Int J Digit Libr (2010) 11:125–140
DOI 10.1007/s00799-011-0071-y

Evaluation of visual video summaries: user-supplied constructs and descriptions

Stina Westman

Published online: 10 September 2011

Abstract Evaluation of video summarization approaches requires more information on the user-perceived qualities of different types of summaries. Also, evaluation measures need to be further developed in a user-led manner. This article reports on a user-centered evaluation of visual video summaries. Four types of summaries (fastforward, user-controlled fastforward, scene clips, and storyboard) were evaluated with a set of existing performance and satisfaction measures. A repertory grid elicitation was conducted with our participants gathering evaluation constructs related to both video summary content and controls. Results showed a lack of correlation between performance and satisfaction measures. User-supplied evaluation constructs were shown to span both the performance and satisfaction dimensions of the video summary evaluation space. Most constructs achieved moderate to good inter-rater agreement in a consequent survey. Free descriptions of videos and respective summaries showed that while users are able to interpret object- and event-related information from short summaries, thematic inference lacked, leading to worse descriptions than for the full videos.

Keywords Video summarization · Video summaries · Evaluation measures · Repertory grid · Video attributes

1 Introduction

The goal of video summarization is to create video surrogates, i.e., summaries that facilitate access to video content. Video summaries are kinds of metadata akin to abstracts that stand for the full video object and are useful for the purposes of browsing and making sense of retrieved video objects. Visual video summaries may be employed in result sets where people aim to make relevance judgments about whether to look further or download the video. They could also assist in the retrieval process by functioning as navigation aids through a collection or a video stream. Evaluating the quality of video summaries in these different contexts presents significant challenges. Within the interactive evaluation paradigm, summaries may be evaluated in a task-specific manner (e.g., does the summary help in making relevance assessments) which aims for ecological validity [4]. Another viewpoint into the evaluation of video summaries focuses on their effectiveness in retaining the gist of the original video (i.e., their informative function) [40]. Useful measures for human performance related to video summaries are being investigated [23]. The full range of features that viewers use to evaluate video summaries is unknown and evaluation methodologies are still developing.

It is known that current subjective measures of summaries do not correlate strongly with users' performance on information retrieval tasks, and might be ill-defined (e.g., usefulness without a specified context) [39]. As an alternative to current measures, Taskiran and Bentley [39] suggest interviewing users after direct system use, to uncover richer attributes related to how users perceive video summaries. Addressing

This paper is a substantially revised and extended version of a paper (Evaluation Constructs for Visual Video Summaries) which originally appeared in the Proceedings of the 14th European Conference on Digital Libraries (ECDL 2010).

S. Westman (✉)
Department of Media Technology, Aalto University School of Science, P.O. Box 15500, 00076 Aalto, Finland
e-mail: stina.westman@aalto.fi

this suggestion, a user study on video summarization was conducted which combined existing performance and satisfaction measures with an exploratory analysis in the form of repertory grid analysis. The aim was to see how currently suggested performance and satisfaction measures of video summaries relate and what additional evaluation constructs for visual video summaries could be elicited from users after exposure to several types of summaries. Furthermore, the ability of summaries to inform free description of video content was evaluated, comparing descriptions of summaries to those of full videos.

This study contributes to the understanding of the relative advantages and disadvantages of different types of video summaries, and describes a user-centered methodological approach for collecting users' criteria for summary evaluation via the repertory grid method. Results are presented on existing and new measures, and user-led criteria for visual video summary evaluation is discussed.

2 Related work

In this section, common approaches to video summarization and review evaluation measures used in user studies so far are reviewed and the repertory grid approach introduced.

2.1 Types of video summaries

Video summaries are created in order to give viewers access to video content without having to watch the entire video. This enables users to browse large video collections and otherwise interact with video sequences in a non-linear manner. Summaries may be used to find out if a specific video is the one we want to watch. Summaries are useful in a variety of domains, e.g., cinema, online video databases, distance education, mobile delivery of video, video conferencing [40]. For most applications, video summaries serve two functions: an indicative function, where the summary is used to indicate what topics are contained in the original video; and an informative function, where summaries are used to cover the information in the full video as much as possible, subject to summary length [40].

This article investigates various types of visual summaries of video content, using documentary videos. Summaries may be constructed for different video genres, and using various types of information. Several types of audiovisual cues may be utilized when selecting content for the summary: key frames, segments, graphics, and text [28]. The resulting video summary can thus take various forms: e.g., textual keywords, static keyframe mosaics, and dynamic video skims [41]. Automatic video summarization may utilize internal or external techniques [28]. Internal video summarization techniques identify segments of video for inclusion in the summary by analyzing low-level features in the video stream (e.g., color, shape, object motion, speech, on-screen text). External video summarization techniques collect and analyze contextual information (e.g., time and location in which video was recorded) and user-based information (e.g., descriptions of content, and browsing and viewing activity). Furthermore, hybrid techniques which use a combination of these techniques are used. The resulting summaries may further be typified according to what type of content they focus on, and the functionality offered to the user (interactivity, personalization) [28]. According to a recent review [28], most summaries are generic and internally-produced, relying on events and features rather than user perception of content. Also, hybrid summaries typically utilize object and event data specific to the domain.

Different modalities contribute differently to the information users gain from video summaries. Marchionini et al. [24] found that while audio summaries were thought to be less ambiguous and provide keywords, visual summaries provided the overall gist of the video topic. User controls to video summaries vary greatly, and users have expressed wishes toward more control over the summary playback [44]. In test settings, video summaries may be shown to users only once, combined with unlimited pausing [32] or both the capability to pause and replay [24].

2.2 Evaluation of video summaries

The need to evaluate video summarization methods and resulting summaries is clear. Borrowing from the field of text summarization, Taskiran et al. [40] divide evaluation efforts into intrinsic and extrinsic. In intrinsic evaluation the quality of summaries is evaluated directly, e.g., by judging the fluency of the summary, coverage of key ideas, or similarity to an ideal manually prepared summary. In extrinsic evaluation, the summary is evaluated with respect to its impact on the performance for a specific information task. Both types of evaluations have been conducted within the domain of video summarization, yet no standard methodology has emerged. Early on in the study of video summaries, He et al. [15] listed four desirable qualities for video summaries: conciseness, coverage, context, and coherence. In a similar vein, Taskiran [38] notes that a good video summary must be considerably shorter than the original video sequence; contain the important information of the original video, easy for users to grasp and follow. There clearly exist tradeoffs between these requirements, illustrating the internally conflicting goals of the summarization task.

Owing to the lack of a standard evaluation scheme, a variety of measures has been used in individual studies. Table 1 reviews the different evaluation setups, performance, and satisfaction measures used in some studies in the domain. Performance measures utilized in video summary

Table 1 Video summary evaluation approaches in various studies

Study	Evaluation setup	Performance measures	Satisfaction measures
[3]	Intrinsic and extrinsic evaluation of informativeness and utility of feature film summaries, including manually created skims	Informativeness by open-ended questions	Informativeness, enjoyability, usefulness for deciding to watch, overall quality, utility for understanding genre, atmosphere, narration pace, and characters
[6]	Extrinsic valuation of summary performance in different types of information tasks	Accuracy and speed of fact-finding by browsing and visual/textual gisting	Usability scales of e.g., terrible-wonderful, frustrating-satisfying, dull-stimulating, video and audio quality, ability to communicate essence, ability to inform question answering, preference
[9]	Intrinsic comparison of user-created summaries to system output on videos of different genres	Accuracy and error rate	
[14]	Intrinsic comparison of summaries to a manually created reference summary generated by several judges	Recall and precision in shot detection	
[15]	Extrinsic and intrinsic evaluation of coverage of key presentation ideas by four summary types, including author-generated summaries	Improvement to presummary scores on author-generated inference and factual questions	Conciseness, coverage of key points, coherence, clarity, ability to replace original, choppiness, quality
[16]	Extrinsic and intrinsic evaluation of storyboards and skims from news video and unedited footage	Precision and recall of shot change detection, processing time in summary generation	Information coverage, visual pleasantness, satisfaction
[17]	Intrinsic evaluation of sports video summaries generated based on user preferences and event metadata	Precision and recall in event detection	
[21]	Intrinsic evaluation of a skimming system	Accuracy of identification of events and speakers	Visual and audio quality, semantic continuity, ability to browse content, how well content was summarized, ability to replace original
[22]	Intrinsic evaluation of static and moving summaries of varying length videos		Satisfaction with number and selection of keyframes, enjoyability and informativeness of skims
[24]	User study measuring task performance with five summary types, elicitation of qualitative evaluations, summaries	Object recognition (textual/graphical), action recognition, gist determination (free text/multiple choice/visual)	Usability, usefulness, enjoyment, engagement
[27]	Intrinsic satisfaction evaluation for five summary types after seeing the original video		Satisfaction, representativeness, visual pleasantness, compactness, conveyance of story line, usefulness as poster
[29]	Intrinsic evaluation with comparison to full video		Informativeness, enjoyability
[31,32]	Evaluation campaign judging summaries from unedited video material against ground truth	Speed of summary creation, speed of judging against ground truth, summary size	Percentage of desired segments found, presence of junk, amount of redundancy, satisfaction with tempo, ease of finding content
[33]	User study with several types of information tasks performed using different types of skims (including manually created summaries) and full video	Accuracy and speed of fact-finding, text summarization, image recall	Preference
[34,35]	Intrinsic evaluation of three types of visual and audiovisual summaries of movie content		Coherence, ability to replace original, confidence in answering questions (who/what/where/when), preference
[40]	Evaluation of key information of documentaries retained by summary types by intrinsic and extrinsic methods	Accuracy in answering multiple choice factual questions	Ease of understanding, ability to replace original
[42]	Intrinsic and extrinsic evaluation of summaries of television series episodes, including full episode	Understanding by a written summary and multiple-choice questions	Ease of following story line, consistency, amount of scenes that of no added value, ability to replace original

evaluation include standard information retrieval measures of precision and recall. It is also possible to conduct evaluations relative to an optimal summary [14]. One may utilize either user-generated summaries [9,17] or professionally created summaries [36] as the benchmark. Participants may be asked fact-finding or inference questions about video

content which are used to calculate summary performance [23].

Measures of user satisfaction pertaining to video summaries are still largely lacking [44]. Subjective evaluations may be based on participants watching the full video and straightforwardly indicating their preference [6,33,34], or choosing the best option out of alternative summaries [8]. Li et al. [21] employed measures related to visual and audio quality, semantic continuity, ability to browse content, how well the summary summarized the content, and the degree to which the summary replaced the need to see the original. Ma et al. [22] had participants evaluate summaries according to their enjoyability (if perceptually enjoyable video segments were selected) and informativeness (capability of maintaining content coverage while reducing redundancy). Recently, a set of subjective measures on usability, usefulness, enjoyment, and engagement have been used [24]. Open-ended comments about summaries have been elicited in various studies [6,24]. Kopf et al. [20] obtained feedback from 17 users by means of informal discussions regarding the quality of the video summaries and the content included in them. Goodrum [13] had users perform multidimensional scaling of alternate summaries to investigate the similarities and differences of the summary types. Even large-scale surveys regarding the perceived level of quality of video summaries have been conducted [12].

In the last years, several suggestions for a joint evaluation paradigm have emerged. The TRECVID rushes (unedited video footage) summarization campaign is the first common evaluation campaign for video summarization [31,32]. It provides a common test data set, summarization task, and a set of evaluation measures. Objective performance measures include elapsed time for summary creation, time-on-task for judging against the ground truth, and the size of summary. Subjective measures include the percentage of desired segments found, presence of junk (color bars, clapboards, empty frames), amount of near redundancy, satisfaction with tempo and rhythm of presentation, and ease of finding desired content. TRECVID relies on ground truth from manual assessors, although automated approaches have been suggested [11].

Wildemuth et al. [25,44] have presented an evaluation framework with four classes of variables thought to influence performance and satisfaction in video summarization: user tasks, user characteristics, video characteristics, and summary characteristics. The tasks or human performance measures [23] were further defined into two classes of cognitive measures based on perceptual and conceptual facets of video viewing. Recognition measures evaluate subjects' recall about what they saw. Object recognition may be evaluated by ability to recognize by textual or visual stimuli (keyframes) objects seen in the summary. Action recognition is evaluated similarly by visual stimuli (video clips). Inference measures evaluate how subjects understood the aboutness of what they saw, i.e., the gist of the video. Linguistic gist is evaluated by the accuracy and coverage of a written summary or by having participants select the best written summary. For visual gist evaluation, participants are to select objects that "belong" in the video represented by the summary from still images not seen in the summary but present in the original.

2.3 Repertory grid analysis

The repertory grid is a cognitive mapping technique designed to reveal the personal constructs individuals use to structure and interpret phenomena [37]. The technique utilizes a set of elements considered important within the study domain, to elicit a corresponding set of constructs using one of a variety of interview methods.

For example, in an investigation in video summary evaluation, the different types of summaries would form the set of elements. When interviewed, participants would provide bipolar statements reflecting their perception of the summary types. Constructs represent participant's interpretations of the elements, forming measures along which the participant perceives the elements of the domain. These constructs are founded on perceptions of likeness and difference between the elements, e.g., easy to use versus hard to use. All members of the element set could then be evaluated by each participant along the elicited construct statements.

Various methods may be employed to link elements and constructs. Most often rating scales are used to differentiate between elements on each elicited construct. The elements themselves may be supplied by the researcher or elicited from participants. The results of rated repertory grids may be analyzed as individual grids or across several grids via multivariate methods (e.g., cluster analysis, factor analysis, correspondence analysis).

The repertory grid theory and technique is applicable to studies focusing on user-based evaluations because of its underlying assumption that we perceive our surroundings according to a personal system built on subconscious evaluations. It is also suitable for exploratory studies since, unlike a conventional questionnaire, the repertory grid utilizes constructs that originate from the participant. While the personal construct theory was developed in the field of clinical psychology, the method has been applied in a variety of domains. In information science, the repertory grid technique has been used to gain insight into e.g., mental models of information spaces [26], information retrieval systems [48], information assets [30], document types [10], and search engines [18].

3 Methodology

A user test on video summarization was conducted to answer the following questions: How do current performance and satisfaction measures for video summaries relate to each

Table 2 Summary types

Type	Description
Fastforward	Every 16th frame of the original video resampled
User-controlled fastforward	Every 16th frame of the original video with drag and drop controls for playback speed
Storyboard	Selected 16 keyframes in a grid
Scene clips	Compilation of 500 ms clips of original speed video from keyframes

other? What additional constructs for summary evaluation can be elicited from viewers? Are short visual summaries able to inform video description (e.g., for annotation) when compared to describing the full video?

The focus was on informative visual summaries for documentaries. Research results indicate that when only one modality with automatic summary generation is used, visual summaries fare better than audio [4]. We evaluate the ability of the summaries to convey information about the original video. This does not preclude their indicative function which was assessed in a subjective manner.

3.1 Participants

A total of 28 participants (12 female) were recruited through postings in university newsgroups. They were undergraduate and graduate engineering students between ages 20 and 37 (mean 24.8 years). None had previous experience working with moving video summaries.

3.2 Material

Four videos from the Open Video (OV) Project [1] were selected for the study: (1) A New Horizon, segment 5,[1] (2) Challenge at Glen Canyon, segment 5,[2] (3) Exotic Terrane, segment 10,[3] and (4) Hurricane Force—A Costal Perspective, segment 2.[4] They were all documentary videos with similar content dealing with forces of nature. These were full color videos of approximately 2 min long.

3.3 Summaries

Four types of summaries (Table 2) were produced for each video. Both static and moving summaries were devised, and one summary type included user controls.

Storyboard keyframes were taken from the Open Video library, and in two cases supplemented by manually selected keyframes to total 16 frames for each video. The supplementary keyframes were selected from shots not represented in the OV keyframe set. All keyframes were visible at once in a 4×4 grid.

Scene clips summary type was devised to be the moving counterpoint of the storyboard. It covered the same shots but instead of still keyframes, consisted of a compilation of 16 half-second clips of original normal-paced video, starting from the keyframes.

Fastforwards [44] consisted of a resampling of every 16th frame of the original video, the result of which was shown at regular video speed of 25 fps. The fastforward rate in this study was relatively slow due to the focus on informative summaries.

User-controlled fastforwards were based on the same resampled video as the fastforward but instead of automated playback, it was controlled by the user. User controls for fastforwards have been previously suggested [44]. The drag and drop functionality allowed for pausing and restarting as well as slower or faster playback of the video.

Summaries were roughly 6% of the length of the original videos. The summarization rate is considerably lower than the 4% and even 2% in TRECVID settings [5]. This is due to differences in the video material to be summarized. The fully edited documentary video excerpts merit different summarization rates than unedited, rushes footage.

Two viewings per summary (or respective time) were allowed due to findings that users in the unlimited viewing condition view visual summaries multiple times [24]. Time with summary was kept the same across the different summary types. In the experiment, scene clips and fastforwards were shown twice for a total of 16 sec of viewing time. Storyboards and user-controlled fastforwards were visible for 16 sec each. In addition, 2 extra seconds were given to access the controls on the user-controlled fastforward based on observations from the pilots conducted. The summary would not start playing back until controls were accessed.

3.4 Measures

The visual recognition and inference measures suggested in [23] were employed as performance measures (Table 3). Further added was an inference measure related to action based on the importance of actions and activities in visual gisting [46]. Action inference was evaluated in this study by the

[1] http://www.open-video.org/details.php?videoid=689,1:59.

[2] http://www.open-video.org/details.php?videoid=566,2:02.

[3] http://www.open-video.org/details.php?videoid=728,2:13.

[4] http://www.open-video.org/details.php?videoid=837,2:13.

Table 3 Performance measures

Measure	Stimuli and scoring
Object recognition	Accuracy in recognizing keyframes
Object inference	Accuracy in inferring content by keyframes
Action recognition	Accuracy in recognizing clips
Action inference	Accuracy in inferring content by clips
Text inference	Accuracy of free text descriptions

ability to visually infer actions and activities in the original video based on the summary.

Object recognition and *object inference* (visual gist by visual stimulus [23]) were evaluated by having participants indicate for 18 frames whether they:

1. saw the frame in the summary (recognition) [6 correct frames]
2. did not see but thought it belonged to the original video (inference) [6 frames]
3. did not see and thought it did not belong to the original video [6 frames; 3 from other segments of the same video, 3 from unrelated videos]

Recognition and inference questions were thus integrated into a multiple choice screen. Action recognition and action inference were evaluated in an analogous manner from a set of six two-second clips.

Text inference was also evaluated by gathering free text descriptions of both summaries and full videos. These were prompted for immediately after summary viewing, before any other recognition or inference tasks, so as to avoid any learning effects. The descriptions were scored according to guidelines from [47]. Descriptions were scored for both accuracy and detail, for objects and events as well as the overall theme or topic of the video. The final score for textual inference was the sum of these partial scores.

Satisfaction measures were derived and combined from literature. All the questions were rated on a 5-point Likert scale:

1. The summary was easy to understand [15,32,40]
2. The summary was enjoyable [3,22,29]
3. The summary was informative [3,22,29,41]
4. The summary was interesting [41]
5. The summary was coherent [15,34]
6. The summary represented the video well [27,40]
7. The summary would aid in deciding whether to watch the full video [3,40]
8. The summary would replace watching the full video [15,21,34,40,42]
9. The summary would be useful in browsing video archives [4,21]

All abovementioned performance results are reported as average accuracies in performance (%). All satisfaction results are reported as average ratings on the Likert scale between 1 and 5.

3.5 Procedure

An online system based on PHP/MySQL and JavaScript was used to administer the study and collect data. Summary type and video instance order were counterbalanced through a within-subjects Greco-Latin design of 4 blocks. The following procedure was used, with phases 2–7 repeated for the four summary type trials:

1. Introduction to test and practice viewing all summary types and questionnaires
2. Watch video summary
3. Describe video based on summary
4. Answer performance questions and satisfaction questions 1–5
5. Watch corresponding full video
6. Describe full video
7. Answer satisfaction questions 6–9 comparing summary and full video
8. Exit interview and repertory grid elicitation/survey

The experiment ran for an average of 57 min per participant (SD = 9 min). Out of this time, the four trials took an average of 38 min (SD = 8), where variance comes from the amount of time participants used to fill the questionnaires. The interview with the repertory grid elicitation lasted on average 9 min (SD = 2 min) and the interview together with the repertory grid survey took an average of 14 min (SD = 2 min).

3.6 Elicitation of evaluation measures

Additional constructs for the evaluation of visual video summaries were gathered from participants by two methods. First, in the exit interview participants were encouraged to reflect across summary types and video instances and answering e.g., Which summary type did you prefer? Which did you think was most informative? What did you pay attention to when evaluating the summary? Could you see yourself using these summaries? Was some video content more interesting than others?

Second, the repertory grid technique was used to study the participants' mental model of the summaries and to elicit evaluation constructs. The grid elements are different

Table 4 Accuracy of performance (%)

Type	Object recognition***	Object inference	Action recognition	Action inference	Text inference*
Fastforward	51.2		60.7		60.3
User-controlled fastforward	57.1		57.1		46.9
Storyboard	67.9	58.3	71.4	50.0	57.1
Scene clips	77.4	65.5	78.6	50.0	57.6

*** $p < .001$, ** $p < .01$, * $p < .05$

video summary types and the constructs related to these were elicited from one set of participants. Another set of participants then proceeded to rate the elements according to the constructs.

Two pilot participants and the first eight test participants took part in a repertory grid interview designed to elicit constructs. At the end of the test, they were presented with a screen displaying all four summaries they had seen in the trials. They were to name features shared between any two of those summaries or features distinguishing them. The interviewer noted down the constructs (e.g., usability) and asked further questions to establish its endpoints on the semantic differential scale (easy to use vs. difficult to use). The items pooled from the ten interviews revealed 20 distinct constructs. A single participant contributed between 1 and 5 constructs. No new constructs emerged after the 8th interview participant. The 20 items were listed in a survey instrument, designed to gather rating data for all summary types on a 5-point scale. The rest of the participants ($n = 20$) filled out the repertory grid survey with these 20 constructs as items. The survey data was used to evaluate both the summary types and the constructs gathered.

3.7 Analysis

Significant differences are reported here based on conducting one-way analyses of variance (ANOVA), assuming unequal variances. Post hoc comparisons were conducted upon obtaining a significant F value. All possible pairwise comparisons were analyzed via Games-Howell post hoc tests. Chi square tests were used to compare the distributions of different attributes in the video descriptions obtained. Only significant results at three different levels (***$p < .001$, **$p < .01$, *$p < .05$) are reported.

At the end of the summary trials, participants re-answered the recognition questions for the last video. This enabled us to gather a (between-subjects) baseline for the full videos. There was no statistically significant difference between the videos so absolute scores have been used. As the fastforward and user-controlled fastforward summaries included nearly all shots, object and action inference were disregarded for these and only recognition is reported.

Inter-rater agreement in scoring the free descriptions for text inference was evaluated. Between two independent raters there was a Pearson correlation of $r = .91$ which was deemed good. In addition to scoring for text inference, content analysis was performed on the free descriptions of full videos and summaries in order to see what types of video content they referred to. Full videos and summaries, as well as different summary types were compared in their descriptions based on an existing video attribute typology [46].

The construct rating data was evaluated for consistency via calculating an average linearly weighted Cohen's kappa [7] for each construct across all rater pairs. Correspondence analysis was conducted on the construct rating data in order to show interrelations between (1) summary types, (2) constructs and (3) summary types and constructs. These results are visualized in three dimensions and correlations are presented between axis coordinates and the construct ratings in order to characterize the dimensions resulting from the analysis.

4 Results

Quantitative results on user performance and satisfaction are presented by summary type, enhanced by qualitative results on the advantages and disadvantages of them, as perceived by our participants. Furthermore, a set of evaluation constructs elicited from our participants is presented, and the descriptions they gave of video content are analyzed. A section is devoted to assessing the connections between current and new measures.

4.1 Performance and satisfaction by summary type

For performance, summary types differed in terms of object recognition and text inference (Table 4). Summary type had an effect on object recognition ($F = 9.448$, df $= 3$, $p < .001$). According to Games-Howell post hoc tests, object recognition was better for scene clips than for either type of fastforwards ($p < .001$). Object recognition was also better for storyboards than fastforwards ($p < .05$). Performance in text inference also differed ($F = 3.015$, df $= 3$, $p < .05$) as fastforwards supported free text inference better than user-controlled fastforwards ($p < .05$).

Table 5 Satisfaction ratings (1–5)

Type	Understandable	Enjoyable**	Informative	Interesting	Coherent
Fastforward	3.15	2.69	3.15	3.27	3.04
User-controlled fastforward	3.27	3.38	3.04	3.54	3.04
Storyboard	3.46	3.69	3.31	3.23	2.65
Scene clips	3.54	3.15	3.50	3.58	3.27

*** $p < .001$, ** $p < .01$, * $p < .05$

Table 6 Additional satisfaction ratings (1–5)

Type	Representative*	Decision aid	Replacement***	Browsing tool
Fastforward	3.00	3.31	1.81	3.54
User-controlled fastforward	3.31	3.73	2.31	3.62
Storyboard	2.77	3.15	1.46	3.35
Scene clips	3.38	3.54	2.15	3.62

*** $p < .001$, ** $p < .01$, * $p < .05$

Among the satisfaction measures there were significant differences in enjoyability, representativeness, and ability to replace original (Tables 5, 6). The enjoyability of the summary types differed ($F = 5.050$, df $= 3$, $p < .01$) as storyboards were more enjoyable than fastforwards ($p < .01$). Differences were also found in how well the summary represented ($F = 3.005$, df $= 3$, $p < .05$) or could replace the full video ($F = 5.782$, df $= 3$, $p < .001$). Scene clips were more representative than storyboards ($p < .05$). Scene clips ($p < .01$) and user-controlled fastforwards ($p < .01$) would be better able to replace full video than storyboards.

Video instance had effects on the ease of understanding ($F = 12.802$, df $= 3$, $p < .001$), enjoyability ($F = 4.862$, $p < .01$), and informativeness ($F = 5.220$, $p < .01$) of the summaries. The content also influenced the summaries' representativeness ($F = 4.843$, df $= 3$, $p < .01$) and ability to inform decisions on whether to watch the full video ($F = 3.302$, df $= 3$, $p < .05$).

Presentation order had effects on both performance and satisfaction. Order had an effect on the ease of understanding a summary ($F = 3.175$, df $= 3$, $p < .05$), informativeness ($F = 3.324$, df $= 3$, $p < .05$), and decision aiding ($F = 4.014$, df $= 3$, $p < .01$). In all cases, the fourth and last summary received lower ratings as it was not as easy to understand ($p < .05$) nor as able to inform decisions on whether to watch ($p < .05$) than the third summary. Also, the fourth summary was rated as less informative than the first ($p < .05$). Scores in text inference differed by presentation order ($F = 3.509$, df $= 3$, $p < .05$) as they were higher for both first and third summary described when compared with the last summary ($p < .05$).

No effects of participant gender or age group (under 24 vs. older) were discovered.

4.2 Summary qualities

In the exit, interview participants were asked to indicate which summary type they preferred and which they thought most informative. Table 7 summarizes the findings and offers justifications in the form of comments from participants. The fastforward and scene clips were preferred as summary types and thought to be most informative by most. There seemed to be a tradeoff between the continuity offered by the fastforward and the slower pace of the selected clips. User-controlled fastforward was preferred by some but most indicated that it required too much of their concentration to control the playback and find a suitable speed. The storyboard was deemed by most not to be representative enough of the original videos. However, it was thought to be good for locating particular information and fast overviews of video content and parts.

Participants were asked which content they preferred. Most preferred the hurricane ($n = 12$) or volcano (9) video content with fewer mentioning dam (5) or water (2) as their favorite. The ease of understanding a summary ($F = 9.845$, df $= 1$, $p < .01$) and its interestingness ($F = 11.627$, df $= 1$, $p < .001$) were higher for preferred content. Also, the scores on how well the summaries were able to represent the content were higher for preferred content ($F = 4.159$, df $= 1$, $p < .05$).

4.3 Elicited evaluation constructs

The evaluation constructs elicited from the first set of participants are given in Table 8. The average linearly weighted Cohen's kappa [7] value for each construct reflects the agreement across the participant set in the use of the construct for

Table 7 (P)reference and (I)nformativeness (n) of summary types with reasoning

Type	P	I	Advantages and disadvantages
Fastforward	8	10	Best conveyance of plot, best overall picture, displays most information in a short time, most effective, shows whole video, offers continuity, no need to control, too fast
User-controlled fastforward	5	7	One can choose what to focus on, ability to review interesting spots, most control over playback, most useful, nothing gets missed, choice of playback speed, difficult to find suitable speed, must focus on controlling, good for unlimited viewing time
Storyboard	3	2	One can see overall picture at once, gives an idea about different parts of video, one can watch what one wants, good for searching for particular information, ability to spend more time on a frame, ability to focus on details, most informative, not representative of moving content, must browse between frames, does not form a whole, confusing, does not show everything, not enough information
Scene clips	12	9	Clearest, gives an overall idea of content, best usability, easy to follow, offers continuity, closest to original, best for recognizing scenes, time to process, suitable pace, missed parts, shots changed too fast

Table 8 Constructs scales and kappa (k) values

Construct	Scale endpoints 1–5	k
Adjustable speed	Standard speed–Adjustable speed	.86
Automated	Can stop playback–Cannot stop playback	.80
Controllable	Automated playback–Controllable playback	.59
Multiple images	Shows one frame at a time–Multiple frames at a time	.58
From video	Constructed from still images–Constructed from video	.53
Small space	Takes up lot of screen space–Takes up little screen space	.52
Moving imagery	Still images–Moving images	.49
Chosen shots	Covers whole content of the video–Contains selected spots	.48
Slow playback	Presentation speed high–Presentation speed slow	.47
Focusable	Have to focus on what is shown–Can select focus	.42
Increased speed	Normal paced video–Increased pace video	.41
Reviewable	Difficult to view a part closer–Easy to view a part closer	.37
Fidelity	Shots missing–Repeats the content of the video with fidelity	.36
Continuity	One has to piece up the continuation–Shows the continuation	.36
Usability	Difficult to use–Easy to use	.30
Skipping	Have to watch completely–Can skip parts	.26
Searchable	Difficult to locate a certain spot–Easy to locate a certain spot	.19
True content	Content deformed–Content not deformed	.19
Overall picture	Does not provide an overview–Provides an overview of video	.08
Fast aboutness	Slow to find out what video is about–Fast to find out what video is about	.04

the different summary types. A kappa value of 0 denotes no agreement and a value of 1 indicates perfect agreement.

Correspondence analysis was conducted on the construct rating data to show interrelations between (1) summary types, (2) constructs and (3) summary types and constructs. In correspondence analysis results, similar constructs and summaries plot close together. Results are presented for the first three axis (Figs. 1, 2), accounting for 68% of variance in individuals' summary rating data. To characterize each axis, Pearson correlations between axis coordinates and the construct ratings were calculated. In the visualizations confidence ellipses enclosing 95% of individual summary ratings have been

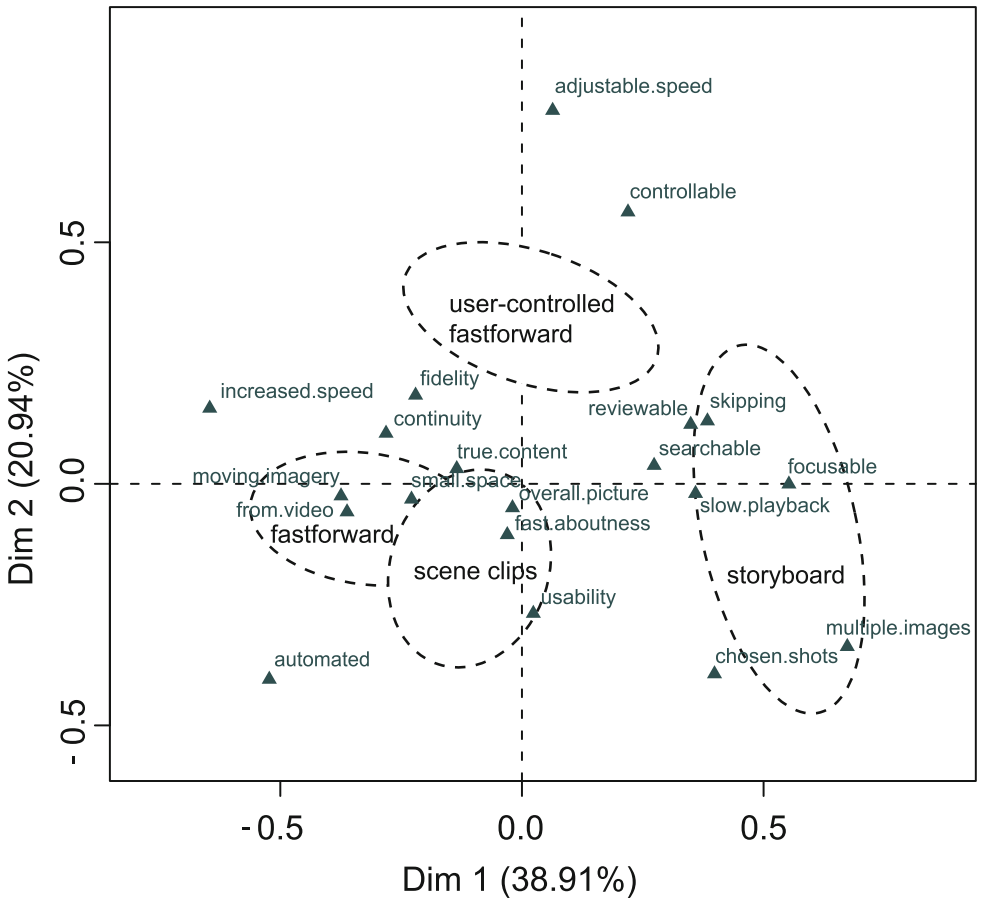

Fig. 1 Correspondence analysis results in dimensions 1 and 2

plotted instead of individual data points in order to simplify the illustrations.

Dimension one was related to the summary content (named here "coverage"). It was significantly correlated with e.g., not moving images ($r = -.81$), not increased speed ($r = -.78$), the ability to choose what to focus on ($r = .77$), multiple images ($r = .70$), and the inclusion of chosen parts from the video ($r = .64$). Dimension two was related to summary controls ("effort"). It was correlated with e.g., adjustability of playback speed ($r = .92$), controllability ($r = .81$), low usability ($r = -.65$), not automated ($r = -.43$), and not chosen shots ($r = -.54$). Dimension three was related to summary's depiction of the original video ("constancy"). It was correlated with e.g., not slow playback ($r = -.64$), multiple images ($r = .52$), fidelity ($r = .29$), overall picture ($r = .29$), and fast aboutness ($r = .26$).

4.4 Connections between measures

In a correlation circle (Fig. 3), the coordinates of each measure represent its correlations with the axes. While the explained variance in the two-dimensional plot for our data set is low (34%), the visualization serves to highlight connections between the types of measures (performance, satisfaction, constructs). Dimension one builds upon the performance measures while most satisfaction measures correlate more with dimension two.

As was to be expected, there were strong correlations between the performance measures: object and action recognition measures correlated ($r = .42$, $n = 112$, $p < .001$) as did the object and action inference measures ($r = .38$, $n = 56$, $p < .001$). Textual inference did not correlate significantly with other performance measures. There was significant correlation between object-oriented measures ($r = .24$, $n = 56$, $p < .05$) but almost none between action-oriented measures. All satisfaction measures correlated strongly ($p < .01$) amongst each other.

There were weaker correlations between performance and satisfaction measures. Object recognition correlated with informativeness ($r = .14$, $n = 112$, $p = .076$). Text inference also correlated with informativeness of the summary ($r = .18$, $n = 112$, $p < .05$). Action inference correlated with coherence ($r = .24$, $n = 56$, $p < .05$) and ease of understanding ($r = .20$, $n = 56$, $p = .067$).

Correlations between constructs and existing measures showed that some constructs correlated with performance measures and others with satisfaction measures. Inclusion of chosen shots into the summary correlated with object recog-

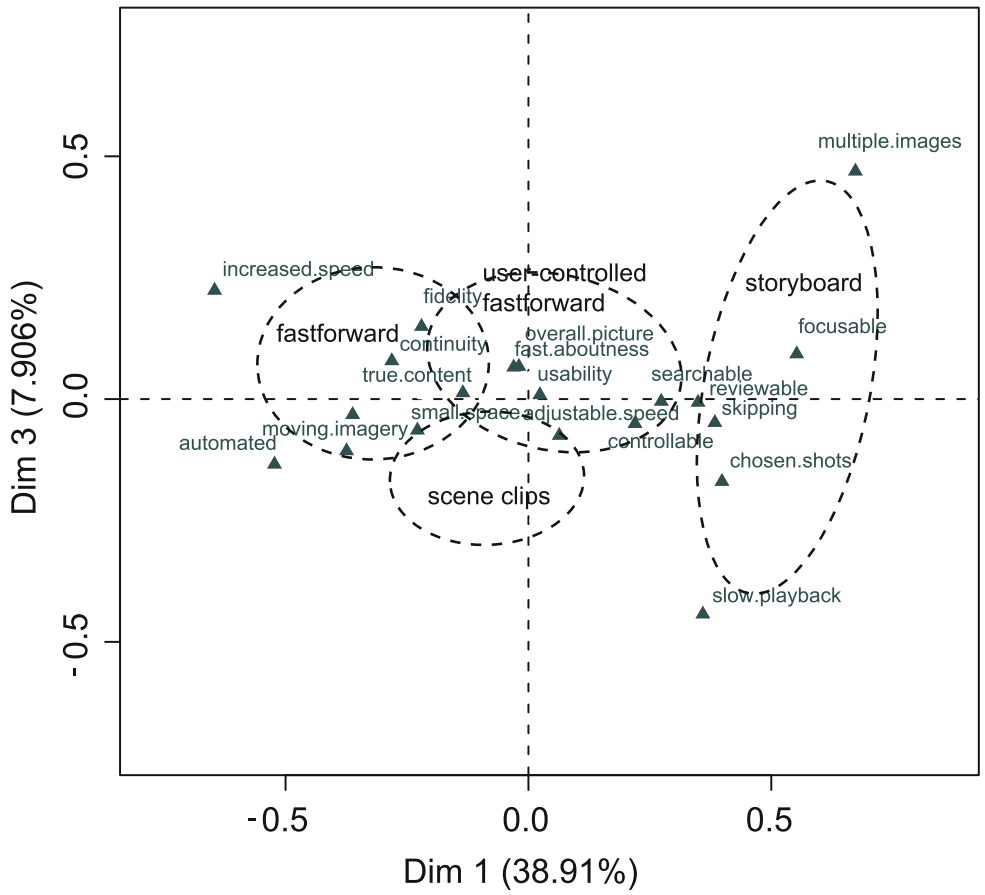

Fig. 2 Correspondence analysis results in dimensions 1 and 3

nition ($r = .38$) and action recognition ($r = .20$). Fidelity, fast aboutness, and overall picture all correlated with various satisfaction measures. The inclusion of moving imagery into the summary correlated negatively with enjoyability ($r = -.25$), but positively with the utility of the summary as measured by the last four satisfaction questions. Automated summary playback correlated with lessened enjoyability ($r = -.23$) but was also correlated with higher text inference scores ($r = .19$). All these correlations are significant at the $p < .05$ (df = 80) level. Various constructs had no significant correlation with existing measures.

4.5 Descriptions of summaries and videos

Free descriptions of the summaries and the corresponding full videos were gathered directly after participants were done viewing them. Descriptions of full videos were on average 26 (SD = 13) words long, while summary descriptions were shorter at an average of 13 (SD = 6.7) words.

Content analysis was performed on the descriptions using an established coding scheme for visual inference from videos [46]. The attributes were related to the visual qualities of the video, objects or people present, and actions or activities conducted, the setting of the video and its overall topic and plot. Furthermore, the attribute *affective* was added to account for expressions of uncertainty in the description. This enabled us to also unobtrusively evaluate the confidence the participants had in their descriptions.

Figure 4 illustrates the shares of all descriptions which included a reference to an attribute. Descriptions of full videos included more geographical and thematic attributes, and descriptions of summaries included more visual attributes and expressions of uncertainty. Across all attributes however, there was no significant difference in the types of attributes mentioned for full videos and summaries ($\chi^2 = 12.95$, df = 9, $p = .17$).

In both the full and summary conditions our participants described the objects present in the video, the actions and activities performed, settings and geographical attributes as well as the theme of the video. To a lesser degree, the descriptions also included references of video plot (e.g., genre, temporal narrative) and visual attributes (shooting angle, colors, graphical elements). There were mentions of uncertainty in the descriptions of both full videos and summaries, more so for the summarized versions.

There was no significant difference in the attribute distribution by summary type ($\chi^2 = 15.14$, df = 27, $p = .97$).

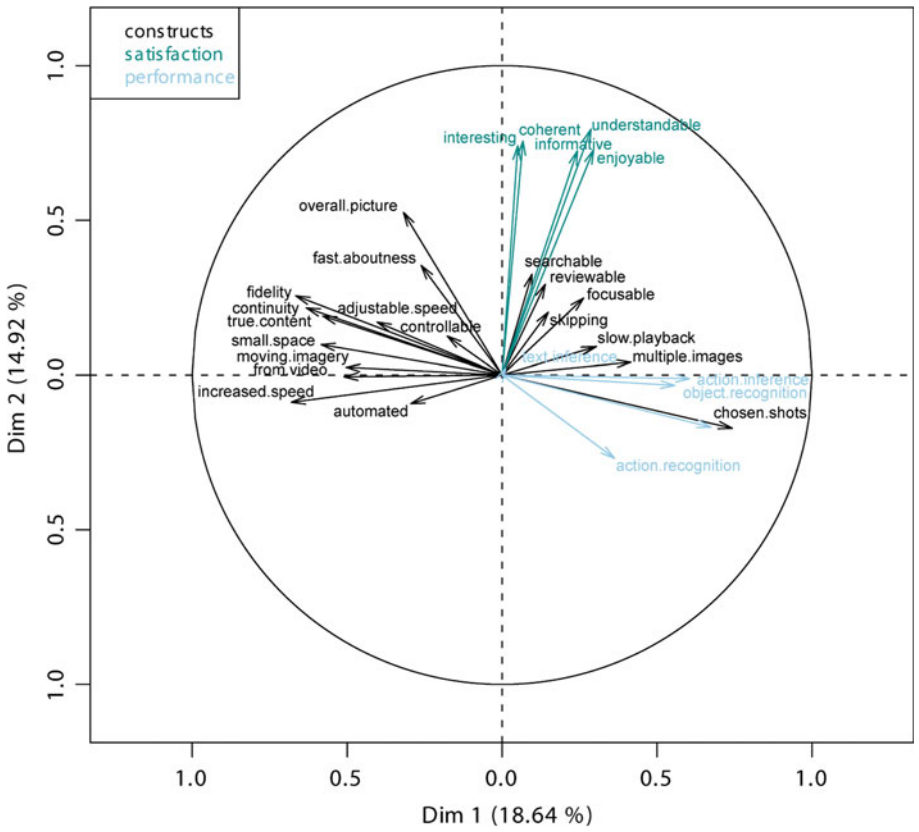

Fig. 3 Correlations between measures visualized in a correlation circle

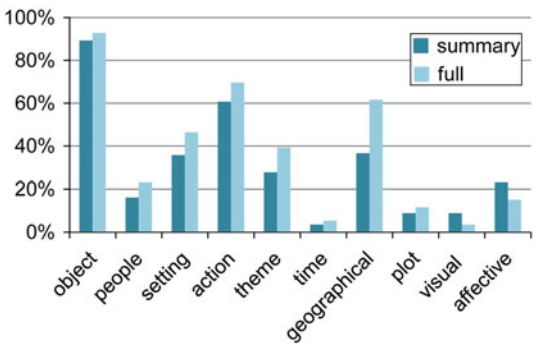

Fig. 4 Percentage of descriptions referencing an attribute

There were most mentions of people in scenes clips and of action in fastforwards. Theme was referenced in 43% of storyboards but only 14% in user-controlled fastforwards. Most mentions of uncertainty occurred in user-controlled fastforwards.

The text inference scores for full videos were on average 7.0 (SD = 1.3) while scores for summaries were on average 4.4 (SD = 1.5). Most of the difference in the scores was due to accuracy in thematic description (average 1.7 points for full vs. .71 for summaries), and thematic detail (1.6 vs. .72). The differences for objects and event information were not as high, neither for accuracy (1.9 vs. 1.6) nor for detail (1.8 vs. 1.4). All these subscore differences were statistically significant at the $p < .001$ level.

5 Discussion

5.1 Summary qualities

There was little difference between still images and respective video clips (storyboard vs. scene clips) or between automated and user-controlled summaries (fastforward vs. user-controlled fastforward). For all summary types, there were participants who preferred that type out of the four compared here. The performance of the summary types differed only for object recognition and text inference. Satisfaction evaluations differed for more measures, e.g., enjoyability, representativeness, and ability to replace original.

Storyboards were on par with dynamic summaries on performance and were most enjoyable but lacked representativeness and ability to replace the original. Most participants preferred moving summaries, stating that storyboards had too low fidelity to be considered good video summaries. Storyboards were thus informative but not indicative in relation to the original video content.

Scene clips supported object recognition and text inference the best and were deemed representative and able to

replace the original video. They were liked due to the clarity in presentation and the normal pace of the moving imagery. Previous findings suggest that static storyboards support image recall better than dynamic summaries [19]. However, in our study scene clips scored the highest in object recognition and the average score was better than for the static counterpart storyboards.

Fastforwards were thought to be most effective in conveying the information and plot of the original video but the sampling rate of every 16th frame was thought to be too fast. This was evident in the satisfaction ratings where fastforwards were less enjoyable than storyboards. Participants noted upon the fact that using a fastforward summary ensures that nothing gets missed in the summarization process.

User-controlled fastforwards were appreciated for the ability to choose focus and review spots. Participants commented that the need to control the summary required focus, possibly distracting them from the viewing, and that it was difficult to find a suitable playback speed. Many participants commented that user-controlled summaries would be more useful in a realistic use setting with unlimited viewing opportunities. On the other hand, some noted that one might take too long to view one, thus possibly rendering the idea of using a surrogate moot.

Elicited evaluation constructs related to summaries' ability to convey the aboutness or overall picture of the video polarized participants resulting in low kappa values. For some, the constant frame rate of the fastforwards represented continuity. Others preferred the normal-paced scene clips whose continuous shots were also useful for object recognition performance. Existing results show that displaying the context of shots makes video summaries more useful in the retrieval process [45]. For our users the issue of gaining an overall, coherent picture of the video contents seemed key.

5.2 Evaluation measures

There were few correlations between the performance and satisfaction measures, making both necessary in evaluation setups. This finding confirms and extends previous results [39]. Performance in object recognition and text inference correlated with subjectively rated informativeness. The introduced action inference measure correlated with subjective evaluations of summary quality as measured by coherence and informativeness. Keyframes and clips could be combined as test stimuli for visual inference measurement in contexts where both object and action inference are important.

Several effects of presentation order were discovered. Scores for text inference were lower for the last summary, possibly indicative of fatigue. Also several satisfaction measures were negatively biased for the fourth and last summary, making it important to counterbalance summary type order in future studies as well.

The evaluation constructs elicited show that users are able to distinguish between multiple qualities of summaries. The correspondence analysis of the construct data shows that both content (modality, selection of shots) and presentation (user controls) of metadata surrogates [2] are recognized and reflected upon by video summary users. Users also explicitly compared the summary to the original video in constructs. The construct evaluations of the storyboards exhibited more variance than those of dynamic summaries, as they were spread out more in the second and third dimensions of the correspondence analysis results. This may be taken to indicate that there was disagreement on whether or not storyboards are easy to use, and whether or not they offer an overall picture of the video.

User-supplied constructs mapped between the performance and satisfaction dimensions in the correlation plot indicating that the constructs reflect both evaluation modes. Similar issues have been raised in free comments in previous studies but were now gathered as reliable rating scale items. The low degree of variance explained by the two-dimensional correlation plot reflects the multidimensional issue of summary quality. Multivariate methods have been used early on in video summary evaluation [13] and have the potential to highlight important dimensions in video summaries.

5.3 Content descriptions

Summarization had no statistically significant effects on the distribution of content attributes in descriptions. While text inference scores calculated based on the descriptions were lower for summaries than full videos, they included the same type of descriptive attributes. Lower scores were mostly due to inaccuracies and deficiencies in thematic descriptions of the content. Summary type had no effect on the overall attribute distribution. However, dynamic summaries had higher shares of action-related attributes and storyboards lent themselves to most thematic descriptions.

In a previous study [46], object and people were the most frequently used attributes of visual inference. As our test videos included less people-centered material, people as an attribute was referenced to a lesser degree in this study. Also, the prevalence of geographical attributes in this study may be attributed to our test videos which dealt with nature in various locations. Still, four out of five top attributes among the two studies were common: object, setting, theme, and action. These attributes seem crucial for making sense of visual video content. The first three are also important in the description of still images [43].

5.4 Limitations

The participants spent roughly a third of their time during the experiment interacting with summary and video content,

and the rest answering questionnaires and describing the content. The inclusion of various measures and the length of the experiment meant that utilizing repeated measures was unfortunately not possible. Owing to the relatively high frame rate, used data on inference measures for fastforwards and user-controlled fastforwards was omitted. In order to maintain the test procedure identical for all summary types, the inference answer option of "did not see but is included in the original video" was displayed for all trials. The inclusion of this non-relevant answer alternative might have created a negative bias toward fastforward types of summaries. Users did select more recognition choices for the fastforwards and user-controlled fastforwards ($\chi^2 = 9.93$, df $= 3$, $p = .02$) so they did not select the inference alternatives in a forced manner despite this issue.

Christel [4] warns against the generalization of results as some types of summaries match certain genres better than others. In this study, satisfaction and, to a lesser degree, performance were affected by the video instance. Differences were related to ease of understanding, enjoyability, informativeness, and text inference based on the summaries. The constructs elicited here might, to some degree, be specific to these video and summary types, thus needing further evaluation. A wider range of summary types need to be investigated, including modalities other than the visual one. For future studies aiming to confirm and complement the results obtained here, different original video lengths and genres need to be included.

6 Conclusions and future work

Video summaries may be utilized as surrogates for full videos or displayed to end-users in conjunction to the full version, requiring study on the distinct contributions of the two. Based on our results on free descriptions of summaries and corresponding full videos, summaries support inference of object and event-related knowledge but full videos are required for accurate and detailed thematic inferences.

It remains important to develop and utilize both performance and satisfaction measures in evaluations of video summaries as these do not correlate in a straightforward manner. Correlation within measure groups (performance, satisfaction) enables the streamlining of evaluation procedures. The constructs obtained here through repertory grid analysis spanned both performance and satisfaction dimensions and showed acceptable inter-rater agreement. They could be used as basis when developing novel evaluation measures for visual video summaries.

For future studies, these user-supplied constructs could be utilized as measures in task-focused studies on video summaries, e.g., in situations of browsing and relevance assessments. This set of constructs could be supplemented by conducting a construct elicitation with another set of users, content, and summary types. Factor analysis could be performed on a larger data set of construct ratings, resulting in a internally consistent measure set.

Acknowledgments This study was supported by the Academy of Finland MOTIVE grant Visual Commons (129357). The authors like to thank the anonymous ECDL and IJDL reviewers for their constructive comments and valuable suggestions. Also appreciated are the efforts at the Interaction Design Laboratory at the School of Information and Library Science, University of North Carolina Chapel Hill for maintaining the Open Video project.

References

1. Anon: Open video digital archive (2010). http://www.open-video.org
2. Balatsoukas, P., Morris, A., O'Brien, A.: An evaluation framework of user interaction with metadata surrogates. J. Inf. Sci. **35**, 321–339 (2009). doi:10.1177/0165551508099090
3. Benini, S., Migliorati, P., Leonardi, R.: Statistical skimming of feature films. Int. J. Digital Multimedia Broadcast. **2010**, 1–12 (2010)
4. Christel, M.G.: Evaluation and user studies with respect to video summarization and browsing. In: Proceedings of SPIE Multimedia Content Analysis, Management and Retrieval, vol. 6073 (2006)
5. Christel, M.G., Lin, W.H., Maher, B.: Evaluating audio skimming and frame rate acceleration for summarizing bbc rushes. In: Proceedings of the 2008 International Conference on Content-based image and video retrieval, CIVR '08, pp. 407–416. ACM, New York, NY, USA (2008). doi:10.1145/1386352.1386405
6. Christel, M.G., Smith, M.A., Taylor, C.R., Winkler, D.B.: Evolving video skims into useful multimedia abstractions. In: Proceedings of the SIGCHI Conference on Human factors in computing systems, CHI '98, pp. 171–178. ACM Press/Addison-Wesley Publishing Co., New York, NY, USA (1998)
7. Cohen, J.: Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. Psychol. Bull. **70**, 213–220 (1968)
8. Corchs, S., Ciocca, G., Schettini, R.: Video summarization using a neurodynamical model of visual attention. In: Proceedings of IEEE 6th Workshop on Multimedia Signal Processing, pp. 71–74 (2004). doi:10.1109/MMSP.2004.1436419
9. de Avila, S.E.F., Lopes, A.P.B., da Luz, Jr. A., de Albuquerque Araújo, A.: Vsumm: A mechanism designed to produce static video summaries and a novel evaluation method. Pattern Recogn. Lett. **32**(1), 56–68 (2011). doi:10.1016/j.patrec.2010.08.004
10. Dillon, A., McKnight, C.: Towards a classification of text types: a repertory grid approach. Int. J. Man-Mach. Stud. **33**, 623–636 (1990). doi:10.1016/S0020-7373(05)80066-5
11. Dumont, E., Mérialdo, B.: Rushes video summarization and evaluation. Multimedia Tools Appl. **48**, 51–68 (2010). doi:10.1007/s11042-009-0374-9
12. Fayzullin, M., Subrahmanian, V.S., Albanese, M., Picariello, A.: The priority curve algorithm for video summarization. In: Proceedings of the 2nd ACM International Workshop on Multimedia databases, MMDB '04, pp. 28–35. ACM, New York, NY, USA (2004). doi:10.1145/1032604.1032611
13. Goodrum, A.A.: Multidimensional scaling of video surrogates. J. Am. Soc. Inf. Sci. Technol. **52**, 174–182 (2001). doi:10.1002/1097-4571

14. Guironnet, M., Pellerin, D., Guyader, N., Ladret, P.: Video summarization based on camera motion and a subjective evaluation method. J. Image Video Process. **2007**, 60245 (2007)
15. He, L., Sanocki, E., Gupta, A., Grudin, J.: Auto-summarization of audio-video presentations. In: Proceedings of the 7th ACM International Conference on Multimedia (Part 1), MULTIMEDIA '99, pp. 489–498. ACM, New York, NY, USA (1999). doi:10.1145/319463.319691
16. Herranz, L., Martinez, J.: A framework for scalable summarization of video. IEEE Trans. Circuits Syst. **20**(9), 1265–1270 (2010). doi:10.1109/TCSVT.2010.2057020
17. Jaimes, A., Echigo, T., Teraguchi, M., Satoh, F.: Learning personalized video highlights from detailed mpeg-7 metadata. In: Proceedings of International Conference on Image Processing, vol. 1, pp. I–133 – I–136 vol.1 (2002). doi:10.1109/ICIP.2002.1037977
18. Johnson, F.C., Crudge, S.E.: Using the repertory grid and laddering technique to determine the user's evaluative model of search engines. J. Doc. **63**, 259–280 (2007). doi:10.1108/00220410710737213
19. Komlodi, A., Marchionini, G.: Key frame preview techniques for video browsing. In: Proceedings of the 3rd ACM Conference on Digital libraries, DL '98, pp. 118–125. ACM, New York, NY, USA (1998). doi:10.1145/276675.276688
20. Kopf, S., Haenselmann, T., Farin, D., Effelsberg, W.: Automatic generation of video summaries for historical films. In: Proceedings of IEEE International Conference on Multimedia and Expo, ICME '04. 2004, vol. 3, pp. 2067–2070 (2004). doi:10.1109/ICME.2004.1394672
21. Li, Y., Narayanan, S., Kuo, C. : Movie content analysis, indexing and skimming via multimodal information. In: Rosenfeld, A., Doermann, D., Dementhon, D. (eds.) Video Mining, Chapt. 5, Kluwer Academic Publishers, Boston (2003)
22. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: Proceedings of the 10th ACM International Conference on Multimedia, MULTIMEDIA '02, pp. 533–542. ACM, New York, NY, USA (2002). doi:10.1145/641007.641116
23. Marchionini, G.: Human performance measures for video retrieval. In: Proceedings of the 8th ACM International Workshop on Multimedia information retrieval, MIR '06, pp. 307–312. ACM, New York, NY, USA (2006). doi:10.1145/1178677.1178720
24. Marchionini, G., Song, Y., Farrell, R.: Multimedia surrogates for video gisting: Toward combining spoken words and imagery. Inf. Process. Manage. **45**, 615–630 (2009). doi:10.1016/j.ipm.2009.05.007
25. Marchionini, G., Wildemuth, B.M., Geisler, G.: The open video digital library: A möbius strip of research and practice. J. Am. Soc. Inf. Sci. Technol. **57**, 1629–1643 (2006). doi:10.1002/asi.v57:12
26. McKnight, C.: The personal construction of information space. J. Am. Soc. Inf. Sci. **51**, 730–733 (2000). http://dx.doi.org/10.1002
27. Mei, T., Yang, B., Yang, S.Q., Hua, X.S.: Video collage: presenting a video sequence using a single image. Vis. Comput. **25**, 39–51 (2008). doi:10.1007/s00371-008-0282-4
28. Money, A.G., Agius, H.: Video summarisation: A conceptual framework and survey of the state of the art. J. Vis. Commun. Image Represent. **19**, 121–143 (2008). doi:10.1016/j.jvcir.2007.04.002
29. Ngo, C.W., Ma, Y.F., Zhang, H.J.: Automatic video summarization by graph modeling. In: Proceedings of the 9th IEEE International Conference on Computer Vision, vol. 2, ICCV '03, p. 104. IEEE Computer Society, Washington, DC, USA (2003)
30. Oppenheim, C., Stenson, J., Wilson, R.M.S.: Studies on information as an asset i: Definitions. J. Inf. Sci. **29**(3), 159–166 (2003). doi:10.1177/016555150302930003
31. Over, P., Smeaton, A.F., Awad, G.: The trecvid 2008 bbc rushes summarization evaluation. In: Proceedings of the 2nd ACM TRECVid Video Summarization Workshop, TVS '08, pp. 1–20. ACM, New York, NY, USA (2008). doi:10.1145/1463563.1463564
32. Over, P., Smeaton, A.F., Kelly, P.: The trecvid 2007 bbc rushes summarization evaluation pilot. In: Proceedings of the international workshop on TRECVID video summarization, TVS '07, pp. 1–15. ACM, New York, NY, USA (2007). doi:10.1145/1290031.1290032
33. Smith, M., Kanade, T.: Video skimming and characterization through the combination of image and language understanding. In: Proceedings IEEE International Workshop on Content-Based Access of Image and Video Database, pp. 61 –70 (1998). doi:10.1109/CAIVD.1998.646034
34. Sundaram, H., Chang, S.F.: Condensing computable scenes using visual complexity and film syntax analysis. Proceedings of IEEE International Conference on Multimedia and Expo **0**, 70 (2001). doi:10.1109/ICME.2001.1237709
35. Sundaram, H., Xie, L., Chang, S.F.: A utility framework for the automatic generation of audio-visual skims. In: Proceedings of the 10th ACM international conference on Multimedia, MULTIMEDIA '02, pp. 189–198. ACM, New York, NY, USA (2002). doi:10.1145/641007.641042
36. Takahashi, Y., Nitta, N., Babaguchi, N.: Video summarization for large sports video archives. In: Proceedings of IEEE International Conference on Multimedia and Expo, vol. 0, pp. 1170–1173. IEEE Computer Society, Los Alamitos, CA, USA (2005). doi:10.1109/ICME.2005.1521635
37. Tan, F.B., Hunter, M.G.: The repertory grid technique: A method for the study of cognition in information systems. MIS Q. **26**(1), 39–57 (2002)
38. Taskiran, C.: Evaluation of automatic video summarization systems. In: Proceedings of SPIE Multimedia content analysis, management and retrieval (2006)
39. Taskiran, C.M., Bentley, F.: Automatic and user-centric approaches to video summary evaluation. In: A. Hanjalic, R. Schettini, N. Sebe (eds.) Multimedia Content Access: Algorithms and Systems, vol. 6506, p. 650607. SPIE (2007). doi:10.1117/12.713913
40. Taskiran, C.M., Pizlo, Z., Amir, A., Ponceleon, D.B., Delp, E.J.: Automated video program summarization using speech transcripts. IEEE Trans. Multimedia **8**(4), 775–791 (2006)
41. Truong, B.T., Venkatesh, S.: Video abstraction: A systematic review and classification. ACM Trans. Multimedia Comput. Commun. Appl. **3** (2007). doi:10.1145/1198302.1198305
42. Tsoneva, T., Barbieri, M., Weda, H.: Automated summarization of narrative video on a semantic level. In: Proceedings of the International Conference on Semantic Computing, pp. 169–176. IEEE Computer Society, Washington, DC, USA (2007). doi:10.1109/ICSC.2007.16
43. Westman, S., Laine-Hernandez, M., Oittinen, P.: Development and evaluation of a multifaceted magazine image categorization model. J. Am. Soc. Inf. Sci. Technol. (2010). doi:10.1002/asi.21463
44. Wildemuth, B.M., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., Gruss, R.: How fast is too fast?: Evaluating fast forward surrogates for digital video. In: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries, JCDL '03, pp. 221–230. IEEE Computer Society, Washington, DC, USA (2003)
45. Wildemuth, B.M., Russell, T., Ward, T., Marchionini, G., Oh, S.: The influence of context and interactivity on video browsing. Tech. Rep. SILS Technical Report 2006-01, University of North Carolina, School of Information and Library Science (2006). http://www.ils.unc.edu/ils/research/TR-2006-1.pdf
46. Yang, M., Marchionini, G.: Deciphering visual gist and its implications for video retrieval and interface design. In: CHI '05 extended abstracts on Human factors in computing systems, CHI '05, pp. 1877–1880. ACM, New York, NY, USA (2005). doi:10.1145/1056808.1057045
47. Yang, M., Wildemuth, B.M., Marchionini, G., Wilkens, T., Geisler, G., Hughes, A., Gruss, R., Webster, C.: Measures of user per-

formance in video retrieval research. Tech. Rep. SILS Technical Report 2003-02, University of North Carolina, School of Information and Library Science (2003). http://www.ils.unc.edu/ils/research/TR-2003-02.pdf

48. Zhang, X., Chignell, M.: Assessment of the effects of user characteristics on mental models of information retrieval systems. J. Am. Soc. Inf. Sci. Technol. **52**, 445–459 (2001). doi:10.1002/1532-2890