Journal on Multimodal User Interfaces
https://doi.org/10.1007/s12193-018-0268-0

ORIGINAL PAPER

An approach for exploring a video via multimodal feature extraction and user interactions

Fahim A. Salim[1] · Fasih Haider[2] · Owen Conlan[1] · Saturnino Luz[2]

Received: 28 October 2017 / Accepted: 28 June 2018

Abstract
Exploring the content of a video is typically inefficient due to the linear streamed nature of its media and the lack of interactivity. Video may be seen as a combination of a set of features, the visual track, the audio track and transcription of the spoken words, etc. These features may be viewed as a set of temporally bounded parallel modalities. It is our contention that together these modalities and derived features have the potential to be presented individually or in discrete combination, to allow deeper and effective content exploration within different parts of a video in an interactive manner. A novel system for video exploration by offering video content as an alternative representation is proposed. The proposed system represents the extracted multimodal features as an automatically generated interactive multimedia webpage. This paper also presents a user study conducted to learn its (proposed system) usage patterns. The learned usage patterns may be utilized to build a template driven representation engine that uses the features to offer a multimodal synopsis of video that may lead to efficient exploration of video content.

Keywords Multimedia analysis · Video representation · Multimodal video processing · Human media interaction

1 Introduction

Content consumption is increasingly becoming video oriented. Whether we want to entertain ourselves in free time or learn something new, we are relying on larger amount of video content than ever. Take YouTube[1] as an example: over a billion hours of video content are watched daily. The reasons for this are obvious, high speed internet has made access to high quality video very convenient and new devices have lowered the barriers to publish video content. But easy availability is not the only reason for our increasing reliance on video content.

Video is one of the most versatile forms of content in terms of multimodality. Multimodality is the greatest strength of video content. By multimodality we mean that videos are composed of a set of features, namely the moving video track, the audio track and other derived features, such as a transcription of the spoken words. Together these modalities can give a very effective way of communicating information. The content value of these modalities as a whole far exceeds their values separately.

However, richness both in terms of modalities and the amount of available video content create a challenge. It is very difficult, if not impossible, for users to fully view every piece of video that could be useful or important to them. Therefore, it is important that users are able to find the right content conveniently. However, finding the right

This research is supported by The ADAPT Centre for Digital Content Technology. ADAPT is funded under the SFI Research Centres Programme (Grant 13/RC/2106) and is co-funded under the European Regional Development Fund at School of Computer Science and Statistics, Trinity College Dublin, Ireland and EU H2020 project SAAM under Grant No. 769661 at the University of Edinburgh, UK and Science Foundation Ireland (Grant 12/CE/I2267).

✉ Fahim A. Salim
salimf@tcd.ie

Fasih Haider
Fasih.Haider@ed.ac.uk

Owen Conlan
owen.conlan@scss.tcd.ie

Saturnino Luz
S.Luz@ed.ac.uk

[1] ADAPT Centre, Trinity College Dublin, Dublin, Ireland

[2] Usher Institute, University of Edinburgh, Edinburgh, UK

[1] https://techcrunch.com/2017/02/28/people-now-watch-1-billion-hours-of-youtube-per-day/—last verified: October 2017.

Published online: 13 July 2018

video among many is just part of the problem as videos can vary in length, and some might run several hours in length. Many techniques have been proposed to aid users in finding the right content within videos, including linked-data based approaches [43], frame trees [21], and semantics-based approaches [13].

While these approaches tend to improve video browsing performance, pointing to the right portion in a video still requires users to watch a fair amount of the video to decide which content is relevant to their needs. As Lei et al. [22] observed, due to its linear nature, it might take longer to evaluate the content of a video than a textual document. Therefore it is reasonable to say that identifying the right content within video is still a cumbersome process.

Researchers have long identified the importance of user control in the process of video search [9]. However, as it will be discussed in Sect. 2, the focus has been on creating optimal user interfaces of a predominantly visual character. While these systems do add value to the search process they are quite limited in their use cases. Customizing the interactivity in these systems usually produces mixed result, as the way a user interacts with video content is highly dependent on the context of the task [10,16,29].

It is our contention that there is a need to look at video content differently. Video content can be viewed as a diverse multimodal content source by breaking the tight bond between the different modalities. By the tight bond we mean that in video content different modalities i.e. visual, textual and audio content are presented in a continuous linear stream. Breaking the tight bond between modalities can open up new opportunities for exploration of video content.

This paper is an extended version of our SPECOM 2017 paper [37]. In term of novel aspects, it proposes the following.

- An approach to automatically transform video content into an interactive multimedia document.
- The transformation is achieved through representation of multimodal features automatically extracted from the different modalities of video.
- Representing the content of video by automatically extracted feature have the following advantages
  1. Minimize if not eliminate the need for prior curation i.e. minimal or no human effort for video content curation.
  2. Allow end users to control not only the choice of modality but also the level of detail to consume video content.

Following are the contributions of this paper.

- A novel system and methods to allow users to explore a video using five different tabs in a flexible manner.
- Prototyping of the proposed system.
- Conducting of a user study for a subjective evaluation of the prototype.
- Statistical evaluation of user study for male and female users.
- Correlation among the usage of different tabs of the system.
- Evaluation of user preference of different tabs using Mallows–Bradley–Terry (MBT) model.

The proposed approach may be implemented as a template driven representation engine to represent the video as a multimedia web page.

The purpose of this research is to streamline the design of that template engine. A user study was conducted by utilizing a novel system prototype developed to learn the usage pattern of participants. This involved extracting multimodal features from TED videos and representing them to 29 users for evaluation.

2 State of the art

Current techniques allow users to explore the content within a video either by providing the ability to search for something in particular, or by giving an overall synopsis of a video i.e. video summarization. In video summarization, importance is usually attributed to visual features [2,7]. However multimodal features are also getting considerable attention due to the added value they bring in terms of identifying important segments [12,20] both in produced videos and in recordings of natural interaction where the interplay among multiple modalities is often core to understanding the content [3,4].

Searching for a piece of information within video assets is still predominantly text based, as seen on the main commercial offerings (such as YouTube, etc.). However, there has been a great deal of research on multimodal methods to extract information from videos. In Haesen et al. [18], the authors use face recognition along with name tags to find footage of certain people in videos, while Nautiyal et al. [31] use text recognition techniques to identify textual information appearing in videos. Matejka et al. [28] facilitate searching for relevant sections in baseball videos based on metadata attributes, while Schoeffmann et al. [38] facilitate navigation and searching within a video through visualization of low-level features and frame surrogates. In order to enhance the efficiency of search within video, Dong and Li [11] use ontologies. Waitelonis and Sacks [43] extend the idea of semantic-based search by using linked data to provide a time-based video index, which allows searching within the video. Rafailidis et al. [34] devise a framework for retrieving information using multimedia queries. Another interesting example of multimodal video search is Zhang et al. [44], where the authors devise a cross-media retrieval approach to search for video by giving audio samples and vice versa. In order to make the exploratory search process in video more interactive researchers long identified the importance of giving user more control as well as more responsibilities [9]. Moumtzidou et al. [30] use several content-based analysis and retrieval modules and a custom user interface to support the user in browsing through a video collection.

It is believed that giving users the ability to interact with video in a non-linear manner helps them in the process of exploration. Different systems have been proposed which help users explore time-based media [24], and video more specifically, in a modular and non-linear manner. Pavel et al. [32] use a chapter/section structure to provide skimmable summary and thumbnail of a video segments to a user, while Galuscakova et al. [15] customize a regular video player to anchor segments based on topic similarity. Others have created Hypervideos which give user the ability to interact with a video in a non-linear manner [17,39,41,42].

Current exploration within video approaches do utilize multimodal features of video in their processing but underutilized the potential of multimodal features in the representation. With the use of key-frames and other textual information, such as keywords, these approaches allow the user to search a particular item and to jump to a particular position in the video, but they lack the ability to provide an overall synopsis. Video summarization techniques provide the overall synopsis of the video, but the ability to get something particular is limited by the overall visual focus of the representation. While Hypervideos and other interactive video systems do give users more control in the exploration process, they require some degree of human curation and a purpose built user interface which ties them to the intended use cases, thus limiting their flexibility for evolving user needs.

It is our contention that an exploration approach can utilize multimodal features more widely to enhance the users exploration experience within a video by automatically curating its content on demand and by representing them in a configurable manner thereby transforming the video into an interactive multimedia webpage.

Such a transformative representation may have certain advantages, for example Calumby et al. [6] observed that since video content is multimodal in nature, multimodality in representation of content enhances the user exploration experience. However, while multimodality in representation is good, it is important to let the user choose the modality to explore content based on the task and their personal preference for improved task performance [10,16,29]. Automatically generating interactive webpage, would enhance the exploration opportunities of video content. This is because users can explore the content by directly interacting with the extracted features and reconfiguring the rendered webpage according to their evolving needs.

3 Proposed approach

The proposed approach is based on the idea of considering video a diverse multimodal content source. It works by using known techniques to segment a video and automatically extract multimodal features from its different modalities.

Upon a content exploration request, content of video is represented as an interactive multimedia webpage instead of a typical video stream, thereby transforming it into an interactive document. The interactive webpage is automatically curated by representing the extracted features to the user so that user can consume the different segment of the video in the modality of his/her choice and the amount of detail preferred.

This form of representation gives the user more flexibility and greater control over how they choose to interact with and consume the content of the video. Since the proposed approach creates the webpage automatically, it does not limit the interactivity to the designer's anticipated use cases, and allows the user to personalize the rendering according to the evolving task needs and personal preferences.

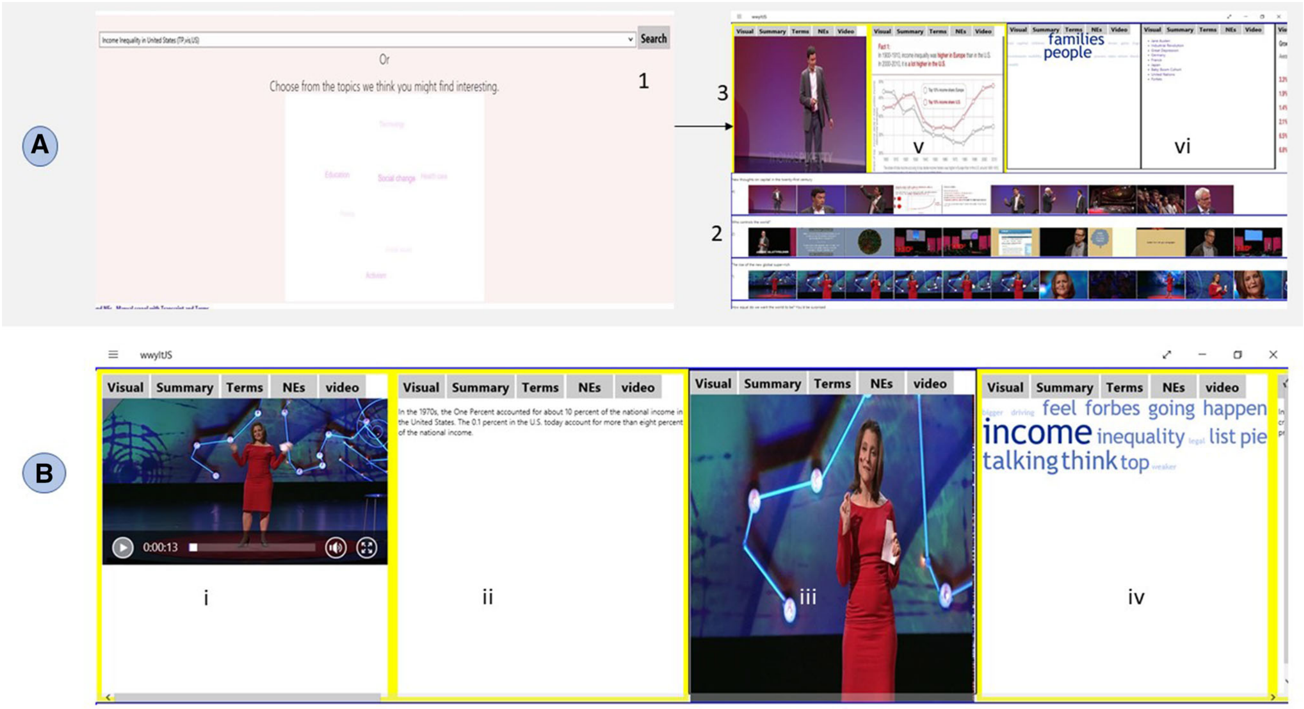
A user study was performed to validate the proposed approach by developing a prototype system to let users explore the content within a video.

4 Prototype design

A prototype system was developed to perform the user study. The prototype was built using HTML5 and JavaScript and designed to work with both traditional and touchscreen interfaces. Figure 1 shows the prototype. Figure 1a(1) shows the search box to enter textual queries. Figure 1a(2) shows results of the query. The results videos are shown as a filmstrip while the top row Fig. 1a(3) shows the detailed multimodal representation of a video.

Figure 1b shows the representation of the TED talk by Chrystia Freeland [14]. The C99 algorithm [8] divided the transcript of this video in 16 segments. Figure 1b shows the first four segments; the users can swipe or scroll to the right to see the rest of the segments. Highlighted in yellow are the segments where the key-terms searched by the user occurred for example Fig. 1b(i).

Each segment has five tabs, and each tab contains a rendering of extracted features. A description of the segmentation approach and each tab is given in the following sections.



Or

Choose from the topics we think you might find interesting.

In the 1970s, the One Percent accounted for about 10 percent of the national income in the United States. The 0.1 percent in the U.S. today account for more than eight percent of the national income.

Fig. 1 Screen shots of the prototype system. **a**(1) upper left, shows the search box. **a**(2) Upper right shows the results. **a**(3) Upper right shows the detail representation of a video. **b** Top row of results page i.e. detail representation of a video. **b** Screen shot of the video representation, showing 4 out of 16 segments. Users can swipe or scroll right to see the other 12 segments. Each square represents a segment, and each segment has five tabs, where each tab contains the rendering of automatically extracted features. Segment one and two are highlighted by a yellow border in the figure. User can tap or click on the tabs to see different renderings For example **b**(ii) currently shows the summary. In short, to explore this video, users have 16 segments and 80 tabs (16 segments ×5 tabs) to choose from

4.1 Segmentation

A video can be segmented in many ways utilizing different modalities (see Sect. 2). The choice of modalities is often domain dependent. For the current study, we segmented the video by utilizing the textual modality. Thus, the transcripts of TED videos were first split into sentences using StanfordNLP toolkit [25] and fed into the C99 text segmentation algorithm [8] in order to produce video segments.

Once segments are identified, to facilitate exploration, they can be represented to users in different ways. Users can not only choose which segments they want to explore but also what extracted features or their combination they would use to explore a chosen segment.

4.1.1 Highlighted segments

The segments in which the query terms appeared are highlighted with a thick yellow border to be distinguished from other segments. Our assumption is that users might want to interact more with highlighted segments than other segments. Figure 1b(i, ii, iv) shows the first two segments and the fourth as highlighted.

4.2 Visual

The *Visual* tab shows the key frames of a segment. A custom tool was developed using openCV [5] to detect camera shot changes. From those scene changes, one frame from each shot was selected. From those selected frames, frames with and without a face were identified using a HAAR cascade [23]. Speakers often use visual aids, such as presentation slides in a TED presentation. The heuristic was that shot without a face after a shot with a face might contain some images of visual aid used by the presenter which could contain useful information. Users can tap or click on the frame to see all the selected frames to get a visual synopsis of a particular segment. Figure 1a(v) and b(iii) shows examples of extracted key frames.

4.3 Summary

The *Summary* tab shows the automatic text summary generated from the transcript of the segment. An online summarization tool for generating text summaries [1] was utilized. Figure 1b(ii) shows an example of extracted summary.

4.4 Terms

The *Terms* tab shows the word cloud generated from the transcript of the segment. We used the online tool TagCrowd [40] for the word cloud generation. Figure 1b(iv) shows an example of word cloud.

4.5 NE

The *NE* (Named Entity) tab shows the list of extracted named entities from the transcript of the segment. This prototype used the Named Entity Recognizer tool [35]. Figure 1a(vi) shows an example of extracted named entities.

4.6 Video

The *Video* tab shows the video snippet of a segment. The text in transcript comes with timestamps. Once the segmenter segmented the text, timestamps were used to determine the start and end time of a particular segment and its video snippet was offered to users to watch. Figure 1b(i) shows an example of video segment.

5 User study

The proposed approach was validated by a user study which lets users explore the content within a video. The study is based on a simulated task scenario [19]. 29 users (21 males and 8 females) were asked to perform an exploratory search task to explore a video using the prototype (Sect. 4). Exploratory search is defined as a complex search task in which the user has to retrieve some facts first, which enable further search queries solving the overall search problem. Often the user is not sure about his/her search goal and sometimes, he/she is not very familiar with the topic of the search [26,43].

5.1 Study participants

A total of 29 users (21 males and 8 females) participated in the user study. They all have postgraduate degrees in computer science, digital humanities or related disciplines. Since TED talks are produced for a general audience therefore all the users were chosen not to be from an economics background as all the test videos are on the topic of economics.

5.2 Participant instructions

At the beginning of each session the user is given a briefing on the functionality of the developed prototype. After the introduction to the prototype, user is asked to perform the exploratory search task Fig. 1a(1). The user is asked to imagine that they want to get some knowledge about a particular topic e.g. economic inequality.

Users were asked to perform an exploratory search query. For this study, the query was pre-selected to be "Income inequality in the United States". The result of their query are TED videos Fig. 1a(2).

Users were asked to concentrate only on the top row i.e. the detail representation of one video (Fig. 1b). But instead of watching the video, users explored the video using our custom representation (Sect. 4). It was left to the user's discretion to choose the combination of segments and tabs they thought sufficient to have the overall synopsis of the video in regard to the query.

Each user performed the query twice. That is, each user explored two TED talks using the prototype one by one.

5.3 Test videos

Because of their general and storytelling nature, TED videos appeal to a wide and diverse audience, and therefore are a good candidate for our research. While our experiment is performed on TED videos, we assume that this approach can be extended to other content types such as life-logging videos, product launches and video messages. Due to its information seeking focus, the proposed approach is not expected to be effective for other video types such as movies, songs or entertainment oriented videos.

5.3.1 Video one: "New thoughts on capital in the twenty-first century" by Thomas Piketty

This TED talk [33] is approximately 21:00 min in length and consists of two parts: the first part consist of a presentation while the second part is an interview. The presenter used slides and charts extensively during the presentation. It is our opinion that the information presented in the video is on average more technical than a typical TED video.

5.3.2 Video two: "The rise of the new global super-rich" by Chrystia Freeland

This TED video [14] is approximately 15:20 min in length and it is different than the first video in a number of ways. Firstly, the video solely consists of the presentation and contains no interview, furthermore the presenter does not use any slides or any other visual aid during her presentation. It is our opinion that this video was easier to comprehend than the first one due to the general nature of the information provided.

5.4 Feedback capturing

Following are the types of feedback captured during the user study.

1. User interaction (to analyze the user interaction with the representation)
2. Verbal feedback (to understand the user's decision making while interacting with the representation)
3. User satisfaction questionnaire (to gather quantitative data on the user's experience with the representation)

Following sections provides the details for each of the feedback type.

5.4.1 User interactions with tabs

Both audio recording and screen capture footage were analyzed and annotated manually by the researchers. The following heuristic was employed to record feedback: the more a user selects a particular tab for a segment, the more interested they are in consuming the information using that particular feature rendering. Therefore, each interaction with tabs was noted down, thus when a user chose to view *Terms* of a particular segment, this counted as a user interaction with the representation.

5.4.2 Think out loud

We employed a think aloud protocol [36] approach to elicit and analyze user feedback. The following procedure was followed in order to learn exploration patterns in the recorded data. By encouraging users to think out loud we intended to record their thought process in exploring the content of the represented video. We were particularly interested in user comments regarding the effectiveness of different features in terms of efficiency and usability for exploring the video. Feedback such as "the speaker is talking about the 70's here" was not deemed as important as "I find this summary more useful than the last one" or "I find short summaries useful" etc. Therefore the latter two examples were noted down. Users were encouraged to provide their opinion about the video representation during the post experiment debriefing. They were asked open-ended questions such as what kind of information they would like to see in such content representations.

5.4.3 User satisfaction questionnaire

After the experiment, users were asked to complete a questionnaire to provide their feedback on the different aspects of the representation. The questionnaire contained 10 questions to be answered on 5-point Likert scale (5 denoting strong agreement). The first three questions regarded ease of use of the prototype. Questions 4–6 regarded the segmentation of the videos, and questions 7–10 concerned the user's perception of efficiency in viewing the video through the proposed representation compared to watching the video in a conventional manner. The full list of questions can be seen in Fig. 8.

6 Analysis

This section describes the analysis of the data collected from the user study.

- Analysis of user satisfaction through questionnaire: We have asked users to fill out a questioner after interacting with the system. The response obtained from users is reported and analyzed using Kruskal–Wallis test to demonstrate the differences for male and female users.
- Relationship between the usage of different tabs in the representation to identify any pattern that could help in streamlining the design of such systems.
- A subjective analysis of the verbal feedback to get suggestions from participants in order to improve the design.

6.1 Analysis of user satisfaction questionnaire

We have in total 30 responses from users (21 males and 9 females). The data of males and females is not balanced. So that we divided the males data into three folds. First fold have in total 9 males data, second fold has other 9 males data and third fold has remaining 3 males plus 3 males from fold 1 and 3 from fold 2 which were randomly selected. As a result, we have three male folds and one female fold. The mean and standard deviation values of all the male folds including female fold is depicted in Figs. 2 and 3, while Table 1 depicts the overall mean and standard deviation for all the participants. The motivation behind in dividing the male data set into folds is to balance it against female data for statistical evaluation. From Fig. 2, we observed that the females score has a higher mean value than male scores except for question one where male-fold2 has a higher mean. Later we performed Kruskal–Wallis test to compare the mean values and found that the difference between male-folds and female fold is not statistically different ($p > 0.5$) except only in one case i.e. question 4 ($p_{Female--Male-fold-1} = 0.04$) as depicted in Fig. 4.

Figure 2 shows the average score given to each question by participants. As it can be seen that users liked the representation in general, as the lowest average score is 3.5 out of 5. Female participants gave higher scores to the representation compared to their male counterparts, but the difference is not statistically significant ($p > 0.05$, obtained using Kruskal–Wallis analysis as depicted in Fig. 4). It can also be seen

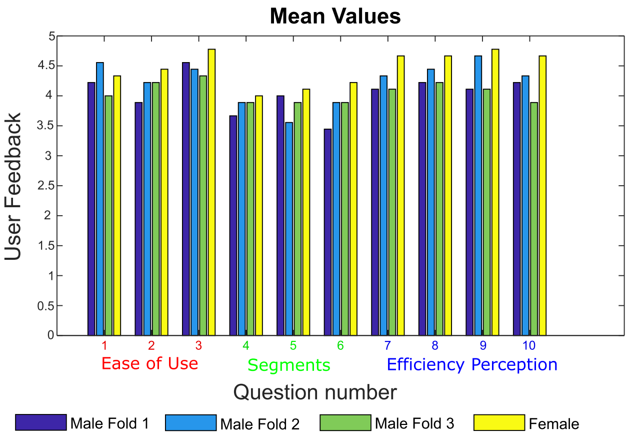


Fig. 2 Mean values of feedback by 29 users

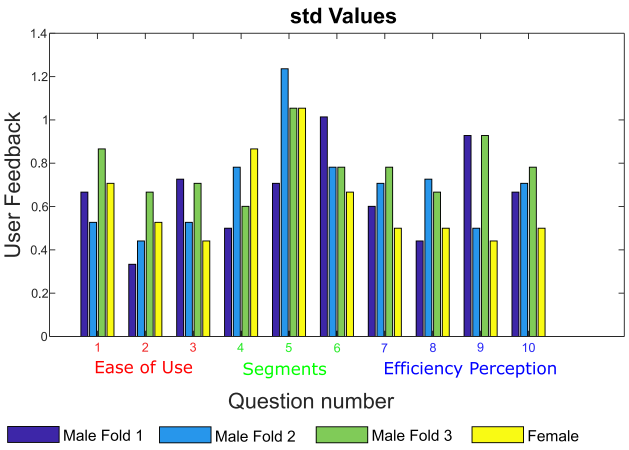


Fig. 3 Standard deviation values of feedback by 29 users

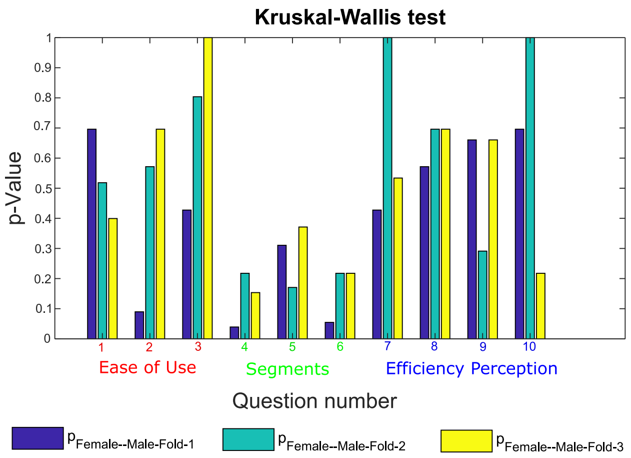


Fig. 4 *p* values of feedback by 29 users

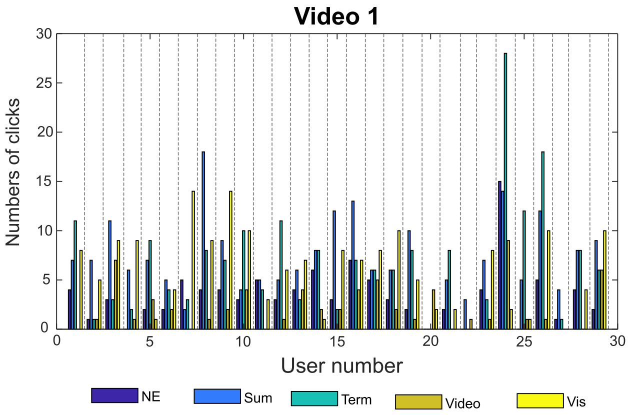


Fig. 5 Number of clicks by 29 users on the systems tabs for the first video

in Fig. 2 that while the overall ease of use and perceived efficiency was scored quite positively by the users, their satisfaction with the segmentation of test videos is lower than the rest (questions 4–6).

The numbers of clicks on the system's tabs for both videos are shown in Figs. 5 and 6.

6.2 Analysis of user interactions

We conducted statistical significant test of a null hypothesis that the number of clicks on each tab is same for both type of videos. As user number one did not explore the second video, we ignored his dataset for this statistical test. In addition, as the second video does not have presentation slides (vis), we also removed the visual tab from this evaluation. As a result a data set of 28 users and number of clicks on 4 tabs (video, summary, NE, Term) is used for statistical evaluation. The Kruskal–Wallis test did not reject the null hypothesis in 3 out of 4 cases $p_{term} = 0.82$, $p_{NE} = 0.68$ and $p_{Video} = 0.44$, but rejected it for the summary tab ($p_{Sum} = 0.04$). This is an indication that mean value of user clicks is statistically different for both type of videos but only for the summary tab. One of the possible reason is that the video 1 has presentation slides and video 2 has no presentation slide which make users to use the summary tab more for video 2 than video 1.

Table 1 Mean and standard deviation values of feedback of all the participants

	Q.1	Q.2	Q.3	Q.4	Q.5	Q.6	Q.7	Q.8	Q.9	Q.10
Mean	4.27	4.20	4.57	3.80	3.87	3.83	4.37	4.37	4.40	4.37
SD	0.74	0.48	0.57	0.71	0.97	0.83	0.61	0.61	0.81	0.67

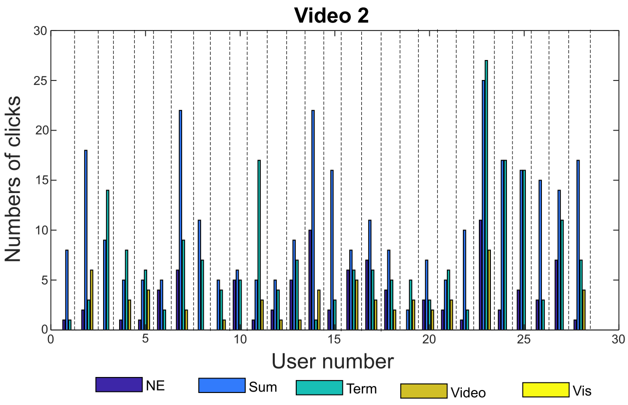
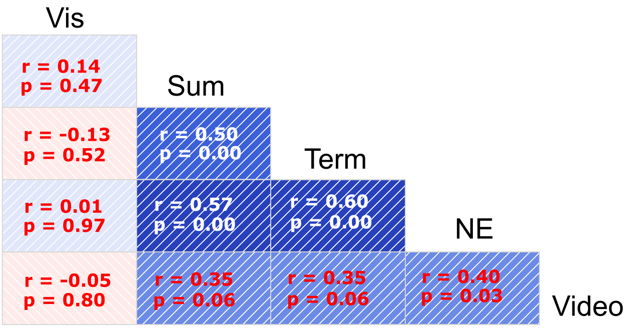


Fig. 6 Number of clicks by 28 users on the systems tabs for the second video



	Vis	Sum	Term	NE

Sum	r = 0.14
	p = 0.47			

Term	r = -0.13
	p = 0.52
		r = 0.50
		p = 0.00		

NE	r = 0.01
	p = 0.97
		r = 0.57
		p = 0.00
			r = 0.60
			p = 0.00	

Video	r = -0.05
	p = 0.80
		r = 0.35
		p = 0.06
			r = 0.35
			p = 0.06
				r = 0.40
				p = 0.03

Fig. 7 Pearson correlation between usage of tabs (representation)

We performed the Pearson correlation test on user interactions with tabs (Sect. 5.4.1), with a null hypothesis that there is no relationship between the usage of tabs (where usage of tabs is defined as number of clicks on tabs) in the proposed system. Figure 7 shows that the usage of tabs Sum, Term and NE is correlated pairwise, and that these correlations are statistically significant ($p < 0.05$). For example, usage of 'Sum tab' is correlated with usage of 'NE tab' and this correlation ($r = 0.57$) is statistical significance($p < 0.05$).

Finally, we ranked the tabs for each user by counting the number of times a user clicked on a tab. Sometimes users have the same number of clicks for two tabs. In that case, we calculated an aggregated response from other users who have a different number of clicks for those tabs. This response is calculated in terms of how many users rank one of two tabs as higher/lower than other tab. In case most users rank a tab $t1$ higher than another tab $t2$, then $t1$ is assigned a higher rank than $t2$. As there are 5 tabs, and number of possible rank orderings (permutation elements) is 120. Letting rankings be represented as 5-tuples of the form (visual, summary, term, NE, video), the most often chosen ranking is (4, 1, 2, 3, 5), chosen by 6 users. We employed the Mallows-Bradley-Terry (MBT) model to estimate parameters for the ranked data. The computed MBT parameters were (0.11071, 0.46184, 0.26687, 0.09943, 0.06115). Based on these parameters, the estimated order of ranking is (3, 1, 2, 4, 5), which means that users prefer mostly the Summary tab, then the Term tab, then Visual tab, then the NE tab and, lastly, the video tab.

6.3 Subjective analysis of verbal feedback

We analyzed the verbal and observational (Sect. 5.4.2) feedback of the users using grounded theory [36]. While the details of our use of this method is beyond the scope of this paper, we briefly summarize the results here. The analysis revealed an overall tendency of users to make time versus depth decisions. Different renderings (tabs) (Sect. 4) have different capabilities in terms of efficiency and effectiveness. Consider the rendering of a word cloud of key terms. This rendering is very efficient to consume as it can be glanced very quickly to give the user an idea of what a segment might be about, compared to a textual summary which would require more time to read but would give a deeper synopsis of the segment.

6.3.1 Time versus depth decisions

While watching, the video snippet of a particular segment would be the most effective way to consume a segment. However, it may also be the most time-consuming. Users not only chose certain tabs because of the length of the segment, but they also chose them according to their personal preferences. Some of the users identified themselves as "detail-oriented" in verbal feedback, they opted for tabs that showed more detail, such as the video snippets, while others opted more often for word cloud of *Terms*. This shows the personalization potential of the proposed representation.

7 Discussion and future work

Statistical analysis of the questionnaire revealed that while users found the representation easy to use and effective in helping them to comprehend the content within a video, they did not always agree with the segmentation performed by the chosen algorithm. They particularly disliked the fact that some segments were either too short or too long. In the verbal feedback, they often expressed their desire for more balanced segments of the video. However, since users had the flexibility of choosing the representation of their choice they rated other aspects higher (see Fig. 2).

Analysis of user interaction revealed that most users preferred textual representation compared to visual representation. The MBT model parameters showed that the *Summary* and word clouds of *Terms* were the most chosen representations by the users while watching the video snippet was the last. Analysis of verbal feedback backed this up. Many users reported themselves as fast readers and hence found textual representations more useful. The ranking model (MBT) gives user preference ranking patterns which can be used to further streamline the design of the system. Pearson correlation test results show that the usage of the Visual tab is less correlated with the use of other tabs, while usage of the other four tabs is more closely correlated pairwise [statistically significant ($p < 0.05$) 4 out of 6 cases]. This can be used in designing a system which offers fewer tabs than the current version because from the correlation results it is observed that the usage of the *Summary* tab is correlated more with the usage of *NE* tab than others. Therefore in streamlining the system design, the other tabs can be removed in order to provide a less cluttered interface.

7.1 Choice factors

The main aim of this study was to learn user interaction patterns with represented features in order to design a better representation approach for users. Our assumption was that the user would be interacting more with the highlighted segments in the representation. Even though 18 out of 29 users did most of their exploration in highlighted segments, user also interacted with non-highlighted segments quite often as well. This finding seems consistent with the definition of exploratory search in that user needs evolve during the process [27].

While relevance was a factor in choosing one segment over an other, non-relevant cannot be completely ignored in the representation as user might want to consume them to get the context. User personal preference played an important role in their choice of features (tabs) for interaction.

In the recorded feedback and also in the post experiment interview, users unanimously reported that they were missing the information regarding the length of each segment versus the length of the whole video. They considered that information as an important factor in their choice of rendering to consume the information. Informing them about the length of a particular segment influenced their choice of tabs for that segment. For example, for a long segment they preferred a rendering such as a word cloud of *Terms* to quickly get the information, while for a shorter segment they might read the summary or watch the video snippet.

Displaying relevant meta-data appears to be a desirable feature for users in a exploration system.

Based on the above discussion we can identify the following factors as a useful guide in designing a system for interactive search within video:

- Relevance (relevance with respect to the test query and also user's personal interest)
- User Preference (in terms of modality and amount of detail)
- Length and other relevant meta-data.

Currently the prototype simply represents all extracted features for each segment to the user to choose from (see Fig. 1). We believe that by using the above factors we can design a representation engine that can automatically create a personalized webpage for the user. This would reduce the amount of clicking needed by the user to explore the content of video more effectively.

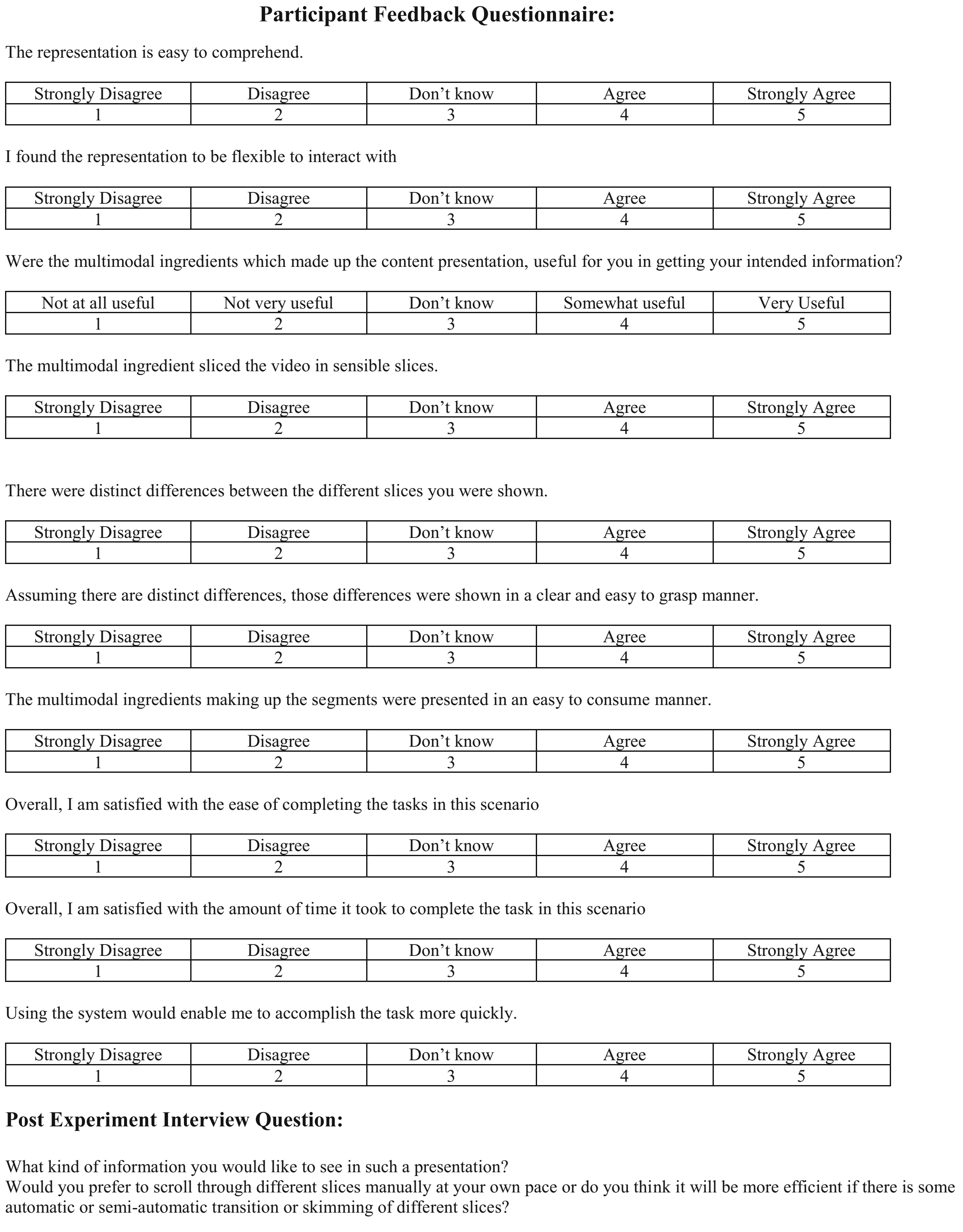
7.2 Future work

Our aim is to reduce the number of choices for the user while exploring the video by multimodal representation. Based on the results of the user study we have learned some usage patterns. We aim to utilize these usage patterns to develop some representation templates for video exploration. Those templates will be used to develop a representation engine which will offer an automatically generated multimodal synopsis of a video to the user for efficient exploration. Some users did not like the segmentation of the video using the current technique. We will also experiment using different segmentation techniques by utilizing semantic uplift of the topics discussed in the video.

Although the current experiment was performed as a query based search scenario, we believe that our proposed approach for video's exploration can be applied to other scenarios. For example, where the user has to consume content within video recordings such as a quick synopsis of meeting recorded by a smart conferencing system or getting a key point in a technical tutorial video.

Appendix

See Fig. 8.



Participant Feedback Questionnaire:

The representation is easy to comprehend.

Strongly Disagree	Disagree	Don’t know	Agree	Strongly Agree
1	2	3	4	5

I found the representation to be flexible to interact with

Strongly Disagree	Disagree	Don’t know	Agree	Strongly Agree
1	2	3	4	5

Were the multimodal ingredients which made up the content presentation, useful for you in getting your intended information?

Not at all useful	Not very useful	Don’t know	Somewhat useful	Very Useful
1	2	3	4	5

The multimodal ingredient sliced the video in sensible slices.

Strongly Disagree	Disagree	Don’t know	Agree	Strongly Agree
1	2	3	4	5

There were distinct differences between the different slices you were shown.

Strongly Disagree	Disagree	Don’t know	Agree	Strongly Agree
1	2	3	4	5

Assuming there are distinct differences, those differences were shown in a clear and easy to grasp manner.

Strongly Disagree	Disagree	Don’t know	Agree	Strongly Agree
1	2	3	4	5

The multimodal ingredients making up the segments were presented in an easy to consume manner.

Strongly Disagree	Disagree	Don’t know	Agree	Strongly Agree
1	2	3	4	5

Overall, I am satisfied with the ease of completing the tasks in this scenario

Strongly Disagree	Disagree	Don’t know	Agree	Strongly Agree
1	2	3	4	5

Overall, I am satisfied with the amount of time it took to complete the task in this scenario

Strongly Disagree	Disagree	Don’t know	Agree	Strongly Agree
1	2	3	4	5

Using the system would enable me to accomplish the task more quickly.

Strongly Disagree	Disagree	Don’t know	Agree	Strongly Agree
1	2	3	4	5

Post Experiment Interview Question:

What kind of information you would like to see in such a presentation?
Would you prefer to scroll through different slices manually at your own pace or do you think it will be more efficient if there is some automatic or semi-automatic transition or skimming of different slices?

Fig. 8 User satisfaction questionnaire

References

1. autosummarizer.com (2016) http://autosummarizer.com/
2. Belo L, Caetano C, do Patrocínio Z, Guimarães SJ (2016) Summarizing video sequence using a graph-based hierarchical approach. Neurocomputing 173:1001–1016. https://doi.org/10.1016/j.neucom.2015.08.057
3. Bouamrane MM, King D, Luz S, Masoodian M (2004) A framework for collaborative writing with recording and post-meeting retrieval capabilities. In: Proceedings of the sixth international workshop on collaborative editing systems, Chicago, November 6, 2004. IEEE distributed systems online journal on collaborative computing

4. Bouamrane MM, Luz S (2007) An analytical evaluation of search by content and interaction patterns on multimodal meeting records. Multimed Syst 13(2):89–103. https://doi.org/10.1007/s00530-007-0087-8
5. Bradski G (2000) The OpenCV Library. Dr. Dobbs J Softw Tools 120:122–125
6. Calumby RT, André M, Torres S (2017) Neurocomputing diversity-based interactive learning meets multimodality. Neurocomputing 259:159–175. https://doi.org/10.1016/j.neucom.2016.08.129
7. Chen F, De Vleeschouwer C, Cavallaro A (2014) Resource allocation for personalized video summarization. IEEE Trans Multimed 16(2):455–469. https://doi.org/10.1109/TMM.2013.2291967
8. Choi FYY (2000) Advances in domain independent linear text segmentation. In: Proceedings of NAACL 2000, Stroudsburg, PA, USA, pp 26–33
9. Cobârzan C, Schoeffmann K, Bailer W, Hürst W, Blažek A, Lokoč J, Vrochidis S, Barthel KU, Rossetto L (2017) Interactive video search tools: a detailed analysis of the video browser showdown 2015. Multimed Tools Appl 76(4):5539–5571. https://doi.org/10.1007/s11042-016-3661-2
10. Craig CL, Friehs CG (2013) Video and HTML: testing online tutorial formats with biology students. J Web Librariansh 7(3):292–304. https://doi.org/10.1080/19322909.2013.815112
11. Dong A, Li H (2008) Ontology-driven annotation and access of presentation video data. Estudios de Economía Aplicada 26(2):840–860
12. Evangelopoulos G, Zlatintsi A, Potamianos A, Maragos P, Rapantzikos K, Skoumas G, Avrithis Y (2013) Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. IEEE Trans Multimed 15(7):1553–1568. https://doi.org/10.1109/TMM.2013.2267205
13. Farhadi B, Ghaznavi-Ghoushchi MB (2013) Creating a novel semantic video search engine through enrichment textual and temporal features of subtitled YouTube media fragments. In: Proceedings of the 3rd international conference on computer and knowledge engineering, ICCKE 2013 (Iccke), pp 64–72 https://doi.org/10.1109/ICCKE.2013.6682857
14. Freeland C (2013) The rise of the new global super-rich. https://www.ted.com/talks/chrystia_freeland_the_rise_of_the_new_global_super_rich
15. Galuščáková P, Saleh S, Pecina P (2016) SHAMUS: UFAL search and hyperlinking multimedia system. Springer, Cham, pp 853–856. https://doi.org/10.1007/978-3-319-30671-1_80
16. Ganier F, de Vries P (2016) Are instructions in video format always better than photographs when learning manual techniques? The case of learning how to do sutures. Learn Instr 44:87–96. https://doi.org/10.1016/j.learninstruc.2016.03.004
17. Girgensohn A, Marlow J, Shipman F, Wilcox L (2015) HyperMeeting: supporting asynchronous meetings with hypervideo. In: Proceedings of the 23rd annual ACM Conference on multimedia conference, pp 611–620. https://doi.org/10.1145/2733373.2806258
18. Haesen M, Meskens J, Luyten K, Coninx K, Becker J, Tuytelaars T, Poulisse G, Pham T, Moens M (2011) Finding a needle in a haystack: an interactive video archive explorer for professional video searchers. Multimed Tools Appl 63(2):331–356. https://doi.org/10.1007/s11042-011-0809-y
19. Halvey M, Vallet D, Hannah D, Jose JM (2014) Supporting exploratory video retrieval tasks with grouping and recommendation. Inf Process Manag 50(6):876–898. https://doi.org/10.1016/j.ipm.2014.06.004
20. Hosseini MS, Eftekhari-Moghadam AM (2013) Fuzzy rule-based reasoning approach for event detection and annotation of broadcast soccer video. Appl Soft Comput 13(2):846–866. https://doi.org/10.1016/j.asoc.2012.10.007
21. Hudelist MA, Schoeffmann K, Xu Q (2015) Improving interactive known-item search in video with the keyframe navigation tree. Springer, Cham, pp 306–317
22. Lei P, Sun C, Lin S, Huang T (2015) Effect of metacognitive strategies and verbal-imagery cognitive style on biology-based video search and learning performance. Comput Educ 87:326–339. https://doi.org/10.1016/j.compedu.2015.07.004
23. Lienhart R, Kuranov A, Pisarevsky V (2003) Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In: Proceedings of the 25th DAGM pattern recognition symposium, pp 297–304. https://doi.org/10.1007/978-3-540-45243-0_39
24. Luz S, Masoodian M (2004) A mobile system for non-linear access to time-based data. In: Proceedings of the working conference on advanced visual interfaces, ACM, pp 454–457
25. Manning C, Surdeanu M, Bauer J, Finkel J, Bethard S, McClosky D (2014) The Stanford CoreNLP natural language processing toolkit. In: ACL system demos, pp 55–60
26. Marchionini G (2006) Exploratory search: from finding to understanding. Commun ACM 49(4):41–46. https://doi.org/10.1145/1121949.1121979
27. Marchionini G (2006) From finding to understanding. Commun ACM 49(4):41–46
28. Matejka J, Grossman T, Fitzmaurice G (2014) Video lens : rapid playback and exploration of large video collections and associated metadata. In: Proceedings of UIST'14, pp 541–550. https://doi.org/10.1145/2642918.2647366
29. Merkt M, Schwan S (2014) Training the use of interactive videos: effects on mastering different tasks. Instr Sci 42(3):421–441. https://doi.org/10.1007/s11251-013-9287-0
30. Moumtzidou A, Avgerinakis K, Apostolidis E, Aleksić V, Markatopoulou F, Papagiannopoulou C, Vrochidis S, Mezaris V, Busch R, Kompatsiaris I (2014) VERGE: an interactive search engine for browsing video collections. Springer, Cham, pp 411–414
31. Nautiyal A, Kenny E, Dawson-Howe K (2014) Video adaptation for the creation of advanced intelligent content for conferences. In: Irish machine vision and image processing conference, pp 122–127
32. Pavel A, Reed C, Hartmann B, Agrawala M (2014) Video digests: a browsable, skimmable format for informational lecture videos. In: Symposium on user interface software and technology, USA, pp 573–582. https://doi.org/10.1145/2642918.2647400
33. Piketty T (2014) New thoughts on capital in the twenty-first century. https://www.ted.com/talks/thomas_piketty_new_thoughts_on_capital_in_the_twenty_first_century
34. Rafailidis D, Manolopoulou S, Daras P (2013) A unified framework for multimodal retrieval. Pattern Recognit 46(12):3358–3370. https://doi.org/10.1016/j.patcog.2013.05.023
35. Ratinov L, Roth D (2009) Design challenges and misconceptions in named entity recognition. In: Proceedings of CoNLL '09, ACL, Stroudsburg, pp 147–155
36. Rogers Y (2012) HCI theory: classical, modern, and contemporary, vol 5. Morgan & Claypool Publishers, San Rafael
37. Salim FA, Haider F, Conlan O, Luz S (2017) An alternative approach to exploring a video. In: Karpov A, Potapova R, Mporas I (eds) Speech and computer. Springer, Cham, pp 109–118
38. Schoeffmann K, Taschwer M, Boeszoermenyi L (2010) The video explorer a tool for navigation and searching within a single video based on fast content analysis. In: Proceedings of the ACM conference on Multimedia systems, pp 247–258. https://doi.org/10.1145/1730836.1730867
39. Shipman F, Girgensohn A, Wilcox L (2008) Authoring, viewing, and generating hypervideo. ACM Trans Multimed Comput Commun Appl 5(2):1–19. https://doi.org/10.1145/1413862.1413868
40. Steinbock D (2016) http://tagcrowd.com/

41. Tian Q, Sebe N, Qi GJ, Huet B, Hong R, Liu X (2016) MultiMedia modeling. 22nd international conference, MMM 2016 Miami, FL, USA, January 4–6, 2016 proceedings, part I. Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics) vol 9516, pp 382–394. https://doi.org/10.1007/978-3-319-27671-7

42. Tonndorf K, Handschigl C, Windscheid J, Kosch H, Granitzer M (2015) The effect of non-linear structures on the usage of hypervideo for physical training. IN: Proceedings—IEEE international conference on multimedia and expo, August 2015. https://doi.org/10.1109/ICME.2015.7177378

43. Waitelonis J, Sack H (2012) Towards exploratory video search using linked data. Multimed Tools Appl 59(2):645–672. https://doi.org/10.1007/s11042-011-0733-1

44. Zhang H, Liu Y, Ma Z (2013) Fusing inherent and external knowledge with nonlinear learning for cross-media retrieval. Neurocomputing 119:10–16. https://doi.org/10.1016/j.neucom.2012.03.033

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.