Character-Based Video Summarization

Abhijith Suresh
Information Technology
Rajagiri School of Engineering & Technology
Kakkanad, Kerala

Aevin Tom Anish
Information Technology
Rajagiri School of Engineering & Technology
Kakkanad, Kerala

Simon Alexander
Information Technology
Rajagiri School of Engineering & Technology
Kakkanad, Kerala

Sreelakshmi S
Information Technology
Rajagiri School of Engineering & Technology
Kakkanad, Kerala

Tinku Soman Jacob
Asst.Professor, Information Technology
Rajagiri School of Engineering & Technology
Kakkanad, Kerala

Abstract:- The environment of video production and consumption on social media platforms has undergone a significant change as a result of the widespread usage of the internet and reasonably priced video capture devices. By producing a brief description of each video, video summarizing helps viewers rapidly understand the content of videos. However, standard paging techniques can result in a severe burden on computer systems, while artificial extraction might be time-consuming with a high quantity of missing data. The multi-view description of videos can also be found in the rich textual content that is typically included with videos on social media platforms, such as subtitles or bullet-screen comments. Here, a novel framework for concurrently modelling visual and textual information is proposed in order to achieve that goal. Characters are found randomly using detection techniques, identified by re-identification modules to extract probable key-frames, and then the frames are aggregated as a summary. The subtitles and bullet-screen remarks are also used as multi-source textual information to create a final text summary of the target character from the input video.

Keywords:- *Video, Summary, Subtitles, Bullet-Screen, Frames.*

I. INTRODUCTION

The widespread use of the internet and affordable video capturing devices have dramatically changed the landscape of video creation and consumption in social media platforms the purpose of video summaries varies considerably depending on application scenarios for sports viewers want to see moments that are critical to the outcome of a game whereas for surveillance video summaries need to contain scenes that are unusual and noteworthy the application scenarios grow as more videos are created eg there are new types of videos such as video game live streaming and video blogs vlogs this has led to a new problem of video summarization as different types of videos have different characteristics and viewers have particular demands for summaries such a variety of applications have stimulated heterogeneous research in this field.

Unfortunately, manual summarization could be time-consuming with a high missing rate. Usually, fans are keen on summarizing the clips of their favourite star from movies or TV series but printed summaries of content are hard to create and read such that it is highly time-consuming and video summaries will provide a visual treat. Therefore, adequate techniques to automatically extract the character-oriented summarization are urgently required. Indeed, the character-oriented video summarization task is quite different from the traditional video summarization. An ordinary video summary is expected to consist of important or interesting clips of a long video. However, a character-oriented video summary should consist of the clips in which the specific character appears. Thus, to fulfill character-oriented summarization, it is indispensable to identify the clips of a specific person from a long video to perform a person search. The person search task has been tackled by, for example, joint modelling of detection and re-identification, or the memory-guided model. However, these prior arts are designed for a different scenario, e.g. surveillance video analysis. Thus, they mainly focus on the person search with a relatively static pose and background, or even similar clothing. In the scenario of character-oriented summarization, both background and poses of characters are always changing, as well as the different clothing in different scenarios, which extremely increases the difficulty. Obviously, current person search techniques should be further enhanced. At the same time, on social media platforms, videos are usually accompanied by rich textual information, which may benefit the understanding of media content. Especially, with the so-called “bullet-screen comments” (i.e., comments flying across the screen like bullets), namely the time-sync feedback of massive users, more comprehensive or even subjective description could be achieved, which results in more explicit cues to capture the characters.

To that end in this particular project a novel framework for jointly modelling visual and textual information for character-oriented summarization is proposed to be specific we first locate characters indiscriminately through detection methods then identify these characters via the re-identification module to extract potential key-frames and finally aggregate the frames as summarization moreover as multi-source textual information ie the subtitles and bullet-screen comments are utilized. Experiments on real-world datasets validate that this solution outperforms several state-of-the-art baselines and further reveal some rules of the semantic matching between characters and textual information.

II. LITERATURE SURVEY

A. *CNN and HEVC Video Coding Features for Static Video Summarization*

In this paper video summarization by automatically selecting keyframes, which is the most common method among the two main techniques for summarizing videos. Examining all frames within a video for selection can be slow and waste time and computational resources on redundant or similar frames. Additionally, it is important to use space reduction to accelerate the process and ensure that only meaningful features are considered. To address these issues, this study will utilize deep learning techniques, such as conventional neural networks and random forests, which have gained popularity in recent years for generating tasks in image and video processing.

With the advancement of the High-Efficiency Video Codec (HEVC) video standard, the deep learning community often overlooks the potential of the information contained in the HEVC video bitstream. This research aims to address this gap by using low-level HEVC features in combination with CNN features to aid in video summarization. This study also introduces a new method for reducing the dimension of the feature space obtained from CNNs using stepwise regression, as well as a novel approach for eliminating similar frames based on HEVC features.

B. *A memory network approach for story-based temporal summarization of 360 videos*

This paper proposes a method for creating a summary of a 360-degree video based on the story being told in the video. The method involves using a memory network to analyze the content and structure of the video and select keyframes that are important for telling the story. The selected keyframes are then assembled into a summary video that tells the story in a condensed form.

The authors propose using their method for summarizing 360-degree videos that are too long or contain too much excess motion to be easily interpretable. They also suggest that their method could be used to create summaries of other types of videos, such as traditional flat videos or live-streaming videos. Overall, the proposal put forward in this paper is a method for creating summaries of 360-degree videos that are able to focus on the content of the video and the story being told, while also maintaining a coherent and coherent narrative structure.

C. *Scene Summarization via Motion Normalization*

This paper proposes a method for creating a summary of a 360-degree video based on the story being told in the video. The method involves using a memory network to analyze the content and structure of the video and select keyframes that are important for telling the story. The selected keyframes are then assembled into a summary video that tells the story in a condensed form.

The authors propose using their method for summarizing 360-degree videos that are too long or contain too much excess motion to be easily interpretable. They also suggest that their method could be used to create summaries of other types of videos, such as traditional flat videos or live-streaming videos. Overall, the proposal put forward in this paper is a method for creating summaries of 360-degree videos that are able to focus on the content of the video and the story being told, while also maintaining a coherent and coherent narrative structure.

D. *Similarity Based Block Sparse Subset Selection for Video Summarization*

This paper proposes a method for creating a summary of a 360-degree video based on the story being told in the video. The method involves using a memory network to analyze the content and structure of the video and select keyframes that are important for telling the story. The selected keyframes are then assembled into a summary video that tells the story in a condensed form.

The authors propose using their method for summarizing 360-degree videos that are too long or contain too much excess motion to be easily interpretable. They also suggest that their method could be used to create summaries of other types of videos, such as traditional flat videos or live-streaming videos. Overall, the proposal put forward in this paper is a method for creating summaries of 360-degree videos that are able to focus on the content of the video and the story being told, while also maintaining a coherent and coherent narrative structure.

E. *HSA-RNN: Hierarchical structure-adaptive RNN for video summarization*

This paper proposes a method for creating a summary of a 360-degree video based on the story being told in the video. The method involves using a memory network to analyze the content and structure of the video and select keyframes that are important for telling the story. The selected keyframes are then assembled into a summary video that tells the story in a condensed form.

The authors propose using their method for summarizing 360-degree videos that are too long or contain too much excess motion to be easily interpretable. They also suggest that their method could be used to create summaries of other types of videos, such as traditional flat videos or live-streaming videos.Overall, the proposal put forward in this paper is a method for creating summaries of 360-degree videos that are able to focus on the content of the video and the story being told, while also maintaining a coherent and coherent narrative structure.

III. METHODS

A. *Identification*

The detection module seeks to identify any character-containing regions of interest (RoIs). It makes use of the Faster R-CNN detector, which is renowned for its capacity to find objects of various sizes in a variety of scenarios. Additionally, to solve overfitting and inference-time mismatch concerns and improve performance, the cutting-edge Cascade R-CNN detector is used. It's crucial to remember that the detectors can be changed as necessary. All characters are positioned at this point without difference. The Detection module determines whether or not a character is present in a RoI in order to increase performance. The classifier's performance is ensured by enough character-oriented frame training data. As shown in Figure 3.4, the Region Proposal Network (RPN) is used by the Faster R-CNN to create region proposals. Every proposal goes through ROI pooling to . A fixed-length feature vector is extracted from each proposal using ROI pooling, and it is subsequently categorised using Fast R-CNN. Results provide class scores and bounding boxes for items that were identified. By being trained and tailored particularly for the detection job, the RPN delivers better region recommendations than generic approaches. The RPN and Fast R-CNN share convolutional layers, making it possible to combine them into a single network and streamline training. The RPN creates candidate suggestions by applying a sliding window of size nxn to the feature map, which are then filtered according to their objectness score. The Faster R-CNN model for person detection in frames is trained for the proposed task using a customised dataset.

B. *Re-Identification*

Use The FaceNet technique, which generates precise face mappings using deep learning architectures like ZF-Net or Inception Network, is used by the system's re-identification module. 1x1 convolutions are introduced to reduce the number of parameters. These models normalise the output embeddings at the L2 level. In order to optimise the distance between embeddings of various identities while decreasing the squared distance between embeddings of the same identity, the architecture is then trained using the triplet loss function. The convolutional neural network (CNN) extracts information from the input face shot after pooling and fully connected layers. The CNN's output is a feature vector that represents the input face. The triplet loss function instructs the network to map input faces to a high-dimensional embedding space where faces of the same identity are grouped together and faces of different identities are dispersed. The embedding layer transfers the CNN output to the embedding space using this method. These embeddings are utilised for facial recognition tasks including identification and verification.

C. *Abstractive Summarization*

The natural language processing method called abstractive summarising goes beyond basic copying of important phrases to provide a succinct and useful summary of a given text using an lstm-based encoder-decoder model is one method for abstractive summarization the neural network design known as lstm or long short-term memory consists of four neural networks and memory units known as cells information is stored in the cells and memory operations are carried out via gates the forget gate input gate and output gate are three gates that are shown in figure 36 information that is no longer necessary in the cell state is deleted via the forget gate it requires two inputs ht-1 output from the preceding cell and xt input at certain time which are multiplied by weight matrices and combined with biases an activation function that creates a binary output is applied to the outcome information is lost if the output is 0 but is saved for use in the future if it is 1 the input gate enriches the cell state with important data in a manner similar to the forget gate it controls the information using the sigmoid function and filters the values to be remembered the tanh function produces a vector with all feasible values from ht-1 and xt ranging from -1 to 1 then to provide usable information the vector and the controlled values are multiplied the output gate takes information that is pertinent from the current cell state and outputs it the tanh function is first used to transform the cell into a vector following that the data is controlled by the sigmoid function and filtered using the inputs ht-1 and xt the vector is then multiplied with the controlled values before being provided as output to the next cell

D. *System Architecture*

- First the input video will be passed to the identification module where each frame may contain several regions of interest (RoIs) in which each RoI indicates a bounding box that contain a specific character. Definitely, $\{RoI\} = \emptyset$ indicates that no character appear in the frame. Next we have the re-identification module where a user query will be inputted. The user query is the query where the character to be summarized will be inputted. In the re-identification module, the FaceNet module is used to generate feature difference maps for similarity estimation between the inputted image and images in the gallery. Multi-source textual information is leveraged to enhance the distinction in the re-ID module. To perform summarization in the re-identification module, the first step is document vectorization, where each document (subtitle or bullet-screen comment) is vectorized for semantic embedding. Subtitles can be easily vectorized given their strong logic and normality. Bullet-screen comments, on the other hand, are generated by massive users and may contain short text with informal expressions or even slang. To vectorize them, a character-level LSTM is used. After distinction in the re-ID module, the potential keyframes containing the target character C are obtained. If following the strict definition of character-oriented summarization, all the keyframes are simply spliced together to form the summarization video. Additionally, the subtitles of the keyframes are processed using an LSTM to generate the final written summary. Subtitles could be direct descriptions of character status and behaviors in the first-person perspective, with relatively formal expression. On the contrary, bullet-screen comments are always subjective comments in the third-person perspective, with informal expressions or even slang generated by massive users. Thus, it is necessary to select the appropriate source of textual information based on the visual feature of a certain frame, so that the selected textual information could better reflect the identity of the character. After the distinction in re-ID module, the potential key-frames which contain the target character C is now obtained. If following the strict definition of character-oriented summarization, all the keyframes is simply spliced up as the summarization video. However, considering the visual effect that viewers may prefer to fluent videos with a continuous story, some frames are tolerated without target characters that connect two extracted clips.

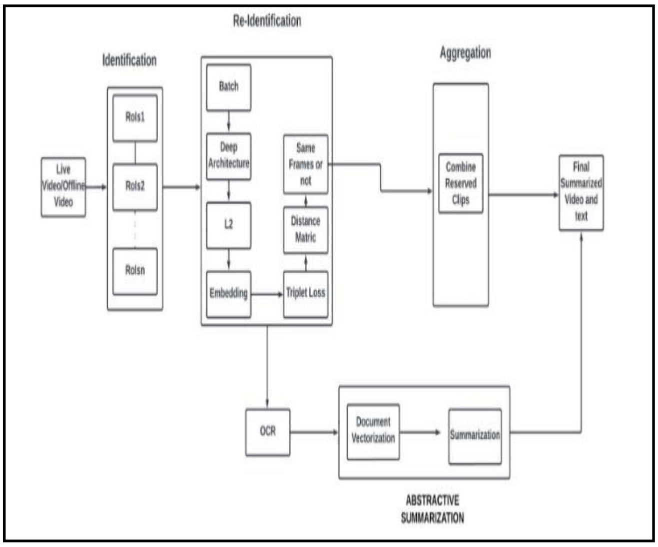


Fig 1. System Architecture

IV. DATASET

The COCO (Common Objects in Context) dataset and the SAMSum dataset are two prominent datasets that are utilized for various reasons. In computer vision research, the COCO dataset is frequently used for tasks including object identification, segmentation, and captioning. It has more than 330,000 annotated photos that are divided into 80 item types and each image has five illustrative captions. Modern algorithms for object identification and segmentation have been trained and tested using this dataset.

The photos themselves and the related annotations are the two primary parts of the COCO dataset. A hierarchical directory structure is used to organize the photos, with distinct folders for the train, validation, and test sets. Each file representing an annotation is in JSON format and refers to a distinct picture. The name of the picture file, size (width and height), object class (such as "person" or "car"), bounding box coordinates (x, y, width, height), segmentation mask (in polygon or RLE format), and keypoints with their locations, if available, are all included in each annotation. Each image in the collection also has five subtitles that provide background information about the scene. The SAMSum dataset, on the other hand, is made up of around 16,000 textual exchanges that appear to be messenger-like discussions. English-speaking linguists were tasked with creating speech that would accurately match the subjects and lengths of their own real-life messenger exchanges. The collection includes slang terms, emoticons, typos, and a variety of styles and registers, including casual, semi-formal, and formal English. After the dialogues were created, they were annotated with succinct summaries that were written in the third person with the intention of giving quick summaries of the topic.

In conclusion, the SAMSum dataset provides a broad collection of conversational interactions accompanied by third-person summaries, enabling the COCO dataset's role as a complete resource for image identification tasks. studies on conversation synthesis.

V. RESULTS AND DISCUSSION

Our approach yielded promising outcomes in both text summary generation and summary video creation. The text summaries precisely captured vital information about each character, encompassing their roles, actions, and relationships. Concurrently, the abstract summary videos effectively showcased pivotal moments involving the identified characters, providing a concise overview of the video content. The amalgamation of textual clues and re-identification proved to be a formidable strategy for character-based identification and summary generation. By integrating natural language processing techniques, we successfully extracted pertinent information from the textual clues, thereby enhancing the quality of the generated text summaries. Furthermore, our re-identification module proficiently tracked characters across frames, guaranteeing the inclusion of their crucial moments in the summary video. Nonetheless, challenges persist in our methodology. The accuracy of the re-identification module heavily depends on the video footage's quality and the complexity of the scenes. Factors such as noisy or low-resolution videos, as well as crowded scenes with occlusions, can potentially impact the module's performance. Consequently, future research should prioritize refining the re-identification algorithms and exploring advanced video processing techniques to address these challenges.

In conclusion, our approach combining textual clues and re-identification for character-based identification and summary generation exhibited promising results. The text summaries and abstract summary videos derived from this methodology provide valuable insights into the video content, enabling applications in various domains, such as video analysis, content summarization, and film studies.

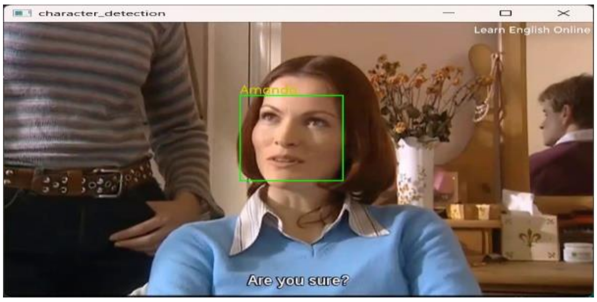
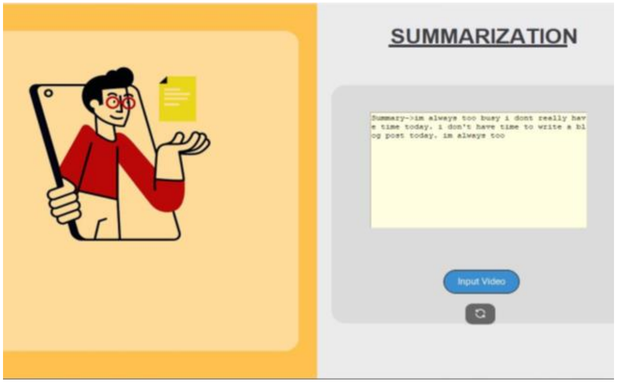


Fig 2 Target Character



```
Summary->im always too busy i dont really hav
e time today. i don't have time to write a bl
og post today. im always too
```

Fig 3 Text Summarization

As creating and watching videos is more active than ever, video summarization now has the potential to have a larger impact on real-world services and people's daily lives. A printed summary of any particular character or instance is tiring to read so a visual summary is essential as it is easy to understand. Video summarization will be useful for creating a compact summary of their favorite character in a movie or sitcom industry or even for analyzing a particular character for literature.

In this paper, a novel framework to jointly model the visual and textual cues for character-oriented summarization is proposed. Video summarization will help people to understand a story or a movie up to the particular instance it has occurred, helping out people to get the continuity of the story if they have missed any portions. In addition, textual information may help to describe the target characters more completely, and even provide some more clues which could hardly be revealed by visual features. This phenomenon also inspires us to fully utilize the potential of time-sync textual information with the help of audio-to-speech converter to identify textual cues even if the video doesn't have any textual information.

Experiments on real-world data sets validated that the solution outperformed several state-of-the-art baselines on both search and summarization tasks, which proved the effectiveness of the framework on the character-oriented video summarization problem. Problems with manual summarization such as excess time consumption and missing rate were solved. A novel framework to jointly model the visual and textual cues for character-oriented summarization was proposed.

REFERENCES

[1]. O. Issa and T. Shanableh, "CNN and HEVC Video Coding Features for Static Video Summarization," in IEEE Access, vol. 10, pp. 72080-72091, 2022, doi: 10.1109/ACCESS.2022.3188638.

[2]. S. Lee, J. Sung, Y. Yu, and G. Kim, "A memory network approach for story-based temporal summarization of 360 videos," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 1410–1419.

[3]. S. Wehrwein, K. Bala and N. Snavely, "Scene Summarization via Motion Normalization," in IEEE Transactions on Visualization and Computer Graphics, vol. 27, no. 4, pp. 2495-2501, 1 April 2021, doi: 10.1109/TVCG.2020.2993195.

[4]. S. Mingyang Ma; Shaohui Mei; Shuai Wan; Zhiyong Wang; David Dagan Feng; Mohammed Bennamoun, "Similarity-Based Block Sparse Subset Selection for Video Summarization," IEEE Transactions on Circuits and Systems for Video Technology, Volume: 31, Issue: 10, October 2021, pp. 3967 - 3980, DOI: 10.1109/TCSVT.2020.3044600

[5]. B. Zhao, X. Li, and X. Lu, "HSA-RNN: Hierarchical structure-adaptive RNN for video summarization," in *Proc. IEEE/CVF Conf. Comput. Vision Pattern Recognit.*, 2018, pp. 7405–7414.

[6]. Y. Shen, T. Xiao, H. Li, S. Yi, and X. Wang, "End-to-end deep Kroneckerproduct matching for person re-identification," IEEE/CVF Conf. Comput. Vision Pattern Recognit., pp. 6886–6895, 2018.

[7]. S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137–1149, Jun. 2017

[8]. P. Varini, G. Serra and R. Cucchiara, "Personalized Egocentric Video Summarization of Cultural Tour on User Preferences Input," in IEEE Transactions on Multimedia, vol. 19, no. 12, pp. 2832-2845, Dec. 2017, doi: 10.1109/TMM.2017.2705915.

[9]. A. Sharghi, J. S. Laurel, and B. Gong, "Query-focused video summarization: Dataset, evaluation, and a memory network based approach," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2017, pp. 2127–2136.

[10]. M. Gygli, H. Grabner, and L. V. Gool, "Video summarization by learning submodular mixtures of objectives," in Proc. IEEE Conf. Comput. Vision Pattern Recognit., 2015, pp. 3090–3098