

Storyboard Based Video Browsing: Optimizing human visual inspection for video archive navigation and search

Algernon Ip Vai Ching
ICA-3361071



Department Computer Science
Utrecht University

16-06-2017

Supervisor: Hürst, W.
2nd supervisor: Wiering, F.

Abstract

With the size increase of video databases it is important to create an efficient video browsing system. We investigate how to optimize the interface, relying solely on human visual inspection. This thesis sets its focus on the latter because research has shown the tremendous power of such human abilities. Yet, despite the resulting potential advantages, these facts are often neglected in current system designs. We go into the fundamentals of a storyboard system which has been proven to perform well in the VBS 2015 by testing the size and layout of its frames, and research the contribution of the storyboard system, when integrating it with a video browsing system relying on filtering of results. The results yielded statistically significant differences between certain size and layout combinations, while the contribution test yielded less strong, yet positive evidence. Suggestions for future work consist of researching variables that could improve the storyboard interface, such as the color contrast, that could enhance object recognition as well as ways to optimally integrate filtering with an intelligent interface by testing the effects of different filters as well as experimenting with different frame rates.

Table of contents	
1. Introduction	3
2. Related work	5
3. Design options & research questions	7
3.1. Experiment 1: size and layout	7
3.2. Experiment 2: Segmentation versus brute force	9
4. Experiments	11
4.1. Experiment 1: size and layout	11
4.1.1. <i>Independent variables</i>	11
4.1.2. <i>Dependent variables</i>	11
4.2. Experiment 2: Segmentation versus brute force	12
4.2.1. <i>Independent variables</i>	12
4.2.2. <i>Dependent variables</i>	13
4.3. Tasks and databases	13
4.3.1. <i>Experiment 1: size and layout</i>	13
4.3.1.1. Subjects & experiment design	14
4.3.2. <i>Experiment 2: Segmentation versus brute force</i>	14
4.3.2.1. Subjects & experiment design	15
4.4. Procedure:	18
4.4.1. <i>Experiment 1: size and layout</i>	18
4.4.1.1. <u>Material & environment</u>	19
4.4.2. <i>Experiment 2: Segmentation versus brute force</i>	20
4.4.2.1. <u>Material & environment</u>	20
5. Analysis and discussion of results	21
5.1. Experiment 1: layout and size	21
5.1.1. <i>Performance</i>	22
5.1.2. <i>Discussion performance</i>	25
5.1.3. <i>User experience</i>	26
5.1.4. <i>General observations thumbnail sizes</i>	27
5.1.5. <i>Discussion user experience</i>	28
5.2. Experiment 2: Segmentation versus brute force	28
5.2.1. <i>Performance</i>	28
5.2.2. <i>Discussion performance</i>	30
5.2.3. <i>User experience</i>	30
5.2.4. <i>Discussion user experience</i>	32
6. Conclusion and future work	33
6.1. Conclusion	33
6.2. Future work	33
7. Reference	34
8. Appendix	35

1. Introduction

Browsing through a large video archive can be quite challenging now that the number of videos available to the public anywhere and at any time has become very large. Finding a specific video among hundreds of hours can require a lot of time, let alone a specific clip within a particular video. The need for an efficient and effective search system thus increases. An effective search is obviously realized by an effective search engine, but another (often forgotten) factor is a good interface; whereas the search engine seeks the searched items and sorts the search result, the interface is concerned with presenting a good visual representation of the results, making it as efficient as possible to go through them and identifying the (most) correct items within the results. Naturally an inefficient interface will lead to longer search times.

Therefore, the general aim of this research is:

to contribute to the ultimate goal of creating the best interface for accessing large video archives.

The quality of the interface can be determined by two factors that need optimization. Firstly, there is the quantitative data: the most important factor is generally **performance** i.e., finding the relevant content as fast as possible with a minimal number of mistakes. Thus, the relevant quality measurements are **speed** and **accuracy**. We will use this definition for performance, which has its basis in the VBS (Video browser showdown), a video browsing competition held annually to test different video browsing systems against each other [1]. The many years of using this definition have proven its effectiveness to test the ability for a video browsing system to fulfill its purpose. The values of the relevant quality measurements that make up the performance factor are determined for so-called Known Item Search tasks (KIS; see section three). The second factor to improve the quality of a vid search engine is qualitative: the **user experience** (e.g., enjoyment, perceived performance and workload, ease of use, etc.). Here, we focus on two sub criteria: the **workload** and **perceived performance**. The importance of the user experience lies with the fact that too much workload can lead to mental fatigue which would demotivate the end users to use the system as well as influence their performance. Even if the performance is good, it is also important that the user experiences it as such, because it could demotivate the user otherwise; therefore, in this thesis the perceived performance is also tested. The workload and perceived performance are measured using the **NASA TLX**, which is a standard for workload measurement in research. Additionally, the perceived performance data will be complemented with a **comparative questionnaire**.

One method of creating an efficient and effective interface for video browsing is the use of a **storyboard**, which is similar to the ones used in animations and video editing. An animation or video is essentially represented by a set of extracted images shown one after another in temporal order. Taking these images according to a certain sample rate and laying them out in a certain order that makes sense in a natural way which is usually temporal, a storyboard is created. The positioning of the frames (i.e., the images that make up the video) do not necessarily have to be in a horizontal line from left to right like a traditional storyboard (cf. animation and video editing). With a system used in the VBS 2015 event, the observation was made that simple storyboards based purely on visual inspection can work equally effective as complex querying systems [10]. It has led to a series of studies in which the important design factors have been analyzed in order to optimize the system [12, 14]. These have shown promising results in terms of effectiveness of the

storyboard, however they have not yielded solid evidence for an optimal configuration in terms of the above-mentioned quality factors.

One of these experiments [11] raised the assumption that the interaction factor (i.e., scrolling in the case of the system used in the VBS 2015 event) could have a strong influence on the measured effectiveness of the different layouts used for the storyboard. Therefore, to improve the system used at the VBS 2015 and in the subsequent tests we further investigate this issue. In particular, we present an experiment which isolated the involved variables (i.e., size and layout; see section three) in order to clearly investigate their influence on the quantitative and qualitative factors as specified above.

With increasing video archive sizes, it is impossible to solely rely on human visual inspection while maintaining the performance. However, that does not automatically make it inessential to the effectiveness of a video browsing system. To find the contribution of an optimized interface design, it has to be tested against additional ways of filtering. In a second experiment, we have thus shifted the focus to this balance between filtering and optimizing the interface design. This was done by performing tests similar to the first experiment using the same factors for quality determination, but instead comparing a system using only an optimized interface design (in this case a storyboard based video browser) against a system that uses a form of filtering additional to this design (see section three). In order to make an accurate comparison the video archive had to have a size at which the system was proven to still be effective. Since the concept of the storyboard layout originated from the VBS 2015 system (where it had proven itself), the related database has been considered the upper limit. To keep the second experiment as consistent with the first experiment as possible, both experiments have approximately the same database size, which falls well under the limit of the VBS 2015 database. The comparison made in the test was between a full database, and a database that has been filtered using the same **shot detection** as used in [13].

With these two experiments, we will contribute to the aforementioned general aim by adding these concrete issues of the following research problem:

to develop an optimal storyboard design that relies purely on human based search in the form of visual inspection.

In this thesis, we will be going to review the two experiments in the following way: first we will go through the related work that has led us to the setup of these two experiments. After that we will discuss our design decisions and the setup, followed by the results and our interpretation of our findings. We close off with a conclusion at which we will evaluate the results and the setup, followed by suggestions for future work.

2. Related work

Video browsing is a very complex interactive task to perform efficiently. The use of frames as a means to create an abstraction of a video has been used widely in different ways in attempts to solve this problem. For example, Arman [2] created so called Rframes, which were representative frames for shots, in order to make an abstract of the videos which could be navigated through using a side scrolling browser. Zhong [3] had a different approach by using hierarchy based clustering, to be able to get a rough overview of the contents of the video, and to go deeper into the details if needed. Rather than using static key frames, so called dynamic key frames have also been used to create an effective video browsing system. Ding [4] experimented with several frame rates for dynamic key frames, and found that eight to twelve fps was the limit for proper object recognition. Tse [5] compared several keyframe extraction techniques, both static and dynamic, and tested the possibility of using multiple dynamic frames simultaneously laid out. The conclusion for the latter was that the most notable performance drop started between three and four simultaneous dynamic key frames.

In [6] Komlodi compared the difference between static and dynamic key frames using a storyboard; in their experiment their static storyboard display scored statistically significantly better at object identification than the dynamic slideshow. However, in the other factors that were established as performance indicators the differences did not yield any significant differences.

A hybrid was presented by Jackson [7] where a storyboard system of video instances of one particular video, named Panopticon is presented with a consistent time interval between each video instant looped so that every scene can be processed uninterruptedly. The study compared the system against Youtube and Videoboard (mimicking video editing software) searching information within videos, and found that statistically the use of the Panopticon system was significantly faster, as were the search times. The reason that we have decided to not make use of dynamic keyframes stems from the fact that video browsing requires not only searching within videos, but also between them. In order to minimize the cognitive load (and conserve object identification performance as stated above), we have decided to use static keyframes instead.

Herranz [8] compared the effects of different multiscale sampling methods for extracting frames to create a storyboard for video collections. The objective and subjective evaluation both suggested a better performance for the “scalable” approach (based on hierarchical clustering) in terms of providing a useful video abstract.

Hürst [9] proposed two interface designs for mobile devices for video browsing, one using a storyboard, which would later lead to the VBS 2015 submission, and ultimately the basis for the experiments reported in this thesis. The storyboard grid interface was based on preceding research on, among others, the size influence on search performance using thumbnails [16,17]. The clustered layout, with cluster size five (see section three), was introduced to the storyboard [10] to optimize the interface for quick scanning through the storyboard, in a rough but efficient manner (i.e., relatively large chunks can be quickly filtered out). This strategy is in line with hierarchical clustering in the sense that the human visual inspection in this system works in a top down approach when it comes to filtering the data. Despite a change in tasks, the storyboard system scored above expectation, reaching third place in the 2015 competition.

A follow up study researched the combinations of different sizes and layouts on the mobile devices [11], particularly tablets. In terms of user experience, the results showed a general preference for the cluster layout. Performance wise the results did not yield a significant

difference, although they showed a slight advantage, particularly in the number of correctly completed tasks, and to some degree in the VBS score. After evaluating the results, we have decided to perform a follow up study, which we will discuss in this thesis (experiment 1: Size and layout).

As indicated in [10], the human visual inspection limit will most likely not be able to deal with large databases, and as the sizes of those rise, at some point will be outperformed by automatic filtering simply because there is too much data to process in too little time for the human brain. Hence, a collaborative system was created, which tackled the problem of the limited amount of data that can be processed by human computational power by doing advanced computer filtering using a pc, as well as compensating for the lack of semantic understanding of the computer, which would be done by human visual inspection, using a tablet. This concept proved to work very well, as it had reached second place in the VBS 2016 competition [12]. In light of this success, the second experiment reviewed in this thesis, will investigate the contribution of the human visual inspection to the collaboration.

3. Design options & research questions

There are many design options that can be considered that could potentially optimize a storyboard interface. In this thesis, we consider some important ones with two experiments. In this section, per experiment we will go into what the design options are and the motivation behind them. From there we will go into what decisions we have made regarding the design options for the experiments and formulate the relevant research questions that we will explore.

3.1. Experiment 1: size and layout

The power of VBS 2015 system lies in the data presentation to the user, having the thumbnail size and layout as the main design options that decide how the storyboard is presented. Using the VBS system as starting point we want to further optimize these components to improve the system and in that way, contribute to solving our research problem. As mentioned in section one, in order to get a clearer image of the influence of the size and layout on the above-mentioned factors, these design options need to be isolated properly when evaluating their influence on performance and experience. To achieve this goal, the number of screens has been minimized to one per task, meaning that the videos in the storyboard get cut off at that point.

While small thumbnail sizes provide an easy overview as it takes less eye movement to scan over an entire scene or a collection of scenes, it is also likely to take more time to recognize the contents of a single thumbnail. For larger thumbnails, this works the other way around. Having larger thumbnails also means having less thumbnails per screen and vice versa. In the experiment performed in [11] this resulted in having more screens to navigate through the entire video archive when using larger sizes of thumbnails. It is still implied in this first experiment, but due to the isolation of the variables, which involved removing the scrolling interaction, not taken into account. With this first design decision comes the first research question:

RQ1. What is the *optimum thumbnail size* (with respect to performance indicators specified above)?

The second design decision to be evaluated is the layout for the storyboard. The linear layout is widely used in several video-editing applications; the direction in which the chronological order is arranged (i.e., left to right, bottom to top) is the same as the reading direction in western writing, which suggests that people are strongly conditioned to this way of processing similar types of visual information. A possible disadvantage for this is the fact that it may require a lot of eye movement when searching through the content (in this case the storyboard). As an alternative the other layout used in this experiment is the cluster layout. The reason for choosing this layout as an alternative lies with the results of the VBS 2015 system mentioned above from which it originates.

Because it scored well during the competition, this possible solution proved to be worth investigating. The idea behind this arrangement is grouping of similar thumbnails (i.e., the ones that belong to the same scene) together providing more contextual information in a single spot compared to spreading out an entire scene over 1 or more lines. This might possibly make it easier to filter out entire scenes that are irrelevant for the search task (see figure 1a and 1b). However, since it might be less intuitive, it may not have the desired effect. The design decision of layout brings up the second research question:

RQ2. What is the *optimum thumbnail arrangement* (with respect to performance indicators specified above)?

With these two design decisions, there are several possible configurations. Each of these configurations can have a different outcome in terms of performance. The effects of size could for example be different with a linear layout than with a cluster layout. It is therefore important to also evaluate these different configurations and to see whether a relation between these two decisions exist, and if so how they influence each other. This brings us to the third research question:

RQ3. What is the *mutual dependency of thumbnail size and arrangement* (with respect to performance)?

Besides the performance, the user experience can also be influenced by the design options. Therefore, it is important to evaluate their effect on the workload. Layout wise there is a trade-off between the advantages of the clustered scenes versus the natural way of reading; if it takes too much effort to adapt to a different reading direction the advantage of clustering might not weight up to that disadvantage.

As explained in section one, the system should also be experienced as good performing besides having an actual good performance for if it does not, then the motivation for using it will be low, which negatively affects the usability. The optimization of the qualitative factors (i.e., the user experience) that are related to the research problem raise the following research questions:

RQ4. How does *thumbnail arrangement* influence *experienced workload*?

RQ5. How does *thumbnail arrangement* influence *perceived performance*?

In addition to this we are interested in general insight into usability with respect to thumbnail sizes, in particular about how the latter impacts user experience. The object identification should not be too intensive, while at the same time it should not be too hard to keep the overview of the events within a scene or between different scenes. However, we expect that the influence to be minimal, hence the greater focus is set on the layout, meaning that the experiment will be set up in such a way that the tests will be done per layout (see section 4.4.1).

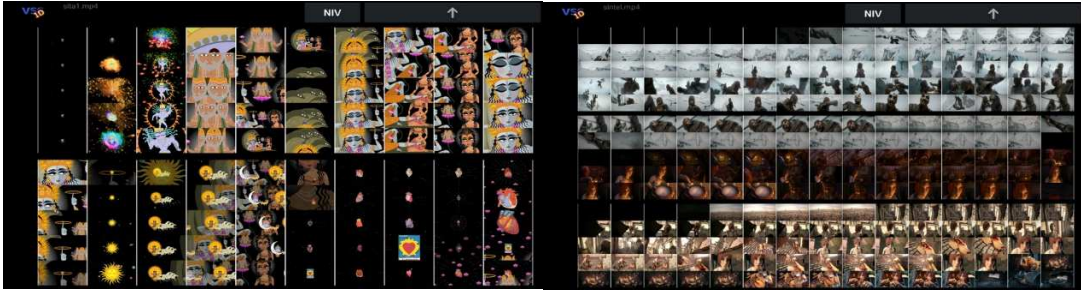


Figure 1: a) cluster layout b) linear layout

3.2. Experiment 2: Segmentation versus brute force

Despite the success of a video browser based solely on human visual inspection with smaller databases, it only scales to a certain extend. At some point a form of filtering is inevitable to maintain usability. The initial idea was to implement several filtering features: color filtering, and two forms of concept filtering, one being a list of keywords, and a more complex concept database using a search bar. Additionally, the database would be pre-filtered using a shot detection. This system was planned to be used to participate in the VBS 2017, however due to sickness, we have not been able to meet the deadline and we have therefore shifted the focus to the shot detection.

In our second experiment, we investigate the effects of an optimized interface design onto a filtering system. We compare the system based purely on human visual inspection using an unfiltered database to a system that has filtering applied to it (i.e., shot detection) in terms of the same performance and user experience criteria as the first experiment. We will refer to the systems as either the **brute force** or the **shot segmentation** approach respectively (see Figure 2a and 2b). This design decision raises the following research question:

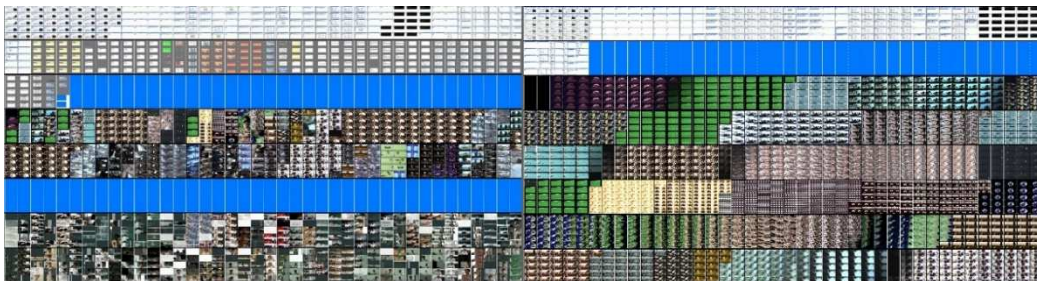


Figure 2¹: a) brute force b) segmentation

¹ Video data under creative commons license [18]

RQ6. How much does a *storyboard interface design* contribute to the *performance* of a video browsing system that relies on both *human visual inspection* and *filtering* (with respect to the performance indicators specified above)?

Consistent with the first experiment, the user experience is also covered in the second experiment between segmented and the brute force approach. This leads to the following two research questions:

RQ7. How much does a *storyboard interface design* contribute to *minimizing experienced workload* of a video browsing system that relies on both *human visual inspection* and *filtering* (with respect to the user experience indicators specified above)?

RQ8. How much does a *storyboard interface design* contribute to the *perceived performance* of a video browsing system that relies on both *human visual inspection* and *filtering* (with respect to the user experience indicators specified above)?

4. Experiments

4.1. Experiment 1: size and layout

4.1.1. Independent variables

For this experiment, the independent variables were the **thumbnail sizes** and the **layout**. We have taken three thumbnail sizes, with the following number of thumbnails per screen: from smallest to largest thumbnails this is 100, 225 and 625 thumbnails per screen. As the last two have been used in the VBS 2015, chosen after informal testing and convenience purposes, in order to evaluate the trade-off mentioned in section three, a larger size (thus less thumbnails per screen) had been added to further investigate the effects of sizes. Because the smallest size was barely recognizable, it made more sense to investigate a larger size. There were two thumbnail arrangements/layouts, which were the linear layout as the standard, and the cluster layout as the alternative solution with cluster size five, meaning that this layout has columns of five from left to right with the chronological order going from top to bottom and then continuing at the top of the next column. The reason for choosing this configuration is the previous success from the VBS 2015 submission which used a similar set up. In order to keep the test cases minimal only one cluster size has been tested, which seems a good size, since smaller sizes tend to lose the grouping effect and become more similar to the linear layout whereas the larger sizes tend to make the reading direction focus too much on the vertical direction, and this also reduces the grouping effect (the cluster size should ideally be as long as the number of clusters that the scene contains). The storyboard, and thus the video archive, consists purely out of an unfiltered extraction from the video, at one frame per second. This step width has been chosen because there needs to be a sufficient amount of footage represented to get a proper overview of the video.

4.1.2. Dependent variables

In order to verify the performance there are two dependent variables; these are the components of the VBS score. The first variable is the **speed**: the average time to solve a task should be as low as possible for each combination of size and layout. The second variable would be the **accuracy**, requiring the most amount of correctly completed tasks, with the least amount of incorrect answers (i.e., trying to avoid a brute-forced correct answer).

The user experience is also verified by two dependent variables: **the workload**, and **the perceived performance**. The former is measured with the *NASA TLX*, which quantifies the workload using a 7-point scale. Despite the numerical characteristic of the TLX, translating a qualitative feature (such as workload) to quantitative values is prone to inaccuracies and therefore significance testing on this would not make sense, so this has been omitted. The perceived performance will be tested using the relevant questions of the *NASA TLX*, as well as a comparative questionnaire. The comparative questionnaire will ask the test subjects about the layouts, and how practical they were in their experience and which one they would prefer in a concrete given situation.

All scores and actions, provided with a timestamp, were logged to a file for each user, with each score divided by the size/layout combinations in order to determine the score, and to make sure that the user did not perform any irregular activity (e.g., no scrolling at all, while not close to the correct video, picking the wrong video multiple times etc.). Actions consisted out of selecting a thumbnail, and either pressing the **submit button** or the **NIV** (not in video) button. Furthermore,

standard info about each round (thumbnail size, layout, correct or incorrect and the occurrence of a time out) are being logged along with a timestamp.



Figure 3: NIV and submit button

4.2. Experiment 2: Segmentation versus brute force

4.2.1. Independent variables

In the segmented versus brute force experiment, amount of sizes and layouts has been reduced to only one configuration. However, the independent variable in this experiment **is the use of the shot segmentation** (or not). Instead of a tablet, we decided to implement the system on a laptop, which provided more screen space. However, after doing some informal testing it became clear that spreading the thumbnails across the entire screen required too much head movement, making it physically more demanding than the tablet version, as with the tablet head movement could be decreased by moving the tablet. Using a 40x40 grid covering the entire height of the screen was a good compromise, so the maximum number of thumbnails per screen came down to 1600. The reason for this choice stems from the original plan to participate in the VBS 2017, which had a significantly larger database than previous years; this introduced storage issues for the tablet, so a laptop or pc would make more sense. Since the extra screen space advantage was still present and it ensured a static screen (cf. tablet; see section 4.4.1.1) with the laptop implementation, it made sense to not go back to a tablet. The layout used in both systems for the experiment was the cluster layout with cluster size five due to the results from the size and layout experiment, and in regards to minimizing the eye movement this also made sense. The system without any filtering had a database of a 1fps extraction from the videos. The shot detection has been applied by using a 25fps extraction of the videos. Each shot interval was then divided that shot by 25. For each $1/25^{\text{th}}$ of that interval, the frame of the 25fps extraction that was closest to the time that was represented by the $1/25^{\text{th}}$ was taken and placed in the storyboard. Because of that, each shot was represented by a block of 5x5 frames in theory. However, since the shot detection does not have a 100% accuracy. In this configuration, the user often had more frames of a specific scene, and in most cases, more frames that mapped to the correct scene, but as a trade-off this also meant that the test subject had more frames to process per video (see figure 4a and 4b for screenshots of the application).

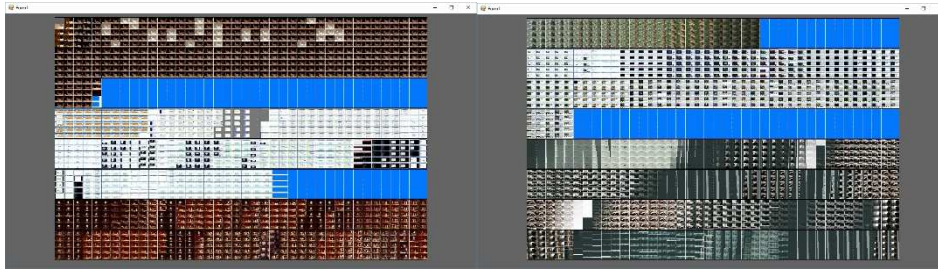


figure 4²: a) segmentation b) brute force screenshots

4.2.2. *Dependent variables*

For the dependent variables, it made sense to remain consistent with the performance guidelines of the VBS so therefore **speed** and **accuracy** were also the variables to be measured in this experiment. To stay consistent with the size and layout experiment, we also used the NASA TLX and the corresponding questionnaire to measure the **workload** and **perceived performance** respectively.

Correct answers, incorrect attempts and scrolling were again all logged to measure accuracy, with all of these actions provided with a timestamp to keep track of the speed.

4.3. **Tasks and databases**

4.3.1. *Experiment 1: layout and size*

The size and layout experiment consisted of finding a five second video clip in a storyboard with a one frame per second video extract. In the size and layout experiment the video clip would either be an excerpt of the video that was represented in the storyboard, in which case the video clip had to be selected in the storyboard, or it could be an unrelated clip. In the latter case the NIV button had to be pressed. The selection of the video clip in the storyboard consisted out of pressing one of the five thumbnails that contained the video clip; which of the five thumbnails was selected was irrelevant. Each task had a time limit of 10 seconds. If the submitted answer was incorrect, then the test subject would get notified about the answer being incorrect, after which another attempt could be done. There was an infinite number of attempts possible, as long as it was within the time limit; however, from two mistakes on, every next mistake would lead to a point reduction for the respective task to be completed. In the case that the submitted answer was correct, the app would indicate that the answer was correct and the user would get a five second break until the next task was presented, along with the remaining number of tasks for the current size/layout combination. If the NIV was pressed, the test subject would not be notified on the correctness of the task, but would instead immediately be given next task after the five second break. This was done since the NIV button can only be pressed once; if it was incorrect it would not be logical anymore to present that option. Every test round (which covered three sizes) would start out with a demonstration from the tester. After that the test subject had to perform four tasks

² see footnote 1

per size/layout combination, the first being a test for the test subject to get used to the size, so this would not count for the score. The total amount came down to 26 tasks (of which two are done by the tester). For the score the formula of the VBS score has been used as this can be considered as a standard for test like such (see section five for the formula). An incorrect NIV press would count as a time out. In the segmented versus brute force experiment, the rules for searching the video were similar (see section 4.3.2).

The video archive consisted out of a mix of different types of videos, both animated and live action, all of them from royalty free for legal purposes (see table 1 for a detailed list of the videos). There was a total of 6 videos used in the experiment, but the videos have been split up into a total of 13 video excerpts (one excerpt from *Sita sings the blues* is dedicated to the tester demo). The tasks have been spread over the entire storyboard, in order to prevent testers from being able to estimate where the next task will be. The tasks themselves were pseudo-randomized in the scripts.

Title	Type (animated/live)	Genre	Length	Number of thumbnails	#tasks "Real"
Big Buck Bunny	Animation	Comedy	9:56	595	4
Elephants Dream	Animation	Sci-Fi	10:53	653	4
La chute d'une plume	Stop motion	Bricolage	10:23	622	4
Sintel	Animation	Fantasy	14:48	887	4
Sita Sings the Blues	Animation	Comedy/fantasy	1:21:31	4891	13
Valkaama	Live action	Drama	1:33:13	5585	20

Table 1: video archive size and layout experiment

4.3.1.1. Subjects & experiment design

In the size and layout experiment, each test person has tested all combinations of size and layout. This way there is a stronger comparison between the combinations performance wise, but also in the preferences of the test subjects. In order to eliminate a possible learning effect of having a certain order of combinations, the test subjects were divided between six different categories of different orders to in terms of size and layouts. The total amount of test subjects was 42 (i.e., seven test subjects per category).

In this experiment the test subjects consisted of university students that voluntarily participated in the test.

4.3.2. Experiment 2: segmentation versus brute force

Per video the test subject had 40 seconds of search time, being allowed infinite attempts within the time limit. The reason that the amount of time was increased is because the second experiment, contrary to the size and layout experiment, had more than one screen per video, as well as more videos in the storyboard each task, thus this also meant that some time was lost due to scrolling. After some informal testing, 40 seconds seemed to be a reasonable amount for a

search task. The penalty for more than two mistakes also applied to this experiment. The KIS tasks also consisted of 5 second videos. The video content consisted of a subset of the videos used in the VBS 2017 competition. Since the amount of video material had to be minimized (see section one), this experiment had approximately the same number of hours. The content existed of live action footage almost exclusively, ranging in different topics. Contrary to the first experiment, all videos of the video archive were relatively short, and the length of each video did not differ a lot from one another. See table 2 for the overview of videos used in the segment vs brute force experiment.

4.3.2.1. Subjects & experiment design

In the segment versus brute force experiment, the number of test persons was 20. The difference in number of test subjects lies with the fact that the segmentation versus brute force experiment had less variables that could influence the results, which made the results by themselves stronger relative to the first experiment, thus requiring less test subjects. Each test person has tested both systems (one half starting with segmented, and the other with brute force).

In the segmented versus brute force experiment, the test subjects also consisted of university students that voluntarily participated in the test.

Title	Type	Genre	Length	Number of thumbnails (1fps/25fps)	Number of tasks
12_Hour_Rugby._-o- _.12_Hour_Rugby_VCD_PA L_512kb.mp4	Live action	Sports	08:44	524/3100	1
12Lvideoshow._-o- _.12LNK3_512kb.mp4	Live action	documentary	07:19	439/7575	1
12raweggs._-o- _.eggshigh_512kb.mp4	Live action	Sketch comedy	06:40	400/2950	1
15JaarBallongezelschap._-o- _.Ballongezelschap25Jaar_w m384x28825Fps512k_512kb. mp4	Live action	Documentary	08:35	515/2075	2
16maggio2005b._-o- _.skytg24_pinternet_1405200 5_512kb.mp4	Live action	News show	07:07	427/625	2
19_11_05_ArmandoGuevara Ochoa._-o- _.ArmandoGuevaraOchoa_51 2kb.mp4	Live action	Documentary	07:53	473/925	2
19_11_2005_Cyberadiccio. -o- _.cyberadiccio_512kb.mp4	Live action	Documentary	06:39	399/3525	2

42ID_End_of_Tour_Music_Video_original._-o- _42ID_End_of_Tour_Music_Video_original_512kb.mp4	Live action	Public affairs	08:43	523/5625	2
75_anniversaire_3_saint_hierarques._-o- _Saint_3_docteurs_512kb.mp4	Live action	Religious	08:55	535/1525	2
01-AWalletADollarAndTheExistenceOfGod1theism._-o- _01-AWalletADollarAndTheExistenceOfGod1theism_512kb.mp4	Slides	Religious	06:30	390/2025	2
1almostPure._-o- _1_almost_Pure_512kb.mp4	Live action	Art Television	06:30	390/3925	1
1PeoplesVideo.tv._-o- _UrbanEssence_60Kbps_512kb.mp4	Live action	Documentary	07:24	444/925	2
1s1cous._-o- _Cours_1s1.3_512kb.mp4	Live action	Documentary	06:39	399/5575	1
1stVideoConferenceInFrench._-o- _1st_french_videoconference_512kb.mp4	Live action	Reality	06:40	340/425	2
002Rockbox._-o- _002_Rockbox_512kb.mp4	Computer stream	Instructional	07:02	422/225	2
2.ManifestacionFrenteAlHotelRitzPorLaPrivatizacionDeLaSanidad._-o- _2_manifestacin_frente_al_hotel_ritz_por_la_privatizacion_de_la_sanidad_publica_en_Madrid._23092008_512kb.mp4	Live action	News show	06:29	389/1125	1
3_sa_t1_13h_6-Block._-o- _3_sa_t1_13h_6-Block_lo_512kb.mp4	Live action	News show	07:20	440/650	1
3rdOct2008salaats._-o- _MakkahIsha3rdOct08ledbySHeikhGhamdi_512kb.mp4	Live action	Documentary	07:08	428/675	1
3rdPartyInterview._-o- _3rdParty_512kb.mp4	Live action	Documentary	07:03	423/3300	1
4.ManifestacionFrenteAlHotelRitzPorLaPrivatizacionDeLaSanidadPublica._-o- _4_manifestacin_frente_al	Live action	News show	06:42	402/2250	1

hotel_ritz_por_la_privatizaci n_de_la_sanidad_pblica_en_ Madrid._23092008_512kb.m p4					
005OrbitDownloader._-o- _.005_Orbit_Downloader_51 2kb.mp4	Computer stream	Instructional	08:34	514/575	2
5_pashto_video_nasheeds._- o-_.5_512kb.mp4	Live action	Documentary	07:48	468/3825	1
5ica_5._-o-_.5ica_512kb.mp4	Live action	Documentary	07:23	443/2700	1
006Jumpcut._-o- _.006_jumpcut_512kb.mp4	Computer stream	Instructional	09:15	555/425	1
9-11Homemade2001._-o- _.911HomeMade_0001_512k b.mp4	Slides	Documentary	06:48	408/3375	1
9antsEliseBakkertheintegratio nofartandquality._-o- _.relaisrotterdam1_512kb.mp 4	Live action	News show	07:17	437/2050	1
10_abj_rogerio._-o- _.10_rogerio_512kb.mp4	Live action	News show	06:43	403/575	1
11_Ghosts_II_and_18_Ghost s_II.mpg._-o- _.11_Ghosts_II_and_18_Gho sts_II_512kb.mp4	Live action	Art Television	07:24	444/5775	1
11-9-misteri-da-vendere._-o- _.wtc7-misteri2-improvviso- tempi-pull-1.2- 640x480_512kb.mp4	Live action	Documentary	07:04	424/2250	1
12_11_2005_sociedad_de_la _informacion._-o- _.sociedad1_512kb.mp4	Live action	Documentary	07:42	462/1500	2

Table 2: video archive

4.4. Procedure:

4.4.1. Experiment 1: layout and sizes

Upon entrance to the test room, the test subject would receive a form containing

- Demographic data questions
- NASA TLX 2x
- Comparative questionnaire

The experiment started off with the test subjects filling in the demographic data. The questions consisted of age, gender, study subject, and the amount of experience in watching videos on mobile devices and video editing on either mobile devices or PC's (see appendix A, B for details). Then they would get a short introduction about the experiment and its goals (comparing different layouts and sizes) and how the experiment works. After that the actual experimenting took place, each layout starting with a demonstration of one KIS task to make sure they understood how the layouts were set up. The test subjects would get nine known item search tasks (i.e., three per size). Between each task there was a five second period, meant for a quick ease of mind. To finish up the test for one layout the test subject had to fill in the NASA TLX. After repeating the previous steps for the other layout, we finished off with the comparative questionnaire. Because qualitative comparisons of layouts seem more important than size (e.g., for the latter we expect differences between performance and user preferences, e.g., users will likely prefer bigger ones despite high performance with small ones) we only focused on that. The total time is estimated at roughly 20 minutes (see table 3).

Time(minute)	Activity
1.5	Gather demographic data
1.5	Short introduction into goals of the experiment (compare different layouts)
2	Demo of the layouts
±8.3	24 tasks*(5 sec countdown + 5 second target clip + 10 sec max search time)
3	fill in NASA TLX (first time might take longer, so on average a minute extra)
3	Comparative questionnaire
Total	±19.3 minutes

Table 3: procedure overview and time estimation per test subject layout and size experiment

Because the first experiment had two variables to evaluate, the number of tasks was split in such a way that each combination had a sufficient number of tasks, whereas the second experiment only had one variable to evaluate; using 10 tasks per system in the second experiment yielded a strong comparison between the segmented and the brute force approach. Other than the difference in task distribution and shift in focus of the test (i.e., different variables), the procedures were virtually the same.

4.4.1.1. Material & environment:

For the first experiment, we have used an 8-inch tablet. Due to focus on mobile, 8 inch seems a good choice for video watching and browsing (not too small, but also not too big) and size wise this is comparable to the one used in the VBS.

Some details on the tablet:

- Sony Xperia z3 tablet compact
- Type: TFT LCD capacitive touchscreen, 16M colors
- Size: 8.0 inches (~70.4% screen-to-body ratio)
- Resolution: 1200 x 1920 pixels (~283 ppi pixel density)

Although the distance between the user and the screen is an important issue, we did not expect much variation among different tasks done by a single user due to the 2 practice tasks. Informal observations across users would be noted and tracked though whenever possible. The subjects would be seated at a table in a “neutral room”, comfortable and quiet, without distraction, and optionally something to lean the tablet against (see figure 5a).

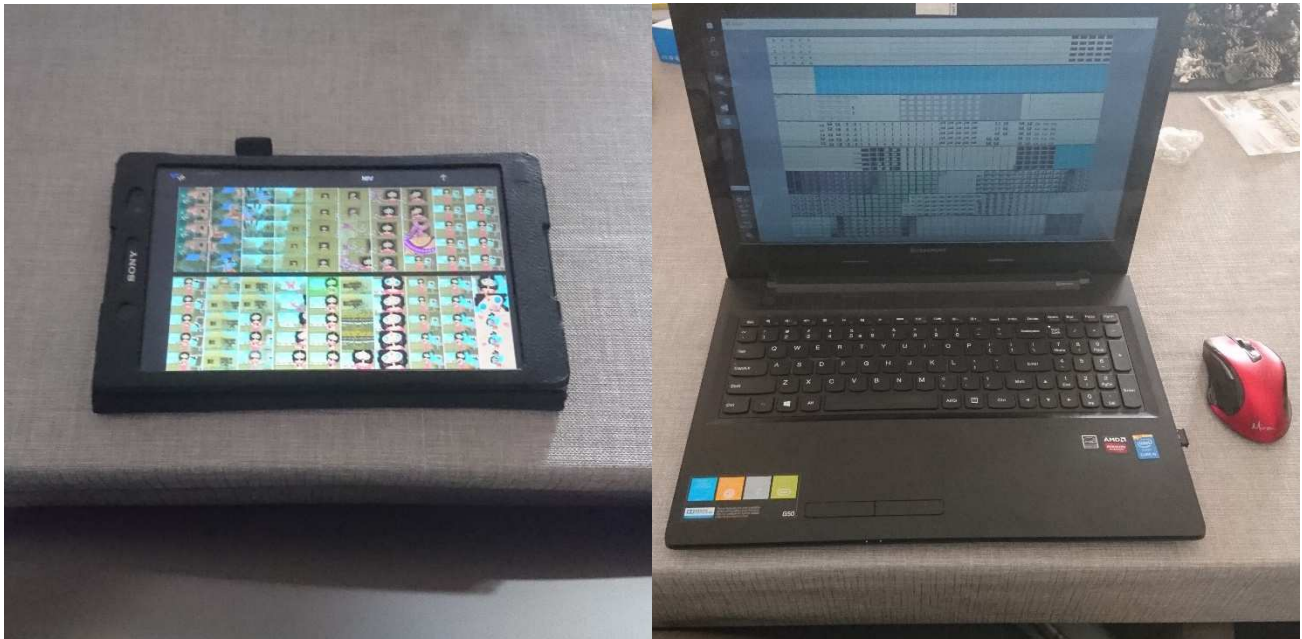


Figure 5: a) size and layout b) segmented versus brute force

4.4.2. Experiment 2: segmentation versus brute force

The procedure for the second experiment was fairly similar to the first. Upon entrance to the test room, the test subject would receive a form containing

- Demographic data questions
- NASA TLX 2x
- Comparative questionnaire

The experiment again started off with the test subjects filling in the demographic data, which was similar, with the only difference being the experience that was asked for was more aimed at pc usage in combination with videos. Then the test subjects would get the introduction about the experiment, explaining the difference between the two versions. After that the actual experimented started, with each version having a demonstration of one KIS to make sure they understood how the layouts were set up. Then the test person would get 10 known item search tasks. Each task again had a 5 second period afterwards, meant for a quick ease of mind. To finish up the test for one version the test subject had to fill in the NASA TLX (see appendix D), which was followed by the repetition of the previous steps regarding the actual experiment for the other version (the versions being either the **segmented** or the **brute force** approach). The second experiment was also concluded with a comparative questionnaire. The total time is estimated at roughly 28 minutes (see table 4).

Time(minute)	Activity
1.5	Gather demographic data
1.5	Short introduction into goals of the experiment (compare different layouts)
2	Demo of the layouts
±16.6	20 tasks*(5 sec countdown + 5 second target clip + 40 sec max search time)
3	fill in NASA TLX (first time might take longer, so on average a minute extra)
3	Comparative questionnaire
Total	±27.6 minutes

Table 4: procedure overview and time estimation per test subject for segmentation versus brute force experiment

4.4.2.1. Material & environment:

For the second experiment, we have used a laptop. As stated in section 4.2.1 the reason for choosing a laptop in this case was to keep the screen static (i.e., minimizing movement during testing) and having slightly more room for displaying the thumbnails.

The details of the laptop:

- Lenovo IdeaPad G50-70
- CPU: Intel Core i5 4210U
- GPU: AMD Radeon R5 M230
- Screen size: 15,6 inch
- Resolution: 1366x768

The subjects would again be seated at a table in a “neutral room”, comfortable and quiet, without distraction (see figure 5b).

5. Analysis and discussion of results

In both experiments for each of the research questions regarding performance, the most important criterion is the VBS score. The VBS score per task has been defined as [15]:

$$s_i^k = \frac{100 - 50 \frac{t}{T_{max}}}{p_i^k}$$

$$\text{where } p_i^k = \begin{cases} 1, & \text{if } m_i^k \leq 1 \\ m_i^k - 1, & \text{otherwise} \end{cases}$$

And the total VBS score is calculated as:

$$s^k = \sum_{i=1}^n s_i^k$$

Here t and T_{max} represent the time needed to complete the task, and the time limit per task. p_i^k represents the penalty given for the mistakes made per task, with m_i^k being the number of mistakes made during the corresponding task. n is the total number of tasks done per setup (i.e., a combination of layout and size in the *size and layout experiment*, or either the segmentation or the brute force approach in the *segmentation versus brute force experiment*).

The analysis will build up to this by discussing the individual components of this score before considering the VBS score. For significance test we use $\alpha = 0.05$ as threshold. Because the time has a limit of 10 and 40 seconds respectively, this means that in practice not all task will be completed, thus the time cannot be averaged properly and will therefore not have significance testing performed on it. However, because it might offer some contextual information towards the performance, we have also taken time into consideration. After the analysis of the performance the user experience will be discussed, providing context on the performance. Due to the qualitative nature of this data there will not be any statistical significance testing on it either. After going through the results, the effects of them will be interpreted and discussed.

5.1. Experiment 1: layout and size

The results show a clear difference between both layout and size in terms of performance. This is supported through statistical analysis which will be discussed in this section. The analysis will be split into three sections to distinguish between the three performance research questions. The sizes of the thumbnails are referred to as size 2, 3 and 5 with the numbers representing the number of groups of 25 thumbnails used for generation of the storyboard per row; 2 being the largest thumbnails and 5 the smallest.

5.1.1. Performance

RQ1. What is the optimum thumbnail size (with respect to performance indicators specified in section 3)?

The thumbnail sizes show to have a major influence on the performance. Despite the fact that image recognition is possible even at really small thumbnail sizes [16], the test has shown that a larger size generally yields higher scores. The number of correctly completed tasks per size/layout combination went up from small to larger images (1.70, 1.95, 2.64 on average from small to large). Significance testing using ANOVA single factor yielded a significant difference ($F_{\text{stat}} = 37.41$, $F_{\text{crit}} = 3.03$, $P = 6.20E-15$).

The number of mistakes per size is generally low, as on average the number of mistakes was only slightly below 1 (1.05, 1.01, 0.18 from small to large). Due to violation of the normality assumption of the ANOVA test, the significance test used for this test is the Kruskal-Wallis test. This also yielded a significant difference between the sizes ($K = 42.43$, critical value = 5.99, $P = 6.11E-10$).

Figure 6a indicates the averages amount of correct and mistakes per size category, as well as the average over the entire population. It indicates the highest improvement at size 2, hence the average column resembles size 3 and 5 most.

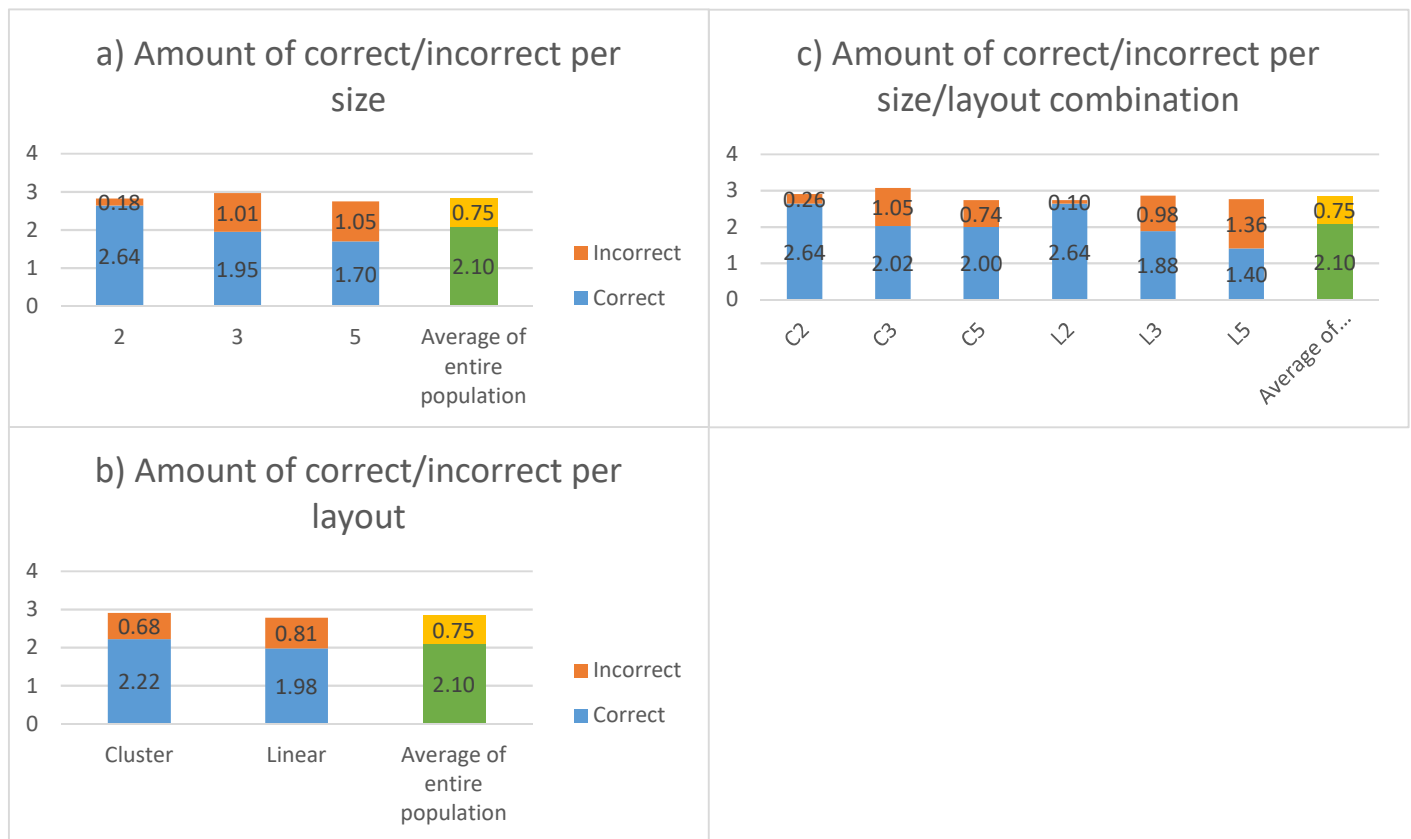


Figure 6: correct/incorrect per a) size b) layout c) size/layout combination. The components for the VBS score (time omitted due to the time-out factor)

The average time for the correctly performed tasks seems to be consistent with the number of correctly completed tasks, as well as the number of mistakes for each size (5.54, 5.12, 4.03 from small to large)

With its components being consistent in the scores rankings it is not surprising that the VBS scores yield the same order of sizes in score. With a limit of 300 the VBS scores were not bad on average (187.60, 196.45, 228.58 from small to large thumbnails). The significance test (ANOVA: single factor; $F_{\text{stat}} = 73.08$, $F_{\text{crit}} = 3.03$, $P = 1.07E-25$) confirmed a significant difference between the different sizes. The data thus seems to imply that the larger sizes are preferred in terms of performance, which also becomes visible upon looking at figure 7a with size 3 being close to the average, and size 2 noticeably higher than the average score.

RQ2. What is the *optimum thumbnail arrangement* (with respect to performance indicators specified in section 3)?

The difference in layout seems to have less of an effect on the search tasks; the results are closer to each other. To begin with the number of correct answers: with an average of 2.22 for the cluster layout and 1.98 for the Linear layout there is not a whole lot of difference. However statistical analysis did indicate a significant difference (paired t-test: $t_{\text{stat}} = 2.80$, $t_{\text{crit}} = 1.66$, $P = 0.0059$; $\text{variance}_{\text{clus}}: 0.542222222$, $\text{variance}_{\text{lin}}: 0.807428571$).

The number of mistakes was even closer to each other. The results from the tests were 0.68 and 0.81 for the cluster and linear layout respectively. Using Kruskal Wallis, due to skewness, significance testing yielded $K = 0.14$, Critical value = 3.84 and $P = 0.71$, so a statistical significant difference was not found at $\alpha = 0.05$. Figure 6b confirms the small difference between the correctly completed tasks and the number of mistakes.

Time wise the data speaks in favor of the cluster layout (cluster 4.70 and linear 5.06 seconds on average), though the evidence is not very strong, and as stated earlier, significance testing is not possible since uncompleted tasks cannot be averaged in, and only counting the completed tasks would yield an incomplete representation of either layout.

The VBS score had a small difference per layout, 208.71 and 199.71 for the cluster and linear layout respectively on average (see figure 7b), which was in line with expectation considering its components. Significance testing however has shown that the difference in score between the two layouts is significant (Paired t-test: $t_{\text{stat}} = 3.56$, $t_{\text{crit}} = 1.98$, $P = 0.00052$).

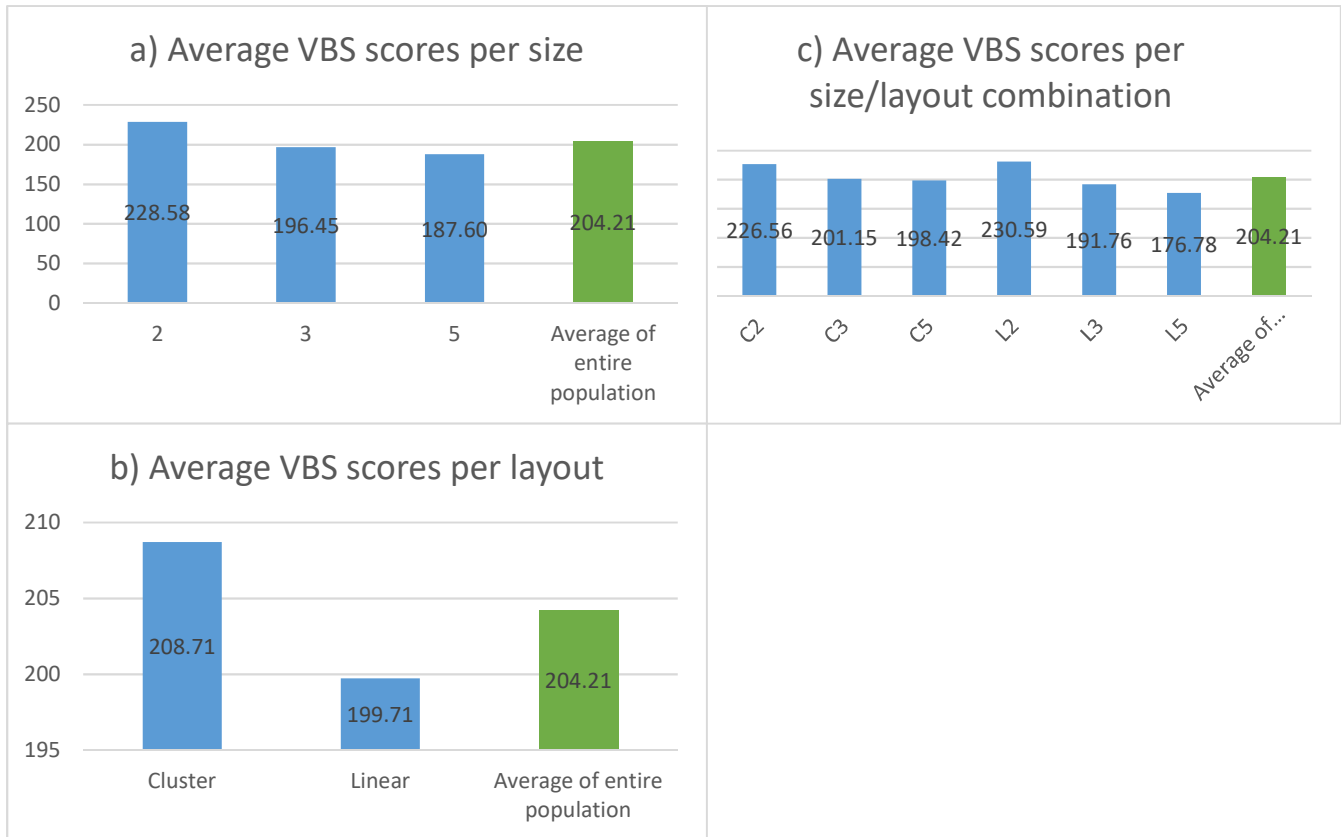


Figure 7: Average VBS scores per a) size b) layout c) size/layout combination. How were the scores per different categories?

RQ3. What is the *mutual dependency of thumbnail size and arrangement (with respect to performance)*?

Looking at the combinations of layouts and sizes, the influences of both variables are noticeable. The amount of correctly performed tasks (ANOVA: single factor $F_{\text{stat}} = 18.82$, $F_{\text{crit}} = 2.25$, $P = 7.79\text{E-}16$) and the number of mistakes (Kruskal-Wallis $K = 48.83$, Critical value = 5.99, $P = 2.49\text{E-}11$) both showed significant differences (for all the averages see table 5). Time wise from slowest to fastest the order for the combinations is mostly consistent with RQ1 and RQ2 except for L2, which scored better than the expected C2. Both the amount of errors and the VBS scores were distributed in the same way, which seemingly forms a pattern. As with RQ1 and RQ2 the significance test showed that for the VBS not all populations are distributed the same way (ANOVA: single factor $F_{\text{stat}} = 36.64$, $F_{\text{crit}} = 2.25$, $P = 5.5\text{E-}28$). See figure 6c and 7c for the averages of the number of correct and mistakes, and the averages on the VBS score.

The next sub section will discuss the ranking of the size/layout combinations, and how it relates to layout and the size individually (i.e., RQ1 and RQ2).

	Correct	Mistakes	Time correct	VBS
C2	2.64285714	0.26190476	4.177236988	226.5639
C3	2.02380952	1.04761905	4.773614134	201.1479
C5	2	0.73809524	5.199617248	198.4213
L2	2.64285714	0.0952381	3.883739635	230.5895
L3	1.88095238	0.97619048	5.502435675	191.7601
L5	1.4047619	1.35714286	5.887049774	176.7834
average over all	2.09920635	0.74603175	N/A	35.05945

Table 5: averages over components for each combination and the VBS score averages

5.1.2. Discussion performance

The results related to the research questions 1 and 2 both indicated a significant increase as the larger sizes and the cluster layout yielded better scores. For research question 3 there are three possible scenarios regarding the influence of the independent variables on the performance: advantages of variables increase total performance, they affect each other differently (e.g. subtraction), or no coherent effect. For the most part the data seems to be consistent with the first scenario. But in the case of the layout, the number of mistakes did not yield a significant difference. This is visible in figure 6c which indicates that for the lower sizes the number of mistakes is lower at the cluster layout compared to their counterpart, but with the larger sizes this is the other way around. The strong influence of the size is well visible in the number of correct tasks for each size/layout combination. In other words, the influence of the size is larger than that of the layout; with the size increasing the score increases, but comparing layout counterparts does not yield a consistent increase or decrease. The P-values in from the statistics confirm this, as with RQ2 the values are significantly larger than in RQ1.

The consequence of this phenomenon is that the VBS score also increases as the size goes up, but the largest linear size has the best score. The middle size however does increase consistently in all components of the VBS-score as well as the score itself. This might suggest that the optimal increase due to the layout sits around that size. So, from the data we can conclude that the largest size works better, independent of layout (i.e., the cluster layout does not provide a performance benefit). However, with smaller sizes, the cluster layout does increase the performance.

5.1.3. User experience

The user experience will be split into the above-mentioned components (workload and perceived performance)

Using the Nasa TLX the following aspects have been measured (see appendix D for the full NASA TLX):

- Mental demand
- Physical demand
- Temporal demand
- Performance
- Effort
- Frustration

The workload used all components except for the performance. The perceived performance consists of the performance aspect of the NASA TLX together with the questionnaire (see appendix B for the questionnaire). The user experience focus lies on the layouts, hence the NASA TLX was filled in after all tasks related to one layout were completed (and then the same for the other). The reason for that was the expected unique contribution of the layouts.

RQ4. How does *thumbnail arrangement* influence *experienced workload*?

All the TLX scales go from positive to negative. Almost all the questions scored in favor of the cluster layout. The effort seems to have most noticeable edge. The only exception in this is the physical demand, which scored about 0.1 lower on average. Table 6 shows the averages of the Nasa TLX questions. All in all, the differences between the two layouts per workload varies between 0.1 and 0.67, which is an improvement of about 1.4% to 9.5%.

		Averages
Mental Demand	Clustered	3.85
	Linear	4.116666667
Physical Demand	Clustered	1.45
	Linear	1.375
Temporal Demand	Clustered	4.541666667
	Linear	4.641666667
Effort	Clustered	3.391666667
	Linear	4.058333333
Frustration	Clustered	2.7
	Linear	3.025

Table 6: Workload results

RQ5. How does *thumbnail arrangement* influence *perceived performance*?

The comparative questionnaire started off with a question about the preference of the layouts and a motivation for their choice. To get a more concrete image about their view on the layouts the test subjects were asked what, according to them, were the advantages of both layouts (see appendix B for the questionnaire).

The most used arguments in favor of the Linear layout was the intuitive reading direction from left to right, top to bottom. After that it was also very often stated that the chronological order was easier to follow; the subtle difference between the two is eye movement versus the intuitiveness helping to make following the story more easily though they are closely related. Most other arguments were along the same line as the above mentioned, like the simplicity or comfort that it offers, as well as it being the standard layout for text reading and in modern video editing software. One of the salient remarks that had been made is that the linear layout works like a stretched-out cluster in long scenes (but horizontally instead of vertically). All in all, the reasons for preference towards a linear layout stems from the familiarity of the reading direction.

For the cluster layout, there are several arguments as to why this layout would be preferred. The most prominent one is the block structure which makes it easier to distinguish scenes. Another often mentioned argument is that the chunks of 5 (i.e., the cluster size) is very convenient to read through; the test subjects indicated that it reads very comfortably as well. What is very remarkable is the fact that a lot of subjects claimed that the cluster layout requires less eye movement, which is also part of the question physical demand. The TLX averages for physical demand contradict this, even though by a very small difference; while the argument itself makes sense, it does not show in the TLX. The perceived performance scored in favor of the clustered layout which is a 10% difference.

The questionnaire was closed off by asking for general comments about the test. A very common comment was that the smallest thumbnail size was too small, and in some cases this has led to a shortage of time according to the test subjects. This seems to be consistent with the performance, and is explainable through the fact that the smallest sizes are slightly smaller than the experiment done in [14]. For other sizes this did not seem to be the case. The difference between this test and the previous one is the fact that every subject has tested each size in the latest test, while in the previous test all test subjects just did one, giving them more time to get used to the size they were testing, and not having to experience the change in sizes. Another thing that is perceived as difficult is a video with a lot of similar scenes, which makes sense for the fact that there are relatively little recognition points to remember, and finding those back can be quite difficult.

Other suggestions were often in the direction of building a higher contrast between thumbnails by either using dynamic thumbnails or drawing boundaries around clusters, which would increase the focus on searching discrete chunks of information.

Weighting out their arguments in favor and against each layout most test subjects did choose in favor of the clustered layout. The amount came down to exactly 28 test subjects preferring the cluster layout and 14 test subjects the linear layout.

5.1.4. General observations thumbnail sizes

Despite not having direct data for comparison on user experience of thumbnail sizes, we did get general comments on the experiment itself, including the sizes. Our observations suggest that the

test subjects generally preferred the middle and the largest size. A common statement that has been made is that the smallest was experienced as too small, although there was also a prominent number of test subjects among those who experienced it as too small that indicated that it was, despite it being too small, not too hard. The size disadvantage did make the processing of the image more intensive, and some of them indicated that the time limit was therefore often too short. Further research is required to verify this data, however because pixel density, resolution, display type etc. influences the visual data that can be processed on a specific thumbnail of a given size, the importance of this factor compared to the layout seems less significant.

5.1.5. Discussion user experience

It appears that the preferred layout for the vast majority is the clustered layout. Despite the relatively small differences in the workload the layout works more conveniently for most subjects in the mentioned aspects. The difference between the two groups (linear preferred vs cluster preferred) seems to be the ability to adapt to the alternative layout, as the vast majority that prefers the linear layout indicates that the reason for their preference is the intuitiveness of the storyboard. This raises the question whether the learning curve might be slightly longer than the time that the users have had to use the system, or rather whether learning to work with such a layout would not be worth it for some users.

Looking at the numbers, exactly $2/3^{\text{rd}}$ of the test subjects have chosen the cluster layout, which might not be an overwhelmingly convincing result, but nevertheless proves that the cluster layout is most preferred. The test only lasted approximately 15 minutes, yielding roughly 7.5 minutes of testing time for one layout, and with the previous test this was roughly 15 minutes per layout, which might be too short. Getting more used to might possibly increase both the performance and the user experience. Further evaluation on this matter would be needed in order to draw a solid conclusion on this. It is however safe to conclude that in terms of user experience the cluster layout has the edge, especially at larger sizes (i.e., size 2 and 3).

When comparing the results of the performance and the user experience we observed one salient detail: despite that improvement in performance using the cluster layout only applied for the smaller sizes, the cluster layout was overall most commonly preferred, which seems to be a contradiction between the qualitative and quantitative factors. It should be noted that the largest sizes had the highest performance, however this does not mean that the larger the size the better since the increase in performance has lowered at the largest tested size, meaning that it is possible that the optimum lies between the middle and the largest size.

5.2. Experiment 2: Segmentation versus brute force

5.2.1. Performance

RQ6. How much does a *storyboard interface design* contribute to the *performance* of a video browsing system that relies on both *human visual inspection* and *filtering* (with respect to the performance indicators specified above)?

Performance wise, the brute force approach seemed to be more successful. The brute force

approach scored statistically significantly higher than the segmentation approach in terms of number of correctly completed tasks. A paired T-test between the number of correct averages of both approaches yielded a statistical significant difference in favor of the brute force approach (paired t-test: $t_{stat} = 5.64$, $t_{crit} = 2.09$, $P = 1.95E-05$). In figure 8a the averages of each approach have been laid out. Although the difference is quite apparent from the chart, the scores by themselves have scored lower than initially expected. Despite the difference number of correctly completed tasks, the speed at which the tasks were completed was lower on average for the segmentation than for the brute force approach. In the discussion, we will review the results and into possible explanations how they came to be.

The number of mistakes however did not show any statistical significant difference. Due to the violation of the assumptions for a paired t-test, a non-parametric test was performed on the data. With the averages for the brute force and segmentation approach being 19.65 and 22.35 respectively (see figure 8b), the H_0 was not rejected, implying that the difference in mistakes cannot be proven to be not by chance with a certainty of 95% (Wilcoxon signed rank test: $W_{stat} = 80.5$, $W_{crit} = 46$, $P = 0.56$).

Comparing the VBS scores the differences on average were not very large either: 449.5 for the brute force and 406.7 for the segmentation approach (see figure 8c). Statistical analysis of the data confirmed that the difference between the two approaches is not statistically significant (paired t-test: $t_{stat} = 1.86$, $t_{crit} = 2.09$, $P = 0.08$).

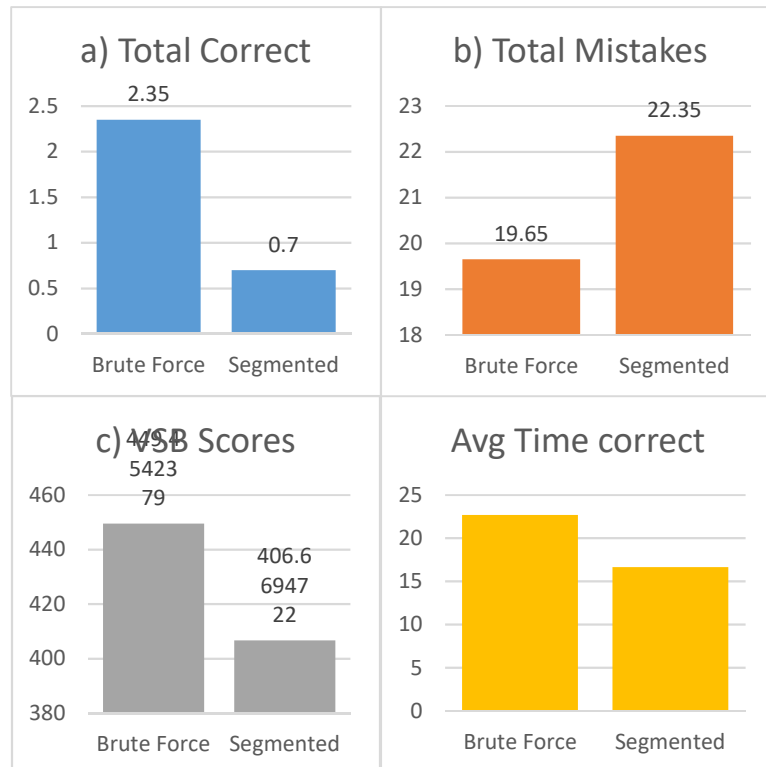


Figure 8: a) total correctly completed tasks
 b) number of mistakes per task
 c) VBS score
 per approach on average

Based directly on the statistics there is an advantage towards the brute force approach in terms of being able to completing a task correctly. However, due to the little difference in mistakes, the actual VBS score, and with that the overall performance advantage is not statistically significantly higher. Since the research question investigates the contribution of the human visual inspection to a system that combines both human visual inspection and filtering, the following must be considered: not having a lower performance suggests a considerable amount of usefulness of the interface design.

5.2.2. Discussion performance

As shown in the previous section, the only component that has shown a statistical difference (in favor of the brute force approach) is the number of correctly performed tasks. However, looking at the VBS score, there are a few things worth noting. Firstly, analyzing the formula for the VBS score: the time is the factor that that gets subtracted from the theoretical max of 100 points. On top of that the score gets divided by the number of mistakes made after the second mistake. Even though the formula was not explained to the test subjects, it was mentioned that a penalty was given after 2 mistakes, therefore it is likely that the participants were hesitant to choose after 2 incorrect attempts. Conserving can be a lot more beneficial unless you are very certain that you are close to the actual video. The fact that the averages are both around 20 might be an indication that the participants have been (subconsciously) aware of this.

Another thing to note, compared to the size and layout experiment, is the content of the video archive. The archive used in this experiment was a subset of the VBS 2017, which mostly contained live action videos. The archive used in the size and layout experiment was more diverse in the sense that it also contained animations, which typically have brighter colors and a higher contrast, while having less different colors overall. The instructional videos that contained computer streams, although technically not animations somewhat similar, did well too generally according to the log, which speaks in favor of this theory.

Lastly it might be worth noting that one of the effects of the shot detection works along the same lines of the cluster layout as it enhances the cluster effect by grouping per shot. Because the cluster layout already has this kind of effect in of itself, it might explain the small difference in performance; the major difference between the two being that the shots are segmented in similar (mostly rectangular) sizes, whereas the brute force approach has different cluster sizes. Some inaccuracies in the shot detection also resulted in shots that were represented longer in the storyboard than necessary.

5.2.3. User experience

RQ7. How much does a storyboard interface design contribute to minimizing experienced workload of a video browsing system that relies on both human visual inspection and filtering (with respect to the user experience indicators specified above)?

The experienced workload measurement speaks in favor of the brute force approach. On all questions that the NASA TLX contains, the brute force scored higher than the segmentation approach. In most cases the difference is not very high, but it is worth noting that both Mental demand and frustration scored more than 10% lower with the brute force system. Table 7 shows the workload averages from the test subjects.

		Averages
Mental Demand	Brute force	5.145
	Segmented	5.8975
Physical Demand	Brute force	2.3975
	Segmented	2.905
Temporal Demand	Brute force	5.145
	Segmented	5.775
Effort	Brute force	5.53
	Segmented	6.055
Frustration	Brute force	4.48
	Segmented	5.39

Table 7: Workload results segmentation versus brute force

RQ8. How much does a *storyboard interface design* contribute to the *perceived performance* of a video browsing system that relies on both *human visual inspection* and *filtering* (with respect to the user experience indicators specified above)?

The questionnaire had the same structure as the size and layout questionnaire, only focusing on the segmentation and the brute force elements (see appendix C). In the questionnaire, the test subjects had an overwhelming preference for the brute force approach. With 17 in favor of the brute force approach and 3 for the segmentation, the difference was very clear. As for the argumentation, the most common reason stated for the brute force was the fact that there were less frames per screen, and in total. The number of thumbnails to process per screen was simply too intensive, according to most test persons (this phenomenon will be discussed in the next section). They also indicated that the similarity of the thumbnails also caused a lot of distraction when searching for a specific frame (set), because subconsciously they still processed a lot of similar frames, knowing it wasn't relevant. Having less thumbnails per screen also lead to easier searching between videos, making it easier to locate the actual video needed.

However, the test subjects did indicate that on the other hand searching within a video was easier using the segmentation approach. The many thumbnails, and particularly how they were sorted, made it easier to find potentially correct shots. It has also been indicated by multiple test subjects that this had the effect of re-assuring that the test person was searching within in the correct video.

All in all, the perceived performance also scored better in favor of the brute force approach, which was also confirmed by the NASA TLX, which had an 11% difference between the two systems for the performance question.

5.2.4. Discussion user experience

Contrary to the performance, the user experience has a straightforward result, though not with an extreme difference, with a favor towards the brute force approach. As stated in the previous section, the main issue between these two systems was the length of the one, with the advantage of having rectangular blocks of similar frames, versus having short videos, making it easier to find the right video. In principle, it would be expected that the shot segmentation database would not be longer than the brute force approach, which suggest a high influence of the video data. The database contained a lot of videos with short shots, which might have caused the brute force representation shorter than the 25 fps shot representation. A possible solution to the problem of having too much thumbnails could be to use a lower fps rate for the shot detection. However, this would have to be done with care, as there should be enough available thumbnails, which means that it is likely that the shots would have to be spread over a smaller area. With a cluster size 5, a 5x5 area seems ideal as the total thumbnails per screen is a manifold of 25; this ensures that a shot will not be broken off at the end of the screen to the next line.

Although the user experience testing pointed out a preference towards the brute force approach, there have also been certain advantages pointed out that the segmentation approach has. One remark that has been made multiple times was that the test subjects expected that a combination of the two systems as they are implemented in the experiment would work conveniently, using a brute force representation to search between videos, and the shot segmentation within a video.

6. Conclusion and future work

6.1. Conclusion

In this thesis, we have performed two experiments, isolating the essential variables from to investigate their effects, and researching the contribution of intelligent interface design based on human visual inspection within a video browsing system that utilizes both human visual inspection and a form of filtering. The first experiment consisted of a series of KIS-tasks that had to be completed under 6 different configurations of size and layout. The data has shown that there is an increase in performance over with the use of the clustered layout over the linear, but that the increase is most likely at its optimum at the middle size. For the effects of different sizes, it seems that the optimum lies higher, but the exact optimum cannot be concluded from this experiment. The cluster layout seems to yield a lower workload, and with a 1:2 preference in favor of the cluster layout, the cluster layout yields the better user experience overall. The qualitative and quantitative results oddly enough do not fully align in the sense that they contradict one another at larger sizes (i.e., no benefit in performance from the cluster layout, yet the preference still goes to that layout). This suggests that there is not an optimum for both the qualitative and quantitative factors, but instead they seem to require different designs in order to satisfy either optimally.

The second experiment compared two systems against each other where one made use of both a shot detection and a storyboard interface and the other only on the storyboard (i.e., human visual inspection). Performance wise the number of correct answers was statistically significantly higher, and the user experience scored in favor of the storyboard.

In this thesis we have investigated the research problem of developing an optimal storyboard design that relies purely on human based search in the form of visual inspection. In order to do so, a clear definition of “optimal” has to be made, as the quantitative factor is fulfilled differently than the qualitative one. Our results suggest that the goal for the system to be built (performance versus user experience) determines the different design choices.

Our second experiment has validated the importance of the optimization of a good interface design. Reaching the ultimate goal of building the best interface will require a good integration with a filter system. A storyboard interface design contributes to this ultimate goal, by offering different design choices which can increase the performance or user experience, while the database can be manipulated or filtered, without a significant loss in usability.

6.2. Future work

The contradictory data has been quite interesting and a flexible system might be able to satisfy both the qualitative and the quantitative factors, or at least come to a good balance. In this regard, it might also be important to set focus on the influence of size towards the user experience. Future research may also include different (lower) frame rates for the shot detection and the storyboard design to investigate whether there is a better configuration for a good integration of both systems. Different filters might also yield a different result regarding the role of the human interface. The lack of animation in the archive and the relatively lower score may be connected, and might be interesting to look into, perhaps to go in the direction experimenting with creating a better contrast within the storyboard without increasing the thumbnails significantly.

7. Reference

- [1] K. Schoeffmann (2014). A user-centric media retrieval competition: the video Browser showdown, *IEEE Multimedia* vol. 21 no. 4, 8-13
- [2] F. Arman, R. Depommier, A. Hsu, and M-Y. Chiu (1994). Content-based Browsing of Video Sequences, *proc. ACM Multimedia 94*, 97-103
- [3] D. Zhong, H. Zhang and S. Chang (1996). Clustering Methods for Video Browsing and Annotation, *proc. SPIE volume 2670, Storage and Retrieval for Still Image and Video Databases IV*, 239-246
- [4] W. Ding, G. Marchionini (1997). A Study on Video Browsing Strategies, *Technical Report. University of Maryland at College Park*
- [5] T. Tse, G. Marchionini, W. Drag, L. Slaughter, A. Komlodi (1998). Dynamic Key Frame Presentation Techniques for Augmenting Video Browsing, *Proc. of the working conference on Advanced visual interfaces*, 185-194
- [6] A. Komlodi, G. Marchionini (1998). Key frame preview techniques for video browsing, *Proc. of the third ACM conference on Digital libraries*, 118-125
- [7] D. Jackson, J. Nicholson, G. Stoekigt, R. Wrobel, A. Thieme, P. Olivier (2013). Panopticon: A Parallel Video Overview System, *Proc. of the 26th annual ACM symposium on User interface software and technology*, 123-130
- [8] L. Herranz (2012). Multiscale Browsing through Video Collections in Smartphones Using Scalable Storyboards, *Proc. of the 2012 IEEE International Conference on Multimedia and Expo Workshops*, 278-283
- [9] W. Hürst, R. van de Werken, and M. Hoet (2015). A Storyboard-Based Interface for Mobile Video Browsing, *MultiMedia Modeling. MMM 2015. Lecture Notes in Computer Science*, vol 8936. Springer, Cham, 261-265
- [10] W. Hürst, R. van de Werken (2015). Human-Based Video Browsing - Investigating Interface Design for Fast Video Browsing, *IEEE ISM 2015*, 363-368
- [11] A. Ip Vai Ching (2016). Size vs layout: utilizing the human visual inspection to optimize video browsing, *Small project, Utrecht University, Utrecht, The Netherlands*
- [12] W. Hürst, A. Ip Vai Ching, M. Hudelist, M. Primus, K. Schoeffmann, C. Beecks (2016). A New Tool for Collaborative Video Search via Content-based Retrieval and Visual Inspection, *Proc. 2016 ACM on Multimedia Conference*, 731-732
- [13] K. Schoeffmann, M. J. Primus, B. Muenzer, S. Petscharnig, C. Karisch, Q. Xu, W. Hürst (2017). "Collaborative Feature Maps for Interactive Video Search." *International Conference on Multimedia Modeling. Springer, Cham*, 457-462
- [14] R. van Werken (2015). Human-based Video Search Finding the optimal mobile storyboard layout for searching, *MSc Thesis, Utrecht University, Utrecht, The Netherlands*
- [15] C. Cobârzan, K. Schoeffmann, W. Bailer, W. Hürst, A. Blažek, J. Lokoč, S. Vrochidis, K.U. Barthel, L. Rossetto (2016). Interactive video search tools: a detailed analysis of the video browser showdown 2015, *Multimedia Tools Appl.*, 1-33
- [16] W. Hürst, C.G.M. Snoek, W.-J. Spoel, M. Tomin (2011). Size Matters! How Thumbnail Number, Size, and Motion Influence Mobile Video Retrieval, *Proceedings of the 17th international conference on Advances in multimedia modeling*, 230-240
- [17] W. Hürst, C.G.M. Snoek, W. Spoel, M. Tomin(2010). Keep moving!: revisiting thumbnails for mobile video retrieval, *Proc. of the ACM Multimedia 2010 International Conference*, 963-966
- [18] <http://www-nlpir.nist.gov/projects/tv2016/tv2016.html#IACC.3>, *Guidelines for TRECVID 2016, IACC.3*

8. Appendix

A: Introduction

VIDEO BROWSING

Participation in this experiment will take place only with the full permission. Withdrawal from participation is possible at all times and the experiment can be terminated at any moment.

DEMOGRAPHICS

General

Subject nr: ...

Name

Age

Gender

M / F

Study subject

Experience

Watching videos on mobile devices

Yes/no

If yes, how often?

Rarely (e.g., a few times per year)

Sometimes (e.g., a few times per month)

Often (e.g., a few times per week)

Very often (e.g., daily)

Video editing on mobile devices or PCs

Yes/no

If yes, how often?

Rarely (e.g., a few times per year)

Sometimes (e.g., a few times per month)

Often (e.g., a few times per week)

Very often (e.g., daily)

Comparative Questionnaire

In, for example, a video player or a video editing application, which one of the two layouts would you prefer?

Linear
Clustered

Why?

Can you name an advantage for the linear layout?

Can you name an advantage for the clustered layout?

DEMOGRAPHICS

script:

General

Subject nr: ...

Name

Age

Gender

M / F

Study subject

Experience

Watching videos on a PCs?

Yes/no

If yes, how often?

Rarely (e.g., a few times per year)

Sometimes (e.g., a few times per month)

Often (e.g., a few times per week)

Very often (e.g., daily)

Video editing on PCs

Yes/no

If yes, how often?

Rarely (e.g., a few times per year)

Sometimes (e.g., a few times per month)

Often (e.g., a few times per week)

Very often (e.g., daily)

Comparative Questionnaire

In, for example, a video player or a video editing application, which one of the two approaches would you prefer?

Brute forced
Segmented

Why?

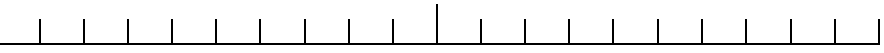
Can you name an advantage for the *brute force* approach?

Can you name an advantage for the *segmented* approach?

NASA Task Load Index


Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

Mental Demand How mentally demanding was the task?



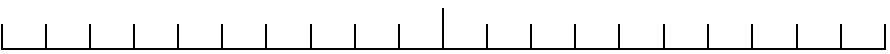
Very Low Very High

Physical Demand How physically demanding was the task?




Very Low Very High

Temporal Demand How hurried or rushed was the pace of the task?




Very Low Very High

Performance How successful were you in accomplishing what you were asked to do?



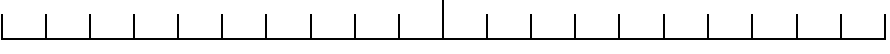
Perfect Failure

Effort How hard did you have to work to accomplish your level of performance?



Very Low Very High

Frustration How insecure, discouraged, irritated, stressed, and annoyed were you?



Very Low Very High