UNIVERSITY OF SCIENCE
ADVANCED PROGRAM IN COMPUTER SCIENCE

HUỲNH LÂM HẢI ĐĂNG - HỒ THỊ NGỌC PHƯỢNG

ENHANCING VIDEO SUMMARIZATION WITH CONTEXT AWARENESS

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

HO CHI MINH CITY, 2023

UNIVERSITY OF SCIENCE
ADVANCED PROGRAM IN COMPUTER SCIENCE

HUỲNH LÂM HẢI ĐĂNG 19125003
HỒ THỊ NGỌC PHƯỢNG 19125014

ENHANCING VIDEO SUMMARIZATION WITH CONTEXT AWARENESS

BACHELOR OF SCIENCE IN COMPUTER SCIENCE

THESIS ADVISOR
ASSOC. PROF. TRẦN MINH TRIẾT
DR. LÊ TRUNG NGHĨA

HO CHI MINH CITY, 2023

COMMENT OF THESIS'S ADVISOR

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

COMMENTS OF THESIS'S REVIEWER

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

. . . . . .

ACKNOWLEDGEMENTS

Authors

Huỳnh Lâm Hải Đăng & Hồ Thị Ngọc Phượng

THESIS PROPOSAL

Thesis title: ENHANCING VIDEO SUMMARIZATION WITH CONTEXT AWARENESS

Advisors: Assoc.Prof. Trần Minh Triết, Dr. Lê Trung Nghĩa

Duration: January 1st, 2023 to August 14th, 2023

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

Theme of Thesis: theoretical research, proposed improvements.

Content:

We aim to propose a novel approach for improving video summarization quality by integrating context awareness. We also aim to propose an evaluation metric that better suits the practical use of problem in real life.

The details include:

- Literature Review and Proposal Writing
  - Conduct a comprehensive literature review on video summarization, identifying the current state-of-the-art techniques and their limitations, as well as opportunities for improvement.
  - Analyze the importance of context in video summarization and compare existing methods and tools for context extraction in videos, in terms of performance and applicability for video summarization.
  - Develop a research proposal, including research questions, hypothesis, and methodology, based on the findings from the literature review.

- Dataset Collection
  - Collect datasets suitable for training and testing.
  - Analyze the current evaluation metrics for video summarization and identify their flaws.
  - Define relevant performance metrics for evaluating the effectiveness of the context awareness in improving the quality of video summarization.
- Model Development
  - Develop baseline model for the sake of benchmarking.
  - Develop different models to prove the proposed hypothesis.
  - Train and optimize the model using the collected datasets.
- Comparison with Existing Video Summarization Techniques
  - Conduct experiments on proposed enhancements with a thoroughly designed ablation study.
  - Analyze the strengths and weaknesses of the proposed approach.
  - Conduct surveys based on the proposed evaluation metric.
- 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.

Implementation plan:

- 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

	December 26th 2022
Advisors	**Authors**
Assoc. Prof. Trần Minh Triết	Huỳnh Lâm Hải Đăng
Dr. Lê Trung Nghĩa	Hồ Thị Ngọc Phượng

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

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 key frames, 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.

Despite the wide-ranging usage of video summarization, there are only a few datasets available for this task, with the two most prominent are SumMe [**?**] and TVSum [**?**]. This limitation hinders the comprehensive evaluation and benchmarking of video summarization algorithms. The scarcity of diverse and representative datasets restricts the generalizability and effectiveness of developed techniques. Additionally, the evaluation metrics employed for video summarization are also flawed, as they fail to fully capture the inherent challenges and complexities involved in generating high-quality video summaries. This inadequacy hampers the accurate assessment of different algorithms and inhibits the advancement of the field.

However, the inherent nature of the video summarization task poses challenges in evaluating the quality of generated summaries without human involvement. It is difficult to determine objectively whether one video summary is superior to another without relying on subjective human judgment. Recognizing this limitation, we propose a self-supervised model that mitigates the issues associated with the data-intensive nature of video summarization. By moving away from fixed ground truth annotations and instead leveraging the inherent structure

and information within the video data itself, our self-supervised model learns to generate informative and representative summaries.

In addition to addressing the data scarcity challenge, we also introduce an innovative evaluation pipeline specifically tailored for the video summarization task. To ensure that our generated summaries effectively capture the essence of the original videos, we conduct a comprehensive survey involving human participants. The survey participants are provided with the original videos, ground truth summaries, and our generated summaries. They are then asked to evaluate and compare the informativeness of the generated summaries against the ground truth summaries. This human-centric evaluation approach enables us to obtain valuable insights into the performance and effectiveness of our proposed video summarization techniques.

By proposing a self-supervised model and an evaluation pipeline that incorporates human judgment, this thesis not only addresses the data scarcity and evaluation challenges but also provides a more realistic and meaningful assessment of the video summarization task. The experimental results and feedback obtained from the survey validate the efficacy and relevance of our proposed approaches, highlighting their potential for improving the accuracy and reliability of video summarization in practical applications.

CHAPTER 1
INTRODUCTION

In this chapter, we provide general information about our work in four sections before getting into details in the following chapters. Section **??** *introduces the practicality and applicability of Video Summarization. We then discuss our motivation for applying self-supervision and introducing a new evaluation metrics in Section* **??**. *Section* **??** *presents our objectives in developing the model as well as the evaluation pipeline. Finally, we describe the outline content of our thesis in Section* **??**.

1.1 Overview

In recent years, the consumption of video content has experienced a remarkable upsurge, driven by the proliferation of multimedia platforms such as TikTok, YouTube, Instagram, and others. A striking example of this growth can be observed in the case of YouTube, where the number of video content hours uploaded per minute has witnessed a substantial increase. Between 2014 and 2020, there was an approximate 40 percent rise in the rate of uploads, with over 500 hours of video being uploaded every minute as of June 2022 [**?**]. This surge in video content on platforms like YouTube reflects the expanding demand among consumers for online video consumption. With an approximation of 2.5 quintillion bytes of data created every day [**?**], there is a pressing need for effective methods that can automatically generate concise and informative summaries of videos, enabling users to quickly comprehend the content without having to watch the entire video.

Video summarization, as a research area, focuses on generate concise summaries that effectively capture the temporal and semantic aspects of a video, while

preserving its salient content. Achieving this objective involves addressing several fundamental challenges, such as identifying key frames or representative shots, detecting important events, recognizing significant objects or actions, and preserving the overall context and coherence of the video.

The task plays a crucial role in facilitating efficient browsing, indexing, and retrieval of video data, offering users the ability to preview and comprehend video content without investing significant time and effort. Moreover, video summarization finds applications in various domains, including video surveillance, multimedia retrieval, video archiving, and online video platforms, where it serves as a valuable tool for enhancing user experience and content accessibility [**?**].

1.2 Motivation

Despite the wide-ranging usage of video summarization, there are only a few datasets available for this task, with the two most prominent being SumMe [**?**] and TVSum [**?**]. This limitation hinders the comprehensive evaluation and benchmarking of video summarization algorithms. The scarcity of diverse and representative datasets restricts the generalizability and effectiveness of developed techniques.

The nature of video summarization task poses a challenge for supervised approaches. Traditional metrics, such as F-measure and precision-recall curves, rely heavily on frame-level matching and do not adequately account for the temporal coherence and semantic understanding of the summary. These kind of metrics fail to fully capture the inherent challenges and complexities involved in generating high-quality video summaries.

Recognize the difficulty of evaluating video summaries solely based on fix ground truths, we propose an innovative evaluation pipeline tailored specifically for the video summarization task. In order to ensure that our generated summaries

effectively capture the essence of the original videos, we conduct a comprehensive survey involving human participants. The survey participants are provided with the original videos, ground truth summaries, and our generated summaries. They are then asked to evaluate and compare the informativeness of the generated summaries against the ground truth summaries. This human-centric evaluation approach allows for a more realistic and meaningful assessment of our proposed video summarization techniques.

In addition to the novel human-based evaluation metric, this thesis introduces a self-supervised model that overcomes the challenges associated with the data-intensive nature of video summarization. Instead of relying on supervision with ground truth annotations, our model leverages the inherent structure and information within the video data itself to generate informative and representative summaries. By moving away from the limitations of traditional annotation-based approaches, our self-supervised model aims to enhance the quality and generalizability of video summarization techniques.

1.3 Objectives

In this thesis, we aim to propose a self-supervised model for video summarization task as well as a human-centric evaluation pipeline with the following main contributions:

-

1.4 Thesis Content

After **Chapter ??: Introduction**, the remainder of our thesis is composed of 5 chapters as follows:

Chapter ??: Background

In this chapter, we present fundamental knowledge, from Machine Learning and Deep Learning, Neural Networks, Convolutional Neural Networks, to Transformers, which will help us to comprehend the next chapters.

Chapter ??: Related Work

In this chapter, we first provide an overview of three main deep learning approaches for solving video summarization task: supervised method, weakly supervised method, and unsupervised method. At each approach, we discuss the leading paper and explain how the follow-up papers could improve the baseline in many aspects. Finally, we analyze the advantages and disadvantages of each method.

Chapter ??: Proposed Methods

In this chapter, ...

Chapter ??: Experiments

In this chapter, ...

Chapter ??: Conclusions

In this chapter, ...

CHAPTER 2

BACKGROUND

This is content.

CHAPTER 3

RELATED WORK

This chapter describes the related work.

3.1 Preliminary

This is content.

3.2 Supervised approaches

This is motivation.

3.3 Unsupervised approaches

This is overview.

3.4 Weakly Supervised approaches

This is section:intro-objectives.

CHAPTER 4
PROPOSED METHODS

This chapter describes the proposed method.

4.1 Self-Supervised Pipeline for Summarization Learning

This is content.

4.2 Clustering-based Video Partitioning and Summarization

This is motivation.

4.3 Summarization Evaluation with Human Feedback

This is section:intro-objectives.

CHAPTER 5
EXPERIMENTS

This chapter describes the experiments conducted to evaluate the proposed method.

5.1 Datasets

This is content.

5.2 Evaluation methods

The evaluation process of our experiments is two-fold with the first part being the automatic evaluation of raw performance between the proposed method and baselines presented in previous works while the second part is the evaluation of the proposed method against human's perspective of the summarization's quality. For the first part of the evaluation procedure, we employ the metric of F-measure which is widely adopted among prior evaluations to measure and compare the performance of our proposal with selected baselines. For the second part, we specifically construct a dedicated survey to collect human's feedback on the quality of the generated summaries in the form of . The following subsections will describe the details of the evaluation process in which the subsection **??** provides information about the automatic evaluation process and the **??** contains details for human-centric evaluation.

5.2.1 Automatic evaluation

The automatic evaluation process compares the performance of the method that we proposed in **??** with the baselines selected among previous works that were mentioned in **??**. The datasets used for this comparison are already described in previous section **??**, hence this section is dedicated to the description of the

evaluation metric and the baselines.

5.2.1.1 F-measure

Prior evaluations have adopted F-measure as the main comparison metric for evaluating performance of several video summarization approaches. We would like to revisit the definition of F-measure in this section to provide a better understanding of the metric as well as its nature.

F-measure nature Given a data consists of several samples and a binary label for each of them, the F-measure metric evaluates a set of predictions on the labels of those samples. To compute the value of the F-measure for a prediction, each of the sample in the dataset is assigned to one of the four categories: true positive (TP), false positive (FP), true negative (TN), and false negative (FN). The assignment is done by evaluating the value of binary label and the prediction with the following rules:

- If the label and the prediction are both positive, the sample is assigned to TP.
- If the label is positive but the prediction is negative, the sample is assigned to FN.
- If the label is negative but the prediction is positive, the sample is assigned to FP.
- If both the label and the prediction are negative, the sample is assigned to TN.

Afterward, the precision value P is computed as the ratio of the number of true positives to the sum of true positives and false positives, meaning that

$P = \frac{|TP|}{|TP|+|FP|}$. The recall value R is computed as the ratio of the number of true positives to the sum of true positives and false negatives, or $R = \frac{|TP|}{|TP|+|FN|}$. The F-measure is then defined as the harmonic mean of precision and recall with the formula as follows:

$$F = 2 \times \frac{P \times R}{P + R} \tag{5.1}$$

The combination of precision and recall in the evaluation of F-measure is to ensure that the metric is able to capture the performance of the prediction in both positive and negative cases.

F-measure in video summarization In the case of video summarization, a value of F-measure is calculated for each of the video in the dataset. The prediction is the summary generated by the method being evaluated while the label is the ground truth summary of the video.

Depending on the type of annotations used by the dataset, the groundtruth summaries created by users may be of several different forms that are mentioned below:

- Sets of keyframes: Video's frames are used as main pieces of information to be selected . The ground truth summary of a user is a set of keyframes that are selected by users as the most important frames in the video, meaning that the grountruth U_i of the i-th user is a set of keyframes $U_i = \{U_i^{(1)}, U_i^{(2)}, ..., U_i^{(k)}\}$ where $U_i^{(j)}$ is the index of the j-th keyframe in the video sequence selected by the i-th user, and k is the number of keyframes selected by the user.

- Sets of key-fragments: The video sequence is partitioned into multiple non-overlapping fragments containing meaningful information in each of them

with the method of partitioning depends on the nature of the dataset and is usually proposed along with dataset's publications by its authors. The ground truth summary of a user is a set of key-fragments that are selected by users as the most important fragments in the video, meaning that the grountruth U_i of the i-th user is a set of key-fragments $U_i = \{U_i^{(1)}, U_i^{(2)}, ..., U_i^{(k)}\}$ where $U_i^{(j)}$ is the index of the j-th key-fragment in the video sequence selected by the i-th user, and k is the number of key-fragments selected by the user.

- Fragment-level scores: Similar to the form of key-fragments, the video sequence is partitioned into multiple non-overlapping fragments containing meaningful information in each of them. The ground truth summary of a user is a set of scores that are assigned to each of the fragments in the video, meaning that the grountruth U_i of the i-th user is a set of scores $U_i = \{U_i^{(1)}, U_i^{(2)}, ..., U_i^{(k)}\}$ where $U_i^{(j)}$ is the score assigned to the j-th fragment in the video sequence by the i-th user, and k is the number of fragments in the video.

To evaluate a method using F-measure on a dataset with one of the above forms of ground truth summary, the method's generated summaries are usually converted to the same form as the ground truth summary. For example, if the ground truth summary is a set of keyframes, the generated one is converted to a set of keyframes as well. The conversion process is usually done by selecting the most important frames or fragments in the video according to the scores assigned to them by the method. Therefore, the pre-evaluation summary of a method for a video is usually of the form $G = \{G^{(1)}, G^{(2)}, ..., G^{(k)}\}$ where $G^{(j)}$ is the index of the j-th frame or fragment in the video sequence selected by the method as key information, and k is the number of frames or fragments selected by the method.

With this set theoretic formulation of generated summary and user's summary, one can define the four categories in preparation of F-measure calculation as follows:

- True positives: Frames or fragments that are selected by both the method and the i-th user, or $TP_i = G \cap U_i$.
- False positives: Frames or fragments that are selected by the method but not by the i-th user, or $FP_i = G \setminus U_i$.
- False negatives: Frames or fragments that are selected by the i-th user but not by the method, or $FN_i = U_i \setminus G$.
- True negatives: Frames or fragments that are not selected by both the method and the i-th user, or $TN_i = \{1, 2, ..., n\} \setminus (G \cup U_i)$ where n is the number of frames or fragments in the video.

From the above formulation, the F-measure of the method for the i-th user is calculated using the formula **??** with $P_i = \frac{|TP_i|}{|TP_i|+|FP_i|}$ and $R_i = \frac{|TP_i|}{|TP_i|+|FN_i|}$, leading to $F_i = 2 \times \frac{P_i \times R_i}{P_i + R_i}$. The F-measure of the method for the video is then calculated as the average of F-measure values for all users in the dataset, or $F = \frac{1}{u} \sum_{i=1}^{u} F_i$ where u is the number of users in the dataset.

Baselines

5.2.2 Human-centric evaluation

5.3 Implementation details

This is section:intro-objectives.

5.4 Experimental results

This is overview.

5.5 Discussion

This is overview.

CHAPTER 6

CONCLUSIONS

This chapter concludes the thesis.

6.1 Future Directions

This is content.

6.2 Final Conclusions

This is motivation.

APPENDICES