An Alternative Representation of Video via Feature Extraction (RAAVE)

A thesis submitted to the

University of Dublin, Trinity College

for the degree of

Doctor of Philosophy

ADAPT Centre and CNGL

School of Computer Science and Statistics

Trinity College Dublin

Author: Fahim A. Salim

Supervised by: Prof Owen Conlan

Co-Supervised by: Prof Nick Campbell

2019

Declaration

I, the undersigned, declare that this work has not been previously submitted as an exercise for a degree at this or any other University, and that, unless otherwise stated, it is entirely my own work.

I agree to deposit this thesis in the University's open access institutional repository or allow the library to do so on my behalf, subject to Irish Copyright Legislation and Trinity College Library conditions of use and acknowledgement.

____

Fahim A. Salim

Permission to Lend or Copy

I, the undersigned, agree that the Trinity College Library may lend or copy this thesis upon request.

____

Fahim A. Salim

Acknowledgements

Many people have provided support to me during the course of my Ph.D. studies. Firstly, I would like to thank my supervisor Professor Owen Conlan for his guidance, support and encouragement over the years. And to my co-supervisor Professor Nick Campbell.

I would like to express my gratitude to the members of the ADAPT Centre, both student and staff members who participated in the user trials conducted in this thesis, for their generosity with their time and their invaluable feedback.

I would like to thank my mom and dad for their continuous encouragement and support. And finally, to my friends and colleagues for their continuous feedback.

Abstract

This thesis proposes a novel approach called RAAVE to transform video from a linear stream of content into an interactive multimedia document and thereby enhance the exploration potential of video content. Exploring the content of a video is typically inefficient due to the linear streamed nature of its media and the lack of interactivity i.e. video content is defined as a sequence of consecutive images with or without a parallel audio component. While researchers have proposed many approaches to enhance the exploration experience of video content; the general view of video content is still the same i.e. a continuous stream of images. It is the contention of this thesis that such a constrained view on video is limiting its potential value as a content source. For this reason, the presented thesis explores the idea of viewing video as a diverse content source, opening new opportunities and applications to explore and consume video content.

RAAVE transforms a video by representing its content as an automatically curated multimedia document. It does so by utilizing a template driven representation engine. Multimodal features are extracted automatically from the different modalities of video content, along with their timestamp, and stored in a repository. Upon receiving a content request, the representation engine utilizes a template collection to represent the content of a video in an appropriate configuration. By configuration it is meant that the presence and granularity of certain features are used to compose a representation of the source video. A video can have multiple multimodal representations. By automatically curating the content, the proposed approach allows users to not only configure the content in terms of the amount of detail, but also in terms of choice in the combination of different modalities.

A modular framework and algorithm for the representation engine and template collection is described. The framework design is influenced by the state of the art and user studies conducted to streamline the design. The representation engine-based approach is evaluated by developing a prototype system grounded on the design of the proposed approach, allowing users to perform different content exploration within a video, tasks. The evaluation demonstrated the benefits of the proposed approach in terms of enhancing the user's exploration experience with video content.

Table of Contents

List of Figures

List of Tables

1. Introduction

Video. The Merriam Webster[1] dictionary defines it as a digital recording of a set of images. Similarly Finke & Balfanz call it a sequence of consecutive images (Finke and Balfanz, 2004). Wikipedia[2] goes a step further and defines video as an electronic medium for the recording, copying, playback, broadcasting, and display of moving visual media. A more detailed definition is provided by Business Dictionary[3] , which defines it as a visual multimedia source that combines sequence of images that usually have corresponding audio components, transmitted to a screen and processed in order.

One is justified in wondering why this thesis starts with definitions which define the obvious. This is precisely the premise of the presented thesis i.e. despite all the research on video content, the general view of video content is that which has been described in the paragraph above: a sequence of moving images with or without an audio component. It is the contention of this thesis that such a constrained view on video is limiting its potential value as a content source. For this reason, the presented thesis explores the idea of viewing video content as a diverse content source, opening new opportunities and applications for exploring and consuming video content. In short, this thesis describes the need, design and evaluation of an approach to transform video content into “something more” and thereby enhance its exploration potential.

1.1. Motivation

Content consumption is becoming increasingly video oriented (Hong *et al.*, 2011; Mujacic *et al.*, 2012a; Masneri and Schreer, 2014). Whether a person wants to entertain him/herself in their free time or learn something new, one ends up relying on more video content than ever. Take

[1] https://www.merriam-webster.com/dictionary/video -- last verified: October 2017

[2] https://en.wikipedia.org/wiki/Video -- last verified: October 2017

[3] http://www.businessdictionary.com/definition/video.html -- last verified: October 2017

YouTube[4] as an example: over a billion hours of video content is watched daily. In a white paper on global internet trends, CISCO estimates that video traffic will account for 82% of all internet traffic by 2021, up from 73% in 2016 (CISCO, 2017). The reasons for this are obvious. High speed internet has made access to high quality video very convenient and new devices have lowered the barriers to publishing video content (Shen and Cheng, 2010; Halvey *et al.*, 2014; Schoeffmann and Hudelist, 2015). However, the ease of availability is not the only reason for the increasing reliance on video content.

Video is one of the most versatile forms of content in terms of multimodality (Sorin, Petan and Vasiu, 2014). Multimodality is video content's greatest strength. The phrase multimodality refers to video's composition as a set of features, namely: the moving video track, the audio track and other derived features, such as a transcription of spoken words. Together these modalities provide an effective means of communicating information. The content value of these modalities as a whole, far exceeds their separate values.

Richness, both in terms of modalities and the amount of available video content, presents a challenge. Firstly, in terms of volume, there is an unprecedented amount of video uploaded every minute. YouTube's CEO told Fortune that 400 hours of video was being uploaded every minute on the platform in 2014[5]. It is safe to say that it is very difficult, if not impossible, for users to view every piece of video which could be useful or even important for them (Hong *et al.*, 2011). Due to the high volume of video content available, it is becoming increasingly difficult for users to get to the relevant content with respect to the context or immediate search need. A recent study by Ericson reports that an average American spends more than a year, over their lifetime, looking for something to watch on TV (Ericsson, 2016).

However, finding the right video among many is just part of the problem. As videos vary in length (up to several hours long), it is possible that a viewer need not consume the whole video, particularly if it is several hours long. It is possible that only certain parts of a video are of importance or interest to the user. So it is not only important to find the relevant video, but also to verify if the whole video is actually important to the viewer, or only a portion (Masneri

[4] https://techcrunch.com/2017/02/28/people-now-watch-1-billion-hours-of-youtube-per-day/ -- last verified: October 2017

[5] http://fortune.com/2014/10/07/youtube-ceo-wojcicki-youtube-today-is-like-google-ten-years-ago/ -- last verified: October 2017

and Schreer, 2014). To put it a different way, it is desirable to find not only relevant videos, but also the relevant portions of a relevant video. Waitelonis & Sack observed that relevance is a highly subjective sentiment of the user which is dependent on context and pragmatics (Waitelonis and Sack, 2012).

In addition, there may be other characteristics that are important for effective video exploration. That is, users might wish to explore video content on multiple devices with different form factors and modalities (e.g. a mobile device or a home assistant device without a visual interface).

In essence, the increasing variety and amount of video content, its mass availability and the proliferation of differing 'always-on, always-connected' devices, are creating new scenarios in which a user might consume video content. These new scenarios bring new challenges and opportunities with them. As it will be elaborated in Section 2 that current video exploration approaches, while providing interesting use cases, are limited in fully harnessing the exploration potential of video content. To provide users with an effective exploration experience, a better approach might be to utilize the multimodality of video content in its representation and provide users with:

- The relevant content (the relevant portion of video)
- The right manner or modality (due to device or personal preference)
- The right amount of detail (due to time constraints or personal preference)
- The segments surrounding the segment of interest (to get a better idea of the narrative)

The area of research which deals with this problem is referred to in the literature as *exploratory search*. Which is defined as a complex search task in which the user has to first retrieve some facts which then enables further search queries to solve the overall search problem. Often the user is not sure about his/her search goal and sometimes, he/she is not very familiar with the topic of the search (Marchionini, 2006; Waitelonis and Sack, 2012).

Current techniques approach video exploration by enhancing the video selection capability from a large collection, either by listing search query results based on indexing of multimodal attributes (Waitelonis and Sack, 2012; Matejka, Grossman and Fitzmaurice, 2014) or by listing video recommendations (Tan *et al.*, 2014). However, finding the relevant video among many is just part of the problem as videos can vary in length, potentially running for up to several hours. Many techniques have been proposed to aid users in finding the relevant content within videos, including linked-data based approaches (Waitelonis and Sack, 2012), frame trees (Hudelist, Schoeffmann and Xu, 2015) and semantics-based approaches (Farhadi and Ghaznavi-Ghoushchi, 2013).

While the above-mentioned approaches tend to improve video browsing performance, pointing to the relevant portion in a video still requires the user to watch an unnecessary amount of the video to decide which content is relevant to their needs. As Lei et al. observed, due to its linear nature, it might take longer to evaluate the content of a video rather than a textual document (Lei *et al.*, 2015). Therefore, it is reasonable to say that identifying the relevant content within video is still a cumbersome process.

To solve that problem, researchers have proposed different techniques, e.g. video navigation (Schoeffmann, Taschwer and Boeszoermenyi, 2010), hypervideos (Mujacic *et al.*, 2012a) or video summarization (Evangelopoulos *et al.*, 2013). However, current research is mainly focused on one or more of the following:

- Creating a new video by adjusting the source video, i.e. new artefact from another artefact
- Recommending one video over another.
- Providing links to navigate to different parts of a video.

Researchers have long identified the importance of user control in the process of video search (Cobârzan *et al.*, 2017a). However, the focus has been on creating optimal user interfaces of a predominantly visual character. While these systems do add value to the exploration process, they are quite limited in terms of usage flexibility. Customizing the interactivity in these systems usually produces mixed result as the way a user interacts with video content is highly dependent on the context of the task (Craig and Friehs, 2013; Merkt and Schwan, 2014; Ganier and de Vries, 2016).

It is the contention of this thesis that there is a need to look at video content differently. Video can be viewed as a diverse multimodal content source by breaking the tight bond between the different modalities. By the tight bond it is meant that, in video content, different modalities i.e. visual, textual and audio content are presented in a linear stream. Breaking this linearity between modalities can open up new opportunities for video content exploration.

1.2. Research Question

The main question this thesis aims to evaluate is the following:

To what extent can multimodal features extracted from a video be utilized to transform video content in order to enhance a user's video exploration experience in navigation, synopsis and engagement with the exploration process?

Video exploration is a complex task which is defined as a combination of tasks such as video retrieval (searching for videos in a collection), video navigation (search within a single video) and video summarization (synopsis of a video) (Schoeffmann and Hudelist, 2015; Cobârzan *et al.*, 2017a). The presented thesis is focused on the navigation and synopsis parts of video exploration.

By exploration experience it is meant the efficiency, effectiveness and user engagement with the video exploration approach. O'Brien & Toms observed that due to the complex nature of exploratory search, traditional measures of information retrieval such as efficiency and effectiveness are not adequate to evaluate an exploratory search approach (O'Brien and Toms, 2013). They consider engagement a key quality of the process. Interactive video exploration research (Section 2.2) stipulates the importance of flexibility in an exploration approach. Therefore, it is the contention of this thesis that an approach to explore video content should not just be efficient and effective; it should also be engaging and flexibly interactive.

This thesis describes an approach to represent the content of video to users in order to enhance its exploration potential. The approach works in two phases. Firstly, state of the art tools are used to extract features along with timestamps from different modalities of the video stream. Then, upon receiving a content request, a representation engine utilizes a template collection to represent the content of a video in an appropriate configuration.

A configuration determines the presence and granularity of certain features in order to compose a representation of the source video. A video can have multiple multimodal representations. Therefore, a representation may only have a subset of all available multimodal features.

As noted in experiment 2 (section 5), different users tend to consume different portions and different feature sets while consuming the content of the video. Therefore, it is unlikely that there is one perfect (one size fits all) representation for a video.

1.3. Research Objectives

The following are the objectives of this research:

1. Review of the state of the art in video exploration to understand what approaches have been utilised to enhance the user experience for exploring video content to date.
    - Interactive video explorations
    - Interactive video exploration within a video
    - Non-linear video exploration and Hypervideos
    - Video summarization
2. Examination of techniques to break the tight bond between parallel modalities and extract features from them.
3. Present the extracted multimodal features to users in an interactive manner so they can explore the content of a video.
    - Learn usage patterns.
4. Design and develop a template driven representation engine approach based on the usage patterns that automatically generate multimedia interactive document from video content for users.
5. Evaluate the representation engine performance with respect to content exploration tasks.

1.4. Thesis Contribution

This research proposes a new approach for multimodal video representation based on a template driven representation engine capable of transforming video to enhance the user's exploration experience with video content.

The primary contribution of this thesis is the proposed approach named *RAAVE. RAAVE* helps users explore video content effectively. In short, it can be described by the following:

1.4.1. Transforming video content to create new ways to explore and interact with it

According to research (Lei *et al.*, 2015), it is slower to get the essence of a video than a textual document because of its linear nature. Multimodal information is tightly bound within the video. By tightly bound, it is meant the continuous linear stream of parallel modalities (the moving video track, the audio track and other derived features, such as a transcription of spoken words) is intended to be played sequentially. While it is an effective means of communicating information, in certain situations, watching a video to get the desired content is a bit cumbersome, compared to skimming through a text document or a webpage with text and images etc.

This thesis proposes an approach named RAAVE which transforms the video by breaking the tight bond between the parallel modalities opening up new opportunities for exploring a video.

Transforming video content by tearing it apart and showing it as a multimedia document is more flexible to consume because the viewer is no longer limited to watching the video, but can consume the content in a modality which might be more suitable for the given content, or the user may prefer it due to personal choice e.g. fast reader or prefer visuals etc.

1.4.2. Finding the relevant video and the relevant portion(s) within a long video

Given a list of videos resulting from a search query, e.g. a simple query on YouTube or TED.com, a user can end up with dozens of videos and some of those videos could be hours long.

RAAVE is designed to help users in making the following decisions:

1) Is this video of interest to me?
2) If yes, then what portions should I consume and what portions of this long video can I skip without missing out on something important.

1.5. Research Methodology

A detailed analysis of the state of the art was conducted to see different approaches for exploration of video content (section 2). Even though there have been many diverse approaches covering different aspects of the problem. One thing common to all of them was the process of extracting features from the video. Therefore, the first step in this research was to conduct a detailed analysis of multimodal feature extraction from video content. Researchers have used many different toolsets to extract a variety of multimodal features from video content. However, the choice of the toolset used, and features extracted are highly dependent on the genre of video and the application.

Due to the diverse nature of the task, this research focused only on informational and infotainment videos. The experimentation is performed on TED[6] presentation videos.

[6] https://www.ted.com/

Multimodal features were extracted from the different modalities of TED videos by utilizing different tools. After identifying multimodal features and toolsets, those features were analysed for their correlation with user engagement criterion. Feature extraction and engagement assessment was the first phase.

In the second phase, the extracted multimodal features were presented to the user in an interactive manner to support exploration. A user study was conducted by utilizing a novel system prototype developed to learn the usage patterns of participants. From the study, certain usage patterns were learned. These usage patterns were utilized to design a template driven representation engine.

In the third phase, the representation engine was developed and evaluated with user trials. Users performed exploratory search tasks by utilizing the representation engine. Users performed two types of tasks:

- Finding a particular piece of information within video content.
- Evaluating the essence of a video in a limited amount of time and writing a synopsis.

1.6. Thesis Overview

The current chapter (chapter 1) explains the motivations for this work and has provided an overview of the thesis.

In chapter 2, the state of the art of video exploration is reviewed. It describes the nuance of different approaches proposed by researchers to enhance the exploration experience of users with video content. It also contains a discussion on the limitations of current approaches and the need for a new approach to video exploration.

Chapter 3 describes the design of the proposed approach. The approach is proposed as a framework of a template driven representation engine. The chapter also describes the design of the representation engine.

Chapter 4 describes the first phase of the research i.e. the feature extraction and toolset identification. The extracted multimodal features along with their timestamp information are utilized in developing a prototype to represent the content of video as interactive webpages to users.

Chapter 5 discusses the design of the prototype and the results of the user study. From the user study, some usage patterns were learned. Based on the user study and learned usage patterns a template driven representation engine was designed. Chapter 3 describes the design of the engine.

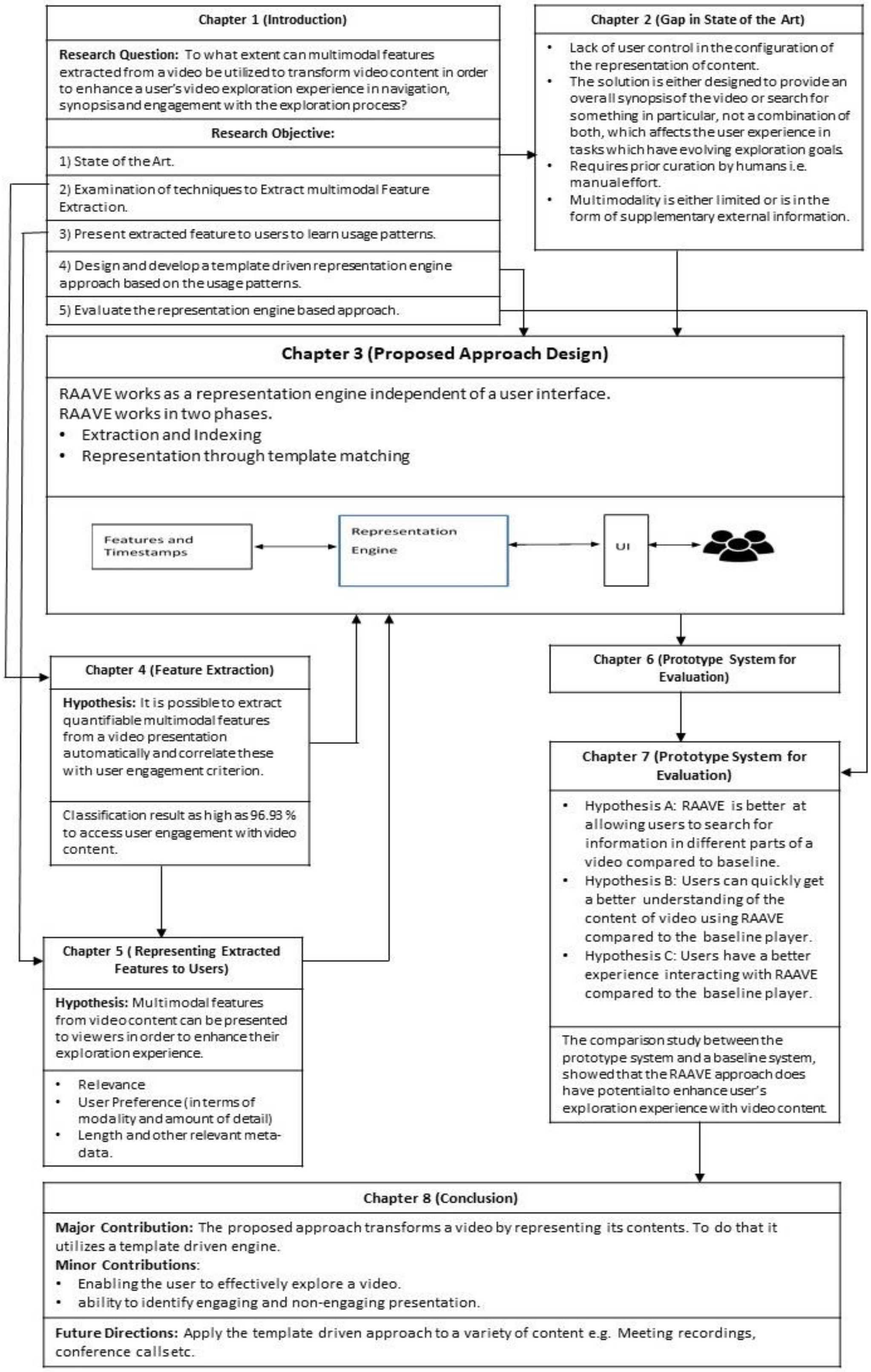
Chapter 6 details the developed prototype based on that design.

Once the prototype of the representation engine was developed, user studies were conducted to assess the user experience in terms of exploration in video content while utilizing the proposed representation engine. Chapter 7 discusses the evaluation and the results.

Chapter 8 concludes the thesis and contains details about future work and some applications of the proposed approach.

Figure 1 shows a visual map of the chapters of the thesis.

Visual Map of the Thesis Chapters



Chapter 1 (Introduction)

Research Question: To what extent can multimodal features extracted from a video be utilized to transform video content in order to enhance a user's video exploration experience in navigation, synopsis and engagement with the exploration process?

Research Objective:

1) State of the Art.

2) Examination of techniques to Extract multimodal Feature Extraction.

3) Present extracted feature to users to learn usage patterns.

4) Design and develop a template driven representation engine approach based on the usage patterns.

5) Evaluate the representation engine based approach.

Chapter 2 (Gap in State of the Art)

- Lack of user control in the configuration of the representation of content.
- The solution is either designed to provide an overall synopsis of the video or search for something in particular, not a combination of both, which affects the user experience in tasks which have evolving exploration goals.
- Requires prior curation by humans i.e. manual effort.
- Multimodality is either limited or is in the form of supplementary external information.

Chapter 3 (Proposed Approach Design)

RAAVE works as a representation engine independent of a user interface.
RAAVE works in two phases.

- Extraction and Indexing
- Representation through template matching

Chapter 4 (Feature Extraction)

Hypothesis: It is possible to extract quantifiable multimodal features from a video presentation automatically and correlate these with user engagement criterion.

Classification result as high as 96.93 % to access user engagement with video content.

Chapter 5 (Representing Extracted Features to Users)

Hypothesis: Multimodal features from video content can be presented to viewers in order to enhance their exploration experience.

- Relevance
- User Preference (in terms of modality and amount of detail)
- Length and other relevant meta-data.

Chapter 6 (Prototype System for Evaluation)

Chapter 7 (Prototype System for Evaluation)

- Hypothesis A: RAAVE is better at allowing users to search for information in different parts of a video compared to baseline.
- Hypothesis B: Users can quickly get a better understanding of the content of video using RAAVE compared to the baseline player.
- Hypothesis C: Users have a better experience interacting with RAAVE compared to the baseline player.

The comparison study between the prototype system and a baseline system, showed that the RAAVE approach does have potential to enhance user's exploration experience with video content.

Chapter 8 (Conclusion)

Major Contribution: The proposed approach transforms a video by representing its contents. To do that it utilizes a template driven engine.
Minor Contributions:

- Enabling the user to effectively explore a video.
- ability to identify engaging and non-engaging presentation.

Future Directions: Apply the template driven approach to a variety of content e.g. Meeting recordings, conference calls etc.

Figure 1: Visual map of the chapters

2. State of the Art

The goal of this research is to enhance the user experience in exploring video content by proposing a new approach for exploring content within video. As described in section 1.1, video exploration is a complex task in which user informational needs are either imprecise or constantly evolving, and the mass availability and omnipresence of video content creates a challenge. Therefore, in the literature, video exploration is seen as a combination of different tasks, such as video retrieval (exploration in a video collection), video navigation (exploration within a single video) and video summarization (quick overview or skim of a video) (Schoeffmann and Hudelist, 2015; Cobârzan *et al.*, 2017a).

This chapter provides a review of video exploration to identify the best practices in the area and the limitations of current approaches. The review is organized as follows.

A brief overview of video exploration is presented. Limitations of retrieval-based approaches and characteristics to enhance user experience are discussed next. After identifying the characteristics, this review then focuses on interactive exploration within a video. Finally, the limitations of current exploration within video approaches are discussed and the gap in the state of the art, in terms of the identified characteristics, is discussed.

2.1. Overview of video exploration approaches

Traditionally, video exploration approaches are built around retrieval engines that use multimodal low-level features, for example visual features (colours, edges, textures etc.), audio features (Fourier transform or pitch etc.), automatic speech recognition (ASR) or optical character recognition (OCR), to find relevant videos within a large collection and present them as a ranked list (Halvey *et al.*, 2014; Munzer *et al.*, 2017; Tsukuda, Masahiro and Goto, 2017). However, as mentioned in section 1, the exponential growth in video content has created challenges to explore this massive amount of content effectively (Zhang and Nunamaker, 2004; Hong *et al.*, 2011; Masneri and Schreer, 2014). It is simply too distracting for the user if they are shown a large list of videos in response to a query (Hong *et al.*, 2011).

Researchers have proposed different techniques to mitigate the problem, for example Waitelonis & Sack proposed a semantic search approach based on linked data to show relevant videos (Waitelonis and Sack, 2012). Dong et al. use morphological analysis and image matching between slides and video frames to annotate video presentations and use a combination of ontologies to show semantically relevant video presentations (Dong, Li and Francisco, 2008). This idea of ontology reasoning is extended by Bertini et al. in their approach, the authors use an ontology reasoning engine and a multi-touch interface to allow users to perform semantic search and organize the results in ontology graphs and a list view (Bertini *et al.*, 2011). The graph based approach is also used by Halvey et al. in which authors use a soft graph based approach to let users group videos based on semantic concepts (Halvey *et al.*, 2014). The graph is later used to show more relevant video results.

One way of solving the problem of too many videos is to choose a video that could be interesting or useful for the user and present the suggestion to him/her. Therefore, creating recommender systems for users based on their viewing habits and commenting patterns also attracts a lot of interest. For example, Brezeale & Cook attempt to predict user movie preferences by clustering movie subtitles, low level visual features and ratings given by users (Brezeale and Cook, 2009). Anwar et al. tried to sort videos into different categories based on their features (Anwar, Salama and Abdelhalim, 2013), while Tan et al. use heterogeneous data from different sources to create a better recommender system based on user video preferences (Tan *et al.*, 2014).

In addition to querying a video collection with text, researchers have proposed approaches which allow users to query a video collection using different modalities. For example, Zhang et al. devise a cross-media retrieval approach to search for video by providing audio samples and vice versa (Zhang, Liu and Ma, 2013). Rafailidis et al. extend this idea and propose a unified framework to allow retrieving information using multimedia queries based on semantic similarity. Their framework searches for semantic similarity based on a weighting scheme trained to match media objects as unified sets of different modalities (image, audio, 3D, video and text). Mauceri et al. allow users to retrieve relevant videos through dynamically creating sample key frames by capturing their motion and use the key-frames as templates to search for video clips with similar motions (Mauceri *et al.*, 2015). To make the retrieval process of motion videos efficient Qin et al. propose a hashing scheme to efficiently handle the high dimensionality of video data (Qin *et al.*, 2017).

Ramezani and Yaghmaee, combine motion based multimedia queries and a recommender system to efficiently retrieve relevant videos from a large collection (Ramezani and Yaghmaee, 2016). Similarly, Choi et al. propose a video recommendation approach based on capturing and analyzing viewer’s facial expressions (Choi *et al.*, 2016).

Researchers have also experimented with making video exploration a collaborative activity. For example, Tsukuda et al. use data mining methods on time synchronized user comments to allow users to query a video collection based on viewer’s emotions (Tsukuda, Masahiro and Goto, 2017). The idea of enhancing exploration of a video collection by collaborative annotation is also used by Nicolaescu & Siddiqui who propose a widget-based system which shows semantic annotations like place, object, agent and event on maps and other interface elements to allow the users to explore the video collection (Nicolaescu and Siddiqui, 2017). Schoeffmann et al. make video exploration a colloborative task by allowing multiple users to search a video collection by displaying a heatmap of user interest (segments of videos visited by other users) along with video search results (Schoeffmann *et al.*, 2017).

Munzer et al. extend the idea of the collaborative exploration system proposed above (Schoeffmann *et al.*, 2017) by allowing multimodal queries and allowing expert users to help non-experts (Munzer *et al.*, 2017).

2.1.1. Limitations of video exploration approaches and focus of the thesis

It can be seen in the above overview of the state of the art of video exploration that researchers have proposed many interesting approaches to explore video collections. However, finding the right videos among a collection is just part of the problem given that videos can vary in length. It is tedious and cumbersome for a user to sequence through a long video to get the desired content (Hong *et al.*, 2011; Khan and AlSalem, 2012).

It is the contention of this thesis that the video exploration experience in video content can be enhanced by providing the user with the ability to explore the content within a video. The remainder of the review focuses on the state of the art approaches in exploring the content within a video.

Before the review of exploration within video approaches itself. It is worthwhile to first review the literature to identify the characteristics of an exploration approach that may provide an enhanced experience for the user.

2.1.2. Characteristics for an enhanced user experience in exploration within a video

Exploration within a video is a complex task in which a user’s search needs can evolve quickly (Waitelonis and Sack, 2012). While content-based retrieval approaches can achieve good results in terms of searching through the content of a video, they suffer from various deficiencies such as: failing to evolve to changing human needs, the semantic and usability gap between system representation of content and user understanding of the exploration task (Hong *et al.*, 2011;

Munzer *et al.*, 2017). To solve these problems researchers have long identified the need for user interactivity in the video content exploration process (Mackay and Davenport, 1989; Hauptmann *et al.*, 2006; Sorin, Petan and Vasiu, 2014; Girgensohn *et al.*, 2016a; Cobârzan *et al.*, 2017b; Munzer *et al.*, 2017).

While researchers have proposed many techniques to interactively explore the content of a video (section 2.2) most require the user to watch a fair amount of the video to decide which content is relevant to their needs. Due to its linear nature (a continuous linear stream of consecutive images with or without a parallel audio component) it might take longer to evaluate the content of a video rather than a textual or a hypertext document (Zhang and Nunamaker, 2004; Lei *et al.*, 2015).

In addition to the ability to non-linearly explore content of a video, users also require different amounts of detail depending on the context (Shipman, Girgensohn and Wilcox, 2008; Girgensohn *et al.*, 2016a; Gravier *et al.*, 2016).

In the discipline of digital humanities, researchers refer to exploration within a single document or in corpus of documents with concepts like distant and close reading (Drouin, 2014; Jänicke *et al.*, 2015; Jin, 2017; Mehta *et al.*, 2017).

Close reading is defined as a set of practices that involves analyses of spatial and temporal interactions between the syntactic, semantic, structural, rhetorical features within a document to uncover layers of meaning that lead to deep comprehension (Boyles, 2013; Mehta *et al.*, 2017).

While distant reading is defined as a set of approaches to text analysis, leveraging computational tools for detecting patterns in a corpus as a whole. It aims is to generate an abstract view by shifting from observing textual content to visualizing global features of a single or of multiple text(s) (Jänicke *et al.*, 2015).

While traditionally close reading and distant reading have been different and often opposing school of thoughts (Bode, 2017). However, researchers in digital humanities now stress on the importance of a hybrid approach to exploration i.e. distant and close reading techniques complementing each other in order to provide an enhanced user experience (Jin, 2017). Referred in literature as detail on demand (Shneiderman, 2003) or hierarchal representation (Koch *et al.*, 2014), researchers have proposed many approaches to represent textual content which combines distant and close reading practices to enhance user's exploration experience (Shneiderman, 2003; Koch *et al.*, 2014; Jänicke *et al.*, 2015; Mehta *et al.*, 2017).

Apart from hierarchal or detail on demand exploration needs, video content is multimodal by nature. Multimodality in representation of content enhances the user exploration experience (Calumby *et al.*, 2017). However while multimodality in representation is good, it is important to let the user choose the modality to explore content based on the task and their personal preference for improved task performance (Craig and Friehs, 2013; Merkt and Schwan, 2014; Ganier and de Vries, 2016).

In short, to enhance a user's exploration experience with video content, an exploration approach should utilize content-based retrieval and user interaction in a complementary way. It should allow users to interact with video content at their own pace, with their own strategies, navigating the content autonomously (Munzer *et al.*, 2017; Sauli, Cattaneo and van der Meij, 2017).

The above discussion can be summarized in the form of the following characteristics:

1) Interactivity.
2) Ability to navigate the content of a video in a non-linear manner.
3) Abilities to search for piece of information and a quick overview.
4) Different level of details in the representation.
5) Let user choose the modalities to consume the content of video.

It is the contention of this thesis that by possessing these characteristics, an exploration approach would have the potential to enhance the user's exploration experience.

2.2. Review of Interactive exploration within video

Searching some piece of information within video assets is still text based as far as commercial offerings such as YouTube[7], etc. are concerned. However, researchers have worked a great deal on multimodal methods to extract information from videos. Brachmann & Malaka use a video player with a scroller showing navigational blocks and a magnification slider (Brachmann and Malaka, 2009). Increasing the magnification expands the scroller for the particular block and displays additional information such as speech transcript.

Schoeffmann et al. facilitate navigation and searching within a video through visualizing low-level features and frame surrogates, i.e. frames which are visually similar to a sample frame (Schoeffmann, Taschwer and Boeszoermenyi, 2010). While the approach provides a quick

[7] www.youtube.com

overview of the video and lets the user find visually similar portions of the video the representation focused only on visual information. Adcock et al. use Optical Character Recognition (OCR) techniques to detect slides and slide changes in video lectures, extract useful keyframes and represent them to users as links during video play (Adcock *et al.*, 2010). Khan & AlSalem use a similar approach, but augment their system with a Natural Language Understanding engine (Khan and AlSalem, 2012). Cooper et al. proposed a collaborative system which allows two users to search for relevant shots (Cooper *et al.*, 2011). They do so by providing an additional shared display to continuously show updated information to both users, in addition to the individual exploration interface.

In Haesen et al. authors use facial recognition along with name tags to find footage of certain people in a video collection (Haesen *et al.*, 2011). The videos are segmented and annotated by measuring lexical cohesion of adjacent text passages and examining the repetition of named entities. The user can navigate a video from a list by either using: a clock visualization; or a video timeline, with indication of relevant portions and a special video player. Monserrat et al. use object detection to track changes in blackboard style videos and display the frames in a spatial layout to allow the user to navigate the video by directly interacting with it (Monserrat *et al.*, 2013). The idea of direct interaction is extended by Denoue et al. in their system, the authors use image processing and OCR to magnify the frames in video and directly select text from it in a manner similar to a text document (Denoue *et al.*, 2013).

Moumtzidou et al. apply a modular approach for interactive video exploration (Moumtzidou, Avgerinakis and Apostolidis, 2014). They employ different modules to segment and cluster videos based on predefined high-level concepts and represent the shots of video content in a grid like interface. Matejka et al. propose a faceted search approach to search relevant sections of a baseball video based on external metadata attributes (Matejka, Grossman and Fitzmaurice, 2014). To make interaction with video content flexible, Hudelist et al. propose a hierarchal representation of key frames (Hudelist, Schoeffmann and Xu, 2015). They use a navigation tree structure to show key frames, differing in levels of detail, at each branch.

Yadav et al. use multimodal analysis and a customized time aware word-cloud to enhance video navigation experience in lecture videos (Yadav *et al.*, 2015). In their system they use visual analysis techniques to extract key frames with maximum written content. They also use speech-to-text conversion and acoustic analysis techniques to detect stressed words. They used this information to create a customized word cloud which shows keywords in a time aware manner and use colour coding to represent emphasis. This allows users to navigate the video based on topics of interest and see more detail in a section of interest by going through its key frames. Similarly, Balasubramanian et al. use multimodal features such as transcript and slide images to extract keywords from video lectures and present them as links to different parts of a video (Balasubramanian, Doraisamy and Kanakarajan, 2016).

Galuscakova et al. use 60 second segments of video to represent relevant segments and timestamps in search results to allow the user to read the transcript without watching the video and to jump to the relevant segment (Galuscakova, Saleh and Pecina, 2016).

In order to perform different types of search tasks, Schoeffmann et al. propose a system which allows users to choose from different visualizations of keyframes and show metadata in the form of a heatmap to allow collaborative search of relevant segments (Schoeffmann *et al.*, 2017). The idea of enhancing exploration in video collaborative annotation is also used by Nicolaescu & Siddiqui, in which authors propose a widget based system which shows semantic annotations like place, object, agent and event on maps and other interface elements to allow the user to navigate to the segment of interest in a video (Nicolaescu and Siddiqui, 2017).

Though search within video approaches do help the user find the right portion of a video, it is still cumbersome to explore content in video. It is because of its linear nature it takes longer to get the essence of a video compared to a textual document (Lei *et al.*, 2015). To solve this, researchers have proposed nonlinear methods for exploring video content.

2.2.1. Nonlinear video exploration

Different systems have been proposed which help users explore time-based media (Luz and Masoodian, 2004), and video more specifically, in a modular and non-linear manner. Barthel et al. propose a collaborative approach that enables different users to create a video which provides alternative paths to navigate the content to learn about a topic (Barthel, Ainsworth and Sharples, 2013). Merkt et al. provide a table of content style links to navigate to different sections of a video (Merkt *et al.*, 2011). This idea is extended by Pavel et al. in their study, authors use a chapter/section structure to provide a textual summary and a thumbnail image to present video segments (sections of video) to a user (Pavel *et al.*, 2014). Similarly, Meixner & Gold design and evaluate an approach to create a table of content structure to non-linearly navigate a video for smart-phones and tablet devices (Meixner and Gold, 2016).

A widely used approach to enable users to non-linearly explore video content is creating hypervideos.

2.2.1.1. Hypervideos

Hypervideos are based on the same notion as hypertext, i.e. hypertext ideas applied on a video (Sauli, Cattaneo and van der Meij, 2017) with earliest method for video branching proposed as early as 1965 (Nelson, 1965). Boissiere proposes a system to identify topic changes in news broadcasts by searching for special characters like “>>”, placed by transcribers in news transcripts to segment the video and provide hyperlinks to the identified segment (Boissiere, 1998). This idea is enhanced by Finke & Balfanz, in which authors propose a modular architecture for a hypervideo system for an interactive TV portal that consists of an annotation engine, hotspot identification, a metadata format and presentation engine (Finke and Balfanz, 2004). Stahl et al. apply the idea of hotspots and link nodes in educational settings and extend it with the ability to link additional material such as external web pages (Stahl, Finke and Zahn, 2006). The idea of supplementary material is used by Hoffmann & Herczeg in their study, the authors use the hypervideo principle to create a personalized and interactive storytelling system which consist of a customized video player (Hoffmann and Herczeg, 2006). Aubert et al. extends the idea of storytelling by hypervideo with the use of structured metadata schemas such as RDF annotations (Aubert *et al.*, 2008), similarly RDF annotation combined with automatic entity extraction is proposed by Hildebrand & Hardman to generate annotations for interactive TV programs (Hildebrand and Hardman, 2013).

Interactivity is an important aspect of hypervideos. Leggett & Bilda experimented with alternative designs to allow users to navigate a hypervideo by reference images or following a line on a map or grid etc. (Leggett and Bilda, 2008).

Shipman et al. devised an approach to automatically create navigation links for hypervideo by proposing a hierarchical summary generation method to provide detail on demand, video content browsing (Shipman, Girgensohn and Wilcox, 2008). In order to generate hierarchical summaries, authors used low level multimodal features such as: colour histograms and closed captions, clustering algorithms and heuristics dependent on the genre of videos. These features were used to segment video clips, determine the number of levels for the hierarchy and generate hyperlinks for navigation. The authors also designed a custom interface to search the collection and a specialized video player to browse the content. Tiellet et al. use the idea of detail on demand in an educational setting by offering a multimedia presentation of content (Tiellet *et al.*, 2010). In their study, authors use a hypervideo system which offer links to more detailed information to students in the form of high definition images, supplementary text and annotation to learn surgical procedures.

Sadallah et al. observed that prior hypervideo approaches were based on ad-hoc specifications and hypermedia standards such as SMIL (Bulterman and Rutledge, 2009) and NCL (Neto and Soares, 2009) are not well-suited for hypervideos (Sadallah, Aubert and Prié, 2012). Authors propose an annotation driven and component-based model for hypervideo, inspired from other multimedia standards but more suited for hypervideos. While Mujacic et al. proposed a different approach, they propose to use an hypervideo generation approach based on the SMIL (Bulterman and Rutledge, 2009) specification (Mujacic *et al.*, 2012a). In order to simplify the process of authoring hypervideos, Meixner et al. propose an authoring system to allow non-technical users to create XML based annotations for hypervideo systems (Meixner *et al.*, 2014).

Girgensohn et al. use a hypervideo system and collaborative annotation to offer dynamically generated links to consume the content of meeting recordings asynchronously (Girgensohn *et al.*, 2016b).

While the above-mentioned approaches create hypervideo using video content, Leiva & Vivó took a different approach. In their study authors use web page interaction logs to synthesize an interactive hypervideo to allow a user to visualize webpage usage (Leiva and Vivó, 2013). Similarly, Petan et al. propose a similar approach to synthesize interactive hypervideos for corporate training scenarios (Petan, Petan and Vasiu, 2014).

Recent survey papers (Meixner, 2017; Sauli, Cattaneo and van der Meij, 2017) define the following as the primary aspects of all hypervideo based approaches:

- An authoring environment for annotations and setting up navigation paths.
- A meta-data structure for annotated data and navigation links.
- A specialized environment including but not limited to a customized video player for consuming the hypervideo.

Both Meixner and Sauli et al. consider the complexity of hypervideo systems, both in terms of production (authoring systems) and in consumption environments, to be an issue which is affecting the value of such systems in exploration tasks (Meixner, 2017; Sauli, Cattaneo and van der Meij, 2017).

While hypervideos give more flexibility to the viewer in consuming the content, the flexibility is still limited to the extent to which the author can anticipate it. The viewer cannot go beyond that and while they do provide means to consume information in a multimodal manner. The multimodality comes from additional artefacts embedded by the curators instead of utilizing the potential of video as a multimodal content source.

2.2.2. Video Summarization

To allow users to get the essence of a video in a shorter time, researchers have proposed many approaches which are referred to in literature as video summarization.

Video summarization is defined as a technique that facilitates video content consumption by extracting the essential information of a video to produce a compact version (Guan et al. 2014).

It would not be farfetched to say that in video summarization, importance is usually attributed to visual features. For example, Benini et al. propose an approach to build a video summary or video skim by Logical Story Units (LSU) (Benini, Migliorati and Leonardi, 2010a). They create LSUs with salient features such as motion intensity in Mpeg I-frames and P-frames and face detection in frames trained over Hidden Markov Models (HMM). Users are provided with a video skim (highlights of the video) to decide if they would like to see the full video or not. De-Avila et al. create a video summary by extracting colour features from frames trained by unsupervised clustering, via the k-means clustering algorithm, and represent the key frames to users (de Avila *et al.*, 2011), while Almeida et al. use colour histograms in I-frames[8] of MPEG encoding and a noise filtering algorithm to generate a summary of videos (Almeida, Leite and Torres, 2013). Filtered I-frames are then used to generate video skims for users. Zhang et al. try to create a multi video summary of user generated videos based on aesthetic-guided criterion and generate a single video skim for users (Zhang, Zhang and Zimmermann, 2015). Belo et al. extends the idea of clustering key-frames by proposing a graph based hierarchal approach to extract key frames from video footage (Belo *et al.*, 2016).

However, multimodal features are also getting considerable attention due to the added value they bring. For example, Chen et al. use both visual and audio features to propose a hybrid approach combining content truncation and adaptive fast forwarding to offer users a summary of a video by allowing a fast forwardable video skim (Chen, Vleeschouwer and Cavallaro, 2014). Kim et al. use low level multimodal features and a fusion algorithm to create clusters that are then utilized to create video summaries (Kim, Frigui and Fadeev, 2008). Wang et al. utilized multimodal features to develop an approach to segment program boundaries by getting program-oriented informative images (POIM) (Wang *et al.*, 2008). They then used these POIMs

[8] I-frame is an abbreviation for Intra-frame, because they can be decoded independently of any other frames. In MPEG encoding, I-frames are used as key-frames in conjunction with P-frames (Predicted frames) to compress the size of the video.

as a basis to get keyframes and representative text to offer as visual and textual summaries of the segmented programs. Hosseini & Eftekhari-Moghadam use multimodal features and a fuzzy logic base rule set to extract highlights (skim) of soccer games (Hosseini and Eftekhari-Moghadam, 2013). A comprehensive multimodal feature extraction can be seen in Evangelopoulos et al. in which authors take advantage of all three visual, audio and linguistic modalities as well as different data fusion techniques to create video skims with different ratios (Evangelopoulos *et al.*, 2013).

It can be seen in the above review that the goal of video summarization is to create a new video artefact from the source which is shorter in length. In video summarization, the applied methodologies choose the content to be included in the output artefact autonomously. The user is not part of the decision-making process and is only shown the final output. The output is represented to users either as static summaries (key frames) or dynamic summaries (video skims) (Guan et al. 2014) . The ability to search is generally not provided and user interactivity is limited to: links, in the case of static summaries; or ability to choose playback speeds, in the case of dynamic video skims.

Video summarization approaches tend to focus on just the summarization or synopsis aspect of video exploration and are expected to be used in conjunction with retrieval-based approaches. For example Hong et al. and Munzer et al. shows the summary of retrieved video as a filmstrip of key frames i.e. a static summary (Hong *et al.*, 2011; Munzer *et al.*, 2017). However, the issues of lack of user control and choice of modalities in the representation remain and limit the exploration potential of the retrieved video.

2.2.3. Gap in exploration within video approaches

In terms of the five characteristics of video exploration approaches (section 2.1.2), it can be seen from the review above that there are many approaches which let users explore video content interactively. Approaches like hypervideos help users navigate the content of video in a nonlinear manner. And with approaches like video summarization techniques, there has been a lot of focus on searching within video and quick overviews of a video. However, there has not been much focus on the multimodality of videos while representing its content to the users and giving users the ability to control the level of detail of the content presented i.e. there is a gap in state of the art approaches with regard to points 4 and 5 of section 2.1.2.

The remainder of this state of the art review will focus on the approaches which, to some extent have tried to address this gap. Table 1 lists those approaches.

An approach is included in the focused list if it followed the following criteria.

- Allows different level of details in the representation.
- Enables the user to choose the modalities to consume the content of video

2.2.3.1. Brachmann & Malaka

Brachmann & Malaka use an interactive video player with a scroller showing navigational blocks and a magnification slider (Brachmann and Malaka, 2009). Increasing the magnification expands the scroller for the particular block and displays additional information, such as speech transcript (Figure 2). The magnification levels are determined manually in the presented version. While their system is interactive and provides users with the ability to see more detailed information in the form of speech segments, users are not provided with the ability to search for information. Users need to manually jump to particular blocks to search for information or use a magnifier bar to display the customized seek bar to get detailed information. The multimodal representation is limited to manually annotated speech transcripts of a certain length and thumbnail images of magnified video block.

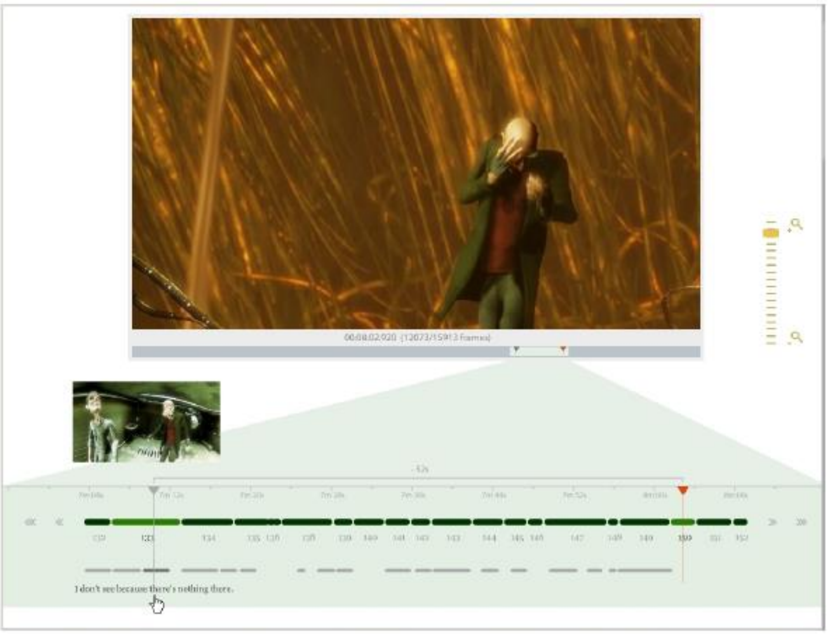

00:08:02.920 (12073/15913 frames)
I don't see because there's nothing there.

Figure 2: Customized video player with interactive timeline magnification to show speech and scene information (Brachmann and Malaka, 2009)

2.2.3.2. Yadav et al.

Yadav et al. use multimodal analysis and a customized time aware interactive word-cloud to enhance the video navigation experience in lecture videos (Yadav *et al.*, 2015). In their system, authors use visual analysis techniques to extract key frames with maximum written content. They also used speech to text conversion and acoustic analysis techniques to detect stressed words. They used this information to create a customized word cloud which shows keywords in a time aware manner and use colour coding to represent emphasis (Figure 3). This allows users to navigate the video based on topics of interest and see more detail in a section of interest by going through its key-frames. While the users have the choice of multimodality in the representation by using the customized word cloud or interact with the slide show of key-frames, they still need to navigate to the portion of video by going through a series of word clouds.

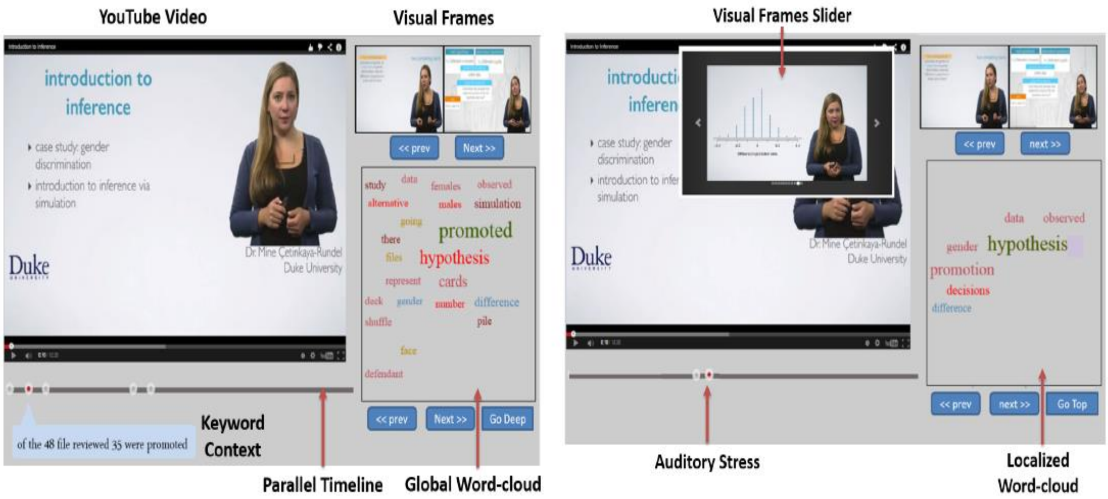
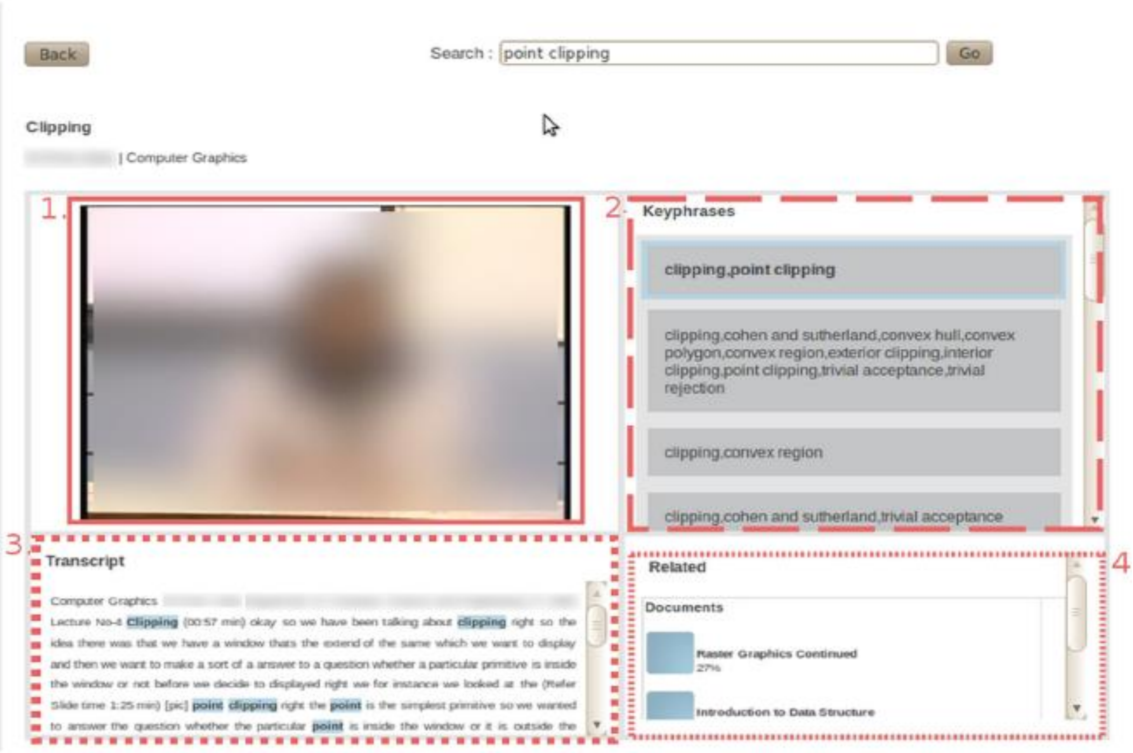


Figure 3:time-aligned word cloud and keyframes for navigation within a video (Yadav *et al.*, 2015).

2.2.3.3. Balasubramanian et al.

Balasubramanian et al. use multimodal features such as transcript and slide images to extract keywords from video lectures and present them as links to different parts of a video (Balasubramanian, Doraisamy and Kanakarajan, 2016). Upon a search request, users are presented with highlighted keywords in a transcript, to navigate to the point, and also an overview of the video, with keywords appearing in the different segments of a video lecture. Figure 4 shows the system. Users have a choice of seeing the keywords or scroll through the whole transcript and non-linearly navigate to certain points in the video.


Back
Search : point clipping
Go
Clipping
| Computer Graphics
1.
2.
Keyphrases
clipping,point clipping
clipping,cohen and sutherland,convex hull,convex polygon,convex region,exterior clipping,interior clipping,point clipping,trivial acceptance,trivial rejection
clipping,convex region
clipping,cohen and sutherland,trivial acceptance
3.
Transcript
4
Related
Documents
Raster Graphics Continued
27%
Introduction to Data Structure

Computer Graphics

Lecture No-4 Clipping (00:57 min) okay so we have been talking about clipping right so the idea there was that we have a window thats the extend of the same which we want to display and then we want to make a sort of a answer to a question whether a particular primitive is inside the window or not before we decide to displayed right we for instance we looked at the (Refer Slide time 1:25 min) [pic] point clipping right the point is the simplest primitive so we wanted to answer the question whether the particular point is inside the window or it is outside the

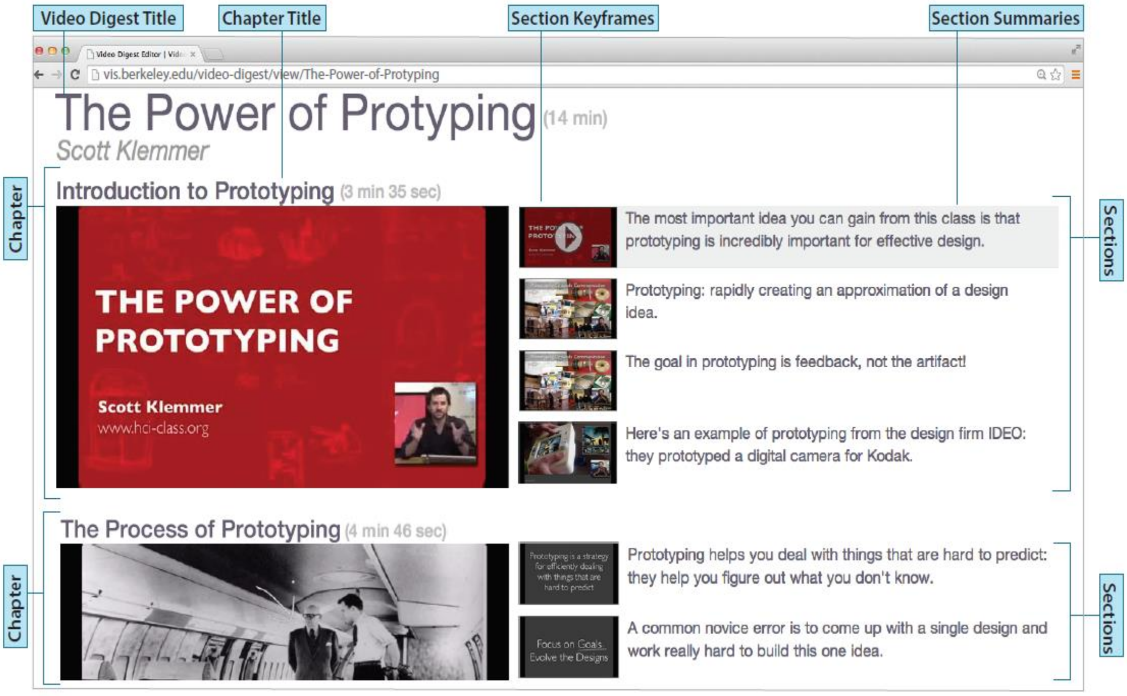
Figure 4: 1. Video player, 2. Keywords in each video segment, 3. Interactive transcript, 4. Related documents. (Balasubramanian, Doraisamy and Kanakarajan, 2016)

2.2.3.4. Galuscakova et al.

Galuscakova et al. use 60 second segments of video to represent relevant segments and timestamps in search results to allow the user to read the transcript without watching the video and to jump to the relevant segment (Galuscakova, Saleh and Pecina, 2016). Interactivity is limited to a customized scrub bar and navigating to the particular segment shows more detailed information in the form of segment transcript. However, multimodality in the representation is limited to either watching the video or reading the transcript of the selected segment.

2.2.3.5. Pavel et al.

Pavel et al. use a chapter/section structure to provide a textual summary and a thumbnail image to present video segments (sections of video) to a user (Pavel *et al.*, 2014). Their system is divided into two interfaces. The first interface is for content curation in which editors create the chapter/section structure. The other interface is designed for end users to consume the video content in a nonlinear manner. Figure 5 shows the end user interface. Users can see a thumbnail image and summary of the section to decide if they want to navigate to that section of the chapter to consume the segment in a more detailed manner.



vis.berkeley.edu/video-digest/view/The-Power-of-Protyping

The Power of Protyping (14 min)

Scott Klemmer

Introduction to Prototyping (3 min 35 sec)

THE POWER OF PROTOTYPING

Scott Klemmer
www.hci-class.org

The most important idea you can gain from this class is that prototyping is incredibly important for effective design.

Prototyping: rapidly creating an approximation of a design idea.

The goal in prototyping is feedback, not the artifact!

Here's an example of prototyping from the design firm IDEO: they prototyped a digital camera for Kodak.

The Process of Prototyping (4 min 46 sec)

Prototyping helps you deal with things that are hard to predict: they help you figure out what you don't know.

A common novice error is to come up with a single design and work really hard to build this one idea.

Figure 5: video is divided into chapters and each chapter is divided into sections. Section is shown with a thumbnail and a text summary clicking on the section plays the video from that particular timestamp (Pavel and Reed, 2014)

2.2.3.6. Meixner & Gold

Meixner & Gold design and evaluate an approach to create a table of content structure to non-linearly navigate a video for smart-phones and tablet devices (Meixner and Gold, 2016). Users can use the curated table of content to consume the video in a nonlinear manner (Figure 6).

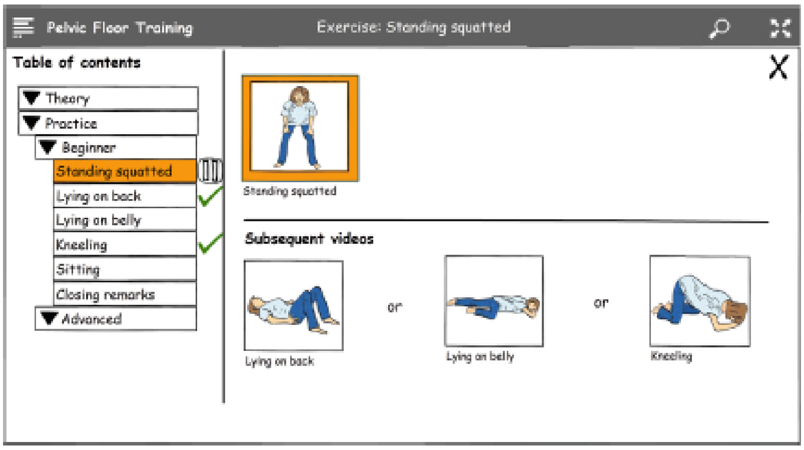
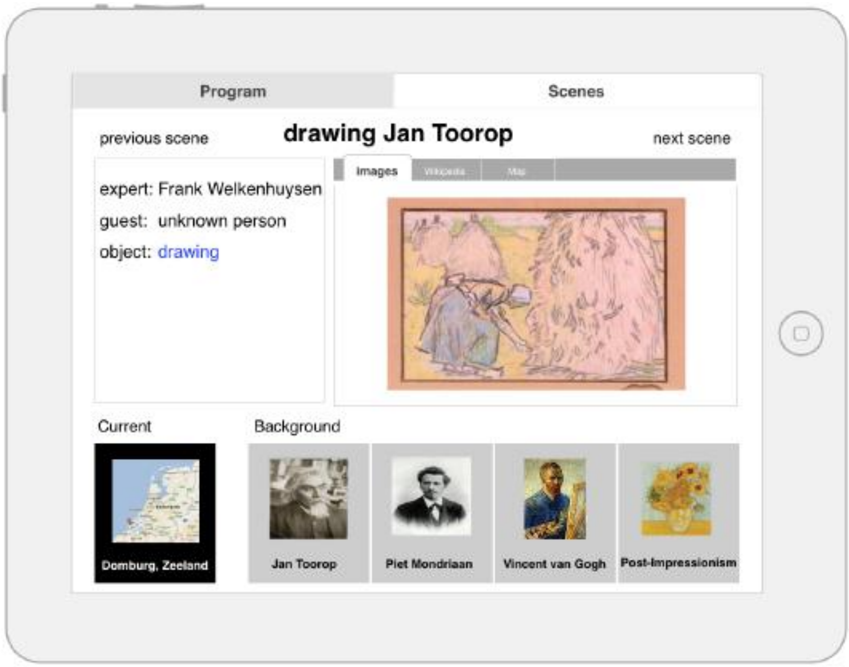


Figure 6: table of contents in the hypervideo player. (Meixner and Gold, 2016).

2.2.3.7. Hildebrand & Hardman

RDF annotation combined with automatic entity extraction is proposed by Hildebrand & Hardman to generate annotations for interactive TV programs (Hildebrand and Hardman, 2013). Users are presented with annotated multimodal information on a second screen with the help of a custom application (Figure 7).



expert: Frank Welkenhuysen
guest: unknown person
object: drawing

Figure 7: Concept design of second screen application for TV programs (Hildebrand and Hardman, 2013).

2.2.3.8. Shipman et al.

Shipman et al. devised an approach to automatically create navigation links for hypervideo by proposing a hierarchical summary generation method to provide detail on demand, video content browsing (Shipman, Girgensohn and Wilcox, 2008). In order to generate hierarchical summaries, authors used low level multimodal features such as: colour histograms and closed captions, clustering algorithms and heuristics dependent on the genre of videos, so as to: segment video clips, determine the number of levels for the hierarchy and generate hyperlinks for navigation (Figure 8 left). The authors also designed a custom interface to search the collection and a specialized video player to browse the content (Figure 8 right).

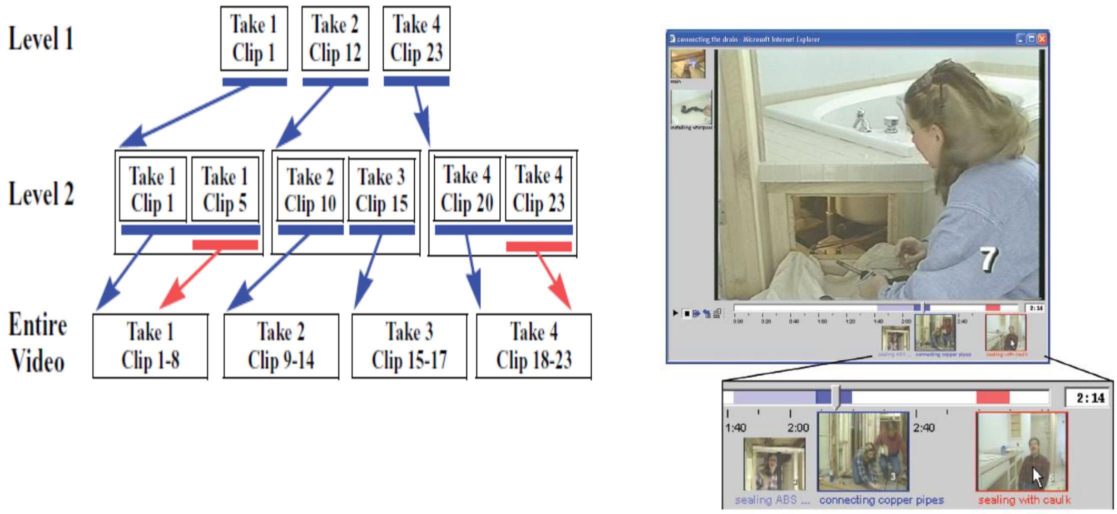
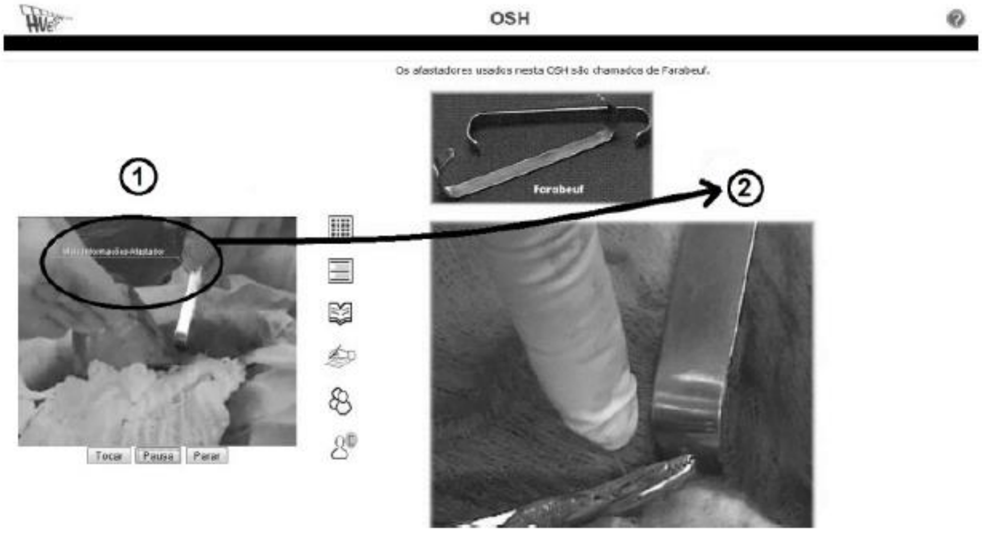


Figure 8: (left) different levels of links for generated video skim. (right) customized hypervideo player. (Shipman, Girgensohn and Wilcox, 2008).

2.2.3.9. Tiellet et al. and Sadallah et al.

Tiellet et al. use the idea of detail on demand in an educational setting by offering a multimedia presentation of content (Tiellet *et al.*, 2010). In their study, authors use a hypervideo system which offers links to more detailed information to students in the form of high definition images, supplementary text and annotation to learn surgical procedures (Figure 9).

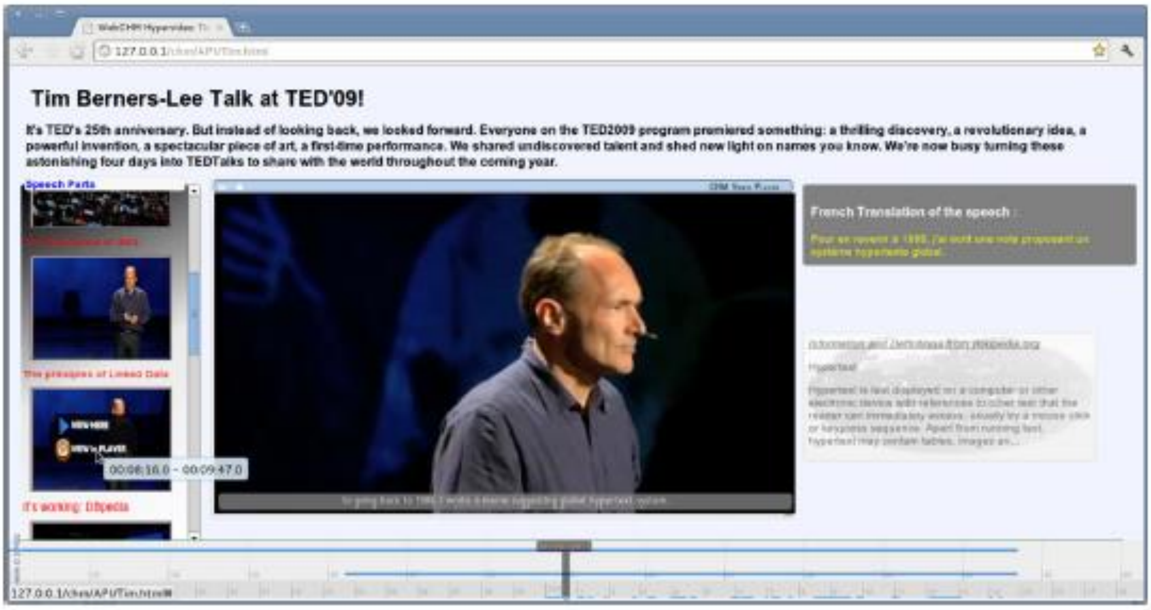


Os afastadores usados nesta OSH são chamados de Farabeuf.

Figure 9: (Left) hypervideo player. 1. hyperlink. (Right) Supplementary Information. 2. target media (Tiellet *et al.*, 2010).

Similarly Sadallah et al. propose an annotation driven and component based model for hypervideo inspired from other multimedia standards, but more suited for hypervideos

(Sadallah, Aubert and Prié, 2012). Figure 10 shows the representation of a curated hypervideo to users. On the left is the graphical representation of the video segments. In the centre is the video player while on the right is the supplementary information, such as the translation to other languages and useful information from external sources e.g. Wikipedia.



Tim Berners-Lee Talk at TED'09!

It's TED's 25th anniversary. But instead of looking back, we looked forward. Everyone on the TED2009 program premiered something: a thrilling discovery, a revolutionary idea, a powerful invention, a spectacular piece of art, a first-time performance. We shared undiscovered talent and shed new light on names you know. We're now busy turning these astonishing four days into TEDTalks to share with the world throughout the coming year.

French Translation of the speech :

Figure 10: (Left) A graphical map of the video. (Centre) Hypervideo player. (Right) Translated Content and supplementary information from Wikipedia (Sadallah, Aubert and Prié, 2012).

Table 1: state of the art for detail on demand and multimodality of content representation for exploring within a video.

Paper	**Year**	**interactive**	**Nonlinear**	**Both Search and Synopsis**	**Detail on Demand**	**Multimodal Representation**
Brachmann & Malaka	2009	Yes	Partial	No	Magnify the scrubber bar to see time aligned speech blocks	Sub titles along with video.
Yadav et al.	2015	Yes	Yes	Partial	Dynamic time aligned word cloud and slide show	Word cloud and slide images
Balasubramanian et al.	2016	Yes	Partial	Yes	Key words and interactive transcript	Keywords and transcript along with video
Galuscakova et al.	2016	Yes	Partial	Yes	Interactive transcript	Transcript along with video.
Pavel et al.	2014	Yes	Yes	No	Summary of transcript.	Text summary and a thumbnail in Table of Content format.
Meixner & Gold	2016	Yes	Yes	Yes	Multilevel Table of Content	Annotated hypertext
Hildebrand & Hardman	2013	Yes	Yes	No	Supplementary Information	Supplementary information on second screen
Shipman et al.	2008	Yes	Yes	Yes	Hierarchal links to clips in video	Supplementary information in custom video player
Tiellet et al.	2010	Yes	Yes	No	Supplementary images and annotation.	Supplementary images and annotation.
Sadallah et al.	2012	Yes	Yes	No	Links to external content	Links to external content

2.2.4. Discussion regarding interactive exploration approaches within a video

The state of the art review has revealed that while current approaches do provide interesting applications, they are limited in utilizing the potential of video while representing the content. For example, hypervideos and interactive video navigation systems do allow users to explore video content in a nonlinear and interactive manner, and there has been some attempts to allow user to explore content at different level of detail (Shipman, Girgensohn and Wilcox, 2008; Tiellet *et al.*, 2010; Hudelist, Schoeffmann and Xu, 2015; Yadav *et al.*, 2015), however, the multimodality of video content is still underutilized in the presentation of content to the user. In essence, current state of the art approaches are limited in either one or more of the following aspects:

- Lack of user control in the configuration of the representation of content (section 2.2.2).
- The solution is either designed to provide an overall synopsis of the video or search for something in particular, not a combination of both, which affects the user experience in tasks which have evolving exploration goals (Hong *et al.*, 2011; Ruotsalo *et al.*, 2015).
- The user's ability to interact with the content is either limited or the interface is designed to be either:
    - Content dependent for example (Monserrat *et al.*, 2013).
    - Overly complex (Cobârzan *et al.*, 2017b; Meixner, 2017; Sauli, Cattaneo and van der Meij, 2017)
- Requires prior curation by humans i.e. manual effort (section 2.2.1.1).

It can be seen in Table 1 and in the description above (section 2.2.3), that researchers have experimented with presenting multimodal content in different level of details to a user to enhance the exploration experience within video. The multimodality of the representation is either in the form of links to supplementary content or a display of transcript of the spoken word and/or keywords shown along with video play. The review of the approaches in Table 1, revealed that multimodal potential of video as a multimedia content source is underutilized and users have limited if any control in the process of content curation i.e. the approaches:

- Require some degree of human curation for multimodal representation.
- Are highly customized systems which limit the choices for users in terms of how they want to consume the content.
- Multimodality is either limited or is in the form of supplementary external information.

It is the contention of this thesis that an exploration approach can utilize multimodal features more widely to enhance the user's exploration experience with video content. The approach should automatically curate content on demand by representing the content in a configurable manner. By configuration, it is meant that content representation may be configurable not just in terms of the amount of detail but also in the choice of combination of different modalities. Automatic curation of extracted features would minimize the dependence on prior human curation and supplementary material and would allow users to get more value out of video content. Providing the ability to change the configuration of the representation would enable users to go beyond the anticipation of designers and customize the content to their evolving exploration needs.

The following chapter details the proposed approach of this thesis. The proposed approach addresses the gap in the state of the art described above.

3. Proposed Approach Design

This chapter describes the proposed approach of this thesis. The proposed approach is presented as a framework of a template driven representation engine that is designed to answer the research question described in section 1.2 and fulfil the characteristic of an exploration approach, outlined in section 2.1.2.

This chapter describes the theoretical design of the proposed engine-based approach named RAAVE. The experimentation with the developed prototype, based on the design and its evaluation, is discussed in later chapters. The design of the approach is influenced by the state of the art and from the learnings of the experiments performed in phase 2 (section 5).

3.1. Introduction

The purpose of this thesis is to evaluate the research question described in section 1.2 i.e. to evaluate the extent to which multimodal features extracted from video can enhance the user's video exploration experience. As described in section 2, video exploration can be a combination different tasks namely:

- Retrieval
- Navigation
- Synopsis or summarization

Therefore, the proposed approach named RAAVE is designed to be able to perform all of the above tasks. It has also been detailed, in section 2.2.3, that out of the five characteristics of an exploration approach, a major gap exists in the state of the art when it comes to the following:

- Detail on demand representation of content.
- Multimodality in the representation of the content.

Current approaches underutilize the multimodality in content representation and granting user control in the level of detail of the information, in an interactive manner.

This is because current researchers approach the problem by creating highly customized interfaces which augment supplement informational and/or control elements around a video. The problem with such solutions is that they cannot change with evolving user needs thereby limiting the exploration experience of the user.

The proposed approach (RAAVE) solves this issue by taking a different strategy. Instead of supplementing information or customizing the end interface, the proposed approach represents automatically extracted multimodal features in a configurable manner, independent of the end user interface. By configuration, it is meant that content representation is configurable not just in terms of the amount of detail, but also in the choice of combination of different modalities. The configuration of extracted features in the representation is done with the help of templates. Hence, representation of content can be modified by changing the template selection. In order to select templates for representing the content, the proposed approach utilizes a representation engine which uses a template collection and a template matching process. The template collection and template matching process are inspired by the finding of experiment 2 (section 5). The following section describes the design of the proposed approach.

3.2. Approach Design

The proposed approach (RAAVE) utilizes the fact that a video is not just a single/homogenous artefact but, it is a combination of different temporally bound parallel modalities (visual, audio, linguistic/ textual). As described in the state of the art, current approaches to represent video content are customized for a particular use case and cannot be reconfigured for evolving user needs. To solve this, RAAVE works as a representation engine independent of a user interface. Figure 11 shows the overview of the approach. RAAVE works in two phases.

- Extraction and Indexing
- Representation through template matching

Both phases work independently. Extracted features are stored in a repository. Representation is done independently of the extraction so that the configuration of the representation can be dynamic and flexible.

3.2.1. Extraction and Indexing

The steps involved in the phase are as follows:

- Video Segmentation
- Multimodal Feature Extraction

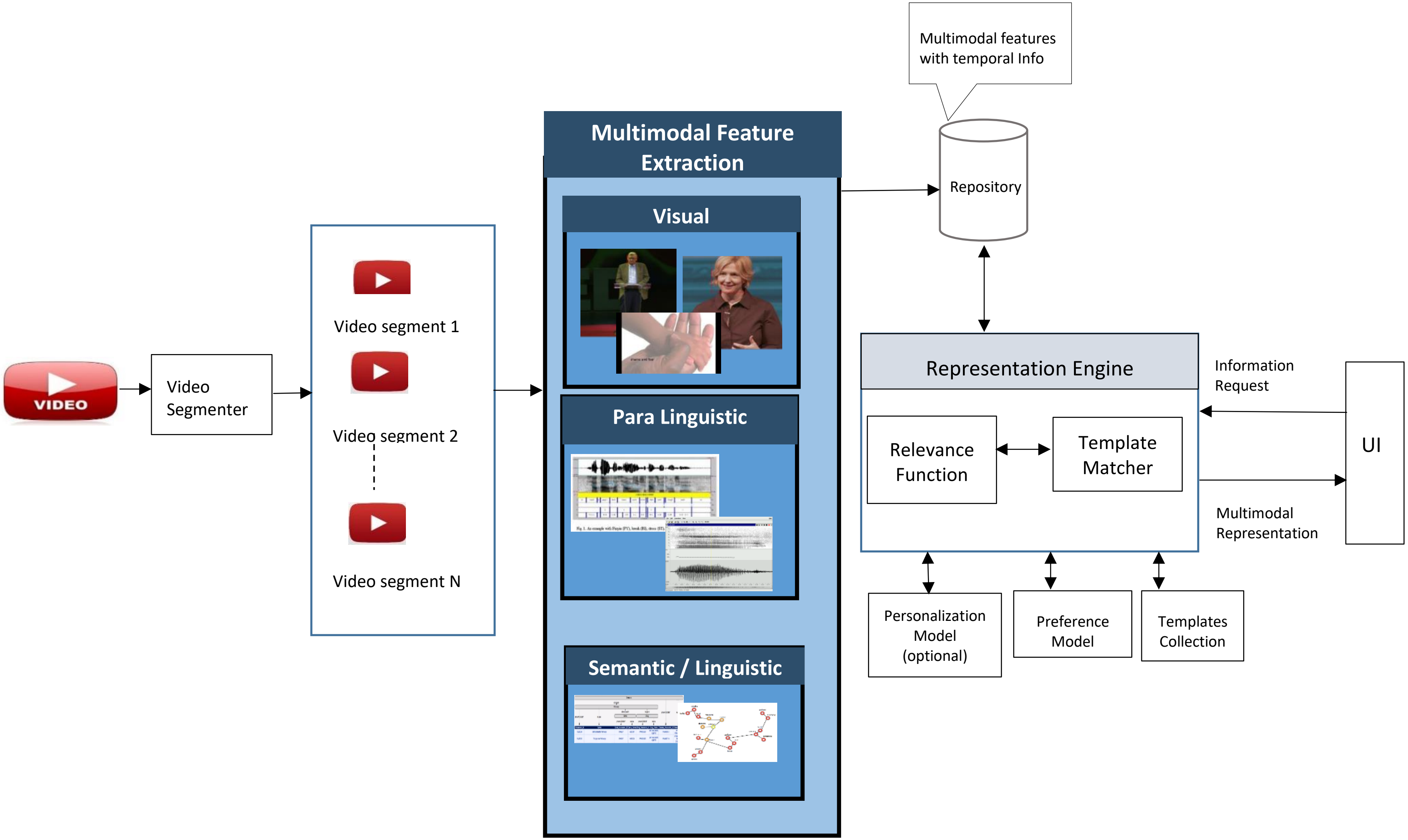


Figure 11: Overview of the proposed approach (RAAVE)

3.2.1.1. Segmentation

To generate an appropriate representation, the representation engine needs segments regardless of how they are segmented. The focus of this approach is not to devise a new segmentation technique but to utilize already existing video segmentation techniques to segment the video and then represent the segment in a multimodal and configurable manner.

Researchers have developed many techniques to segment videos (section 2). The choice of a particular segmentation approach depends on many factors e.g. the genre of the video.

This thesis is focusing on presentation style information video e.g. TED talks. Even with presentational style videos there can be multiple ways in which a video can be segmented.

Systems can choose from already developed off the shelf techniques to segment a video into smaller units based on the genre of the video. As an example, consider the segmentation algorithm used on TED style informational videos in the experiments: the C99 text algorithm (Choi, 2000).

Other segmenters which can be used are:

- visual
- multimodal and
- semantic/linguistic etc.

3.2.1.2. Feature Extraction

After segmenting the video, the next step is to extract multimodal features from the video segments along with their timestamps. To this end, the video needs to be decomposed into different modalities i.e. visual, audio, textual and video itself. State of the art tools and techniques can be utilized to extract features from these modalities.

By *multimodal feature extraction* it is meant: the characteristics or features within the different modalities by which a video delivers its message to the viewer (Figure 12). These may include the following:

- The visual modality i.e. anything visually interesting or engaging to the viewer, e.g. visual features such as a camera close up or visual aid, etc.
- The paralinguistic modality, i.e. the audio features, e.g. laughter, applauses and other audio features.

- The linguistic modality, i.e. the spoken words, any text within the video as well as any supporting resources, e.g. human written synopsis etc.

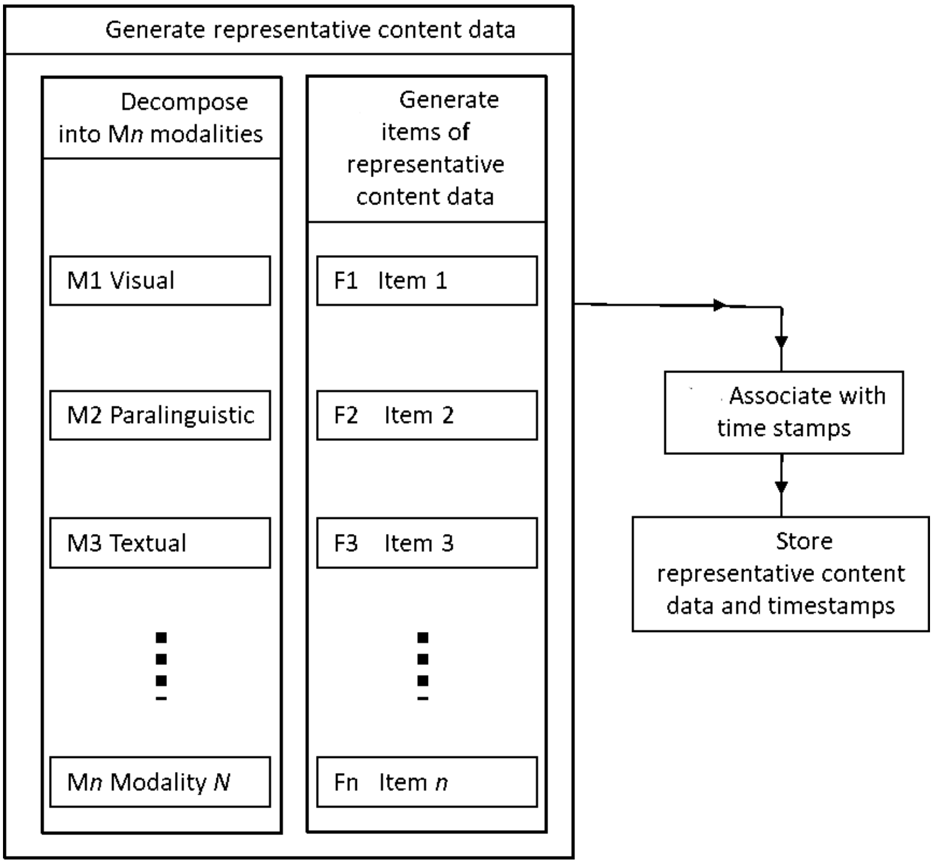

Generate representative content data
Decompose into Mn modalities
M1 Visual
M2 Paralinguistic
M3 Textual
Mn Modality N
Generate items of representative content data
F1 Item 1
F2 Item 2
F3 Item 3
Fn Item n
Associate with time stamps
Store representative content data and timestamps

Figure 12: Feature Extraction from parallel modalities of video. Where M_1, M_2 ...M_n are different Modalities and F_1, F_2...F_N are different features extracted from the modalities.

3.2.1.3. Expanse of Feature

The need for different feature representations in certain configurations depends on the fact that different features have a different expanse. Some features would offer more detailed content to the user i.e. they would have a deeper expanse in terms of information value, while others would offer less detailed content to the user. However, the less detailed features would be more efficient to consume in terms of time.

For example, consider the actual video footage of a particular segment. It would offer the full content of that segment but require more time to view it, compared to an automatic text summary generated from the segment transcript, for instance. The text summary would require less time for the user to consume but its expanse in terms of content value would be limited compared to the video footage. Similarly, consider key frames from the video footage or a word cloud of key terms from textual transcript. Both will be efficient in terms of time but limited in terms of depth of information.

Hence different features may have different expanse of depth of information and they would also belong to different modalities.

Hence different features may have different expanse of depth of information and they would also belong to different modalities.

As mentioned above, the goal is to represent the content of video in different configurations. By configurations, it is meant different combinations of extracted features which would internally offer a different expanse of information to users, and they would do so in different modalities.

3.2.1.4. Feature Availability

It is possible that certain features are not present in a particular video. As an example, consider a TED presentation video in which the presenter does not use any PowerPoint slides or any other visual aid. For such a video, the tools designed to extract slides from a video would not return any output, hence the keyframes feature would not be available for that video which would affect the choice of potential representations for that video. For the sake of simplicity, the discussion onwards considers two modalities only i.e. visual and textual.

3.2.2. Representation through template matching

Once the video is segmented and multimodal features extracted, the next step is to represent the segment in an appropriate configuration. To do that, a representation engine utilizing a template collection is proposed.

For each segment of the video, the representation engine chooses a suitable template. A template is essentially a configuration setting to represent the extracted features.

The engine works on a request response cycle. Upon receiving a request for information from a UI (User Interface), the engine does the following activities.

For each segment of the video it:

1) Determines the degree of relevance of the segment with the current request for information.
2) Based on the relevance, chooses an appropriate template for representing the segment.

In order to perform the two tasks, the engine requires the following:

1) Determining the relevance using a Relevance Function
2) Template Matching by utilizing a Template Collection

3.2.2.1. Template collection

The Template Collection contains the list of templates from which the engine can choose an appropriate template.

A template is a configuration setting to represent the extracted features of a video segment. A template basically determines which extracted feature or combination of features shall be included in the representation of a video segment.

A template has the following dimensions:

- Expanse
- Primary Modality

The number of templates is dependent upon the extracted features, as a template is a possible permutation of available features. Not all the permutations are included in the collection.

Table 2: Dimensions of template matching

	Dimensions	
	Expanse	**Primary Modality**
Possible Values	Efficient	Visual
	Deep	Textual
		Mixture

The following is an example of a template collection:

Table 3: An example of a template collection

Template ID	Expanse	Primary Modality	Feature(s)
1	Efficient	Textual	Word Cloud
2	Efficient	Visual	Key Frames
3	Deep	Textual	Text summary
4	Deep	Visual	Video Snippet
5	Deep	Textual	Word Cloud, Text Summary
6	Deep	Visual	Key frames, Video snippet
7	Deep	Mixture	Key frames, text summary
8	Deep	Mixture	Word Cloud, Video snippet
. . .			
N	Deep	Mixture	A permutation of extracted features.

3.2.3. Template Matching Criteria

The template matching process is based on the following 3 criteria:

- Relevance
- Expanse
- Primary Modality

3.2.3.1. Relevance

Determining what segment is relevant or important in a given context. The hypothesis is that relevant segments need to be represented in greater detail over non-relevant segments.

3.2.3.2. Expanse

Different features offer different amounts of information within a video segment i.e. they have a different expanse. They are either deep or efficient. If a feature is efficient to consume, then it would not offer detailed content. Alternatively, if it is deep in terms of content then it would require more time to be consumed.

Take, as an example of textual features, a word cloud of keywords generated from a transcript of a video segment. The word cloud is efficient to consume in the sense that it can be glanced at quickly, but it does not offer much detail of what is discussed in the segment. Alternatively, if a text summary is generated from the transcript then it would take longer to read it, however it will also give a more detailed understanding of the video segment.

Therefore, a word cloud is an efficient feature while a textual summary is a deep feature of linguistic modality.

3.2.3.3. Primary Modality

A segment can be represented in either a single modality or a mixture of modalities depending on the context.

3.2.4. Template Matching

The Template Matching process is essentially a 3-dimensional problem. Given a video to represent, the engine has to find an appropriate template from the collection for each segment. The engine does so base on 3 criteria (section 3.2.3).

It is the hypothesis of this research that relevant segments would require a deeper exploration. Therefore, the first step in matching a template is finding the relevance value of a segment. Determining relevance determines the depth of the segmentation representation i.e. the choice of template.

3.2.5. Determining the Relevance

RAAVE transforms video content based on the context. In order to choose a template for relevant segments, the representation engine needs to determine the relevance of each segment.

Whether a segment is relevant or not in a given context can be determined by many factors. For representation purposes it makes sense to assume that given a user query, if a segment contains the keywords of the user query, then it is relevant.

Apart from the query, personalized interest can also determine the relevance of a segment. A segment may be relevant if the query terms appear or user’s topic of interest appears in the segment. Similarly, the segments surrounding the segment in question may also determine its relevance.

The following are some of the factors which determine relevance of a segment:

- Request context
- Personalization model
- Segments preceding and following the segment.

3.2.5.1. Request Context

By request context, it is meant the current informational need of the requesting entity. It could include, but is not limited to, the search query.

The context may also be information such as time, location or the device initiating the request.

3.2.5.2. Personalization Model (Optional)

There can be topics which the user might be interested in. So even if a video segment does not contain the info required by the current query request, it might have info which might be of interest.

In the case of absence of a search query, the personalization model becomes more important to determine the degree of relevance of a segment.

3.2.5.3. Relevance Function

The Relevance Function is the component of the engine which takes as input a segment and relating factors and returns the relevance value of that segment.

Segment_Relevance = getRelevance (seg, Pseg, Fseg, pm, co).

Where:

- *Seg* is the segment in question
- *Pseg* is the segment preceding the segment
- *Fseg* is the segment following the segment
- *pm* is the personalization model
- *co* is the request context

Segment_Relevance may be a Boolean or an overloaded version of *getRelevance* may return more than binary values for relevance.

3.2.6. Choosing a Template

After determining the relevance of a segment, what remains is selecting an appropriate template for the segment. The representation engine must choose a template based on expanse and primary modality.

The engine determines the expanse of the template based on the segment's relevance i.e. deeper exploration is desirable if the segment is relevant.

Once the depth value has been determined, the only thing left to determine is the modality of the template. Now the modality can be singular (visual or textual etc.), or it can be a mixture. Based on experiments, it is the assumption of this thesis that a mixed modality is appropriate for relevant segments. For this reason, a mixed modality is used only for relevant segments and single modality for others.

The next step in narrowing down the choice of template is choosing a particular modality value for the segment template. The engine must choose a modality value in the case of a single modality template. In the case of a mixed modality, the engine has even more things to consider i.e. it needs to decide if all modalities will be represented by deep features or a combination e.g. visually deep and textually efficient etc.

The engine narrows down the choice of modality based on two criteria. They are:

- Segment Suitability
- Preference Model

3.2.6.1. Segment Suitability

As discussed in section 3.2.1.2, multimodal features are extracted from video segments using different tools. It is possible that certain features were either not extracted from the segment for no particular reason or they were not suitable from the point of view of content value.

3.2.6.2. Preference Model

In experiment 2 (section 5) it was found that a user may prefer a particular modality i.e. fast reader or prefer visual information. The representation engine utilizes the personalization model to narrow down the choice of template for the segment.

In summary, the engine determines the modality through a combination of a segment's feature suitability and use preference model. It gives priority to segment suitability and, if suitable, it uses user preference to narrow down the choice of template to represent the segment.

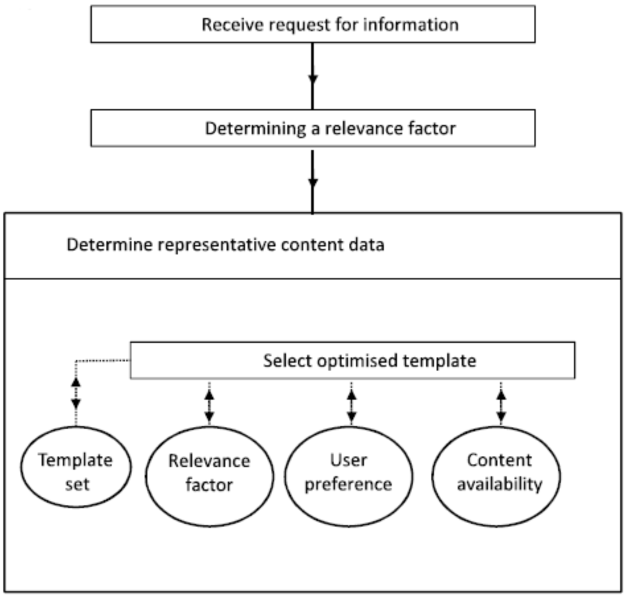

Receive request for information
Determining a relevance factor
Determine representative content data
Select optimised template
Template set
Relevance factor
User preference
Content availability

Figure 13: Template matching process overview

Figure 13 shows the components involved in the template matching process. The pseudo code for the representation engine is presented in the following section. This pseudo code is used to develop the prototype which evaluates the proposed.

3.2.7. Pseudo Code

Preconditions

1) Video has been segmented
2) Features are extracted and indexed
3) Template collection, preference/personalization model is available

```
For each segment do:
    Determine if segment is relevant to a given context by relevance function.
    If segment is relevant choose deep templates
    Else choose efficient template
    If segment is efficient then choose single modality
        Find segment suitability for modality
        If only one suitable template to choose from then represent segment
        Else see user preference
        Choose template with users preferred modality
    Else choose mix modality
        Choose visually efficient and textually deep or vice versa based on segment suitability and
        User preference giving priority to segment suitability.
```

3.2.8. Design Summary

Effective exploration of video means that the user may not want to or need to consume all the video. To put it in another way not all parts of video may be equally relevant in each context. Therefore, it would make sense to show in the representation, the relevant parts of the video using more detailed representation. However, user may not want to completely ignore the other parts of the video, to better comprehend the video or for any other reason. Therefore, it would make sense to represent those parts in a less detailed representation. That way the user can focus on the relevant parts of the video without completely losing the information in the other parts.

To represent the segments to the user, RAAVE utilized a representation engine. The representation engine is essentially the main component of the whole approach. It is the link between the user interface (UI) and the extracted features of the video segments.

In summary the engine determines the modality by combination of segment's feature suitability and use preference model. It gives priority to segment suitability and if suitable it uses user preference to narrow down the choice of template to represent the segment.

In short:

- a segment of video is either relevant or not in a given context
- a segment is represented by feature(s) which are deep or efficient
- the representation belongs to modality visual, textual or mixture.

Hence, we can describe effective video representation as representing each segment based on appropriate value of three dimensions which are expanse, relevance, and primary modality.

The following chapter presents the first phase of the experiments performed in this thesis. It details the multimodal feature extraction phase of the thesis i.e. point 2 of the research objectives (section 1.3).

4. Phase 1: Multimodal Feature Extraction and User Engagement Assessment

The main question this research evaluates is the effectiveness of representing extracted features from video content to users to enhance their exploration experience. The first phase of the research was to examine the techniques to extract multimodal features from video content i.e. point 2 of the research objectives (section 1.3).

State of the art review (section 2) made it clear that the choice of extracted feature is highly dependent on the genre of video content. Therefore, considering the diverse nature of video exploration, this research focused only on informational and infotainment videos. The experimentation is performed on TED presentation videos.

4.1. Video Dataset

The experiments are performed using TED presentation videos. TED videos have become very successful with over 24,00 talks as of March 2016, since they were first published online in June 2006.[9] Whilst aimed at a more general audience they nonetheless tackle a wide variety of topics.

4.1.1. Why Presentation Style Videos

A video presentation such as a TED video typically involves one speaker presenting a topic, often accompanied by supporting media, such as still images or further video. This form of video may be seen as both simple and sophisticated at the same time. Presentations are simple in the

[9] http://en.wikipedia.org/wiki/TED_%28conference%29 -- last verified: November 2017

sense that there is usually just one person continuously talking to an audience with or without audio/visual aids; they are sophisticated in the sense that they can deliver semantically diverse kinds of messages to their viewers and engage them in a variety of ways. The relative simplicity (visual, audio, linguistic) of presentation style videos compare to movies or songs etc. makes them a good candidate for experimenting for the proposed approach.

4.1.2. Why TED talks

Mostly research on presentation style videos focuses on educational presentations such as course lectures, Massive Online Open Courses (MOOC) or e-learning scenarios (section 2). The effectiveness of formal course specific videos and any developed techniques is usually measured by comparing the performance of students using the system with students not using the system (Mujacic *et al.*, 2012b). Since e-learning videos are task or goal oriented they are only consumed by a particular audience for specific purposes.

It is the goal of this thesis to propose an approach which empowers consumer to explore content which is beyond specific task and goal thereby having a wider applicability. TED talks thanks to their general and storytelling nature, appeal to a wider and more diverse audience and therefore are an ideal candidate for the experimentation. While the experimentation is performed on TED presentation videos, it is the assumption of this thesis that the approach will be extendable to other content type such as lifelogging videos, product launches and video messages etc. Due to its information seeking focus, the proposed approach is not expected to be effective for video type such as movies, songs or entertainment-oriented videos.

While they are task specific TED videos are more general purpose therefore engage a larger more diverse viewer base.

4.2. Analysing Multimodality of Video for User Engagement Assessment

4.2.1. Hypothesis

It is possible to extract quantifiable multimodal features from a video presentation automatically and correlate these with user engagement criterion.

4.2.2. Motivation

4.2.2.1. Primary Motivation

The goal of this research is to enhance the user experience while exploring content of video. User engagement is a key quality of this process (O'Brien and Toms, 2013). Therefor this experiment evaluates the relationship between multimodal features and user engagement. The extracted features form this experiment were utilized in representing video content to users to enhance their exploration experience i.e. phase two of the research (section 5).

4.2.2.2. Secondary Motivation

Apart from identifying the value of extracted features in the user experience, identifying the relationship between extracted features and user engagement also have potential in other applications. For example, there is an enormous amount of audio-visual content available on-line in the form of talks and presentations. The prospective users of the content face difficulties in finding the right content for them. Automatic detection of interesting (engaging vs. non-engaging) content can help users to find the videos according to their preferences. It can also be helpful for a recommendation and personalized video segmentation system.

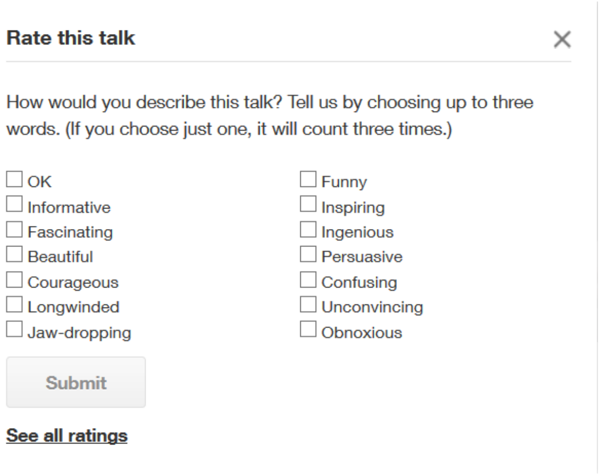
4.2.3. User Engagement with video content

In order to identify the relationship with user engagement, it is important to first define user engagement with video content in the current context. According to (O'Brien and Toms, 2013) user engagement is based on six factors such as Perceived Usability (PUs), Aesthetics (AE), Novelty (NO), Felt Involvement (FI), Focused Attention (FA), and Endurability (EN) aspects of the experience. Specifically for videos (Dobrian *et al.*, 2011) and (Guo, Kim and Rubin, 2014) analyse user engagement by measuring for how long a user watched a video. Questionnaires are also a very common method for analysing engagement factors in video artefacts as in (Benini, Migliorati and Leonardi, 2010b; Haesen *et al.*, 2011). As it can be seen there is not much agreement in measuring engagement. It is because engagement with content is highly context dependent (Attfield, Piwowarski and Kazai, 2011).

Therefore, this research views engagement as the elaborate feedback system described in detail in the following section.

4.2.4. TED talks and user feedback

The interesting thing about the user feedback on the TED website is that in addition to asking users to simply tell if they like or dislike a particular presentation, it also asks viewers to describe the video in terms of particular words. A user can choose up to three words among the choice of 14 to rate a video. Figure 14 shows the choices available to the user to rate a particular TED talk.



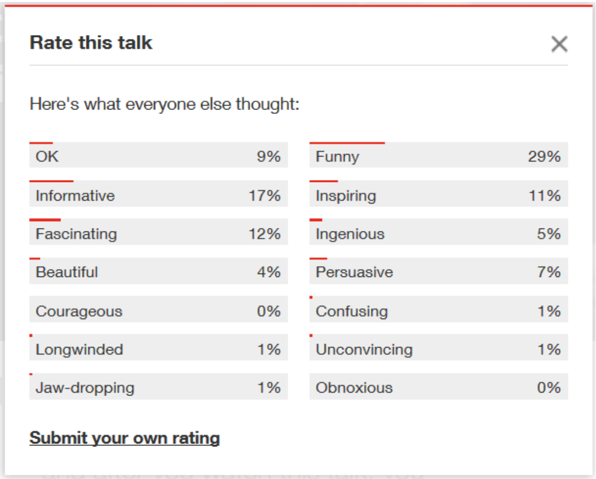
Rate this talk ✕

How would you describe this talk? Tell us by choosing up to three words. (If you choose just one, it will count three times.)

- □ OK
- □ Informative
- □ Fascinating
- □ Beautiful
- □ Courageous
- □ Longwinded
- □ Jaw-dropping
- □ Funny
- □ Inspiring
- □ Ingenious
- □ Persuasive
- □ Confusing
- □ Unconvincing
- □ Obnoxious

Figure 14:Ted.com rating criterion.

The website shows the overall feedback about a video in terms of percentages i.e. among all the ratings given to the video what percentage of ratings found the video to be inspiring and what percentage found it to be ``Longwinded'' etc. (Figure 15).



Rate this talk ✕

Here's what everyone else thought:

OK	9%	Funny	29%
Informative	17%	Inspiring	11%
Fascinating	12%	Ingenious	5%
Beautiful	4%	Persuasive	7%
Courageous	0%	Confusing	1%
Longwinded	1%	Unconvincing	1%
Jaw-dropping	1%	Obnoxious	0%

Figure 15:Overall ratings of a TED video

By giving these choices to users, the TED website provides elaborate feedback on a given video. Therefore, there is no binary feedback to learn from, but a rather fuzzy description of what viewers thought about a particular video. The rating system for user feedback thus provides a more nuanced characterization of user engagement and non-engagement with the video presentation. Since the ratings consist of voluntarily information given by the users, in terms of semantically positive and negative words, it provides good basis for analysis of the relevant factors of engagement for TED talks listed in section 4.2.3.

4.2.5. Data Collection

A collection of 1340 TED presentation videos along with subtitles files were downloaded from the TED website. A custom crawler written in python language was developed for this purpose. Apart from video and subtitle files metadata such as corresponding ratings given to each video was downloaded for analysis.

4.2.6. Analysis of User Rating

Since TED user ratings are not binary, but rather descriptive terms representing a user's feedback. It requires some pre-processing before any analysis could be performed on them, i.e. some kind of normalization is needed.

A rating given to an individual video cannot be simply relied upon. It is because the TED website reports what percentage of viewers rated the video as saying “Inspiring” or “Longwinded” etc. But if a particular video is rated by 1000 people and another by 10, then percentages of different rating criterion may not give the whole story. Though one can argue that since viewers can potentially rate any video they want, if a video is getting more viewers to rate it then it must have something which makes its viewers express their opinion. While it is possible to get an absolute number of rating count against each criterion for an individual video, the percentage may not adequately predict engagement of a video presentation.

Among the 14 rating criteria provided to users, 9 were identified as being positive words, 4 as being negative words, and 1 as neutral (Table 4).

Table 4:Rating Word Classification.

Rating Word	Classification
Beautiful, Courageous, Fascinating, Funny, Informative, Ingenious, Inspiring, Jaw-dropping, Persuasive.	Positive
Confusing, Longwinded, Obnoxious, Unconvincing	Negative
OK	Neutral

As it can be seen in Table 5, ratings tend to be overwhelmingly positive. Both count and percentage, positive criterion tend to score much higher than negative or neutral ones. Even the highest scoring negative ratings “Unconvincing” has average count and percentage 51 and 3.73 less than the lowest scoring positive rating “Funny” with average score of 106 and 4.73.

If only the ratings of an individual video are considered, then it would seem like all the videos only positively engage the viewers. Since there was not any video to which a negative rating word got the highest count by the viewers. So, to deduct which video is found to be "Obnoxious" or "Longwinded" by viewers, some kind of normalization is required. In order to do that the following definitions were used for the experiment.

For a video to be considered "Beautiful" or "Persuasive" etc. it must have a rating count more than average rating count for that particular rating word. With this, TED talks were categorized as "Beautiful and not Beautiful", "Inspiring and not Inspiring", "Persuasive and not Persuasive" etc. giving two classes for classification for each of the 14 rating words.

Table 5:Global Average per each Rating Criteria

Rating	Average (Count)	Average (%)
Beautiful	120	6.67
Confusing	**15**	**1.17**
Courageous	122	6.08
Fascinating	234	12.64
Funny	106	4.73
Informative	246	15.24
Ingenious	134	7.64
Inspiring	384	18.16
Jaw-dropping	118	5.45
Longwinded	**28**	**2.23**
Obnoxious	**23**	**1.62**
OK	*65*	*4.88*
Persuasive	188	9.70
Unconvincing	**51**	**3.73**

TED talks provide topic tags with each video. For example, a TED talk could have topic tags like "Culture, poverty, history and photography". These topics tags are quite diverse i.e. out of the 1340 videos used in the experiment, there were a total of 321 unique tags found many of them only appearing once while a single video had up to 10 different topic tags.

4.3. Experiment 1.1: Higher Level Features (Visual + Paralinguistic) for user engagement assessment

4.3.1. Hypothesis

A Multimodal feature set is better for classifying engaging videos than single modality feature set.

4.3.2. Feature Extraction

4.3.2.1. Visual Features

As far as visual features are concerned most researchers tend to focus on low level generic features without associating any semantic meaning to it. This is useful for detecting scene changes and similar things (section 2). Since this thesis is looking at presentations and for this study specifically at Ted talks, there are not many visual scene changes. Instead this experiment took an approach similar to Hosseini & Eftekhari-Moghadam and Haesen et al. (Haesen *et al.*, 2011; Hosseini and Eftekhari-Moghadam, 2013) by taking higher level visual features. The approach of this experiment is perhaps closer to Haesen et al. since they also look for faces within videos (Haesen *et al.*, 2011).

HAAR cascades (Lienhart, Maydt and Lienhartintelcom, 2002) in OpenCV library (Bradski, 2000) were used to detect when the speaker is on the screen or not. For this study, the number of seconds in which there was a close up shot of the speaker and when there was a distant shot and when the speaker was not on the screen were calculated.

4.3.2.2. Paralinguistic Features

For paralinguistic features, the number of laughter and applauses and laughter applause ratio within TED talks were calculated. Since TED talks come with subtitles, getting this information was a simple process and was obtained with a simple program written using the python programming language.

For all extracted features, it was also measured whether for each video, the value of each feature was greater or less than the average value for that feature. For example, if the number of close up face seconds for a given video was greater than the average number of close up face seconds, value 1 was assigned to the feature “Above average close up shots” and 0 otherwise. The same was done for other features thereby doubling the number of visual and paralinguistic features to 12 for the experimentation.

4.3.3. Classification Method

For correlating features with user ratings to see some potential patterns the WEKA toolkit which allows easy access to a suite of machine learning techniques (Hall *et al.*, 2009) was utilized. Machine Learning Algorithm Logistics Regression, and tenfold cross-validation testing for the analysis was used on 1340 TED talk videos, to see how feature values affected user ratings. Both percentage count and actual count for each rating were tested.

4.3.4. Results and Discussion

The aim of this experiment is to see the value in the multimodality of video content. Experiments were performed by removing visual features to see if this will affect the correct classification of video for the ratings. Figure 16 shows that the accuracy of correctly classified instances increased with the inclusion of visual features for most of the rating words, 7 to be precise. While for 3 it remained equal but for 4 rating words it actually decreased.

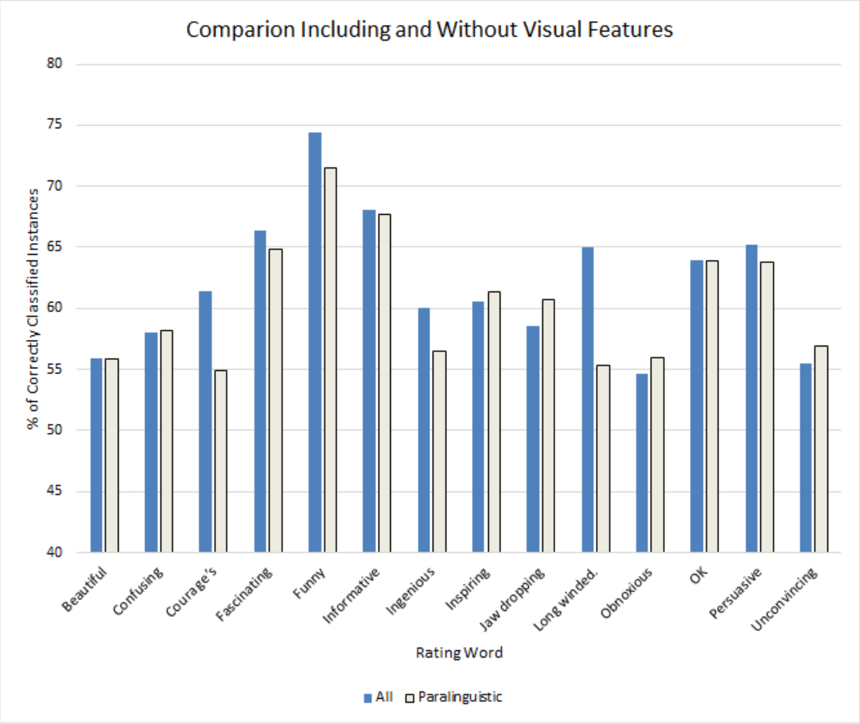


Figure 16:Comparison including and without visual features for classification of Ratings.

Results of the experiment are interesting in many regards. Firstly, it is the preliminary step towards the hypothesis about the value of different modalities within a video stream. Another interesting aspect of this study is that all the features were automatically extracted, i.e. no manual annotation was performed. So, any model based on the feature set could be easily used for new content and any advancement in computer vision and paralinguistic analysis technology would help in making the model better. This model has the potential to become a component of personalization systems to enhance contextual quires.

The experiment had the following limitations

- The feature set is limited.
- Accuracy is not the most robust measure of classifier performance.

The next experiment attempted to address some of these limitations.

4.4. Experiment 1.2: Utilizing both High and Low-level feature set for assessing user engagement with videos

4.4.1. Hypothesis

Combination of High and Low-level features is better at classifying engaging vs. non-engaging videos compared to just using High level features.

4.4.2. Feature Extraction

All the features of experiment 1.1 as described in section 4.3.2 are used for this experiment. In addition to these, prosodic features are also extracted from the audio stream.

4.4.2.1. Prosodic Features Extraction

The openSMILE toolkit (Eyben *et al.*, 2013) was used for prosodic feature extraction. This extraction was performed on the audio files extracted from TED talks videos using FFMPEG (Bellard, Niedermayer and Others, 2012). The extracted audio files have a sampling frequency of 44.1 KHz with a resolution of 16 bits. In this study, the prosody feature set of ComParE challenge used in The INTERSPEECH 2013 computational paralinguistic challenge (Schuller *et al.*, 2013) (6373 features in total) was chosen. The ComParE feature set include the Energy, Spectral, Mel-Frequency Cepstral Coefficients (MFCCs), and Voicing Related Low-Level Descriptors (LLDs). A few LLDs including logarithmic harmonic-to-noise ratio (HNR), Voice quality features (harmonic to noise ratio), Viterbi smoothing for F0, Spectral harmonicity and Psychoacoustic spectral sharpness.

4.4.3. Classification Method

Firstly, normalization of the feature set was performed by using z-score normalization technique and then they were scaled in the range of [0 1]. To reduce the high dimensionality of features PCA (Principle Component Analysis) was employed over the feature set to reduce the number of dimensions to number of instances. After that the data was mapped to the reduced dimensions.

From the statistical significance(***p***) of the transformed feature set with the rating (yes or no), Transformed features with ***p*** value less than 0.05 were selected.

MATLAB[10] (Statistics and Machine Learning Toolbox) was used to perform classification and apply the discriminant analysis method with 10-fold cross validation. The classification method works by assuming that the feature sets of the classes to be discerned are drawn from different Gaussian distributions and adopting a pseudo-linear discriminant analysis (i.e. using the pseudo-inverse of the covariance matrix.)

4.4.4. Results and discussion

In addition to automatically assessing user engagement in videos, this experiment also investigated the value within different modalities of video for application purposes. Therefore, experiments were performed comparing prosodic features alone and prosodic features combined with visual and paralinguistic features to see if there is any difference in classification results.

Classification was performed as both a 2-class mentioned in section 4.2.6, and assessed the results using the F score statistic (with the β parameter set to 1). In this setting, the F score is equivalent to the harmonic mean of the precision and recall scores. For two-class "Yes" and "No" classification, F-scores as high as 96.93% (Figure 17), were obtained. Although the highest F-score was achieved with the fusion of prosodic and other features, it was not the case that fusion improved the score, in all cases. The number of cases where fusion increased the F-score are equal to the cases where prosodic features alone resulted in a higher score (see Figure 17 for details).

[10] http://uk.mathworks.com/products/matlab/

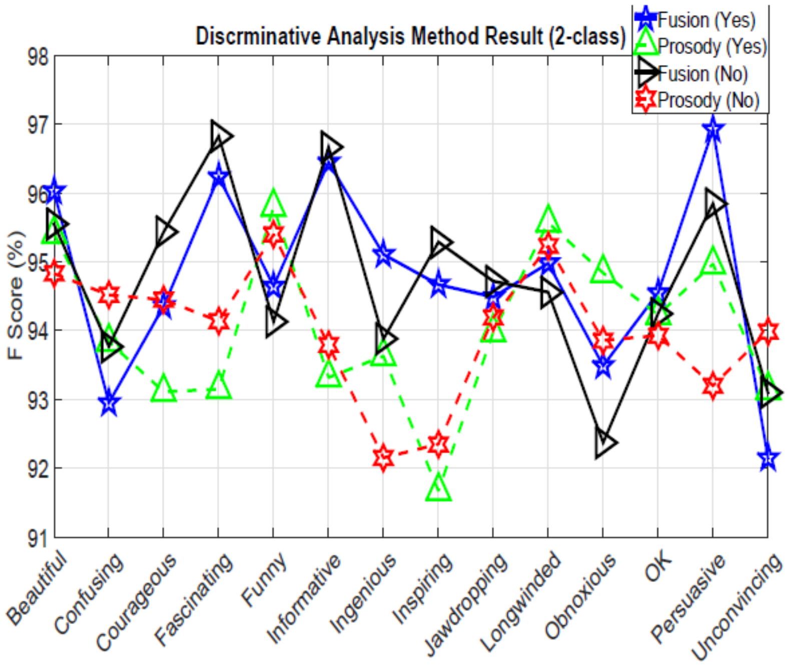


Figure 17: Classifier results (F Score) for 2-class problem, comparison of Fusion and Prosody features.

These results are widely applicable. The resulting model can become a component of a personalized video recommender system. It can also enhance contextual search queries; since extracted features for the study together with their timestamps in the video can be indexed to support segmentation and add value to any other metadata of the video collection. This multimodal meta-data and engagement correlation, combined with other extracted meta-data would help a representation engine in transforming a video form a single artefact into a customizable and context aware information retrieval source.

4.5. Concluding remarks on Phase 1

Experiments in phase 1 provided the multimodal features to enable exploration of video in a configurable and interactive manner. Apart from being the feature extraction phase of the proposed approach. Phase 1 gave valuable information regarding the value within different modalities of video content.

4.5.1. Other uses of user engagement assessment

Other researchers have utilized the dataset and extracted features of phase 1 for different applications. For example, Haider et al. used the configuration of experiment 1.2 (see section 4.4 for details) to perform analysis of variance (ANOVA) test to analyse viewer engagement with presentations for the purpose of providing feedback to presenters to help them improve the engagement level of a talk (Haider *et al.*, 2016). In a recent study, the same authors (Haider *et* *al.*, 2017) extends the idea by segmenting TED presentations based on speech expressions to identify the user-engagement at segment level.

Phase 1 ratified the value of automatically extracted multimodal features. In phase 2 of the research, extracted features are represented to users in an interactive manner. The following chapter describes the experiment performed in phase 2 i.e. point 3 of the research objectives (section 1.3).

5. Phase 2: Representing Extracted Multimodal Features to Users in an Interactive Manner

The goal of this research is to enhance the user experience in exploring video content. Researchers have long identified the importance of user control in the process of video search (Cobârzan *et al.*, 2017b). Therefore, in order to streamline the design of the proposed approach, in phase 2, a user study is performed by representing automatically extracted multimodal features to users (research objective 3 section 1.3). Users get the ability to control the way they consume the contents of video in the exploration session. This chapter describes the details and outcome of the user study.

5.1. Representing multimodal features to users to learn usage patterns

5.1.1. Hypothesis

Multimodal features from video content can be presented to viewers in order to enhance their exploration experience.

5.1.2. Motivation

It is the contention of this thesis, that giving users the ability to interact with video in a non-linear manner helps them in the process of exploration. Different systems have been proposed which help users explore time-based media For example (Luz and Masoodian, 2004) and (Pavel and Reed, 2014). While Hypervideos (section 2.2.1.1) and other interactive video systems (section 2.2) do give users more control in the exploration process, they either require some degree of human curation or they have a purpose-built user interface which ties them to the intended use cases, thus limiting their flexibility for evolving user needs.

By transforming video into an automatically generated interactive webpage, the exploration opportunities of video contents can be enhanced. This is because users can explore the content by directly interacting with the extracted features and reconfiguring the rendered webpage according to their evolving needs.

5.2. Experiment Design Overview

This experiment is based on the idea of considering video a diverse multimodal content source. It works by using known techniques to segment a video and automatically extract multimodal features from its different modalities.

Upon a content exploration request, content of video is represented as an interactive multimedia webpage instead of a typical video stream, thereby transforming it into an interactive document. The interactive webpage is automatically curated by representing the extracted features to the user so that user can consume the different segment of the video in the modality of his/her choice and the amount of detail preferred.

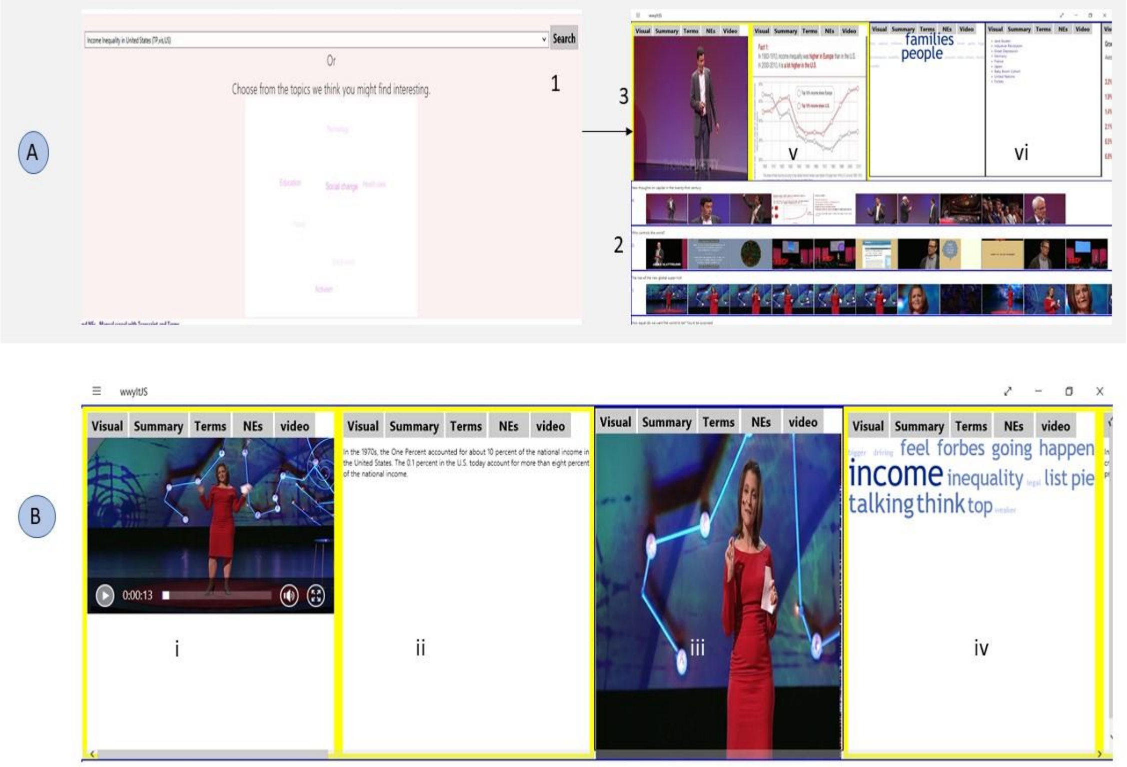
This form of representation gives the user more flexibility and greater control over how they choose to interact with and consume the content of the video. Since the proposed approach creates the webpage automatically, it does not limit the interactivity to the designer's anticipated use cases and allows the user to personalize the rendering according to the evolving task needs and personal preferences.

5.3. Prototype Design

To perform the user study, a prototype system was developed. The prototype was built using HTML5 and JavaScript and designed to work with both traditional and touchscreen interfaces. The semi-functional prototype is designed to simulate a typical exploratory search task i.e. users put a query in a search box and they get a list of videos as a result to choose from. Once the user chooses a video the prototype instead of playing the video enable the user to explore the content of the video in multimodal manner.

Figure 18A(1) shows the search box to enter textual queries. Once the user presses the search button Figure 18A(1) the results of the search. The results is essentially a list of videos for the user to explore. Figure 18A(2) shows the video list. Each entry contains video title with a filmstrip display of the keyframes (section 5.3.3) of the video. User can choose a video from the list to explore its content by clicking or tapping.

Once the user chooses a video by clicking or tapping on it. The content of the video are represented as an interactive multimedia document. Figure 18B shows a sample representation. The video is transformed into a multimedia document by segmenting it into smaller parts and extracting multimodal features from its different modalities. Following sections describe the techniques and toolset used for segmentation and feature extraction.



Choose from the topics we think you might find interesting.

In the 1970s, the One Percent accounted for about 10 percent of the national income in the United States. The 0.1 percent in the U.S. today account for more than eight percent of the national income.

Figure 18: Screen shots of the prototype system. A (1) upper left, shows the search box. A (2) upper right shows the results. A (3) upper right shows the detail representation of a video. (B) shows the top row of results page i.e. detail representation of a video. (B) shows the screen shot of the video representation, showing 4 out 16 segments. Users can swipe or scroll right to see the other ten segments. Each square represents a segment, and each segment has five tabs, where each tab contains the rendering of automatically extracted features. Segment one and two are highlighted by a yellow border in the figure. User can tap or click on the tabs to see different renderings for example B(ii) currently shows the summary. In short, to explore this video, users have 16 segments and 80 tabs (16 segments X 5 tabs) to choose from.

5.3.1. Segmentation

A video can be segmented in many ways utilizing different modalities (see section 2). The choice of modalities is often domain dependent. For the current study, the video is segmented by utilizing the textual modality. Thus, the transcripts of TED videos were first split into sentences using StanfordNLP toolkit (Manning *et al.*, 2014) and fed into the C99 text segmentation algorithm (Choi, 2000) in order to produce video segments.

Once segments are identified, to facilitate exploration, they can be represented to users in different ways. Users can not only choose which segments they want to explore but also what extracted features or their combination they would use to explore a chosen segment.

5.3.2. Highlighted Segments

The segments in which the query terms appeared are highlighted with a thick yellow border to be distinguished from other segments. The assumption is that users might want to interact more with highlighted segments than other segments. Figure 18B (i, ii, iv) shows the first two segments and the fourth as highlighted.

5.3.3. Visual

The *Visual* tab shows the key frames of a segment. A custom tool was developed using openCV (Bradski, 2000) to detect camera shot changes. From those scene changes, one frame from each shot was selected. From those selected frames, frames with and without a face were identified using a HAAR cascade (Lienhart and Maydt, 2002). Speakers often use visual aids, such as presentation slides in a TED presentation. The heuristic was that shot without a face after a shot with a face might contain some images of visual aid used by the presenter which could contain useful information. Users can tap or click on the frame to see all the selected frames to get a visual synopsis of a particular segment. Figure 18A(v) and B(iii) shows examples of extracted key frames.

5.3.4. Summary

The *Summary* tab shows the automatic text summary generated from the transcript of the segment. An online summarization tool for generating text summaries (Autosummarizer, 2016) was utilized. Figure 18B(ii) shows an example of extracted summary.

5.3.5. Terms

The *Terms* tab shows the word cloud generated from the transcript of the segment. We used the online tool TagCrowd (Steinbock, 2016) for the word cloud generation. Figure 18B(iv) shows an example of word cloud.

5.3.6. NE

The *NE* (Named Entity) tab shows the list of extracted named entities from the transcript of the segment. This prototype used the Named Entity Recognizer tool (Ratinov and Roth, 2009) Figure 18A(vi) shows an example of extracted named entities.

5.3.7. Video

The *Video* tab shows the video snippet of a segment. The text in transcript comes with timestamps. Once the segmenter segmented the text, timestamps were used to determine the start and end time of a particular segment and its video snippet was offered to users to watch. Figure 18B(i) shows an example of video segment.

5.3.8. Representation Details

Figure 18B show a sample representation of the TED talk by Chrystia Freeland (Freeland, 2013). The video is segmented into 16 segments (see 5.3.1). Each segment is represented by a square with each square containing 5 tabs. Figure 18B shows the first 4 segments. User can use the scroll or swipe with finger to view the rest of the segments.

Each segment contains 5 tabs showing the multimodal features extracted (see 5.3.3 to 5.3.7) to represent it content so that the user can explore the content in different modalities and amount of detail. This gives the user more control on the way they want to explore the content. For example the user get a quick overview of a segment using the word cloud tab (Figure 18Biii) and get more detail by using the summary tab (Figure 18Biv). if user prefers to explore the content visually than he/she can either choose the keyframes (see section 5.3.3 and Figure 18Bi) to get a quick visual overview or use the video tab (see Figure 18Bi) to explore the segment in detail. In addition to multimodality the user can explore the content in a nonlinear manner for example user can play the video tab in segment# 14 while swapping around and see the keywords of other segments to get a quick overview of the other segments of the video.

In summary, by using the multimodal representation, the user can control the modality(visual, textual) and the amount of detail (text summary or word cloud) in any combination or sequence based on his/her personal preference to better suit the evolving need of the exploration task.

The prototype is used in a user study to validate the hypothesis (section) and learn the usage patterns to stream line the design the proposed approach. Following sections describes the user study and its results.

5.4. User Study

To validate the hypothesis (section 5.1.1) and to learn the usage patterns (research objective 3 of section 1.3), a user study was performed by developing a prototype system to let users explore the content within a video.

5.4.1. Experiment Task

The study is based on a simulated task scenario (Halvey *et al.*, 2014). 29 users (21 males and 8 females) were asked to perform an exploratory search task to explore a video using the prototype (section 5.3). Exploratory search is defined as a complex search task in which the user has to retrieve some facts first, which enable further search queries solving the overall search problem. Often the user is not sure about his/her search goal and sometimes, he/she is not very familiar with the topic of the search (Marchionini, 2006; Waitelonis and Sack, 2012).

5.4.1.1. Study Participants

A total of 29 users (21 males and 8 females) participated in the user study. They all had postgraduate degrees in computer science, digital humanities or related disciplines. Since TED talks are produced for a general audience therefore all the users were chosen not to be from an economics background as all the test videos are on the topic of economics.

5.4.1.2. Participant Instructions

At the beginning of each session the user is given a briefing on the functionality of the developed prototype.

After the introduction to the prototype, user is asked to perform the exploratory search task Figure 18A (1). The user is asked to imagine that they want to get some knowledge about a particular topic e.g. economic inequality.

Users were asked to perform an exploratory search query. For this study, the query was pre-selected to be ``Income inequality in the United States''. The result of their query are TED videos Figure 18A (2).

Users were asked to concentrate only on the top row i.e. the detail representation of one video (Figure 18B). But instead of watching the video, users explored the video using our custom representation (section 5.3). It was left to the user's discretion to choose the combination of segments and tabs they thought sufficient to have the overall synopsis of the video in regard to the query.

Each user performed the query twice. That is, each user explored two TED talks using the prototype one by one.

5.4.2. Test Videos

Because of their general and storytelling nature, TED videos appeal to a wide and diverse audience, and therefore are a good candidate for our research. While our experiment is performed on TED videos, we assume that this approach can be extended to other content types such as life-logging videos, product launches and video messages. Due to its information seeking focus, the proposed approach is not expected to be effective for other video types such as movies, songs or entertainment-oriented videos.

5.4.2.1. Video one: "New thoughts on capital in the twenty-first century" by Thomas Piketty

This TED talk (Piketty, 2014) is approximately 21:00 minutes in length and consists of two parts: the first part consists of a presentation while the second part is an interview. The presenter used slides and charts extensively during the presentation. It is our opinion that the information presented in the video is on average more technical than a typical TED video.

5.4.2.2. Video two: "The rise of the new global super-rich" by Chrystia Freeland

This TED video (Freeland, 2013) is approximately 15:20 minutes in length and it is different from the first video in a number of ways. Firstly, the video solely consists of the presentation and contains no interview, furthermore the presenter does not use any slides or any other visual aid during her presentation. It is our opinion that this video was easier to comprehend than the first one due to the general nature of the information provided.

5.4.3. Feedback Capturing

Following are the types of feedback captured during the user study.

- User interaction (to analyse the user interaction with the representation).
- Verbal feedback (to understand the user's decision making while interacting with the representation).
- User satisfaction questionnaire (to gather quantitative data on the user's experience with the representation).

Following sections provides the details for each of the feedback type.

5.4.3.1. User Interactions with Tabs

Both audio recording and screen capture footage were analysed and annotated manually by the researchers. The following heuristic was employed to record feedback: the more a user selects a particular tab for a segment, the more interested they are in consuming the information using that particular feature rendering. Therefore, each interaction with tabs was noted down, thus when a user chose to view *Terms* of a particular segment, this counted as a user interaction with the representation.

5.4.3.2. Think out loud

A think aloud protocol (Rogers *et al.*, 2012) approach was employed to elicit and analyse user feedback. The following procedure was followed in order to learn exploration patterns in the recorded data. By encouraging users to think out loud we intended to record their thought process in exploring the content of the represented video. We were particularly interested in user comments regarding the effectiveness of different features in terms of efficiency and usability for exploring the video. Feedback such as “the speaker is talking about the 70's here” was not deemed as important as “I find this summary more useful than the last one” or “I find short summaries useful” etc. Therefore, the latter two examples were noted down. Users were encouraged to provide their opinion about the video representation during the post experiment debriefing. They were asked open-ended questions such as what kind of information they would like to see in such content representations.

5.4.3.3. User Satisfaction Questionnaire

After the experiment, users were asked to complete a questionnaire to provide their feedback on the different aspects of the representation. The questionnaire contained 10 questions to be answered on 5-point Likert scale (5 denoting strong agreement). The first three questions regarded ease of use of the prototype. Questions 4 to 6 regarded the segmentation of the videos, and questions 7 to 10 concerned the user's perception of efficiency in viewing the video through the proposed representation compared to watching the video in a conventional manner. The full list of questions can be seen in Appendix A (section 10.1).

5.5. Analysis

This section describes the analysis of the data collected from the user study.

- Analysis of user satisfaction through questionnaire: We have asked users to fill out a questioner after interacting with the system. The response obtained from users is reported and analysed using Kruskawalis test to demonstrate the differences for male and female users.

- Relationship between the usage of different tabs in the representation to identify any pattern that could help in streamlining the design of such systems.
- A subjective analysis of the verbal feedback to get suggestions from participants in order to improve the design.

5.5.1. Analysis of User Satisfaction Questionnaire

There are in total 30 responses from users (21 males and 9 females). The data of males and females is not balanced; therefore, the male data is divided into three folds. The first fold has responses from 9 males, second fold has the other 9 males data and third fold has remaining 3 males plus 3 males from fold1 and 3 from fold2 which were randomly selected. As a result, there are three male folds and one female fold. The mean and standard deviation values of all the male folds including female fold is depicted in Figure 19 and Figure 20, while Table 6 depicts the overall mean and standard deviation for all the participants. The motivation behind in dividing the male data set into folds is to balance it against female data for statistical evaluation. From Figure 19, it is observed that the females score has a higher mean value than male scores except for question one where male-fold2 has a higher mean. Later Kurkawalis test was performed to compare the mean values and found that the difference between male-folds and female fold is not statistically different (p>0.5) except only in one case i.e. question 4 (\$$p_{\text{Female--Male-fold-1}}$=0.04) as depicted in Figure 21.

Figure 19 shows the average score given to each question by participants. As it can be seen that users liked the representation in general, as the lowest average score is 3.5 out of 5. Female participants gave higher scores to the representation compared to their male counterparts, but the difference is not statistically significant (p>0.05, obtained using Kruskal-Wallis analysis as depicted in Figure 21). It can also be seen in Figure 19 that while the overall ease of use and perceived efficiency was scored quite positively by the users, their satisfaction with the segmentation of test videos is lower than the rest (questions 4 to 6).

Table 6: Mean and Standard deviation values of feedback (Questionnaire) of all the participants.

	Q.1	**Q.2**	**Q.3**	**Q.4**	**Q.5**	**Q.6**	**Q.7**	**Q.8**	**Q.9**	**Q.10**
Mean	4.27	4.20	4.57	3.80	3.87	3.83	4.37	4.37	4.40	4.37
SD	0.74	0.48	0.57	0.71	0.97	0.83	0.61	0.61	0.81	0.67

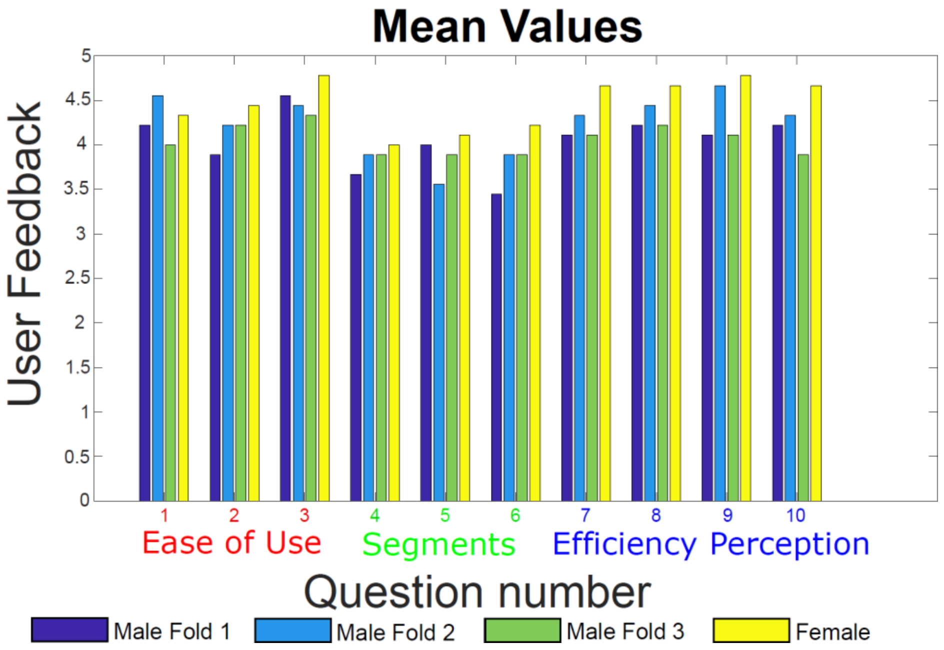


Figure 19: Mean values of feedback by 29 users.

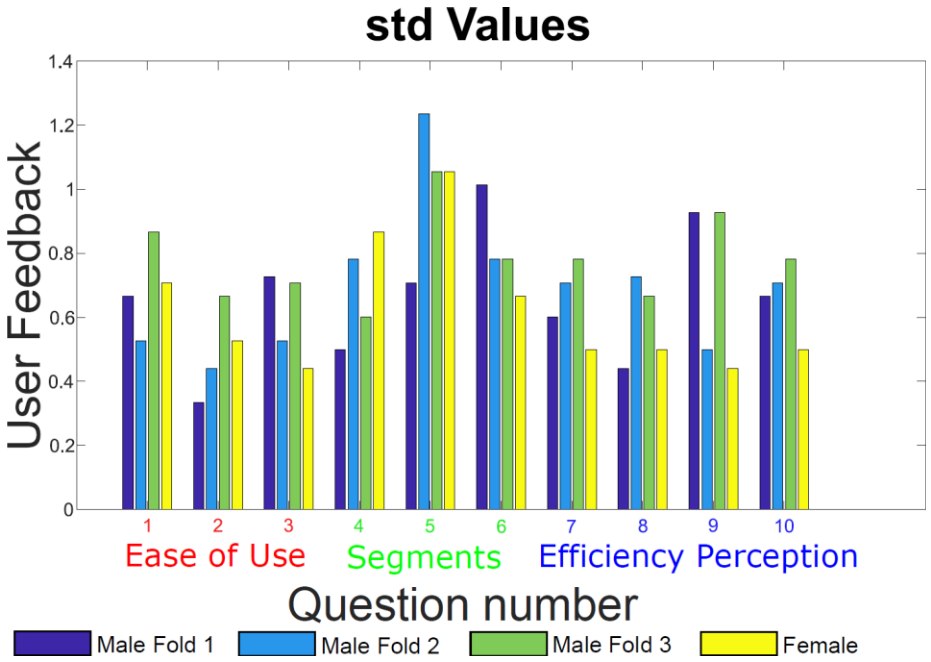


Figure 20: Standard deviation values of feedback by 29 users.

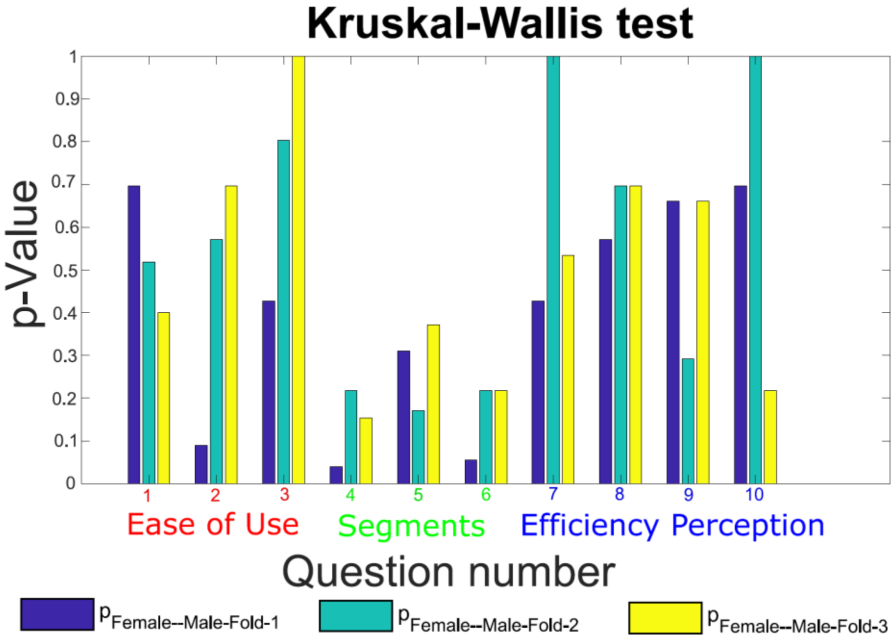


Figure 21: p-values of feedback by 29 users.

The numbers of clicks on the system's tabs for both videos are shown in Figure 22 and Figure 23.

5.5.2. Analysis of User Interactions

Statistical significant test was conducted with a null hypothesis that the number of clicks on each tab is same for both type of videos. As user number one did not explore the second video, we ignored his dataset for this statistical test. In addition, as the second video does not have presentation slides (vis), we also removed the visual tab from this evaluation. As a result, a data set of 28 users and number of clicks on 4 tabs (video, summary, NE, Term) is used for statistical evaluation.

The Kruskal-Wallis test did not reject the null hypothesis in 3 out of 4 cases p_{term}=0.82, p_{NE}=0.68 and p_{Video}=0.44, but rejected it for the summary tab (p_{Sum}=0.04). This is an indication that mean value of user clicks is statistically different for both type of videos but only for the summary tab. One of the possible reason is that the video 1 has presentation slides and video 2 has no presentation slide which make users to use the summary tab more for video 2 than video 1.

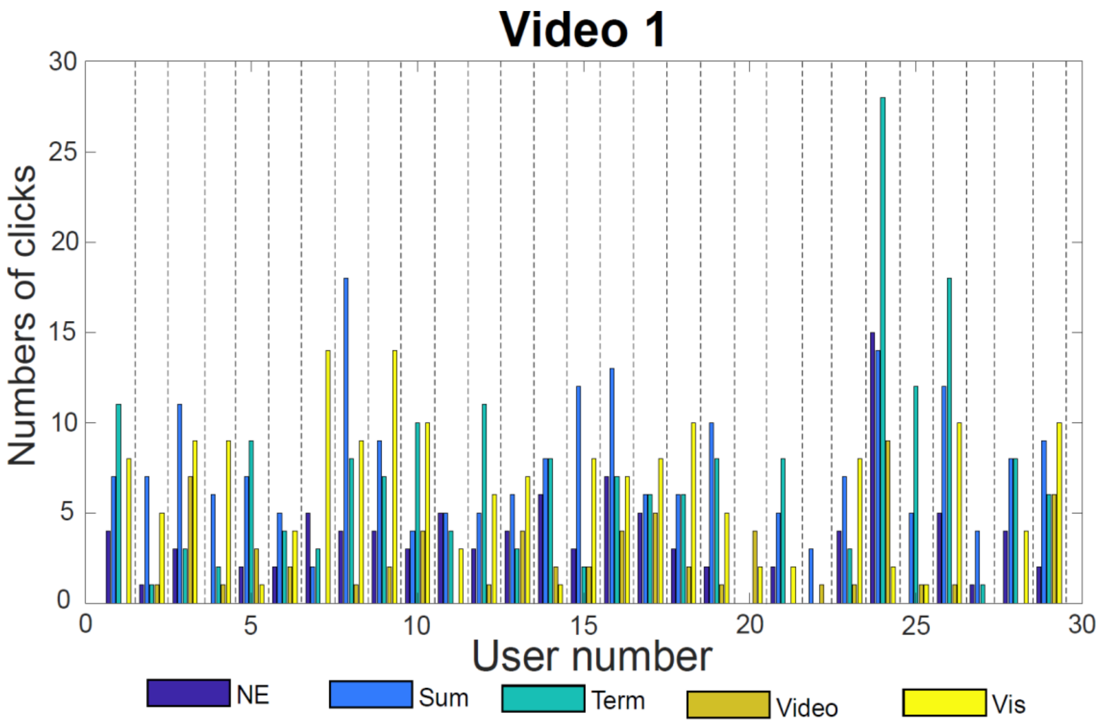


Figure 22: Number of clicks by 29 users on the systems tabs for video 1 Thomas Piketty.

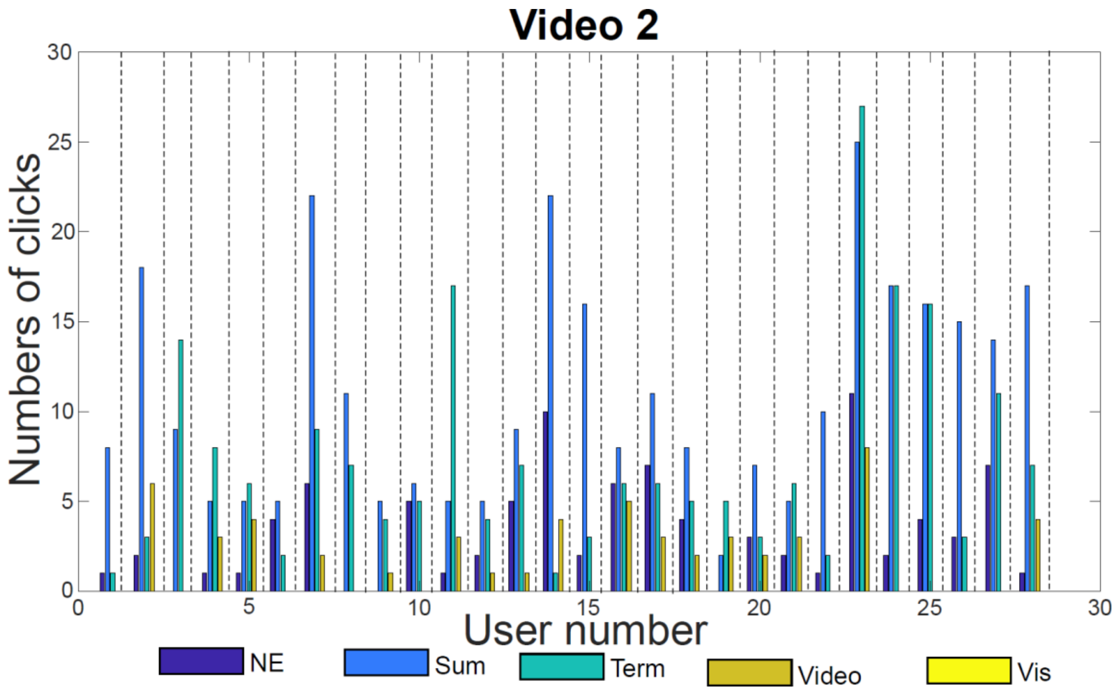
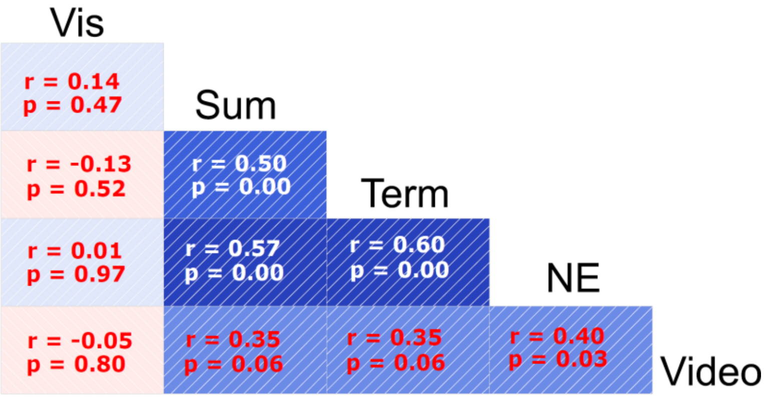


Figure 23: Number of clicks by 28 users on the systems tabs for the second video.

Pearson correlation test was performed on user interactions with tabs (section 5.4.3.1), with a null hypothesis that there is no relationship between the usage of tabs (where usage of tabs is defined as number of clicks on tabs) in the prototype system. Figure 24 shows that the usage of tabs Sum, Term and NE is correlated pairwise, and that these correlations are statistically significant ($p < 0.05$). For example, usage of `Sum tab' is correlated with usage of `NE tab' and this correlation (r=0.57) is statistical significance(p<0.05).

Finally, the tabs were ranked for each user by counting the number of times a user clicked on a tab. Sometimes users have the same number of clicks for two tabs. To rank in this case, we calculated an aggregated response from other users who have a different number of clicks for those tabs. This response is calculated in terms of how many users rank one of two tabs as higher/lower than other tab. In case most users rank a tab *t1* higher than another tab *t2*, then *t1* is assigned a higher rank than *t2*. As there are 5 tabs, and number of possible rank orderings (permutation elements) is 120. Letting rankings be represented as 5-tuples of the form (visual, summary, term, NE, video), the most often chosen ranking is (4, 1, 2, 3, 5), chosen by 6 users. We employed the Mallows-Bradley-Terry (MBT) model to estimate parameters for the ranked data. The computed MBT parameters were (0.11071, 0.46184, 0.26687, 0.09943, 0.06115). Based on these parameters, the estimated order of ranking is (3,1,2,4,5), which means that users prefer mostly the Summary tab, then the Term tab, then Visual tab, then the NE tab and, lastly, the video tab.



	Vis	Sum	Term	NE

Sum	r = 0.14
	p = 0.47			

Term	r = -0.13
	p = 0.52
		r = 0.50
		p = 0.00		

NE	r = 0.01
	p = 0.97
		r = 0.57
		p = 0.00
			r = 0.60
			p = 0.00	

Video	r = -0.05
	p = 0.80
		r = 0.35
		p = 0.06
			r = 0.35
			p = 0.06
				r = 0.40
				p = 0.03

Figure 24:Pearson Correlation between usage of tabs (representation).

5.5.3. Subjective Analysis of Verbal Feedback

Verbal and observational (section 5.4.3.2) feedback of the users was analysed using grounded theory (Rogers *et al.*, 2012). While the details of the use of this method is beyond the scope of the research objectives, following is a brief summarization the results here. The analysis revealed an overall tendency of users to make time versus depth decisions. Different renderings (tabs) (section 5.3) have different capabilities in terms of efficiency and effectiveness. Consider the rendering of a word cloud of key terms. This rendering is very efficient to consume as it can be glanced very quickly to give the user an idea of what a segment might be about, compared to a textual summary which would require more time to read but would give a deeper synopsis of the segment.

5.5.3.1. Time versus Depth Decisions

While watching the video snippet of a particular segment would be the most effective way to consume a segment, it may also be the most time-consuming. Users not only chose certain tabs because of the length of the segment, but they also chose them according to their personal preferences. Some of the users identified themselves as “detail-oriented” in verbal feedback, they opted for tabs that showed more detail, such as the video snippets, while others opted more often for word cloud of *Terms*. This shows the personalization potential of the proposed representation.

5.6. Discussion

Statistical analysis of the results revealed that while users found the representation easy to use and effective in helping them to comprehend the content within a video, they did not always agree with the segmentation performed by the chosen algorithm. They found some segments were either too short or too long. In the verbal feedback, they often expressed their desire for more balanced segments of the video. However, since users had the flexibility of choosing the representation of their choice they rated other aspects higher (see Figure 19).

Analysis of user interaction revealed that most users preferred textual representation compared to visual representation. The MBT model parameters showed that the *Summary* and word clouds of *Terms* were the most chosen representations by the users while watching the video snippet was the least. Analysis of verbal feedback backed this up. Many users reported themselves as fast readers and hence found textual representations more useful. The ranking model (MBT) gives user preference ranking patterns which can be used to further streamline the design of the system. Pearson correlation test results show that the usage of the Visual tab is less correlated with the use of other tabs, while usage of the other four tabs is more closely correlated pairwise (statistically significant ($p<0.05$) 4 out of 6 cases). This can be used in designing a system which offers fewer tabs than the current version because from the correlation results it is observed that the usage of the *Summary* tab is correlated more with the usage of *NE* tab than others. Therefore, in streamlining the system design, the other tabs can be removed in order to provide a less cluttered interface.

5.6.1.1. Choice Factors

The main aim of this study was to learn user interaction patterns with represented features in order to design a better representation approach for users. The assumption was that the user would be interacting more with the highlighted segments in the representation. Even though 18 out of 29 users did most of their exploration in highlighted segments, user also interacted with non-highlighted segments quite often as well. This finding seems consistent with the definition of exploratory search in that user needs evolve during the process (Marchionini, 2006).

While relevance was a factor in choosing one segment over another, non-relevant segments cannot be completely ignored in the representation as users might want to consume them to get the essence of the video. User personal preference played an important role in their choice of features (tabs) for interaction.

In the recorded feedback and also in the post experiment interview, users unanimously reported that they were missing the information regarding the length of each segment versus the length of the whole video. They considered that information as an important factor in their choice of rendering to consume the information. Informing them about the length of a particular segment influenced their choice of tabs for that segment. For example, for a long segment they preferred a rendering such as a word cloud of *Terms* to quickly get the information, while for a shorter segment they might read the summary or watch the video snippet.

Displaying relevant meta-data appears to be a desirable feature for users in an exploration system.

Based on the above discussion we can identify the following factors as a useful guide in designing a system for interactive search within video:

- Relevance (relevance with respect to the test query and also user's personal interest)
- User Preference (in terms of modality and amount of detail)
- Length and other relevant meta-data.

Currently the prototype simply represents all extracted features for each segment to the user to choose from (see Figure 18). The aim is to reduce the number of choices for the user while exploring the video by multimodal representation. Based on the results of the user study some usage patterns have been learned. They are utilized to develop some representation templates for video exploration.

The design of the proposed representation engine is described in section 3. The prototype based on the design and its evaluation as per research objective 5 (section 1.3) is presented in the following chapter.

6. Prototype System for Evaluation in Phase 3

In order to evaluate the proposed system with respect to the research question (section 1.2) a prototype based on the design described in section 3 is needed to represent the content of video to users. Following is a description of the system developed to evaluate the proposed approach.

As per the design mentioned in section 3 the prototype system entails two main phases; extraction and indexing, and representing extracted multimodal features in an interactive manner.

6.1. Extraction and Indexing

6.1.1. Segmentation and Feature Extraction

Segmentation and feature extraction for the prototype system is the same as per experiment 2 (see section 5.3.1 for video segmentation and section 5.3 for multimodal feature extraction) with one exception that “Named Entities” (section 5.3.6) are no longer extracted or represented separately. It is because the information they contain is also present in the Wordcloud of keywords (section 5.3.5).

6.1.2. Indexing

For the representation engine to offer users the extracted features, they (the extracted features) along with their timestamp information need to be stored in an efficiently retrievable manner. To do this all the data related to video segments and the multimodal features along with the timestamp are stored as documents in tables or cores in Solr search platform (Velasco, 2016). Solr is written in JAVA using Lucene indexing and searching engine (McCandless, Hatcher and Gospodnetic, 2010). Following is an example of a video segment indexed by Solr.

```
{
        "id":"ThomasPiketty_2014S-480p_c99_2",
        "video":["ThomasPiketty_2014S-480p"],
        "num":[2],

        "segmenter":["C99"],

        "Start_time":["00:02:42"],

        "End_time":["00:07:58"],

        "seg_text":["\n\tSo there is more going on here, but I'm not going to
talk too much about this today, because I want to focus on wealth inequality.
\nSo let me just show you a very simple indicator about the income inequality
part. \nSo this is the share of total income going to the top 10
percent…………………… "],

        "_version_":1573803391248760832},
```

Similarly, information about multimodal feature is indexed for the representation engine to quickly search it. Following is an example of a multimodal feature indexed in Solr.

```
{

        "id":"ChrystiaFreeland_2013G-480p_c99_1_Keyframes",

        "video":["ChrystiaFreeland_2013G-480p"],

        "segmenter":["C99"],

        "Expanse":["Efficient"],

        "Modality":["Visual"],

        "Name":["Keyframes"],

        "seg_num":[1],

        "FeatureValue":["CF/CF_1_vis.html"],

        "_version_":1579238985017851904}
```

6.2. Representation through template matching

6.2.1. Representation Engine

The representation engine is implemented as a server application developed using ASP.net MVC framework. The server-side work on a request response cycle. The engine receives standard HTTP request for video content in the form of a query and it responds by sending the multimodal data to the requesting application.

The representation engine does that by implementing the pseudo code of section 3.2.7 in C# language.

6.2.2. Determining Relevance

In the prototype, the relevance of a segment is determined solely based on the query request. Once the representation engine receives the request it passes the query to the Solr system which returns segments as relevant, in which the keywords of the request query appears.

6.2.3. Template Matching

Templates are configurations of available multimodal features. In theory, any permutation of available features can be represented (section 3.2.2.1). However, not all permutation may make sense for a representation. Therefore, an implementation may not have some possible templates in the collection. Furthermore, an implementation may only include a few permutations because of application design choices.

Template collection can be implemented in any format depending on the technology used.

For the prototype, the template collection is embedded in the representation engine source code since the design of the user study (section 7) only required a subset of possible permutations of feature set. Following is an example of template embedding in the representation engine source code in C# language.

The feature set entailed within this implementation included the following (see section 5.3 for details):

- Extracted Keyframes from video recording of the segment.
- Text transcript and textual summary generated from the transcript.
- Word Cloud (generated from transcript).
- Video recording of the segment.

This enabled the implementation of template matching as a series of simple if-else statement. The following code snippet shows two of the if-else branches implemented in C# language.

```
            if (videoSeg.Relevant == true && userPreference ==
ModalityName.Textual && hasSummary == true)
            {
                PrimFeature = videoSeg.AvailableFeatures.Where(F => F.Name ==
FeatureName.Text_Summary).FirstOrDefault();


                PrimFeature.RepPlace = RepresentationPlace.Primary;
                segRep.FeatureSet.Add(PrimFeature);
```

……

```
else if (videoSeg.Relevant == false && userPreferce == ModalityName.Visual &&
hasKeyFrames == true)
            {
                PrimFeature = videoSeg.AvailableFeatures.Where(F => F.Name ==
FeatureName.Keyframes).FirstOrDefault();
                PrimFeature.RepPlace = RepresentationPlace.Primary;

                segRep.FeatureSet.Add(PrimFeature);
```

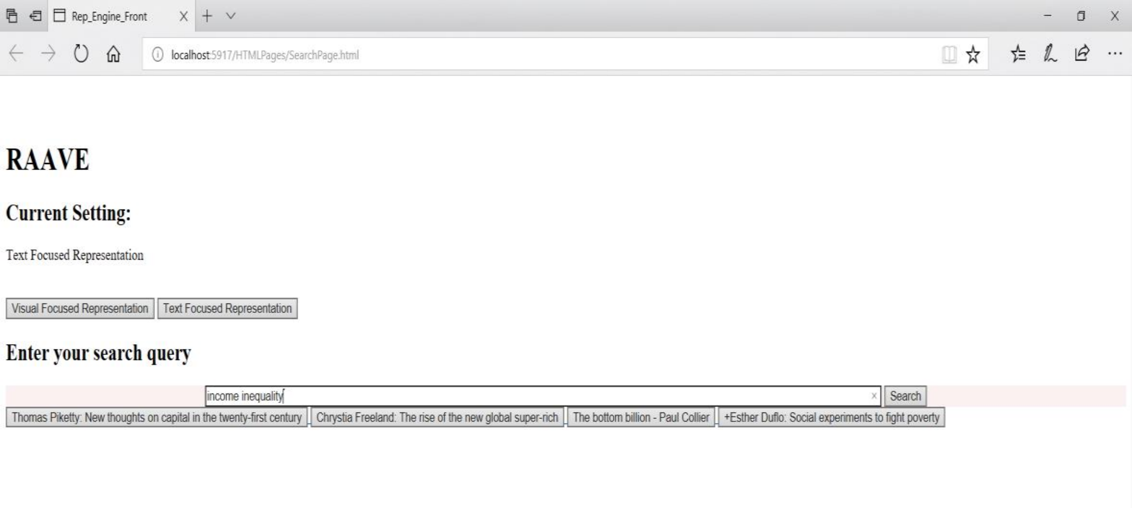
…….

The above code snippet shows the implemented template selection. As per the pseudo code in section 3.2.7, relevant segments are represented with deep features like text summary/transcript or video recording of that segment, depending upon the user preference for modality while non-relevant segments get efficient features such as word cloud or key-frames depending upon the user preference for modality.

The actual placement of the represented features and any other information is dependent upon the user interface. Section 6.3 describes the implemented user interface.

6.3. Video search system

The prototype is built on query-based video search. The prototype is essentially a web application which allows users to search for videos based on their queries. The first webpage is a simple query box where users can enter their query text and search for relevant videos (Figure 25).


Rep_Engine_Front
localhost:5917/HTMLPages/SearchPage.html
Visual Focused Representation
Text Focused Representation
Search
Thomas Piketty: New thoughts on capital in the twenty-first century
Chrystia Freeland: The rise of the new global super-rich
The bottom billion - Paul Collier
+Esther Duflo: Social experiments to fight poverty

RAAVE

Current Setting:

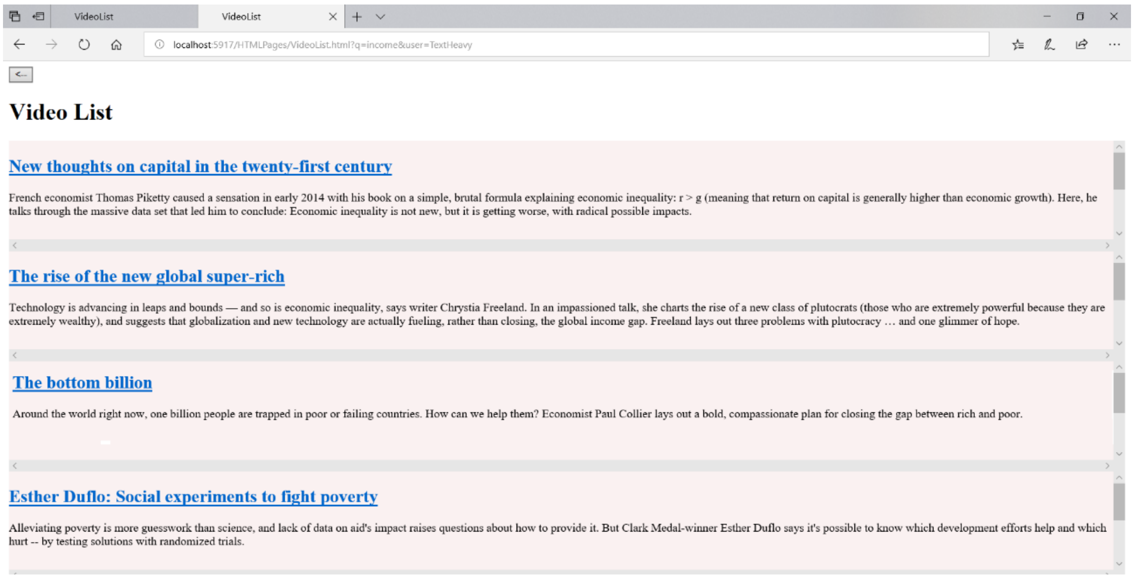
Text Focused Representation

Enter your search query

income inequality

Figure 25: Front Page search box

Once the user enters a query the next page shows the list of videos relevant to the query (Figure 26). Just like any video service it gives the title and a small description of each video. The user can click on the video to explore its content.


VideoList
VideoList
localhost:5917/HTMLPages/VideoList.html?q=income&user=TextHeavy
<--

Video List

New thoughts on capital in the twenty-first century

French economist Thomas Piketty caused a sensation in early 2014 with his book on a simple, brutal formula explaining economic inequality: r > g (meaning that return on capital is generally higher than economic growth). Here, he talks through the massive data set that led him to conclude: Economic inequality is not new, but it is getting worse, with radical possible impacts.

The rise of the new global super-rich

Technology is advancing in leaps and bounds — and so is economic inequality, says writer Chrystia Freeland. In an impassioned talk, she charts the rise of a new class of plutocrats (those who are extremely powerful because they are extremely wealthy), and suggests that globalization and new technology are actually fueling, rather than closing, the global income gap. Freeland lays out three problems with plutocracy … and one glimmer of hope.

The bottom billion

Around the world right now, one billion people are trapped in poor or failing countries. How can we help them? Economist Paul Collier lays out a bold, compassionate plan for closing the gap between rich and poor.

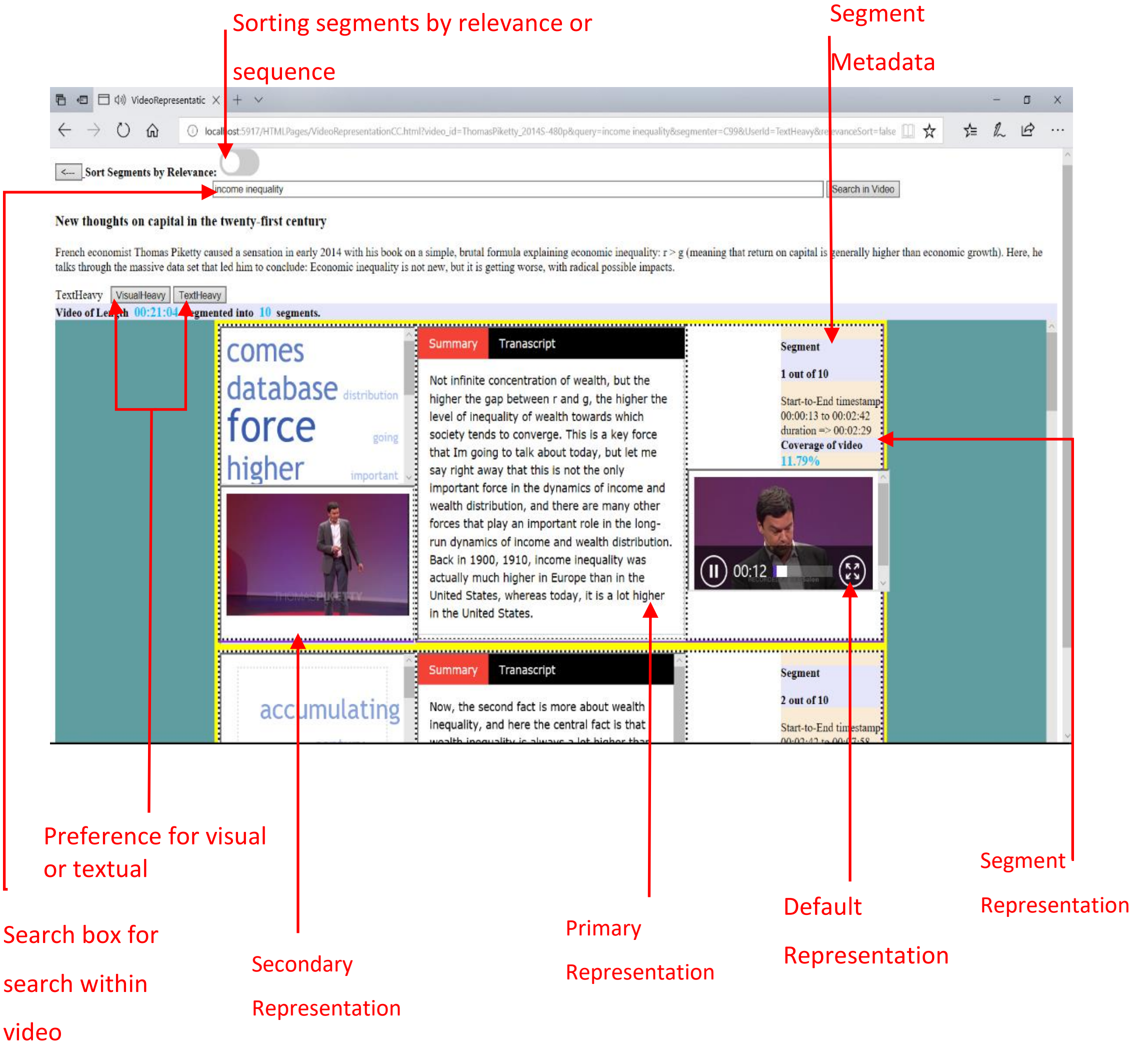
Esther Duflo: Social experiments to fight poverty

Alleviating poverty is more guesswork than science, and lack of data on aid's impact raises questions about how to provide it. But Clark Medal-winner Esther Duflo says it's possible to know which development efforts help and which hurt -- by testing solutions with randomized trials.

Figure 26: List of videos relevant to the query

6.3.1. Video Representation (sample output of RAAVE engine)

Figure 27 shows the automatically generated representation of the content of the video by the RAAVE prototype engine. The user can use the provided search box to search for information within the video. The video is divided into segments and each segment is represented based on the value of the relevance function (section 6.2.2) which currently is a binary value. For relevant segments, the primary representation area shows features with deep expanse while for non-relevant segments efficient segments are placed in the primary representation area. Depending on user preference which can be selected by the preference buttons, primary representation area shows either visual or textual features.



localhost:5917/HTMLPages/VideoRepresentationCC.html?video_id=ThomasPiketty_2014S-480p&query=income inequality&segmenter=C99&UserId=TextHeavy&relevanceSort=false

<--- Sort Segments by Relevance:

income inequality | Search in Video

New thoughts on capital in the twenty-first century

French economist Thomas Piketty caused a sensation in early 2014 with his book on a simple, brutal formula explaining economic inequality: r > g (meaning that return on capital is generally higher than economic growth). Here, he talks through the massive data set that led him to conclude: Economic inequality is not new, but it is getting worse, with radical possible impacts.

TextHeavy | VisualHeavy | TextHeavy

Video of Length 00:21:04 segmented into 10 segments.

Summary | Transcript

Not infinite concentration of wealth, but the higher the gap between r and g, the higher the level of inequality of wealth towards which society tends to converge. This is a key force that Im going to talk about today, but let me say right away that this is not the only important force in the dynamics of income and wealth distribution, and there are many other forces that play an important role in the long-run dynamics of income and wealth distribution. Back in 1900, 1910, income inequality was actually much higher in Europe than in the United States, whereas today, it is a lot higher in the United States.

Segment

1 out of 10

Start-to-End timestamp: 00:00:13 to 00:02:42 duration => 00:02:29

Coverage of video

11.79%

Summary | Transcript

Now, the second fact is more about wealth inequality, and here the central fact is that

Segment

2 out of 10

Start-to-End timestamp:

Figure 27: Video representation by representation engine.

6.4. Template Based Representation (additional use-case examples)

Figure 27 shows a sample output of the implemented representation engine. As explained in section 3 and 6.3, the template driven representation engine works independently of the user interface. However a user interface is needed to for users to explore the content and evaluate the representation engine based approach. Both the implemented representation engine (section 6.1 and 6.2) and the user interface (section 6.3) is designed to be a minimum viable product. The system is implemented specifically for the evaluation performed and reported in section 7. However the capabilities of the template driven approach goes far beyond what has been demonstrated in section 7.

As an example consider that the pseudo code described in section 3.2.7. It can be used in cases where users explore the content on different devices like a smart speaker without an video display. The representation engine in this case can choose a template only with textual modality instead of visual or mix modalities to represent the information which can then be read out by the end interface. Similarly as templates are chosen from a template collection which is independent of the relevance and template matching modules, specifically designed templates can be added based on the target devices used in the application e.g. a template based on textual summary + keyframes for a mobile device and a template based on text summary for a smart speaker. While the templates implemented in section 6.2.3 may be used for exploration on a personal computer.

Moreover the implementation described in 6.1 and 6.2 can be applied in many uses cases in addition to the one shows in section 6.3.1. As an example consider a video which is very different from a TED presentation i.e. video of a football match. The commentary can be used to segment the match and extract multimodal features for indexing (see 6.1 and 6.2).

A user may query "Goal" using the search box (Figure 27) in order to get all the goals in the match i.e. relevant segments of the football match (Oskouie, 2012; Hosseini and Eftekhari-Moghadam, 2013). The representation engine in this case would show the segments with "Goal" as relevant and would use feature-set such as textual summary or video footage to represent the segment based on the preference chosen, while the non-relevant segments would be represented by key-words or keyframes depending on the choice of modality (see Figure 27).

The prototype described here is used to perform the experiments to evaluate the proposed approach. The following chapter describes the experiments in detail.

7. Phase 3: Evaluating the Representation Engine

The question this thesis evaluates is, the effectiveness of multimodal feature representation in video exploration experience. Therefore, the third and final phase of this thesis is about evaluating the representation engine with the help of user studies (point 5 of the research objectives in section 1.3). The prototype developed in section 6 is used for this purpose. This chapter details the experiment and its result.

7.1. Evaluating the effectiveness of representation engine compare to baseline video player.

The goal of this experiment is to evaluate the proposed approach (section 3). The evaluation is done by conducting user studies, utilizing a prototype implementation of the proposed approach (section 6). The experiment is performed as a comparison study to evaluate the performance of the proposed approach named RAAVE against a baseline system which is a standard video player.

7.1.1. Motivation

The goal of the thesis is to evaluate the question (section 1.2) of the extent to which multimodal features can enhance the user experience in video content exploration. As stated in section 1.2 and 2, exploration in a video is not just the ability to search something, but it is also the ability to have an overall synopsis in an efficient manner while providing an engaging and flexible user experience. Therefore, the performance of the representation engine needs to be evaluated not only for both kind of tasks (search and synopsis) but also to evaluate the user's experience while performing those tasks.

7.1.2. Hypotheses

The main hypothesis of the experiment is "RAAVE engine provides an enhanced exploration experience to user compared to a baseline video player by enabling efficient and effective video navigation, synopsis and better engagement".

In order to evaluate the main hypothesis i.e. exploration experience, it is divided in to 3 sub hypotheses as per the discussion in section 1.2 and 2, following lists the sub hypotheses.

- Hypothesis A: RAAVE is better at allowing users to search for information in different parts of a video compared to baseline.
- Hypothesis B: Users can quickly get a better understanding of the content of video using RAAVE compared to the baseline player.
- Hypothesis C: Users have a better experience interacting with RAAVE compared to the baseline player.

7.2. Experiment Design

In order to evaluate the above hypotheses, the experiment is designed as a comparison study between two systems (RAAVE and Baseline). Participants "users" performed two types of tasks; the answer search task for the evaluation of Hypothesis A and the synopsis writing task to evaluate Hypothesis B. After performing the tasks users were asked to give feedback about their experience with both systems. This was done to evaluate Hypothesis C.

This thesis is about evaluating the extent to which multimodal features can enhance the exploration experience of user within video content. In order to enhance the experience this thesis has proposed a template driven representation engine which represents multimodal extracted features in different configurations. The goal of experiment 3 is to test the proposed representation engine with users performing exploration tasks with video content. The proposed approach does not propose a user interface (UI).

The representation engine is designed to be UI agnostic as the end interface may be customized for the end user device and other factors. However, a UI is needed to perform the experiment as users need it to interact with the represented feature set. The experiment is intended to evaluate the user experience with automatically created documents, not a particular user interface and compare it to the baseline video player which only plays video. Therefore, the user interface for the prototype is designed to be minimalist and barebone in order to make sure that users evaluate the approach and not the UI.

Apart from evaluating the hypotheses, another goal of the experiment was to assess the user behaviour while performing different tasks. By user behaviour it is meant the usage patterns of users, with the different modalities and feature set while performing searching for answers for some particular piece of information or trying to get the overall synopsis quickly. For this reason, the experiment is designed to stress test the RAAVE system.

7.2.1. Experiment Systems

7.2.1.1. RAAVE System

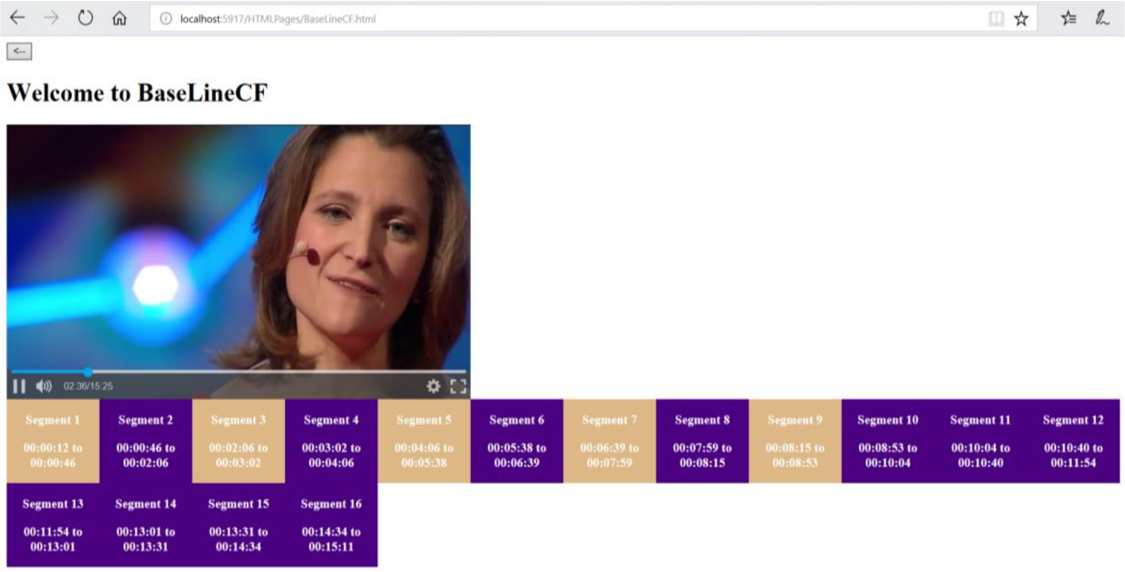
RAAVE system is the system based on the proposed approach described in section 3. The details of the system are described in section 6.

7.2.1.2. Baseline System

To evaluate the hypothesis (section 7.1.2) RAAVE is compared with a baseline system which in this case is a simple video player. Admittedly the choice of using a simple video player as a baseline is an unusual one. Traditionally researchers compare their approach with an approach from the state of the art. However it was not feasible for the current experiment. The proposed approach (RAAVE) automatically transforms video into a multimedia document i.e. RAAVE transforms video into something more than a video. State of the Art (SOTA) approaches that do that are Hypervideo based approaches (section 2.2.1.1). As detailed in the SOTA review (section 2.2.1.1 and 2.2.4), Hypervideo systems require human curation by utilizing specialized authoring environments and video players. Since RAAVE utilizes automatically extracted multimodal features therefore it is not feasible to have a direct comparison with Hypervideo systems due to this fundamental difference in the approaches. Moreover while Hypervideo systems do represent multimodal information with video content, they utilize supplementary content from sources other than the source video, while the goal of the RAAVE approach is to utilize the content within the source video in novel ways to enhance the user experience thereby making a direct comparison not feasible due to the fundamental difference in the approaches.

Another reason to choose a simple video player as a baseline goes to the main goal of the presented thesis. As described in section 1 and 2 despite all the research; the current view on video content is essentially a continuous stream of images with or without an audio component. As RAAVE proposes to consider video content as a diverse content source by transforming it automatically, it is natural to compare the transformation based approach with the continuous stream of images with a parallel audio component i.e. a regular video player as a baseline which is also ubiquitous due its familiarity and common usage.

Figure 28 shows the baseline system which users utilize to interact with the video. It contains a standard video player with pause/play button and a scrubber so that user can drag it across to watch different portion of a video. Underneath the video player is information regarding the start and stop time of different segments of the video. It is provided to aid the user in performing the answer search task.



Welcome to BaseLineCF

Segment 1 00:00:12 to 00:00:46

Segment 2 00:00:46 to 00:02:06

Segment 3 00:02:06 to 00:03:02

Segment 4 00:03:02 to 00:04:06

Segment 5 00:04:06 to 00:05:38

Segment 6 00:05:38 to 00:06:39

Segment 7 00:06:39 to 00:07:59

Segment 8 00:07:59 to 00:08:15

Segment 9 00:08:15 to 00:08:53

Segment 10 00:08:53 to 00:10:04

Segment 11 00:10:04 to 00:10:40

Segment 12 00:10:40 to 00:11:54

Segment 13 00:11:54 to 00:13:01

Segment 14 00:13:01 to 00:13:31

Segment 15 00:13:31 to 00:14:34

Segment 16 00:14:34 to 00:15:11

Figure 28: Baseline video player and information regarding the start and end times of segments of the video.

Following are the two tasks performed by users in the experiments.

7.2.2. Answer Search Task

The answer search task is designed to evaluate the ability to find information within video. Users are given a set of questions which have to be answered by utilizing the content of the video as quickly as possible.

The answer search task was performed by 12 users. They performed it on all 4 videos for the total of 24 times. Each user performed the task twice, once using the proposed system and once the baseline system. The order of the system and the video was always changed i.e. some user performed the task using RAAVE first and baseline the second time while others did it vice versa. Following is an example of questions for video (Piketty, 2014). The full list of questions can be seen in Appendix D (section 10.4).

1. Jane Austen is mentioned in.

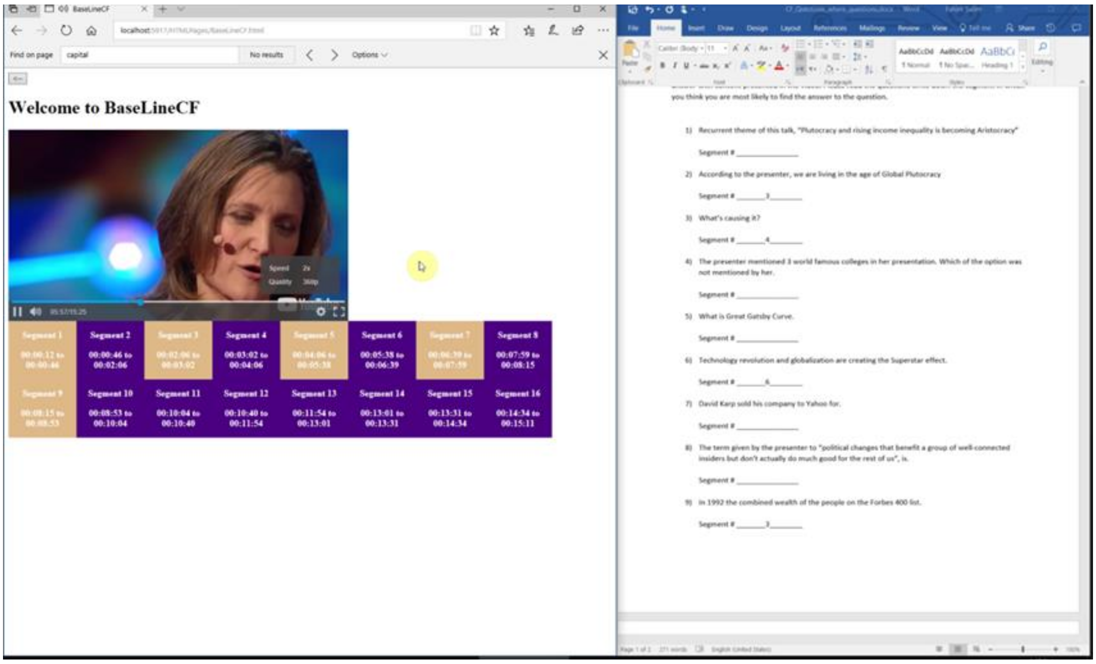
   Segment # ________________

2. What made the swiss show flexibility in bank secrecy?

   Segment # ________________

For each video, there are 14 questions. Figure 29 shows a screenshot of answer search task using baseline system.

Since hypothesis A is about comparing the ability to search for information within video, therefore in this task, instead of providing the actual answer to the question users where asked to provide the segment number which contains the information needed to answer the question that is identify the portion of video in which they think contains the relevant information.



Welcome to BaseLineCF

Segment 1 00:00:12 to 00:00:46	Segment 2 00:00:46 to 00:02:06	Segment 3 00:02:06 to 00:03:02	Segment 4 00:03:02 to 00:04:06	Segment 5 00:04:06 to 00:05:38	Segment 6 00:05:38 to 00:06:39	Segment 7 00:06:39 to 00:07:59	Segment 8 00:07:59 to 00:08:15
Segment 9 00:08:15 to 00:08:53	Segment 10 00:08:53 to 00:10:04	Segment 11 00:10:04 to 00:10:40	Segment 12 00:10:40 to 00:11:54	Segment 13 00:11:54 to 00:13:01	Segment 14 00:13:01 to 00:13:31	Segment 15 00:13:31 to 00:14:34	Segment 16 00:14:34 to 00:15:11

you think you are most likely to find the answer to the question.

1) Recurrent theme of this talk, "Plutocracy and rising income inequality is becoming Aristocracy"

Segment # ______________

2) According to the presenter, we are living in the age of Global Plutocracy

Segment # _______3_______

3) What's causing it?

Segment # _______4_______

4) The presenter mentioned 3 world famous colleges in her presentation. Which of the option was not mentioned by her.

Segment # ______________

5) What is Great Gatsby Curve.

Segment # ______________

6) Technology revolution and globalization are creating the Superstar effect.

Segment # _______6_______

7) David Karp sold his company to Yahoo for.

Segment # ______________

8) The term given by the presenter to "political changes that benefit a group of well-connected insiders but don't actually do much good for the rest of us", is.

Segment # ______________

9) In 1992 the combined wealth of the people on the Forbes 400 list.

Segment # _______3_______

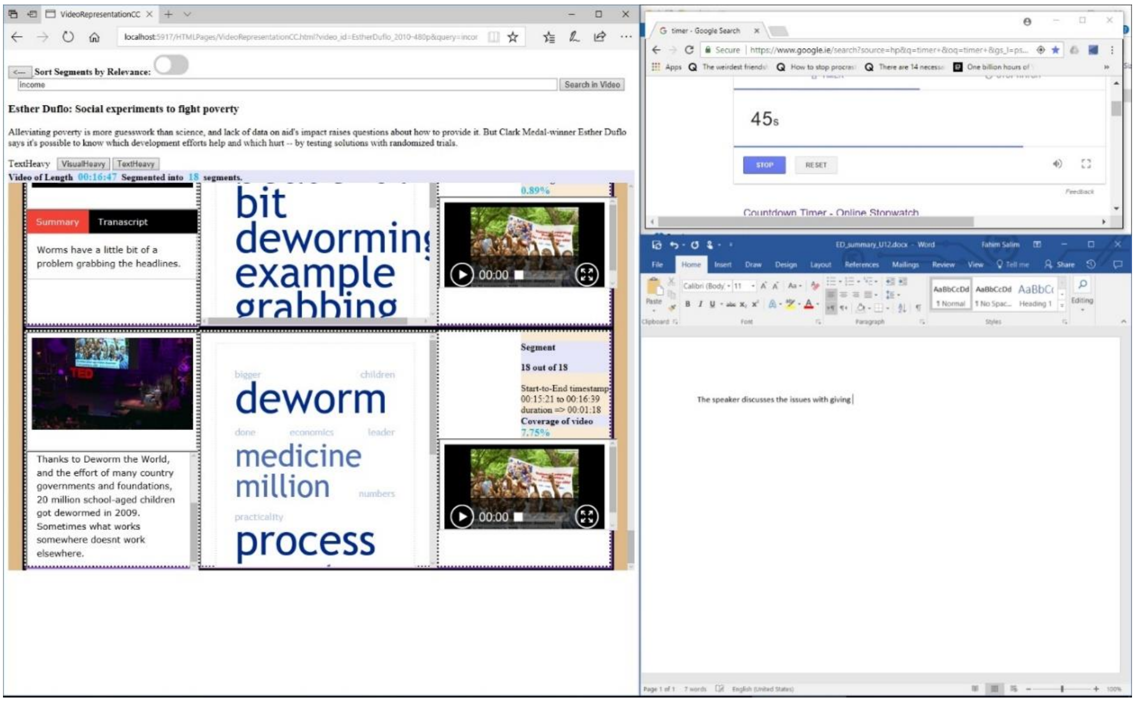
Figure 29: Answer search task using baseline system (screen shot).

7.2.3. Synopsis writing task

The second task corresponds to the goal of enabling the user to get the essence of the video effectively. This task is designed as a comparison study to evaluate user's ability to get the essence of video effectively that is to evaluate Hypothesis B. Each of the 12 participants performed the synopsis writing task twice, once using the RAAVE system and once the baseline player. Figure 30 shows an example of synopsis writing task using the RAAVE system.

In this task, participants consume the content of video for a shorter amount of time compared to the length of the video and write a synopsis of the video. Throughout the video summarization literature (section 2.2.2) researchers have used the ratio 0.2 to test their video summary generation approach.

Therefore, this experiment also uses 0.2 as ratio for the amount of time given to user to consume the content of video so that they can write a synopsis. For example, for a video of 15 minutes user could consume the content for up to 3 minutes. Note that this is not the amount of time to write the synopsis but to consume the content of the video, users were allowed to take as much time they wanted to write the synopsis. It was left on the user's discretion if they wanted to start writing the synopsis or take notes while they were consuming the content and continue writing the synopsis after the allowed time passed or they consume the content first and write the synopsis later.



Esther Duflo: Social experiments to fight poverty

Alleviating poverty is more guesswork than science, and lack of data on aid's impact raises questions about how to provide it. But Clark Medal-winner Esther Duflo says it's possible to know which development efforts help and which hurt -- by testing solutions with randomized trials.

Video of Length 00:16:47 Segmented into 18 segments.

Worms have a little bit of a problem grabbing the headlines.

Thanks to Deworm the World, and the effort of many country governments and foundations, 20 million school-aged children got dewormed in 2009. Sometimes what works somewhere doesnt work elsewhere.

Segment

18 out of 18

Start-to-End timestamp
00:15:21 to 00:16:39
duration => 00:01:18
Coverage of video
7.75%

The speaker discusses the issues with giving

Figure 30: Synopsis writing task using RAAVE system (screen shot).

To compare user performance with the two systems, another set of participants "reviewers" were asked to evaluate the synopsis produced by the users (details in section 7.3.2).

Hence the experiment has two systems, two types of tasks and two types of participants. Table 7 summarizes the configuration.

Table 7: Experiment Items

Systems	Participants	Tasks
RAAVE	Users	Answer Search
Baseline	Reviewers	Synopsis writing

To summarize: participants (users) use two systems (RAAVE and Baseline) so that their performance could be compared for the two tasks (Answer search task for Hypothesis A and Synopsis writing for Hypothesis B). In order to compare their performance users were asked to perform both tasks twice. Each individual user session had 4 attempts. Table 8 lists the attempts for two users an as example to make things clear.

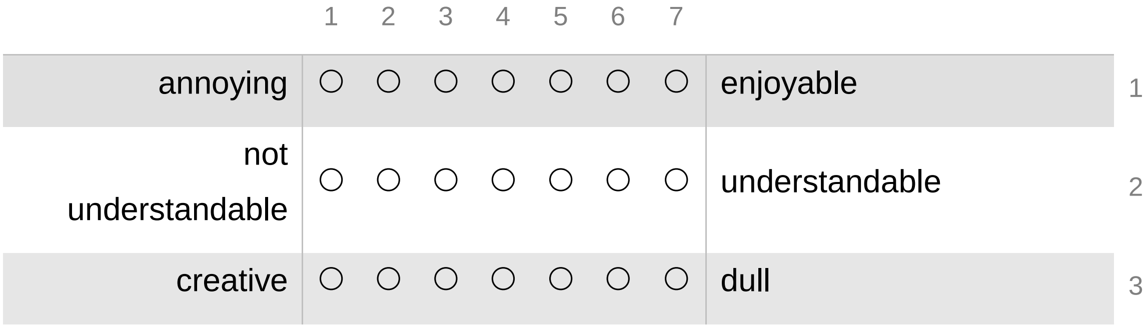
Table 8: User attempts

User	Attempt 1	Attempt 2	Attempt 3	Attempt 4
1	RAAVE system to perform Answer search using (video 1)	Baseline system to perform synopsis writing (video 2)	Baseline system to perform Answer Search (video 3)	RAAVE to perform Synopsis writing. (video 4)
2	Baseline system to perform Synopsis writing (video 4)	RAAVE system to perform Answer search using (video 3)	Baseline system to perform Answer Search (video 2)	RAAVE to perform Synopsis writing. (video 1)
3	..	..	..	..
n	..	…	..	..

7.2.4. User Experience Questionnaire

In order to evaluate Hypothesis C i.e. comparison of user experience with the two systems, after performing the experiment tasks, users were asked to fill the user experience questionnaire (Laugwitz, Held and Schrepp, 2008). The user experience questionnaire (UEQ) is designed to compare the user experience with two systems. Users are asked to fill a questionnaire consisting of 26 questions. Each question consists of a pair of contrasting attributes and 7-point Likert scale between them. Table 9 shows 3 of the 26 questions in UEQ. (see Appendix E (section 10.3) for the full list).

Table 9: Example questions UEQ first 3 of 26.



	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3

The 26 questions of the UEQ can be categorized in to following categories.

- Attractiveness
- Perspicuity
- Efficiency
- Dependability
- Stimulation
- Novelty

User filled the UEQ twice, once for RAAVE and once for the baseline.

Since each user perform 4 attempts, the experiment uses 4 TED talks. Following section describes the test videos.

7.2.5. Test Videos

For the user study, a total of 4 TED videos were utilized. The number was chosen to insure that each user explores a different video for each of the experiment attempts (see Table 8). As described in section 4.1.2, TED videos are chosen due to their general purpose nature and appeal to a wider audience therefore it made the selection of experiment participants simpler as any person familiar with informational style videos such TED was qualified (see 7.2.6). All sample videos belong to the same topic i.e. "Economic Inequality". This topic was chosen due the fact for the fact that none of the study participants had an educational background in economics. It is to ensure consistency in the experiment tasks (see section 7.2.2 and 7.2.3) (Hong *et al.*, 2011; Halvey *et al.*, 2014).

While TED presentation videos have a consistent structure (Scotto di Carlo, 2014) there can be slight variations. For example some presenters use visual aids while other prefer to talk without any slides. Similarly some presentation ends with a supplementary item such as an interview etc. while others only consists of a presentation. The sample video were chosen to be representative of the general TED videos collections. Two of the TED videos used, contain visual aid i.e. the presenters use slides and pictures in their presentation (Duflo, 2010; Piketty, 2014) while the other two presenters do not use any visual aid (Collier, 2008; Freeland, 2013).

All videos are of the same subject area which is Economic Inequality. The reason for choosing two videos with slides and the other two without slides follows the idea of content value of different modalities (section 1.1). For example in the video *ED* (Duflo, 2010) , the presenter speaks about mumps causing deaths in NY. She never tells the listener about the total number but a slide in her presentation shows the number. Now a question related to this can only be answered by utilizing the visual modality. Listening to audio only or running a text search on the transcript of the presentation would not yield the answer. In the researcher's opinion, videos without visual aid were relatively easier to comprehend than the other two.

In TED presentations, the presenter often presents the main idea of the presentation in the beginning of the video as in Freeland's presentation (Freeland, 2013), the presenter gives the main idea of the presentation early on and gives some details to reiterate it. This makes it easier for users to get the overall synopsis because users can still get the main idea of the video even if they did not consume all the portions. Whereas Collier and Duflo (Collier, 2008; Duflo, 2010) describe a problem and then offers a solution later and summarize the discussion in the end. Therefore, users are more likely to miss the essence of the presentation if they do not consume all the portions compared to the first video. The 4th video (Piketty, 2014) is an interesting case. While the presenter does give the main idea in the beginning, the technical nature of it makes it difficult for viewers to fully get the point, especially if the viewer does not have an economics background, it is only by watching the middle and the end parts of it that a viewer can get the essence, making it a relatively difficult video to comprehend.

By choosing videos with and without visual aid and easy and difficult videos, the exploration experience of users with different type of content can be evaluated.

7.2.5.1. "New thoughts on capital in the twenty-first century" by Thomas Piketty and "The rise of the new global super-rich" by Chrystia Freeland

These two videos are the ones which are used in phase 2 (section 5) and are previously described in section 5.4.2.1 and section 5.4.2.2. Throughout this chapter these videos would be referred with their identifier *TP* for Thomas Piketty and *CF* for Chrystia Freeland.

7.2.5.2. “Social experiments to fight poverty” by Esther Duflo

This TED video (Duflo, 2010) is approximately 16:40 minutes in length. The presenter uses slides and charts extensively during the presentation. It is the researcher’s opinion that this video while not as technical in nature as (Piketty, 2014) does contain a lot of information. Throughout this chapter this video would be referred as *ED*.

7.2.5.3. “The bottom billion” by Paul Collier

This TED video (Collier, 2008) is approximately 16:51 minutes in length. The presenter does not use any visual aid during the presentation. In a manner similar to (Freeland, 2013) this presentation, in the researcher’s opinion does not contain too much technical information and is easy to comprehend for a general audience. Throughout this chapter this video would be referred as *PC*.

Table 10 summarizes the information regarding the test videos.

Table 10: Information about test videos.

Vid. #	Title	Identifier	Slides
1	“The rise of the new global super-rich” by Chrystia Freeland	CF	No
2	“Social experiments to fight poverty” by Esther Duflo	ED	Yes
3	“The bottom billion” by Paul Collier	PC	No
4	“New thoughts on capital in the twenty-first century” by Thomas Piketty	TP	Yes

7.2.6. Experiment Participants

The experiment has two types of participants.

- Users
- Reviewers

7.2.6.1. Users

A total of 12 users performed the two tasks in the experiment, 7 males and 5 females. A sample size of 12 users seems adequate compared to studies on this subject (Meixner *et al.*, 2014;

Gravier *et al.*, 2016). They all have postgraduate degrees in computer science, digital humanities or related disciplines. Since TED talks are produced for a general audience therefore all the users were chosen not to be from an economics background as all the test videos are on the topic of economics. Admittedly participants are a rather cohesive group in a sense that they are all of an academic background. However they are an adequate representation of TED audience which is described as highly educated[11] with an interest in scientific and intellectual pursuits (Sugimoto *et al.*, 2013; Scotto di Carlo, 2014).

7.2.6.2. Reviewers

A total of 7 reviewers participated in judging the summaries created by users. In a similar manner to the users; reviewers were also chosen not to be from an economics background and had postgraduate degrees in computer science, linguistics or related disciplines.

7.3. Feedback Capturing, Annotations and Data for Analysis

7.3.1. Screen Capturing

User actions and their interaction with the representation was recorded via screen capturing and audio recording.

7.3.2. Summary Evaluation Task (Performed by Reviewers)

In synopsis writing task users produced a total of 24 summaries 12 of them were created by using the RAAVE system while the rest were created using the baseline video player. Since 4 TED talks were used in the experiment there are 6 summaries created for each video.

There are two types of techniques to compare summaries, researchers often use automatic tools such as ROUGE (Lin, 2004). The other technique is to use human evaluators. Bayomi et al. observed that automatic techniques fall short in evaluating certain factors of summary quality compared to humans (Bayomi, Levacher and Ghorab, 2016). Therefore in order to evaluate the user produced synopses, current experiment used a similar approach (Bayomi, Levacher and Ghorab, 2016).

[11] https://www.ted.com/about/our-organization/how-ted-works/debunking-ted-myths; last verified: 1-02-2019

All 7 reviewers evaluated the summaries of all 4 TED talks. For each TED talk, they were provided with the video and the 6 summaries created of that TED talk. They were asked to do the following:

1) Watch the TED talk.
2) Evaluate each summary individually according the characteristics listed in Table 11.
3) Rank the summaries in order of preferences (1 and 6).

Table 11: Summary evaluation criteria

Readability and Understandability: Whether the grammar and the spelling of the summary are correct and appropriate						
Extremely Bad -	1	2	3	4	5 -	Excellent
Informativeness: How much information from the source video is preserved in the summary.						
Extremely Bad -	1	2	3	4	5 -	Excellent
Conciseness: As a summary presents a short text, conciseness means to assess if this summary contains any unnecessary or redundant information.						
Extremely Bad -	1	2	3	4	5 -	Excellent
Overall: The overall quality of the summary.						
Extremely Bad -	1	2	3	4	5 -	Excellent

Reviewers were asked to assign a rank between 1 to 6 to the synopsis of each video with the most preferred summary being 1st and the least preferred summary being 6th.

7.3.3. Answers and Durations

For the answer search task, the following items were recorded for each user:

- The number of questions attempted.
- The number of questions correctly answered.
- Time taken to complete the task.
- Duration per correct answer (by normalization).

7.3.3.1. Normalizing duration per correct answer

To assess the efficiency of answer searching, measuring average time to find an answer would not be appropriate since the videos are of different length. To normalize that; percentage of video length is used. For example, if average if a user took 5 minutes to search for answers for a 20 minutes video, it is considered as 25% of length of video was required by user to find all the answers. Dividing it by the number of correct answers gave the length of video required per correct answer.

7.3.4. User Interactivity with the two systems

7.3.4.1. Baseline player Logs

For the baseline video player , a log of user interactions with the video player was recorded using the SocialSkip system (Chorianopoulos, 2011). SocialSkip logs the standard interaction such as play, pause and seek etc. along with timestamps and other relevant meta-data on a server which can be downloaded as a csv file. Appendix F (section 10.6) shows an example of the log.

7.3.4.2. Annotation of user interactions

For the analysis, user interactions were annotated manually from video recordings. Table 12 shows an example for user# 12 interacting with the baseline system while performing the question search task. Top row shows the minute of the experiment session. The left column shows user's interaction with the system. In the example user is increasing the video play speed to 1.5x. The right column shows user's action for the task at hand. In the example user spent the minute reading the questions and wrote the answer for a question.

Table 12: Annotation example for user interactions using baseline system.

Minute: 26-27	
Action System	Action task
Play speed 1.5x at 26:11	Reading questions Ans. Q.13 at 26:21

Table 13 shows the example for user #12 interacting with RAAVE system. In the example user interacts with the wordcloud, summary, and transcript of segment no 4,6 and 7. User also wrote the answer of Q.12 in that minute. Appendix E (Section 10.5) shows the full annotation for user# 12 with both RAAVE and Baseline system as an example.

Table 13: Annotation example for user interaction using RAAVE system

Minute: 12-13	
Action System	Action task

Scrolling Seg 4 word cloud sci Seg 6 sum (sci) Seg 6 trans Seg 7 trns (sci) scolling	Ans. Q.12 at 12:03

7.4. Expected Outcomes

- Evaluation of the set of hypotheses (section 7.1.2).
- Usage patterns with both systems.
- Differences in user interactions with the two systems while performing the tasks.
- Are there particular feature representations offered by RAAVE which were more useful than others?

7.5. Experiment Results

7.5.1. Results Hypothesis A (Answer Search Task)

Hypothesis A: RAAVE is better at allowing users to search for information in different parts of a video compared to baseline.

By better it is meant the following

- Users were able answer question more accurately with RAAVE compared to Baseline.
- Users were able search the answers efficiently with RAAVE compared to Baseline

Table 14, Table 15 and Table 16 shows the results of answer search task. Table 14 shows the overall performance of users using both systems, while Table 15 shows the detailed comparison of each user's performance using both systems.

Table 16 reports results based on interaction strategies employed by users.

Overall results in Table 14 shows that in terms of answering correctly users did better using the baseline system 9.5 correct answers compared to 9.16 using RAAVE. However, for the two videos containing visual aid (ED+TP) users were able to perform better using RAAVE i.e. on average user performed better with RAAVE for difficult videos whereas their performance was better using the baseline system for easier videos. (see section 5.4.2 and 7.2.5 for discussion about easy and difficult videos).

In terms of efficiency users seems to perform better using the RAAVE system, as the normalized duration per correct question is lower for RAAVE system compared to Baseline. The duration per question is calculated by measuring the normalized duration (section 7.3.3.1) spent by user performing the task, divided by correct answers (lower is better).

Table 14: answer search task results (overall performance)

Row#	Videos	RAAVE		Baseline	
		Correct	**Per Question**	**Correct**	**Per Question**
1	Overall Avg. (avg. corr. / avg. dur)	9.16	9.57	9.50	10.23
2	Average all users	9.17	10.04	9.50	10.75
3	Median all users	10	8.80	10	10.60
4	CF+PC (Avg.) (no slides)	10	8.12	11.16	8.46
5	ED+TP (Avg.) (with slides)	8.33	11.21	7.83	12.75

Table 15 shows the individual performance of each user while performing answer search task. It also lists the information about how each user interacted with the systems. For the baseline system the number of interaction (play, pause, seek) is listed in “# Interac” column. For the RAAVE system, user’s interactions with feature set and modalities is presented in “Mod.” Column while “# qur.” List the number search in video queries executed by each user. The number of attempted questions (Att.), number of correct answers (Corr.) is also listed for each user. The column “Time” shows the time taken by each user to complete the task while “% of vid” shows normalized duration of task and “per ques %” columns shows the normalized duration per correct answer.

It can be seen in Table 15 that while on average user answer more correctly using the baseline system (9.50 using baseline vs. 9.17 using RAAVE) in terms of efficiency user spent less time per correct answer using RAAVE compared to the baseline (10.04 using RAAVE vs. 10.75 using Baseline (lower is better)). The Student T-Test p-value of the result is 0.27 which means the results are not statistically significant.

Table 15: User performance in answer search task. “#Interac” column list the number of interaction performed by the user with the baseline player. #qur. Column list the number of queries executed by user in RAAVE. “Att.” Refers to the number of question attempted while “Corr.” lists the number of correct answers provided by each user. “% of Video” refers to the time it took to complete the task with respect to the length of video. “per ques %” refers to the percentage of video needed to consume per correct answer (lower is better).

Answer Search Task															
	Baseline							RAAVE							
User	Vid.	# Interac	Att.	Corr.	Time	% of vid.	per ques. %	Vid.	Mod.	# qur.	Att.	Corr.	Time	% of Vid.	per ques. %
1	ED	28	10	7	14:01	83.52	11.93	PC	ST	12	10	7	10:35	62.93	8.99
2	CF	92	12	10	16:20	106.06	10.61	TP	STV	25	12	8	17:57	85.21	10.65
3	ED	195	12	9	13:44	81.83	9.09	PC	ST	24	14	10	7:50	50.87	5.09
4	TP	8	13	10	24:25	115.90	11.59	ED	STWVP	26	11	6	20:02	119.36	19.89
5	CF	31	13	12	11:42	75.97	6.33	TP	STWVP	17	12	11	24:20	115.51	10.50
6	PC	2	12	11	17:30	104.06	9.46	CF	ST	6	12	12	12:15	79.55	6.63
7	ED	167	12	7	12:28	74.28	10.61	PC	STWV	19	13	10	14:05	83.75	8.37
8	CF	8	12	10	16:19	105.95	10.60	TP	STWVP	12	10	8	14:31	68.91	8.61
9	ED	48	10	6	15:06	89.97	15.00	PC	STW	10	12	10	21:08	125.67	12.57
10	TP	26	14	8	32:22	153.64	19.20	ED	STWV	9	7	7	15:53	94.64	13.52
11	CF	4	12	12	16:12	105.19	8.77	TP	STW	8	11	10	16:06	76.42	7.64
12	PC	10	14	12	11:39	69.28	5.77	CF	STW	0	14	11	13:40	88.74	8.07
	Avg.	51.58	12.17	9.50		97.14	10.75		Avg.	14.00	11.50	9.17		87.63	10.04
	Med.	27.00	12.00	10.00		97.02	10.60		Med.	12.00	12.00	10.00		84.48	8.80

Table 16:User performance based on interaction techniques and the nature of videos.

Row#	System	Settings	# of Users	Avg. Qur/interac	Attempted	Correct	% of Video	Per Ques. %
1	RAAVE	Text+Vid	6	18.00	10.83	8.33	94.56	11.93
2	RAAVE	Text	6	10.00	12.17	10.00	80.70	8.16
3	RAAVE	Less queries	7	8.14	10.86	9.29	85.27	9.43
4	RAAVE	More queries	5	22.20	12.40	9.00	90.94	10.90
5	RAAVE	TP+ED	6	16.17	10.50	8.33	93.34	11.80
6	Baseline	TP+ED	6	78.67	11.83	7.83	99.86	12.90
7	RAAVE	CF+PC	6	11.83	12.50	10.00	81.92	8.29
8	Baseline	CF+PC	6	24.50	12.50	11.17	94.42	8.59

While Table 15 showed details about individual use performance, Table 16 summarizes the results. Row 1 to 4 shows performance of users while applying different strategies. Users who only used textual features performed better than those who used both textual and visual (Row 1 and 2 of Table 16). The reason for this can be that users who interacted with multiple modalities and feature set got distracted with the abundance of options while other users focused their efforts on just the textual features. Users who executed less search in videos queries performed better than those who performed more queries. (Rows 3 and 4 of Table 16). It could be because the questions were spreads across different segments. Users who employed a mixture search and consumption had a better idea of the narrative of the video which helped them quickly find the right segment quicker than those who relied more on the search box.

7.5.2. Results Hypothesis B (Synopsis Writing Task)

Hypothesis B: Users can quickly get a better understanding of the content of video using RAAVE compared to the baseline player.

Since the time allowed to users to consume the content was fix at 20% of the video length. The comparison of user performance for synopsis writing task is done as:

- Comparison of reviewer ratings to the synopsis produced by using both systems.
- The likelihood of a synopsis produced by RAAVE be given top rank by reviewers is higher compared to the Baseline.

Table 17 shows the average score for each characteristic. The first row shows the overall scores i.e. the average of scores against all the summaries of 4 videos. The next four rows show the average score of the summaries of an individual video. Last two rows combine the average score of videos with and without visual aid. “CF+PC” shows the average score of (Collier, 2008; Freeland, 2013) since these two videos do not contain any visual aid and “ED+TP” shows the average score of (Duflo, 2010; Piketty, 2014).

Table 17: Average of score given by 7 reviewers

	Readability & Understandability		**Informativeness**		**Conciseness**		**Overall**	
	RAAVE	**Baseline**	**RAAVE**	**Baseline**	**RAAVE**	**Baseline**	**RAAVE**	**Baseline**
All	3.50	3.64	3.01	3.12	3.50	3.36	3.19	3.20
CF	3.32	3.43	2.61	3.21	3.43	3.71	3.04	3.00
ED	3.65	4.00	3.15	2.93	3.46	3.07	3.23	3.13
PC	3.64	3.89	2.86	3.00	3.50	3.36	3.00	3.21
TP	3.42	3.25	3.83	3.33	3.75	3.33	3.67	3.33
CF+PC	3.43	3.74	2.69	3.07	3.45	3.48	3.02	3.14
ED+TP	3.58	3.54	3.37	3.18	3.55	3.23	3.37	3.26

The results in Table 17 follow the same pattern as in the question search task. The scores are slightly better with the Baseline system overall. But better results are achieved using RAAVE system with videos which contain visual aid and are relatively more technical and difficult to comprehend than the other two.

In addition to rating of the user produced summaries individually, reviewers were also asked to rank the summaries in order of their preference. Users produced 6 summaries per video. For each video, reviewers assigned a rank between 1 and 6 to produced summaries (1 for most preferred and 6 to the least preferred).

Table 18 shows the results for each video. Est. column lists the estimated likelihood for the summary produced to be ranked no.1 by reviewers. For two of the videos CF and ED the RAAVE has the higher probability to produce the top summary while for the other two Baseline scored higher.

Table 19 shows the overall likelihood for each system instead of individual summaries. It can be seen that the RAAVE has scored higher for both kind of videos i.e. the likelihood that the top rank will be assigned to a summary produced using the RAAVE system.

Table 18: Probability that reviewer will rank this as no.1

TP			**CF**			**ED**			**PC**		
Est.	**User**	**Sys**	**Est.**	**User**	**Sys.**	**Est.**	**User**	**Sys.**	**Est.**	**User**	**Sys.**
0.1076	1	B	0.125	1	R	0.1891	2	B	0.08249	2	R
0.2314	3	R	0.1647	3	B	**0.232**	5	R	**0.50695**	4	B
0.164	6	B	0.1125	4	R	**0.1984**	6	R	0.10486	5	B
0.0382	7	B	**0.2443**	7	R	0.1574	8	B	**0.13291**	8	R
0.1873	9	R	**0.1887**	9	B	0.1179	11	R	0.04768	10	B
0.2716	12	B	0.1647	10	R	0.1052	12	R	0.12511	11	B

Table 19: probability that reviewers are likely to rank this as no.1

Videos	RAAVE	Baseline
TP+ED (slides)	0.54	0.46
CF+PC (no slides)	0.57	0.43

Table 20 shows each user's interactions with both the baseline and RAAVE system. On average reviewers rank the synopsis produced by RAAVE system slightly better than the baseline system, 3.50 compared to 3.56 of the baseline. The Student T-Test p-value of the result is 0.45 which means the results are not statistically significant.

The purpose of the analysis of this section is to identify what techniques employed by users enabled them to produce better synopsis. Table 21 shows the performance of user's using different modalities and feature sets. 3 users which only consumed the automatically generated summaries of segment seemed to perform better than the rest (see top row of Table 21). 5 users who used both text and video were able to create better summaries compared to the users who only use text i.e. summary and transcript. It might be because different modalities can be consumed in parallel e.g. a user can play the video snippet of segment 1 and read through the summaries of the remaining segments while continue listening to the video snippet of the first segment.

Table 20: User Interaction with Baseline and RAAVE for synopsis writing task. The column "Interac." lists the number of interaction each user did with the base line player logged by socialSkip tool. "Beg, Med, End" shows if the user consume the beginning, middle and end of the TED talk, Y denotes Yes while N denotes no. "Modality" list the type of modalities and feature set, "S" stands for summary, "T" for Transcript, "V" for Video and "W" stands for word clouds. The column "Avg. Rank" list the average of rank given to synopsis by 7 reviewers (lower is better).

	Baseline									**RAAVE**								
User	**Interac.**	**Beg**	**Med**	**End**	**Read.**	**Info.**	**Conci.**	**Overall**	**Avg. Rank**	**Modality**	**Beg**	**Med**	**End**	**Read.**	**Info.**	**Conci.**	**Overall**	**Avg. Rank**
1	21	Y	Y	N	3.17	3.17	3.50	3.17	4.33	S+T	1	0	0	3.00	2.14	3.29	2.71	4.29
2	14	Y	Y	N	4.43	3.29	3.14	3.43	3.43	S+V	1	1	0	3.71	2.43	3.57	2.86	4.29
3	2	Y	N	N	3.86	3.00	3.29	3.14	3.43	S+V	1	1	0	3.67	3.83	3.33	3.83	2.50
4	14	Y	N	Y	4.71	4.00	4.14	4.43	1.29	S+T	1	1	0	3.00	1.57	3.86	2.43	4.57
5	13	Y	Y	N	3.86	2.86	3.43	3.14	3.71	S	1	1	1	4.00	3.71	4.00	3.86	2.43
6	13	Y	Y	N	3.00	3.67	3.17	3.67	3.33	S+T	1	0	0	4.00	3.43	3.43	3.71	2.86
7	37	Y	Y	Y	2.50	2.50	3.33	2.33	5.67	S+W	1	1	1	4.00	3.57	3.43	4.00	2.29
8	102	Y	Y	Y	3.57	2.71	3.00	2.86	3.57	S+T+V	1	1	1	3.57	3.29	3.43	3.14	3.14
9	12	Y	Y	Y	3.00	3.43	4.14	2.86	3.00	S+T	1	1	1	3.17	3.83	4.17	3.50	3.00
10	9	Y	N	N	2.86	2.29	2.71	2.29	5.29	S+T+V	1	0	0	3.29	3.14	3.14	3.00	3.43
11	6	Y	Y	N	4.14	2.86	3.14	3.00	3.29	S+T	1	1	1	2.83	2.67	3.33	2.57	4.43
12	18	Y	Y	Y	4.29	3.71	3.29	4.00	2.43	S	1	1	1	3.67	2.67	3.00	2.50	4.83
	Avg.				3.62	3.12	3.36	3.19	3.56	Avg.				3.49	3.02	3.50	3.18	3.50

Table 21: user performance based on the choice of modality and feature sets.

System	Setting	No. of Users	Readability	Informativeness	Conciseness	Overall	Avg. Rank
RAAVE	Sum	**3**	**3.89**	**3.32**	**3.48**	**3.45**	**3.18**
RAAVE	Sum + trans	5	3.20	2.73	3.61	2.99	3.83
RAAVE	Sum + Vid	4	3.56	3.17	3.37	3.21	3.34
RAAVE	End	6	3.54	3.29	3.56	3.26	3.35
RAAVE	No end	6	3.44	2.76	3.44	3.09	3.65
Baseline	Over all (B)	12	3.62	3.12	3.36	3.19	3.56
RAAVE	Overall (R)	12	3.49	3.02	3.50	3.18	3.50

7.5.3. Results Hypothesis C (User Experience)

Hypothesis C: Users have a better experience interacting with RAAVE compared to the baseline player.

Users were asked to fill the UEQ (Leggett and Bilda, 2008) twice, once for RAAVE and once for the Baseline system. The 26 questions of the UEQ can be categorized in to following categories.

- Attractiveness
- Perspicuity
- Efficiency
- Dependability
- Stimulation
- Novelty

Figure 31 shows the comparison of user experience with both systems. Users scored RAAVE better in all categories except "Perspicuity" which is not surprising since the baseline system is much more familiar and simpler than RAAVE system. Table 22 shows the T-Test score to assess if the difference between the two systems reported by users is statistically significant or not, as it can be seen that results are statistically significant in all but one category (Table 22).

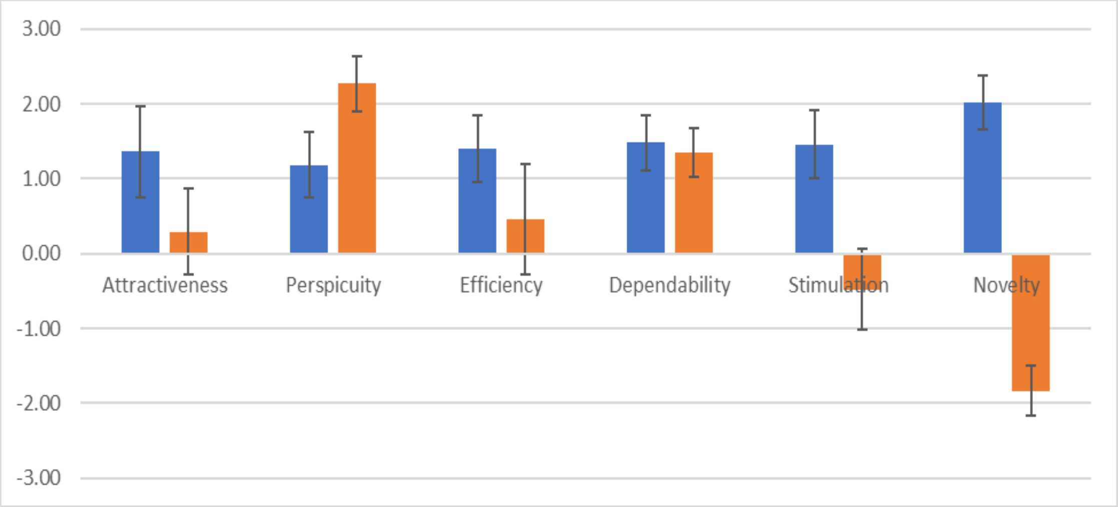


Figure 31: Blue (darker) bars represent RAAVE while orange (lighter) bars represent the baseline system.

Table 22: Simple T-Test to check if the scale means of the two systems differ significantly.

Alpha level: 0.05		
Category	**P-Value**	**Statistical Significance**
Attractiveness	0.0206	Significant Difference
Perspicuity	0.0012	Significant Difference
Efficiency	0.0465	Significant Difference
Dependability	0.6057	No Significant Difference
Stimulation	0.0000	Significant Difference
Novelty	0.0000	Significant Difference

7.6. Discussion

As explained in section 7.2, experiment 3 is designed to stress test the prototype system thereby the proposed template-based approach, still the results of the experiment are encouraging.

In efficiently searching for information within video content, users were spending less time per correct answer compared to the baseline system (Table 15). Overall users were able to answer more correctly using baseline system, it is because due to the design of the experiment it was easier for them to watch the whole video and answer the questions parallelly. As it can be seen in Table 16 on average users spent 97.14% of video duration to perform the task considering the fact that each video included promos at both the beginning and end (which users skipped as they had no content value relative to the task). RAAVE got better results for difficult videos compared to the easy ones (Table 15). In terms of search strategies user who used a mixture of query box and content interaction were able to answer more accurately while spending less time on the task (row 3 and 4 of Table). This can be used in further streamlining the representation of the video i.e. using templates which encourage users to interact more with content.

Regarding quickly getting a better understanding of the essence of the video. The results are moderately encouraging for RAAVE system. The likelihood that the synopsis creates using RAAVE would be ranked 1st by reviewers was higher compared to the baseline player even though the margin is not very wide (Table 18). In terms of quality criterion overall synopsis produced using the baseline system were scored higher compared to RAAVE. However, as it was the case in answer search task RAAVE scored better for difficult videos (TP+ED) on all the 4, quality criterion (last row of Table 17). In terms of interaction strategies, the 3 users who only consumed the automatically generated summaries on average scored better than others.

In the synopsis writing task, while the users who consumed both textual and video modality scored better than those who consumed the summary and transcript i.e. the textual modality, their performance was lower than the 3 who only use the summary. A similar trend can be seen in the answer search task (Table 16 row 1 and 2). The 6 users who interacted with both text and video modality took longer and answer less questions compared to the 6 users who only interacted with the text modality.

Hence the proposed approach provide advantages in terms of providing an flexible and engaging experience to user during exploration tasks and provides advantages in terms spending less time searching for information and have a better understanding of video by choosing both the modality and amount of detail to consume the content.

While it was initially assumed that giving users ability to parallelly consume different modalities would be beneficial, e.g. it is possible for the user to listen to one segment while reading the summary of another. However, the results suggest that such a strategy does not always yield optimal performance.

In terms of user experience with the system, despite the lack of familiarity and other limitations of RAAVE users had a productive experience with the RAAVE system. Users rated the RAAVE system more favourably by a wide margin except in the category of dependability although RAAVE's score is still higher than Baseline player. It is not surprising as due to its familiarity, simple nature, and wide availability, the regular video player is very dependable i.e. it does the simple things it does quite well, whereas RAAVE provided a lot of options and the UI was not fully matured. With a better UI and more practice, the user exploration experience with RAAVE is bound to get better.

7.7. Concluding Remarks

The goal of the experiment was to evaluate the potential of the template driven representation approach RAAVE in video exploration tasks. The comparison study between the prototype system and a baseline system, showed that the RAAVE approach does have potential to enhance user's exploration experience with video content.

One might consider the gain in efficiency (answer search task) and quality (synopsis writing task) to be marginal as reported in the results. This is not a complete surprise as the experiment was designed to stress test the proposed prototype by deliberately introducing some disadvantages for the prototype system.

Consider as an example the answer search task, the test videos were in the range of 15 to 21 minutes and there were 14 questions to be answered and they were well spread across the content of whole video. This made simply watching the whole video to answer the questions a viable strategy. In a real exploration scenario, it is more likely that the user needs to find fewer number of questions whose answer might be scattered across different portion of a longer video for example 3 questions in a 30-minute video. In such a scenario watching the whole video would definitely decrease the overall efficiency of the task. The template driven approach of RAAVE would provide for an efficient alternative.

Combined with a more polished interface and increased familiarity (more practice), RAAVE's value in enhancing the user exploration experience is bound to improve.

Apart from testing the prototype system, another goal of the experiment was to analyse user usage and interaction strategies with the prototype to see if there can be some usage pattern that could help improve the design of the proposed representation engine. The analysis has been useful in this regard as well.

The template selection process in the prototype is based on 3 factors. These are segment relevance, feature set availability and finally user preference of a modality and expanse (section 3.2.3). Analyses of the experiment suggest that the task should also be a factor, i.e. users need to consume the content differently for different tasks. While RAAVE did provide the ability to alter the representation per user's preference, users were more likely to use the representation as it is (the way the engine offered it) even if it was not optimal. So, it would be beneficial for users if an implementation of the representation engine also uses the nature of task as a factor in choosing the template.

Secondly the representation of meta-data. In the experiment of phase 2 (section 5) user feedback insisted on showing more meta-data about the representation. While the suggested meta-data was presented in the new prototype used in phase 3, analysis of experiment suggests that user may benefit from more meta-data. Specifically, information regarding not only what segment is relevant but what makes the segment relevant or the information regarding which modality of that segment makes it relevant could help users search the relevant content more efficiently. However, how best to offer that information to users without creating too much clutter and whether users will utilize that information in an efficient manner would require more testing.

In short, experiment 3 has shown that the template driven approach does have the potential to enhance the user exploration experience. As results have shown that the proposed approach does improve the exploration experience even if the improvements are modest due to the intentional disadvantages designed in the experiment.

8. Conclusion

The main question this thesis evaluated is, the extent to which automatically extracted multimodal features can be leveraged to enhance the exploration experience in video content. This has the following aspects:

- Exploration in video content.
- Definition of experience and its enhancement.
- Evaluate the potential of automatically extracted features in enhancing the experience.

Exploration in a video is not just the ability to search for something, but it is also the ability to have an overall synopsis in an efficient manner while providing an engaging and flexible user experience (section 1.2).

The following objectives stemmed from the research question (section 1.3).

1. Review of the state of the art in video exploration.
2. Examination of techniques to break the tight bond between parallel modalities and extract features from them.
3. Presentation of the extracted multimodal features to users in an interactive manner so they can explore the content of a video and learn usage patterns.
4. Design and development of a template driven representation engine approach based on the usage patterns that automatically generate multimedia interactive document from video content for users.
5. Evaluation of the representation engine's performance with respect to content exploration tasks.

The review of the state of the art in video exploration to understand the approaches utilized by researchers, revealed the following limitations.

- Representation of the content lacks user control in its configuration.
- A solution is either suitable for searching for something in particular or to provide overall synopsis of the video and not both and it affects the user experience in tasks with evolving exploration goals.

- The user’s ability to interact with the content is either limited or the interface is designed to be either content dependent or overly complex.
- A solution often requires prior curation by humans i.e. manual effort.

Multimodal features were extracted from TED presentation videos and their relationship with user engagement evaluated. Phase 1 (chapter 4) identified a toolset to automatically extract multimodal features and ratified their value in term of user engagement.

Once it was established that automatically extracted multimodal features do indeed have a relationship with user engagement criterion, the next phase was to represent them to users in order to learn some usage pattern and to observe how the user interact with different features while exploring video content. Chapter 5 reported on the user study conducted for that purpose.

The design of the proposed approach that utilized multimodal features more widely to enhance the user’s exploration experience with video content is outlined in Chapter 3. The design is influenced by the state of the art and from the findings of the experiments performed in phase 2 (chapter 5). The design is based on the contention that a video is not just a single/homogenous artefact but, it is a combination of different temporally bound parallel modalities (visual, audio, linguistic/ textual). As described above, SOTA approaches to represent video content are highly customized and cannot be reconfigured for evolving user needs. To solve this the proposed approach RAAVE is designed as a representation engine independent of a user interface that automatically transforms video into an interactive multimedia document (Chapter 3). A prototype was developed to test the proposed approach (chapter 6).

The performance of the prototype and thereby the proposed approach was evaluated as a comparison study in Chapter 7.

A comparison study was performed to evaluate the performance of the proposed approach in different exploration tasks. The experiment showed encouraging results regarding the applicability of the proposed approach in enhancing users’ exploration experience in video content. Participants in the user study performed two types of task; the answer search task to evaluate RAAVE’s ability to better enable users to find information within a video compared to a baseline system and synopsis writing task to evaluate RAAVE’s ability to enable users to get a quicker understanding of the content of video compared to a baseline system. Apart from comparing the performance of RAAVE in exploration tasks the user study in chapter 7 also compared the feedback of user experience with RAAVE and the baseline system. Results of the user study showed that RAAVE did enhance the performance and user experience in terms of exploration in video tasks and demonstrated the potential of the proposed approach.

8.1. Thesis Contribution

This thesis proposed an approach to transform a video into an interactive multimedia document. Transforming a video into an interactive document opens up new ways to explore the content of video as users in addition to watching the video, can consume it in a combination of different modalities and amount of detail, better suited to the context.

The proposed approach transforms a video by representing its contents. To do that it utilizes a template driven engine. Hence proposing, designing and evaluating a template driven representation engine-based approach to transform video content is the major contribution of this PhD thesis.

One of the minor contributions of this thesis is enabling the user to effectively explore a video. That is, the approach enables the user to effectively, both search information in video and also to get an overall essence of the video in a configurable manner.

The other minor contribution is the engagement assessment system described in phase 1. The ability to identify engaging and non-engaging presentation has the potential to be used in a variety of applications as demonstrated in section 2.

8.1.1. Contribution to the SOTA

Following are the contributions to the state of the art.

8.1.1.1. Journal and Conference Papers

Following is the doctoral consortium paper that discussed the idea of considering video as a multimedia content source in order to enhance its exploration potential.

- **Salim F.A.** From artifact to content source: Using multimodality in video to support personalized recomposition. In User Modelling, Adaptation and Personalization 2015. UMAP, 2015

The results of the experiments performed in phase 1 (section 4) were reported in the following conference papers.

- **Salim F.A.,** Haider F., Conlan O., Luz S., and Campbell N. 2015. Analyzing Multimodality of Video for User Engagement Assessment. In Proceedings of the 2015 ACM on

International Conference on Multimodal Interaction (ICMI '15). ACM, New York, NY, USA, 287-290.

- **Salim F.A.,** Levacher K., Conlan O., Campbell N. Extending Multimodal Characteristics of Video to Understand User Engagement and Potential Segmentation. In User Modelling, Adaptation and Personalization 2015. UMAP, 2015.

Following conference paper reported the preliminary results of experiment performed in phase 2 (section 5).

- **Salim F.A.**, Haider F., Conlan O., Luz S. (2017) An Alternative Approach to Exploring a Video. In: Karpov A., Potapova R., Mporas I. (eds) Speech and Computer. SPECOM 2017. Lecture Notes in Computer Science, vol 10458. Springer.

The following journal paper is the extended version of the above conference paper which reported updated analysis of the users study of phase 2 (section 5).

- **Salim, F.A.**, Haider F., Conlan O., Luz S. (2018). An Approach for Exploring a Video via Multimodal Feature Extraction and User Interactions. Journal on Multimodal User Interfaces. https://doi.org/10.1007/s12193-018-0268-0.

Following article reports the experiment performed to evaluate RAAVE in phase 3 (section 7).

- **Salim, F.A.**, Conlan O. (2018). Introducing RAAVE; an Approach for Multimodal Video Exploration. Multimedia Tools Appl. (under review).

8.1.1.2. Other Associated Publications

Other researcher have utilized the ideas presented in this thesis. Following are conference papers produced by other researchers with the help of the author of this thesis.

- Haider F., **Salim F.A.**, Luz S., Conlan O., Campbell N. "High level visual and paralinguistic features extraction and their correlation with user engagement," in Signal Processing and Information Technology (ISSPIT), 2015 IEEE International Symposium on, vol., no., pp.326-331, 7-10 Dec. 2015.

In the above paper authors used the configuration of experiment 1.2 (see section 4.4 for details) to perform analysis of variance (ANOVA) test to analyse viewer engagement with presentations for the purpose of providing feedback to presenters to help them improve the engagement level of a talk. While in the following papers authors extended the idea by segmenting TED presentations based on speech expressions to identify the user-engagement at segment level.

- Haider F., **Salim, F.A.**, Luz, S., Vogel, C., Conlan, O., Campbell, N. (2017) Visual, Laughter, Applause and Spoken Expression Features for Predicting Engagement Within TED Talks. Proc. Interspeech 2017, 2381-2385, DOI: 10.21437/Interspeech.2017-1633.
- Haider F., **Salim, F.A.**, Luz, S., Vogel, C., Conlan, O., Campbell, N. (2018) An Active Feature Transformation Method for Attitude Recognition of Video Bloggers Interspeech.

Following paper describes the project which utilized ideas presented in this thesis to express multimedia content of video using semantically uplifted information.

- Debattista J.**, Salim F.A.**, Haider F., et al., "Expressing Multimedia Content Using Semantics — A Vision," 2018 IEEE 12th International Conference on Semantic Computing (ICSC), Laguna Hills, CA, USA, 2018, pp. 302-303.

8.1.1.3. Patents

Following is a patent application based on the template driven representation engine-based approach described in section 3.

- UK Patent Application No: 1714592.1. (under review)

  Determining representative content to be used in representing a video

  A computer implemented method of determining representative content to be used in representing a video, a data processing apparatus comprising one or more processors adapted to perform the method and a computer readable storage medium comprising instructions which, when executed by a computer, cause the computer to carry out the method.

8.2. Future Directions

This thesis has presented the design and evaluation of RAAVE which transform a video into an interactive multimodal document to open up new ways to explore its content. The experiments performed in this thesis currently are narrowed down in the following ways:

- Only presentation style TED videos are used.
- Only query-based scenario for exploration was evaluated.

There are many possible directions to pursue further research work. Future directions can be divided into two broad categories:

- Enhancing feature extractions.
- Applying RAAVE in different scenarios.

8.2.1. Enhancing the RAAVE pipeline by using different Machine Learning methods to extract multimodal features.

Multimodal features are a vital part of the proposed approach. In phase 2 and 3 of the experiments, a limited number of features were represented to users to evaluate the potential of the approach. In the future it is intended to incorporate more multimodal features in the representation and evaluate the potential enhancement in the exploration. Some examples are:

- Visual features such as facial expressions, body movements.
- Audio/paralinguistic features.
- Linguistic features such as semantic uplift of topic concepts etc.

8.2.2. Applying RAAVE in different application scenarios.

In the future, the plan is to expand the scope of content i.e. to apply the proposed approach on a variety of video content e.g.

- Massive Online Open Courses (MOOC) videos.
- Training videos. (Lynda, misc. corporate training videos).
- Instructional tutorials (how to fix a bike, how to apply makeup etc.).
- News footage and documentaries.
- Life logging videos (meeting recordings, conference calls, video chats)

As an example consider life logging videos particularly meeting recording or recordings of conference calls or video chats. The proposed template based multimodal representation can be used to provide the ability to explore the content of meeting in a nonlinear and multimodal manner. Girgensohn et al. proposed a hypervideo based approach to explore meeting recordings (Girgensohn *et al.*, 2016c). As detailed in state of the art (section 2), RAAVE extends the idea of nonlinear exploration of video by transforming it into multimedia document. Similarly for meeting recordings, the templating approach can be applied to create a multimedia brief based on not only topic of interest or speaker choice (Luz and Masoodian, 2004; Girgensohn *et al.*, 2016c) but also choose the amount of detail and choice of modality by choosing an appropriate template based on user preference or end user device (see 6.4).

Apart from the variety of content, the other dimension is the application of the approach in exploration task scenarios. In the future, it is intended to test the approach in a variety of situations e.g. automatic curation of news article or multimedia essay from video footage based on a personalization model instead of waiting for the user to execute a query.

Another interesting use of the proposed approach is to transform the video content for professional use cases. An example could be allowing the ability to search for information within long video footage and automatically or semi-automatically curating a new multimedia artefact which may be a video, or it may be a multimedia document.

Finally, it is a hope of the author of this thesis that the proposed idea of transforming content based on context can be expanded to content other than video. The design factors of the representation engine can be used to search a heterogenous data-source which could be structured or semi-structured and the information extracted can be automatically represented or curated as an interactive multimedia document and help create digital narratives on demand.

9. References

Adcock, J. *et al.* (2010) 'TalkMiner : A Lecture Webcast Search Engine', *Interfaces*, (21), pp. 241–250. doi: 10.1145/1873951.1873986.

Almeida, J., Leite, N. J. and Torres, R. D. S. (2013) 'Online video summarization on compressed domain', *Journal of Visual Communication and Image Representation*. Elsevier Inc., 24(6), pp. 729–738. doi: 10.1016/j.jvcir.2012.01.009.

Anwar, A., Salama, G. I. and Abdelhalim, M. B. (2013) 'Video Classification And Retrieval Using Arabic Closed Caption', in *ICIT 2013 The 6th International Conference on Information Technology VIDEO*.

Attfield, S., Piwowarski, B. and Kazai, G. (2011) 'Towards a science of user engagement (Position Paper)', in *WSDM Workshop on User Modelling for Web Applications*.

Aubert, O. *et al.* (2008) 'Canonical processes in active reading and hypervideo production', *Multimedia Systems*, 14(6), pp. 427–433. doi: 10.1007/s00530-008-0132-2.

Autosummarizer (2016) *autosummarizer.com*. Available at: http://autosummarizer.com/.

de Avila, S. E. F. *et al.* (2011) 'VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method', *Pattern Recognition Letters*. Elsevier B.V., 32(1), pp. 56–68. doi: 10.1016/j.patrec.2010.08.004.

Balasubramanian, V., Doraisamy, S. G. and Kanakarajan, N. K. (2016) 'A multimodal approach for extracting content descriptive metadata from lecture videos', *Journal of Intelligent Information Systems*, 46(1), pp. 121–145. doi: 10.1007/s10844-015-0356-5.

Barthel, R., Ainsworth, S. and Sharples, M. (2013) 'Collaborative knowledge building with shared video representations', *International Journal of Human Computer Studies*. Elsevier, 71(1), pp. 59–75. doi: 10.1016/j.ijhcs.2012.02.006.

Bayomi, M., Levacher, K. and Ghorab, M. R. (2016) 'Natural Language Processing and Information Systems', 9612, pp. 187–199. doi: 10.1007/978-3-319-41754-7.

Bellard, F., Niedermayer, M. and Others, A. (2012) 'FFmpeg', *ht tp://ffmpeg. org*.

Belo, L. dos S. *et al.* (2016) ‘Summarizing video sequence using a graph-based hierarchical approach’, *Neurocomputing*, 173, pp. 1001–1016. doi: 10.1016/j.neucom.2015.08.057.

Benini, S., Migliorati, P. and Leonardi, R. (2010a) ‘Statistical Skimming of Feature Films’, *International Journal of Digital Multimedia Broadcasting*, 2010, pp. 1–11. doi: 10.1155/2010/709161.

Benini, S., Migliorati, P. and Leonardi, R. (2010b) ‘Statistical Skimming of Feature Films’, *International Journal of Digital Multimedia Broadcasting*, 2010, pp. 1–11. doi: 10.1155/2010/709161.

Bertini, M. *et al.* (2011) ‘Interactive video search and browsing systems’, in *Proceedings - International Workshop on Content-Based Multimedia Indexing*. doi: 10.1109/CBMI.2011.5972543.

Bode, K. (2017) ‘The Equivalence of “Close” and “Distant” Reading; or, Toward a New Object for Data-Rich Literary History’, *Modern Language Quarterly*, 78(1), pp. 77–106. doi: 10.1215/00267929-3699787.

Boissiere, G. (1998) ‘Automatic Creation of Hypervideo News Libraries for the World Wide Web’, in *HYPERTEXT ’98 Proceedings of the ninth ACM conference on Hypertext and hypermedia*, pp. 279–280.

Boyles, N. (2013) ‘Closing in on close reading’, *Educational Leadership*, pp. 36–41. doi: 10.1126/science.7504323.

Brachmann, C. . b and Malaka, R. . (2009) ‘Keyframe-less integration of semantic information in a video player interface’, *EuroITV’09 - Proceedings of the 7th European Conference on European Interactive Television Conference*, pp. 137–140. doi: 10.1145/1542084.1542109.

Bradski, G. (2000) *The OpenCV Library*, *Dr. Dobb’s Journal of Software Tools*. Available at: http://www.drdobbs.com/open-source/the-opencv-library/184404319.

Brezeale, D. and Cook, D. J. (2009) ‘Learning video preferences using visual features and closed captions’, *IEEE Multimedia*, 16(3), pp. 39–47. doi: 10.1109/MMUL.2009.51.

Bulterman, D. and Rutledge, L. (2009) *SMIL 3.0*. 2nd edn. Springer-Verlag Berlin Heidelberg.

Calumby, R. T. *et al.* (2017) ‘Diversity-based interactive learning meets multimodality’, *Neurocomputing*, 259, pp. 159–175. doi: 10.1016/j.neucom.2016.08.129.

Chen, F., Vleeschouwer, C. De and Cavallaro, A. (2014) ‘Resource Allocation for Personalized

Video Summarization', 16(2), pp. 455–469.

Choi, F. (2000) 'Advances in Domain Independent Linear Text Segmentation', in *Proceedings of NAACL 2000*. Seattle: Association for Computational Linguistics, pp. 26–33.

Choi, I. Y. *et al.* (2016) 'Collaborative filtering with facial expressions for online video recommendation', *International Journal of Information Management*. Elsevier Ltd, 36(3), pp. 397–402. doi: 10.1016/j.ijinfomgt.2016.01.005.

Chorianopoulos, K. (2011) 'SocialSkip: pragmatic understanding within web video', *EuroTV '11*, pp. 25–28. doi: 10.1145/2000119.2000124.

CISCO (2017) 'The Zettabyte Era: Trends and Analysis', *Cisco*, (June 2017), pp. 1–29. doi: 1465272001812119.

Cobârzan, C. *et al.* (2017a) 'Interactive video search tools: a detailed analysis of the video browser showdown 2015', *Multimedia Tools and Applications*. doi: 10.1007/s11042-016-3661-2.

Cobârzan, C. *et al.* (2017b) 'Interactive video search tools: a detailed analysis of the video browser showdown 2015', *Multimedia Tools and Applications*, 76(4), pp. 5539–5571. doi: 10.1007/s11042-016-3661-2.

Collier, P. (2008) *The bottom billion*, *TED*. Available at: https://www.ted.com/talks/paul_collier_shares_4_ways_to_help_the_bottom_billion (Accessed: 22 December 2017).

Cooper, M. *et al.* (2011) 'Multimedia Information Retrieval at FX Palo Alto Laboratory', *SPIE 7881, Multimedia on Mobile Devices 2011; and Multimedia Content Access: Algorithms and Systems V*, 7881. doi: 10.1117/12.878615.

Craig, C. L. and Friehs, C. G. (2013) 'Video and HTML: Testing Online Tutorial Formats with Biology Students', *Journal of Web Librarianship*, 7(3), pp. 292–304. doi: 10.1080/19322909.2013.815112.

Denoue, L. *et al.* (2013) 'Real-time direct manipulation of screen-based videos', *Proceedings of the companion publication of the 2013 international conference on Intelligent user interfaces companion - IUI '13 Companion*, p. 43. doi: 10.1145/2451176.2451190.

Dobrian, F. *et al.* (2011) 'Understanding the impact of video quality on user engagement', *ACM SIGCOMM Computer Communication Review*, p. 362. doi: 10.1145/2043164.2018478.

Dong, A., Li, H. and Francisco, S. (2008) ‘ONTOLOGY-DRIVEN ANNOTATION AND ACCESS OF’.

Drouin, J. (2014) ‘Close- And Distant-Reading Modernism: Network Analysis, Text Mining, and Teaching The Little Review’, *The Journal of Modern Periodical Studies*, 5(1), pp. 110–135. doi: 10.1353/jmp.2014.0001.

Duflo, E. (2010) *Social experiments to fight poverty*, *TED*. Available at: https://www.ted.com/talks/esther_duflo_social_experiments_to_fight_poverty (Accessed: 22 December 2017).

Ericsson (2016) *TV AND MEDIA 2016, An Ericsson Consumer and Industry Insight Report*.

Evangelopoulos, G. *et al.* (2013) ‘Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention’, *IEEE Transactions on Multimedia*, 15(7), pp. 1553–1568. doi: 10.1109/TMM.2013.2267205.

Eyben, F. *et al.* (2013) ‘Recent developments in openSMILE, the munich open-source multimedia feature extractor’, *Proceedings of the 21st ACM international conference on Multimedia - MM ’13*, (May), pp. 835–838. doi: 10.1145/2502081.2502224.

Farhadi, B. and Ghaznavi-Ghoushchi, M. B. (2013) ‘Creating a novel semantic video search engine through enrichment textual and temporal features of subtitled YouTube media fragments’, *Proceedings of the 3rd International Conference on Computer and Knowledge Engineering, ICCKE 2013*, (Iccke), pp. 64–72. doi: 10.1109/ICCKE.2013.6682857.

Finke, M. and Balfanz, D. (2004) ‘A reference architecture supporting hypervideo content for ITV and the internet domain’, *Computers and Graphics (Pergamon)*, 28, pp. 179–191. doi: 10.1016/j.cag.2003.12.005.

Freeland, C. (2013) *The rise of the new global super-rich*, *TED*.

Galuscakova, P., Saleh, S. and Pecina, P. (2016) ‘SHAMUS: UFAL Search and Hyperlinking Multimedia System’, in *Advances in Information Retrieval: 38th European Conference on IR Research, ECIR 2016, Padua, Italy*. Padua, pp. 853–856. doi: 10.1007/978-3-319-30671-1.

Ganier, F. and de Vries, P. (2016) ‘Are instructions in video format always better than photographs when learning manual techniques? The case of learning how to do sutures’, *Learning and Instruction*. doi: 10.1016/j.learninstruc.2016.03.004.

Girgensohn, A. *et al.* (2016a) ‘Guiding Users through Asynchronous Meeting Content with Hypervideo Playback Plans’, in *Proceedings of the 27th ACM Conference on Hypertext and*

Social Media - HT '16, pp. 49–59. doi: 10.1145/2914586.2914597.

Girgensohn, A. *et al.* (2016b) 'Guiding Users through Asynchronous Meeting Content with Hypervideo Playback Plans', *Proceedings of the 27th ACM Conference on Hypertext and Social Media - HT '16*, pp. 49–59. doi: 10.1145/2914586.2914597.

Girgensohn, A. *et al.* (2016c) 'Guiding Users through Asynchronous Meeting Content with Hypervideo Playback Plans', *Proceedings of the 27th ACM Conference on Hypertext and Social Media - HT '16*, pp. 49–59. doi: 10.1145/2914586.2914597.

Gravier, G. *et al.* (2016) 'Shaping-Up Multimedia Analytics: Needs and Expectations of Media Professionals', in, pp. 303–314. doi: 10.1007/978-3-319-27671-7.

Guan, G., Wang, Z., Mei, S., Ott, M., *et al.* (2014) 'A {Top}-{Down} {Approach} for {Video} {Summarization}', *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(1), pp. 1–21. doi: 10.1145/2632267.

Guan, G., Wang, Z., Mei, S., Ott, M. A. X., *et al.* (2014) 'A Top-Down Approach for Video Summarization', 11(1).

Guo, P. J., Kim, J. and Rubin, R. (2014) 'How Video Production Affects Student Engagement : An Empirical Study of MOOC Videos', in *L@S 2014 - Proceedings of the 1st ACM Conference on Learning at Scale*. doi: 10.1145/2556325.2566239.

Haesen, M. *et al.* (2011) 'Finding a needle in a haystack: an interactive video archive explorer for professional video searchers', *Multimedia Tools and Applications*, 63(2), pp. 331–356. doi: 10.1007/s11042-011-0809-y.

Haider, F. *et al.* (2016) 'High level visual and paralinguistic features extraction and their correlation with user engagement', in *2015 IEEE International Symposium on Signal Processing and Information Technology, ISSPIT 2015*. doi: 10.1109/ISSPIT.2015.7394353.

Haider, F. *et al.* (2017) 'Visual , Laughter , Applause and Spoken Expression Features for Predicting Engagement within TED Talks', in *Interspeech*. Stockholm, pp. 2381–2385.

Hall, M. *et al.* (2009) 'The WEKA data mining software', *ACM SIGKDD Explorations*, 11(1), pp. 10–18. doi: 10.1145/1656274.1656278.

Halvey, M. *et al.* (2014) 'Supporting exploratory video retrieval tasks with grouping and recommendation', *Information Processing and Management*. Elsevier Ltd, 50(6), pp. 876–898. doi: 10.1016/j.ipm.2014.06.004.

Hauptmann, A. *et al.* (2006) ‘Extreme video retrieval: joint maximization of human and computer performance’, in *Proceedings of the 14th ACM international conference on Multimedia*. Santa Barbara: ACM, pp. 385–393. doi: 10.1145/1180639.1180721.

Hildebrand, M. and Hardman, L. (2013) ‘Using Explicit Discourse Rules to Guide Video Enrichment’, *Www*, pp. 461–464.

Hoffmann, P. and Herczeg, M. (2006) ‘Hypervideo vs. Storytelling: Integrating Narrative Intelligence into Hypervideo’, in *Proceedings of the Third International Conference on Technologies for Interactive Digital Storytelling and Entertainment (TIDSE 2006)*, pp. 37–48. doi: 10.1007/11944577_4.

Hong, R. *et al.* (2011) ‘Beyond search Event Driven summarization of web videos’, *ACM Transactions on Multimedia Computing, Communications, and Applications*, 7(4), pp. 1–18. doi: 10.1145/2043612.2043613.

Hosseini, M.-S. and Eftekhari-Moghadam, A.-M. (2013) ‘Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video’, *Applied Soft Computing*. Elsevier B.V., 13(2), pp. 846–866. doi: 10.1016/j.asoc.2012.10.007.

Hudelist, M., Schoeffmann, K. and Xu, Q. (2015) ‘Improving interactive known-item search in video with the keyframe navigation tree’, *MultiMedia Modeling*, 8935, pp. 306–317. doi: 10.1007/978-3-319-14445-0_27.

Jänicke, S. *et al.* (2015) ‘On Close and Distant Reading in Digital Humanities : A Survey and Future Challenges’, *Eurographics Conference on Visualization (EuroVis) (2015)*, pp. 1–21. doi: 10.2312/eurovisstar.20151113.

Jin, J. (2017) ‘Problems of Scale in “Close” and “Distant” Reading’, 1, pp. 105–130.

Khan, E. and AlSalem, A. (2012) ‘Ivia: Interactive Video Intelligent Agent Framework for Instructional Video Information Retrieval’, *Procedia - Social and Behavioral Sciences*, 64, pp. 186–191. doi: 10.1016/j.sbspro.2012.11.022.

Kim, D.-J., Frigui, H. and Fadeev, A. (2008) ‘A generic approach to semantic video indexing using adaptive fusion of multimodal classifiers’, *International Journal of Imaging Systems and Technology*, 18(2–3), pp. 124–136. doi: 10.1002/ima.20147.

Koch, S. *et al.* (2014) ‘VarifocalReader &amp;#x2014; In-Depth Visual Analysis of Large Text Documents’, *IEEE Transactions on Visualization and Computer Graphics*, 20(12), pp. 1723–1732. doi: 10.1109/TVCG.2014.2346677.

Laugwitz, B., Held, T. and Schrepp, M. (2008) 'Construction and Evaluation of a User Experience Questionnaire', *HCI and Usability for Education and Work*, pp. 63–76. doi: 10.1007/978-3-540-89350-9_6.

Leggett, M. and Bilda, Z. (2008) 'Exploring design options for interactive video with the Mnemovie hypervideo system', *Design Studies*. Elsevier Ltd, 29(6), pp. 587–602. doi: 10.1016/j.destud.2008.07.008.

Lei, P. L. *et al.* (2015) 'Effect of metacognitive strategies and verbal-imagery cognitive style on biology-based video search and learning performance', *Computers and Education*. Elsevier Ltd, 87, pp. 326–339. doi: 10.1016/j.compedu.2015.07.004.

Leiva, L. A. and Vivó, R. (2013) 'Web browsing behavior analysis and interactive hypervideo', *ACM Transactions on the Web*, 7(4), pp. 1–28. doi: 10.1145/2529995.2529996.

Lienhart, R. and Maydt, J. (2002) 'An extended set of Haar-like features for rapid object detection', *Proceedings. International Conference on Image Processing*, 1. doi: 10.1109/ICIP.2002.1038171.

Lienhart, R., Maydt, J. and Lienhartintelcom, R. (2002) 'An extended set of Haar-like features for rapid object detection', *Proceedings. International Conference on Image Processing*, 1, pp. 900–903. doi: 10.1109/ICIP.2002.1038171.

Lin, C. Y. (2004) 'Rouge: A package for automatic evaluation of summaries', *Proceedings of the workshop on text summarization branches out (WAS 2004)*, (1), pp. 25–26.

Luz, S. and Masoodian, M. (2004) 'A mobile system for non-linear access to time-based data', in *Proceedings of Advanced Visual Interfaces AVI'04*, pp. 454–457. doi: 10.1145/989863.989950.

Mackay, W. E. and Davenport, G. (1989) 'Virtual video editing in interactive multimedia applications', *Communications of the ACM*, 32(7), pp. 802–810. doi: 10.1145/65445.65447.

Manning, C. *et al.* (2014) 'The Stanford CoreNLP Natural Language Processing Toolkit', *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 55–60. doi: 10.3115/v1/P14-5010.

Marchionini, G. (2006) 'From finding to understanding', *Communications of the ACM*, 49(4), pp. 41–46.

Masneri, S. and Schreer, O. (2014) 'SVM-based Video Segmentation and Annotation of

Lectures and Conferences', in *Proceedings of the 9th International Conference on Computer Vision Theory and Applications*. Lison: IEEE, pp. 425–432. Available at: https://mail-attachment.googleusercontent.com/attachment/u/0/?ui=2&ik=32ee55125a&view=att&th=1428087d8bf5cacb&attid=0.1&disp=safe&realattid=f_hobm4c3m0&zw&saduie=AG9B_P-3dKmx79XijjC8coaDupgh&sadet=1385544126278&sads=lWMQ5gZzJKQyB0O-mKXXcW80Hvw.

Matejka, J., Grossman, T. and Fitzmaurice, G. (2014) 'Video Lens : Rapid Playback and Exploration of Large Video Collections and Associated Metadata', in *Uist*, pp. 541–550. doi: 10.1145/2642918.2647366.

Mauceri, C. *et al.* (2015) 'Evaluating visual query methods for articulated motion video search', *International Journal of Human Computer Studies*. Elsevier, 77, pp. 10–22. doi: 10.1016/j.ijhcs.2014.12.009.

McCandless, M., Hatcher, E. and Gospodnetic, O. (2010) *Lucene in Action, Second Edition: Covers Apache Lucene 3.0*. Greenwich, CT, USA: Manning Publications Co.

Mehta, H. *et al.* (2017) 'Metatation: Annotation as Implicit Interaction to Bridge Close and Distant Reading', *ACM Transactions on Computer-Human Interaction*, 24(5), pp. 1–41. doi: 10.1145/3131609.

Meixner, B. *et al.* (2014) 'Towards an easy to use authoring tool for interactive non-linear video', *Multimedia Tools and Applications*, 70(2), pp. 1251–1276. doi: 10.1007/s11042-012-1218-6.

Meixner, B. (2017) 'Hypervideos and Interactive Multimedia Presentations', *ACM Computing Surveys*, 50(1), pp. 1–34. doi: 10.1145/3038925.

Meixner, B. and Gold, M. (2016) 'Second-Layer Navigation in Mobile Hypervideo for Medical Training', in *MultiMedia Modeling: 22nd International Conference, MMM 2016*. Miami, pp. 382–394. doi: 10.1007/978-3-319-27671-7.

Merkt, M. *et al.* (2011) 'Learning with videos vs. learning with print: The role of interactive features', *Learning and Instruction*, 21(6), pp. 687–704. doi: 10.1016/j.learninstruc.2011.03.004.

Merkt, M. and Schwan, S. (2014) 'Training the use of interactive videos: Effects on mastering different tasks', *Instructional Science*. doi: 10.1007/s11251-013-9287-0.

Monserrat, T. *et al.* (2013) 'NoteVideo: facilitating navigation of blackboard-style lecture videos', in *CHI '13 Proceedings of the SIGCHI Conference on Human Factors in Computing*

Systems. Paris, pp. 1139–1148. doi: doi: 10.1145/2466110.2466147.

Moumtzidou, A., Avgerinakis, K. and Apostolidis, E. (2014) ‘VERGE : An Interactive Search Engine’, in *MultiMedia Modeling. MMM 2014. Lecture Notes in Computer Science*. Dublin, pp. 411–414.

Mujacic, S. *et al.* (2012a) ‘Modeling, design, development and evaluation of a hypervideo presentation for digital systems teaching and learning’, *Multimedia Tools and Applications*, 58(2), pp. 435–452. doi: 10.1007/s11042-010-0665-1.

Mujacic, S. *et al.* (2012b) ‘Modeling, design, development and evaluation of a hypervideo presentation for digital systems teaching and learning’, *Multimedia Tools and Applications*, 58(2), pp. 435–452. doi: 10.1007/s11042-010-0665-1.

Munzer, B. *et al.* (2017) ‘WHEN CONTENT-BASED VIDEO RETRIEVAL AND HUMAN COMPUTATION UNITE: TOWARDS EFFECTIVE COLLABORATIVE VIDEO SEARCH’, in *Proceedings of the IEEE International Conference on Multimedia and Expo Workshops (ICMEW) 2017*, pp. 214–219.

Nelson, T. H. (1965) ‘Complex information processing’, *Proceedings of the 1965 20th national conference on* -, pp. 84–100. doi: 10.1145/800197.806036.

Neto, C. de S. S. and Soares, L. F. G. (2009) ‘Reuse and imports in Nested Context Language’, *Proceedings of the XV Brazilian Symposium on Multimedia and the Web - WebMedia ’09*, pp. 1–8. doi: 10.1145/1858477.1858497.

Nicolaescu, P. and Siddiqui, A. (2017) ‘Emerging Technologies for Education - iPad’, pp. 533–543. doi: 10.1007/978-3-319-52836-6.

O’Brien, H. L. and Toms, E. G. (2013) ‘Examining the generalizability of the User Engagement Scale (UES) in exploratory search’, *Information Processing and Management*. Elsevier Ltd, 49(5), pp. 1092–1107. doi: 10.1016/j.ipm.2012.08.005.

Oskouie, P. (2012) ‘Multimodal feature extraction and fusion for semantic mining of soccer video: a survey’, *Artificial Intelligence …*, pp. 173–210. doi: 10.1007/s10462-012-9332-4.

Pavel, A. *et al.* (2014) ‘Video digests: A browsable, skimmable format for informational lecture videos’, in *Symposium on User interface software and technology, USA*, pp. 573–582. doi: 10.1145/2642918.2647400.

Pavel, A. and Reed, C. (2014) ‘Video Digests : A Browsable , Skimmable Format for

Informational Lecture Videos’.

Petan, A. S., Petan, L. and Vasiu, R. (2014) ‘Interactive Video in Knowledge Management: Implications for Organizational Leadership’, *Procedia - Social and Behavioral Sciences*, 124, pp. 478–485. doi: 10.1016/j.sbspro.2014.02.510.

Piketty, T. (2014) *New thoughts on capital in the twenty-first century*, *TED*. Available at: https://www.ted.com/talks/thomas%5C_piketty%5C_new%5C_thoughts%5C_on%5C_capital%5C_in%5C_the%5C_twenty%5C_first%5C_century (Accessed: 1 November 2017).

Qin, J. *et al.* (2017) ‘Fast action retrieval from videos via feature disaggregation’, *Computer Vision and Image Understanding*. Academic Press, 156, pp. 104–116. doi: 10.1016/j.cviu.2016.09.009.

Ramezani, M. and Yaghmaee, F. (2016) ‘A novel video recommendation system based on efficient retrieval of human actions’, *Physica A: Statistical Mechanics and its Applications*. Elsevier B.V., 457, pp. 607–623. doi: 10.1016/j.physa.2016.03.101.

Ratinov, L. and Roth, D. (2009) ‘Design challenges and misconceptions in named entity recognition’, in *Proceedings of the Thirteenth Conference on Computational Natural Language Learning - CoNLL ’09*, p. 147. doi: 10.3115/1596374.1596399.

Rogers, Y. *et al.* (2012) *HCI Theory: Classical, Modern, and Contemporary*, *Synthesis Lectures on HumanCentered Informatics*. doi: 10.2200/S00418ED1V01Y201205HCI014.

Ruotsalo, T. *et al.* (2015) ‘Interactive Intent Modeling: Information Discovery Beyond Search’, *Communications of the ACM*, 58(1), pp. 86–92. doi: 10.1145/2656334.

Sadallah, M., Aubert, O. and Prié, Y. (2012) ‘CHM: an annotation- and component-based hypervideo model for the Web’, *Multimedia Tools and Applications*, pp. 1–35. doi: 10.1007/s11042-012-1177-y.

Sauli, F., Cattaneo, A. and van der Meij, H. (2017) ‘Hypervideo for educational purposes: a literature review on a multifaceted technological tool’, *Technology, Pedagogy and Education*. Routledge, 5139, pp. 1–20. doi: 10.1080/1475939X.2017.1407357.

Schoeffmann, K. *et al.* (2017) ‘Collaborative Feature Maps for Interactive Video search’, in *MMM 2017*. Springer, pp. 457–462. doi: 10.1007/978-3-319-51814-5.

Schoeffmann, K. and Hudelist, M. A. (2015) ‘Video Interaction Tools : A Survey of Recent Work’, *ACM Computing Surveys*, 48(1). doi: 10.1145/2808796.

Schoeffmann, K., Taschwer, M. and Boeszoermenyi, L. (2010) ‘The Video Explorer – A Tool for Navigation and Searching within a Single Video based on Fast Content Analysis’, in *Proceedings of the first annual ACM SIGMM conference on Multimedia systems - MMSys ’10*, pp. 247–258. doi: 10.1145/1730836.1730867.

Schuller, B. *et al.* (2013) ‘The INTERSPEECH 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism’, *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pp. 148–152.

Scotto di Carlo, G. (2014) ‘The role of proximity in online popularizations: The case of TED talks’, *Discourse Studies*, 16(5), pp. 591–606. doi: 10.1177/1461445614538565.

Shen, J. and Cheng, Z. (2010) ‘Personalized video similarity measure’, *Multimedia Systems*, 17(5), pp. 421–433. doi: 10.1007/s00530-010-0223-8.

Shipman, F., Girgensohn, A. and Wilcox, L. (2008) ‘Authoring, viewing, and generating hypervideo’, *ACM Transactions on Multimedia Computing, Communications, and Applications*, 5(2), pp. 1–19. doi: 10.1145/1413862.1413868.

Shneiderman, B. (2003) ‘The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations’, *The Craft of Information Visualization*, pp. 364–371. doi: 10.1016/B978-155860915-0/50046-9.

Sorin, A., Petan, L. and Vasiu, R. (2014) ‘Interactive video in knowledge management : Implications for organizational leadership’, *Procedia - Social and Behavioral Sciences*. Elsevier B.V., 124(2001), pp. 478–485. doi: 10.1016/j.sbspro.2014.02.510.

Stahl, E., Finke, M. and Zahn, C. (2006) ‘Knowledge Acquisition by Hypervideo Design : An Instructional Program for University Courses’, *Journal of Educational Multimedia and Hypermedia*, 15, pp. 285–302.

Steinbock, D. (2016) *http://tagcrowd.com/*. Available at: http://tagcrowd.com/.

Sugimoto, C. R. *et al.* (2013) ‘Scientists Popularizing Science: Characteristics and Impact of TED Talk Presenters’, *PLoS ONE*, 8(4). doi: 10.1371/journal.pone.0062403.

Tan, S. *et al.* (2014) ‘Cross domain recommendation based on multi-type media fusion’, *Neurocomputing*. Elsevier, 127, pp. 124–134. doi: 10.1016/j.neucom.2013.08.034.

Tiellet, C. A. B. *et al.* (2010) ‘Design and evaluation of a hypervideo environment to support veterinary surgery learning’, in *HT ’10: Proceedings of the 21st ACM conference on Hypertext* *and hypermedia*, pp. 213–222. doi: 10.1145/1810617.1810656.

Tsukuda, K., Masahiro, H. and Goto, M. (2017) ‘SmartVideoRanking: Video Search by Mining Emotions from Time-Synchronized Comments’, *IEEE International Conference on Data Mining Workshops, ICDMW*, pp. 960–969. doi: 10.1109/ICDMW.2016.0140.

Velasco, R. (2016) *Apache Solr: For Starters*. CreateSpace Independent Publishing Platform.

Waitelonis, J. and Sack, H. (2012) ‘Towards exploratory video search using linked data’, *Multimedia Tools and Applications*, 59(2), pp. 645–672. doi: 10.1007/s11042-011-0733-1.

Wang, J. *et al.* (2008) ‘A multimodal scheme for program segmentation and representation in broadcast video streams’, in *IEEE Transactions on Multimedia*, pp. 393–408. doi: 10.1109/TMM.2008.917362.

Yadav, K. *et al.* (2015) ‘Content-driven Multi-modal Techniques for Non-linear Video Navigation’, *Proceedings of the 20th International Conference on Intelligent User Interfaces - IUI ’15*, 2015–Janua, pp. 333–344. doi: 10.1145/2678025.2701408.

Zhang, D. and Nunamaker, J. F. (2004) ‘A natural language approach to content-based video indexing and retrieval for interactive e-Learning’, *IEEE Transactions on Multimedia*, 6(3), pp. 450–458. doi: 10.1109/TMM.2004.827505.

Zhang, H., Liu, Y. and Ma, Z. (2013) ‘Fusing inherent and external knowledge with nonlinear learning for cross-media retrieval’, *Neurocomputing*. Elsevier, 119, pp. 10–16. doi: 10.1016/j.neucom.2012.03.033.

Zhang, Y., Zhang, L. and Zimmermann, R. (2015) ‘Aesthetics-Guided Summarization from Multiple User Generated Videos’, *ACM Transactions on Multimedia Computing, Communications, and Applications*, 11(2), pp. 1–23. doi: 10.1145/2659520.

10. Appendices

10.1. Appendix A: Phase 2, User Satisfaction Questionnaire.

Participant Feedback Questionnaire:

The representation is easy to comprehend.

Strongly Disagree	Disagree	Don't know	Agree	Strongly Agree
1	2	3	4	5

I found the representation to be flexible to interact with

Strongly Disagree	Disagree	Don't know	Agree	Strongly Agree
1	2	3	4	5

Were the multimodal ingredients which made up the content presentation, useful for you in getting your intended information?

Not at all useful	Not very useful	Don't know	Somewhat useful	Very Useful
1	2	3	4	5

The multimodal ingredient sliced the video in sensible slices.

Strongly Disagree	Disagree	Don't know	Agree	Strongly Agree
1	2	3	4	5

There were distinct differences between the different slices you were shown.

Strongly Disagree	Disagree	Don't know	Agree	Strongly Agree
1	2	3	4	5

Assuming there are distinct differences, those differences were shown in a clear and easy to grasp manner.

Strongly Disagree	Disagree	Don't know	Agree	Strongly Agree
1	2	3	4	5

The multimodal ingredients making up the segments were presented in an easy to consume manner.

Strongly Disagree	Disagree	Don't know	Agree	Strongly Agree
1	2	3	4	5

Overall, I am satisfied with the ease of completing the tasks in this scenario

Strongly Disagree	Disagree	Don't know	Agree	Strongly Agree
1	2	3	4	5

Overall, I am satisfied with the amount of time it took to complete the task in this scenario

Strongly Disagree	Disagree	Don’t know	Agree	Strongly Agree
1	2	3	4	5

Using the system would enable me to accomplish the task more quickly.

Strongly Disagree	Disagree	Don’t know	Agree	Strongly Agree
1	2	3	4	5

Post Experiment Interview Question:

What kind of information you would like to see in such a presentation?
Would you prefer to scroll through different slices manually at your own pace or do you think it will be more efficient if there is some automatic or semi-automatic transition or skimming of different slices?

10.2. Appendix B: Phase 2, Experiment 2, Sample of Annotated Data

Feedback captured for User # 7.

0-1.	
Speech.	Action.
I think I will chose this pic (graph) because it is a graph.	Sum of seg1 highlighted. Go to next seg. Highlighted. Vis of seg. Go to next seg. Not highlighted.

1-2	
Speech.	Action.
	Sum of seg not highlighted. Go to next seg. Not highlighted. Vis of seg. Go through vis of seg. Sum of seg. Vid of seg.

2-3	
Speech.	Action.
I will not choose this vid because I do not understand the accent. I will choose this (vis with graph).	Go through vis of seg. Not highlighted. Go to next seg. Not highlighted. Vis of seg. Go through vis of seg.

3-4	
Speech.	Action.
It is quick and it is catchy.	Go to next seg not highlighted. Vis of seg. Go through vis of seg. Sum of seg. Go to next seg. Highlighted. Go through vis of seg. Go to next seg. Highlighted. Go through vis of seg. Sum of seg.

4-5	
Speech.	Action.
This summary is interesting because it resume what inequality.	Sum of seg. Highlighted. Go to next seg. Highlighted. Go through vis of seg. Sum of seg. Term of seg.

	Go to next seg. Not highlighted. Last Term of seg. Go through vis of seg. //move to next. Vid 2.

5-6	
Speech.	Action.
	//vid 2. Term of seg1 highlighted. Try to go through vis of seg. Go to next seg. Highlighted. Term of seg. Sum of seg. Go to next seg. Not highlighted. Sum of seg. Term of seg. Sum of seg.

6-7	
Speech.	Action.
	Sum of seg. Not highlighted. Go to next seg. Highlighted Term of seg. Sum of seg. Go to next seg. Highlighted. Term of seg. Sum of seg.

7-8	
Speech.	Action.
This one is good because it give a reason to inequality	Sum of seg. Highlighted. Go to next seg. Not highlighted. Term of seg. Sum of seg. Go to next seg. Not highlighted. Term of seg. Go to next seg. Not highlighted. Term of seg. Sum of seg. Go to next seg. Not highlighted. Term of seg. Sum of seg.

8-9	
Speech.	Action.
I like this one too because it give idea about poverty and middle class.	Sum of seg. Not highlighted. Go to next seg. Not highlighted.

The terms are good that is why I am looking at this summary.	Term of seg. Go to next seg. highlighted. Term of seg. Go to next seg. highlighted. Term of seg. Sum of seg.

9-10	
Speech.	Action.
Summary is good too it speaks about education. It relies to the subject searched for.	Sum of seg. highlighted. Go to next seg. Not highlighted. Term of seg. Sum of seg.

10-11	
Speech.	Action.
	Sum of seg. highlighted. Go to next seg. Not highlighted. Term of seg. Go to next seg. Not highlighted. Term of seg. Go to next seg. Not highlighted. Term of seg.

Questionnaire

Diff in slice could be better. How terms are differencing in slices.

Term to topic should be bigger.

1) Visual good for video. Term should be filtered. Term should be more relevant.
2) Manual.

Vis with graph is better in absence terms are good.

10.3. Appendix C: Phase 3, Experiment 3 User Experience Questionnaire

Please make your evaluation now.

For the assessment of the product, please fill out the following questionnaire. The questionnaire consists of pairs of contrasting attributes that may apply to the product. The circles between the attributes represent gradations between the opposites. You can express your agreement with the attributes by ticking the circle that most closely reflects your impression.

Example:

attractive	○	⊗	○	○	○	○	○	unattractive

This response would mean that you rate the application as more attractive than unattractive.

Please decide spontaneously. Don't think too long about your decision to make sure that you convey your original impression.

Sometimes you may not be completely sure about your agreement with a particular attribute or you may find that the attribute does not apply completely to the particular product. Nevertheless, please tick a circle in every line.

It is your personal opinion that counts. Please remember: there is no wrong or right answer!

Please assess the product now by ticking one circle per line.

	1	2	3	4	5	6	7		
annoying	○	○	○	○	○	○	○	enjoyable	1
not understandable	○	○	○	○	○	○	○	understandable	2
creative	○	○	○	○	○	○	○	dull	3
easy to learn	○	○	○	○	○	○	○	difficult to learn	4
valuable	○	○	○	○	○	○	○	inferior	5
boring	○	○	○	○	○	○	○	exciting	6
not interesting	○	○	○	○	○	○	○	interesting	7
unpredictable	○	○	○	○	○	○	○	predictable	8
fast	○	○	○	○	○	○	○	slow	9
inventive	○	○	○	○	○	○	○	conventional	10
obstructive	○	○	○	○	○	○	○	supportive	11
good	○	○	○	○	○	○	○	bad	12
complicated	○	○	○	○	○	○	○	easy	13
unlikable	○	○	○	○	○	○	○	pleasing	14
usual	○	○	○	○	○	○	○	leading edge	15
unpleasant	○	○	○	○	○	○	○	pleasant	16
secure	○	○	○	○	○	○	○	not secure	17
motivating	○	○	○	○	○	○	○	demotivating	18
meets expectations	○	○	○	○	○	○	○	does not meet expectations	19
inefficient	○	○	○	○	○	○	○	efficient	20
clear	○	○	○	○	○	○	○	confusing	21
impractical	○	○	○	○	○	○	○	practical	22

organized	○	○	○	○	○	○	○	cluttered	23
attractive	○	○	○	○	○	○	○	unattractive	24
friendly	○	○	○	○	○	○	○	unfriendly	25
conservative	○	○	○	○	○	○	○	innovative	26

10.4. Appendix D: Experiment 3, Sample of Multiple Choice Questions for Answer Search Task

MCQ for Ted Talk Titled: “New thoughts on capital in the twenty-first century” by Thomas Piketty

The Video has been segmented into 10 smaller segments. Following are some questions which can be answer with content presented in the video. Please read the questions write down the segment in which you think you are most likely to find the answer to the question.

1. In what segment(s) does the author reiterates the recurring theme of this talk.

   Segment # _______________

2. The factors in income inequality being higher in US compared to Europe.

   Segment # _______________

3. What made the swiss show flexibility in bank secrecy?

   Segment # _______________

4. Which one is a criticism on the presenter thesis

   Segment # _______________

5. What is the Data source used by the presenter?

   Segment # _______________

6. Some economists argue in support of inequality that it’s an engine of capitalism

   Segment # _______________

7. The least efficient way of decreasing inequality is starting wars

   Segment # _______________

8. The growth rate of economy has been unusually high in certain countries during

   Segment # _______________

9. The World wars and their aftermath has Decreased inequality

   Segment # _______________

10. Decrease in Capital Gains has caused an increase in economic growth

    Segment # _______________

11. In Pre-Industrial society the growth rate of economy was traditionally close to zero.

    Segment # _______________

12. In 21st century the Top 10% population has the following share of global income.

    Segment # _______________

13. Jane Austen is mentioned in.

    Segment # _______________

14. In 21st century the Top 10% population has the following share of global wealth.

    Segment # _______________

10.5. Appendix E: Experiment 3, Sample of User Action Annotations.

Minute: 0-1	
Action System	Action task
System: RAAVE Seg 1 high sum (pri)	Task: Questions, Video: CF Attempt 1 starts at 00:44

Minute: 1-2	
Action System	Action task
Seg 2 high sum (pri) Seg 3 not high word cloud Seg 4 high sum (pri) Seg 5 high sum (pri)	

Minute: 2-3	
Action System	Action task
Seg 6,7,8,9 slow scroll word cloud (pri) Seg 10 high sum (pri) Seg 12 high sum (pri) Scroll to end Scrolling Seg 12 high sum (pri) Seg 9 not high sum (sci) Seg 6 not high sum (sci)	

Minute: 3-4	
Action System	Action task
Scrolling around Seg 10 high sum (pri) Scrolling Seg 12 high sum (pri) Seg 12 word cloud (sci) Seg 12 sum (pri) Scollring around Seg 9 not high word clould (pri)	Ans. Q.1 at 03:49

Minute: 4-5	
Action System	Action task
Seg 9 sum (sci) Seg 7 not high word cloud (pri) Seg 7 sum (sci) Seg 6 not high word cloud (pri)	Ans. Q.2 at 4:17 Ans. Q.3 at 4:49

Seg 6 sum (sci) Scrolling Seg 2 high word cloud (sci) Slow scrolling Seg 6 not high sum (sci) scrolling	

Minute: 5-6	
Action System	Action task
Scrolling Scrolling around sums and word clouds Seg 6 not high sum sci Scrlling around Seg 1 high trans pri Seg 2 high trans pri Seg 3 not high trasn (sci)	Ans. Q.4 at 05:06

Minute: 6-7	
Action System	Action task
Seg 4,5 high trans (pri) Seg 6,7,8 word cloud (pri) Seg 10 high trans (pri) Seg 11 word cloud pri Seg 12 high trans pri Seg 12 sum pri Scoll Seg 10 high trans pri Seg 10 word cloud sci	Ans. Q.5 at 06:37

Minute: 7-8	
Action System	Action task
Seg 10 word cloud sci cont.. Seg 9 word clould pri Seg 8 word clould pri Seg 8 sum sci Seg 8 trans Seg 7 word cloud pri Scrolling through word cloud Seg 7,8,9,10 Seg 11,16 word clouds Seg 15 sum (sci)	Q.6 at 07:28
Minute: 8-9	
Action System	Action task
Seg 16 sum (sci) Seg 15 sum (sci)	

Seg 14 trans (sci) Seg 13 trans (sci) Seg 12 trans (pri) Seg 11 trans (sci) Seg 10 trans (pri) Seg 7 word cloud (pri)	

Minute: 9-10	
Action System	Action task
Seg 7 trans (sci) Seg 6 sum (sci) Seg 6 trans Scrolling up swtich to summary along to way Seg 1 high sum Seg 1 trans pri	Ans. Q.7 at 09:28

Minute: 10-11	
Action System	Action task
Seg 1 trans pri Seg 2 tran pri Seg 2 word cloud Scrolling Seg 3 tran (sci)	Ans. Q.9 at 10:54

Minute: 11-12	
Action System	Action task
Seg 4 sum pri Seg 5 trans pri Scrolling up Seg 2 high sum (pri) Seg 1 trans (pri) Seg 1,2,3 wordclouds Seg 2 trans (pri)	Ans. Q.8 at 11:17 Ans. Q.10 at 11:32 Ans. Q.12 at 11:39

Minute: 12-13	
Action System	Action task
Scrolling Seg 4 word cloud sci Seg 6 sum (sci) Seg 6 trans Seg 7 trns (sci) scolling	Ans. Q.12 at 12:03

Minute: 13-14	
Action System	Action task
Seg 10 sum pri Seg 11 word cloud pri Seg 12 word clould sci Seg 12 trans pri Seg 13 tans (sci) Seg 14 not high Seg 14 word pri Seg 14 sum sci	Ans. Q.13 at 13:47

Minute: 14-15	
Action System	Action task
Seg 14 trans sci Seg 15 trans (sci) Seg 16 trans (sci)	Ans Q.13 at 14:06 (second) Ans. Q.14 at 14:24 Attempt 1 ends at 14:27

Minute: 15-16	
Action System	Action task
System: Baseline Video play at 15:47	Task: Summary, Video: TP Attempt 2 starts at 15:44

Minute: 16-17	
Action System	Action task
Play speed 1.5x at 16:02 Play speed 1.25x at 16:36	Taking notes

Minute: 17-18	
Action System	Action task
 seek	Taking notes

Minute: 18-19	
Action System	Action task
 seek	Taking notes

Minute: 19-20	
Action System	Action task
Seek to end part Video pause at 19:57	Taking notes Time up at 19:57

Minute: 20-21	
Action System	Action task
	Continue writing synop at 20:03

Minute: 21-22	
Action System	Action task
	Continue writing synop

Minute: 22-23	
Action System	Action task
	Continue writing synop Attempt 2 finish at 23:25

Minute: 23-24	
Action System	Action task

Minute: 24-25	
Action System	Action task
System: Baseline Video play at 24:37 Play speed at 1.25x at 24:42 seek	Task: Questions, Video: PC Attempt 3 starts at 24:35

Minute: 25-26	
Action System	Action task
Pause at 25:03 Play at 25:58	Reading questions

Minute: 26-27	
Action System	Action task
Play speed 1.5x at 26:11	Reading questions Ans. Q.13 at 26:21

Minute: 27-28	
Action System	Action task
	Ans. Q.5 at 27:31 Ans. Q.6 at 27:40

Minute: 28-29	
Action System	Action task
Pause play	Ans. Q.8 at 28:32 Ans. Q.4 at 28:52

Minute: 29-30	
Action System	Action task
	Ans. Q.12 at 29:27

Minute: 30-31	
Action System	Action task
	Ans. Q.3 at 30:19

Minute: 31-32	
Action System	Action task
	Ans. Q.2 at 31:27

Minute: 32-33	
Action System	Action task
	Ans. Q.11 at 32:25 Ans. Q.10 at 32:49

Minute: 33-34	
Action System	Action task
	Ans. Q.9 at 33:50

Minute: 34-35	
Action System	Action task
Play speed 2x at 34:13	Ans. Q.8 at 34:47

Minute: 35-36.	
Action System	Action task

Minute: 36-37	
Action System	Action task
Video ends at 36:19	Ans. Q.14 at 36:12 Attempt 3 ends at 36:21

Minute: 37-38	
Action System	Action task
System: RAAVE Seg 1 sum sci Seg 2,3,4,5,6,7 not high sum (sci)	Task: Summary, Video: ED Attempt 4 starts at 37:27

Minute: 38-39	
Action System	Action task
Seg 7 sum sci cont.. Seg 8,9,10,11 not high, sum (sci)	

Minute: 39-40	
Action System	Action task
Seg 11 sum sci cont.. Seg 12,13,14,15,16,17,18 not high sum sci	Taking notes

Minute: 40-41	
Action System	Action task
Scrolling up slow Scrolling slow	Taking notes cont.. Time up at 40:48

Minute: 41-42	
Action System	Action task
	Continue writing synop at 41:02

Minute: 42-43	
Action System	Action task

Minute: 43-44	
Action System	Action task
	Attempt 4 ends at 43:23

10.6. Appendix F: Experiment 3: Sample of Social Skip User Action Log (Baseline System)

	Time	TransactionId	TransactionTime	Transaction	SkipTime
301	183	1	19-09-17 0:23	Backward	7
301	174	1	19-09-17 0:23	Backward	11
301	56	1	19-09-17 0:26	Backward	247
301	586	1	19-09-17 0:30	Backward	17
301	573	1	19-09-17 0:30	Backward	15
301	683	1	19-09-17 0:31	Backward	11
43	95	1	25-09-17 4:44	Backward	9
43	132	1	25-09-17 4:45	Backward	13
43	121	1	25-09-17 4:45	Backward	23
43	215	1	25-09-17 4:46	Backward	7
43	245	1	25-09-17 4:47	Backward	0
43	245	1	25-09-17 4:47	Backward	0
43	252	1	25-09-17 4:47	Backward	0
43	338	1	25-09-17 4:48	Backward	0
43	379	1	25-09-17 4:48	Backward	0
43	455	1	25-09-17 4:49	Backward	0
43	490	1	25-09-17 4:49	Backward	7
43	490	1	25-09-17 4:49	Backward	0
43	667	1	25-09-17 4:50	Backward	17
43	656	1	25-09-17 4:50	Backward	27
43	646	1	25-09-17 4:50	Backward	18
43	635	1	25-09-17 4:50	Backward	29
43	711	1	25-09-17 4:51	Backward	12
43	700	1	25-09-17 4:51	Backward	23
43	799	1	25-09-17 4:52	Backward	0
43	887	1	25-09-17 4:53	Backward	0
43	921	1	25-09-17 4:53	Backward	0