Generating semantic video thumbnails

Master Thesis Information Studies

Jorick van Hees
*
Blue Billywig
jorick.vanhees@student.uva.nl
UvA Student Nr.: 10894020
VU Student Nr.: 2567527

Abstract: Conclusion in abstract

ABSTRACT

We present a novel way to provide a visual preview of a video in the form of video thumbnails. This video thumbnail is an alternative to the static image thumbnail, which is much used in todays interfaces. These thumbnails are often the only visual clue for the user to get a sense of the contents of the video. The video thumbnail is designed to improve the user experience of video navigation structures. We present a method that automatically generates these video thumbnails in an ambiguous domain using clustering techniques on concept features, metadata analysis to generate topics and a comparison to manually selected static thumbnails in order to evaluate the results. The resulting video thumbnails are tested against static thumbnails in a user study where we measure information and engagement in user experience. Based on the significant improvement in information perceived by the user, we consider video summarisation techniques to provide a solid groundwork when generating video thumbnails. The characteristic intent of the video thumbnail emphasises the need for alterations to these techniques in order to improve engagement toward the audience.

CCS Concepts

• **Computing methodologies** → **Video summarization;**
• **Human-centered computing** → *User interface design;* Graphical user interfaces;

Keywords

Video thumbnail generation, interfaces

1. INTRODUCTION

Thumbnail images for videos are used all over the web. They are static representations for videos, and provide a visual preview of the video itself. In combination with a title and description, they form one of the most common interfaces when dealing with a collection of videos. They increase the accuracy when conducting searches in video databases, improve the aesthetics of an overview page, and can increase engagement when using appealing thumbnails.

The video thumbnail is a preview of the full video in the form of a five-second excerpt containing no audio, which conveys the contents of the video in a more expressive manner. In this work, we describe a method that automatically generates these video thumbnails using state-of-the-art techniques like concept feature extraction and topic-dependent metadata analysis. With the addition of manually-selected static thumbnails as training data, we provide a novel way of conveying the purposes of the static thumbnail to the video thumbnail. This way, our approach has the ability to adjust to the specific requirements in engagement and information regarding the end-user, providing drop-in replacements for existing thumbnails.

In this paper, we analyse the effect of video thumbnails on user experience, and discuss the use of video summarisation techniques in the automatic video thumbnail generation. We hypothesise that the use of a video thumbnail instead of a static thumbnail will improve the user experience on two fronts. The user is more engaged to watch the full video, and will be more informed about the contents of the video. Furthermore, we hypothesise that video summarisation techniques will provide solid ground to generate a thumbnail, but require adjustments to better suit the intent of a video thumbnail. In order to test these hypothesis, two research questions are formulated:

- How can we build a system that automatically generates video thumbnails that have a positive effect on the user?
- How do video thumbnails created by the system affect users regarding engagement and information?

Using a real world dataset in an ambiguous domain, the video thumbnail results of our method are challenged in a user survey against their static counterparts, where we test video previews with and without a variant of the thumbnail. These video previews are tested on information and engagement towards the user.

The user study highlight the fact that the video thumbnail can match the performance of a static thumbnail without any problems. In addition, the results also point out

*Blue Billywig, Catharina van Renneslaan 20, 1217 CX Hilversum, The Netherlands

ACM ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

that our approach of using a static thumbnails as training data is very effective in conveying the essence of the video, compared to a manually selected static thumbnail by a professional news editor. Finally, we conclude that our findings pose interesting opportunities in interface design and video analysis.

1.1 Basic concepts & related work

User engagement

Work in the field of video summarisation has a lot of common ground with the generation of video thumbnails. Both use similar data to extract the desired information from a video and its metadata, the same techniques in computer vision is used to process the data and the end result could be very similar.

One could argue that the results from video summarisation could be used in some implementations of the video thumbnails. However, the use cases in both domains are vastly different in terms of user engagement. In general, a summary tries to accurately describe the contents of the video, which eliminates the need to view the full video. In turn, the goal of the thumbnail is to engage the user to view the full video. It tries to show just enough to trigger the user to view the remaining content. Since the implementation scenario and intention of the video thumbnail are different from video summaries, the methods used to generate video summarisation require alterations to meet the specific demands of the video thumbnail.

User engagement.

Event detection.

Existing methods for video event detection based on the TRECVID Media Event Detection Task often use the provided training data. In our approach, no training data is available. However, we can use a number of techniques in order to extract the most characteristic frames from the video. Clustering frame vectors is a technique that is used by many in order to find related parts of the video. Concept features are extracted from frames with a standard frame interval, which are then clustered using unsupervised algorithms like K-means or a hierarchical variant of K-means. These clusters can be used in video summarisations to show a variety of video contents.

Video navigation.

There are a lot of interfaces designed to assist humans in navigating a video library. A summary by Schoeffmann et al. [1] describes over 40 different interfaces that use different techniques to allow convenient browsing through a collection of videos. These range from displaying a key frame best describing the video, or a collection of keyframes with their sizes related to the importance of the frame. Most of these interfaces feature a system that automatically determines what to display on the screen, taken into consideration the different conditions in which the system is used. A study by Hürst et al. [2] describes a user study to the recognition of video using different thumbnail sizes, numbers and various movement in the thumbnails. The study shows that users are able to handle multiple small thumbnails on mobile devices, especially when the thumbnails included motion. Since one of the goals of the video thumbnail is to improve the navigation of users in a news media website, video navigation literature could especially prove useful in the design and implementation of video thumbnails in overview pages.

Static thumbnail generation.

The issue of video navigation using thumbnails has been an active topic of research. A 2015 study by Kim et al. [3] describes a method of automatically combining video frames to generate a thumbnail containing more information that a single frame. Another study describes thumbnail candidate selection using image quality evaluation [4]. A combination of internal and external analysis of the video content to select thumbnails is used by a study by Liu et al. [5]. The techniques and analysation methods that are featured in these methods can possibly be of use when evaluating the generated video thumbnails.

Many approaches to generating thumbnails use a ranking of different frames to propose a suitable thumbnail [6, 7, 8]. In static thumbnail generation, this ranking can be used to select the best thumbnail. In video thumbnail generation (or other video navigation interfaces) the ranking can be used to create a composition of the video. This creates opportunities to generate video thumbnails that consist of different shots from the original video.

2. GENERATING VIDEO THUMBNAILS

The main body of this paper consists of the design of a system that is able to generate video thumbnails, based on textual metadata, an editor-selected static thumbnails and the video itself. There are a number of requirements set for system:

- The generated thumbnail only contains a single video segment extracted from the video, not a compilation of video segments.
- Contents of the video thumbnail should be appropriate for the intention of a thumbnail.
- The system should be able to generate a video thumbnails for an ambiguous domain.

The system can be divided into five major stages: Feature extraction, candidate selection, candidate evaluation and topic-based model training. In the first stage, a number of moments in the video are selected and labeled as candidates. In the second stage, candidates are labeled with a positive or negative value based on their similarity to the editor-selected thumbnail. In the third stage, all videos in the dataset are clustered based on topics derived from their metadata. In the fourth and final stage, an SVM is trained for each cluster using the labeled candidates from each video in that cluster.

2.1 section: dataset requirements

In order to generate a meaningful video thumbnail, training examples in the form of static thumbnails are required in the dataset to train models that can evaluate and rank video thumbnail candidates. We use human-selected images that are used as a thumbnail for each video, since they contain the concepts that represent the video best from a human perspective. Another advantage of using the static thumbnail as evaluation data in the system, is that it has the same purpose as the video thumbnail the system generates. Since

our approach should be suitable for all sorts of news videos, the dataset is not limited to a single domain. The required size of the dataset should be around the hundreds of videos, based on earlier research in video summarisation [9, 10, 11] and the fact that we are clustering our samples based on topics (section 2.5).

2.2 section: feature extraction

Features from the video are extracted from individual frames with 1 frame per second using a convolutional neural network, trained on the ImageNet dataset [12] using CaffeNet [13]. The result is a high-level sparse feature representation of the frame with 1000 dimensions. This representation is used throughout the system as a representation of the frame. This method of extracting features proves to be very effective for event-detection [14, 15, 16] and video summarisation [17, 18].

2.3 section: candidate selection

Instead of evaluating all possible frames in a video using a sliding window, a number of candidates are selected from the video. The selection algorithm is inspired by the bag-of-fragments as described in [19]. Work in video summarisation and event detection use similar clustering techniques in order to establish a semantic structure in the video [20, 21]. This semantic structure can be utilised to discover the most representing window for every cluster, which can be used as candidates.

A comprehensive list of video thumbnail candidates is generated using a sliding window of 5 seconds with 1 fps. The video frames in the window are max-pooled into a vector representation that represents every concept in the video thumbnail. In order to prune the list of candidates, the video thumbnail is compared to a bag-of-fragments representation of the full video. The bags of fragments are created with the features extracted from all frames, clustered using K-means. The number K is calculated using the length of the video $K = \frac{frames}{20 * fps}$. The similarity between the video thumbnail and each BoF is measured using cosine similarity. The resulting similarity vector (with K dimensions) is then normalised, to prioritise the most representing video thumbnail for each BoF. Finally, a ranking is made for each BoF where the top result in each ranking is selected as thumbnail candidate.

Include visual explaining the clustering of similar frames using real example

2.4 section: candidate evaluation

The resulting list of candidates is ranked based on the human-selected static thumbnail for that specific video. The same method of extracting features from the video is used on the static thumbnail, which results in a concept vector that can be compared to concept vectors extracted from the video frames. For every candidate, the cosine similarity between the candidate vector representation (as described in section 2.3) and the thumbnail concept vector is calculated.

Since the resulting similarity values are very diverse across the dataset, a top N % percentile selection is used to determine positive $(+1)$ and negative (-1) values which can later be used to train the SVM models. As shown in figure 3, the similarity values between candidates and the thumbnail are highly irregular across different videos. A threshold based on a percentile is therefor more suited to create a balanced

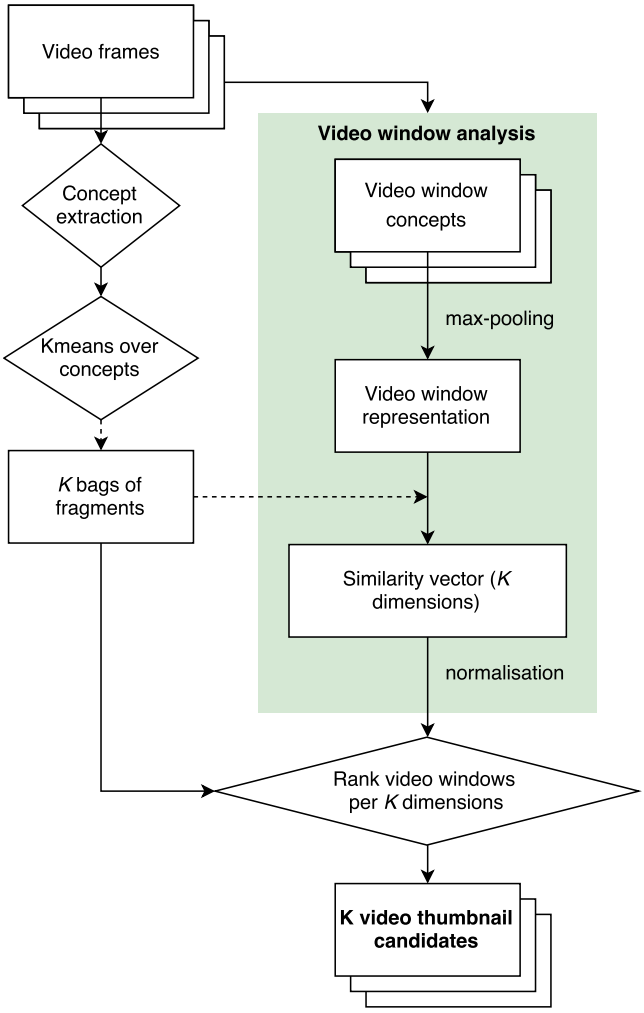


Figure 1: A schematic overview of the candidate selection.

training set that includes data from all videos. The N value is selected at 20, since it provides a balanced distribution of positive and negative examples used as training data.

2.5 Topic clustering using metadata

As described in section 2.1, the videos in the dataset are not limited to a single domain and manually selected thumbnails vary in contents throughout the whole dataset. In order to improve the accuracy of the system and provide a better prediction of positive video thumbnails, the dataset is clustered into specific topics using the metadata available.

Since specific topic categorisation is not available and the tags available in the metadata are not discriminative, a textual analysis of all metadata is used. The title, description and tags are concatenated and a stoplist is used to remove regular words. A bag-of-words corpus is formed with the resulting documents, which is used to create a latent Dirichlet allocation model with a number of topics $T = 25$. This value for T is chosen because our domain only includes news related documents, while a value of 100 is reasonable for the whole Wikipedia dataset [22, 23].

With the LDA model, a sparse vector representation of the video metadata is generated which can be used to cluster

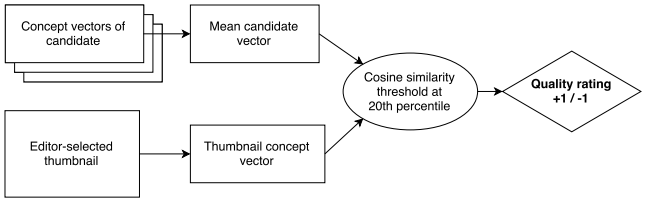


Figure 2: Evaluation the generated candidates using thumbnails selected by an editor.

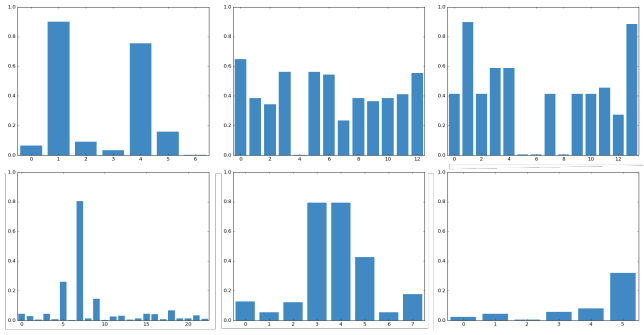


Figure 3: Visualisation of the cosine similarity between the editor selected thumbnail and candidates for multiple videos.

the videos in the dataset. Clustering is done using K-means with M clusters, where M is selected based on the graph in figure (insert figure). A larger M would mean more specific topics and more accurate SVM predictions, but will also reduce the number of samples for each topic. We found that the ideal value of M would be around 10, which results in a good balance between the number of samples per topic and inertia of the clusters.

insert figure about SVM accuracy versus number of K

2.6 Model training

For each topic created in the topic clustering, an SVM is trained in order to rank new video thumbnail candidates. The data used for training an SVM consists of all the videos that are classified in that specific cluster (as described in section 2.5), along with the labels generated in section 2.4. In order to avoid overfitting the data, we use the same parameters for each SVM. An RBF kernel is used with $C = 10$ and $\gamma = 100$ on an average dataset of X training samples. The average accuracy of all SVMs was around 0.75.

2.7 Predicting new videos

New videos with related textual metadata can be processed by the system in order to generate a video thumbnail. First, the frames are analysed with 1 fps to concept vectors as described in section 2.2. These frames are then used to create a list of candidates as described in section 2.3. The textual metadata is converted to a bag-of-words, which is then converted to an LDA vector with the model generated in 2.5. This vector can then be used to decide on the model to use. The concept vector representation of the video thumbnail candidates are then applied to the model (trained in section 2.6), which classifies the vectors. The final ranking is based on the scores that are associated with the classification.

3. EXPERIMENTAL SETUP

The TRECVID Multimedia Event Detection track provides a number of datasets that can be used for event-detection training tasks, which are often used in video summarisation [17, 10, 11]. Other previous techniques in video highlighting and summarisation use data gathered from online platforms as YouTube and Facebook [24, 9].

The dataset that is used in our experimental setup is retrieved from the Irish online publisher 'Landmark Digital'. The websites that publishes the videos reports news articles on a broad number of domains like world news, sports, business, life and local news. The videos are often published alongside an article that could be categorised in any domain. The dataset, however, does not contain any references to the related articles or domains.

The videos in the dataset are accompanied by metadata in the form of a title, description and an unspecified number of (free-form) tags. This (editor created) metadata is primarily used for search engine optimisation and does not contain any structure other than the three values specified. Since these values are created by a professional news editor, we can assume that the data in these fields is a good representation of the content.

As stated in section 2.1, the system requires thumbnails associated with the videos in the dataset to serve as positive examples for training purposes. All videos in our dataset contain a thumbnail, of which only a portion of these thumbnails is manually selected. The remaining thumbnails are automatically generated and cannot serve as reliable training examples, since an editor hasn't made a conscious decision about the contents in the thumbnail. This means that only a portion of our dataset can be used in the training stage of our system. These statistics can be viewed in table 1.

Table 1: Overview of the dataset used in the experimental setup.

Total clips	2162
With title	2162
With description	1912
With tags	1690
With editor selected thumbnail	956

4. SECTION: USER STUDY

The video thumbnails generated by the system described in section 2 have been tested in an A/B user survey against a baseline in the form of static thumbnails: 50% of the respondents received the survey which included video thumbnails, while the other 50% received a version with static thumbnails. The survey was conducted via a custom built website to ensure compatibility across multiple devices.

The user study is conducted to answer the following questions about the use of thumbnails in video previews:

- **Q1.1** What is the effect of a thumbnail on the information received by the user?
- **Q1.2** What is the effect of a thumbnail on the engagement of the user?
- **Q2.1** What is the effect of a video thumbnail compared

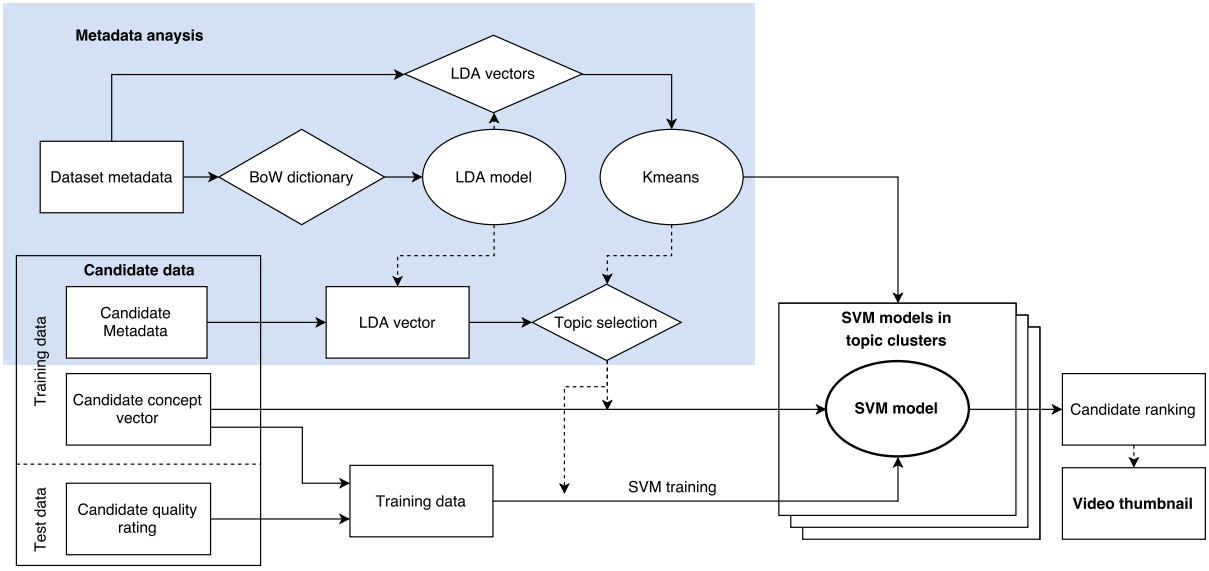


Figure 4: Clustering videos in the dataset based on metadata using KMeans clustering on LDA topic vectors.

to a static thumbnail on the information received by the user?

- **Q2.2** What is the effect of a video thumbnail compared to a static thumbnail on the engagement of the user?

With the answers to Q1.1 and Q1.2, we can confirm that our survey aligns with our rationale based on related work on thumbnails described in 1.1. We then can evaluate the answers to Q2.1 and Q2.2 and depict a conclusion on the effect of a video thumbnail in comparison to a static thumbnail.

4.1 Hypothesis

The following hypotheses can be formulated for the questions stated in section 4:

- **H1.1** The use of a thumbnail in a video preview increases the information received by the user.

Based on the related work in section 1.1, a visual preview increases the accuracy of the user finding relevant content. Thus, we expect that the addition of the only visual element in our survey (either a video thumbnail or static thumbnail) has a positive effect on the information that a user receives from the thumbnail.

- **H1.2** The use of a thumbnail in a video preview increases the engagement of a user.

Research on interfaces and websites for online news shows a high influence of images on the engagement of users, as described in section 1.1. We expect that a similar influence is visible in our experiment by the introduction of any thumbnail.

- **H2.1** The use of a video thumbnail increases the information received by the user compared to the use of a static thumbnail.

The addition of moving images to the thumbnail increases the amount of raw information displayed in the preview, so one could argue that the amount of information conveyed to the user is also increased. Related work discussed in section 1.1 also highlights an improvement of user accuracy in search results when comparing video summaries to a series of static video frames.

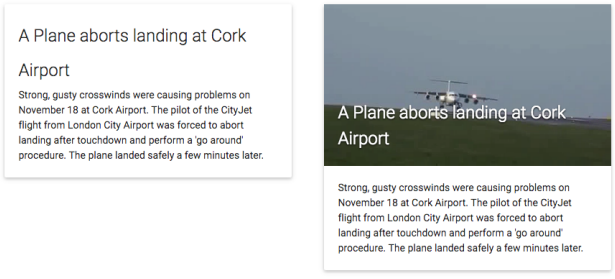
- **H2.2** The use of a video thumbnail increases the engagement of a user compared to the use of a static thumbnail.

Based on the increased attention and interest towards certain items caused by visual elements as described in 1.1, we hypothesise that the addition of motion in the thumbnail increases the interest of a user even further.

4.2 Survey setup

For each video, two previews were shown in successive order: The first preview contained only a title and description, while the second preview contained a title, description and a (static or video) thumbnail. An example of the preview with thumbnail is shown in figure 5.

Figure 5: An example of a video preview. A version without thumbnail is displayed on the left, while a version with thumbnail is displayed on the right.



After each preview, two statements were made about the video preview to measure the engagement of the participant towards the video, and wether the participant felt informed about the contents of the video (we will refer to these statements as *(*context):

- I am interested in viewing the video (engagement).
- I know what to expect from the video (informative).

By comparing the difference in answers between the version with thumbnail and without thumbnail, we are able to measure the impact of using a thumbnail in the preview. This difference can then be compared between the static thumbnail and video thumbnail, allowing us to analyse the effect of a video thumbnail. This way, any preconception from the user about certain topics or videos can be taken into account.

The videos used in this survey were manually picked from the dataset based on number of views. Early feedback on the survey setup revealed that randomly selected videos would be uninteresting and yield no change in response, regardless of the form of the video preview. The difference in target audience between the dataset source and survey respondents would be the primary explanation. The age of most of the videos is a second explanation, since most of the news videos are outdated at the point of conducting the survey. Static thumbnails for the videos were manually picked at the time of publishing, while the video thumbnails were generated by the system described in section 2.

4.3 Responses

A total of 54 respondents participated, of which 27 received the version with video thumbnails, and 26 received the version with static thumbnails. Each participant received previews for a total of three videos, resulting in a total of 137 responses, of which 68 refer to the static thumbnail and 69 refer for the video thumbnail. An overview of these numbers can be found in table 2.

Table 2: Response overview

	Total	**Static**	**Video**
Participants	54	26	27
Preview 1	47	23	24
Preview 2	46	23	23
Preview 3	44	22	22
Total	137	68	69

The difference between the expected number of responses ($54 * 3 = 162$) and the actual number of responses (137) can be explained by the fact that the participants could interrupt the survey at any time. Responses where the participant interrupted the survey between a preview without a thumbnail, and a preview with a thumbnail were ignored.

4.4 Result analysis

The data is gathered from two tests, one with static thumbnails and one with video thumbnails. The research questions are deliberately divided into two groups. In the first group of questions (Q1), we can test our survey setup with established work and further investigate the effect of the video thumbnail (Q2).

4.4.1 Excluding thumbnail versus including thumbnail

Using the Wilcoxon signed-rank test for ordinal, paired data, a highly significant difference is found between previews including thumbnails and excluding thumbnails regarding information and engagement. The P-values are depicted in table 3.

Table 3: Significance of the difference between responses on previews with a thumbnail and without a thumbnail.

	P-value
Information (Q1.1)	1.00×10^{-14}
Engagement (Q1.2)	3.19×10^{-10}

The responses reveal an overal increase in engagement when using thumbnails. These results align with the findings in other work described in 1.1, and allow us to further investigate the differences between the static thumbnail and the video thumbnail in this topic in 4.4.2. The responses are visualised in figure 6.

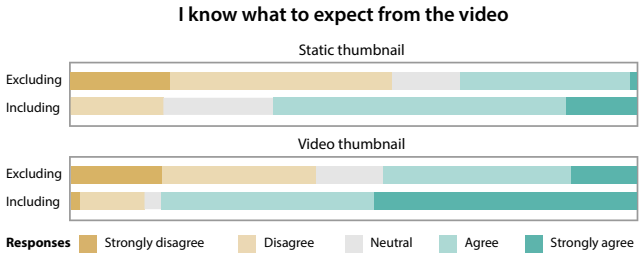


Figure 6: isualisation of the responses regarding **information** on static and video thumbnails, grouped by presence of the thumbnail.

The same Wilcoxon signed-rank test is used for the responses regarding engagement, where a highly significant difference is found. The same patterns that were found for information are visible, an overal increase is found when including thumbnails, though the change is more subtle. The responses are visualised in figure 7.

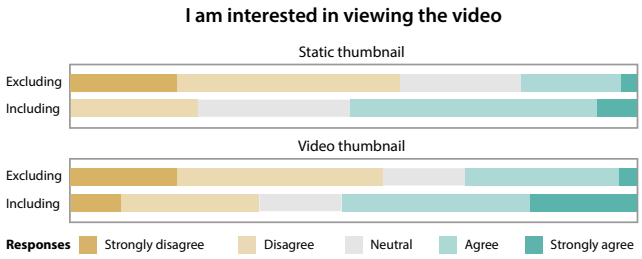


Figure 7: Visualisation of the responses regarding **engagement** on static and video thumbnails, grouped by presence of the thumbnail.

4.4.2 Static thumbnail versus video thumbnail

With the highly significant differences between static thumbnails and video thumbnails found in 4.4.1, we can further

investigate the differences between them. We use a Mann-Whitney U test on the ordinal data from these unpaired groups.

In order to analyse the responses in an objective manner, we need to compare the differences between the static group and video group for previews excluding a thumbnail. Since these previews were exactly the same among these two groups, we hypothesise that there is no significant difference between these groups.

Table 4: Significance of the differences between static thumbnails versus video thumbnails, including and excluding the thumbnail.

P-value	Including	Excluding
Information (Q2.1)	0.0006	0.0414
Engagement (Q2.2)	0.7641	0.3935

The results stated in table 4 tell us a few things. First of all, there is **no significant difference** regarding engagement between the groups, including or excluding the thumbnail. Regarding information, a **highly significant difference** was found when a thumbnail was included, and a **significant difference** was found when the thumbnail was excluded.

While there is a highly significant difference between static thumbnails and video thumbnails regarding information, we cannot directly draw any conclusions, since a significant difference was also found between these groups when the thumbnail was absent. This means that the hypothesis stating that there is no significant difference between the two groups when excluding the thumbnail is rejected, and an in-depth analysis is required to gain any meaningful insights about the effects of the video thumbnail.

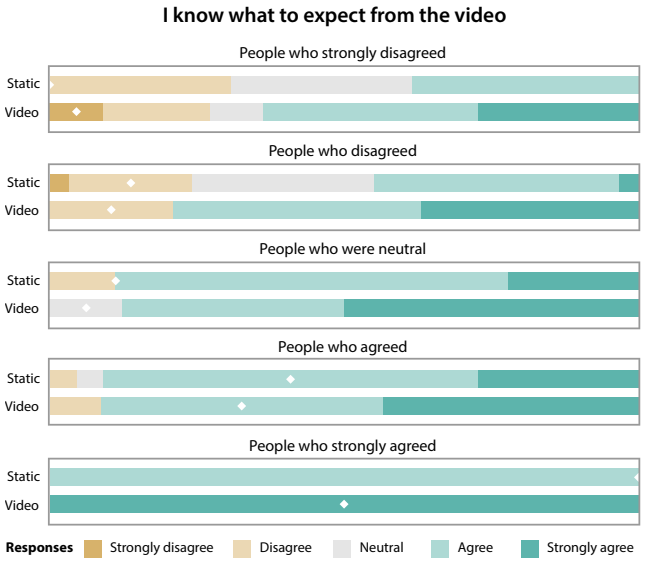
Since the Likert responses are categorised as ordinal data, we cannot calculate the differences or mean values between these groups. Instead, the responses are categorised into five different groups by the response on the preview without a thumbnail. This way, we can analyse the effect of a thumbnail while any initial response on a topic is taken into account. A detailed visualisation of this data is shown in 8, while the uncategorised responses are shown in figure 6 (static thumbnail: including - video thumbnail: including).

In the 'strongly disagree' category, more than half of the respondents shifted to a positive answer when they received a preview with video thumbnail. The same shift is also visible in the 'disagree' category, both categories show a reduction in the amount of negative responses. Looking at the neutral and positive responses, a clear contrast is apparent in the amount of 'neutral' and 'strongly agree' responses. This increase shows that the positive effect of a thumbnail is amplified with the use of a video variant.

In the 'neutral' category, a positive change in response is clearly visible in both static and video thumbnails. Again, we see the amplified positive effect in the 'strongly agree' responses for video thumbnails. The 'agree' category shows a similar trend, although the the overall positive change is less noticeable. The number of responses in the category 'people who strongly agreed' is too limited to draw any conclusions (both thumbnail variants contained only a single response for 'strongly agree'), but is included in the visualisation for the sake of completeness.

Overal we can conclude that the use of a video thumbnail amplifies the positive effect caused by the use of a thumbnail in regard to the information that a user obtains from a video preview. While our responses show a hint of the same effect on the engagement (figure 7, we are unable to draw any conclusions since they are not significantly different. With an in-depth study containing more controlled variables, a larger sample size and a real-world environment, a better understanding of the effect on engagement can be developed.

Figure 8: Effect of static thumbnails versus video thumbnails, regarding information, categorised by response on the preview without a thumbnail. Responses that align with the initial response are marked with a white indicator to highlight the baseline for each category.



5. CONCLUSION & RECOMMENDATIONS

In this paper, we conclude that it is possible to automatically generate video thumbnails that have a positive effect on users compared to static thumbnails which are commonly used today. Based on the results from our user study, we conclude that the implementation of the techniques from the field of video summarisation play a crucial part in the improved information that users experienced. We believe that our method of candidate selection results in a conceptually diverse set candidates, which play an important role in the improvement of information.

User engagement is a highly complex topic where many variables are at play in complex situations. Although our results do not directly show a significant difference between video thumbnails and static thumbnails, we have reasons to believe that the video thumbnail will affect user engagement when an approach is used that better fits the specific intent and purpose of the video thumbnail. Our method of extracting video thumbnails is largely based on techniques used in video summarisation, where the primary concern is to extract information instead of improving user engagement. Using an editor-selected thumbnail as evaluation data is a step in the right direction, and we suggest that alterations like these to video summary techniques will greatly improve the engagement of the video thumbnails.

New interfaces that facilitate video thumbnail selection

for editors will allow us to further analyse the choices made for specific video thumbnails, which could lead to new insights and allow us to adjust our methods of selection and evaluation. The effects of large scale, real-world implementations of video thumbnails could also be utilised to gain new insights in the behaviour of users with this new interface element. Finally, we conclude that the promising effects of video thumbnails should be utilised in the application of previews for videos in order to improve user experience.

References

[1] Klaus Schoeffmann et al. "Video browsing interfaces and applications: a review". In: *Journal of Photonics for Energy* (2010),

[2] Wolfgang Hürst et al. "Size Matters! How Thumbnail Number, Size, and Motion Influence Mobile Video Retrieval". English. In: *Advances in Multimedia Modeling*. Springer Berlin Heidelberg, Jan. 2011, pp. 230–240.

[3] Jongdae Kim et al. "Comprehensible Video Thumbnails". English. In: *Computer Graphics Forum* 34.2 (May 2015), pp. 167–177.

[4] Weigang Zhang et al. "Web video thumbnail recommendation with content-aware analysis and query-sensitive matching". English. In: *Multimedia Tools and Applications* 73.1 (2014), pp. 547–571.

[5] Wu Liu et al. "Multi-Task Deep Visual-Semantic Embedding for Video Thumbnail Selection". In: *Proceedings of the IEEE …* (2015), pp. 3707–3715.

[6] Jiwon Choi and Changick Kim. "A framework for automatic static and dynamic video thumbnail extraction". English. In: *Multimedia Tools and Applications* (2015), pp. 1–17.

[7] Weigang Zhang et al. *A Novel Framework for Web Video Thumbnail Generation*. English. IEEE, 2012.

[8] Yuli Gao, Tong Zhang, and Jun Xiao. "Thematic video thumbnail selection". English. In: *Image Processing (ICIP)* (2009), pp. 4333–4336.

[9] Jurandy Almeida, Neucimar J Leite, and Ricardo da S Torres. "VISON: VIdeo Summarization for ONline applications". English. In: *Pattern Recognition Letters* 33.4 (Mar. 2012), pp. 397–409.

[10] Michael Christel and Neema Moraveji. *Finding the right shots: assessing usability and performance of a digital video library interface*. ACM, Oct. 2004.

[11] Arthur G Money and Harry Agius. "Video summarisation: A conceptual framework and survey of the state of the art". English. In: *Journal of Visual Communication and Image Representation* 19.2 (Feb. 2008), pp. 121–143.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "ImageNet Classification with Deep Convolutional Neural Networks". In: *Advances in neural …* (2012), pp. 1097–1105.

[13] Yangqing Jia et al. *Caffe: Convolutional Architecture for Fast Feature Embedding*. ACM, Nov. 2014.

[14] Amirhossein Habibian, Koen E A van de Sande, and Cees G M Snoek. *Recommendations for video event recognition using concept vocabularies*. ACM, Apr. 2013.

[15] Tim Althoff, Hyun Oh Song, and Trevor Darrell. *Detection bank: an object detection based video representation for multimedia event recognition*. ACM, Oct. 2012.

[16] Lu Jiang, Alexander G Hauptmann, and Guang Xiang. *Leveraging high-level and low-level features for multimedia event detection*. ACM, Oct. 2012.

[17] Muhammad Ajmal et al. "Video Summarization: Techniques and Classification". English. In: *Computer Vision and Graphics*. Springer Berlin Heidelberg, Sept. 2012, pp. 1–13.

[18] Masoud Mazloom et al. "Encoding Concept Prototypes for Video Event Detection and Summarization". In: *ICMR '15: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. Columbia University. ACM, June 2015.

[19] Pascal Mettes et al. "Bag-of-Fragments: Selecting and Encoding Video Fragments for Event Detection and Recounting". In: *ICMR '15: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. University of Amsterdam. ACM, June 2015.

[20] Zheng Yuan et al. *Video summarization with semantic concept preservation*. ACM, Dec. 2011.

[21] Amirhossein Habibian, Thomas Mensink, and Cees G M Snoek. *Composite Concept Discovery for Zero-Shot Video Event Detection*. ACM, Apr. 2014.

[22] David Newman et al. "Distributed Algorithms for Topic Models". In: *Journal of Machine Learning Research* 10.Aug (2009), pp. 1801–1828.

[23] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent Dirichlet Allocation". In: *Journal of Machine Learning Research* 3.Jan (2003), pp. 993–1022.

[24] Huan Yang et al. "Unsupervised Extraction of Video Highlights via Robust Recurrent Auto-Encoders." In: *ICCV* cs.CV (2015), pp. 4633–4641.