What are the Most Eye-catching and Ear-catching Features in the Video? Implications for Video Summarization

Yaxiao Song
University of North Carolina at Chapel Hill
100 Manning Hall, CB #3360
Chapel Hill, NC 27599, USA
yaxiaos@email.unc.edu

Gary Marchionini
University of North Carolina at Chapel Hill
100 Manning Hall, CB #3360
Chapel Hill, NC 27599, USA
march@ils.unc.edu

Chi Young Oh
University of North Carolina at Chapel Hill
100 Manning Hall, CB # 3360
Chapel Hill, NC 27599, USA
cyoh@email.unc.edu

ABSTRACT
Video summarization is a mechanism for generating short summaries of the video to help people quickly make sense of the content of the video before downloading or seeking more detailed information. To produce reliable automatic video summarization algorithms, it is essential to first understand how human beings create video summaries with manual efforts. This paper focuses on a corpus of instructional documentary video, and seeks to improve automatic video summaries by understanding what features in the video catch the eyes and ears of human assessors, and using these findings to inform automatic summarization algorithms. The paper contributes a thorough and valuable methodology for performing automatic video summarization, and the methodology can be extended to inform summarization of other video corpuses.

Categories and Subject Descriptors
H.3.1 [**Information Storage And Retrieval**]: Content Analysis and Indexing – *abstracting methods, indexing methods.*

General Terms
Measurement, Performance, Design, Experimentation, Human Factors.

Keywords
Video summarization, visual salience, audio salience.

1. INTRODUCTION
With the rapid growth in computing technology and explosive proliferation of digital videos online, it is imperative to give web users effective summarization and skimming tools to facilitate finding and browsing videos.

Video summarization, a mechanism for generating short summaries of videos, has generated substantial research and development effort that aims to aid users in browsing and retrieving relevant videos from large video collections. Although video summarization techniques have been well established for video genres such as News video and sports video, relatively few techniques focused on instructional documentary video. This paper examines a set of video summaries created by multiple human assessors for a corpus of instructional documentary video – the NASA K-16 Science Education Programs – and uses statistical procedures to characterize the most eye-catching and ear-catching features in these manually generated video summaries. The paper provides insights into what features automatic algorithms should attempt to extract when performing automatic video summarization. The valuable methodology used in this paper can be extended to inform summarization of other video corpuses.

The rest of the paper is organized as follows. Section 2 reviews existing work in automatic video summarization. Section 3 discusses the procedure of acquiring manually generated video summaries as well as feature indexing for the summaries. Section 4 analyzes the indexing results, and discusses the most important features that draw people's attention when watching the video. Section 5 concludes the paper.

2. RELATED WORK
There has been a great deal of work on video summarization [1, 4, 16, 24]. In general, video summarization techniques can be classified into three different categories. First, video summarization can be performed by looking at the text transcripts of the video, and using text summarization techniques to generate text summaries for the video. Then the video segments corresponding to the text summaries can be selected to form a video summary. Secondly, video summaries can be extracted by processing the audio stream, and detecting audio features such as speech emphasis, pitch, excitement level, and so on. Thirdly, video summaries can be created by exploiting both high-level and low-level visual features.

Intuitively, videos can be time compressed by speeding-up, or by pause shortening or removal. For example, [2] selectively shortened and removed pauses in the speech audio. However, time compression via speeding-up and/or pauses shortening or removing, even when used together, can hardly lead to compaction rates[1] of more than 2:1 [2], and often increases users' cognitive load greatly. In many real-life video retrieval or audio/video summarization applications, a compaction rate of 10:1 or above is desirable.

To further reduce the playback time of the audio, *skimming* techniques can be used. A simple and straightforward method for

WWW 2010, April 26-30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

[1] Compaction rate refers to ratio of full-length play time to the human time to consume the summary as distinguished from compression rate that refers to real-time to machine transfer time.

creating video summaries is by systematic subsampling: Extracting fixed-duration excerpts of the original video at fixed intervals. For example, select the first 10 seconds of the video, skip the next 50 seconds, select another 10 seconds, and skip another 50 seconds, and so on and so forth. Then the selected 10-second segments can be joined together to form a video summary and played back to the viewer at the original frame rate, which yields a compaction rate of 6:1. Although the summaries created by systematic subsampling are likely to exclude some important segments, they are easy and inexpensive to implement. Thus, systematic subsampling has often been adopted as the default or baseline method in evaluating other automated video summarization techniques [6].

Some videos, such as instruction or presentation videos, are dominated by a talking head, and the important information is mostly contained in the audio stream. [4] used a Hidden Markov Model (HMM) to recognize speech emphasis to create summaries for natural, conversational speech such as recorded telephone or interview conversations. Even for sports videos, where the visual play of actions are more attractive to the viewers than the audio, the important visual events are often accompanied by great audience excitement and sharp increase in the audio volume, hence video summarization can be performed by detecting sudden changes in excitement levels [3, 18].

An intuitive and practical approach to summarizing videos like news programs, instruction or presentation videos, and teleconferences, is to analyze the speech transcript [1, 5, 19, 26]. Closed captions are readily available for most broadcast videos, such as the news programs. For other videos where closed captions are not available, automatic speech recognition (ASR) techniques can be used to generate the speech transcript. ASR is clearly more useful for retrieval where term frequency based bag of word techniques are used than for human-consumable summarizations that demand smoothly connected text.

Another common strategy in summarizing videos is to segment the videos and extract one or more keyframes from each segment, then concatenate the keyframes to form static or dynamic summaries. The keyframe extraction is generally determined based on visual features, such as clustering using color histograms [9, 10, 14, 24, 27]. With the extracted keyframes, static summaries such as the storyboard can be created. Dynamic summaries also can be created. A simple and straightforward method for generating skims is to include the contiguous neighborhood frames of the selected keyframes and concatenate them together to form continuous segments. Note that care must be taken at spoken sentence boundaries, as users find it annoying when audio begins in the middle of a sentence or phrase [26].

In some videos (e.g., News, sports, video rushes), important visual and speech materials are often repeated multiple times in adjacent shots, which creates a certain level of redundancy in the video. The redundancy phenomenon has been incorporated in many video retrieval approaches. For instance, [29] investigated visual redundancy between two adjacent shots in the video to calculate the transitional probability of a shot being visually relevant given that the previous shot was visually relevant. [15] proposed a method for automatically summarizing unedited video rushes which have a lot of repeated shots by removing unusable shots and clustering the remaining frames using k-means clustering to identify repeated shots.

Some summarization algorithms exploit specific domain knowledge for certain genres of videos. For example, interesting events in the soccer games are often limited to goals and goal attempts, which only occupy a small portion of the entire game. Incorporating the specific domain knowledge about these important events in the sports videos, the long program can be condensed into a compact summary [3, 7, 18]. News programs are usually composed of alternative concatenation between anchor shots and news segments. [20] proposed an algorithm for summarizing news videos by retrieving the anchor audio and summarizing the visual parts of the news segment by classifying shots into special and normal events. Then the anchor audio is overlaid with the visual summaries for news sequences to form a summary, allowing the viewers to understand the story headlines as well to perceive motion activity of the story.

Most of the video summarization techniques discussed above focus on processing single data stream, either text, or audio, or visual, with a few exceptions [5, 13, 18, 20]. By combining approaches from more than one modality, video summarization has the potential to be performed with better coverage, context, and coherence. For example, [21] developed the MoCA video abstracting system, which produces movie trailers automatically. The system detects special events in the movie, such as faces, text in the title sequence, and close-up shots of the main actors from the visual features based on a few heuristics, and identifies events like explosions and gunfire using audio parameters (e.g., loudness, frequencies, pitch, frequency transition etc.). Then the text, video clips, and audio clips containing those events are selected and assembled by adding dissolves and wipes to make the final movie trailer sequence. [11] employed *similarity analysis* techniques to automatically extract informative audio excerpts, and augment the visual surrogates (i.e. storyboards) with the audio excerpts to create so-called "Manga summaries". [8] described a multi-modal scheme for automatically summarizing meeting videos based on *audio and visual event detection* together with *text analysis*. [6] designed video skimming techniques that used (1) tf-idf measure and audio analysis based on audio amplitude, and (2) audio analysis combined with image analysis based on face/text detection and camera motion.

Although video summarization techniques have been well established for a variety of video genres such as News video, presentation video, sports video, and video rushes, relatively few techniques focused on summarizing instructional documentary video. For example, the NASA K-16 Science Education Programs are a collection of instructional documentary videos, which aim at K-16 students, teachers, and parents. Some of the programs are grade school programs that include hands-on activities, web activities and resources, and some are technology-based programs for lifelong learners. Abbreviated and effective video summaries of these programs are needed for the teachers to make lesson plans as well as for general video users to browse the program collections and make relevance judgments. Unfortunately, the instructional documentary videos may not have visual or audio redundancy as in video rushes, nor may they have explicit domain knowledge or heuristics as in sports videos based on which video summarization can be effectively and automatically performed, hence video summarization remains an open and challenging research area for these videos.

Borrowing terminology developed for text summarization evaluation, [23, 26] classified video summary evaluation methods

into two categories: *intrinsic* and *extrinsic*. In intrinsic evaluation methods, the quality of the generated summaries may be judged directly, based on the user judgment of *fluency* of the summary, *coverage* of key ideas of the source material, or *similarity* (e.g., fraction of overlap) to ground truth summaries prepared by human experts. In extrinsic evaluation methods, the video summaries are evaluated in terms of their impact on the performance for a specific information retrieval task. Since automatically generated summaries are often evaluated by comparing them to manually generated ground truth summaries, in order to derive good summarization algorithms, it is important to first learn how people extract summaries from the full videos. This paper examines the video summaries for the instructional documentary videos created by multiple human judges and how good these summaries are rated by other human assessors, and uses statistical procedures such as Pearson's Chi-square test and Ordinal Logistic Regression to identify the most eye-catching and ear-catching features in these manually generated video summaries which make the summaries informative and salient.

3. PROCEDURE

3.1 Phase 1: Creating the video summaries

A group of 12 human judges were recruited to manually extract the most salient segments from videos to form video summaries for a set of four instructional documentary videos. The 12 judges were 2 senior undergrad students, 9 master students, and 1 faculty member in a digital video class, among whom 3 were females and 9 were males. All the judges were familiar with video editing tools but none had experience with video indexing.

The videos were selected from the NASA Connect video collection. The titles of the videos are:

- NASAConnect: Virtual Earth
- NASAConnect: Proportionality-Modeling The Future
- NASAConnect: Wired For Space
- NASAConnect: Dancing In The Night Sky

Each video is about 28.5 minutes. We provided three viewing conditions for each video – audio only, visual only, and combined (i.e., with both visual and audio streams). FFmpeg was used to strip the visual or audio streams from the full videos to create the audio only and visual only versions of the video. In particular, the visual only versions were played back at the same frame size and frame rate as the full videos, and the audio only versions were played at the same speed as the full videos.

Each judge was assigned three different videos out of the four videos, and the three videos are of different conditions: one audio only, one visual only, and one combined. After viewing each assigned video, each judge extracted the most salient segments from the video to be included in the video summary. Specific instructions on extracting the segments were provided to the human judges as follows:

"You will be assigned three media streams. One will be the soundtrack of a 30 minute video; one will be the visual track of a different 30 minute video; and one will be a different full 30 minute video. Use your favorite editor to experience the stream and select the five most salient extracts that summarize the gist, recording the time stamp in the original stream for each one. The extracts (surrogates) should be about 5 to 10 seconds long. Save the surrogates and time stamps and write a short paragraph that describes your selection strategy. Repeat this for the other two streams."

Note that each human judge selected 5 extracts for the audio only video, 5 extracts for the visual only video, and 5 extracts for the full video condition. In total, the 12 judges extracted 178 individual segments for the three viewing conditions of the 4 videos. The total number of segments is not 12(*judges*) * 5(*segments*) * 3(*viewing conditions*) = 180 because one person selected 3 segments instead of 5 segments for one of the assigned videos (i.e., "Dancing In The Night Sky", viewing condition: combined). Overall, there are 58 segments with both audio and visual, 60 segments with audio only, and 60 segments with visual only.

It takes about 3 minutes to play all the 15 reference (ground-truth) summary segments for each 30-min video, yielding a compaction rate of about 10:1. In most video retrieval systems, a higher compaction rate (e.g., more than 30:1) is desirable, hence the reference summaries need to be further compacted to serve as good video summaries or video surrogates.

The intuitive approach to select the best segments for the video summary is to select the reference summary segments that express the majority of the judges' opinions. For each video, the segments selected by the judges were first sorted by their time stamps, and segments overlapped by more than one judge's selections were identified. Here, an overlap is identified if some segments selected by different judges had more than 3 seconds in common, and the union of the overlapped segments was included in the gold standard surrogates. For example, for the visual and audio combined version of "NASAConnect: Proportionality-Modeling The Future", Segment "30s - 40s" was selected by one judge, and Segment "33s - 45s" was selected by another judge. Thus, the union segment "30s – 45s" was included in the gold standard surrogate.

However, there were only a small number of segments that overlap in the sets of 15 manually exacted segments for each video each version (the chance of 3 people selecting the same 10 second segment at random from 1710 seconds of video is less than 2 in 10 billion). Studies of overlap in professional indexer term assignment provide a severe upper bound for summarization. [12] found that overlaps for medical subject headings selected by two indexers was only 33% and free text overlaps are even lower (e.g., pick any random two tags assigned to images on Flickr). Figure 1 (a) - (c) show the overlaps among the summary segments selected by the judges for the video "NASAConnect: Virtual Earth" under the audio only, visual only, and combined conditions respectively. For instance, the three horizontal bars with 5 red blocks each in Figure 1 (a) represent the summary segments selected by three assessors for the audio only version of the video. We observed an overlap of segments among all three assessors at the beginning of the video, and an overlap of segments between assessor 2 and assessor 3 around the end of the first quarter of the video. For the visual only version of the same video, there was one small overlap of segments between assessor 1 and assessor 3 as shown in Figure 1 (b). And for the combined version, there was one small overlap of segments between assessor 2 and assessor 3, as shown in Figure 1 (c).

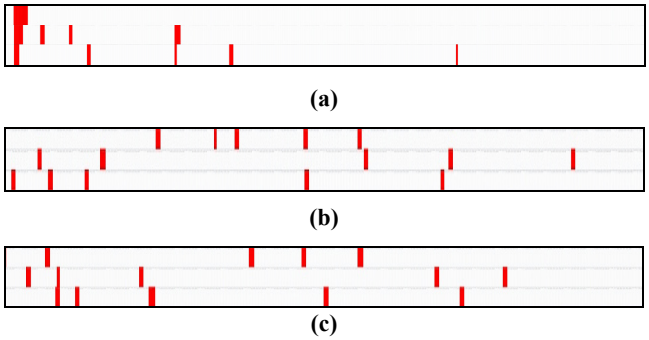

Figure 1. Overlaps among summary segments: (a) overlaps in the audio only version, (b) overlaps in the visual only version, (c) overlaps in the full video.

To investigate the audio, visual, and subjective features that occur in the selected extracts, and to determine a principled way to select the best extracts from the set of 15, two additional phases of evaluation were conducted.

3.2 Phase 2: Indexing the video summaries

Four human assessors were recruited to index the video summaries created by the group of 12 human judges in Phase 1. Note that none of the judges in Phase 1 participated in this assessment in Phase 2. The authors of this paper created an indexing template containing a list of features as an indexing rubric to be used by the 4 human assessors. Audio and visual features were binary choices (i.e., whether the assessor heard/saw them or not) and the subjective extract types were forced choices of which type best characterized the extract.

- Six audio features: music, single human voice, multiple people talking, proper nouns (e.g., human names, object names, location names), natural sound, and artificial sound.
- Eight visual features: text (superimposed names, locations), faces, graphics & logos, graphs, equations, animals, human built artifacts, and natural scenes.
- Three extract intention options: indicative, descriptive, or cannot tell? (choose only one).
- Four extract functions: context, definition, example / illustration, or summary / overview? (choose only one).

Each human assessor watched and/or listened to each of the 178 summary segments carefully, and marked the corresponding items in the template. For example, if a face appears in a summary segment, "face" is scored as "1" for the segment; otherwise, it is scored as "0". Furthermore, if a summary segment is about "example/illustration", that segment is marked as "1" in "example/illustration" category, and "0" in the other three categories.

Thus, the indexing result data on audio or visual features are *binary* responses. And the indexing results on extract intentions or extract functions are *nominal* responses, as there is no natural order among the 3 or 4 response categories.

One video was done by the four assessors together as a group to establish a baseline of rating for the features. Not surprisingly, there was generally good consensus on the visual and audio features and considerable debate on the subjective types, and no requirement was made to reach consensus across the four assessors but all were encouraged to be consistent within their own ratings. After working on one video together as a group, each assessor worked on the rest videos on him or herself. Phase 2 took assessors about 5 hours to complete.

3.3 Phase 3: Rating the manually extracted video summary segments

After the completion of Phase 2, the same group of human assessors in Phase 2 was asked to complete Phase 3. In this phase, each of the four assessors was assigned one of the four NASA Connect videos used in Phase 1. Each assessor watched one 28.5 minute full video with both audio and visual streams. After watching the assigned videos, the assessors were asked to watch and/or listen to the corresponding summary segments extracted by the judges in Phase 1, and rate each segment on a 1-7 scale (where 1 is very bad, 2 is bad, 3 is somewhat bad, 4 neutral, 5 is somewhat good, 6 is good, and 7 is very good). Each assessor rated the summary segments extracted from all three viewing conditions for the assigned video, and it took the assessors about one hour to complete their ratings in Phase 3.

4. RESULTS AND DISCUSSION

This section describes how the collected data were analyzed and discusses the statistical analysis results. First, a multi-rater variation of free-marginal kappa was used to measure the inter-rater reliability of the four sets of indexing results collected in Phase 2. Pearson's Chi-square tests of independence were used to evaluate whether the percentage of the summary segments having a certain feature differs across the viewing conditions, i.e., audio only, visual only, and both. Finally, Ordinal Logistic Regression analyses were conducted to model the relationship between the summary goodness and various audio or/and visual features.

4.1 Inter-rater Reliability

Inter-rater reliability (also known as inter-rater agreement) is the extent to which two or more raters (or coders) agree. Inter-rater reliability addresses the consistency of the implementation of a rating system.

To measure the inter-rater reliability of the four sets of indexing results in Phase 2, a multi-rater variation of free-marginal kappa was used. Values of kappa can range from -1.0 to 1.0, with -1.0 indicating perfect disagreement below chance, 0.0 indicating agreement equal to chance, and 1.0 indicating perfect agreement above chance. According to [17], kappa values in the 0.81-1.0 range indicate almost perfect agreement, those in the 0.61-0.8 range indicate substantial agreement, those in the 0.41-0.6 range indicate moderate agreement, and those in the 0.21-0.4 range indicate fair agreement.

Kappa values were computed for each indexing category (i.e., visual features, audio features, extract intention, and extract function) over each video viewing condition (i.e., audio, visual, or combined), and their averages were calculated by category and viewing condition. These values are shown in Table 1.

These results show that the reliability of indexing results across the four assessors was relatively high on the visual features (0.73) and audio features (0.78), and was fair to moderate in the more

subjective categories of extract intention (0.42) and extract function (0.37).

Kappa values in almost all indexing categories were higher in the "combined" condition than in the "audio only" or "visual only" conditions. These results coincide with our intuitions that people will agree more easily about specific audio and visual objects than about subjective judgments and provide one kind of face validity for the overall summarization value of the process used in phase 1. Because the reliability of indexing results across the four assessors was satisfactory, we were confident to use all the data from four assessors in the following analyses.

Table 1. Inter-rater reliability (Kappa) of the indexing results.

Indexing Category		Video viewing condition			Average
		Audio	Visual	Combined	
Visual features	Text		0.55	0.55	0.55
	Faces		0.66	0.79	0.73
	Graphics & logos		0.56	0.54	0.55
	Graphs		0.93	0.78	0.86
	Equations		0.84	0.98	0.91
	Animals		0.97	0.98	0.98
	Human built artifacts		0.47	0.63	0.55
	Natural scenes		0.70	0.70	0.70
	Average		0.71	0.75	0.73
Audio features	Proper noun	0.57		0.51	0.54
	Single human voice	0.86		0.84	0.85
	Multiple human voice	0.87		0.86	0.87
	Music	0.80		0.93	0.87
	Natural sound	0.91		0.91	0.91
	Artificial sound	0.63		0.70	0.67
	Average	0.77		0.79	0.78
Extract intention		0.50	0.24	0.51	0.42
Extract function		0.39	0.30	0.43	0.37
Grand Average		0.69	0.62	0.73	0.69

4.2 Eye-catching visual features

In Phase 1 of this study, video summaries were manually created by multiple human judges in three video viewing conditions: audio only, visual only, and combined. In Phase 2 of this study, each of the summary segments selected in Phase 1 was manually indexed by multiple assessors according to the pre-defined rubric. To investigate whether the distributions of the visual features differ for the visual only and the combined conditions, Pearson's Chi-square (X^2) tests of independence were conducted to compare the indexing results based on the percentages of the summary segments having certain visual features in the visual only summaries and the video summaries containing both visual and audio. For example, does the distribution of having "text (names, locations)" differ between the visual only and the combined conditions? Below are the null and alternative hypotheses for the Chi-Square tests of independence tested in Section **4.2**:

Ho: Whether a summary segment has a certain visual feature or not is independent of the video viewing conditions.

Ha: Whether a summary segment has a certain visual feature or not is related to the video viewing conditions.

For example, Table 2 is the contingency table reporting the numbers and percentages (i.e., the numbers in the parentheses) of segments having or not having the visual feature "text (names, locations)" for the visual only and the combined video viewing conditions. The data show if people viewed the visual only versions of video, it is more likely that they would select a summary segment with "text (name, location)" than if they viewed the video with both audio and visual streams. The Pearson's chi-square value is 29.237 (df = 1), which is statistically significant at the conventionally accepted significance level of $\alpha = 0.05$. Thus, we can reject the null hypothesis, and conclude that whether a manually selected salient summary segment has "text (name, location)" features or not is significantly related to the viewing condition, i.e., visual only summaries contain more "text" than "combined" summaries with both audio and visual.

Table 2. Contingency Table of "Text (names, location)" for Visual only and Combined conditions.

		Visual	Combined	Total
Text (names, location)	**1**	177 (73.75%)	115 (49.57%)	292 (61.86%)
	0	63 (26.25%)	117 (50.43%)	180 (38.14%)
	Total	240 (100%)	232 (100%)	472 (100%)

The contingency tables were computed for the remaining 7 visual features for the visual only and the combined conditions. These tables are not included in this paper due to the space limit, but Figure 2 summarizes the differences in the percentages of each of the eight visual features in the visual only and the combined summary segments.

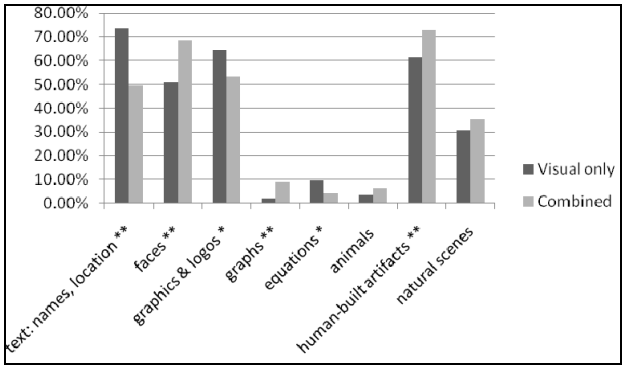

Figure 2. Different Distributions of Visual features in the Visual only and the Combined Summary Segments.

The probability of having a certain visual features is statistically reliably related to the viewing conditions – visual or combined – for six out of the eight visual features (note the asterisks in Figure 2, where * denotes significant at 0.05 probability level, and ** denotes significant at 0.001 probability level). In other words, we observe significantly reliable differences between the visual and the combined conditions for these six visual features. Note that the probability of a summary segment having visual features such as "text (names, location)", "graphics & logos", or "equations" in the visual only condition is statistically significantly higher ($\alpha =$

0.05) than the probability of having these features in the visual and audio combined condition.

The results make sense. According to previous studies [22, 25, 28], the text, visual, and audio modalities of video surrogates have different roles and effects in video sense-making. The text or audio carry semantic information in video and help people understand the content of the video, while images add affective and confirmatory value. When people watch videos with just the visual streams, they have to pay more attention to the visual features that also provide semantic information, such as text overlays, graphics & logos, and equations. Features like "faces" are informative when accompanied by human voices, which exist in the "combined" condition but not as much in the "visual only" condition. Thus, it is reasonable that significantly reliably more summary segments in the "combined" condition have "faces" than segments in the "visual only" condition. Natural scenes and animals may be easily discerned visually and not need textual/verbal accompaniments.

4.3 Ear-catching audio features

Pearson's Chi-square (X^2) statistics were computed for the indexing results on the audio features for the audio only summaries and the video summaries containing both visual and audio. The null and alternative hypotheses for the Chi-Square tests of independence are as follows:

Ho: Whether a summary segment has a certain audio feature or not is independent of the video viewing conditions.

Ha: Whether a summary segment has a certain audio feature or not is related to the video viewing conditions.

Figure 3 presents the different percentages for the six audio features by audio only and combined conditions. Ear-catching features are not as distinctive across the two conditions. The probability of having a certain audio features is statistically reliably related to the viewing conditions – audio only or combined – for only 3 out of the 6 audio features (note the asterisks in Figure 3, where * denotes significant at 0.05 probability level).

Music is almost ubiquitous in these particular educational videos that aim to motivate middle school students. Music is played softly in the background even when a narrator or character is speaking (and the volume increases during scene transitions). Sound effects and natural sounds (e.g., waves at a beach) are used during transitions and sometimes as background during voiceovers but nowhere as frequently as music.

No statistically reliable differences were found for single human voice, which was quite important for both conditions. And no statistically reliable differences were found for natural sounds and artificial sounds between the audio only and combined conditions. Proper nouns were more important in the combined conditions, perhaps because they are tied visually to the person or place whereas they stand alone in the audio only condition. Multiple human voices tend to not be selected often in either condition; however, they were more significant to the audio only condition than to the combined condition, perhaps because there was less alternative information to catch the judges' ears.

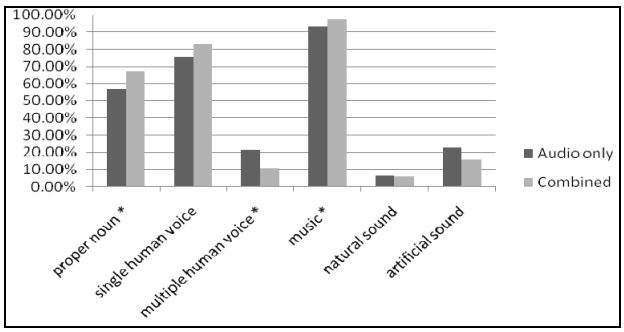

Figure 3. Different Distributions of Audio features in the Audio only and Combined Summary Segments.

4.4 Extract intention

Pearson's Chi-square (X^2) statistics were computed for the indexing results on the extract intention of the video summaries in all three video viewing conditions: audio only, visual only, and combined. The null and alternative hypotheses for the Chi-square tests of independence are as follows:

Ho: The extract intentions of the segments are independent of the video viewing conditions.

Ha: The extract intentions of the segments and the video viewing conditions are related.

Figure 4 shows the percentages of segments marked by the human assessors as indicative, descriptive, or cannot tell. Note that 7 segments with missing values (i.e. the assessors did not provide responses) were excluded from the analysis.

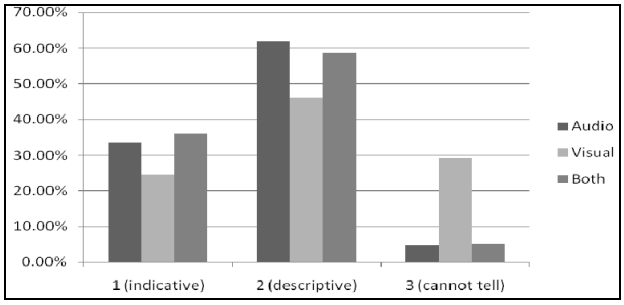

Figure 4. Extract Intention by viewing condition.

Figure 4 shows that the summary segments selected by human judges have a higher probability of being descriptive or indicative if only the audio streams were available or if both audio and visual were available than if only the visual streams were available. When only the visual streams were available to the human judges, they selected more summary segments that they "cannot tell" whether they were indicative or descriptive than if only the audio or both audio and video streams were available.

The Pearson's chi-square value for the comparisons for the extract intention is 82.4393, which is statistically significant at α = 5.29723E-17 for df = 4. Thus the differences of the distribution of the extract intention of the segments among the three video viewing conditions are significantly reliable, and the chi-square test has rejected the null hypothesis of equal population proportions. Next, we ran multiple post hoc pair-wise contrasts to determine which (if any) pair-wise proportions in the three viewing conditions resulted in the significant differences.

Table 3 shows the results of the pair-wise contrasts between any two pairs out of the three viewing conditions. Note that the proportions of indicative or descriptive segments selected by human judges are statistically significantly different for audio only and visual only conditions, and are also statistically significantly different for the visual only and the combined conditions, while the proportions of the indicative or descriptive segments are very comparable for the audio only and combined conditions. These results are consistent with the results for the eye-catching and ear-catching features in sections 4.2 and 4.3. Clearly, visual information contributed less evidence for determining overall function for a summary than audio information alone or audio in conjunction with a visual channel.

Table 3. Pair-wise comparisons of "Extract intention".

	Audio vs. Visual	Visual vs. Combined	Audio vs. Combined
Pearson X^2	50.92	47.48	0.50
df	2	2	2
Asymp. Sig. (2-sided)	8.76967E-12	4.9046E-11	0.780

4.5 Extract function

The extract functions of the video summaries were investigated for all three video viewing conditions: audio only, visual only, and combined. The null and alternative hypotheses for the Pearson's Chi-square tests of independence are as follows:

Ho: The extract functions of the segments are independent of the video viewing conditions.

Ha: The extract functions of the segments and the video viewing conditions are related.

The analysis of how assessors categorized the extract function required a number of different pair-wise comparisons because there were four possible categories for every judgment. Figure 5 shows the percentages of segments marked as context, definition, example/illustration, or summary/ overview for each of the three summary conditions. Note that we recoded these four variables into a new variable called "function", where "context" →1, "definition" → 2, "example/illustration" → 3, and "summary/ overview" → 4.

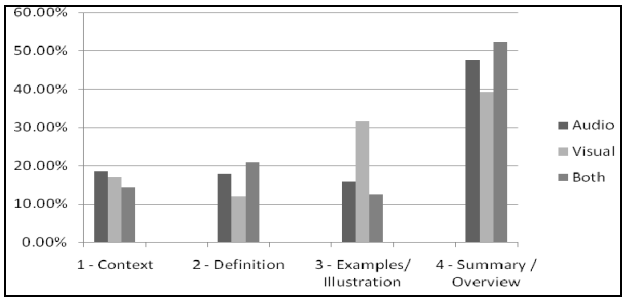

Figure 5. Extract Function by viewing condition.

The overall Pearson's Chi-square value for the extract function among all three viewing conditions is 35.91433526, which is significant at $\alpha = 2.86434E\text{-}06$ for df = 6. As with the analysis of the summary function, pair-wise comparisons were conducted between any two video viewing conditions (see Table 4). Only the contrasts between visual only and combined video viewing conditions are statistically significantly reliable at a 0.05 level.

One important result that jumps from Figure 5 is the strong coding of visual only summaries into the "example/illustration" function type. This result agrees with previous results that suggest that visual features are useful to augment and illustrate rather than carry the primary intentionality in video [28] as well as with intuitions about video sense making.

Table 4. Pair-wise comparisons of "Extract function"

	Audio vs. Visual	Visual vs. Combined	Audio vs. Combined
Pearson X^2	2.6483054	8.3170162	1.970944
df	3	3	3
Asymp. Sig. (2-sided)	0.4490838	0.0398947	0.578459

4.6 Important features and summary goodness

It is important to understand how the different features contribute to a good video summary. In Phase 3 of the study, human assessors were asked to watch full videos and then rate the summary segments on a 1-7 scale (where 1 is very bad, and 7 is very good).

Ordinal Logistic Regression analyses were conducted to model the relationship between the summary goodness and various audio or visual features. The response variable is the evaluation score (i.e., ordinal responses, 1-7), and the predictors are a number of audio or visual features (i.e., binary values, 0 or 1).

4.6.1 Goodness of the audio only summary and the Audio features

Since we identified six audio features, the ordinal logistic model for the relationship between the ratings of the audio only summary segments and the six independent variables (i.e., the audio features) can be written as:

$$\ln(\theta_j) = \alpha_j + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_6 X_6,$$

where j goes from 1 to 6 (i.e., the number of categories minus 1, because the ratings were on a 1-7 scale). X_1, X_2,..., X_6 are independent variables described as follows:

X_1: music (0 or 1),
X_2: single human voice (0 or 1),
X_3: multiple people talking (0 or 1),
X_4: proper nouns (0 or 1),
X_5: natural sound (0 or 1),
X_6: artificial sound (0 or 1),

and $\beta_1, \beta_2, \ldots, \beta_6$ are their corresponding coefficients.

Note that each logit $\ln(\theta_j)$, where $j = 1, 2, \ldots, 6$, has its own α_j terms but the same set of coefficients β_k, where $k = 1, 2, \ldots, 6$, because the effect of each independent variable is the same for different logit functions. We are mainly interested in the values of β_k, and the values of α_j are often not of much interest.

Table 5 summarizes the β coefficients of the regression model for the audio only summary segments, where the response variable is the evaluation scores and the predictors are the 6 audio features.

Table 5. Coefficients of the regression model for audio only summary segments.

	β	Std. error	Sig.
Proper noun (city, person, place name)	.194	.5119	.705
Single human voice	3.485	1.5142	.021
Multiple human voice	1.959	1.5359	.202
Music	.997	1.0729	.353
Natural sound	2.478	1.3470	.066
Artificial sound effects	-1.350	.5716	.018

The data show that single voice segments correlated most highly with the goodness ratings of assessors (i.e., the highest beta value in the regression model). Natural sounds, such as animal sounds or natural environmental sounds (e.g., water, earthquake) also contribute positively to the goodness of the summary segments, yet the relation is not statistically significant (i.e., the p-values is greater than 0.05). Given these videos are highly produced Science Education Programs aiming at K-16 students, it is reasonable that these natural sounds are important audio indicators of the salient content of the video.

The other statistically significant beta value, artificial sound effects, was negatively correlated with goodness, which suggests that it is a distractor when only audio stream is available. Perhaps in television news this would not be the case (e.g., explosions), however, in these highly produced educational videos, the frequent use of sound effects seem to be less useful than other more distinctive audio features.

4.6.2 Goodness of the visual only summary and the Visual features

Similarly, an Ordinal Logistic Regression model was constructed to characterize the relationship between the ratings of the visual only summaries and the eight visual features (with the modal equation omitted here). Table 6 summarizes the β coefficients of the regression model for the visual only summary segments.

Table 6. Coefficients of the Regression model for visual only summary segments.

	β	Std. error	Sig.
Text (names, location)	1.017	.7039	.149
Faces	1.379	.6124	.024
Graphics & logos	.075	.6104	.902
Graphs	-.857	.8772	.329
Equations	.971	1.1572	.402
Animals	-.615	1.2011	.609
Human built artifacts	-1.457	.6461	.024
Natural scenes	1.770	.6542	.007

Natural scenes and faces were found the strongest correlates to goodness in the regression model for visual only summaries. Natural scenes are attractive to people because the videos are instructional documentary videos on scientific topics where natural scenes appear frequently, and people are naturally attentive to faces. Text, faces, and equations also contribute positively to a good visual only summary segment. Graphs and animals appeared to be negatively correlated to the goodness of a summary segment. However, due to the fact that graphs and animals only appear in a small number of segments (i.e., less than 10%), and their effects on the regression model are not significant, it is not statistically reliable to conclude that having graphs and animals in a summary segment will decrease the goodness of the segment.

4.6.3 Goodness of the summary and the audio and visual features

Furthermore, Ordinal Logistic Regression analyses were conducted to model the relationship between the ratings of the "combined" summary segments (with both audio and visual) and the six audio features and eight visual features. Table 7 summarizes the β coefficients of the regression model for the "combined" summary segments.

Table 7. Coefficients of the regression model for audio and visual combined summary segments.

	β	Std. error	Sig.
Proper noun (city, person, place name)	-.075	.6350	.906
Single human voice	2.414	1.1152	.030
Multiple human voice	2.423	1.6139	.133
Music	2.126	1.5081	.159
Natural sound	.864	1.4529	.552
Artificial sound effects	.358	.9640	.710
Text (names, location)	-1.210	.7053	.086
Faces	.243	.7479	.745
Graphics & logos	-.767	.6824	.261
Graphs	1.487	.9877	.132
Equations	-1.026	1.5752	.515
Animals	21.125	15495.8600	.999
Human built artifacts	-1.666	.8086	.039
Natural scenes	1.544	.7402	.037

Similar to the regression model for the audio only segments, single human voices and multiple human voices appeared to be most highly correlated with the goodness ratings of segments. Music correlates positively with the goodness of the combined segments more than it does for the audio only segments, but in neither condition, its influence is statistically significant.

As with the regression analysis for the visual only segments, human built artifacts are negatively correlated with the goodness of the combined summaries. What is surprising in the results is that the "graphs" appeared to contribute positively to the goodness of "combined" summary segments with both audio and visual streams, as opposed to their negative correlation with the goodness of the visual only summary segments. In the combined segments, the graphs were mostly accompanied with single or multiple human voices, which makes the graphs easier to understand as well as more informative. But note in both regression models (i.e., visual only and combined), the coefficients of the "graphs" feature are not statistically significant.

Also note the huge standard error for the "animal" feature in Table 7. The huge standard error is probably due to the very small mean of 0.05 for the "animal" feature (i.e., only 5% of the segments have "animal"). Analyses based on variables which exist in 5% of the data are not statistically reliable. Therefore, we decided to remove the "animal" variable as a predictor from the

regression model. Likewise, the "equations" and "natural sound" features also have very small means, i.e., 0.03 and 0.07 respectively. And the p values for these three features are all greater than 0.05. Hence, we removed these three features with means less than 0.1 (i.e., animal, equation, and natural sound), and re-modeled the ordinal logistic regression.

Table 8 summarizes the coefficients of the regression model for the combined summary segments, where the response variable is the evaluation score and the predictors are the 5 audio features and 6 visual features (i.e., with 1 audio feature and 2 visual features dropped from the model). Natural scenes and single human voice are shown to be most positively related to the ratings of the summary segment, whereas human built artifacts seem to be negatively related to the ratings of the segments. Graphs, and audio features such as multiple voice and music, are also positively related to the goodness of a summary segment, yet the relation was not found statistically significant.

Table 8. Coefficients of the revised regression model for audio and visual combined summary segments.

	β	Std. error	Sig.
Text (names, location)	-.989	.6698	.140
faces	.166	.7159	.817
Graphics & logos	-1.089	.6604	.099
graphs	1.225	.9568	.201
Human built artifacts	-1.735	.7660	.024
Natural scenes	1.522	.6940	.028
Proper noun (city, person, place name)	.210	.5727	.714
Single human voice	2.328	1.1060	.035
multiple human voice	2.091	1.5687	.183
music	1.558	1.2982	.230
Artificial sound	.176	.9504	.853

Based on the regression models obtained for the audio only, visual only, and combined summary segments, the implications for automated video summarization might be to automatically determine single human voice in these particular instructional or educational videos, since so much of these videos is about narrator explaining some key concepts. In fact, a lot of automated video summarization techniques are based on text summarization and audio emphasis (or excitement level) detection. The findings of this paper are consistent with the previous works in video summarization. Single human voices with the highest tf-idf or emphasized speech should be extracted to form audio summaries. For the particular genre of instructional educational videos investigated in this paper, visual features such as natural scenes and graphs are important visual information carriers. Therefore, automated video summarization techniques should also focus on extracting these visual features.

Note that the ordinal logistic regression conducted in this study considered main effect models with all independent variables. Future work may consider models with interaction among the independent variables as well as consider different model selection methods.

Also note that this paper only deals with one corpus of instructional documentary video, hence the results and implications may not apply to other video genres (which may have different salient audio or visual features from the NASA videos). However, the methodology is very valuable, and can be easily extended to inform summarization of other video corpuses.

5. CONCLUSIONS

A set of manually created video summaries for instructional documentary videos were examined by multiple human assessors. The summary segments were indexed by the assessors according to a rubric defined by the authors of the paper, and the segments were rated by the assessors on a 1-7 scale based on their goodness in helping people make sense of the full videos. Then the most eye-catching and ear-catching features in these summary segments were identified using statistical procedures.

The results demonstrate that for instructional documentary videos, the bulk of the content useful for summaries is carried in the audio channels with the visual channel providing supporting examples or illustrations. Within the visual channel, text, equations, and graphs are important if audio is not available, otherwise, human faces and natural scenes are often selected to form video summaries. All of these visual features except natural scenes can be easily detected with automatic techniques. Within audio channels, single human voices and natural sounds tend to be selected for summaries for these instructional documentary videos, whereas multiple human voices are not as commonly selected. This also makes automatic summarization easier because human voices can be easily recognized and pattern matching should be useful for recognizing many natural sounds.

Clearly, video indexing is complex and many factors influence both how people select salient segments. Through understanding how people create video summaries, this paper provides guidance for important features that automatic algorithms should be looking for when performing automatic video summarization. More importantly, the valuable methodology in this paper can be extended to inform summarization of other video corpuses.

6. ACKNOWLEDGMENTS

We would like to thank the human judges and assessors who participated in the study and thank the anonymous reviewers for the valuable comments.

7. REFERENCES

[1] L. Agnihotri, K. Devera, T. McGee, and N. Dimitrove. Summarization of video programs based on closed captions. In Proc. SPIE. Conf. Storage and Retrieval for Media Databases, page 599-607, San Jose, CA, Jan. 2001.

[2] B. Arons. Speechskimmer: A system for interactively skimming recorded speech. ACM Transactions onComputer Human Interaction, 4:3-38, 1997.

[3] R. Cabasson and A. Divakaran. Automatic extraction of soccer video highlights using a combination of motion and audio features. In *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases 2003*, volume 5021, pages 272-276, Santa Clara, CA, 2003.

[4] F. R. Chen and M. Withgott. The use of emphasis to automatically summarize a spoken discourse. In *Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992 IEEE International Conference on*, volume 1, pages 229-232 vol.1, 1992.

[5] M. Christel, S. Stevens, T. Kanade, M. Mauldin, R. Reddy, and H. Wactlar. Techniques for the creation and exploration of digital video libraries. In *Multimedia Tools and Applications, B. Furht, Editor*. Kluwer Academic Publishers, 1996.

[6] M. G. Christel, M. A. Smith, C. R. Taylor, and D. B. Winkler. Evolving video skims into useful multimedia abstractions. In *CHI '98: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 171-178, New York, NY, USA, 1998. ACMPress/Addison-Wesley Publishing Co.

[7] A. Ekin and A. M. Tekalp. Automatic soccer video analysis and summarization. *IEEE Trans. on Image Processing*, 12:796-807, 2003.

[8] B. Erol, D.-S. Lee, and J. Hull. Multimodal summarization of meeting recordings. In *ICME '03: Proceedings of the 2003 International Conference on Multimedia and Expo - Volume 3 (ICME '03)*, pages 25-28, Washington, DC, USA, 2003. IEEE Computer Society.

[9] D. Farin, W. Effelsberg, and P. H. N. deWith. Robust clustering-based video-summarization with integration of domain-knowledge. In *Proc. IEEE Int. Conf. Multimedia and Expo 2002 (ICME'2002)*, pages 89-92, Lausanne, Switzerland, 2002.

[10] A. M. Ferman and A. M. Tekalp. Two-stage hierarchical video summary extraction to match low-level user browsing preferences. *IEEE Trans. Multimedia*, 5(2):244-256, Jun. 2003.

[11] J. Foote, M. Cooper, and L. Wilcox. Enhanced video browsing using automatically extracted audio excerpts. *IEEE,* 2000.

[12] M.E. Funk and C.A. Reid. Indexing consistency in MEDLINE. *Bull Med Libr Assoc.* 1983;**71**:176–183.

[13] Y. Gong. Summarizing audiovisual contents of a video program. *EURASIP J. Appl. Signal Process.*, 2003:160-169, 2003.

[14] A. Hanjalic and H. Zhang. An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. *IEEE Trans. Circuits Syst. Video Technol*, 9(8):1280-1289, 1999.

[15] A. Hauptmann, M. Christel, W. Lin, B. Maher, J. Yang, R. Baron, and G. Xiang. Summarizing bbc rushes the informedia way. In *TVS '07: Proceedings of the international workshop on TRECVID video summarization*, New York, NY, USA, 2007. ACM.

[16] L. Kennedy and D. Ellis. Pitch-based emphasis detection for characterization of meeting recordings. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Virgin Islands, Dec. 2003.

[17] J. R. Landis and G. G. Koch. The measurement of observer agreement for categorical data. Biometrics, Vol. 33, No. 1, pages 159-174, Mar. 1977.

[18] B. Li, H. Pan, and I. Sezan. A general framework for sports video summarization with its application to soccer. In *Proc. IEEE Int. Conf. Acoustic, Speech and Signal Processing*, pages 169-172, Hong Kong, 2003.

[19] Y. Li, C. Dorai, and R. Farrell. Creating magic: system for generating learning object metadata for instructional content. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 367-370, New York, NY, USA, 2005. ACM.

[20] W.-N. Lie and C.-M. Lai. News video summarization based on spatial and motion feature analysis. In *Proceedings of the 5th Pacific Rim Conference on Multimedia. Lecture Notes in Computer Science*, volume 3332, pages 246-255, 2004.

[21] R. Lienhart, S. Pfeiffer, and W. Effelsberg. Video abstracting. *Commun. ACM*, 40(12):54-62, 1997.

[22] G. Marchionini, Y. Song, and R. Farrell. Multimedia surrogates for *video gisti*ng: Toward combining spoken words and imagery. *In Journal of Information & Process Manage*. 45(6): 615-630.

[23] P. Over, A. F. Smeaton, and P. Kelly. The trecvid 2007 bbc rushes summarization evaluation pilot. In *TVS '07: Proceedings of the international workshop on TRECVID video summarization*, pages 1-15, New York, NY, USA, 2007. ACM.

[24] K. Ratakonda, I. M. Sezan, and R. J. Crinon. Hierarchical video summarization. In *Proc. SPIE Conf. Visual Communications and Image Processing*, volume 3653, pages 1531-1541, San Jose, CA, Jan. 1999.

[25] Y. Song and G. Marchionini. Effects of audio and visual surrogates for making sense of digital video. In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 867–876, New York, NY, USA.

[26] C. Taskiran, Z. Pizlo, Amir, D. A., Ponceleon, and E. J. Delp. Automated video program summarization using speech transcripts. *IEEE Transactions on Multimedia*, 8(4):775-791, 2006.

[27] S. Uchihashi, J. Foote, A. Girgensohn, and J. Boreczky. Video manga: Generating semantically meaningful video summaries. In *ACM Multimedia'99*, pages 383-392. ACM Press, 1999.

[28] B. M. Wildemuth, G. Marchionini, T. Wilkens, M. Yang, G. Geisler, B. Fowler, A. Hughes, and X. Mu. (2002). Alternative surrogates for video objects in a digital library: Users' perspectives on their relative usability. In *ECDL '02: Proceedings of the 6th European Conference on Research and Advanced Technology for Digital Libraries*, pages 493–507, London, UK. Springer-Verlag.

[29] J. Yang and A. G. Hauptmann. Exploring temporal consistency for video analysis and retrieval. In *MIR' 06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 33-42, New York, NY, USA, 2006. ACM.