Human Computation in Online Video Storytelling

Philo D. I. van Kemenade

Submitted in partial fulfillment of the requirement of the degree of
Master of Science in Cognitive Computing

Goldsmiths College
University of London

Supervisor
Dr. Marian Ursu

Department of Computing
Goldsmiths, University of London
New Cross, London SE14 6NW

September 21, 2012

Abstract

Tasks like retrieval, filtering and reconfiguration of digital video are difficult to solve using current computational techniques. An important cause of this difficulty is the semantic gap between visual representations and the meaning we address to them. A solution commonly sought in AI research is to reduce the gap by feature extraction followed by supervised learning of semantical concepts that are labelled to content. These methods often fail to work both reliably and generally on the unpredictable content found in the large video libraries of user generated content that account for much of the internet traffic these days. Another way of hunting down meaning in visual content is to step over the gap altogether and ask people for a meaningful interpretation one wishes to acquire for an item of content. By accessing many people's interpretations in small bite-sized tasks, collectively grounded annotations can be established. This form of accessing human computational power has seen a major increase in attention and application, for a large part because of the increased connectivity of individuals to the web and the surging amount of visual content that is uploaded. This thesis investigates how tasks involving meaningful interpretation of video content can benefit from the use of human computation. In order to test the validity of these approaches 'wePorter' is developed, a system with the purpose of finding local intervals of interest within videos in a set of topically related content. We also investigate how such a system can be used for reconfiguration of content into new and informative stories. We introduce 'parallel play' as a useful method for user interest elicitation in time-based media and present our results of reconfiguration of video parts, filtered based on users' attentional data.

Contents

Chapter 1

Introduction

This section introduces the problem of visual information overload, hints at current methods to solve the problem and indicates why they are not satisfactory for the wide domain of user generated video content that accounts for unprecedented amounts of data and traffic on the web. The idea of human computation is introduced and hinted to as a possible solution. The particular problem of filtering segmented video parts based on interest is introduced.

Since the increase of bandwidth and connectivity to the web as well as proliferation of online tools and platforms to share content online, much of the web's content is visual. In contrast to textual data that is symbolic and machine readable, the meaning of visual content resides mostly in sensory data (such as sound and visuals). Sensory data needs to be interpreted and thus needs advanced techniques from fields like computer vision and machine learning to process the data.

Chapter 2

The Quest for Meaning in Video

2.1 Introduction

Interpreting moving images is not a hard task. The medium film is often described as 'dictatorial' because of the way the audience is immersed in a multi-modal experience controlled by the content's creators. When watching a film, we sit back and relax, passively taking in the presented information without much effort. A similar ease is reflected in our use of the word 'couch potato' to describe the passive role of television audiences. Watching film or video gives us almost immediate access to a wide range of information about what is presented on screen. We recognise objects on screen and understand words that spoken in a language we know. We are also quick to infer a larger picture around the things we perceive, like personality traits of characters on screen and our emotional stance towards them. While most of these things happen extremely quickly and seemingly automatically to us humans, computers often have a hard time even starting to perceive a visual representation of an object.

When we attend to visual content depicting parts of the world around us, we can't help ourselves from seeing its parts as separate entities. We recognise objects as if they stand out from their background even though they are simply patterns of colours on a two dimensional surface. To a computer, tasks like object segmentation and recognition are hard because visual information needs to be interpreted in some form of sequential processing. Digitally, images are usually represented by collections of numbers indicating local intensities (e.g. colour or brightness) at the different points that make up the image. How to calculate from this information, which objects are present, and what other concepts can be assigned to an image is studied in the field of computer vision. The task most related to finding computational interpretations of video content is video concept detection. Although recent years have seen important advances in the use of high-level semantical concepts in tasks like concept detection and concept-based video retrieval [37, 36, **?**, **?**], computational methods commonly have difficulties in performing both reliably and generally.

Because of the often elusive character of concepts like meaning and understanding, goals for this chapter are kept intensionally modest. The intension is not to give an accurate explanation of daunting concepts like meaning or semantics, nor is it to give an accurate account of the diverse work on signifying systems such as in the field of semiotics. This first chapter is meant to briefly introduce the difficulties that current computational methods have in arriving at meaningful interpretations of visual content. To this purpose we formulate a framework of computational interpretation of visual content that serves to establish terminology to work with in this work, rather than to make claims about

the deeper functioning of human understanding or signifying systems. The next section addresses two high-level challenges to the goal of finding meaning in video and indicates how they arise. Of these, the *semantic gap* is the most poignant and we take a look at how computational approaches aim to overcome this problem. The chapter concludes by pointing out outstanding challenges and hinting at a different solution that might step across the semantic gap altogether.

2.2 Challenges in Computational Interpretation of Visual Content

As video content possesses most of its information in the visual stream, most research into the interpretation of video has focussed on the analysis of visual content [37, ch. 2]. To better understand what is going on in the interpretation of visual content by both humans and computers, it helps to model the process from start to end. Figure 2.1 shows in a high-level model how objects in the world are sensed and consequently rendered in a visual representation. We can think of this process taking place when we photograph a car and end up with a picture of that car as a result. When the representation of an object is next interpreted by someone, we can think of this person as establishing semantical concepts relating to aspects of the depiction. A situation to which this part of the model applies would be someone looking at the picture and recognising the car.

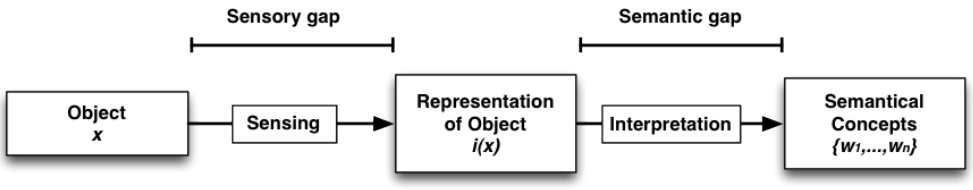


Figure 2.1: A high-level model of the interpretation of visual media content

A first source of complication in the process from object to its interpretation, is the *sensory gap*, described by Smeulders et al. as follows:

> "The sensory gap is the gap between the object in the world and the information in a (computational) description derived from a recording of that scene."[35]

The sensory gap makes accurate description of objects in the world difficult as it introduces uncertainty about what aspects of the object are represented. Characteristics of illumination, occlusion, clutter and camera viewpoint all affect the representation of a sensed object. When detailed knowledge about the recording conditions is absent, it is impossible to know which parts of the sensory information should be attributed to the state of the object and which are due to incidental artefacts. Different 3D objects can yield the same 2D representation and differently coloured objects might be represented by identical colour values. This also works the other way, as one object may appear very different in shape and colour on different images depending on illumination and camera viewpoint.

A second and more challenging issue that hinders meaningful computational interpretation of visual content is the *semantic gap* that lies between a digital representation and

the conceptual interpretation we address to it. Snoek and Worring adapt the original definition from [35] to specifically fit the medium video when they describe the semantic gap as:

> "The lack of correspondence between the low-level features that machines extract from video and the high-level conceptual interpretations a human gives to the data in a given situation."[37]

One of the causes of the semantic gap is that the way people perceive images is mostly contextual[35]. We look for concepts that are already familiar from our environment or earlier encounters with visual content. Our perception of a simple object is determined by our vast background of personal experience and cultural upbringing. In contrast to these contextual interpretations, computational image descriptions rely purely on data-driven features that can be extracted from the content. Difficulties arise when there is a mismatch between the two.

Another cause of the gap are interpretations that are subjective in nature. Semantical concepts relating to feelings and emotions can vary widely across different people. Deciding computationally whether concepts such as "romantic" or "funny" apply to a piece of content is hard when there is no agreement about the interpretation to begin with.

Perceived concepts are also combined to infer a larger story around the things we actually see. These knowledge-based interpretations enable us to perceive deeper layers of meaning that are not in itself explicitly represented. An important example of this is the way 'readers' of narrative texts or moving images combine elements in their aim for *coherence*[**?**, p. 38] [**?**, **?**]. High-level concepts like coherence over time are usually not explicitly represented in digital content and can be hard to compute algorithmically.

Even if there is little context dependency in the perception or recognition of an object in video, it might still be hard for computational methods to produce appropriate semantical labels. This is due to the wide variety in appearance of visual concepts. Determining whether a clock is present in a video can be difficult because of the many different sizes, shapes and colours clock can have.

All of these issues contribute to the gap between the low-level features extracted from video and the interpretations humans give to them. Challenges posed by the semantic gap are of mayor concern to the research community focussing on multimedia retrieval based on querying by user defined semantical concepts. The challenge is thus relevant to different scientific disciplines such as computer vision, information retrieval, machine learning and human-computer interaction. The next section briefly reviews computational strategies that aim to narrow the semantic gap.

2.3 Computational Undertakings of the Quest for Meaning

This section gives a short overview of strategies to narrow the semantic gap that is apparent in the computational interpretation of video content. At the core of this quest for meaning is the task of concept detection[37], where video clips are analysed to automatically detect whether a certain concept is present. Another step that is commonly taken to go from low-level video features to semantical interpretations is a classification of the type of content. This classification can be done at different levels, ranging from general and conceptually low-level (a scene containing music) to specific and conceptually high-level (a rock concert at an outdoor festival)[**?**].

Before starting the processes of classification or concept detection, videos are usually segmented into smaller clips. The most common unit for temporal video segmentation is

the *shot*, one continuously recorded interval in the same setting of time and place. Shot segmentation is a well-understood problem and efficient automatic methods exist [TODO ref 'automatic partitioning of full-motion video', 'a formal study of shot boundary detection']. Another form of partitioning is to segment the video into *scenes*, possibly consisting of multiple shots, signifying a unit within a story[**?**]. While shot segmentation can be done automatically thanks to data-driven procedures, the task of scene segmentation relies on semantical and narrative interpretations of the content and is thus a lot harder to solve computationally.

The tasks of video classification and video concept detection, are generally organised as follows. For a video segment or keyframe i, represented by n-dimensional feature vector x_i, a measure is calculated that indicates whether conceptual label ω_j applies to shot i (concept w_j is present in i or i can be classified as being of type w_j). The common paradigm to find the relation between x_i and ω_j is supervised learning. Supervised learning methods use a large number of examples in a training phase to find an optimal combination of features that codes for the presence of a particular concept. Using the found relationship from features to conceptual label, previously unseen instances can then be classified with a certain accuracy. This section briefly addresses different features that can be extracted from video content and explains the general framework of supervised learning.

2.3.1 Feature Extraction

Video content has a multi-modal nature, and may consist of a recorded visual stream, animations, recorded or synthesised sounds, spoken language and textual information in (sub)titles, all presented in a sequential format over time. This rich nature of the medium makes that there are many different types of features that can be extracted from a piece of content.

To help alleviate the semantic gap between low-level features and high-level interpretations, features should have enough discriminatory power to distinguish between the appearances of different concepts. Due to the sensory gap, variations in appearance also exist that are not caused by a difference in semantics, but are rather induced by the recording conditions. Features need so have a sufficient level of *invariance* to these accidental visual distortions introduced by the sensory gap[35]. A higher level of invariance in the description of concept w_j means the concept will be detected across a variety of different recording conditions. On the other hand the invariance might cause concept w_j to be detected in the representation of other concepts with a similar appearance. Invariance thus comes at the cost of discriminatory power. In the choice of a feature set a balance should be sought between invariance and discrimination that is suitable for the particular domain of content and application. Most focus in feature extraction is on visual features, and we will start by indicating the types of features that are in use.

Visual Features

Despite the different modalities that can collectively make up a piece of video, it's defining characteristic is the presence of a sequence of images. Most efforts to narrow the semantic gap in video systems focus on the visual modality and try to make use of the features that can be extracted from it. In their wide-ranging overview of concept-based video retrieval techniques, Snoek and Worring point to the following types of visual features that are used in video concept detection[37].

- *Colour* - Colour can generally be represented in different 3D colour spaces (e.g. rgb, hsv or lab) and has discriminating potential superior to the single dimensional

greyscale domain. In [35] Smeulders et al. indicate two aspects that have to be considered when working with colour features. First is the considerable variability in appearance of coloured surfaces under different recording circumstances, contributing to the sensory gap. Second is the intricacy of human colour perception that has to be accounted for in addressing visual interpretations approaching those brought forth in human experience.

- *Texture* - While colour features can be calculated for every pixel in an image, texture features look at regions of multiple pixels to determine local patterns. Texture features are used to describe different materials or surfaces, for example the fine grained texture of sand versus the linear texture of hairs. A common practice is to capture directional patterns of texture using localised derivatives of changes in colour[TODO ref Gabor filters: unsupervised texture segmentation using Gabor filters]. An example application of such methods is the detection of edges within an image.

- *Shape* - When colours and textures of an image have been analysed, the resulting features can be used to partition images into smaller homogeneous areas. The shape of these areas can next be represented by features that either describe the shape's region or contour. Data-driven methods are used for *weak segmentation*, where an image is deconstructed into shapes that share a visual property. [TODO ref R. C. Veltkamp and M. Hagedoorn, State-of-the-art in shape matching,] *Strong segmentation* on the other hand, uses knowledge about the shapes of objects to delineate contours of semantic concepts in the image.

- *Temporal* - Besides addressing aspects of single video frames, visual features can also capture how characteristics of frames develop over time. By following how sequential images change over time, patterns of motion can be tracked to describe camera motion[TODO ref Y. Tonomura, A. Akutsu, Y. Taniguchi, and G. Suzuki, Structured video computing, IEEE MultiMedia, vol. 1, pp. 3443, 1994.], motion of regions or points [TODO refJ. Sivic, F. Schaffalitzky, and A. Zisserman, Object level grouping for video shots, International Journal of Computer Vision, vol. 67, pp. 189210, 2006.] or even the movement of segmented objects[TODO ref H. T. Nguyen, M. Worring, and A. Dev, Detection of moving objects in video using a robust motion similarity measure].

Auditory Features

While video semantics might be most prominently expressed in the visual domain, auditory signals can also be used for segment-type classification and concept detection. In fact, auditory features may have the important advantage of being computationally cheaper relative to their visual counterparts. Considering this benefit, it can be strategic to start with an initial analysis of audio signals, and only proceed with more costly video analyses if further disambiguation is required. Different types of audio signals can be used for analyses

As is the case for visual features, video content is first segmented before audio features are extracted. Generally, auditory feature extraction is done on two levels: short-term frame level and long-term clip level [**?**]. Frames are usually very short sample intervals spanning 10 to 40 ms, for which auditory signals are assumed to be stationary. Clips have longer durations that span multiple frames and are used to capture the changes in frame features over time.

In their overview of the merit of audio and visual features in the characterisation of semantic content, Wang et al. categorise audio features into groups based on the type of information they are extracted from. They distinguish features based on *volume*, *zero crossing rate*, *pitch* and *spectral features*. Description of these features and their respective merit are left out of this thesis. Interested readers are referred to [**?**] for a detailed explanation. With frames analysed, features at clip level can be calculated to reflect the changes in frame level features over time. The different types of features on clip level can be characterised by the categories of the frame level features they are based on.

[TODO also see G. Lu, Indexing and retrieval of audio: A survey, Multimedia Tools and Applications, vol. 15, pp. 269290, 2001.]

Textual Features

While some works investigate the role of auditory[**?**] and textual features[23] in video analysis,

" text search against transcript narrative text provides almost all the retrieval capability, even with visually oriented generic topics." [Addressing the Challenge of Visual Information Access from Digital Image and Video Libraries][12]

Structural Features

[cite bordwell & Thompson: analysis of context dependency] [HyperCafe: Narrative and Aesthetic Properties of Hypervideo [32]]

2.3.2 Supervised Learning

Once features are extracted for a collection of videos, they can be studied to see how they relate to the concepts that are detected within the videos. The paradigm of supervised learning looks to find a relation between features describing data instances and classifications that can be attributed to them. The approach in supervised learning is to provide a large set of training examples along with the known classification or labelling of the content. Different methods can be used to determine a function that optimally describes how features are combined into a calculation of the probability that an instance should be labeled with a particular classification.

2.4 Outstanding Challenges

2.5 Discussion

2.5.1 H

uman interaction [Relevance feedback: A power tool for interactive content-based image retrieval][31] [collaborative filtering] `http://en.wikipedia.org/wiki/Collaborative_filtering`

Chapter 3

Human Computation towards Visual Meaning

3.1 Characterising Human Computation

In a recent survey paper, Quinn and Bederson present a taxonomy of Human computation systems[29]. They sketch out the trend of this new method for intelligent problem solving by the increase of academic papers featuring the term 'human computation' and its relative 'crowd-sourcing'. They summarise the myriad of definitions given in recent works by several different authors in two key points:

- "The problems fit the general paradigm of computation, and as such might someday be solvable by computers."
- "The human participation is directed by the computational system of process"

The first point introduces an interesting question whether storytelling is a computable process. In 1950, Alan Turing envisioned in his seminal paper 'Computing Machinery and Intelligence' that a computer program would be able to successfully play a game now known as the Turing Test. His work also mentions that

> "[t]he idea behind digital computers may be explained by saying that these machines are intended to carry out any operations which could be done by a human computer"[38]

The notion of 'human computer' benefits from some contextualisation, as in the last few decades we've become unaccustomed to the term. Human computers were not uncommon in the time of Turing and before that from the 18th century, when 'computer' was used to signify 'one who computes'[16]. People bearing the function title were involved in the execution of calculations produced by strictly following mathematical theories. The activity that these computers were involved in was a process of rote, not requiring any human creativity. While working on the design for the first ever mechanical computer, Charles Babbage called it "mental labour"[4, Ch. 20].

Quinn and Bederson further present a classification along six dimensions they see as the most salient distinguishing factors: [29]

3.2 Humans Computing Visual Meaning

3.2.1 Examples

ESP Game

Peek-a-Boom

reCaptcha

3.3 Computation in Interaction

Clicking from one video to the next (choosing from a set of related videos) these inter-video links could be seen as indicators for relatedness and relevance, much like google's page rank algorithm use links across webpages to establish a notion of the most significant site on a particular topic.

There is an important difference here though. Whereas the links used by Google's search algorithms are embedded in machine readable hyperlinks, the path of clicking on from one video to the next is a characteristic of a person's interaction.

3.3.1 The web as platform for creation

Many media scholars have written about the role of the web [refs New Media Reader]. Important trend of the web as platform of creation. In terms of video creation for example, the last few years have seen the development of online video editing tools and environments such as popcorn.js, WeVideo and Kaltura.

3.4 Deriving Meaning from Video Via Human Factors

[Personalized online document, image and video recommendation via commodity eye-tracking][40]

[VideoReach: an online video recommendation system][27]

3.5 Collaborative Filtering

The idea of using user's past interactions within a system hosting digital content for the filtering of items that might be of interest is not a new one and usually goes by the name of collaborative filtering. Collaborative filtering can generally take two forms: User-based, Item-based

Information filtering agents and collaborative filtering both attempt to alleviate information overload by identifying which items a user will find worthwhile.

3.6 A Characterisation of Human Computation Systems

3.6.1 Purpose

3.6.2 Motivation

3.6.3 Task

Chapter 4

Human Computed Stories in wePorter

[draft]

4.1 Introduction

This section describes the interaction design of the wePorter system, built to exemplify how the paradigm of human computation can be used in tasks like filtering of video segments and semi-automated video reconfiguration. The wePorter system runs an interactive webpage that functions as the main source for data acquisition, presentation of results and general proof of concept. In this chapter the system is analysed along the axes of *Purpose*, *Motivation* and *Task*, that were introduced in the analysis of Human Computational systems in chapter 3. Next to these guidelines for analysis, some remarks are made about the specifics in the functioning of the system. Lastly we discuss the implementation of the web application that is central in wePorter.

4.2 User Generated Video Content

> Since the dawn of YouTube, weve been sharing the hours of video you upload every minute. In 2007 we started at six hours, then in 2010 we were at 24 hours, then 35, then 48, and now...60 hours of video every minute, an increase of more than 25 percent in the last eight months. In other words, youre uploading one hour of video to YouTube every second. Tick, tock, tick, tock thats 4 hours right there!

These astonishing figures of the amount of video that is uploaded to YouTube are nothing short of mind blowing, but will most likely sound dated in a matter of years or even months. Looking at the increase of content uploaded to the video platform in past years, the growth does not seem likely to come to a halt soon [ref table]. All these videos are great for online video junkies, and are increasingly part of the online journalism landscape [30]. At the same time, all these videos being put online beg the question which ones of them to watch.

[graph of YouTube content uploads]

The increasing amounts of user-generated video content (UGVC) being put online, lead to an information overload and present both challenges and opportunities in search, retrieval[39] and recommendation[43] tasks [more refs]. There is an increased need for

ways of aggregation and filtering. Both of these tasks rely heavily on an at least a shallow understanding of what is presented in these media, which, as we've seen in chapter 2, is a hard problem to solve via current computational techniques. With so much content being uploaded, how can we find our way in the already enormous ocean of online videos?

4.3 The Purpose

With more than an hour of new content per second it is no wonder that YouTube has come to be viewed as the go-to for online video, much like "the digital video repository for the Internet" [1] that was envisioned by its founders in their first official blog post. An important activity on video platforms like YouTube is searching and much work has focussed on video retrieval[25, 36, 12, 17, 22, 18, 15, 37, 34]. We review the different aims in users' interaction in video retrieval tasks and point to an aspect of video retrieval that has not received much attention in recent research. This is the challenge of segmented video recommendation and we will explain why it forms a problem task to address within a Human Computation System like wePorter.

4.3.1 Different Aims

Annotations reflecting the content of a video can, along with other meta data of the video, be used for retrieval of videos in response to textual queries[ref]. The effectiveness of such a retrieval task varies depending on the information that is used in the search algorithm[refs] and the type of content that is searched for [18][more refs]. A third characteristic that determines the effectiveness of a video retrieval system is the goals that users have in their usage of the system [18]. User goals can vary widely from more to less specific[13]. We expand on this latter point, as it forms an important context for the wePorter system.

Direct Navigation

The most specific goal is exemplified by a user who is drawn to a video platform by a direct link from an external website. Links can either be in the form of actual hyperlinks or playable embedded videos that are followed through to the platform. Navigations via such links form a direct mapping between a user's intention to the desired piece of content. In this case, users have a very specific reason to come and watch. Their desire, at least of knowing the contents of the video, is satisfied after the viewing. YouTube's system engineers call this way of video viewing *direct navigation*[13].

Search and Goal-oriented browse

When users have not obtained a direct link to a potentially relevant piece of content, they might still have a specific goal in mind when visiting an online video platform. Reasons to visit might be the wish to see a particular music video or to find an instance of a series by a particular producer. This goal of discovering a rather specific video is referred to as *"search and goal-oriented browse"*[13]. Provided that the desired piece of content exists and the video platform has an appropriate search function in place, these 'narrow queries', will result is a result set of search results from which the user is likely to handpick the sought-after result fairly quickly. Here the user's desired result often lies within a single

[1] `http://YouTube-global.blogspot.co.uk/2005/07/greetings-everyone-thanks-for-visiting.html`

item of content. Perhaps a few misses are required, but after a couple of clicks the user hits the desired video.

Unarticulated Want

Yet a less specific goal is seen in users who come to a video platform "to just be entertained by content that they find interesting" [13]. These users mainly browse from one piece of content to the next, often aided by the platform's recommendations of related content. It has been found that YouTube's related video recommendation functionality, which recommends videos that are related to the video currently being watched, is one of the most important view sources of videos. In fact, traffic received from these recommendations is the main source of views for the majority of videos on YouTube [43]. Features derived from users' navigations such as 'click-through rate' have been used to improve content-based video recommendation [41].

Goal of a person's query in this kind of navigation is no longer defined in a single returnable item of content or even a containable set of items. Rather, the interactive pathway through the a set of interesting bits of content is what represents a user's aim. This broader, exploratory goal of finding different parts of interesting content has been termed 'unarticulated want'[13].

Encapsulated Wander

Considering the three categories of user motivation above, another, composite motive can be imagined. Users often start with a query for a particular topic, followed by a journey across many videos relating to their search term. Their navigation seems unarticulated but it is encapsulated by the topic of their query. Think of someone who wants to get an overview of a large music festival she recently attended. Big events where many people record videos, are often massively covered on UGVC platforms, resulting in an overload of visual information. Searching YouTube for this month's videos from the participatory festival Burning Man, two weeks after it ended, returns "About 7,660 results"[2]. A similar large set of topically related UGVC can be imagined at a website that asks participants to contribute their videos recorded at a recent event or centred around a particular topic.

This kind of 'broad queries' returns a result set of related content in which a user will probably consider many items as a successful retrieval. Furthermore, one could even say that the desired result of a user's query is spread across the multiple pieces of content. By traversing the space of different videos in the result set, users interactively construct the desired answers to their own queries. We call this motivation for discovery within a topically-related set of videos 'encapsulated wander'.

Interactivity is generally agreed to play an important role in the task of video retrieval, as is reflected by the separate category in the annual TRECVid challenge for interactive video retrieval[33]. Several works have indicated the importance of interactivity in the task of video retrieval to filter through a set of initially returned results [14, 11, 14, 15]. While most of these systems are aimed at retrieval of clearly specified queries, exemplified by the TRECvid retrieval task, the need for interactive exploration is even more apparent for the broader oriented goal of users engaged in 'encapsulated wander'.

4.3.2 Serving the Purpose of Encapsulated Wander

The answer to a user's query now lies as much in the journey through the content as in the returned content itself. By traversing from one piece of content to the next, users

construct a sequence of concatenated items. This self-constructed story is an important concept that wePorter capitalises on, as will soon become apparent.

The task at hand of recommending a larger group of interesting videos is radically different compared to the more narrow queries that could be answered by a small set of true positives in an information retrieval task. Besides the spread of the searched for result across different pieces of content, there is a second important difference that lies in the nature of the majority of UGVC.

User-contributed videos commonly consist of raw, unedited footage. In [30] Rosentiel and Mitchell report that within the collection they investigated only 39% of the news-related footage contributed by citizens was edited. It should be noted that this collection contained only the most popular videos per week and that a different distribution will be found in the complete set of news-related videos or all the videos hosted on YouTube.

Users with broad expectations will not only want to be presented with multiple relevant items from a complete repository, they are also looking for the most interesting parts within these relevant items. This issue is particular to time-based media, and especially relevant for video. Other temporal media, like audio in general and music in particular, have less of a need for segmentation because of their common usage in multimedia applications. People usually tend to listen to a song entirely and if they which to experience an album in part, constituent songs are already units on their own that can easily be reconfigured. Tag-a-tune is a game with a purpose used to acquire tags for clips of music. Although it could be employed for labelling of smaller audio sub-clips within songs, the games only aims at global labelling of a sounds[24].

Because of the unstructured nature of the majority of UGVC it is desirable to establish local recommendations that point to 'sub-clips' within a video that are of particular interest. Whereas digital music albums shared online consist of a collection of songs that can each easily be made to stand alone, video currently suffers from a less malleable identity online. Online videos are currently much like black boxes that can be played, paused, rated, commented on, tagged and shared only in its entirety. What if a piece of raw, unedited UGVC features something spectacular for ten seconds halfway along its timeline, but shows much of the same for the rest of the time? Answering this question will be the first part of the purpose of the wePorter system.

4.3.3 Storytelling as Structured Recommendation

[TODO; based on section in Meaning chapter. Indicate necessity for segmented recommendation + segmented annotation in general. Point to little attention to the retrieval of specific video segments [37][22].]

The ten significant seconds in a two-minute video become a needle in a haystack when an initial set of videos relating to your query includes tens to hundreds of possibly relevant videos with lengths between some tens of seconds and a couple of minutes. The aggregation and reconfiguration of several of these 'needles' into a meaningful new whole is another non-trivial task. We present wePorter as a test case for new methods that address both these issues of information overload in video libraries of UGVC. More precisely, wePorter's purpose is two-folded:

From a set of topically related unstructured user-generated videos:

1. Filter localised intervals of interest within each of the source videos
2. Reconfigure interesting video parts into a meaningful new story

4.4 The Motivation

How to get a group of unrelated people to contribute their efforts to solving the tasks set in our two-folded purpose? This section looks at the reasons people might have to contribute their computational powers to a system with a purpose like wePorter. Looking at the way people engage with online video content on platforms like YouTube, we identify patterns in their behaviour that can be matched to a task in a human computation system. This behaviour that is characterised by a more active role in multimedia consumption, can be seen as a larger trend in the development of new media. The end of this section indicates how the motivations of users of the wePorter system can link in with this larger trend.

Jain and Hampapur indicate entertainment, information or communication as purposes for the creation of consumer videos [19]. Although most UGVC is probably not as purposefully produced as the professional productions that Jain and Hampapur report on, these different purposes give an indication as to what user's motivations might be when interacting with a video system.

4.4.1 Information Provision through Online Video

Since the proliferation of mobile video recording devices, it has become common practice for large-scale (semi-)public events to be covered by UGVC that gets uploaded to the web. While some are critical[21] to the often heralded democratisation and empowerment of people by the new media production and distribution tools, it is clear that the UGVC at places like YouTube attracts a lot of traffic from people looking to be informed about recent events. After all, UGVC can have its advantages over traditional media when it comes to video news coverage, especially for unexpected events where traditional media do not have the immediacy of user-generated 'reports' recorded by coincidental passersby.

In a recent study as part of the the Pew Research Centers Project for Excellence in Journalism, the most popular video's from YouTube's 'News and Politics' were analysed for a period of 15 months[30]. The authors of the study exemplify the power UGVC can have in news provision by showcasing frequently viewed videos detailing scenes from the earthquake and subsequent tsunami that hit Japan in March 2011. The week following the disaster, the 20 most viewed news-related videos on YouTube all related to the catastrophic event and were together viewed more than 96 million times. Most of these videos were recorded by individuals who happened to be in the affected areas when the disaster struck, either uploaded by themselves, or by TV channels who appropriated the content. The study furthermore reports that in the studied period, the most searched term of the month on the YouTube platform as a whole was a news-related event 5 out of 15 months.

While the journalism study above focusses on videos with the 'News & Politics' label, information provision about current events might span a larger set of categories. Someone looking for footage in order to get a sense of the atmosphere at a recent music festival or public demonstration, might very well find relevant videos in categories like 'Entertainment', 'Travel & Events' or 'Nonprofits & Activism'. Across all of these categories, we are able to find examples of vast collections of UGVC, uploaded in the period following up newsworthy events.

4.4.2 Informative Entertainment

The wePorter system focusses on these kinds of topically related sets that people are currently exploring interactively by browsing from one video to the next. This way of navigation is an intermediary between the goals of *goal-oriented browse* and *unarticulated*

want. . The apparently aimless browsing is now encapsulated by the event but users still roam freely within this topicalised set of content. By navigating from video to video, watching some and skipping others, users leave attentional traces that give valuable insight into a user's intentional standpoint.

It is this kind of interactions that are already taking place at a large scale that we like to make use of in the wePorter system. Motivated by the wish to explore informative content, users will instinctively and implicitly contribute their human knowledge to a system that is set up appropriately. This kind of motivation fits the category of 'implicit work' as it involves activities that people already engage in for their own reasons [29]. Considering users' wish to be informed and the interactive way in which they navigate, there is most likely also a factor of entertainment involved though. We expect though that the more specific motivation of information provision might show to become a valid categorisation for the motivation of people in a HCS as it is a common activity on the web and inherently linked with the hard problem of meaningful interpretation of content.

4.5 The Task

In this section we take a look at how the larger goal of finding intervals of regional interest across time within a single video can be branched out into bite-sized tasks executable by a person in a single interaction. We begin by introducing some conceptual considerations that influenced the interface design. Then a detailed overview of the wePorter web interface is presented. We end with a section focussing on the implementation of the system.

4.5.1 Design Considerations

Below are included several points that have been instructive in the development of the interactive task central to the wePorter system. Some of these point are system requirements, others are more guiding design principles or thoughts that have been inspiring and formative in the development.

wePorter is a web interface

The power of a HCS that relies on data from many interactions is truly unleashed in an online setting, where many people can easily participate and interact. For this goal alone already, wePorter must be a web-based system. Besides the obvious choice of staying in the realm of the online video content, it makes sense to embed the theoretical explorations of this research in the practicalities of current web technologies. With the ongoing development of technology like HTML5, many new possibilities for a user's web browser are unleashed. The implementation of a research tool concerning online video is a good opportunity for the exploration of the technological possibilities of present day web technologies.

Hypervideo

The power of digital content on the internet lies for a great part in its capacity to be hyperlinked. Linking to externally hosted content alleviates the burden of having to host or recompile pieces of media. Instead, files can be played and remixed by reference, leaving their respective sources intact and where they are. In the presentation of their digital video repurposing system 'Diver', Pea et al. indicate the advantages of using a virtual camera

controlled by XML-based files that reference parts of source video instead of rendering new video clips[28]:

- "Virtual video clips eliminate the generation of redundant video files, greatly reducing disk storage requirements."
- "No rendering time means vastly improved performance. Users can instantly create and play back dynamic path videos without long video-rendering delays."

An implementation of a system where users interact with content that is dynamically reconfigured in real-time will benefit considerably from a hyperlinked functionality, especially when this takes place in an online setting where bandwidth will be limited.

The idea of hypervideo in the context of interactive narratives has been proposed by Sawhney et al., where users were invited to navigate a virtual cafe by means of *temporal* and *spatio-temporal* and *textual* links present in the video interface. A temporal link is "[a] time-based reference between different video scenes, where a specific time in the source video triggers the playback of the destination video scene"[32]. wePorter utilises temporal links to link sequences of video scenes together.

Localised Interest

To answer to the first purpose of wePorter, we wish to distinguish parts of videos based on their level of interest. In order to elicit users' preference for particular parts within a video, we divide each *'source video'* in our initial set of topic-related content into smaller *'video parts'* and present a selection of these in a user interaction. Slicing up source videos virtually and playing their parts by reference is made possible by a hyperlinked implementation of the video player.

Forced Feedback

The user interaction design should enable means to learn about a user's interest in a video at a particular moment in time. This to the purpose of discovering localised regions of interest within separate videos. To make a user's interaction as enjoyable as possible, implicit data acquisition should be preferred over explicit questions. In other words, user will be more likely to repeat a task that implicitly logs their behaviour during interaction, than one where they are presented with a questionnaire after every click. By making the acquisition of user feedback an integral part of an interaction, users are directed into contributing their computational power without even being aware of it. In wePorter, this enables *'curation through interaction'* by using the *'forced feedback'* to maintain and improve a dynamic story space.

Preference Elicitation by Parallel Play

Considering measures that could indicate how people's interest varies across different videos, an idea that quickly surfaced is that interest is closely linked to attention. When a piece of content contains something that is interesting to many people, this will most likely result in an increase in views, provided the content is accessible to a variety of people. This simple notion is the idea behind global recommendations that show most popular or 'trending' content. Whether the trending item is a video on a sharing platform or a phrase on a microblogging service, when there is a large number of people attending to it, this is a reason to suspect the item to be of interest for people who haven't engaged with it yet.

An obvious limitation of these global recommendations is the lack of personalisation. Personalised recommendations are offered because the content that is globally popular may not be related to the topics of my interest. For the purpose wePorter is serving however, focus around a particular topic is already in place and we are in the first instance mostly interested in picking out the parts that share a high level of interest globally.

The wePorter system uses a new method of user preference elicitation, that makes explicit choice of attention an integral part of the user's task. A user is concurrently presented with two videos for which we would like to elicit preference, and is forced to make an explicit choice of attending to one or the other. During this 'parallel play' of two video parts we capture the amount of time attended to each of the parts and store this for later analysis. Parallel play is useful for eliciting preference for time-based media like audio and video as it lets users express their preference in the time they attend to an item.

Recurrent Interaction

Because of the reliance on data, user's should be able (and encouraged) to engage in the interaction more than once.

Users Between Consumers and Producers

Studies reflecting on new media technology and its incorporation in our everyday life are in recent years often speaking of a media convergence, where multimedia content flows dynamically across multiple media platforms and media audiences take an active, participatory role in their search for entertainment experiences. In his book 'Convergence Culture', Jenkins writes:

> "This circulation of media content - across different media systems, competing media economies, and national borders - depends heavily on consumers' active participation. I will argue here against the idea that convergence should be understood primarily as a technological process bringing together multiple media function within the same devices. Instead, convergence represents a cultural shit as consumers are encouraged to seek out new information and make connections among dispersed media content. [...] The term *participatory culture* contrasts with older notions of passive media spectatorship. Rather than talking about media producers and consumers as occupying separate roles, we might now see them as participants who interact with each other according to a new set of rules that none of us fully understands."[20]

Surveying the diverse body of research into interactive TV, Cesar and Chorianopoulos propose a new way of looking the life cycle of digital content that considers content editing, content sharing and content control as an alternative to the more hierarchical 'produce-deliver-consume' paradigm associated with traditional media[8]. The movement from passive consumers to (inter)active contributors indicates new expectations by users of new media applications. The trend of users' more active engagement in new media technology fits well with the approach of interest defined by users' interaction and our proposal of storytelling as structured recommendation.

The Death of the Author, the Birth of Collective Creation

The first part of the title above was originally voiced by Roland Barthes in a 1968 essay bearing the phrase as its title, in which he contemplates the role of the author in literary

writing[5]. His title 'the death of the author' points to his criticism to the importance that typically gets associated to the author in the analysis of literary work. To Barthes, a text is not "the message of the Author-God", but rather "a multi-dimensional space in which a variety of writings, none of them original, blend and clash. The text is a tissue of quotations drawn from the innumerable centers of culture". Instead of the author, for Barthes, it is the reader where the diverse backgrounds of text become unified. Barthes' criticism is resolute:

> "the birth of the reader must be at the cost of the death of the Author"

In our aim of reconfiguring interesting sub-clips into a new arrangement, the changing role of the author surfaces in a new context.

Less restrictive forms of digital content licensing, like Creative Commons (CC), mean that it is now possible for content uploaded by its original creator, to be used under specified conditions in a new piece of work by someone else. This kind of licences has been noted to be an important facilitator of research into Human Computation[24]. They make it possible for works not only to be used and remixed by other individuals, but also to be incorporated in algorithmically constructed reconfigurations of user generated content.

Different video platforms are currently offering less-restrictive CC licensing as an integrated part of their services. YouTube currently offers the option of choosing a most basic attribution licence and reports 4 million videos licensed this way [7]. The video platform Vimeo focusses on letting video and animation producers share and showcase their original work. The platform has internalized the use of CC from 2010[3] and many of their users licence their videos such that they can be remixed by others. Figure 4.1 shows that a large part of the licences on the Vimeo platform allow derivatives to be made[1].

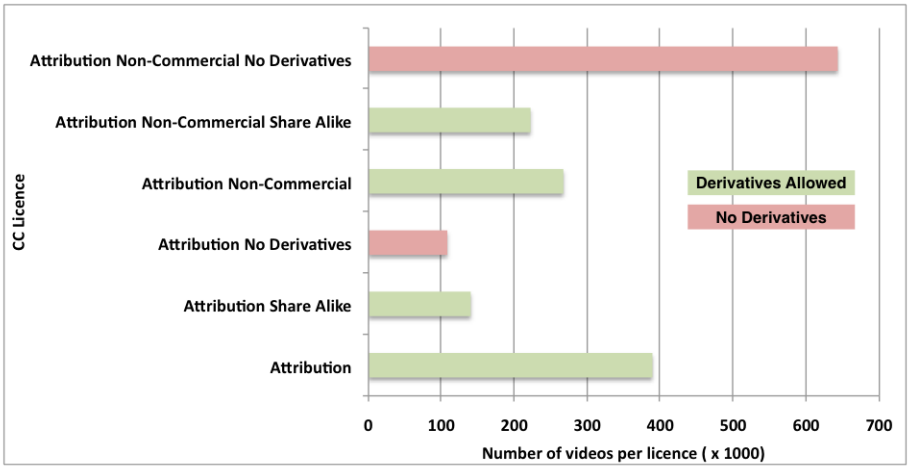


Figure 4.1: Number of videos for each of the Creative Commons licences on Vimeo

Besides the collective actions of the multiple users that help shape the creation of new configurations of content, there is a further level of collaboration between the users and the tools they interact with. Manovich even extends this relation to the tools' designers in his view on collaborative new media authorship:

> "Authoring using [Artificial Life] or [Artificial Intelligence] is the most obvious case of human-software collaboration. The author sets up some general rules but s/he has no control over the concrete details of the work these emerge as

a result of the interactions of the rules. More generally, we can say that all authorship that uses electronic and computer tools is a collaboration between the author and these tools that make possible certain creative operations and certain ways of thinking while discouraging others. Of course humans have designed these tools, so it would be more precise to say that the author who uses electronic/ software tools engages in a dialog with the software designers [...]." [26]

4.5.2 The Interface

This sections describes the user interface that directs participation of wePorter users towards solving the purpose of distinguishing local intervals of interest within videos. After a conceptual overview of the functioning we include a walkthrough to explain precisely how the interaction takes place.

We hypothesise that interesting parts of content will attract a relatively large amount of attention compared to less interesting parts.

To force a users to make an explicit choice between parts of content for which we would like to elicit their preference, we present two pieces of video playing concurrently and force users to attend to one or the other. In order to get an idea of the variation of interest across a video, *'source videos'* are divided into smaller *'video parts'*, each of which is presented separately in interactions over time. During this 'parallel play' of two video parts we capture the amount of time attended to each of the parts and store this for later analysis.

A Walkthrough

When a user opens the wePorter web interface he is welcomed by a short introduction to the project and successively guided to further instructions explaining the experiment. The instructions as they are presented to the user are shown in figure 4.2.

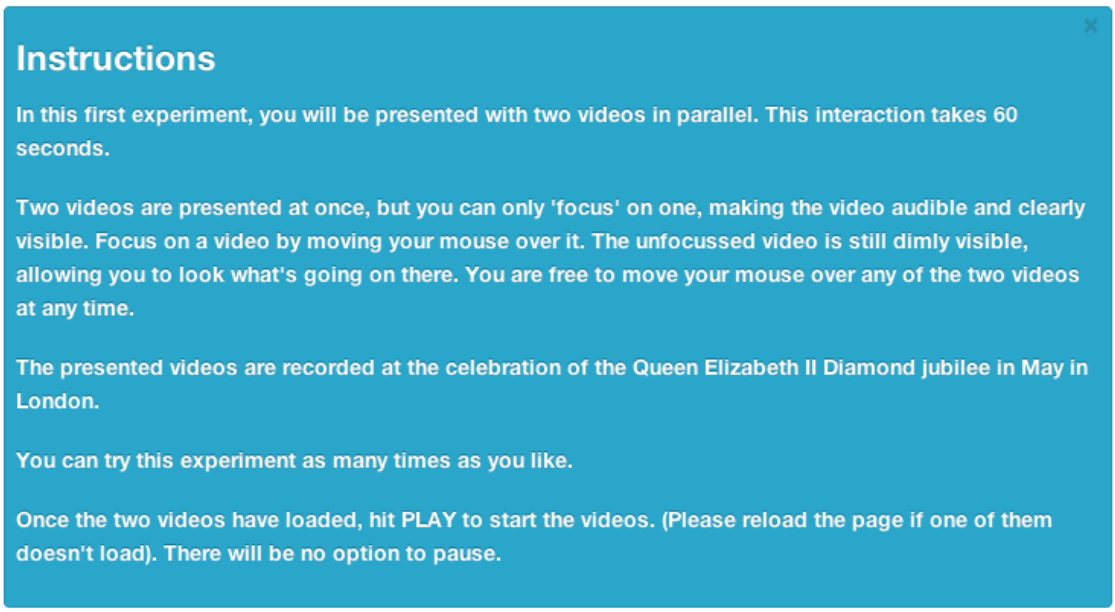
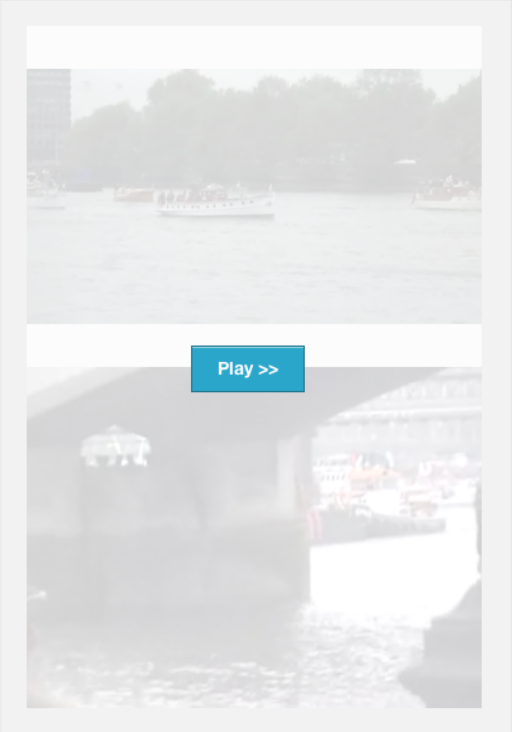
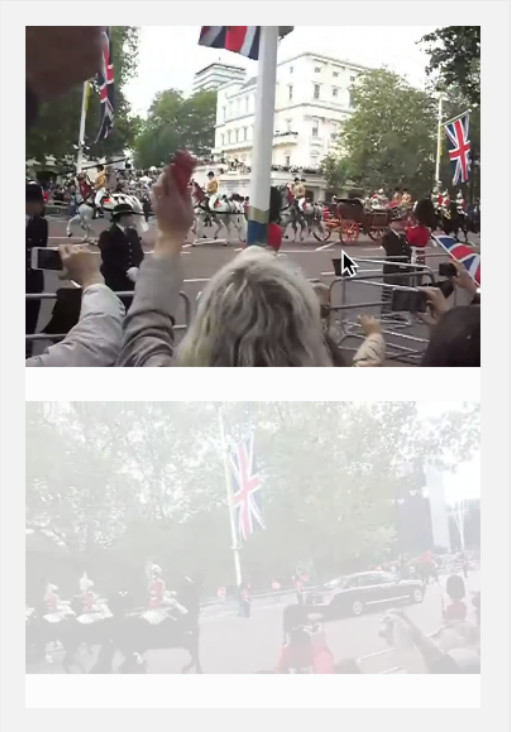


Figure 4.2: wePorter Instructions



(a) Upon load



(b) While playing, focus on top video

Figure 4.3: The wePorter parallel play interface

Upon loading the webpage, a database is queried for a pair of sequences made up from different video parts for which the system would like to elicit a user's preference. The two sequences both consist of six video parts that each have a duration of 10 seconds. The interaction of the two sequences playing in parallel thus has a total duration of 60 seconds, reflecting the short time span common across UGVC at YouTube[9, 10]. A detailed description of the algorithm used for the construction of these sequences is given in section 4.6.

After reading the instructions, the user scrolls down to the interactive parallel video player that displays two videos on top of each other. By clicking the 'Play' button, the user starts the interaction and sets in motion the consecutive playback of both sequences in parallel.

During the parallel play of the two sequences of video parts, the user triggers which of the two videos is in 'focus' by placing the mouse cursor over it. When focus is placed on a video, this makes it audible and clearly visible. The unfocussed video is silent and still dimly visible. This partial visibility allows the user to discern to a limited extend what is displayed in the unfocussed video. Seeing something that attracts interest can lead the user to change focus from one video to the other. The limitation of only one of the two videos being in focus at once, gives users incentive to explore the narrative space of the parallel sequences. The aspect of focus lets users spread their attention between concurrent parts by:

- making a choice to attend to a video part they find most interesting.
- changing from time to time to check what is being played in the unfocussed video.

We record which video a user is attending to by keeping a count for each of the two

video parts playing concurrently and increasing the count for the focussed video every 100 milliseconds. When a video part ends, the count is logged internally on the user's browser side before the next video part is started with its own count. When the parallel sequences are played back completely, the end of the interaction is reached and the counts for each of the 12 video parts are stored in a server-side database. Each pair of counts for two concurrently presented parts, represents a distribution of the user's attention over those parts.

Note that we never explicitly ask anyone to point at the video that is most interesting. Users are simply instructed as to how the interface works and then left to explore the videos as they like. By recording users' behaviour this way, we achieve a detailed insight into which of a pair of videos a user has attended to at what time. In section 5 we report on how these measure can be telling in the process of filtering and reconfiguration.

4.6 Implementation

This section describes in detail the technologies used in the wePorter system and how components relate to each other. We begin by illustrating how video content is prepared for presentation in the wePorter web interface, and next describe the system framework.

4.6.1 Preparation of Content

In order to get localised feedback on distinct temporal intervals within videos, we present users with a sequence of 'video parts' of equal duration, each originating from their own respective 'source video'. A initial step is thus to prepare video parts so they can be presented in a user interaction. What is played back to a user is a part of the source video referenced by hyperlinks to start and end points. This hyperlinked implementation means slicing up source videos does not involve cutting up video content or recompilation of any sort.

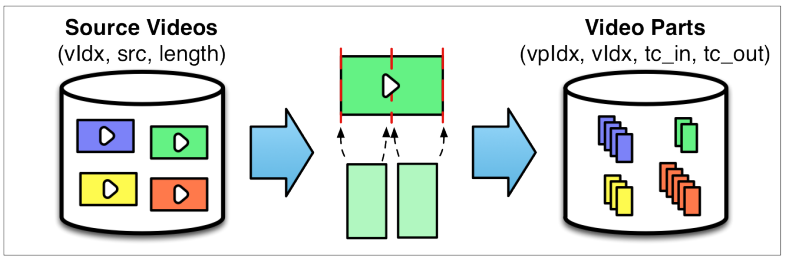


Figure 4.4: Slicing Source Videos into Video Parts by Reference

The wePorter system takes as a starting point a set of topically related source videos, representing a result set that could be acquired by querying a large UGCV platform for video from a large scale public event. We keep a database of source videos, storing their source path and length:

$$video = (vIdx, srcPath, length) \tag{4.1}$$

where $vIdx$ is the video's index. Next, we define video parts as tuples of source video and two time codes referencing start and end:

$$videoPart = (vpIdx, vIdx, tc_{in}, tc_{out}) \tag{4.2}$$

where tc_{in} references the time code within the video indexed by $vIdx$ that is the start of $videoPart$ and tc_{out} references the time code within the video indexed by $vIdx$ that is the end of $videoPart$.

Algorithm 1 shows the procedure to generate video parts from source videos. The algorithm starts at the beginning of a video and extracts a video part for every consecutive window of duration d.

There might be an interval with a duration less than d at the end of a source video that is not included in the resulting set of video parts. Because the parallel sequence player expects video parts of equal length, these end bits are discarded and will not be presented during user interaction. Our assumption is that because of the raw, unedited nature of the videos used in wePorter, disregarding the final few seconds of videos will be tolerable. People recording video in a point and shoot fashion usually stop recording when a phenomenon that prompted them to start filming has ended and so the final bit of their videos does not commonly contain the most important content.

Algorithm 1 Generate Video Parts

```
1:  procedure PARTITION(sourceVideos, d) ▷ partition videos into parts with duration d
2:      i ← 1
3:      for all video ∈ sourceVideos do
4:          tc_in ← 0
5:          while tc_in ≤ video.length − d do
6:              tc_out ← tc_in + d
7:              videoPart_i ← (video.src, tc_in, tc_out)
8:              tc_in ← tc_in + d
9:              i ← i + 1
10:         end while
11:     end for
12: end procedure
```

4.6.2 The live system

Figure 4.5 shows the data framework of the wePorter system. The system runs as a web interface and is accessible online for multiple users at the same time. The server-side functionality is implemented in PHP making use of connections to a mySQL database. On the client-side, Javascript deals with the playback of video sequences as well as keeping track of all interaction data. Upon a user's navigation to the wePorter web page, two sequences are loaded in the parallel video player. A user triggers the interaction by clicking a play button. During the interaction, Javascript running in the user's browser keeps track of the accumulating attentional ratings per video part. Once an interaction has finished, all interaction data is added to the database on the server. The updated interaction data and counts of video parts are subsequently used in the generation of sequences for new interactions either by the same user or a new visitor.

For experimentation purposes, the current implementation maintains a single set of topically related source videos representing the context of 'encapsulated wander' for a single query. An extension of the system would be to let users query a live database like YouTube for content of their interest and thus define a dynamic collection of content to explore. Another option would be to give users a choice of which set of videos they would like to interact with. These methods could prove to be useful as they let users interact with content they have chosen themselves, which might make their interaction

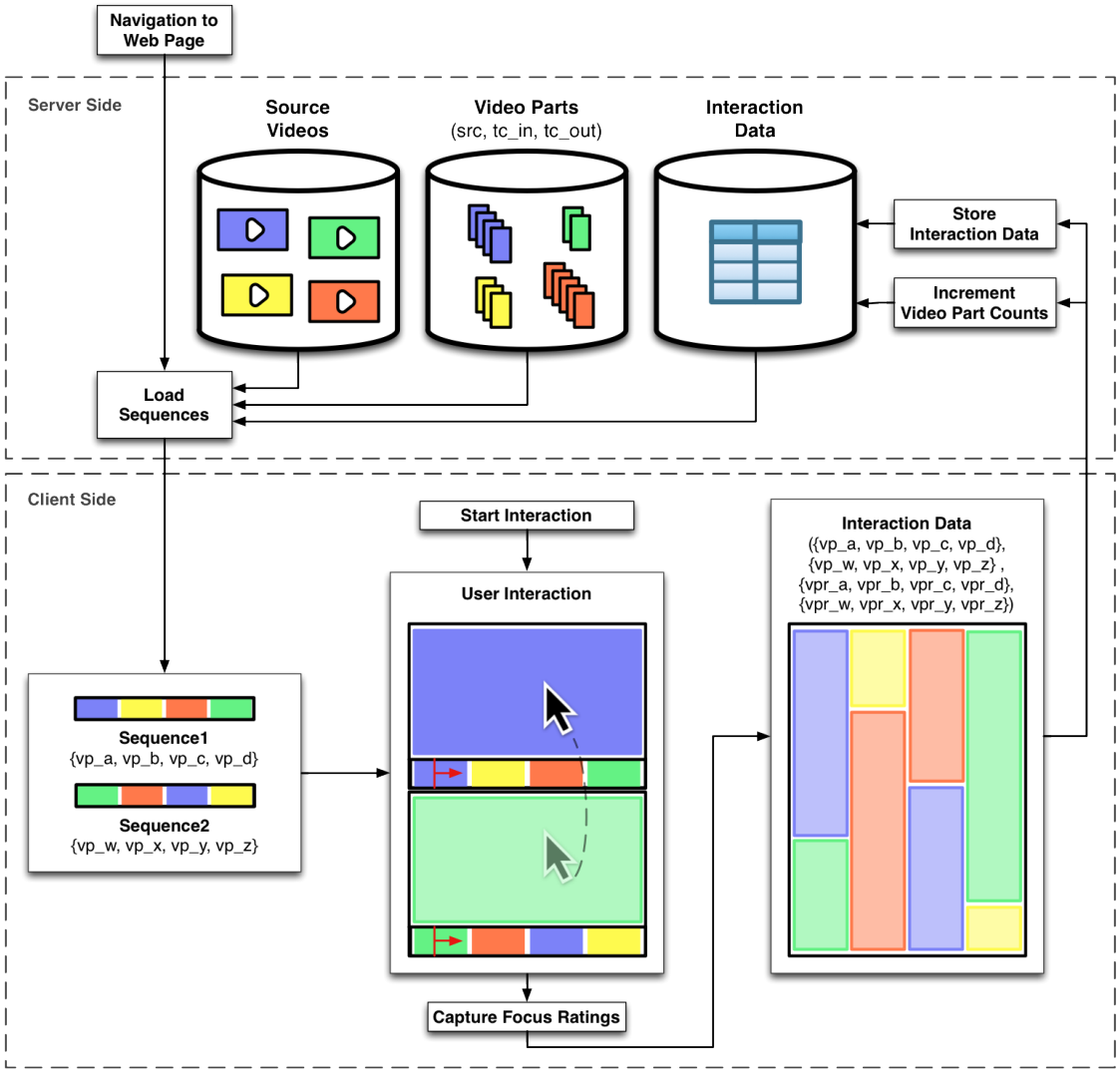


Figure 4.5: System Framework of wePorter

more interest-driven. In our current implementation, we have chosen to use a fixed set of video content for more controllable experimentation, and left these extensions for future work.

Loading Sequences

The procedure of loading two sequences for a user interactions is detailed in algorithm 2. Given the set of video parts, we iteratively construct two sequences of *n_parts* video parts to be played in parallel. The sequences have equal length and satisfy the following constraints:

1. **Horizontal source constraint:** Video parts within a sequence all originate from different source videos.

2. **Vertical source constraint:** Two video parts that are played concurrently one above the other, originate from different source videos.

These constraints guarantee variety in the interaction, both within a single sequence and across sequences for concurrent parts. On one hand it enables a more varied editing of the video story, which is desired for the user experience. On the other hand it makes sure that each interaction elicits user preference for a variety of sources, which leads to diversified data acquisition.

Amongst the interaction data that is stored in the database, we keep a count for every video part of how many times it has been presented. The counts are used to select from the set of video parts that satisfy the constraints, the ones that are least presented so far. This ensures that all video parts will be presented roughly an equal number of times.

After the two sequences have reached the desired size of *n_parts*, they are randomly shuffled to make sure video parts are presented at different positions in sequence roughly equal amounts of time. Shuffling happens in unison which means correspondence is kept across both sequences so that the constraints still hold. The process of shuffling in unison is described in pseudo code in algorithm 3.

[TODO comment on combinatorics and complexity. all video parts are presented more or less equal, except long ones]

4.6.3 Hypervideo in a Web Browser

Playing back parts of different online videos in a single video experience is a common feature of any video editing system, but this functionality has only recently become available to code that runs in a web browser. Today most videos that live on the web are much like black boxes and this is not just because computers are having a hard time understanding the visuals. When we interact with video online it is almost always on the high level of the entire video. Whether it's playing, sharing, commenting on or linking to video, we lack the functionality of referring to parts that lie within or interact with components like (sub)titles, images or audio as separate entities. [... ref video vortex] calls this type of video "hard" and "flat" [check quotes].

This is starting to change. An important player in the movement of treating video like the web by hyperlinking, cross referencing and remixing it in code, is Mozilla, whose foundation runs 'Popcorn' [2], a project that makes video work much like the web. Part of the Popcorn project is 'Popcorn.js' [3], a Javascript library that intends to open up videos

[2] `http://mozillapopcorn.org/`
[3] `http://popcornjs.org/`

Algorithm 2 Load Sequences Random Shuffled

```
1: procedure LOAD SEQUENCES(n_parts, videoParts, vpCounts)
2:     seq_1 ← []
3:     seq_2 ← []
4:     for i ← 0, n_parts do
5:         selectionH_1 ← []          ▷ keep selections of parts that satisfy constraints
6:         selectionH_2 ← []
7:         for all vp ∈ videoParts do                          ▷ Horizontal constraint
8:             if vp.vIdx ≠ part.vIdx for all part ∈ seq_1 then
9:                 add vp to selectionH_1
10:            end if
11:            if vp.vIdx ≠ part.vIdx for all part ∈ seq_2 then
12:                add vp to selectionH_2
13:            end if
14:        end for
                                                                ▷ For seq_1:
15:        minSelection_1 ← []  ▷ select from selectionH_1 parts that have minimal count
16:        minCount_1 ← min(count) from selectionH_1     ▷ look up counts in vpCounts
17:        for all vp ∈ selection_1 do
18:            if vp.count = minCount_1 then
19:                add vp to minSelection_1
20:            end if
21:        end for
22:        selected_1 ← random from minSelection_1
23:        append selected_1 to seq1                          ▷ add video part to seq_1
                                                                ▷ For seq_2:
24:        selectionV = []
25:        for all vp ∈ selection_2 do                         ▷ Vertical constraint
26:            if vp.src ≠ selected_1.src then
27:                add vp to selectionV
28:            end if
29:        end for
30:        minSelection_2 ← []  ▷ select from selectionV parts that have minimal count
31:        minCount_2 ← min(count) from selectionV      ▷ look up counts in vpCounts
32:        for all vp ∈ selection_2 do
33:            if vp.count = minCount_2 then
34:                add vp to minSelection_2
35:            end if
36:        end for
37:        selected_2 ← random from minSelection_2
38:        append selected_2 to seq2                          ▷ add video part to seq_2
39:    end for
40:    shuffle_in_unison(seq_1, seq_1)
41:    return (seq_1, seq_2)
42: end procedure
```

Algorithm 3 Shuffle Sequences in Unison

1: **procedure** SHUFFLE IN UNISON(seq_1, seq_2)
2: $seed \leftarrow make\_seed()$ ▷ to seed $random\_generator$ twice identically
3: $random\_generator.set\_seed(seed)$ ▷ set seed
4: $random\_generator.shuffle(seq_1)$ ▷ shuffle
5: $random\_generator.set\_seed(seed)$ ▷ set seed
6: $random\_generator.shuffle(seq_2)$ ▷ shuffle
7: return (seq_1, seq_2)
8: **end procedure**

to the web, give it back it's depth and make it more soft. The wePorter system relies for a big part on the 'Popcorn.js' library for the playback of video parts in sequence.

Chapter 5

Evaluation

This chapter presents results of number of experiments we have run. This includes We report on the feedback that is received from users' interaction through the use of the wePorter web interface. We show how it can be used for the purposes of filtering segments of interest within a video and the configuration of a new video story from the initially unstructured content. We present results from a user evaluation study on the merit of filtered and reconfigured content. We end with a reflection on our results in the discussion section.

5.1 Preliminary Experiments

5.1.1 Positional Bias

Positioning two video's one on top of the other, might inflict a bias for users in their attentional behaviour. It might be the case that videos on the top are systematically more attended to than videos displayed below. We've experimented to see whether such positioning bias effects occur.

To see whether the positioning of a video has effect on users' attentional behaviour, we've presented two groups of users the same two videos playing in parallel and varied their relative positioning. In two trials, participants were divided by random into control group and test group. The first trial was conducted with 37 participants (24 control, 13 test), the second with 32 (18 control, 14 test). In each trial, test group participants were show the same two videos as the control group, but their positioning was flipped. Figure 5.1 shows a visual explanation of the experimental setup.

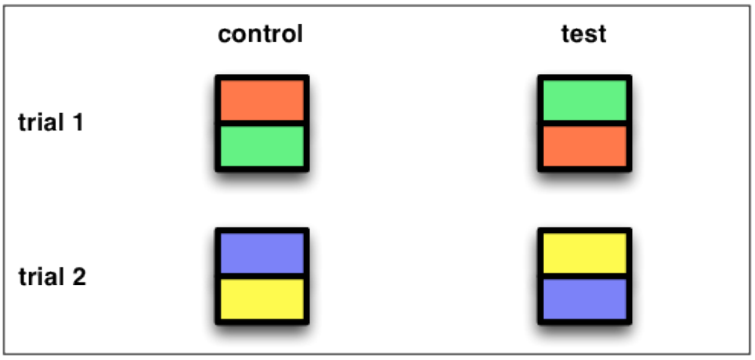


Figure 5.1: Setup for Positional Bias Experiment

5.1.2 Context Dependency

We propose a statistical analysis of interaction data to inform the reconfiguration of initially unstructured video parts. We hypothesise that:

1. Users' attentional behaviour depends on the sequential ordering of video parts
2. Data about users' attentional behaviour can indicate what are preferred orderings.

Before we investigate the second hypothesis in section 5.2, we must scrutinise the first. In order to test whether users' attentional behaviour is dependent on the sequential ordering of video parts we have run a experiment in two trials. The first of these had 35 participants, the second 28. In each trial participants were presented with with a two sequences of three video parts playing back continuously using the parallel video player described in section 4.5.2. Based on random picks, roughly half of the participants were labelled as control group the others as test group. Participants in this group where shown two sequences where all video parts originated from different sources. This was changed for the second group, where participant were shown parts across the sequences that clearly belong to the same source video.

The first trial presented test group users with a first part in the bottom sequence that was followed by a part from the same source video as the third part in the top video. The second trial presented test group users with pieces from the same source video in the first and third part of the top sequence and second part of the bottom sequence. The baseline sequences shown to control group participants had video parts of different sources on all positions and had final parts of both sequences identical to test group users. A visual description of these conditions is shown in figure 5.2. Where we use the same conventions to display sequences, video parts and colouring to indicate source videos as in figure 4.5.

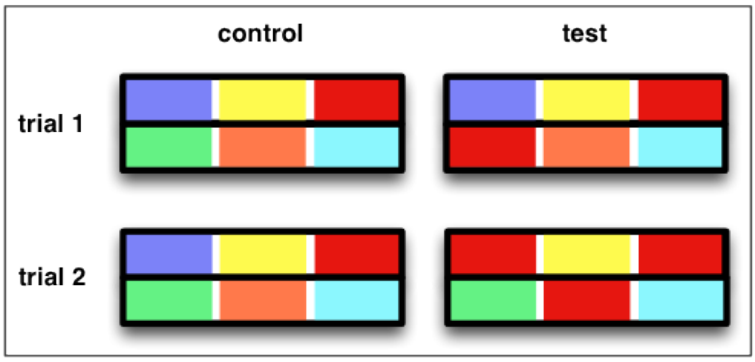


Figure 5.2: Setup for Context Dependency Experiment

5.2 Main Experiments

Our main experiments use the system framework described in section 4.6 for the preparation and presentation of video content, data capture from user interaction and storage of interaction data. The experiments we report on here are based on the interactions from a total of 68 persons over a period of a week.

Id	File Name	Title	Length (sec)	Views
1	"jubilee_01.webm"	"Diamond Jubilee London 5th June 2012 The Mall Video 3"	186	45
2	"jubilee_02.webm"	"Diamond jubilee London 5th June 2012 The Mall"	47	108
3	"jubilee_03.webm"	"London Thames - Queens Diamond Jubilee Pageant - Dunkirk Little Ships, June 2012"	234	2304
4	"jubilee_04.webm"	"My Diamond Jubilee video!"	54	28
5	"jubilee_05.webm"	"Queens Barge. Diamond Jubilee London 2012 VIDEO Ursula Maxwell-Lewis 0053"	163	88
6	"jubilee_06.webm"	"Queens diamond jubilee London 5th June 2012"	30	43
7	"jubilee_07.webm"	"Queens Diamond Jubilee Procession, 5th June 2012, London, UK"	133	58
8	"jubilee_08.webm"	"Queens Elizabeth 60th Diamond Jubilee London 2012. 1st"	73	203
9	"jubilee_09.webm"	"Queens Elizabeth 60th Diamond Jubilee London 2012. 2nd"	165	246
10	"jubilee_11.webm"	"Queens Elizabeth 60th Diamond Jubilee London 2012. 14th"	390	88

Table 5.1: Source Videos used for the main experiments (Views count accessed on 23-09-2012)

5.3 Setup

For our main experiments we use a set of 10 unedited, user-generated videos that were returned in response to a query for "Diamond Jubilee London" on YouTube. The videos along with their title, length and current view count is shown in table 5.1. As mentioned in section 4.6, we divide source videos into video parts of equal length and generate two sequences consisting of an equal number of video parts to present in parallel. In our main experiments

The videos used for the main experiments have a total length of 25 minutes and their individual view counts range from 28 to 2304.

5.3.1 Clean Data

Data returned from the capturing of attentional behaviour in the parallel player interface might have several deficiencies that we account for by preprocessing the data before we start analysis. These range from inaccurate artifacts induced by the interface to users' behaviour patterns that make their data less revealing.

5.3.2 Landscapes of Attention

The distinction to be made, between interesting intervals on one hand and less striking parts of a video on the other is not likely to be a very strict one. After all, an unedited video captures a single stretch of space and time, so any event that is of particular interest will unlikely have hard cut-off points in time. Rather, if we see interest as a function

of time in a particular video, we would expect a somewhat continuous flowing line with spikes every now and then when an interesting event occurs.

Looking at interest at a more global level, aggregating over a large group of users, would result in an even a more smooth landscape of interest. This kind of data reveals mountains and valleys that can be used for attention-based segmentation. From the thus segmented parts, the ones with high attentional scores can be returned as candidates for interesting parts within a video.

5.3.3 Evaluating Focus

By capturing the amount of focus time, we receive measures for user attention. We hope these measures for attention help point to patterns in user preference. If the two dimensions turn out to be linked, it makes sense to base segmentation of potentially interesting parts of video on the focus data captured in our parallel play interface.

To see how focus measures relate to user preference, we compare video parts which have received low focus rates

5.4 Questionnaire

5.5 Discussion

This section presents a discussion of the results presented in this chapter as well as considerations about wePorter at system level. We also relate the findings from our experimentation with the wePorter system back to the broader ideas concerned Human Computation towards meaningful video analysis.

5.5.1 Potential Extensions

Using the length of individual videos and the number of times they have been viewed, we have calculated that the total time spent on watching the 10 videos above, amounts to over 185 hours. That is a lot of user interaction that could help in the computation of the most interesting video segments. If directed through our one minute parallel play interface, all this user interaction would result in more than 11000 complete interactions. With 12 video parts presented in each interaction, it would amount to more than 130000 attentional ratings for video parts or over 900 ratings per part.

This is not yet regarding the system's filtering functionality. Once reliable estimates of user interest have been established from the collection of captured attentional ratings, we can filter for the parts that seem most interesting, that way enabling a convergence towards the most interesting video segments. A simple way of doing this would be to rank all video parts according to their average attentional rating and discard the parts that are systematically less attended to, on the assumption that they are considered less interesting. More complex procedures are of course possible, for example taking into account the number of ratings a part has received, the positions in sequence at which parts have been presented or the kind of users that have submitted the interactions (e.g. first time users versus experienced users). Once the initial set of source videos has been filtered, more data can be acquired for the remaining set, after which filtering can be applied again. This iterative process of filtering can continue until the process yields a small subset of video part that have acquired most attention in aggregate.

Filtering video parts to converge to a subset of the source videos that is iteratively narrowed down, means that most interactive computation will be focussed on the video parts

that receive most attention. There is an obvious issue of exploration versus exploitation here. When is filtering applied and what part of the current set of content is discarded in an iteration.

5.5.2 Feedback from Comments

> "[...]the technology may though have many uses like the crowd sourcing of video editing or the training of AI to mimic human audio visual focus and attention" - Mia

As part of the user feedback form that was presented during the experiments, participants were given the option to leave their comments. Thinking of the potential of the interface, some comments, like the one above, hit the nail right on the head. Others touched upon different aspects of the experiment, the interface and the content of the videos. Overall a number of salient points emerged from the comments:

- **"I would like to see news this way"** - People were positive about the way video was presented in parallel and were enthusiastic about the possibility to interact.
- **"I believe there was a bug"** - A number of people reported difficulties in the playback of videos. Most issues concerned one or more video parts not immediately playing after the preceding one had finished. Besides causing data to be less clean, this caused some users to be confused.
- **"The subject and content of the videos was uninteresting"** - Many people said they were not particularly interested in the topic of the presented videos. This meant that often they weren't drawn strongly to a particular video and did not feel a strong reason to shift focus from one or another.

Some participants also offered ideas as to how they saw the project could be extended:

> "Idea is great and project full of potential, in particular for big events which are well covered and allow multi-angle views of an action. It would be interesting to have information about the content producer displayed discretely on the player. That way the audience could vote on the quality of a source, and in time reward the owner for it's content, encouraging him to submit more videos in this system." - Marc

Chapter 6

Future Directions

This is the Future Directions chapter. We discuss possible extensions in terms of both system capabilities and ideas for future research.

Creations thus constructed will embody an interesting aspect of collective creation through the merging of content creators, people contributing their human computational capacity and algorithmic aggregation of these components. It will be interesting to see how these developments take form and how they might shape future ideas of authorship.

In its current version the system uses aggregate counts for the time a user focussed on a particular sub-clips. Since the interactive player is set up to capture users' focus every 100 milliseconds, this could also become the resolution of recording. That way each video parts would have $length/\delta$ time bins in a interest histogram.

Chapter 7

Conclusions

This is the conclusions chapter.

- Human Computation might be especially applicable in UGVC (and multimedia) applications
- Human interaction can be a valuable source of information for the filtering of media content, even without explicit use of semantical concepts.
- We have proposed the basis of a system that allows for experimentation with implicit capturing of user interaction data and algorithmic reconfiguration of video segments based on the acquired data. The system can be used for filtering based on attentional data, but interaction opportunities can be extended.
- Specifically the method of Parallel Play may prove useful in allowing for segmented user preference elicitation in time-based media.
- Applied methods are very data-reliant and therefore need more experimentation for results to be further corroborated. An initial analysis though, hints that the methodology can enable meaningful filtering of unstructured video content in a way that is in line with user interests.
- Configuration of video parts to be tested
- Parallel sequences (= reconfiguration) are experienced as informative and entertaining.

Chapter 8

Acknowledgements

I would like to thank a number of people who've been instrumental in the development of this thesis as well as my own development along the way.

My supervisor Dr. Marian Ursu for providing focus, support and inspiration by asking the right questions every time again.

Prof. Mark Bishop for leading the programme, guiding us far and wide along different ideas in cognitive science, always with enthusiasm and personal attention.

'her Smallness' Lucia for her never-ending positive support during the structuring of our ideas.

The 'Chicken' of Batavia and its inhabitants for making the place like a second home to me. In particular Louise for keeping me up late with great music from all over the world.

The people together with whom we made 'Picnic' an inclusive, friendly and safe learning and teaching community, for sharing their knowledge and enthusiasm to make things happen.

Simone for being a truly nourishing host and dear partner in life and crime during my stay in beautiful Rennes les Baines, where much foundational work for this research was done.

Jan and Pim for bringing out the best in me, no matter how many kilometres our friendship spans.

Goldsmiths with her unconventional Department of Computing for having me experience the creative side of Computer Science and letting me merge my passions for AI and video.

Bibliography

[1] Creative commons licensed videos on vimeo. http://youtube-global.blogspot.co.uk/2012/07/heres-your-invite-to-reuse-and-remix-4.html. Accessed: 15/09/2012.

[2] Searching youtube for "burning man, this month. http://www.youtube.com/results?search_query=burning+man. Accessed: 17/09/2012.

[3] Vimeo timeline. https://vimeo.com/about/timeline. Accessed: 15/09/2012.

[4] Charles Babbage. On the economy of machinery and manufactures. 1832.

[5] R. Barthes. The death of the author (1968). *Image-music-text*, pages 142–148, 1977.

[6] E. Bruno and D. Pellerin. Video structuring, indexing and retrieval based on global motion wavelet coefficients. *Pattern Recognition, 2002. Proceedings. 16th International Conference on*, 3:287–290 vol. 3, 2002.

[7] Cathy Casserly. Heres your invite to reuse and remix the 4 million creative commons-licensed videos on youtube. https://vimeo.com/creativecommons, June 2012. Accessed: 15/09/2012.

[8] Pablo Cesar and Konstantinos Chorianopoulos. The Evolution of TV Systems, Content, and Users Toward Interactivity. *Foundations and Trends® in Human-Computer Interaction*, 2(4):373–95, 2009.

[9] M. Cha, H. Kwak, P. Rodriguez, Y.Y. Ahn, and S. Moon. I tube, you tube, everybody tubes: analyzing the world's largest user generated content video system. *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement*, pages 1–14, 2007.

[10] X. Cheng, C. Dale, and J. Liu. Understanding the characteristics of internet short video sharing: YouTube as a case study. *Arxiv preprint arXiv:0707.3670*, 2007.

[11] M. Christel and N. Moraveji. Finding the right shots: assessing usability and performance of a digital video library interface. *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 732–739, 2004.

[12] M.G. Christel and R.M. Conescu. Addressing the challenge of visual information access from digital image and video libraries. *Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 69–78, 2005.

[13] J. Davidson, B. Liebald, J. Liu, P. Nandy, T. Van Vleet, U. Gargi, S. Gupta, Y. He, M. Lambert, and B. Livingston. The YouTube video recommendation system. *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296, 2010.

[14] O. De Rooij, C G M Snoek, and M Worring. Query on demand video browsing. *Proceedings of the 15th international conference on Multimedia*, pages 811–814, 2007.

[15] O. De Rooij, C G M Snoek, and M Worring. Balancing thread based navigation for targeted video search. *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 485–494, 2008.

[16] D.A. Grier. *When Computers Were Human*. Princeton University Press, 2007.

[17] M.J. Halvey and M.T. Keane. Analysis of online video search and sharing. *Proceedings of the eighteenth conference on Hypertext and hypermedia*, pages 217–226, 2007.

[18] L Hollink, G P Nguyen, D C Koelma, A Th Schreiber, and M Worring. Assessing user behaviour in news video retrieval. *IEE Proceedings - Vision, Image, and Signal Processing*, 152(6):911, 2005.

[19] R. Jain and A. Hampapur. Metadata in video databases. *ACM Sigmod Record*, 23(4):27–33, 1994.

[20] H. Jenkins. *Convergence Culture: Where Old and New Media Collide*. ACLS Humanities E-Book. NYU Press, 2006.

[21] Anna Maria Jönsson and Henrik Örnebring. USER-GENERATED CONTENT AND THE NEWS. *Journalism Practice*, 5(2):127–144, April 2011.

[22] M.S. Kankanhalli and Y. Rui. Application potential of multimedia information retrieval. *Proceedings of the IEEE*, 96(4):712–720, 2008.

[23] H. Kuwano, Y. Taniguchi, H. Arai, M. Mori, S. Kurakake, and H. Kojima. Telop-on-demand: Video structuring and retrieval based on text recognition. *Multimedia and Expo, 2000. ICME 2000. 2000 IEEE International Conference on*, 2:759–762 vol. 2, 2000.

[24] E. Law and L. Von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. *Proceedings of the 27th international conference on Human factors in computing systems*, pages 1197–1206, 2009.

[25] Ying Liu, Dengsheng Zhang, Guojun Lu, and Wei-Ying Ma. A survey of content-based image retrieval with high-level semantics. *Pattern Recognition*, 40(1):262–282, January 2007.

[26] Lev Manovich. Who is the author? sampling / remixing / open source. `http://www.manovich.net/DOCS/models_of_authorship.doc`. Accessed: 16/09/2012.

[27] T. Mei, B. Yang, X.S. Hua, L. Yang, S.Q. Yang, and S. Li. VideoReach: an online video recommendation system. *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 767–768, 2007.

[28] R. Pea, M. Mills, J. Rosen, K. Dauber, W. Effelsberg, and E. Hoffert. The diver project: Interactive digital video repurposing. *Multimedia, IEEE*, 11(1):54–61, 2004.

[29] A.J. Quinn and B.B. Bederson. Human computation: a survey and taxonomy of a growing field. *Proceedings of the 2011 annual conference on Human factors in computing systems*, pages 1403–1412, 2011.

[30] Tom Rosenstiel and Amy Mitchell. YouTube & the News. Technical report, July 2012.

[31] Y. Rui, T.S. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *Circuits and Systems for Video Technology, IEEE Transactions on*, 8(5):644–655, 1998.

[32] N. Sawhney, D. Balcom, and I. Smith. HyperCafe: narrative and aesthetic properties of hypervideo. *Proceedings of the the seventh ACM conference on Hypertext*, pages 1–10, 1996.

[33] A.F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. *Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330, 2006.

[34] Alan F Smeaton, Peter Wilkins, Marcel Worring, Ork de Rooij, Tat-Seng Chua, and Huanbo Luan. Content-based video retrieval: Three example systems from TRECVid. *International Journal of Imaging Systems and Technology*, 18(2-3):195–201, August 2008.

[35] A.W.M. Smeulders, M Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(12):1349–1380, 2000.

[36] C G M Snoek, B Huurnink, L Hollink, M de Rijke, G Schreiber, and M Worring. Adding Semantics to Detectors for Video Retrieval. *IEEE Transactions on Multimedia*, 9(5):975–986.

[37] Cees G M Snoek and Marcel Worring. Concept-Based Video Retrieval. *Foundations and Trends® in Information Retrieval*, 2(4):215–322, 2009.

[38] AM Turing. Computing machinery and intelligence. *Mind*, 1950.

[39] A. Ulges, M. Koch, D. Borth, and T.M. Breuel. Tubetagger-youtube-based concept detection. *Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on*, pages 190–195, 2009.

[40] S. Xu, H. Jiang, and F. Lau. Personalized online document, image and video recommendation via commodity eye-tracking. *Proceedings of the 2008 ACM conference on Recommender systems*, pages 83–90, 2008.

[41] B. Yang, T. Mei, X.S. Hua, L. Yang, S.Q. Yang, and M. Li. Online video recommendation based on multimodal fusion and relevance feedback. *Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 73–80, 2007.

[42] M. Yang, B.M. Wildemuth, and G. Marchionini. The relative effectiveness of concept-based versus content-based video retrieval. *Proceedings of the 12th annual ACM international conference on Multimedia*, pages 368–371, 2004.

[43] R. Zhou, S. Khemmarat, and L. Gao. The impact of YouTube recommendation system on video views. *Proceedings of the 10th annual conference on Internet measurement*, pages 404–410, 2010.