UNIVERSITY OF SCIENCE
ADVANCED PROGRAM IN COMPUTER SCIENCE

Thesis Proposal

(submitted by students)

Thesis title:

ENHANCING VIDEO SUMMARIZATION WITH CONTEXT AWARENESS

Thesis advisor: Assoc. Prof. Trần Minh Triết and Dr. Lê Trung Nghĩa

Students: **Huỳnh Lâm Hải Đăng** (19125003) - **Hồ Thị Ngọc Phượng** (19125014)

Type of thesis: Research with demo application

Duration: 01-01-2023 to 14-08-2023

Content of thesis:

1. Introduction

Video summarization is an emerging research field that addresses the need for efficient video browsing and retrieval in today's vast and ever-expanding video collections. With the exponential growth of multimedia data, the ability to effectively analyze and extract relevant information from video content has become crucial. Video summarization techniques aim to automatically generate a concise and meaningful representation of a video by selecting keyframes, shots, or segments that capture the essence of the content. This process can significantly reduce the time and effort required to review and analyze video data, thereby improving the efficiency and accuracy of various applications, including video surveillance, education, entertainment, and social media.

Video summarization is a challenging task that requires selecting and organizing the most informative and representative segments from a video. Various approaches have been proposed for this task, such as keyframe extraction, video skimming, and keyshot selection. Keyframe extraction selects a subset of frames that represent the most salient and diverse scenes or events [**?**]. Video skimming creates a shorter video that preserves the continuity and coherence of the original video [**?**]. Keyshot selection divides the video into shots based on temporal boundaries or transitions and selects the most representative shots from each segment [**?**]. These methods are based on different criteria, such as visual saliency, motion analysis, and semantic relevance. However, there is still room for improvement and innovation in this field, as different methods may have different advantages and disadvantages depending on the video content, the user's needs and preferences, and the application scenarios.

Recent advances in computer vision and machine learning have enabled the development of more sophisticated video summarization techniques that can handle complex video content and capture semantic relations between different segments. For example, deep learning-based methods have been used to learn feature representations of video frames and generate summaries that are more informative and coherent. Deep learning is a branch of machine learning that uses neural networks to learn from data and perform complex tasks. Deep learning has been widely used for video summarization in recent years, as it can extract high-level features from videos and generate summaries in an end-to-end manner [**?**, **?**].

Despite the recent advanced developments, video summarization remains a difficult task, as it involves comprehending the semantic and temporal aspects of the video content, along with the user's focus and preference. A potential solution to this challenge is to leverage saliency detection information to enhance the video summarization quality. Saliency detection is a task that aims to predict where people look in videos, i.e., the salient regions that attract human attention [?]. Saliency detection can provide useful cues for video summarization, as it can highlight the important and informative parts of the video content, and filter out the irrelevant and redundant parts [?].

The primary objective of this thesis proposal is to advance the field of video summarization by introducing novel techniques that facilitate the creation of informative and concise video summaries that encapsulate the most pertinent and substantial content. The proposed research methodology involves incorporating saliency detection outcomes as a supplementary cue for the summarization model. The efficacy and efficiency of the proposed approach will be evaluated on multiple benchmark datasets and benchmarked against existing state-of-the-art video summarization techniques to demonstrate the effectiveness of the saliency detection module.

2. Related Works

In recent years, deep learning has been widely applied to both video summarization and saliency detection, achieving state-of-the-art performance on various benchmarks. In this section, we review some of the existing works that use deep learning for video summarization, saliency detection, and the combination of both.

For video summarization, deep learning methods can be categorized into two types: supervised and unsupervised. Supervised methods learn to generate summaries that are similar to human-annotated summaries using various loss functions, such as reconstruction loss [?], ranking loss [?], diversity loss [?], or reinforcement learning [?]. Unsupervised methods learn to generate summaries that are representative and diverse without using human annotations, by exploiting self-supervised signals, such as reconstruction error [?], determinantal point processes [?], or generative adversarial networks [?].

For saliency detection, deep learning methods can also be divided into two types: static and dynamic. Static methods focus on predicting salient regions or objects in individual frames of a video, using convolutional neural networks (CNNs) that can capture local and global features [?]. Dynamic methods aim to model the temporal coherence and motion cues of salient regions or objects across frames of a video, using recurrent neural networks (RNNs) [?], optical flow [?], or non-local neural networks [?].

Some works have explored the use of saliency detection to enhance video summarization, by incorporating saliency maps as additional inputs or constraints for the summarization models. For example, [?] proposed a joint optimization framework that integrates visual saliency and semantic information for video summarization. [?] proposed a saliency-aware interestingness model that learns to rank video segments based on their visual saliency and semantic relevance. [?] proposed a deep summarization network that fuses static and dynamic saliency maps with CNN features for video summarization.

In summary, deep learning has shown great potential for video summarization and saliency detection, as well as their combination. However, there are still some challenges and open problems that need to be addressed, such as how to handle complex and diverse video content, how to evaluate the quality and usefulness of summaries and saliency maps, how to leverage large-scale unlabeled video data, and how to incorporate user preferences and feedback.

3. Motivation

Video summarization is an essential task in video analysis, as it can significantly reduce the time and effort required to browse through a large video dataset. With the proliferation of video data in various domains, such as surveillance, social media, and entertainment, the need for efficient and effective video summarization techniques has become increasingly important. Video summarization can help users quickly navigate through large amounts of video content, and identify the most relevant and important information without having to watch the entire video. Moreover, video summarization can facilitate tasks such as video retrieval, browsing, and archiving, by providing a concise representation of the video content.

Despite the importance of video summarization, it remains a challenging task due to the complexity of video content and the subjective nature of summarization. Video content can vary in terms of its visual complexity, temporal dynamics, and semantic structure, which makes it difficult to identify the most salient parts of the video. Moreover, video summarization can be subjective, as different users may have different preferences and goals for the summary. Therefore, developing effective and efficient video summarization techniques requires addressing these challenges and exploring new approaches that can accurately capture the most relevant and important information from the video while considering the user's attention and interest.

Visual noise, such as blurry backgrounds and irrelevant objects, can pose a significant challenge for accurately extracting important information and summarizing videos. One potential solution is to integrate salient object maps as a cue to guide the model's attention toward relevant content while filtering out redundant information. Saliency detection identifies visually prominent regions in an image or video based on contrast, texture, and other visual features. By integrating saliency detection into the visual analysis process, the model can prioritize salient regions and allocate resources to process them while suppressing background noise. This approach has shown promising results in various computer vision tasks, including object detection, image segmentation, and visual tracking. In this study, we aim to explore the potential of saliency detection for video summarization and evaluate its effectiveness in enhancing visual processing accuracy and efficiency.

4. Objectives

We aim to study and propose a novel approach for improving video summarization quality by integrating saliency detection results. In this thesis, our main objectives are as follows:

- Literature Review and Proposal Writing
    - Conduct a comprehensive literature review on video summarization and saliency detection, identifying the current state-of-the-art techniques and their limitations, as well as opportunities for improvement.
    - Analyze the importance of saliency detection in video summarization and compare existing methods and tools for saliency detection and extraction in videos, in terms of performance and applicability for video summarization.

  - Develop a research proposal, including research questions, hypotheses, and methodology, based on the findings from the literature review.

- Dataset Collection
  - Collect datasets suitable for training and testing the saliency detection and video summarization modules.
  - Define relevant performance metrics for evaluating the effectiveness of the saliency detection module in improving the quality of video summarization.

- Saliency Detection Model Development
  - Develop a saliency detection model capable of accurately identifying the most visually prominent regions of a video.
  - Train and optimize the model using the collected datasets.

- Video Summarization Model Development
  - Develop a video summarization model capable of effectively extracting the importance level of each frame in a video.
  - Train and optimize the model using the collected datasets.

- Integration of Saliency Detection into Video Summarization
  - Integrate the saliency detection module into the video summarization framework.
  - Evaluate the performance of the new video summarization model with and without the saliency detection module using the defined performance metrics.

- Comparison with Existing Video Summarization Techniques
  - Compare the performance of the new video summarization model with existing state-of-the-art video summarization techniques.

    - Analyze the strengths and weaknesses of the proposed approach.

- Demo Application Development
    - Develop a demo application that can demonstrate the functionality and usability of the proposed framework for video summarization.

- Thesis Writing and Submission
    - Write up the thesis, including an introduction, literature review, methodology, results, discussion, and conclusion.
    - Submit the thesis for review and evaluation by the thesis committee.

Research timeline:

5. References

- Literature Review and Proposal Writing: 01-01-2023 to 31-01-2023
- Dataset Collection: 01-02-2023 to 15-02-2023
- Saliency Detection Model Development: 16-02-2023 to 15-03-2023
- Video Summarization Model Development: 16-03-2023 to 15-04-2023
- Integration of Saliency Detection into Video Summarization: 16-04-2023 to 15-05-2023
- Comparison with Existing Video Summarization Techniques: 16-05-2023 to 31-05-2023
- Demo Application Development: 01-06-2023 to 30-06-2023
- Thesis Writing and Submission: 01-07-2023 to 31-07-2023

Approved by the advisors ***Signatures of advisors***	**Ho Chi Minh city, 30-03-2023** ***Signatures of students***
Assoc. Prof. Trần Minh Triết	Huỳnh Lâm Hải Đăng
Dr. Lê Trung Nghĩa	Hồ Thị Ngọc Phượng