A COMPARATIVE ANALYSIS OF VIDEO SUMMARIZATION TECHNIQUES

MOHANAD ALI GASHOT

JUNE 2019

A COMPARATIVE ANALYSIS OF VIDEO SUMMARIZATION TECHNIQUES

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES
OF
ÇANKAYA UNIVERSITY

BY

MOHANAD ALI GASHOT

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR
THE DEGREE OF MASTER OF SCIENCE
IN
COMPUTER ENGINEERING

JUNE 2019

Title of the Thesis: **A COMPARATIVE ANALYSIS OF VIDEO SUMMARIZATION TECHNIQUES**

Submitted by **MOHANAD ALI GASHOT**

Approval of the Graduate School of Natural and Applied Sciences, Çankaya University.

Prof. Dr. Can ÇOĞUN

Director

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

Prof. Dr. Erdoğan DOĞDU

Head of Department

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assist. Prof. Dr. Roya CHOUPANI

Supervisor

Examination Date: 25/06/2019

Examining Committee Members:

Assist. Prof. Dr. Roya CHOUPANI
Computer Engineering Department, Çankaya University

Assist. Prof. Dr. Yuriy ALYEKSYEYENKOV
Computer Engineering Department, THK University

Prof. Dr. Erdoğan DOĞDU
Computer Engineering Department, Çankaya University

STATEMENT OF NON-PLAGIARISM

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Name, Last Name : MOHANAD ALI GASHOT

Signature :

Date : 25/06/2019

ABSTRACT

A COMPARATIVE ANALYSIS OF VIDEO SUMMARIZATION TECHNIQUES

Gashot, Mohanad Ali

M.Sc., Department of Computer Engineering

Supervisor: Assist. Prof. Dr. Roya CHOUPANI

June 2019, 47 pages

With the rapid evolution of the digital era, devices equipped with digital cameras are being more popular and widely used to capture digital videos. The huge number of frames in each video poses challenges toward processing these videos, as the result of the need for intensive processing to handle these frames. Hence, several techniques have been proposed to provide more efficient summaries for videos. An efficient summary is required to provide maximum information about the contents of the video using the minimum number of its selected frames. Such a summary can significantly reduce the complexity of the computations while maintaining high-quality results in the applications that rely on these summaries. Thus, more attention is being attracted by video summarization techniques to be employed in different fields of applications, such as predicting the genre of the video or measuring the similarity between two videos.

Video summarization techniques rely on two main steps, finding the boundaries of the shots in the video and selecting the frame in each shot that holds the maximum information for that shot. A video shot is a sequence of frames that are collected without any cuts or transitions during the capturing of the video.

However, recognizing the boundaries of these shots is a challenging process, due to the rapid development in digital video processing techniques that are used to merge the shots in videos. When these boundaries are recognized, a frame is selected to represent the information in that shot. However, the use of different techniques to find the boundaries of the shots and select the frames to represent them produce different summaries in different methods. Thus, it is important to compare the performance of these methods in order to select the one that is most appropriate for the application the method is required for. Moreover, recognizing the factors that can improve the performance of video summarization techniques can assist future researchers to propose video summarization methods with significantly improved performance.

In this study, the methodologies of recent state-of-the-art methods are reviewed, and their performances are evaluated, so that, a comprehensive review and a reasonable comparison are presented. The evaluation results show that the use of a single pass over the frames of a video reduces the complexity of computations required to find the boundaries of the shots in that video. Such an approach has proven to provide high-quality summaries compared to methods that use multiple passes over the video frames. Moreover, the use of clustering techniques has also shown significant improvement to the quality of the produced summary. Thus, the methods that combine these approaches have shown relatively better performance.

Keywords: Image Processing; Keyframes Clustering; Video Summarization

ÖZ

VİDEO ÖZET TEKNİKLERİNİN KARŞILAŞTIRMALI BİR ANALİZİ

Gashot, Mohanad Ali

Yüksek Lisans, Bilgisayar Mühendisliği Anabilim Dalı

Tez Yöneticisi: Dr. Öğr. Üyesi Roya CHOUPANI

Haziran 2019, 47 sayfa

Dijital çağın hızla evrimleşmesiyle, dijital kameralarla donatılmış cihazlar daha popüler hale geliyor ve dijital videoları çekmek için yaygın olarak kullanılıyor. Her videodaki çok sayıda kare sayısı, bu kareleri işlemek için yoğun işlem yapma ihtiyacının sonucuna göre, bu videoların işlenmesine yönelik zorluklar getirmektedir. Bu nedenle, bu videolar için daha verimli özetler sağlamak için birkaç teknik önerilmiştir. Videodan seçilen minimum kare sayısını kullanarak videoların içeriği hakkında maksimum bilgi sağlamak için verimli bir özet gereklidir. Böyle bir özet, bu özetlere dayanan uygulamalarda yüksek kaliteli sonuçları korurken hesaplamaların karmaşıklığını önemli ölçüde azaltabilir. Bu nedenle, videonun türünü tahmin etmek veya iki video arasındaki benzerliği ölçmek gibi farklı uygulama alanlarında kullanılacak video özetleme teknikleri daha fazla dikkat çekmektedir.

Video özetleme teknikleri, videodaki çekimlerin sınırlarını bulmak ve bu çekim için maksimum bilgiyi tutan her çekimdeki kareyi seçmek için iki ana basamağa dayanır. Bir video çekimi, videonun çekilmesi sırasında herhangi bir kesinti veya geçiş olmadan toplanan bir kare dizisidir.

Ancak, bu görüntülerin sınırlarını tanımak, videolardaki çekimleri birleştirmek için kullanılan dijital video işleme tekniklerindeki hızlı gelişme nedeniyle zorlu bir süreçtir. Bu sınırlar tanındığında, çekimdeki bilgiyi temsil edecek bir kare seçilir. Ancak, çekimin sınırlarını bulmak için farklı tekniklerin kullanılması ve onları temsil edecek karelerin seçilmesi, farklı yöntemlerde farklı özetler üretmektedir. Bu nedenle, yöntemin ihtiyaç duyduğu uygulama için en uygun olanı seçmek için bu yöntemlerin performansını karşılaştırmak önemlidir. Ayrıca, video özetleme tekniklerinin performansını artırabilecek faktörleri tanımak, gelecekteki araştırmacılara önemli ölçüde geliştirilmiş performansla video özetleme yöntemleri önerme konusunda yardımcı olabilir.

Bu çalışmada, son teknoloji ürünü yöntemler gözden geçirilmiş ve performansları değerlendirilmiş, böylece kapsamlı bir inceleme ve makul bir karşılaştırma sunulmuştur. Değerlendirme sonuçları, bir videonun kareleri üzerinden tek bir geçişin kullanılmasının, o videodaki çekimlerin sınırlarını bulmak için gereken hesaplamaların karmaşıklığını azalttığını göstermektedir. Böyle bir yaklaşımın, video kareleri üzerinden çoklu geçişler kullanan yöntemlerle karşılaştırıldığında yüksek kaliteli özetler sağladığı kanıtlanmıştır. Ayrıca, kümeleme tekniklerinin kullanımı da üretilen özetin kalitesinde önemli bir gelişme olduğunu göstermiştir. Dolayısıyla, bu yaklaşımları birleştiren yöntemler nispeten daha iyi performans göstermiştir.

Anahtar Kelimeler: Görüntü işleme; Anahtar Kareler Kümeleme; Video Özeti

TO MY BELOVED FAMILY . . .

ACKNOWLEDGMENTS

I would like to express my sincere gratitude and appreciation to my thesis advisor Assist. Prof. Dr. Roya Choupani, who encouraged and guided me throughout my M.Sc. studies at Çankaya University.

I am deeply grateful to all the jury members for agreeing to read my thesis and to participate in the defense of this thesis. Their careful reading of the thesis and valuable comments are greatly acknowledged.

I would like to express my warmest thanks to my mother, father and brothers, Fouad, Osama and their families for their love, prayers and moral support without which the completion of this work would have been impossible. Furthermore, I would like to express my appreciations to my wife, Farah, for her continuous support and patience throughout my M.Sc. studies. I would like to express my deepest love to my son Mohammed who made my life much more meaningful.

Last, but certainly not the least, I am grateful to my relatives and friends both in Libya and Turkey, who provided me with moral and spiritual support.

TABLE OF CONTENTS

LIST OF TABLES

TABLES

LIST OF FIGURES

FIGURES

CHAPTER 1

INTRODUCTION

1.1. Overview

The analysis of a story or a video required the extraction of features at key-point “frames” [1]. The meaning of the story or the video scene can be encoded using these features. An action in a certain interval of the video or a temporal series of events can be made of multiple scenes. Deep Neural Networks (DNN) are widely used to analyze the contents of each frame in the video, especially Convolutional Neural Networks (CNNs), according to the existence of multi-dimensional that can detect local features. The use of more categories or objects in the training of these neural networks can significantly improve the descriptors they generate for the video frames. Then, an overview of the internet video can be provided by the video summarization technique. Hence, these techniques are considered as one of the key techniques in browsing and managing videos. To improve the efficiency of the produced video summary, summarization techniques attempt to represent as much information in the video as possible using the minimum number of frames selected from that video. Thus, more attention is attracted by these techniques according to the rapidly increasing number of videos being available in recent years [2].

According to the definition and aim of video summarization technique, the use of these techniques reduces the resources consumption required to process the videos, such as energy, bandwidth and storage. Hence, these techniques are being widely employed in different types of applications, such as entertainment, military and security [3]. Moreover, videos are being favored among other types of media, especially with the growth of the internet and the ease of access to these videos by different viewers. Advent of social media and growth of video sharing websites, YouTube, has only contributed to the increasing importance of video graphic content.

More content is uploaded to YouTube a day, than a person can watch in his/her whole lifetime. With the emergence of video content as an effective mode of information propagation, automating the process of summarization a video has become paramount [4].

Recently, the problem of video summarization has shown certain challenges toward producing an efficient summary of the video, where machine learning techniques are being employed to overtake such problem. These techniques aim to evaluate the contents, automatically, and select the most relevant contents from the video for the summary. The employment of machine learning techniques has eased certain video processing tasks, such as producing movies trailers, highlighting sport events, or to shorten the contents of a video, in general. Hence, less amount of information is being processed, instead of processing the entire enormous amount of information in the entire video.

The video summarization method is required to find the informative or important parts of the original video, which requires comprehending the contents of the video. However, with the digital era and the ease of access to the internet, several types of videos are becoming available, varying from home videos to movies and documentaries. Such variation imposes more difficulties toward comprehending the video contents by the video summarization method, especially with the absence of any prior knowledge [5].

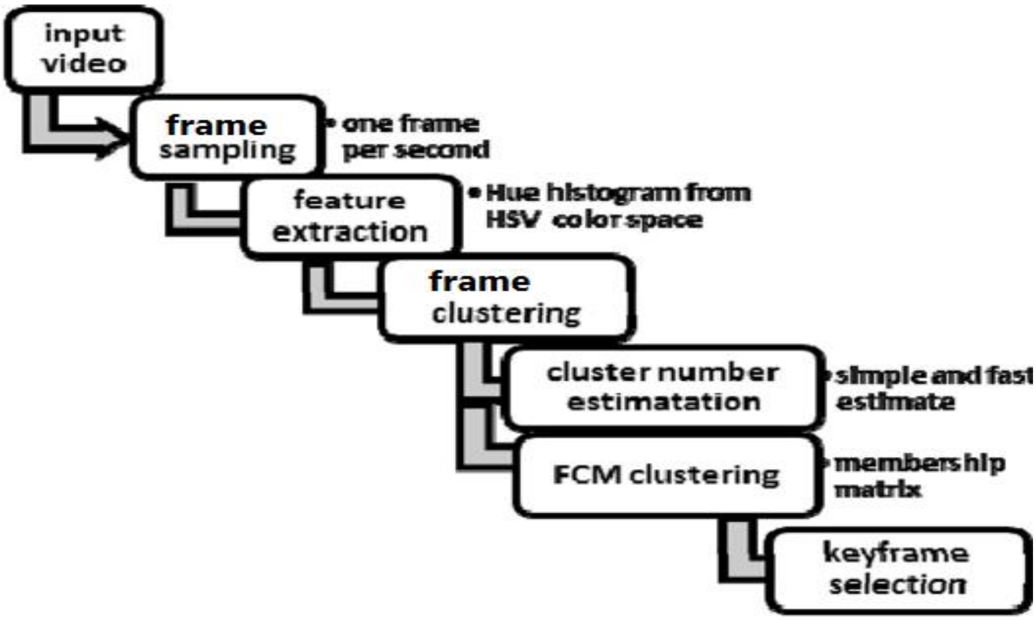

Figure 1.1 Video summarization approach adapted from [3].

With the expanding types of applications and use of video format to represent information, the popularity of digital videos has exponentially increased. Hence, a limitless number of videos are becoming available. Such exponential growth attracts more attention to the video summarization methods by wide range of research efforts. The novel video summarization method presented in [6] tracks the movement of local features in each frame, by tracking the changes in their positions in consequent frames.

Multimedia analysis is required by the video summarization method so that redundancies are recognized; unique and important parts of the video are selected for the summary. Many of the partners from the K-Space European Network of Excellence have cooperated to produce a video summarization method that relies on combination of features analysis. The three steps of the method are:

- Segmenting the video into its original shots, by defining the boundaries of each segment based on several indicators, including the segmentation confidence values.
- Detecting redundancy in the generated segment, so that, a single segment is selected to improve the efficiency of the produced summary, by avoiding redundant information in the output summary.
- The selected top-ranked segments are concatenated in order to produce the video summary, depending on the duration defined as the maximum for the summary.

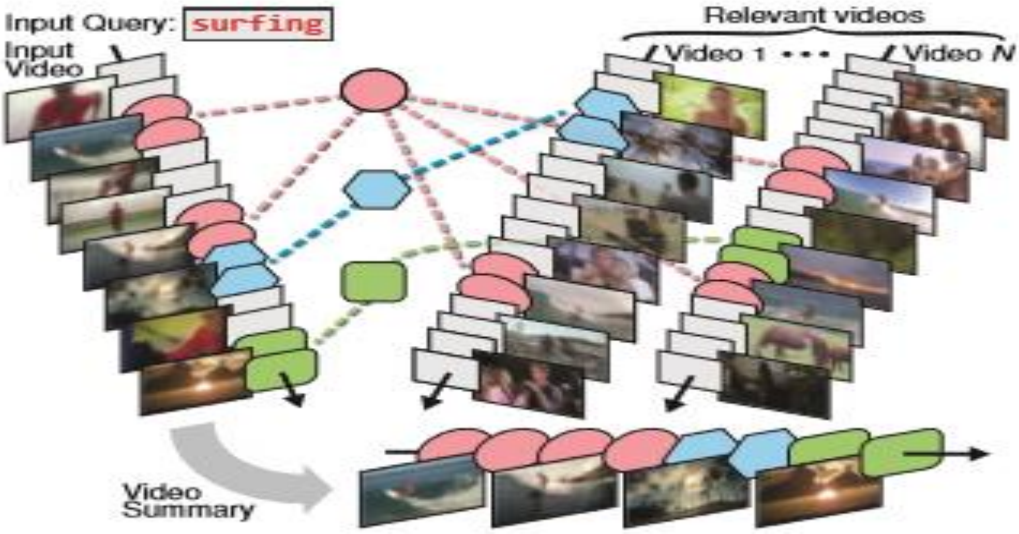

Figure 1.2 An illustration of video co-summarization as identifying visually most similar events shared across N videos.

Despite the huge number of videos and the enormous amount of information in each one, multimedia users have shown continuous interest in these videos, especially those available through direct networks. These videos are being collected and viewed from different sources, such as social networks, digital libraries and media networks. Such interest also increases the importance and attention brought to the video summarization techniques, to reduce the amount of data being processed to brose and manage these videos, in addition to search and retrieval applications. Without video summarization, processing the huge amount of information in each video is a time- and resource-consuming process that can severely reduce the performance of the application. Moreover, video contents can be accessed by many terminals, which requires the use of video summarization methods to provide the users with an overview of the idea being represented in the video. Hence, video summarization can assist adapting the multimedia contents, so that, users are provided with suitable version of the contents, depending on the environment being used.

1.2. Background and motivation

The number of videos contents in multimedia systems has exponentially increased in the recent years. Every user of multimedia system has become a possible source of digital videos, instead of being limited to commercial sources a decade ago [7]. Video contents can be found in different sources, such as digital libraries (e.g. Internet Archive1, Open Video2), personal collections, social media (e.g. YouTube), websites (e.g. ABC6, TVE7), on-demand video providers, digital television broadcasting, optical storage discs (e.g. DVD, Bluray). Thus, multimedia users have shown continuous interest in these videos, especially those available through direct networks. To reduce the time and resources required to process these videos, video summarization techniques are being used to reduce the amount of data required to represent the information in the digital video. Such reduction can improve the efficiency in accessing and navigating through the video contents, which can provide a better experience to the multimedia users, by providing a quicker and more efficient overview of the contents of the video.

1.3. Objectives

The main objective of this study is to review the utilization of several characteristics by the video summarization techniques. This fundamental issue may further be characterized in the general objectives illustrated below, which provides an overview of the core sections of this thesis:

- To observe the utilization of the employment in scalable video coding for the purpose of video summarization, and the way they are being utilized together to summarize digital videos.
- To observe the idea of several characteristics as an intrinsic property of the summary.
- The qualified summary principle must be described alongside with the use of methods used for the key-steps in video summarization techniques.
- To observe potential employment of the video summarization developed previously.

1.4. Video abstraction

Videos are one of the types of contents that require large time-consumption for visualization. A video clip may have a duration that varies from few seconds to hours, where complete visualization of the video is required from users to access all the semantic information in that video [8]. However, most of the applications that process digital videos are required to perform in a very limited time window, where the time required to process such huge amount of data affects the efficiency of the application. Hence, a surrogate summary is used instead of the entire video data, so that, the same information can be represented using less amount of data, i.e. a more efficient representation of the video contents.

1.4.1. The information and browsing time ratio

When a video summary is used as a surrogate for the contents in that video, a significant reduction in the amount of information in that video is imposed [9]. However, the visualization time required to go through that information can be also significantly reduces. When huge number of videos exist, the use of several summaries can reduce the time and effort required to go through those videos, by users. This time reduction imposes a reduction in the amount of information presented in the summary, compared to the amount of information available in the full-length video [8]. A good video summary balances the ratio between the information maintained from the original video in the summary and the length of the produced summary. For illustration purposes, the word shuttle is used in a search query executed in the Open Video Project interface.

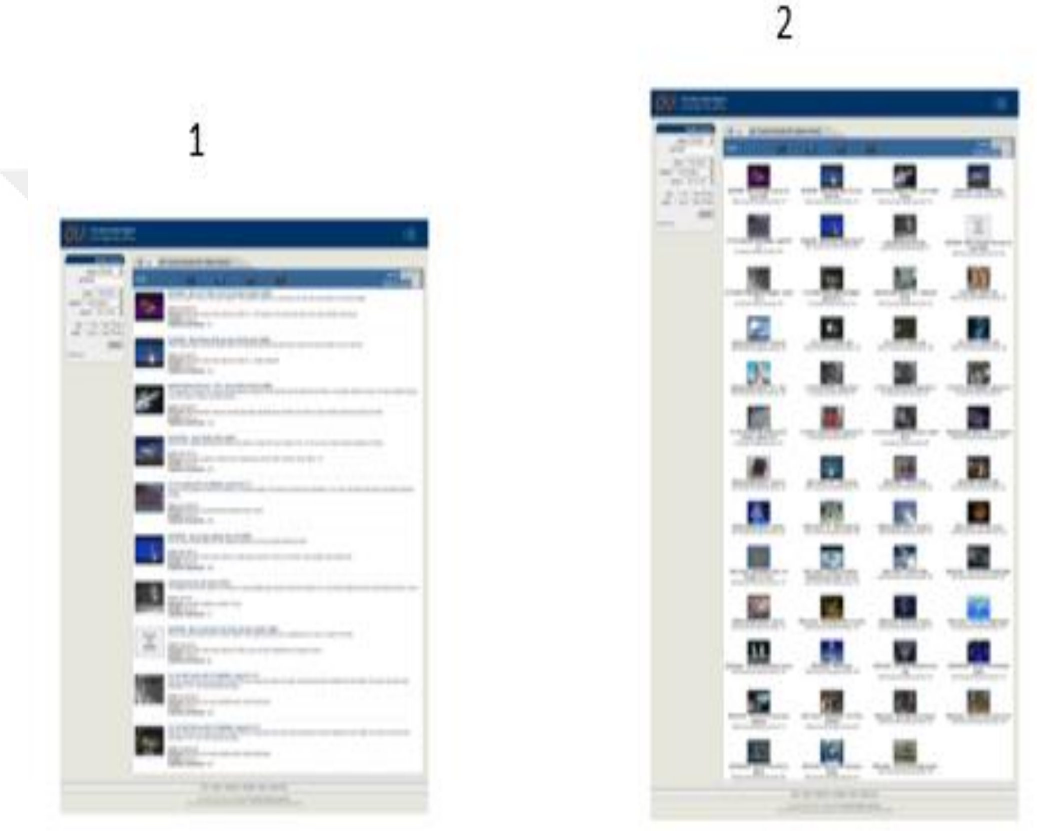

Figure 1.3 The Open Video interface with different layouts: (1) keyframe and description, (2) keyframe and title adapted from [10].

CHAPTER 2

LITERATURE REVIEW

As the usage of unmanned aircraft frameworks increases, the volume of potentially useful video information that unmanned aircraft frameworks capture on their missions is straining the resources of the U.S. military that are desirable to process and usage of this information [11]. The video from [11] is an example of captured via unmanned aircraft frameworks in Iraq. The video shows ISIS fighters herding civilians into a building. U.S. forces did not fire on the building because of the presence of civilians. The video footage remained processed via U.S. Central Command (CENTCOM) prior to release to the public to the important activities within the video, for example, ISIS fighters carrying weapons, civilians being herded into the building to serve as human shields, and muzzle flashes emanating from the building [11].

These video sets take up a huge space for amassing information and takes a long duration to ascertain the content that requires a higher cognitive process for content search and retrieval according to [1]. The efficient method for storing video information is to remove high-degree of redundancies and for creating an index of important events, objects and a preview video based on vital key-frames. These requirements cover the essentials to build algorithms that can meet the space and time requirements for videos and properties of adequate approaches to be developed to solve the desires of summarization. The 3 effective characteristics for a semantic summarized video framework are un-supervision, efficient and scalable framework that can help in reducing time and space complexities.

Video summarization can be improved through the usage of vision-language embeddings trained on image features paired with text annotations, whether from the same domain (i.e., videos of a similar type) or from a quite different one (still images with diverse content) [12]. The feature representation in the embedding space has the potential to better capture the story elements and enable utilizers to directly guide summaries with freeform text input.

According to Dumont et al. [13], video summarization of information sets accompanied via rich text annotations, like the ones released via Yeung et al. as part of their Video SET framework, it also shows their limitations. These information sets have only a few videos that can be highly variable. Thus, the amounts of training and test information are not necessarily sufficient to draw firm conclusions about the relative advantages of different summarization methods. Compounding the problematic are the inconsistencies in the sorts of annotations that are available for different information sets as notation that can be utilized to train good interestingness objectives. While efforts like Video SET are a good start, they want to be greatly expanded in scope.

Chu et al., [14] has reported that the main technical challenge is dealing with the sparsity of co-occurring patterns, out of hundreds to possibly thousands of irrelevant shots in videos being considered. To deal with this challenge, they developed a Maximal Biclique Finding algorithm that remains optimized to find sparsely co-occurring patterns, discarding less co-occurring patterns even if they are dominant in one video. Their algorithm remains parallelizable with closed-form updates, thus can in a simple way scale up to handle many videos simultaneously. They also demonstrate the effectiveness of the proposed approach on motion capture and self-compiled YouTube information sets. Results suggest that summaries generated via visual co-occurrence tend to match more closely with human generated summaries, when compared to several popular unsupervised techniques. A video summarization and optionally summarization processes are presented in the Figure 2.1.

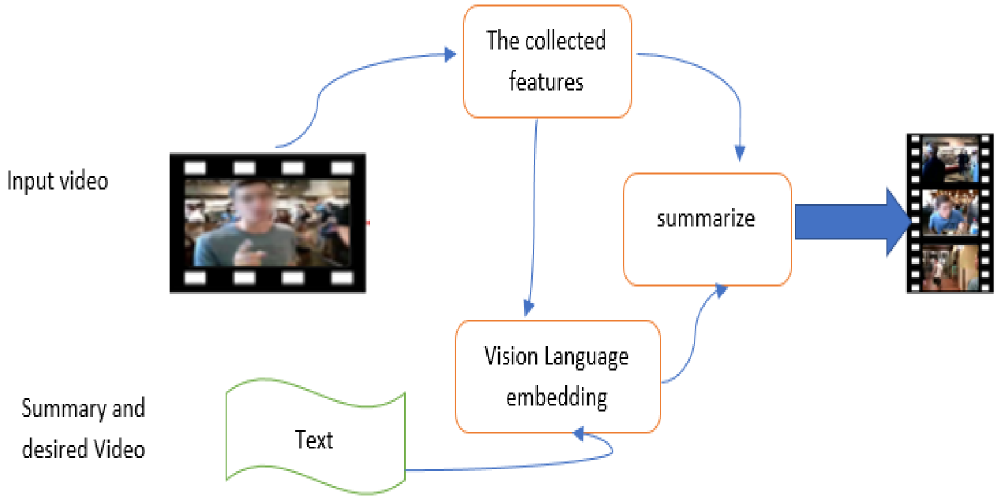

Figure 2.1 A video and optionally summarization processes.

.

Jacob et al., [15] have announced that a new computational visual attention model, inspired on the anthropological visual framework and based on computer vision methods such as face detection, and saliency map computation, to estimate static video abstracts, that remains, collections of salient images extracted from the original videos. Videos from the Open Video Project have been used to indicate that the approach represents an effective solution to the problem of automatic video summarization, producing video summaries with similar quality to the ground-truth manually created via a group of 50 utilizers.

According to Herranz Arribas [10], video summarization and adaptation are time and resource consuming tasks. He also utilizes efficient methods to generate output bitstreams with a very low delay and with low resource requirements. These methods are based on scalable approaches. A bitstream is scalable if, selecting certain packets, a basic version of the content can be obtained, while via including another set of packets an enhanced version can be obtained (e.g. higher resolution).

When dealing with video content, often it is more useful and meaningful to present the summary as a short video sequence, instead of independent frames [10]. Segments provide dynamic information about the events and actions in the video sequence, that isolated images cannot provide. This representation, usually known as video skim, is obtained via selecting certain segments of the original sequence. An additional advantage of video skims is that they can include audio.

According to Li et al., [16], most of earlier work in video summarization chose to select keyframes via randomly or uniformly sampling the video frames from the original sequence at certain time intervals, which was applied in the Video Magnifier, MiniVideo framework. Although this is the simplest way to extract keyframes, the disadvantages are that an arrangement may cause some short yet important segments to have no representative frames while longer segments could have multiple frames with similar content, thus failing to capture and represent the actual video content.

Truong and Venkatesh [17] has declared that several multimedia applications remains rapidly maximizing due to the advance in the computing and network structure, together with the widespread utilization of electronic video technology. Among the key elements for the success of these applications is how to efficiently manage and store a huge amount of audio-visual data, while at the same time providing utilizer-friendly access to the stored data. Video abstraction is a method for generating a short summary of a video, which can either be a sequence of *keyframes* or *video skims*. In terms of browsing and navigation, a good video abstract will enable the utilizer to gain a huge amount of information about the target video sequence in a specified time constraint or sufficient information in time.

Over et al., [18] have maintained that the necessity of video processing applications is to deal with abundantly available videos. Video Summarization aims to create a summary of video to enable a quick browsing of a collection of video maximal size database. It is useful for allied video processing applications such as video indexing, retrieval etc. Video Summarization is a process of creating and presenting a meaningful abstract view of entire video within a short period of time.

Khan and Pawar [19] have announced that sports videos contain some fascinating events that capture attention of the utilizer. People prefer summarized version of sports video rather than watching full videos. Full version of the video may contain many non-significant events such as advertisement, unnecessary playbacks, replays etc. Even if a generic sports video summarization framework remains efficient and useful, the summarization technique in a domain-specific way, such as soccer videos, may present conveniences to utilizers. Many sports broadcasters and web sites utilize editing effects for example super-imposed text captions and slow-motion replay scenes to discriminate the key events. A high-level semantics can be perceived via using these editing effects.

Yuan et al., [2] has reported that a novel Deep Side Semantic Embedding method to generate video summaries via leveraging the freely available side information. The Deep Side Semantic Embedding methods construct a latent subspace via correlating the hidden layers of the two unit-modal autoencoders, which embed the video frames and side information, respectively. Specifically, via interactively minimizing the semantic relevance loss and the feature reconstruction loss of the two unit-modal autoencoders, the comparable common information between video frames and side information can be more completely learned. Their semantic relevance can be more effectively measured. The segments are selected from videos via minimizing their distances to the side information in the constructed latent subspace. They also conducted that experiments on two datasets and demonstrate the superior performance of Deep Side Semantic Embedding methods to several state-of-the-art approaches to video summarization.

Gianluigi and Raimondo [20] have reported that video summarization is reducing the size of data that must be checked to retrieve the desired data from a video, which is an essential task in video analysis and indexing applications. An innovative approach for the selection of representative key frames of a video sequence for video summarization has been proposed. Via analyzing the differences between 2 consecutive frames of a video sequence, the algorithm determines the complexity of the sequence in terms of changes in the visual content expressed via different frame descriptors. The algorithm, which escapes the complexity of existing methods based, such as on clustering or optimization strategies, dynamically and rapidly selects a variable number of key frames within each sequence. The key frames are extracted via detecting curvature points within the curve of the cumulative frame differences. Another advantage is that it can extract the key frames on the fly. The curvature points can be determined by using computer technology to the frame differences and the key frames can be extracted as soon as a second-high curvature point has been detected.

Pritch et al., [21] have proposed the clustered summaries methodology as an efficient method to browse and search surveillance video. Surveillance videos are very long and include many thousands of objects. Regular browsing is practically impossible. In clustered summaries, multiple objects having similar motion are shown simultaneously. This enables to view all objects in a much shorter time, without losing the ability to discriminate between different activities. Summaries of thousands of objects can be created in a few minutes. They conducted a small utilizer study, containing the same objects as a normal summary and a clustered summary. Later, given an activity to search for an efficient viewing of all objects in the surveillance video, the summarized cluster are creating examples for classifiers. Multiple examples can be prepared and given to the learning mechanisms very quickly utilizing unsupervised clustering and clustered summaries. Even a simple nearest neighbor classifier can initially be utilized, cleaned up using clustered summaries, and the results given to learning classifiers.

De Avila et al., [22] have declared that the proposed approach of video summarization has a good performance to generate automatic video summarization, within budget and low time. The video summarization approach tested with the Open Video storyboards. The summaries generated via the proposed algorithm are as good as Open Video storyboards and even better. A huge number of tests must be done to confirm its applicability into video summarization. Nowadays, it is acceptable that this video summarization may be a viable alternative way to video summarization difficulties. They also designing a utilizer study methodology to qualitative evaluate the summaries. Although the issues of keyframe extraction and video summarization have been intensively investigated, there is not a standard or an optimal solution to evaluate their performance. The strategy is based on utilizers' perception. A video summarization approach can be easily utilized to generate video skims.

Zhu et al., [23], proposed a novel unsupervised framework to learn jointly from both visual and independently-drawn non-visual data sources for discovering meaningful latent structure of surveillance video data. They also investigate ways to cope with discrepant dimension and representation whilst associating these heterogeneous data sources and derive effective mechanism to tolerate with missing and incomplete data from different sources. They showed that the proposed multi-source learning framework not only achieves better video content clustering than state-of-the-art methods, but also was capable of accurately inferring missing non-visual semantics from previously unseen videos. In addition, a comprehensive utilizer study was conducted to validate the quality of video summarization generated using the proposed multi-source model.

Herranz and Martinez [24] have reported that Video summaries offer compact images of video sequences, with the length of video summaries playing an important role, rating off the amount of data conveyed and how fast it can be presented. The scalable summarization as a model to easily adapt the summary to a suitable length, according to the requirements in each case, along with a suitable framework. The analysis algorithm utilizes a novel iterative ranking procedure is the result of the extension of the previous one, balancing data coverage and visual pleasantness, overcoming a ranked list, a scalable picture of the sequence beneficial for summarization. The output was efficiently generated from the bitstream of the sequence utilizing bitstream extraction.

Liu et al., [25] have proposed a method that utilizes web pictures for calculating frame interestingness of a lightweight video. Web pictures collections, such as those on Flickr, tend to contain interesting pictures due to their pictures being more carefully taken, composed, and selected. For the reason that these pictures have already been chosen via way of subjectively interesting, they serve as evidence that similar pictures are also interesting. So, using these web pictures researchers calculate the interestingness of video frames. The interestingness of each video frame according to it is similarity to web pictures. The similarity is defined based on the scene content and composition. Interestingness of a video frame is measured via considering how many photos it looks like, and how similar it is to them. Via measuring frame interestingness of videos from YouTube utilizing photos from Flickr show the initial success of the method.

CHAPTER 3

METHODOLOGY

To achieve the objectives of this study, the studies in the literature related to the video summarization topic are investigated by extracting the information required to accomplish the goals of the study. This information is extracted by answering a set of the research question, defined based on the goals of the study. The articles are selected from the most popular digital libraries, be defining a search query suitable for each library, and filtering the results of these searches using a set of including/excluding rules. Data are extracted from each study that passes the including/excluding phase, in order to answer the selected research question.

3.1. Research Question

To extract information relevant to the goals of this study, from earlier methods proposed for video summarization, the goals that this study aims to achieve are illustrated below:

- G1: To classify the existing video summarization techniques according to the methods employed by them to achieve their tasks. This classification can assist recognizing the popular methods employed in video summarization and which of them are getting more focus in the recent years.
- G2: To identify the applications that require video summarization, which aims to determine the fields that make use of video summarization techniques and how these techniques are being employed in different applications over the studied interval.
- G3: To determine the type of video summarization whether to be in real-time or off-line, which also indicates the complexity of the method employed by the video summarization technique, where real-time applications require less-complex methods.

- G4: To identify the user that the proposed method is intended for, which reflects the usability of these methods, as methods requiring specific users or experts are less usable than those that can be used by any user.
- G5: To recognize the popularity of video summarization studies in different countries, by identifying the country that each study is conducted in.

Based on these goals, the following research questions are defined to be answered based on the information extracted from each study in the field of video summarization:

- **RQ1: What is the methodology used in the proposed video summarization method?**
  The answers of this question can be used to measure the popularity of each method, as well as recognizing the fading and rising methodologies used in video summarization.

- **RQ2: What types of inputs does the proposed method accept?** Methods can accept direct images from a video capturing device or videos stored on disks. The answers of this question can assist recognizing the type of applications that the proposed method can handle, real-time or offline applications.

- **RQ3: What are the outputs provided by the proposed method?** Does the summarization method provide images extracted from the video, text to describe the contents in the video or any other types of output, including many possible combinations.

- **RQ4: What is the application that the proposed method is intended for?**
  The answers of this question can be used to identify the field where the video summarization techniques are being widely used, and how such deployment changes occur over time.

- **RQ5: Who are the users that are supposed to interact with the proposed method?**
  Depending on the type of application the proposed method is intended to be used for, users of the system may vary from specialist to any computer user. The answers of this question over the studied interval can illustrate the change in the types of users and the experience required to use these methods.

- **RQ6: When was the study published?**
  By determining the date that each study is published in, the change in any of the previous questions over the studies interval can be investigated.

- **RQ7: Where was the study conducted?**
  The country that the study has been conducted, to illustrate the geographic interest in video summarization techniques.

3.2. Studies Search

Studies related to the video summarization topic are collected from three of the most popular digital libraries, which are the Science Direct, ACM Digital and IEEE Xplore. Initially, the query is set to search for video summarization phrase in the titles of the articles in these libraries. Then, more complex search queries are used in order to improve the results retrieved from these libraries. The final search query used with these libraries can be describes as “Video **AND** (Summary **OR** Summarization **OR** Abstract **OR** Abstraction)” in the titles of the articles. However, as each library uses its own query format, Table 3.1 shows the search queries used with each digital library.

Table 3.1 Search queries for related articles in the selected digital libraries.

Library	Search Query
Science Direct	TITLE("video" AND ("summary"OR"summarization"OR"abstract"OR"abstraction"))
ACM Digital Library	"query": { acmdlTitle:(+video +(summary summarization abstract abstraction) }
IEEE Xplore	("Document Title":video AND (Document Title:summary OR "Document Title":summarization OR "Document Title":abstract OR "Document Title":abstraction))

3.3. Inclusion/Exclusion Criteria

The studies retrieved from the digital libraries are passed through a set of inclusion/exclusion rules. Data are extracted from studies that fulfill the requirements of the inclusion rules and does not match any of the exclusion rules. The rules of the inclusion criteria are:

- The topic of the study is in the video summarization field.
- The study presents a well-described method for video summarization.
- The study is published in the interval between 2008 and 2017.

Moreover, rules in the exclusion criteria are:

- The language that the study is presented in is other than English.
- The study presents a summary of a conference.
- Gray literature and books.
- A duplicate study.
- The full text of the study is inaccessible.

As illustrated in the flowchart shown in Figure 3.1, a study is excluded immediately when it does not match any of the inclusion rules or matches any of the exclusion rules. Studies that are forwarded to the data extraction phase must agree with all the inclusion rules and does not match any of the exclusion rules.

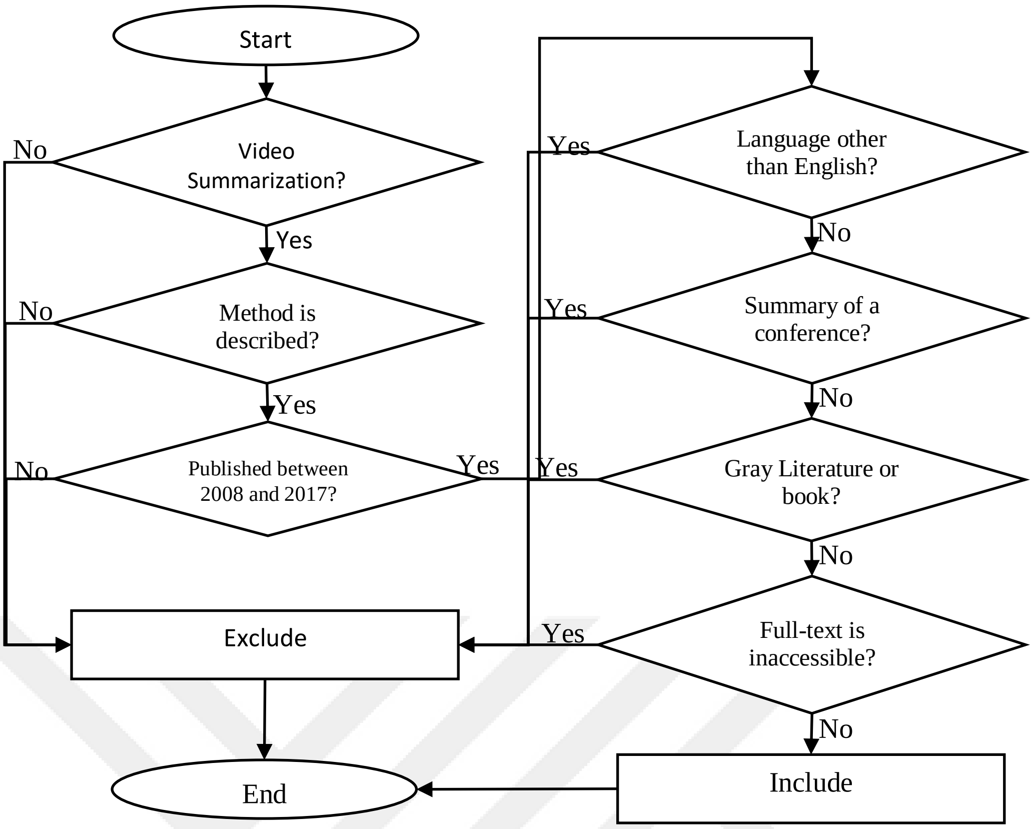

Figure 3.1 Flowchart of the inclusion/exclusion criteria application.

3.4 Performance evaluation

In order to evaluate the performance of the methods selected from the literature using the illustrated criteria, a dataset set of videos acquired from the Open Video digital library [26]. This dataset contains 50 videos of different genres, such as lecture, ephemeral, educational, historical and documentary. The durations of these videos vary from one to four minutes, with constant 30 frames per second. In addition to the videos, the dataset also contains summaries collected based on the recommendations of 50 user, each user summarizes five videos, resulting in five summary recommendations per each video [27].

The summary generated by each of the investigated methods is compared to the summary recommended by the dataset provides, in order to evaluate the performance of the method. The evaluation is conducted using the recall and precision measures, which are combined in a single measure known as the F1-Score. The recall represents the number of the frames in the generated summary that exist in the recommended summary to the total number of frames in the recommended summary. The precision represents the number of frames in the generated summary that are similar to those in the recommended summary to the total number of frames provided by the method. These measures are combined, as shown in Equation 3.1. to calculate the F1-Score, which is the overall measure of the method.

$$F1\_Score = \frac{2 \times Recall \times Precision}{Recall + Precision} \tag{3.1}$$

The recall measure reflects the performance of the method in terms of how good it has been able to extract the frames of the summary, where larger recall values indicate the extraction of more of the required frames. However, it is possible to extract multiple frames that are similar to each other, which reduces the efficiency of the method. This efficiency is reflected by the precision, where higher precision values indicate less duplicates in the extracted frames for the summary. Finally, a method with a high F1-Score indicate that it has high recall and high precision, which reflects high overall performance.

As the frames adjacent to each other may be very similar visually, the position of the extracted frame cannot be an indication of the quality of the summarization method. Thus, the frames in the extracted summary are compared to those in the recommended summary visually, using Speeded-Up Robust Feature (SURF) method. Each extracted frame is compared to those in the recommended summary in order to calculate the recall and precision of the method, hence, calculate the F1-Score. Moreover, the time required to detect and extract the summary from the video frames is also calculated. Two execution times are measured, the first calculated the time required to process each frame in the input video, and the other is the time required for each frame in the summary.

CHAPTER 4

EXPERIMENTAL RESULTS AND DISCUSSION

To select the studies related to video summarization and evaluate the performance of the methods proposed in these studies, the search queries described in Section 3.2 are executed against the corresponding digital libraries. Then, by applying the inclusion/exclusion criteria, shown in Section 3.3, the studies shown in Table 4.1 are selected for evaluation.

Table 4.1 Titles of the studies selected for the evaluation.

Study	Title	Evaluation Dataset
Fayk et al. [3]	Particle Swarm Optimization Based Video Abstraction	This method is evaluated using 20 videos from different genres, such as talk shows, cartoon and news.
Ji et al. [4]	Video Abstraction Based on The Visual Attention Model and Online Clustering	Videos that are used to evaluate the performance of this method are collected from the open-video.org, which also provides a summary for each video generated by combining computer-based and manual methods.
Mei et al. [5]	Video Summarization Via Minimum Sparse Reconstruction	Two datasets, 50 videos from the Open Video Project and 50 videos from [2] that covers several genres, such as sports, commercials, home videos and news.

Study	Title	Evaluation Dataset
Peng and Xiao-Lin [6]	Keyframe-Based Video Summary Using Visual Attention Clues	This method is also evaluated using a combination of two dataset, one collected from the open-video.org and the other from cvrr.ucsd.edu/aton/testbed.
Badre and Thepade [7]	Novel Video Content Summarization Using Thepade's Sorted n-ary Block Truncation coding	The VSUMM dataset [2] is used to evaluate the performance of this method.

The methodology of each of the selected method is summarized as follows:

- **Fayk et al. [28]:** The input video is segmented into a set of shots of a constant interval, in order to select the keyframe per each shot. Per each of these shots, the similarity between every two consequent frames is calculated, by measuring the average red, green and blue color values in these frames. Then, Particle Swarm Optimizer is used to select the keyframes per each segment, so that, the average difference between every two sequential keyframes is maximized in that segment. Binary discrete optimization is used to select these keyframes, where each particle in the swarm has a vector with a size equal to the number of frames in the segment. Each of these frames can be assigned with zero or one, where zero indicates not selecting the frame in the summary and vice versa. However, as the segments are selected based on a constant number of frames, a single segment can contain frames from multiple shots and a certain shot may have its frames distributed into more than one segment. To avoid the extraction of similar keyframes in the outputted summary, this method executes post processing steps that identify and eliminate such frames. First, an intra-merge is executed, where the first frame in a group with average similarity less than 10% is selected as the summary keyframe of the entire group. Then, intra-merge is executed, by measuring the difference between the first keyframe selected in a certain group with the last one from the previous group. If the difference between these two frames is found less than 10%, the first keyframe in the late group is eliminated. This process is repeated for all the frame in the late group, until a difference greater than 10% is found. The experimental results of the study suggest that the use of 150 frames per segment produces the best performances in most of the cases, i.e. different video lengths.

- **Ji et al. [29]:** To reduce the complexity of the computations in this method, the size of the frames in the video is reduced to 240 pixels height, while maintaining the aspect ratio to calculate the width. A vector of 20 values is calculated per each of the frames in the video to predict the shot boundaries in these frames, where half of these features are used to detect abrupt cuts and 10 to detect gradual transitions. Being a boundary of shot is determined by a decision tree classifier, which predicts these boundaries based on the calculated vectors [33]. The keyframes per each shot are selected by measuring the difference between each frame and the last selected keyframe, where the first frame in the shot is selected as the first keyframe. If the difference is larger than the threshold, the frame is selected as a keyframe and the upcoming frames are compared to it, until the end of the shot is reached. Two filtering schemes are used to deal with the key-frames that do not match the inclusion criteria. The first filter drops monochrome frames, by calculating the histogram of the frame’s color, using 10 bins, and compare the ratio of maximum frequency to the image size, to a threshold of 0.75. The second scheme extracts the region of interest based on the saliency map of the frame and compare the ratio between size of that region to the total size of the frame to a threshold of 0.95. Frames with ratio less than this value are considered to attract less visual attention from humans, hence, have no significant value in the outputted summary. To calculate the region of interest, the saliency map is calculated for the frame [34]. Then, the saliency value per each region is compared to two thresholds, in order to update their values and extract the ROI. If the saliency value at that region is larger than 0.5, the region is considered of significant interest and the threshold value is subtracted from the pixels in that region, which produces positive numbers. If the saliency of the region is found to be less than 0.25, the region is considered of no interest and the threshold value is subtracted from each saliency value in the region. For regions with saliency value between these values, the summation of the adjacent regions is calculated and added to the saliency value, after subtracting 0.5 from it, so that, regions within this range adjacent with interesting regions are selected, while those surrounded by uninteresting regions are neglected. To avoid the complexity of the existing clustering techniques, which requires the entire data to create the required clusters, this method proposes an online clustering method that is used to cluster the filtered keyframes to select the ones for the outputted summary. The clustering is based on the similarity between the entire image, the ROI and the background, so that, frames are considered in the same cluster when their overall and the ROI or the background are similar. The similarity is measured by calculating the Euclidean distance of the three-color histograms, with 64 bins, using Bhattacharyya distance. Two values are calculated to determine the cluster of a keyframe. The first value is calculated by subtracting 0.21 from the minimum distance of the ROI and background between the keyframe and the previous one. The second value is equal to the subtraction of 0.13 from the distance between the entire keyframes. If both values are negative, the new keyframe is considered to be in the same cluster as of the previous one. Otherwise, a new cluster is created for the new keyframe. Finally, the keyframe that has the minimum summation of distances to other keyframes in the same cluster is selected for the output summary to represent the frames in that cluster.

- **Mei et al. [30]:** Each of the video frames is represented using a 360-dimensional feature vector, in this method. This vector is created by combining two vectors, one of 252-dimentional, extracted using CENTRIST [10, 11], while the remaining 108-dimensional features are calculated based on the moment of the colors in the frame. Without considering the color information in the frame, a spatial pyramid is created by the CENTRIST to extract the features. The last two levels, where each contains six image patches, are considered in the vector creation process. Per each of these patches, a 40-dimentional eigenvector is created and the mean and variance value of that patch are appended to it, creating a 42-dimentional vector per each patch. Thus, a total of 252 dimensions are created for six patches in the selected spatial levels. Additionally, the frame is divided into 12 regions, 3×4, and the mean, standard deviation and skewness are calculated for each of these patches, per each color channel, producing 108-dimentional feature vector. The methodology of this method is to select a subset of frames that can be used to reconstruct all the frames with maximum Percentage of Reconstruction (POR). The POR between two frames is calculated using the L2 norm distance between the vectors of these frames, where the minimum POR value between a single frame and each of the selected keyframes for summary is selected to represent that frame. After calculating the minimum POR per each frame in the video, the minimum of these values is compared to a predefined threshold value, which is selected to be 70%. If the minimum of these values is less than the threshold value, this frame is appended to the selected keyframes for the summary and the operation is repeated. Otherwise, the selection process is terminated and the selected keyframes are outputted as the summary of the video.

- **Peng and Xiao-Lin [31]:** This method relies on calculating the visual attention index for each frame, by detecting the static and dynamic attentions of these frames. To do so, each frame is divided into 8×8 regions to calculate the dynamic and static attention indices for each of two consequent frames, where the number of these regions depends on the size of the frame. The dynamic attention index is calculated by monitoring the movement of these subregions, where each region is represented using eight vectors for the intensity features, color features, horizontal and vertical positions, vertical and horizontal deviation from the previous frame. The motion coherency is then estimated using kernel-density estimation [37]. The static attention index is calculated using the intensity and color features of these regions, by generating a saliency-based visual attention model using center-surround differences. These attention indices are fused into a single attention index, where priority is given to the dynamic attention index value. The difference between the fused attention indices for every two consequent frames is calculated and compared to a threshold value equal to the summation of the average index and the weighted (multiplied by 1.5) standard deviation. Frames with indices greater than the threshold are selected as candidates keyframes. Then, using K-means clustering technique, a single frame per each shot is selected for the output video summary.

- **Badre and Thepade [32]:** Per each frame in the video, the intensity value of each color channel, red, green and blue, are collected and flattened into a single vector. This vector is ordered ascendingly and divided into N divisions. The average of each division is collected, producing an $N\times3$ values that represent the Thepade's array of the frame. Then, eight similarity measures are calculated for every two consequent frames, which are the Intersection, Wavehedge, Canberra, Sorenson, Chebyshev, Mean Square Error, Fidelity, and Square-chord distances, and averaged to represent the similarity between these two frames. A threshold value is calculated by adding the standard deviation of the calculated similarity measures to their average, where the latter frame of similarities greater than the threshold are outputted as keyframes that represent the summary of the video. Different number of divisions per vector, N, are evaluated in the study, where the results show that the use of five divisions has produced the best performance.

To evaluate the performance of these methods, each method is implemented using Microsoft's C# programming language [38] developed using Visual Studio 2017 Community Edition integrated development environment. These implementations are executed using a computer with Windows 10 operating system with Intel Core-i5 processor, which has 2.4GHz frequency, and 4 GB of memory. OpenCV image and video processing library is used to process the input videos and execute any image processing task, using the EmguCV [39] library, which is a .NET wrapper for the OpenCV library. Tasks related to machine learning, such as clustering, are implemented using the Accord machine learning library for .NET development. The evaluation procedure described in Section 3.4 is used for each of the implemented methods. Table 4.2 reviews the methodology of each method and the main components used for the implementation of the method.

4.1. Experiment A - Particle Swarm Optimization Based Video Abstraction

In this experiment, the method proposed by Fayk et al. [28] is evaluated using the selected dataset. This method divides the video into segments of equal number of frames, then measures the similarity among frames per each segment based on the edges and color distribution in these frames. Then, Particle Swarm Optimization (PSO) algorithm is executed on the frames per each segment to select the keyframe from that segment. As the segments are selected based on the number of frames, a segment may have more than one shot, or a shot may be distributed over more than a segment. Thus, post-processing is executed by measuring the similarity among the frames selected from one segment, in order to select a single frame, which is then compared to the frame selected from the next segment. If the similarity between frames from two consequent segments is found less than a threshold value, the frame from the earlier segment is neglected. Table 4.2 shows the performance measures calculated for the output from this method for each of the videos in the dataset.

Table 4.2 Performance measures of the summaries generated by the method proposed by Fayk et al. [28] .

Video	**Frames**				**Precision**	**Recall**	**F1-Score**	**Execution Time (s/Frame)**	
	Total	**Recom-mended**	**Extracted**	**Match**				**Input Frames**	**Output Frames**
V21.mpg	3290	9	10	5	0.5	0.5556	0.5263	0.0303	9.9689
V22.mpg	2118	5	4	4	1	0.8	0.8889	0.0032	1.702
V23.mpg	1759	11	7	6	0.8571	0.5455	0.6667	0.0251	6.3184
V24.mpg	1815	10	6	6	1	0.6	0.75	0.0211	6.3835
V25.mpg	1797	5	3	2	0.6667	0.4	0.5	0.0137	8.2216
V26.mpg	6269	4	22	4	0.1818	1	0.3077	0.014	3.9867
V27.mpg	3192	6	12	5	0.4167	0.8333	0.5556	0.0125	3.3315
V28.mpg	3561	14	15	9	0.6	0.6429	0.6207	0.0172	4.0855
V29.mpg	1944	5	4	2	0.5	0.4	0.4444	0.028	13.6095
V30.mpg	1815	7	6	5	0.8333	0.7143	0.7692	0.0131	3.97
V31.mpg	2517	7	6	3	0.5	0.4286	0.4615	0.0132	5.5342
V32.mpg	2689	8	3	1	0.3333	0.125	0.1818	0.0272	24.3807
V33.mpg	3261	13	6	4	0.6667	0.3077	0.4211	0.033	17.9098
V34.mpg	4205	17	9	8	0.8889	0.4706	0.6154	0.0022	1.04
V35.mpg	2336	15	8	7	0.875	0.4667	0.6087	0.0208	6.0606
V36.mpg	4223	11	12	6	0.5	0.5455	0.5217	0.0113	3.9692
V37.mpg	3413	4	8	3	0.375	0.75	0.5	0.003	1.279
V38.mpg	4570	10	17	7	0.4118	0.7	0.5185	0.0108	2.9025
V39.mpg	5254	13	22	11	0.5	0.8462	0.6286	0.0094	2.2425
V40.mpg	2940	18	12	12	1	0.6667	0.8	0.0139	3.4025
V41.mpg	2765	14	6	5	0.8333	0.3571	0.5	0.0354	16.3018
V42.mpg	2674	14	10	10	1	0.7143	0.8333	0.0121	3.2328
V43.mpg	4796	28	11	10	0.9091	0.3571	0.5128	0.0344	15.0002
V44.mpg	2429	10	9	5	0.5556	0.5	0.5263	0.0032	0.8504

V45.mpg	2428	12	9	7	0.7778	0.5833	0.6667	0.0028	0.7561
V46.mpg	3591	7	10	5	0.5	0.7143	0.5882	0.0147	5.2733
V47.mpg	2166	7	6	4	0.6667	0.5714	0.6154	0.0196	7.0876
V48.mpg	3705	13	11	9	0.8182	0.6923	0.75	0.0147	4.9517
V49.mpg	3614	22	17	17	1	0.7727	0.8718	0.0106	2.2517
V50.mpg	4829	15	12	9	0.75	0.6	0.6667	0.0041	1.6307
V51.mpg	2934	15	5	3	0.6	0.2	0.3	0.0113	6.6364
V52.mpg	3615	14	8	8	1	0.5714	0.7273	0.0108	4.8906
V53.mpg	1883	6	3	3	1	0.5	0.6667	0.0212	13.3263
V54.mpg	2886	8	5	2	0.4	0.25	0.3077	0.0192	11.0568
V55.mpg	1740	2	4	2	0.5	1	0.6667	0.0068	2.9778
V56.mpg	2325	6	7	4	0.5714	0.6667	0.6154	0.011	3.6457
V57.mpg	3449	7	8	4	0.5	0.5714	0.5333	0.0141	6.0616
V58.mpg	3186	5	11	5	0.4545	1	0.625	0.0166	4.7945
V59.mpg	2106	7	10	5	0.5	0.7143	0.5882	0.0206	4.3384
V60.mpg	2093	10	7	6	0.8571	0.6	0.7059	0.0187	5.5939
V61.mpg	2275	12	9	8	0.8889	0.6667	0.7619	0.0086	2.1738
V62.mpg	2618	7	2	2	1	0.2857	0.4444	0.0123	16.1603
V63.mpg	2310	6	7	4	0.5714	0.6667	0.6154	0.0114	3.767
V64.mpg	5309	12	19	9	0.4737	0.75	0.5806	0.0186	5.1928
V65.mpg	2739	6	8	4	0.5	0.6667	0.5714	0.0205	7.0165
V66.mpg	2187	6	8	3	0.375	0.5	0.4286	0.0109	2.9818
V67.mpg	2722	6	10	5	0.5	0.8333	0.625	0.0041	1.1062
V68.mpg	1949	4	4	2	0.5	0.5	0.5	0.0275	13.3953
V69.mpg	3616	5	13	4	0.3077	0.8	0.4444	0.0096	2.6825
V70.mpg	1407	5	1	1	1	0.2	0.3333	0.0429	60.3888
Average	**2986.28**	**9.66**	**8.84**	**5.5**	**0.658334**	**0.59208**	**0.577184**	**0.015826**	**7.316438**

These results show that the average time required by this method to process an input frame in 0.016s, and an average of 7.32s is consumed per each output frame in the summary. Moreover, the average F1-Score is 0.577 with its values distributed as shown in the histogram illustrated in Figure 4.1, which shows that the summaries of most of the videos are within the F1-Score range between 0.5 and 0.8.

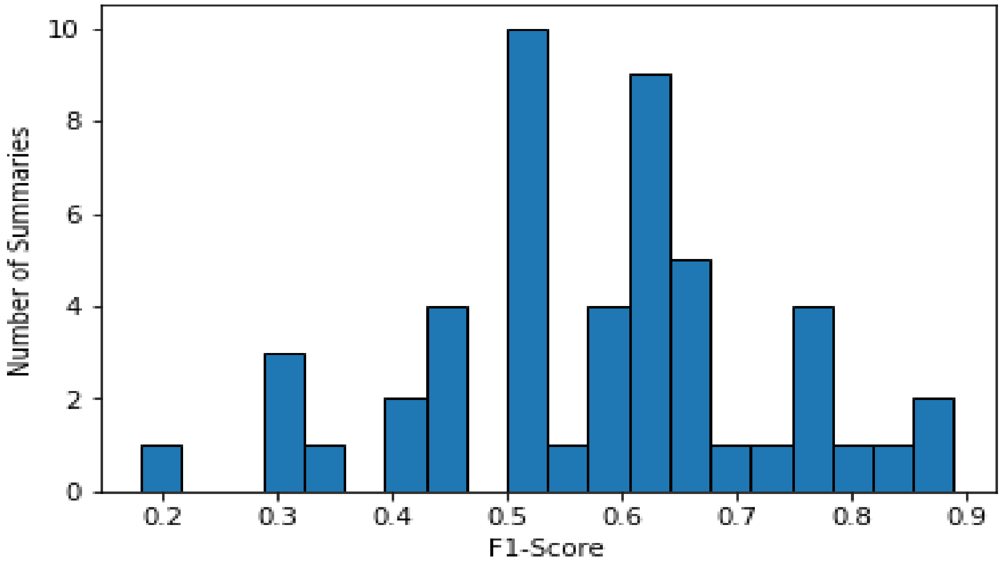

Figure 4.1 Histogram of F1-Scores for the summaries generated by the method proposed by Fayk et al. [28].

4.2. Experiment B - Video Abstraction Based on The Visual Attention Model and Online Clustering

This experiment evaluates the summaries generated using the method proposed by Ji et al. [29]. The similarity among video frames, in this method, is calculated based on Regions of Interest (ROI) extracted from the frame, using saliency maps created for the frames. This approach ensures the negligence of information normally neglected by the viewer, which improves the quality of the frames selected for the summary. The selected frames are then clustered based on the distances among their similarities to select a keyframe from each cluster. As video frames in a single cluster are visually similar to each other, the selection of the frame most similar to others can represent the entire cluster, which increases the efficiency of the method. Table 4.3 summarizes the performance measures calculated for the summaries generated using this method.

Table 4.3 Performance measures of the summaries generated by the method proposed by Ji et al. [29].

Video	Frames				Precision	Recall	F1-Score	Execution Time (s/Frame)	
	Total	Recom-mended	Extracted	Match				Input Frames	Output Frames
V21.mpg	3290	9	15	7	0.4667	0.7778	0.5833	0.0491	10.7667
V22.mpg	2118	5	2	2	1	0.4	0.5714	0.0068	7.1513
V23.mpg	1759	11	6	4	0.6667	0.3636	0.4706	0.0263	7.7
V24.mpg	1815	10	5	5	1	0.5	0.6667	0.0255	9.26
V25.mpg	1797	5	6	4	0.6667	0.8	0.7273	0.0261	7.8162
V26.mpg	6269	4	9	3	0.3333	0.75	0.4615	0.0098	6.8107
V27.mpg	3192	6	8	3	0.375	0.5	0.4286	0.0165	6.6
V28.mpg	3561	14	10	9	0.9	0.6429	0.75	0.02	7.1202
V29.mpg	1944	5	2	2	1	0.4	0.5714	0.0073	7.1004
V30.mpg	1815	7	1	1	1	0.1429	0.25	0.0045	8.2022
V31.mpg	2517	7	3	3	1	0.4286	0.6	0.0085	7.1335
V32.mpg	2689	8	4	2	0.5	0.25	0.3333	0.032	21.5312
V33.mpg	3261	13	10	3	0.3	0.2308	0.2609	0.0279	9.09
V34.mpg	4205	17	1	1	1	0.0588	0.1111	0.0019	8.1994
V35.mpg	2336	15	7	5	0.7143	0.3333	0.4545	0.0947	31.6002
V36.mpg	4223	11	3	3	1	0.2727	0.4286	0.005	7.0666
V37.mpg	3413	4	3	2	0.6667	0.5	0.5714	0.0062	7.1
V38.mpg	4570	10	5	4	0.8	0.4	0.5333	0.0075	6.8393
V39.mpg	5254	13	4	3	0.75	0.2308	0.3529	0.0052	6.799
V40.mpg	2940	18	6	6	1	0.3333	0.5	0.0138	6.7833
V41.mpg	2765	14	11	9	0.8182	0.6429	0.72	0.0376	9.4636
V42.mpg	2674	14	7	5	0.7143	0.3571	0.4762	0.0298	11.3724
V43.mpg	4796	28	15	13	0.8667	0.4643	0.6047	0.0235	7.5064
V44.mpg	2429	10	9	5	0.5556	0.5	0.5263	0.014	3.767
V45.mpg	2428	12	9	7	0.7778	0.5833	0.6667	0.0166	4.4666
V46.mpg	3591	7	8	5	0.625	0.7143	0.6667	0.0334	14.9996
V47.mpg	2166	7	3	3	1	0.4286	0.6	0.0125	9.0002
V48.mpg	3705	13	4	4	1	0.3077	0.4706	0.0073	6.776
V49.mpg	3614	22	14	13	0.9286	0.5909	0.7222	0.0253	6.5214
V50.mpg	4829	15	6	5	0.8333	0.3333	0.4762	0.0084	6.7666
V51.mpg	2934	15	6	5	0.8333	0.3333	0.4762	0.018	8.7827
V52.mpg	3615	14	2	2	1	0.1429	0.25	0.004	7.2496
V53.mpg	1883	6	1	1	1	0.1667	0.2857	0.0112	21
V54.mpg	2886	8	5	4	0.8	0.5	0.6154	0.0288	16.5996
V55.mpg	1740	2	3	2	0.6667	1	0.8	0.0303	17.5668
V56.mpg	2325	6	4	4	1	0.6667	0.8	0.0147	8.5485

V57.mpg	3449	7	7	4	0.5714	0.5714	0.5714	0.0168	8.3001
V58.mpg	3186	5	9	5	0.5556	1	0.7143	0.0207	7.3115
V59.mpg	2106	7	6	3	0.5	0.4286	0.4615	0.0314	11.0333
V60.mpg	2093	10	3	3	1	0.3	0.4615	0.0128	8.9335
V61.mpg	2275	12	6	6	1	0.5	0.6667	0.0203	7.7003
V62.mpg	2618	7	2	2	1	0.2857	0.4444	0.0081	10.5496
V63.mpg	2310	6	3	2	0.6667	0.3333	0.4444	0.0091	7.0333
V64.mpg	5309	12	10	8	0.8	0.6667	0.7273	0.0145	7.68
V65.mpg	2739	6	6	3	0.5	0.5	0.5	0.0212	9.6827
V66.mpg	2187	6	4	3	0.75	0.5	0.6	0.0184	10.0749
V67.mpg	2722	6	3	2	0.6667	0.3333	0.4444	0.0148	13.4327
V68.mpg	1949	4	2	1	0.5	0.25	0.3333	0.0175	17.0496
V69.mpg	3616	5	5	3	0.6	0.6	0.6	0.0128	9.2406
V70.mpg	1407	5	3	3	1	0.6	0.75	0.0376	17.6324
Average	**2986.28**	**9.66**	**5.72**	**4.14**	**0.773386**	**0.45833**	**0.530058**	**0.01932**	**9.934234**

The higher precision of the frames selected by this method for the summary, compared to the method proposed by Fayk et al. [28], indicates that the number of redundant frames in the output is lower in this method. However, the lower recall indicates that many of the recommended frames are missing in the summary. The lower number of frames outputted from this method, with an average of 5.72, also confirms that the method has been missing frames in its outputs, compared to 8.84 in the previous experiment. Moreover, the histogram of the F1-Score values from this experiment, shown in Figure 4.2, shows that the highest quality in the summaries do not exceed 0.8 of F1-Score.

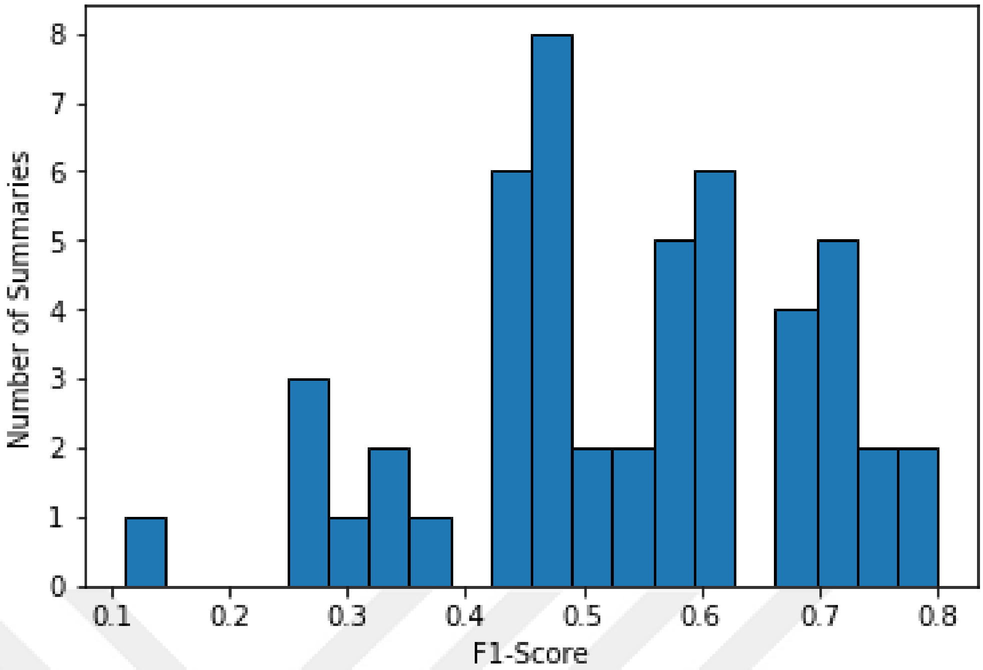

Figure 4.2 Histogram of F1-Scores for the summaries generated by the method proposed by Ji et al. [29].

4.3. Experiment C - Video Summarization Via Minimum Sparse Reconstruction

The performance of the method proposed by Mei et al. [30] is evaluated in this experiment, by evaluating the quality of the summary generated by the method. This method relies on calculating the Percentage of Reconstruction (POR) for each frame by comparing the frame reconstructed from the previous frames to the actual frame in that position. Frames with POR less than a threshold value are then selected as keyframes for the summary of the video. Frames with visual information that cannot be retrieved from previous frames are considered of high importance and must be included in the summary, while those that can be reconstructed from the previous frames are considered to be less important and neglected from the summary, to improve the summarization efficiency without reducing the accuracy. The quality measures of the summary generated per each video is shown in Table 4.4.

Table 4.4 Performance measures of the summaries generated by the method proposed by Mei et al. [30].

Video	Frames				Precision	Recall	F1-Score	Execution Time (s/Frame)	
	Total	Recommended	Extracted	Match				Input Frames	Output Frames
V21.mpg	3290	9	23	5	0.2174	0.5556	0.3125	0.0793	11.3415
V22.mpg	2118	5	6	4	0.6667	0.8	0.7273	0.019	6.7138
V23.mpg	1759	11	8	6	0.75	0.5455	0.6316	0.0812	17.8445
V24.mpg	1815	10	8	5	0.625	0.5	0.5556	0.0652	14.7925
V25.mpg	1797	5	11	4	0.3636	0.8	0.5	0.0448	7.3215
V26.mpg	6269	4	30	4	0.1333	1	0.2353	0.0426	8.9045
V27.mpg	3192	6	17	6	0.3529	1	0.5217	0.0464	8.7106
V28.mpg	3561	14	27	13	0.4815	0.9286	0.6341	0.0503	6.6359
V29.mpg	1944	5	3	3	1	0.6	0.75	0.0836	54.1803
V30.mpg	1815	7	14	3	0.2143	0.4286	0.2857	0.0567	7.3445
V31.mpg	2517	7	13	6	0.4615	0.8571	0.6	0.0408	7.8929
V32.mpg	2689	8	16	7	0.4375	0.875	0.5833	0.077	12.9404
V33.mpg	3261	13	11	7	0.6364	0.5385	0.5833	0.0891	26.4202
V34.mpg	4205	17	9	8	0.8889	0.4706	0.6154	0.0049	2.3116
V35.mpg	2336	15	21	13	0.619	0.8667	0.7222	0.053	5.8982
V36.mpg	4223	11	25	8	0.32	0.7273	0.4444	0.0324	5.48
V37.mpg	3413	4	23	3	0.1304	0.75	0.2222	0.0125	1.8582
V38.mpg	4570	10	33	9	0.2727	0.9	0.4186	0.0326	4.5101
V39.mpg	5254	13	42	12	0.2857	0.9231	0.4364	0.0318	3.9815
V40.mpg	2940	18	22	12	0.5455	0.6667	0.6	0.0547	7.307
V41.mpg	2765	14	7	6	0.8571	0.4286	0.5714	0.0908	35.8753
V42.mpg	2674	14	25	12	0.48	0.8571	0.6154	0.054	5.7717
V43.mpg	4796	28	16	9	0.5625	0.3214	0.4091	0.0894	26.803
V44.mpg	2429	10	9	5	0.5556	0.5	0.5263	0.0067	1.7965
V45.mpg	2428	12	9	7	0.7778	0.5833	0.6667	0.0063	1.702
V46.mpg	3591	7	18	4	0.2222	0.5714	0.32	0.0464	9.2614
V47.mpg	2166	7	15	5	0.3333	0.7143	0.4545	0.066	9.5332
V48.mpg	3705	13	20	9	0.45	0.6923	0.5455	0.0405	7.5008
V49.mpg	3614	22	23	18	0.7826	0.8182	0.8	0.037	5.8184
V50.mpg	4829	15	23	9	0.3913	0.6	0.4737	0.0378	7.9395
V51.mpg	2934	15	21	7	0.3333	0.4667	0.3889	0.0432	6.0345
V52.mpg	3615	14	25	8	0.32	0.5714	0.4103	0.0381	5.5067
V53.mpg	1883	6	9	5	0.5556	0.8333	0.6667	0.0607	12.701
V54.mpg	2886	8	17	3	0.1765	0.375	0.24	0.0624	10.5991
V55.mpg	1740	2	3	0	0	0	0	0.055	31.876
V56.mpg	2325	6	20	6	0.3	1	0.4615	0.0474	5.5094

V57.mpg	3449	7	14	6	0.4286	0.8571	0.5714	0.0438	10.7794
V58.mpg	3186	5	13	3	0.2308	0.6	0.3333	0.0583	14.2823
V59.mpg	2106	7	14	6	0.4286	0.8571	0.5714	0.0674	10.1369
V60.mpg	2093	10	14	7	0.5	0.7	0.5833	0.0596	8.9139
V61.mpg	2275	12	19	12	0.6316	1	0.7742	0.0336	4.0188
V62.mpg	2618	7	5	4	0.8	0.5714	0.6667	0.0289	15.1215
V63.mpg	2310	6	17	6	0.3529	1	0.5217	0.0401	5.4538
V64.mpg	5309	12	29	10	0.3448	0.8333	0.4878	0.0583	10.6753
V65.mpg	2739	6	8	3	0.375	0.5	0.4286	0.0557	19.0829
V66.mpg	2187	6	7	4	0.5714	0.6667	0.6154	0.0421	13.1594
V67.mpg	2722	6	24	4	0.1667	0.6667	0.2667	0.0221	2.5044
V68.mpg	1949	4	7	2	0.2857	0.5	0.3636	0.0839	23.3493
V69.mpg	3616	5	26	4	0.1538	0.8	0.2581	0.0369	5.131
V70.mpg	1407	5	1	1	1	0.2	0.3333	0.0998	140.423
Average	**2986.28**	**9.66**	**16.4**	**6.46**	**0.4554**	**0.676372**	**0.494102**	**0.050202**	**13.793**

The results show that despite the higher average recall, which indicates that higher number of recommended frames are being extracted by this method, the lower precision indicates high redundancies among the outputted frames in the summary. Additionally, the average execution time per each frame in the outputted summary is relatively high, compared to the earlier experiments. The histogram of the F1-Score shown in Figure 4.3 shows that most of the summaries have F1-Score value in the range between 0.4 and 0.7, which is lower than those in the previous experiments.

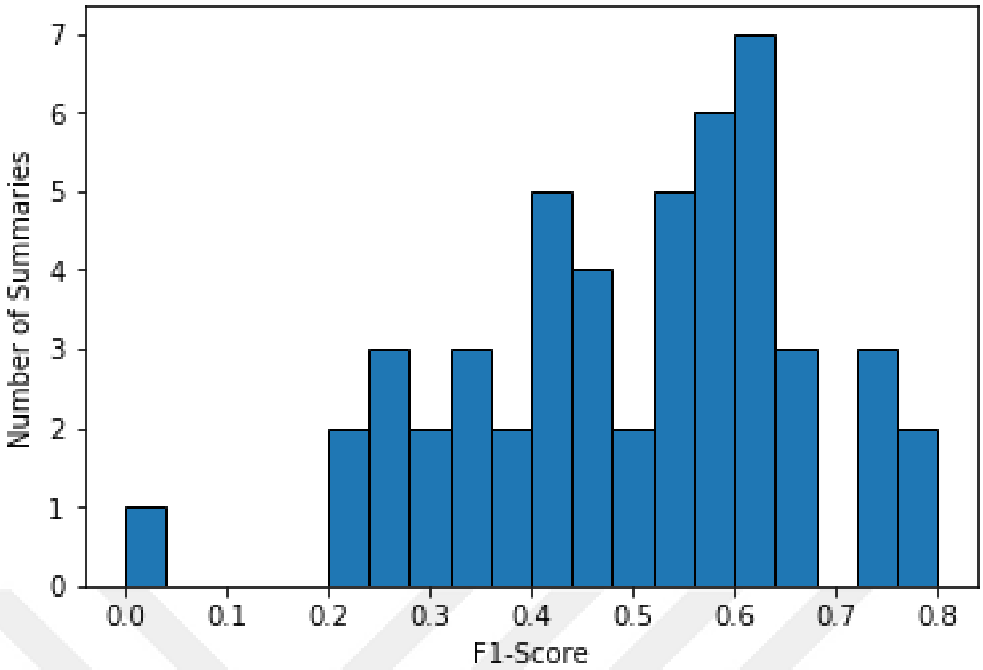

Figure 4.3 Histogram of F1-Scores for the summaries generated by the method proposed by Mei et al. [30].

4.4. Experiment D - Keyframe-Based Video Summary Using Visual Attention Clues

The method proposed by Peng and Xiao-Lin [31] is evaluated in this experiment. The frames of a video being summarized using this method are analyzed to extract the static and dynamic object in the scene. Visual Attention Index is calculated for the static and dynamic parts in order to calculate the overall Visual Attention Index (VAI) for the frame. Frames with VAI larger than a threshold value are selected as the candidates to the summary keyframes. Then, KMeans clustering method is used to cluster these frames in order to select one frame per each cluster for the output. Per each cluster, the frame with the highest VAI is selected to summarize the cluster, which represents the keyframes of a single shot, so that, all the important information per each cluster are included, while a keyframe per each shot is selected. The qualities of the summaries provided by this method are summarized in Table 4.5.

Table 4.5 Performance measures of the summaries generated by the method proposed by Peng and Xiao-Lin [31].

Video	Frames				Precision	Recall	F1-Score	Execution Time (s/Frame)	
	Total	Recom-mended	Extracted	Match				Input Frames	Output Frames
V21.mpg	3290	9	22	8	0.3636	0.8889	0.5161	0.0291	4.3434
V22.mpg	2118	5	4	3	0.75	0.6	0.6667	0.003	1.5405
V23.mpg	1759	11	13	9	0.6923	0.8182	0.75	0.021	2.8356
V24.mpg	1815	10	12	10	0.8333	1	0.9091	0.0117	1.7643
V25.mpg	1797	5	11	5	0.4545	1	0.625	0.0111	1.8267
V26.mpg	6269	4	27	4	0.1481	1	0.2581	0.0093	2.1678
V27.mpg	3192	6	13	5	0.3846	0.8333	0.5263	0.0066	1.6365
V28.mpg	3561	14	22	12	0.5455	0.8571	0.6667	0.0123	2.0145
V29.mpg	1944	5	11	4	0.3636	0.8	0.5	0.0177	3.1194
V30.mpg	1815	7	6	5	0.8333	0.7143	0.7692	0.0105	3.1404
V31.mpg	2517	7	10	6	0.6	0.8571	0.7059	0.0045	1.1364
V32.mpg	2689	8	19	7	0.3684	0.875	0.5185	0.018	2.5308
V33.mpg	3261	13	25	9	0.36	0.6923	0.4737	0.0225	2.9169
V34.mpg	4205	17	6	6	1	0.3529	0.5217	0.0015	0.9981
V35.mpg	2336	15	14	9	0.6429	0.6	0.6207	0.0252	4.2204
V36.mpg	4223	11	18	9	0.5	0.8182	0.6207	0.0075	1.7853
V37.mpg	3413	4	6	3	0.5	0.75	0.6	0.0018	1.0629
V38.mpg	4570	10	21	9	0.4286	0.9	0.5806	0.0078	1.6938
V39.mpg	5254	13	20	10	0.5	0.7692	0.6061	0.0054	1.4439
V40.mpg	2940	18	15	14	0.9333	0.7778	0.8485	0.0093	1.8015
V41.mpg	2765	14	14	10	0.7143	0.7143	0.7143	0.0174	3.4581
V42.mpg	2674	14	10	8	0.8	0.5714	0.6667	0.0096	2.5425
V43.mpg	4796	28	23	16	0.6957	0.5714	0.6275	0.0186	3.8751
V44.mpg	2429	10	9	5	0.5556	0.5	0.5263	0.0042	1.1502
V45.mpg	2428	12	9	7	0.7778	0.5833	0.6667	0.0033	0.8787
V46.mpg	3591	7	12	6	0.5	0.8571	0.6316	0.0117	3.5175
V47.mpg	2166	7	6	5	0.8333	0.7143	0.7692	0.0048	1.6935
V48.mpg	3705	13	9	9	1	0.6923	0.8182	0.003	1.2909
V49.mpg	3614	22	17	16	0.9412	0.7273	0.8205	0.0096	2.067
V50.mpg	4829	15	10	8	0.8	0.5333	0.64	0.0024	1.1937
V51.mpg	2934	15	12	9	0.75	0.6	0.6667	0.0075	1.836
V52.mpg	3615	14	11	8	0.7273	0.5714	0.64	0.0033	1.0929
V53.mpg	1883	6	7	6	0.8571	1	0.9231	0.0102	2.748
V54.mpg	2886	8	11	7	0.6364	0.875	0.7368	0.0129	3.3639
V55.mpg	1740	2	3	2	0.6667	1	0.8	0.0057	3.2529
V56.mpg	2325	6	9	5	0.5556	0.8333	0.6667	0.0078	2.0331

V57.mpg	3449	7	10	5	0.5	0.7143	0.5882	0.009	3.0888
V58.mpg	3186	5	14	5	0.3571	1	0.5263	0.0099	2.223
V59.mpg	2106	7	13	5	0.3846	0.7143	0.5	0.0126	2.0607
V60.mpg	2093	10	14	9	0.6429	0.9	0.75	0.0081	1.1898
V61.mpg	2275	12	11	9	0.8182	0.75	0.7826	0.0057	1.1805
V62.mpg	2618	7	2	2	1	0.2857	0.4444	0.0135	17.5755
V63.mpg	2310	6	9	5	0.5556	0.8333	0.6667	0.0054	1.3599
V64.mpg	5309	12	22	11	0.5	0.9167	0.6471	0.0144	3.5085
V65.mpg	2739	6	9	5	0.5556	0.8333	0.6667	0.0117	3.5685
V66.mpg	2187	6	12	5	0.4167	0.8333	0.5556	0.0078	1.4223
V67.mpg	2722	6	9	5	0.5556	0.8333	0.6667	0.0051	1.5498
V68.mpg	1949	4	6	3	0.5	0.75	0.6	0.0093	3.0471
V69.mpg	3616	5	13	4	0.3077	0.8	0.4444	0.0051	1.4376
V70.mpg	1407	5	3	3	1	0.6	0.75	0.039	18.3531
Average	**2986.28**	**9.66**	**12.28**	**7**	**0.62214**	**0.760258**	**0.643732**	**0.010488**	**2.830764**

According to the average F1-Score achieved in this method, which is 0.64, this method has been able to outperform the method evaluated in the previous experiments. Figure 4.4 also shows that the summary of only one video is less than 0.4 F1-Score, while two summaries have more than 0.9 F1-Score. This method has also relatively lower execution time per each output frame, compared to the methods evaluated in the previous experiments, which makes it more suitable for multiple applications, especially real-time applications that require fast performance.

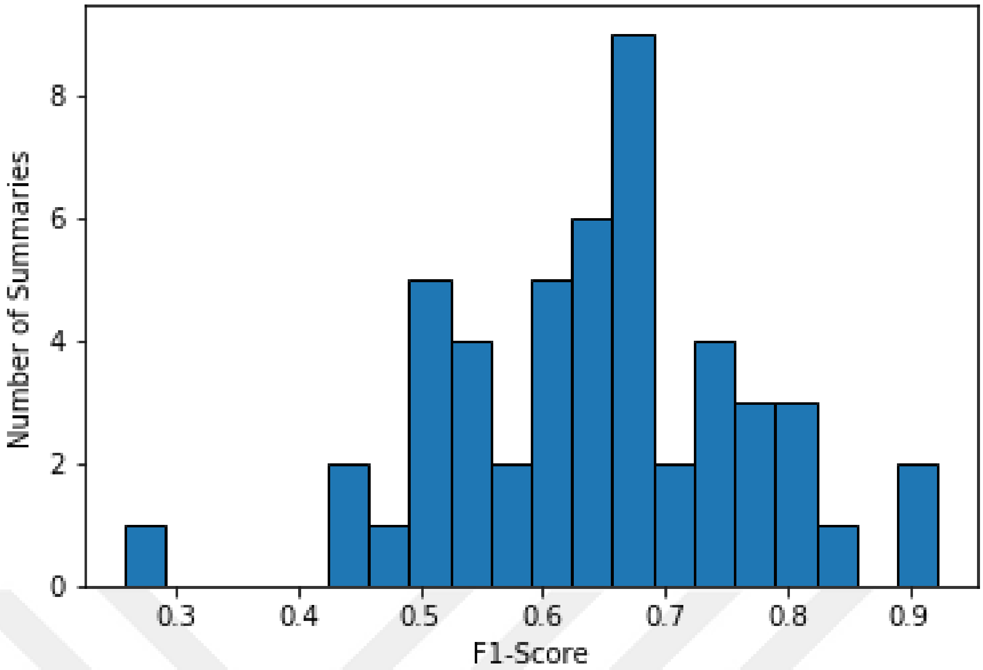

Figure 4.4 Histogram of F1-Scores for the summaries generated by the method proposed by Peng and Xiao-Lin [31].

4.5. Experiment E - Novel Video Content Summarization Using Thepade's Sorted n-ary Block Truncation coding

The performance of the method proposed by Badre and Thepade [32] is evaluated in this experiment, which measures the similarity between frames using eight metrics, namely the Intersection Distance, Square-chord distance, Chebyshev Distance, Mean Square Error, Canberra Distance, Wavehedge Distance, Sorensen Distance and Fidelity Distance. Then, the mean and standard deviations of these similarities for every two consequent frames are calculated, in order to find out the threshold that is used to select the frames in the summary of the video.

Table 4.6 Performance measures of the summaries generated by the method proposed by Badre and Thepade [32].

Video	Frames				Precision	Recall	F1-Score	Execution Time (s/Frame)	
	Total	Recommended	Extracted	Match				Input Frames	Output Frames
V21.mpg	3290	9	24	8	0.3333	0.8889	0.4848	0.0485	6.639
V22.mpg	2118	5	5	4	0.8	0.8	0.8	0.005	2.058
V23.mpg	1759	11	13	8	0.6154	0.7273	0.6667	0.035	4.7295
V24.mpg	1815	10	8	6	0.75	0.6	0.6667	0.0195	4.4145
V25.mpg	1797	5	10	4	0.4	0.8	0.5333	0.0185	3.354
V26.mpg	6269	4	24	4	0.1667	1	0.2857	0.0155	4.085
V27.mpg	3192	6	18	6	0.3333	1	0.5	0.011	1.9735
V28.mpg	3561	14	17	11	0.6471	0.7857	0.7097	0.021	4.3495
V29.mpg	1944	5	10	5	0.5	1	0.6667	0.0295	5.748
V30.mpg	1815	7	9	6	0.6667	0.8571	0.75	0.0175	3.494
V31.mpg	2517	7	13	5	0.3846	0.7143	0.5	0.0075	1.461
V32.mpg	2689	8	20	7	0.35	0.875	0.5	0.03	4.012
V33.mpg	3261	13	17	7	0.4118	0.5385	0.4667	0.0375	7.1615
V34.mpg	4205	17	9	8	0.8889	0.4706	0.6154	0.0025	1.1135
V35.mpg	2336	15	12	9	0.75	0.6	0.6667	0.042	8.2105
V36.mpg	4223	11	15	6	0.4	0.5455	0.4615	0.0125	3.5755
V37.mpg	3413	4	9	3	0.3333	0.75	0.4615	0.003	1.186
V38.mpg	4570	10	24	9	0.375	0.9	0.5294	0.013	2.4845
V39.mpg	5254	13	26	11	0.4231	0.8462	0.5641	0.009	1.855
V40.mpg	2940	18	13	13	1	0.7222	0.8387	0.0155	3.469
V41.mpg	2765	14	13	8	0.6154	0.5714	0.5926	0.029	6.211
V42.mpg	2674	14	11	10	0.9091	0.7143	0.8	0.016	3.8565
V43.mpg	4796	28	19	11	0.5789	0.3929	0.4681	0.031	7.824
V44.mpg	2429	10	9	5	0.5556	0.5	0.5263	0.007	1.917
V45.mpg	2428	12	9	7	0.7778	0.5833	0.6667	0.0055	1.4645
V46.mpg	3591	7	14	6	0.4286	0.8571	0.5714	0.0195	5.029
V47.mpg	2166	7	15	5	0.3333	0.7143	0.4545	0.008	1.1325
V48.mpg	3705	13	13	10	0.7692	0.7692	0.7692	0.0055	1.5065
V49.mpg	3614	22	18	18	1	0.8182	0.9	0.016	3.2575
V50.mpg	4829	15	19	10	0.5263	0.6667	0.5882	0.004	1.0515
V51.mpg	2934	15	14	10	0.7143	0.6667	0.6897	0.0125	2.627
V52.mpg	3615	14	9	9	1	0.6429	0.7826	0.0055	2.2295
V53.mpg	1883	6	6	4	0.6667	0.6667	0.6667	0.017	5.3475
V54.mpg	2886	8	13	6	0.4615	0.75	0.5714	0.0215	4.7475
V55.mpg	1740	2	5	2	0.4	1	0.5714	0.0095	3.257
V56.mpg	2325	6	8	4	0.5	0.6667	0.5714	0.013	3.8175

V57.mpg	3449	7	11	4	0.3636	0.5714	0.4444	0.015	4.684
V58.mpg	3186	5	15	5	0.3333	1	0.5	0.0165	3.4615
V59.mpg	2106	7	11	6	0.5455	0.8571	0.6667	0.021	4.0635
V60.mpg	2093	10	14	8	0.5714	0.8	0.6667	0.0135	1.987
V61.mpg	2275	12	15	12	0.8	1	0.8889	0.0095	1.4645
V62.mpg	2618	7	3	3	1	0.4286	0.6	0.0225	19.5335
V63.mpg	2310	6	9	5	0.5556	0.8333	0.6667	0.009	2.271
V64.mpg	5309	12	22	9	0.4091	0.75	0.5294	0.024	5.851
V65.mpg	2739	6	10	5	0.5	0.8333	0.625	0.0195	5.357
V66.mpg	2187	6	13	5	0.3846	0.8333	0.5263	0.013	2.192
V67.mpg	2722	6	10	5	0.5	0.8333	0.625	0.0085	2.328
V68.mpg	1949	4	10	3	0.3	0.75	0.4286	0.0155	3.051
V69.mpg	3616	5	20	4	0.2	0.8	0.32	0.0085	1.5605
V70.mpg	1407	5	5	2	0.4	0.4	0.4	0.0655	18.363
Average	**2986.28**	**9.66**	**13.18**	**6.82**	**0.55258**	**0.74184**	**0.59491**	**0.01751**	**4.13634**

Despite the lower execution time required per each frame and the high average recall of 0.74, the low average precision of 0.55 indicate that most of the frames selected for the summary are redundant, i.e. almost half of the frames can be removed from the output summary without affecting the information represented by the outputted summary. The low precision has also reduced the average F1-Score of the summaries provided by this method to 0.61, despite the recall rate of 0.74. The histogram of the F1-Scores shown in Figure 4.5 also shows that most of the outputted summaries have F1-Scores between 0.4 and 0.7, which is relatively lower than the summaries provided by the method proposed by Peng and Xiao-Lin [31], which is evaluated in the previous experiment.

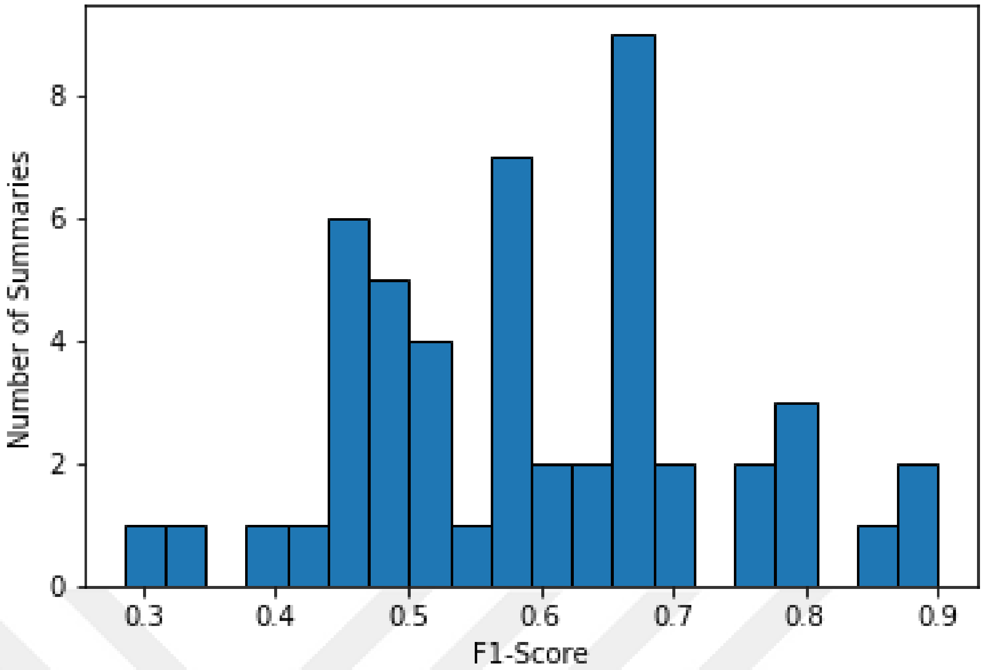

Figure 4.5 Histogram of F1-Scores for the summaries generated by the method proposed by Badre and Thepade [32].

4.6. Comparison and Discussion

In order to compare the qualities of the summaries outputted by the investigated methods, the average precision, recall and F1-Score values of these methods are summarized in Table 4.7. These measures are also illustrated visually in Figure 4.6, which shows that the method proposed by Peng and Xiao-Lin [31] has the highest overall performance. Despite the higher precision of the summaries provided by the method proposed by Ji et al. [29], the low recall rate, compared to the other methods, indicates that the number of frames included in the summary is insufficient to represent the entire video.

Table 4.7 Average performance measures for the summaries provided by the investigated methods.

Study	Precision	Recall	F1-Score
Fayk et al. [28]	0.66	0.59	0.58
Ji et al. [29]	**0.77**	0.46	0.53
Mei et al. [30]	0.46	0.68	0.49
Peng and Xiao-Lin [31]	0.62	**0.76**	**0.64**
Badre and Thepade [32]	0.55	0.74	0.59

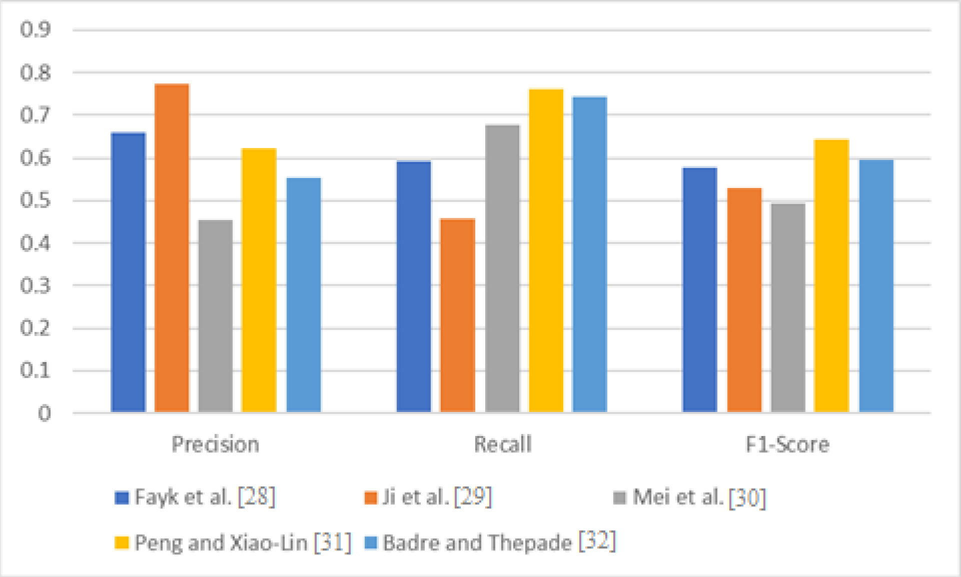

Figure 4.6 Illustration of the average performance measures for the summaries provided by the investigated methods.

Moreover, the lowest performance among the investigated methods is achieved by the method proposed by Mei et al. [30], which has the lowest precision rate in comparison with the other methods. Moreover, the high precision rates of the methods proposed by Ji et al. [29] and Peng and Xiao-Lin [31] illustrate the benefits of using clustering in order to select one frame per a set of candidates for each shot. Such approach reduces the number of redundant frames selected for the summary. However, the low recall rate of the method proposed by Ji et al. [29] show that this method has poor shots segmentation approach, which is based on the visual attention measure extracted from recognized regions of interest. On the other hand, the deployment of visual attention index by Peng and Xiao-Lin [31] has been able to achieve the highest recall, where these indices are calculated for the static and dynamic parts of the frame, before being combined into a single measure to recognize the proper selection of keyframe candidates for the clustering.

Another important performance measure for the video summarization methods is the execution time required to process the frames in the input video and output its summary. These times are summarized in Table 4.8 and shown in Figure 4.7. This comparison shows that the method proposed by Peng and Xiao-Lin [31] has the lowest execution time, which makes it more suitable for different applications. Moreover, the highest execution time is consumed by the method proposed by Mei et al. [30], which requires reconstructing the frames based on the previous ones and measure the similarity between the reconstructed frame and the actual one, in order to select the frames. Additionally, the methods proposed by Fayk et al. [28] and Ji et al. [29] has relatively higher gap between the time required to process the input frames and the time required per each output frames. This gap is a result of the lower number of frames in the provided summary, which reduces the efficiency of the method as more time is required to process frames that are not included in the output summary.

Table 4.8 Comparison of the average execution times consumed by the investigated methods per each input and output frames.

Study	**Time per Input Frame (*m*S)**	**Time per Output Frame (S)**
Fayk et al. [28]	15.83	7.32
Ji et al. [29]	19.32	9.93
Mei et al. [30]	50.20	13.79
Peng and Xiao-Lin [31]	**10.48**	**2.83**
Badre and Thepade [32]	17.51	4.14

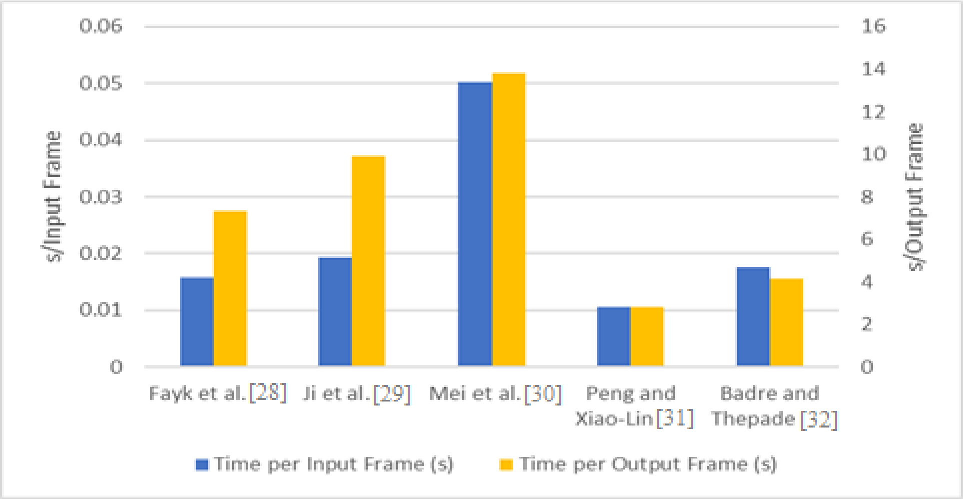

Figure 4.7 Illustration of the average execution times consumed by the investigated methods per each input and output frames.

4.7 Summary

The evaluation results of the investigated methods are used to answer the research question set for this thesis. These answers can be summarized as follows:

1. **Methodologies:** Video summarization includes two main tasks. The first task is to segment the video frames according to the shots, so that, frames that belong to the same shot are included in a single segment. The second task is to extract one frame per each segment, so that, the information in that frame can represent the overall information of the frames in that segment. Clustering has shown better results in recognizing similar frames, which most probably belong to the same shot, in order to select one of these frames as the output. Moreover, the use of visual attention index to select a frame from the cluster has shown relatively better results.

2. **System Inputs:** Although all the video summarization methods accept videos as inputs, the features extracted from the frames in these videos and forwarded to the other processes in the methods are different from one method to another. Color histograms and edges are widely extracted from these frames in order to segment the videos and select a keyframe per each segment. Some of the approaches investigated in this study extract this

information from certain regions of the frames, instead of the entire frame, in order to increase the efficiency of the extracted features.

3. **System Outputs:** The investigated methods attempt to output a frame per each segment extracted from the video, where these methods attempt to extract the frame that contain the most amount of information in that segment. The aim of video summarization methods is to provide a preview of the entire video using the least possible number of frames, i.e. images, so that, a reviewer can predict the contents of the video by only viewing the summary of that video.

4. **System Employment:** The investigated methods can be deployed in any type of application that requires video summarization using images. However, thumbnails selection and search for related videos, by searching for videos with similar summaries, are the applications that mostly use video summarization techniques.

5. **System Users:** According to their deployments, video summarization is achieved by computer for computers, instead of human users. However, in some applications, such as thumbnails selection, the outputted summary is displayed for users in order to select the suitable thumbnail for their video. Moreover, even if summaries are displayed for users to provide an overview of the video, they are still achieved by the computers, while human users are only viewers of the results.

6. **Dates and Locations:** The rapid growth in the number of digital videos being captured in the recent years has brought significant attention to the video summarization techniques. Thus, more studies are being presented to propose video summarization methods. Moreover, the interest in such techniques is found to be worldwide, where the studies are being proposed from different countries and no significant country or region with more interest in this topic is found, which illustrate the importance of video summarization.

CHAPTER 5

CONCLUSION AND FUTURE WORK

More attention is being attracted toward video summarization techniques with the growing number of digital videos being stored and processed in the recent years. A video summarization technique aims to extract the important images from the video, so that, the entire contents of the video can be reflected using these images. Such techniques can reduce the time and storage required to process these videos in order to achieve different tasks. Measuring the similarity among videos is one of the widely used applications that make use of the video summarization. Instead of comparing these videos frame by frame, the summaries are compared in order to recognize their similarities. However, such applications highly depend on the quality of the frames selected from the videos in the summary, where the selection of less important frames can produce wrong results. Thus, the quality of the summary provided by the video summarization techniques is a key feature to evaluate such methods. Moreover, the time required to process the video and output the summary is another important measure of the performance of the methods, where methods that require longer execution time can process less videos and may delay the entire procedure that relies on such method.

In this study, recent video summarization methods are selected from the literature and investigated in order to recognize the methodologies that have the higher performance than others. Such investigation can assist future researches in selecting the appropriate technique employed in newer video summarization methods.

Studies are selected from three of the most popular digital libraries, where certain inclusion/exclusion criteria are followed in order to select the studies investigated in this study. A video summarization dataset is selected from the Open Video Project videos dataset, which contains 50 videos with their recommended summaries. The performances of the selected methods are evaluated by measuring the ratio of images in the recommended summary that the method has been able to extract to the total number of recommended frames, which is known as the recall, and the ratio of the recognized recommended images to the total number of extracted frames, which is known as precision. These measures are then combined in a single overall performance measure, known as the F1-Score. Moreover, according to the importance of the execution time required to process the videos and provide the summary, the average time required to process each input frame from the video and output image in the summary are also measured.

A video summarization method, in general, consists of two phases. The first phase segments the video into shots, where each segment contains the frames from a single shot. The second phase is to select a single frame from that segment to represent the information in that segment. However, different methods use different approaches to achieve these goals, which produce different performances. The use of Visual Attention Index extracted from the static and dynamic parts of each frame and combined into a single measure is found to be the approach with the highest segmentation accuracy. Moreover, the use of clustering methods has also shown significantly better performance, where the selected frames are distributed into clusters depending on their similarities, and one frame per each cluster is selected to represent the frames in that cluster. Methods that use clustering techniques have shown better precision than other techniques, i.e. less redundancy among the selected frames exists. Additionally, the use of visual attention index has been able to achieve higher recall rate, i.e. extracts most of the images in the recommended summary. The method that combines these two approaches, i.e. Keyframe-Based Video Summary Using Visual Attention Clues proposed by Peng and Xiao-Lin [6], has been able to achieve the highest overall performance measures with 64.37% F1-Score and the lowest execution times, with 10.48*m*S per input frame and 2.83S per output frame.

In future work, a new method is going to be implemented based on the findings of this study, where the Visual Attention Index is going to be deployed for keyframes selection and different clustering techniques are going to be evaluated to filter the candidate frames into the output summary.

REFERENCES

[1] **Ma, Y. F., Lu, L., Zhang, H. J., & Li, M. (2002, December),** *"A user attention model for video summarization"*, In Proceedings of the tenth ACM international conference on Multimedia (pp. 533-542). ACM.

[2] **Yuan, Y., Mei, T., Cui, P., & Zhu, W. (2017),** *"Video Summarization by Learning Deep Side Semantic Embedding"*, IEEE Transactions on Circuits and Systems for Video Technology.

[3] **Doulamis, A. D., Doulamis, N. D., & Kollias, S. D. (2000),** *"A fuzzy video content representation for video summarization and content-based retrieval"*, Signal Processing, 80(6), 1049-1067.

[4] **Wang, M., Hong, R., Li, G., Zha, Z. J., Yan, S., & Chua, T. S. (2012),** *"Event driven web video summarization by tag localization and key-shot identification"*, IEEE Transactions on Multimedia, 14(4), 975-985.

[5] **Otani, M., Nakashima, Y., Rahtu, E., Heikkilä, J., & Yokoya, N. (2016, November),** *"Video summarization using deep semantic features"*, In Asian Conference on Computer Vision (pp. 361-377). Springer, Cham.

[6] **Iparraguirre, J., & Delrieux, C. A. (2014),** *"Online Video Summarization Based on Local Features"*, International Journal of Multimedia Data Engineering and Management (IJMDEM), 5(2), 41-53.

[7] **Ejaz, N., Tariq, T. B., & Baik, S. W. (2012),** *"Adaptive key frame extraction for video summarization using an aggregation mechanism"*, Journal of Visual Communication and Image Representation, 23(7), 1031-1040.

[8] **Nam, J., & Tewfik, A. H. (1999, October),** *"Dynamic video summarization and visualization"*, In Proceedings of the seventh ACM international conference on Multimedia (Part 2) (pp. 53-56). ACM.

[9] **Chen, B. W., Wang, J. C., & Wang, J. F. (2009),** *"A novel video summarization based on mining the story-structure and semantic relations among concept entities"*, IEEE Transactions on Multimedia, 11(2), 295-312.

[10] **Herranz Arribas, L. (2010),** *"A scalable approach to video summarization and adaptation"*.

[11] **Collins, R. T., Lipton, A. J., Kanade, T., Fujiyoshi, H., Duggins, D., Tsin, Y., & Wixson, L. (2000),** *"A system for video surveillance and monitoring",* VSAM Final Report, 1-68.

[12] **Plummer, B. A., Brown, M., & Lazebnik, S. (2017, July),** *"Enhancing video summarization via vision-language embedding",* In Computer Vision and Pattern Recognition.

[13] **Dumont, E., Merialdo, B., Essid, S., Bailer, W., Byrne, D., Bredin, H., ... & Sikora, T. (2004),** *"A collaborative approach to video summarization".*

[14] **Chu, W. S., Song, Y., & Jaimes, A. (2015),** *"Video co-summarization: Video summarization by visual co-occurrence",* In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3584-3592).

[15] **Jacob, H., Pádua, F. L., Lacerda, A., & Pereira, A. C. (2017),** *"A video summarization approach based on the emulation of bottom-up mechanisms of visual attention",* Journal of Intelligent Information Systems, 49(2), 193-211.

[16] **Li, Y., Zhang, T., & Tretter, D. (2001),** "*An overview of video abstraction techniques (Vol. 6)",* Technical Report HPL-2001-191, HP Laboratory.

[17] **Truong, B. T., & Venkatesh, S. (2007),** *"Video abstraction: A systematic review and classification",* ACM transactions on multimedia computing, communications, and applications (TOMM), 3(1), 3.

[18] **Over, P., Smeaton, A. F., & Kelly, P. (2007, September),** *"The TRECVID 2007 BBC rushes summarization evaluation pilot",* In Proceedings of the international workshop on TRECVID video summarization (pp. 1-15). ACM.

[19] **Khan, S., & Pawar, S. (2015),** *"Video summarization: survey on event detection and summarization in soccer videos",* International Journal of Advanced Computer Science and Applications, 6(11).

[20] **Gianluigi, C., & Raimondo, S. (2006),** *"An innovative algorithm for key frame extraction in video summarization",* Journal of Real-Time Image Processing, 1(1), 69-88.

[21] **Pritch, Y., Ratovitch, S., Hendel, A., & Peleg, S. (2009, September),** *"Clustered synopsis of surveillance video. In Advanced Video and Signal Based Surveillance, 2009. AVSS'09",* Sixth IEEE International Conference on (pp. 195-200). IEEE.

[22] **de Avila, S. E., da Luz, A., & Araujo, A. D. A. (2008, June),** *"Vsumm: A simple and efficient approach for automatic video summarization",* In Systems, Signals and Image Processing, 2008. IWSSIP 2008. 15th International Conference on (pp. 449-452). IEEE.

[23] **Zhu, X., Loy, C. C., & Gong, S. (2016),** *"Learning from multiple sources for video summarization"*, International Journal of Computer Vision, 117(3), 247-268.

[24] **Herranz, L., & Martinez, J. M. (2010),** *"A framework for scalable summarization of video"*, IEEE Transactions on Circuits and Systems for Video Technology, 20(9), 1265-1270.

[25] **Liu, F., Niu, Y., & Gleicher, M. (2009, July),** *"Using Web Photos for Measuring Video Frame Interestingness"*, In IJCAI (pp. 2058-2063).

[26] **G. Marchionini and G. Geisler (2002),** "*The open video digital library*", D-Lib Magazine, vol. 8, pp. 1082-9873.

[27] **S. E. F. De Avila, A. P. B. Lopes, A. da Luz Jr, and A. de Albuquerque Araújo, "VSUMM (2011),** *"A mechanism designed to produce static video summaries and a novel evaluation method"*, Pattern Recognition Letters, vol. 32, pp. 56-68.

[28] **M. B. Fayk, H. A. El Nemr, and M. M. Moussa (2010),** *"Particle swarm optimisation based video abstraction"*, Journal of Advanced Research, vol. 1, pp. 163-167.

[29] **Q.-G. Ji, Z.-D. Fang, Z.-H. Xie, and Z.-M. Lu (2013),** *"Video abstraction based on the visual attention model and online clustering,"* Signal Processing: Image Communication, vol. 28, pp. 241-253.

[30] **S. Mei, G. Guan, Z. Wang, S. Wan, M. He, and D. D. Feng (2015),** *"Video summarization via minimum sparse reconstruction"*, Pattern Recognition, vol. 48, pp. 522-533.

[31] **J. Peng and Q. Xiao-Lin (2009),** *"Keyframe-based video summary using visual attention clues"*, IEEE MultiMedia, pp. 64-73, 2009.

[32] **S. R. Badre and S. D. Thepade (2016),** *"Novel Video Content Summarization Using Thepade's Sorted n-ary Block Truncation coding"*, Procedia Computer Science, vol. 79, pp. 474-482.

[33] **J. Bescós, G. Cisneros, J. M. Martínez, J. M. Menéndez, and J. Cabrera (2005),** *"A unified model for techniques on video-shot transition detection"*, IEEE transactions on multimedia, vol. 7, pp. 293-307.

[34] **L. Itti, C. Koch, and E. Niebur (1998),** *"A model of saliency-based visual attention for rapid scene analysis"*, IEEE Transactions on Pattern Analysis & Machine Intelligence, pp. 1254-1259.

[35] **J. Wu, H. I. Christensen, and J. M. Rehg (2009),** *"Visual place categorization: Problem, dataset, and algorithm",* in 2009 IEEE/RSJ International Conference on Intelligent Robots and Systems, pp. 4763-4770.

[36] **J. Wu and J. M. Rehg (2011),** *"CENTRIST: A visual descriptor for scene categorization",* IEEE transactions on pattern analysis and machine intelligence, vol. 33, pp. 1489-1501.

[37] **P. Bhattacharya (1967),** *"Estimation of a probability density function and its derivatives",* Sankhyā: The Indian Journal of Statistics, Series A, pp. 373-382.

[38] **A. Hejlsberg, S. Wiltamuth, and P. Golde (2006),** *"The C# programming language"*, Adobe Press.

[39] **C. Emgu (2013),** *"Emgu CV: OpenCV in .NET (C#, VB, C++ and more)",* Online: http://www.emgu.com.

APPENDICES

CIRRICULUM VITAE

PERSONAL INFORMATION

Surname, Name: GASHOT, Mohanad Ali

Nationality: Libyan

Date and Place of Birth: 30 July 1986, Yefren, Libya

Marital Status: Married

Mobile: +90 544 553 71 67

e-mail: mohanad86a@gmail.com

EDUCATION

Degree	**Institution**	**Year of Graduation**
B.Sc.	Al Ryayna Higher Institute for Comprehensive Professions	2008
High School	14th of January Engineering High School	2005

WORK EXPERIENCE

Year	**Place**	**Enrollment**
2008 - 2010	Al Ryayna Higher Institute for Comprehensive Professions	Instructor
2010 – 2012	Administration Higher Technical Institutes	Employee (Student Affairs)
2012 – 2015	Administration Higher Technical Institutes	Head of Department (Student Affairs)

LANGUAGES

Arabic (Native)

English (Fluent)

Turkish (Good)

HOBBIES

Basketball, Swimming, Table Tennis.