User-Centred Video Abstraction

A thesis submitted for the degree of Doctor of Philosophy

By

Kaveh Darabi

Department of Computer Science

Brunel University London

June 2015

Abstract

The rapid growth of digital video content in recent years has imposed the need for the development of technologies with the capability to produce condensed but semantically rich versions of the input video stream in an effective manner. Consequently, the topic of Video Summarisation is becoming increasingly popular in multimedia community and numerous video abstraction approaches have been proposed accordingly. These recommended techniques can be divided into two major categories of automatic and semi-automatic in accordance with the required level of human intervention in summarisation process. The fully-automated methods mainly adopt the low-level visual, aural and textual features alongside the mathematical and statistical algorithms in furtherance to extract the most significant segments of original video. However, the effectiveness of this type of techniques is restricted by a number of factors such as domain-dependency, computational expenses and the inability to understand the semantics of videos from low-level features. The second category of techniques however, attempts to alleviate the quality of summaries by involving humans in the abstraction process to bridge the semantic gap. Nonetheless, a single user's subjectivity and other external contributing factors such as distraction will potentially deteriorate the performance of this group of approaches. Accordingly, in this thesis we have focused on the development of three user-centred effective video summarisation techniques that could be applied to different video categories and generate satisfactory results.

According to our first proposed approach, a novel mechanism for a user-centred video summarisation has been presented for the scenarios in which multiple actors are employed in the video summarisation process in order to minimise the negative effects of sole user adoption. Based on our recommended algorithm, the video frames were initially scored by a group of video annotators 'on the fly'. This was followed by averaging these assigned scores in order to generate a singular saliency score for each video frame and, finally, the highest scored video frames alongside the corresponding audio and textual contents were extracted to be included into the final summary. The effectiveness of our approach has been assessed by comparing the video summaries generated based on our approach against the results obtained from three existing automatic summarisation tools that adopt different modalities for abstraction purposes. The experimental results indicated that our proposed method is capable of delivering remarkable outcomes in terms of *Overall Satisfaction* and *Precision* with an

acceptable *Recall* rate, indicating the usefulness of involving user input in the video summarisation process.

In an attempt to provide a better user experience, we have proposed our personalised video summarisation method with an ability to customise the generated summaries in accordance with the viewers' preferences. Accordingly, the end-user's priority levels towards different video scenes were captured and utilised for updating the average scores previously assigned by the video annotators. Finally, our earlier proposed summarisation method was adopted to extract the most significant audio-visual content of the video. Experimental results indicated the capability of this approach to deliver superior outcomes compared with our previously proposed method and the three other automatic summarisation tools.

Finally, we have attempted to reduce the required level of audience involvement for personalisation purposes by proposing a new method for producing personalised video summaries. Accordingly, SIFT visual features were adopted to identify the video scenes' semantic categories. Fusing this retrieved data with pre-built users' profiles, personalised video abstracts can be created. Experimental results showed the effectiveness of this method in delivering superior outcomes comparing to our previously recommended algorithm and the three other automatic summarisation techniques.

Acknowledgments

In the first place, I would like to thank my dear supervisor, Dr. George Ghinea, for his continued support and guidance in the course of my PhD programme. He is an absolutely professional and determined instructor who helped me to complete this thesis with his valuable comments.

I would like to extend my gratitude to my adorable parents and lovely, beautiful sister for being there and making it possible with their endless love. Baba, Pari and Golnar, I dedicate this work to you and I truly appreciate what you have done for me. Hopefully, I will have an opportunity to make it up for you.

In addition, I would like to convey my thanks to all my family members especially to Babaie, Ozra, Amoo, Afsaneh, Ali, Judith, Joseph, Saghi, Sohrab, Mehran, Sepideh, Bardia, Majid and dear Reza, for their support in these years.

Last but not least, I also thank all my friends, colleagues and staff in Department of Computer Science at Brunel University and to those who gave me their valuable times to participate in my experimental studies.

Thank you very much all.

List of Publications

The following papers have been published or have been accepted for publication as a direct result of the research discussed in this thesis:

1- Darabi, K. and Ghinea, G. (2014) "Video summarization by group scoring," *Multimedia Computing and Systems (ICMCS), 2014 IEEE International Conference on, Marrakech, Morocco*, IEEE, pp.112-116.

2- Darabi, K. and Ghinea, G. (2014) "Personalized video summarization based on group scoring," *Signal and Information Processing (ChinaSIP), 2014 IEEE China Summit & International Conference on, Xi'an, China*, IEEE, pp.310-314.

3- Darabi, K. and Ghinea, G. (2014) "Personalized video summarization by highest quality frames," *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on, Chengdu, China*, IEEE, pp.1-6.

4- Darabi, K., Ghinea, G., Kannan, R., Kannaiyan, S. (2014) "User-Based Video Abstraction Using Visual Features*", Signal Processing and Information Technology (ISSPIT), 2014 IEEE International Symposium on, Noida, India*, IEEE, pp.xxx-xxx.

5- Darabi, K., Ghinea, G. (2015) "Personalized Vide Summarization Using SIFT", *Applied Computing (SAC), The 30th ACM/SIGAPP 2015 Symposium On, Salamanca, Spain* ACM, pp. xxx-xxx.

List of Abbreviations

$\mathbf{AvgFrame_N}$	Average Assigned Score to N^{th} Frame
ASR	Automatic Speech Recognition
BEG	Best Educated Guess
CBAC	Content Based Adoptive Clustering
CFN	Common Feature Number
DI	Digital Item
DSR	Design Science Research
FMRI	Functional Magnetic Resonance Imaging
$\boldsymbol{Fr_i}$	i^{th} Frame
$\mathbf{FrameScore_{NM}}$	Assigned Score to N^{th} Frame by M^{th} Operator
GA	Genetic Algorithm
GMM	Gaussian Mixture Model
GRIF	Generalized Robust Invariant Feature
HIP	Heterogeneity Image Patch
$\boldsymbol{Im_j}$	j^{th} Image
IS	Information Systems
LLC	Locality-Constrained Linear Coding
LSU	Logical Story Unit
MFCC	Mel-Frequency Cepstrum Coefficients
MM&P	Metadata Management and Personalisation
MI	Mutual Information
POS	Part of Speech
PME	Perceived Motion Energy
RAD	Rapid Application Development

ReqNO	Required Number of Frames
RIFT	Rotation Invariant Feature Transform
ROI	Region of Interests
SIFT	Scale Invariant Feature Transform
SOMP	Simultaneous Orthogonal Matching Pursuit
SMQT	Successive Mean Quantization Transform
$\mathbf{SS_n}$	Relevancy Score of $\boldsymbol{N^{th}}$Scene of Movie for a User
SVM	Support Vector Machine
SM1	First Video Summarisation Method
SM2	Second Video Summarisation Method
SM3	Third Video Summarisation Method
SM4	Proposed Video Summarization Method in Chapter 4
SM5	Proposed Video Summarization Method in Chapter 5
SM6	Proposed Video Summarization Method in Chapter 6
SURF	Speeded Up Robust Features
TarVidTime	Required Video Summary Time
VIPP	Video Portal with Personalisation

Table of Contents

List of Tables

List of Figures

Chapter 1
Introduction

1.1. Multimedia Content Expansion

In recent years, the generation and availability of digital videos has been increasing at an exponential rate (Money and Agius, 2007). This has been largely due to advent of internet and online multimedia technologies. The rapid development of cheap storage media, advanced compression methods, higher quality and faster output devices are just a group of other influential factors that have played major roles in accelerating this growth. This has accordingly led to the arrival of multiple multimedia applications which have affected the users' level of demands. Furthermore, viewers are facing an enormous collection of multimedia information that is extremely difficult to manage and extract the required content from. As a result, researchers have been inspired to explore the potential techniques to store, browse and retrieve different multimedia content such as audio, images and videos in the most efficient and profitable ways.

Thus, this has imposed the need for the development of mechanisms with the capability to reflect the most visually, auditory and semantically valuable multimedia content in more compact and efficient forms. This can be simply justified in regards to the great amount of time and cost that will be saved in the presence of such techniques (Ajmal et al., 2012). Considering digital video as the most pertinent media content, a significant amount of research has been devoted to the Video Summarisation topic. In a nutshell, this entails the production of condensed versions of full length videos through the identification and extraction of the most admissible content of input stream.

The general topic of summarisation or abstraction has been under investigation for quite some time since the arrival of Information Retrieval theory. In the mentioned field, the textual documents have to be analysed in order to extract the essential segments that represent

the entire document concisely at an acceptable level (Manning et al, 2009). This can be mapped into the context of video summarisation by considering the original video as the entire text document, which should be assessed in order to retrieve the most imperative partitions (Li and Merialdo, 2010).

However, the abstraction of video can be considered as a more complicated task due to its multimodal nature. In fact, digital videos are usually composed of a number of different media including audio, image and text that should be considered simultaneously for most of the content retrieval and summarisation objectives. This can be articulated to the fact that each of these mentioned media can be regarded as an important information resource with potentially valuable data which can determine the content of any eventual video digest. Moreover, the final outcome can be integrated into other video processing-related applications such as interactive multimedia browsing and searching systems, which further highlights the importance of effective video summaries.

1.2. Video Summarisation

As will be discussed in the next chapter extensively, video summarisation is the process of extracting the most valuable aural, visual and textual content of an input video in order to provide end-users with shorter but semantically rich versions of the original stream.

Generally, video summaries are categorised into two groups, namely, static video abstracts and dynamic video skims, based on the nature of extracted content (Truong and Verkatesh, 2007). The highest quality representative frames solely form the content of the first type, while retrieving the highlights from the original sequence is the basis for producing dynamic video digests (Money and Agius, 2007). An extended discussion in this regard will be provided in the next chapter.

The algorithms that are being employed for these purposes can be categorised into two major groups of automatic and semi-automatic in accordance with the required level of human involvement in the abstraction task. There are a number of positive and negative attributes that can be associated with each of these categories, which will be explained comprehensively in chapter 2. Abstraction methods can also be classified based on the modalities that they adopt for the analysis and obtaining the valuable video segments.

Furthermore, domain dependency can be considered as another determining factor in grouping the video summarisation techniques.

1.3. Personalisation

This concept has been vastly adopted in different areas of computer science during recent years (Mobasher et al., 2000; Fukumura et al., 2008). This can be substantially articulated to the businesses' demands in capturing and fulfilling their customers' expectations in order to gain an edge in competitive markets. In general, personalisation can be defined as the procedure of customising the output data in respect of the audiences' priorities and inclinations in order to meet their requirements (Fukumura et al, 2003). Interactive video systems, e-commerce websites and search engines are just some examples of the fields that integrate personalisation modules. For instance, users' online shopping habits and information regarding their selected items are obtained and processed in an attempt to tailor the content of offers and webpages based on their captured interests (Wong et al, 2005). Personalisation in the context of multimedia can be defined as the attempt to tailor and output the content in accordance with the viewers' perceived requirements and interests towards different multimedia content. This theory will be analysed in more detail in the following chapter.

1.4. Personalised Video Summarisation

One of the emerging topics in multimedia that has received a great amount of attention by researchers in recent years has been personalised video summarisation. This concept was formed by fusion of the two earlier discussed research fields. The primary objective in producing personalised video abstracts is to address the end-users' priorities in extracting the most important video segments (Takahashi et al., 2005b). Similar to any other multimedia tool with a personalisation component, there should be a mechanism in place to understand the viewers' preferences and expectations. These retrieved data should be further incorporated into a summarisation module in order to produce satisfactory results.

Different algorithms and techniques have been proposed by researchers in furtherance to produce user-tailored video digests that will be evaluated comprehensively in chapter 2. It should be mentioned that analogous to normal video summarisation systems, these techniques

can be automatic or semi-automatic in regards to the abstraction phase. In addition, these tools can extract the users' data explicitly or in an implicit manner.

1.5. Research Motivation

In spite of numerous algorithms suggested by researchers, video summarisation is still a challenging topic. In fact, a group of major drawbacks could be associated to the existing abstraction tools, which have considerably and negatively affected the effectiveness of these methods. Domain-dependency, media-reliance and user's subjectivity are just a number of these issues that will be discussed in detail in the next chapter. Accordingly the main motivation for current research is thus the development of an abstraction technique which can produce semantically rich video summaries with minimal dependency to a particular domain and users' personal preferences. In response to this motivation, a number of research objectives have been defined that we will address in the corresponding chapters of this thesis.

1.6. Research Aim and Objectives

Accordingly, the main aim of this research has been defined as **To develop three effective video summarisation techniques that could be applied to different video categories and generate satisfactory results in terms of *Recall, Precision, Timing* and *Overall Satisfaction*.**

This is in response to the remarks from the previous section in regards to a noticeable number of shortcomings that have hindered the existing summarisation techniques from achieving high-quality results.

In order to fulfil this aim, as chapter 2 will show, three studies will be carried out in order to test if the 4 identified research objectives have been achieved:

Objective 1: To investigate the exiting video summarisation techniques in order to identify the limitations and barriers against of this technology. A secondary research into the existing literature should be adopted to achieve this objective.

Objective 2: To design, develop and evaluate a user-centred video summarisation algorithm based on group scoring in accordance to the findings from the previous investigation. The accomplishment of this objective will be described in chapter 4.

Objective 3: To extend the work of previous objective and design, develop and evaluate a personalised video summarisation algorithm based on group scoring. The achievement of this objective will be discussed in chapter 5.

Objective 4: To extend the work of previous objective and design, develop and evaluate a personalised video summarisation system with reduced end-user involvement. The fulfilment of this objective will be described in chapter 6.

1.7. Thesis Outline

In the preceding sections, the main thrust of this research, which was to propose and evaluate a number of video summarisation techniques based on the identified objectives (derived from limitations and strengths of the existing models), was discussed. Thereafter, in this section, the structure of the research work, as will be carried out in this thesis, is shown.

Chapter 2: In this chapter some background information in regards to digital video and its structural units will be initially provided. This is followed by an overview of a number of video processing tasks that are widely adopted in different video abstraction mechanisms. Further, the theory of video summarisation and its main components are discussed comprehensively. Classifying the existing techniques based on their nature, a number of currently proposed abstraction models are categorised and reviewed. Thereafter, the concept of personalisation and its applications within the multimedia context will be discussed. Moreover, the integration of personalisation modules in video summarisation models in an attempt to produce user-tailored summaries is subsequently investigated. Additionally, a group of methodologies and metrics that have been utilised by researchers to assess the quality of the video summaries are reviewed. Finally, a number of shortcomings in respect of the discussed methods are disclosed and the identified research objectives will be accordingly pointed out.

Chapter 3: Primarily, a brief review on different research paradigms and their characteristics are carried out in this chapter. The justification behind the adoption of Positivism as the chosen research perspective in accordance to its attributes and the nature of our research is provided. Later, in response to the required research activities in the course of our work, Design Science Research (DSR) will be introduced as the employed methodology for conducting the research. Furthermore, the adopted methodologies for performing the necessary tasks within each phase of our DSR-based approach will be justified. The rationale behind the adoption of Rapid Application Development (RAD) as the software development methodology will also be detailed. Finally, the experimental procedure to evaluate the effectiveness of our proposed methods will be described extensively.

Chapter 4: Our proposed user-centred video summarisation technique will be described and evaluated in this chapter. According to our suggested approach, video frames should be scored by a panel of video annotators based on their visual and semantic saliency. Further, the assigned scores are averaged in order to produce a singular representative value for each frame. Subsequently, the most salient video frames will be transferred into a final video digest alongside their articulated audio and textual content. Lastly, an experimental study will be carried out to investigate the effectiveness of our approach by comparing the summaries generated by our system against the versions produced by other selected automatic tools.

Chapter 5: In this stage, our user-centred personalised video summarisation method will be discussed in detail. In this technique, an approach to capture the viewers' priorities towards existing scenes in a video sequence will be introduced. The obtained information will be incorporated into a summarisation module to upgrade the initially achieved saliency scores by the frames belonging to the scenes by a higher level of priority. Finally, akin to the recommended method in chapter 4, the highest quality video segments containing audio, visual and textual content will be extracted and inserted into the final video skims. Similarly to chapter 4, a comparative study will be carried out to evaluate the quality of generated summaries by our novel system against the ones produced by automatic tools. However, in the current research, the created output will be checked against the outcome of our proposed algorithm in chapter 4 as well as to investigate the potential improvement caused by integrating the personalisation component.

Chapter 6: A novel personalised video abstraction technique will be proposed in this chapter with the primary goal of facilitating the process of understating the users' priorities. In order to do so, Scale Invariant Feature Transform (SIFT) which is the basis for our personalisation task, will be reviewed initially. Further, a mechanism for measuring the viewers' interests into 103 high-level visual categories is introduced in an attempt to generate the personalised video skims. In addition, the relevancy of each video scene towards each of these 103 video categories will be calculated in order to figure out the preference level of a particular user towards that scene. Finding out the priority level of a particular end-user towards different video segments in an input video stream, the previously assigned scores for the frames residing in the scenes with higher degree of importance will be improved upon. Finally, analogous to the adopted methods in chapter 4 and 5, the most significant video frames (with highest scores) are selected to be inserted into the final summary. Lastly, a comparative study will be employed to verify the effectiveness of our approach by testing the video digests created by our recent technique against those retrieved from automatic tools and our previous method.

Chapter 7: A summary of our research findings are provided in this chapter alongside the highlights of the knowledge contributions that have been made by conducting this research. In addition, a number of limitations that can be associated with this work will be outlined and proposals for future work will be made accordingly.

Chapter 2
Literature Review

2. Overview

This chapter starts off by presenting a brief description regarding digital videos, their structural units and a set of prominent related topics. Later, feature retrieval, video segmentation and keyframe extraction, which are fundamental concepts in video abstraction tasks, will be reviewed. This will be followed by an introduction to *Video Summarisation* as one of the recent trendiest video-processing-related research topics. Further into the chapter, a group of different techniques that has been applied to produce video summaries will be explained. Having discussed the concept of personalisation in general and specifically in the multimedia context, the topic of *Personalised Video Summarisation* will be then elaborated upon. In addition, a number of existing methods that create personalised video abstracts will be expounded. These techniques will be analysed in terms of their strengths and shortcomings and the research objectives will be clarified accordingly.

2.1. Digital Videos

Video is a generic term used for a story told with moving images and sound. This media is tied an attribute of human perception, namely the persistence of vision. This is our visual system's capability to combine consecutive still images into one fluid moving image (Burg, 2009). The fast expanding applications of videos have imposed an increasing demand for development of technologies and tools with the capability of efficient video processing tasks such indexing, browsing and retrieval of video content. In general, any advanced and complex video processing operation requires understanding of the structure of video as a pre-requisite (Zhang et al., 2007). As a result, a brief description of digital video construction units will be presented below (Figure 2.1):

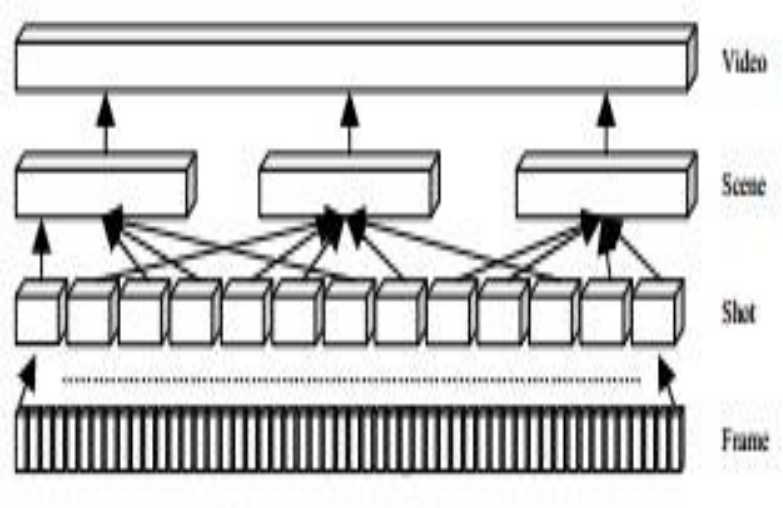


Figure 2.1.Video structural units (Pan et al., 2009)

A digital video is composed of subsequent *frames* that are displayed to viewers in fast succession to induce the impression of movement. A single frame is equivalent to a still image that is composed of hundreds of pixels. A *pixel* is the smallest addressable element of any still picture and contains binary data regarding the colour information of its corresponding physical point in a readable format for computers (Graf, 1999). As a result, any single frame within a video can be specifically addressed by its frame index and in accordance to its temporal location. Moreover, the speed at which these still images are displayed is called *frame rate* (Burg, 2009).

In higher levels, a video can be described based on two construction units: *Shots and Scenes*. A shot is a sequence of frames that continues for an uninterrupted period of time taken by a single camera, representing a continuous action in time and space (Sklar, 1990), while a video scene is composed of a sequence of semantically and visually correlated shots. In other words, all the constituting shots of a scene share the same content in terms of action, place and time (Corridoni and Bimbo, 1998).

There are generally two types of transitions between consecutive shots: *Abrupt and Progressive*; an *Abrupt* shot change is a cut or camera break and happens in one frame, while in the case of progressive, the transition occurs gradually and many editorial frames such as fading or dissolves are employed (Sorwar et al., 2002).

In the following sections, primarily, a set of the tightly relevant topics that mainly form the fundamental elements of *Video Summarisation* techniques will be reviewed.

2.1.1. Video Feature Extraction

In information retrieval theory, feature extraction is a form of diminishing the dimensions of textual data. The objective is to transform the input text document into a set of representative words (features) in significantly reduced dimensions, while relevant information of the entire document could be reflected optimally (Manning et al., 2009).

In the context of image processing, feature extraction can be defined as the process of retrieving basic visual characteristics from an input image without any prior knowledge regarding the shapes and subsequently presenting them in an assessable format (Nixon and Aguado, 2008).

In audio analysis, the attributes of a single segment from an audio signal which was portioned temporally, can be adopted as the representative features (Wen et al., 2012). Finally, in video content analysis, feature extraction is the task of mining and elicitation of expressive data from the available information resources inside the video (visual, textual and auditory).

Since each of these modalities could be considered as a valuable and distinctive data source, many researchers have been utilising them for feature extraction solely or collectively (Zhu and Zhou, 2003). The core idea of this task is to simplify the selection and classification of large data collections across all the noted fields (Choras, 2007).

In text analysis, the frequency of the words is the key factor in retrieving the representative words as the key features (Manning et al., 2009), however, a large set of aural features are being employed for audio processing purposes. Energy features (harmonic energy and noise energy), spectral shape features (roll-off frequency and Mel-Frequency Cepstral coefficient), temporal shape features (zero-crossing rate) and perceptual features (loudness and sharpness) are all used for information retrieval purposes (Peeters, 2004). Considering frames as the fundamental element of any video, visual feature extraction has received a higher level of attention from the video analysis communities. Visual characteristics in general can be computed at three distinct layers (Choras, 2007), including pixel, local (blocks) and global level (entire image). Colour, texture and shapes are just a group of these attributes being considered at these levels in image processing tasks. It is also necessary to point out that, based on their semantic-clarity, visual features can be divided into two groups of high-level and low-level. Low-level features can be gained directly from the original images, while

high-level features are constructed on the basis of captured low-level components (Saber and Teklap, 1998).

2.1.2. Video Segmentation

As was previously mentioned, a video shot is composed of a set of subsequent visually-approximate still images that have been taken by a single camera. In addition, a scene is a collection of semantically and visually related shots. Thus, as shown in Figure 2.2, shot boundary detection, also known as temporal video segmentation, is defined as the process of identifying transitions between adjacent video segments (Yuan et al., 2007; Wu and Xu, 2013). Considerable research has been allocated to this topic recently as it plays a crucial role in development of any further video processing tools (Li et al., 2009a).

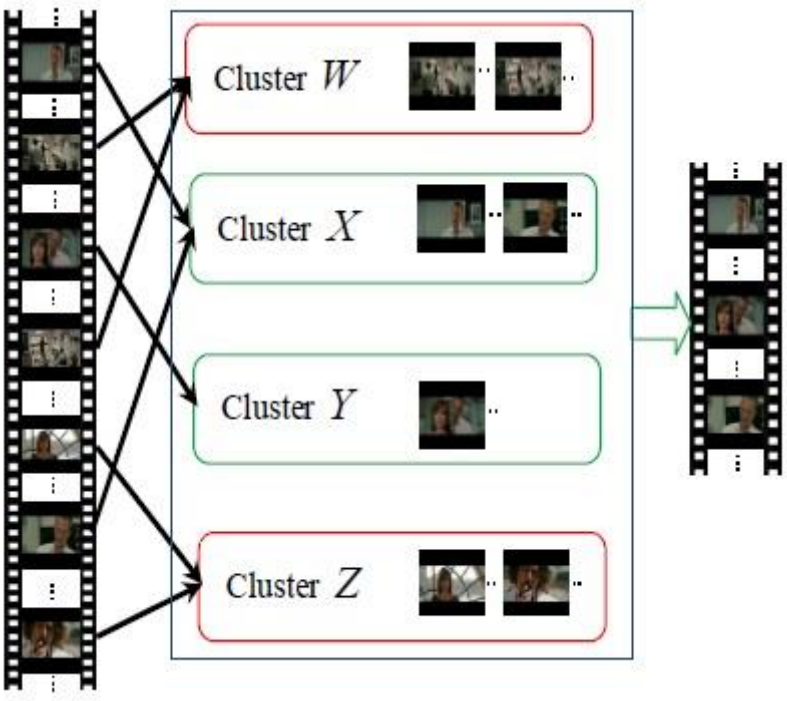


Figure 2.2.Shot boundary detection (Hari et al., 2013)

Video is constructed of multi-streams of auditory, visual and textual information. Therefore, various techniques using these modalities have been recommended in different studies to fulfil this objective. However, visual content has been the major resource for this task due to its potential to produce better outcomes (Dimitrova et al., 2002). In visually-oriented segmentation approaches, the main idea is to extract the features from the video frames and classify these images into different shots based on the determined differences of their

retrieved features (Wu and Xu, 2013). Pixel values, luminance/colour histograms, edge change ratio and motion information are just examples of the low-level visual characteristics that are being employed for these purposes (Li et al., 2009a).

In the most basic methods, the difference in pixel values (in different colour spaces) of two consecutive frames is calculated to check their visual dissimilarity. The timestamp for the pairs whose visual distinction exceeded a predefined threshold value could be potentially identified as a temporal location of a shot change (Boreczky and Rowe, 1996). However, these kinds of techniques are mainly used for abrupt change detection. This is due to the inefficiencies of these systems to identify progressive changes (Boreczky and Rowe, 1996). In a related attempt, the fact that consecutive frames in a short temporal segment of videos are often visually correlated has been the fundamental idea of Li et al. (2009). As a result, the video was portioned into segments with 21 frames initially and the pixel-wise distance (based on luminance component) of the first and last frame of each segment was computed. Thereafter, every ten segments were then grouped together to form a unit with a singular threshold value, which was measured adaptively in accordance with local and global statistics of distance values. Moreover, those segments exceeding the threshold figure were further analysed for detection of any possible transition.

In a more advanced work, SIFT (Scale Invariant Feature Transform) features were the basis for detection of progressive shots transitions (Li et al., 2009b). In the first stage, the colour histograms of two subsequent frames were compared to remove those that are not clearly shot boundaries. Later, the reliability similarity value of two frames were calculated using SIFT features. Thus, the number of the common key-points of SIFT between two frames were counted to determine the similarity of these two images. This was then utilised to detect abrupt shot changes. However, the progressive transitions could be identified by the variance transition of the number of the SIFT key-points in a frame.

The concept of video segmentation is not limited to shot boundary detection however. In fact, in accordance to the use case scenario, any number of frames whose total temporal length can be potentially longer or shorter than an actual shot could be considered as a video segment. As was discussed earlier, a collection of semantically and temporally correlated video shots construct a video scene as a higher level and a more meaningful video unit. Thus, scene boundary detection has become another prominent area of research in recent years.

In related research, auditory and visual features were integrated in video structure parsing task (Baek et al., 2005). An SVM (Support Vector Machine) model was applied to classify frames into different shots according to their luminance values in the HSV colour-domain. Further, a clustering algorithm based on visual and temporal proximity of shots was employed to form candidate scenes. These nominated scenes were further corrected based on audio analysis. Therefore, scene detection was carried out based on retrieved audio information assuming that most of the shots in a single scene should contain same background music.

In a very similar attempt, an audio segmentation model was developed to categorise the audio tracks into speech, music, environmental sound and silence segments for scene detection purposes (Jiang et al, 2000). In this work, an expanding window technique was used to cluster visually correlated shots into a candidate scene. The audio class change detection was performed on audio segments with intervals of one second. Finally, once a shot break and an audio break were detected simultaneously within those intervals, the boundary of that sequence of shots could be labelled as a scene boundary.

2.1.3. KeyFrame Extraction

KeyFrames are the semantically and visually valuable video frames that can be used to reflect the main elements of the video shots. Therefore, keyframe extraction, which is the procedure of identifying the most representative frame-set with the capability to represent the whole videos content in a precise and concise way, has been put forward by the research community. Most of the earlier methods selected the keyframes randomly or based on a certain time intervals. However, the chosen frames based on these types of methods are not necessarily capable of representing the video content at an acceptable level (Sun et al., 2008). Therefore, several methods have been recently proposed for keyframe extraction tasks, which mainly utilise mathematical concepts and low-level visual characteristics of the video frames (Liu, et al., 2003).

The low-level visual features of video frames have been the essential elements in keyframe selection tasks. For instance, an alpha-trimmed average histogram can be used to describe the colour distribution of an entire shot. Comparing the histogram of each frame of the shot with the average histogram, the keyframes can then be identified (Ferman and Tekalp, 2003).

In another similar study, three low-level visual features, namely, colour histogram, edge direction histogram and wavelet statistics were utilised to form the frames visual difference curve. Each two consecutive frames of a video shot were then compared based on the mentioned elements solely. The generated figures of their comparisons were further fused into a compound value representing the visual dissimilarity of all consecutive frames in a particular shot. Finally, the frames that are located in the middle of two local maxima points are selected as the representative frames for that particular shot (Ciocca and Schettini, 2006). However, most of the colour histogram based approaches do not capture the underlying dynamics when the level of camera or object motion is high (Sun et al., 2008). In further research entailing a computationally expensive work, the mean and variance of three colour components of each frame were analysed and used to detect any sudden or gradual change in the frames content (Qiang and Sen, 2006).

In another attempt, a hierarchical clustering algorithm was adopted to merge similar frames into a new category. This was carried out by applying multiple partitional clustering to all contributing frames of a video segment. In the final stage, keyframes were chosen as centroids of generated optimal clusters. However, there are some drawbacks with this method in terms of threshold determination and computation load (Hanjalic and Zhan, 1999).

In related work, a Perceived Motion Energy (PME) value was computed for each frame and the frames at the turning point of the motion acceleration or deceleration were chosen as the keyframes. The PME values could be computed by a combination of pre-calculated figures for average magnitude of motion vectors in the whole frame and the percentage of dominant motion direction (Liu et al., 2003). However, the triangle model for keyframe selection can only be applied to the shots with motion patterns, while for the shots with no pattern the first frame of the shot simply was chosen as the keyframe.

2.2. Video Summarisation:

The growing amount of multimedia content has imposed the need for development of systems, which are able to summarise videos of different genres in an effective way. This can be considered as significantly cost efficient, as they reduce the required space for the storage of the videos. Additionally, they have the potential to provide users with a more convenient and faster access to the content of the original video. Therefore, the primary objective of video abstraction should be regarded as an attempt to provide users with a quick idea about

the content of the input video by delivering them with a concise video representation (Furini, 2010). Consequently, a tremendous amount of research work has been allocated to this topic and various abstraction techniques have been developed. Generally, there are two main groups of summarisation approaches: in semi-automatic methods, as opposed to full-automatic ones, human involvement for abstraction purposes at some level is necessary. Nevertheless, an effective abstraction model should be able to consider multiple high-level or low-level factors such as sound, illumination of the scenes and psychological features of the human perceptual system (Iparraguirre and Delrieux, 2013).

Broadly, two basic types of video summaries exist, namely static keyframes abstracts and dynamic video skims (Truong and Verkatesh, 2007; Money and Agius, 2007). The earlier is also known as still image abstraction or static storyboard, while the second is also regarded as moving image abstraction or dynamic storyboard (Almeida et al., 2012). In the first type (Sujatha and Mudenagudi, 2011), the most informative frames are selected as the representatives for the whole video, while in video skimming, a short highlight of the original video is produced. In other words, video skims are composed of small dynamic partitions of audio and video which are semantically, visually and auditory valuable (Gao et al., 2009). Nevertheless, both approaches have to preserve the most salient and significant content of the videos in order to reflect a comprehensible description of the original video. As opposed to a static storyboard, dynamic video summarisation methods are more likely to provide the end-users with satisfactory results since they often have the capability to combine the auditory and moving visual elements. On the other hand, static techniques are more efficient and easier to develop (Oh, et al., 2004).

Generally, video summarisation approaches (as shown in Figure 2.3) comprise three major phases: firstly, video segmentation in which a system aims to detect video segment boundaries; secondly, feature extraction from the video portions, and, thirdly, selection of the most significant partitions using the retrieved features (Ren and Zhu, 2008).

Various techniques have been applied to distinguish the most significant segments of an input video. Mainly, the differentiation of low-level characteristics between adjacent frames or a holistic view over the entire video is adopted for this purpose. In both strategies, an importance score should be computed for each segment by analysing their various attributes, including visual, audio and textual features (Taskiran et al., 2006). These computed scores

are then used to rank the segment of videos and to select the most significant ones as a video digest.

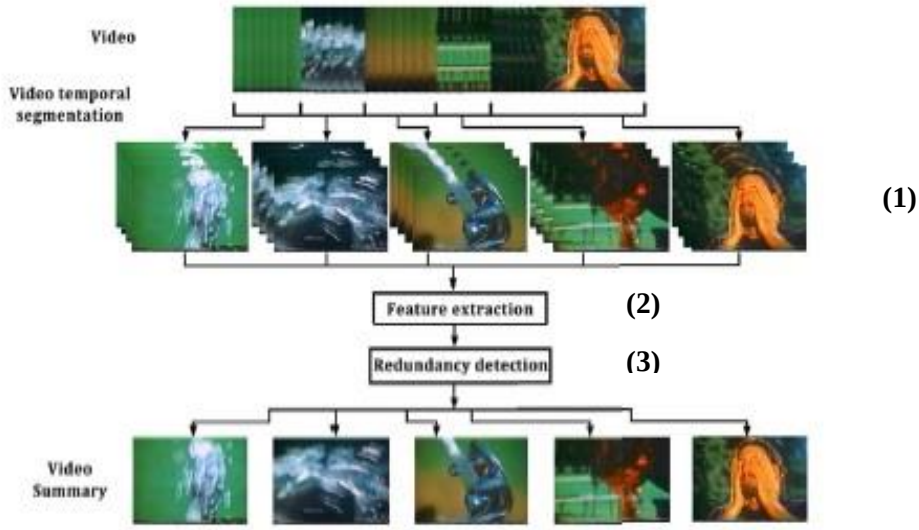


Figure 2.3. Video summarisation phases (Cahuina and Chavez, 2013)

2.3. Automatic Video Summarisation:

As a result of advanced audio-visual capturing tools, developing effective techniques to generate static and dynamic video skims is becoming increasingly popular (Ngo et al., 2005). In order to produce perfect summaries, some content-based summarisation approaches have been suggested to extract semantics of the video (Takahashi et al., 2005a). However, understanding the semantic content of the video in an acceptable rate is still beyond the capabilities of today's intelligent systems. Therefore, most of the current methods rely on low-level feature extraction (Iu et al., 2004), including colour histogram, edge histograms, textual and aural features (Guo et al., 2012).

Sequential clustering algorithms (Li et al., 2003), dynamic programming techniques like MINMAX and Iso-content (Datta et al., 2005; Hays and Efros, 2007) and motion patterns (Li et al., 2006) are among the approaches that have been employed alongside low-level feature extraction methods for video abstraction purposes. As was mentioned earlier, two main categories of summarisation approaches are being developed based on the human's level of involvement, namely, automatic and semi-automatic. These will now be further elaborated upon.

2.3.1. Summarisation Based on Low-Level Features

In some techniques, low-level visual or aural features alongside mathematical concepts like graphs and clustering have been adopted for summarisation purposes. In some works, different clustering algorithms were applied to partition the video frames into shots in order to produce static or dynamic video summaries. For instance, Yeung and Yeo (1997) proposed an approach in which, the shots were clustered based on their temporal adjacency, while Uchihashi et al. (1999) adopted YUV colour histograms to cluster the shots using the supervised clustering algorithms. In both methods however, representative frames from the highest scored segments were chosen using a frame packing algorithm. Additionally, in an attempt to use unsupervised clustering techniques, each video frame was labelled by a compressed chromaticity signature; a multi-level hierarchical clustering algorithm in conjunction with trained Hidden Markov Models were then used on the videos to extract the keyframes (Lu et al., 2001).

In another graph-based method, colour features and texture analysis were the basis for generation of static video abstracts. After pre-sampling the video frames to one frame per second, the shots boundaries were detected according to the consecutive frames' pairwise distances. As a result, the HSV colour histogram of each frame was computed for this purpose. In the next stage, shots with the size of one frame were eliminated as potential noise, while for the remaining shots the second frame was selected as the representative keyframe. Thereafter, the Discrete Haar Wavelet Transform was applied to the reduced HSV colour space of each representative keyframe to retrieve the texture features. Next, a reverse nearest neighbour graph was built utilising the Bhattacharya distance between the extracted features of the frames. Finally, all the frames that are mutually reachable were portioned into a video cluster and the initial frame of each cluster was chosen as delegate for that video segment (Mahmoud et al., 2013).

However, these clustering-based summarisation methods are not capable of understating the semantics of videos, as they cluster the frames solely based on their low-level visual features. As a result, there is a high possibility for this type of algorithms to accumulate the semantically irrelevant frames into a single group due to their visually similarities.

In a novel technique, the Bag-of-Importance model alongside Locality-Constrained linear Coding (LLC) was adopted for static video summarisation (Lu et al., 2014). In the first step, the LLC method was applied to convert the raw visual local descriptors into anchor points in

the transformed space. So, the similarity of the features could be computed simply by comparing their transformed codes. Later, the contribution level of each individual feature is assessed in the context of the video content and an individual frame. In fact, the importance of each weight could be computed from its distribution among all the features. Therefore, a video could be represented as the Bag-of-Importance which explains the relative frequency of transformed features over the entire corpus. Thereafter, the representative score of each frame is produced by aggregating the grades of important code-words for all the extracted interest points. Following the filtration of the most frequent terms (stop-words), a representative curve along the frames is built and the local maxima points are chosen as the final static summary.

In a recent work (Mei et al., 2014), the video summarisation problem was viewed as a sparse reconstruction problem in which all the existing frames in the original video could be recreated from a subset of them chosen as the keyframes. As a result, a $L_{2,0}$norm based sparse dictionary selection model was adopted to identify the representative frames. Simultaneous Orthogonal Matching Pursuit (SOMP as a typical Greedy Algorithm) (Tropp et al., 2006) was applied to solve the $L_{2,0}$ norm model. According to a similar work (Liu et al., 2014a), a dissimilarity-based sparse modelling method was suggested for generation of static summaries of user generated videos. In this work, smart device sensor data was utilised instead of frames visual features. Further, in a more recent attempt (Liu et al., 2014b) the collaborative sparse coding method alongside information captured by the accelerometer sensors embedded in smart phones were retrieved to enhance the performance by constricting the effects of the outliers. The larger the acceleration of smart phones, the smaller weight values was assigned to the corresponding frames. In spite of some noticeable results, these techniques solely rely on the low-level features of the video sequence without considering the semantics of the input content.

In a Genetic-Algorithm (GA) based summarisation technique, video abstraction was defined as a search problem in a space of all possible abstractions, where each video abstract could be represented as data point or in other words, chromosome (Ashwin-Raju and Velayutham, 2009). Thus, a GA algorithm started with a population of randomly generated chromosomes (a sorted list of randomly selected frame numbers in ascending format). The evolution procedure was carried out iteratively by selecting pairs of chromosomes and applying crossover and subsequent mutation operations to reproduce the next generation. Finally, the

fitness function was utilised to analyse each individual in the population. As a result, a simple colour histogram, the Gong colour histogram and colour correlogram measures were adopted to formulate the fitness function. In fact, both Euclidean and city-block distance measures between all the constituting frames of each chromosome were computed and the GA search was defined as a maximisation problem to insert those frames combination in the abstract that maximise the distance among them. However, this should be regarded as an extremely computationally expensive algorithm.

In another related attempt, web-images were utilised to facilitate the process of selecting the most informative frames from user-generated videos (Khosla et al., 2013). In the first stage, web-images related to each object class were clustered into 100 canonical viewpoints by K-means clustering and their decision boundaries were learned using a multi-class Support Vector Machine (SVM) over multiple iterations. Later, additional examples to each identified viewpoint were assigned from the collection of training video frames by applying the same procedure. Thereafter, for a given test video, each frame was assigned to one of the subclasses using the learned classifiers and an average decision score of the positive examples was calculated to rank the subclasses. Finally, to generate a summary with the length of K, the K frames from the original video that are closest to centroids of the top K ranked subclasses were chosen (shown in Figure 2.4). This algorithm managed to achieve some impressive results although its performance is tightly linked to the availability of a comprehensive collection of training images.

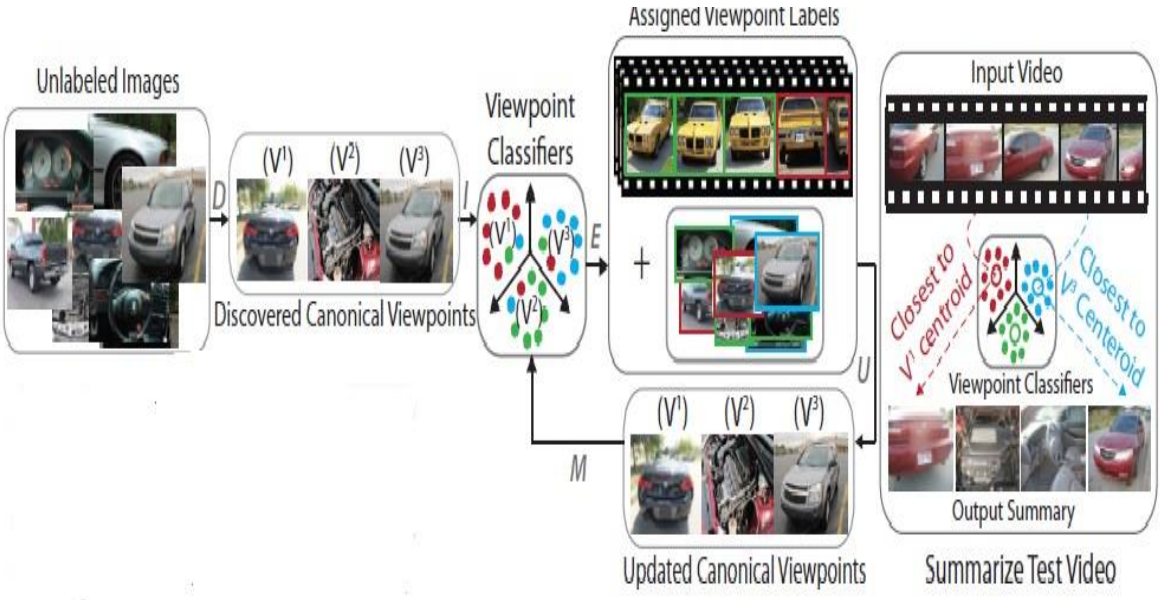


Figure 2.4. Video summarisation based on web-images (Khosla et al., 2013)

In another SVM-based work, videos with overlapping views were summarised into storyboards (Li et al, 2011a). As a result, a multi-keyframe correlation map was constructed to display the videos with overlapping views. Thus, any input video was initially parsed into content-approximate shots using the motion activity descriptors alongside optical flow and colour features. Discarding the lower-activity video shots, the remaining ones were assessed to be represented by a keyframe that had the least mean square difference of features set with the other contributing frames of that shot. Therefore, an importance score for each keyframe was computed by the fusion of visual features such as luminance, colour, edge, wavelet and gradient features and subsequent subtraction of the noise information (computed through subtracting Laplacian pyramid). Thereafter, a correlation map among the keyframes was created according to their temporal adjacency, visual similarities and their probabilistic correlation using the neighbouring frames' feature values. After the construction of the correlation map of keyframes from different views, the support vector machine was adopted to classify the event-based multi-keyframe on the map. Later, rough set theory was applied to identify the most crucial and significant frames in each identified class. However, the effectiveness of this method is tightly linked to availability of overlapping views of the same video content.

A fuzzy rule based approach has been the essence for another recent video abstraction research (Kapoor et al, 2013). Primarily, the video was partitioned into 10 frames segments and the first frame of each segment was extracted for further processing. Next, each retrieved frame was transformed into the CIELab colour space due to its proximity to human visual perception. In the next stage, the difference in luminance pixel values of any two subsequent selected frames was computed to detect any significant change. However, a dynamic membership function (dependent on the mean value of all consecutive segment differences) alongside a number of fuzzy rules were adopted to take into account the diversity in type and content of the video before the diagnosis of a significant change between two consecutive segments. Further, in order to detect the unique frames, a histogram analysis between the adjacent segments in the same colour space was performed and the fuzzified output was intersected with the results generated from the pixel level analysis to extract the final keyframes.

Heterogeneity Image Patch (HIP) is a recent image feature which was introduced in another work as a new concept for video abstraction task. This is a novel mechanism to evaluate the

heterogeneity of existing patches inside any photo. As a result, this metric is measured for all the frames in a video segment in order to generate the corresponding HIP map for that partition. The proposed index could be utilised to select a number of candidate keyframes from a large collection of frames using the HIP curve. These nominated frames were further filtered out into the final keyframe list by applying a min-max based algorithm to an affinity matrix. Additionally, the same method was employed for generating the dynamic video digests. In doing so, the video skimming task is mapped into an optimisation context by attempting to minimise the HIP-based distance of the retrieved video excerpts with the input video (Dang and Radha, 2014).

In another earlier graph-based attempt (Benini et al., 2007), after segmentation of the input videos into story units using a time-constrained clustering algorithm (Yeung and Yeo, 1996), each connected sub-graph was considered as a Logical Story Unit (LSU), where each node for these sub-graphs represents a cluster of visually and temporally approximate shots and their connecting edges show the temporal flow inside the LSU. Considering the role of motion activity as a measure of entropy in video segmentation, the motion vector field of all frames from the compressed MPEG-stream for each shot had to be calculated. Further, each individual visual concept (each node in LSU sub-graphs) can be assumed as a state in Hidden Markov Models where the transition states translate the temporal dependencies between the shots and available shots are existing observations. Considering the temporal length of each LSU in original video, a corresponding time-reduced match is then generated in the video skim in an attempt to provide an abstract representation of the whole video. Finally, the probability distribution (motion intensity of each shot in each state) and state transition probability (relative frequency of transitions between clusters) are adopted in order to identify and concatenate all the observed candidate shots and thus generate the summaries.

The Bag-of-Visual-Words approach was adopted in another attempt to produce static summaries (Cahuina and Chavez, 2013). Initially, the original video was temporally segmented into shots by detecting any abrupt changes in dissimilarity vectors of any two consecutive colour histograms (pre-computed for each frame). Later, the false positives were filtered out by comparing each value in the dissimilarity vector against the neighbourhood and giving importance to high values that are not around noisy areas. Further, the variance of each frame was used to detect and discard the monochrome frames that usually lead to dissolve effects. After clustering the identified shots based on their colour histograms information and identifying the sample frames, the Bag-of-Word approach was adopted to

create a visual word dictionary of detected feature vectors from the detected frames. Furthermore, a histogram of visual words was created for each representative frame by counting the occurrence of each visual word. Subsequently, the frames were clustered again according to their visual words histograms and those closest to the clusters centroids were chosen as the keyframes. Finally, in order to remove possible duplicates, a pairwise Manhattan distance of any two consecutive frames based on their colour histograms was calculated. In spite of some good results, since this algorithm requires multi-level clustering, it can be considered as considerably computationally expensive.

The texture and colour features have been the basis for still image abstraction in another recent work (Carvajal et al., 2014). After sub-sampling an input video into one frame per second, the colour-homogenous frames (with standard deviation below a threshold) that are mainly uninformative were discarded. The local texture information for each available block in the remaining frames was retrieved by applying the 2D discrete cosine transform. Later, a dictionary of texture features was trained by applying a k-means clustering algorithm on a set of training images. Therefore, a multi-dimensional histogram representing the relative frequency of each participating local texture feature per frame could be generated. After clustering the frames according to their fused texture and HSV colour (hue component) histograms, the closest frames to the cluster centroids were selected as the representative keyframes. Finally, visually similar frames (comparing the Euclidean difference of all keyframes pairwise) were filtered out. However, summarising the videos adopting only texture and colour features is not capable of extracting semantically significant content of the videos.

Low-level visual features and information theory were joined in another work for production of static video summaries (Jian et al., 2010). As a result, a two dimensional histogram was generated for each frame based on colour features in HSV space and on texture characteristics. This was utilised to quantify the visual similarity of successive frames. Using Mutual Information (MI) theory (which explains the relevance of two event sets), the extent of information delivered between two consecutive frames was measured to check their visual and semantic similarity. This is due to the essence of information theory which describes the information as the main object to study. This information propagation was characterised by detection of the change in RGB colour among consecutive frames. As a result, a matrix of changes for each colour component was created in which each element retains the probability of the change from one gray-scale value to another colour level between two subsequent

frames. The values generated from two patterns were fused into a singular inter-frame difference figure. Finally, a sliding window with a dynamically adaptive threshold value was employed through the statistics of the average inter-frame differences to identify the most salient ones. However, the level of transmitted information between two consecutive frames is not capable of determining the semantic importance of a particular frame in the context of whole video.

In related research (Iu et al., 2004), video structure analysis and graph optimisation were combined to generate the video summaries. Here, the shot boundary detection was carried out by measuring the similarity between consecutive slice images. A shot is an uninterrupted segment of video frame sequence from a single camera view. Accordingly, video shot groups were constructed based on visually similar and temporally adjacent video shots using H-S histogram correlation between the shots constituting keyframes. Intersecting the video shot groups, the video scene boundaries and skimming time could be determined. A scene is a series of coherent shots from a narrative point of view. Further, the video hierarchical structure had to be analysed in detail. Considering the different types of scenes (loop or progressive) and the desired skimming time, the corresponding summary length for each identified video scene is determined. For each scene, a full spatio-temporal graph is constructed, whereas the weights on the vertexes show the length of shots and the weights on the edges describe the dissimilarity functions between shots. This graph is built based on the shots keyframes colour histogram correlation and their temporal adjacency. Finally, in order to generate the video summaries, the path with the maximum summation value on the vertexes and closest aggregation scores on the edges to the pre-calculated skimming time for each scene is captured.

In another mathematical approach, a hierarchical video structure summarisation using a Laplacian Eigenmap was proposed (Jiang et al., 2009). Considering Laplacian Eigenmap as an efficient way of information extraction (Belkin and Niyogi, 2001), a reference frame subspace approach was applied to select a number of reference frames to measure the dissimilarity between any two frames. This is modelled based on the difference of their dissimilarity vectors (a vector representing the dissimilarity between an image and all of selected reference images) in Laplacian subspace. In the second phase, video structure analysis was carried out in three levels as following: scene level, shot level and sub-shot level. By using an adaptive threshold, video shot boundaries was determined and the middle frame of each shot was chosen as the representative frame. Subsequently, a K-mean

clustering algorithm was applied to the identified shots and the keyframe at the cluster centre is chosen as the representative scene frame. The sub-shot keyframes can be identified by comparing the dissimilarity of all frames within a shot.

In another clustering-based method (Zemcik et al., 2007), colour histogram and gradient distribution were used as the image descriptors for each frame. Each video is divided into two second sequences with 1.84s overlap. In the pre-processing stage any video sequence containing any unwanted frame (determined based on the set of unwanted descriptors) was discarded. Among the remaining shots, the concatenation of mean and standard deviation of the feature vector for existing frames in each shot was calculated to subsequently form the feature vector for that shot. Principal component analysis was applied to each shot to reduce the dimensionality and K-mean clustering algorithm was adopted to cluster the visually similar shots. Changing the K, the desired length of the video summary can be determined considering the equal length of each video shot. At the end, for each cluster, the nearest video shot to the centroid of the cluster was chosen as the representative for that cluster.

These types of algorithms are capable of considering only visual features of the current frames and a group of closest frames in order to estimate the level of attention and perception of viewers. Consequently, the computed peak points are at maximum level locally, while the corresponding frames might not be the most important in the context of whole video (You et al., 2009).

In a more complicated approach, a combination of graph model and clustering algorithm in multi-view video summarisation was adopted, which considers the content correlations between dynamic shots within each view and across multiple views. A hyper graph was then constructed to represent the different types of correlations (visual similarities, temporal adjacency and semantic relationship) between the shots in each view and across the different views, where each hyper-edge models a different type of correlation between the shots. The hyper-graph model was adopted to eliminate the side effects of inappropriate fusion weights selection. This hyper-graph was then transformed into a spatio-temporal graph in the next stage where the weights on the edges could be calculated by summing the weights of the hyper-edges that those shots belong to. Later a random walk clustering algorithm and multi-objective optimisation was utilised for summary generation (Guo et al., 2012).

Based on a collaborative approach, the results from different abstraction techniques were merged in furtherance to incorporate their individual strengths in the summary production

task (Dumont and Merialdo, 2008). According to this technique, a video sequence was first segmented adopting several methods including shot boundary detection based on colour histograms and hard cut detection based on SVM (Bailer et al., 2007). In the next phase the segmentation results retrieved from different methods were fused together. As a result, a clustering algorithm based on boundary time was used to categorise the closest boundaries. Further, a time boundary with the highest confidence value was selected for a video segment. In the last phase, each summarisation algorithm individually assessed the common segment to establish the extent of relevancy and redundancy. Relevant segments were identified using two methods: the first measured the visual activity and face detection results, while the second divided the original video into one second segments, clustered those segments and finally selected the most common partitions with capability to cover the maximum content. In order to determine the redundant video segments, colour bars identification, grouping several takes of one scene, as well as pattern models were applied. The fusion step then merged the different produced lists of relevant and redundant segments in order to produce the final selection list (Bailer et al., 2008).

In another clustering-oriented attempt, the proposed algorithm tried to extract a specific number of representative frames to generate the abstract of a particular digital video. As a result, a Content Based Adaptive Clustering (CBAC) was employed for this purpose. Unlike most of the common existing methods, the shot boundary detection was not carried out in this algorithm. Instead, video samples were projected as some points in the multi-dimensional characteristic space representing a group of low-level features such as colour, texture, motion and shape. The difference in their distances was assessed globally for a selection of representative frames. The time-based sequence of the video was mapped to a trajectory of points in the feature space. This spatial distribution of the points (video frames) corresponding to a video segment could be explained as clusters linked by abrupt or gradual changes. Moreover, the trajectory moved around in a small cluster. It is impossible for all the frames in a video to be spatially distributed far from each other and therefore to have irrelevant content, due to the nature of video in which frames are put together to express meaningful information. This characteristic of the distribution of points provides an essential basis for this clustering technique. As a result, frames are divided into equal units and then the difference between the first and the last frame in the unit had to be calculated in order to partition the frames into two different sets based on the level of change (large-change and small-change). The frames from large-change clusters were retained, while all those

belonging to small-change clusters were omitted except the first and last frames. If the total number of remaining frames was enough then these representative frames were put together to constitute representative sequences, which could be used for temporal summarisation of video. Otherwise, this clustering process should be reiterated several times till the required number of frames could be achieved (Sun and Kankanhalli, 2000).

2.3.2. Video Summarisation by Various Modalities

In most of the methods that were described earlier, some good results were achieved but a prefect video summary can only be produced by extraction of semantic content of video. Since the previously mentioned methods are highly tied to low-level visual features of videos, they are unlikely to fully reflect the semantic content of the videos. Thus, a different research strand involves other modalities (in addition to Visual content) in the summarisation process as a potential information source.

2.3.2.1. Audio Data

Accordingly, Bhatt et al (2009) adopted auditory features solely in an attempt to generate dynamic video skims. After portioning the input audio into one second segments and removing DC component from all partitions, each section was further divided into frames with the length of 320 audio samples (20 msecs). Later, each segment was initially tested for silence or environmental noise, speech, music, and music with speech. Primarily, silence regions detection was carried out by measuring short time energy of each segment through aggregating the sum of squares of the signal samples. Segments with short time energy below a predefined threshold were identified as silent segments. Further, non-silent partitions were tested for environmental noise using short time entropy and the modified autocorrelation peak values. Non-environmental-noise audio segments were further assessed for detection of speech only versus non-speech (further to music only and music with speech) sounds using a number of auditory features including low short time energy ratio, Mel-Frequency Cepstrum Coefficients (MFCC) and variance of log energy. A Gaussian Mixture Model (GMM) and Fuzzy decision trees were adopted for training purposes. Finally, based on the identified category for each audio segment, video abstracts in accordance to that particular video genre were generated. However, the proposed algorithm can potentially fail to include many of the visually and semantically rich video content into the final summary due to silence of its corresponding video segment.

2.3.2.2. Audio-Visual Data

Audio analysis was the basis for another multi-modal technique in which keyframes were selected based on semantic analysis of shots, scenes and frames in a holistic structure (You et al., 2009). A video was preliminary segmented into scenes using audio features assuming the prolonged consistency of the audio track of a scene in terms of signal characteristics. Classification on the non-silent clips of audio was performed to fit each clip in one of five existing genres. Later, these scenes were segmented into shots using the luminance histogram. All audio clips were weighted based on the class which they belong to, and the computed average score of all clips in a scene was measured as the semantic audio importance of that scene. All the representative histograms of a scene are then compared in order to classify the shots into two groups of related and unrelated shots. A shorter scene with more unrelated content is better. The size of a face or text together with its region in the frame was adopted to produce the importance index for that frame. Additionally, the number of occurrences of detected faces or text in a single scene could generate the text and face saliency value for that particular scene. Affective features (pitch, loudness, motion speed and luminance) in one scene were measured and then compared to the whole sequence to show the level of semantic relevance of that scene in regards to the overall sequence. Shots were further semantically measured using the above semantic audio importance and face and text importance as well. Hence, other factors including camera motion, object motion, temporal motion coherence were all taken into account to build a semantic shot importance model. However, the existing video processing techniques for face and text recognition purposes still suffer shortcomings in terms of accuracy and scalability, which can directly affect the performance of the explained approach. Furthermore, the results produced by this approach are highly dependent on the audio-visual quality and noise level. For instance, a noisy audio environment or cluttered scenes can undermine the accuracy and performance of face recognition systems (Herranz and Martinez, 2008).

According to another multimodal technique based on audio-visual features, colour, motion and MFCC features of the audio signal were all analysed to generate the video abstracts (Jiang et al., 2000). Initially, the entire video was segmented into a number of one second length temporary partitions and colour histograms were calculated for each frame. The produced histograms were then averaged over a segment to produce a reference histogram for that partition. Subsequently, motion features were computed for each frame using the SIFT

algorithm and the Euclidean distance between these features were used in the computation of a singular motion vector for each video segment. For auditory analysis, MFCC features were calculated for each video segment. However, considering the vulnerability of this type of features against noisy conditions, tensor subspace analysis was adopted to extract audio characteristics for one second audio frames. Afterwards, the dynamic time warping algorithm was used to calculate the similarity measure between two audio segments. In order to perform segmentation, a dissimilarity matrix was computed, in which each element represented the pairwise distinction between two segments. The calculated values of colour, motion and sound for each segment were further normalised and adaptively weighted in order to be fused into a single value. In the final stage, a Fuzzy C-Means clustering algorithm alongside a maximum likelihood estimation approach was employed to cluster the video segments in an optimal manner. Finally, the video segments closest to the centroids of the clusters were extracted to be inserted into video summary.

2.3.2.3. Audio-Visual-Textual Data

In another multimodal summarisation technique, the saliency of auditory, visual and textual information was analysed separately and then integrated into a multi-modal saliency curve (Evangelopoulos et al., 2009; 2013). For audio saliency detection task, the primary objective was to build a data-driven and time-dependent function with capability to change in accordance to the importance level of auditory sensory information. Therefore, the audio frames were decomposed into a set of equally separated frequency bands (frequency components) and each band was modelled by an AM-FM signal. Gabor filters were further utilised to perform band-pass filtering, while the Teager-Kaiser energy operator and energy separation algorithm were all adopted to decompose each signal into instantaneous energy, amplitude and frequency signals. However, only one frequency component, which dominates the signal spectrum, was employed as a dominant modulation component (the one which produced the maximum energy response over the time frame) and provided the basis for yielding a feature vector comprising instant amplitude, frequency and source energy respectively. Then, each feature was normalised over a long-term window to scalar values that sum to one and the results formed a one-dimensional temporal saliency map. For visual analysis, the frames pixels were considered as the voxels whose saliency was analysed based on their intra-feature, inter-scale and inter-feature interactions. Each frame as a volume was decomposed into 3 conspicuity volumes (intensity, colour and orientation), after which each

volume was further decomposed into multiple scales representing a Gaussian volume pyramid. Intensities were then calculated based on the difference between RGB value of a point and the average value of the surrounding region; colour opponent theory was then used to generate a colour conspicuity score. Finally, orientation was computed employing spatiotemporal steerable filters adjusted to respond to a moving stimulus. Consequently, the outcome was a set of updated multi-scale volumes; the saliency for each point is the average of all volumes over all features and scales. At the end, a single saliency value for each frame was generated by multiplying the normalised feature scores with the calculated saliency value from the last step. In order to evaluate the textual content, forced segmentation on the audio stream was performed using speech transcripts generated by a Sonic ASR (automatic speech recognition) system and phone-based acoustic models as the pre-processing stage. The timestamps inside the provided subtitles can present the rough location of the text in an audio stream and were useful to start the forced segmentation procedure. Then, time-aligned transcripts were analysed using a decision-tree-based probabilistic tagger to carry out part of speech (POS) tagging and the highest scored POS tags were assigned respectively to proper nouns, common nouns, noun phrases, adjectives, verbs and the remaining parts of speech. Therefore, each frame could be scored based on its textual saliency. In the last step, the produced outcomes from different modalities were integrated to produce a single, composite saliency curve. Thus, an intra-model fusion scheme was adopted in which each individual saliency feature was normalised to a value [0,1] and weighted based on its variance. The most salient audio and video sub-clips based on a predefined skimming percentage were then chosen for inclusion in final summary. The proposed approach can produce some impressive results for a number of video categories. However, its performance degrades when the fluctuation in aural or visual features remains at a minimum over the course of video.

2.3.3. Domain-specific Video Summarisation

Another category of video summarisation techniques is that of domain-specific methods with capability to generate summaries for particular video genres by utilising the exclusive features and attributes available in those categories. These methods are frequently being employed in summarisation of sport videos and are mainly based on the fusion of low-level and object level features in order to identify the most valuable events (Ekin et al., 2003; Zhang and Chang, 2002).

In earlier work, the low-level features of the video were analysed for event-detection purposes, while more recently studies employ ontology based approaches (Bertini et al., 2005). Accordingly, a formal ontology reasoning approach was proposed to produce semantic abstraction of sport videos (Ouyang and Liu, 2013). As a result, sport videos are annotated with ontologies in order to build a three-level hierarchy sports abstraction (keyframe, representative shots and video clips). In order to build the required knowledge infrastructure for semantic analysis, the sports video model was divided into an upper ontology and a domain-specific ontology. While the first one represented the general features of basic sport videos, the second was adopted to depict the details of general concepts. An XML scheme was utilised for describing and reasoning of the video ontology. An interactive keyframe selection technique was adopted to generate static video abstracts. While the semantic information of shots and keyframes was obtained directly through the users' annotations, the semantic results for the representative scenes could be gained from the inference engine. This proposed algorithm requires a great extent of user involvement which can potentially affect its scalability.

In a proposed approach to summarise documentary movies, the generated summaries were represented in the format of a set of contiguous audio-visual segments that were homogeneous in a cross media space (Perez-Daniel et al., 2014). Adopting the Data Cube concept (Gray et al., 1997), several partitions of the same data set could be generated by employing various possible combinations of the audio-visual features space. In order to describe the visual features, a number of colour-based MPEG7 features including Scalable Colour Descriptor and Colour Structure Descriptor alongside a texture-based (representing a pyramid of blocks with the histogram of oriented gradients) feature were adopted. In addition, MFCC and chroma vectors were utilised to denote the auditory information. As a result, a consensus clustering algorithm with the capability of incorporating various combinations of dimensions of the description space was utilised to build such partitions. A consensus clustering is a procedure to merge agreements over several clustering on a similar data set with different dimensions. The median frame of each cluster was chosen to be inserted into the summary. Despite some considerable outcomes, the practicality of this algorithm is linked to the availability of MPEG7 data. Furthermore, presence of aural noise can increasingly deteriorate the quality of final summary.

In contrast to visual methods, there has been an attempt to summarise sport videos using the audio features, considering the fact that interesting events can lead to changes in the speech

excitement level (Otsuka et al., 2006). Accordingly, the percentage of excited speech in each audio segment is calculated alongside its energy level enabling the system to compute the importance level of each video segment.

Interestingly, in a combination model (Taskiran et al., 2006), the textual content of movies alongside its audio characteristics were both used for video abstraction. Using a speech recognition system, transcripts of the video are retrieved and subsequently an inverted word index alongside a phrase glossary index is created. In this system, it is audio pauses instead of shot boundaries which are used for segmentation purposes of the video. The importance score for each video segment is computed by applying information retrieval techniques. Each video segment is considered as a document and term frequencies within segments as well as the distribution of pairs of words within it could both potentially determine the importance of each segment. However, this type of summarisation technique does not generate satisfactory results when speech signals are noisy (Ngo et al., 2005). Moreover, this proposed algorithm is not applicable to silent videos.

As opposed to audio-visual oriented techniques, in a text-based approach sport video events are detected by analysing and alignment of webcast text and broadcast video (Xu et al., 2008). After filtering out the stop words and names of players, a probabilistic latent semantic analysis is applied to cluster the webcast text into different categories. Later, words with the highest number of occurrence in each category are chosen as keywords to represent the event types. Sentences containing these keywords are text events. In order to synchronise the webcast text and corresponding event in the video, a conditional random field model algorithm is employed to detect the start and end boundary of the event. However, the proposed algorithm can only function in presence of webcast data.

In contrast to the previous method, a visual-oriented approach was proposed for football video summarisation using an improved algorithm for the detection of replay shots. Shot boundary segmentation was carried out by detection of differences in the dominant colour pixel ratios and colour histograms. In the next phase, the shots were fed into the event detection engine to be examined for identification of the logos (TV logos are recently being adopted as a visual-effect before showing the slow-motion shots), score board, Goal-Mouth and shot classification. Finally, a rule-based classifier was used for interesting events detection (Eldib et al., 2009). Nonetheless, high quality video summaries could be generated adopting this method only in the presence of carefully developed replay shots.

In another domain-specific approach, a summarisation method for a basketball game was proposed based on monitoring the temporal changes in the score. For this purpose, a scoreboard region detection method was used and a text area detection algorithm was applied to identify the areas of an image with many vertical strokes. Only the regions which remained static for a second were then chosen as candidates for scoreboard (Kim at al., 2005). In the next step, a number recognition algorithm was applied to the filtered result in an attempt to determine the score regions. Simultaneously, the video shots were classified into play shots and non-play shots based on the ratio of dominant coloured pixels. Finally, by defining some semantic templates for exciting scenes, the importance score of changing score frames could be computed and important shots were included into the video digest. Unsurprisingly, the availability of scoreboard in the original video is a prerequisite for good performance of this algorithm.

Motion features at different video levels was the basis for a proposed framework to summarise the surveillance videos (Sujatha et al., 2014). Initially, the original video was divided into a number of blocks, each containing a non-uniform number of segments. The optical flow at frame level was computed and was further propagated to the segments and blocks levels. The frame motion was derived from the overall motion of the existing feature points in that particular frame. Those are the points where strong derivatives were observed in two orthogonal directions. Later, the motion entropy for each block is obtained by computing the probability of possible motion in a segment and therefore, the most salient blocks with the highest motion activity can then be extracted for the final summary.

2.3.4. Online Summarisation methods

As opposed to most of the summarisation methods which operate in an offline manner, only a limited number of studies have explored the topic of online video summarisation. This is a result of difficulties in producing summaries in real-time based on impartial information (Almeida et al., 2013; Zeng et al., 2011). However, in many applications, such as video sensor networks, capturing the entire video before summary generation is inefficient due to the limitations in terms of time and memory resources (Ou, 2014). As a result, in some recent methods, the summarisation engine operates directly with the video stream in real time without a need to assess the entire video sequence.

For instance, in related research, online videos were summarised progressively as they are recorded (Ou et al., 2014). In the first stage, the MPEG-7 colour layout descriptor as a feature was extracted for each input frame. Later, the extracted feature was clustered using an online GMM based clustering (Stauffer and Grimson, 1999). After the clustering, the decision regarding the inclusion or discarding the input frames were made based on the frame corresponding cluster weight. As a result, the frames belonging to the cluster with large weights (more frequent content) and small variance (lower activity level in that cluster) were filtered out (as shown in Figure 2.6).

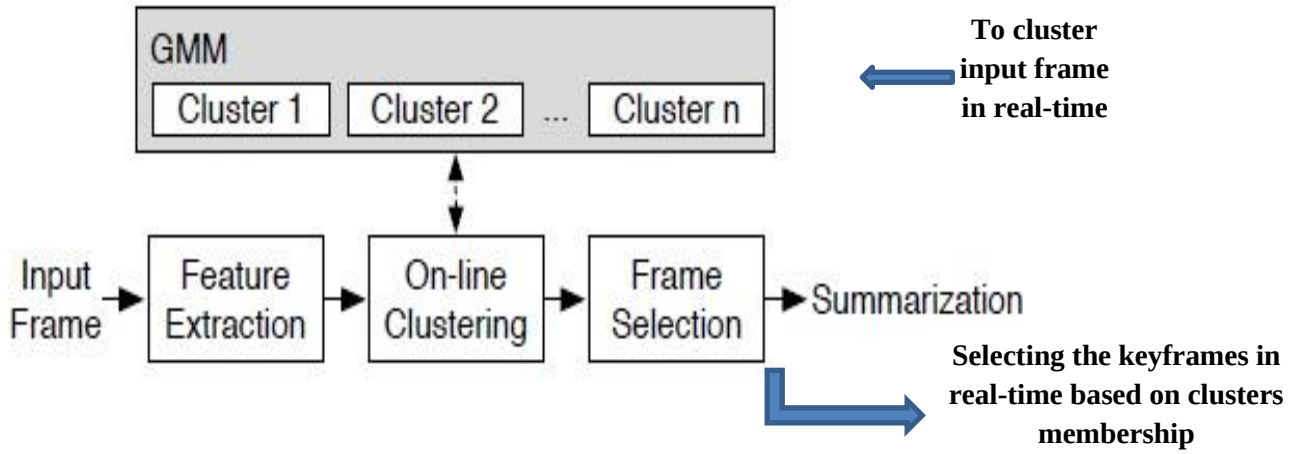


Figure 2.5. Online video summarisation using GMM (adopted from Ou et al., 2014)

Other of this type of methods is the work of Valdes and Martinez (2008), who employed a decision tree as the basis for the generation of online video summaries. Firstly, the original videos were divided into equal-length shots. Later, a binary decision tree whose nodes represent the shot condition (inclusion or exclusion) and each branch explains a result of summarisation was developed. This was a binary tree which models all the possible video summaries that could be generated from an original video considered as collection of *n* basic units. Thereafter, a score demonstrating the quality of the video summary is calculated. In order to minimise computation costs, splitting the trees into partial sub-trees was undertaken and the score for each branch was computed considering three major weighted factors: shot continuity, activity and redundancy. Finally, online tree pruning using a time sliding window was performed so as to generate the real-time abstracts.

Tracking the local features among the consecutive video frames was used in another technique to summarise the videos in real time (Iparraguirre and Delrieux, 2013). Here, the

number of similar Speeded Up Robust Features (SURF) visual features between any two consecutive frames was defined as a Common Feature Number (CFN). In order to identify any potential keyframe, this absolute value was compared against the average of 10 prior CFNs to calculate their change ratio. Any value above a pre-set threshold will nominate the second frame of the pair as a candidate keyframe. The nominated keyframe later was checked against the last detected keyframe in a similar way, for noise amount evaluation. Finally, in order to produce a video skim, after detection of a keyframe, the 30 following frames were also added to the final abstract. However, such an algorithm can only evaluate the saliency of video segments locally.

2.4. Semi-Automatic Video Summarisation:

As was explained earlier, these types of techniques require human intervention at some stage. This will result in bridging the semantic gap between the low-level features and the human's perceived responses (Han et al., 2014). Thus, the probability of generating semantically stronger video abstracts will be boosted. Here, the user's spontaneous behaviour while viewing the videos is captured in order to generate video summaries. These systems can thus potentially determine the importance level of different shots by measuring camera motion level plus movement of eyes and facial expressions of users, while they are interacting with the video segments (Yoshitaka and Sawada, 2012).

Users' high level collaborative activities such as textual tags can also be used to generate video digests. For instance, user-created video bookmarks can be employed for video abstraction (Chung et al., 2011). The proposed model system consists of two correlated modules: a bookmark server and a bookmark analyser (video summary generator). The database (bookmark server) contained several queues, each of them maintaining the different bookmarks created for the same video. As time goes forward, this database keeps on growing till it stabilises. At this point, the bookmark analyser attempts to construct a bookmark histogram by dividing a video into a number of equal time intervals and trying to assign each of those weighted bookmarks to one of the existing bins (weights for older bookmarks are decreased exponentially). These histograms are then smoothed using Gaussian filters and their peaks are identified. In the last stage, homogeneous (location and visual) peaks are merged to generate the final summary. However, this type of method is extremely costly and time consuming.

In another related work (Ngo et al., 2005) a video is initially partitioned into shots and a similarity graph produced where the weights of the edges connecting the shots reflect the visual similarity and temporal distance of constituting frames. Later, a normalised cut algorithm is used recursively to decompose these shots into sub-graphs (clusters). As a result, the scene boundaries are determined by segmenting a graph into sub-graphs, each correlating to a scene. Omitting the edges along the shortest path (using Dijkstra's algorithm) from the cluster containing the first shot in a video to the cluster that contains its last shot, these sub-graphs could be identified. In the next phase, the motion attention mode is employed to measure the attention level of users when watching the videos. The prior probability and attention value for each cluster is computed to define the quality of the scenes.

In the Click2SMRY framework, crowdsourcing was adopted as the basis for video summarisation (Wu et al., 2011). Here, each video was partitioned into equally-sized sub-segments (5 seconds each) and thereafter video workers were asked to identify potential video highlights by holding the SPACE key on the keyboard while they were watching the original videos. Therefore, each click was assigned to one corresponding sub-segment. Finally, based on the required length of summary, a number of these sub-segments with the highest selection rates by different workers was extracted to be inserted into the final summary. However, segmentation of a video shot solely based on the time element can increase the possibility of generating false results. This is due to the inability of this method to address the dramatic change in the visual and semantic content within each sub-segment.

In a very recent study, the two fields of brain imaging and visual attention modelling were utilised to produce semantically rich video abstracts (Han et al., 2014) as shown in Figure 2.5. Accordingly, the Functional Magnetic Resonance Imaging (FMRI) technique was adopted to identify and monitor the main brain areas involved in visual information perception and cognition called Regions of Interests (RoIs). Then, the attentional engagement of brain to different video content stimuli for generation of benchmark attention curve was measured (using a spectral graph representing RoIs interactions). As a result, an fMRI-driven visual attention model with the capability to optimise the low-level features combination under the supervision of a smaller training FMRI data was presented. The optimised attentional model could increase the correlations between the low-level visual features and brain responses. Once the FMRI-driven attention model was learned at the training stage, the identified patterns could be generalised for summarising any new input video at the application stage.

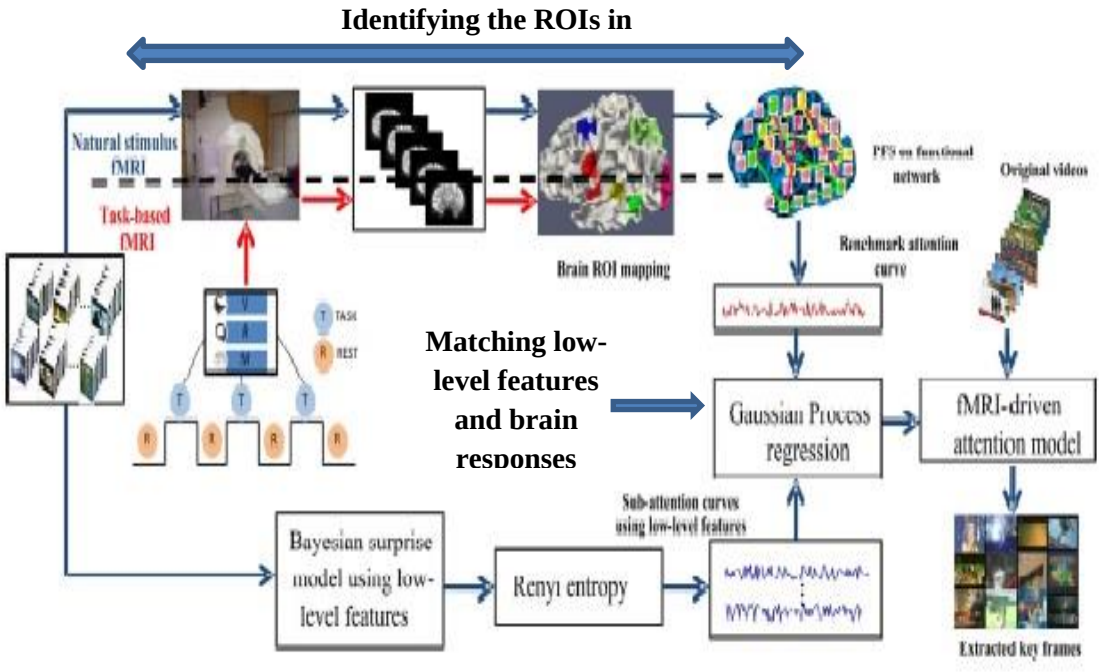


Figure 2.6. Video summarisation using FMRI method (adopted from Han et al., 2014)

2.4. Multimedia Personalisation

Personalisation has been an area of research interest in the computer science community during recent years. Capturing the user's interests and modifying the output in a way that meets the user's requirements in the best possible way is the main objective of personalisation systems. Personalisation has been adopted in different areas of computer science including information retrieval systems, video portal systems and e-commerce websites. In all mentioned areas, the personalisation module within the system tries to acquire useful information regarding users and usage environment. For instance, Fukumura et al. (2003) detailed how a personalised website could be developed that presents the digital content based on the user's browsing history. In this system, it is assumed that a website consists of three components: 1) Content 2) Container 3) Relationship; therefore, the browsing history of a user is a sequence of components that a visitor had followed earlier. A component extraction algorithm was then used in later stages to choose the most valuable elements for the presentation.

A user's generated sessions can also be used for personalising the output (Mobasher et al., 2000). The offline component of the system carried out on two tasks. In the data preparation

phase, a session file of viewed pages by each user containing attributes about each page-view was generated. Subsequently, the output was fed into a data mining algorithm for pattern discovery purposes. In the last phase, the recommendation engine considers the conjunction of active server sessions and discovered patterns to construct the personalised content.

Unsurprisingly, the core idea of personalisation in the context of multimedia is similar. In fact, the most effective and impressive approach to provide users with a fast and convenient access to any multimedia content is to integrate their preferences and the characteristics of the environment around them into their queries (Caschera and D'Ulizia, 2007). Generally, these features can be divided into five categories (Skondras et al., 2011): a) users' general information; b) users' preferences, representing the user browsing, filtering and search habits; c) usage history, describing a user's interaction with digital items; d) presentation preferences, showing their favourite multimedia presentation means; and e) accessibility characteristics, concerning the end-users with audio and visual limitations.

Personalising multimedia content is an extensive process, which in most cases is achieved in three consecutive steps: semantic and structural information extraction from the available resources (including the original files or supplemented metadata); creating profiles for end-users automatically or manually according to their priorities; filtering the content based on user's profiles to fit user's preferences. These will now be looked at in more detail.

2.4.1. Information Extraction and Representation

Information extraction and semantic annotation of media content can be done both manually and automatically. Multimedia documents can be enriched and annotated with metadata expressing the (i) semantic and (ii) affective/emotional/rhetoric content of any multimedia document (Ren and Zhu, 2008). While semantic annotations determine the conceptual content of a document, the affective annotations are applied to express their perspective. For instance, a multimedia document can represent an attitude or a bias in their content by virtue of language, movement, juxtaposition, colour, rhythm, etc. Affective annotations provide information like "video sequence with dramatic presentation", "expresses negative opinion vehemently", "evocative music sequence", "visually stimulating picture" (Ren and Zhu, 2008). There are a number of different elements that exists in a multimedia content and should be extracted and be presented in a content model. The most important items which should be included in the content model are: objects in the media stream and their clear

properties; the spatial relationship between those objects; Events in the video segments involving those objects and, lastly, the temporal relationship between those video segments (Angelides, 2008).

In many systems a manual scheme to enrich the media with metadata using different tools and standards is being adopted. For instance, in a multimedia content sharing system the MPEG-7 and MPEG-21 standards could be used to represent the content framework. The i2CAT Machine Project was the continuance of a project called Integrated Project in which the main objective was to define an advanced environment for sharing media content. The MM&P (Metadata Management and Personalisation) module embedded in this system was responsible for storing and handling the metadata. In this module, the DI (Digital Item) structure was created by utilising MPEG-21. The DI was the main element which was defined by the MPEG-21 standard to project the structure of content and bind descriptions to them. A DI structure comprises three types of elements, namely, container, items and components. These DI elements were bound to thumbnail, semantic (title, creators, genre, etc.) and technical (bit rate and file format) descriptions, which were defined by the MPEG-7 standard to constitute the machine project metadata (Rovira et al., 2007). However annotating media content manually can be extremely time-consuming.

2.4.2. Profiling

Creating a user profile is an approach to capture the users' evolving information needs. User preferences can be captured implicitly or explicitly for creation of their profiles. The profiling of viewers is continuously carried out by the systems and is usually created based on all available user interactions, such as user selected media streams, navigation patterns and social networking (Malheiro et al, 2011). However, there are potential drawbacks to a user's profile identification, such as the variety in the users' interests, which can lead to a sparse data representation that should be addressed.

Until now, a number of methods for creating user profiles have been suggested by researchers. In general, in most of these techniques, a model to translate the users' implicit information resources to their unique profiles was suggested. In one related work, a general profiling model for personalisation of multimedia content was used, in which the user profiles, context profiles and their combination were all employed in an attempt to make the original user query more precise and to produce a better outcome. In this method, spatial,

temporal, semantic and structural characteristics were utilised to describe user, context and multimedia data. Therefore, user profiles, context profiles and data profiles were built using the available descriptions. These profiles were further applied to filter a user query in order to narrow the results to those which optimally meet the users' expectations (Caschera, M.C. and D'Ulizia, 2007).

Several user and context adoption methods have also been presented in the existing literature. For instance, ontology-oriented techniques can be adopted to model a user profile (Beckett, 2001). One of these is UBIcomp, which uses ontologies to represent contextual data in regards to users (Christopoulou et al., 2005). While in an ontology-based project called Smart Push, professional editors were asked to enrich user's information with semantic metadata (Jokela et al. 1999). In similar work, user profiles were created by developing knowledge graphs to model the correlation between various concepts in the Linked Open Data Cloud where concepts with similar semantic context are connected to each other (Hanckok and Walker, 1992). Based on another related attempt, user profiles were designed to store the users' personal and semantic data alongside their tag cloud, resulting from their social interactions on the web or from previous selections of video streams (Malheiro et al., 2011).

In another implicit data extraction method, relevance feedback was used as the essence for creating user profiles (Hopfgartner et al, 2010). When a particular user interacts with a result, he/she leaves a semantic finger print that represents the extent that the corresponding content is interesting to him/her. The described system applied a weighted story vector approach to acquire this finger print and to update the weight of that story. As each news story includes one or more broad categories, which are identified beforehand, a user's interest in those categories can be measured by combining the acquired weights of the different stories belonging to that category over various user iterations. By performing hierarchical clustering and extracting transcripts from clustered stories and utilising information retrieval techniques, sub-categories can then be generated.

The user profiling task can also be carried out in both explicit and implicit manners within the same solution. For instance, the goal of the EignNews system (Daneshi et al., 2013) was to provide end-users with a personalised playlist of news videos. Accordingly, explicit preferences, personal information, as well as implicit preferences were gathered and fused into a singular value to clarify the end-users' priorities towards different news stories.

Previously assigned scores by each end-user to a list of pre-defined news categories and their viewing history of other related news videos were then used to form such a priority grade.

Data representation of created profiles is another essential topic that should be taken into consideration. For instance, in a p2p personalisation system, a user-profile was viewed as a mapping of users and multimedia tags to a set of interest weights (Nalin et al., 2011). In this work, the interest weight of a multimedia tag was expressed in a numerical value format reflecting the user's level of interest in that particular item. These could then be integrated into the user profiles which are mainly represented as a vector of keywords and weights. These vector-style profiles have the potential to be represented in a Bag-of-Word format in order to calculate peer-users' similarities. As a result, selection of an appropriate data representation model for generated profiles can facilitate further operations such as filtering, which we now proceed to describe.

2.4.3. Filtering:

To date, a group of different methods has been employed for content filtering purposes. In content-based filtering, recommendations will be made to users based on the content similarity to (implicitly or explicitly) the obtained user's profile and previous recommendations as shown in Figure 2.7.

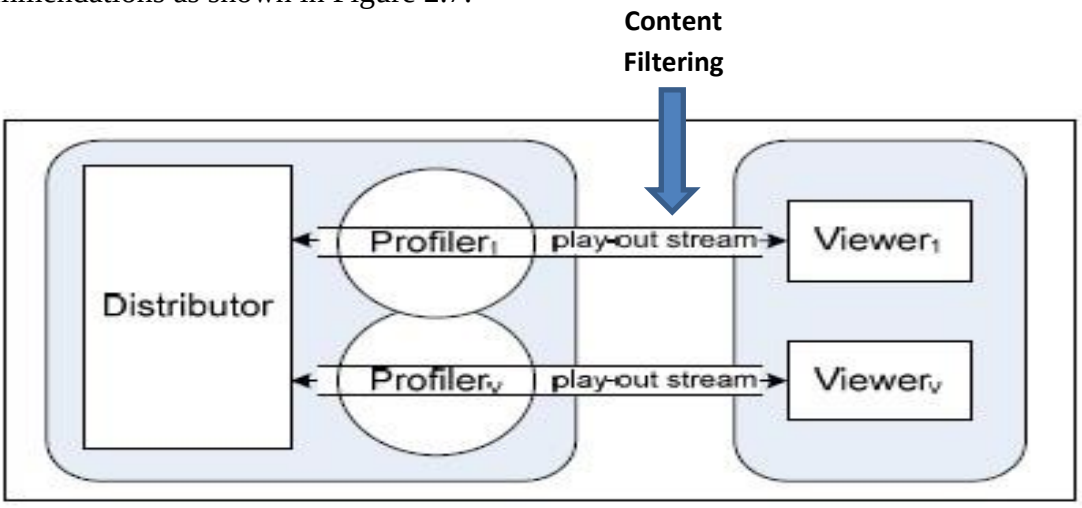


Figure 2.7. Content filtering personalisation (adopted from Malheiro et al., 2011)

This contrasts with manual rule-based systems, where rules derived from user demographics and static profiles (collected through a registration process), determine the content for the users. Furthermore, in collaborative filtering a user is assigned to a group in which other

members' content ratings (concerning their profiles) are used to retrieve filtered contents (Mobasher et al., 2000). Additionally, in collaborative tagging, users are also allowed to enrich contents by means of tags and to share such descriptions. The structures captured from these data sources can then be utilised for recommendation improvement (Malheiro et al, 2011).

In many systems, relevance feedback approaches are currently being used as a method to filter multimedia content according to users' interests. For instance, a group of candidate images could be presented to users who are then asked whether the shown images are relevant or irrelevant. The main concept is to develop an interactive system with the capability to save and process the user's interactions with the system's recommendations. As a result, the decision making process can be adjusted dynamically rather than using a set of pre-defined formulas (Patrikakis et al., 2011).

A different approach is that taken in a multimedia retrieval system called PIDALION, in which the user's querying and clicking behaviours were analysed to generate feedback. By keeping track of their choices and applying them to future searches, the users could find their content of interest easily. This system tried to monitor the user's behaviours from four perspectives: category with the highest rate of appearance; initial search criteria that mainly characterise the validated results; precision results, and proximity to the initial search criteria (Markaki, 2009). The results derived from the analysis of this information will lead the system to filter the content for a specific user. This method can, however, face lots of difficulties due to the reluctance of users to provide enough feedback.

2.4.4. Personalised Content Evaluation

In order to evaluate the effectiveness of a personalised multimedia content, two different approaches can be applied. In system-centred approaches, the results are compared to a list of analysed documents and precision and recall rate are accordingly computed. However, this method is not suitable for the systems that are focused around the users (Vorhees, 2004). Accordingly, in a user-centred approach, the satisfaction level of the user is measured in an interactive way. However, user satisfaction can be also biased. In addition, it can be impractical to test all the variables involved in an interaction (Hopfgartner et al., 2010). As a result, a mechanism should be adopted to reduce the effect of users' subjectivity.

In the next section, the concept of personalised video summarisation with the primary objective to incorporate the end-users' priorities in video summary generation task will be investigated.

2.5. Personalised video summarisation

A personalised video summarisation system is designed to generate a shorter version of a video based on the user's preferences and interests, while maintaining the significant semantic content of the original video stream (Takahashi et al., 2005b).

Generating useful metadata, extracting the most valuable user preferences and applying them to generate video abstracts to address the users' needs should be regarded as an important research area. Furthermore, exploiting appropriate summarisation techniques, which can produce effective summaries based on the learned user's profiles, is another challenging research topic.

The major components of a personalised video abstraction system are illustrated in Figure 2.8. As was mentioned earlier, the personalisation module is responsible for extracting the users' preferences and generating the required metadata, while the summarisation component should effectively incorporate the captured data in the video abstraction task in an attempt to maximise the viewers' satisfaction level.

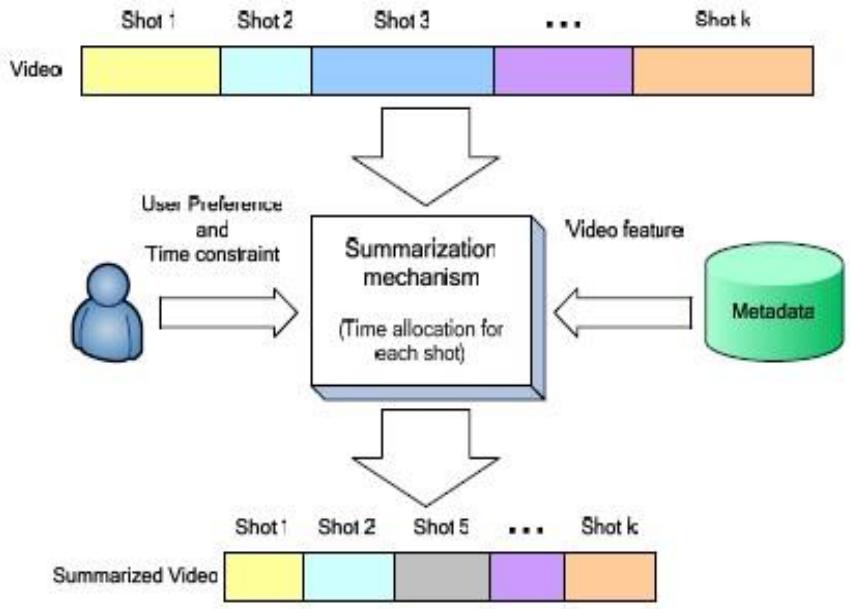


Figure 2.8. Personalised video summarisation modules (Lie and Hsu, 2010)

Designing an effective personalised summarisation system is tightly linked to the adoption of required descriptions for each video segment. Generally, there are two broad types of personalised summarisation algorithms. The first type requires the users to interact with the system directly to train it with their preferences and interests. For instance, the content provider component in this type of systems provides users with the facilities to annotate and enrich the multimedia content. These annotations can be utilised in later stages in order to validate or improve the metadata, which is provided automatically by the system. These user-made descriptions can also be used alongside other audio, visual or textual feature extraction techniques to compensate their inefficiencies in terms of semantic analysis. However, manual annotation should be regarded as time-consuming task due to the explosive growth of video archives. In addition, in an interactive system, users are involved with some technical details, which may be undesirable for them (Ferman and Tekalp, 2003). Consequently, many systems employ automatic video semantic annotation to foster personalised video summarisation and retrieval (Otsuka et al., 2006).

As was discussed earlier, efficiently capturing user interests plays an important role in designing any kind of personalisation system. Since the personalisation methods currently being applied in online applications are mostly text-based, these approaches are not effective enough to be directly applied for multimedia environments. Therefore, several new approaches have been proposed to understand and apply the viewers' preferences. For example, in a system called P-BNN (Personal Broadcast Navigator News) user interests can be inferred both implicitly and explicitly. Accordingly, the keywords that are used by a user to retrieve a video can be captured as a user profile. Moreover, this system is capable to acquire other user priorities during the keyword expansion process (Maybury et al., 2004).

The application of personalisation methods in video summarisation approaches was employed in order to generate a personalised abstract of broadcasted American football video (Babaguchi et al., 2004). Here, important events were initially detected in the video stream by matching the textual captions appearing in a video frame with the descriptions of game-stats or employing the webcast text, in which highlights of the game were described. Consequently, any significant event could be extracted based on these comparisons. All the retrieved events were then analysed in accordance to their importance level and users' level of interests and preferences towards them. The most salient video shots were selected and concatenated as the video highlights. The quality of the generated videos was further

improved by enriching the visual highlight shots with their corresponding audio-textual content from the original video stream. For personalisation purposes in the mentioned approach, a profile was adopted to collect the personal preferences and interests of different viewers. To this end, the following group of elements were included in the user profiles: favourite teams; favourite players; favourite event category. Additionally, other user's required specifications for generating the abstract, such as their required summary length were collected as well. This method aims to personalise the summary content by employing the information provided in these acquired user profiles. In the next section, a group of personalised video summarisation techniques are discussed.

2.5.1. Personalised Video Summarisation Techniques

In this section a group of personalised approaches that are currently being applied to generate video abstracts are discussed. As was mentioned earlier, in some personalisation systems, users are employed to enrich multimedia content in an attempt to capture their preferences directly.

Accordingly, in a personalised content selection system for news video (Merialdo et al., 1999), a video was segmented assuming that a unique scene is the whole presentation of a piece of news. A module converts the closed-caption into a subtitle text and assigns each piece of news to one or more categories; thereafter, multiple keyframes are extracted from each scene. In the next step, a face recognition algorithm is applied to each extracted keyframe and recognised faces alongside the person's name are stored. In addition, an enrichment module provides users with interaction facilities like making hand-writing strokes, spoken commentaries and assigning tags to selected frames. These tags could be utilised to learn a user's preferences towards different elements in input videos. Thus, a personalised video summarisation or content selection system can be developed. A main drawback of the mentioned algorithm is domain-dependency, however.

Human involvement in the abstract generation process of the following systems does not include content enrichment and annotation. In one study (Hopfgartner et al., 2010), shot boundaries were detected using colour histograms and a set of keyframes were chosen based on proximity to the average colour histogram for its corresponding shot. The video was segmented into stories by applying a Semantic Latent Dirichlet Allocation approach to the text stream of the video. In the next step, named entities in the transcripts are chosen as the

representatives of the story content using OpenCalais (a toolkit to distinguish the semantic category of a text). Afterwards, these identified entities were put into their context in Linked Open Data Cloud (SKOS) data models so their semantic categories were elaborated. The user's interactions with the interface could then assist the system to identify the topics of interests. Performing user profiling, a user's categories and subcategories of interests can accordingly be recognised. The main downside of the above system is that the quality of the video abstract is highly dependent on the availability of knowledge in the SKOS database regarding the various concepts.

In related research, a video portal system called VIPP (video portal with personalisation) was designed to deal with sport videos containing semantic metadata. Viewer operations such as video browsing (video segment selection, playback, fast-forward, and rewind) and retrieval (by text or presentation of highlight scenes) were analysed and associated directly with keywords in the metadata. This was followed by updating of the importance weight of those keywords for that particular user. These weighted keywords could then be adopted for creating and updating the user profiles. The sum of the weights of keywords of interest in each profile was calculated as the fitness value of each video segment for that specific user. Finally, a personalised video digest could be generated considering temporal order, attributes and fitness values of the video segments (Babaguchi et al., 2003).

A system with three-tier server-middleware-client architecture was developed to address the issue of personalisation and summarisation under a heterogeneous usage environment (Tseng and Lin, 2002). The client tier provided a user with facilities to specify his/her requests and usage environment. These user requests can include information regarding preference topics, some keywords and time constraints. Furthermore, the client tier could receive and deliver the customised content to the end-user. In the server, each content source was associated with a set of corresponding MPEG-7 descriptions, which was created by IBM's VideoAnnEx annotation tool and a group of content adoptability declarations. The personalisation engine in the middleware then matches user requests with the media descriptions and consequently selected the optimal set of contents to generate the summary. The adoption engine then retrieves the content for the user based on the adoptability declarations.

In another study employing MPEG-7 metadata, user profiling alongside a supervised learning algorithm were the basis for generation of the personalised content. In the first stage, a video operator annotated the equally segmented video segments with a number of high-level

semantic features. Learning the user's preferences was carried out in two stages of training and classification. During the training phase, users were asked to watch the videos and label the interesting events. The high-level features belonging to that specific time window were extracted to produce a training set for a particular user. The training set was then utilised by a supervised learning algorithm to train a binary classifier that can detect the highlight preferences for a specific user. This learned classifier was applied to any new input video during the classification stage to classify each video segment to a relevant or irrelevant group for that specific user (Jaimes et al., 2002).

In a proposed method to generate the personalised summaries of life-log in office environment, the input videos from multiple views were segmented into a single event sequence. The users had to input their degrees of interest in each event, person and objects to assist the system in retrieving the target video. Therefore, the optimal event sequence was constructed accordingly by selection of the best candidate views. Subsequently, summaries of the event sequence were generated by considering the user-entered degrees of interests into contributing elements. The domain knowledge of the existing elements in the office environment and rules obtained from questionnaires were used to facilitate view selection. A fuzzy rule-based system to approximate the human decision-making process was applied for summary production purposes (Park and Cho, 2011).

Crowd intelligence can also be used to verify the identified video highlights. Initially, a multimedia content analysis module is responsible for generating a list of highlight candidates, called best educated guess (BEG), of a football video (Smits and Handjalic, 2010). The BEGs were selected based on the temporal variation of user's excitement level using the concept of arousal. Affective low-level features including motion activity, sound energy, motion entropy, zero crossing rate and shot change rate were all analysed individually and collectively to identify potentially exciting scenes. In the next step, the BEG set was refined by the users' collaborative tagging. By analysing the tags assigned by a specific user during enrichment of a video content and validation of the candidate highlights, user-tailored summaries could accordingly be generated.

In related work, human physiological responses such as respiration rate and blood volume pulse were monitored in order to measure changes in the user's affective state. Video segments, which elicit significant physiological responses in the users, are more likely to be interesting to a specific user and to be therefore included in in the summary. The temporal

location of the corresponding video segment should then be identified in order to produce personalised affective video abstracts (Money and Agius, 2009). However, external factors such as distraction can affect the outcomes negatively.

More recently, sketches have been the basis for generating personalised video summaries (Zhang et al., 2013). Using an interactive selection method (users can select the chosen subject in any frame), similar keyframes were extracted from the video. For identification of the keyframes with identical content to the query frame, the keyframes were segmented and the similarities between the corresponding segments and their neighbours segments were measured. Later, the F-DOG method was applied to the selected keyframes in order to generate the sketches (Kang et al., 2007). After elimination of the abundant points and smoothing the hard boundaries, the Camsift method was utilised to track the objects and identify the motion path, followed by construction of a similarity graph to represent the attributes of each sketch alongside the visual and temporal proximity to the other sketches. For sketch video summarisation in regards to a specific object, the weights of 0 or 1 were assigned to the nodes and edges based on the inclusion of that chosen object. Further, using a minimum weight vertex algorithm, specific sketches based on the weight of each node were obtained. To finalise the layout of the chosen sketches, an energy function was defined considering the five factors of temporal sequence, length of shot, balanced distribution of sketches on the canvas, continuity of the motion track and overlapping sketches. So, a dynamic programming was used in the final stage to acquire the minimal value of energy function.

In an integrated personalised summarisation and retrieval system, the users' relevance feedback was captured to produce the final storyboard summary (Shafeian and Bhanu, 2012). The process was initiated by the users' submitted video query. As a result, a list of the most similar videos to the textual query (known as top videos) was returned. In order to produce the personalised summaries, the similarity scores for all the frames in a top video to the query were computed. Thereafter, all the frames belonging to the top videos were clustered based on their visual features from potentially different videos alongside their similarity scores. The similarity scores for the frames were computed by comparing their visual features in the HSV colour space. Next, the clusters and their constituting frames were sorted in a descending format according to their weights respectively. As a result, a list of ranked frames from all top videos was generated. In the next step, a factor graph was generated for each top movie in which the nodes represent the frames and the edges show the visual affinity of the

corresponding frames. Applying the Affinity Propagation (AP) algorithm (Frey and Dueck, 2003) to the constructed graph, the exemplars (keyframes) were extracted to be inserted into the storyboard summary of each top video. Finally, for personalisation purposes, the generated summaries were presented to the end-users according to their relevancy to the query. Their relativity was measured in accordance to their visual features (HSV colour space) and virtual features similarities. The second criterion was generated during the online stage and in regards to the users' given relevance feedbacks.

In a resource-allocation-based framework, playback speed and perceptual comfort have been the key elements for generation of personalised summaries (Chen et al., 2014). First, a shot-boundary detector divided the original video into short clips, followed by grouping these identified clips into video segments. Later, a number of candidate sub-summaries were generated for each segment by assigning different combinations of playback speeds (from a set of discrete options) to each of a set of contributing clips. The benefit for each sub-summary was computed by calculation of the base benefits of the corresponding clips and extra gain through satisfying specific preferences (inclusion of the user's favourite object, time duration and story continuity). Information regarding the still content of the scenes (to evaluate the relevance of video clip) and information associated with scene activities (to assess visual comfort) were adopted to determine the base benefit for each clip. Finally, the duration resource was allocated between available sub-summaries using Lagrangian relaxation and Convex-hull optimisation methods (Everett, 1963). Adopting these methods, a convex hull for each segment was constructed based on the benefit and cost (length of sub-summary) of all candidate sub-summaries of that segment. All the points from all of the convex-hulls were sorted in decreasing order of the increment of benefit per unit of length. Finally, the ordered points were collected until the summary length exceeded the time constraint.

In another attempt (Hari et al., 2013), human face features were adopted for identifying the keyframes and generation of personalised movie abstracts. In the initial phase, the shot boundaries were detected using the Mutual Information model. In the next step, face detection was carried out on all existing frames using Successive Mean Quantization Transform (SMQT) features and a Sparse Network of Winnows classifier. As a result, a set of face and non-face sample images was adopted for SMQT feature extraction and training the system. Subsequently, for any input video, the existing faces were detected adopting the retrieved features and the previously trained model. For personalisation purposes, the end-

users were provided with a list of detected faces to select their favourite ones. Furthermore, a face recognition algorithm for identifying the matching faces from the identified collection using eigenfaces was adopted. Finally, the shots containing the users' selected faces were included into final digest. However, the effectiveness of this algorithm is bound to limited video categories.

In a semi-automatic, manifold embedding based approach (Han et al., 2011), human subjects were asked to choose their preferred keyframes in an input video sequence so as to overcome the barriers against detection of semantically rich video frames. Then, the visual summaries were constructed based on the inter-frame visual similarity to the pre-selected keyframes. Firstly, a graph based visual saliency algorithm (Harel et al., 2007) was used to assign a weight for each pixel within a frame. Next, the bidirectional similarity between all pairs of frames within a same fixed time window was calculated. The figures were generated by measuring the Sum Squared Distance of two patches and their saliency weights. Later, the distance matrix was projected into the Euclidean space using a manifold learning technique. Thereafter, each frame was assigned a weight based on the features in the embedded manifold and the user's chosen keyframes. Lastly, the key-segments were identified by agglomerative clustering followed by application of a 0-1 knapsack algorithm to the generated clusters.

In another closely related study, the behaviour of viewers is the determining factor in selection of the personalised content (Yoshitaka and Sawada, 2012). The attention level of users was measured, while they were watching the videos by monitoring their operations on the remote controller of the video player and also their eye movements. The video segments which were replayed or played back were labelled as salient to be included into the final summary. Additionally, eye movement data of the users were stored as a sequence of coordinate data of gazing points with a capture rate of 60 samples per second. Finally, the segments of the video, which comprised eye fixations, were chosen to be incorporated into the personalised summary. In such a system the way that human behaves in viewing a raw video can be a clue for the selection process. However there are limitations in the accuracy of the eye and face tracking technique (Peng et al., 2009).

Table 2.1 compares the discussed summarisation techniques from different categories based on their performance and some of their identified shortcomings.

Category	*Sample Techniques*	*Observations/Remarks*
Low-Level Features Based Methods	(Mahmoud et al., 2013);(Khosla et al., 2013); (Carvajala et al., 2014)	-Average results -Incapability to understand the semantic of videos -Computationally expensive
Multi-Modality Methods	(Evangelopoulos et al.,2103);(Jiang et al., 2009);(Bhatt et al., 2009)	-Good results in presence of required information resources - Dependent on availability of the information resource -Noise sensitivity
Domain-Specific Methods	(Daniel et al., 2014);(Taskiran et al., 2006);(Eldib et al., 2009)	-Some acceptable results for specified categories -Not generalizable methods - Dependent on availability of the information resource -Noise sensitivity
Semi-Automatic Methods	(Yoshitaka and Sawada, 2012);(Wu et al., 2011);(Chung et al., 2012)	-Good results in controlled conditions -User's subjectivity -External Factors (distraction) -High expense
Online Summarisation Methods	(Valdes and Matninez,2008);(Ou et al., 2014);(Ipparraguirre and Delrieeux, 2013)	-Average results -No semantic analysis -Local summarisation
Personalised Methods	(Chen et al., 2014);(Hari et al., 2014);(Han et al., 2011)	-Some satisfactory results -Great extent of users involvement- -Incapable of semantic analysis

Table 2.1. Comparison of summarisation techniques from different categories

2.6. Evaluation Methods in Video Summarisation

It is absolutely essential to devise a feasible scheme that can appropriately evaluate the effectiveness of a video summarisation technique. However, evaluation of the quality of automatically generated video summaries can be considered a complicated task due to difficulties in deriving objective quantitative measures for summary quality (Taskiran et al., 2006). Nevertheless, numerous automatic, semi-automatic and manual methodologies have been designed for video summarisation performance analysis purposes. In general, precision and recall and are the two major determining factors in establishment of degree of effectiveness of a summarisation tool (Li et al., 2011b). In information retrieval, the precision rate explains the capability of a system in terms of identification and returning the most relevant and important documents in regards to a user's query, while the recall rate demonstrates the ability of a system to reflect a wider range of documents (Manning et al., 2009). These concepts can be easily propagated to the area of video summarisation to assess the quality of produced summaries. The entire documents collection is equivalent to the input video in the context of video abstraction, while the identified video segments play the role of each single document. Moreover, the duration of the summary and rhythm (tempo) of the generated abstracts are yet other criteria which have been considered by other researchers (Liu et al., 2008). In the next section, a group of suggested models for video summarisation evaluation will be reviewed briefly.

2.6.1. Evaluation Methods

As was mentioned above, there is no standard evaluation model for video summarisation available yet. However, various automatic, semi-automatic and manual evaluation methodologies have been presented. In manual models the main idea is to determine to what extent the machine-generated recounting summaries can potentially capture information from a multimedia content in comparison to the human-generated ones (Metze et al., 2013). Creating a ground-truth list of significant video segments and manually counting the number of similar shots in both the list and in the summary is the most common technique in this category of evaluation methods. A related approach entails providing a list of important topics and textual descriptions of important scenes to the assessors who are then asked to rank the summaries based on these measures (Cunha et al., 2012). Based on another similar method, the temporal location of each extracted keyframe was compared to the temporal

location of user-extracted keyframes and those with a lower difference than a pre-defined threshold were selected as the true detection. Afterwards an F-score (combinational metric composing of recall and precision) value was computed for each generated summary (Lu et al., 2014).

In a questionnaire-based approach, the quality of the summaries were analysed based on the Quality of Perception metric (Gulliver and Ghinea, 2006), which are broadly adopted in assessing the effectiveness of multimedia fields, namely, Information assimilation and satisfaction. The first one measures the extent that the users assimilate the information from the summaries, while the later denotes the effectiveness of an approach to satisfy the users' expectations. Accordingly, a questionnaire was designed to test the quality of summaries from these two perspectives (Ghinea et al, 2014).

In semi-automatic approaches, the evaluation was carried out by comparing the user generated summary with ones generated automatically. In one proposed method, the keyframes were extracted from both versions of the summaries and their visual similarities were measured in the HSV colour space. Finally, the ratio of matched frames to the number of keyframes from the user-generated summaries could be an indicative for the efficiency of the system (Cahuina and Chavez, 2013).

In automatic approaches, on the other hand, the certain indices inside the generated summary were assessed against a pre-defined threshold to measure the suitability of produced abstracts. For instance, Mutual Information and face detection ratio were utilised in one study to analyse the quality of the summaries (Hari et al., 2013). Based on another automated approach, annotations were adopted for the purpose of automatic evaluation using the notion of average precision. In general, this procedure was initiated by generating multiple summaries of a single video using crowdsourcing provided by Amazon Mechanical Turk, and subsequently comparing those summary versions against the ones produced by applying various algorithms. As a result, a group of precision-recall curves were constructed that could be employed for comparison of algorithms against one another (Khosla et al, 2013). In another related system, which functions on surveillance videos, the condensed ratio of produced abstracts (an indicator for the amount and type of motion activities) was used to analyse the efficiency of method (Sujatha et al., 2014). The error rate, defined as the ratio of selected outliers (set of frames with motion blur) to the chosen keyframes, was used in another methodology to evaluate the quality of the summaries (Liu et al., 2014b).

In a combinational model, two multimedia experts were asked to extract the most important video segments in their opinion, while a third person was responsible for intersecting the chosen segments by the other two so as to identify overlapping partitions. Afterwards, another summary version was generated automatically based on subsampling. Finally, a group of individuals were recruited to compare the summary generated by their system with the other two versions generated later by assigning satisfaction scores (Wu et al., 2011).

2.7. Summary and Discussion

This chapter initially provided some background information in the area of digital video and video processing. This was followed by an introduction to the video summarisation concept and a comprehensive review of existing methods available for video abstract generation.

According to the literature, video summarisation is a process of identification and selection of the most significant and valuable auditory, visual and semantic segments of the original video. In general, most of the discussed methods are composed of three major phases: input video segmentation, feature extraction and selection of the most attractive segments (according to their corresponding low-level or high-level extracted features).

It should be reminded that all of these approaches were either fully-automated or in many cases human intervention was necessary. In automatic methods, low-level visual, aural or textual features alongside complicated mathematical concepts are mainly adopted for the selection procedure, while in semi-automatic models, due to presence of the human factor, higher level attributes are considered.

Notwithstanding some acceptable results, automatic methodologies mainly suffer two major issues: they are either domain-specific or largely sensitive to changing conditions. Due to the domain-dependency, a summarisation methodology could only be utilised for a sole video category and cannot be generalised for other genres. Vulnerability to changing conditions can be defined as the incapability to cope with diversity in local or environmental factors such as modified lighting condition, external noise, etc. As a result, a slight transition in any of these conditions can potentially deteriorate the outcome significantly. This is due to the direct effect of these changes on low-level features and the inability of these algorithms to address such situations.

In addition, methods utilising complicated statistical and mathematical concepts are often quite time-consuming and computationally expensive. Methods involving graphs, clustering and statistical classification models are usually very costly to be applied to large volumes of low-level multimedia data. Moreover, it should also be mentioned that these models can only achieve successful results under certain restrictive conditions.

Another major downside in regards to the explained methods is their ineffectiveness and incompetence in understanding the semantics of video segments. As it is almost far beyond today's technologies to interpret low-level features into high level semantically meaningful concepts, it is quite impractical for these systems to compare different video segments based on their semantic and contextual values. Furthermore, the lack of potential to contrast the importance of video partitions in the context of whole video is another drawback that should be regarded in relation to these methods. This is due to the nature of these methodologies which mainly concentrate on low-level characteristics of temporally approximate (neighbour) video segments.

The second group of summarisation tools reviewed were those with human involvement to bridge the semantic gap between low-level features and perceived high-level conceptual categories. In most of the approaches belonging to this category, a user was solely employed to explicitly or implicitly determine the importance of different video partitions. Adopting a single user for this purpose can be considered problematic as well. The personal inclinations and preferences of different people can be dramatically different. As a result, this subjectivity will have a direct influence on their content selection, which can be potentially negative. In addition, some external factors such as distraction or noise can significantly undermine the quality of the eventual product.

In the second section of this chapter we focused on the topic of personalisation in multimedia. This was defined as the procedure of integrating the end-users' preferences and priorities in multimedia content presentation. As was discussed, the main objective is to understand end-users' characteristics and interests in order to tailor the final product in a format to meet the identified expectations.

The concept of personalised video summarisation was discussed accordingly in the following section. This was defined as the task of integrating end-users' interests and preferences into the video summarisation process. Therefore, a scheme should be devised with the ability to discover users' priorities towards different video segments or existing aural, visual and

textual objects in the videos. Thereafter, these elaborated elements should be combined with video summarisation techniques in order to generate personalised abstracts.

A group of developed personalised summarisation techniques was explained afterwards. These abstraction techniques are either automatic or semi-automatic. Users' information extraction in both of these methodologies was performed either explicitly or implicitly. The basic idea for these systems is the fact that superior video summaries can be delivered to end-users only when their expectations have been considered.

However, all the fully-automated video summarisation methods still suffer from the shortcomings previously mentioned in this section in terms of cost and semantic importance detection, notwithstanding the fact that the personalisation process has improved their results significantly. On the other hand, most the developed semi-automatic approaches require considerable end-user involvement for understanding their priorities, which can be very inconvenient and time-consuming. This can potentially reduce the likelihood of their optimal participation in their interests' retrieval process. Finally, a collection of evaluation methods that has been devised by researchers to analyse their developed summarisation tools was described. Moreover, a number of criteria that should be considered for evaluation purposes were mentioned.

2.8. Problem Statement:

As explained previously, the existing automatic video summarisation techniques are suffering a number of issues such as domain-dependency, noise sensitivity and high computation expenses. On the other hand, the semi-automatic approaches involve human in abstraction process to overcome these issues, however, sole user subjectivity or distraction can potentially deteriorate the final summaries. Accordingly, a number of research objectives are identified and explained in the next section in furtherance to address the discussed research gaps.

2.9. Research Objectives:

Our initial research objective is the investigation of the existing video summarisation approaches in order to identify their shortcomings and strengths. This objective was addressed in this chapter.

Having described the shortcomings of the existing automatic video summarisation techniques, **our second research objective can be described as to design, develop and evaluate a user-centred video summarisation algorithm based on group scoring in accordance to the findings from the previous investigation.** In the proposed approach, the negative effects of employing sole user such as subjectivity should be minimised. Our proposed method will be explained comprehensively in chapter 4.

Considering the varieties in preferences and inclinations of the users who are going to watch the generated summaries, a mechanism should be devised with capability to distinguish the end-users' priorities towards different video segments with potentially distinct auditory, visual and semantic content. As a result the third research objective is defined as **to extend the work of previous objective and design, develop and evaluate a personalised video summarisation algorithm based on group scoring**, which will be explained in chapter 5.

The fourth research objective subsequently is defined as to extend the work of previous objective and design, develop and evaluate a personalised video summarisation system with reduced end-user involvement. The recommendation of an approach to personalise the video summaries through creating more generic user profiles that can be applied effectively to any input video in an attempt to provide users with a better and more satisfactory experience will be discussed in chapter 6.

In the next chapter, the methodology that has been adopted throughout this research will be explained. Furthermore, the way that this identified research objectives are correlated with the chosen methodology will be discussed.

Chapter 3

Research Methodology

3. Overview

In the previous chapter, the four main objectives of our study were identified. In this chapter the adopted methodology of this research that is used throughout our work to achieve the identified objectives from previous chapter will be explained and justified.

Initially, the concept of positivism as our chosen research paradigm and its major assumptions are explained. In the next section, the use of the Design Science Methodology (DSR) throughout our research, according to its attributes and definitions, is justified. Later, the employed approaches to carry out our research in its different phases are discussed and finally a list of tools and materials that is used during our study is explained and justified. The Figure 3.1 demonstrates our adopted DSR methodology and the methods employed at each stage within the process.

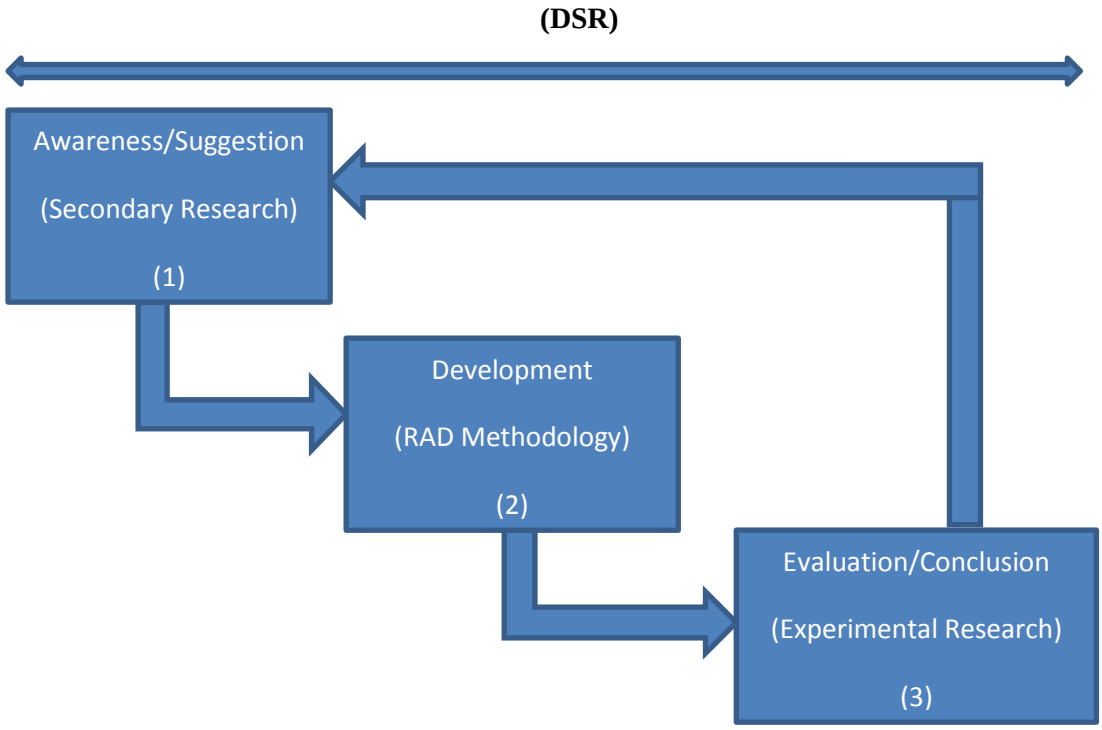


Figure 3.1. Adopted methodologies in course of research

3.1. Research Definition

The word *research* has been defined as "The systematic investigation into and study of materials and sources in order to establish facts and reach new conclusions" (Oxford Dictionary, 2013). It also has been described as a set of activities to figure out a phenomenon (Kuhn, 1996). In spite of various identified applications for research, expanding and improvement of knowledge should be regarded as the primary objective. In other terms, generating the new knowledge that is generally applicable should be considered as the main goal in a research activity (Dawson, 2002). The research methodology then, explains the ways that a research project should be undertaken optimally or, in other words, the best applicable practices to a specific issue (Howell, 2013). However, a research method is a collection of activities that a research community accepts as suitable for production of their knowledge. It should be added that a research method also has the responsibility to maximise the validity and accuracy of the conducted research (Vaishanvi and Kuechler, 2009).

3.2. Research Perspective

The term paradigm is used to demonstrate a shared conceptual framework by a community of researchers. In other words, it describes a research culture with a set of beliefs, values, and assumptions that a group of researchers has in common in regards to the nature and conduct of research (Kuhn, 1996). Thus, these similar elements can be used in order to elaborate the required activities in the course of a research. In general, a paradigm can be demonstrated from four philosophical views of ontological, epistemological, methodological and axiological. (Mingers, 2001; Vaishnavi & Kuechler, 2009). These worldviews can be defined as below:

Ontology (nature of reality): investigates the nature or form of the research area in accordance to its reality and possibility.

Epistemology (nature of Knowledge): explains the relationship between the knowers (Researcher) and what can be known in which way.

Methodology (theory of reasoning): examines the possible ways that a researcher can choose in order to obtain the knowledge.

Axiological (ethics): defines the value extent of a given subject and effects of such values on conduct of research.

There are four research paradigms, namely positivist, interpretive, critical and research design ,the latest of which is becoming popular in information systems (Fallman, 2003; Stolterman, 2008). In the positivist research paradigm, the assumption and hypotheses should be supported by data collection. On the other hand, the basis of the interpretive paradigm is on the fact that there is no single reality and therefore, data collection should be used for retrieving knowledge. Similarly, the critical research paradigm assumes that reality is constructed socially; however such construction is affected by a number of power relations between people. Finally, design science is associated with human-made artefacts in terms of construction and evaluation in order to enhance system elements (Myers, 1997; Vaishnavi and Kuechler, 2009). The philosophical assumptions of these four paradigms are illustrated in Table 3.1.

	Philosophical assumptions			
Research paradigms	**Ontology**	**Epistemology**	**Methodology**	**Axiology**
Positivist	- Single, stable reality - Law-like	- Objective - Detached observer	- Experimental - Quantitative - Hypothesis testing	- Truth (objective) - Prediction
Interpretive	- Multiple realities - Socially constructed	- Empathetic - Observer subjectivity	- Interactional - Interpretation - Qualitative	- Contextual understanding
Critical/ Constructionist	- Socially constructed reality - Discourse - Power	- Suspicious - Political - Observer constructing Versions	- Deconstruction - Textual analysis - Discourse analysis	- Inquiry is value-bound - Contextual understanding - Researcher's values affect the study
Design	- Multiple, contextually situated realities	- Knowing through making - Context-based construction	- Developmental - Impact analysis of artefact on composite system	- Control - Creation - Understanding

Table 3.1. Different research paradigms from philosophical point of views (Vaishnavi & Kuechler, 2009)

In the context of our research, positivism is adopted for the following reasons. Initially, from an ontological point of view, our research topic is development of the most effective and efficient video summarisation technique which should be considered as a sole reality in this world, thus it can be obviously associated to the positivist philosophy. In addition, a new technique will be proposed based on the existing scientific facts, which through subsequent experiments could be converted into knowledge. This is totally compatible with the positivist epistemology. Moreover, from a methodological point of view, the effectiveness and efficiency of our proposed novel video summarisation approaches will be validated based on the experimental studies that are carried out in the evaluation phase. Finally, a quantitative user-based evaluation study will be performed to demonstrate the value level of our recently proposed summarisation methods (axiology), which again conforms to the positivist paradigm.

3.3. Research Type

Generally, a group of underlying attributes and characteristics is tied to the positivist paradigm from a methodological perspective, as described below (Mertens, 1998; Kane and O'Reily, 2001):

1- Only the facts that are observable and provable should be considered as science.
2- The existing relations in both natural and social worlds can be examined using experimental studies.
3- There should be a value-free method to investigate the world.

As was discussed earlier, the chosen philosophy for this research in accordance to their compatible natures is positivist. In addition, the essence of this paradigm consists in the interpretation of the results based on the evidence that has been collected and assessed in a systematic manner. The main aim of this research is to propose an effective video summarisation technique. As a result, three subsequent research questions could be formed:

RQ1) what is the best technique to generate video summaries in order to minimise the shortcomings of the other techniques?

RQ2) what is the best way to develop (implement) the proposed techniques?

RQ3) are the developed tools effective enough?

The first research question corresponds to the first identified research objective, while the second and third questions should be addressed in achieving the second, third and fourth research objectives.

Accordingly, three approaches for generation of video summaries alongside a set of variables including *Recall*, *Precision*, *Timing* and *Overall Satisfaction* associated with the effectiveness of our generated video abstracts will be recommended. These variables should be empirically observable and will be explained comprehensively in the following sections. Since our work deals with three characteristically varied research questions then, there is a necessity for adoption of a research strategy that can potentially address all of them.

In dealing with the first question, an exploratory research will be employed to develop the best potential summarisation techniques on the basis of the existing literatures. This is due to the nature of exploratory research, which investigates a problem that has not been clearly defined (Shields and Rangarjan, 2013). Thus, within an exploratory research, the research questions that have not been answered previously are examined. In addition, in most cases, this type of research will assist researchers in formulating more relevant hypotheses and further investigations.

In order to tackle the second research question, a mechanism should be proposed in order to design and develop the recommended techniques into the form of the actual software products (in context of this research) a priori to the final experiments. In fact, these artefacts will produce the independent variables which will be studied later during the experimental phase. The main objective of a design theory is to support the achievement of goals. These theories are a mixture of natural, social and mathematical sciences with the aim to put the explanatory, predictive or normative theories into practice (Walls et al., 1992). The adopted methodology for design and development of these artefacts (software products) will be explained in section 3.5.

Furthermore, in order to address the third research question, the developed techniques from the last stage of research should be empirically investigated. Therefore, a confirmatory type of research should be adopted at this stage for verification purposes. Confirmatory studies try to acknowledge a pre-specified relationship as opposed to exploratory ones in which potential correlations could be extracted (Boudreau et al., 2001). In fact, this type of studies concentrates less on elaboration of theories or mechanisms; instead they tend to verify the validity of extracted hypotheses (Jenkins, 1985). Finally, deductions can be made by

evaluating a large number of observations in order to generalise a theory and form a universal law. The adopted hypotheses in the course of research will be explained in section 3.7.4.2.

3.4. Research Method

Further to our earlier discussion in this chapter, three naturally distinct research types were identified for this study that should be addressed at different stages. Therefore, as mentioned in the last section, a research method should be employed with the capability to address all of these distinct research types. As a result, the Design Science Research (DSR) methodology was utilised to carry out this research according to its attributes. A brief review of this methodology alongside the rationale behind its adoption is carried out further in the next two sections.

3.4.1. Design Science Research Methodology

Generally, Design Science is an outcome-based methodology, which is mainly being adopted for research in the information systems arena. In fact, the general objective of design-science research is to devise innovative and purposeful artefacts for a specific problem domain. The artefacts should be designed, implemented and evaluated in an effective manner in order to provide the solutions to unsolved problems or enhance a phenomenon or service (utility). The mentioned fact represents this methodology as an adoptable approach for the positivist philosophy since truth and utility can be regarded as two sides of the same coin (Hevner et al., 2004). In DSR, the word “purposeful” reflects the idea that the developed artefacts should be able to deliver useful and efficient services since they are supposed to upgrade the existing practices, or to recommend better solutions (Kuechler & Vaishnavi, 2008). The different phases that a Design Science researcher has to go through in order to produce knowledge are illustrated in Figure 3.2 in the next page. In this figure, the correlations between each stage of this methodology and our developed research questions (RQs) alongside the corresponding required research type are shown.

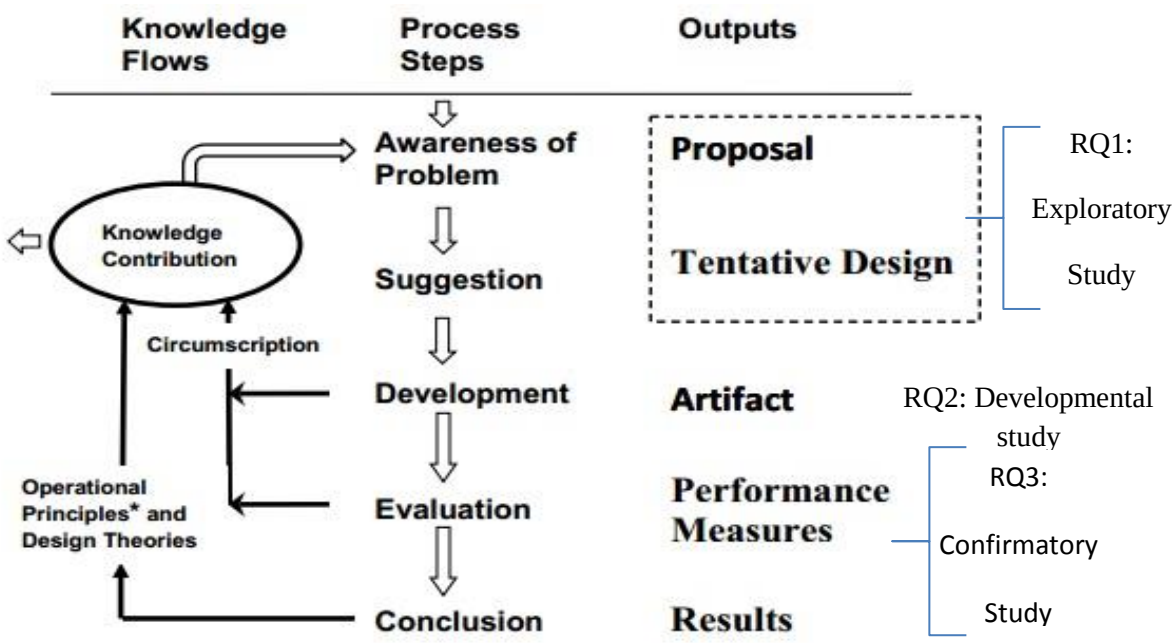


Figure 3.2. The different stages in DSR methodology (Vaishanvi and Keuchler, 2004)

In the *Awareness* stage, the opportunity for innovation in a related field is identified using multiple sources, such as new developments in industry, and the outcome is produced in the form of a proposal. Subsequently, the *Suggestion* phase will result in development of a tentative design (model) in accordance to the identified requirements. Later, in the *Development* stage, the devised tentative design from the previous level will be developed into actual artefacts. This will be followed by a comprehensive *Evaluation* of the developed artefacts in the next step to investigate its generalisability. Finally, the *Conclusions* are drawn according to the artefacts' behaviours in order to make contributions to knowledge or, indeed, to develop new proposals.

3.4.2 Why Design Science Research?

Further to our discussion in section 3.3 the main objective for our research was identified as proposing effective video summarisation techniques. Since we are trying to propose new video abstraction approaches with the potential to produce higher quality video summaries compared to the existing tools, therefore, this can be regarded as an attempt to develop the artefacts to improve the current solutions to an existing problem. This is in exact compatibility to the application scenarios that have been denoted for this methodology.

In addition, in section 3.3 a number of research questions were extracted in relation to our primary objectives, each of which should be addressed by a different type of research activity as was accordingly described. Exploration, development and confirmation were outlined as these activities that should be carried out subsequently. Further, each of these identified research types and their corresponding objectives can be associated with at least one of the phases described for DSR methodology. The required exploratory research to develop the abstraction techniques can be covered in the two initial stages, while implementation of the software products should be carried out in the development phase. Finally, our confirmatory study to investigate the effectiveness of our proposed methods prior to knowledge generation will be performed in the evaluation and conclusion steps respectively. This can be used as another justification for the utilisation of this methodology.

Moreover, since different iterations in the DSR methodology will result in identification of potential gaps and development of new theories, this framework can be considered as being consistent with our research objectives. The identified research objectives mentioned in section 2.9 were formed in accordance to these cycles. In particular, the experimental results and derived conclusions from the initial summarisation technique could reveal the opportunity for proposing a new approach in which end-users' preferences are addressed.

As was explained earlier, different stages of our methodology should be addressed by distinct research types. An exploratory research is required for the initial two stages in order to form the tentative designs (techniques/RQ1), while a developmental study is required for implementation of artefacts (summarisation tools/RQ2), and eventually a confirmatory study is necessary for investigation purposes (tools evaluation/RQ3). The adopted methodology for each of these research types will be described in following sections.

3.5. Methodology for Proposal and Tentative Design

As previously mentioned, our first research question is to identify the best potential video summarisation technique, which is answered through an exploratory study. Having said that, a secondary research based on a comprehensive study of the existing literature and related work should be carried out in order to identify a knowledge gap in the video summarisation field and to form the observable approaches on that basis. These methods will all be described in detail in the upcoming chapters.

3.6. Artefacts Design and Development

Design can be defined as a process to create new artefacts. According to another definition, design consists of a collection of instructions that should be applied for producing things (Hevner and Chatterjee, 2010). On the basis of existence of the required knowledge a priori, a design can be categorised as routine or innovative. As opposed to an innovative design, the required knowledge for producing the artefacts is available in a routine design. However, innovative designs are mainly adopted to fill a knowledge gap or to improve the existing condition by conducting researches (Vaishanvi and Keuchler, 2009). A design activity can also be explained in accordance to its distinctive scientific nature. As opposed to natural science in which a body of knowledge regarding the relations and interactions between some class of objects and phenomenon should be explained, the design science is a body of knowledge in regards to creation of artefacts and objects (Simon, 1996). In this section of our study, designing the software products that can represent our methods in an appropriate manner is the main objective. In the context of this work, the artefacts are the software products that supposedly deliver the functionalities of our proposed approaches. In fact, these artefacts (independent variables for our experimental research) are responsible for generation of the video summaries in accordance to their fundamental models.

In Information Systems (IS), the importance of an appropriate design has been widely discussed. The applicability of design has been directly linked to the relevance of IS research (Glass, 1999; Winograd, 1996). In the system design stage, the primary goal is to choose the best options among the possible candidates in order to limit the resources and utility (Hervner and Chatterjee, 2010). These previously mentioned artefacts can be developed based on various software development methodologies which should be justifiable in regards to the main objectives and characteristics of the project.

3.6.1. Software Development Methodology

A software development methodology is defined as a set of procedures, techniques, tools and documentation assets with the aim to assist the software developers in their attempts to design and implement a new information system (Avison and Fitzgerald, 2006).

The success rate of a software development project can be potentially boosted if the adopted methodology is chosen in accordance to the nature and characteristics of the project. In fact, various attributes of a project in terms of technical, organisational and available resources

should be considered in advance of the selection of a particular framework (Geambasu et al., 2011). Two main categories of approaches for a software development process could be considered based on the extent that their development phases are separated, namely, *Traditional* and *Agile* (Boehm & Turner, 2004; Nilsson, 2005). In traditional models such as Waterfall, the project is divided into sequential phases and a final deliverable can be prepared at the end of each phase, while in Agile ones such as Spiral, the final product will be ready after a number of iterations (Thayer and Boehm, 1986). In addition, there are methodologies that combine some essential elements from both categories. The most well-known one is arguably Rapid Application Development (RAD) (Geambasu et al., 2011).

Rapid Application Development is a software methodology that has been developed on the basis of prototyping approaches with the objective to produce faster and cheaper software products (Martin, 1991). This is the framework that has been employed for the design and development of artefacts in the development phase of our adopted DSR methodology. Some of the characteristics of this software development methodology alongside the justification beyond selection of this framework are explained further. In Figure 3.3 the correlation between different stages in RAD methodology is illustrated.

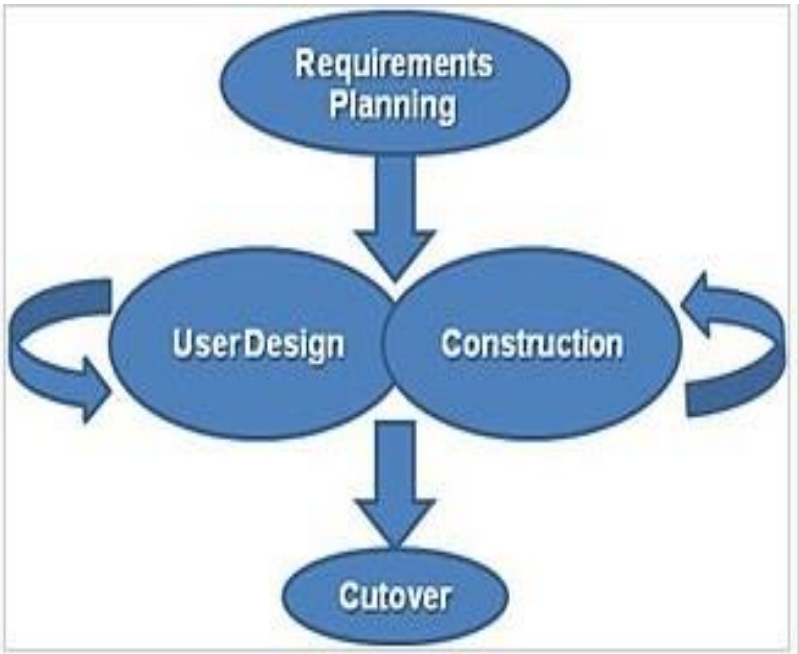


Figure 3.3. The proposed model for RAD (Martin, 1991)

In the initial step, which is *Requirements Planning*, the key requirements of a system alongside the existing constraints are elaborated. Further, in the *User-Design* and *Construction* phase the users are engaged with the aim to design and develop the final tool, which fulfils the end-users' requirements optimally. Finally, in the *Cutover* stage, the testing and training procedures are carried out in a compressed manner in order to shorten the development temporal length (Martin, 1991).

In fact, RAD can be considered as a structured methodology which is mainly adopted by small-teams with more limited financial and temporal resources. This approach to software development concentrates more on the development phase rather than planning using evolutionary and participative prototyping mechanisms. In addition, this framework has been regarded as an appropriate option for the user-interface based software products that require a high level of human interaction (Martin, 1991; Mortimer, 1995). Moreover, this framework divides the eventual system into smaller segments. Thus, it can appropriately address any potential changes in initial requirements (Geambasu et al., 2011).

In the context of our research, the main objective of this phase is the development of a tool in the quickest available way so that it has the capability to reflect different elements of our proposed approach for video summarisation in an optimal manner. As a result, this methodology is well-suited for our research objectives due to fact that it emphasises more on development of artefacts rather than on the creation of detailed planning or design documents. In fact, the implementation of the artefacts in the fastest way as a prerequisite to our experimental studies in the evaluation stage should be regarded as the primary goal of development phase. Furthermore, our proposed summarisation approaches, as will be discussed in later chapters, are all user-centred and require a high level of human involvement. Therefore, this methodology can be described as the best choice since software products with a large amount of human interaction are considered as the main use case scenarios for this framework. Finally, since our recommended methods are evolving in the course of the research, the applied methodology should be capable of addressing these changes in the most efficient way. In fact, during each cycle of the DSR approach, which is initiated based on a newly identified knowledge gap, a novel technique will be presented in our work, which will impose new requirements that should be dealt with optimally. This has been denoted as one of the main attributes of the RAD methodology. In the next section, the methodologies used to carry out our confirmatory research to support our recommended techniques will be explained.

3.7. Methodology for Evaluation and Conclusion

Further to our earlier discussions, a positivist paradigm is adopted for the entire research and this perspective is fundamentally associated with generating knowledge through testing and supporting the theories. Therefore, in this section the adopted methodology for our evaluation and conclusion phases of the DSR approach is explained in more detail. Generally, in this stage, the hypotheses formed based on our recommended theoretical methods have to be empirically studied in a confirmatory style. In doing so, a mechanism should be applied to measure the effectiveness of our nominated approaches. However, assessing the quality of a summarisation technique is characteristically a complicated task. Nonetheless, a number of parameters that potentially have the capacity to reflect the effectiveness of our methods could be identified and examined instead. These criteria are all products of our initial exploratory studies and are extracted from secondary research. Therefore, a research method with the capability to assess and verify the validity of our introduced methods should be employed. Thus, experimental research, which characteristically has the ability to distinguish any significant differences between controlled conditions (Gulliver, 2004) should be highlighted as the right research method for this part of study. Experimental research forms part of so called Fixed Research Design, which will be discussed in the next section.

3.7.1. Fixed Research Design

Before we carry on with our adopted experimental methodology, some of the essential concepts in designing an experiment will be elaborated. To this end, there are some variable-based aspects that should be considered in any experiment (Kirk, 1995):

1. Manipulation: in any experiment there should be at least one independent variable which is manipulated by the researcher.
2. Measurement: There is at least one criterion in each experiment that should be measured and it is called the dependent variable.
3. Control: There is a possibility for the existence of other influential factors (apart from independent variables) that can affect the experimental subject's behaviour and they are known as extraneous variables.

In our study, as it will later be explained, the independent variables to be studied are a set of video summarisation techniques and our dependent variables are a set of indices (*Recall, Precision, Time* and *Satisfaction*) that indicate the quality level of the generated summaries.

However, in our work, a number extraneous variables including noise and lighting conditions that can potentially be a decisive factor in formation of those outcome variables will be controlled in a laboratory-style condition to minimise their influence.

In general, a fixed design research can be categorised from the manipulation point of view into Experimental Research and Correlational Research. In correlational research, there is no manipulation of independent variables and the main objective is to investigate the correlations between two or more variables. However, in an experimental research the independent variables should be modified to monitor the possible consequences on the outcome variables (Creswell, 2008).

In the context of this work, different video summarisation approach as the independent variables should be tested in order to observe their impacts on our selected dependent variables. As a result, our research in nature is experimental rather than correlational.

3.7.2. Validity and Generalisability

Based on positivism's underlying attributes, the results of an experimental study are considered internally valid as long as a set of pre-specified conditions are met; however, external validity cannot be guaranteed. In other words, using the laws of cause and effect, generalisation is possible as long as the pre-determined conditions are met (Kane and O'Reily, 2001). On the other hand, external validity demonstrates the level that our research results can be generalised to other samples and conditions. It can be concluded that the more internally valid an experiment is, the less it can be validated externally. In fact, the results retrieved from one experiment can never be generalised universally regardless of its quality. As a result, the major objective should be set to test and support the proposed theory instead of attempting to get generalisable results (Leary, 1995). Nonetheless, in our study, we try to improve the external validity through the use of multiple viewing devices and probability sampling of the participants in order to make more generalisable findings. Presenting the video summaries retrieved from different techniques (independent variables) from a number of genres using various devices and selecting the participants randomly from different demographical categories can potentially boost the external validity of our experimental results.

3.7.3. Experimental Research

Experimental techniques can be categorised into two groups: structured or non-structured based on their level of pre-elected experimental factors. In structured experiments, the changes in a collection of pre-defined experimental elements are measured consistently using questionnaires, as opposed to unstructured ones in which open interviews or participant observations are utilised for conclusion purposes. In our work, structured experiments have been adopted using a questionnaire to assess the effectiveness of our approaches and tools suggested for video abstraction purposes.

In addition, different techniques can be used for performing the experimental research, such as field studies, case studies, simulations, laboratory studies and field experiments. However, the most prominent methods, in which the potential influence of any change of independent variable on the dependent variables is monitored appropriately, should be regarded as being either a field study or a laboratory study. In laboratory studies, the experiments are carried out in an artificial environment which is created by researchers in order to control different types of variables and tasks strictly (Boudreau et al., 2001). Although, the field experiments provide the researchers with the opportunity to investigate phenomena in a more realistic environment, there are external metrics, involved in this type of study, playing key roles in determining the final outcome, which should be measured and monitored.

As was discussed earlier, the external validity and as a result, the generalisability of an experimental research can be improved if the outcomes are observed and investigated in natural occurring system. However, as mentioned previously, adopting a field study necessitates observing and measuring other participating external variables (such as level of noise, etc.), which is not the purpose of this study.

Therefore, our experiments are carried out in a more controlled laboratory-style condition. Consequently, a number of vacant rooms (except one participant and an observer) with similar lighting conditions and minimised external auditory noises are adopted for our studies. However, different experimental devices are used for the experiments as will be discussed later, in order to produce realistic and more generalisable results.

3.7.4. Experimental Methodology

In accordance to the recommended techniques (each corresponding to one of the research objectives identified in chapter 2, three experiments will be carried out to observe their effectiveness. In this section, the general experimental design specifications that are common to all three studies will be discussed. Later, more detailed experimental designs that are devised specifically for each proposed approach will be explained in the evaluation section of each chapter. However, we first introduce the video clips that are used in the course of our studies in the next section.

3.7.4.1. Participants Recruitment

The participants for our research will be recruited based on the convenience sampling technique, in which the samples are drawn from the population with highest level of accessibility and availability. Their participation is not incentivised by any mean and they are fully informed in regards to ethical aspects of the experiments.

3.7.4.2. Experimental Videos

In our experiments, we used a total of six videos each belonging to one of the six different video categories namely, *Music Video*, *Sport*, *Movie*, *Advertisement*, *News* and *Documentary*. All the videos were digitised in MP4 format with a resolution of 640*360 pixels, a 25 frames/seconds play rate and bit rate of 2193 Kbit/second. The original length of each clip was two minutes, whereas the summarised versions were in the region of 30 seconds (the exact summary time depends on the adopted summarisation technique). The main justification for the temporal length of chosen summaries is due to the limitations of a user's memory. In fact, the probability of the human's brain tendency to forget the information that was shown to it at the early stages of the clip will be largely increased if the duration of the video is considerably more than 30 seconds (Aldridge et al., 1995). In addition, the longer the video is, the probability of some negative external factors on participants (such as distraction), which can potentially deteriorate the final outcome, will be boosted.

Furthermore, the videos are selected from the mentioned categories in order to measure the effectiveness of different summarisation techniques on different genres, as the task of video abstraction can be very domain-specific. Additionally, selecting the videos from a broader

spectrum of infotainment is likely to make our experimental task more interesting to the participants, since it covers a wider range of interests and inclinations. It should be mentioned that there are numerous video clips that mainly belong to a particular genre; however, they also incorporate scenes that can also be related visually and semantically to other categories. As a result, selection of an input video from this cross-domains category of videos can be essential. Moreover, there is a probability that the input video for an abstraction task had been previously summarised either automatically or manually. Therefore, assessing the effectiveness of our approach on summarising the previously skimmed video clips should be considered important as well. Further, the initial videos that have been adopted for our experimental research are described.

Music Video: This shows the Linkin-Park music band playing and singing. Whilst the band appears in some scenes, in others the story line of a girl being ignored by her peers is developed. A number of landmarks and buildings are also displayed. There are both visually (such as landmarks) and auditory valuable (singing parts) information in this video clip.

Sport: This comprises the highlights of a football match from the French league. This clip contains goal scoring scenes as well as some other critical moments of the match. These are shown from a variety of viewing angles, both at normal playback speed and slow motion.

Documentary: This is a National Geographic clip of an eagle hunting a sea snake to feed its eaglets. Some critical scenes are shown at slow motion speed as well from different camera angles. There is valuable auditory information in the clip as well, which is not covered visually, since the narrator provides some information regarding the eagles' general hunting habits.

Movie: This is a two minute trailer for the Avengers Assemble movie. This clip includes some action scenes alongside a number of dialogues between the actors. Valuable textual information regarding the name of the movie and the producers is also given.

Advertisement: This is a cross-domain commercial video for Pepsi in which two groups of well-known football players in a western-movie style clip play football over of a can of soda. There is minimal aural information present in the clip.

News: This is a news video clip regarding an European Parliament MP who has recently resigned; in it, he is supporting his successor, who is under attack in the media because of his book about aliens. The most important information source for this video can be found in its aural and textual content; visual data play a less important informational role in this particular video. Table 3.2 shows a sample frame from each of the experimental videos.

		
Music Video clip	**Movie clip**	**News clip**
		
Sport clip	**Advertisement clip**	**Documentary clip**

Table 3.2. Experimental videos from 6 different categories

3.7.4.3. Experimental Procedure

As was noted in section 3.3, confirmatory research will be undertaken to investigate the third research question of our study, which targets the effectiveness of our proposed summarisation techniques. Since confirmatory research necessitates the formulation of hypotheses, four hypotheses underpinning this stage of our study were formed as follows:

H1- Our technique will generate video summaries at an **acceptable** *Recall* rate.

H2- Our technique will generate video summaries at a **high** *Precision* rate.

H3- Our technique will generate summaries by **strictly** meeting the *Time* constraints.

H4- Our technique will generate summaries with the **highest** *Overall Satisfaction*.

Acceptable: Our technique will have a higher score than the average of mean scores obtained by the other techniques for at least half of the video categories. Additionally, for the genres in which our approach does not manage to exceed, the difference between scores achieved by our technique and average of those obtained by the other tools, should not exceed 1.5 units.

High: The mean score achieved by our approach will be higher than the average of scores gained by other summarisation methods for all video categories. In addition, for at least one genre, our method has to receive the highest score in comparison to other techniques.

Strictly: The generated summaries should be exactly 30 seconds.

Highest: Our summarisation technique should achieve the highest scores for all the video categories in comparison to other available approaches.

Each of these hypotheses will form the metrics indicating the dependent variables which will assist us in observing the consequences of changes on independent variables. In the context of our research, the independent variables are different video summarisation techniques that are employed to generate different summary versions of the experimental videos. These approaches will be explained in detail in the next chapter. However, in this section, the four identified dependent variables namely, *Recall, Precision, Time* and *Overall Satisfaction* will be detailed. Participants express their opinions on each of the mentioned criteria on a scale of 0 (lowest level of interest) to 10 (highest level of interest) based on their perceived experiences. Further, each of these variables will be elaborated upon.

Recall measures the extent to which the generated summaries reflect all the existing scenes from the original videos.

Precision evaluates the ability of the generated summaries to include the most important scenes of the initial videos into the summaries.

Timing explains the level of temporal proximity of the built abstracts to the required summary length.

Overall Satisfaction score represents the extent to which the end users are satisfied with the summaries from different points of views, namely visual and aural coherency, continuity and adjustability.

The experiments involved a group of participants who firstly watched the six previously described multimedia video clips. This is done in order to familiarise the participants with the content of the experimental material. In the next step, after displaying each original video clip once more, a number of 30 seconds summary versions, which are generated using our proposed technique alongside other abstraction tools, are presented to the users in a randomised order to minimise any potential order effects. It should be reminded that participants have no prior knowledge regarding the adopted tool for each summary version. This is in an attempt to reduce any user bias towards a specific abstraction tool.

Presenting all the generated summary versions for an input video, the users are then asked to complete a questionnaire relating the previously discussed metrics in respect of measuring the effectiveness of a video abstract. Each statement in the questionnaire corresponds directly to one of the dependent variables. It should be reminded that prior to start of the experiments, each of these four questions are explained in detail to the participants in order to ensure that they fully understand the basis for the scoring mechanism. Further, the users are asked to score each of these four questions on a scale of 0 to 10 according to their perceived audio-visual content of the summaries. A score of zero signifies total disagreement, while a score of 10 shows total agreement with the statement. This questionnaire is shown in the Table 3.3 below. It should be mentioned that the same questionnaire will be adopted for carrying out all three investigative studies in the course of our research.

Participant No:	**Video Category:**	**Agreement score**
1- The video summary covers appropriately all the existing scenes of the original video clip. **(Recall)**		
2- The video summary successfully extracts the most semantically important segments of the original video. **(Precision)**		
3- The generated video summary is exactly 30 seconds and the time constraint is met. **(Timing)**		
4- I am satisfied with generated video summaries. **(Overall Satisfaction)**		

Table 3.3. Questionnaire used for measuring the opinions of participants towards the generated summaries

3.7.4.4. Analysis of Results

An average score is generated for each dependent variable of each video summary which reflects the mean opinion of all the participants. Thereafter, these mean scores can be utilised to compare the effectiveness of different abstraction tools (including ours) from four different perspectives.

Our proposed method can be described as effective if all of the initially formed hypotheses are addressed properly based on our statistical analysis. In response to the first hypothesis, we expect that the average assigned scores from the *Recall* point of view achieves acceptable results in accordance to its provided definition.

In relation to the second hypothesis, the generated summaries should be scored highly in terms of *Precision*. This can is due to the personalisation concept, which will be described comprehensively in chapter 5. As has been discussed when reviewing in information retrieval theories, there is always a trade-off between the *Precision* and *Recall* rates. This is because of the fact that the more precise the returned results, the less capable of covering the broader spectrum of data they are (Manning et al, 2009). Thus, we attempt to create a balance between these two indices.

As regards the third hypothesis, we expect that our proposed approach achieves the maximum score (10) across all video categories. This will reflect the fact that the pre-specified time constraint (30 seconds) for the generated summaries is met strictly.

Finally, in regards to the last and the most important hypothesis, our method should be capable of achieving the highest scores from the *Overall Satisfaction* point of view across all six video genres. This metric is tied to the users' perceived experience and their satisfaction levels from a number of angles, as previously mentioned.

3.7.4.5. Statistical Significance

The assigned scores by different participants should be averaged for comparison purposes as was noted, but their statistical significance should be checked further. The independent samples t-test is used in our study in order to check if there are significant statistical differences between the measured metrics associated with each of those four independent variables pairwise. This tool is an appropriate mechanism to compare the achieved means for two groups of data for statistical purposes. As a result, we have adopted t-test analysis in all three studies to compare the mean opinions of participants in regards to different versions of a video summary. In our statistical analysis the results could be considered significant if $p<0.05$. It should be added that these mean scores are used in order to reflect the opinions of the whole sampled population.

3.8. Summary

In this chapter we firstly described the various paradigms that are being employed by different researchers. Initially, the concept of positivism as our chosen research paradigm and its major assumptions which justify its adoption were discussed. Further, in response to the identified research objectives in chapter 2, three research questions were formed. Since these extracted research problems were characteristically different, three distinct research types had to be chosen to address them. As a result, the use of Design Science Methodology capable of potentially covering all the undertaken research-based activities was justified. Later, in respect to the different phases of DSR methodology, a number of research methodologies were proposed and justified.

Secondary research was suggested to identify the knowledge gap and to develop a new technique for video summarisation purposes. Moreover, the Rapid Application Development methodology was adopted to design and develop the previously formed approaches into actual artefacts. Finally, experimental research was utilised to measure the effectiveness of our algorithms.

Chapter 4

Video Summarisation Based on Group Scoring

4. Overview

In this chapter, we address the first research objective of our research. To this end, initially, a brief review of video summarisation and some of the related abstraction techniques discussed earlier in chapter 2 are provided. This is followed by introduction of our user-centred video abstraction approach. Involvement of more than one video operator should be regarded as the main use case of this algorithm. The summarisation task will be performed in three steps as will be detailed later. An experimental study, as described in section 4.3, is then carried out in order to evaluate the effectiveness of the proposed tool. As a result, the video summaries from different categories generated using our summarisation tool are compared against the results produced by a number of automatic summarisation systems that adopt different approaches for abstraction. Lastly, conclusions are drawn in in section 4.4.

4.1. Video Summarisation

As was extensively discussed in chapter 2, in dynamic video summarisation (skimming), the most significant video segments are identified and extracted in order to produce shorter versions of an input video. Therefore, each original video should be segmented into a number of structural partitions, which can be as short as a frame or as long as a video scene. Thereafter, these constructional units should be compared against each other using available information sources. The required data for this comparison task could be obtained directly from their low or high level audio, visual and textual content or can be provided as metadata through human involvement.

According to our earlier discussion in that chapter, a number of automatic and semi-automatic video summarisation tools have been suggested by researchers in recent years to address this task. However, development of a technique with ability to produce effective summaries with a high level of users' satisfaction is still an unsolved problem. Thus, we

further investigated and highlighted some of the issues in regard to these existing systems and subsequently proposed our user-centred abstraction approach in response to the second research objective, namely "**To design, develop and evaluate a user-centred video summarisation algorithm based on group scoring".**

4.2. Proposed Video Summarisation Technique

In our work, a human-based group-based approach has been adopted to find the most valuable video segments to be included into a final summary. The Figure 4.1 illustrates the steps that should be taken based on our proposed approach to generate user-centered video summaries.

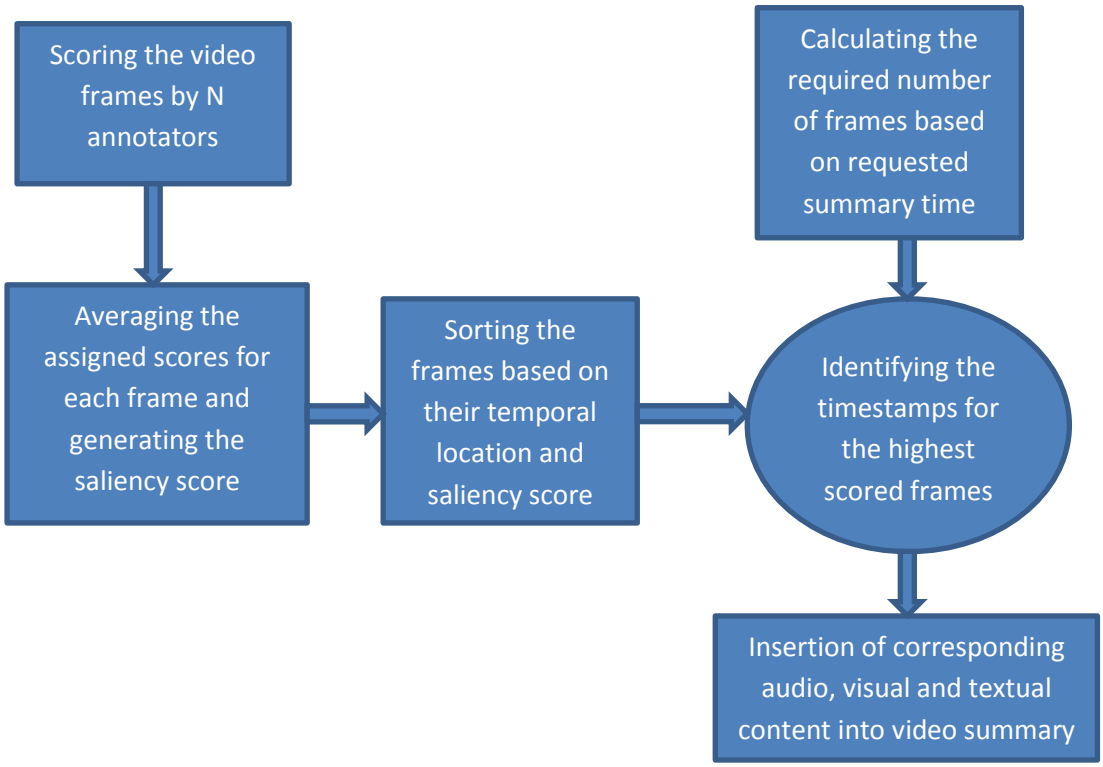


Figure 4.1. Chart describing the stages in our proposed summarisation approach

Considering the shortcomings of the existing fully-automatic summarisation techniques to detect semantic concepts to a satisfactory level and their drawbacks in terms of domain dependencies, user-based methods should be regarded as the best available option for

determining the most salient video segments. This is due to the human's capabilities to determine semantically meaningful video content. Furthermore, the human brain has the ability to assess and compare the quality of each section of the video in the context of the whole video, as opposed to most of the automatic abstraction systems reviewed in chapter 2.

However, the personal inclinations and preferences of different people can be dramatically different, which will have a direct influence on their content selection. Therefore, adopting a group of operators can increasingly reduce the effect of subjectivity of sole actors and it will smooth the final video summary towards more satisfactory results for a wider range of audiences. Moreover, there are a number of scenarios in which more than one operator is required and engaged in the video abstraction process. In order to create a sole video abstract, the video summaries generated by each of these operators should be compared to each other and a third party has to select the overlapping segments, which can be a very time-consuming procedure.

For instance, in the "Match of the Day" TV show in which highlights of the English Premier League football matches are shown, football pundits extract the most interesting scenes of a football match and include them into a summary. However, each of these pundits can produce their own version of summaries based on their personal interests and perceived significance of different sections of the video. Generating a single final summary, in which the views and choices of different experts have been contemplated and reflected aggregately, is another application of this method.

Accordingly, our three-step multi-annotators video summarisation method is proposed to address the mentioned issues. Initially, the video frames are labeled with the scores representing their saliency from a particular video operator's (annotator) point of view. Then, the assigned scores to each video frame by different annotators are averaged in order to produce a singular value for that frame as a saliency score. Lastly, the video segments with the highest saliency scores are extracted to be placed into a video skim in respect of a pre-set summary time constraint. Each of these stages is described in detail in the following sections.

4.2.1. Frames Scoring

In our approach, a group of short videos from different categories were presented to different operators. In the first instance, the operators watch the videos with the sole purpose of

familiarising themselves with the subject matter and do not score them. In the next step, the same individuals are asked to score those videos whilst watching them. To do this, they indicate the scores using a slider with the value range of 0 to 10. The operators score the video frames based on their personal interests and the perceived significance of the content they were watching. Figure 4.2 shows the interface of the scoring process of the video frames.

This group was also advised to consider the different available modalities (audio, visual and textual) for scoring purposes. Therefore, per each N available frames in the original video there will be N assigned scores between 0-10 per each operator. Thus, the most satisfying frames will be scored with 10 and the least important sections are graded 0. $FrameScore_{NM}$ represents the value allocated to the N^{th}frame of the video by the M^{th} scorer. As opposed to the Click2SMRY framework (Wu et al., 2011) in which the video sub-shots had to be categorised as either highlights or non-highlights, in our proposed approach, the panel of scorers is able to express their perceived importance level of each video frame.

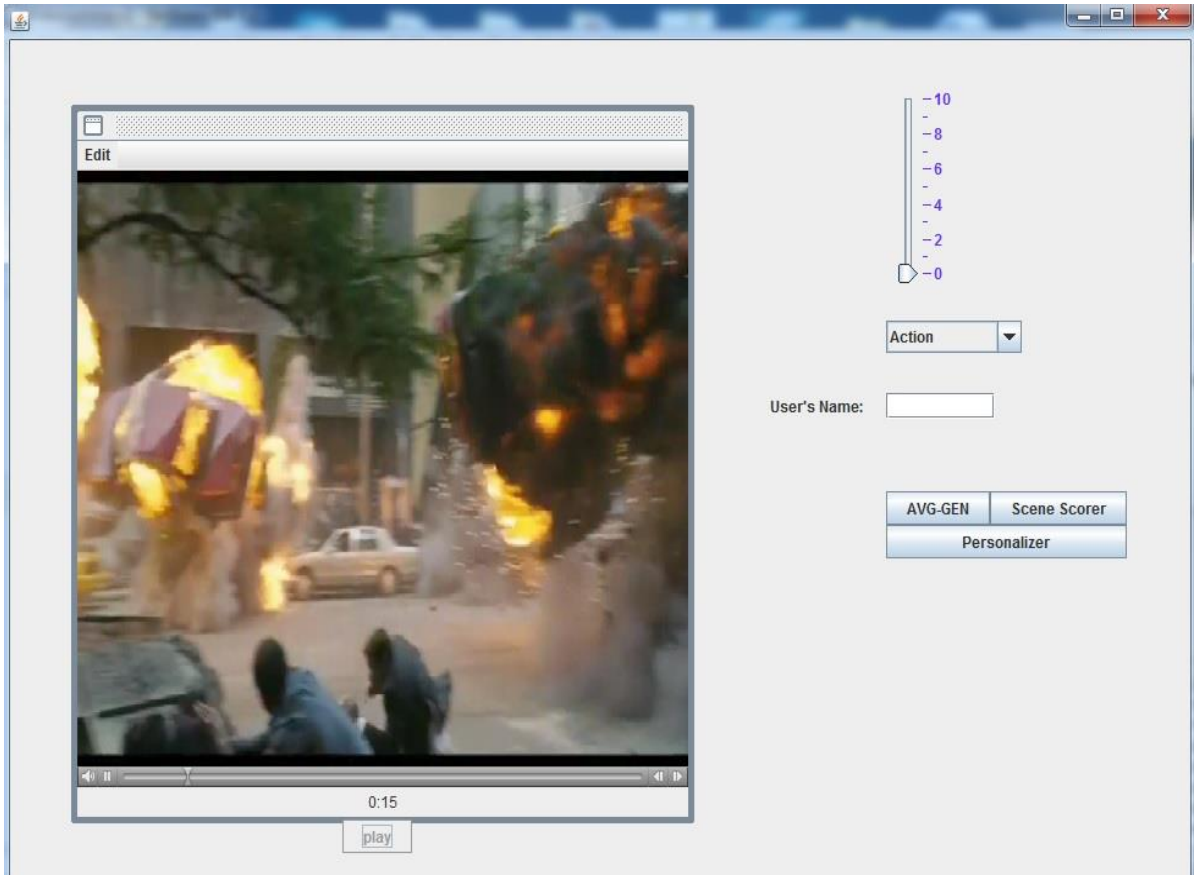


Figure 4.2. Video Frame Scoring

4.2.2. Frames Saliency Detection

In the next step, the scores generated by all operators for all frames are averaged and a single value is computed for each sole frame inside the original video. This represents the overall perceived quality of that particular frame across all M operators. $AvgFrame_N$ is therefore computed as:

$$AvgFrame_N = \frac{\sum_{N=1}^{M} FrameScore_{NM}}{M} \qquad (4.1)$$

The averaging process is thus employed to smooth the frame scores towards a less biased result by reducing the effect of dramatic differences in assigned scores to a particular frame.

4.2.3. Summary Generation

The target video summary time and the video frames frequency scale (number of frames in 1 second) are the elements to determine the number of extracted frames. $ReqNO$ calculates the required number of frames for extraction (according to equation 4.2) while $TarVidTime$ shows the required video summary time.

$$ReqNO = \text{TarVidTime(seconds)} \times \text{FramesFrequencyScale} \qquad (4.2)$$

In the final stage, the highest scored frames alongside the audio and textual content are selected and inserted into the final video digest. Thus, all frames are sorted based on their $ReqNO$ values. Considering the required number of frames, those highest scored frames will be selected to be added to a final list and to be sorted based on their time order in the original video. So, if *K* represents the frame number in the original video, *L* is a list of chosen frames while $SortedFrames$ is a collection to keep the chosen frames in ascending format according to their temporal locations and T_{Fj} represents the time-related position of frame F_j in original video.

$$L = \left\{ \mathrm{F_K} \,\middle|\, 0 < K < ReqNO \;\&AvgFra \geq AvgFrame_{\mathrm{U}_{i=1}^{\mathrm{N-ReqNo}}\mathrm{L'(i)}} \right\} \qquad (4.3)$$

$$SortedFrames = \left\{ F_j \,\middle|\, 0 < j < ReqNo \;\&\; T_{F_j} > T_{F_{j-1}} \right\} \qquad (4.4)$$

Using this sorted list, the temporally corresponding audio and text segments with those elected frames are copied from the original tracks into the video summary. Considering that

semantically and temporally approximate frames are mainly given similar scores, the number of sudden cuts in the generated summary drops significantly. As a result, more meaningful and comprehensible auditory and visual contents could be included in the final digest.

The algorithm that has been adopted for selection of the highest quality video segments is illustrated further in pseudo-code style (Figure 4.3) in the next page. In this approach, the original time-related locations for all the frames in $SortedFrames$ are retrieved in the first place. As a result, the timestamps for those frames in the input video, which should be transferred to the summary, are obtained. However, the algorithm initially checks if there are any temporally adjacent frames in the $SortedFrames$ list. Therefore, for those neighbour frames the timestamps (Start and Finish time) of the corresponding transferable content from the original video will be expanded from the first to the last frame of the group. Thus, the entire audio, visual and textual content (instead of the sole frame) from the original video in accordance to identified timestamps are copied into the abstract. This results in generation of video summaries with more coherent and continuous visual content in parallel to high quality adjacent aural data. In addition, the summary generation task will be accelerated noticeably by reducing the number of insertion tasks.

```
Create a new movie // to insert our extracted content inside
Create a new Audio Track // to insert the auditory content inside
Add the new Audio Track to new movie
Input a new Frame to List-of-Frames
Sort List-of-Frames // based on average scores
Set Required-frames-ratio to Required Summary length divided by Total Video length
Set Required-Frames-No to Total number of frames multiplication by Required-frames-ratio
  Do while counter is less than Required-Frames-No
  Input the next frame from sorted List-of-Frames into Final-Frame-List
  Increment counter
  End Do;
Sort Final-Frame-List // based on temporal location
  Do while there is still frame in Final-Frame-List
  Input a frame from Sorted Final-Frame-List
  Set timestamp1 to the current frame's temporal location in original video
    Do while distance of the Frame with the former is less or equal to frames' time interval
    Input the next Frame
    End Do;
  Set timestamp2 to the last extracted frame's temporal location in original video
  Insert the segments of the input video from timestamp1 to timestamp2 into new video
  Copy Audio track of the input video from timestamp1 to timestamp2 into new audio track
  End Do;
Convert the new movie into a file
```

Figure 4.3. The algorithm for selection of highest quality video

4.3. Experimental Evaluation

In this section, the experimental methodology which was adopted to assess the effectiveness of our recommended video skimming tool in response to the formed hypotheses is explained. The experimental procedure is composed of two distinct phases with two completely separate groups of participants. In the first stage, the summary for each input video clip is generated based on our suggested approach. This is followed by a comparison-based method to check the quality of the produced abstracts employing our tool against those generated by some other systems.

4.3.1. Generating the Summaries

A group of short videos (two minutes each) from six different video categories comprising, *Movie, Sport, Documentary, Advertisement, Music* and *News* genres were utilised to assess the quality of the proposed method. These are the videos that were previously discussed content-wise in chapter 3. 10 operators with different demographic details (six males and four females between the ages of 27 and 60 years old) were asked to watch each of these six videos first and to score each frame of the videos based on their personal interests and preferences. User-assigned scores for each frame were then averaged and a 30 seconds video summary was generated by aggregating the top-scoring frames of the respective video. In Table 4.1, sample frames from each of the video clips alongside the corresponding assigned scores, assigned by the first three operators, are presented.

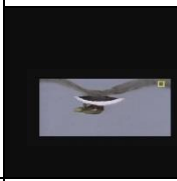
	DOC	MOV	ADV	NEWS	MUS	SPO
						
User1	8	7	5	4	6	9
User2	3	2	9	1	6	7
User3	7	7	5	5	4	8

Table 4.1. Assigned scores to sample frames by 3 users

4.3.2. Evaluation of Generated Summaries

To measure the quality of the generated summaries, a comparison method has been adopted. So, our generated summaries were compared against the abstraction results of the same videos, which were built by three automatic video summarisation systems. These systems perform the video abstraction task by analysing different modalities and employing different algorithms.

In the first technique (You et al., 2009), summarisation is based on audio-visual analysis. Shots are semantically measured using semantic audio importance analysis. This is complemented by face and text importance detection. Hence, other factors including camera motion, object motion and temporal motion coherence are also taken into account to build a semantic shot importance model.

In the second method (Evangelopoulos et al., 2013), the saliency of auditory, visual and textual information is analysed separately and integrated into a multi-modal saliency curve. Then, the most salient audio and video sub-clips based on a predefined skimming percentage are chosen for inclusion in the final summary.

However, in the third system (Boem et al., 2013), low-level visual features are adopted solely for abstraction purposes. The similarity between adjacent frames, face region, and frame saliency are computed to analyse the spatiotemporal saliency in a video clip. The spatial saliency is calculated based on ltti saliency and local entropy of the video and face detection measurement using the Viola Jones algorithm.

In order to assess the video summaries employing the mentioned techniques, all 6 input videos were submitted to the developers of these three techniques. They were asked to generate and return a 30 seconds video summary for each of the submitted videos using their developed tools. Subsequently, four summary versions for each input video were generated including the digests that were produced based on our tool.

The four created summaries of each of these six video categories were represented to 20 end users (10 Female and 10 Male between the ages of 25 and 55 years old). These users were different to the initial 10 operators used to create the user-centric summaries and had no prior knowledge regarding any of the systems, which had produced the summaries. After watching the original video and the summaries (the summaries were presented in a randomised order to

participants, to avoid order effects), they are asked to score each of these abstracts from four different perspectives consisting of *Recall* (Re), *Precision* (Pe), *Timing* (Ti) and *Overall Satisfaction* (OS) by scoring a corresponding statement in the questionnaires as detailed in Table 3.3.

The given scores for each of these measures are averaged over 20 users and the final figures for each of the video categories are used for comparison purposes. Table 4.2 illustrates the average results achieved by each tool across different categories (alongside the standard deviation). SM1, SM2, SM3 and SM4 show the results generated by, respectively, the first, second, third summarisation methods, as well as our proposed algorithm.

	SM1				SM2				SM3				SM4			
	Re	**Pe**	**Ti**	**OS**	**Re**	**Pe**	**Ti**	**OS**	**Re**	**Pe**	**Ti**	**OS**	**Re**	**Pe**	**Ti**	**OS**
DOC	8.1 (1.0)	7.5 (1.3)	9.3 (0.7)	**4.1 (0.8)**	7 (1.0)	6.5 (1.0)	8.1 (0.7)	**5.7 (1.4)**	6.3 (1.3)	6 (1.6)	6.7 (0.9)	**4.8 (1.1)**	6 (1.2)	6.9 (1.2)	10 (0)	**7 (0.7)**
MOV	8.2 (1.3)	8.7 (0.9)	9.2 (0.6)	**4.4 (1.1)**	7.5 (1.3)	7.2 (1.0)	7.7 (0.7)	**6.1 (1.5)**	4.3 (1.0)	4.1 (1.2)	6.2 (1.1)	**4.5 (1.3)**	7.1 (1.4)	7.1 (1.6)	10 (0)	**7.3 (1.2)**
ADV	7.6 (1.6)	7.6 (1.3)	9.3 (0.9)	**4.2 (1.4)**	6.6 (1.2)	5.9 (1.3)	8.2 (0.6)	**6.1 (1.8)**	7.4 (1.3)	7 (1.1)	6.3 (1.1)	**5.1 (1.6)**	7.6 (1.3)	8.6 (0.7)	10 (0)	**8.4 (0.8)**
NEW	7.3 (1.5)	7.1 (1.5)	9.1 (0.7)	**2.1 (1.1)**	6.5 (1.0)	6.2 (0.8)	7.6 (0.7)	**3.8 (1.2)**	6.1 (1.3)	5.2 (1.5)	6.4 (0.7)	**2.4 (1.5)**	5.9 (1.0)	6.9 (1.2)	10 (0)	**6.3 (1.6)**
MUS	7.1 (0.9)	7.6 (1.5)	8.6 (0.8)	**2.7 (1.0)**	6.7 (1.0)	6.4 (1.2)	7.4 (0.9)	**5.3 (1.6)**	6 (1.4)	6.1 (1.6)	5.8 (1.0)	**3.5 (1.1)**	6.4 (1.0)	6.8 (1.1)	10 (0)	**6.4 (1.7)**
SPO	7.7 (1.4)	6.7 (2.1)	8.6 (1.3)	**3 (0.9)**	6.1 (1.1)	5.8 (1.6)	7.9 (0.8)	**5.2 (1.8)**	4.5 (1.0)	3.5 (1.2)	6 (0.9)	**3.4 (1.4)**	6.7 (1.3)	7 (1.1)	10 (0)	**6.8 (1.0)**

Table 4.2. Average assigned scores to each summary from 4 perspectives

Generally, *Recall*, *Precision* and *Timing* rates for the first system across all six categories have been high. However, the *Overall Satisfaction* has been the lowest between all six videos. It could be due to the nature of this method, in which the audio and video are summarised separately. However, the extracted static keyframes are concatenated in a slide-show style and will only be combined with the summarised audio later. The second method

achieves some good results for particular categories including for the *Movie* and *Music Video*; however the performance was considerably domain-dependent. The results generated by our proposed method scored the highest marks in terms of *Overall Satisfaction* and *Timing* in all six categories in spite of some average *Recall* results for a number of categories. A more detailed investigation of these measures is provided in the next section.

4.3.3. Results

The research question that we are trying to deal with in this stage is to understand if our proposed method is effective enough or not? In response to this question, four hypotheses were formed relating to the measured metrics through questionnaires namely:

1- Our proposed method will generate the video summaries at an **acceptable** *Recall* rate.

2- Our proposed method will generate the video summaries at a **high** *Precision* rate.

3- Our proposed method will generate summaries by **strictly** meeting the *Time* constraints.

4- Our proposed method will generate the summaries with the **highest** *Overall Satisfaction*.

In this section we assess whether or not the hypotheses have been verified as a result of the experiment undertaken.

4.3.3.1. Recall

In this section we check the acceptability of *Recall* rate that has been achieved by our summarisation tool. As noted in the last chapter, we expect that the achieved scores by our system from this point of view to exceed the average scores of the other three tools in a number of categories. The comparison of our achieved results with the average scores of the other systems across six categories is displayed in Figure 4.4.

As shown in the chart, our approach managed to exceed the average score of the other three systems across three video categories namely, *Movie*, *Advertisement* and *Documentary*. In addition, the achieved score for *Advertisement* category is the highest among all the produced summaries (Figure 4.4).

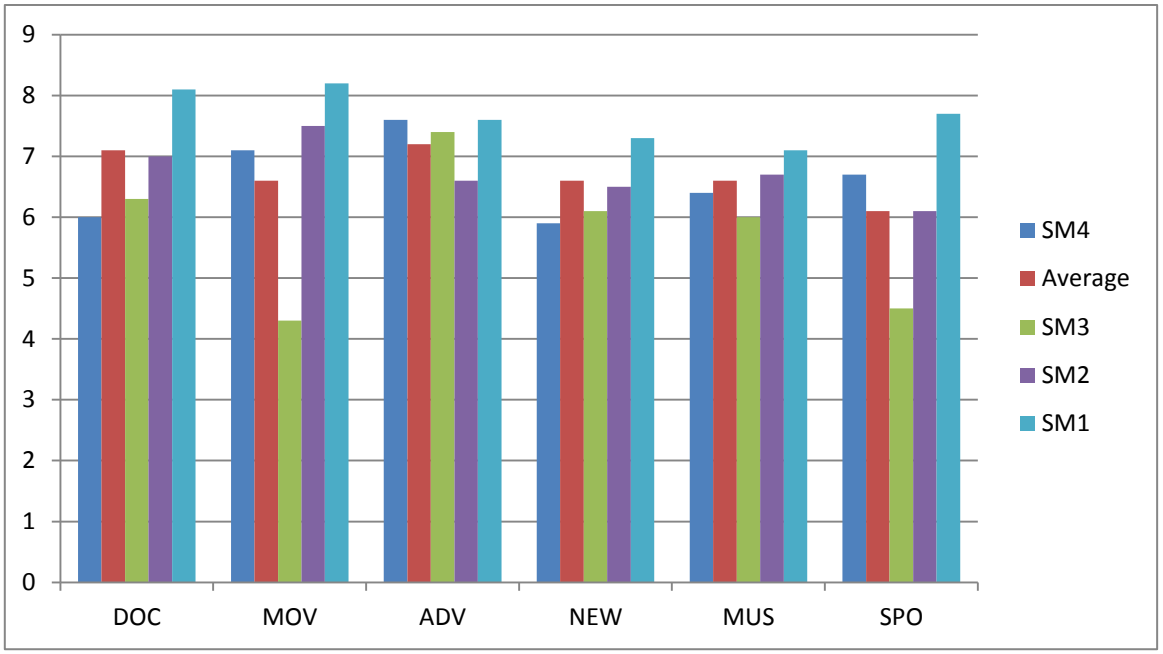


Figure 4.4. Comparison of our results against the other 3 tools for the *Recall* metric

Although in the other three categories we did not manage to better the average score, however, the differences between SM5 and average of the other three tools for these genres are significantly less than 1.5 units. Therefore, in response to the previously defined hypothesis, it can be claimed that our technique is capable of producing video summaries with **acceptable** *Recall* rate and the first hypothesis (H1) is verified. From this point of view, the first summarisation technique has achieved the best results compared to the other tools; this that can be attributed to the nature of its generated summaries. Since static keyframes were extracted from the original videos and were displayed in a slideshow style, a wider range of video segments could be covered therefore. However, this will impose a negative effect on overall quality of the video skims as will be discussed later. In contrast, the third system, which utilises the visual characteristics solely for generation of dynamic video skims, has been assigned the lowest grades in this respect.

Statistical Significance Analysis

The scores which were given to our video summaries from the *Recall* point of view is compared pairwise with the grades set achieved by any other tool to check if their population means are different. The significance level (p value) that is adopted to investigate the null hypothesis is set to $p=0.05$. Therefore, the achieved scores by any two tools are significantly different from each other if $p<0.05$.

A t-test analysis was adopted to investigate the statistical significance ($p<0.05$) of mean scores achieved by SM4 relative to the average marks obtained by the other three methods for *Recall* metric. The cells containing the significant values are shaded in Table 4.3. The scores assigned to the first summarization method are significantly better than the marks earned by ours statistically in all categories except *Advertisement*. In fact, our proposed technique did not manage to gain significant difference comparing to the other three tools (at the same time) for any of the categories. Nevertheless, our tool managed to achieve higher marks in comparison to average scores of the other three tools and subsequently the first hypothesis (H1) is verified.

	SM4-SM1 (Re)		SM4-SM2 (Re)		SM4-SM3 (Re)	
	t	p	t	p	t	p
DOC	-7.18	**<0.05**	-4.78	**<0.05**	-0.87	>0.05
MOV	-2.62	**<0.05**	-0.84	>0.05	9.45	**<0.05**
ADV	0	>0.05	3.44	**<0.05**	0.63	>0.05
NEW	-3.55	**<0.05**	-1.98	**<0.05**	-0.86	>0.05
MUS	-5.55	**<0.05**	-1.45	>0.05	1.22	>0.05
SPO	-2.16	**<0.05**	1.57	>0.05	5.80	**<0.05**

Table 4.3. Investigation of the statistical difference between the results obtained by our method and the other 3 systems from the *Recall* perspective

4.3.3.2. Precision

Measuring the effectiveness of our approach from the *Precision* angle is the aim of this segment. Achieving higher scores than the other tools' average marks in all genres alongside attaining the highest grade in a number of categories were defined as the indicator for a **high** *Precision* rate.

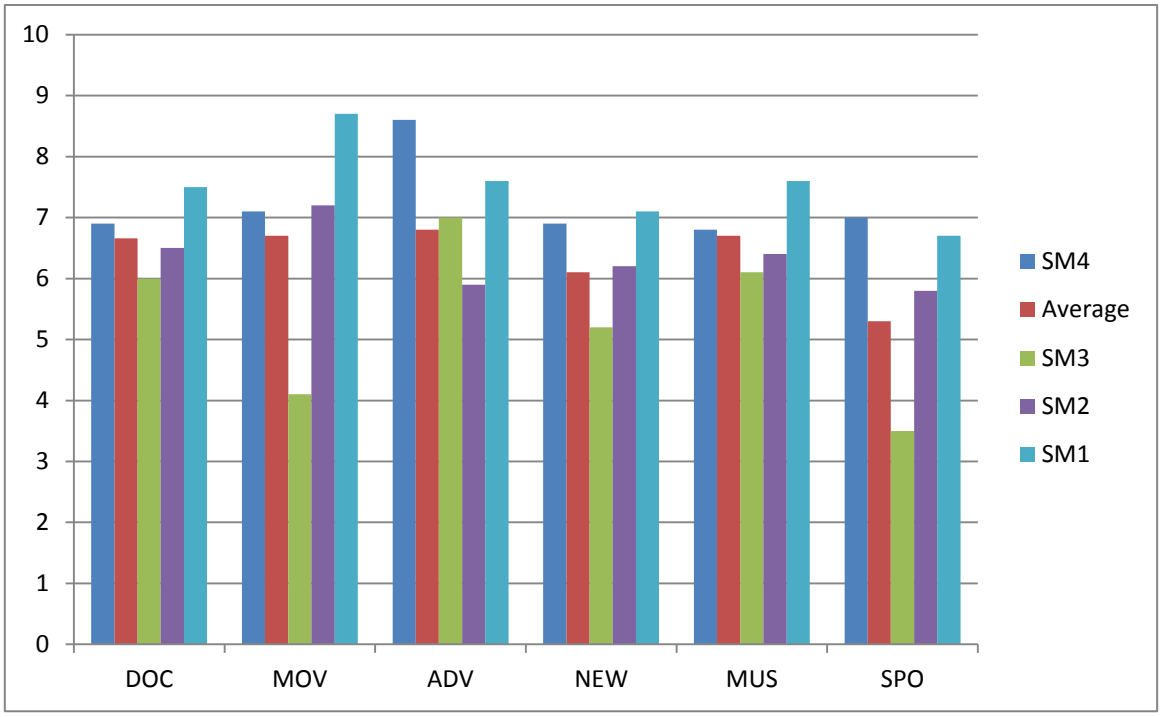


Figure 4.5. Comparison of our results against the other 3 tools for the *Precision* metric

As is depicted in Figure 4.5, our summarisation method managed to exceed the average scores achieved by the other three tools across all six video categories. In addition, the abstracts generated for *Advertisement* and *Sport* clips received the highest scores from the *Precision* perspective based on same table.

Since *Precision* is defined as a scale representing the percentage of essential retrieved video segments, it is conceptually closet to the main objectives of video summarisation tasks, in contrast to *Recall* rate. Consequently, in our work, a higher priority has been considered for this metric in comparison to the earlier index. The higher *Precision* scores in our system can be justified in accordance to the nature of our method. Since the highest quality frames (semantically and visually) are likely to receive the higher grades by the annotators, the probability to reflect the most significant visual contents will be boosted.

In addition, the availability of audio and textual data during the scoring process will assist the annotators in measuring the importance of a particular video partition. Similar to *Recall*, the first abstraction technique managed to receive some good results in this respect by scoring the highest in four video categories namely, *Documentary, Movie, News* and *Music videos*. This can be identically associated with the characteristic noted in the last section for this abstraction method. In fact, extracted keyframes from different segments of the video will

substantially increase the probability to reflect the most valuable visual content of the input clip. On the other hand, the third system has been marked the lowest for this metric akin to the *Recall* metric. Concluding, we can say that as far as H2 is concerned, SM4 delivered video summaries with **high** *Precision* and the H2 is verified.

Statistical Significance Analysis

A t-test analysis was adopted to investigate the statistical significance ($p<0.05$) of the mean scores achieved by SM4 pairwise compared with the average marks achieved by the other three approaches. The marks assigned to the summaries generated by our tool are significantly different to those produced by the third summaristion method for all the video categories from *Precision* perspective. In addition, SM1 achieved higher mean score with significant difference only for *Documentary* category for this metric. Although SM4 managed to obtain the highest grades for two video categories (*Sport* and *Advertisement*), only the results for the *Advertisement* video are statistically significant (Table 4.4). Nonetheless, H2 is verified.

	SM4-SM1 (Pe)		SM4-SM2 (Pe)		SM4-SM3 (Pe)	
	t	p	t	p	t	p
DOC	-2.29	**<0.05**	1.37	>0.05	2.90	**<0.05**
MOV	-1.72	=0.05	-0.31	>0.05	8.05	**<0.05**
ADV	2.93	**<0.05**	7.91	**<0.05**	5.11	**<0.05**
NEW	-0.54	>0.05	-2.00	**<0.05**	4.01	**<0.05**
MUS	-1.61	>0.05	1.37	>0.05	2.23	**<0.05**
SPO	0.40	>0.05	3.06	**<0.05**	8.23	**<0.05**

Table 4.4. Investigation of the statistical difference between the results obtained by our method and the other 3 systems from the *Precision* perspective

4.3.3.3 Timing

Here, we investigate whether the generated summaries by our proposed method meet the time constraint **strictly** or not? Further, we analyse how effective our system has been from this

point of view compared with the other techniques. The temporal length for our produced video abstracts should be exactly 30 seconds in order to satisfy the related hypothesis. Thus, the highest possible score (10) should be achieved by our summaries across all experimental video clips. The position of our approach from this angle against the other methods is demonstrated in following chart. The green bars show the highest score achieved by any of the three existing tools.

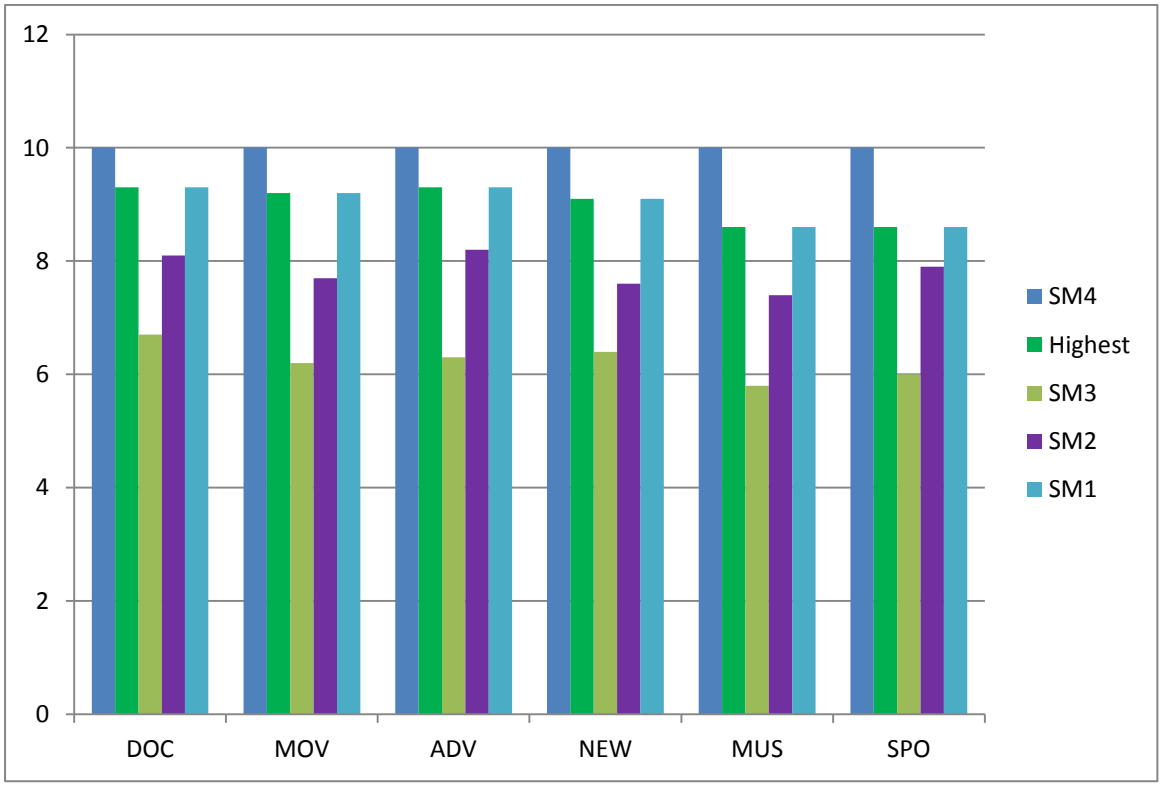


Figure 4.6. Comparison of our results against the other 3 tools for the *Timing* metric

Our video abstraction tool is the only one with the capability to generate the summaries that are rigorously meeting the pre-specified time constraint. The third and second tools have been graded the lowest respectively. Their produced abstracts exceeded the required summary length for each clip by at least three seconds, that is an extra 10 % of total demanded length. In fact, this will help the summaries to gain better *Recall* and *Precision* marks since they can potentially reflect more of the input video content. However, this is in total contrast with the fundamental parameters that have been defined for an effective video summarisation approach. Thus, the third hypothesis (H3) is verified.

4.3.3.4. Overall Satisfaction

This measure can be considered as the most important element in the determination of a video summary's effectiveness. This is due to the fact that it covers a broad spectrum of characteristics that can be attributed to a high quality video summary. As was noted before, continuity, connectivity and adjustability of audio and visual content alongside the clarity and informativeness of the presented subjects are all contributing factors, which can improve the end-user's perceived experience of the final video digest. In addition, this metric can express the effectiveness of our approach in respect of the other discussed elements as well. This can be justified based on the fact that a satisfactory video summary should be able to cover the different segments of the original video at an acceptable rate, while the most crucial ones are reflected at a high percentage. This should be done while respecting the predefined summary time constraint.

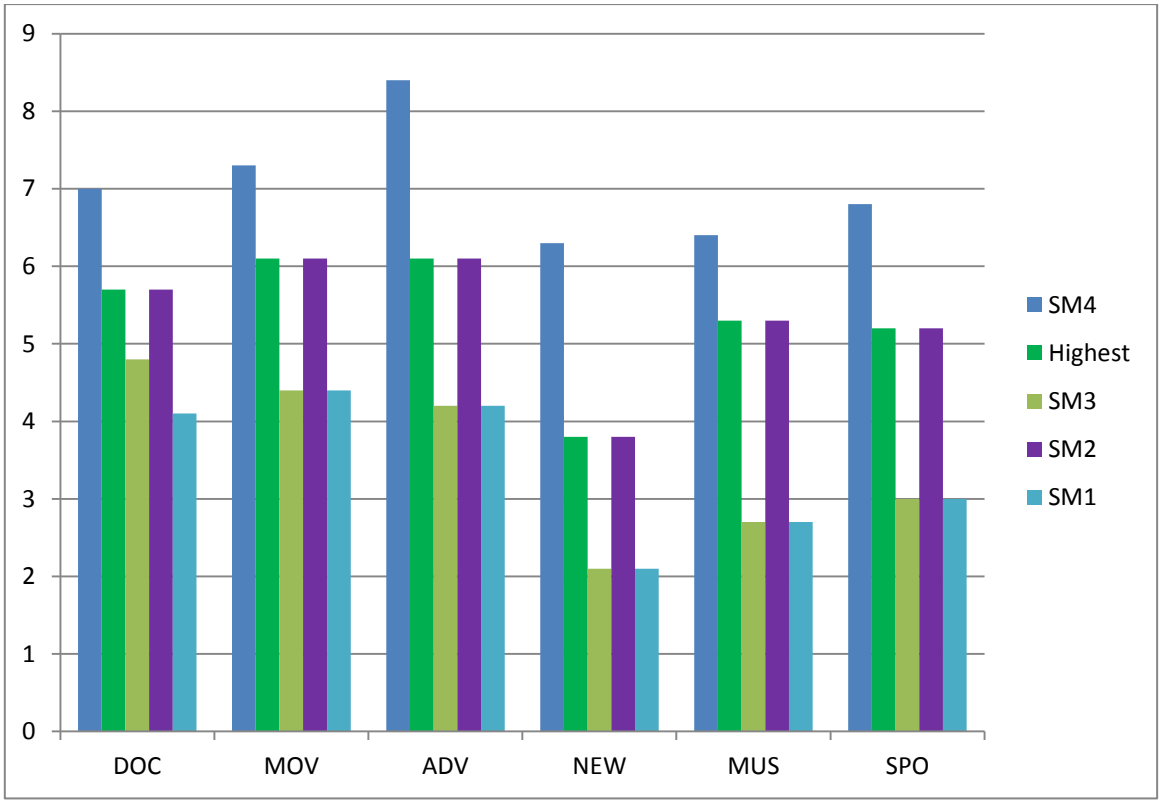


Figure 4.7. Comparison of our results against the other 3 tools for the *Satisfaction* metric

As a result, in this work, more emphasis is allocated to achieving high *Overall Satisfaction* scores of the created video skims. Further to our earlier discussion in section 4.3, we expect our recommended technique to obtain the highest results for all six experimental materials.

The significant difference between the results generated based on our technique and the highest score of those produced by the others (which is clarified in Figure 4.7) can potentially reflect the effectiveness of the suggested algorithm from this point of view. Based on Figure 4.7, our summaries have the highest rankings in all six categories and the nearest scores from this perspective have been achieved by the second summarisation tool, in which the video skims are produced based on aural, visual and textual data. On the other hand, the lowest grades have been assigned to the first abstraction tool, which despite attaining the highest scores in terms of *Recall* and *Precision*, the lack of audio data and its slideshow style have noticeably negatively affected participants' perceived quality.

In regards to our summary versions, a number of factors could be considered that can potentially justify these highest achieved scores. Initially, there is a good balance between *Recall* and *Precision* rates in the built abstracts, while the temporal limit is addressed properly. In addition, since temporally approximate frames are allocated similar scores then the contents can be represented in a more continuous manner.Another influential factor, which should be regarded, is the audio quality of our summaries which is much higher than the other three tools in terms of adjustability to visual content and its clarity. Eventually SM4 by achieving the highest scores in terms of *Overall Satisfaction* across all six video categories verified the fourth hypothesis (H4). In the next section, the statistical validity of the users' perceived quality of our technique will be validated.

Statistical Significance Analysis

Our method managed to achieve the highest scores for all the video categories from the users' *Overall Satisfaction* aspect (as the most important metric). Therefore, a mechanism should be employed to check the statistical validity of its assigned marks. In order to check the statistical significance between the grades achieved by our system and the other three tools from *Overall Satisfaction* perspective, t-test analysis was undertaken. The results from this test (Table 4.5) show that there are statistically significant differences between user ratings given to the summaries created by SM4 (our user-centric summarisation tool presented in this paper) and those created by the three other automatic summarisation tools from users' perceived overall quality of the video (i.e. *Overall Satisfaction*). The significance of these differences between the grades achieved by our tool and those gained by SM1 and SM3 are noticeable considering the generated t-values. In addition, among all the categories, News assigned scores from this perspective has the highest significant difference.

	SM4-SM1 (OS)		SM4-SM2 (OS)		SM4-SM3 (OS)	
	t	p	t	p	t	p
DOC	10.72	**<0.05**	4.95	**<0.05**	7.78	**<0.05**
MOV	7.02	**<0.05**	2.53	**<0.05**	9.22	**<0.05**
ADV	11.15	**<0.05**	5.51	**<0.05**	9.26	**<0.05**
NEW	14.63	**<0.05**	9.56	**<0.05**	15.58	**<0.05**
MUS	7.24	**<0.05**	3.94	**<0.05**	6.46	**<0.05**
SPO	11.10	**<0.05**	5.05	**<0.05**	12.49	**<0.05**

Table 4.5. Investigation of the statistical difference between the results obtained by our method and the other 3 systems from the Overall Satisfaction perspective

4.4. Conclusion

In this chapter, a number of existing summarisation methods was reviewed and a novel, user-centric technique for video summarisation was proposed. Thus, in our work, a group of operators are employed to score the video scenes as they are watching the videos. This scoring procedure is based on the available information from different modalities. In the next step, their scores are combined to come up with a single value for each video frame. This is the score which represents the saliency of a particular video frame in the context of the whole video. The highest scored frames alongside the associated audio and textual data (while meeting the 30 seconds time constraint) are extracted to be inserted into the final summary.

Our recommended method was evaluated by employing 20 end-users to compare its generated results against the summaries created by three existing automatic summarisation systems. The experimental results indicated that the proposed approach is capable of delivering superior outcomes in terms of *Overall Satisfaction* and *Precision* with an acceptable *Recall* rate, indicating the usefulness of involving user input in the video summarisation process. Finally, we also identify the production of personalised summaries as part of our future endeavours. In fact, with incorporation of the end-users' priorities towards different video segments, more satisfactory video summaries with higher precision can be generated. This will be explained comprehensively in the next chapter.

Chapter 5

Personalised Video Summarisation Based on Group Scoring

5. Overview

In this chapter, we address the second research objective identified in chapter 2. Firstly, we concisely review the concept of personalised video summarisation in section 5.1 then explain how the user-centred approach introduced in the last chapter will be the basis for generation of personalised video summaries. Accordingly, the summarisation task will be achieved through a number of consecutive stages described in section 5.3. Later, in section 5.4, the procedure to understand the end-users preferences towards the different video scenes is explained. Afterwards, the frames scores for different video segments will be updated based on the captured end-users' priorities towards these video partitions. Eventually, based on the pre-defined skimming time, the highest scored video frames will be extracted and included into the personalised video summaries in accordance to the approach demonstrated in the previous chapter. In order to evaluate the effectiveness of our proposed method, we employ the same methodology used in chapter 4. However, we will compare video summaries generated by our recent system against the results produced by our previous method (SM4) in addition to the summaries from the other three automatic tools.

5.1. Personalised Video Summarisation

The concept of personalisation in the context of video summarisation was defined as a process to incorporate end-users' personal interests and inclinations into the summary generation task, according to our earlier discussion in chapter 2.

As we highlighted on a number of occasions before, one of the major components in any video summarisation model is the segmentation of the input video into structural units. Further to our extensive discussions in chapter 2, different approaches have been employed

by researchers in an attempt to personalise the video summaries' content. However, in all of them, a mechanism has been applied in order to understand the users' priorities towards the identified video segments. The required information for these purposes can be extracted from the users explicitly or in an implicit manner. These retrieved data could be applied to initial video summaries to customise the results in accordance to the audiences' preferences. In spite of numerous models that have been suggested for this research topic, establishing a technique in which viewers' priorities are effectively integrated in the generated video summaries is still a challenging topic. thus, we further introduce our user-based personalised video abstraction technique in respect of the third research objective elaborated in chapter 2, namely, "**to extend the work of previous objective and design, develop and evaluate a personalised video summarisation algorithm based on group scoring**''.

5.2. Personalised Video Summarisation by Group Scoring

In the previous chapter, we described an approach to video summarisation based on a group scoring method, in which original video frames are scored by a number of video scorers and the assigned scores averaged to produce a singular value for each frame. A group of frames with the highest average scores are then chosen to be inserted into the final summary. In this approach, the required number of video annotators could be varied based on the different use-case scenarios. The proposed method was evaluated in an experimental study and showed the capability to yield promising results (vis. a vis. machine-generated approaches) in six different video categories. However, the generated summaries for all of the end-users were identical and their individual preferences were not considered in the summarisation process. Accordingly, in this chapter, we try to develop an approach to personalise the final summaries to the individual end-user's expectations, and thus to produce a better user experience. The new recommended video abstraction method is composed of three major phases which will be explained comprehensively in their corresponding sections of the chapter. Firstly, the video segments are enriched by video annotators through annotations and scoring. Later, a mechanism will be adopted to capture the end-users' priorities towards the identified video partitions. Finally, the video summaries will be generated in respect to the updated frame scores. The diagram representing the different stages in our novel approach is shown in Figure 5.1.

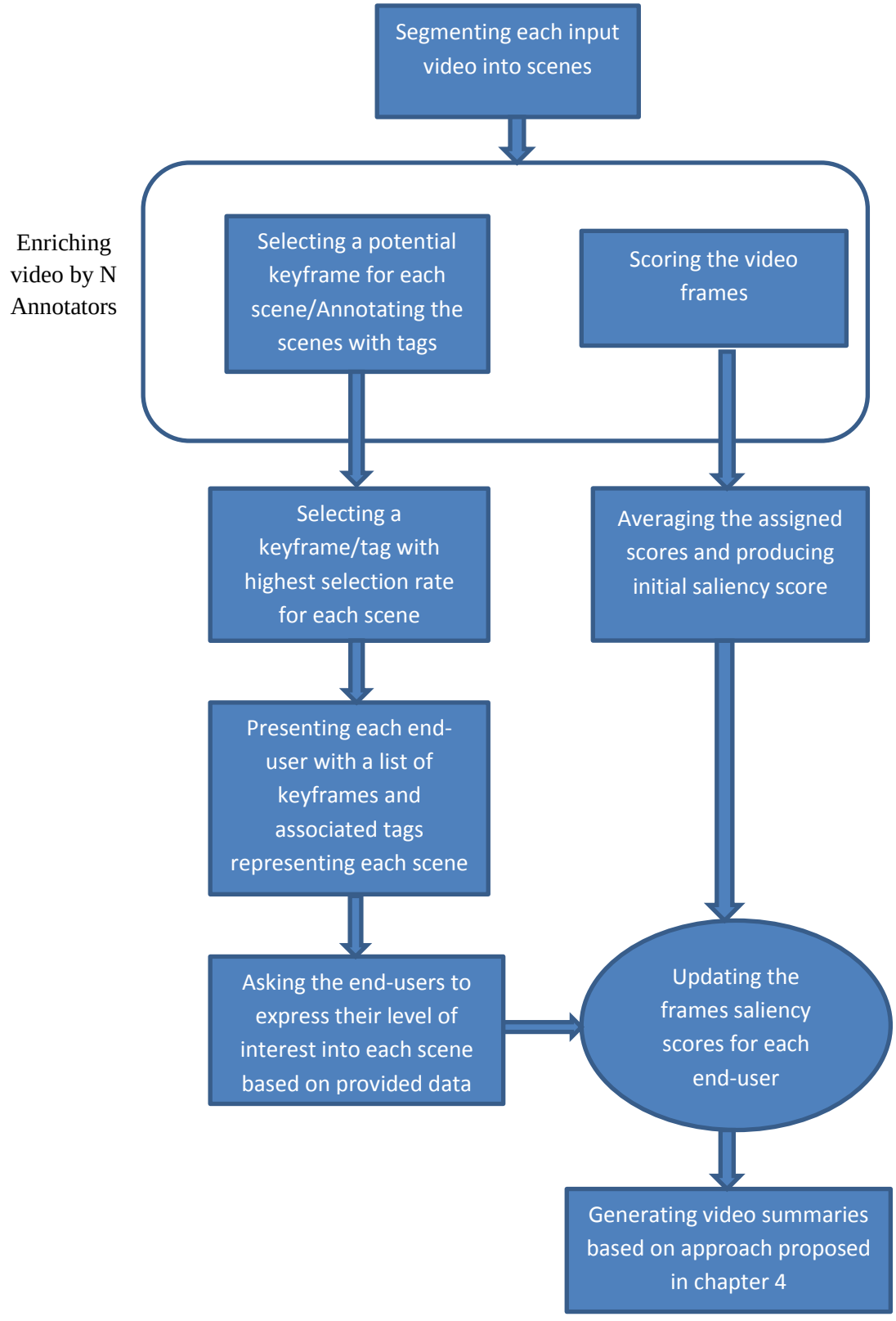


Figure 5.1. chart describing the stages in our personalised video summarisation approach

5.2.1. Video Segments Enrichment

In this stage, a semi-automatic procedure is applied for enrichment and scoring purposes. In the first step, the original videos are segmented into a number of scenes (group of semantically and visually similar frames). Later, each scene is enriched with a group of audio and visual tags and the appointment of a representative keyframe. The following section reveals the approach that is used for the video segmentation task.

5.2.1.1. Scene Boundary Detection

AVcutty (avycutty.de 2014), as a typical scene boundary detection tool, has been adopted to determine the timestamps for each contributing scene. It should be reminded that each scene in the context of a complete video plays the same role as a paragraph in a whole text. Therefore, there should be a semantic and visual correlation and cohesion between the constituting frames of a particular scene. The mentioned tool utilises the colour and motion features of the video frames for scene change detection purposes. The required minimum time length for each scene will be set to three seconds. Thus, any identified video scene with shorter length will be added to the next scene. This facilitates scoring and annotating of the original video by reducing the number of unnecessary pauses during the enrichment process, which is defined in the next section.

5.2.1.2. Video Scenes Annotation and Scoring

In this stage, video operators are asked to score and enrich the video segments based on the auditory, visual and textual content of the video. The video annotators score video frames 'on the fly' in a 0-10 range using the slider tool (similar to the approach in chapter 4). Using the identified timestamps for the scene boundaries, the videos will be paused automatically at the end of each scene and the video annotators immediately will be prompted to annotate the video scene using the provided graphical user interface (while the scoring process is stopped). The user interface that will be employed for this purpose is shown in Figure 5.2. The video scorers can optionally enrich the video scenes while the videos are halted, by assigning audio and visual tags to each scene. These tags could contain information regarding the significant events, objects and any activities in the corresponding video segment. The video scorers have the possibility to choose the previously assigned tags (by former scorers) or to add new ones based on their personal perceptions and priorities to the scenes. Once the annotation process

for one scene is finished, the scorers will then be engaged in scoring the video frames for the following scene using the slider tool. By re-starting the video, the initial frames from the upcoming scene are likely to be scored with unwanted grades. This is due to a predictable minor delay from the time, in which video annotators have to observe and evaluate the contextual significance of the opening frames (of the following scene), till the point they can actually start scoring. Therefore, to minimise the negative effect of this lag, a new pre-computed value was dynamically calculated and assigned to the slider tool each time that a scene starts. In order to produce this value, a score was computed for each scene, by averaging the previously assigned scores from the former annotators to the whole frames of that particular scene. Any recent assigned scores from new scorers will update these computed average scores.

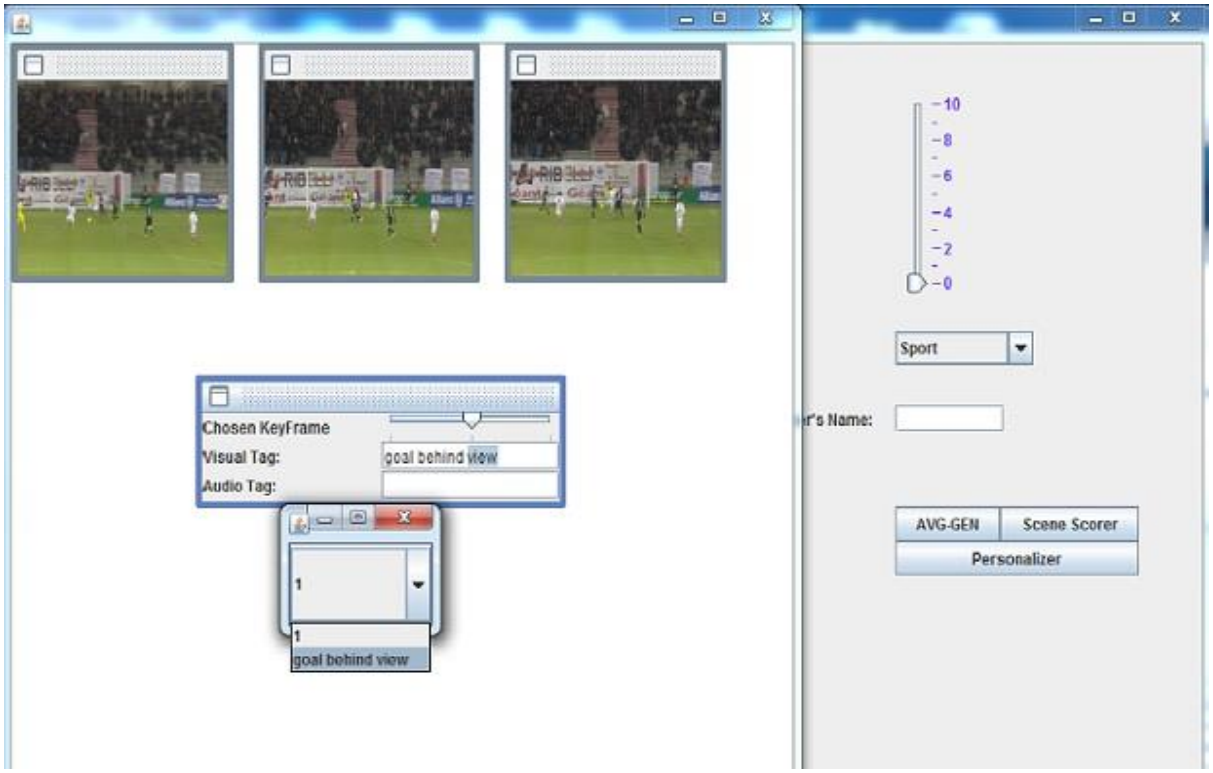


Figure 5.2. The adopted tool for annotating the video scenes

5.2.1.3. Scenes Key-Frames Selection

During the scene enrichment stage, the annotators are also presented with a set of three candidate keyframes at the end of each scene. The video annotators are asked to elect the one that they personally perceive as having the highest quality to represent and summarise the

semantic and visual content of that scene. For extraction of these three nominated keyframes, each video scene has to be fragmented into three temporally equal shots in the first place; each shot will be represented by a keyframe (to improve the coverage rate of any visual content changes in whole scene). In order to select a keyframe for each of these three identified video shots, two criteria should be considered. First, the frame has the highest assigned score between all the existing frames of that shot. Second, the candidate frame is temporally located in the middle of each shot. Therefore, between all the previously highest scored frames of each shot, the frame which is temporally closer to the centre of that shot will be introduced as a potential keyframe for that video shot (to increase the likelihood of extracting more visually significant and stable frames). These three nominee frames from each scene are then compared against each other from two different perspectives. First, their visual content attractiveness and richness should be considered. Second, their capabilities to reflect the semantic concepts of the corresponding video scene have to be taken into account. Finally, for each scene, the candidate frame that has the highest selection rate awarded by the different annotators (video operators) will be selected as the representative keyframe. In Figure 5.3, the graphical interface that will be used by the video scorers for audio-visual tags assignment and keyframes extraction is illustrated.

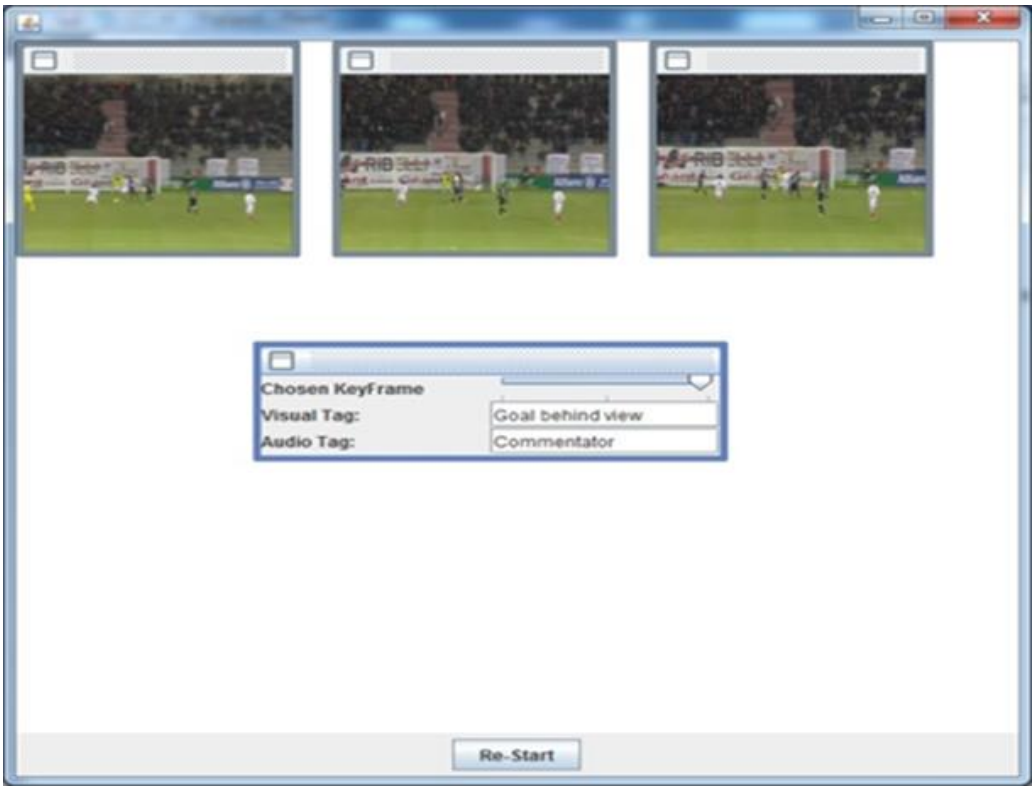


Figure 5.3. Interface for tags assignment and key-frame extraction

5.2.2. Capturing the Users' Priorities

In the section we propose an approach, which provides the foundation for personalising our video summaries. As a result, in this phase a mechanism, which is responsible for capturing an end-user's priorities towards different video partitions in a particular video, is introduced. Therefore, prior to the generation of the final video abstracts, the end-users will be provided with visual and textual information regarding the content of the existing video scenes. The primary goal here is to prioritise the video segments based on user-obtained preferences and priorities.

Accordingly, in our method, we try to provide the audience with a concise overview of all the contributing scenes of a video in order to produce the user-tailored summaries. Therefore, in the first stage, a list of representative keyframes with their associated visual and audio tags is presented to the end users. Here, each of the displayed representative frames corresponds to a single video scene (these are the delegate keyframes chosen by most of the video operators in the previous stage), while attached auditory and visual information to each keyframe correspond to the mostly verified tags for that scene by the different video scorers (one audio content tag and one visual content tag per each scene). Therefore, an overview of various audio, visual and textual contents of a clip that occur in the different video segments can be presented to the viewers in prior to the generation of any video digest. As a result, users will have the option to check and select their favourite segments in a fast way without the need to watch the entire video.

End users will be asked to express their level of interest in each video scene, based on the displayed video frames and the corresponding tags, using the provided slider tool (Figure 5.4). The representative keyframes alongside the explanatory associated tags can potentially inform the audience in regards to the semantic, aural and visual content of each participating video scene. Thereafter, the end-users are provided with a facility to choose three different priority levels in respect of each scene. Level 0 has been considered for the scenes with the lowest level of significance to them, while level 1 is awarded to scenes with medium importance that are preferred to be included into final abstract. Level 2 designates the scenes that users found the most attractive and they feel should be definitely included into their version of the video summary. In the next section, the mechanism adopted to assimilate these extracted data in the summarisation task will be clarified.

Figure 5.4. A GUI for understanding the users' priorities towards scenes

5.2.3. Updating the Frame Scores

In this phase, the initial generated average scores of the frames, assigned by the video scorers are updated based on the personal interests of each end-user, as captured in the last stage. This is the saliency score, which was computed for each video frame in accordance to the approach introduced in the previous chapter. Therefore, based on the user-selected priority level for each scene, the primary average scores are updated accordingly.

The scores of frames belonging to scenes with a level 0 of interest will not be altered at all since they correspond to scenes with the least degree of significance for that particular viewer. However, for scenes with a level 1 of priority, the grades for the frames, whose primary assigned scores are the highest among the frames of that scene, will be increased by 20 percent (to a maximum value of 12). This is done in order to potentially escalate the probability of incorporation of the highest quality frames of those scenes into the eventual video digest.

The updated mark for the frames belonging to the scenes with the highest level of priority (i.e. level 2) for a particular end-user will be recalculated in a different format. The grades for

the frames who were scored the highest in each scene, will be upgraded to the maximum possible value (12). This would considerably increase the chance of definite inclusion of the highest quality segments of those particular scenes in the final summary. However, the marks for the frames of these scenes, whose scores are not the highest but nonetheless manage to exceed the respective scene's average scores, will be boosted by 20 percent as well (to a maximum of 12). The scores for the remaining frames of these scenes will remain unchanged. The algorithms that have been adopted for updating the frame scores are shown in Figures 5.5.a and 5.5.b. As will be revealed in the next section, the highest scored video frames are selected for insertion into the final summary.

```
Do For total number of identified scenes
Input the next scene's time boundary
Set the current scene's Average-Score to Zero
Set the current scene's Highest-Score to Zero
Set current scene's Total-Score to Zero
  Do while the frame belongs to the current scene's time boundary
  Input next frame
  Add the frame's score to the current scene's Total-Score
  Increment the counter
      Do if the frame's score is larger than the scene Highest-Score
      Set Highest-Score to the current frame's score
      End Do;
  Set the current scene's Average-Score to Total-Scores divided by counter
  Store the current scene's Average-Score in database
  Store the current scene's Highest-Score in database
  End Do;
Set counter to Zero
End Do;
```

Figure 5.5.a. Pre-processing stage prior to upgrading the frames' score for a user

```
Input the user's priority list for a particular clip
Input List-of-Frames
 Do for all the frames existing in List-of-Frames
  Do if the frame belongs to the scene with priority level 2
   Do if the frame's score is equal to its corresponding scene's Highest-Score
   Set frame's score to 12 // maximum possible score
   End Do;
   Do Else if the frame's score is larger than its corresponding scene's Average-Score
   Upgrade the frame's score by 20%
   End Do;
   Do Else
   Do not alter the frame's score
   End Do;
  End Do;
  Do Else if the frame belongs to the scene with priority level 1
   Do if the frame's scene is equal to its corresponding scene's Highest-Score
   Upgrade the frame's score by 20%
   End Do;
   Do Else
   Do not alter the frame's score//only upgrades the frames with highest score in that scene
  End Do;
  Do Else// if the frame belongs to priority level '0'
  Do not alter the frame's score
```

Figure 5.5.b.The algorithm for updating the frames score based on priority level

5.2.4. Generating the Personalised Summary

In the final step, the personalised video summaries are produced based on the updated frame scores. In accordance to the summarisation method based on group scoring described in chapter 4, the video summaries will be generated on the basis of the most visually and semantically salient frames. As a result, considering the required number of frames which can be established based on the pre-specified summary time, the highest scored frames alongside the corresponding audio and textual content will be selected to be included into the final list. Later, the contributing segments will be sorted in respect of their temporal locations in the input video clip.

Considering the fact that the frames belonging to more popular scenes (for a particular audience) can potentially achieve better scores, we can expect that more personally satisfactory outcomes can be obtained using our technique. In addition, as was mentioned in the last chapter, since the semantically and temporally close video frames are usually scored similarly, the consistency and continuity level in the generated abstracts could be improved noticeably.

In the next section, we undertake another experimental study to evaluate the effectiveness of our recent video summarisation approach. As will be discussed later, we have compared our new method against our previous one in addition to the other three automatic summarisation tools.

5.3. Experimental Evaluation

In this section, the experimental procedure followed to evaluate the effectiveness of our currently proposed video abstraction technique will be explained. Similar to the previous chapter, the evaluation process is composed of two distinct phases with two different groups of participants, namely video annotators and end-users. Video annotators are responsible for scoring and enriching the video segments, while end-users are those for whom we attempt to generate personalised abstracts in accordance to their expectations. Each of these two phases will be discussed comprehensively in the following sections.

5.3.1. Video Segments Enrichment and Scoring

Akin to our work in chapter 4, 10 operators with different demographic details (5 males and 5 females between the ages of 25 and 55 years old) were adopted to score and enrich the video frames based on the procedure introduced in section 5.2.1. The experimental videos are identical to those, which were discussed in chapter 3. The video operators' assigned scores were averaged and a singular value was produced for each frame. In addition, a representative keyframe alongside the informative audio-visual tags will be chosen for each video scene based on the annotators' selections.

5.3.2. Users' Priorities Extraction

In order to assess the quality of our personalised video summarisation approach, the generated results have been compared against the video abstracts produced by four other systems. However, prior to summarisation of the input videos, the end-users' priorities towards the identified scenes in each of those six videos should be captured. As a result 30 users from different demographic backgrounds (18 males and 12 females with age ranging between 24 and 58 years old) are recruited to express their level of interests towards the scenes in each input video. As a result, their priorities towards exiting video scenes will be captured according to the method explained in section 5.2 and video summaries accordingly generated for each of those six video clips.

5.3.3. Evaluation of Generated Summary

The same videos were skimmed by four other abstraction tools: three of them summarise the videos automatically by applying statistical and mathematical algorithms, while the fourth tool functions semi-automatically based on human involvement (our proposed tool in the last chapter). The six original video clips alongside their five summary versions created by five existing tools (including the personalised summaries generated for each specific user using our proposed technique) were presented to the same 30 end-users that we previously employed for personalisation purposes.

These five summaries from each category were shown to the users in a random order so as to minimise order effects. Moreover, no information regarding the underlying summarisation tools, which was employed to generate the video summaries, was revealed to participants.

After watching the original video and the summaries the users were asked to score each of the generated abstracts awarding marks between 0 (worst video summary possible) to 10 (best video summary possible), from 4 different perspectives consisting of *Recall* (Re), *Precision* (Pe), *Timing* (Ti) and *Overall Satisfaction* (OS), analogously to the adopted mechanism in chapter 4. The given scores for each of these measures were averaged over 30 users and their mean values (alongside their standard variation) for each of the video categories are given in Table 5.1. SM1, SM2, SM3, SM4 and SM5 indicate the average achieved scores by, respectively, the first, second, third, fourth and our currently proposed summarisation methods.

	SM1				SM2				SM3				SM4				SM5			
	Re	Pe	Ti	OS	Re	Pe	Ti	OS	Re	Pe	Ti	OS	Re	Pe	Ti	OS	Re	Pe	Ti	OS
MOV	7.8 (1.7)	7.6 (1.1)	9.1 (1.0)	**4.1 (1.5)**	7.0 (1.2)	7.5 (1.1)	7.8 (0.8)	6.3 (1.7)	4.3 (1.1)	4.4 (1.3)	6.5 (1.2)	**4.0 (1.8)**	7.1 (1.3)	6.8 (1.7)	10 (0)	**7.2 (1.4)**	6.5 (1.4)	8.3 (1.1)	10 (0)	**7.9 (0.9)**
ADV	7.5 (1.9)	7.7 (1.0)	9.0 (0.9)	**3.9 (1.7)**	6.0 (1.4)	5.6 (1.5)	7.2 (1.2)	5.4 (2.0)	6.8 (1.5)	6.5 (1.5)	6.3 (1.2)	**4.1 (1.4)**	7.1 (1.2)	8.2 (1.1)	10 (0)	**7.8 (0.9)**	7.5 (1.1)	8.7 (0.8)	10 (0)	**8.3 (0.9)**
DOC	7.7 (1.3)	7.1 (1.4)	9.1 (0.7)	**4.3 (1.2)**	7.3 (1.1)	6.9 (1,2)	7.9 (0.8)	5.8 (1.7)	5.1 (1.5)	6.1 (1.5)	6.7 (0.8)	**4.5 (1.1)**	6.7 (1,2)	7.1 (1.0)	10 (0)	**7.2 (0.8)**	6.8 (1.5)	7.9 (1.0)	10 (0)	**8.0 (1.0)**
NEW	6.4 (1.7)	6.7 (1.4)	8.6 (1.0)	**2.0 (1.1)**	6.1 (1.4)	5.8 (1.2)	7.7 (1.0)	3.4 (1.4)	5.3 (1.6)	5.1 (1.7)	5.9 (0.8)	**1.9 (1.3)**	6.4 (1.1)	6.7 (1,3)	10 (0)	**6.1 (1.5)**	6.6 (1.5)	7.5 (1.3)	10 (0)	**7.1 (1.3)**
SPO	6.9 (2.4)	6.0 (1.5)	8.3 (1.5)	**3.4 (0.8)**	5.8 (1.2)	5.8 (1.4)	7.8 (1.0)	5.4 (1.9)	4.5 (1.3)	3.8 (1.1)	5.7 (0.9)	**4.1 (1.5)**	6.9 (1.3)	7.4 (1.1)	10 (0)	**6.9 (1.2)**	6.5 (1.5)	7.8 (1.0)	10 (0)	**7.4 (1.2)**
MUS	7.7 (1.0)	6.8 (1.4)	8.5 (1.2)	**3.1 (1.4)**	6.8 (1.2)	6.4 (1.2)	7.9 (0.9)	5.4 (1.7)	5.8 (1.7)	5.7 (1.7)	6.2 (1.1)	**3.5 (1.3)**	6.5 (1.3)	6.8 (1.1)	10 (0)	**6.3 (1.5)**	6.2 (1.3)	7.6 (1.5)	10 (0)	**7.2 (1.5)**

Table 5.1. Average assigned scores to each summary from 4 perspectives

Our proposed method has been scored highest from the *Precision* and *Overall satisfaction* point of views across all six existing categories. High *Precision* scores can justify the effectiveness of our method in producing personalised result, as it can indicate that the video segments with higher priorities to each individual end-user have been identified and inserted into the final digest. In addition, our current technique managed to deliver the best quality video digest among all six categories based on the average *Overall Satisfaction* marks. Generally, the SM1 tool generates some good results in terms of *Recall* and *Precision*, however, the nature of this method leads to lower grades in terms of *Overall Satisfaction*. The results from our previous technique enjoyed acceptable user ratings over six different

categories. However, lower scores for *Precision* and *Overall Satisfaction* can be attributed to the inability of this method (SM4) to actually generate personalised content.

5.3.4. Results

In this section we are seeking to assess the effectiveness of our proposed approach in response to the four previously identified hypotheses.

5.3.4.1. Recall

Based on the pre-specified hypothesis in chapter 2, the *Recall* ratio for generated video summaries based on our recently suggested approach should exceed the average scores achieved by the other four techniques for at least three categories, while the distance for

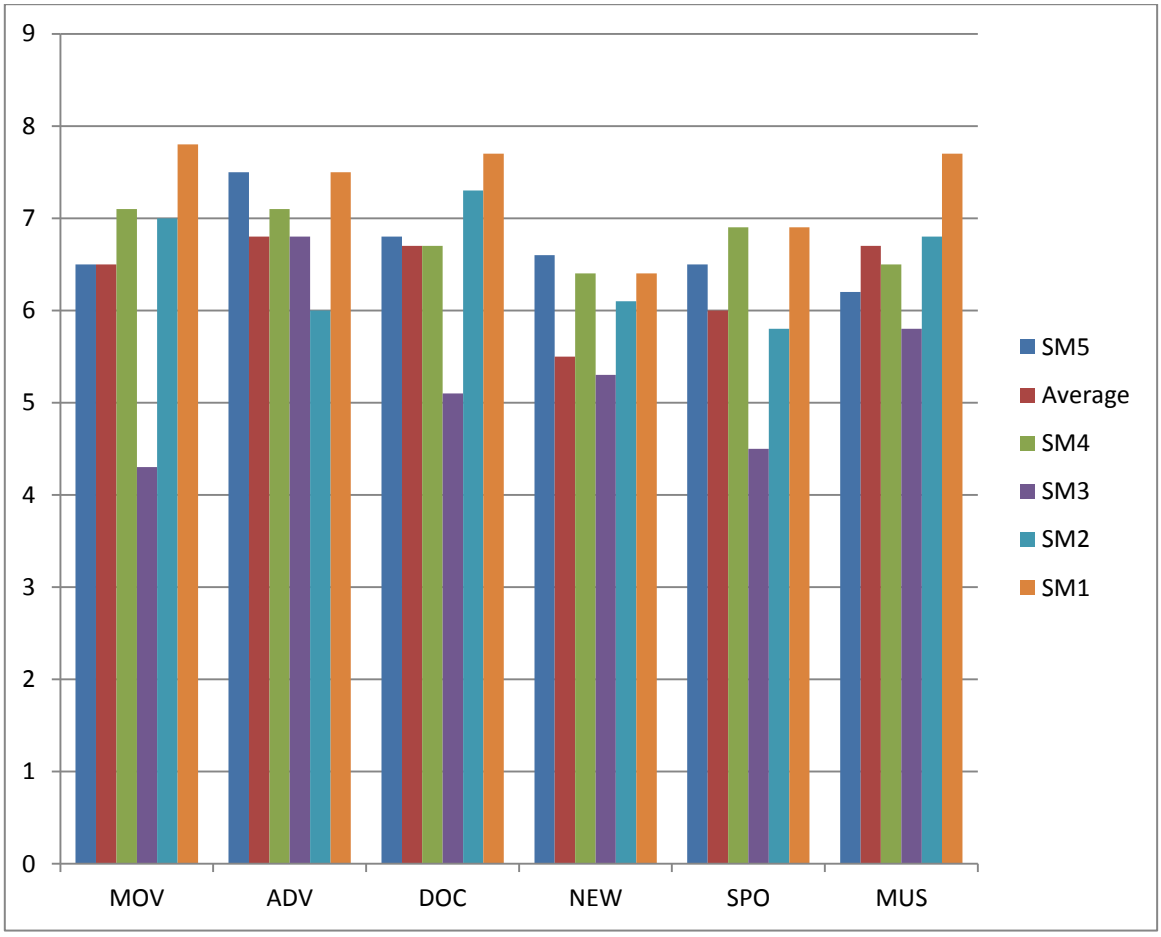


Figure 5.6. Comparison of our results against the other 4 tools for the *Recall* metric

the remaining video genres (those for which SM5 has lower grades in comparison to average of the other four tools) should be less than 1.5 unit. As is depicted in Figure 5.6, the video summaries deliver good results in terms of the *Recall* metric. In fact, in five out of six existing categories the achieved scores by SM5 are higher or equal to the average of mean grades gained by the other four versions, while it managed to obtain the highest grade amongst all 5 summary versions for *News* video clip. Moreover, the difference of grades achieved by SM5 for *Sport* video is only 0.5 units below the average marks obtained by the other four approaches. Therefore, the **Acceptability** of *Recall* rate across all six video categories can be acknowledged and subsequently the first hypothesis (H1) is verified.

Statistical Significance Analysis

In order to validate the statistical significance ($p<0.05$) of the assigned scores for our new proposed tool from this perspective a t-test analysis has been adopted. The *Recall* indicator was compared pairwise between the scores gained by our method and the achieved grades by the other four systems and the results are displayed in Tables 5.2. Akin to our previous method, our personalised summarisation method could not gain statistically significant better results for any of the categories in comparison to the other four tools (simultaneously). Nonetheless, the results have improved significantly against SM3 by producing statistically significant better results for four out of six categories.

	SM5-SM1 (Re)		**SM5-SM2 (Re)**		**SM5-SM3 (Re)**		**SM5-SM4(Re)**	
	t	**p**	**t**	**p**	**t**	**p**	**t**	**p**
SPO	-0.75	>0.05	2.61	**<0.05**	9.42	**<0.05**	-2.16	**<0.05**
DOC	-2.39	**<0.05**	-1.45	>0.05	2.89	**<0.05**	0.39	>0.05
NEW	0.42	>0.05	1.38	>0.05	3.26	**<0.05**	0.79	>0.05
ADV	-0.15	>0.05	5.11	**<0.05**	1.85	**<0.05**	-1	>0.05
MUS	-4.02	**<0.05**	-1.76	**<0.05**	1.29	>0.05	-1.07	>0.05
MOV	-1.88	**<0.05**	6.22	**<0.05**	-1.28	>0.05	-3.66	**<0.05**

Table 5.2. Investigation of the statistical difference between the results obtained by our approach and the other 3 systems from the *Recall* perspective

In addition, as opposed to our previous approach, the generated summaries by SM5 are statistically significant better for three of the video categories in comparison to SM2 (*Sport*, *Advertisement* and *Movie*). Although SM5 managed to obtain higher score comparing to the other three tools for News video, the difference is not statistically significant.

5.3.4.2. Precision

Assessing the quality of the generated abstracts from the *Precision* criterion is the goal of this section. As was argued in the last chapter, the concept of personalisation is tightly correlated with the primary objectives of video summarisation. In the context of our new work, this metric plays an even more fundamental role since it is directly linked to personalisation. Further to our earlier discussions, *Precision* can be defined as the capability of the abstraction tool to extract the highest quality video segments. However, due to viewer subjectivity, the most significant parts of a particular video clip can be dramatically different for two different end-users. Therefore, a high *Precision* score should be regarded as an indicator of the ability of a tool to incorporate its audience's preferences and therefore can potentially justify the effectiveness of an approach in terms of personalisation.

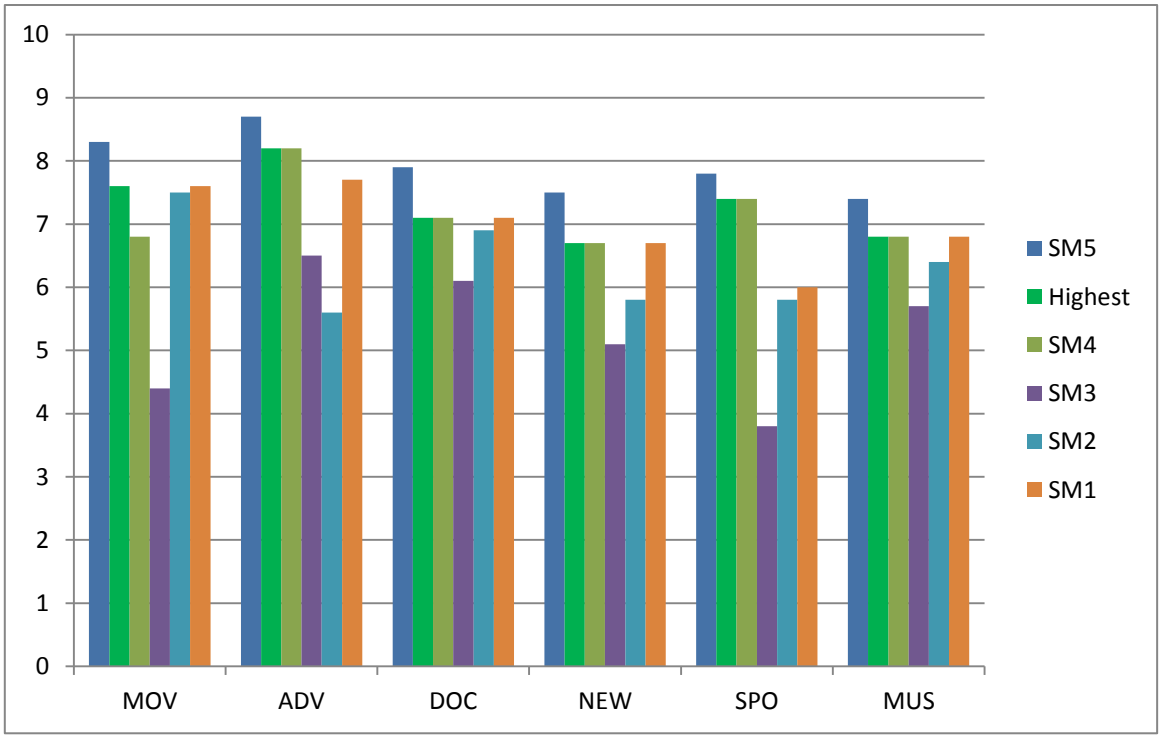


Figure 5.7. Comparison of our results against the other four tools for the *Precision* metric

In Figure 5.7, the scores achieved by our recent summarisation tool were compared to the highest scores obtained by any of the other four summary versions, which are displayed in green bars. As can be observed, our technique managed to receive the best grades across all six video categories from this point of view. This can be directly related to the inclusion of a personalisation mechanism into our recent summarisation tool. In fact, this can be further acknowledged by comparing the results achieved by our recent method against those delivered by the previous one. This chart therefore, can clearly confirm the effectiveness of our suggested video summarisation tool from the *Precision* perspective. Achieving the highest scores by SM5 among all the approaches for this index across all the video categories verified the second hypothesis (H2).

Statistical Significance Analysis

In order to validate the statistical significance of the assigned scores for our new proposed tool from this perspective a t-test analysis has been adopted. The *Precision* as the main indicator of the effectiveness of the personalisation module were compared pairwise for the grades obtained by SM5 against and the other four systems and the results are displayed in Tables 5.3. The outcome of this test highlights statistically significant differences (at the $p<0.05$ level) between the scores obtained by SM5 (our new tool) and the other four summarisation systems concerning *Precision.*

	SM5-SM1 (PR)		**SM5-SM2 (PR)**		**SM5-SM3 (PR)**		**SM5-SM4(PR)**	
	T	**P**	**T**	**P**	**T**	**P**	**T**	**P**
SPO	3.23	**<0.05**	7.68	**<0.05**	13.88	**<0.05**	2.3	**<0.05**
DOC	2.11	**<0.05**	3.68	**<0.05**	5.57	**<0.05**	3.15	**<0.05**
NEW	2.06	**<0.05**	4.96	**<0.05**	6.39	**<0.05**	3.31	**<0.05**
ADV	4.25	**<0.05**	10.84	**<0.05**	6.89	**<0.05**	2.46	**<0.05**
MUS	2.10	**<0.05**	3.19	**<0.05**	5.4	**<0.05**	4.32	**<0.05**
MOV	2.15	**<0.05**	2.14	**<0.05**	12.64	**<0.05**	4.96	**<0.05**

Table 5.3. Investigation of the statistical difference between the results obtained by our approach and the other 3 systems from the *Precision* perspective

5.3.4.3. Timing

Based on the set hypotheses in chapter 3, the generated results should meet the temporal requirement **strictly**. Our second set of experiments showed again that only our recommended methods are capable of addressing this constraint. The time lengths of the generated summaries from both of our techniques were 30 seconds and consequently received the highest possible scores.

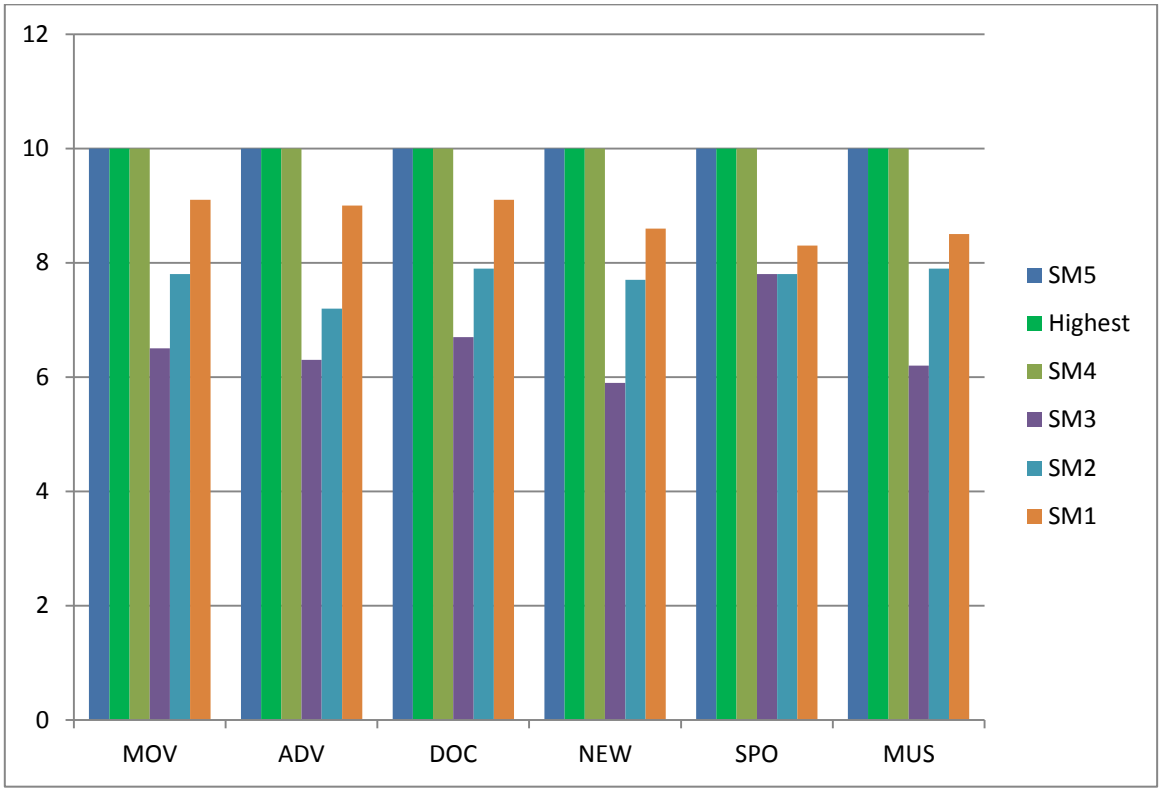


Figure 5.8. Comparison of our results against the other four tools for the *Timing* metric

As illustrated in Figure 5.8, the *Timing* scores for both of our recent summaries and the highest achieved scores from the other tools are equal. Nonetheless, the third hypothesis (H3) can be verified as well since the maximum possible score is obtained.

5.3.4.4 Overall Satisfaction

The analysis of the effectiveness of our proposed approach from the most important and comprehensive parameter is the objective of this section. Based on the results depicted in Figure 5.9, our recent approach managed to be rated the best for all six video clips from this aspect. Considering the essential role that this metric has in the formation of the overall

perceived quality of a video summary, the improved scores in comparison to our earlier technique from this angle (which was ranked the top based on the first set of experiments) can be justified based on the higher *Precision* scores that our recent method could deliver.

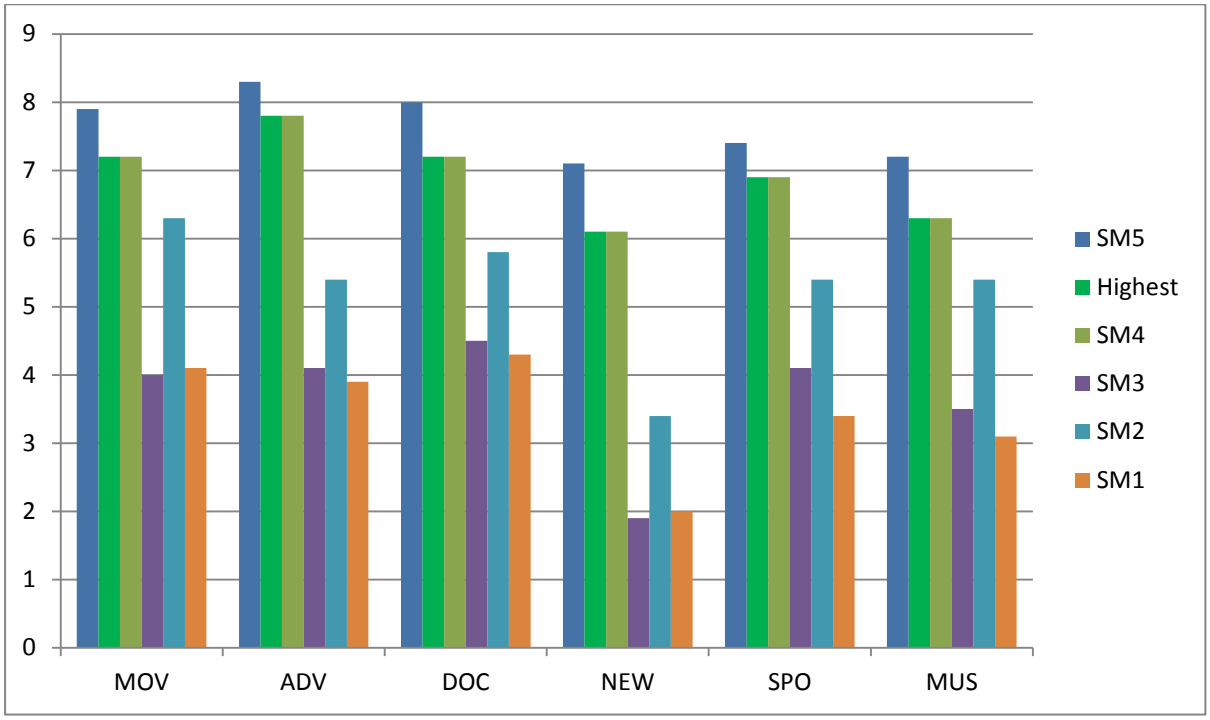


Figure 5.9. Comparison of our results against the other four tools for the *Satisfaction* metric

In addition, our second summarisation technique is built on the basis of the fundamental concepts of our first algorithm. Therefore, the main characteristics of the approved earlier method such as continuity and audio-visual clarity of the contents are inherited in the latter. To sum up, the fourth hypothesis (H4) is thus validated since the **highest** scores were achieved by our new method from the *Overall Satisfaction* point of view.

Statistical Significance Analysis

In order to determine the statistical significance of the assigned scores for our current proposed tool from the user's perceived quality perspective, a t-test analysis has been pursued. The *Overall Satisfaction* as the main indicator for quality of video summaries was compared pairwise between the scores achieved by SM5 and each of the other four systems, and the results are displayed in Table 5.4. The outcome of this test highlights statistically significant differences (at the p=0.05 level) between the scored obtained by SM5 (our new tool) and the other four summarisation systems for this measure across all six video genres.

	SM5-SM1 (OS)		SM5-SM2 (OS)		SMS-SM3 (OS)		SM5-SM4(OS)	
	T	**P**	**T**	**P**	**T**	**P**	**T**	**P**
SPO	12.87	**<0.05**	6.02	**<0.05**	15.14	**<0.05**	2.34	**<0.05**
DOC	15.0	**<0.05**	5.93	**<0.05**	13.25	**<0.05**	3.37	**<0.05**
NEW	16.5	**<0.05**	11.48	**<0.05**	14.98	**<0.05**	4.11	**<0.05**
ADV	11.67	**<0.05**	8.02	**<0.05**	12.5	**<0.05**	2.64	**<0.05**
MUS	9.91	**<0.05**	6.22	**<0.05**	12.51	**<0.05**	3.88	**<0.05**
MOV	11.03	**<0.05**	4.05	**<0.05**	10.14	**<0.05**	2.91	**<0.05**

Table 5.4. Investigation of the statistical difference between the results obtained by our technique and the other 3 systems from the Overall Satisfaction perspective

5.4. Conclusion

In this chapter, a new method for producing personalised video summaries has been proposed. User priorities towards different video scenes were captured by providing the audience with an overview of the content of each partition. Afterwards, the retrieved data was utilised for updating the average scores previously assigned by the video annotators. Finally, our summarisation approach introduced in the last chapter was adopted to extract the most significant audio-visual content of the video. In order to assess the effectiveness of our developed tool a comparative study was adopted. Experimental results indicate the capability of this approach to deliver superior outcomes compared with our previously proposed method and the three other automatic summarisation tools. However, this algorithm demands a high-level of participant intervention since users have to go through all the existing video scenes and express their level of interests towards each. Therefore, proposing a method which requires less end-user involvement is a topic for our future work and will be addressed accordingly in the next chapter.

Chapter 6

Personalised Video Summarisation Based on SIFT Features

6. Overview

In this chapter, we address the third research objective by proposing an approach for generating personalised video abstracts. Initially, the concept of Scale Invariant Feature Transform (SIFT), which is the basis for our technique, is briefly reviewed in section 6.1. Our proposed method for summarising the videos in accordance to user-created profiles will then be explained in section 6.2. In the first stage, video frames are scored by a group of video annotators (operators) according to the audio, visual and textual content of the video similar to our two former summarisation approaches. Then, as will be discussed in section 6.2.2, a matrix that contains the relevancy scores of each video scene into a number of pre-defined categories is computed using the SIFT features of the representative keyframes. In the next phase, the end-user's interest levels towards those high-level visual concepts (categories) are captured in the form of a vector. As a result of combining these two groups of data, the user's priorities in respect of different video segments can be determined. In the next stage, the initial averaged scores of the frames are updated based on the identified end-user's interest level into the corresponding video segments utilising the algorithm elaborated upon later in the chapter. Eventually, the highest scored video frames alongside the auditory and textual content are inserted into the final digest. Akin to our two former studies, the effectiveness of this approach has been evaluated through an experimental study, which compared the video summaries generated by this system against the results produced by a number of automatic and semi-automatic summarisation tools that use different modalities for abstraction. The results of this experimental research will be discussed in section 6.3 and finally the conclusion are presented in section 6.4.

6.1. SIFT

The Scale Invariant Feature Transform (SIFT) is an algorithm which is increasingly being adopted by Computer Vision researchers due to its capability to detect and describe local features of images, which can be further employed for identifying objects and identical photos. This approach is used to transform images into a large set of feature vectors for the identified 'interesting points'. These vectors are all invariant to scaling, rotation and illumination with a high level of robustness to local geometric distortion (Lowe, 2004).

There are a number of competing methods for SIFT, which are employed by researchers for invariant object recognition purposes. Rotation Invariant Feature Transform (RIFT) (Lazebnik et al., 2004), Generalized Robust Invariant Feature (GRIF) (Kim et al., 2006) and Speeded Up Robust Features (SURF) (Bay et al., 2006) are all prominent competitors of the SIFT method. However, an extensive study attempted to measure the effectiveness of these approaches against SIFT and SIFT-based methods (Mikolajczyk and Schmid, 2005). The evaluation results confirmed that SIFT-based descriptors have the highest level of robustness and distinctiveness and therefore are the best options for feature-matching tasks. For instance, imposing an affine transformation of 50 degrees, the SIFT-based identified features showed the highest rate of matching accuracies among all the descriptors. In addition, the tested descriptors demonstrated the highest level of distinctiveness for SIFT-based features. Furthermore, these descriptors had the best performance on both textured and structured scenes between all the options. These SIFT-like methods were also recognised as the most robust when imposing blur to the images and changing the illumination conditions. As will be discussed in section 6.4, we are employing this technique to compare the representative frames from the input video clips with our large training images in order to determine their high-level conceptual categories. In the next section, our novel algorithm for producing the personalised video summaries is explained.

6.2. Personalised Video Summarisation

In the previous chapter, an algorithm was introduced with the capability to include end-user's priorities towards different video scenes in the video summarisation task. The proposed approach's effectiveness was acknowledged by comparing the summary versions that were generated by that tool against the video abstracts produced by four other summarisation methods. In spite of promising results in terms of *Overall Satisfaction* and *Precision*, the

previously suggested approach requires a high-level of end-user involvement. In this section, we propose a new personalised video summarisation technique with the ability to function based on users' pre-built profiles. This approach is composed of three major phases which are now described in detail. Figure 6.1 represents the different stages in our novel personalised approach.

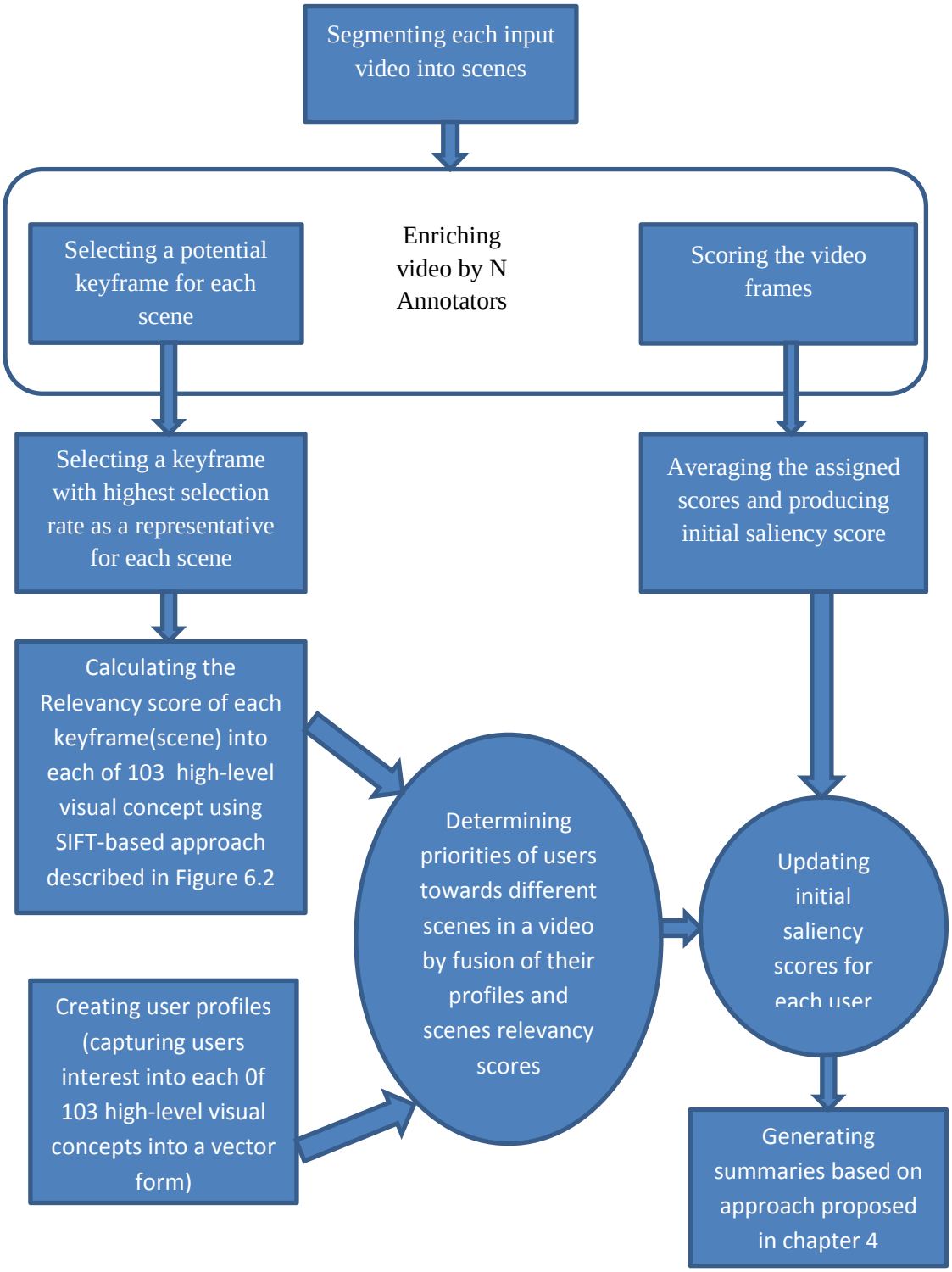


Figure 6.1. chart describing the stages in our novel personalised video summarisation approach

6.2.1. Video Enrichment

In this phase, video frames will be scored by operators and a representative keyframe will be chosen for each scene.

6.2.1.1. Video Scene Detection

Similar to our previous algorithm, AVCutty (www.avycutty.com, 2014) is used as a typical scene boundary detection tool in order to detect the time boundaries for each video scene. The minimum temporal length for each video scene is selected as three seconds in an attempt to reduce the number of pauses and interruptions for the keyframe selection task. Therefore, any identified video scene with shorter length will be added to the next scene.

6.2.1.2. Video Scenes Enrichment

According to our adopted approaches in the former two chapters, video annotators (operators) are responsible to score the video frames based on their personal interests and the perceived significance of the content they are watching. The scoring procedure is very similar to the one used in chapter 5; however, video operators are not required to annotate the identified scenes with audio-visual tags anymore. In fact, in our current approach, using the identified timestamps for the scene boundaries, the videos will be paused automatically at the end of each scene and video scorers immediately will be prompted to select one of the three candidate keyframes of each scene (based on the approach introduced in chapter 5) as the potential representative keyframe. They do this using the provided graphical user interface, while the scoring process is stopped. Therefore, per each N available frames in the original video there will be N assigned scores between 0-10 per each operator. The most satisfying frames will be scored with 10 and the least important sections are graded 0. The assigned scores of each frame by different operators will be averaged in the next step in accordance to our mentioned method in chapter 4. Therefore, a singular value will be produced for each frame inside the original video. The averaging process is thus employed to smooth the frame scores towards a less biased result, by reducing the effect of dramatic differences in assigned scores to a particular frame. In addition, based on our suggested method in chapter 5, for each

video scene, the frame that has the highest selection rate by different operators will be selected as the representative keyframe for that particular scene.

6.2.2. Personalisation

In this segment, a novel mechanism for incorporation of the viewers' preferences in summary generation is proposed. At the first step, it is attempted to measure the relevancy level of each video scene to a group of pre-defined high-level visual concepts. In addition, the end-users' priorities and preferences into any of those concepts can be extracted directly. Combining these two sets of data can assist in determining a particular end-user level of interest into a specific video scene. The 5-step personalisation procedure is explained below. In the next section, the algorithm and training images that are adopted for determination of visual concepts correlation in regards to each keyframe is elaborated upon.

6.2.2.1 Clustering Training Images

117743 images of the Image-net database (www.image-net.org) are adopted as a training collection to perform high-level visual category detection task. This large-scale database is specifically designed as a resource for researchers in multimedia content browsing and retrieving fields. In this collection, an average of 1000 quality-controlled photos for each meaningful concept (described by multiple words belonging to a common "synonym set") is provided from numerous angles. In context of our work, these images are categorised into seven major groups and a total of 103 high-level visual concepts (Table 6.1). In order to facilitate the process of concept detection, the images belonging to each category are clustered into 10 sub-categories based on their visual similarities. The similarity metric that is adopted for this purpose is their colour features in RGB space. As a result, the K-mean clustering algorithm is adopted to cluster the photos inside each category based on their retrieved RGB color histograms. Finally, the list of corresponding clusters for each training image alongside the set of cluster centroids is generated for each category.

Machinary	Natural Scenes	Public Figures	Special Scenes	Species	Sports	Urban Scenes
Aircrafts	Branch water	Actors	Adverts	Aquatic	Archery	Airports
Boats	Caves	Businessman	Concerts	Birds	Blood sport	Amusement parks
Buses	Cliffs	Criminals	Courts(trials)	Fishes	Boxing	Buidlings
Cars	Coast	Instructors	Crime scenes	Insects	Cycling	Cinemas
Coach horses	Delta	Lawyers	Earthquacks	Mammals	Funambulism	Down towns
Construction machines	Forest	Magicians	Funerals	Pets	Gymnastics	Highways
Gadgets	Ice mass	Medical team	Hurricanes	Plants	Racing	Hotel-casinos
Medical Machines	Natural elevation	Military men	Landslides	Reptiles	Riding	Libraries
Musical instruments	shore	Musicians	Large crowds	Trees	Rock climbing	Monuments
Observatory	Sky	Performers	Medical assistance		Rowing	Nuclear stations
Robots	Spring	Police officers	Tornados		Skating	Power Lines
ships	Waves	Politicians	Volcanos		Skiing	Religious places
Space machines		Religious leaders	War scenes		Team sports	Stadiums
Submarine		Sales person			Track and field	Stations
Trains		Scientists			Water sports	Streets
War machines		Singers			Wresteling	Water chutes
		Sports men				windmills
		Technicians				
		Trainers				
		workers				

Table 6.1. List of high-level visual categories adopted for our personalisation module

6.2.2.2. Scenes Conceptual Category Relevancy

The diagram for measuring the relevancy score of each video scene is illustrated in Figure 6.2. At this stage, the dependency level of each detected video scene to each of those 103 visual categories is measured and expressed in the form of a 103-length vector. Firstly, each video scene is associated with a keyframe that has the highest potential to represent that particular video segment visually and semantically. Therefore, in accordance to the procedure explained in the last chapter, the candidate frame that has the highest selection rate (among the three possible candidates) for each scene will be elected as the representative keyframe for that particular video segment. Thereafter, assessing the visual similarity between the chosen frames and training images from those 103 pre-defined categories will be the basis for this purpose. The process to calculate their visual proximity is carried out in two stages.

Initially, for each representative keyframe, the most visually similar sub-category will be identified among 10 candidates of each category. This is achieved by computing colour histograms in the RGB space for each representative keyframe in the first place. In the next stage, these computed histograms are compared against the generated list of cluster centroids

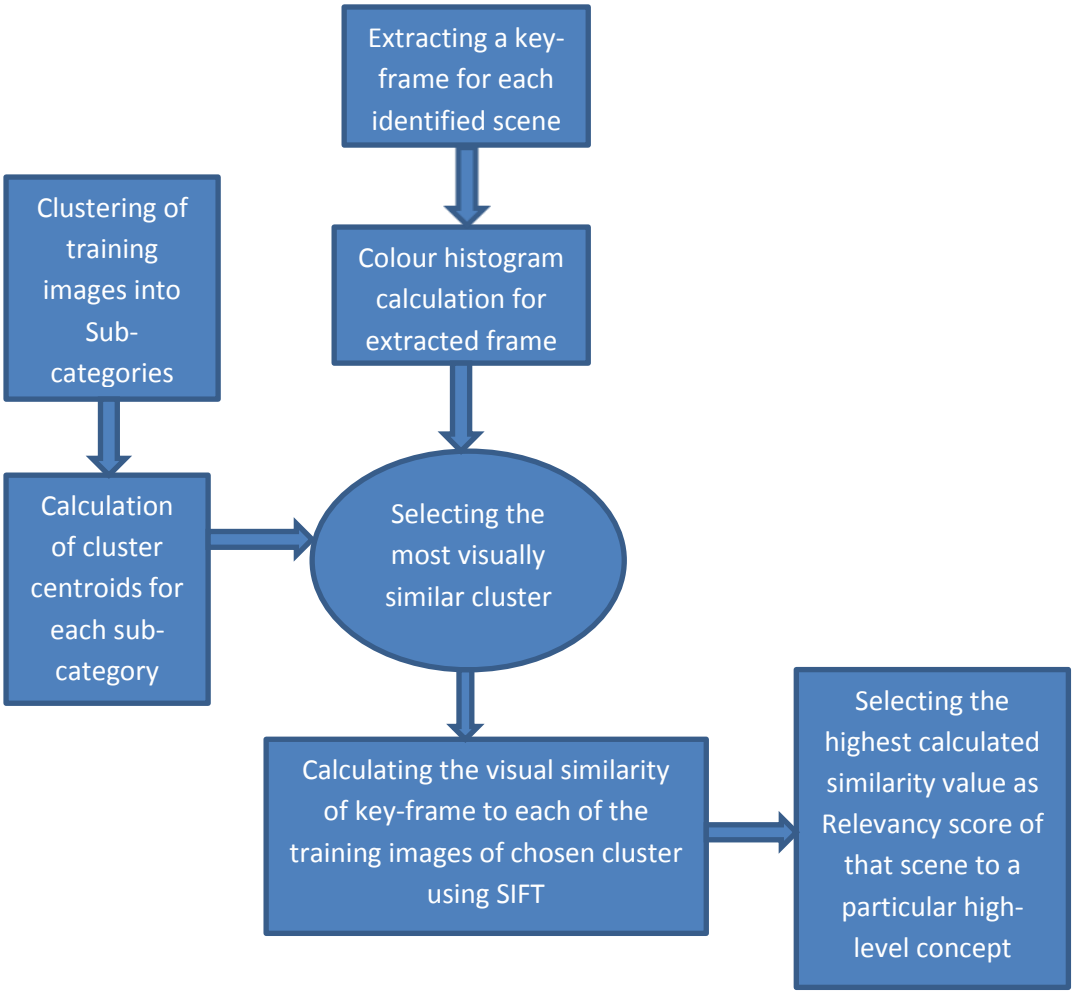


Figure 6.2. Chart describing the stages for calculation of relevancy score of a scene to each of 103 high-level visual categories

(each cluster corresponds to one sub-category) in the previous stage. The cluster with the minimum Euclidean distance to the keyframe produced histogram will be chosen as the delegate sub-category for that particular high-level visual concept. This can significantly optimise the speed and efficiency of the algorithm by reducing the number of training images that the keyframes should be examined against.

In the second stage, all the training images belonging to that chosen sub-category are extracted to be compared pairwise against the keyframe. These bilateral comparisons are performed utilising their SIFT visual features. Consequently, SIFT features for all of the selected training images alongside the keyframe are extracted for visual resemblance testing. The obtained visual features from the keyframe are contrasted pairwise against those features retrieved from the elected sub-category to identify any potential matches between these two

sets of data. Accordingly, an analogy value is computed for each pair (the keyframe and one of the selected training images), as shall now be described.

If Fr_i and Im_j denote the i^{th} frame (i^{th} scene) and the j^{th} Image of a chosen sub-category respectively, the visual similarity ratio between them is calculated based on the equation below. It should be mentioned that *SIFT* (Fr_i) represents the set of identified SIFT features for the i^{th} frame.

$$VSim(Fr_i, Im_j) = \frac{|SIFT(Fr_i) \cap SIFT(Im_j)|}{|SIFT(Fr_i) \cup SIFT(Im_j)|} \qquad (6.1)$$

Consequently, in order to calculate the relevancy degree of a video scene to one of the pre-defined high- level concepts, the visual similarity value between the representative keyframe of that scene and all of the images belonging to the chosen sub-category (cluster) should be computed pairwise. Hence, the maximum generated similarity value will be assigned to the dependency score of the scene to that particular visual concept as shown in equation (6.2). Thus, if the chosen cluster (sub-category) in category *r* has *n* images then the dependency score is calculated as below:

$$VSim(Sc_i, Cat_r) = Max_{j=1}^{n}\{\frac{|SIFT(Fr_i) \cap SIFT(Im_j)|}{|SIFT(Fr_i) \cup SIFT(Im_j)|}\} \qquad (6.2)$$

Eventually, the dependency scores between each scene and all of the 103 visual concepts are computed based on the mentioned technique. As a result, a dependency matrix D is generated, which has 103 columns (one column per high-level concept) and m rows (m is the number of identified scenes in a movie), as shown in equation 6.2. Each row represents the conceptual category relevancy of a particular scene to all of the predefined conceptual categories in a format of a vector ($Cat_1, Cat2_2, \dots, Cat_{103}$) . The algorithm for calculation of the matrix D=$\begin{bmatrix} 1 & \cdots & 103 \\ \vdots & \ddots & \vdots \\ m_1 & \cdots & m_{103} \end{bmatrix}$ for each movie is given in Figure 6.3. In the next section, we try to create profiles for the end-users in order to retain their level of interest in any of those 103 high-level visual categories.

```
Do for all 103 High-level visual categories

Set Highest-Visual-Similarity score to zero

  Do for all 10 clusters inside each category

  Input the next cluster centroid// corresponding to the next sub-category

  Compute the Euclidean distance between the key-frame histogram and cluster centroid

  Insert Computed distance value into array's index associated with the sub-category number

  End Do;

Set Chosen-Cluster-Number to array's index with the minimum value

Set Identified-SIFT-No1 to SIFT( Input Key-frame)// this functions calculate the number of
identified //SIFT features for a particular image

  Do for all the images belonging to the sub-category number of Chosen-Cluster-Number

  Input next Image// next image from training set

  Set Identified-SIFT-No2 to SIFT( Image)

  Set Number-of-Shared-Features to Match ( Input Key-frame, Image)// this function counts
//the number of identical SIFT features exist in both of the images

  Set Total-Identified-Features to Summation(Identified-SIFT-No1, Identified-SIFT-No2)

  Set Visual-Similarity to Number-of-Shared-Features divided by Total-Identified-Features

    Do if Visual-Similarity is larger than Highest-Similarity-Score

    Set Highest-Similarity-Score to Visual-Similarity

    End Do;

    Do else;

    Keep the current value for Highest-Similarity-Score

    End Do;

  End Do;

  Print the Highest-Similarity-Score for that input Key-frame as its relevancy score

End Do;
```

Figure 6.3. Algorithm for computing the relevancy score of an input keyframe to any of the 103 high level visual categories

6.2.2.3. User Profiling

At this stage, in order to generate the customised summaries, user profiles should be created. These profiles contain information regarding the end-users' level of interest in any of those 103 high-level visual categories mentioned earlier. Per sub-category, one representative image is selected randomly from the training images database. Thereafter, each category is represented by 10 attached images (one from each sub-category) and end-users are asked to express their level of interest in any of these categories by scoring the displayed representative images in accordance to Figure 6.4. The users are required to score each category on a scale from 0 to 10 based on their preferences and priorities toward the viewed visual concept using the provided graphical user interface. The captured data can be stored in a form of a vector ranging from 1 to 103 as well, where each element of the vector is used to represent an end-user's priority level in regards to one of the high-level concepts. Moreover, these generated vectors can be utilised in the following stage for understanding a particular user level of interest in regards to a video scene.

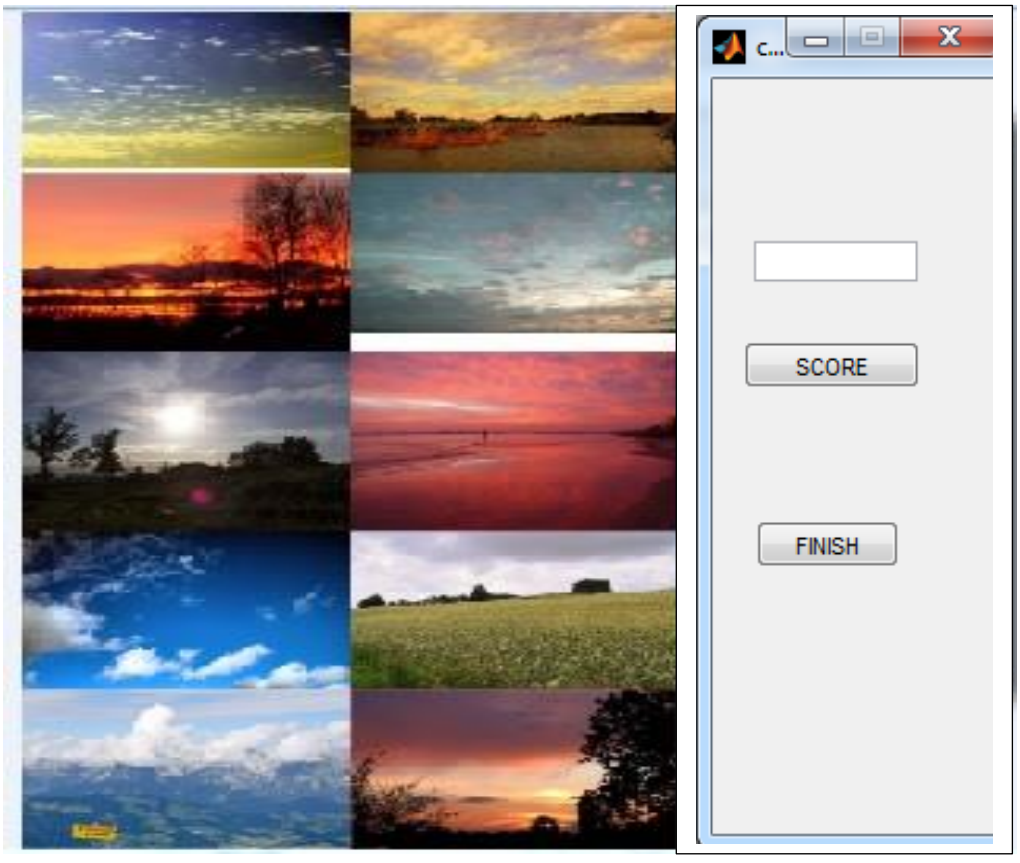


Figure 6.4. Representative images from 10 sub-categories of SKY

6.2.2.4. Determining Users' Priorities towards Scenes

According to the method explained in section 6.2.2.2, a relevancy matrix is built for each movie where each row of data demonstrates the dependency level of a video scene to one of the 103 pre-defined concepts. In order to discover the priorities of the users regarding different scenes, as was mentioned earlier, firstly, each row should be extracted in the form of a vector. This retrieved data should be combined with the obtained user profiles in the next step to develop the required input for tailoring the video abstracts. This data fusion can be achieved by producing the dot product of a scene dependency vector and a user's profile vector as shown in equation (6.3).

Accordingly, If $\vec{S}$ and $\vec{U}$ denote the vectors built from the scenes dependency scores and users' profile respectively then, SS_n computes the relevancy grade of the n^{th} scene of a movie for a specific user profile:

$$SS_n = \vec{S}.\vec{U} \qquad (6.3)$$

Using the mentioned method, a singular value is computed for each scene which translates the priorities that a user has towards that video segment in comparison to the others. Finally, per user and movie, a border is constituted, whose length is equal to the number of identified video scenes. The elements of this border (calculated dot products) can be used for prioritising the video scenes.

6.2.2.5 Prioritising the Video Scenes

At this stage, the priority levels are assigned to video scenes based on their achieved scores. The higher the score is, the more the viewer is interested in a specific video scene. Thus, the generated results in the previous stage are sorted in a descending manner. The scenes whose grades are located in the first quarter of the sorted list are given priority level two, while those in the second quarter are attributed the priority level one; the rest of the scenes are all marked as level 0 (the lowest degree of importance).

6.2.2.6. Updating the Initial Frame Scores

In this phase, the primary generated average scores of the frames, assigned by the video scorers and averaged over the number of operators are updated on the basis of the approach introduced in chapter 5. Thus, based on the computed priority level for each video segment for a particular user, the primary average scores are updated. The scores of frames belonging to the scenes with level 0 of interest will not be altered at all. However, in the scenes with a level one priority (those scenes located in the second quarter of scenes' relevancy scores list), the grades for the frames whose primary assigned scores are the highest among the frames of that scene will be increased by 20 percent (up to a maximum value of 12).

This will potentially increase their probability of inclusion of the most significant video frames belonging to those scenes into the final video abstract. As a result, it will affect positively on the *Recall* and *Precision* of the generated abstracts. The updated grade for the frames belonging to the scenes with the highest level of priority for a particular end-user will be recalculated in a different format. These are the scenes whose dependency scores are among the first quarter of the computed relevancy scores list. The grades for the frames, which initially were scored the highest in each scene, will be upgraded to the maximum possible value (12). In fact, this would escalate the chance of definite inclusion of the highest quality segments of those particular scenes (with level two priority) in the final summary. However, the marks for the frames of those scenes whose scores are not the highest but nonetheless manage to exceed the respective scene's average scores will be boosted by 20 percent as well (to a maximum of 12).

6.2.2.7. Generating the Personalised Summaries

In the final step, the personalised video summaries are produced based on the updated frames scores in the last stage. In keeping with the summarisation method based on group scoring, which has been discussed extensively in previous chapters, the highest scored frames alongside the audio and textual content are selected and inserted into the final video digest. In next section we undertake an experimental evaluation to compare the quality of generated summaries based on our recent approach against those produced by three automatic tools and our recommended summarisation technique described in chapter 5.

6.3. Experimental Evaluation

The confirmatory research adopted for this stage will be discussed comprehensively in this section. Accordingly, the experimental procedure will be carried out in two different stages analogous to the two former chapters. In the first phase, video annotators score the frames and select the representative keyframe based on the process explained in section 5.2. Afterwards, the participants are used for profile creation tasks based on the suggested algorithm in this chapter. In addition, the same group of participants is required for generation of video abstracts based on the suggested technique in chapter 5. This is due to our desire to compare the results from our new technique against those from the previous one. The experimental process is explained thoroughly below.

6.3.1. Frames Scoring and Scenes Enrichment

Six short videos (two minutes each) belonging to a different video genre (*Movie, Sport, Documentary, Advertisement, Music* and *News*) identical to those used in the two preceding chapters were utilised for the evaluation purpose of our proposed method. In the initial phase, 10 operators (video annotators) with different demographic details (five females and five males within an age range of 25-45 years old) were asked to watch each video and score the frames using the provided slider tool. In addition, at the same time, they are also asked to annotate the video segments and to choose the representative keyframe of each scene based on the method proposed in previous chapter. Here however, the video operators are not required to annotate the scenes with audio-visual tags, since, we are comparing the results from this technique against the former one; the annotators are asked to score and annotate the videos only once to avoid repetition. This is due to the fact that, the single output from this stage can be effectively utilised for both summarisation approaches. The assigned scores for each frame were then averaged to generate a singular value for that frame according to the method detailed in chapter 4.

6.3.2. Users' Profiling and Priorities Extraction

In this stage, we try to discover the priorities of 30 end-users (15 females and 15 males within an age range of 24-60 years old) towards the different scenes of each video using the proposed method described in chapter 5. It should be reminded that these participants are different to those who annotated the original videos. Based on the previous method, the users

have to express their level of interest to each video scene explicitly for personalisation purposes and the video summaries are generated accordingly.

Later, the same group of participants is employed for profiling purposes. In accordance to the method explained in section 5.2, they are asked to express their level of interests to those representative images (from the image database) from the high level categories by scoring them. Therefore, by applying this technique a generic profile is built for each end-user based on his/her degree of interest in each high-level visual category. Summary versions for the same video clips are then generated employing our novel summarisation technique.

6.3.3. Evaluation of Generated Summaries

In order to evaluate the effectiveness of our personalised video summarisation approach, the generated summaries using the method described in this paper have been compared against the abstracts generated based on four other approaches. Three of these tools summarise the videos automatically (as explained extensively in chapter 4) by applying statistical and mathematical algorithms, while the fourth tool is based on our former semi-automatic personalised video summarisation approach (proposed in chapter 5). The six original videos alongside their five summary versions created by the five existing tools (including the personalised summaries generated for each specific user using the currently proposed technique) were presented to the same 30 end-users on the basis of whose inputs their personalised summaries were created.

After watching the original video and the summaries, users were asked to score each of the generated abstracts awarding marks between 0 (worst video summary possible) to 10 (best video summary possible), from four different perspectives consisting of *Recall* (Re), *Precision* (Pe), *Timing* (Ti) and *Overall Satisfaction* (OS).These measures were described in detail in chapter 4. The given scores for each of these measures were averaged over the30 users and their mean values for each of the video categories are given in Table 6.1. SM1, SM2, SM3, SM5 and SM6 indicate the average achieved scores (alongside the standard deviation) by, respectively, the first, second, third, fourth and our recently proposed video abstraction methods.

Our current technique (SM6 in Table 6.2) has the highest scores from the *Overall Satisfaction* point of view across all six video categories, while the quality of its generated summaries is still significantly better than those produced with automatic tools.

	SM1				SM2				SM3				SM5				SM6			
	Re	Pe	Ti	**OS**	Re	Pe	Ti	**OS**	Re	Pe	Ti	**OS**	Re	Pe	Ti	**OS**	Re	Pe	Ti	**OS**
MOV	7.1 (1.5)	6.2 (1.5)	8.2 (1.0)	**4.2 (1.9)**	6.2 (1.4)	6.3 (1.5)	6.7 (1,2)	**6.2 (1.5)**	3.7 (1.1)	3.9 (1.6)	5.5 (1.2)	**4.4 (1.4)**	6.3 (1.2)	6.4 (1.3)	10 (0)	**6.3 (1.4)**	7.1 (0.9)	6.8 (1.1)	10 (0)	**7.1 (1.1)**
ADV	7.0 (1.7)	6.9 (1.5)	8.4 (1.8)	**3.8 (1.4)**	5.5 (1.5)	5.6 (1.2)	6.9 (0.9)	**6.1 (1.4)**	6.2 (1.3)	6.2 (1.4)	6.0 (1.0)	**4.6 (1.2)**	6.4 (0.9)	7.0 (1.0)	10 (0)	**7.1 (0.8)**	7.6 (0.9)	7.6 (0.9)	10 (0)	**7.6 (0.9)**
DOC	7.2 (1.1)	6.2 (1.3)	8.3 (1,1)	**3.8 (1.3)**	6.2 (1.0)	6.0 (1.1)	7.5 (1.0)	**5.2 (1.4)**	5.1 (1.3)	5.6 (1.7)	6.5 (1.0)	**4.4 (1.4)**	6.3 (1.2)	6.4 (1.1)	10 (0)	**6.4 (0.8)**	6.7 (1.3)	6.9 (1.3)	10 (0)	**7.0 (1.3)**
NEW	6.0 (2.0)	6.2 (1.9)	8.6 (1.1)	**2.0 (1.1)**	5.7 (1.6)	5.6 (1.1)	6.7 (1.2)	**3.3 (1.9)**	4.9 (1.3)	4.9 (1.7)	5.7 (1.2)	**2.4 (1.9)**	5.5 (1.3)	6.5 (1.2)	10 (0)	**5.7 (1.5)**	6.8 (1.0)	6.2 (1.1)	10 (0)	**6.6 (1.0)**
SPO	6.6 (1.7)	6.4 (2.0)	8.1 (1.4)	**2.9 (1.3)**	5.0 (1.5)	5.4 (1.6)	7.1 (1.0)	**5.2 (1.8)**	3.7 (1.3)	3.5 (1.3)	6.0 (1.0)	**3.3 (1.8)**	6.3 (1.1)	6.5 (0.9)	10 (0)	**6.5 (1.1)**	7.3 (0,9)	6.3 (1.1)	10 (0)	**7.0 (0.8)**
MUS	7.4 (1.1)	6.8 (1.9)	7.9 (1.3)	**2.9 (1.1)**	6.1 (1.3)	6.3 (1.0)	6.8 (0.9)	**5.2 (1.8)**	5.3 (1.5)	5.6 (1.4)	5.6 (1.0)	**3.5 (1.3)**	6.5 (1.2)	7.3 (1.1)	10 (0)	**6.3 (1.3)**	7.3 (1.0)	7.0 (0.9)	10 (0)	**7.1 (1.2)**

Table 6.2. Average assigned scores to each summary from 4 perspectives

In addition, SM6 was ranked the highest in terms of *Precision* (corresponding to the effectiveness of an approach in regard to generation of personalised results) on three categories while it only came second to our previous method (SM5) in respect of the other three (*Sport, News* and *Music*). Moreover, the achieved scores for *Recall* among all six categories have improved significantly in comparison to our previous algorithm.

6.3.4. Results

In this segment, similar to two previous chapters, we try to analyse the effectiveness of our proposed video summarisation approach from the four identified hypotheses in chapter 3. It should be mentioned again that the generated summaries should have an **Acceptable** *Recall* rate, while their *Precisions* achieve a **High** ratio. In addition, while the *Time* constraint should be met **Strictly**, the *Overall Satisfaction* has to be rated as the **Highest** among all the available versions. We start our analysis by evaluating the Recall rate in the next section.

6.3.4.1. Recall

In this section, the effectiveness of this approach in regards to the *Recall* metric is assessed. We expect that our summaries achieve better scores than the average of scores assigned to the other four tools for at least half of the categories, while the difference in scores for those with lower grade is less than one unit.

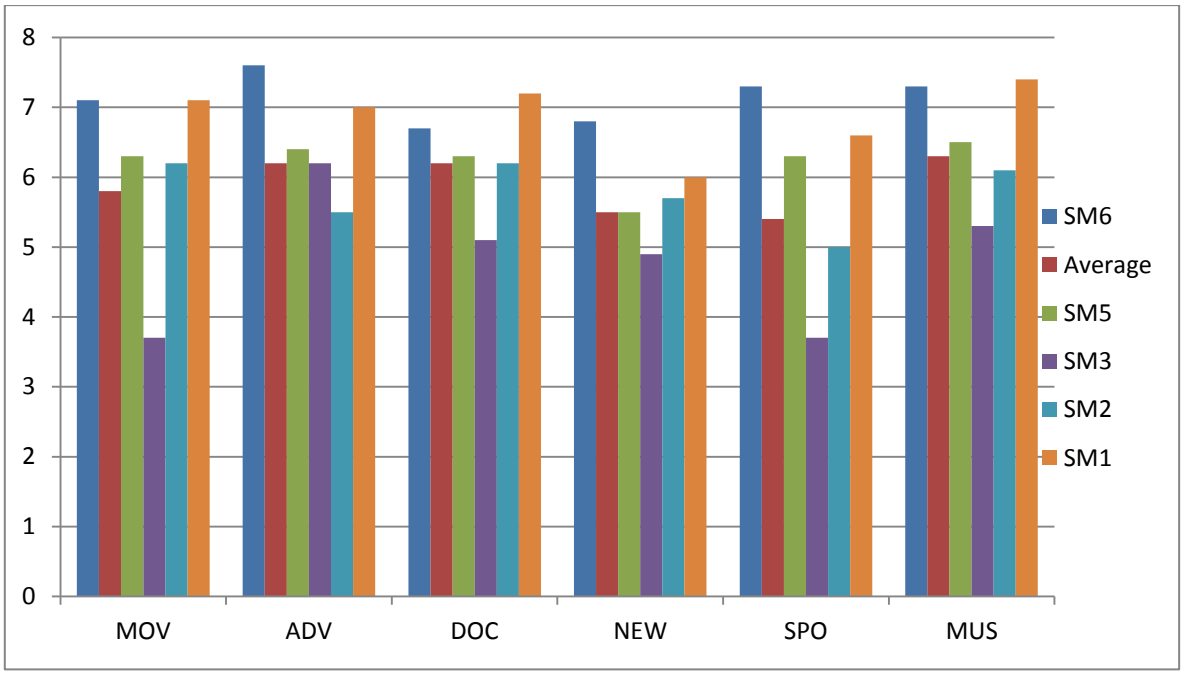


Figure 6.5. Comparison of our results against the other 4 tools for the *Recall* metric

As is shown in Figure 6.5, the *Recall* score has improved significantly for our novel video summarisation approach comparing to the last two techniques. This algorithm managed to exceed the average scores achieved by the other four tools across all the video categories. In addition, for three video categories (*Movies*, *Advertisement* and *Music Video*) the highest scores for this metric obtained by our latest technique. Thus, the corresponding hypothesis (H1) is confirmed. The justification for this growth can be attributed to the employed mechanism for updating the frame scores in our recent method. Since frames scores (a group of them) for half of the video scenes will be upgraded automatically, the possibility of different video scenes to have representative segments into the final summary will be boosted noticeably. Therefore, more segments of the video can be potentially covered in the final summary.

On the other hand, based on our last technique, scores will be updated for the frames that belong to those segments in which the end-users have explicitly expressed their interest in. The less the viewers are interested in various video scenes, the less frames scores will be altered. Therefore, fewer video scenes will have delegate frames into the final summary.

Statistical Significance Analysis

According to Table 6.3, there is a significant statistical ($p<0.05$) difference between the results achieved by our algorithms and those obtained by the other tools for three video categories (*Sport*, *News* and *Documentary*) in regards to this index. Among these three categories, the generated t-values for *Sport* video are considerable. The assigned scores for the other three categories exceeded the average mean marks obtained by the other four approaches although in some cases the pairwise differences between our latest approach and SM1 are not statistically significant. Nevertheless, our tool managed to achieve higher marks in comparison to average scores of other three tools and subsequently the first hypothesis (H1) is verified.

	SM6-SM1 (Re)		**SM6-SM2 (Re)**		**SM6-SM3 (Re)**		**SM6-SM5(Re)**	
	t	**p**	**t**	**p**	**t**	**p**	**t**	**p**
DOC	3.63	**<0.05**	2.05	**<0.05**	5.16	**<0.05**	1.81	**<0.05**
MOV	0.23	>0.05	3.17	**<0.05**	14.38	**<0.05**	3.39	**<0.05**
ADV	1.63	>0.05	6.83	**<0.05**	3.89	**<0.05**	5.68	**<0.05**
NEW	2.10	**<0.05**	4.22	**<0.05**	7.30	**<0.05**	5.74	**<0.05**
MUS	-0.12	>0.05	3.92	**<0.05**	6.21	**<0.05**	3.18	**<0.05**
SPO	2.34	**<0.05**	7.13	**<0.05**	14.20	**<0.05**	4.50	**<0.05**

Table 6.3. Investigation of the statistical difference between the results obtained by our method and the other 3 systems from *Recall* perspective

6.3.4.2. Precision

In this section, we assess the quality of generated summaries by our latest algorithm from the *Precision* point of view, i.e. how effective our technique is in extracting the most significant

video segments in regards to a specific audience. The chart below compares the results received by the current method's (SM6) generated summaries against the highest grades achieved by the other four tools across the six video genres.

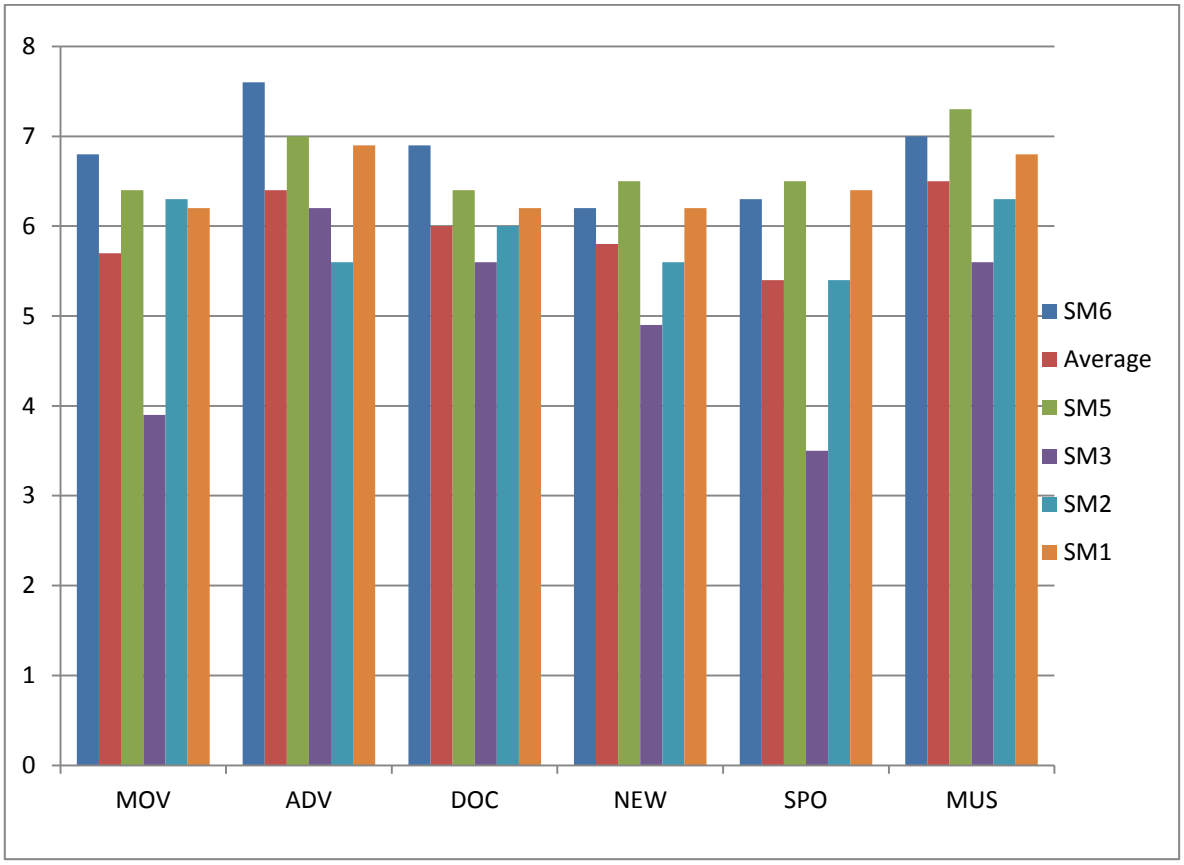


Figure 6.6. Comparison of our results against the other 4 tools for the *Precision* metric

Based on Figure 6.6, our novel method obtained the highest scores in three out of six video genres, whilst in the other three categories the obtained scores are higher than the average grades of the other four systems. Our recommended algorithm in chapter 5 (SM5) attained better results in terms of this metric for the *News*, *Sport* and *Music Video* categories. This can be justified in accordance to the nature of these two types of video.

The first one is due to inability of our system to distinguish the end-users' priorities towards the different events in a context of a sport match. Additionally, the lack of a mechanism to incorporate auditory information (essential for the *News* and *Music videos*) for personalisation purposes has reduced the *Precision* scores for News genre. However, since the video annotators in the first stage have scored the video frames using all available modalities (audio, visual and textual) then it minimises the risk of missing the most valuable segments in the final summary. Moreover one should also bear in mind, the significant

amount of reduced time and cost using the generic user profiles (instead of employing viewers to score scenes) for personalisation. Nonetheless, the proposed method has addressed the articulated hypothesis by gaining **high** results in this respect.

Statistical Significance Analysis

As it illustrated in Table 6.4, there is a significant statistical ($p<0.05$) difference between the results achieved by our latest techniques in comparison to those generated by the other four tools for the first three video categories namely, *Movie, Advertisement* and *Documentary*.. Therefore, our latest algorithm produced the highest quality summaries from this perspective for the mentioned categories. Although our previous approach produced better results for the other three categories, the differences are not statistically significant. In fact, no other summarisation method managed to produce results with higher average mean with statistically significant difference. Finally, the gained marks for the latest algorithm comparing (SM6) to the average grades achieved by the other four tools is higher for the other three genres. Consequently, the second hypothesis (H2) is verified.

	SM6-SM1 (Pr)		SM6-SM2 (Pr)		SM6-SM3 (Pr)		SM6-SM5(Pr)	
	t	p	t	p	t	p	t	p
DOC	2.52	**<0.05**	2.84	**<0.05**	4.33	**<0.05**	2.10	**<0.05**
MOV	2.07	**<0.05**	1.77	**<0.05**	11.37	**<0.05**	1.81	**<0.05**
ADV	2.14	**<0.05**	6.67	**<0.05**	4.52	**<0.05**	2.37	**<0.05**
NEW	0.37	>0.05	1.83	**<0.05**	4.12	**<0.05**	-0.94	>0.05
MUS	-1.54	>0.05	3.19	**<0.05**	4.79	**<0.05**	-0.89	>0.05
SPO	-0.17	>0.05	2.26	**<0.05**	9.15	**<0.05**	-0.86	>0.05

Table 6.4. Investigation of the statistical difference between the results obtained by our method and the other 4 systems from the *Precision* perspective

6.3.4.3. Time

Just like our two former techniques, this method has also produced the summaries that fulfill the pre-determined *Time* constraint **Strictly**. As can be seen on Figure 6.7, the only other system that could respect this requirement is our previously suggested approach (SM5).

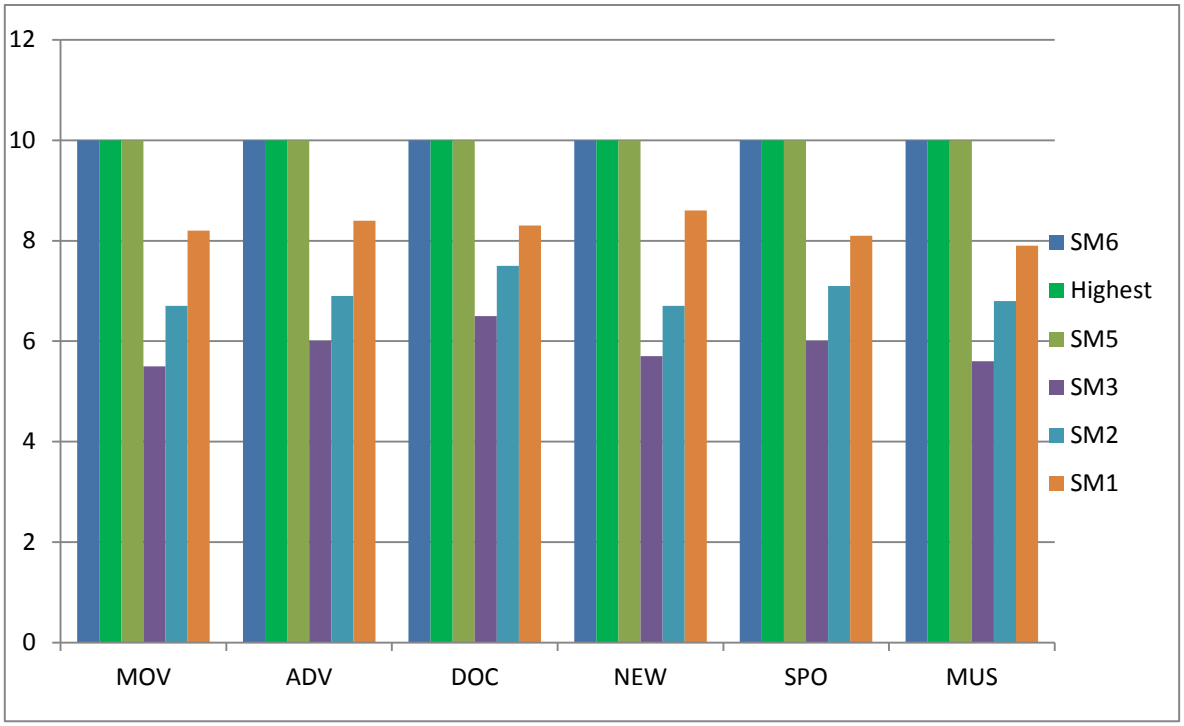


Figure 6.7. Comparison of our results against the other 4 tools for the *Timing* metric

Accordingly, by receiving the highest possible mark from this perspective, the hypothesis (H3) is verified.

6.3.4.4. Overall Satisfaction

The most important measure for assessing the effectiveness of a video summarisation technique will be investigated in this section. Figure 6.8 exhibits the overall perceived quality (i.e. *Overall Satisfaction* metric) of the generated summaries by the current approach against the other methods.

As can be seen, the highest results for this index have been obtained by our novel algorithm (SM6). In addition, the highest scores achieved by the other four tools all belong to SM5 (our previous technique), as is shown. This can be explained based on the fact that the combinational scores of *Recall* and *Precision* are more balanced in our more recent attempt. In spite of lower results for a number of video categories comparing to the last approach, the capability of the new technique to cover larger segments of the video affected the participants positively in terms of their overall perceived quality of the video. Moreover, another influential factor in achievement of better outcomes can be due to the fact that the audiences' level of participation has been reduced significantly in the current algorithm.

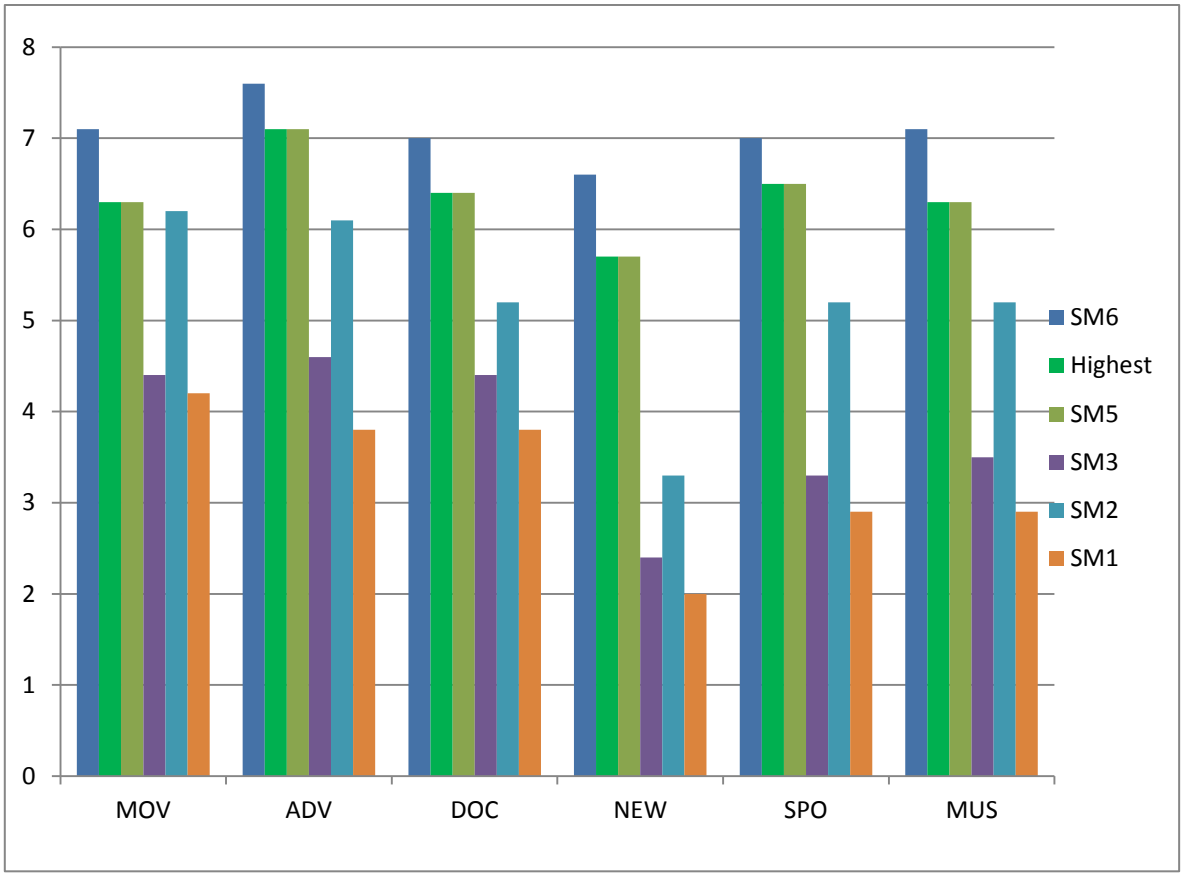


Figure 6.8. Comparison of our results against the other 4 tools for the *Satisfaction* metric

Statistical significance Analysis

We further analysed our results through a t-test. The *Overall Satisfaction* results as the main indicator were compared pairwise against the achieved scores of the other four systems and the results are displayed in Table 6.5. The outcome of this test highlights statistically significant differences (at the p=0.05 level) between the scores obtained by SM6 (our new tool) and the other four summarisation systems for the mentioned measure across all categories. The level of differences for a number of categories including *Music* and *News* videos are considerable. In addition, the least level of significance differences can be associated to the comparison of the results generated from SM6 and SM5 (our recent and previous summarisation methods respectively). Finally, our technique generates the highest quality video summaries in this regards and accordingly the fourth hypothesis (H4) can be verified.

	SM6-SM1 (OS)		SM6-SM2 (OS)		SM6-SM3 (OS)		SM6-SM5(OS)	
	t	**p**	**t**	**p**	**t**	**p**	**t**	**p**
DOC	11.3	**<0.05**	5.26	**<0.05**	8.45	**<0.05**	2.13	**<0.05**
MOV	7.06	**<0.05**	2.52	**<0.05**	9.57	**<0.05**	2.94	**<0.05**
ADV	11.20	**<0.05**	4.15	**<0.05**	10.51	**<0.05**	2.63	**<0.05**
NEW	19.92	**<0.05**	14.38	**<0.05**	14.15	**<0.05**	4.30	**<0.05**
MUS	11.89	**<0.05**	5.74	**<0.05**	11.60	**<0.05**	3.51	**<0.05**
SPO	13.80	**<0.05**	4.46	**<0.05**	10.36	**<0.05**	2.10	**<0.05**

Table 6.5. Investigation of the statistical difference between the results obtained by our method and the other four systems from the Overall Satisfaction perspective

6.4. Conclusion

In this chapter, a new method for producing personalised video summaries has been proposed. Accordingly, SIFT visual features were adopted to identify the video scenes' semantic categories. Fusing this retrieved data with pre-built users' profiles, personalised video abstracts can be created. Experimental results indicate the effectiveness of this approach in delivering superior outcomes comparing to our previously proposed method and three other automatic summarisation tools.

Chapter 7
Conclusion and Future work

7. Overview

Video Summarisation is one of the most challenging topics in the multimedia domain that has received a great extent of coverage by researchers during recent years. Providing viewers with concise though rich versions of an input video sequence through identification and extraction of the most valuable segments is the primary objective of this research area. However, the difficulty of bridging the existing gap between low-level textual, visual and aural features of video streams and high-level semantically meaningful concepts is one of the major contributing factors in delaying the introduction of a definitive abstraction technique. The previous chapters of this work in response to our established research aim have inspected the research carried out in relation to video summarisation in order to determine the shortcomings of the existing methods and accordingly proposed and investigated three summarisation algorithms to address these pinpointed limitations. In this chapter, we finalise our research by providing a summary of our recommended techniques, the experimental findings, the subsequent knowledge contributions and some proposal for future work based on identified limitations of our research.

7.1. Research Domain

This research has focused on design and development of video summarisation algorithms with the capability to enhance the users' experience and perceived quality of the abstracts in comparison to previously proposed techniques. Although various abstraction methods have been suggested by researchers in recent years, a review of the existing studies highlighted that the primary focus has been allocated to the fully-automatic techniques, which do not require human's intervention. The foundation of these approaches is to understand the most semantically important video units through analysing the low-level visual, textual and aural

characteristics. However, as was extensively discussed in chapter 2, mapping these features into high-level semantics and concepts is still a pretty much challenging exercise with limited success. On the other hand, the performance of human-centred propounded models has been negatively affected by the subjectivity of the users and other external factors such as distraction. In light of this, we attempted to address the shortcomings in this field by answering the following research questions:

RQ1- What approach should be proposed to reduce the limitation of the existing techniques?

RQ2- What is the best way to develop the proposed approach?

RQ3- Is the proposed approach effective enough?

Accordingly, the research aim was defined as: **To develop three effective video summarisation techniques that could be applied to different video categories and generate satisfactory results in terms of *Recall, Precision, Timing* and *Overall Satisfaction*.** In order to achieve this research aim, four research objectives were defined and achieved, which are listed below.

Objective 1: To investigate the exiting video summarisation techniques in order to identify the limitations and barriers against of this technology. This objective was addressed in chapter 2 and the shortcomings and limitations of the existing methods were revealed by studying the existing literature.

Objective 2: To design, develop and evaluate a user-centred video summarisation algorithm based on group scoring in accordance to the findings from the previous investigation: This objective was covered in chapter 4.

Objective 3: To extend the work of previous objective and design, develop and evaluate a personalised video summarisation algorithm based on group scoring: This objective was met in chapter 5 by proposing and analysing a novel mechanism to understand and incorporate the viewers' priorities towards different video scenes in prior to abstracting the video.

Objective 4: To extend the work of previous objective and design, develop and evaluate a personalised video summarisation system with reduced end-user involvement: This

objective was accomplished in chapter 6 with presenting a profile-based personalisation module according to SIFT visual features.

Accordingly, the chapters of this thesis were designed respectively to address the accomplishment of the identified research objectives. The overview of the former six chapters is provided in the next section.

7.2. Summary of Findings

- In chapter 2, a group of existing abstraction methods were assessed and classified from different perspectives.. The inability to understand the semantics of video content, domain-dependency, sensitivity to changing conditions and complexity (computational expense) of fully automatic methods were established as the major barriers against the adoption of techniques with no human involvement. In addition, user subjectivity and other external factors such as distraction were identified as the limitations of user-centred approaches. Later, a review of the most common evaluation methodologies adopted by researchers to measure the quality of generated summaries revealed *Recall* and *Precision* as the two mainly employed metrics. The review also highlighted that user perceived quality of video is the most important metric to determine the effectiveness of a multimedia content, which not properly recognised by earlier studies of video abstraction. Accordingly, in our work, we also employed *Overall Satisfaction* metric as a success index for our proposed video summarisation algorithms.

- A novel user-centred video summarisation approach was proposed chapter 4 based on group scoring. A panel of video scorers was recruited to grade video frames as they watch on the fly based on the significance of available modalities. In order to downgrade the negative effects of the subjectivity of the users, their assigned marks were averaged and a singular saliency score was computed for each video frame. Eventually, the highest scored segments of the video alongside the visual, textual and aural content in respect of a pre-specified time constraint were copied into the video digest. The proposed algorithm was later evaluated by comparing the video digests produced by our method against the summary versions of the same input videos

generated by three different automatic systems that utilise different modalities and mathematical algorithms for abstraction purposes. In accordance to pre-established hypotheses, our method managed to deliver superior outcomes in comparison to the other three tools. Achieving the highest scores from the participants' perceived quality (*Overall Satisfaction)* perspective across all the experimental videos was the major element in confirming the effectiveness of our approach. Obtaining acceptable results in comparison to the other system in terms of *Recall* and *Precision* confirmed the strength of our system.

- In chapter 5, a personalised video summarisation approach on the basis of our previously recommended group scoring method was introduced. The original video was primarily segmented into video scenes according to the visual similarity of neighbouring frames. In our new approach, the panel of video annotators had to enrich the video segments by selecting a keyframe (among the three choices for each scene) and annotating each scene with audio and visual tags in addition to scoring the frames (similar to our earlier method). Prior to summarisation of a video, the end-users were provided with a list of representative keyframes, each corresponding to one scene and their associated tags. These are the frames and tags that had the highest selection rate by the video operators. Furthermore, the viewers were asked to express their level of interest in each scene based on the provided information on a scale of zero to two. As a result, the saliency scores for each frame, based on the users' priorities and their membership to different scenes were updated and users-tailored summaries were accordingly generated. Finally, the generated summaries based on our personalised method were evaluated in comparison with the outputs from three automatic tools as well as those produced by our former summarisation technique. The results managed to confirm our study's hypotheses in regards to the effectiveness of our recommended algorithm. Achieving the highest scores in terms of *Overall Satisfaction* and *Precision* were the two most important criteria that acknowledged the quality of video digests. Incorporating the personalisation module and the consequent improvement of the summaries in comparison to our previous approach were confirmed as well.

- A new algorithm with the primary objective of streamlining the process of extraction and integrating the audience's priorities was proposed in chapter 6. Initially, a group of training images from 103 high-level visual concepts were obtained. The representative keyframe from each video scene was selected in accordance to our past method and was compared against each of the 103 high-level concepts in order to measure their visual similarity. The results of these comparisons were captured in the form of a vector, whose constituent elements represented the relevancy of the keyframe to a particular concept. Later, user profiles were built by displaying the users' representative images from our training collection and asking them to score each category based on their level of interest. As a result, any user profile thus built could be represented as a 103-sized vector as well. Finally, the inner product of these two arrays could generate a priority value indicating the preference level of a particular user for a specific scene. These produced values were further adopted to accordingly update the saliency scores of video frames and thus generate the final summaries. The effectiveness of our recent approach was validated by evaluation of the video summaries retrieved by our novel method against those achieved from the other tools as well as our previous personalised system. The users' perceived quality of abstracts across all experimental videos had the best scores for our latest proposed technique. In addition, the *Precision* marks were highest for three genres, while it only came second to our previous technique in respect of the other three. Additionally, there was a significant improvement in terms of *Recall* scores compared to our earlier algorithm.

7.3. Research Contributions

The research aim and objectives developed in our first two chapters provided the foundation for our knowledge contributions of the study described in this thesis. A summary of research contributions will be discussed in this section.

- ***Video Summarisation Based on Group Scoring:*** As opposed to earlier user-centred approaches, our proposed techniques adopt multiple users to minimise the negative effects of a single user's employment such as subjectivity and distraction. Accordingly, a novel domain-independent technique for summarising videos was proposed in the fourth chapter. Here, a number of scenarios in which more than one

operator is required and engaged in the video abstraction process. In the most closely related work to ours, the Click2Summary framework (Wu et al., 2011), a video abstraction model based on crowdsourcing was proposed in which the input video sequence was primarily segmented into five seconds shots and those segments with highest selection rate by operators were inserted into video digests. However, segmentation of a video shot based solely on the time element can increase the possibility of generating false results. This is due to the inability of this method to address the dramatic change in the visual and semantic content within each sub-segment. In addition, this method is not capable to produce video summaries at different levels incorporating timing requirements. This can be attributed to the fact that there is a binary label assigned by operators for each shot; therefore many shots with average quality (that should be included considering the time constraint and context of video) could be missed. These shortcomings were all addressed in our proposed approach by providing video workers with facilities to assign scores (rather than binary labelling) to the frames (rather than temporally segmented shots) and extraction of highest scored frames. In addition, our method is more practical compared to the previously recommended fully-automatic approaches, given their significant computational expense and time overhead**.**

- ***Novel Evaluation Methodology:*** In this work, a novel method to assess the effectiveness of video summaries was introduced. As opposed to earlier methodologies, a combinational framework was devised and utilised in our research to evaluate the capability of our proposed summarisation approaches based on statistical metrics (Recall and Precision) alongside the users' perceived quality of summaries (Overall Satisfaction) as the most important indicator to represent the quality of a multimedia content. Earlier work either concentrated on mathematical concepts or took into account the users' perceived experience. Later, a comparison-based technique was developed to assess the quality of summaries generated based on different modalities. This was carried out by feeding the input clip to summarisation tools that were employing visual, visual-aural and visual-aural-textual characteristics to skim the videos. To the best of our knowledge, this approach had never been used for evaluation purposes of video abstraction techniques.

- ***Personalised User-Centred Video Summarisation:*** In chapter 5, a new mechanism was introduced to understand users' priorities towards different video scenes and customising the final summaries based on the retrieved data. As a result, viewers were provided with a list of keyframes and associated audio-visual tags corresponding to the content of each scene. Therefore, the audience could potentially comprehend the subject matter of each scene in a fast manner prior to expressing their level of interest into that particular segment. Moreover, the approach employed for extraction of the representative keyframe for each scene should be considered as another contribution. In order to do so, each video scene was initially segmented into three shots of equal temporal length. The highest scored video frame closest to the centre of each shot was nominated to represent that scene. Finally, the keyframe that had the highest selection rate by different video operators between the three available choices during the annotation process was chosen as the representative keyframe for that scene. In addition, the proposed method for updating the particular frames residing in scenes with higher level of importance is novel.

- ***Personalised Video Summarisation Using SIFT:*** In chapter 6, a novel mechanism was proposed to tailor the video summaries in accordance with end-users' interests. SIFT features of the representative keyframe for each scene were used as the basis for the personalisation module. Thus, the number of common SIFT features between a keyframe and training images was used as a metric to show the relevancy of that scene to a particular high-level concept. Hence, each video scene (keyframe) could be represented as a vector, whose element represented the dependency level of that scene to a particular high-level concept based on the SIFT similarity measure. Furthermore, a unique mechanism was adopted in order to create the profiles for the users and to fuse those profiles into the summarisation modules for generating the final summaries. Thereafter, the dot product of the border representing the users' interest into the high-level concepts and the array indicating the relevancy of each video scene into the same categories was employed as a metric to assess the priority level of each scene for a particular user. To the best of our knowledge, the proposed algorithm for personalising the video summaries based on SIFT features has not been previously adopted.

7.4. Research Limitations and Future Work

In this section some directions for future research in order to reduce the limitations and assumptions corresponding to our research will be briefly discussed.

In spite of some promising results achieved by our proposed video summarisation models, the extra cost of human involvement can be considered rather high, as video operators still have to be actively engaged in the frame scoring process. Therefore, devising a scheme in which the saliency scores for the frames can be calculated based on a video operator's perceived interest without requiring him/her to directly mark the video segments could be regarded as an interesting topic for future research.

Based on our second approach, the end-users' priorities towards different video scenes were obtained in order to customise the final summary in respect to each user's interests. However, the audience has to go through the list of representative keyframes and corresponding tags manually in order to prioritise the scenes. Thus, gathering information historically regarding users during different iterations in an attempt to minimise their level of direct intervention can be considered as another topic for additional study in the future. Moreover, extracting the audio-visual information of the scenes in a more convenient fashion than direct annotation should be explored in later work.

In chapter 6, user profiles were formed based on their assigned scores to each high-level visual concept. As a result, the personalisation of video abstracts was carried out by considering solely visual information (no audio-textual data), which can deteriorate the effectiveness of our personalised video summarisation approach. Accordingly, creating profiles for the viewers based on different information resources (audio, visual and textual) and trying to integrate them with different modalities retrieved data from input video sequence is another direction for future research.

Last but not least, our recommended techniques were evaluated against a group of automatic summarisation systems in order to measure their effectiveness. Comparing our methods against the additional techniques, which involve humans in the abstraction task, can be considered as more generalisable comparison that can be addressed in future work. However, the reluctance of other researchers to either provide us with their developed tools or to summarise our videos using their systems themselves directly should be mentioned as a

major barrier. Moreover, using experimental videos of varied length for our evaluation purposes can also further increase the applicability of our research. All our valuable directions for future endeavours.

References

Ajmal, M., Ashraf, M.H., Shakir, M., Abbas, Y. and Shah, F.A. 'Video Summarization: Techniques and Classification', *Computer Vision and Graphics, International Conference, ICCVG 2012, Warsaw,* Poland, pp. 1-13.

Aldridge, R., Davidoff, J., Ghanbari, M., Hands, D., and Pearson, D., (1995) 'Recency effect in the subjective assessment of digitally coded television pictures', *In Proc. IPA, Edinburgh, UK,* pp. 336-339.

Almeida, J., Leite, N.J. and Torres, R.S. (2013) 'Online video summarization on compressed domain', *Journal of Visual Communication and Image Representation,* 24 (6), pp. 729-738.

Almeida, J.; Leite, N.J. and Torres, R.S. (2012) 'VISON: Video Summarization for Online applications', *Pattern Recognition Letters,* Vol. 33, pp.397-409.

Angelides, M.C. (2004) 'Multimedia content modelling and personalization, *IEEE Multimedia,* 10(4), pp. 12-15.

Ashwin Raju, D.K. and Velayutham, C.S. (2009) 'A study on Genetic Algorithm based video abstraction system', *Nature & Biologically Inspired Computing. NaBIC 2009. World Congress on, Coimbatore, India, IEEE,* pp. 878-883.

Avison, D. and Fitzgerald, G. (2006) Information Systems Development: Methodologies, Techniques & Tools, *4th Edition, McGraw-Hill Education,* UK.

Babaguchi, N., Kawai, Y., Ogura, T. and Kitahashi, T. (2004) "Personalized abstraction of broadcasted American football video by highlight selection," Multimedia, IEEE Transactions 6(4), pp.575-586.

Babaguchi, N., Ohara, K. and Ogura, T. (2003) 'Effect of personalization on retrieval and summarization of sports video', *Conference on Multimedia information, communication and signal processing,* pp. 940-944.

Baek, J.S., Lee, S.T. and Baek, J.H. (2005) 'Scene Boundary Detection by using Shot Clustering and Music Detection', *Artificial intelligence, Portuguese conference on, Covilha, Portugal, Springer,* pp. 94-97.

Bailer, W., Dumont, E., Essid, S., Merialdo, B. (2008) 'A collaborative approach to automatic rushes video summarization', *Image Processing, 2008. ICIP 2008. 15th IEEE International Conference on, California, USA, IEEE,* pp. 29-32.

Bailer, W., Lee, F. and Tallinger, G. (2007) 'Skimming rushes video using retake detection, In proceedings of the international workshop on TRECVID Video Summarisation, ACM, Augsburg 2007.

Bay, H., Tuytelaars, T., Van Gool, L. (2006) 'SURF: Speeded Up Robust Features', *Proceedings of the ninth European Conference on Computer Vision*, Berlin, Germany, ACM, Vol. 3951, pp 404-417.

Beckett, D. (2012) 'RDF/XML syntax specification, www.w3c.org/TR, Accessed October 2012.

Bhatt, R.B., Krishnamoorthy, P. and Kumar, S. (2009) 'Efficient general genre video abstraction scheme for embedded devices using pure audio cues', *ICT and Knowledge Engineering, 2009 7th International Conference on, Bangkok, Thailand, IEEE,* pp. 63-67.

Belkin, M. and Niyogi, P. (2001) 'Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering', *Advances in Neural information Processing Systems,* pp. 14.

Benini, S., Migliorati, P. and Leonardi, R. (2007) 'Hidden Markov models for video skim generation', *Eighth International Workshop on Image Analysis for Multimedia Interactive, Santorini, Greece, IEEE,* pp. 6-6.

Bertini, M., Cucchiara, R., Del Bimbo, A. and Torniai C. (2005) 'Ontologies enriched with visual information for video annotation', *in Proc. 2nd Eur. Semantic Web Conf., Heraklion, Greece. ACM,* pp. 1-6.

Boehm, B.R. and Turner, R. (2004) Balancing Agility and Discipline: A Guide for the Perplexed, *Addison-Wesley.*

Beom, M., Williem, L. and Park, I. (2013) 'Spatiotemporal Saliency-Based Video Summarization on a Smartphone, *JBE,* 18(2), pp.185-195.

Boreczky, J. and Rowe, S. (1996) 'Comparison of video shot boundary detection techniques', *in Proceeding of Conference On Visual Communication and Image Processing*, pp. 1-6.

Boudreau, M., Gefen, D. and Straub, D.W. (2001) 'Validation in information systems research: A state-of-the-art assessment', *MIS Quarterly: Management Information Systems,* 25(1), pp. 1-16.

Burg, J. (2009), The science of digital media, *Pearson International edition*, New Jersey.

Cahuina, E.J.Y.C. and Chavez, G. (2013) 'A New Method for Static Video Summarization Using Local Descriptors and Video Temporal Segmentation', *Graphics, Patterns and Images (SIBGRAPI), 2013 26th SIBGRAPI - Conference on, Brazil, IEEE,* pp. 226-233.

Carvajal, J., McCool, C. and Sanderson, C. (2014) 'Summarisation of short-term and long-term videos using texture and colour', *Applications of Computer Vision (WACV), 2014 IEEE Winter Conference on, Steamboat Springs , IEEE,* pp.769-775.

Caschera, M.C. and D'Ulizia, A. (2007) 'Information extraction based on personalisation and contextualization models for multimodal data', *Database and Expert Systems Applications, 2007. 18th International Workshop on, Regensburg, Germany, IEEE,* pp.114-118.

Chen, F., De Vleeschouwer, C. and Cavallaro, A. (2014) 'Resource Allocation for Personalized Video Summarization', *Multimedia, IEEE Transactions on*, 16(2), pp. 455-469.

Christopoulou, E., Goumopoulos, C. and Kameas, A. (2005) 'An ontology-based context management and reasoning process for Ubicomp application', *Proceedings of the 2005 joint conference on Smart objects and ambient intelligence: innovative context-aware services: usages and technologies, France, IEEE,* pp. 265-270.

Choras, R.S. (2007) 'Feature extraction for CBIR and biometrics applications', *Proceedings of the 7th Conference on 7th WSEAS International Conference on Applied Computer Science*, Vol. 7, ACM, pp. 1-9.

Chung, M. G., Wang, T. and Sheu, P. (2011) 'Video Summarization Based on Collaborative Temporal Tags', *Journal of Online information review*, 35(4), pp. 653-668.

Ciocca, G. and Schettini, R. (2006) 'An innovative algorithm for key-frame extraction in video summarisation', *Journal of Real-Time Image Processing (Springer),* 1(1), pp. 69-88.

Corridoni, J.M. and Bimbo, A.M. (1998) 'Structured representation and automatic indexing of movie information content', *Pattern Recognition*, 31(12), pp. 2027-2045.

Creswell, J. (2008) Educational research: Planning, conducting, and evaluating quantitative and qualitative research, *Pearson: Merrill Prentice Hall*, New Jersey.

Cunha, T.O., Souza, F.G.H, Araujo, A.A. and Pappa, G.L. (2012) 'Rushes video summarization based on spatio-temporal features', *Proceedings of the 27th Annual ACM Symposium on Applied Computing, ACM*, pp. 45-50.

Daneshi, M., Vajda, P., Chen, D.M., Tsai, S.S., Yu, M.C., Araujo, A.F., Chen, H. and Girod, B. (2013) 'Eigennews: Generating and delivering personalized news video', *Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on, San Jose, USA, IEEE*, pp. 1-6.

Dang, C.T., Radha, H. (2014) 'Heterogeneity Image Patch Index and Its Application to Consumer Video Summarization', *Image Processing, IEEE Transactions on*, 23(6), pp. 2704-2718.

Datta, R., Li, J. and Wang, J. (2005) 'Content-Based Image Retrieval—Approaches and Trends of the New Age', *Proc. ACM Multimedia Workshop Multimedia Information Retrieval, New York, USA, ACM*, pp. 253-262.

Dawson, C. (2002) Practical Research Methods, *UBS Publishers' Distributors*, New Delhi.

Dimitrova, N., Zhang, H.J., Shahraray, B., Sezan, I., Huang, T. and Zakhor, A. (2002) Applications of video content analysis and retrieval', *IEEE Multimedia*, 9(3), pp. 42–55.

Dumont, E and Merialdo, B. (2007) 'Split-screen dynamically accelerated video summaries', *In proceedings of the international workshop on TRECVID Video Summarisation, ACM, Augsburg 2007.*

Eldib, M.Y., Zaid, B., Zawbaa, H.M., El-Zahar, M., El-Saban, M. (2009) 'Soccer video summarization using enhanced logo detection', *Image Processing (ICIP), 2009 16th IEEE International Conference on, Cairo, Egypt, IEEE,* pp.4345-4348.

Ekin, A., Tekalp, A. M. and Mehrotra, R. (2003) 'Automatic soccer video analysis and summarization', *IEEE Trans. Image Processing*, 12(5), pp. 796-807.

Evangelopoulos, G., Zlatintsi, A., Potamianos, A., Maragos, P., Rapantzikos, K., Skoumas, G. and Avrithis, Y. (2013) 'Multimodal Saliency and Fusion for Movie Summarization based on Aural, Visual, and Textual Attention', *IEEE Transactions on Multimedia,* 15(7), pp.1553-1568.

Evangelopoulos, G., Zlatintsi, A., Skoumas, G., Rapantzikos, K., Potamianos, A., Maragos, P. and Avrithis, Y. (2009) 'Video Event Detection and Summarization Using Audio, Visual and Text Saliency', *Proc. IEEE Int'l Conf. on Acoustics, Speech and Signal Processing (ICASSP-09), Taipei, Taiwan, IEEE,* pp.1-6.

Everett, H. (1963) 'Generalized lagrange multiplier method for solving problems of optimum allocation of resources', *Operation Research.,* 13(3), pp. 399-417.

Fallman, D. (2003) 'Design-oriented Human—Computer Interaction', *In: SIGCHI Conference on Human Factors in Computing Systems, 2003 USA,* pp. 225-232.

Ferman, A. and Tekalp, A. (2003) 'Two-stage hierarchical video summary extraction to match low-level user browsing preferences', *IEEE Transactions on Multimedia,* Vol. 5, pp.244-256.

Frey, B. J. and Dueck, D. (2007) 'Clustering by passing messages between data points', *Journal of Science,* Vol. 315, pp. 972-976.

Fukumura, S., Nakano, T. , Harumoto, K., Shimojo, S., Nishio, S. (2003) 'Realization of personalized presentation for digital contents based on browsing history', *Communications, Computers and signal Processing, 2003. PACRIM. 2003 IEEE Pacific Rim Conference on, IEEE,* Vol. 2, pp. 605-608.

Furini, M., Geraci, F., Montangero, M. and Pellegrini, M. (2010) 'Stimo: Still and moving video storyboard for the web scenario', *Multimedia Tools Applications,* 46(1), pp. 47–69.

Gao, Y., Wang, W.B., Yong, J.H. and Gu, H.J. (2009) 'Dynamic video summarization using two-level redundancy detection', *Multimedia Tools and Applications,* 2(2), pp. 233–250.

Geambasu, C., Jianu, I., Jainu, I. and Gavrila, A. (2011) 'INFLUENCE FACTORS FOR THE CHOICE OF A SOFTWARE DEVELOPMENT METHODOLOGY', *Journal of Accounting and Management Information Systems,* 10(4), pp. 479-494.

Ghinea, G., Kannan, R., Swaminathan, S. and Kannaiyan, S. (2014) 'A novel user-centered design for personalized video summarization', *Multimedia and Expo Workshops (ICMEW), 2014 IEEE International Conference on, Chengdu, China, IEEE,* pp. 1-6.

Glass, R. (1999) 'Computing calamities: lessons learned from products, projects, and companies that failed', *Design IEEE Software*, 16(2), pp. 103-104.

Graf, R.F. (1999), Modern Dictionary of Electronics, *Newnes*, Oxford.

Gray, J., Chaudhuri, S., Bosworth, A., Layman, A., Reichar, D., Venkatro, M., Pellow, F. and Pirahesh, H. (1997) 'Data cube: A relational aggregation operator generalizing group-by, cross-tab, and sub totals', *Data Mining and Knowledge Discovery,* 1(1), pp. 29-53.

Gulliver, S.R and Ghinea, G. (2006) 'Defining the users perception of distributed multimedia quality', *ACM Transactions on Multimedia Computing, Communications, and Application*, 2(4), pp. 241-257.

Gulliver, S.R. (2004) Distributed Multimedia Quality: The User Perspective, Doctor of Philosophy *edn*, *Brunel University*, United Kingdom.

Guo, Y., Zhu, Y., Liu, F., Song, C. and Zhou , H. (2012) 'Multi-view video summarisation', *Multimedia, IEEE*, 12(7),pp. 717 – 729.

Han, J., Li, K., Shao, L., Hua, X., He, S., Guo, L., Han, J. and Liu, T. (2014) 'Video abstraction based on FMRI-driven visual attention model', *Information sciences*, Vol. 281, pp. 781-796.

Han, B., Hamm, J. and Sim, J. (2011) 'Personalised video summarization with human in the loop', *Applications of Computer Vision (WACV), 2011 IEEE Workshop on, Hawaii, IEEE,* pp. 51-57.

Hanckok M. and Walker, S. (1992) 'An evaluation of automatic query expansion in an online library catalogue, *J-documentation*, pp. 406-421.

Hanjalic, A. and Zhang, H. (1999) 'An Integrated Scheme for Automated Video Abstraction based on Unsupervised Cluster-Validity Analysis', *Circuits and Systems for Video Technology, IEEE Transaction on*, 8(9), pp.1280-1289.

Harel, J., Koch, C. and Perona, P. (2007) 'Graph-based visual saliency', *In Advances in Neural Information Processing Systems 19, MIT press,* pages 545-552.

Hari, R., Roopesh, C.P. and Wilscy, M. (2013) 'Human face based approach for video summarization', *Intelligent Computational Systems (RAICS), 2013 IEEE Recent Advances in, Trivandrum, India, IEEE*, pp. 245-250.

Hays, J. and Efros, A.A. (2007) 'Scene Completion Using Millions of Photographs', *ACM Trans. Graph.* 26 (3), Article 4.

Herranz, L. and Martinez, J.M. (2008) 'Integrated summarization and adaptation using H.264/MPEG-4 SVC', *Visual Information Engineering, 2008. VIE 2008. 5th International Conference on, Xi'an, China, IEEE,* pp.729-734.

Hevner, A. and Chatterjee, S. (2010) Design Research in Information Systems: Theory and Practice 1st edition, *Springer*.

Hopfgartner, F., Jose, J.M., Yu, Z., Lugmayr, A., Chorianopoulos, K. and Mei, T., (2010) 'Semantic user profiling techniques for personalised multimedia recommendation, *MULTIMEDIA SYSTEMS*, 16(5), pp. 255-274.

Howell, K. E. (2013) Introduction to the Philosophy of Methodology, *Sage Publications*, London.

Http://www.avcutty.de/english/ (Accessed 25 March 2014)

Http://www.image-net.org (Accessed 28 April 2014)

Http://www.oxforddictionaries.com (Accessed June 2013)

Iparraguirre, J. and Delrieux, C. (2013) 'Speeded-Up Video Summarization Based on Local Features', *Multimedia (ISM), 2013 IEEE International Symposium on, California, USA,IEEE,* pp. 370-373.

Iu, s., King, I. and Lyu, M.R. (2004) 'video summarisation by video structure analysis and graph optimization', *ICME 2004, Multimedia and Expo 2004, IEEE International Conference on*, IEEE, pp.1959-1962.

Jaimes, A., Echigo, T., Teraguchi, M. and Satoh, F. (2002) 'Learning personalized video highlights from detailed MPEG-7 metadata', *Image Processing 2002, International Conference on, IEEE,* Vol.1, pp.I-133-I-136.

Jenkins, A. M. (1985) 'Research Methodologies and MIS Research', *In Research Methods in Information Systems, Elsevier Science Publishers B.V., Amsterdam, Holland,* pp. 103-117.

Jian, Y., Meng, D. and Xiu, Y. (2010) 'Application of information theory in video abstraction extraction', *Environmental Science and Information Application Technology (ESIAT), 2010 International Conference on, Wuhan, China, IEEE,* pp. 112-115.

Jiang, R.M., Sadka, A.H. and Crookes, D. (2009) 'Hierarchical video summarization in reference subspace', *Consumer Electronics, IEEE Transactions on*, 55(3), pp.1551-1557.

Jiang, H., Zhang, H.J., and Lin, T. (2000) 'Video segmentation with the Support of Audio Segmentation and classification', *Microsoft Research China*, pp. 1-4.

Jokela, S.; Sulonen, R. and Turpeien, M. (1999) 'Agents in delivering personalised content based on semantic metadata', *In Proceeding of intelligent in cyberspace, Citeseerx,* pp 89-93.

Kang H., Lee, S. and CH, C. K. (2007) 'Coherent Line Drawing', *In Proceedings of the 5th international symposium on Non-photorealistic animation and render in, ACM,* pp.43-50.

Kane, E. and O' Reily, M. (2001) Doing your Own Research, *Marion Boyars*, London.

Kapoor, A, Biswas, K.K. and Hanmandlu, M. (2013) 'Fuzzy video summarization using key frame extraction', *Computer Vision, Pattern Recognition, Image Processing and Graphics (NCVPRIPG), 2013 Fourth National Conference on, Gujarat, India, IEEE,* pp.1-5.

Khosla, A., Hamid, R., Lin, C.J., Sundaresan, N. (2013) 'Large-Scale Video Summarization Using Web-Image Priors', *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, Portland, USA, IEEE,* pp. 2698-2705.

Kim, S., Yoon, K.J. and Kweon, I.S. (2006) 'Object Recognition Using a Generalized Robust Invariant Feature and Gestalt's Law of Proximity and Similarity', *Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'06), IEEE,* pp. 193-193.

Kim, E.J., Lee, G.G., Jung, C., Kim, S.K., Kim, J.Y. and Kim, W.Y. (2005) 'A video summarization method for basketball game', *In Proceedings of the 6th Pacific-Rim conference on Advances in Multimedia Information Processing, Springer,* Vol. 1, pp. 765-775.

Kirk, R.E. (1995) Experimental Design: Procedures for the Behavioral Sciences, *Pacific Grove,* CA: Brooks/Cole.

Kuhn, T. (1996) The Structure of Scientific Revolutions, Chicago: *University of Chicago Press.*

Kumar, S. and Singh, S. (2011), Business Research Methods, *Thakur Publishers,* New Delhi.

Lazebnik, S., Schmid, C., and Ponce, J. (2004) 'Semi-Local Affine Parts for Object Recognition', *Proceedings of the British Machine Vision Conference, Kingston, UK,* pp.1-10.

Leary, M. (1995) Behavioural Research Method, *Pacific Grove, Brooks/Cole Publishing*.

Li, P., Guo, Y. and Sun, H. (2011a) 'Multi-keyframe abstraction from videos', *Image Processing (ICIP), 2011 18th IEEE International Conference* on, *Brussels, Belgium, IEEE,* pp. 2473-2476.

Li, L., Zhou, K., Xue, G.R., Zha, H. and Yu. Y. (2011b) 'Video Summarization via Transferrable Structured Learning', *Proceedings of the 20th international conference on World wide web, Hyderabad, India, ACM,* pp. 287-296.

Li, Y. and Merialdo, B. 'Multi-video summarization based on Video-MMR', *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, pp.1-4.

Li, Y. N., Lu, Z. M. and Niu, X.M. (2009a) 'Fast video shot boundary detection framework employing pre-processing techniques', *Image Processing, IET* , 3(3), pp. 121-134.

Li, J, Ding, Y., Shi, Y. and Li, W. (2009b) 'Efficient Shot Boundary Detection Based on Scale Invariant Features', *Image and Graphics, 2009. ICIG '09. Fifth International Conference on,* pp.952-957.

Li, X., Chen, L., Zhang, L., Lin, F. and Ma, W. (2006) 'Image Annotation by Large-Scale Content-Based Image Retrieval', *Proc. ACM Int'l Conf. Multimedia, ACM,* pp. 607-610.

Li, J. and Wang, J. (2003) 'Automatic Linguistic Indexing of Pictures by Statistical Modelling Approach', *IEEE Trans. Pattern Analysis and Machine Intelligence*, 25(9), pp. 1075-1088.

Lie, W.N. and Hsu, K.C. (2008) 'Video Summarisation Based on Semantic Feature Analysis and User Preference', *Sensor Networks, Ubiquitous and Trustworthy Computing, 2008. SUTC '08. IEEE International Conference on, IEEE,* pp. 486-491.

Liu, Y., Liu, H., Liu, Y. and Sun, F. (2014a) 'User-generated-video summarization using Sparse Modelling', *Neural Networks (IJCNN), 2014 International Joint Conference on*, pp. 3909-3915.

Liu, Y., Liu, H., Liu, Y. and Sun, F. (2014b) 'Outlier-attenuating summarization for user-generated-video', Multimedia and Expo (ICME), *2014 IEEE International Conference on, , Chengdu, China, IEEE,* pp. *1*-6.

Liu, Y., Liu, Y., Ren, T. and Chan, K. (2008) 'Rushes video summarization using audio-visual information and sequence alignment', *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop, ACM,* pp. 114-118.

Liu, T., Zhang, H.J. and Qi, F. (2003) 'A novel video key frame extraction algorithm based on perceived motion energy model', *IEEE transactions on circuits and systems for video technology,* 13(10), pp.1006-1013.

Lowe, D. G. (2004) 'Distinctive Image Features from Scale-Invariant Keypoints', *International Journal of Computer Vision*, 60 (2), pp. 91-110.

Lu, S., Wang, Z., Mei, T., Guan, G. and Feng, D. (2014) 'A Bag-of-Importance Model with Locality-Constrained Coding based Feature Learning for Video Summarization', *Multimedia, IEEE Transactions on,* 16(6), pp.1497-1509.

Lu, C., Drew, M. and Au, J. (2001) 'Classification of summarized videos using hidden Markov Models on compressed chromaticity signatures', *Proceeding of 9th ACM international conference on multi-media, Ottawa, Canada, ACM,* pp. 479-482.

Mahmoud, K.M., Ghanem, N.M., Ismail, M.A (2013) 'VGRAPH: An Effective Approach for Generating Static Video Summaries', *Computer Vision Workshops (ICCVW), 2013 IEEE International Conference on, Sydney, Australia, IEEE,* pp.811-818.

Malheiro, B., Foss, J., Burguillo, J.C., Peleteiro, A and Mikic, F.A (2011) 'Dynamic Personalisation of Media Content', *Semantic Media Adaptation and Personalization (SMAP), 2011 Sixth International Workshop on, Vigo, Spain, IEEE,* pp. 21-26.

Manning, C., Raghvan, P. and Schutze, H. (2009), Introduction to information retrieval, *Cambridge*, UK.

Markaki, O.I. (2009) 'Personalization Mechanisms for Content Indexing, Search, Retrieval and Presentation in a Multimedia Search Engine', *16th International Conference on Systems, Signals and Image Processing, 2009, Duborvnik, Croatia, IEEE,* pp.1-9.

Martin, J. (1991) *Rapid Application Development*. *Macmillan,* USA.

Maybury, M., Greiff, W., Boykin, S., Ponte, J., Mac henry, C. and Ferro, L. (2004) 'Personalcasting: Tailored broadcast News', *User Modelling and User-Adapted Interaction,* Vol.14, pp.119-144.

Merialdo, B., Lee, K.T., Luparello, D. and Roudaire., J. (1999) 'Automatic construction of personalized TV news programs', *In Proceedings of the seventh ACM international conference on Multimedia (MULTIMEDIA '99). ACM, New York,* pp. 323-331.

Mertens, D., (1998) Research Methods in Education and Psychology. *Thousand Oaks,* California, Sage.

Mei, S., Guan, G., Wang, Z., He, M., Hua, X. and Dagan F. D. (2014) 'L2,0 constrained sparse dictionary selection for video summarization', *Multimedia and Expo (ICME), 2014 IEEE International Conference on, Chengdu, China, IEEE,* pp.1-6.

Metze, F., Ding, D., Younessian, H. and Hauptman, A. (2013) 'Beyond audio and video retrieval: topic-oriented multimedia summarization', *International Journal of Multimedia Information Retrieval,* Vol. 2, pp. 131-144.

Mikolajczyk, K., and Schmid, C. (2005) 'A performance evaluation of local descriptors', *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 10(27), pp. 1615-1630.

Mingers, J. (2001) 'Combining IS Research Methods: Towards a Pluralist Methodology', *Information Systems Research,* 12 (3), pp. 240-259.

Mobasher, B., Cooley, R. and Siravstava, J. (2000) 'Automatic personalization based on Web usage mining', *Communications of the ACM,* 43(8), pp.142-151.

Money, A.G. & Agius, H. (2009) 'Analysing User Physiological Responses for Affective Video Summarisation', *Displays (Elsevier),* 30(2), pp. 59-70.

Money, A.G. & Agius, H. (2007) 'Video summarisation: A Conceptual Framework and Survey of the State of the Art', *Journal of Visual Communication and Image Representation (Elsevier),* 19(2), pp.121-143.

Mortimer, A.J. (1995) 'Project management in rapid application development', *Project Management for Software Engineers, IEE Colloquium on,* pp.5/1-5/3.

Myers, M. D. (1997) Qualitative Research in Information Systems, *MIS Quarterly.* Available: http://www.qual.auckland.ac.nz/ (Accessed September 2014).

Nalin, C., Jonghun, H. and Saewoong, B. (2011) Context-aware ontological schemes for multimedia personalization," *ICT Convergence (ICTC), 2011 International Conference on*, *Seoul, Korea, IEEE*, pp. 288-289.

Ngo, C., Ma, Y. and Zhang, H.J. (2005) 'Video summarisation and scene detection by graph modelling', *Circuits and Systems for Video Technology, IEEE Transactions on*, 15(2), pp. 296-305.

Nilsson, A. G. (2005) Information Systems Development, *Springer*, US.

Nixon, M.S. and Aguado, A.S. (2008), Feature extraction and image processing, *Newnes*, UK.

Oh, J., Wen, Q., Lee, J. and Hwang, S. (2004) 'Video abstraction', *In Video Data Management and Information Retrieval, IRM press*, pp. 321-346.

Otsuka, I., Radhakrishnan, R., Siracusa, M., Divakaran, A. and Mishima, H. (2006) 'An enhanced video summarization system using audio features for a personal video recorder', *Consumer Electronics, IEEE Transactions on*, 52(1), pp.168-172.

Ou, S., Lee, C. , Somayazulu, V.S., Chen, Y.K. and Chien, S.Y. (2014) 'Low complexity on-line video summarization with Gaussian mixture model based clustering', *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, *Florence, Italy, IEEE,* pp.1260-1264.

Ouyang, J.Q and Liu, R. (2013) 'Ontology reasoning scheme for constructing meaningful sports video summarisation', *Image Processing, IET*, 7(4), pp.324-334.

Pan, L., Wu, W. and Shu, X. (2009) 'Key Frame Extraction Based on Sub-Shot Segmentation and Entropy Computing', *Pattern Recognition, 2009. CCPR 2009. Chinese Conference on*, pp. 1-5.

Park, H.S.and Cho, S.B. (2011) 'A personalized summarization of video life-logs from an indoor multi-camera system using a fuzzy rule-based system with domain knowledge, *Information Systems*, 36(8), pp. 1124-1134.

Patrikakis, C., Papaoulakis, N., Papageorgiou, P., Pnevmatikakis, A, Chippendale, P., Nunes, M.S., Cruz, R.S., Poslad, S. and Zhenchen W. (2011) 'Personalized Coverage of Large Athletic Events', *MultiMedia, IEEE*, 18(4), pp. 18-29.

Peeters, G. (2004) 'A large set of audio features for sound description (similarity and classification)', *in the CUIDADO project*, France.

Peng, W.T., Huang, W.J., Chu, W.T., Chou, C.N.; Chang, W.Y., Chang, C.H. and Hung, Y.P. (2009) 'A User Experience Model for Home Video Summarization', *In Proceedings of the 15th International Multimedia Conference on Advances in Multimedia Modelling*, *France, Springer*, pp. 484-495.

Perez-Daniel, K.R., Nakano Miyatake, M., Benois-Pineau, J., Maabout, S. and Sargent, G. (2014) 'Scalable video summarization of cultural video documents in cross-media space based on data cube approach', *Content-Based Multimedia Indexing (CBMI), 2014 12th International Workshop on, Klagenfurt, Austria, IEEE,* pp. 1-6.

Qiang, Y. and Sen, M. (2006) 'The Method of Key Frame Distilling Based on Similar Ratio Between Frames', *China Cable Television,* pp.1681-1683.

Ren, W. and Zhu, Y. (2008) 'Video summarisation approach based on machine learning', *Intelligent Information Hiding and Multimedia Signal Processing, 2008. IIHMSP '08 International Conference on, Harbin, China, IEEE,* pp. 450-453.

Rovira, M.M., Garcia, S., Berengue, M.C. and Fernandez, G. (2007) 'Metadata Management and Personalisation of Audio-visual Content Using MPEG-7 and MPEG-21 Standards in a Distributed Framework, Automated Production of Cross Media Content for Multi-Channel Distribution', *AXMEDIS '07, Third International Conference on, IEEE,* pp. 7-10.

Saber, E. and Tekalp, A.P. (1998) 'Integration of color, edge and texture features for automatic region-based image annotation and retrieval', *Electronic Imaging,* Vol. 7, pp. 684–700.

Shafeian, H. and Bhanu, B. (2012) 'Integrated personalized video summarization and retrieval', *Pattern Recognition (ICPR), 2012 21st International Conference on, Tsukuba, Japan, IEEE,* pp. 996-999.

Shields, P. and Rangarjan, N. (2013) A Playbook for Research Methods: Integrating Conceptual Frameworks and Project Management, *New Forum press*, US.

Simon, H. A. (1996) The Sciences of the Artificial (3rded.*), MIT Press,* Cambridge.

Sklar, R. (1990), Film: An International History of the Medium, *Thames and Hudson,* London.

Skondras, E., Louta, M. and Sarigiannidis, P. (2011) 'A Personalized Audio Web Service Using MPEG-7 and MPEG-21 Standards', *Informatics (PCI), 2011 15th Panhellenic Conference on, Kastoria, Greece, IEEE,* pp. 199-204.

Smits, E. and Hanjalic, A. (2010) 'A System Concept for Socially Enriched Access to Soccer Video Collections', *MultiMedia, IEEE,* 17(4), pp.26-35.

Sorwar, G., Dooley, L. and Murshed, M. (2002) 'Integrated Technique with Neuro-computing for Temporal Video Segmentation', *Advances in Soft Computing,* Vol.14, pp. 159-167.

Stolterman, E. (2008) 'The Nature of Design Practice and Implications for Interaction Design Research', *International Journal of Design,* 2**,** 55-65.

Sun, X. and Kankanhalli, M. (2000) Video Summarization Using R-Sequences", *Journal of Real-Time Imaging*, Vol.6, pp. 449-459.

Stauffer, C. and Grimson, W.E.L. (1999) 'Adaptive background mixture models for real-time tracking', *in Proc. IEEE Computer Soc. Conf. Computer Vision and Pattern Recognition (CVPR), IEEE*, Vol. 2, pp. 246-252.

Sun, Z., Jia, K. and Chen, H. (2008) 'Video Key Frame Extraction Based on Spatial-Temporal Color Distribution', *2008 International Conference on Intelligent Information Hiding and Multimedia Signal Processing, IEEE*, pp. 196- 199.

Sujatha, C., Chivate, A.R., Tabib, R.A. and Mudengudi, U. (2014) 'Multilevel framework for summarization of surveillance videos', *Signal and Image Processing (ICSIP), 2014 International Conference on, Paris, France, IEEE,* pp. 265-270.

Sujatha, C. and Mudenagudi, U. (2011) 'A study on key frame extraction methods for video summary', *in Proc. Int. Conf. Computational Intelligent and Communication Networks, Gwalior, India, IEEE,* pp. 73-77.

Takahashi, Y., Nitta, N. and Babaguchi, N. (2005a) 'Video Summarisation for Large Sports Video Archives', *Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on,* pp.1170-1173.

Takahashi, Y., Nitta, N. and Babaguchi, N. (2005b) 'Automatic Video Summarization of Sports Videos Using Metadata', *Advances in Multimedia Information Processing* , Vol. 3332, pp. 272-280.

Taskiran, C.M., Pizlo, Z., Amir, A., Ponceleon, D. and Delp, E.J. (2006) 'Automated video program summarization using speech transcripts, *Multimedia, IEEE* , 8(4), pp. 775 - 791.

Thayer, R.H. and Boehm, B.W. (1986) Tutorial: software engineering project management, *Computer Society Press of the IEEE*. p.130.

Tong L., and Hong Z.H. (2000) 'Automatic video scene extraction by shot grouping', *Pattern Recognition, Proceedings. 15th International Conference on*, Vol.4, pp. 39-42.

Tropp, J.A., Gilbert, A.C. and Strauss, M.J. (2006) 'Algorithms for simultaneous sparse approximation, part I: Greedy pursuit', *Signal Processing*, 86(3), pp. 572-588.

Truong, B.T. and Venkatesh, S. (2007) 'Video abstraction: A systematic review and classification', *ACM Transactions on Multimedia Computing, Communications and Applications*, 3(1), Article 3, pp. 1-37.

Tseng, B.L. and Lin, C.Y. (2002) 'Personalized video summary using visual semantic annotations and automatic speech transcriptions', *Multimedia Signal Processing, 2002 IEEE Workshop on, Virgin Islands, USA, IEEE*, pp. 5-8.

Uchihashi, S., Foote, J., Girgenson, A. and Boreczky, J. (1999) 'Generating semantically meaningful video summaries', *Proceedings of ACM Multimedia'99, Orlando, FL, ACM,* pp. 383,-392.

Vaishnavi, V. and Kuechler, W. (2009) *Design Science Research in Information Systems.* DESRIST.org. Available at: http://desrist.org/desrist [Accessed October, 2014].

Valdes, V. and Martinez, J.M. (2008) 'Binary tree based on-line video summarization', *in Proc. ACM TRECVid Video Summarization Workshop, ACM,* pp. 134-138.

Vorhees E.M. (2008) 'On test collections for adaptive information retrieval', *information processing management,* 44(6), pp. 1879-1885.

Walls, J., Widmeyer, G. R. and El Sawy, O. A. (1992) 'Building an information systems design theory for vigilant EIS', *Information Systems Research,* 3, pp**.** 36 - 59.

Wang, J., Yang, J., Yu, K., Lv, F., Huang, T.S. and Gong. Y. (2010) 'Locality-constrained linear coding for image classification', *In Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 3360-3367.

Wen, G., Tuo, J., Jiang, L. and Wei, J. (2012) 'Audio feature extraction for classification using relative transformation', *Audio, Language and Image Processing (ICALIP), 2012 International Conference, Shanghai, China,* IEEE, pp. 260-265.

Winograd, T. (1996) Bringing Design to Software, *Addison-Wesley,* Reading.

Wolf, C., Jolion, J.M. and Chaassing, F. (2002) 'Text Localization , Enhancement and Binarization in Multimedia Document', *16th International conference on Pattern Recognition* , Vol. 2, pp.11-15.

Wong, K.W., Fung, C.C., Xiao, X. and Wong, K.P. 'Intelligent Customer Relationship Management on The Web', TENCON 2005 2005 IEEE Region 10, pp.1-5.

Wu, Z. and Xu, P. (2013) 'Shot boundary detection in video retrieval', *Electronics Information and Emergency Communication (ICEIEC), 2013 IEEE 4th International Conference on, Beijing, China,* IEEE, pp. 86-89.

Wu, S., Thawonmas, R. and Chen, K. (2011) 'Video Summarization via Crowdsourcing', *Proceeding CHI '11 Extended Abstracts on Human Factors in Computing Systems, Vancouver, Canada, ACM,* pp. 1531-1536.

Xu, C., Zhang. Y.F., Zhu, G., Rui, Y., Lu, H. and Huang, Q. (2008) 'Using Webcast Text for Semantic Event Detection in Broadcast Sports Video', *Multimedia, IEEE Transactions on,* 10(7), pp.1342-1355.

Yoshitaka, A., Sawada, K. (2012) 'Personalized Video Summarization Based on Behavior of Viewer', *Signal Image Technology and Internet Based Systems (SITIS), 2012 Eighth International Conference on , Marrakech, Morocco, IEEE,* pp.661-667.

You, J., Hannuksela, M. and Gabbouj, M. (2009) 'Semantic audio-visual analysis for video summarisation', *IEEE Region 8 EUROCON 2009 Conference (2009),IEEE,* pp.1358-1363.

Yeung, M.M. and Yeo, B.L. (1997) 'Video visualisation for compact presentation and fast browsing of pictorial content', *IEEE Transactions on Circuits and Systems for Video Technology,* 7(5), pp. 771-785.

Yeung, M.M. and Yeo, B.L. (1996) 'Time-constrained clustering for segmentation of video into story units', *In Proc. of ICPR'96*, *IEEE*, Vol. 3, p. 375.

Yuan J., Wang, H., Xiao, L., Zheng, W.; Li, J.; Lin, F. and Zhang, B. (2007) 'A Formal Study of Shot Boundary Detection', *Circuits and Systems for Video Technology, IEEE Transactions on*, 17(2), pp.168-186.

Zemcik, P., Potucek, I., Sumec, S., Herout, A., Hradis M., Beran V., Chmela P., Lanik A. and Mlich J. (2007) 'Video summarization at Brno University of Technology', *ACM, In proceeding of TRECVID 2007 rushes video summarisation*.

Zeng, X., Xie, X. and Wang, K. (2011) 'Instant video summarization during shooting with mobile phone', *in Proc. ACM Int. Conf. Multimedia Retrieval (ICMR), Trento, Italy*, *ACM,* pp. 40:1–40:8.

Zhang, Y., Ma, C., Zhang, J., Zhang, D. and Liu, Y. (2013) 'An interactive personalized video summarization based on sketches', *In Proceedings of the 12th ACM SIGGRAPH International Conference on Virtual-Reality Continuum and Its Applications in Industry, Hong Kong, ACM,* pp. 249-258.

Zhang, S., Zhang, Y., Chen, T., Hall, P. M. and Martin, R. (2007) 'Video structure analysis', *Tsinghua Science and Technology*, 12(6), pp. 714-718.

Zhang, D. and Chang, S.F. (2002) 'Event detection in baseball video using superimposed caption recognition', *in Proc. ACM Multimedia*, pp. 315-318.

Zhu, Y. and Zhou, D. (2003) 'Video browsing and retrieval based on multimodal integration', *Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on*, IEEE, pp.650-653.