ThumbReels: Query-Sensitive Web Video Previews Based on Temporal, Crowdsourced, Semantic Tagging

Barnaby Craggs
HighWire DTC
Lancaster University, UK
b.craggs@lancaster.ac.uk

Myles Kilgallon Scott
HighWire DTC
Lancaster University, UK
m.kilgallonscott@lancaster.ac.uk

Jason Alexander
SCC
Lancaster University, UK
j.alexander@lancaster.ac.uk

ABSTRACT

During online search, the user's expectations often differ from those of the author. This is known as the 'intention gap' and is particularly problematic when searching for and discriminating between online video content. An author uses description and meta-data tags to label their content, but often cannot predict alternate interpretations or appropriations of their work. To address this intention gap, we present ThumbReels, a concept for query-sensitive video previews generated from crowdsourced, temporally defined semantic tagging. Further, we supply an open-source tool that supports on-the-fly temporal tagging of videos, whose output can be used for later search queries. A first user study validates the tool and concept. We then present a second study that shows participants found ThumbReels to better represent search terms than contemporary preview techniques.

Author Keywords

ThumbReels; video; thumbnails; crowdsourcing; video summarisation; video surrogates; metadata

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Thumbnail previews are surrogate objects used by Video-Sharing Websites (VSWs) to provide users with concise representations of content. Video generated over half of all IP traffic in 2012; Cisco predict this will account for 80-90% of global consumer traffic by 2017[1]—underlining the importance of techniques that allow users to discover and discrimintate relevant content.

VSWs, in response to user-defined queries, typically return search results as a populated list of surrogate objects, composed of a thumbnail and author defined metadata such as a title and description. Problems with discovery and discrimination arise when the returned surrogate objects do not accurately represent the content of the video, creating an 'intention gap' [6]; a discrepancy between the information sought by the user and the actual content of the video. For example, of the top five YouTube search results for the term "apple juice" (see Figure 1) only one returns a contextually relevant thumbnail (result 2). Only by watching the video can the user know the actual content; undermining the purpose of the surrogate objects in providing a mechanism for discrimination.

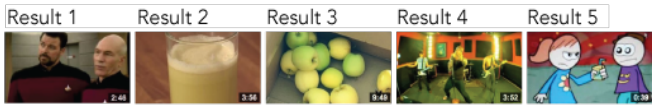

Figure 1. Top five results for the query "apple juice" on `youtube.com`.

To address this problem, we created ThumbReels. The ThumbReel concept involves two steps. First, viewers (as a crowd) temporally tag videos whilst watching them—these tags are stored on the search server. Second, when a search query is issued, the tag data, along with traditional meta-data is used to create a result list. A ThumbReel is then generated for each result. ThumbReel frames are extracted from points on the video timeline where the crowd's tags match the query terms. These are then animated to create a query-sensitive preview of the video.

We validate the ThumbReel concept in two steps. First, we crowdsource the tagging of three videos using our open-source tool. Second, we evaluate the generated ThumbReels against traditional preview mechanisms. This paper therefore contributes: (1) The ThumbReel concept, (2) An open-source tool for crowdsourcing temporal video tags, with accompanying proof of operation, and (3) An evaluation of ThumbReels; highlighting a significant proportion of participants found ThumbReels to better represent search terms in comparison to contemporary preview techniques.

RELATED WORK

The process of producing surrogate objects to represent previews of video content is widely studied. Christel [1] highlights that surrogates "quickly and accurately ... focus attention on the relevant material" and defines three key types of traditional surrogate objects for video content [2]. Poster frames are contemporary thumbnails and filmstrips are animations of multiple poster frames extracted at set intervals from the video. Most VSWs present static thumbnails, however adult websites commonly use filmstrips to preview content. There is no standardised approach to creating such

[1] `http://newsroom.cisco.com/release/1197391/`

CHI'14, April 26 - May 01 2014, Toronto, ON, Canada

ACM 978-1-4503-2473-1/14/04...$15.00.

filmstrips, with variance in the number of extracted frames and the portions of the video from where they are drawn. Skims are genre specific surrogates; a temporal video-edit, closely resembling movie trailers, where key moments are used to promote a movie. Again, skims are not commonplace online.

'Intention Gaps'

When seeking online videos, users engage a search engine with the intention of discovering specific content using query terms rather than explicitly stating their intent [5]. When authors provide video content and meta-data (as titles, descriptions and tags) their intent is to gain an audience. Allowing the author to curate thumbnails returned in a search may present an 'intention gap': a discrepancy between the intended information sought by the user and the perceived intention by the content author [6]. While the observed intention gap may purely be due to the author mistakenly choosing the wrong thumbnail, or indeed that only one thumbnail is available for the whole content, it is well known that the affordance of author curated surrogates can invite deliberate misrepresentation or intentional abuse, such as the meme of "Rick-Rolling."

Previous work has shown that users would discover relevant content more rapidly when thumbnails are dependent on context (query-sensitive) and that these are preferable to thumbnails chosen by the video authors [2, 6]. Despite this, none of the popular VSWs appear to return query-sensitive thumbnails when populating search results, employing algorithmic or author-curated thumbnails instead.

Video Metadata

Video content on VSWs contains both surface and semantic features, the latter including metadata such as the title and description. There are additional semantic metadata schemes for captioning, subtitles and categorisation; but "these elements are not constrained to any framework or ontology, making automated interpretation difficult" [8]. Whilst researchers have applied "considerable effort" into making video content on the Web more accessible, it still remains opaque on websites [8]. The lack of semantic information (in metadata tags) creates problems related to synonymy and polysemy and can reduce the efficiency of content search [7]. Given the use of semantic query terms when searching for content, the need for better metadata about the actual video content is critical to discoverability.

Crowdsourcing Video Tags

Enriching semantic data algorithmically is problematic. Image processing solutions can identify pixel-arrangements to a great degree of accuracy, however they are unable to tell us whether a picture is beautiful or whether an instruction is useful. Crowdsourcing has successfully enriched the semantic metadata of video [8] by asking users to describe the content through tagging. Estellés and Guevara [3] suggest that people will volunteer their support in return for personal esteem or wellbeing.

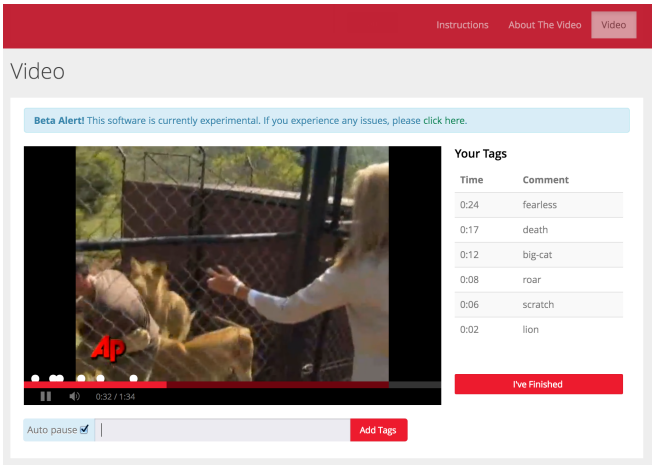

Figure 2. Crowdsourcing video tagging tool.

THUMBREELS

ThumbReels address the 'intention gap' by providing a crowdsourced set of semantic temporal tags for the video content. These tags (as metadata) can be used to create a preview that is tailored to a user's query terms when searching for content. To create a ThumbReel we need be able to both gather tags about a video and to generate a preview.

Crowdsourced Video Tagging Tool

To source tags we created a Web based tool using HTML and Javascript and hosted this as a Web page on publicly accessible servers (see Figure 2). The tool enables a viewer to watch a video and describe what is happening, using keywords. When they start to type, the video is paused and the tag along with timestamp of the current frame is recorded. Throughout the process the viewer can see a list of their previous tags along with respective timestamps; the tags of other viewers are hidden so as to not distract or influence. We have made this tool open-source and is available for download from: `http://thumbreels.com`.

Creating a ThumbReel

ThumbReels consist of frames that are extracted from points on the video timeline where a search term matches previous viewer's tags. These frames are then animated to create a query-sensitive preview of the video. With no previously formalised method for constructing ThumbReels, we derived a specification by analysing and replicating the frame rates, sizes, and inter-frame delays used in filmstrips. Further, observations suggest that users primarily search for videos using semantic query terms [9] and so we only use these in our preview generation. We break ThumbReel creation into two steps: (1) Tag Processing and, (2) Preview Generation.

Tag Processing

The user-generated temporal tags need processing before they are suitable for search and preview generation. The process we followed was: (1) Split compound and comma-separated tags into separate keywords, (2) Card-sort keywords to remove erroneous data, (3) Lemmatise keywords, (4) Remove non-semantically categorised keywords, and (5) Plot a matrix of unique lemma, sorted by frequency, indicating

Video	YouTube ID	First Term	Second Term
V1	fAZpWctgZrY	"Water" ($f = 15$)	"Cannon" ($f = 7$)
V2	kjWBgA81LM	"Lion" ($f = 23$)	"Man" ($f = 10$)
V3	U82Xc0Oedmw	"Apple" ($f = 25$)	"Juice" ($f = 19$)

Table 1. Selected YouTube videos & search terms. (f = tag frequency)

	V1	V2	V3	ALL
No. Participants	28	36	31	95
Sort Correlation	89%	83%	85%	86%
Separated Tags	211	238	455	904
Discards	6	5	4	15
Total Tags	205	233	451	889
Unique Tags	95	109	186	
Uniques as %	46%	47%	41%	

Table 2. Descriptive tagging statistics.

chronological timestamps. This matrix identifies potential frame extraction points and is used to compile a ThumbReel.

Generating a Preview

Tag processing would typically happen 'offline' from the user query, while preview generation would occur 'on-the-fly' so ThumbReels could best match the user's search terms. This process consists of: (1) Extracting a single frame from each point on the video timeline where the query term matches the crowdsourced tags stored in the matrix generated above, (2) Selecting 10 evenly distributed frames from this extracted frameset in chronological order, (3) Compile these frames into an animation using an inter-frame delay of 500 ms (for our user studies we generated animated GIFs).

THUMBREEL EVALUATION

To evaluate the ThumbReel concept, we conducted two user evaluations. The first aimed to assess the effectiveness of gathering temporal, semantic video tags. The second aimed to compare the ThumbReel visualisation against traditional preview techniques. For both evaluations we used three short raw news footage videos. These were selected to require only a small time commitment from participants given the average YouTube video is approximately 5 minutes. The videos we used are listed in Table 1.

Tag Sourcing

Anonymous participants were recruited by invitation using social networks. This attracted 95 participants who were directed to the tool. Participants were offered the ability to tag, using keywords, one or all of three of the videos.

The participants created a total of 904 tags on the three videos. The crowdsourced tags were manually processed independently by the two primary investigators. Interpretation disagreements were negotiated where possible; providing an 86% correlation of the 889 usable tags, after erroneous responses were removed (N = 15). Those tags not agreed upon (N = 124) were used as provided by participants. A summary of the collected tags is presented in Table 2. Lemmatisation was performed using NorthWestern University's MorphAdorner[2]. A meta-analysis of the was tags was conducted to leave only those categorised as semantic (N = 724).

[2] http://morphadorner.northwestern.edu/

Overall, the crowd-sourced gathering of temporal tags was deemed successful. Participants created a large number of tags which (as discussed below) were suitable to generate ThumbReels. While real-world deployment is necessary to understand true crowd dynamics, this study shows promise for this within-video tagging technique.

Comparing Preview Representations

The second evaluation compared ThumbReels against traditional previews to find which better addressed the intention gap by more closely representing query search terms. To do this, we ran an online user evaluation, again with anonymous, self-selecting respondents.

Video Surrogates

For each of the test videos, we selected a pair of tags from the top ten most frequent. This pair: (1) Represented a reasonable set of query terms when searching for the test video, and (2) Yielded search results that did not return the test video on YouTube. The chosen terms are shown in Table 1.

Three preview surrogates were produced for each video, with following formats: (1) A copy of the static thumbnail from YouTube, (2) A filmstrip by extracting 10 equally spaced frames from the total length of each video, animated at 500 ms intervals, and (3) A ThumbReel constructed using the chosen query terms and the method described above.

Procedure

Task group one (TG1) compared thumbnails with filmstrips; task group two (TG2) compared filmstrips with ThumbReels. Each task group consisted of three preview comparisons, six in total. Respondents were assigned to groups with a formula that balanced responses across tasks and allowed completion of up to six tasks. In each task the participant was asked to observe a pair of surrogates matching the query terms and to state whether the second was 'worse', 'the same', or 'better' than the first. The respondent was also asked to supply a qualitative statement elaborating the reason(s) for this rating.

Analysis & Results

As this was a repeated-measures design with non-trivial within-subject correlations across comparison tasks, Generalised Estimating Equations (GEEs) allowed valid statistical inference. We collapsed the 'worse' and 'the same' responses into a single 'not better' category and estimated, for each task group separately, the GEE model of the resulting binary dependent variable using a logit link function.

Task Group 1: Across all three tasks, 81% of responses rated the filmstrip as being the better representation of the query terms than the thumbnail displayed by YouTube (see Table 3). Across-task working correlation matrix entries were high: $\rho_{12} = \rho_{21} = 0.490, \rho_{13} = \rho_{31} = 0.998, \rho_{23} = \rho_{32} = 0.507$. Exp($\beta$) is the estimated odds of the 'better than' response, which would take the value of 1 if the probability of a 'better than' response were equal to a 'not better than' response (i.e. if both probabilities were equal to 0.5). However, the GEE model's two-sided Wald tests strongly rejected the null hypothesis of H_0: Exp(β)=1 across all three TG1 tasks.

Task	'Better than'	n	Exp(β)	p-value
Task Group 1 (TG1)				
V1	21 (84%)	25	4.528	0.001
V2	19 (79%)	24	3.435	0.008
V3	19 (79%)	24	4.398	0.001
Task Group 2 (TG2)				
V1	18 (75%)	24	2.674	0.028
V2	11 (46%)	24	0.759	0.507
V3	21 (81%)	26	4.427	0.003

Table 3. ThumbReel evaluation results.

Task Group 2: Across all three tasks, 68% of responses rated the ThumbReel as being the better representation of the query terms than the filmstrip (see Table 3). Across-task working correlation matrix entries are moderate: $\rho_{12}=\rho_{21}=0.066, \rho_{13}=\rho_{31}=0.346, \rho_{23}=\rho_{32}=0.274$. The GEE model's two-sided Wald tests strongly rejected the null hypothesis of H_0: Exp(β)=1 for V1 and V3, but not for V2.

Overall: This study showed that participants found ThumbReels to better represent query serch terms than FilmStrips, which were in turn where better than thumbnails. Whilst it did not test 'real world' searching, it does provide positive support for the ThumbReels surrogate concept.

DISCUSSION

Scalability: Given time, and as the number of tags naturally increases, it may become difficult to identify distinct clusters of a specific tag within a video. One solution is to apply a weighting to clusters of tags, so that these become the primary area for frame extraction. This results in the discarding of lesser-clustered tags until they reach a determined level of significance as the tag distribution stabilises [4]. A stabilised tag distribution would also help to mitigate the risk that malicious users will join the crowd and corrupt the work of other volunteers.

Generalisability: Inviting participants to the study via social networking websites, has both a selection bias (given that we have targeted a particular group) and a self-selection bias (given that the study is voluntary). We acknowledge there may be an element of video selection bias present and future work is required to examine whether the preference for ThumbReels can be generalised to videos of different genre, varying length, across different demographics, and different varieties of VSWs.

The use of thumbnail surrogates is not exclusive to VSWs. Thumbnails are used online by a wide variety of digital content providers, with video and crowdsourced images for product descriptions prevalent amongst online retailers. These retailers increasingly rely on customer-sourced content to provide thumbnail previews for their expanding catalogues of products. This presents an alternative use case for ThumbReels by returning previews, images or video frames that are contextually relevant.

CONCLUSION

In this paper we introduced ThumbReels, a new type of surrogate for video content that helps bridge the 'intention gap'. ThumbReels are an animated preview, constructed of frames taken from a full video, marked by a crowdsourced tagging process and matched with query terms. We provided an open-source tool for end-user temporal tagging of videos. Two user studies evaluated the effectiveness of crowd-sourcing temporal video tagging (with good success) and participant perception of ThumbReels in representing query search terms (showing a significant improvement over other contemporary techniques for the test videos).

Planned future work includes algorithmically creating ThumbReels based on natural-language interpretation of crowd-sourced tags and further evolution to using query-sensitive (video) skims. Work is also planned to test the viability of dynamic creation versus cached ThumbReels.

ACKNOWLEDGEMENTS

We wish to thank Kim Kaivanto for his invaluable help with statistical evaluation and Curtis Kennington for help with the creation of the software tools used in this study. This research was supported by HighWire, a post-disciplinary Doctoral Training Centre at Lancaster University funded by the RCUK Digital Economy Programme through the EPSRC (Grant Reference EP/G037582/1).

REFERENCES

1. Christel, M. G. Evaluation and user studies with respect to video summarization and browsing. *Multimedia Content Analysis, Management, and Retrieval* (Jan. 2006), 60730M–60730M–15.
2. Christel, M. G., Winkler, D. B., and Taylor, C. R. Improving Access to a Digital Video Library. In *Proc. Interact 1997*, Chapman & Hall, Ltd (1997).
3. Estellés-Arolas, E., and Gonzalez-Ladron-de Guevara, F. Towards an integrated crowdsourcing definition. *Journal of Information Science 38*, 2 (Apr. 2012), 189–200.
4. Halpin, H., Robu, V., and Shepherd, H. The complex dynamics of collaborative tagging. *Proc Int. Conf. on World Wide Web* (2007), 211–220.
5. Haubold, A., and Natsev, A. Web-based information content and its application to concept-based video retrieval. In *Proc. Content-based Image and Video Retrieval*, ACM Press (2008).
6. Liu, C., Huang, Q., and Jiang, S. Query sensitive dynamic web video thumbnail generation. In *18th IEEE International Conference on Image Processing (ICIP)* (2011), 2449–2452.
7. Marchetti, A., Tesconi, M., and Ronzano, F. Semkey: A semantic collaborative tagging system. *Workshop on Tagging and Metadata for Social Information Organization at WWW* (2007), 8–12.
8. Steiner, T., Verborgh, R., Van de Walle, R., Hausenblas, M., and Vallés, J. G. Crowdsourcing event detection in YouTube video. *Proc. Detection, Representation, and Exploitation of Events in the Semantic Web* (2011), 58–67.
9. Yamamoto, D., Masuda, T., Ohira, S., and Nagao, K. Video Scene Annotation Based on Web Social Activities. *IEEE Multimedia 15*, 3 (2008), 22–32.