International Journal of Internet, Broadcasting and Communication Vol.17 No.4 53-66 (2025)
http://dx.doi.org/10.7236/IJIBC.2025.17.4.53

IJIBC 25-4-7

Video summarization with adjustable ratio of highlights and story

Eunbi Kim, Ahrin Jun, Danbi Yoon, Kitae Hwang*

Undergraduate Student, School of Computer Engineering, Hansung University, Korea
** Professor, School of Computer Engineering, Hansung University, Korea*

*jenny7551@naver.com, wjsdkfls03@naver.com, yoondb1128@naver.com, calafk@hansung.ac.kr**

Abstract

Most current video summarization techniques focus on converting long videos into short summary clips by detecting semantically important parts of the original video. However, many summarization applications, such as personal summaries of weddings, travel videos, movie trailers, or short summaries for search, require summaries that include the entire content of the original video, not just the important parts. This paper proposes a technique for generating summary videos by combining highlight scenes representing important parts of the video with story-conveying scenes capturing the overall flow of the video. To achieve this objective, we introduce the concept of 'diversity contribution,' which numerically quantifies how much each individual scene contributes to creating a diverse scene composition in the summary video. The higher the proportion of scenes with high diversity contribution values, the more comprehensively the summary video incorporates the entire content of the original video. In this study, we developed an algorithm to create a summary video by adjusting the diversity contribution and importance of scenes in proportion, and verified its operation by implementing an actual summary application system. Furthermore, we confirmed through experiments that the accuracy of the summary technique presented in this paper was over 90%.

Keywords: *Video Summarization, Video Highlight, Video Story*

1. Introduction

Recently, personal videos such as video lectures and video messages have been mass-produced and shared across social networks[1,2,3]. Furthermore, video content has proliferated on various video hosting platforms like YouTube, social networks such as Facebook and Instagram, and online video repositories. To enable users to easily search for desired videos within this vast volume of video content, it is essential to provide short summary videos that condense lengthy original videos[4,5,6].

To achieve this objective, various studies for automatic video summarization have been conducted. To date, most video summarization techniques have focused on selecting important parts of the original video to capture

Manuscript Received: September. 29, 2025 / Revised: October. 28, 2025 / Accepted: November. 2, 2025
Corresponding Author: calafk@hansung.ac.kr
Tel: +82-2-760-4300, Fax: +82-2-760-5771
Professor, School of Computer Engineering, Hansung University, Korea

its complete meaning in a concise format. Currently, the most widely adopted method for generating summary videos utilizes visual attributes of the video, such as color, motion, objects, and user attention, to identify its important parts[4,5,6]. Recently, LLM(large language model) techniques have been actively studied to convert the visual characteristics of a video into text and use this to evaluate the semantics and importance of frames in the video to create a summary video[7].

These existing studies are focused only on assigning importance scores to scenes called frames or segments and composing summary videos with parts with high importance. However, real-world video summarization applications have very different requirements[5]. For example, summaries of personal videos, such as weddings or travel videos, should be created to encompass the entire content, including key moments. Sports videos should include highlights of key events and a comprehensive overview of the entire game. Movie trailers should also focus on highlights while incorporating elements of the story. Even summaries of longer videos, designed for quick browsing, should incorporate a balanced mix of content. Likewise, existing video summarization applications require both important scenes and less important scenes for understanding the story.

This study explores a technique for generating summary videos by combining highlight scenes that represent important parts of the video with scenes conveying the story that captures the overall flow and storyline of the video. The core contributions of this paper are as follows. First, we introduce a weighting parameter w that enables adjustment of the ratio between the proportion of highlight scenes and story-conveying scenes. Second, we employ a method that utilizes the visual attributes of videos to detect highlight scenes. Third, we introduce a new concept called diversity contribution to incorporate story progression. Scenes with higher diversity contribution enhance the overall diversity of the summary video. As the diversity of scenes comprising the summary video increases, the proportion of highlight scenes decreases, while the proportion of varied scenes across the entire summary video increases. This effect causes the summary video to be made up of scenes that are relatively evenly divided across the entire time of the original video. Fourth, this paper presents a framework that evaluates scene fitness score by combining diversity contribution with scene importance(highlight scores) through the weighting parameter w, thereby generating the summary video with the scenes with highest fit. Fifth, we present an algorithm to resize segments to avoid audio interruptions in the summary video.

We implemented the proposed method and conducted performance evaluation. As a result of the performance evaluation, the performance of extracting highlights from the original video was evaluated at 93.7%. In addition, to assess the story inclusion performance of the proposed approach, we generated and compared synopses of originals and summary videos using an LLM. The average similarity score was evaluated to 92.8%, demonstrating that the summarization system proposed in this paper achieves very high performance.

2. Related Works

Over the past few decades, numerous studies have been conducted on video summarization. Recently, systems such as Google Gemini, ScreenApp, and NoteGPT that summarize videos into text have been developed and are being utilized[8]. However, our research is about techniques to summarize long videos into short videos, and these are not related to our study.

Video summarization has applications in various domains, including personal video summaries, sports

highlights, movie trailers, and summary videos for rapid video search [5]. Video summarization is the problem of selecting important video frames from among each frame included in a video, and the summarized video is a set of representative video frames from the original video. The most widely adopted approach to video summarization is selecting frames deemed representative or important based on the visual properties of video frames. Some methods utilize color histograms to select important frames. For specific applications such as surveillance cameras, sports events, or exam monitoring, important frames are often selected based on sudden changes, events, or suspicious attributes within the video. Alternatively, representative frames may be chosen by focusing on the movement of specific objects such as people, cars, or dogs. As another approach, attention -based video summarization models have been proposed that are tailored to user interests[9,10].

Meanwhile, many studies utilize models trained on datasets where humans manually assigned importance score to each frame of a video to identify important frames. There was also a study that indirectly interpreted the most viewed areas of YouTube videos as frames with high importance and created a summary model using these data sets [11]. Furthermore, to address the limitations of semantic interpretation of videos using visual features and the high cost of creating datasets with manually annotated importance scores, techniques employing Large Language Models (LLMs) are being researched. These methods convert each video frame into text captions and extract important frames from these captions using trained multimodal LLMs[7].

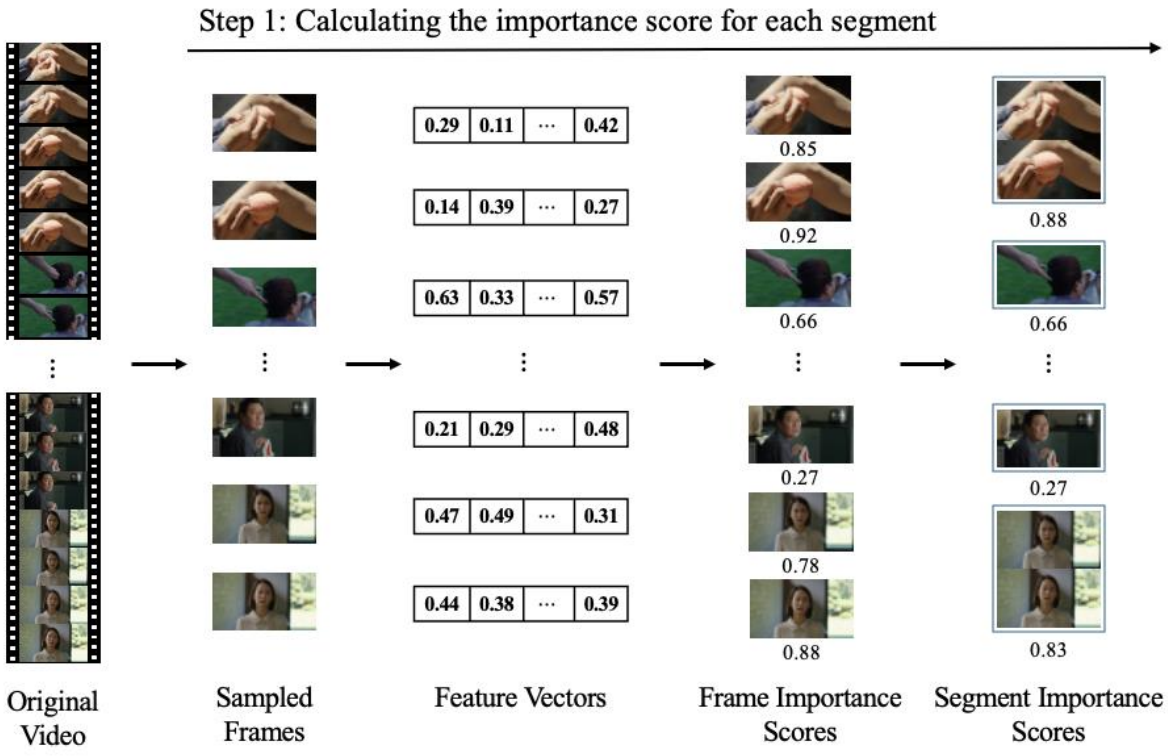
Most summarization techniques studied so far have focused on selecting important frames in a video. However, real-world video summarization applications require summary videos that capture not only the important parts but also the overall context of the entire video. In this paper, we propose a technique for creating a summary video that includes both important frames and frames containing the context of the entire video, and implement an application system that allows users to control the ratio of the two.

3. Proposed Approach

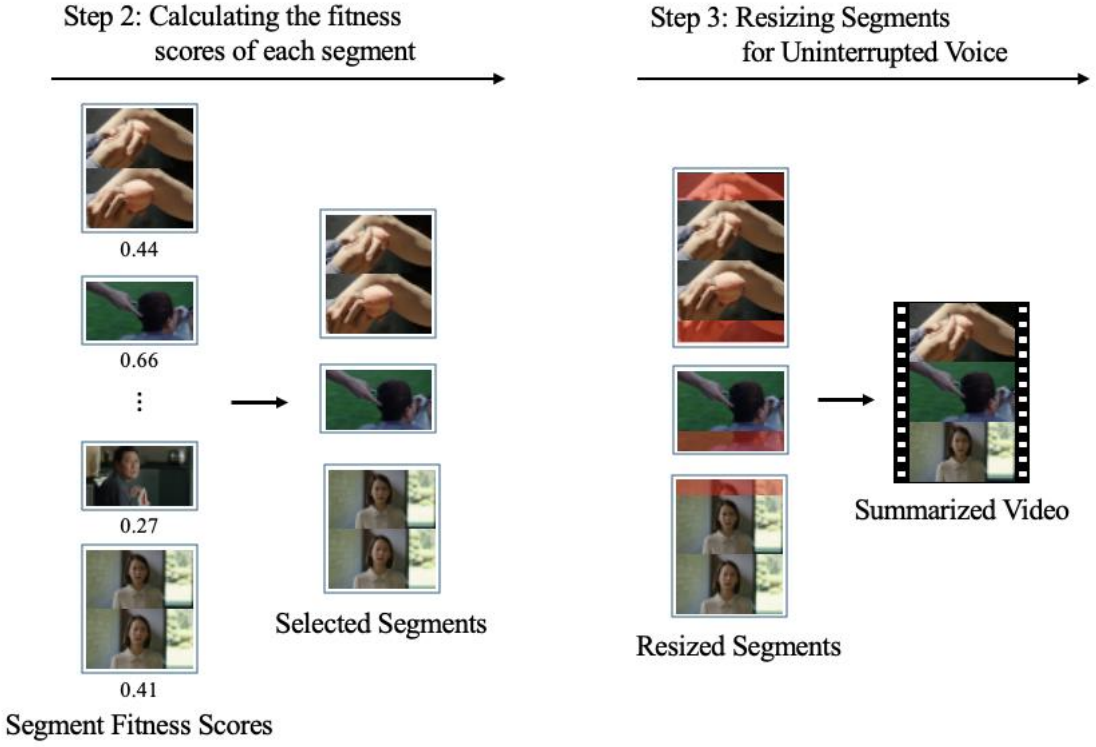
In this section, we describe the process and algorithm for video summarization by adjusting the ratio between highlight-centric and story-centric summarization.

3.1 Overview

Figure 1 depicts the video summarization model proposed in this study, which consists of three main steps. Figure 1(a) shows the first step which involves dividing the original video into multiple segments and determining the importance of each segment. Figure 1(b) shows the second step which calculates the fitness score of each segment by considering both the importance and diversity of each segment, selecting the segments with the highest fitness score while ensuring the summary video does not exceed the length constraint. Finally Figure 1(c) shows the third step which resizes all selected segments to minimizes audio discontinuities and connects them to generate the final summary video.



(a) Step 1: Calculating Importance Score for Each Segment



(b) Step 2: Calculating Fitness Score **(c) Step 3: Resizing Segments for Uninterrupted Voice**

Figure 1. Overview of our approach

The first step is to divide the video into segments and determine their importance. In this paper, a segment is a single scene composed of several consecutive frames. In this step, we assign importance scores to representative frames and simultaneously segment the video into scenes using the TransnetV2 model. Then, the importance score of each scene is calculated by adding up the importance scores of the representative frames belonging to each scene. Representative frames are selected at one-second intervals, and scores are assigned only to them. Subsequently, the importance score for each scene is calculated by aggregating the importance scores of all representative frames belonging to that scene. The importance score of a scene is the highlight score of the scene. Scenes are selected in order of importance score so that the length of the summary video is less than the target length L, and these are connected to create a summary video with 100% highlights. Connecting these selected scenes generates a summary video composed entirely of highlights. Since this paper summarizes the video based on the user-specified highlight and story summary ratio value w, the importance value of each scene is recalculated by reflecting the w value and scenes with high importance scores are selected. During this process, it ensures that the selected scenes do not exceed the target summary length L. Finally, it examines the audio of each selected scene and adjusts the scene length to prevent audio discontinuities, thereby generating the summary video.

3.2 Step1: Calculating the importance score for each segment

3.2.1 Extracting visual features

First, we utilized the pretrained Inception V3 model[12], trained on the ILSVRC 2012 (ImageNet) classification benchmark, to analyze the visual features of each representative frame and generated a visual feature vector consisting of 2048 values representing color, texture, shape, objects, and other visual elements. The original training of the Inception V3 model relied on the RMSProp optimizer with a decay of 0.9 and $\varepsilon = 1.0$, utilizing a starting learning rate of 0.045 that was exponentially decayed by 0.94 every two epochs. To ensure stability, the training incorporated gradient clipping with a threshold of 2.0. This information is utilized in subsequent steps to determine the importance of frames and assess the similarity of segments. Since the trained Inception V3 model achieves optimal performance with input images of 299×299 pixels, all frames were standardized to 299×299. Additionally, the vector size was reduced to 1024 dimensions to match the 1fps, 1024-dimensional Inception-V3 feature specification used in the Mr.HiSum dataset[13].

3.2.2 Calculating the importance score of each frame

Video summarization is a matter of selecting which frames from the original video. Therefore, the process begins with assigning importance scores to individual representative frames. In this paper, we utilized the PGL_SUM model[11] trained on the Mr.HiSum dataset[13] to assign importance scores ranging from 0 to 1 to each representative frame in the video. Figure 2 illustrates an example of importance score assignment for individual representative frames. In this paper, we will refer to the representative frame simply as a frame from now on.

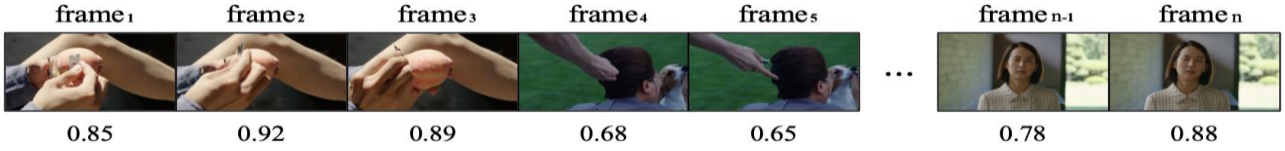


Figure 2. The importance score for each frame

3.2.3 Segmentation

If a summary video is created by simply connecting the selected frames based on the importance scores assigned to the frames, the summary video will not be natural visually or semantically. Therefore, it is desirable to recognize multiple frames that are visually, temporally, or spatially continuous as a single scene and create a summary video by connecting appropriate scenes. In this paper, we employed the pretrained TransNet V2 model[14] for scene detection in videos. Scene changes were identified using the model's default threshold of 0.5. We refer to these scenes as segments throughout this paper.

3.2.4 Calculating the segment score

After dividing the video into segments, we calculate the importance score of each segment based on importance scores assigned to frames belonging to each segment. For this, we first compute the mean μ_i, maximum m_i, and standard deviation σ_i of importance scores of all frames contained within s_i for all segments $\{s_1, s_2, \cdots, s_n\}$. We then define the importance score $I(_i)$ for s_i using the following formula.

$$I(_i) = \alpha \cdot \mu_i + (1 - \alpha) \cdot m_i - \beta \cdot \sigma_i \quad (1)$$

In formula (1), the coefficient $\alpha \in [0,1]$ is a value to balance the relative contribution between mean and maximum values, and a coefficient $\beta \geq 0$ is a value to mitigate excessive fluctuations in importance evaluation by incorporating the variability of frame importance scores. The larger the α value, the more the importance of the frames belonging to the segment is emphasized on average, and the smaller the α value, the more the importance of a specific frame belonging to the segment is reflected. This approach ensures that frames reflecting critically important situations receive higher importance scores. The parameter σ adjusts the segment score based on the standard deviation of importance scores among frames, preventing overestimation of segments with high score variability.

In our system, we set $\alpha = 0.7$ to prioritize stable evaluation based on mean values, and set $\sigma = 0.3$ as the deviation adjustment coefficient. These parameters represent an optimized configuration that prioritizes segments with consistently high importance while mitigating the adverse effects of excessive variation among frames within segments. Figure 3 shows a sample of the assignment of importance scores to each segment.

Figure 3. The importance score for each segment

3.3 Step2: Calculating the fitness score of each segment

3.3.1 Calculating the diversity contribution of each segment

To create a highlight-focused summary video, one can simply concatenate segments with high importance scores. However, a well-summarized story-based video is a collection of evenly selected segments throughout

the video. To create a strong story-driven video, one need to select segments across the entire video, selecting segments that have low similarity to each other.

For this purpose, this paper introduces the concept of diversity. Diversity refers to the degree to which segments selected for the summary video contain distinct visual information from one another. The goal of diversity is to include as many different scenes as possible by reducing the proportion of similar footage within the summary video.

To achieve diversity in summary videos, we must first calculate the diversity contribution each segment makes to the entire summary video. The diversity contribution is computed using cosine similarity based on the visual feature vectors between segments. For this purpose, the following notations are defined.

- $V = \{s_1, s_2, \ldots, s_N\}$: The entire set of video segments
- $A \subset V$: The set of summary segments selected so far
- $j \in V \setminus A$: The candidate segment under consideration for selection
- c_j : The length of segment j
- $I(j)$: The importance score of segment j
- $D(j \mid A)$: The diversity contribution that segment j increases when it is added to the current set A.
- $sim(i,j)$: The cosine similarity between segments i and j
- $w \in [0,1]$: A weight that adjusts the balance between story-based summarization and highlight-based summarization
- L : The maximum allowable length of the summary video
- $v(j)$: The final value of candidate segment j
- $f(j)$: The final fitness of candidate segment j

Now, let us calculate the diversity contribution of each segment. Let A be the set of summary segments selected so far from the entire segment set $V = \{s_1, s_2, \ldots, s_N\}$, and candidate segment j is repeatedly searched until the length of A does not exceed L. The importance score $I(j)$ and length c_j and of each candidate segment has been calculated in the previous step. The diversity contribution $D(j \mid A)$ represents how much segment j enhances the diversity of the summary. This paper utilizes the concept of coverage to calculate the diversity contribution. Coverage refers to how well a specific segment i is represented by the set A of currently selected summary segments and is defined as follows.

$$Coverage(i, A) = max_{s \in A}\,(sim(i, s)) \tag{2}$$

In formula (2), $sim(i,s)$ is the visual similarity between segments i and s, which is the value calculated by cosine similarity between vectors using the previously extracted visual feature vectors. That is, the similarity value with the already selected segment s that is most similar to i becomes the coverage of i.

The coverage increase is a value that indicates how much the coverage for all segments i increases compared to the existing one when selecting candidate segment j, and is defined as follows.

$$\Delta Coverage(i, j, A) = max\big(Coverage(i, A), sim(i, j)\big) - Coverage(i, A) \tag{3}$$

In formula (3), $\max\big(Coverage(i,A), sim(i,j)\big)$ means the larger value between the cosine similarity of segment i and segment j and the similarity between segment i and the segment with the highest similarity among the set A of currently selected segments. The formula (3) indicates how much the representativeness of segment i is improved compared to before by adding segment j to set A. The sum of these coverage increases, each multiplied by the importance $I(i)$ of segment i, constitutes the diversity contribution $D(j \mid A)$, which is defined as follows.

$$D(j|A) = \sum_{i \in V} I(i) \cdot \Delta Coverage(i,j,A) \tag{4}$$

In formula (4), the reason why we multiply by importance $I(i)$ is that more important segments have a greater impact on the story. Even among segments with the same coverage increase, selecting those with higher importance better conveys the story.

In this way, the diversity contribution quantitatively represents how much new information a candidate segment adds to the existing selected set. Segments with high diversity contribution have less information overlap with already selected segments, and greatly improve the representativeness of important segments. Therefore, segment i with high $I(i)$ and low similarity to the existing selected set A has a higher probability of being selected.

Existing diversity measures primarily focus on maximizing dissimilarity between segments to avoid selecting visually similar scenes together. While this approach allows for the inclusion of seemingly diverse scenes, it fails to capture the contribution of new information each segment adds to the overall summary, i.e., the expansion of meaningful content. As a result, the summary retains diversity but suffers from a loss of representativeness. In contrast, our proposed diversity contribution quantifies how much new information each segment contributes to the existing summary set as a coverage increase(ΔCoverage), and scales this contribution by the importance of each segment. This enables the inclusion of new information while ensuring that scenes crucial to the storyline are maintained, thereby achieving a higher level of semantic and representational balance than simple dissimilarity-based diversity.

3.3.2 Calculating the diversity contribution of each segment

Now, the value $v(j)$ of representing the value of candidate segment j is defined by combining the importance $I(j)$ and the diversity contribution $D(j \mid A)$ in a weighted sum, as follows.

$$\mathrm{v}(j) = (1-w) \cdot I(j) + \mathrm{w} \cdot D(\mathrm{j}|\mathrm{A}) \tag{5}$$

In formula (5), $w \in [0,1]$ is a weighting parameter that controls the balance between highlight-centric (importance-based) and story-centric (diversity-based) summarization. As w approaches 0, segments with high importance are increasingly prioritized; as w approaches 1, segments with high diversity contribution are increasingly prioritized. The weighting parameter $w \in [0,1]$ is manually determined by the user to control the balance. For story-centric summarization, one might consider simply selecting segments across the entire video at equal time intervals. However, this method increases the tendency for important segments to be omitted, resulting in poor story representation. Therefore, this paper selects segments based on diversity contribution that reflects importance, ensuring that segments crucial for story formation are included while

also achieving balanced segment selection across the entire video.

To reflect the efficiency of short and high-value segments, we define the final fitness score $f(j)$ of segment j as the value per unit length (seconds).

$$\mathrm{f(j)} = \frac{v(j)}{c_j} \tag{6}$$

In formula (6), $v(j)$ is the value of the segment j and c_j is the length of the segment j. This value serves as the criterion for selecting segments that can most efficiently pack information within the maximum allowable time L of the summary video.

3.3.3 Final selection of segments for summary video

The final segment selection process utilizes a method for solving the 0-1 Knapsack problem, with the algorithm depicted in Figure 4. This process begins with an empty set A of summary video segments and iteratively adds segments with the highest $f(j)$ values to this set, ensuring the total video length does not exceed L. The algorithm represented by Figure 5 has a time complexity of O(nxm) for selecting m segments from n when the number of segments in the original video is n and the number of segments in the summarized video is m. And The requirements for the size of memory or auxiliary storage devices are not particularly large. Even if this algorithm is run at the user's request in a system with a large number of videos, it is not considered a major problem because video summarization does not require real-time processing that requires an extremely immediate response. We leave improving the algorithm to reduce time complexity for future research.

Algorithm Final Selection of Segments for Summary Video

Require: Candidate segment set $V = \{s_1, s_2, \ldots, s_N\}$
Require: Segment lengths $\{c_1, c_2, \ldots, c_N\}$
Require: Maximum summary length L
Ensure: Selected segment set A

$A \leftarrow \emptyset$
$T \leftarrow 0$
while $T < L$ **do**
  $U \leftarrow V \setminus A$
  if $U = \emptyset$ **then break**
  end if
  for each $j \in U$ **do**
    Compute $v(j)$ for current A
    $f(j) \leftarrow v(j)/c_j$
  end for
  $j^* \leftarrow \arg\max_{j \in U} f(j)$
  if $T + c_{j^*} \leq L$ **then**
    $A \leftarrow A \cup \{j^*\}$
    $T \leftarrow T + c_{j^*}$
  else
    break
  end if
end while
return A

Figure 4. Algorithm for final selection of segments

3.4 Step3: Segment resizing for uninterrupted voice

Since segments were divided based on visual features, segment boundaries may not align with the start and end points of speech utterances. Therefore, concatenating segments without considering speech boundaries would result in an unnatural summary video with broken voices. In this study, we performed segment resizing for each segment in the summary set A, resizing segments both forward and backward to encompass the start and end points of speech utterances. This section describes this process.

3.4.1 Speech segment extraction

First, the audio portion is extracted from the original video. Then, using the pretrained Whisper and pretrained Silero-VAD models, noise is removed, speech segments are detected, and the start and end points of each segment are identified and stored. Silero-VAD uses a manually set threshold of 0.3, and no additional parameter tuning or model optimization was applied. Whisper transcribed the detected segments using default inference parameters with word-level timestamps.

3.4.2 Segment Boundary Adjustment and Overlap Resolution

Each segment s_i in the summary set $A = \{s_1, s_2, \ldots, s_N\}$ consists of a tuple $[s_i{}^{start}, s_i{}^{end}]$ of a start point $s_i{}^{start}$ and an end point $s_i{}^{end}$, and all speech regions are extracted from the original video and are constructed as a set $V = \left[[\, v_1{}^{start}, v_1{}^{end}], \cdots, [v_N{}^{start}, v_N{}^{end}]\right]$. Then, for each segment, we compare whether it falls within the set V. If it does, the segment's start and end points are adjusted to match the speech region according to the following equation

$$s'^{start} = min(\, s^{start}, v^{start}) - \delta_{before} \quad (7)$$

$$s'^{end} = max(\, s^{end}, v^{end}) + \delta_{after} \quad (8)$$

In formula (7) and (8), s'^{start} and s'^{end} are the adjusted start and end boundaries for segment s, respectively. In formula (7), $min(\, s^{start}, v^{start})$ means the smaller value of s^{start} and v^{start}, and the start point of segment s is moved back by δ_{before}. In formula (8), $max(\, s^{end}, v^{end})$ means the larger value of s^{end} and v^{end}, and the end point of segment s is moved forward by δ_{after}. δ_{before} and δ_{after} represent the margin times for boundary adjustment, and we set them to 0.2 seconds and 0.35 seconds, respectively.

4. Evaluation

This section presents the evaluation of the performance of the summarization technique proposed in this paper.

4.1 Overview

The performance evaluation of the summarization technique proposed in this paper was conducted from two perspectives. First, we evaluated whether the highlight sections of the original video were well reflected in the summary video. Second, we assessed whether the story of the original video was well captured in the summary

video. For this evaluation, we performed experimental analysis using 50 randomly selected YouTube videos. To evaluate the first aspect of performance, we examined whether the YouTube Most Replayed sections representing highlight areas were included in the summary videos generated using our proposed method. Second, to verify performance if the latter, we evaluated the similarity between synopses generated by the LLM for the original videos and those generated for the summary videos.

4.2 Evaluation Result

4.2.1 Evaluation of YouTube Most Replayed Section Matching

YouTube provides 'Most Replayed' section information for each video based on repeat playback data. This information can be considered highlight data representing the important parts of a video, as it constitutes objective data based on the behavior of a large number of viewers.

We selected 50 videos in 9 genres from YouTube and summarized them, as shown in Table 1. Then, we evaluated the performance by counting how many Most Replayed Sections in the original video is comprised in the summarized video for each video based on the information provided by YouTube. The evaluation results showed that the average match was 93.7%. This result demonstrates that our video summarization technique captures points deemed important by users.

Table 1. Dataset used for experiments

Item	Details
Number of videos	50
Original video length	20 min ~ 35 min
Summary ratio	20%
Video genres	Movie (10), Variety (8), Vlog (8), Drama (6), Documentary (5), Animation (5), Cooking (3), Review (3), Game (2)
Video source	YouTube

Figure 5 is a part from a short YouTube clip of the movie 'My Neighbor Totoro', illustrating this comparison. The gray areas in the unfolded video frames indicate YouTube's Most Replayed sections, while the red bars represent segments included in the summary video. This video has four Most Replayed sections (S1, S2, S3, S4), and you can see that they are all selected as segments of the summary video.

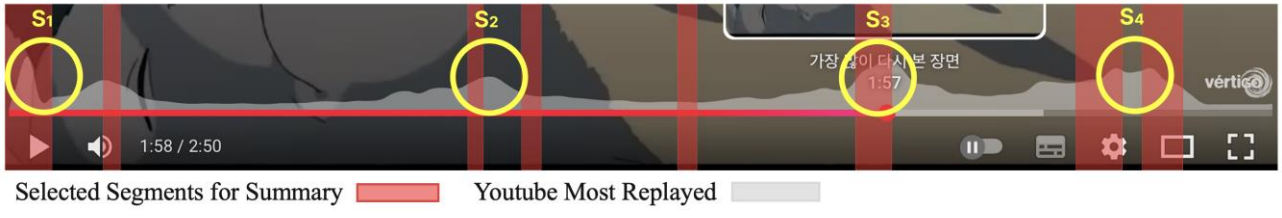


Figure 5. Comparison of Most Replayed Sections and segments in the summary video

4.2.2 LLM-Based Synopsis Consistency Evaluation

To evaluate whether the video summarization technique proposed in this paper effectively captures the overall narrative structure and context of the original video in the summary video, we evaluated a synopsis consistency using Google Gemini 2.5 Pro, an LLM. First, we generated synopsis texts for 50 original YouTube videos and summarized videos using Google Gemini 2.5 Pro., respectively. The same dataset described in Table 1 was used for this evaluation. Then, we had Google Gemini 2.5 Pro evaluate the consistency of the core topics, flow of events, and key information between the two texts, and achieved high performance with an average of 92.8% agreement. Figure 6 shows the results of comparing the synopsis of the original video and the synopsis of the summary video for a sample of the movie 'My Neighbor Totoro'. Figure 6(a) is the synopsis of the original video, Figure 6(b) is the synopsis of the summary video, and Figure 6(c) is five sample video frames from which text synopses were derived. These experimental results demonstrate that the story-centric summary video proposed in this paper effectively preserves the overall meaning and context of the original video.

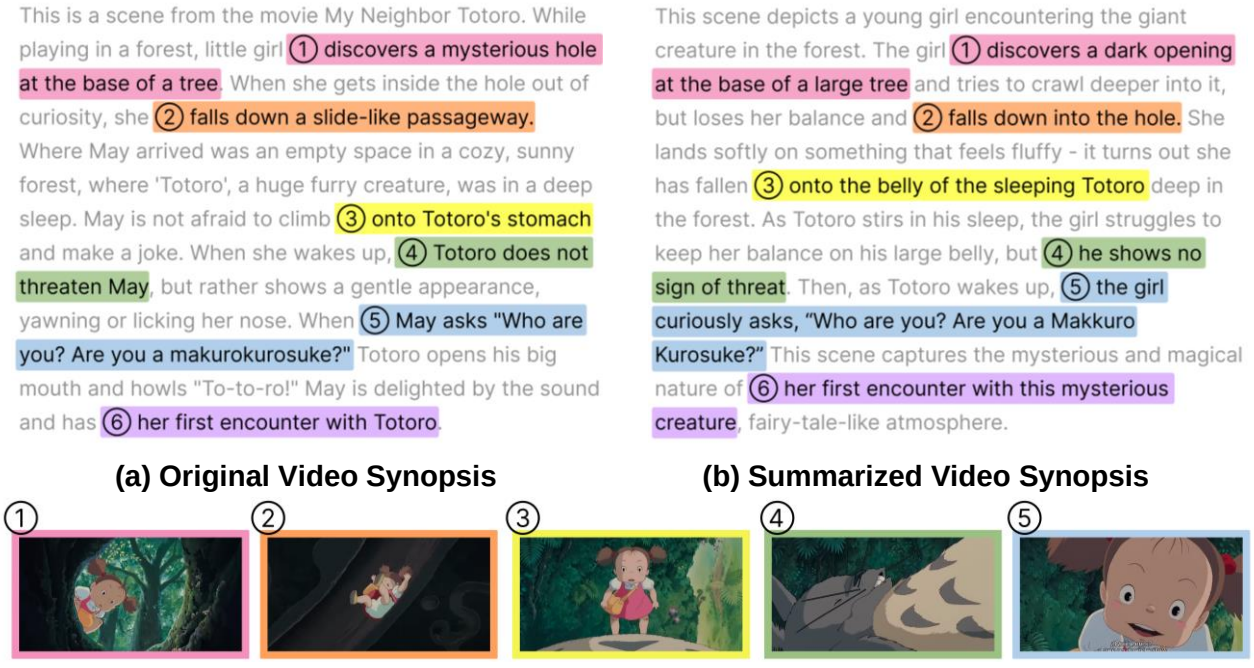


Figure 6. Comparison of synopses of original and summarized videos generated by LLM

4.3 Discussions and future works

We have some points to discuss when considering the results of the performance evaluation. First, our method does not work for all videos. The videos we selected as experimental subjects in Section 4.2.1 were videos with highlights. Our method does not work effectively for boring videos with no distinctive features. Also, it does not work effectively for videos that require expertise in interpreting highlights, such as sports videos. In addition, it does not work effectively news or product reviews where the highlights are not strong and the audio continues seamless, making it difficult to segment them without interruption. Additionally, even within the same genre, summary performance varies greatly depending on the content of the video. We believe that further detailed research and experiments are needed on these issues.

Second, although the diversity distribution method proposed in this paper for story-centered summarization is based on visual features and cosine similarity between segments rather than the content of the video, the experimental results in Section 4.2.2 confirm that the context of the original video is relatively well reflected in the summarized video. This method, which utilizes cosine similarity, has the advantage of saving processing time because it uses the visual feature data of the initially analyzed video to find the highlight segments. f additional methods such as the LLM are used to understand the content of the original video, the summarization time may increase further. However, we believe that future work is needed to study methods that directly and deeply capture the context of a video, such as LLM, to compare their performance with the current method or to include them in our system.

5. Conclusion

Most current video summarization techniques focus on identifying important segments within long videos and connecting them to create short videos. However, many video summarization applications require not only important segments but also the inclusion of the entire story from the original video. Therefore, this paper proposed and implemented a video summarization technique that can adjust the ratio between highlights and story.

The contributions of this paper are as follows. First, to include the storyline of the original video in the summary video, we introduced the concept of diversity contribution for each segment and selected segments with high diversity contribution to ensure even distribution across the entire duration. Second, we introduced a weight value w to adjust the ratio of highlight segments and story-inclusive segments, allowing users to set this value. Third, we proposed and implemented a segment resizing algorithm within the summary video so that the voice does not break up. Fourth, we implemented a video summarization system and validated its summarization performance through experimental evaluation. Performance evaluation results demonstrated that the summary video not only effectively incorporate important parts of the original video but also effectively preserve the original video's context.

However, this system has limitations in its use in all application areas due to its uncertain performance depending on the genre and content of the video. There are still challenges to be solved through further experiments and the combination of techniques that more closely analyze the context of the video.

Acknowledgement

This research was financially supported by the Hansung University

References

[1] Yifan Cui, Xinyi Shan, Jeanhun Chung, "A Feasibility Study on RUNWAY GEN-2 for Generating Realistic Style Images," International Journal of Internet, Broadcasting and Communication(IJIBC), vol. 16, no. 1, pp.99-105, 2024. DOI:https://doi.org/10.7236/IJIBC.2024.16.1.99

[2] Jiyoon Ok, Juyeon Soung, Chaewon Park, and Kitae Hwang, "Implementation Details of EPUB Reader using GraphRAG," International Journal of Internet, Broadcasting and Communication(IJIBC), vol.17, no.2, pp.223-231, 2025, DOI: https://doi.org/10.7236/IJIBC.2025.17.2.223

[3] Zong-Yi Zhu, Sun Congxi, Hyeon-Cheol Kim, "Short Video Influence on Food Tourism: Examining the Role of TikTok in Shaping City Image and Travel Intentions of Chinese Gen Z," International Journal of Internet,

Broadcasting and Communication(IJIBC), vol. 17, no. 2, pp.172-182, 2025. DOI:https://doi.org/10.7236/IJIBC.2025.17.2.172

[4] E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, "Video Summarization Using Deep Neural Networks: A Survey," Proceedings of the IEEE, vol. 109, no. 11, pp. 1838-1863, Nov. 2021. DOI: https://doi.org/10.1109/JPROC.2021.3117472

[5] H. B. U. Haq, M. Asif, and M. B. Ahmad, "Video Summarization Techniques: A Review", International Journal of Scientific & Technology Research, vol. 9, issue 11, pp. 146-152, Nov. 2020. DOI: http://doi.org/10.1109/CVPR.2013.350

[6] A. Rav-Acha, Y. Pritch, and S. Peleg, "Making a Long Video Short: Dynamic Video Synopsis," in Proc. 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR), New York, NY, USA, pp. 435-441, 2006. DOI: https://doi.org/10.1109/CVPR.2006.179

[7] M. J. Lee, D. Gong, and M. Cho, "Video Summarization with Large Language Models", in Proc. Computer Vision and Pattern Recognition Conference (CVPR), pp. 18981-18991, 2025. DOI: https://doi.org/10.48550/arXiv.2504.11199

[8] S. Yeung, A. Fathi, and L. Fei-Fei, "Videoset: Video summary evaluation through text," arXiv preprint arXiv:1406.5824.2014. DOI: https://doi.org/10.48550/arXiv.1406.5824

[9] Y.-F. Ma, X.-S. Hua, L. Lu, and H.-J. Zhang, "A generic framework of user attention model and its application in video summarization," IEEE Transactions on Multimedia, vol. 7, no. 5, pp. 907-919, Oct. 2005. DOI: https://doi.org/10.1109/TMM.2005.854410

[10] J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, "Summarizing videos with attention," in Proc. Asian Conference on Computer Vision (ACCV), pp. 39-54, 2018. DOI: https://doi.org/10.1007/978-3-030-21074-8_4

[11] E. Apostolidis, G. Balaouras, V. Mezaris, and I. Patras, "Combining Global and Local Attention with Positional Encoding for Video Summarization," in Proc. 2021 IEEE International Symposium on Multimedia (ISM), Naples, Italy, pp. 226-234, 2021. DOI: https://doi.org/10.1109/ISM52913.2021.00045

[12] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," in Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, pp. 2818-2826, 2016. DOI: https://doi.org/10.1109/CVPR.2016.308

[13] J. Sul, J. Han, and J. Lee, "Mr. HiSum: A large-scale dataset for video highlight detection and summarization," in Proc. Advances in Neural Information Processing Systems (NeurIPS), vol. 36, pp. 40542-40555, 2023. DOI:https://doi.org/10.48550/arXiv.2504.18689

[14] T. Soucek and J. Lokoc, "Transnet v2: An effective deep network architecture for fast shot transition detection," in Proc. 32nd ACM International Conference on Multimedia, pp. 11218-11221, 2024. DOI: https://doi.org/10.1145/3664647.3685517