual contributions. We collected customer engagement content from YouTube videos featuring technical electronic products, which we used to generate product review video synopses as reported in the article. Our proposed model achieved an accuracy of 75%. We incorporated verbal features in the proposed baseline, including key-phrase score, word-cloud-based, and sentiment-based features. Notably, the only verbal features of our proposed model are outer forms, as important segments are extracted from original videos. We specifically asked

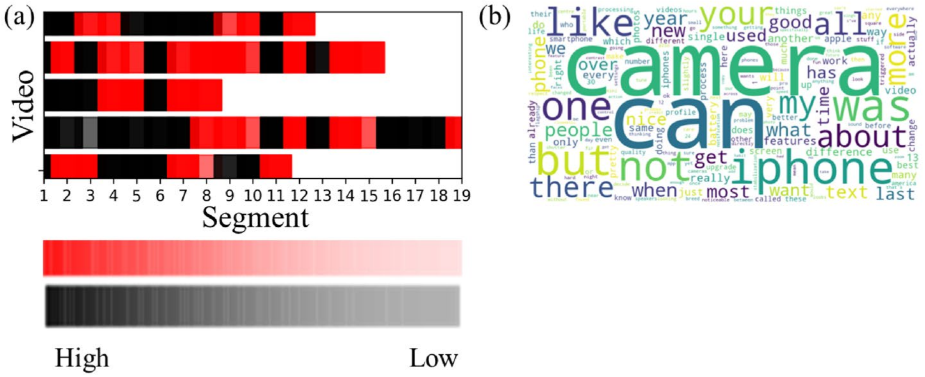

Fig. 3 Here, **a** represents the sentiment and the sentiment score of each segment of the videos. The red indicates positive sentiment, while black indicates negative sentiment. **b** shows words of varying sizes according to the frequency of occurrence in each segment of the videos

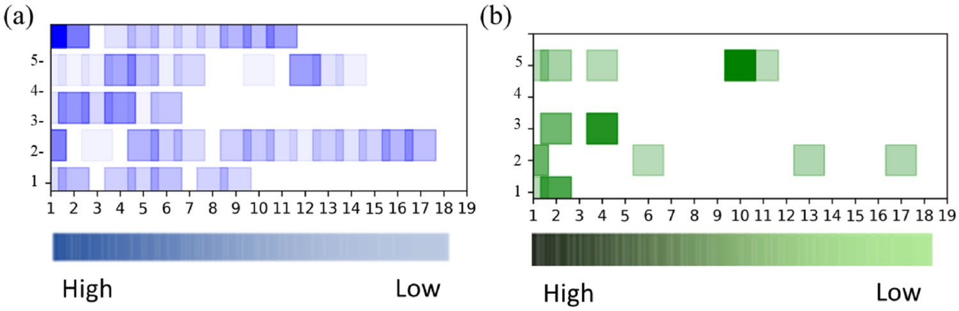

Fig. 4 Both **a**, **b** show the keyword-based score of each segment. In this case, the darkness of the color represents the score

Table 4 Ablation study of the proposed method

Method	F-measure
Key-phrase score	0.38
Word-cloud score	0.42
Sentiment score	0.48
Proposed (without average score of key-phrase and word-cloud score)	0.52
Proposed (without word-cloud score and sentiment score)	0.61
Proposed (without key-phrase score and sentiment score)	0.52
Proposed	0.75

human volunteers to verify the video summaries produced by our approach. The observed outcomes are documented in Table 4.

Conclusion

In conclusion, our work contributes significantly to enhancing online shopping accessibility for visually impaired individuals through the following key contributions:

Dataset Creation: We introduced a novel dataset specifically designed for summarizing product review videos, addressing a crucial gap in existing resources. This dataset serves as a valuable foundation for future research endeavors in this field.

Baseline Summarization Method: Our proposed baseline method offers a starting point for developing and refining video summarization techniques tailored to improve accessibility for the visually impaired.

Integration of Verbal Features: Through empirical investigations, we demonstrated the effectiveness of incorporating verbal features into the video summarization process, highlighting their potential to enhance accessibility for visually impaired users.

Future research directions: Involving visually impaired users in collecting reviews and insights can guide the development of more intuitive and effective summarization tools. Participatory design techniques can ensure that the solutions satisfy the target audience's real needs and preferences. We aim to expand on our preliminary contributions and further advance the field of accessible online experiences for people with visual impairments through this future research direction.

Efforts to enhance AI assistive accessibility frequently emphasize the necessity for models that are interpretable and comprehensible to users, particularly those with visual impairments. However, our manuscript does not delineate the methodologies to ensure that video summarization technology for online shopping incorporates explainability features specifically designed for individuals with low vision. The task of video summarization encompasses complexities such as object recognition, interaction analysis, diverse content types, and the requirement for contextual understanding. Traditional methods like Grad-CAM [74] and LIME [75] are inadequate due to their reliance on visual cues, which are inaccessible to visually impaired individuals. In the context of video summarization, initiatives are being planned or implemented to elucidate the internal mechanisms of our proposed pipeline by converting 'black box' [76] systems into more transparent frameworks with embedded explainability features.

By providing a robust dataset and foundational methodology, our work advances efforts toward fostering inclusivity and accessibility in online shopping. We aim to catalyze innovation and progress, ultimately striving for a more inclusive online experience for all users, irrespective of visual impairment.

Acknowledgements The hardware resource (GPU) of the project is funded by the Science and Engineering Research Board (SERB), Govt. of India, Project No: SRG/2022/000122, executed in XIM University, Bhubaneswar, India, supervised by Arif Ahmed Sekh.

Funding Open access funding provided by UiT The Arctic University of Norway (incl University Hospital of North Norway).

Data Availability The datasets generated during and/or analyzed during the current study will be available in the GitHub repository, (https://github.com/Ratnabali-Pal/EAOS).

Declarations

Conflict of interest Authors declare no conflict of interest.

References

1. Barra S, Bisogni C, De Marsico M, Ricciardi S. Visual question answering: which investigated applications? Pattern Recognit Lett. 2021;151:325–31.
2. Joshi RC, Yadav S, Dutta MK, Travieso-Gonzalez CM. Efficient multi-object detection and smart navigation using artificial intelligence for visually impaired people. Entropy. 2020;22(9):941.
3. Felix SM, Kumar S, Veeramuthu A. A smart personal ai assistant for visually impaired people. In: 2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI). IEEE; 2018. p. 1245–50.
4. Basavarajaiah M, Sharma P. Survey of compressed domain video summarization techniques. ACM Comput Surv (CSUR). 2019;52(6):1–29.
5. Hussain T, Muhammad K, Ding W, Lloret J, Baik SW, Albuquerque VHC. A comprehensive survey of multi-view video summarization. Pattern Recognit. 2021;109:107567.
6. Hussain T, Muhammad K, Ullah A, Cao Z, Baik SW, Albuquerque VHC. Cloud-assisted multiview video summarization using CNN and bidirectional LSTM. IEEE Trans Ind Inform. 2019;16(1):77–86.
7. Sharma V, Gupta M, Kumar A, Mishra D. Video processing using deep learning techniques: a systematic literature review. IEEE Access. 2021;9:139489–507.
8. Muhammad K, Obaidat MS, Hussain T, Ser JD, Kumar N, Tanveer M, Doctor F. Fuzzy logic in surveillance big video data analysis:

comprehensive review, challenges, and research directions. ACM Comput Surv (CSUR). 2021;54(3):1–33.

9. Bohr A, Memarzadeh K. The rise of artificial intelligence in healthcare applications. In: Artificial intelligence in healthcare. Elsevier; 2020. p. 25–60.
10. Li B, Xu X. Application of artificial intelligence in basketball sport. J Educ Health Sport. 2021;11(7):54–67.
11. Ji Z, Zhang Y, Pang Y, Li X, Pan J. Multi-video summarization with query-dependent weighted archetypal analysis. Neurocomputing. 2019;332:406–16.
12. Gaikwad D, Sarap S, Dhande D. Video summarization using deep learning for cricket highlights generation. J Sci Res. 2022;14(2):533–44.
13. Liu T, Meng Q, Vlontzos A, Tan J, Rueckert D, Kainz B. Ultrasound video summarization using deep reinforcement learning. In: International conference on medical image computing and computer-assisted intervention. Springer; 2020. p. 483–92.
14. Mäntylä MV, Graziotin D, Kuutila M. The evolution of sentiment analysis-a review of research topics, venues, and top cited papers. Comput Sci Rev. 2018;27:16–32.
15. Do HH, Prasad P, Maag A, Alsadoon A. Deep learning for aspect-based sentiment analysis: a comparative review. Expert Syst Appl. 2019;118:272–99.
16. Bi J-W, Liu Y, Fan Z-P, Zhang J. Wisdom of crowds: conducting importance-performance analysis (ipa) through online reviews. Tour Manag. 2019;70:460–78.
17. Yuan CW, Hanrahan BV, Lee S, Rosson MB, Carroll JM. Constructing a holistic view of shopping with people with visual impairment: a participatory design approach. Univ Access Inf Soc. 2019;18:127–40.
18. Alagarsamy S, Kusuma B, Mohan CVN, Sukumar MV, Sujan DVVSS, Devendrareddy M, et al. Smart system for reading the bar code using Bayesian deformable algorithm for blind people. In: 2022 6th International Conference on Trends in Electronics and Informatics (ICOEI). IEEE; 2022. p. 424–9.
19. Tapu R, Mocanu B, Zaharia T. Wearable assistive devices for visually impaired: a state of the art survey. Pattern Recognit Lett. 2020;137:37–52.
20. Fernandes H, Costa P, Filipe V, Paredes H, Barroso J. A review of assistive spatial orientation and navigation technologies for the visually impaired. Univ Access Inf Soc. 2019;18:155–68.
21. Gurari D, Li Q, Stangl AJ, Guo A, Lin C, Grauman K, Luo J, Bigham JP. Vizwiz grand challenge: answering visual questions from blind people. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. p. 3608–17.
22. Holanda GB, Souza JWM, Lima DA, Marinho LB, Girão AM, Frota JBB, Rebouças Filho PP. Development of ocr system on android platforms to aid reading with a refreshable braille display in real time. Measurement. 2018;120:150–68.
23. Boorugu R, Ramesh G. A survey on nlp based text summarization for summarizing product reviews. In: 2020 Second International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE; 2020. p. 352–6.
24. Jiao Y, Qu Q-X. A proposal for kansei knowledge extraction method based on natural language processing technology and online product reviews. Comput Ind. 2019;108:1–11.
25. Fan Z-P, Li G-M, Liu Y. Processes and methods of information fusion for ranking products based on online reviews: an overview. Inf Fusion. 2020;60:87–97.
26. Shah J, Sagathiya M, Redij K, Hole V. Natural language processing based abstractive text summarization of reviews. In: 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC). IEEE; 2020. p. 461–6.
27. Doğan E, Kaya B. Deep learning based sentiment analysis and text summarization in social networks. In: 2019 International Artificial Intelligence and Data Processing Symposium (IDAP). IEEE; 2019. p. 1–6.
28. Patel D, Shah S, Chhinkaniwala H. Fuzzy logic based multi document summarization with improved sentence scoring and redundancy removal technique. Expert Syst Appl. 2019;134:167–77.
29. Siledar T, Rangaraju R, Muddu SSRR, Banerjee S, Patil A, Singh SS, Chelliah M, Garera N, Nath S, Bhattacharyya P. Product description and qa assisted self-supervised opinion summarization. 2024. arXiv preprint arXiv:2404.05243
30. Rochan M, Ye L, Wang Y. Video summarization using fully convolutional sequence networks. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018. p. 347–63.
31. Muhammad K, Hussain T, Tanveer M, Sannino G, Albuquerque VHC. Cost-effective video summarization using deep cnn with hierarchical weighted fusion for iot surveillance networks. IEEE Internet Things J. 2019;7(5):4455–63.
32. Ji Z, Xiong K, Pang Y, Li X. Video summarization with attention-based encoder–decoder networks. IEEE Trans Circuits Syst Video Technol. 2019;30(6):1709–17.
33. Zhang K, Grauman K, Sha F. Retrospective encoders for video summarization. In: Proceedings of the European Conference on Computer Vision (ECCV). 2018. p. 383–99.
34. Rafiq M, Rafiq G, Agyeman R, Choi GS, Jin S-I. Scene classification for sports video summarization using transfer learning. Sensors. 2020;20(6):1702.
35. Guntuboina C, Porwal A, Jain P, Shingrakhia H. Deep learning based automated sports video summarization using yolo. Electron Lett Comput Vis Image Anal (ELCVIA). 2021;20(1):99–116.
36. Emon SH, Annur A, Xian AH, Sultana KM, Shahriar SM. Automatic video summarization from cricket videos using deep learning. In: 2020 23rd International Conference on Computer and Information Technology (ICCIT). IEEE; 2020. p. 1–6.
37. Khan H, Hussain T, Khan SU, Khan ZA, Baik SW. Deep multi-scale pyramidal features network for supervised video summarization. Expert Syst Appl. 2024;237:121288.
38. Kawamura K, Rekimoto J. Fastperson: enhancing video-based learning through video summarization that preserves linguistic and visual contexts. In: Proceedings of the Augmented Humans International Conference. 2024. p. 205–16.
39. Benoughidene A, Titouna F, Boughida A. Static video summarization based on genetic algorithm and deep learning approach. Multimed Tools Appl. 2024;2024:1–26.
40. Deng D. Dbscan clustering algorithm based on density. In: 2020 7th International Forum on Electrical Engineering and Automation (IFEEA). IEEE; 2020. p. 949–53.
41. Zhang Y, Liu Y, Wu C. Attention-guided multi-granularity fusion model for video summarization. Expert Syst Appl. 2024;249:123568.
42. Otani M, Nakashima Y, Rahtu E, Heikkilä J, Yokoya N. Video summarization using deep semantic features. In: Asian Conference on Computer Vision. Springer; 2016. p. 361–77.
43. Gygli M, Grabner H, Riemenschneider H, Gool LV. Creating summaries from user videos. In: European Conference on Computer Vision. Springer; 2014. p. 505–20.
44. Benkhelifa R, Laallam FZ. Opinion extraction and classification of real-time Youtube cooking recipes comments. In: International Conference on Advanced Machine Learning Technologies and Applications. Springer; 2018. p. 395–404.
45. Im J, Kim M, Lee H, Cho H, Chung S. Self-supervised multimodal opinion summarization. 2021. arXiv preprint arXiv:2105.13135.
46. Stangl AJ, Kothari E, Jain SD, Yeh T, Grauman K, Gurari D. Browsewithme: an online clothes shopping assistant for people with visual impairments. In: Proceedings of the 20th International ACM SIGACCESS Conference on Computers and Accessibility. 2018. p. 107–18.

47. Kostyra E, Żakowska-Biemans S, Śniegocka K, Piotrowska A. Food shopping, sensory determinants of food choice and meal preparation by visually impaired people. obstacles and expectations in daily food experiences. Appetite. 2017;113:14–22.
48. Yamaguchi F, Li D, Ueda M, Nakajima S. A product feature mentioned timestamp extraction method in review videos for online shopping. In: 2024 International Conference on Computing, Networking and Communications (ICNC). IEEE; 2024. p. 157–62.
49. Yao Y, Duan J, Xu K, Cai Y, Sun Z, Zhang Y. A survey on large language model (llm) security and privacy: the good, the bad, and the ugly. High-Confidence Comput. 2024;24:100211.
50. Kanafani H, Ghauri JA, Hakimov S, Ewerth R. Unsupervised video summarization via multi-source features. In: Proceedings of the 2021 International Conference on Multimedia Retrieval. 2021. p. 466–70.
51. Sadman N, Gupta KD, Haque A, Poudyal S, Sen S. Detect review manipulation by leveraging reviewer historical stylometrics in Amazon, Yelp, Facebook and Google reviews. In: Proceedings of the 2020 the 6th International Conference on E-Business and Applications. 2020. p. 42–7.
52. Rochan M, Wang Y. Video summarization by learning from unpaired data. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019. p. 7902–11.
53. Pal R, Kar S, Sekh AA. Artificial eye: online video browsing guide for visually impaired. In: International conference on computer vision and image processing. Springer; 2023. p. 410–21.
54. Campos VP, Araújo TM, Souza Filho GL, Gonçalves LM. Cinead: a system for automated audio description script generation for the visually impaired. Univ Access Inf Soc. 2020;19(1):99–111.
55. Campos VP, Gonçalves LM, Ribeiro WL, Araújo TM, Do Rego TG, Figueiredo PH, Vieira SF, Costa TF, Moraes CC, Cruz AC, et al. Machine generation of audio description for blind and visually impaired people. ACM Trans Accessible Comput. 2023;16(2):1–28.
56. Manojkumar V, Mathi S, Gao X-Z. An experimental investigation on unsupervised text summarization for customer reviews. Procedia Comput Sci. 2023;218:1692–701.
57. Gomes L, Silva Torres R, Côrtes ML. Bert-and tf-idf-based feature extraction for long-lived bug prediction in floss: a comparative study. Inf Softw Technol. 2023;160: 107217.
58. Fang K, Wu T-L, Yang D, Savarese S, Lim JJ. Demo2vec: reasoning object affordances from online videos. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. p. 2139–47.
59. Hassan A, Mahmood A. Deep learning approach for sentiment analysis of short texts. In: 2017 3rd International Conference on Control, Automation and Robotics (ICCAR). IEEE; 2017. p. 705–710.
60. Ramasamy LK, Kadry S, Nam Y, Meqdad MN. Performance analysis of sentiments in twitter dataset using svm models. Int J Electr Comput Eng. 2021;11(3):2275–84.
61. Onan A. Deep learning based sentiment analysis on product reviews on twitter. In: International conference on big data innovations and applications. Springer; 2019. p. 80–91.
62. Khurana D, Koli A, Khatter K, Singh S. Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl. 2023;82(3):3713–44.
63. Giarelis ., Kanakaris N, Karacapilidis N. A comparative assessment of state-of-the-art methods for multilingual unsupervised keyphrase extraction. In: IFIP international conference on artificial intelligence applications and innovations. Springer; 2021. p. 635–45.
64. Khan MQ, Shahid A, Uddin MI, Roman M, Alharbi A, Alosaimi W, Almalki J, Alshahrani SM. Impact analysis of keyword extraction using contextual word embedding. PeerJ Comput Sci. 2022;8:967.
65. Kabir AI, Ahmed K, Karim R. Word cloud and sentiment analysis of amazon earphones reviews with r programming language. Inform Econ. 2020;24(4):55–71.
66. Hearst MA, Pedersen E, Patil L, Lee E, Laskowski P, Franconeri S. An evaluation of semantically grouped word cloud designs. IEEE Trans Visual Comput Graph. 2019;26(9):2748–61.
67. Naldi M. A review of sentiment computation methods with r packages. 2019. arXiv preprint arXiv:1901.08319
68. Niwattanakul S, Singthongchai J, Naenudorn E, Wanapu S. Using of Jaccard coefficient for keywords similarity. In: Proceedings of the international multiconference of engineers and computer scientists, vol. 1. 2013. p. 380–4.
69. Fernandez RC, Min J, Nava D, Madden S. Lazo: a cardinality-based method for coupled estimation of Jaccard similarity and containment. In: 2019 IEEE 35th International Conference on Data Engineering (ICDE). IEEE; 2019. p. 1190–201.
70. Plummer BA, Brown M, Lazebnik S. Enhancing video summarization via vision-language embedding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. p. 5781–9.
71. Khosla A, Hamid R, Lin C-J, Sundaresan N. Large-scale video summarization using web-image priors. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2013. p. 2698–705.
72. Otani M, Nakashima Y, Rahtu E, Heikkilä J, Yokoya N. Video summarization using deep semantic features. In: Asian Conference on Computer Vision. Springer; 2017. p. 361–77.
73. Du J, Rong J, Michalska S, Wang H, Zhang Y. Feature selection for helpfulness prediction of online product reviews: an empirical study. PLoS ONE. 2019;14(12):0226902.
74. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. Int J Comput Vis. 2020;128:336–59.
75. Liu X, Han B, Qian F, Varvello M. Lime: understanding commercial 360 live video streaming services. In: Proceedings of the 10th ACM Multimedia Systems Conference. 2019. p. 154–64.
76. Akkem Y, Biswas SK, Varanasi A. Streamlit-based enhancing crop recommendation systems with advanced explainable artificial intelligence for smart farming. Neural Comput Appl. 2024;2024:1–15.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.