

**SUMMARIZATION OF LARGE-SCALE VIDEOS TO TEXT
FORMAT USING SUPERVISED BASED SIMPLE RULE-
BASED MACHINE LEARNING MODELS**

U.K.H.A. Sugathadasa

199488F

Department of Computational Mathematics,

Faculty of Information Technology

University of Moratuwa

Sri Lanka

March 2022

**SUMMARIZATION OF LARGE-SCALE VIDEOS TO TEXT
FORMAT USING SUPERVISED BASED SIMPLE RULE-
BASED MACHINE LEARNING MODELS**

Udage Kankanage Harindu Ashan Sugathadasa

199488F

Thesis/Dissertation submitted in partial fulfilment of the requirements for the degree
Master of Science in Artificial Intelligence

Department of Computational Mathematics

University of Moratuwa

Sri Lanka

March 2022

Declaration

I declare that this is my own work, and this thesis/dissertation does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to University of Moratuwa the non-exclusive right to reproduce and distribute my thesis/dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Name of the Student:	Signature:	Date:
Sugathadasa U.K.H.A.

The above candidate has carried out research for the Masters/MPhil/PhD thesis/dissertation under my supervision.

Name of the Supervisor:	Signature of the supervisor:	Date:
Dr. Subha Fernando

Name of the Co-Supervisor:	Signature of the co-supervisor:	Date:
Dr. Varuna De Silva

Abstract

Video Summarization has been one of the most interested research and development field since the late 2000s, thanks to the evolution of social media and the internet, due to the influence to provide a concise and meaningful summary of large-scale video. Even though the video summarization has been elongated through several non-ML and traditional based techniques and ML-based techniques, generation of correct and required summaries from the video is yet a limitation. To overcome this concern, different techniques have been attempted including vision-based approaches and NLP related approaches. With the inspiration of NLP related Transformer networks, researchers are looking to integrate such sequence-based learning algorithm into the video dimension as to apply spatiotemporal extractions. Despite the VS implementations, another extension of VS has been exponentially emphasized, namely TVS which generates the summaries of the video via a text format.

Simply the evolution of VS towards TVS is not a straightforward journey since a lot of blockers have been eliminated using UL, RL, and SL based frameworks. When it comes to the STOA methods in TVS, Transformer based methods are eventually highlighted along the T5 based NLP frameworks. Since this area is still at the ground level, a lot of unknow facts and issues can be explored. Especially the attention-based sequence modelling of the learning algorithm should be carefully imitated to achieve the best accuracy improvements. All the improvements are subjected to apply into a real-time application ulteriorly. To tackle such improvements, a novel standalone method should be introduced with the simplest network layout which can be applicable to the embedded devices. This is where the **Simple Rule-based Machine Learning Network to Text-based Video Summarization (SiRuML-TVS)** has been unveiled.

Though the network contains a single input of large-scale video and a single output of meaningful description for the given video, the high-level network layout compromises three ML modules for Video Recognition, Object Detection, and finally Text Generation. Each module is subjected to different evaluation criterions however, the end-to-end full network is evaluated on a single metric. Different combination of each module can be affected to the performance of the entire pipeline however, the

combination of Transformers and CNNs provide the better tradeoff between accuracy and the computational inferencing. This makes a hope to deploy the proposed method in an edged device thus, the gap between theoretical explanation to practical implementation will be filled.

Dedication

I dedicate this thesis to my parents, my sister, my wife, the university lecturers from undergraduate level who are always withstand in my successes and failures.

Acknowledgements

I would like to pay my greatest appreciation to Dr. Subha Fernando, who motivated my enthusiasm on research and provided guidance and advises through the research journey till the end. Her experience on research was tremendously big support to this work, and I be indebted a great deal of its success for herself.

Every person has its own shadow when at be any place. My second appreciation is to Dr. Varuna De Silva, who is from University of Loughborough, for supporting both me and supervisor when we had a doubtful circumstances and unsolvable stakes.

Next, I would like to elapse my thanks and appreciation to the staff members of the Department of Computation Mathematics, for their supportiveness and consideration for my research work.

Lastly, I would also like to express appreciation through my bottom of heart for all my Lecturers from the Undergraduate program, my fellow colleagues from the University, and all my family members, who supported me through the journey of research program in both verbally and mentally. Even though they might not know about this, the impact from them is substantial, and I might not make it to where I am now at without themselves.

Table of Contents

Declaration	i
Abstract	ii
Dedication	iv
Acknowledgements	v
List of figures	x
List of tables	xi
Abbreviations	xii
Introduction	1
1.1 Prolegomena	1
1.2 Background and Motivation	1
1.3 Aims and Objectives	2
1.3.1 Aim	2
1.3.2 Objectives	2
1.4 Problem Definition	3
1.5 Proposed Solution	3
1.6 Resource Requirements	4
1.7 Outline of the Thesis	4
1.8 Summary	5
Summarization of Large-Scale Videos to Text format – Developments and Issues... 6	
2.1 Introduction	6
2.2 Early Developments (Gestation) in Video Summarization	6
2.3 Breakthrough and Trends in Video Summarization – Latest	9
2.3.1 Unsupervised Learning based Video Summarization	9
2.3.2 Reinforcement Learning based Video Summarization	11
2.3.3 Supervised Learning based Video Summarization	13
2.3.4 Latest Trend in Video Summarization	15
2.4 Summary of Past related research	17
2.5 Taxonomy of the Video Summarization Algorithms	18
2.6 Challenges in Video Summarization in Text Format	18
2.6.1 Selection of Dataset	19
2.6.2 Unsuccessful Multimodal approaches	19
2.6.3 Lack of facilitation to change the model	19
2.6.4 Lack of explanation capability	19

2.6.5	Deployment.....	20
2.7	Problem Definition.....	20
2.8	Summary.....	20
Technologies used for Video Summarization.....		21
3.1	Introduction.....	21
3.2	Convolutional Neural Networks – for Object Detection.....	22
3.3	TimeSformer – A new trend.....	23
3.3.1	Mathematical Description – TimeSformer.....	23
3.3.2	Sub variants of TimeSformer.....	26
3.4	Text-To-Text Transfer Transformer – for Text output.....	27
3.4.1	Input and Output of T5.....	27
3.4.2	T5 Model Architecture.....	28
3.4.3	T5 Model Variants.....	28
3.5	Summary.....	29
Approach.....		30
4.1	Introduction.....	30
4.2	Hypothesis.....	30
4.3	Input.....	30
4.4	Output.....	31
4.5	Process.....	31
4.6	Users.....	31
4.7	Features.....	31
4.8	Summary.....	32
Design.....		33
5.1	Introduction.....	33
5.2	High-level Architecture of the Design.....	33
5.3	Small-clip Generator.....	34
5.4	Action Recognition.....	34
5.5	Object Detection.....	34
5.6	NLP Text Creator.....	35
5.6.1	Word Ordering.....	35
5.6.2	NLP Sentence.....	36
5.6.3	Combiner.....	36
5.7	Summary.....	36
Implementation.....		37

6.1	Introduction	37
6.2	System Requirements	37
6.3	Framework for Action Recognition	37
6.4	Framework for Object Detection.....	38
6.5	Dataset Preparation.....	39
6.5.1	Action Recognition and Object Detection	39
6.5.2	NLP Text Creator	40
6.6	Setup Training Process	41
6.6.1	Training Process for Action Recognition.....	41
6.6.2	Training Process for Object Detection.....	43
6.6.3	Training Process for NLP Text Creator	44
6.7	Rule-based Algorithm	46
6.8	Summary	47
Evaluation	48
7.1	Introduction	48
7.2	Evaluation Strategy	48
7.2.1	Evaluation at Training Phase	48
7.2.2	Overall Evaluation to the System.....	49
7.2.3	Model Evaluation.....	50
7.3	Experiment Setup for SiRuML-TVS.....	51
7.4	TVS Models Comparison.....	52
7.5	Use Case of the SiRuML-TVS System	55
7.6	Summary	56
Conclusion and Further Work	57
8.1	Introduction	57
8.2	Conclusion.....	57
8.2.1	Achievements of Project Objectives	57
8.2.2	Overall Conclusion.....	58
8.3	Limitations and Further Works	59
8.4	Summary	59
References	60
Appendix	64
	Appendix I: Video Recognition module for Inferencing.....	64
	Appendix II: Change of DetectM2 class in Object Detection Module, YoloV5 ...	65
	Appendix III: Training Script for T5 Module	66

Appendix IV: Full System of SiRuML-TVS 68

LIST OF FIGURES

Figure 2.1 - High-level diagram of Reinforcement Learning Framework.....	12
Figure 2.2- Taxonomy of Video Summarization	18
Figure 3.1- Technology Stack of Text based Video Summarization	21
Figure 3.2- CNN Architecture in Image Classification task	22
Figure 3.3- Different model architectures for TimeSformer [20]	26
Figure 3.4- Visualization of Divide Space-Time Attention Module	27
Figure 3.5 - Model Architectures of T5	28
Figure 4.1- Input-Process-Output for Text-based Video Summarization.....	30
Figure 5.1 - Top-level Architecture of SSML-TVS.....	33
Figure 5.2 - Sub modules in the NLP Text Creator	35
Figure 6.1- Folder Structure of METEOR Dataset	39
Figure 6.2 - Original classes provided from METEOR dataset.....	40
Figure 6.3 - Extended classes from the original classes.....	40
Figure 6.4 - Shape of Input Text and Target Text for NLP Text Creator Model.....	41
Figure 6.5 - Sample configuration file of Training in MMAAction2.....	42
Figure 6.6 - Comparison of YoloV5-Small original model (left) with the Modified YoloV5-Small (right).....	44
Figure 6.7- Training loss against steps in WebNLG based training	45
Figure 6.8 - Training results with custom created dataset	46
Figure 7.1- Structure of Videos and Text files of custom validation dataset.....	49
Figure 7.2 - Sample frames of a video and corresponding outcome as a sentence....	49
Figure 7.3 - Experiment Setup for TVS evaluation	52
Figure 7.4 - Quantitative Results of the TVS system. In part (a), original video clip is shown. Black color car (highlighted by red box) is overtaking the white car. In part (b), objects of the video has been shown. In part (c), the final result is provided.	55

LIST OF TABLES

Table 2.1- Summary of benefits and issues in the past related researches	17
Table 7.1 - Quantitative Comparison of Object Detection modules	52
Table 7.2 - Quantitative Comparison of Video Recognition modules	52
Table 7.3 - Qualitative Results for NLP Model	53
Table 7.4 - Quantitative Evaluation on different models	54

ABBREVIATIONS

Abbreviation	Definition
VS	Video Summarization
TVS	Text-based Video Summarization
SL	Supervised Learning
UL	Unsupervised Learning
RL	Reinforcement Learning
WSL	Weakly-Supervised Learning
GAN	Generative Adversarial Network
LSTM	Long-Short Term Memory
NLP	Natural Language Processing
BiLSTM	Bi-Directional Long-Short Term Memory
CNN	Convolutional Neural Network
CSN	Channel Separated Network