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 19, 2012

Abstract

Digital video retrieval, filtering and reconfiguration are difficult tasks to solve using current computational techniques. An important cause of this difficulty is the semantic gap between a visual representation and the meaning we address to it. A solution commonly sought in AI research is to reduce the gap by visual analysis and the linking thereof to previously established symbolic representations of semantical concepts. These methods often perform poorly on unpredictable content found in the large video libraries of user generated content that account for much of the global internet traffic these days. A second way of hunting down meaning in visual content is to step over the gap altogether and ask people directly 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. 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 larger set of topically related content. We also investigate how such a system can be used for reconfiguration of content into new and informative stories. [concluding]

Contents

Chapter 1

Introduction

Chapter 2

The Quest for Meaning in Video

[The intention of this thesis is not to give an accurate explanation of daunting concepts like meaning or semantics, nor is it to give an in-depth description of the diverse work on the relationship between signifier and signified in the field of semiotics. This first chapter is meant to briefly introduce the difficulties that current computational methods have in arriving at a meaningful interpretation of visual content. To this purpose we formulate a framework of computational analyses of meaning that serves to establish terminology to work with, rather than to make claims about the internal functioning of human understanding or signifying systems.]

2.1 Computational Undertakings of the Quest for Meaning

2.1.1 Meaning in Visuals

[Video structuring, indexing and retrieval based on global motion wavelet coefficients][2]

2.1.2 Meaning in Concept

2.1.3 Meaning in Structure

[cite bordwell & Thompson: analysis of context dependency] [HyperCafe: Narrative and Aesthetic Properties of Hypervideo [24]]

2.1.4 Meaning in Annotations

[Telop-on-demand: Video structuring and retrieval based on text recognition][16] [Addressing the Challenge of Visual Information Access from Digital Image and Video Libraries][8]

2.2 Computational Undertakings of the Quest

2.2.1 The Semantic Gap

2.2.2 Steps towards Meaning: An Overview

Schematized summary of different steps: indexing automatic metadata annotating Human-Driven Labeling Machine-Driven Labeling multimodal feature fusion concept based content based [Relevance feedback: A power tool for interactive content-based image retrieval][23] [The Relative Effectiveness of Concept-based Versus Content-based Video Retrieval][30]

[collaborative filtering] `http://en.wikipedia.org/wiki/Collaborative_filtering`

2.2.3 Indexing to enable search

Because visual data on it's own provides little machine-readable handles to search and find, repositories of multimedia content need to be index to enable search. Within the task of video indexing several approaches are taken

2.3 Computational Difficulties

Chapter 3

Human Computation towards Visual Meaning

3.1 Characterising Human Computation

[Define GWAP]

In a recent survey paper, Quinn and Bederson present a taxonomy of Human computation systems[20]. 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"[26]

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'[12]. 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"[1, Ch. 20].

Quinn and Bederson further present a classification along six dimensions they see as the most salient distinguishing factors: [20]

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.

differences: public, readable // private, non readable conscious choice // unconcious result of interaction Concluding can be consciously put in place by several people at large scale // dependent on real 'human' traffic.

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][28]

[VideoReach: an online video recommendation system][18]

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. I

already happening at YouTube

3.6 A Characterisation of Human Computation Systems

3.6.1 Purpose

3.6.2 Motivation

3.6.3 Task

Chapter 4

Interactive Storytelling: From database to data-based

4.1 Symbolic approaches

[32, 21, 27]

4.2 Statistic approaches

4.3 Remixing

The reconfiguration of smaller units that carry meaning within themselves is common practice for textual media such as blogs, where it is easy to quote part of another author‘s writing in a new post [ref].

It needs to be said that some reconfiguration of videos is taking place, but even though it often concerns content that was originally sourced online, much of the creative act of remixing happens offline.

4.3.1 Taking the remix online

examples like: Aaron Koblin (johnny cash project, exquisite corps, etc)

there is even word of a true remix culture.

the availability of multimedia content via the internet has meant a surge in

Chapter 5

Human Computed Stories in wePorter

5.1 Introduction

This section describes the interaction design of the wePorter system, built to examplify how human computation can be used in tasks like local video part filtering 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.

5.2 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 [22]. At the same time, all these videos being put online beg the question which ones of them to watch.

[table of YouTube content uploads]

The increasing amounts of content being put online, lead to an information overload and present serious challenges in search and information retrieval tasks [ref]. 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?

5.3 The Purpose

Searching -¿ IR 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 ever blog post [ref]. An important activity on video platforms like YouTube is searching and much attention has been given to different methods of multimedia search and indexing [refs]. Youtube's acquisition by Google in 2006 underlines the platform's role as a video search engine.

5.3.1 Different Goals

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 [13][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 [13]. User goals can vary widely from more to less specific[9]. 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*[9].

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

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

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” [9]. 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 [31]. Features derived from users’ navigations such as ‘click-through rate’ have been used to improve content-based video recommendation [29].

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’[9].

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”[**?**]. 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[25]. Several works have indicated the importance of interactivity in the task of video retrieval to filter through a set of initially returned results [10, 7, 10, 11]. 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’.

5.3.2 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 [22] 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[17].

Because of the raw, unedited 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.

5.3.3 Storytelling as Structured Recommendation

[TODO; based on section in Meaning chapter]

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 unedited user-generated videos:

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

5.4 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.

5.4.1 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 in UGVC that gets uploaded to the web. While some are critical[15] 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[22]. 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.

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 topicalized 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 whish 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 'implicity work' as it involves activities that people already engage in for their own reasons[20]. Considering users' wish to be informed and the interactive way in which they navigate, there is most likely also a factor of enjoyment 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.

5.5 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.

5.5.1 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[19]:

- "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"[24]. 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."[14]

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[4]. 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

Originally voiced by Roland Barthes, who was contemplating a way of literary writing without the use of a clear narrator. He titled his essay 'the death of the author' to signify the lack of presence of an author in written work following this style of writing. In our aim of reconfiguring interesting sub-clips in a new arrangement, the issue of authorship 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[17]. 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 [3]. 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[**?**] and many of their users licence their videos such that they can be remixed by others. Figure 5.1 shows that a large part of the licences on the Vimeo platform allow derivatives to be made[**?**].

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

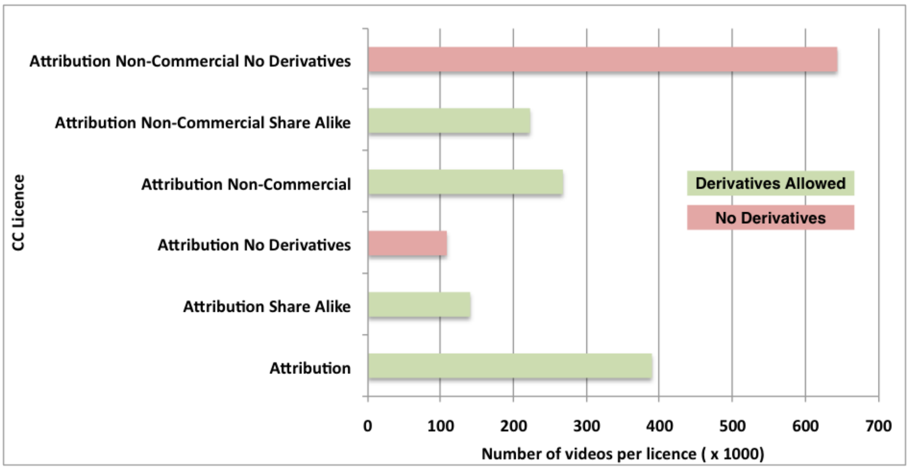


Figure 5.1: Number of videos for each of the Creative Commons licences on Vimeo

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 [...]."[**?**]

5.5.2 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.

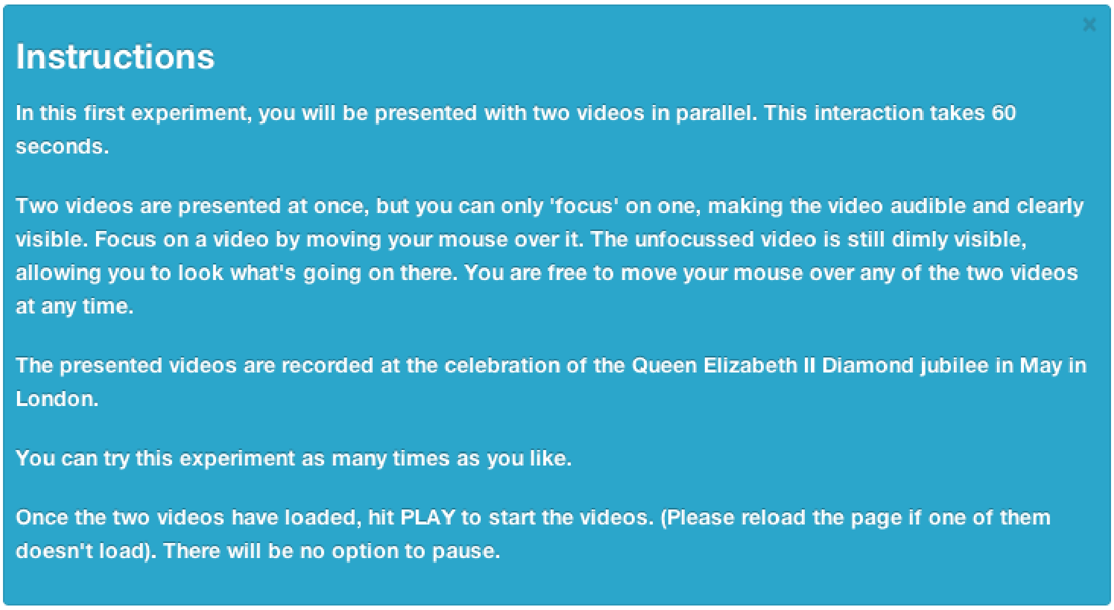
We hypothesise that interesting parts of content will attract a relatively large amount of attention compared to less interesting parts.

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 5.2.



Instructions

In this first experiment, you will be presented with two videos in parallel. This interaction takes 60 seconds.

Two videos are presented at once, but you can only 'focus' on one, making the video audible and clearly visible. Focus on a video by moving your mouse over it. The unfocussed video is still dimly visible, allowing you to look what's going on there. You are free to move your mouse over any of the two videos at any time.

The presented videos are recorded at the celebration of the Queen Elizabeth II Diamond jubilee in May in London.

You can try this experiment as many times as you like.

Once the two videos have loaded, hit PLAY to start the videos. (Please reload the page if one of them doesn't load). There will be no option to pause.

Figure 5.2: wePorter Instructions

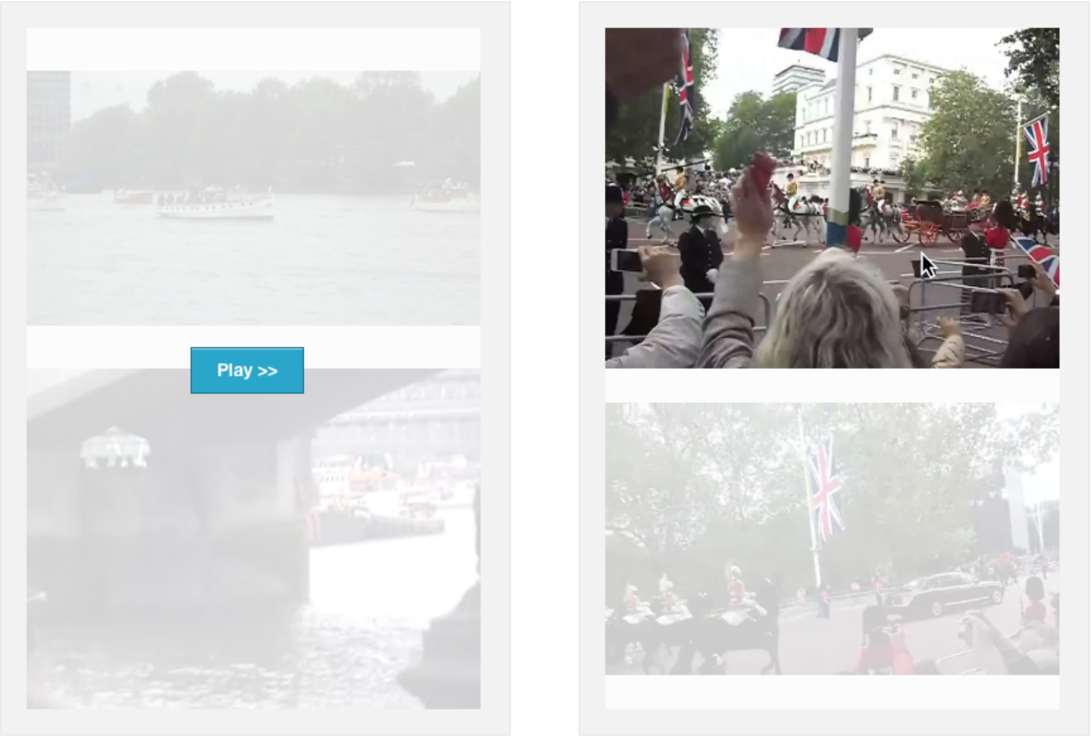
Upon loading the webpage, a database is queried for a pair of sequences made up from different video parts for which the system would like to elicit a user's preference. The two sequences both consist of six video parts that each have a duration of 10 seconds. The interaction of the two sequences playing in parallel thus has a total duration of 60 seconds, reflecting the short time span common across UGVC at YouTube[5, 6]. A detailed description of the algorithm used for the construction of these sequences is given in section 5.6.

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



(a) Upon load

(b) While playing, focus on top video

Figure 5.3: The wePorter parallel play interface

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 **??** we report on how these measure can be telling in the process of filtering and reconfiguration.

5.6 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.

5.6.1 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.

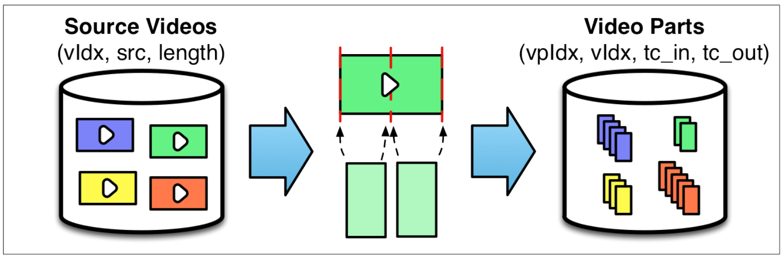


Figure 5.4: Slicing Source Videos into Video Parts by Reference

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:

$$video = (vIdx, srcPath, length) \tag{5.1}$$

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:

$$videoPart = (vpIdx, vIdx, tc_{in}, tc_{out}) \tag{5.2}$$

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 caused them to film has ended and so the final bit of their videos does not commonly contain the most important content.

5.6.2 The live system

Figure 5.5 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 play. Once an interaction have 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 a new interactions either by the same or new visitors.

Algorithm 1 Generate Video Parts

```
procedure SLICE($sourceVideos, d$)        ▷ slice videos into parts with duration $d$
    $i \leftarrow 1$
    for all $video \in sourceVideos$ do
        $tc_{in} \leftarrow 0$
        while $tc_{in} \leq video.length - d$ do
            $tc_{out} \leftarrow tc_{in} + d$
            $videoPart_i \leftarrow (video.src, tc_{in}, tc_{out})$
            $tc_{in} \leftarrow tc_{in} + d$
            $i \leftarrow i + 1$
        end while
    end for
end procedure
```

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. A useful extension of the system would be to let users query a live database like YouTube for content of their interest and thus define a dynamical collection of content to explore. To focus experimentation on a fixed set of video content, this extension is left 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.

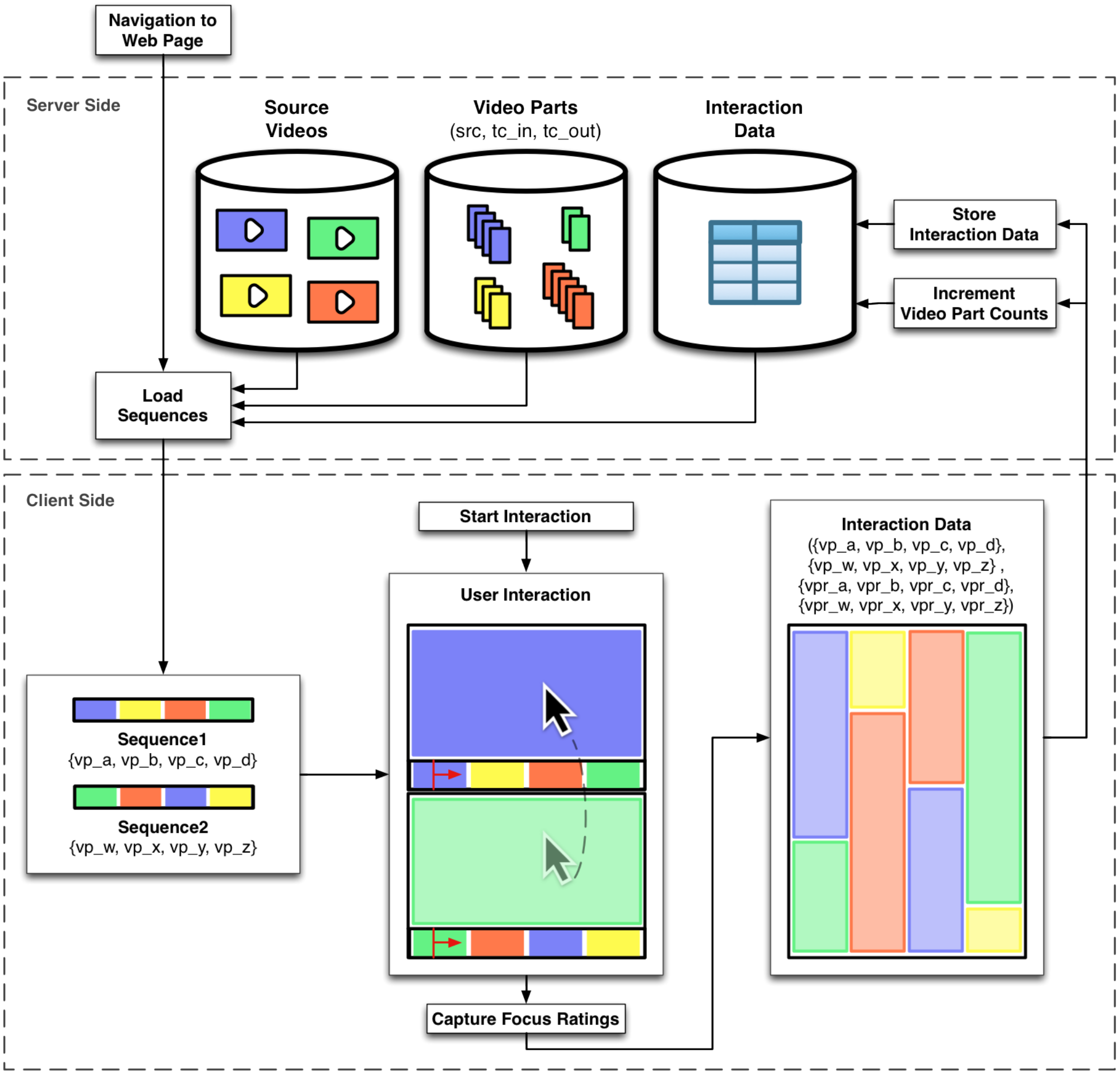


Figure 5.5: System Framework of wePorter

Algorithm 2 Load Sequences Random Shuffled

```
procedure LOAD SEQUENCES($n\_parts, videoParts, vpCounts$)
    $seq_1 \leftarrow []$
    $seq_2 \leftarrow []$
    for $i \leftarrow 0, n\_parts$ do
        $selectionH_1 \leftarrow []$                ▷ keep selections of parts that satisfy constraints
        $selectionH_2 \leftarrow []$
        for all $vp \in videoParts$ do                              ▷ Horizontal constraint
            if $vp.vIdx \neq part.vIdx$ for all $part \in seq_1$ then
                add $vp$ to $selectionH_1$
            end if
            if $vp.vIdx \neq part.vIdx$ for all $part \in seq_2$ then
                add $vp$ to $selectionH_2$
            end if
        end for
                                                                      ▷ For $seq_1$:
        $minSelection_1 \leftarrow []$   ▷ select from $selectionH_1$ parts that have minimal count
        $minCount_1 \leftarrow min(count)$ from $selectionH_1$      ▷ look up counts in $vpCounts$
        for all $vp \in selection_1$ do
            if $vp.count = minCount_1$ then
                add $vp$ to $minSelection_1$
            end if
        end for
        $selected_1 \leftarrow$ random from $minSelection_1$
        append $selected_1$ to $seq1$                            ▷ add video part to $seq_1$
                                                                      ▷ For $seq_2$:
        $selectionV = []$
        for all $vp \in selection_2$ do                               ▷ Vertical constraint
            if $vp.src \neq selected_1.src$ then
                add $vp$ to $selectionV$
            end if
        end for
        $minSelection_2 \leftarrow []$    ▷ select from $selectionV$ parts that have minimal count
        $minCount_2 \leftarrow min(count)$ from $selectionV$       ▷ look up counts in $vpCounts$
        for all $vp \in selection_2$ do
            if $vp.count = minCount_2$ then
                add $vp$ to $minSelection_2$
            end if
        end for
        $selected_2 \leftarrow$ random from $minSelection_2$
        append $selected_2$ to $seq2$                            ▷ add video part to $seq_2$
    end for
    $shuffle\_in\_unison(seq_1, seq_1)$
    return $(seq_1, seq_2)$
end procedure
```

Algorithm 3 Shuffle Sequences in Unison

```
procedure SHUFFLE IN UNISON($seq_1, seq_2$)
    $seed \leftarrow make\_seed()$    ▷ to seed $random\_generator$ twice identically
    $random\_generator.set\_seed(seed)$    ▷ set seed
    $random\_generator.shuffle(seq_1)$    ▷ shuffle
    $random\_generator.set\_seed(seed)$    ▷ set seed
    $random\_generator.shuffle(seq_2)$    ▷ shuffle
    return $(seq_1, seq_2)$
end procedure
```

5.6.3 Hypervideo in your 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, who's foundation is working on '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 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.

[2] http://mozillapopcorn.org/

[3] http://popcornjs.org/

Chapter 6

Evaluation

This is the Evaluation section.

6.1 Experiments

For our experiments described in this section we use a set of 10 unedited, user-generated videos that were returned in response to a query for "Diamond Jubilee London" on YouTube.

6.2 Results

6.2.1 The Interface

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 and report on this in section 6.

6.2.2 Landscapes of Interest

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. Afterall, 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 perhaps even a more smooth landscape of interest. This kind of data could reveal mountains and valleys that can be used for interest-based segmentation. From the thus segmented parts, the ones with high interest scores can be returned as the salient parts within a video.

Chapter 7

Discussion

Chapter 8

Future Directions

This is the Future Directions chapter.

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 9

Conclusions

This is the conclusions chapter.

Bibliography

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

[2] 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.

[3] Cathy Casserly. Here's your invite to reuse and remix the 4 million Creative Commons-licensed videos on YouTube. Technical report, June 2012.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] Pablo Cesar Dick C A Bulterman Vilmos Zsombori Ian Kegel Rodrigo Laiola Guimaraes. Creating Personalized Memories from Social Events: Community-Based Support for Multi-Camera Recordings of School Concerts. Technical report, August 2011.

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

[23] 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.

[24] 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.

[25] 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.

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

[27] Marian F Ursu, Vilmos Zsombori, John Wyver, Lucie Conrad, Ian Kegel, and Doug Williams. Interactive documentaries. *Computers in Entertainment*, 7(3):1, September 2009.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] V. Zsombori, M. Frantzis, R.L. Guimaraes, M.F. Ursu, P. Cesar, I. Kegel, R. Craigie, and D. Bulterman. Automatic generation of video narratives from shared UGC. *Proceedings of the ACM Conference on Hypertext and Hypermedia*, pages 325–334, 2011.