Multi-modal Surrogates for Retrieving and Making Sense of Videos: Is Synchronization between the Multiple Modalities Optimal?

By
Yaxiao Song

A dissertation submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the School of Information and Library Science.

Chapel Hill
2010

Approved by

Dr. Gary Marchionini, Advisor

Dr. Barbara M. Wildemuth, Reader

Dr. Jane Greenberg, Reader

Dr. Bradley M. Hemminger, Reader

Dr. Abby Goodrum, Reader

Abstract

YAXIAO SONG: Multi-modal Surrogates for Retrieving and Making Sense of Videos: Is Synchronization between the Multiple Modalities Optimal?
(Under the direction of Dr. Gary Marchionini)

Video surrogates can help people quickly make sense of the content of a video before downloading or seeking more detailed information. Visual and audio features of a video are primary information carriers and might become important components of video retrieval and video sense-making. In the past decades, most research and development efforts on video surrogates have focused on visual features of the video, and comparatively little work has been done on audio surrogates and examining their pros and cons in aiding users' retrieval and sense-making of digital videos. Even less work has been done on multi-modal surrogates, where more than one modality are employed for consuming the surrogates, for example, the audio and visual modalities. This research examined the effectiveness of a number of multi-modal surrogates, and investigated whether synchronization between the audio and visual channels is optimal. A user study was conducted to evaluate six different surrogates on a set of six recognition and inference tasks to answer two main research questions: (1) How do automatically-generated multi-modal surrogates compare to manually-generated ones in video retrieval and video sense-making? and (2) Does synchronization between multiple surrogate channels enhance or inhibit video retrieval and video sense-making? Forty-eight participants participated in the study, in which the surrogates were measured on the the time participants spent on experiencing the surrogates, the time participants spent on doing the tasks, participants' performance accuracy on the tasks, participants' confidence in their task responses, and participants' subjective ratings on the surrogates. On average, the uncoordinated surrogates were more helpful than the coordinated ones, but the manually-generated surrogates were only more helpful than the automatically-generated ones in terms of task completion time. Participants' subjective ratings were more favorable for the coordinated surrogate C2 (Magic A + V) and the uncoordinated surrogate U1 (Magic A + Storyboard V) with respect to usefulness, usability, enjoyment, and engagement. The post-session questionnaire comments demonstrated participants' preference for the coordinated surrogates, but the comments also revealed the value of having uncoordinated sensory channels.

To Mom, Dad, Feng, and Leo, with love

Acknowledgments

It is my pleasure to thank all the wonderful people who have supported me and made this dissertation possible.

First and foremost, I want to express my sincerest gratitude to my advisor, Dr. Gary Marchionini, for providing guidance, resource, time, support, care, and effort, throughout the entire duration of my research. He enlightened me in the field of information science, and ignited my enthusiasm for research. He is a wonderful mentor.

I also offer my deep appreciation to Dr. Barbara Wildemuth. It was my great pleasure to have Dr. Wildemuth on my committee. I would like to thank her for her constant help and support. I have benefited a lot from the numerous discussions with her during the past few years.

I'm also grateful to the other members of my committee, Dr. Jane Greenberg, Dr. Brad Hemminger, and Dr. Abby Goodrum, for their careful attention to my research, their time and knowledge, and their valuable feedbacks shared with me.

In my daily work, I have been blessed with a friendly and cheerful group of fellow students, faculty, and staff. Thanks to all of them for the help they offered to me and the joyful time we spent together. Special thanks to Dr. Evelyn Daniel, Dr. Chuanshu Ji, Dr. Rob Capra, Dr. Lili Luo, Terrell Russell, Scott Adam, Lara Bailey, and Chiyoung Oh.

I would also like to acknowledge all the participants for their time and diligence, and the financial support from the NSF grant (IIS 0455970).

Finally, I sincerely thank my grandma, my parents, and in-laws for their support and encouragement during my PhD study. Thank my beloved husband, Feng, for his unconditional love and care since we first met. Thank my dearest son, Leo, for the love and joy he has brought to me. Having him has been a true blessing in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1

INTRODUCTION

1.1 Problem Statement

We are living in a digital information world. Most videos today are produced in digital form. Computing technologies have promoted the creation, availability and distribution of a massive amount of digital videos. As digital video cameras and webcams become common household appliances, making videos has become both easier and less expensive. As a result, large collections of digital videos (e.g., YouTube, Internet Archive, Open Video) are increasingly available for people to download and use on various devices ranging from desktop computers to small devices, such as cell phones and PDAs. The tremendous volume of digital videos, in turn, requires effective and efficient access to those videos. While digital video is becoming increasingly ubiquitous, the usability of web-based video retrieval is often quite poor. One problem is that searchers need better summaries, excerpts, or other highly condensed representations of the videos, to make judgments about whether to download and view the full videos, as well as to locate the most relevant minutes from thousands of hours of video content. We refer to these human-consumable summaries as surrogates.

Video surrogates are the key to successful video search systems or large video repositories because they facilitate finding and selecting videos from the large collections, and help people quickly make sense of the content of a video before downloading (which requires not only large storage but bandwidth) or seeking further detailed information. It has been demonstrated by a number of usability studies (Christel et al., 1998; Ding et al., 1999; He et al., 1999; Goodrum, 2001; Yang et al., 2003; Wildemuth et al., 2002, 2003; Lie and Lai, 2004) and some real-life video retrieval and search applications (e.g., Internet Archive, Open Video) that people can quickly make sense of videos by viewing the abbreviated video surrogates.

Different surrogates have a variety of advantages and disadvantages, and the unique advantages of different surrogates can be selectively applied in video retrieval systems. The bulk of surrogates in today's video retrieval systems are still text-based, although visual surrogates which represent visual cues of the videos, such as poster frames, storyboards, and fast forwards, have become available on some digital video repositories (e.g., Internet Archive, Open Video). Many years of TREC Video results in the past demonstrate that linguistic data generally lead to better performance in video retrieval than the visual features (e.g., Smeaton et al., 2004). In 2005, some groups showed better performance with visual features than linguistic features, but under very difficult linguistic conditions, where multiple language translation was automated (Over et al., 2005). Nevertheless, some studies report that people like to have visual surrogates regardless of their performance effects. For example, He et al. (2000) compared four possible ways of summarizing presentations, presentation slides, text transcript, transcript with highlighted points, and a manually created audio-video summary. They reported that users prefer audio-visual summaries to both text transcripts and presentation slides, though the four summarization methods offer comparable results. An eye-tracking study by Hughes et al. (2003) also confirmed that participants liked the pictures and felt that they were necessary and added some value to the search process.

Videos are multi-modal presentations: they are a combination of a series of moving pictures playing at a constant speed (i.e., 25 to 30 frames per second) along with one or more synchronized audio track(s). Visual and audio data as entry points for retrieval are increasingly practical with better broadband access, and visual and audio surrogates might become important components of video retrieval and sense-making. Audio features of the videos, as well as visual features of the videos, are also important information carriers and important cues for understanding the videos. Nevertheless, most surrogates in today's video retrieval systems, are unimodal: they are either text-based surrogates or visual-based surrogates, whereas the notion of audio surrogates has not been well discussed or examined by researchers. Not surprisingly, few systems in practice utilize audio surrogates for aiding video browsing and retrieval. However, audio surrogates for digital videos, either stand-alone or combined with visual surrogates, could be very powerful and promising, because they engage humans' natural ability to hear and require no training for sense-making. If reasonably designed, audio surrogates can very well assist people in understanding the content of the videos. Not only can audio surrogates be downloaded much faster than videos, but they also require less system resources and minimal screen real estate to be useful. They thus have the potential to be successful surrogate alternatives for limited displays on small devices (e.g., PDAs, cell phones). Moreover, audio surrogates can be used by people in situations where they can not use vision to work on videos, for example, when they are driving or walking on the street. Most importantly, audio surrogates can be even more powerful if successfully combined with visual surrogates to leverage multiple sensory channels without significantly increasing people's cognitive loads.

Video surrogates that leverage multiple sensory channels are multi-modal surrogates. For instance, storyboards or slide shows can be augmented by audio narrations to leverage both the hearing and seeing modalities (Wildemuth et al., 2002). Video skims created by the Informedia Project are compact, content-rich video abstractions incorporating both audio and visual information from longer video sources, which preserve frame rate while greatly reducing viewing time (Christel et al., 1998). Movie trailers are a familiar example of multi-modal surrogates for videos; however, they are usually extremely well-made by professionals and are very expensive. To make surrogates available for large video collections, we need inexpensive surrogates which can be created automatically.

1.2 Research questions

The study described here explores multi-modal surrogates for retrieving and making sense of videos. The main questions to be answered are:

- How do automatically generated multi-modal surrogates compare to manually generated ones in video retrieval and video sense-making?
- Does coordination between multiple surrogate channels enhance or inhibit video retrieval and video sense-making?
  - If *unsynchronized* multi-modal surrogates are created by extracting the most salient samples from the audio and visual channels separately and then combining the extracted salient samples, will they convey more useful information about the video per unit time and lead to better retrieval and sense-making performance than extracting the most salient samples from the two channels synchronically?

1.2.1 Multi-modal Surrogates

Multi-modal surrogates are surrogates that employ multiple sensory channels, for example, the visual modality and the audio modality. The obvious gain of using multiple sensory channels is increased usefulness: Weaknesses of one modality are offset by the strengths of another. Possible drawbacks are increased cognitive load.

For instance, movie trailers are temporally coordinated multi-modal surrogates that are easy to comprehend, however, they are usually very expensive.

On the contrary, slide shows can be augmented by audio narrations to leverage both the hearing and seeing modalities. Additionally, it is less expensive to sample from the two channels separately and then create the surrogate. To make surrogates available to be used in large video collections, inexpensive surrogates which can be created automatically are preferred.

This study examined the effectiveness of several multi-modal surrogates, for which human audio and visual sensory channels are both leveraged. The potential value of spoken audio (either extracted from the video or spoken by a speech synthesizer) when combined with some visual surrogates for retrieving and making sense of videos, is of special interest.

This research question leads to the second research question of this study: the need for synchronization between the audio and visual channels of the surrogates. With surrogates consisting of audio and visual abstracts from the video, is it necessary to carefully synchronize the audio and visual channels?

1.2.2 Coordination between Channels

According to Dual Coding Theory (Paivio, 1986), verbal and visual information are simultaneously processed in two separate sub-systems of human cognition. Auditory information, when designed to complement the visual environment, can have additive effects on human recall (Thompson and Paivio, 1994).

Though it has already been accepted by a lot of researchers that for primary information objects, coordinated media channels lead to better understanding, retention, and satisfaction, there are no safe and sound conclusions on whether this coordination requirement for primary information objects also applies to highly abbreviated, condensed information objects, such as video surrogates. In fact, our two preliminary studies (Song and Marchionini, 2007; Marchionini et al., 2009) are among the first studies examining uncoordinated multi-modal surrogates employing multiple sensory channels, yet different conclusions were drawn from these two studies about the synchronization necessity due to the fact that different multi-modal surrogates were examined in the two studies. This study will take lessons from the previous studies, and carry out the evaluation of synchronized multi-modal surrogates against unsynchronized multi-modal surrogates.

1.2.3 Sampling Together or Separately?

In addition to the synchronization issues between the audio and visual channels, the strategies of sampling the most salient abstracts from the two channels will also be of interest in this study. I will distinguish the ***pre-processed integration*** for video surrogates (where the audio and visual channels are sampled simultaneously and thus are pre-coordinated at indexing time), from the ***user-centered integration*** (where the audio and visual channels are sampled independently (i.e., are uncoordinated) and need to be integrated in the consumer's head at consumption time). Going back to the audio augmented slide show example, although users are required to integrate the two channels in their heads at consumption time, which may lead to increased cognitive load, yet perhaps more sense-making is possible with this user-centered integration, rather than pre-coordination at indexing time. Therefore, even though pre-coordination at indexing time may be desired from the consumer's affective point of view because we are more used to synchronized presentations, my hypothesis is that, because the most salient abstracts from the audio channel do not necessarily align temporally with the most salient abstracts from the visual channel, **more** information may be carried in the abstracts than if the audio and visual of the abstracts are temporally coordinated. As a result, independently sampling the most salient samples across different channels and letting users integrate the uncoordinated channels at consumption time may lead to **more** sense-making potential.

1.2.4 Hypotheses

Based on the research questions, a set of hypotheses are proposed as follows:

1. Human interpretation cost

    - **Time**: it will take people longer time to make sense of the surrogates with unsynchronized audio and visual channels than surrogates with the synchronized channels.
    - **Affective measures**: people will prefer the synchronized surrogates to the unsynchronized ones.

2. Performance:

    - **Accuracy**: people will perform the tasks with higher accuracy with unsynchronized multi-modal surrogates than with the synchronized ones.

- **Confidence**: people will have higher confidence in their task responses using the unsynchronized multi-modal surrogates than using the synchronized ones.

1.3 Significance of the Study

Research on multi-modal surrogates can not only enrich the video information seeking literature, but also give recommendations and provide implications for the design of effective video surrogates for video retrieval systems to better serve users' needs. This study will provide some guidelines for creating multi-modal surrogates. Additionally, the dilemma of having synchronized surrogate channels vs. independently extracted unsynchronized surrogates across audio and visual channels will be investigated carefully for a specific important video genre: **instructional documentaries**. The study can help us get a better understanding on how necessary the synchronization between the surrogate channels is and whether unsynchronized sampling of individual channels provides more useful information than synchronized sampling from the channels. These findings will inform the interface design and video representation for video retrieval, browsing, and sense-making. In addition, the evaluation tasks and methods used in this study will offer new approaches to evaluating retrieval techniques and may be useful to other researchers.

Chapter 2

RELATED LITERATURE

This chapter reviews existing literature in four different but related research areas. To study users' use of multi-modal surrogates in video retrieval and sense-making, it is necessary to consider the literature on how people understand videos and images from a cognitive point of view. In addition, previous and current techniques for automatically making video summaries or video surrogates, and video retrieval techniques through metadata and surrogates will be considered. Finally, the methodologies used in evaluating the video summaries or surrogates are examined.

2.1 Psychological Theories of Making Sense of Videos

To study the effectiveness of multi-modal surrogates and the necessity of temporally coordinating the visual and audio channels for the video surrogates, it is necessary to consider the current literature on how people make sense of multimedia contents and, more importantly, how people make sense of videos through surrogates. Since videos, as well as multi-modal surrogates, contain multiple channels of information, i.e., audio, visual, and textual channels, it's also important to look at the literature on Dual Coding Theory. This section reviews the previous work on how people make sense of videos and video surrogates.

2.1.1 Multimedia Sense-making

Viewing images and videos is both a sensory and a mental experience. Hanson (1987, p.36) has written:

> "Sensory and mental experiences are also influenced by a complex process that includes environmental, technical, physiological, psychological, cultural, and social factors."

According to Hanson (1987), environments influence our understanding of content and how much attention we pay to the subject. Whether we watch a film in a movie theater or watch a video on TV, whether we watch it alone or with family and friends, we experience different levels of distraction from the environment and may even perceive the video differently. With the prevalence of personal computers, laptops, cell phones, and PDAs, watching videos has become a personal activity, which can take place almost anywhere at anytime. Within the scope of this study, the video viewing environment is set to be private (for example, one watches videos alone on a personal computer or cell phone), thus the outside distraction from the environment can be considered to be minimal.

2.1.1.1 Physiological Perception of Visuals

Our brain gets 75% of its information from the eyes (Youngblood, 1970, p. 46), and the eyes filter the information to be processed in the brain. When we watch a film or video, we are actually watching a series of still images displayed to us at 24fps (frame per second) or more. Because of the way our brain and eye process the visual information, an image appears to be present (known as the "afterimage") for approximately one twenty-fifth of a second on the retina after the exposure to the image ceases (Wertheimer, 1961). This phenomenon is called *persistence of vision*. Due to this physiological phenomenon, our eye and brain get the illusion that the images are moving.

2.1.1.2 Psychological Perception of Visuals

Once the eye and brain register a visual representation, the process of making sense of icons and symbols in the visual field falls into the realm of psychological perception (Hanson, 1987).

The same image can convey different meanings to different people given their different backgrounds, knowledge, experiences, cultures, specific information needs, or interests. Watching video (or film) requires comprehension performance such that people are required to know what a video is about and make meaningful representations out of the video (Lee and Roskos-Ewoldsen, 2004). Video comprehension is even more complicated than image comprehension due to its temporal and spatial complexity, for example, new information is associated with previous information and modifies the current mental representation. According to the event-index-model by Florida et al. (1998), people are involved with mental representation construction along five situational dimensions–time, space, causation, intentionality, and protagonist– all five dimensions account for how information is integrated when people watch videos. To explore what kinds of video surrogates can help people efficiently and effectively makes sense of videos or make relevance judgments, it is important to learn how people make sense of video, either through perceiving the full video itself, or through perceiving surrogates.

The smallest meaningful unit in a video is the frame (image). To understand how people make sense of multimedia content, we can look at how people perceive visual images as a starting point for corresponding video understanding and sense-making.

People's understanding of visual images involves several different levels. The Renaissance art historian Panofsky identified three levels of imagery comprehension: *pre-iconographical* description, *iconographical* analysis, and *iconographical* interpretation or synthesis (Panofsky, 1955).

- The ***pre-iconographic description*** of an image describes what fundamental visual elements are depicted in the image (i.e., the generic things in the image), such as "red", "women", "dog", "mirror", and "lamp".

  The pre-iconographical level of description does not require any previous in-depth knowledge of either the image or its context. Pre-iconographical description is possible if the viewer understands the factual subject matter shown in the image (e.g., a woman), which "can be identified ... on the basis of our practical experience" (Panofsky, 1972, p. 9).

- The secondary level, ***iconographical analysis***, describes what the image represents (i.e., the specific things in the image).

  This level of comprehension requires the viewer's pre-iconographical understanding of and familiarity (e.g., social and cultural knowledge) with the theme and concepts represented, and involves a deeper understanding of the subject matter. For example, a representation of a haloed woman could be specifically interpreted as the Virgin Mary, and a saint holding keys usually represents St. Peter.

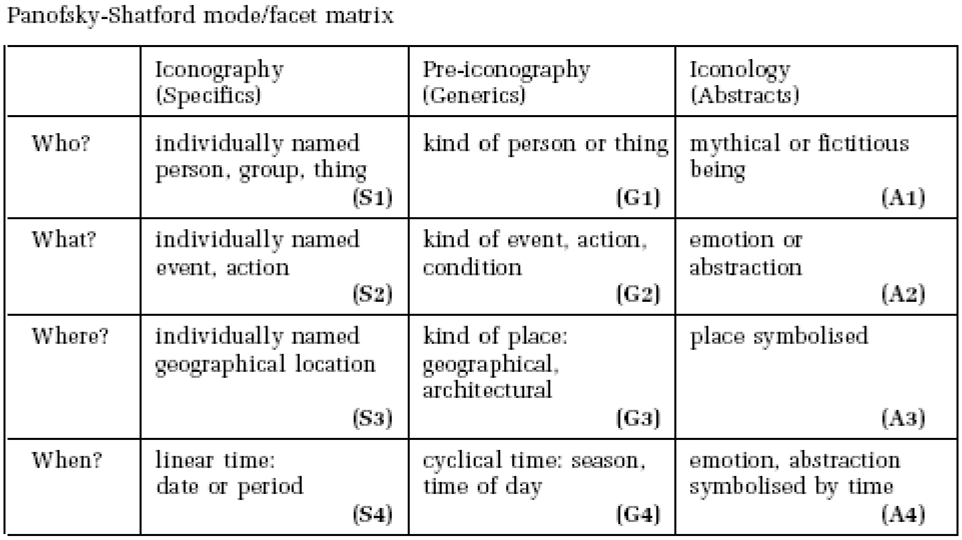
- The third level, ***iconological (iconographical interpretation or synthesis)***, examines an image beyond its face value, and focuses on the symbolic meaning (especially in personal, social, and political terms) of the image, such as "nature", "peace" and "fidelity".

  This level of comprehension requires knowledge of the first two levels as well as adding an *emotional* aspect to the viewers' understanding of the image.

Similar to Panofsky's image comprehension model, Hanson (1987) defined three levels of meaning that we can find from a frame: *icon*, *index*, and *symbol*. As stated in (Hanson, 1987, p. 43)

> "An icon is an image that stands for something else ... [A]n index ... associates the image with a conceptual idea ... [A] symbol...(relates) the image to the actual context..."

Berger (2004) outlined some technical codes that influence our understanding of videos. For example, camera pan down often signifies power and authority, and camera pan up often signifies smallness and weakness; zoom in signifies observation and focus, while fade in and fade out signify beginning and ending, respectively. People's understanding of these codes or symbols are learned through their experiences and influenced by the culture. The same visual representation can convey different meanings to different people.



Panofsky-Shatford mode/facet matrix

	Iconography (Specifics)	Pre-iconography (Generics)	Iconology (Abstracts)
Who?	individually named person, group, thing (S1)	kind of person or thing (G1)	mythical or fictitious being (A1)
What?	individually named event, action (S2)	kind of event, action, condition (G2)	emotion or abstraction (A2)
Where?	individually named geographical location (S3)	kind of place: geographical, architectural (G3)	place symbolised (A3)
When?	linear time: date or period (S4)	cyclical time: season, time of day (G4)	emotion, abstraction symbolised by time (A4)

Figure 2.1: Panofsky-Shatford Mode Facet Model. (Source: Armitage and Enser (1997), p. 290.)

Shatford (1986) applied Panofsky's imagery comprehension model to indexing images, and suggested a mode facet matrix by relabeling Panofsky's three levels as *Generic of* , *Specific of*, and *Abstract (about)*, with each level further divided into four facets: **who**, **what**, **where** and **when** (see Figure 2.1). The "generic of" level describes general objects and actions, such as woman, house or walking. The "specific of" level describes individually named objects and events, such as the Eiffel tower or the Niagara Falls. The "about" level contains moods, emotions, abstractions and symbols, for example, happiness, fear, justice, etc. The 3x4 matrix has often been referred to as the Panofsky-Shatford model by information scientists (e.g., Enser, 1995) and used in a number of studies in indexing both still images (Armitage and Enser, 1997) and moving images (Turner, 1995).

From the Content Based Information Retrieval (CBIR) perspective, Jaimes et al. (1999) incorporated Panofsky and Shatford's models and defined a 10-level pyramid framework for describing the visual content attributes of images or video. The same pyramid can also be used to classify attributes obtained from audio. The pyramid includes 4 syntactic layers and 6 semantic layers (see Figure 2.2).

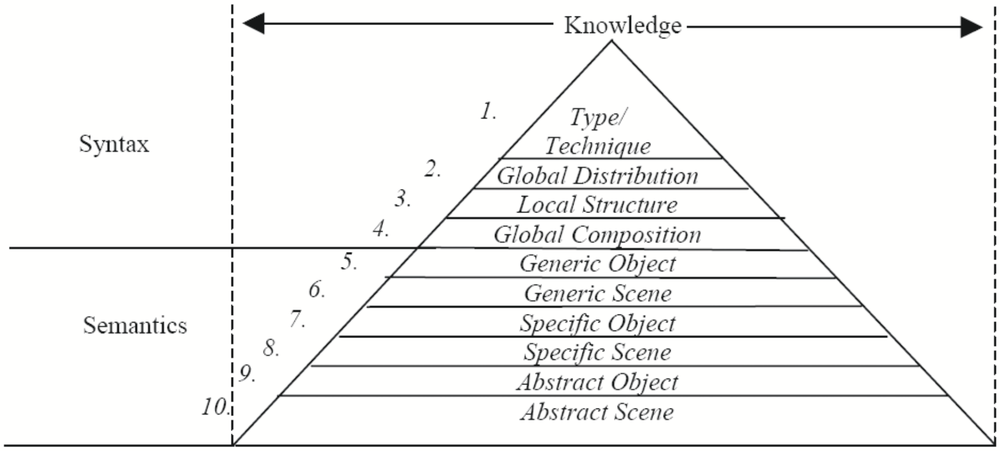


Figure 2.2: The 10-level Jaimes Chang Indexing Pyramid.

The top 4 levels of the pyramid describe the syntactic and perceptual information within the images, i.e., how the content is organized, "but not its meaning" (Jaimes et al., 2000). The first level, *Type/Technique* describes the type of the image, for example photo, X-ray, painting. etc. The second level, *Global Distribution* describes the global attributes of the image, such as color histogram, texture. The third level, *Local Structure* describes components that are local to individual parts within the image, such as dots, lines, and circles, as well as temporal/spatial positions (e.g., start time and centroid) (Jaimes and Chang, 2000). As for the fourth level, *Global Composition*, as noted in Jaimes et al. (2000, p.1), it

> "relates to the way in which those local components are arranged in the image (e.g., symmetry)."

The four syntactic/perceptual levels do not require world knowledge to perform indexing of the images. The six semantic levels describe the meaning of the elements within the images, and closely mirror the Panofsky-Shatford model (i.e., *Generic*, *Specific*, and *Abstract*).

Each of the levels of image comprehension discussed above (Panofsky, 1955; Shatford, 1986; Jaimes et al., 1999) may also apply to video materials, both full videos or video surrogates.

Similar hierarchies of visual understanding are commonly found in studies of content-based video and image retrieval (Eakins and Graham, 1999; Greisdorf and O'Connor, 2002), which suggest that people interact with image or video at three levels. According to Eakins and Graham (1999), at **level 1**, *primitive* features of the image (e.g., color, texture, shape, spatial location) are perceived. Levels 2 and 3 need semantic interpretations. At **level 2**, "derived" (a.k.a. *logical*) features involving some degree of logical inference about the identity of the objects depicted in the image (e.g., people/thing, place/location, action) are perceived, and people draw on their existing knowledge to identify the objects perceived. For example, to comprehend a picture of a "double-decker bus" or "the Eiffel tower" (Eakins and Graham, 1999, p. 7), one needs some prior knowledge to identify an object as a bus rather than a lorry or have the knowledge that a specific building structure has been named "the Eiffel tower". **Level 3** involves a great amount of high-level complex reasoning and subjective judgement about the image's *abstract* attributes, such as the meaning and purpose of the objects or scenes depicted, and emotional or religious significance of the image. Greisdorf and O'Connor (2002, p. 8) described hierarchical levels of image perception which were consistent with Eakins and Graham (1999).

2.1.1.3 Perceiving Films and Videos

Beyond perceiving individual frames, video comprehension is more complicated than image comprehension due to its temporal and spatial complexity. In particular, new information is associated with previous information and also modifies the viewer's current mental representation.

Noel Carroll and David Bordwell are often considered the two main cognitivists in film theory. They proposed a cognitive approach to film theory as an alternative to the mainstream in film studies. Cognitivists are interested in both simple visual information processing and more complex kinds of interpretative and emotional responses. As summarized in Yang (2005), Bordwell (1989) regarded film comprehension as an activity of sense-making and inference-making, and Branigan (1992) interpreted film understanding as a narrative construction process. Ponech (1997, p. 85) writes, "Movie spectatorship involves two kinds of perceptual activities: sensory contact with the cinematic image and epistemic access to the image along with further objects, situations, and events". Thus, the viewers' sense-making process always involves two levels: "sensory seeing and cognitive seeing".

In a work contemporary to Ponech (1997), Grodal (1999, p. 59) described a flow diagram with four main steps in processing of audiovisual input. **Step 1** consists of *basic perception*. The brain makes its first visual analysis of colors, textures, lines and figures. The process creates perceptual intensities without any meaning in the ordinary sense of the word. **Step 2** consists of memory-matching. The brain searches its memory files for possible matches, aided by feelings of familiarity or unfamiliarity. **Step 3** is the construction of narrative scene or universe, and a cognitive-emotional appraisal and motivation phase, which consists of relating and contextualizing the items seen and determined in steps 1 and 2 to a living being (a human or an animal in cartoons) and a scene. For example, snakes represents possible danger, and the viewer may experience strong arousal with increased heartbeats and sweat. Step 3 then leads to **step 4**: reactions at a high level of arousal, such as crying, laughter, or shivering.

Grodal's 4-step model resembles Ponech's 2-level categorization of movie spectatorship. Specifically, step 1 is about low-level visual perception/identification (i.e., sensory seeing), and steps 2 - 4 are about high-level cognitive understanding (i.e., cognitive seeing). Both Ponech's 2-level sense-making process and Grodal's 4-step processing model echo Panofsky's 3-level comprehension model, where pre-iconographical is about sensory seeing, while iconographical analysis and iconographical interpretation are about cognitive seeing.

Moreover, according to the research on information objects (e.g., the Dual Coding Theory) involving more than one modality (take, for example, full motion videos with both audio and visual channels), if the information from different modalities is well integrated, the modalities reinforce rather than interfere with each other and may lead to increased usability and comprehension. A more detailed discussion of the Dual Coding Theory will be provided in Section 2.1.3.

Furthermore, when people watch a video, they may screen out some information and choose to pay attention to only some stimuli. Selective attention and retention are results of the individual's information needs and interests, and may also relate to the genre of the video itself. For example, a person watching a video for the first time may pay different attention to the video than a person who is watching it for the second time. Also, the same person pays different attention when he watches a news video than when he watches a comedy. Next, let's look at the different levels of attention we use when watching videos.

2.1.1.3.1 Levels of attention Not only does watching videos or films require comprehension performance such that people are required to know what a video is about and make meaningful representations out of the video (Lee and Roskos-Ewoldsen, 2004), but watching a film has also been considered as a constant struggle against distractions, either from within the film, or from outside the film (Hutchinson, 2004). Hutchinson summarized three distinct levels of attention we use when watching films: (1) attending to the video other than to the circumstances, (2) attending to the film as a fictional story versus a constructed world consisting of actors, sets, and artificial devices, and (3) attending to the diegetic or internal circumstances of the story itself.

For the first level, attending to the video, it has been found that movie viewing always involves some goal-driven attention (i.e., a viewer sits down with the intention of watching the movie). However, stimulus-driven attention can be activated by peripheral events (e.g. a person stands up and leaves his seat) or auditory stimuli (e.g. chatting and coughing from other audiences).

For the second level, attending to the film as a fictional story, most movie viewers (if not professional critics) will watch a film with attention to the fictional story. Film makers employ “invisible” film editing so as to hide the camera transitions from one angle to another. Artificial devices like cameras, microphones, cables, etc, are intentionally kept outside the camera frame to hide the distracting stimuli from the viewers as much as possible.

The last level of film viewing attention, attending to diegetic circumstance, involves maintaining the viewer’s attention and interest. Like a bird flying in the sky, a movie or film needs to have a great deal of continual changes to be engaging, especially over long periods of time.

In addition, the way people perceive videos is also subject to the differences in cultural and sociological aspects, which falls outside the scope of this review.

2.1.2 Video Sense-making through surrogates

With the boom of digital technologies, large collections of digital videos are increasingly available for people to download and use. The tremendous volume of digital videos, in turn, requires effective and efficient access to those videos. It has been demonstrated by a number of usability studies (Ding et al., 1999; Yang et al., 2003; Wildemuth et al., 2002, 2003) and some real video retrieval and search engines (e.g., Open Video, Internet Archive) that people can quickly make sense of videos by viewing the abbreviated video surrogates. Surrogates facilitate faster lookups of and access to video collections.

Yang et al. (2003) classified video surrogates into two modalities: *textual* and *audiovisual*. In fact, *audio* and *visual* can be further divided into two separate modalities. Better understanding the role of the individual modalities and how to integrate multiple modalities so as to facilitate better video searching and browsing is crucial to the success of a digital video library. Hence, it is important to learn how people react to, perceive, conceive, and integrate these different modalities.

For the visual-based surrogates, people make sense of the videos through understanding the images. The previous sections have already discussed how people perceive images, and these models and theories also apply to how people make sense of the videos through visual surrogates.

2.1.2.1 Rephrasing the literal

Ding et al. (1999) designed an exploratory usability study to compare three types of video surrogates – visual (keyframes), verbal (keywords/phrases), and visual and verbal combined.

After viewing the surrogates (i.e., keywords, key frames, or both), participants were asked to write 2-3 sentences that summarize what the video clip was about. It was found that participants tended to make up a sentence to include *all keywords* they saw in the surrogates or tried to *rephrase* them. The participants also tended to use **specific** terms with iconographic concepts such as *names*, *location*, and *means* to summarize the video. The same rephrasing pattern was also reported by Song and Marchionini (2007).

In addition, the summaries they wrote by viewing the video surrogates were often **people oriented** (Ding et al., 1999), consistent with Massey and Bender (1996). It seems to be easier to make a story about people, hence participants tended to make a story and put a particular person at the core of the story, which may or may not match the video content.

2.1.2.2 The Effects of Individual and Multiple Modalities

Surrogates of different modalities contribute unique values. Christel et al. (1997) found the visual-based representations (i.e., poster frame) of the video documents led to far faster location of the relevant video than the text-based ones. Goodrum (2001) reported that visual-based surrogates support higher congruence in similarity judgments than do text-based ones, while for specific queries, the text-based surrogates force higher congruence in utility judgment than the visual-based ones. Yet many years of TREC video results in the past have demonstrated that linguistic data generally leads to better performance in video retrieval than visual features (Smeaton et al., 2004).

According to the research on information objects involving more than one modality (e.g., full motion videos), if the information from different modalities is well integrated, the modalities reinforce rather than interfere each other and may lead to increased usability and comprehension. The redundant information from different modalities provides cross-references to the target to be understood (Pryluck, 1976), and offsets weaknesses of one modality with the strengths of another. Ding et al. (1999) confirmed these benefits of combining multiple modalities (i.e., text and image), and found that redundant information simultaneously perceived through the two modalities actually sped up processing time.

The advantages of integrating multiple modalities in full video objects also apply to consuming condensed video surrogates. Ding et al. (1999) found that users viewed key frames and keywords **sequentially** and **selectively**: They may first look at key frames or keywords as a whole, then switch to the other modality. The textual surrogates such as keywords or short abstracts are often used by people to set up the baseline of the story. The visual surrogates such as key frames are often used to reinforce, confirm, and adjust the story. The study concluded that users strongly favor the combined surrogates, in which each modality (i.e., verbal, image) makes a unique contribution to the comprehension of a video, and in combination they reinforce each other.

The unique values and additive effects of the textual and visual surrogates were also confirmed by other studies (Wildemuth et al., 2002; Hughes et al., 2003). Wildemuth et al. (2002) reported that participants used keywords to understand the content of the video, as advance organizers for viewing the visual portion of the surrogate, and as a source of ideas for terms to use in future searches. They commented that textual video surrogates can facilitate the process of determining relevance, and non-textual video surrogates can effectively complement textual surrogates. Hughes et al. (2003) conducted an eye-tracking study and demonstrated that text dominates how people make sense of retrieval sets, while images add confirmatory value. Note that these studies concentrated on text and visual modalities, which both fall into the visual sensory modality. In a more recent multi-modal surrogate study, Song and Marchionini (2007) compared the effectiveness of three different surrogates – visual alone (a storyboard), audio alone (spoken description), and visual and audio combined (a storyboard augmented with spoken description) – for making sense of digital video, and showed that combined surrogates that employ both visual and audio modalities are more effective, strongly preferred, and do not penalize efficiency. According to the study participants, the audio and visual reinforced each other. "With the two (modalities) together, the surrogate is more efficient, and understanding the surrogates becomes simpler than when they are apart." In particular, the audio (spoken description) carries semantic information in video, "gives you a concrete outline of what is going on in the video", while the visual (storyboard) anchors the content, aids memory, and motivates people to watch the video.

2.1.3 Dual Coding Theory

Videos have both visual and audio channels. Watching videos employs two sensory modalities: visual (i.e., seeing) and auditory (i.e., hearing). Watching videos also requires simultaneous information processing in verbal and nonverbal systems of human cognition. Therefore, it is necessary to review works and experiments done in Dual Coding Theory (DCT).

2.1.3.1 Dual Coding of Verbal and Nonverbal Systems

According to Paivio's Dual Coding Theory (Paivio, 1986, 2006), humans have the ability to process information in verbal and nonverbal (i.e., visual) channels simultaneously and separately in two subsystems (i.e., verbal and nonverbal) of human cognition.

As Paivio wrote (Paivio, 2006, p.58):

> "The verbal and nonverbal systems, although functionally independent, must coordinate their activities to achieve common goals... Independence means that the systems can be active separately or together. Cooperation is possible because each system can activate the other via their interconnections."

The cooperation between the verbal and nonverbal systems yields **additive benefits** in some verbal and nonverbal activities, which has been proved in many cognitive psychology experiments. For example, on top of Paivio's finding that "pictures were superior to words as retrieval cues" (summarized in Paivio (2006), p.66), Paivio and Csapo (1973) found that free verbal recall (i.e., without cues) of pictures was higher than free recall of concrete words (e.g. piano, house, etc.), which was in turn higher than free verbal recall of abstract words (e.g., ability, grief, etc.). The explanation is simple from the dual coding point of view. Images might be dually coded in both nonverbal (i.e., as images) and verbal (i.e., as words) systems of human cognition; when we see an image, given enough response time, we name it silently in our mind. When the images are shown too fast, dual coding is prevented because there is not enough time to name the images. Similarly, concrete words might be dually coded as well because they may arouse mental images in our mind. It is generally easier to dually code pictures than words. Abstract words, on the contrary, are unlikely to evoke images in our mind; therefore, they can not be dually coded. The higher the probability of dual coding occurring, the higher are the additive effects. This reasonsing explained the higher recall of the concrete words over the abstract words.

Note that the additive benefits of dual coding are distinct from repetition of a single coding system. The additive effects of dual coding were tested further by Paivio (1975), when free recall proportions were compared for pictures presented once, words presented once, pictures presented twice, words presented twice, and pictures presented once followed by their printed names (or vice versa). It was found that successive repetition of the same code (either words or images) increased the recall more than unrepeated code, but by an amount that was lower than additive benefits from dual coding, where the same items were presented in both images and words. Likewise, it can be predicted that dually coded multi-modal surrogates lead to higher recall than simple repetition of unimodal surrogates.

It is worth noting that the contributions of image (nonverbal) and verbal codes to the additive effect are also different. Begg (1972) suggested that "imagery contributed more than the verbal code to their additive effect", as summarized in Paivio (2006, p.75).

In addition to free recall and cued memory, dual coding also has powerful effects on *recognition memory* and search times. Parallel to the above-mentioned finding that "pictures were superior to words as *retrieval cues*" (Paivio, 1971, p.66, *italics* added), recognition memory was also found to be higher for pictures than for their concrete word counterparts, which is in turn, found to be higher than for recognition memory for abstract words. For example, Shepard (1967) investigated recognition memory of long lists of words, sentences, and images, and showed that, from a list of 600 stimuli per session, images came out on top with 98.5% correct recognition, while 90% of words and 88% of sentences were recognized correctly. The results can be explained by the fact that images are more easily dual coded than the concrete words, and the abstract words can only rarely be dual coded into images. However, the superiority of images discussed above may not directly apply to video surrogates consisting of moving images that are played fast, due to the dual coding difficulty given inadequate consumption time. Likewise, text surrogates should be given enough time for consumption.

Although it is, in general, correct that pictures are superior to words in terms of free recall, cued recall, recognition and search times, Paivio and Begg (1974) found that the results were different for faces, such that it is generally slower to search for human faces than searching for corresponding names. Experiments showed the phenomenon of "verbal overshadowing of visual memories" (Schooler and Engstler-Schooler, 1990). Participants first viewed a video of a bank robbery. Half of the group were then asked to write verbal descriptions of the robber's face, while the other half did some task unrelated to the video. It was found that only 1/3 of the participants who wrote the descriptions were able to correctly recognize the robber's face, as opposed to 2/3 of those who did not write the descriptions. The verbal overshadowing effects can be explained in Dual Coding Theory: verbal coding was aroused when participants were asked to write descriptions, and the verbal descriptions composed by the participants became a new memory trace, which may not have been precise. The imprecise memory of the robber's face may not help and might even hinder the identification of the robber. Therefore, contradicting the usual additive effects of dual coding (i.e., naming images), verbal descriptions (though inducing dual coding) were proven to be unhelpful in remembering faces.

2.1.3.2 Dual Coding of Multiple Sensory Modalities

The above paragraphs discussed dual coding of verbal and nonverbal systems. Paivio (1972) suggested that sensory modalities (for example, visual modality and auditory modality) may have additive effects on recall as well. The additive effects of multiple sensory modalities were confirmed in Thompson and Paivio (1994): hearing the sound and seeing the pictures of audiovisual objects (e.g., phone) yields an additive effect on object recall as compared to only having single modality stimulus. Moreover, the additive effects of multiple modalities were higher than simple repetitions of the same modality.

The additive effects of dual coding were also achieved in other sensory modalities. Lyman and McDaniel (1990) found that seeing pictures (i.e., visual) and smelling odors (i.e., olfaction) at the same time led to higher recognition memory and free verbal recall of odors, while having auditory, visual, and olfaction altogether could further increase the recall.

2.1.3.3 Criticisms of Dual Coding Theory

Despite the success of Dual Coding Theory, there have also been a lot of criticisms of DCT. Some criticisms focus on the inconsistent experimental findings, for example, pictures are not always recalled better than words, especially with children (Dilley and Paivio, 1968). Paivio (2006) suggested this might be explained by the labeling difficulties by children. Experiments done by Cole et al. (1971) supported Paivio's explanation about the inconsistency. Cole et al. (1971) explicitly asked the children in Grades 1 through 8 to name pictures, and found that the children indeed recall images better than words, due to the additive effects of verbal and nonverbal dual coding. Also, as summarized in Paivio (2006), when people age, the superiority of images over words will vanish, such that for old people, the advantages of dual coding for images over words may not be found.

Criticisms of DCT also seek alternative attributes or explanations of concrete words superiority to abstract words, or theoretical alternatives to DCT. For a detailed summary of the critiques and rejoinders of DCT, please refer to Paivio (2006, chapter 4, p.82-86).

2.1.3.4 Applications of Dual Coding Theory

Dual Coding Theory has applications in many cognitive domains, and has been used by instructional designers as a theoretical basis for multimedia materials. One DCT hypothesis is that the verbal and nonverbal codes, being functionally independent, can have additive effects on human recall. Audio accompaniment has been found effective in complementing visual cues and "auditory information, when designed to complement the visual environment, is natural and people are innately comfortable with it - its use requires no training" (Gunther et al., 2004, p.435). For example, Kulhavy et al. (1993) conducted two experiments where undergraduates studied a city map and then heard a text which was associated with the map features, and found that having visual and verbal stimuli together added memory for structural properties of a map. Mayer and Moreno (1998) also offered strong evidence for using narrated text with graphics rather than on-screen text with graphics, which they termed the "modality effect". Watching videos with one or more synchronized audio track(s) also leads to increased recall over consuming just the visual or the audio, and observations of the benefits of multi-modal surrogates have also started to be reported (Ding et al., 1999; Goodrum and Spink, 2001; Wildemuth et al., 2002; Boekelheide et al., 2006; Song and Marchionini, 2007; Marchionini et al., 2009).

2.1.4 Summary

This section reviews the previous works on how people make sense of videos and video surrogates. Viewing images and videos is both a sensory and a mental experience, and various models that pose multiple levels of image comprehension and movie spectatorship were reviewed (Panofsky, 1955; Shatford, 1986; Jaimes and Chang, 2000; Eakins and Graham, 1999; Greisdorf and O'Connor, 2002; Ponech, 1997). When people watch videos, they may screen out some information and pay selective attention to only some stimuli. The way people perceive videos is also influenced by cultural and sociological aspects.

Past work shows that people can make sense of videos through abbreviated video surrogates. According to Dual Coding Theory, auditory information, when designed to complement the visual environment, can have additive effects on human recall. It is already known and widely accepted that, for primary information objects, coordinated media channels lead to better understanding, retention, and satisfaction. However, there are no safe and sound conclusions on whether this coordination requirement for primary information objects also applies to highly abbreviated, condensed information objects, such as video surrogates. For video understanding through surrogates, if the audio and visual channels of the surrogates are sampled from the full video independently such that they may not be synchronized, it is possible that the unsynchronized audio and visual channels convey more information about the video per unit time than if the two channels are synchronized. Therefore, it is important to investigate the synchronization necessity between the audio and visual channels of the surrogates, as well as to examine whether multi-modal surrogates with unsynchronized audio and visual channels are more effective than synchronized ones.

2.2 Automated Video Summarization

As more digital videos become available online, it is imperative to give users effective summarization and skimming tools to facilitate finding and browsing of the videos. Surrogates are condensed information for representing the full information objects. Video surrogation, a mechanism for generating this condensed information for full videos, is needed for effective and efficient video acquisition from and indexing for large collections of videos. Video summarization, or video abstraction, as their names imply, are mechanisms for generating short summaries of videos. Therefore, those names are closely related and dependent, if not completely interchangeable.

There is a great deal of work on summarization techniques for text, audio, and videos (e.g., Abracos and Lopes, 1997; Chen and Withgott, 1992; Kennedy and Ellis, 2003; Li et al., 2004). A number of summarization techniques for text documents utilize the tf-idf (term frequency-inverse document frequency) schema or its variations to find the most informative and representative sentences to form text summaries (Abracos and Lopes, 1997; Li et al., 2004).

There are several different approaches for summarization of audio and videos. One approach looks at the text transcripts of audio or video, and uses summarization techniques for text documents to generate text summaries for the audio or video. Then the audio and video segments corresponding to the text summaries can be selected to form an audio or video summary. Another approach extracts audio summaries from presentations or (single- or multi-user) speeches based on audio features such as speech emphasis, pitch, excitement level, and so on. For videos, in particular, the visual features can be exploited to create video summaries.

In this Section, previous and current techniques for automatically making video summaries and video abstracts are reviewed. Since audio and video are closely related, and the audio stream is an important and non-neglectable part of video, some work in automated audio or speech summarization is also reviewed.

Various approaches have been investigated for summarizing videos. Yahiaoui et al. (2003) classified existing video summarization approaches in two main categories: *rule-based* approaches and *mathematically oriented* approaches. The former combine evidence from several types of processing (i.e., audio, video, text) to detect certain configurations of events to include in the summary (Christel et al., 1998; Lienhart et al., 1997). The latter use similarities within the video to compute a relevance value of video segments or frames, for example, using singular value decomposition (Gong and Liu, 2003) and segment importance measures (Uchihashi et al., 1999).

This section categorizes and reviews some existing summarization approaches according to the type of data stream they mainly process.

2.2.1 Video Summarization via Temporal/Spatial Compression

2.2.1.1 Time Compressed Video

2.2.1.1.1 Speeding up video One intuitive way of making compact video surrogates to cover all information contained in the video is to simply speed up the video clip. However, the increased playback speed often comes with decrease in comprehension due to the audio distortion (i.e., a degradation of the speech signal) and a processing overload of short-term memory. Therefore, compression of this kind is limited to a maximum compression factor of 1.5-2.5 depending on the particular program genre and speech speed (Heiman et al., 1986), beyond which the speech audio becomes perturbing and incomprehensible.

2.2.1.1.2 Dropping short sequences One practical way to time-compress the audio is to remove redundant information from the speech signal. The *sampling* methods drops short segments from the speech signal at regular intervals. For example, for the original sequences {1,2,3,4,5,6,7,8,9} of 50 milliseconds each, short sequences {2,4,6,8} can be dropped. By dropping alternate chunks of speech from the original signal, 2x compression can be achieved. Unfortunately, this results in an increase in pitch, making the audio less comprehensible and enjoyable.

An variant of the sampling methods is *dichotic* sampling, where different audio segments are played to each ear. For example, for original sequences {1,2,3,4,5,6,7,8,9} of 50 ms each, short sequences {1,3,5,7,9} are played to the left ear, and short sequences {2,4,6,8} are played to the right ear. *Dichotic* sampling takes advantage of the auditory system's ability to integrate information from both ears (Arons, 1997), which increases intelligibility and comprehension of the compressed audio when compared with the standard sampling methods (Gerber and Wulfeck, 1977).

2.2.1.1.3 Pause shortening or removal In addition to speeding-up the video and dropping short sequences, removing or shortening pauses can be used to further reduce 15%-20% playback time without compromising content (Gan and Donaldson, 1988). Simply removing all pauses in speech results in speech that is "natural, but many people find it exhausting to listen to because the speaker never pauses for breath", as stated in Neuburg (1978). There are two categories of pauses in speech: *Juncture pauses*, average 500-1000 ms, which are under talkers' conscious control, usually occurring at major syntactic boundaries; and *Hesitation pauses*, averaging 200-250 ms, which are not under talker control (Minifie, 1974). Lass and Leeper (1977) suggested that when time compressing the speech, juncture pauses can not be removed or shortened without interfering with comprehension. For example, Arons (1997) time compressed speech audio, such that the pauses are selectively shortened or removed. In particular, pauses less than 500 ms are removed, and pauses more than 500 ms are shortened to 500 ms. With these thresholds, speech audio is sped up while providing the listener with cognitive processing time as well as the pace of the utterance.

However, time compression via speeding-up and/or pauses shortening or removing, even when used together, can hardly lead to compaction rates of more than 2:1 (Arons, 1997). In many real-life video retrieval or audio/video summarization applications, a compaction rate of 10 and above is desirable.

To further reduce the playback time of the audio, **skimming** techniques can be used. For instance, if an audio clip takes 60 seconds to play at normal speed, it may take just 30 seconds when time compressed, while only takes

5 or 10 seconds with higher levels of skimming techniques. The following paragraphs review existing skimming techniques for summarizing videos.

2.2.1.2 Systematic Subsampling Video

Another simple and straightforward method for creating video summaries would simply increase the frame rate across the whole video. A computationally expensive way to get a two-fold video speed-up, is to render the frames at twice the original frame rate. This puts burden on a client's CPU, which has to decode twice as many frames in the same amount of time.

On the other hand, the fast forward, a common summarization approach used in many video retrieval systems, is performed by taking every Nth frame from the original video, and concatenating them as a summary to be played at normal speed.

Fast forwards with audio is equivalent to the *time compressed video* by sampling discussed above: Every Nth image frame is extracted from the visual stream, and the audio stream is time-compressed at the same compaction rate. This approach can not decrease the viewing time by more than five-fold without seriously degrading the audio coherence.

For fast forwards with no audio, the audio stream is not provided with the visual fast forwards. Wildemuth et al. (2003) reported on a study of the use of fast forwards for digital video, and recommended a fast forward default speed of 1:64 of the original video. Although this approach can achieve a much higher compaction/compression rate than fast forwards with audio, yet it can lead to severe coherence degradation and discomfort to the viewer.

Instead of taking the every Nth frame of the video, video summarization can be simply performed 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, 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. This subsampling summarization method by keeping and skipping frames at fixed intervals, will likely produce discontinuities at the interval boundaries and exclude essential information from the summary (Wactlar et al., 1996). To improve the quality of the summaries based on subsampling techniques, a windowing function or smoothing filter, such as a cross-fade, can be applied at the junctions of the selected segments (Omoigui et al., 1999).

Although the summaries created by systematic subsampling are likely subject to exclusion of important segments, they are easy and inexpensive to implement. Therefore, subsampling is often adopted as the default or baseline method in evaluating other automated video summarization techniques (Christel et al., 1998).

2.2.1.3 Split-screen Display

Instead of doing time compression and systematic subsampling, some summarization techniques reduce video playback time by displaying multiple video streams at the same time.

One participating group of the 2007 TRECVID, Institut EURECOM, presented a summarization approach where the most important and non redundant shots selected to appear in the summary were dynamically accelerated and optimally grouped into sets of four and presented simultaneously using a split-screen display, so as to maximize the content included in the summary per time unit. However, the resulting summaries increased the viewers' cognitive load greatly and did not rate highly with the evaluation campaign assessors in terms of ease of use.

2.2.2 Audio-based Video Summarization

The audio stream is an important and non-neglectable part of video. For some videos, especially these featuring presentations or conversations, a great deal of important information is contained in the audio stream. Therefore, not only have the audio summarization techniques focused on processing audio (Chen and Withgott, 1992; Cooper and Foote, 2002a), but also a number of video summarization techniques have also taken the approach of processing the audio stream of video (Chen and Withgott, 1992; Arons, 1997; Taskiran et al., 2002).

2.2.2.1 Emphasis/Pitch Detection

Emphasis or pitch detection is an often used technique for summarizing audio and video. Some videos, such as instruction or presentation videos, are dominant by a talking head, and the important information is mostly contained in the audio stream. Even for sports videos, where the visual play of actions are more attractive to the viewers than the audio, the video highlights of the games are often accompanied by excitements and peaks in the audio stream.

Chen and Withgott (1992) focused on creating summaries for natural, conversational speech such as recorded telephone or interview conversations. A discrete density Hidden Markov Model (HMM) was used to recognize emphasis in the speech, and groups of emphasized words in phrases are selected to form audio excerpt summaries. In other words, acoustic phrases in which emphasis occurred in close proximity are used to form a summary. Experimental results were quite promising and showed that the automatically generated summarizing excerpts have no noticeable difference when compared to human selected excerpts, which suggests that speech emphasis may be useful for summarizing animated conversations or videos dominated by these conversations.

SpeechSkimmer, a speech skimming system developed at the MIT Media Lab (Arons, 1997), used both time-compression and removing portions of the audio. SpeechSkimmer skims speech recordings via time compressed speech, pause shortening, automatic emphasis detection, and non-speech audio feedback. SpeechSkimmer allows audio to be played at multiple levels of speed and detail. Specifically, level 1 is the original audio. Level 2 is the pause-shortened audio. At level 3, pause-based skimming was performed based on a simple heuristic: long juncture pauses tend to indicate either a new topic, some content words, or a new talker. Therefore, only the speech that occurs just after a significant pause in the original recording is included in the summary. Level 4 is based on pitch-based emphasis detection. Moreover, recorded non-speech sound effects were provided as navigation cues, for example, a short tone is played when the user transitions to a new skimming level. The higher the skimming level, the higher frequency of the tone. With the SpeechSkimmer interface, users have continuous real-time control of the speed and detail level of the audio: the speech content can be played at normal speeds, with pauses removed, or restricted to phases emphasized by the speaker.

2.2.2.2 Excitement Levels Detection

For sports videos, the important events are often accompanied by great audience excitement and sharp increase in the audio volume, and video summarization can be performed by detecting the sudden changes in excitement levels (Cabasson and Divakaran, 2003; Li et al., 2003). This approach actually falls within the scope of video summarization based on specific domain knowledge (i.e., sports video), which is discussed in Section 2.2.6.

For videos of informational presentations and seminars, video segmentation can be performed based on *audio pause boundaries*. For example, Taskiran et al. (2002) segmented speech by detecting large inter-word pauses of the speaker in the audio based on the timecode of the extracted transcript. This segmentation method avoids having very long segments selected for the summary. Then the emphasized segments can be selected to form video summaries.

2.2.3 Text-based Video Summarization using Closed-captions or Transcripts

For news programs, instruction or presentation videos, and teleconferences, where the camera is fixed on the speaker for a long time, a large portion of the important information is contained in the audio stream. Therefore, an intuitive and practical approach of summarizing videos is based on analyzing the speech text transcripts (Christel et al., 1996). Closed captions are readily available for most broadcast videos, like news programs. For other video genres, such as presentations and teleconferences, where closed captioning is not available, automatic speech recognition (ASR) techniques can be used to generate the speech transcript.

Agnihotri et al. (2001) presented a summarization system for generating summaries for talk shows using the closed-caption text. The system extracts and analyzes closed-caption text of the talk show videos, uses cue words and domain knowledge of program structure to determine the boundaries of individual guests of the talk show and commercial breaks, and then creates a program summary. The authors experimented with their summarization system with seventeen hours of closed-caption data, and evaluated the system in terms of precision and recall. The summarization system produces high level summary information and a table of contents indexed by topics. The recall of finding the guests in a talk show is 93% (i.e., 25 out of 27 guests in the talk shows were correctly identified), while no guest was incorrectly identified (i.e., precision is 100%).

Taskiran et al. (2002) proposed an algorithm, referred to as FREQ in Taskiran et al. (2006), to automatically generate video summaries based on video transcripts. The FREQ algorithm generates summaries based on word-frequency, word co-occurrence, and dispersion scores derived from program segments. The videos were first segmented into a number of segments based on long inter-word pauses. Then the words in a segment are scored based on a method related to *tf-idf*, and each segment is scored by summing the scores for all words contained in the segment. The *log-likelihood* ratio was used for detecting significant co-occurring words in the program to identify important phrases. To manage the tradeoff between detail and coverage of the summaries while maximizing the coverage of the summaries, a measure of similarity dispersion over the whole video program was derived, where small dispersion is wanted when summaries are clustered in the full video, and large dispersion is wanted when summaries are distributed uniformly across the video. In each iteration of the greedy algorithm for selecting segments, the segment yielding the greatest increase in the dispersion value of the current summary is selected to be included in the summary, until the summarization ratio of 0.1 is reached.

Taskiran et al. (2006) designed a user study to compare the quality of the FREQ generated video summaries and the quality of summaries generated using two other algorithms, RAND and DEFT, which do not utilize word-frequency or dispersion scores (Taskiran et al., 2006). The FREQ algorithm has reliable performance even with transcripts obtained by ASR which has a high error rate. The FREQ algorithm was found to be statistically significantly better than RAND and DEFT in terms of the number of correct answers out of the 10 multiple choice questions, and the number of answers contained in the summaries. The study makes a great contribution in suggesting the use of video transcript to generate video summaries, and further suggested considering generating summaries using more modalities other than just the transcript in future studies.

Also note that Taskiran et al. (2006) used error prone speech transcripts from ASR in the FREQ algorithm to automatically generate summaries. If highly accurate transcripts such as closed-captioning are available, we can expect the automated video summaries using transcript will perform even better. The state-of-art ASR techniques, however, is not sufficient to be used solely to generate closed-captioning with high accuracy. Martone et al. (2004) proposed an algorithm for generating automated closed-captioning using text alignment. The algorithm aligns video transcripts with no time codes with ASR output containing time code for each word. With this technique, if the program transcript is available, highly accurate closed-captions can be automatically generated efficiently, and more effective video summaries can be created from the speech transcript.

Another example of video summarization based on transcripts is the MAGIC (Metadata Automated Generation for Instructional Content) system developed at IBM, which utilizes various content analytics tools to automatically generate metadata for instructional video content (Li et al., 2005). The audiovisual analysis modules recognize semantic sound categories and identify narrators and informative text segments, while the text analysis modules extract title, keywords and summary from video transcripts. In particular, the text analysis tools extract a document title, a set of keywords (ranked by frequency and/or ranked from the most specific to the most generic), topic shift boundaries, and a summary description comprising a few important sentences from the video transcripts.

2.2.4 Visual-based Video Summarization via Frame Clustering

Another common strategy in summarizing videos is to segment the videos and extract one or more keyframes from each segment. Then the extracted keyframes can be concatenated to form static or dynamic summaries.

Video segmentation can be done by *detecting shot boundaries* (Hampapur et al., 1994) or by *detecting changes in the dominant image motion* (Peyrard and Bouthemy, 2003). A shot is an image sequence which presents continuous action from a single operation of the camera without an editor's cut, fade or dissolve. The shot boundary detection can be performed using visual features, such as color histograms (Nagasaka and Tanaka, 1992), and special effects like fades and dissolves (Zhang et al., 1993).

Video segmentation can also be done by *detecting important events*, e.g., cheering crowds and goals in sports videos or guests in the talk shows. Because some video programs contain easily identifiable important events, it is possible to summarize the videos by exploiting domain knowledge for identifying these events. In fact, summarization using domain knowledge will be discussed in Section 2.2.6.

Once the video is segmented, one or more keyframes can be extracted from each segment. The keyframe extraction is generally determined based on visual features, such as clustering using color histograms (Uchihashi et al., 1999; Ratakonda et al., 1999). More recently, two stage clustering techniques have been used to extract keyframes and form video summaries, as done in (Hanjalic and Zhang, 1999; Farin et al., 2002; Ferman and Tekalp, 2003). These works contained similar two-stage clustering structures. For example, Ferman and Tekalp (2003) first selected a non-redundant set of keyframes from each shot using the fuzzy c-means algorithm (a variation of the K-means clustering method) and data pruning methods. The number of keyframes selected from each shot is determined through cluster validity analysis. In the second stage, the number of keyframes from each segment can be reduced and clusters can be merged according to user browsing preferences, so as to update the size of the video summary.

With the extracted keyframes, static summaries like the storyboard can be created. Dynamic summaries can be created too. 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 (Taskiran et al., 2002).

Cooper and Foote (2002b) constructed video summaries based on *similarity analysis*. First, one frame is extracted from each second of the video, and a color feature vector is extracted from each extracted frame. Then a non-negative similarity matrix is calculated by comparing the feature vectors of each pair of frames based on low-order *discrete cosine transform* (DCT) coefficients. The most representative contiguous segments of the video are determined by summing the columns of the similarity matrix. The non-negative matrix factorization (NMF) of the similarity matrix is employed to determine the essential structural components of the video to be included in the summary while avoiding redundancy.

Gong and Liu (2003) used the *singular value decomposition* (SVD) of a 3-D RGB histogram frame feature matrix to construct video summaries. They segmented the whole video sequence into several "clusters", each of which contains "visually similar" shots. The longest shot from each cluster was then chosen and concatenated, in the time order, to form the video summary. The SVD was used to reduce the dimensionality of the frame feature vectors (i.e., from 1125 dimensions down to 150 dimensions) for clustering.

2.2.5 Video Summarization through Redundancy Detection

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 works.

For instance, Yang and Hauptmann (2006) 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 visually relevant shot. Van Gemert et al. (2006) used the visual repetition of the same item in multiple shots in concept detection. Huurnink and de Rijke (2007) explored visual redundancy over many consecutive shots and developed a framework to incorporate redundancy for cross-channel retrieval of visual items using speech. The models were tested in a series of retrieval experiments and it was found that incorporating redundancy in cross-channel video retrieval leads to significant improvements in retrieval performance. Hauptmann et al. (2007a) 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.

2.2.6 Video Summarization using Specific Domain Knowledge

Some special genres of videos, like sports video, talk shows, news videos and audio-video presentations, have special characteristics, such that specific domain knowledge can be exploited in the summarization algorithms.

2.2.6.1 Sports Videos

For a lot of viewers, interesting events in the soccer games are 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. Li et al. (2003) proposed a general framework for indexing and summarizing sports videos. The framework includes a unifying model based on *automatic events detection* for modeling both action-and-stop sports (e.g., baseball and American football) and continuous action sports (e.g., soccer and ice hockey) according to domain-specific knowledge of these sports. Event detection is performed based on a *heuristic*: An exciting action is usually replayed by the broadcaster, preceded by close-up shots of the key players, the audience, the coach, or the referee, after the live version of the action is played. The model was successfully applied to soccer videos, and the exciting events were detected using low-level visual and aural features, such as the dominant color of the shot, a sudden scene cut between the live action and the close-up shots, and the increased level of audience excitement in the audio track.

Cabasson and Divakaran (2003) used the temporal patterns of motion activity (captured by MPEG-7 motion activity descriptor) around each audio peak to detect and capture interesting events in soccer videos. The *heuristic* is: interesting events such as goals in a soccer match are usually associated with a sharp increase in audio volume, and often lead to an interruption of the game for a non-trivial duration. This approach captures important events like goals as well as several other interesting events that such as attempts at goals and major injuries. The method is computationally simple and flexible, and results indicate that the scheme works well for soccer games from different parts of the world including a women's soccer game.

Ekin and Tekalp (2003) proposed a fully automatic and computationally efficient framework for analysis and summarization of soccer videos using *cinematic* and *object-based* features. Three types of summaries tailored very specifically to soccer games were output and evaluated, including (1) all slow-motion segments in a game, (2) all goals in a game, and (3) slow-motion segments classified according to object-based features (e.g., referee and penalty box). The efficiency, effectiveness, and the robustness of the proposed framework are demonstrated over a large data set, consisting of more than 13 hours of soccer video, captured at different countries and conditions.

2.2.6.2 Audio-visual Presentation Videos

He et al. (1999) focused on automatic summarization of audio-video presentations, i.e., informational talks given with a set of slides. Given the nature of these audio-video presentations, new sources of information such as *slide-transition timing* and *user-access logs* may be exploited, in addition to traditional sources such as pitch-based emphasis detection. The automatic summaries were created using three different algorithms: (1) using slide-transition only; (2) identifying emphasis in speech by pitch activity analysis (Arons, 1994); (3) using slide-transitions, pitch activity, and user access patterns (as measured by the ratio between the average-user-count of the current slide and the average-user-count of the previous slide). For all three algorithms, pause detection (Gan and Donaldson, 1988) was used to ensure that audio segments in the summary did not begin in the middle of a phrase.

Furthermore, He et al. designed a user study of 24 participants to compare these three types of automatic summaries to manually generated summaries by the authors of the talks or someone regarded as qualified by the authors. Participants were given pre-study quiz questions before they watch any of the summaries to measure their expertise in the topic areas of the talks. After watching the summaries, participants were given preference surveys to evaluate the summaries along four dimensions: **conciseness**, **coverage**, **context**, and **coherence**. To determine whether the automatically generated summaries had captured the key content of the talk identified by the author, participants were asked to answer pre-study quiz questions before watching any summaries and author-generated quiz questions after watching the summaries.

The questions required participants to draw inferences from the summary or to simply relay information contained in the summary. For the post-summary quiz questions, participants did significantly better with the author generated summaries than with the automatically generated summaries. Nonetheless, all post-summary quiz scores increased significantly comparing to the pre-study quiz scores, even when the automatically generated summaries were played to the participants, suggesting the automatically generated summaries increased the participants' knowledge in the topic areas of the talks. Furthermore, there was no statistically significant difference among the automated methods: using slide-transitions, pitch activity, and user access patterns all together did not generate better summaries than using just slide-transitions or pitch activity.

As for the preference ratings, the author generated talk summaries were rated significantly more favorably than automatically generated summaries though many participants were surprised when being told afterward that some summaries were computer generated. No significant differences were found between users' preferences for the three automatically generated summaries. Particularly, the automatically generated summaries were rated poorly on coherence, but people quickly got used to them. Significant **habituation effects** were observed: Participants' perceptions of summary quality improved over time. The participants were also asked to rate their confidence on whether the summary had covered key points in the talk. Confidence ratings were higher with the author generated summaries, whereas the automatically generated summaries also got a respectable rating of 60%. Moreover, when being asked whether they would skip the talk based on the summary, participants were more convinced that they could skip the talk after hearing author generated summaries. In short, people learn from the automatically generated summaries, although less than from author generated ones. Though people initially find the automatically generated summaries less coherent than the author generated ones, yet they quickly grow accustomed to them, which suggests these automatic summarization techniques may be ready to serve in real-life systems.

2.2.6.3 News Videos

(Gong, 2003) presented an audiovisual summarization system for summarizing news, documentaries, and seminars. The system summarizes the audio and visual contents of the given video separately, and then integrates the two summaries with a **partial** alignment. The audio summary is achieve by selecting spoken sentences that best present the main content of the audio track, and the visual summary is generated by eliminating redundancies and preserving visually rich contents in the visual stream. The system uses a Bipartite graph-based alignment algorithm to align each spoken sentence in the audio summary with its corresponding visual segment displaying the speaker's face and fill the remaining period of the visual summary with other image segments. Since the audio and visual summaries are performed independently, the method maximizes the information coverage for both audio and visual contents of the original video.

Lie and Lai (2004) proposed an algorithm for summarizing news programs. A news program is generally composed of alternative concatenation between anchor shots and news segments. The anchor audio enables the viewers to understand each piece of news in the most efficient manner. Thus the algorithm first performs shot boundary detection and shot classification to detect the anchor shots, based on which the anchor audio is retrieved accordingly. Then the visual parts of the news segment are summarized by classifying shots into special and normal events through the analysis of spatial and motion features and assigning different time-allocation weighting to shots, subject to the length of the corresponding anchor shot. The news video summary is presented in such a way that the anchor audio is overlaid with the visual summaries for news sequences, allowing the viewers to understand the story headlines as well perceive motion activity of the story.

These methods (Gong, 2003; Lie and Lai, 2004) have proven to be effective for summarizing a special genre of videos, i.e., the news programs, which have structured audio and visual features. The techniques may not be generalizable to other video genres where the programs have different structures.

2.2.7 Video Summarization via Multi-modal Integration

Although the methods discussed above fall into different categories, most of them still focus on processing single data stream, either text, or audio, or visual, with a few exceptions (He et al., 1999; Li et al., 2003; Gong, 2003; Lie and Lai, 2004). By combining approaches from more than one modality, video summarization has the potential to be performed with better coverage, context, and coherence.

Lienhart et al. (1997) 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 some 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 some dissolves and wipes to make the final movie trailer sequence.

Foote et al. (2000) employ *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". The combined visual surrogates and audio excerpts comprise a lightweight web interface for browsing digital videos and are not expensive to create. However, the interface requires users to click on each keyframe to play the corresponding audio, which can make users get bored quickly, and it may be time-consuming to play all the audio excerpts extracted that associate with the key frames. Since high compaction/compression rates are desired for good browsing interfaces, the proposed interface may not be very practical for large video libraries. Also, the interface may not easily adapt to limited display on small form factor devices. As we think more about video browsing on different devices, we would like to have browsing interfaces or surrogates as compact (in both time and space), and informative as possible which can be useful on wide ranges of devices.

Erol et al. (2003) described a multi-modal scheme for automatically summarizing meeting videos based on *audio and visual event detection* together with *text analysis*. Text analysis is performed using the *tf-idf* measure, audio activity analysis is based on sound directions and the magnitude of the audio signal, and significant visual events in a meeting are detected by analyzing the localized differences of luminance values in the compressed domain. The audio and visual events and the keyword segments are sorted according to the activity scores and the tf-idf measures respectively. Then the top N important audio, visual, and keyword segments are concatenated in the time order to create meeting summaries. Padding as well as merging temporarily close segments are performed to keep the summaries more comprehensible.

Mihajlovic et al. (2007) presented a case study showing how important events (highlights) can be automatically detected in video recordings of Formula 1 car racing using their multi-modal content-based video retrieval techniques. The techniques processed information from three information sources, i.e., the audio signal (consisting of human speech, car noise, crowd cheering, horns, etc.), the image stream (containing important events such as passing, start, fly-out, and replay), and superimposed (overlay) text in the video (e.g, names of drivers, position in the race, and lap information). The multi-modal techniques can automatically derive interesting events in Formula 1 car racing videos, and help answer queries like “In which lap did Schumacher make a fly-out?” (Mihajlovic et al., 2007, p.292).

As a part of the CMU Informedia project, Smith and Kanade (1997) produced automatic video summaries (i.e., skims) through image and language understanding to extract specific objects, audio keywords and relevant video structure. For language understanding, phrases with total tf-idf values higher than a fixed threshold are selected as keyphrases, and the audio excerpts are extracted according. Image analysis is powered by segmentation of video into scenes, detection of face and superimposition text on screen, and analysis of camera motion. If a scene contains both faces and text, it is likely that an important person is being introduced, and the segment containing text is included in the summary. For each keyphrase, the characterization results of the surrounding video frames are analyzed, and the most appropriate set of frames for skimming may not align with the audio in time. The resulting summaries have a compaction rate as high as 20:1, and yet retain the essential content of the original videos, illustrating the potential power of integrated language and image information for video summarization.

Also part of the Informedia project, Christel et al. (1998) compared video skimming techniques that used (1) tf-idf measure and audio analysis based on audio amplitude, (2) audio analysis combined with image analysis based on face/text detection and camera motion, and (3) systematic subsampling of video sequences. They reported that audio analysis combined with visual analysis yield significantly better results than the skims obtained purely by audio analysis or uniform sampling. Both Erol et al. (2003) and Christel et al. (1998) showed that using transition effects between different segments in video summaries leads to better comprehension of the summaries.

The aforementioned news summarization techniques proposed in Gong and Liu (2003) and Lie and Lai (2004) are also examples of video summarization via multi-modal integration. Gong and Liu (2003) summarize the audio and visual contents of the given video separately, and then integrate the two summaries with a partial alignment. Since the audio and visual summaries are performed independently, the method maximizes the information coverage for both audio and visual contents of the original video. Lie and Lai (2004) also summarize the audio and visual content separately. The anchor audio is retrieved by shot boundary detection and shot classification, and important visual segments are extracted by the analysis of spatial and motion features. The multi-modal summarization allows the user to understand the story headlines as well perceive motion activity of the stories.

Similar to Lie and Lai (2004), Kim et al. (2004) proposed a method for automatically summarizing news videos by multi-modal content analysis. The method exploits the closed captions as a key source to locate semantically meaningful news highlights, and uses speech signals in an audio stream for synchronization between the closed captioning text and video.

In the 2007 TRECVID summarization campaign, several participating groups submitted video summaries with multi-modal approaches (Over et al., 2007). For example, the AT&T labs proposed a video summarization system which relied on speech and face detection (Liu et al., 2007). The video was segmented into shots and within each shot, one continuous segment containing the most speech and face occurrences is selected to be included in the final video summary. Brno University of Technology in the Czech Republic created their summaries using shot boundary detection and removing junk frames, and the summary consists of thumbnails with extra textual information such as shot duration (Herout et al., 2007). The Hong Kong Polytechnic University, used shot bound detection to structure the rushes video, followed by detection of noise shots and blank shots using a combination of visual and audio features (Ngo et al., 2007). Details of the approaches by 17 out of the 22 TRECVID 2007 participating groups can be found in papers in the proceedings of the 2007 TRECVID workshop (available online: http://www-nlpir.nist.gov/projects/tvpubs/tv.pubs.org.html#2007).

2.2.8 Summary

This section reviews a few main categories of existing video summarization approaches. Some approaches are text-based, i.e., summarizing videos based on the video transcript. Some are audio-based, i.e., summarizing videos by detecting emphasis or pitch activity, or detecting excitement levels in the audio stream. Some are visual-based, e.g., shot boundaries are detected, and frames are clustered based on features like color histograms. Some techniques utilize specific domain knowledge or video redundance, for example, sports videos, and news videos which have special video structures. Some techniques simply do compression in the temporal or spatial dimensions, e.g., the *time compressed video*, *systematic subsampling*, or give parallel presentation like *split-screen display*. And some approaches combine approaches from more than one modality, i.e., text, audio, or visual, and therefore have potential to generate multi-modal summaries with better coverage, context, and coherence. Also note that the *time compressed video* and *split-screen display* approaches may be combined with the other approaches to further condense the video summaries.

2.3 Video Retrieval through Metadata and Surrogates

With the explosive growth of the WWW and web video content, searching, browsing, and retrieving relevant video segments from a large archive of digital videos has become a challenging task. One of the mainstream approaches of video retrieval is to use the linguistic cues (i.e., metadata) to index and then to retrieve the videos. Video retrieval can also be done by querying example video clips (Dimitrova and Abdel-Mottaleb, 1997), which falls into the scope of content-based video retrieval. Another common approach of video retrieval is to browse the video collection for relevant videos instead of directly searching in it. For example, browsing through a set of keyframes of a video can give viewers the gisting of the content of the video quickly. Here we refer to this approach as video retrieval through surrogates. This section reviews existing work in video retrieval.

2.3.1 Video Retrieval through Metadata

Linguistic cues, i.e., metadata, are commonly used to index videos. Video retrieval can be done by querying text metadata associated with the video records. The approach is called *concept-based* video indexing, as opposed to *content-based* indexing, where colors, shapes, and textures are analyzed to index videos.

2.3.1.1 Metadata: Definition and Purposes

Metadata is often defined as “data about other data”, “information about information”. The American Library Association collected 27 submitted definitions on metadata, and devised a more sophisticated working definition:

> “Metadata are structured, encoded data that describe characteristics of information-bearing entities to aid in the identification, discovery, assessment, and management of the described entities.” (American Library Association, 2000)

Metadata provide context information for data of any media sort. Examples of metadata regarding a book may include the book title, author, date of publication, publisher, ISBN (International Standard Book Number), language, product dimensions, number of pages, and so on. Examples of metadata for a digital audio file such as an MP3 typically include the album name, song title, artist, genre, year, track number, and so on.

Metadata can either be free text description and keywords, or take the form of controlled vocabularies. JISC Digital Media (formerly called TASI – Technical Advisory Service for Images) summarized a variety of purposes of different metadata types (JISC Digital Media, 2006):

- **Administrative metadata**: describe how an information-bearing entity is managed or to control access to it;

- **Descriptive metadata**: describe the intellectual content of the object, what the information-bearing entity is;

- **Structural metadata**: help us to relate this digital resource with other resources;
- **Technical metadata**: describe how the digital resource was created;
- **Resource discovery metadata**: help us to find the resource;
- **Provenance and rights metadata**: describe where the resource has come from, who owns it and how it can be used.

In the case of digital assets like imagery, metadata is usually used to describe the creation, content, or context of an image (JISC Digital Media, 2006), such as *creator*, *title*, *creation date*, *subject matter*, and *location* of where the image is taken or where the image is kept or shown (JISC Digital Media, 2006; Metadata Working Group, 2009). Metadata for other multimedia objects such as videos, can be created alike. In particular, metadata for digital videos may be used to locate the relevant 5 minutes from thousands of hours video content.

Metadata can describe an information object at different levels granularity or layers. For the painting "Mona Lisa", we can have metadata to describe the original art painting, the replica, or a digital representation of the painting. In addition, not only can we describe individual resources (i.e., single items) like an image, an MP3 audio file, or a video, but we can also describe the resource collections (i.e., the aggregations of resource items), for example, a photo album, a video collection, and so on. We can also use metadata to describe just a portion of a larger whole, e.g. a particular scene or even a single frame from a full video (JISC Digital Media, 2006).

2.3.1.1.1 Why do we need automatic metadata? As summarized in JISC Digital Media (2006), metadata can come from "two sources: (1) it can be automatically derived from the digital resource itself, or (2) can be created and associated with a resource by human beings". In fact, the metadata generation can be a combined approach involving both the automatic process and human efforts. In many situations, text descriptions associated with the video are unavailable or incomplete. Not only is manual annotation of videos a costly, tedious, time-consuming, and error-prone process, but also people may not be willing to take the time and energy to manually label all their videos (e.g., hours of home videos taken with camcorders) with text descriptors. The same problem and difficulties exist for broadcast video programs, where even detailed transcript text such as the closed captions may not include all features shown in the video, such as important visual events and superimposition text.

To ensure easy sharing of, access to, and searching over the fast proliferating online images and videos, a large amount of research has been done on automatic video annotation (Adjeroh and Lee, 1995; Ide et al., 2001; Snoek and Worring, 2005), such that a computer system automatically assigns keywords to a digital image or video. Because a video consists of a series of moving images playing at a constant frame rate along with synchronized audio track(s), an area closely related research to automatic video indexing is automatic image indexing (Rasmussen, 1997; Li and Wang, 2006; Carneiro et al., 2007).

2.3.1.2 Concept-based Indexing

Since text retrieval has been well established, metadata such as keywords (either free-text or from a controlled vocabulary), descriptions, and captions (essentially free-text annotations) have been traditionally used to index the video, such that video retrieval can be performed by searching the associated text. Google Videos, for example, allows people to search for videos by entering text queries. The search results are retrieved by matching text associated with the videos instead of visual processing of the videos. This approach that uses the linguistic cues to provide access to videos (or images) is called *concept-based* visual indexing.

To investigate the problem of automatically creating metadata for videos, it is wise to first understand how we human beings index or describe videos. In generally, we tend to focus on questions such as **what they are**, **what they mean**, and **who made them**.

As previously discussed in Section 2.1, Panofsky (1955) identified three levels of semantic significance of visual images: pre-iconographical description, iconographical analysis, and iconographical interpretation or synthesis. Shatford (1986) applied Panofsky's imagery comprehension model to indexing images, and suggested a 3x4 mode facet matrix (which is often referred to as the Panofsky-Shatford model) by relabeling Panofsky's three levels as *generic* (pre-iconographic), *specific*, and *abstract*, with each level further divided into four facets: **who**, **what**, **where** and **when**. Jaimes et al. (1999) incorporated Panofsky and Shatford's models and defined a 10-level pyramid framework for describing the visual content attributes of the image or the video sequence.

The top four levels of the pyramid describe images with *syntactic* (or "*perceptual*") information, while bottom six levels describe objects or scenes at *generic*, *specific*, and *abstract* levels respectively.

Having only metadata from the top four *syntactic* levels may not satisfy users' searching need: they may want more complicated queries than just querying by color, shape, or texture. Indexing by metadata at *syntactic* levels may also produce the **semantic gap**: visual similarity does not guarantee conceptual matches. For example, searches on "sea" may return blue images of "sky", and searches on "parrot" may return images of "fox", which are far from the true information targets of a user. Thus, it is necessary for the automatic metadata creation techniques to be able to provide metadata at *conceptual* levels. Unfortunately, the difficulty of automatic indexing increases from the top-most levels down to the bottom levels because more knowledge is required when going down the pyramid. While the two abstract semantic levels, abstract objects and abstract scene, are left out of the scope of current TRECVID retrieval tasks, generic and specific objects and events are emphasized in many years of TRECVID evaluations (Over et al., 2005; Smeaton et al., 2006).

2.3.1.2.1 Video Retrieval through Transcripts If transcripts or closed captions of the videos are available, full-text search can be used to retrieval videos in video retrieval systems by matching a user's query text against the video transcripts or closed captions. The CueVideo system used speech recognition technology to provide audio transcripts for technical talk videos, and supported full-text search though the automated transcripts. The full-text search led to acceptable video retrieval performance despite the low accuracy of speech recognition (Ponceleon et al., 1998).

2.3.1.3 Content-based Indexing

Content-based indexing compares an image or a video to other images or videos according to similarities in low-level (pixel-level) image attributes, such as color, shape, edge, texture, and so on. Content-based Image/Information Retrieval (CBIR), also known as Query by Image/Information Content (QBIC), automatically derives these low-level visual features and retrieves images or videos accordingly. Content-based indexing and CBIR have great value, because without the ability to examine the visual content, searches must rely on metadata such as captions, keywords or descriptions, which may be laborious or expensive to produce.

For example, the ALIPR (Automatic Linguistic Indexing of Pictures - Real Time, pronounced a-lip-er, http://www.alipr.com/) system automatically annotates online pictures in real time (Li and Wang, 2006). The system applies advanced statistical modeling and optimization methods to train computers about hundreds of semantic concepts using a large collection of example pictures from each concept.

CBIR has also been used in assisting physicians in disease diagnosis by retrieving visually similar images to a given query image from medical image databases (Zhou et al., 2008), because the visual characteristics of a disease carry diagnostic information, and often visually similar images correspond to the same disease category. The retrieval accuracy is satisfactory by limiting the input genre to medical images.

2.3.1.3.1 Higher-level Feature Classification

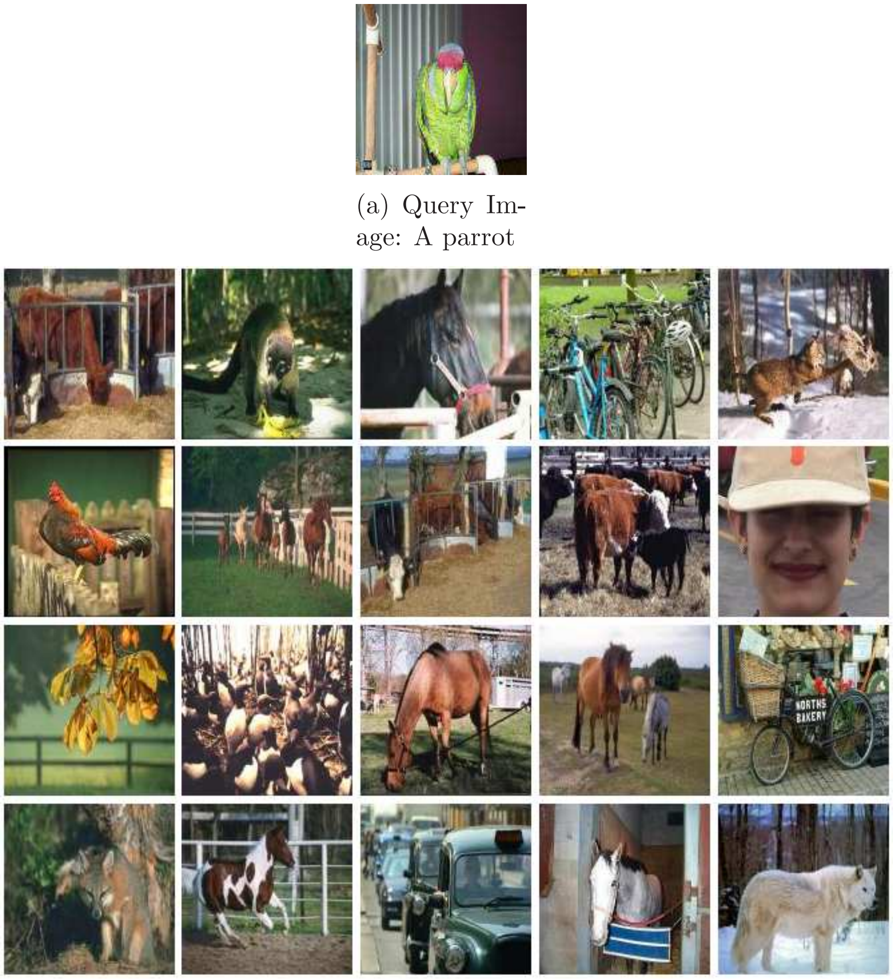
Image searching according to color and texture, in general, has little utility or commercial success, because it often does not bridge the semantic gap, i.e., visual/syntactic similarity may come with conceptually disjoint matches. Therefore, higher levels feature classifications involving meaningful higher level interpretation are desired. Specially, high-level feature classifications on face, people, outdoor, buildings, etc., can be used to filter retrieval results.

CBIR has been successfully used in certain retrieval tasks such as face detection, fingerprint retrieval, and text detection. Although classification on higher-level features in general is less accurate than classification on low-level features, two specific automatic visual feature classifications, face detection and text detection, have been proven to be successful. For example, VOCR (Video Optical Character Recognition, a.k.a. Video Optical Character Reader) can be used to detect text areas in the video. VOCR works on not only letters, numbers, and symbols on visual objects in the video, but also works on text superimposition (text overlay) in the video (Sato et al., 1999).

CIRES (Content Based Image REtrieval System) is an online content-based image retrieval system which utilizes a combination of high-level and low-level vision principles (Iqbal and Aggarwal, 2002). In addition to color analysis, where colors are mapped into a fixed color palette, and texture analysis, CIRES employs perceptual grouping where low-level image features, such as edges, are hierarchically grouped into higher-level feature classifications of face, people, outdoor, buildings, and so on. The system supports both pre-defined sample queries and dynamic user-defined online queries by image examples. The system is able to serve queries ranging from scenes of purely natural objects such as birds, trees, and clouds to “manmade” object images such as buildings, towers, and bridges. Though the system demonstrated some efficacy of using high-level structure in combination with low-level color and texture features, the results of the user-defined queries are often not satisfactory (See Figure 2.3). As shown in Figure 2.3, the example image used for the query is about a green parrot. The retrieved images by the CIRES system range from birds, leaves, horses, and a human face, to a bicycle and cars.

Similarly, video retrieval can also be done by querying example video clips (Dimitrova and Abdel-Mottaleb, 1997). In general, automatic visual concept classification accuracy is too low to be useful. But the accuracy may be improved by limiting the input video genres. The same performance difficulty resides in retrieving audio events in the video. For instance, thunder may trigger false positives of gunshots. By limiting the input videos to surveillance videos, the accuracy will goes up significantly (Valenzise et al., 2007).

2.3.1.3.2 Improving the Metadata Quality Numerous approaches have been implemented to improve the quality of the automatically generated metadata for videos. Certainly, employing some manual efforts can make some improvements in the automatic metadata. Examples include manually restricting the genre of the input video (Gong et al., 1995), manually correcting the au-



(a) Query Image: A parrot

(b) Retrieved Images

Figure 2.3: A CIRES Query of a Parrot based on an Example Image

tomatic speech recognition text, and labeling positive and negative examples to provide *relevance feedback* (Wang et al., 2001). In addition, automatic feature classifications and metadata with higher accuracy can be achieved if we throw in more computational power and more complete knowledge sources (Hauptmann et al., 2007b).

For visual feature detection applications such as face detection, template-based approaches can be used to produce higher accuracy. The drawback of such approaches is the increased computation cost. A template of a given size can be run over every area of each frame in the video, to identify all possible matches to the template. In addition, the template can be scale up and down, which requires even more runs of template matching over the entire video. Thus, various scales of template sizes, various positions in the frames, as well as possible rotations of the templates, etc., can all lead to more complete matches and greater accuracy at the expense of computation.

Instead of separately analyzing different information streams, multi-modal video indexing approaches can be utilized to improve the quality of automatically generated metadata (Satoh et al., 1999; Babaguchi et al., 1999; Alatan et al., 2001). For example, Satoh et al. (1999) developed the Name-It system that successfully associates faces and names in news videos by analyzing multiple information sources, such as video sequences, transcripts, and video captions. The system detects face sequences and does face similarity evaluation from video sequences, and extracts name candidates from the closed captions. Text detection and character-recognition techniques are also used for video-caption recognition so as to obtain face-name association. The state-of-the-art multi-modal video indexing techniques are reviewed in Snoek and Worring (2005).

2.3.1.4 Other Video Indexing Approaches

In addition to manual or automatic indexing of videos, the increase in social web applications and the semantic web have inspired the development of web-based image or video annotation tools, such as social tagging (also known as folksonomy, from folk + taxonomy), where metadata are generated not only by experts but also by creators and consumers of the content. By aggregating the tags of many users, social tagging is intended to make a body of information increasingly easy to search, discover, and navigate over time (Smith, 2007). Social tagging can be used to generated metadata for data of various media sort. For example, Delicious (formerly del.icio.us) is a social bookmarking web service for storing, sharing, and discovering web bookmarks. Flickr is one of the widely cited and most popular photo sharing websites where social tagging is used. And YouTube is one of the most popular video sharing websites where users can upload, share, and tag videos.

2.3.2 Video Retrieval based on Surrogates

Some video retrieval systems not only provide descriptive metadata, which can be used as a basis for searching through the large video collection, but also provide a (visual) preview or summary of the videos. The preview or summary of the video is often called a surrogate. A user may formulate a text query first, and then use the surrogates of video records returned by the search engine to make inferences about the video's content, or make relevance judgments. A user may also use the metadata or surrogates to browse through the video collections for relevant videos.

Surrogates are condensed information representing the full video documents. They are also addressed as video summarization in some research papers (He et al., 1999; Taskiran et al., 2006). Surrogates can be used for a variety of purposes (Taskiran et al., 2006; Goodrum, 2001; Song and Marchionini, 2007), including but not limited to the following:

- Serve as an advertisement, and intrigue the viewer to watch the full videos. Movie trailers, are the best example of surrogates for this purpose.

- Summarize all important information contained in the video. The intent is to replace watching the hour-long full videos. The surveillance videos, if can be summarized such that the important events can be detected, will be extremely useful (Damnjanovic et al., 2008). Sports programs, are also examples where only important instances, such as home runs and exciting pitches in a baseball game, and goals or goal attempts in soccer, and so on, are most interesting to the viewers.

- Provide support for quick and easy searching, browsing, and relevance judgements for videos in large video collections, and give users an efficient overview of an unfamiliar video collection. Surrogates may function as attributes against which a query may be matched, as well as enable people to make the same distinction about the videos as they would make with the full videos.

- Help the viewer decide whether to download and watch the full videos. Surrogates not only save humans' time in watching all the full videos, but with the pervasive mobile devices, surrogates also significantly reduce downloading time and cost.

The above listed purposes of video surrogates fall into two main categories of functions: *indicative* and *informative*. Indicative surrogates are used to indicate what topics are contained in the video, and informative surrogates are used to cover as much information in the full video as possible. For videos of different genres or characteristics, video surrogates can be designed to gear toward one of the two different functions, or a mixture of the two since the two functions are not completely independent.

Video surrogates can also be categorized based on their medium: text, visual (still images or moving images), audio, and multi-modal. Text surrogates containing bibliographic information about the video records, which are often know as metadata, have been discussed in Section 2.3.1. This section summarizes existing visual, audio, or multi-modal surrogates.

2.3.2.1 Visual Surrogates

Video surrogates can take various forms, which mainly fall into two categories: *static* visualization and *dynamic* visualization. There are a variety of non-textual surrogates taking each form that have been used for video retrieval including the following:

- Static visualization: The surrogate are based on keyframes extracted from the video, and the keyframes are presented in a static way.
  - **Poster frame**: (a.k.a. salient still frame) an image selected to represent the video, usually a single frame extracted from the video.
  - **Storyboard**: (alias filmstrips, as in Christel et al. (1999)) a set of keyframes displayed in chronological order, usually in a *static* tabular format.
  - **Collage**: display video data along with related keyframes, maps, and chronological information in response to a user query (Wactlar, 2000).
- Dynamic visualization: Frames or segments are selected from the original videos and concatenated to be played as a dynamic video summary.
  - **Slideshow**: a *dynamic* display of a series of chosen pictures at a certain speed.
  - **Fast-forward**: most simply created by selecting every Nth frame and displaying the selected frames at normal speed (30fps).
  - **Video Skim**: a video clip abstract created by compacting visual and audio information while preserving the original frame rate.
  - **Trailer**: a pre-produced series of clips excerpted from a video.

The static surrogates, for example, storyboards, are not suitable for instruction or presentation videos, where the videos are dominated by a talking head, and the important information is mostly contained in the audio stream.

For example, the skim type of surrogates consisting of the slides and synchronized audio segments were created by He et al. (1999) for summarizing audio-video presentations, i.e., informational talks given with a set of slides.

In addition to the surrogates types above, there are also some novel visualization approaches in support of interactive video browsing. For example, Chen et al. (2004) clustered keyframes extracted from shots and presented them to the user in a hierarchical tree structure called "a similarity pyramid" to enable active browsing. Recently, Mohamad Ali et al. (2009) also proposed a hierarchical structure for browsing scenes, shots, and frames. Amir et al. (2003) provided an efficient video browser with multiple synchronized views, such as storyboards, salient animations, slide shows with audio, full videos and so on. The browser allows users to switch between different views, while preserving the corresponding point within the video among all views.

2.3.2.2 Audio Surrogates

Audio is an important and indispensable component of the video. It's surprising that little work has been done in audio surrogates. This section reviews some audio surrogates which have been proposed or examined by some researchers. Boekelheide et al. (2006) proposed a variety of audio surrogates which may be useful in video retrieval and browsing systems.

2.3.2.2.1 Spoken Metadata Spoken metadata (i.e. spoken keywords, spoken descriptions), which is addressed as "Speech Display of Metadata" in (Boekelheide et al., 2006), can be created with the aid of text summarization tools (if the metadata have not been created by human) and good text-to-speech synthesizers. If the metadata keywords and descriptions are extracted from the transcripts of the video, which were spoken by some actors in the video, "text-to-speech" synthesizers are not necessary. Based on the metadata, the audio segments containing the metadata text can be extracted accordingly to be included in the spoken metadata. Spoken metadata can serve as a good accompaniment to the visual surrogates with the advantage of allowing people to process the audio and visual surrogates concurrently in human cognition systems.

The spoken descriptions based on text summarization tools differ from those based on audio summarizations in that the former rely on text analysis such as *tf-idf* term weight, while the latter rely on signal analysis, such as speech emphasis and pitch. The spoken descriptions created by human also differ from descriptions based on audio summarizations because human may summarize the audio or video using their own words instead of extracting sentences or paragraphs from the audio or video.

The effectiveness of the spoken descriptions or keywords has been studied by some researchers, and the spoken descriptions have proven to be more effective than visual surrogates in video gisting (Hughes et al., 2003; Wildemuth et al., 2002; Song and Marchionini, 2007; Marchionini et al., 2009).

2.3.2.2.2 Audio Snippets An audio or video snippet is a short extract from an audio or video that is substantially shorter in time than the source audio or video. The idea is similar to subsampling video discussed in Section

2.2.1.2.

Boekelheide et al. (2006) proposed the idea of "fast forwards of sound" as a new type of audio surrogate. The audio snippets can be created by sampling small audio segments at intervals across the entire video. To achieve a certain compaction rate, the design decisions revolve around the length of the audio snippets (e.g., 3 seconds, 4 seconds, etc.) and the sampling rate for the snippets (e.g., every 30 seconds, 60 seconds, or 120 seconds). Different sampling techniques (e.g., starting at the beginning of the video or starting from the first "good" audio snippet) can be utilized to improve the quality and usefulness of the extracted audio snippets.

A pilot study shows that 2-second audio snippets are too short and result in very "choppy" tracks, while 3-second audio snippets are acceptable, and 5-second audio snippets are very good, consistent with findings in Christel et al. (1998). It was also found that audio snippets, given the same amount of "preview" time as video snippets containing the same audio segments and the corresponding visual segments, convey much less information than video snippets. However, audio snippets can be very useful surrogates for videos on small devices with limited screen real estates, such as cell phones and PDAs, where video snippets or other visual surrogates become impractical or resource consuming.

2.3.2.2.3 Compressed Audio and Parallel Audio Streams Boekelheide et al. (2006) also proposes compressed audio by "speeding-up" the audio tracks and playing parallel audio streams simultaneously. The idea is similar to speeding up video and split-screen display of video skims. However, as with the time compressed video, it is technically difficult to achieve satisfactory compaction rates even with combination of various audio compression techniques (Heiman et al., 1986), and it's very likely that it will increase users' cognitive load greatly. The same problem exists for parallel audio streams. Humans have the ability to extract information from several audio channels simultaneously, but they can't pay close attention to a large number of channels at the same time. Therefore, with parallel audio streams, it is difficult to obtain high compaction rate. Also, it may place extra burden to human cognition.

2.3.2.2.4 Visual Surrogates for Audio Surrogates for videos can not only be used to help people make sense of the videos before downloading the full videos, but also can be used to help people find and retrieve interesting videos or filter out uninteresting ones in large video collections. The aforementioned audio surrogates are useful for video sense-making, and there are also some possible surrogates useful solely for video retrieval.

According to Boekelheide et al. (2006), "[v]isual surrogates for audio refer to various types of visual representations that can be created to display features of audio tracks." Audio features such as types of sounds (e.g., speech, music, silence, environmental sound, their combinations, etc), types of speakers (e.g., gender, age, number of speakers), speaker alteration patterns, and loudness / excitement levels can be represented to help users understand video content.

For example, bar charts or pie charts can be created to summarize the proportion of each sound feature in the audio track and allow easy comparison between the sound features. Cascading levels of the pie charts (such as “pie of pie” and “bar of pie”) can be used to show finer grained categories. Other types of visual surrogates for audio, such as linear sequence of color blocks, can be used to preserve the temporal relationships between features.

Visual surrogates for audio are visual representations of audio that do not necessarily help people make sense of the audio or videos, but they can be very useful in helping people make inference of the genre or flow of the audio tracks, and in helping people distinguish videos of interest within large collections of videos.

2.3.2.3 Multi-modal Surrogates

There are relatively fewer works in multi-modal surrogates than in text- or visual- based surrogates, and even fewer works in asynchronized multi-modal surrogates. Videos for news programs, lectures, and teleconferences, have special characteristics in the audio and visual presentation, and some multi-modal surrogates have been created for these genres of videos.

As discussed in Section 2.2.6, Gong (2003) proposed a new strategy to present video skims for news, documentaries, seminars, etc. The audio and visual contents of the given video were processed and summarized separately, then the best-representing spoken sentences of the audio track were partially aligned with corresponding visual segments displaying the speakers’ faces and other image segments. With this approach, the audio-visual summary maximizes the information coverage for both audio and visual contents of the original video.

A news program is generally composed of alternative concatenation between anchor shots and news segments. Lie and Lai (2004) proposed a new strategy to present video skims for news videos. The anchor audio highlighting the story headline, is overlaid with the visual summaries for news sequences. The integrated multi-modal summary helps the viewer understand the story headlines as well perceive motion activity of the story. More importantly, the audio and the visual segments included in the integrated summary may not come from the same segments in the original video, hence might not be synchronized.

The above mentioned works focused on a special genre of videos, the news programs. Researchers have only started to investigate multi-modal surrogation for other video genres. Christel et al. (1998) created video skims for video material drawn from three public television series: "The Infinite Voyage", "Planet Earth", and "Space Age". Ding et al. (1999) investigated multi-modal surrogates consisting of both keywords/phrases and keyframes for fourteen 2-3 minute video clips selected from a collection of 24 one-hour Discovery documentaries.

Song and Marchionini (2007) introduced a different approach of representing instructional documentary videos (selected from the NASA Connect Collection), which augmented the storyboards with 1-2 sentence spoken descriptions for the video. The three surrogate conditions evaluated were: visual alone (i.e., storyboard), audio alone (i.e., 1 or 2 sentence spoken descriptions for the video), and visual and audio combined (i.e., storyboard with the spoken descriptions). The descriptions were synthetic audio narrations of the text descriptions of the videos created by human indexers after watching the videos. Different from Manga summaries (Foote et al., 2000), the audio and visual channels in (Song and Marchionini, 2007) are not temporally coordinated. Results from the study showed that the (audio and visual) combined surrogate outperformed visual alone surrogate, but the audio alone surrogate was almost as good as the combined surrogate, which suggested that maybe the temporal coordination between the visual and audio channels is not necessary for highly abbreviated surrogates. But Song and Marchionini (2007) suggested that the issue of synchronicity for surrogates needs careful further investigation.

Marchionini et al. (2009) also evaluated a few unimodal surrogates and multi-modal surrogates consisting of the fast forward and spoken descriptions or spoken keywords, for NASA Connect videos. Two versions of the spoken descriptions and keywords were created for the evaluation, i.e., manually generated and automatically generated. The study showed that when the spoken descriptions were manually generated, they lead to almost as good gisting as combining the spoken descriptions with the fast forwards. Participants also commented that the combined surrogates were sometimes annoying, due to the fact that the audio and visual were not synchronized. Some participants reported that they had to close their eyes or move them away from the screen when listening to the spoken surrogates, and they had to take off their headphones when looking at the fast forwards. Although Marchionini et al. tried to present the surrogates as multi-modal, participants had to separate them at consumption time. It's almost obvious to us that the spoken metadata can be very straightforward and effective way of helping people make sense of the primary video objects and are relatively inexpensive to create. Also, the spoken metadata can be either used alone, or easily combined with visual surrogates, which make them very flexible and promising for video retrieval and browsing interfaces. However, there are still a lot of design questions left to be carefully addressed in the future. For instance, does synchronization between the surrogate channels enhance or inhibit video retrieval and video sense-making? How do the synchronized multi-modal surrogates compare to unsynchronized ones?

2.3.3 Summary

This section reviews some existing work in video retrieval. By using metadata generated manually or automatically, or matching low or high visual features, videos can be retrieved base on their associated text or by example video clips. Video retrieval can also be done by viewing surrogates of the videos. A user may use the surrogates to browse through the video collections for relevant videos, make inferences about the videos' content, or make relevance judgments. Some existing visual, audio, or multi-modal surrogates useful in video retrieval and browsing systems, are reviewed.

2.4 Methodologies Used in Video Retrieval and Surrogation Studies

Yahiaoui et al. (2003) categorized summary evaluation into *user-based* evaluation, where a group of users are asked to provide an evaluation of the summaries or asked to accomplish certain tasks (i.e., answering questions), and *mathematically based* evaluation, where corresponding mathematical values can be used directly as a measure of quality. Unfortunately, these techniques are all subject to some limitations. The user-based evaluation methods are difficult and expensive to set up and their biases are nontrivial to control, whereas mathematically based evaluation methods are difficult to interpret and compare to human judgment.

Borrowing terminology developed for text summarization evaluation, Taskiran et al. (2006); Over et al. (2007) classified video summary evaluation methods into two categories: *intrinsic* and *extrinsic*. This section discusses existing summary/surrogate evaluation methods according to the intrinsic and extrinsic categorization.

2.4.1 Intrinsic Evaluation

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. Most intrinsic evaluations do not compare summaries to the full video being summarized.

2.4.1.1 Coverage of key content / Precision and Recall

For event-based video like sports programs, where events that interest users are generally easy to create and often objective and unambiguous, video summaries may be intrinsically evaluated based on the coverage of important or interesting events in the source videos. For instance, Ekin and Tekalp (2003) introduced algorithms for automatic, real-time soccer video summarization by detecting important events in soccer videos, such as goals, referee, and penalty box. The soccer video summaries can then be intrinsically evaluated based on *precision* and *recall* values for these important events.

2.4.1.2 Similarity to Ground-truth

For videos where important events are often subjective and not easily identifiable or agreed upon by different people, it is more difficult to judge the quality of the summaries, e.g., whether the extracted summary has good coverage of the important segments of the video or not. Therefore, human experts are needed to first identify the important objects or events in the video to generate a ground-truth set of keyframes or events. de Silva et al. (2005) found a considerable proportion of common key frames among the keyframe sets selected by different people. They produced a (ground-truth) *average keyframe set* by averaging the keyframe sets selected by eight subjects. Then summarization evaluation can be performed by comparing the produced keyframe set for summarizing the video with the corresponding average keyframe sets. Ferman and Tekalp (2003) reported an intrinsic study where two neutral observers with knowledge of the target videos determined the number of redundant or missing frames based on whether the frames contained important objects and events identified by the human observers.

2.4.1.3 Questionnaire Method & Subjective Measures

The *questionnaire method* is another commonly used intrinsic evaluation method for video summaries. Likert scale questionnaires (Davis, 1989) are provided to human participants in usability studies to rate their levels of agreement with usefulness and usability statements about summaries, for example, “I found the summary to be clear and easy to understand”, “I feel that I can skip watching the whole program because I watched this summary” (He et al., 1999; Taskiran et al., 2006), or “Using this system helps me better estimate the gist of videos” (Song and Marchionini, 2007).

The 7-point semantic differential scale questionnaires developed by (Ghani et al., 1991) have also been adapted by a number of usability studies to measure the participants’ engagement and enjoyment using the systems or interfaces. For example, Song and Marchionini (2007) asked the participants to rate “Using the video surrogates is not interesting / interesting”, “How you felt using the video surrogates: attention was not focused / attention was focused”, and so on.

2.4.2 Extrinsic Evaluation

In extrinsic evaluation methods, the video summaries are evaluated in terms of their impact on the performance for a specific information retrieval task. There are several sub-categories of the extrinsic evaluation methods.

2.4.2.1 Performance for specific information retrieval tasks

Goodrum (2001) conducted an exploratory study examining the representativeness of both text-based (i.e., title, keywords) and image-based (i.e., salient still frame, multiple keyframes) video surrogates. The study used twelve 10-second short video clips from Cable News Network Image Source as the test videos, and the videos did not contain any human voices or spoken language of any kind. The four surrogate types were evaluated under three task conditions: **no task**, **specific task** (eg., to find information on the Old Faithful geyser in Yellowstone National Park), and **general task** (e.g., to find information that illustrates the fragility of our water resources). Inspired by the multi-dimensional scaling (MDS) approach drawn from earlier experiments (Weis and Katter, 1967) where surrogates for *textual* documents were evaluated, Goodrum used MDS to map the dimensional dispersions of users' judgments of similarity between *videos* and similarity between *surrogates*.

For each of the three task constraints, one group of study participants was asked to render *similarity judgements* (i.e., marking on 5-inch lines) for all pairs of videos in the test collection. Four separate groups of participants were asked to render similarity judgments for all possible pairs of surrogates for one of the four surrogate types respectively. Similarity measures included the degrees of similarity between the stimuli (either video pairs, or surrogate pairs), and the relative degrees of usefulness of the paired stimuli. Then the participants' judgement marks were converted to numeric values and entered into matrices for each group, and analyzed in SPSS. MDS maps were created to display relationships between data in as few dimensions as possible; in

Goodrum (2001), three dimensions provided a good fit with the data. Congruence values were calculated as the 3-D distances between each surrogate and its parent video. The smaller the congruence value, the higher the congruence between the surrogate and the video.

The results showed that, in general, the image-based surrogates demonstrated greater congruity than the text-based surrogates, but each type of surrogate makes a unique contribution to users' perceptions of information content, and should not be excluded from video retrieval systems. Particularly, in the no-task condition, the image-based representations scaled with greater congruity. The specific task resulted in slightly increased congruity for the text-based representations, where the performance of text-based almost tied the performance of the image-based surrogates. And in the general task condition, congruity shifted very slightly to a point of equilibrium between the text-based and image-based surrogates.

However, it is worth noting that the above results were drawn from a study condition where very short video clips were used. Surrogates for the short videos are barely useful in real-life video retrieval applications: Candidate videos can be searched and retrieved using text-based bibliographic information, and the surrogates are expected to further help people select from the candidate videos or make judgments about the videos as people would make using the original videos, but in much less time. For the short video clips used in this study, these advantages of surrogates are hardly visible.

Therefore, the results in Goodrum (2001) can not be generalized to other video datasets and conditions. The author also noted "Had the study utilized a more diverse collection of images, or a wider range of tasks, the results might have been quite different" (Goodrum, 2001, p.181), and suggested future research for examining the effect of combining surrogate types.

There have been a number of exploratory studies that evaluate different types of video surrogates based on specific tasks (Christel et al., 1998; Ding et al., 1999; Goodrum, 2001; Wildemuth et al., 2003; Song and Marchionini, 2007; Marchionini et al., 2009), which will be discussed further in Section 2.4.7.

2.4.2.2 Quiz Method

Quiz questions derived from the full video have been commonly used to evaluate video summaries on the *coverage* of the key content of the video contained in the summaries.

For example, in addition to the intrinsic questionnaire evaluation, He et al. (1999) asked the presentation speakers to write some multiple choice quiz questions that covered the content of their audio-video presentation video. Participants were given pre-study quiz questions before they watch any of the summaries to measure their expertise in the topic areas of the talks, and were asked to answer the author-generated quiz questions after watching the summaries in order to test the coverage of summaries of key ideas from the video. The effectiveness of the video summaries was quantified by the increase in quiz scores.

Taskiran et al. (2006) adopted the quiz approach used in (He et al., 1999). Two independent judges, one of whom was the first author of the work and the other was naive about the summarization algorithms used, independently marked the important points of the programs in the closed-caption transcripts of the programs without watching the summaries. The intersection of the two marked list was used to generate ten multiple choice questions for each program, while the mean number of correct answers out of the ten multiple choice questions for each video by different algorithms was used to evaluate the proposed algorithm along with two random algorithms. The authors also took the approach of an extrinsic evaluation, which will be discussed in detail in Section 2.4.4.

2.4.2.3 Evaluation using simulated user principal

Similar to the quiz method, Yahiaoui et al. (2003) extrinsically evaluated the quality of multi-episode video summaries created by different algorithms according to some *simulated user principal*. After watching all summaries, the user is shown a *randomly chosen* excerpt of a randomly chosen full-length video and is asked to guess which video this excerpt was extracted from. The quality of the summary is quantified by the percentage of correct answers that a user is able to provide when he is shown all possible excerpts of all videos. As with an intrinsic evaluation using precision and recall values, the quality measure in this study may not correlate strongly with users' judgment of summary quality. Also the evaluation method is tied to the summarization algorithms evaluated in the study, such that the evaluation method may not apply to diverse summarization methods.

2.4.3 Limitations of intrinsic and extrinsic evaluation

Taskiran et al. (2006) also pointed out the drawbacks of the quiz methods : "First, it was found that this approach may have difficulty differentiating between different summarization algorithms depending on program content (Taskiran et al., 2002; He et al., 1999). Second, it is not clear how quiz questions can be prepared in an objective manner, except, perhaps, by authors of presentations who are usually not available. Finally, the concept of a "key idea" in a video program is ambiguous and may depend on the particular viewer watching the skim".

Similarly, Taskiran and Bentley (2007) discussed problems with the intrinsic and extrinsic evaluation methods. It is unclear if the precision and recall values in an intrinsic evaluation correlates strongly with users' judgment of summary quality. The subjective assessments in the questionnaire method often do not correlate strongly with users' performance on information retrieval tasks. It is not rare that the discrepancy between performance and satisfaction has been reported in usability studies (Nielsen and Levy, 1994; Wildemuth et al., 2002). As Song and Marchionini (2007) pointed out, several earlier studies (e.g., Christel et al. (1998); Hughes et al. (2003)) have demonstrated that people like to have visual surrogates regardless of their performance effects. And for the quiz method, the quiz questions are usually subjective, depending on who prepares the quiz.

Taskiran and Bentley (2007) not only pointed out the limitations of the precision and recall measures, questionnaire methods, and quiz methods in summary evaluation, but also suggested that the summarization evaluation should be as realistic as possible reflecting user needs and real-world tasks.

The evaluation tasks should be performed by users who will use or who will be interested in using the system in the actual environment to take care of any environmental effects. The authors also suggested that a common summary evaluation data set including test videos and reference summaries should be created.

2.4.4 Evaluation using both Intrinsic and Extrinsic Methods

Taskiran et al. (2006) proposed a method, called FREQ in the paper, which automatically generates video summaries based on video transcripts obtained by ASR, and conducted a user study to judge the quality of the FREQ generated video summaries comparing to the quality of summaries generated using two other algorithms. The work used both extrinsic and intrinsic evaluations. After watching each video skim, the subjects were asked to answer three questions about the quality of the summary (i.e., *intrinsic*), and then answer ten multiple choice questions derived from the original full videos (i.e., *extrinsic*). Two independent human judges created lists of important points of the videos without watching the skims, and the intersection of the two lists were used to generate the questions.

For the *extrinsic* evaluation, the number of correct answers out of the 10 multiple choice questions by each algorithm was summed for all participants, and the FREQ algorithm was found to be consistently and statistically significantly better than RAND and DEFT for all three documentary programs, while the RAND and DEFT algorithms were comparable with no statistically significant differences.

For the *intrinsic* evaluation, the number of answers contained in the skim summaries were compared for the three algorithms, and the participants were asked to rate two subjective assessment statements, i.e., “I found the summary to be clear and easy to understand” and “I feel that I can skip watching the whole program because I watched this summary”, on a 1-5 scale. Summaries generated using the FREQ algorithm contained significantly more answers (or important information about the whole programs) than the other two algorithms. However, the subjective ratings on the assessment questions were almost comparable for all three algorithms. Again, as some of the previous studies found, the subjective assessments do not necessarily correlate strongly with users’ performance on information retrieval tasks, e.g., answering quiz questions.

2.4.5 Automatic Evaluation methods

Huang et al. (2004) proposed an automatic summary evaluation system called SUPERSIEV (System for Unsupervised Performance Evaluation of Ranked Summarization in Extended Videos). First, a set of reference (ground truth) summaries for several videos are gathered in a user study from many assessors. For each video, a single reference summary is generated to express the majority of the assessors’ opinions. Second, the system computes matching scores between each frame in the video and its best target reference frame in the video to form a lookup table that rates each frame. Then the system can quantitatively evaluate a video summary from different aspects by computing recall, cumulated average precision, redundancy rate and average closeness (i.e., an intrinsic evaluation).

Taskiran and Bentley (2007) took a similar approach to Huang et al. (2004), and proposed a pyramid algorithm to calculate a goodness score for automatically generated video summaries based on a set of reference summaries by human judges. The algorithm first identifies summary content units (SCUs) in the set of reference summaries, and assigns a weight to each unique summary content unit (SCU) based on its frequency. The most similar SCU is located for each automatically generated summary segment. Then a disjoint set of automatic summary segments that maximizes overall similarity with the reference set is derived, and a goodness score is calculated for the automatic summary using the weights of the corresponding SCU. The goal of the pyramid algorithm is to automatically produce a goodness score that ranks video summaries close to the ranking produced by human judges, according to a set of human generated reference summaries. Unlike (Huang et al., 2004) where the synthesized mainstream summaries are created automatically through k-means clustering of frames, SCUs in the pyramid algorithm are created manually. Though automatically creating the mainstream summaries is more desirable, the pyramid algorithm does not suffer from breaking SCUs which span multiple shots, which is a common problem with frame clustering.

Although a variety of video summarization or surrogation methods have been proposed whilst corresponding summary evaluations have been done, yet the majority of work is subject to some limitations: first, the video datasets used in the summarization studies have been small, and there is no commonly used video collection to train and test the proposed methods; second, as a result of the first limitation, the evaluation is based on the efforts of just one group, and it is difficult to do cross-group comparisons of different summarization algorithms, or to make statements about the summarization quality with respect to human judgment.

The 2007 TRECVID is an important milestone in the development of video summarization and evaluation techniques. It introduced video summarization evaluation as a new task for the participating groups, and is the first large-scale multi-participant evaluation of video summarization. The evaluation campaign provided a common dataset of rushes videos to be summarized as well as uniform metrics to evaluate the summaries submitted by different groups. Next, the changes and evolution of TRECVID, a common video retrieval benchmark and evaluation forum are reviewed.

2.4.6 TRECVID Fact-Finding Shot-Based Retrieval

The Text REtrieval Conference (TREC) sponsored by the National Institute of Standards and Technology (NIST) was started in 1992 to support the text retrieval research community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. The same needs for the video retrieval research community led to the establishment of the TREC Video Track in 2001.

The TREC Video Retrieval Evaluation (TRECVID) was founded in 2003 as an independent evaluation/workshop from TREC. The TRECVID evaluation workshops focus on a list of different information retrieval research areas in content based retrieval of video. As stated in the guidelines for the TRECVID 2008 evaluation, the main goal of TRECVID is to "promote progress in content-based analysis of and retrieval from digital video via open, metrics-based evaluation", by providing a large common test video collection, uniform scoring mechanisms, as well as a forum for organizations to compare their results. "TRECVID is a laboratory-style evaluation that attempts to model real world situations or significant component tasks involved in such situations".

Over the years, not only did the TRECVID evaluation video data sets grow gradually, but also new tasks were developed and tested. Also for the first few years in the TRECVID history, the evaluation was mainly on shot based retrieval, which was narrower than video retrieval in real applications. The following paragraphs overview the changes and evolution of the TRECVID evaluation in the past few years.

2.4.6.0.1 TRECVID 2003 In 2003, TRECVID participating groups used about 120 hours (241 30-minute programs) of ABC World News Tonight and CNN Headline News recorded by the Linguistic Data Consortium from late January through June 1998, as well as 13 hours of C-SPAN programming (about 30 mostly 10- or 20-minute programs, consisting of government committee meetings and discussions of public affairs, etc.) as the training (i.e., development) set and the search (i.e., test) set. Associated textual data provided with the ABC/CNN video include the output of an automatic speech recognition system and a closed-captions-based transcript. Participating groups performed 4 specific tasks:

- *Shot boundary detection*
- *Story segmentation*
- *Feature extraction*
- *Search*

The story segmentation task is as follows: “given the story boundary test collection, identify the story boundaries with their location (time) and type (miscellaneous or news) in the given video clip(s)”. The feature extraction task asks each participating group to return the list of at most 2000 shots from the test collection for each high-level semantic feature, concept such as “Indoor/Outdoor”, “People”, “Speech”, etc. The search task is defined as follows: “given the search test collection, a multimedia statement of information need (topic), and the common shot boundary reference for the search test collection, return a ranked list of at most 1000 common reference shots from the test collection, which best satisfy the need”, as stated in the TRECVID 2003 evaluation guidelines. Two types of search tasks were performed: *manual search* and *interactive search*, where the former involves humans formulating query based on topic and query interface, and the latter involves humans re-formulating queries based on topic, query and previous query results. For interactive search, the interactive user has no previous knowledge of the search test collection or topics.

The **shot boundary detection** task was evaluated based on *intrinsic* measures such as precision and recall, mean precision and recall for cuts and gradual transitions, and accuracy for gradual transitions. The **story segmentation task** was evaluated by story boundary recall (i.e., number of reference boundaries detected/ total number of reference boundaries) and story boundary precision (i.e., total number of submitted boundaries minus the total amount of false alarms / total number of submitted boundaries). Feature extraction task performance was measured by precision-recall as well as mean average precision (MAP), a measure which combines precision and recall and provides a single-figure measure of quality across recall levels. For the **search task**, the submitted ranked lists of shots found relevant to a given topic by each participating group were judged manually based on average precision and elapsed time (for all runs) per search, and MAP per run.

According to TRECVID 2003, the UNC group found that feature-only systems are outperformed by text-only systems, which in turn are outperformed by text-**plus**-feature systems, as reported in (Smeaton et al., 2004). Some groups, such as CMU, Lowlands, and IBM, explored how to weight and combine the ASR-based retrieval and feature-based retrieval. Other groups developed browsing interfaces for browsing through shots, and found that interactive retrieval was better than automatic retrieval.

2.4.6.0.2 TRECVID 2004 TRECVID 2004 adopted the development and test data for TRECVID 2003 plus various ancillary data created for 2003 (e.g., ASR from LIMSI) as the development data, and 70 hours of CNN Headline News and ABC World News Tonight video captured by the Linguistic Data Consortium during the last half of 1998 from for test data (i.e., a different time window from TRECVID 2003).

The same four tasks were used as for the 2003 TRECVID, with the exception that the story classification (a sub-task in the story segmentation) from 2003 was discontinued because a baseline that always guesses ”News” would get a 50% accuracy. The tasks were evaluated using similar measures as in 2003.

2.4.6.0.3 TRECVID 2005 In 2005, there were four main tasks with some changes from the tasks of the two previous years, in particular, a low-level feature extraction task was added into the evaluation:

- Shot boundary determination
- *Low-level* feature extraction: Given the feature test collection and the common shot boundary reference, identify all shots in which each of the following 3 low-level features (feature groups) is present: pan (left or right) or track, tilt (up or down) or boom, and zoom (in or out) or dolly.
- *High-level* feature extraction: Same as the *feature extraction* task in 2003 and 2004
- Search (*interactive*, *manual*, and *automatic*): In addition to the manual search and interactive search from the previous years, 2005 TRECVID also accepted fully automatic search submissions (no human input in the loop): i.e., system takes topics as input and produces results without human intervention.

and an optional pilot task:

- Explore BBC rushes: Given 50 hours of BBC rushes about vacation spots, a set of keyframes for each video, and minimal metadata per video, (1) build a system to help someone unfamiliar with the details of an archive of rushes and looking for archived video segments for reuse in a new video, browse, search, classify, and summarize the material in the archive; (2) Devise a way of evaluating such a system's effectiveness and usability.

About 160 hours of English, Arabic, and Chinese news from November 2004, several hours of NASA's Connect and/or Destination Tomorrow series, and about 50 hours of BBC rushes on vacation spots were used as the development data and test data. Output of an ASR system, output of a machine translation system (from other language to English), and common shot boundary reference and keyframes were provided to the participating groups as ancillary data associated with the test data.

The four main tasks were evaluated using similar measures as in 2003 and 2004, and the Explore BBC rushes task was an exploratory task, which required no submissions or evaluation at NIST.

2.4.6.0.4 TRECVID 2006 The 2006 TRECVID completed the two-year cycle on English, Arabic, and Chinese news video. The video dataset was similar to TRECVID 2005 but with significant additional data from channels and/or programs not included in the data for 2005.

The tasks were almost identical to the tasks used in 2005 – with one exception that the low-level feature extraction task was removed – and were evaluated using similar measures as in the previous years.

2.4.6.0.5 TRECVID 2007 TRECVID 2007 and 2008 switched the video data from broadcast news to a real archive – Dutch television materials – to see how well the technologies apply to new sorts of data.

Over et al. (2007) presented an overview of the TRECVID 2007 video summarization evaluation campaign pilot. The evaluation used rushes from a BBC dramatic series. In the previous years, the TRECVID evaluation mainly focused on the evaluation of the video information retrieval system using *Shot boundary detection*, *Story bound segmentation*, *Feature extraction*, and *search*, and so on. TRECVID 2006 not only completed the second two-year cycle devoted to automatic segmentation, indexing, and content-based retrieval of digital video - broadcast news in English, Arabic, and Chinese, but also also completed two years of pilot studies on exploitation of unedited video rushes.

TRECVID 2007 was the first year in which rushes summarization gets introduced as a new evaluation task, in addition to the three fundamental tasks: *shot boundary detection*, *high-level feature extraction*, and *search* (interactive, manually-assisted, and/or fully automatic). There have been a number of earlier studies of video summarization (Ding et al., 1997; Christel et al., 1998; He et al., 1999; Wildemuth et al., 2003; Taskiran et al., 2006), but the datasets used by each research group were small and there were no easy and reliable ways to evaluate the summarizations across groups. The 2007 TRECVID provided a common dataset of videos to be summarized as well as uniform metrics to evaluate the summaries submitted by different groups.

Specifically, the campaign provided about 100 hours of news magazine, science news, news reports, documentaries, educational programming video for the three fundamental tasks, and about 100 hours of unedited BBC rushes for the summarization task.

The three fundamental tasks were evaluated using similar measures to previous years. For the summarization evaluation, a set of measures were developed. Over et al. (2007) described what are the guidelines of creating the ground truth lists of important segments, how the ground truth for each video was developed, and how the video summaries submitted by the participants were judged by three human assessors.

The summaries were evaluated by both *subjective* measures (i.e., percentage of desired segments from the full video included in the summary, how easy it was to find the desired content in the summary, and amount of redundancy found in the summary), and *objective* measures (i.e., time taken to assess whether the desired segments appear in the summary, size of the summary (i.e., number of frames), and elapsed time for creating the summary).

Over et al. (2007) also summarized the approaches used by each of the 22 participating groups of the 2007 TRECVID, and the overall results of evaluating the summaries against two baseline systems using simple techniques. The two baseline video summarization systems were created by Carnegie Mellon University (CMU). The first baseline system selected 1 second segments, starting at 12.5 seconds into the current 25 second window and ending at 13.5 seconds, for every 25 seconds of original video. The 1 second chunks were then appended together to generate the summary. The second baseline used a

CMU shot boundary detector. From each shot a keyframe was extracted, and all keyframes for a video were clustered using k-means clustering. From each cluster, one second from the middle of the single shot closest to the centroid was selected to compose the summary.

The participating groups used different approaches to create the summaries. Approaches relied on combinations of some of the following techniques: video segmentation, keyframe/shot clustering algorithms, shot boundary detections, face/speech detection, and redundancy detection, some of which have been discussed in Section 2.2.5 and Section 2.2.7. Details of the approaches by 17 out of the 22 groups can be found in papers in the proceedings of the 2007 TRECVID workshop.

Results of the evaluation pilot found three systems (including City University of Hong Kong(CityU), Laboratoire d'Informatique de Paris 6, and National Institute of Informatics) significantly better than both baselines in terms of the fraction of ground truth included in a summary. With respect to ease of understanding and use, the two baseline systems were indistinguishable from each other, and among all systems, only the CityU system was significantly better than the baselines. No significant difference between the two baselines was found for amount of redundancy, but most systems were significantly better than one or both baselines. In addition, time taken to assess the summaries correlated positively with higher scores on the percentage of included ground truth.

The results suggest that systems were able to do something sensible within the guidelines and perhaps that the 4% compaction rate target could have been even smaller. Baseline systems using simple techniques were surprisingly adequate, although more computational efforts can produce better results.

2.4.6.0.6 TRECVID 2008 TRECVID 2008 continued using the Dutch television materials and the BBC rushes for the evaluation campaign, but the dataset was doubled compared to 2007. Also new datasets, i.e., surveillance video and MUSCLE-VCD-2007 data, were included in the evaluation.

The rush summarization task and two other tasks continued to be evaluated, while the shot boundary detection task was retired from the evaluation. There were also two new tasks introduced in 2008. In total, the participating groups tested systems using the following five tasks.

- **(New)** Surveillance event detection pilot
- High-level feature extraction
- search (interactive, manually-assisted, and/or fully automatic)
- Rushes summarization
- **(New)** Content-based copy detection pilot

TRECVID 2008 guidelines defined the surveillance detection task and the content-based copy detection task respectively as follows: “given 100 hours of surveillance video (50 hours training, 50 hours test) the task is to detect 3 or more events from the required event set and identify their occurrences temporally”, and “given a test collection of videos and a set of about 2000 queries (video-only segments), determine for each query the place, if any, that some part of the query occurs, with possible transformations, in the test collection". There were also two optional tasks which were variants of the content-based (video-only) copy detection task: one using transformed audio-only queries and the other using transformed audio-plus-video queries.

In 2007 the participating groups used about 50 hours of Dutch television materials for development and 50 hours for search and feature test, and used about 18 hours of rushes for development and about 17 hours of rushes for testing in the summarization evaluation. All of these videos and the submitted summaries from 2007 were used as development data for TRECVID 2008, and there were another 100 hours for use as test data for the feature and search tasks and another 18 hours BBC rushes (40 videos) for use as test data for video summarization task, and about 100 hours surveillance video - the output of 5 cameras from the same period of 20 hours, for surveillance event detection task, as well as the MUSCLE-VCD-2007 data for the copy detection task . The submitted summaries from 2007 and the ground truth for 2007 were provided to participants as training truth data for summarization task, and annotations for training data of the surveillance video were provided to participants as training truth data for the surveillance event detection task.

For the **surveillance event detection pilot**, output from participating groups' systems were first aligned to ground truth annotations, then scored for misses and false alarms. In particular, a weighted linear combination of the system's missed detection probability and false alarm rate per unit time, named Normalized Detection Cost Rate (NDCR) measure, is used for evaluating system performance. For details about the evaluation measures, see "2008 TRECVid Event Detection Evaluation Plan" (available at http://www.itl.

nist.gov/iad/mig/tests/trecvid/2008/doc/EventDet08-EvalPlan-v06.htm).

For the **high-level feature extraction task**, the submitted ranked shot lists for the detection of each feature were judged manually based on *precision-recall curves* and *inferred average precision* - a single-valued combination of precision, recall, and ranking ability, which provides a good estimate of average precision.

For the **search task**, the submitted ranked lists of shots found relevant to a given topic by each participating group were judged manually based on inferred average precision and elapsed time (for all runs) per search, and mean inferred average precision per run. Note that starting from TRECVID 2008, inferred average precision instead of MAP was used to measure the high-level feature extraction task and the search task.

For the **summarization task**, the summaries were assessed based on fraction of the ground truth objects/events found in the summary, time needed to check summary against ground truth, duration of the summary, system time to generate the summary, and usability/quality scores. Similar to TRECVID 2007, Carnegie Mellon University provided a simple baseline system to produce summaries within the 2% maximum.

The **content-based copy detection pilot** was evaluated on how many queries the participating systems find the reference data for or whether the systems correctly tell if there is none to find (measured by probability of a miss error and the false alarm rate, and Minimal Normalized Detection Cost Rate combining costs of miss and false alarm errors), copy location accuracy, and copy detection processing time.

2.4.6.0.7 TRECVID 2009 As of August 2009, TRECVID 2009 evaluation is still on-going. The evaluation will use the same video datasets as used in 2008 with as large or larger sizes, i.e., Dutch television materials and the BBC rushes, surveillance video, and the MUSCLE-VCD-2007 data. And the same four tasks of TRECVID 2008 will be performed and evaluated using the same measures as in 2008.

2.4.6.0.8 TRECVID Summary The sections above summarized the changes and evolution of the TRECVID from 2003 to 2009. Smeaton et al. (2006) gave a retrospective overview of the TRECVID campaign with attention to the evolution of the evaluation and participating systems, and open issues. Christel (2009) also reviewed the success and drawbacks of TRECVID over the years.

For the TRECVID series of evaluations, the data set is real, representative, and shared among participating groups, the tasks and search topics reflect many of the types of queries real users pose based on analyzing query logs against the BBC Archives and other empirical data, and the evaluations are performed using open, common metrics and the TRECVID community often offer collaboration and sharing of resources. The search topics include requests for *specific* items or people and *general* instances of locations and events, reflecting the Panofsky-Shatford mode/facet matrix of specific, generic, and abstract subjects of pictures (Christel, 2009). But the user pool for many TRECVID groups comprises mainly of university students and staff because of their easy availability and may not represent a broad set of real-world users. As a matter of fact, Christel and Conescu (2006) did report the differences between expert and novice search behavior when given TRECVID topics. Search by experts establish idealistic upper bounds on performance, because experts have some knowledge not possessed by novices: experts have been working with multimedia information retrieval research for at least a year, have used the tested video retrieval system before the timed runs with the TRECVID, and know about TRECVID evaluation processes and metrics. Ideally, the systems can be evaluated by representatives of the target user group of the systems.

It was also found that interactive search systems with human-in-the-loop have consistently and significantly outperformed their manual counterparts and fully automated systems with no-human-in-the-loop when evaluating the video search tasks in TRECVID (Smeaton et al., 2006).

TRECVID started with a large set of shot boundary determination measurements adopted from previous work but soon adopted precision and recall as the main measures. Frame-precision and frame-recall measures were added to gauge separately the degree of overlap in the matches. The search and feature extraction tasks were measured using precision, recall and precision and recall based measures such as MAP and inferred average precision. In addition to the precision and recall based measures, interactive video search systems were also measure by user characteristics and satisfaction for all years in TRECVID history from 2003 to 2009, though groups were not required to submit this information to NIST.

Contemporary video search engines often rely on filename and associated metadata. Content-based image retrieval by pixel-level attributes such as color, texture, and shape, has been well studied, and many commercial systems support searching by a visual example (Iqbal and Aggarwal, 2002, 2003). However, the underlying semantic gap between the low-level attributes and the image makes content-based query formulation challenging. Therefore, nonlinguistic video are often queried and retrieved using text queries. In the past 6 years, interactive retrieval systems evaluated in TRECVID have almost universally supported query-by-text, query-by-image example is the next most frequently supported strategy across TRECVID participants, and query-by-concept has shown little success. Past years of TREC Video results readily demonstrate the importance of linguistic data (in text format) for retrieval, and 2005 was the first year that some groups showed better performance with features than linguistic features, but under very difficult linguistic conditions i.e., multiple languages with automated translations.

2.4.7 Interactive IR Laboratory Studies

A different approach to video retrieval evaluation is laboratory type of studies. The individual roles that textual and non-textual video surrogates play in making relevance judgment or identifying the contents of a video were investigated by various studies.

An earlier Informedia experiment (Christel et al., 1997) compared the relative effectiveness of three presentations – context-independent poster frame, query-based poster frame, and text title – in a within-subjects study of 30 high school and college students, using a fact-finding task against a documentary video corpus. Participants were shown a result list of video documents on some topics displayed using text title or poster frame (either naively chosen or query-based), and later asked to find a particular result clip which was the answer to a given question by browsing the result set and selecting video clips to play. The three interfaces were evaluated based on *dependent measures* including correctness, task performance time, and subjective satisfaction. The study found that poster frames arranged in a Segment Grid menu, when chosen based on the query context, led to significantly faster location of the relevant video (i.e., faster fact-finding) and greater satisfaction with the interface than using only a plain text menu of document titles or using a context-independent poster frames (i.e., first shot thumbnail for each document).

Similar results were reported by other researchers. Goodrum (2001) performed an exploratory study examining the representativeness of some text-based and image-based surrogates, i.e., title, keywords, (single) salient still frame, and multiple keyframes. The study showed that image-based surrogates performed better than text-based surrogates overall, yet each type of surrogate makes a unique contribution to users perception of information content, and should not be excluded from video retrieval systems.

Ding et al. (1999) designed an exploratory usability study to compare three types of video surrogates–visual (keyframes), verbal text (keywords/phrases), and visual and verbal combined. The study was a *qualitative* investigation of user cognitive processes through *observations* and user's *talking aloud* while performing recognition and comprehension tasks. The results demonstrate that users strongly favor the combined surrogates: not only does each modality make a unique contribution to the comprehension of a video, but in combination they reinforce each other, lead to better comprehension, and may actually require less processing time.

Although the combined surrogates in Ding et al. (1999) employed two information streams, i.e., visual keyframes and verbal text, yet both of them employ only the visual sensory modality. The multimodal surrogates are not multimodal in terms of sensory channels.

2.4.7.1 Open Video Project studies

The Open Video Project also conducted a number of studies to investigate the individual roles of textual and non-textual video surrogates in making relevance judgment or identifying the contents of a video. Wildemuth et al. (2002) evaluated five video surrogates – storyboards with text keywords, storyboards with audio keywords, slide shows with text, slide shows with audio keywords, and fast forward – in relation to their usefulness and usability in accomplishing specific tasks, i.e., gist determination, object recognition, action recognition, and visual gist determination. These performance tasks were closely related to the real-world tasks that users expect to perform with video collections. Participants were also asked to provide comments about the strengths and weaknesses of each surrogate after viewing. Specifically, the keyframes in the storyboard were displayed for a limited amount of time, allowing 500 milliseconds per frame, with either text keywords displayed under the storyboard, or with audio recording of text keywords played and repeated as necessary for the duration of the visual display during the viewing. The slide shows incorporated the same set of key frames as were included in the storyboards, and each frame was displayed for 250 milliseconds. To make the slide shows take the same amount of time as the storyboards, the entire sets of key frames was played twice for the slide shows, with no pause between the two repetitions. Finally, the fast forwards, were playing the every Nth frame of the original video at normal frame rate, so that the fast forwards were N times as fast as the original video, where N was chosen so that the fast forward ran about the same amount of time as the other four surrogates. No audio or text augmented the fast forwards.

According to the results, no surrogate was universally judged "best" by the participants, but the slide show with text keywords was not preferred by anyone, and the fast forward surrogate garnered the most support (i.e., when participants were asked to choose surrogates with which to perform the tasks, the fast forward was chosen in 14 out of 30 trials), particularly from experienced video users. Also note that in this study, storyboards or slide shows with audio keywords were multi-modal surrogates that employed both visual and audio modalities, and the fast forwards were played at about 8x the original speed, which was actually "slow" comparing to the speeds used in the later studies (Wildemuth et al., 2003). User preference also suggested that the fast forward surrogate should be further developed with the addition of audio keywords. In a more recent study, Marchionini et al. (2009) actually developed fast forwards with audio keywords, but the fast forwards were more than 100x, a lot faster than the ones used in Wildemuth et al. (2002), and did not turn out to be very effective surrogates compared to other surrogates used in Marchionini et al. (2009).

Though the viewing compaction rates used in these surrogates supported adequate performance, participants commented that they desired having more control over surrogate speed and sequencing, and they would like to be able to move from surrogate to surrogate. In response to the need for flexibility, Marchionini et al. (2000) developed the AgileViews user interface framework with several different views of a collection, as well as control mechanisms that facilitate low-effort actions and strategies for coordinating the views. Amir et al. (2003) developed an efficient video browser with multiple synchronized views of storyboards, salient animations, slide shows with audio, and full videos, allowing users switch between different views, while preserving the corresponding point within the video among all views. The participants in the study used the keywords to understand the content of the video, as advance organizers for viewing the visual portion of the surrogate, and as a source of ideas for terms to use in future searches. They commented that textual video surrogates can facilitate the process of determining relevance, and non-textual video surrogates can effectively complement textual surrogates. The study also found that both user perceptions and performance could be affected by characteristics of the test video itself. To take care of the effects of the test video characteristics, the recent Open Video usability studies have adopted a set of comparable videos selected from the NASA Connect and NASA Destination Tomorrow collections (Song and Marchionini, 2007; Marchionini et al., 2009).

Hughes et al. (2003) reported an eye-tracking study of digital video surrogates composed of text and three thumbnail images to represent each document. Twelve undergraduate students selected relevant video records from results lists containing titles, descriptions, and three keyframes for ten different search tasks. As they browsed the results page for each search, their eye movements were tracked to determine where, when, and how long they looked at text and image surrogates. It was found that participants looked at and fixated on text statistically reliably more than on images. The text surrogates were used as an anchor point from which the participants made judgments about the search results, and the images were communicating the "feel of the film" and what the video was like and were consistently used to confirm the judgments participants made. Moreover, although text dominates how people make sense of retrieval sets, images add confirmatory value and people like to have them.

Wildemuth et al. (2003) reported on a study of the use of fast forwards for digital video, and recommended a fast forward default speed of 1:64 of the original video with adequate user performance and satisfaction. Although this approach can achieve a much higher compaction/compression rate than fast forwards with audio, yet it still leads to severe coherence degradation and discomfort to the viewer.

Yang et al. (2003) addressed the question what measures could or should be used to test how people perceive and understand video surrogates, and overviewed six user performance measures which were used in two usability studies (Wildemuth et al., 2002, 2003). The six performance measures fall into two categories: *Recognition tasks* (including objection recognition with text stimuli, object recognition with graphical stimuli, and action recognition) and *Inference tasks* (including free-text gist determination, multiple choice gist determination, and visual gist determination). These measures may be useful in evaluating different surrogates in relation to their effectiveness in aiding video retrieval.

The tasks were motivated by the two-level categorization of video comprehension – sensory seeing and cognitive seeing – as discussed in Section 2.1.1. The recognition measures depend on pre-iconographical analysis of the objects and examine whether users remember seeing or hearing particular words, frames or video clips in the surrogates. The inference measures depend on iconographical analysis and iconographical interpretation of the video surrogates, and test how much thematic information users could obtain from the video surrogates and what "story" about the original video users could construct based on the surrogates. The initial field testing of these six measures indicates that they are practical and can differentiate multiple levels of performance with video surrogates (Wildemuth et al., 2002, 2003).

Marchionini (2006) presented a theoretical discussion of several measures of human performance that have been used in developing visual surrogates for the Open Video Digital Library. Two sets of *cognitive performance measures* (i.e., recognition measures and inference measures, as discussed in Yang et al. (2003)) and one set of *attitudinal measures* were described. The cognitive performance measures aim to assess object and action recognition as well as inferences made from gists. The attitudinal measures include a set of twelve Likert-scaled statements (Davis, 1989) to assess usability, usefulness (e.g., This system makes it easier to find information) and learnability (e.g., learning to operate this system was easy for me), and seven-point semantic differential scales adopted from Ghani et al. (1991) to assess engagement (e.g., I felt: absorbed intenselynot absorbed intensely) and enjoyment (e.g., using the system was: interestinguninteresting). These measures have been adopted by a number of later studies (Song and Marchionini, 2007; Marchionini et al., 2009).

Biometric measures can also be investigated as adjuncts to the cognitive measures so that we will have sets of measures for all three classes of human measures: *physical*, *cognitive*, and *affective* (Marchionini, 2006). These measures address different aspects of the search process and human interaction with retrieval systems, and none of them are dispensable to understand the overall effects of video retrieval and sense-making episodes.

2.4.7.2 Think-aloud Protocol

The "Think-aloud protocol" is a method first developed by Ericsson and Simon (1984), which has been used to gather data in usability testing in product design and development, in psychology and a range of social sciences. The method has been adopted as the major method for data collection supplemented by observation and post hoc interviews to understand a person's cognitive processes while he is performing some task of interest. The protocol involves participants thinking aloud as they are performing a set of specified tasks. Participants describe whatever they are thinking and feeling, as they go about their tasks. To compare three types of video surrogates, Ding et al. (1999) used the think aloud protocol and instructed the participants to speak out everything that ran across their mind. By observing what the participants do and what they think aloud, the researchers are able to compare what was said and what was done, and capture implicit information that was not recordable.

A variant of the think-aloud protocol is the "talk-aloud protocol", which involves participants only describing their actions without interpreting or justifying their actions. This method is thought to be more objective than the think aloud method, but the researchers lose the information on why certain actions occur.

2.4.7.3 Interviews

The "interviewing" method is also an indispensable research method in the information science field. Instead of asking participants to perform a real task in a usability study, the interviewing method asks them to recall their own experiences in performing some task. For example, in addition to the think-aloud method, Ding et al. (1999) also performed a post hoc interview, such that misunderstanding or confusion can be clarified or dismissed. Similarly, to elicit users' video relevance criteria, Yang (2005) conducted a semi-structured interview session and an optional real video search session with the think-aloud method. The participants were asked to describe their specific information needs, the information sources selected, video searching questions, results selection process, as well as their final video uses.

2.4.8 Transaction Log Studies

Instead of running laboratory studies, some researchers install specific software on participants' computers for a certain period of time and log their searching interactions to investigate their searching behaviors (Fenstermacher and Ginsburg, 2003). This is client-side monitoring to keep track of searchers' behaviors.

Laboratory studies as well as the client-side monitoring techniques are likely to be limited by subjective elements, small number of participants, and sampling bias, thus usually have large variances in the results. Transaction log analysis (TLA) uses significant amount of useful information about web sites and search engines, and has become a relatively inexpensive technique to investigate user search patterns.

Transaction log analysis has been used widely in analyzing OPAC, digital libraries, and related online applications to provide insight into user search behaviors, and is useful in designing and evaluating search interfaces.

Blecic et al. (1998) uses TLA to compare data from two sets of OPAC transaction logs at a large public university. The first set of data was collected during a four-day period in the middle of the Fall 1995 semester (i.e., students are usually familiar with the OPAC by the middle of the semester). Analysis of the first set of data disclosed some limitations of the OPAC interface, and showed many users experienced difficulty with the basic searching functionality. After making changes to the OPAC interface addressing the problems revealed by the first TLA analysis, the second run of a four-day transaction log was collected 6 months later in the middle of the Spring 1996 semester to achieve comparable results. The second TLA showed statistically significant differences in the search results, which suggests that transaction log analysis of OPACs has the potential to improve the success rate in retrieval.

Jones et al. (2000) performed both quantitative and qualitative analysis on the transaction logs of the Computer Science Technical Reports collection in the New Zealand Digital Library. Jones et. al. not only culled user behavior information from a digital library transaction logs automatically by statistical analysis, but also manually examined the query strings in the transaction logs for searching motivations and searching strategy. Whereas other transaction log studies only focus on a much shorter time period (e.g., a day, as in the case of a Web search engine transaction log analysis of (Zhang et al., 2008)), this work is significant for the large span of time (i.e., 61 weeks) in the transaction logs. The work is also special in that it deals with a more focussed collection whose users are the computer science research community, who can be thought of as the "best case" users of online search engines. It was found that the "best case" users experienced many of the same difficulties with searching as experienced by the general public.

Transaction logs record the interaction between searchers and the search engines, therefore, transaction log analysis can also be used to detect user trends and make predictions about Web searching.Zhang et al. (2008) used time series analysis on a Web search engine transaction log. The transaction log was collected on Dogpile (www.dogpile.com), which is a top 10 ranked Web search engine, over a 24-hour period on 15 May 2006. A total of 4,193,956 transaction log records were collected, and each record contained 13 fields. A sampling strategy was used to select 10% of the original data set, resulting in 419,395 records, to make the statistical analysis feasible with the computing capacity of the current statistical software packages such as SPSS and SAS. The selected sample was then divided into 1080 equidistance groups, with each time slot being 80 seconds. Time series analysis was performed on those time slots. This study is significant in that, in addition to the basic descriptive transaction log analysis, the authors also applied one-step prediction time series analysis along with the Box-Jenkin transfer function models to predict searcher behaviors.

TLA has a number of strengths: The data are collected from a large user base; the analysis is reasonable and non-intrusive; it takes less time than other methods and can be relatively inexpensive. However, TLA is subject to some limitations too: the analysis does not include user demographic and other data, and it lacks data on search reasons and motivations, and there may be incomplete data due to corrupted logging, or network loads.

2.4.9 Summary

This section reviews some existing summary evaluation methods categorized as intrinsic and extrinsic, and previous evaluation work done by some interactive IR laboratory type studies as well as a common video retrieval benchmark and evaluation forum–TRECVID.

An intrinsic evaluation has users judge the quality of the generated summaries directly on fluency of the summary, coverage of key content of the video in the summary, similarity to ground truth summaries, or users' subjective ratings (e.g., usability, usefulness, enjoyment, or engagement). An extrinsic evaluation evaluates the summaries based on the performance for specific information retrieval tasks. Both the intrinsic evaluation and extrinsic evaluation have drawbacks, and they may be used together to evaluate video summaries.

Although a variety of video summarization or surrogation methods have been proposed whilst corresponding summary evaluations have been done, yet the majority of work is subject to some limitations: first, the video datasets used in the summarization studies have been small, and there is no commonly used video collection to train and test the proposed methods; second, as a result of the first limitation, the evaluation is based on the efforts of just one group, and it is difficult to do cross-group comparisons of different summarization algorithms, or to make statements about the summarization quality with respect to human judgment.

Most evaluation methods in the past did not use common datasets of videos, and there were no easy ways to compare the summarization techniques across different groups. Early years TRECVID work mainly focused on fact-finding shot-based retrieval, while TRECVID 2007 was the first year that started summarization evaluation which provided a common dataset and open metrics to evaluate the summaries submitted by different groups.

The laboratory type of studies including think-aloud protocol, interviews, and some previous Open Video usability studies, were reviewed. Finally, transaction log analysis that investigates web users' searching behaviors, detects, and predicts user trends was discussed.

Chapter 3

METHODOLOGY

3.1 Introduction

We have designed a series of usability studies to examine the efficacy of audio alone surrogates and the effectiveness of multi-modal surrogates utilizing both audio and visual channels. Two preliminary studies in the series were conducted previously to address the potential value of a particular type of audio surrogate – spoken audio, either alone or combined with some visual surrogates – for retrieving and making sense of videos.

This study examined the effectiveness of some multi-modal surrogates for retrieving and making sense of videos, and investigated how the automatically generated multi-modal surrogates compared to manually generated ones. A second research question of this study is whether the synchronization between the audio and visual channels of the surrogates enhances or inhibits video retrieval and video sense-making. The strategies for sampling the most salient abstracts from the audio and visual channels are of interest.

Two distinct approaches for creating multi-modal video surrogates were examined: ***pre-processed integration***, where the audio and visual channels were sampled simultaneously and thus were pre-coordinated at indexing time (i.e., ***synchronized*** multi-modal surrogates), and ***user-centered integration***, where the audio and visual channels were sampled independently (i.e., are uncoordinated) and needed to be integrated in the user's head at consumption time (i.e., ***unsynchronized*** multi-modal surrogates). Even though ***synchronized*** multi-modal surrogates and pre-coordination at indexing time may be desirable from the users' affective point of view because we are more used to synchronized presentations, **more** useful information about the video (i.e., better coverage of the key information in the video) may be carried in the ***unsynchronized*** surrogates if the most salient samples extracted from the audio channel do not align temporally with the most salient samples extracted from the visual channel. Therefore, we hypothesize that independently sampling the most salient samples across different channels and letting the users integrate the uncoordinated channels at consumption time may lead to **more** sense-making potential than pre-processed integration.

This Chapter is organized as follows: Section 3.2 summaries two preliminary studies and some experimental results achieved so far (Song and Marchionini, 2007; Marchionini et al., 2009). Section 3.3 describes the study designed and conducted to follow up the two preliminary studies, so as to answer the research questions about the effectiveness of some multi-modal surrogates (automatically generated vs. manually generated), the benefits of synchronized surrogate channels, and the benefits of independent sampling across channels (i.e., unsynchronized surrogate channels).

3.2 Two Preliminary Studies

We conducted two usability studies in 2006 and 2007 respectively to examine the effectiveness of some audio alone surrogates and multi-modal surrogates for making sense of instructional documentary videos.

3.2.1 The 1st Study

The user study done in 2006 (Song and Marchionini, 2007) was a within-subjects study with 36 participants. The study investigated the effectiveness of three different surrogates for making sense of digital videos in digital video libraries. One visual only, one audio only, and one audio and visual combined surrogate condition were examined (see Table 3.1).

Table 3.1: Three Surrogate Conditions for the 1st Study

Visual only	**Storyboard**: A set of keyframes displayed in chronological order in a tabular format.
Audio only	**Spoken Descriptions**: Recording the human written descriptions of the videos taken from the Open Video repository using the AT&T Lab Text-to-Speech synthesizer Online Demo with “Crystal” voice.
Combined	**Storyboard + Spoken Descriptions**: The storyboard is displayed and the spoken description is initiated upon display.

The three surrogate conditions in the 2006 study were evaluated using the following five tasks shown in Table 3.2.

For the **Written Gist Determination Task**, the summaries were scored by two researchers independently on a three point scale (i.e., 0 is wrong, 1 is partially correct, and 2 is correct). The correlation between the respective scores was 0.76,

Table 3.2: Gist & Recognition Tasks Used in the 1st Study

Written Gist Determination Task (open-ended): Write a short summary of the video based on the surrogate they experience in the study.
Keyword Recognition Task: Select keywords that are appropriate for the video from a set of words.
Title Selection Task: Select the most appropriate title for the video segment that the surrogate represents.
Keyframe Recognition Task: Select appropriate keyframes that they think come from a video from a set of keyframes.
Verbal Gist Recognition Task: Select the best description for the video from a set of four descriptions based on the surrogate they experience.

so the two sets of scores were averaged for each trial to yield final scores in the 0-2 range.

For the **Keyword Recognition Task**, some of the keywords came from the keyword field for the video used by the Open Video Project (i.e. were correct), and others were selected from keywords for other videos in the Open Video repository (i.e., were wrong). Some of the words were more concrete (e.g., aircraft for the First Flight video) and some were more abstract (e.g., visualization for the Hurricanes and Computer Simulation video). The number of keywords correctly identified as correct or wrong across each set of four trials was normalized to the 0-1 range for comparison across trials.

For the **Title Selection Task**, the correct title was the title of the video segment used in the Open Video repository, and the wrong ones were selected from titles for other videos in the same video collection in the Open Video. This task was scored as correct or incorrect, thus the score was either 0 or 1 for each trial and the sum taken across the four trials and then divided by four to yield a normalized score between 0 and 1.

For the **Keyframe Recognition Task**, some of the key frames were selected from the storyboard of the video segment in the Open Video repository (i.e., were correct), and others were selected from storyboards of other videos in the same collection (i.e., were wrong). As with the keyword recognition task, the number of keyframes and correct keyframes were varied slightly across trials and the total number of keyframes they saw in each surrogate condition remained the same. The number of key frames correctly identified was normalized to the 0-1 range for comparison across surrogate conditions.

The **Verbal Gist Recognition Task** was scored as correct or incorrect with the correct response the 1-2 sentence description from the video in the Open Video repository, and the distractor responses taken from descriptions for other videos in that series. The task was scored as correct or incorrect, thus the score is either 0 or 1 for each trial and the sum taken across the four trials and then divided by four to yield a normalized score between 0 and 1.

The order of the tasks is important. For example, the open-ended gist writing task should be completed first so that participants will not gain extra information from the other tasks (especially the gist selection task which includes actual descriptions, which should definitely go after all other tasks). The "back" button in the browser was disabled. Once the participants chose to go to the next page, they were no longer able to get back to the previous page. It is important to note that actual video segments were not played at any time during the study and participants had to make sense of the videos merely by consuming the surrogates. In the audio alone condition and the combined condition, participants had control to stop or replay the audio, and the numbers of times the participants replayed and stopped the audio descriptions were recorded as well.

Measures used to compare the three surrogate conditions included performance, confidence, time to consume the surrogates, time to complete the tasks, and a suite of affective measures. Qualitative comments were also used to enrich the interpretation of results.

As shown in Table 3.3, the results demonstrate that combined surrogates are more effective and strongly preferred than both of the individual surrogates, and do not penalize efficiency. Nevertheless, it is found that spoken descriptions alone are almost as good as the combined surrogate and are much better than visual storyboards alone for video gisting.

Table 3.3: Result Summary for the 1st Study

Performance	*Combined* $>$ *Audio only* $>$ *Visual only*
Confidence	*Combined* $\approx$ *Audio only* $>$ *Visual only*
Surrogate Consumption Time *	*Combined* $\approx$ *Audio only* $>$ *Visual only*
Task Completion Time *	*Visual only* $>$ *Audio only* $\approx$ *Combined*
Affective Measures	*Combined* $>$ *Audio only* $>$ *Visual only*

* For *Surrogate Consumption Time* and *Task Completion Time*, the shorter the time is, the more efficient the surrogate is.

Note that for surrogate consumption time, there were very small time differences between the audio only and the combined condition. Examination of the log data suggested that participants replayed the audio portion a lot because they found the audio "hard to understand", and it took the participants longer to consume the combined condition than the visual only condition not because they had two channels of information to integrate, but merely because the audio part was played for a longer time. The small time differences between the audio only and the combined conditions suggest that people are able to integrate two distinct sets of surrogates that use different sensory channels even though they are not temporally coordinated at all. Thus, the expectation that the temporal coordination between visual and audio channels desired for primary information objects is also required for visual and audio surrogates was not borne out.

The open-ended comments from the participants reinforced the results from the quantitative data. Thirty-one out of the 36 participants selected the combined condition as their favorite surrogate among the three conditions. As noted by one participant, “with the two together, the surrogate is more efficient, and understanding the surrogates becomes simpler than when they are apart.” The comments also reconfirmed the power of words in carrying the semantic information in the videos. One participant noted: “Even though a picture is worth a thousand words, a few selected pictures cannot explain the deeper meaning of the subject–audio connected the dots.” Furthermore, the participants also commented on the value the visual surrogates added. They stated “the storyboard was fun and engaging” and “images...helping me be more focused.”

This study had important implications for the design of video retrieval and video library user interfaces. The results recommended incorporating multi-modal surrogates into video retrieval user interfaces, and suggested that audio-only surrogates may have great value for video retrieval especially in small display interfaces. However, it was worth noting that 1/3 of the participants complained that the audio was hard to understand, and 1/3 of the participants complained that the keyframes in the storyboards were too small to see well. Therefore, another design implication from the study is that the audio and visual quality of the surrogates are important. The audio should be clearly articulated and the visual should be easily viewable.

3.2.2 The 2nd Study

In a follow-up study (Marchionini et al., 2009) done in 2007, we compared several different types of non-textual surrogates (i.e., fast forwards, spoken descriptions, and spoken keywords) alone and in combination (see Table 3.4).

Table 3.4: Five Surrogate Conditions for the 2nd Study

Visual only	**Fastforwards** (FFS): Created by selecting roughly every 150th frame from the original video.
Audio only (I)	**Spoken Descriptions** (SD): Recording the manually generated or automatically generated descriptions of the videos using the AT&T Lab Text-to-Speech synthesizer.
Audio only (II)	**Spoken Keywords** (SK): Recording the manually generated or automatically generated keywords of the videos using the AT&T Lab Text-to-Speech synthesizer.
Combined (I)	**Fastforward + Spoken Descriptions** (FFS + SD): The fastforward and the spoken description are played concurrently using the same media player.
Combined (II)	**Fastforward + Spoken Keywords** (FFS + SK): the fastforward and the spoken keywords are played concurrently using the same media player.

Note that the test videos used in the study were about 28 minutes 30 seconds each, which were longer than the test videos used in the 1st study. For videos that are 3 to 5 minutes long, it is not as imperative to create surrogates as for videos that are half an hour each or even longer. In the combined conditions, the audio and the visual start playing at the same time but may end at different time points. When being replayed, the fastforward and the spoken descriptions both start playing from the beginning. The automatically-generated descriptions and keywords were created using the text analysis tool that is part of a system called IBM MAGIC system (Li et al., 2005).

This study investigated the effectiveness of manually and automatically generated spoken descriptions and keywords on six video gisting and recognition tasks (See Table 3.5). Note that Task 3 (Title selection task) in the 1st study was eliminated from the 2nd study because no performance differences were found among all three surrogate conditions using this task. Two new tasks (displayed in ***Bold Italic*** below) were designed and added to the remaining 4 tasks in the 1st study, in order to investigate how people articulate gist on different multimedia evidence.

Table 3.5: Gist & Recognition Tasks Used in the 2nd Study

Written Gist Determination Task (open-ended): Write a short summary of the video based on the surrogate they experience in the study.
Keyword Recognition Task: Select keywords that are appropriate for the video from a set of words.
Keyframe Recognition Task: Select appropriate keyframes that they think come from a video from a set of keyframes.
Visual Excerpt Selection Task (A new task in this study): Select one visual excerpt (with no audio tracks) that they think comes from the video from a set of four visual excerpts based on the surrogate they have experienced. The three visual excerpt distractors will be selected from other videos in the same collections, in similar collections, and in totally different collections respectively in the Open Video repository.
Audio Excerpt Selection Task (A new task in this study): Select one audio excerpt (without visual features) that they think comes from the video from a set of four audio excerpts based on the surrogate they have experienced. Similar strategies of selecting distractors for Visual Excerpt Selection Task will be applied here.
Verbal Gist Recognition Task: Participants select the best description for the video from a set of four descriptions based on the surrogate they experience.

We had four specific research questions related to spoken surrogates for video retrieval purposes:

1. Do automatically generated description and keyword surrogates approach the effectiveness of manually generated ones?
2. How do spoken descriptions and spoken keywords compare for gist-related tasks?
3. How do fast forward surrogates compare with the spoken surrogates?
4. What are the effects of combining fast forwards and spoken surrogates?

For question 4, the study suggested that combining two surrogate media channels must be done carefully. “On one hand, there is a possibility of interference and on the other there is the possibility of leveraging the independent perceptual channels to gather more gisting evidence in the same amount of time. We expected that the possibility of higher cognitive load would rule, and the participants would not like the combined condition, though task accuracy might be higher using the combined surrogates than using the individual surrogates” (Marchionini et al., 2009).

Table 3.6 summarizes some of the experimental results from the 2nd user study (for more details of the results, please refer to Marchionini et al. (2009)). The results demonstrated that manually generated spoken descriptions were significantly reliably better than manually generated spoken keywords and fast forwards for video gisting; whereas when automatically generated, spoken descriptions and keywords either alone or combined with fast forwards are inferior to fast forwards alone. Furthermore, when the spoken descriptions were manually generated, they led to almost as good gist determination and recognition as combining the spoken descriptions with the fast forwards. Participants also commented that the combined surrogates were effective but sometimes annoying, due to the fact that the audio and visual were not synchronized. The study recommended incorporating the spoken description surrogates of good summarizing quality into the video retrieval systems.

Table 3.6: Result Summary for the 2nd Study

Performance & Confidence	$Manual\ SD > FFS > Automatic\ SD$ $Manual\ SK > FFS > Automatic\ SK$ $Manual\ SD > Manual\ SK$ $FFS + Manual\ SD > FFS + Manual\ SK$ $FFS + Manual\ SD \approx Manual\ SD$
Task Completion Time*	$FFS > Manual\ SD$ $FFS > Manual\ SK$ $FFS > Automatic\ SD$ $FFS > Automatic\ SK$
Affective Measures	**Usability & Usefulness for Manual group:** $SD > FFS + SD > FFS > SK$ $SD > FFS + SD > FFS > FFS + SK$ **Engagement & Enjoyment for Manual group:** $FFS + SD > SD > FFS > SK$ $FFS + SD > SD > FFS + SK > SK$ **All 4 affective measures for Automatic group:** $FFS > FFS + SD > FFS + SK > SD > SK$

* For *Task Completion Time*, the shorter the time is, the more efficient the surrogate is.

Note that for both of the two user studies, the audio and visual channels of the combined surrogates were not temporally coordinated, whereas participants from the two studies had completely different feelings about the unsynchronized combined surrogates.

For the first study, because the storyboard was a static representation of the video, augmenting it with a constantly changing audio surrogate like the spoken narration of the description did not make the combination of the two very distracting for the users. In other words, the audio and visual were not terribly out-of-sync. Therefore, storyboard with spoken description can be a **successful** example of multi-modal surrogates whose audio and visual channels are not synchronized.

For the second study, however, because the fast forward was also rapidly constantly changing, playing it together with the spoken descriptions or spoken keywords made the two pieces completely out-of-sync, hence the combined surrogates became increasingly distracting for the users. As reported by the majority of the participants, they could not successfully focus on both channels simultaneously. Some participants closed their eyes when listening to the spoken description, and took off their headphones when viewing the fast forward. Thus, fast forward with spoken descriptions may be a **not-so-successful** example of unsynchronized multi-modal surrogates for videos.

These two preliminary studies both confirmed the effectiveness of multi-modal surrogates for video retrieval and sense-making, but seemed to have distinct conclusions on the necessity of the synchronized surrogate channels. Therefore, a third study was designed to follow up these two studies to investigate the synchronization issues more carefully. The following section discusses the methodology of this research.

3.3 Method

To answer the research questions discussed in Section 3.1 (which were also discussed in more detail in Section 1.2), a range of experiments were performed to evaluate some selected multi-modal surrogates for digital videos in terms of their effectiveness in accomplishing certain inference and/or recognition tasks which relate to real-world users' information needs.

In the study, selected audio surrogates were combined with selected visual surrogates, either carefully coordinated or not coordinated. When the audio and visual channels were not coordinated, the most salient samples were extracted from the two channels separately, and it was possible that the most salient samples across different channels did not occur at the same temporal points. Although users were required to integrate the two channels in their heads at consumption time (i.e., **user-centered integration** was required), which may lead to increased cognitive load, perhaps more sense-making was possible with this user-centered integration, rather than pre-coordination at indexing time.

The surrogates were evaluated based on some **extrinsic** evaluation methods. Details are discussed in the following sections.

3.3.1 Test Videos

A set of 20 comparable videos was selected from the NASA Connect collection as **test videos** for the study. NASA Connect is an award-winning series of instructional programs that supports national mathematics, science, and technology standards. The programs establish the "connection" between math, science, and technology concepts taught in the classroom and NASA research, and are designed to enhance the teaching of math, science, and technology concepts in grades 5-8.

Another set of four videos was selected from the NASA Destination Tomorrow collection that targets general audiences (i.e., lifelong learners) interested in science, and were used as **training videos** for the participants to practice prior to the evaluation tasks (i.e., excluded from data analysis).

Both NASA Connect and NASA Destination Tomorrow programs are produced by Langley Research Center's Office of Education. All of the 24 videos used in the study have a common structural format and are at a similar conceptual level on various science topics, for example, hurricanes, aerodynamics, the Northern lights, and the global water cycle. None of the 24 videos share the same topics. The lengths of the videos are also very similar, ranging from 28 min 15 sec to 29 min 39 sec, and are 28 min 36 sec on average.

The video IDs, titles, and descriptions of these 20 NASA Connect test videos provided in the Open Video repository are shown in Appendix A.

Each of these videos also comes with a SMIL file (including transcripts and timestamp data) that NASA sent to the Open Video team. The transcripts were fed into the MAGIC system to automatically generate descriptions and keywords for the video, which were then used to create the MAGIC audio and visual surrogates discussed in the following sections.

3.3.2 Surrogates

Table 3.7 lists the methods used in this study to extract excerpts from the audio and visual channels respectively. The audio extracts and the visual extracts were combined together to create multi-modal surrogates. In other words, all of the surrogates investigated in this study were multi-modal surrogates with both audio and visual stimuli.

In the study, the storyboard for each video consisted of six keyframes. Each keyframe was 86 pixels x 59 pixels. The visual skims were of .MOV format and played in a QuickTime player. The total screen real estate for the storyboard and the visual skims (either sub-sampling visual or Magic visual) are comparable, and the users (participants) were able to change the sizes of the surrogates. According

Table 3.7: Methods of Extracting Audio and Visual Excerpts for the Multimodal Surrogates

Audio only (I)	**Sub-sampling Audio**: Audio snippets are extracted based on systematic sub-sampling, i.e., extracting n second audio snippet from every N second of the audio track. In this study, we extract 5 seconds out of every 120 second interval.
Audio only (II)	**Magic Audio** : Audio snippets are extracted based on the automatic description sentences extracted from video transcript text by the MAGIC system.
Visual only (I)	**Storyboards**: A set of keyframes are displayed in chronological order in a tabular format.
Visual only (II)	**Sub-sampling Visual** : Visual snippets are extracted from the videos based on systematic sub-sampling, i.e., extracting n second visual snippet from every N second of the visual track. As with sub-sampling audio, in this study we extract 5 seconds out of every 120 second interval.
Visual only (III)	**Magic Visual** : Visual snippets are extracted from the videos based on MAGIC extracted descriptions.

to past study experiences, these surrogate sizes are adequate for the users.

3.3.2.1 Automatically-generated Surrogates

Table 3.8 summarizes the automatically-generated multi-modal surrogate conditions we examined in this study, where **A** denotes audio and **V** denotes visual. In particular, we examined two perfectly temporally coordinated surrogate conditions and two uncoordinated surrogate conditions, and we eliminated two uncoordinated surrogate conditions which were not examined in the study (as shown in Table 3.9).

Note that the condition [Sub-sampling A + Storyboard (V)] was eliminated,

Table 3.8: Automatically-generated Multi-modal Surrogate Conditions

Coordinated (I)	**Sub-sampling A + V**: Video snippets, with synchronized visual and audio channels, extracted by systematic sub-sampling.
Coordinated (II)	**Magic A + V**: Video snippets, with synchronized visual and audio channels, extracted based on MAGIC extracted descriptions.
Uncoordinated (I)	**Magic A + Storyboard (V)**: Storyboard combined with audio snippets extracted based on MAGIC extracted descriptions
Uncoordinated (II)	**Magic A + Sub-sampling V** : Visual snippets extracted by systematic sub-sampling combined with audio snippets extracted based on MAGIC extracted descriptions.

because we could already predict without conducting experiments that it would be outperformed by [Magic A + Storyboard (V)] according to results found in Marchionini et al. (2009). The condition [Sub-sampling A + Magic V] does not make a lot of sense because the Magic summaries are extracted using the textual information, i.e., the transcript, not the visual attributes; hence it was eliminated as well.

The two uncoordinated multi-modal surrogates are examples where the most salient samples are independently extracted from the audio channel and visual channel. Ideally, because the most salient information carried in the audio channel may be different from the most salient information carried in the visual channel, combining the two channels together may provide nearly twice the information about the videos. However, given the limitation of the effectiveness of storyboards and the lack of intelligence of the systematic sub-sampling scheme on the visual channel,

Table 3.9: Automatically-generated Multi-modal Surrogate Conditions Excluded from the Evaluation

Uncoordinated (Eliminated)	**Sub-sampling A + Storyboard (V)** : Storyboard are combined with audio snippets extracted by systematic sub-sampling.
Uncoordinated (Eliminated)	**Sub-sampling A + Magic V** : Audio snippets extracted by systematic sub-sampling with audio snippets extracted based on MAGIC extracted descriptions.

the information carried in the visual channels of the uncoordinated surrogates may be insufficient compared to the visual information of the surrogate condition [Magic A+V] where at least some human intelligence is employed for extracting the text descriptions. To make sure that we do not arrive at wrong conclusions on synchronization and independent sampling due to the uneven comparison, in addition to the above 4 surrogates conditions, we also included one *gold standard* condition of perfectly coordinated multi-modal surrogates and one *gold standard* condition of uncoordinated multi-modal surrogates (see Table 3.10).

Table 3.10: Manually-generated Gold Standard Surrogate Conditions

Coordinated (III)	**Manual A + V** : Video snippets, with synchronized visual and audio channels, extracted by human efforts.
Uncoordinated (III)	**Manual A + Manual V** : Audio snippets extracted by human efforts combined with visual snippets separately extracted by human efforts.

3.3.2.2 Manual Creation of the Gold Standard Surrogates

3.3.2.2.1 Preparing Videos First, four instructional documentary videos were selected from the 20 NASA Connect videos (the test videos to be used in the study) for creating gold standard surrogates. The titles of the videos are as follows:

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

Each video was 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. The audio only and visual only versions have the same durations as the original full videos. 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.

3.3.2.2.2 Phase 1: Generating a Set of Reference Summary Segments 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 included 2 senior undergraduate students, 9 Master's degree students, and 1 faculty member in a digital video course, 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. Each judge was randomly assigned three videos out of the four, and the three videos were of different conditions: one audio only, one visual only, and one both (i.e., full video).

For each of the three videos, regardless of the viewing conditions, the judges watched or/and listened to the entire program. After viewing or hearing each assigned video, each judge was asked to extract the most salient segments from the video to be included in the video summary according to the following specific instructions:

> “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 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 **5** to **10** $+/-$ **2** 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.”

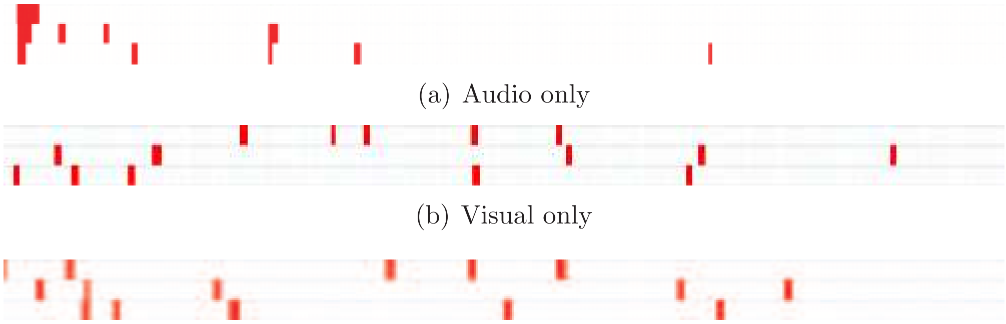
Thus, for each viewing condition of each video, a set of reference (ground-truth) summary segments (containing up to 15 segments) were selected by three independent judges.

Note that each human judge selected 5 extracts for the audio only condition, 5 extracts for the visual only condition, and 5 extracts for the full video condition. In total, the 12 human judges extracted 178 individual segments for the three viewing conditions of the 4 videos, with a small number of overlap segments in each condition. The total number of segments is not $12(judges) \times 5(segments) \times 3(viewing\ conditions) = 180$ 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: both). Overall, there are 58 segments with both audio and visual, 60 segments with audio only, and 60 segments with visual only.

On average, it took 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 surrogates. For each viewing condition of the video, a gold standard surrogate consisting of 5 to 8 summary segments (adding up to around 70 seconds to be comparable to the systematic sub-sampling and MAGIC skims) was created from the reference summary segments to express the majority of the judges' opinions.

The intuitive approach to selecting 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 (audio and visual) both 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.

Unfortunately, there were only a small number of segments that overlapped in the sets of 15 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. For example, Funk and Reid (1983) 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).



(a) Audio only

(b) Visual only

(c) Full video

Figure 3.1: Overlaps among Summary Segments for Video "NASAConnect: Virtual Earth" : (a) Overlaps in the audio only version, (b) Overlaps in the visual only version, (c) Overlaps in the full video.

Figure 3.1 shows the overlaps among the summary segments selected by the judges for the video "NASAConnect: Virtual Earth" under the audio only, visual only, and full video conditions respectively.

Figure 3.1(a) represents 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 two of the three assessors (i.e., 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 two of the three assessors (i.e., assessor 1 and assessor 3) as shown in Figure 3.1(b). And for the full video version, there was one overlap of segments between two of the three assessors (i.e., assessor 2 and assessor 3), as shown in Figure 3.1(c).

To determine a principled way to select the best 5-8 extracts from each set of 15, an additional phase of evaluation was conducted.

3.3.2.2.3 Phase 2: Rating the Manually Extracted Video Summary segments Another group of 4 human assessors were recruited to evaluate the summaries selected by the 12 judges in Phase 1. Note that none of the 4 assessors had participated in Phase 1.

Each of the four assessors was assigned one of the four NASA Connect videos for which the reference summary segments were generated. Each assessor watched the 28.5 minute full video with both audio and visual streams. After watching the videos, the assessors were asked to watch and/or listen to the 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). The assessors rated the 58 segments with both audio and visual first, and then rated the 60 segments with visual only, and finally rated the 60 segments with audio only.

A 5th assessor was recruited as a meta assessor. The meta assessor watched all these four videos and rated the segments for all four videos. Hence, we got two sets of ratings for each segment from two independent assessors – one of the four assessors, and the meta assessor.

The correlation between the respective ratings between the two sets (i.e., by one of the four assessor and the meta assessor) was 0.54, which demonstrated moderately satisfactory inter-rater reliability between the two sets (Landis and Koch, 1977). Hence, the two sets of ratings were averaged for each segment to yield final ratings in the 1-7 range.

For each video and viewing condition, the highest rated 5 to 8 segments (totalling about 70 seconds) out of the 15 summary segments were selected to form the gold standard summary to be used in the study.

3.3.2.2.4 Creating Coordinated and Uncoordinated Gold Standard Surrogates Based on the gold standard summaries we created in Phase 1 and Phase 2, coordinated and uncoordinated gold standard surrogates were created.

The coordinated gold standard surrogates were simply composed of the gold standard segments with synchronized audio and visual streams as selected by at least two independent judges.

For the uncoordinated gold standard surrogates, the gold standard visual segments were overlaid by the gold standard audio segments.

3.3.3 The Extrinsic Evaluation

3.3.3.1 Participants

A within-subjects design was adopted for the study. Forty-eight participants were recruited for the study by posting mass emails to our university-wide LISTSERV. All participants were adult native English speakers with self-assessed adequate listening and visual abilities, who used computers daily and had experience with searching for videos using computers at least occasionally.

3.3.3.2 Tasks

Evaluating the effectiveness of surrogates for videos presents significant challenges. People infer the gist by combining evidence at hand with their personal knowledge and past experience (Boguraev and Neff, 2000; Ponceleon et al., 1999; Spence, 2002), as summarized in Marchionini et al. (2009). In our case, people infer gist of the videos by combining the surrogates with their past video viewing and searching experiences. A series of tasks had been designed to reveal how different forms of video evidence (surrogates) were used in inferring gist. A direct and generative task asked people to articulate the gist of the video based on the surrogates. Other tasks asked people to select salient features from feature sets, such as keywords, keyframes, visual snippets, audio snippets, and textual summaries.

The same six tasks from our 2nd user study (see Table 3.5 in Section 3.2.2) were used in this study as measures of recognition and inference:

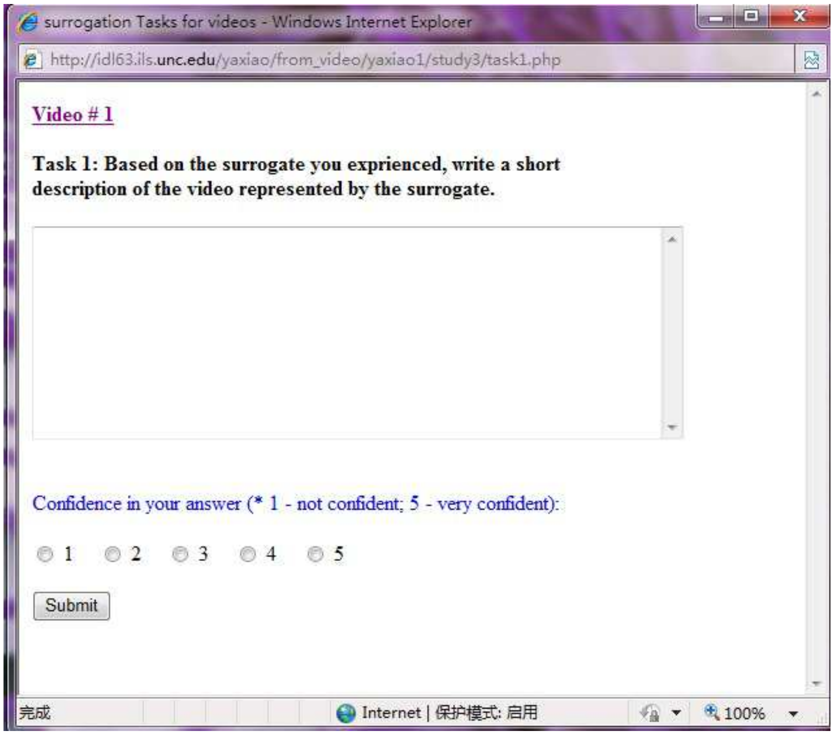
1. Free-text gist written task
2. Keyword determination task
3. Keyframe determination task
4. Visual excerpt determination task
5. Audio excerpt determination task
6. Verbal gist determination task

Note that we renamed the tasks differently from the 2nd user study, because the term "determination" could more precisely indicate both the inference and recognition aspects of some tasks than the term "recognition" or "selection". For example, Task 2 was both recognition and inference task. So was Task 3. Thus, the term "determination" was more precise than the term "recognition" for these two tasks. To be consistent, some of the other inference tasks (i.e., Task 4, Task 5 and Task 6) were named "determination" tasks as well.

For all trials of video surrogates in the study, the six tasks were carefully ordered to minimize potential knowledge gained from previous tasks. For example, the open-ended verbal gist writing task was always performed first, so that participants would not gain extra information from the other five tasks before writing the verbal gist. The verbal gist determination task was always performed last because it disclosed to the participants the correct descriptions, which would significantly deepen participants' understanding of the video segments. For all trials, the tasks were given to the participants in the same order as described below.

3.3.3.2.1 Task 1: Free-text Gist Written Task. Figure 3.2 shows the online stimulus for Task 1. In this open-ended question task, participants were asked to write a short summary of the video based on the surrogate they experienced. Note that the participant only experienced the surrogate and not the full video. To evaluate participants' input for this task, two investigators worked together to identify two main facets for each summary, and the summaries were scored by two coders on a four point scale: for each facet, 2 points were given if the facet was completely addressed, and 1 point was given if the facet was partially addressed, and 0 points were given if the facet was not addressed at all. The points for the two facets were added together to make the final score of each summary. Therefore, "0" meant none of the two main facets were addressed at all; "1" meant only one facet was partially addressed; "2" could either mean one facet completely addressed or two facets both partially addressed; "3" meant one facet completely addressed and the other partially addressed; and "4" meant both facets were completely addressed.

First, the two coders scored a subset of the participants' free text inputs to Task 1 together, to establish a common baseline for scoring the free text gist. Each coder then worked on the rest of the inputs on their own. There was a correlation of `0.691` between the respective scores from the two coders, and the correlation was significant at the 0.001 level (2-tailed). In observance of the high correlation between the scores from the two coders, the two sets of scores were averaged for each trial to yield reasonable and valid final scores in the 0-4 range for each summary. The mean of the scores for the four trials (excluding the training trial) was then computed for


surrogation Tasks for videos - Windows Internet Explorer
http://idl63.ils.unc.edu/yaxiao/from_video/yaxiao1/study3/task1.php
Submit
完成
Internet | 保护模式: 启用
100%

Video # 1

Task 1: Based on the surrogate you exprienced, write a short description of the video represented by the surrogate.

Confidence in your answer (* 1 - not confident; 5 - very confident):

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Figure 3.2: Online Stimulus for Task 1 (Free-text Gist written Task)

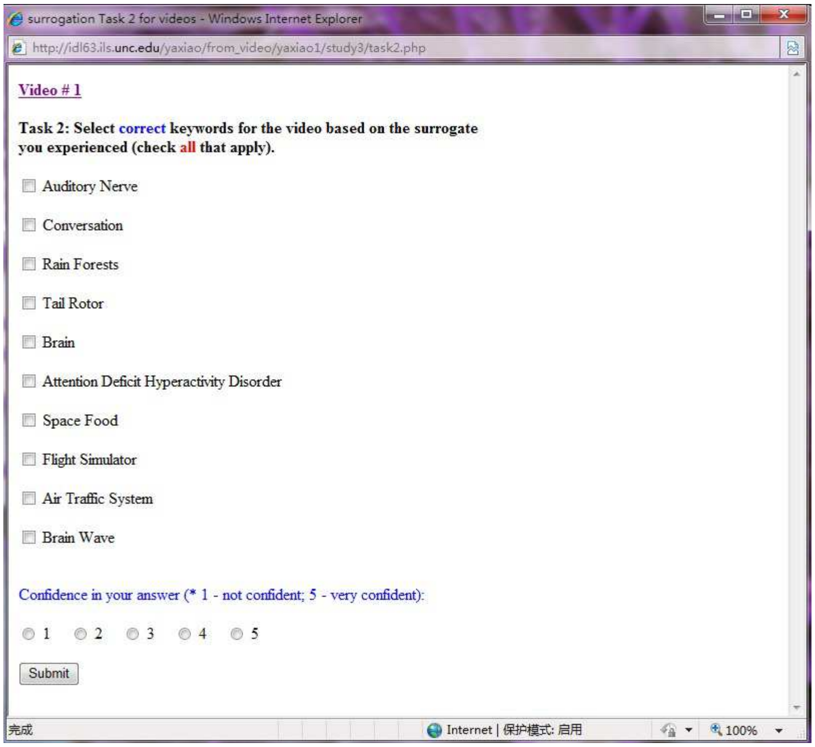
each surrogate condition.

In this and the other five tasks, the participants were also asked to indicate their confidence in their answer for each task trial on a 1-5 rating scale (i.e., 1 - not confident, 5 - very confident). As with the accuracy measures, the confidence ratings were averaged across four trials for comparison between surrogate conditions. The total time to complete each task trial were also recorded and averaged across the four trials.

3.3.3.2.2 Task 2: Keyword Determination Task. In this multiple choice task (choose all that apply), participants were asked to select keywords they believed to be appropriate for the video from a set of 10 words listed. Figure 3.3 shows the online stimulus for Task 2.

For task 2 of each trial, five of the keywords were chosen from the keywords for the video on the Open Video repository (i.e., were correct) and the remaining five were distractors (i.e., were wrong). Of the distractors, two keywords were selected from keywords for other videos in the same video collection (i.e., they had similar structures and were at a similar conceptual level, but were on different topics); two keywords were selected from keywords for videos in a different NASA video collection (i.e., they also had similar structures and were at a similar conceptual level, but again were on different topics); and one keyword was selected from keywords for videos from a totally different video collection (i.e., not only the topics, but also the structures and conceptual levels of the videos were different from the target video) so that we could have both near and far miss examples. Therefore, none of the distractor keywords should also apply to the target video.

For each trial, the fraction of correctly identified keywords was measured. The score was calculated as the number of keywords correctly identified as correct or



surrogation Task 2 for videos - Windows Internet Explorer
http://idl63.ils.unc.edu/yaxiao/from_video/yaxiao1/study3/task2.php

Video # 1

Task 2: Select correct keywords for the video based on the surrogate you experienced (check all that apply).

□ Auditory Nerve

□ Conversation

□ Rain Forests

□ Tail Rotor

□ Brain

□ Attention Deficit Hyperactivity Disorder

□ Space Food

□ Flight Simulator

□ Air Traffic System

□ Brain Wave

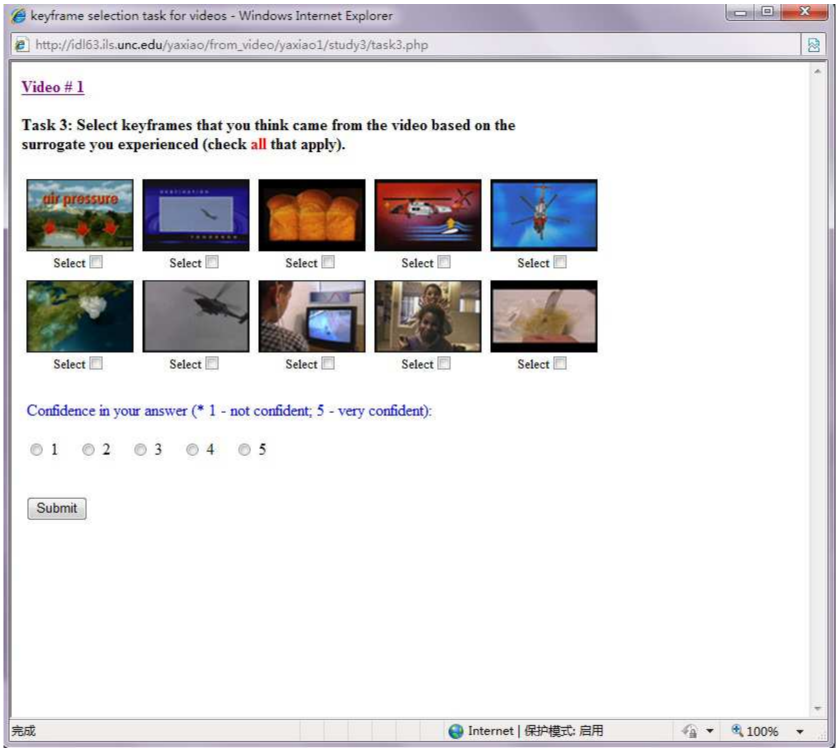
Confidence in your answer (* 1 - not confident; 5 - very confident):

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Submit

完成
Internet | 保护模式: 启用
100%

Figure 3.3: Online Stimulus for Task 2 (Keyword Determination Task)


keyframe selection task for videos - Windows Internet Explorer
http://idl63.ils.unc.edu/yaxiao/from_video/yaxiao1/study3/task3.php
air pressure
完成
Internet | 保护模式: 启用
100%

Video # 1

Task 3: Select keyframes that you think came from the video based on the surrogate you experienced (check all that apply).

Select □ Select □ Select □ Select □ Select □

Select □ Select □ Select □ Select □ Select □

Confidence in your answer (* 1 - not confident; 5 - very confident):

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Submit

Figure 3.4: Online Stimulus for Task 3 (Keyframe Determination Task)

wrong for each trial divided by 10 (i.e., the total number of choices for each trial). The scores (ranging from 0 to 1) across each set of four trials were averaged for comparison between different surrogate conditions, as were the confidence scores and task completion time for each trial.

3.3.3.2.3 Task 3: Keyframe Determination Task. Figure 3.4 shows the online stimulus for Task 3. This was also a multiple choice task (choose all that apply). In each video trial, participants were asked to select appropriate keyframes that they thought came from the video from a set of 10 candidate keyframes provided by the study investigators.

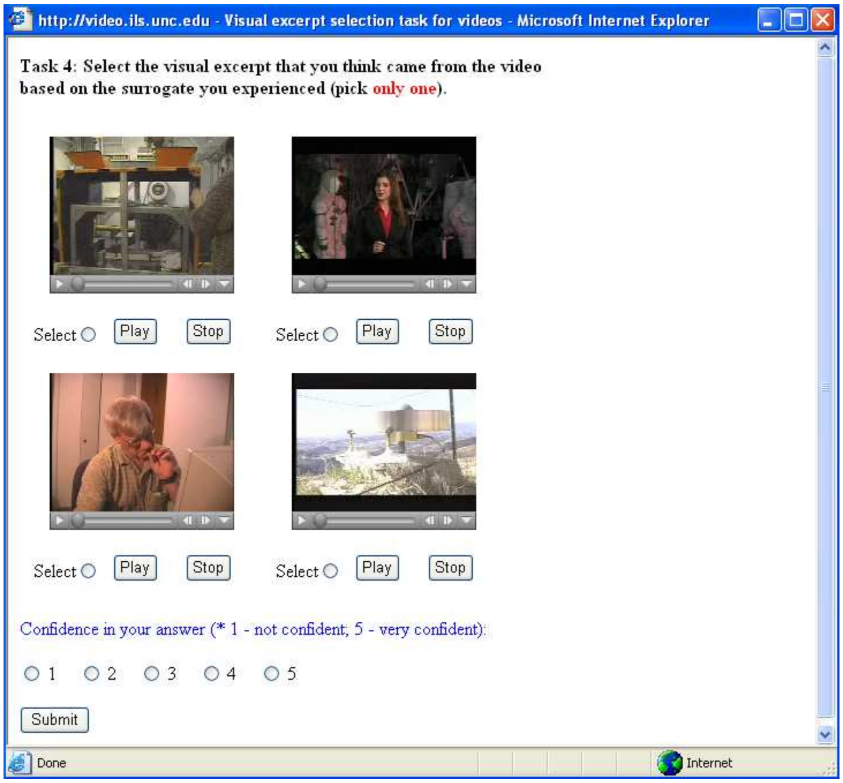
Similar to the keyword identification task, five of the key frames were selected from the keyframes for the video segment on the Open Video repository (i.e., were correct), and five keyframes were selected from keyframes for other videos in the same video collection, in the other NASA collections, and in totally different collections in the Open Video repository (i.e., were wrong). As with the selection of the distractors for the keywords identification task, none of the distractor keyframes should also apply to the target video.

For each trial, the percentage of correct keyframes was computed (i.e., the number of keyframes correctly identified as correct or wrong for each trial divided by 10) and the scores across each set of four trials were averaged. Likewise, participants' confidence ratings and task completion times were computed for comparison between the surrogate conditions.

3.3.3.2.4 Task 4: Visual Excerpt Determination Task. Figure 3.5 shows the online stimulus for Task 4. In this multiple choice task (choose all that apply), participants were asked to play four 7-second visual clips, each of which was extracted from a full video segment but without including audio tracks, and then select the visual excerpt(s) that they thought came from the video segment that the surrogate represents.

The correct visual excerpt for task 4 of each trial was based on the 7-second excerpt of the video in the Open Video repository, but only the visual part was taken and the sound was not included. The three alternatives were visuals extracted from three other videos in the Open Video repository. Of those three distractors, one was selected from other videos in the same collection, one was selected from videos in similar collections, and one was selected from a totally different collection in Open Video.

This task was scored as the fraction of correctly identified visual excerpts (i.e,



http://video.ils.unc.edu - Visual excerpt selection task for videos - Microsoft Internet Explorer

Task 4: Select the visual excerpt that you think came from the video based on the surrogate you experienced (pick only one).

Select ○ Play Stop Select ○ Play Stop

Select ○ Play Stop Select ○ Play Stop

Confidence in your answer (* 1 - not confident; 5 - very confident):

○ 1 ○ 2 ○ 3 ○ 4 ○ 5

Submit

Figure 3.5: Online Stimulus for Task 4 (Visual Excerpt Determination Task)

0, 1/4, 2/4, 3/4, and 1). The mean of the four trials yielded a normalized score between 0 and 1. As with the other tasks, confidence and time to complete tasks were also computed.

3.3.3.2.5 Task 5: Audio Excerpt Determination Task.

This multiple choice task (choose all that apply) was the same as the visual excerpt identification task except that the participants were asked to play four audio excerpts and select excerpt(s) they thought came from the video segment that the surrogate represents.

The correct audio excerpt for task 5 of each trial was based on the 7-second excerpt of the video in the Open Video repository, but only the audio part (without the visual) was extracted. The three distractors were selected using the same strategy as used for the visual excerpt selection task.

Similar to Task 4, this task was scored the fraction of correctly identified audio excerpts (i.e, 0, 1/4, 2/4, 3/4, and 1) for each trial, and the mean score across the four trials (i.e., 0.0 - 1.0) was used to compare different surrogate conditions, along with the mean confidence rating and the mean task completion time.

3.3.3.2.6 Task 6: Verbal gist Determination Task.

In this multiple choice question (choose only one), participants were asked to select the most appropriate textual summary for the video segment from a set of four textual descriptions based on the surrogate they had experienced.

The “correct” answer was the most salient sentences manually extracted from the description for the video available in the Open Video repository by the investigators, and the three alternatives were manually extracted from descriptions for other videos in the repository. Similar strategies for selecting distractor videos as for Task 4 and Task 5 were employed.

This task was scored as correct or incorrect, thus the score was either 0 or 1 for each trial. The sum was taken across the four trials and then divided by 4 to yield a mean score between 0 and 1. Mean confidence and mean task completion time were also computed.

In short, for each of the above six tasks, the accuracy of each trial and the participants' confidence in their answers were recorded as participants progressed in the study. In addition, the time the participants spent experiencing the surrogates, and the time they spent completing each task were also recorded.

3.3.3.3 Rationale for the Tasks

The gist determination tasks used in the study focused on both recognition and inference measures. Recognition tasks test whether the viewers can recall seeing or/and hearing the stimulus words, frames or actions in the video surrogates they have viewed. The inference tasks are based on what "story" the viewers can construct about the video based on surrogate comprehension.

As summarized in Yang et al. (2003), these recognition measures correspond to the **first two steps** (i.e., *basic perception* and *memory matching*) in the flow diagram of processing of the audiovisual input developed by Grodal (1999), while the inference measures correspond to the **last two steps** in the flow diagram (i.e., ***construction of narrative scene or universe***, and ***reactions at a high level of arousal***).

Specifically, the **keyword determination task** and the **keyframe determination task** were both recognition and inference measures. For example, in the keyframe determination task, some of the correct keyframes were shown to the participants when they consumed the surrogate such that the participants needed to recognize them from the list, while some of the correct keyframes were not shown to the participants such that they needed to infer whether they were correct keyframes for the video or not, based on their comprehension of the video via surrogates.

The recognition nature of the keyword determination task and the keyframe determination task was closely related to the users' real-world task of selecting particular frames for later re-use, while the inference nature of the keywords and keyframe determination tasks was closely related to the users' need for quickly making relevance judgments about videos based on the surrogates.

The free-text gist written task (open-ended), visual excerpt determination task (multiple-choice), audio excerpt determination task (multiple-choice), and verbal gist determination task (multiple-choice) were inference tasks. None of the excerpts or gisting descriptions used in the tasks were shown to the participants during the consumption of the video surrogates.

For the **free-text gist written task** and the **verbal gist determination task**, viewers were asked to specify the gist of the video by writing a gist description or selecting the correct gist description. If the surrogate well supported the viewers' ability to infer the gist of the full video from viewing only the surrogate, the viewers were able to make accurate relevance judgments or selection decisions about videos, thus having a successful and efficient browsing experience. The **audio excerpt determination task** and **visual excerpt determination task** were grounded in users' need to select particular visual and audio clips for various types of re-use or relevance judgments.

Most importantly, these tasks have been recommended and used in previous studies (Yang et al., 2003; Song and Marchionini, 2007; Marchionini et al., 2009), and have been proven reliable and valid in measuring the effectiveness of different video surrogates for video retrieval and sense-making.

3.3.3.4 Experimental Design

The surrogates were assigned to participants so that all participants experienced all six surrogate conditions (including the gold standard ones). The surrogate conditions were counterbalanced to minimize order effects.

With a group of 48 participants, it was impossible to completely counterbalance the order of all 6 surrogate conditions. To completely counterbalance all 6 conditions, we would need at least $6! = 720$ participants. Thus, the surrogate conditions were first grouped into coordinated and uncoordinated. Then the groups were counterbalanced, with the conditions within each group counterbalanced as well. Table 3.11 shows how the surrogate conditions were counterbalanced in this study.

Table 3.11: Experimental Design: Counterbalance the Surrogate Conditions

	$\{V_1, ..., V_5\}$	$\{V_6, ..., V_{10}\}$	$\{V_{11}, V_{12}\}$	$\{V_{13}, ..., V_{17}\}$	$\{V_{18}, ..., V_{22}\}$	$\{V_{23}, V_{24}\}$
1	C1	C2	C3	U1	U2	U3
2	C2	C1	C3	U1	U2	U3
3	C1	C2	C3	U2	U1	U3
4	C2	C1	C3	U2	U1	U3
5	U1	U2	U3	C1	C2	C3
6	U2	U1	U3	C1	C2	C3
7	U1	U2	U3	C2	C1	C3
8	U2	U1	U3	C2	C1	C3

Note: V_i denotes the i_{th} video, where $i = 1, 2, ..., 24$. **C1**, **C2**, and **C3** denote Coordinated I (Systematic Subsampling A + V), Coordinated II (Magic A + V), and Coordinated III (Manual A + V), respectively. **U1**, **U2**, **U3** denote Uncoordinated I (Magic A + Systematic Subsampling V), Uncoordinated II (Magic A + Storyboard V), and Uncoordinated III (Manual A + Manual V), respectively.

In the study, we fixed the order of the videos for which surrogates were displayed to the participants. V_1, V_6, V_{13}, and V_{18} were NASA Destination Tomorrow videos for training purposes. Note that the gold standard conditions C3 and U3 always appeared as the last conditions in the coordinated and uncoordinated groups respectively. Because there were only 4 videos (i.e., V_{11}, V_{12}, V_{23}, and V_{24}) for which the gold standard surrogates are available, we could only represent two videos using C3 and two videos using U3, and we could not afford to waste one of the two for the purpose of training. Therefore, we wanted to ensure that the participants got enough practice from the other surrogate conditions before working on the gold standard ones. With this experimental design, there were only 8 different permutations of the surrogate conditions as shown in Table 3.11.

3.3.3.5 Study Procedure

This study was designed to record performance data, as well as user opinions regarding the tasks. First, an online protocol system was developed using PHP/MySQL and JavaScript to administer the study and collect data and manage the counterbalancing of the surrogate condition orderings. Note that the ordering of the six tasks in each trial did not change across different surrogate conditions.

Participants were recruited for the study by posting mass emails to our university-wide LISTSERV. The study participants were scheduled to attend study sessions in groups of 6-10 people based on participant availability until 48 participants successfully completed the study. The participants were seated at alternating workstations in a computer laboratory with enough identical workstations so that participants did not see others' screens. Headphones were provided and the participants were asked to put on the headphones during the study.

Once the participants were seated, the study session started following the procedures below:

Step 1 The investigator briefed the study participants about the study procedure and participants were asked to read and sign the consent forms.

Step 2 The participants were then asked to fill out a pre-session demographic questionnaire about themselves and their computer and video experiences.

Step 3 Study started. During the study, the participant interacted with several browser-like interfaces with different surrogate conditions on a computer. The ordering of the surrogate conditions were counter-balanced by the PHP codes for each participant. Using one interface (surrogate condition) at a time, the user completed a set of tasks designed for selected video segments.

Step 4 Upon completion of the tasks for one interface, the participants completed a short questionnaire about their experience with the interface. Subjective measures such as usefulness and usability were rated on a 5-point Likert scale, while engagement and enjoyment were rated on 7-point scales.

Step 5 The other interfaces were then presented to each participant one by one and the same process (Step 3 and Step 4) was repeated for different sets of video segments.

Step 6 After completing all the tasks for all interfaces, the participants answered a final short questionnaire on their overall experience in the study.

Step 7 Participants were thanked and awarded $20 for their participation.

Note that the first video a participant experienced in each of the surrogate conditions (except for the gold standard conditions) was treated as a training video segment and discarded from the data analysis phase to reduce the first object effect. During the study, the investigator was approachable by the participants, for giving instructions on the test procedures and answering any procedural questions.

The data collection phase ended when 48 participants had successfully completed the study with useful data.

3.3.3.6 Measures

Measures used to compare the six surrogate conditions included participants' task completion accuracy, participants' confidence ratings in their task responses, the time the participants spent experiencing the surrogates, the time the participants spent completing each task, and a suite of subjective measures. Participants' open-ended comments in the post-session questionnaire were used to enrich the interpretation of performance results, too.

The **surrogate consumption time** (i.e., the time the participants spent experiencing the surrogates) was measured in seconds and was recorded before the participants worked on any of the tasks in each trial. The average of the surrogate consumption time across four trials in each surrogate condition was calculated and used as the unit of analysis for comparison between the surrogate conditions.

Participants' task responses were recorded for each task as they worked, and the **task completion accuracy** was computed for each task, as discussed in Section 3.3.3.2. The accuracy scores for each task were averaged across four trials in each surrogate condition, and the means were used as the units of analysis for making comparisons between surrogate conditions. The accuracy means were in the range of 0.0 - 1.0 for all tasks except for Task 1, for which the accuracy means were in the range of 0.0 - 4.0.

Participants' **confidence ratings** in their task responses were collected for each task trial on a 1-5 rating scale (i.e., 1 - not confident, 5 - very confident). As with the accuracy measures, the confidence ratings were averaged across four trials and the means were used as unites of analysis.

For each task, the **task completion time** (i.e., the time the participants spent completing each task) was also recorded. As with the other measures, the recorded time was averaged across the four trials for each surrogate.

After completing tasks for all trials in each surrogate condition, the participants were also asked to rate their experience with the surrogates based on a set of **subjective** measures of *usability*, *usefulness*, *enjoyment*, and *engagement*. Participants first completed twelve 5-point Likert scale questions (Davis, 1989) on usefulness and usability, and then completed eight 7-point semantic differential scales (Ghani et al., 1991) on engagement and enjoyment.

Table 3.12 shows the twelve 5-point Likert scales related to usefulness and usability.

Table 3.12: Subjective Questions related to Usefulness and Usability

Questions related to "Usefulness"
Using this system enables me to understand video content more quickly.
Using this system improves my performance in understanding video content.
This system makes it easier to understand videos.
Using this system enhances my effectiveness in understanding videos.
I find this system useful for understanding videos.
Using this system increases my productivity in understanding videos.
Using this system helps me better estimate the gist of videos.
Questions related to "Usability"
Learning to operate this system was easy for me.
I found this system to be flexible to interact with.
It would be easy for me to become skillful at using this system.
I found this system easy to use.
My interaction with this system was clear and understandable.
I found it easy to get this system to do what I wanted it to do.

Table 3.13 shows the eight 7-point semantic differential scales adapted from Ghani et al. (1991) that focused on enjoyment and engagement.

As with the other data, participants' responses to each of the four subjective measures were averaged for each participant who provided usable responses. For example, the usefulness measure with one surrogate was the average of a participant's

Table 3.13: Subjective Questions related to Enjoyment and Engagement

Questions related to "Enjoyment"	
Using the video surrogates is:	
1 – not interesting	7 – interesting.
1 – not enjoyable	7 – enjoyable.
1 – dull	7 – exciting.
1 – not fun	7 – fun.
Questions related to "Engagement"	
How you felt using the video surrogates:	
1 – not absorbed intensely	7 – absorbed intensely.
1 – attention was not focused	7 – attention was focused.
1 – did not concentrate fully	7 – concentrated fully.
1 – not deeply engrossed	7 – deeply engrossed.

rating to all questions related to "usefulness" under this surrogate condition. The enjoyment measure with one surrogate was the average of a participant's rating to all questions related to "enjoyment" under this surrogate condition.

After the participant had completed trials with all surrogates, they were asked to fill out a post-session questionnaire, which focused on their overall experience with the surrogate conditions. The participants were asked questions such as which surrogates were easier to use, which surrogates were easier to learn to use, which surrogates they liked and disliked the most, and why. Section 4.10 discussed participants' responses and open-ended comments to the post-session questionnaire.

3.3.4 Data Analysis

As discussed in sections above, the means of the accuracy scores, the confidence ratings, and the time to complete each task across the four trials (excluding the training trial) by each participant were computed for each of the six tasks for each surrogate condition, and used as units of analysis. The time the participants spent experiencing the surrogates in each condition was also averaged to compare across surrogate conditions. Subjective measures such as usefulness and usability, engagement and enjoyment, were each averaged for each participant by surrogate condition.

The data were analyzed using SPSS Release 15. One-way repeated-measures ANOVA was used to compare the effectiveness of 6 surrogate conditions (i.e., 3 coordinated and 3 uncoordinated), for each of the measures discussed above – accuracy, confidence, surrogate consumption time, task completion time, and subjective measures.

ANOVA is one of the most widely used statistical procedures, testing whether means on a dependent variable are significantly different among several (i.e. three or more) groups. In situations where there is a large amount of variation between sample members, error variance estimates from standard ANOVAs are large. To eliminate or reduce individual differences as a source of between-group differences (and to create a more powerful test), this study used the repeated-measures design (also known as a within-subject design), which compares measures repeated across more than one condition.

As with the standard ANOVA, repeated-measures ANOVA tests the equality of means. Repeated-measures ANOVA is used when all members of a random sample are measured under a number of different conditions. In a repeated-measures design, the measurement of the dependent variable is repeated, and the data violates the assumption of independence, such that the standard ANOVA will not be able to model the correlation between the repeated measures. Repeated-measures of each sample member provides a way of accounting for the variance between sample members, and reducing error variance.

For each of the six tasks, the means and standard deviations of each measure among all 48 participants were reported for each of the surrogate conditions. The overall main effect F values and probabilities of statistical reliability were reported too. When the main effects were found to be statistically reliable at less than the 0.05 level, pairwise contrasts among the surrogate conditions were conducted and reported. We were interested in seeing whether the data supported the hypothesis that the uncoordinated surrogates were more helpful than the coordinated ones, and that the manually-generated surrogates were more helpful than the automatically-generated surrogates, on the six gist tasks.

Moreover, written comments made by participants about each surrogate condition in the open-ended questions were also examined to complement the quantitative results.

Chapter 4

EXPERIMENTAL RESULTS

This chapter presents the results of the user study conducted to evaluate some multi-modal surrogates, the pros and cons of the uncoordinated surrogates (i.e., mutlimodal surrogates with uncoordinated audio and visual channels), as well as the performance differences between the manually-generated and the automatically-generated surrogates. The findings were described and discussed in terms of the user performance (i.e., accuracy and confidence), surrogate consumption time, task completion time, and users' affective measures of different multi-modal surrogate interfaces.

The study was conducted on Jan. 26, Jan. 27, Jan. 28, and Jan. 29 of 2010 in the computer lab of the School of Information and Library Science at the University of North Carolina at Chapel Hill. The study participants were scheduled to attend study sessions in groups of 4-10 people based on participant availability. Participants were seated at identical alternating workstations in the computer laboratory so they could not see each others' screens and were asked to wear headphones during the study. Each test session ran 1.5 hours on average with a few participants taking up to 2.5 hours.

4.1 Study Participants

Forty-eight participants were recruited for the study by posting mass emails to the university-wide LISTSERV of the University of North Carolina at Chapel Hill. Participants selected from the responses were self-assessed native English speakers with adequate listening and visual abilities, who used computers daily and had experience with searching for videos using computers at least occasionally.

We actually ran 55 participants before we ended up with 48 participants who provided usable data. One of the participants did not meet our participant selection criteria of "Having experience with searching for videos using computers at least occasionally", and we did not find it out until the participant filled out the demographic questionnaire. Another 6 participants failed to provide usable data to the subjective questions (i.e., each had a massive amount of missing entries) after completing each surrogate condition. Therefore, the data from these 7 participants were discarded, and the data from the 48 participants who provided valid data entries were analyzed to address the research questions in this study.

The participants in this study had a mean age of 25, were about 70% female, and half of them were undergraduates. Table 4.1 summarizes the basic demographic characteristics of the study participants.

Table 4.1: Demographic Characteristics of Study Participants

Age		**Gender**		**Academic Status**	
Minimum	18	Female	34	Undergraduate	24
Maximum	63	Male	14	Graduate	11
Mean	25			Staff	9
Median	22			Volunteer	1
Std. dev.	8.5			Former students	3

The participants came from a wide variety of 29 different academic departments

Table 4.2: Department Affiliation of Study Participants (*Note*: Numbers in brackets indicate affiliation with more than one department.)

Department		**Department**	
Information & Library Science	6 (1)	English	4
Journalism and Mass Communication	4	History	3 (2)
Anthropology	2 (1)	Biology	2
Business	2 (1)	Epidemiology	2
Medicine	2	Psychology	2
Chemistry	1	College of Arts and Sciences	1
Developmental Science	1	Economics	1 (1)
Environmental science	1	Exercise and Sport Science	1
French	1	General College	1
Nursing	1	Nutrition	1
Pharmacy	1	Philosophy	1
Public Policy	1	School of Government	1
Statistics and Operations Research	1	University Relations	1
Vocal Performance	1	No Department Affiliation	3
No Affiliation Information Provided	3		

Table 4.3: Participants' Experience with Computers and Video Use. (*Values:* 1 - Never, 2 - Occasionally, 3 - Monthly, 4 - Weekly, 5 - Daily.)

Question	**Mean**	**Std. Dev.**
How often do you use a computer?	4.92	0.45
How often do you watch videos or films?	4.23	0.81
How often do you search for videos or films?	3.81	1.02

in the university, as shown in Table 4.2. Some of the participants were affiliated with more than one department.

In terms of their experience with computers and video use, most of them were quite familiar with using computers and watching videos, but had moderate experience searching for video (see Table 4.3).

As for where the participants search for films or videos, all of the 48 participants (100%) search online, 10 of them (20.8%) search in newspapers or magazines, 10 of them (20.8%) search in film archives, and 2 of them search in video rental stores

Table 4.4: Participants' Video Searching Strategies

Where		**How**	
Online	48	By title	41
Newspaper or Magazine	10	By author or actor	14
Film archives	10	By topic	30
Other: e.g., video rental stores	2	By trailer	13
		Other: e.g., By director	2
		Other: e.g., By keyword	1
		Other: e.g., By language	1

Table 4.5: Participants' Video Searching Purposes

Purposes	
Entertainment only	16
Entertainment & Educational	14
Entertainment & Work	4
Entertainment & Information or news	5
Entertainment & Instruction	2
Entertainment & Social (i.e., videos sent or posted by friends)	4
Entertainment & Edification	1
Entertainment, Politics, Education	1
Information only (i.e., to find out certain things)	1

(see Table 4.4).

Table 4.5 summarizes the purposes for which the participants usually search for videos or films. All of the 48 participants except one (who search for video only "to find out certain things", i.e., information only) search for videos or films for "entertainment". 16 of them search purely for entertainment, others search for videos for entertainment as well as for other purposes, including "knowledge", "education", "work-related help", "news", "instruction", "edification", etc. Some participants search for "video lessons (how-to, DIY, etc.)". A few participants search for videos they "have heard about" from friends.

To investigate the research questions proposed in Chapter 1.2, I compared the 6 surrogate conditions on measures of participants' performance (accuracy) and confidence across the six tasks, measures of time to consume the surrogate and time to complete the tasks, as well as some subjective measures of usefulness, usability, engagement, and enjoyment. Table 4.6 summarizes the measures used in this study. One-way repeated-measures ANOVA was conducted to investigate the differences among the six surrogate conditions on these measures, and the results are discussed in the following sections.

Table 4.6: Measures Used in the Study

	Measures
1.	Surrogate consumption time
2.	Performance scores (Accurary)
3.	Confidence ratings
4.	Task completion time
5.	Subjective measures

4.2 Surrogate Consumption Time

The surrogate consumption times were recorded before the participants worked on the six tasks, hence they are associated with each surrogate condition rather than with each task. Table 4.7 presents the means and standard deviations of the time participants spent consuming each of the six surrogates, as well as the number of times a surrogate condition was replayed and stopped by the participants.

Note that for the coordinated surrogate, the audio and visual channels of the surrogates were sampled simultaneously and thus were pre-coordinated at indexing time, while for the uncoordinated surrogates, the audio and visual channels of the surrogates were sampled independently (i.e., were uncoordinated) and needed to be integrated in the user's head at consumption time. Thus, we expected people would spend more time consuming the uncoordinated surrogates than consuming the

Table 4.7: Surrogate Consumption Time (N = 48, df = 5, 43)

Surrogate	**Consumption time** (sec)		♯ of	♯ of
	Mean	SD	**replays**	**stops**
C1	55.12	18.59	56	5
C2	55.74	14.32	47	10
U1	53.62	14.71	44	5
U2	53.67	17.42	48	13
C3	51.78	8.89	24	10
U3	57.07	18.73	40	16

coordinated ones. But surprisingly, there are no statistically significant differences at the 0.05 level in the surrogate consumption time among the six surrogate conditions (F = 2.187, p = 0.073). The uncoordinated surrogate did not result in longer surrogate consumption time than the coordinated ones as we hypothesized.

As a matter of fact, the mean consumption time for the uncoordinated surrogates U1 (Magic A + Storyboard V) and U2 (Magic A + Systematic subsampling V) were slightly less than the mean consumption time for the coordinated surrogates C1 (Systematic subsampling A + V) and C2 (Magic A+ V), though the differences were not statistically significant.

We also logged the number of replays and stops the participants performed when consuming the surrogates. According to Table 4.7, the coordinated surrogate C3 (Manual A + V) was replayed the fewest times (i.e., 24 times) among all surrogates, which reinforced the result in surrogate consumption time that it was faster to consume C3 than the other surrogates. Note that 24 times only account for 1/8 of all video plays (i.e., 24 (replays) ÷ 48 (participants) ÷ 4 (trials) = 1/8).

The coordinated surrogate C1 (Systematic subsampling A + V) was played the most times (i.e., 56 times), and the uncoordinated surrogate U2 was played the second most times (i.e., 48 times) among all surrogates. The surrogates were stopped for fewer times than the times they were replayed, e.g., 5 times for C1 and

U1, and 16 times for U3.

4.3 Task 1: Free-text Gist Written Task

As discussed previously, participants' responses to the free-text gist written task were scored by two independent coders after the two established a common baseline for scoring. The two coders' scores to the task responses were found to be highly correlated (r = 0.691) and the correlation was significant at the 0.001 level. Therefore, the two sets of scores were averaged for each trial and we took the averages as the scores the participants got for Task 1 in each trial.

For this and the other five tasks, results from each of the four trials (excluding the first one of the five trials) for each task-surrogate pair were averaged and these mean scores were used as the unit of analysis for each participant. One-way repeated-measures ANOVA (SPSS Release 15) was used to test the main effects across the six different surrogate conditions.

Table 4.8: Task 1. Free-text Gist Written Task (N = 48, df = 5, 43)

Surrogate	**Accuracy*** (Max: 4.0)		**Confidence*** (Max: 5.0)		***T*_Time*** (sec)	
	Mean	SD	Mean	SD	Mean	SD
C1	1.451	0.508	3.557	0.854	40.010	25.382
C2	1.638	0.519	3.948	0.682	42.063	24.521
U1	1.602	0.550	3.859	0.787	39.177	23.176
U2	1.638	0.594	3.682	0.993	37.635	18.838
C3	2.026	0.615	4.031	0.925	35.125	18.099
U3	1.797	0.678	3.750	0.899	31.875	19.937

Table 4.8 presents the means and standard deviations for accuracy, confidence, and task completion time for each of the six surrogates for Task 1 (Free-text gist written task) In this and other tables in this chapter, asterisks in column headings

denote statistical significance in the main effects at the 0.05 probability level.

Note that for Task 1 accuracy, the maximum possible score is 4.0. Thus the accuracy for the free-gist written task actually ranged from 36.263% (1.451 out of 4.0) to 50.651% (2.026 out of 4.0) with the six different surrogates. The free-text gist written task had the lowest accuracy among the six tasks, while the completion time for this task was the greatest among the six tasks. The free-text gist written task is unquestionably the most difficult task in this study.

As shown in Table 4.8, strong effects of different surrogates were observed for this task in accuracy (F(5, 43) = 10.300, p < 0.001), participants' confidence ratings (F(5, 43) = 4.572, p = 0.002), and task completion time (F(5, 43) = 5.826, p < 0.001). Figure 4.1, Figure 4.2, and Figure 4.3 present the accuracy, the confidence ratings, and the task completion time on the free-text gist written task across the six surrogates.

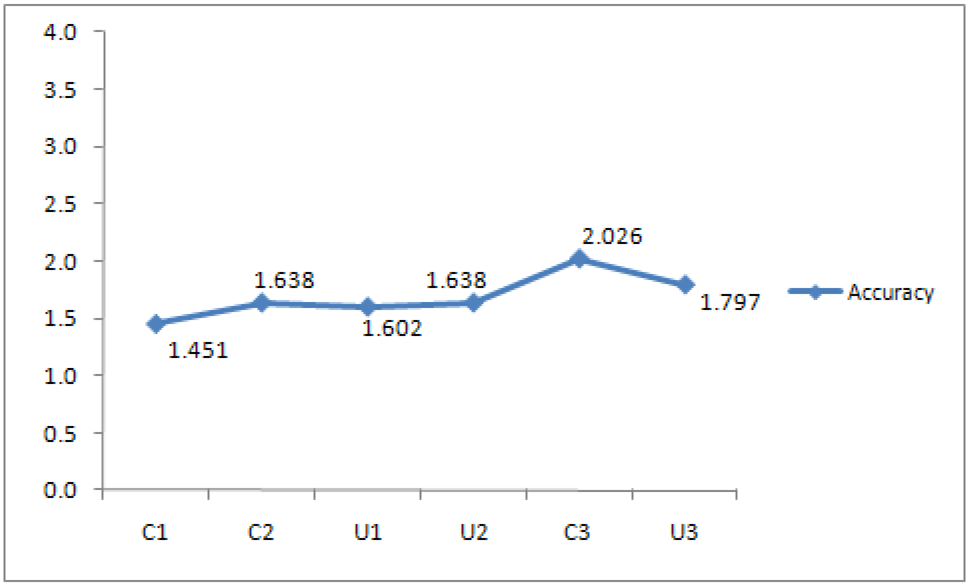


Figure 4.1: Mean Accuracy for Task 1 (Free-text Gist Written Task) by Surrogate Condition. Possible data value range: 0 - 4.0.

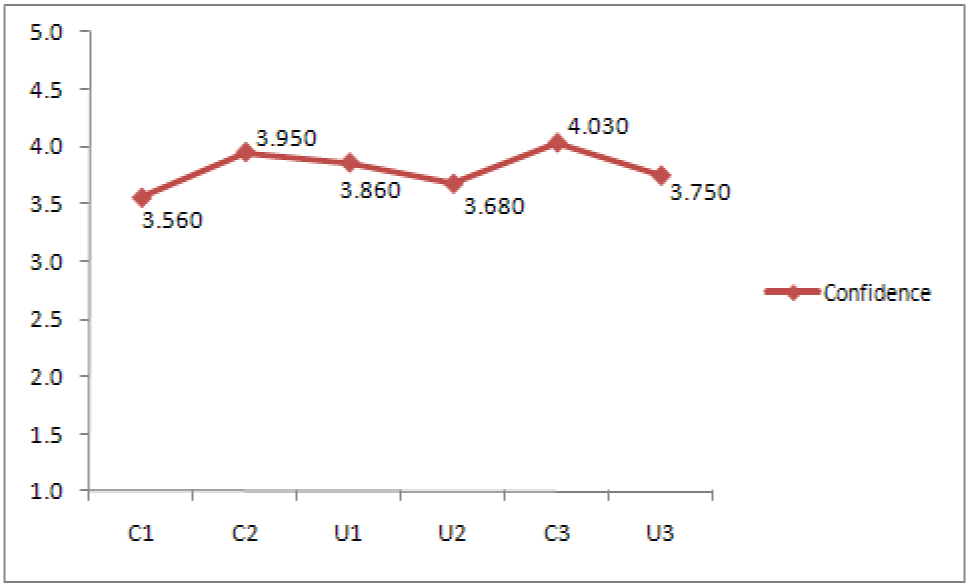


Figure 4.2: Participants' Mean Confidence for Task 1 (Free-text Gist Written Task) by Surrogate Condition. Possible data value range: 1 - 5.

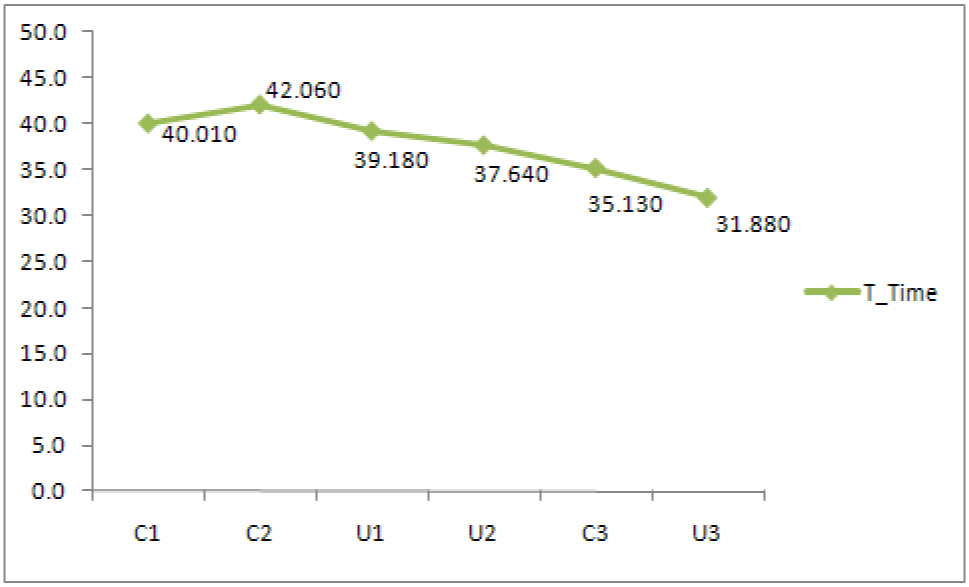


Figure 4.3: Mean Task Completion Time (in seconds) for Task 1 (Free-text Gist Written Task) by Surrogate Condition.

Among all surrogates except two manually-generated surrogates C3 (Manual A + V) and U3 (Manual A + Manual V), the uncoordinated surrogates yielded better overall results than the coordinated surrogates, in terms of accuracy, confidence ratings, and task completion time.

Despite the strong main effects of different surrogates on accuracy for the free-text gist written task, not all pairwise differences in the accuracy means discussed above were statistically reliable at the 0.05 level. Table 4.9 reports the pairwise comparisons of the accuracy means between each surrogate condition pair. The asterisks by the numbers in the table denote mean difference significant at the 0.05 level.

The automatically-generated coordinated surrogate C1 (Systematic subsampling A + V) led to the lowest accuracy, and the manually-generated surrogates C3 (Manual A + V) and U3 (Manual A + Manual V) led to the highest accuracy among all surrogates.

Note that 50% of Table 4.9 is redundant since the pairwise contrasts are symmetric, and we are only interested in the differences which are statistically reliable at the 0.05 level. Therefore, in the following paragraphs, I will omit the full tables for the pairwise comparisons, and report only the pairwise contrasts which were statistically reliable at the 0.05 level.

In this study, the manually-generated coordinated surrogate C3 (Manual A + V) led to the highest accuracy mean (2.026/4 = 50.651% accuracy) and the highest confidence ratings (4.030 out of 5 confidence) on the free-text gist written task across all 6 surrogates, while the manually-generated uncoordinated surrogate U3 (Manual A + Manual V) led to the second highest accuracy mean (1.797/4 = 44.922% accuracy) and reasonable confidence ratings (3.750 out of 5 confidence). Participants were the least confident (3.560 out of 5 confidence) in performing the free-text gist

Table 4.9: Pairwise Comparisons for Task 1 Accuracy

		Mean			**95% Confidence Interval**	
Cond.	**Cond.**	**Difference**	**Std.**	**Sig.**^{a}	**for Difference**	
(I)	(J)	(I-J)	**Error**		Lower Bound	Upper Bound
C1	C2	-0.188*	0.084	0.031	-0.357	-0.018
	C3	-0.576*	0.080	0.000	-0.737	-0.414
	U1	-0.151	0.081	0.070	-0.315	0.013
	U2	-0.188	0.092	0.047	-0.372	-0.003
	U3	-0.346*	0.110	0.003	-0.567	-0.126
C2	C1	0.188*	0.084	0.031	0.018	0.357
	C3	-0.388*	0.095	0.000	-0.579	-0.197
	U1	0.036	0.091	0.691	-0.147	0.220
	U2	0.000	0.083	1.000	-0.167	0.167
	U3	-0.159	0.100	0.117	-0.359	0.041
C3	C1	0.576*	0.080	0.000	0.414	0.737
	C2	0.388*	0.095	0.000	0.197	0.579
	U1	0.424*	0.094	0.000	0.235	0.614
	U2	0.388*	0.089	0.000	0.209	0.567
	U3	0.229	0.115	0.052	-0.002	0.460
U1	C1	0.151	0.081	0.070	-0.013	0.315
	C2	-0.036	0.091	0.691	-0.220	0.147
	C3	-0.424*	0.094	0.000	-0.614	-0.235
	U2	-0.036	0.108	0.738	-0.254	0.181
	U3	-0.195	0.111	0.084	-0.418	0.027
U2	C1	0.188*	0.092	0.047	0.003	0.372
	C2	0.000	0.083	1.000	-0.167	0.167
	C3	-0.388*	0.089	0.000	-0.567	-0.209
	U1	0.036	0.108	0.738	-0.181	0.254
	U3	-0.159	0.119	0.188	-0.398	0.080
U3	C1	0.346*	0.110	0.003	0.126	0.567
	C2	0.159	0.100	0.117	-0.041	0.359
	C3	-0.229	0.115	0.052	-0.460	0.002
	U1	0.195	0.111	0.084	-0.027	0.418
	U2	0.159	0.119	0.188	-0.080	0.398

* - The mean difference is significant at the .05 level.
a - Adjustment for multiple comparisons: *Least Significant Difference* (equivalent to no adjustments).

written task using C1 (Systematic subsampling A + V), which directly parallels the accuracy result. Pairwise comparisons show that the confidence ratings using C1 were significantly lower than using C2 (Magic A + V), C3 (Manual A + V), and U1 (Magic A + Storyboard V).

The coordinated surrogate C2 (Magic A + V) led to statistically reliable higher accuracy and higher confidence levels than the coordinated surrogate C1 (Systematic subsampling A + V) at 0.05 the level. Surrogate C2 (Magic A + V) was created by extracting visual snippets with coordinated visual and audio channels from the video based on MAGIC extracted text descriptions. The MAGIC text summarization process extracts and ranks sentences from the video transcripts using a variety of summarization techniques including discourse segmentation and topic shift detection (Li et al. (2005), as summarized in Song and Marchionini (2007)), while surrogate C1 (Systematic subsampling A + V) was created by systematically extracting a 5-second video clip out of every 120-second interval, with sentences and words often chopped in the middle. This explains the better performance of C2 (Magic A + V) over C1 (Systematic subsampling A + V).

The one-way repeated-measures ANOVA for task completion time showed statistically reliable differences across the surrogates ($F = 5.826$, $p < 0.001$). In general, participants performed the free-text gist written task more quickly with the automatically-generated uncoordinated surrogates than with the automatically-generated coordinated surrogates, and participants performed the free-text gist written task more quickly with the manually-generated surrogates than with the automatically-generated surrogates.

Next, we will report the pairwise comparisons among the surrogates and organize the results on Task 1 by the two research questions we proposed to address in this study.

4.3.1 Coordinated vs. Uncoordinated Surrogates

Table 4.10 shows the mean differences and p values for the pairwise comparisons (Coordinated vs. Uncoordinated) for the free-text gist written task in terms of accuracy, confidence, and task completion time. The differences that are not statistically significant at the 0.05 level are not included in the table.

Table 4.10: Summary of Pairwise Comparisons (Coordinated vs. Uncoordinated) Significant at 0.05 Level for Task 1 Accuracy, Confidence, and Task Completion Time.

	Accuracy	
C1 < U2	mean difference = -.188	p = 0.047
C1 < U3	mean difference = -.346	p = 0.003
C3 > U1	*mean difference = .424*	*p < 0.001*
C3 > U2	*mean difference = .388*	*p < 0.001*
	Confidence	
C1 < U1	mean difference = -.302	p = 0.019
C3 > U2	*mean difference = .349*	*p = 0.035*
C3 > U3	*mean difference = .281*	*p = 0.042*
	T_Time	
C1 > U3	mean difference = 8.135	p = 0.044
C2 > U3	mean difference = 10.188	p = 0.005

Note: Pairwise comparisons which were against the hypothesis were emphasized in *Italic*.

We hypothesized that the uncoordinated surrogates, when carefully designed and sampled, may carry more useful information than coordinated surrogates. For the free-text gist written task, as expected, the participants performed statistically reliably better with the uncoordinated surrogates U2 (Magic A + subsampling V) and U3 (Manual A + Manual V) than with the coordinated surrogate C1 (systematic subsampling A + V).

We also expected that the uncoordinated surrogates U1 (Magic A + Storyboard V) and U2 (Magic A + Systematic subsampling V) would lead to better sense-making than the coordinated surrogate C2 (Magic A + V). But there was no difference. The MAGIC-based coordinated surrogate C2 (Magic A + V) had very high summarizing quality, hence it was able to compete with the uncoordinated surrogates U1 (Magic A + Storyboard V) and U2 (Magic A + Systematic subsampling V) on the gist written task.

There was no statistically reliable difference in Task 1 accuracy (mean difference = .229, p = 0.092) between the two manually-generated surrogates C3 (Manual A + V) and U3 (Manual A + Manual V).

Furthermore, we hypothesized that people would be more confident in their task responses using uncoordinated surrogates than using the coordinated surrogates. As expected, the participants' confidence levels in their gist written responses were statistically reliably higher with the uncoordinated surrogate U1 (Magic A + Storyboard V) than with the coordinated surrogate C1 (Systematic subsampling A + V). Participants were the least confident in performing the free-text gist written task using C1 (Systematic subsampling A + V), which directly parallels the accuracy result.

However, the participants were statistically reliably more confident in their gist written task responses when presented with the coordinated surrogate C3 (Manual A + V) than when presented with the uncoordinated surrogate U2 (Magic A + Systematic subsampling V) or U3 (Manual A + Manual V), which contradicted our hypothesis that people would be more confident in their task responses with the uncoordinated surrogates.

Not only were there no statistically reliable differences in the surrogate consumption time for all surrogate (as discussed in Section 4.2), but it also took the participants statistically reliably less time performing the free-text gist written task with the uncoordinated surrogate U3 (Manual A + Manual V) than the coordinate surrogates C1 (Systematic subsampling A + V) and C2 (Magic A+ V).

In conclusion, the uncoordinated surrogates were effective for the gist written task and did not penalize efficiency in either consuming surrogates or completing the task.

4.3.2 Manually-generated vs. Automatically-generated Surrogates

Automatic surrogation (or summarization) is a difficult challenge (Mani and Maybury, 1999; Marchionini et al., 2009), thus we expected that the manually-generated surrogates would lead to higher accuracy, higher confidence ratings, and less task completion time than the automatically-generated ones. Table 4.11 summaries the mean differences and p values for the pairwise comparisons (Manually-generated vs. Automatically-generated) for the free-text gist written task in terms of accuracy, confidence, and task completion time which are significant at the 0.05 level.

For the free-text gist writing task, the manually-generated surrogates led to uniformly better performance than the automatically-generated surrogates, which is as we expected. As shown in Table 4.11, the manually-generated surrogate C3 (Manual A + V) led to statistically reliably better free-text gist written accuracy than all four automatically generated surrogates: C1 (Systematic subsampling A + V), C2 (Magic A + V), U1 (Magic A + Storyboard V), and U2 (Magic A + Systematic subsampling V). And the manually-generated surrogate U3 (Manual A + Manual V) led to significantly reliably better gist written accuracy than the automatically-generated surrogate C1 (Systematic subsampling A + V).

The participants were statistically reliably more confident in their task responses

Table 4.11: Summary of Pairwise Comparisons (Manually vs. Automatically) Significant at 0.05 Level for Task 1 Accuracy, Confidence, and Task Completion Time.

	Accuracy	
C3 > **C1**	mean difference = .576	p < 0.001
C3 > **C2**	mean difference = .388	p < 0.001
C3 > **U1**	mean difference = .424	p < 0.001
C3 > **U2**	mean difference = .388	p < 0.001
U3 > **C1**	mean difference = .346	p = 0.003
	Confidence	
C3 > **C1**	mean difference = .474	p < 0.001
C3 > **U2**	mean difference = .349	p = 0.035
	T_Time	
C3 < **C2**	mean difference = -6.938	p = 0.025
U3 < **C1**	mean difference = -8.135	p = 0.044
U3 < **C2**	mean difference = -10.188	p = 0.005
U3 < **U1**	mean difference = -7.302	p = 0.003
U3 < **U2**	mean difference = -5.760	p = 0.006

with the manually-generated surrogate C3 (Manual A + V) than with automatically generated surrogates C1 (Systematic subsampling A + V) and U2 (Magic A + Systematic subsampling V), which directly parallels the accuracy result.

Also as expected, the participants completed the free-text gist written task more quickly using the manually-generated surrogates C3 (Manual A + V) and U3 (Manual A + Manual V) than using the automatically-generated surrogates.

4.3.3 Task 1 Summary

Task 1 (Free-text gist written task) was the most difficult task in this study, with accuracy ranging from 36% to 50% for the six surrogates. For pairwise comparisons with significantly reliable differences, participants performed the task more accurately and more quickly with the uncoordinated surrogate than with the coordinated surrogates. Participants also performed the task more accurately and more quickly with the manually-generated surrogates than with the automatically-generated surrogates. These were expected results.

Results on participants' confidence ratings on their task 1 responses were not all as hypothesized. Participants were statistically reliably more confident in their responses with the uncoordinated surrogate U1 (Magic A + Storyboard V) than with the coordinated surrogates C1 (Systematic subsampling A + V), which was expected, and they were statistically reliably more confident in their responses with the coordinated surrogate C3 (Manual A + V) than with the uncoordinated surrogates U2 (Magic A + Systematic subsampling V) and U3 (Manual A + Manual V), which was surprising. Though the uncoordinated surrogate U3 (Manual A + Manual V) yielded the second highest accuracy and the shortest task completion time among all surrogates, participants were not as confident (3.75 out of 5) in their responses.

The automatically-generated coordinated surrogate C1 (Systematic subsampling A + V) was composed of incomplete sentences or phrases of 5 seconds each, hence was not a very informative surrogate for the free-text gist written task. On the contrary, the coordinated surrogate C2 (Magic A + V), which was automatically-generated based on MAGIC text summarization techniques, was actually quite good for the task.

4.4 Task 2: Keyword Determination Task

Table 4.12 summarizes the means and standard deviations for accuracy, confidence, and task completion time for each of the six surrogates for Task 2 (Keyword determination task). Although people did reasonably well on the keyword determination tasks, they had the lowest accuracy mean among all the tasks excluding the free-text gist written task, suggesting that picking the right keywords for the video based on the surrogates of the video is not an easy task, which was further reinforced by people's relatively low confidence levels about their performance in this task.

Table 4.12: Task 2. Keyword Determination Task (N = 48, df = 5, 43)

Surrogate	**Accuracy*** (Max: 4.0)		**Confidence** (Max: 5.0)		***T*_Time*** (sec)	
	Mean	SD	Mean	SD	Mean	SD
C1	0.786	0.085	3.594	0.747	12.990	4.154
C2	0.765	0.086	3.651	0.761	13.307	4.293
U1	0.768	0.087	3.651	0.849	12.792	3.957
U2	0.773	0.095	3.568	0.949	12.443	3.719
C3	0.704	0.098	3.594	0.932	12.083	4.414
U3	0.703	0.095	3.552	0.924	11.042	4.137

Strong effects of different surrogates were observed for this task in accuracy (F(5, 43) = 12.749, $p < 0.01$) and task completion time (F(5, 43) = 4.624, p = .002), but not in participants' confidence ratings (F(5, 43) = .244, p = .941). Figures 4.4, 4.5, and 4.6 present the accuracy, confidence ratings, and task completion time with the six surrogates for task 2.

We hypothesized that the uncoordinated surrogates, when carefully designed and sampled, may carry more useful information than the coordinated surrogates; hence, people would do better on the tasks with the uncoordinated surrogates than with the coordinated surrogates. Also, we hypothesized that people would perform the tasks better with the manually-generated surrogates than the automatically-generated surrogates.

However, for the keyword determination task, the coordinated automatically-generated surrogate C1 (systematic subsampling A+V) yielded the highest accuracy

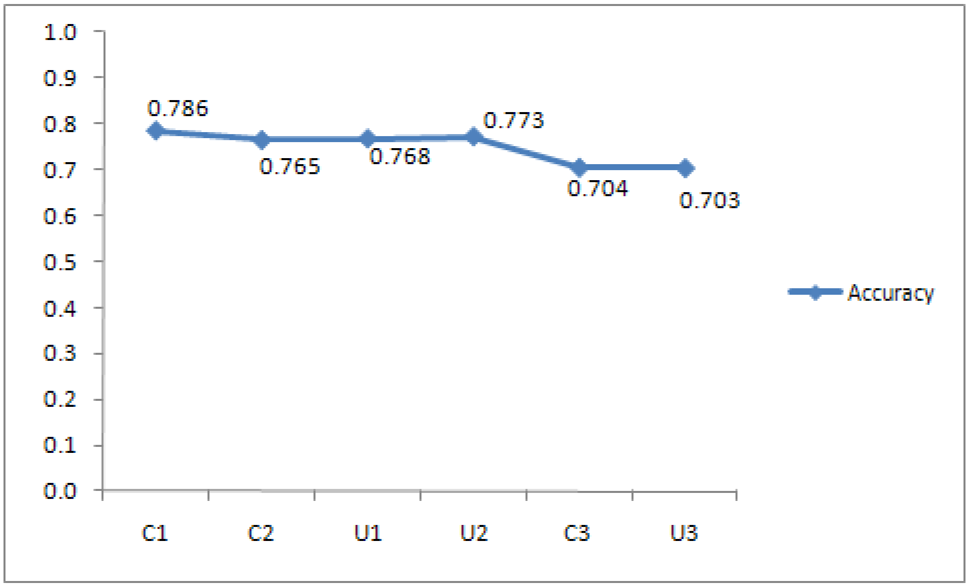


Figure 4.4: Mean Accuracy for Task 2 (Keyword Determination Task) by Surrogate Condition. Possible data value range: 0.0 - 1.0

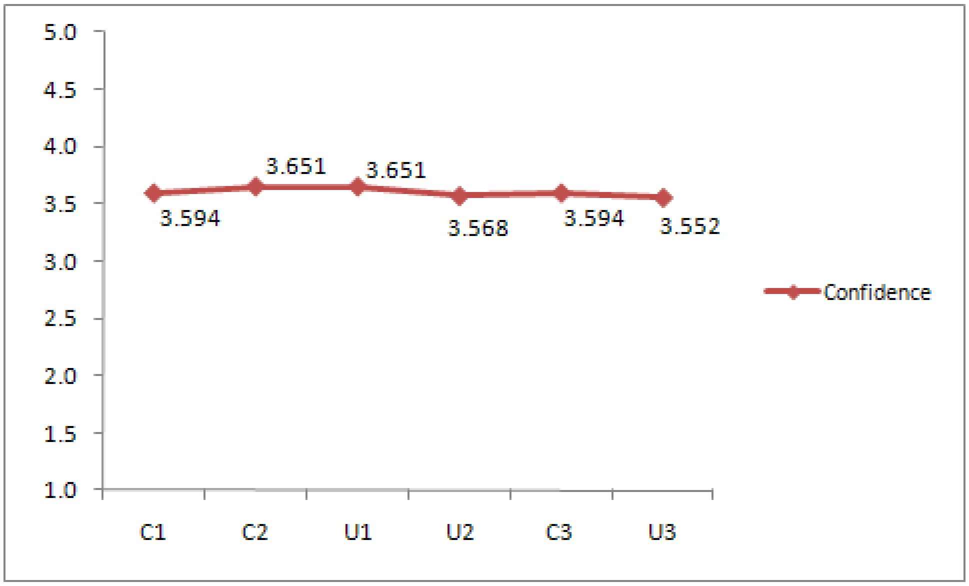


Figure 4.5: Participants' Mean Confidence for Task 2 (Keyword Determination Task) by Surrogate Condition. Possible data value range: 1 - 5.

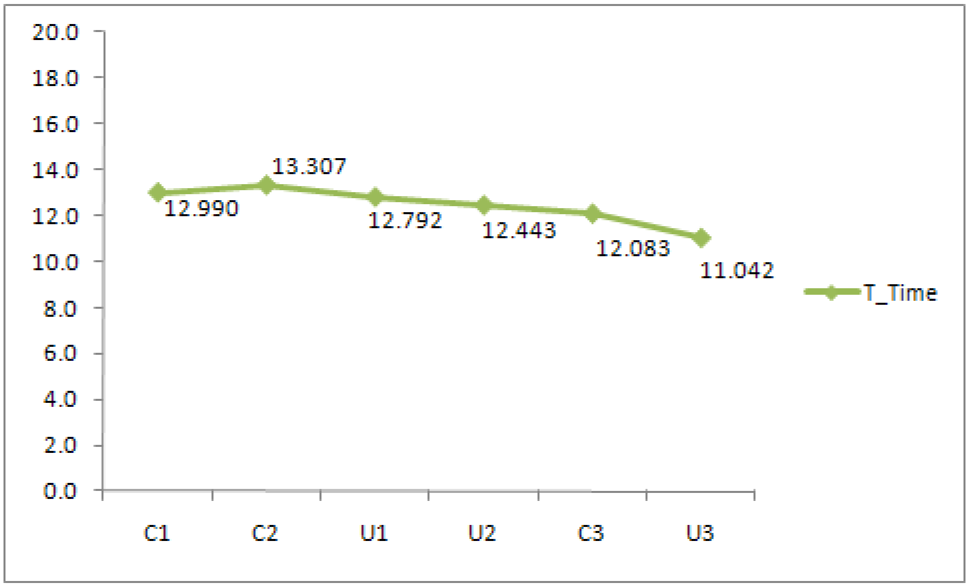


Figure 4.6: Mean Task Completion Time (in seconds) for Task 2 (Keyword Determination Task) by Surrogate Condition

mean values and lowest variability, while the manually-generated surrogates C3 (manual A + V) and U3 (manual A + manual V) yielded the lowest accuracy mean values and highest variability.

Although the results in accuracy and confidence ratings for the keyword determination task were not quite as hypothesized, the task completion time was. In general, the participants performed the keyword determination task more quickly with the uncoordinated surrogates than with the coordinated surrogates, and more quickly with the manually generated one than with the automatically-generated ones.

Next we report the pairwise comparisons among the surrogates and present experimental results organized by the two research questions.

4.4.1 Coordinated vs. Uncoordinated Surrogates

For the keyword determination task, as expected, the participants performed statistically reliably better with the uncoordinated surrogates U1 (magic A + storyboard V) and U2 (magic A + systematic subsampling V) than the coordinated surrogate C3 (manual A + V) in terms of accuracy.

However, it is surprising that the coordinated surrogates C1 (Systematic subsampling A + V) and C2 (Magic A + V) were statistically reliably more helpful than the uncoordinated surrogate U3 (manual A + manual V) in performing the keyword determination task. In fact, the uncoordinated surrogate U3 (manual A + manual V) yielded the lowest accuracy mean, while the coordinated surrogate C2 (Magic A + V) and C1 (Systematic subsampling A + V) yielded the highest and second highest accuracy means for the keyword determination task.

Table 4.13 shows the mean differences and p values for the pairwise comparisons between the coordinated and the uncoordinated surrogates for the keyword determination task, for those that are significant at the 0.05 level. (Note: There were no statistically significant differences in confidence ratings across surrogates.)

Table 4.13: Summary of Pairwise Comparisons (Coordinated vs. Uncoordinated) Significant at 0.05 Level for Task 2 Accuracy

	Accuracy	
U1 > **C3**	mean difference = .064	p = 0.001
U2 > **C3**	mean difference = .069	p < 0.001
C1 > ***U3***	*mean difference = .083*	*p < 0.001*
C2 > ***U3***	*mean difference = .062*	*p = 0.001*
	***T*_Time**	
C1 > **U3**	mean difference = 1.948	p = 0.029
C2 > **U3**	mean difference = 2.266	p = 0.005

Note: Pairwise comparisons which were against the hypothesis were emphasized in *Italic*.

The coordinated surrogates C1 and C2 were surprisingly good in helping people determine the right keywords for the video based on the video surrogates, and participants were reasonably confident about their performance.

The number of choices and fine granularity of keywords made the keyword determination task more of a measure of specific token recognition rather than a measure of inference. Therefore, the most helpful surrogates for the free-gist written task could also be the least helpful surrogates for the keyword determination task, and vice versa. The coordinated surrogate C1 was created by systematically subsampling the audio and visual channels of the video, thus it provided pieces of information throughout the video with fine granularity. Hence it is not difficult to explain the great performance of C1 for the keyword determination task. Also note that the visual stimuli of the uncoordinated surrogate U2 (magic A + systematic subsampling V) was created by systematic subsampling of the visual channel, which made the surrogate very helpful (i.e., yielding the second highest accuracy) for this specific token recognition task.

4.4.2 Manually-generated vs. Automatically-generated Surrogates

The expectations that people would do better on the task with the manually-generated surrogate than the automatically-generated surrogate were not borne out. The automatically-generated surrogates were surprisingly good in helping people perform the keyword determination task. In this study, all four automatically-generated surrogates C1 (systematic subsampling A + V), C2 (magic A + V), U1 (magic A + storyboard V), and U2 (magic A + systematic subsampling V) are significantly more helpful than the manually-generated surrogates C3 (manual A +

V) and U3 (manual A + manual V) in terms of accuracy. The manually-generated surrogates C3 (manual A + V) and U3 (manual A + manual V) yielded the lowest accuracy mean values for the keyword determination task in this study. There was no statistically reliable difference in task 2 accuracy between the two manually-generated surrogates C3 and U3 (mean difference = .001, p = 0.957). Table 4.14 summarizes the mean differences and p values for the pairwise comparisons that are significant at the 0.05 level.

Table 4.14: Summary of Pairwise Comparisons (Automatically vs. Manually) Significant at 0.05 Level for Task 2 Accuracy.

Accuracy		
C1 *>* ***C3***	*mean difference = .082*	*p < 0.001*
C2 *>* ***C3***	*mean difference = .061*	*p < 0.001*
U1 *>* ***C3***	*mean difference = .064*	*p = 0.001*
U2 *>* ***C3***	*mean difference = .069*	*p < 0.001*
C1 *>* ***U3***	*mean difference = .083*	*p < 0.001*
C2 *>* ***U3***	*mean difference = .062*	*p = 0.001*
U1 *>* ***U3***	*mean difference = .065*	*p < 0.001*
U2 *>* ***U3***	*mean difference = .070*	*p < 0.001*
***T*_Time**		
C1 > **U3**	mean difference = 1.948	p = 0.029
C2 > **U3**	mean difference = 2.266	p = 0.005
U2 > **U3**	mean difference = 1.750	p = 0.003
C2 > **C3**	mean difference = 1.224	p = 0.017

Note: Pairwise comparisons which were against the hypothesis were emphasized in *Italic*.

The manually-generated uncoordinated surrogate U3 (manual A + manual V) help the participants perform the keyword determination task statistically reliably more quickly than 3 out of the 4 automatically-generated surrogates C1 (systematic subsampling A + V), C2 (magic A + V), and U2 (magic A + systematic subsampling

V).

4.4.3 Task 2 Summary

To summarize, although people did reasonably well on the keyword determination tasks, they had the lowest accuracy means and the second lowest mean confidence ratings among all the tasks.

The low accuracy and confidence ratings for the keyword determination task may be explained by the fact that many distractors for the keyword determination task were selected from keywords for other videos in the same video collection as the test video, which made some distractors very plausible.

Moreover, the number of choices and fine granularity of keywords made the keyword determination task more of a measure of specific token recognition rather than a measure of gist (inference). The coordinated surrogate C1 (Systematic subsampling A + V) provided pieces of information throughout the video with fine granularity, so did the visual stimuli of the uncoordinated surrogate U2 (magic A + systematic subsampling V). Thus, C1 and U2 were both very helpful for this “specific token recognition” task.

4.5 Task 3: Keyframe Determination Task

Although people did reasonably well on Task 3 (the keyframe determination task), they had the third lowest overall accuracy mean on Task 3 among all the tasks. As with the keyword determination task, many distractors for the keyframe determination task were selected from keyframes for other videos in the same video collection as the test video, which made some distractors very plausible. Thus, keyframe determination is a challenging task. Furthermore, as suggested in Song and Marchionini

(2007), there may be a novelty effect here as people are not used to this type of task.

Moreover, the keyframe selection task yielded the lowest confidence levels across the six surrogates, which reinforces the difficulty and novelty interpretations for the accuracy results.

Table 4.15: Task 3. Keyframe Determination Task (N = 48, df = 5, 43)

Surrogate	**Accuracy*** (Max: 4.0)		**Confidence*** (Max: 5.0)		***T*_Time*** (sec)	
	Mean	SD	Mean	SD	Mean	SD
C1	0.761	0.088	3.448	0.767	15.260	5.001
C2	0.761	0.084	3.417	0.634	17.120	5.839
U1	0.816	0.081	3.688	0.710	14.979	4.510
U2	0.785	0.087	3.370	0.874	15.500	4.629
C3	0.707	0.092	3.375	0.809	13.771	4.948
U3	0.710	0.111	3.198	0.817	13.000	4.440

As shown in Table 4.15, strong effects of different surrogates were observed for this task in accuracy (F(5, 43) = 14.973, p < 0.01), participants' confidence ratings (F(5, 43) = 4.087, p = .004), and task completion time (F(5, 43) = 11.387, p < .001). Figures 4.7, 4.8, and 4.9 present the accuracy, confidence ratings, and task completion time with the six surrogates for task 3.

In this study, the automatically-generated uncoordinated surrogate U1 (Magic A + Storyboard V) and U2 (Magic A + Systematic Subsampling V) yielded the highest accuracy means and reasonably high confidence ratings, while manually-generated coordinated surrogates C3 (manual A + V) and U3 (manual A + manual V) yielded the lowest accuracy and lowest confidence ratings.

Next we present experimental results organized by the two research questions.

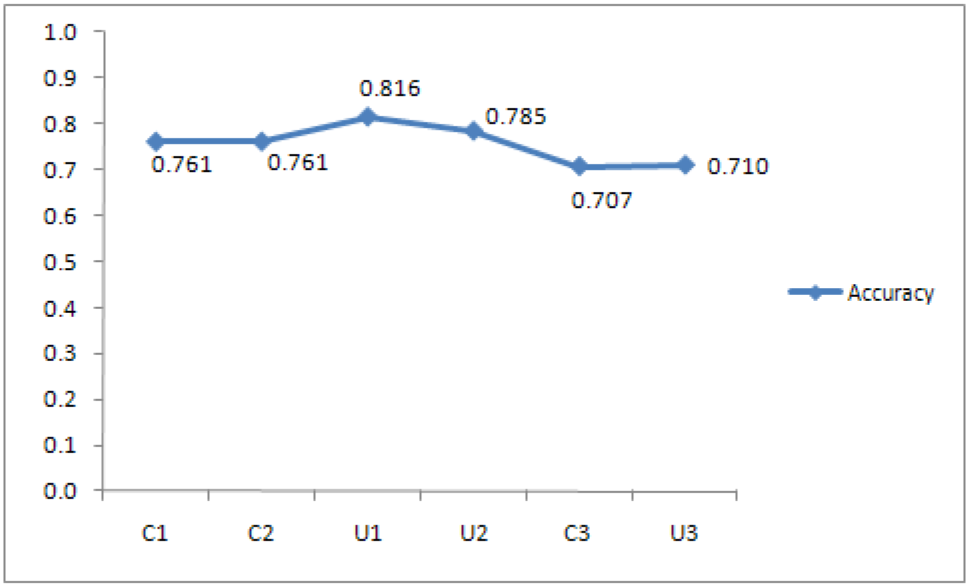


Figure 4.7: Mean Accuracy for Task 3 (Keyframe Determination Task) by Surrogate Condition. Possible data value range: 0.0 - 1.0

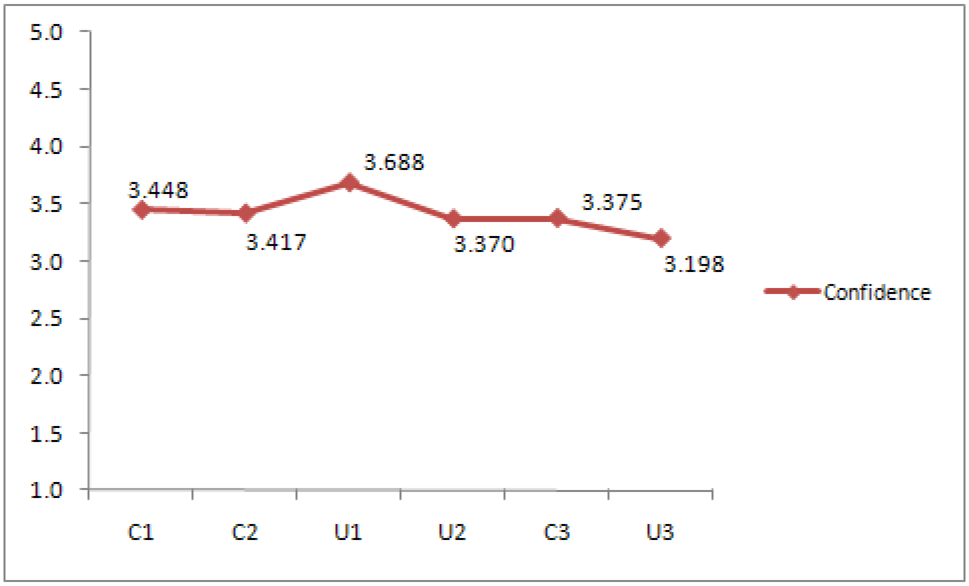


Figure 4.8: Participants' Mean Confidence for Task 3 (Keyframe Determination Task) by Surrogate Condition. Possible data value range: 1 - 5.

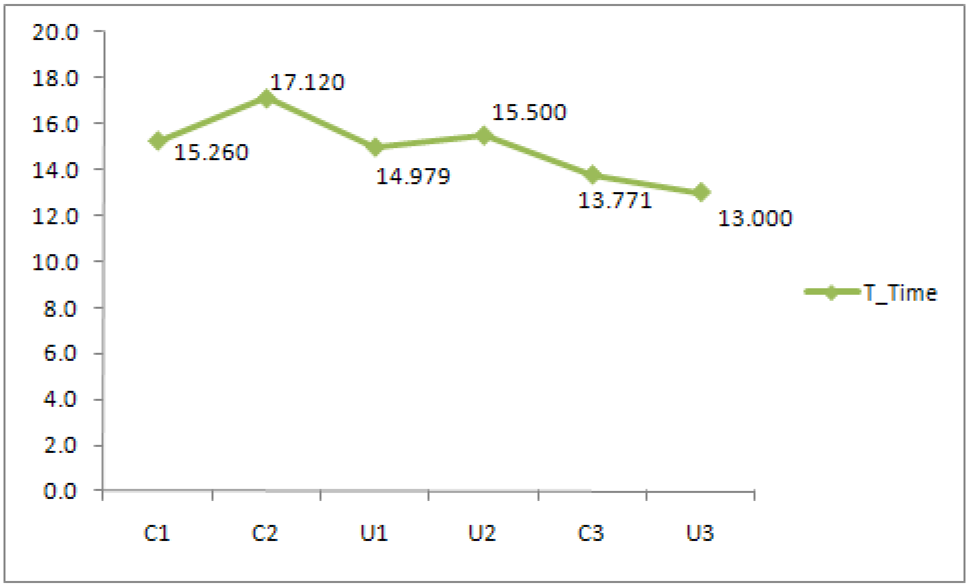


Figure 4.9: Mean Task Completion Time (in seconds) for Task 3 (Keyframe Determination Task) by Surrogate Condition.

4.5.1 Coordinated vs. Uncoordinated Surrogates

Table 4.16 summarizes the mean differences and p values for the Coordinated vs. Uncoordinated pairwise comparisons which are significant at the 0.05 level.

For the keyframe determination task, as expected, the participants performed statistically reliably better with the uncoordinated surrogates U1 (Magic A + Storyboard V) than with all three coordinated surrogate C1 (Systematic subsampling A + V), C2 (Magic A + V), and C3 (Manual A + V) in terms of accuracy and confidence ratings. Participants also performed the keyframe determination task significantly faster with the uncoordinated surrogate U1 (Magic A + Storyboard V) than with the coordinated surrogate C2 (Magic A + V).

The participants also performed statistically significantly better in keyframe determination with the uncoordinated surrogate U2 (Magic A + Systematic subsam-

Table 4.16: Summary of Pairwise Comparisons (Coordinated vs. Uncoordinated) Significant at 0.05 Level for Task 3 Accuracy, Confidence, and Task Completion Time.

	Accuracy	
U1 > **C1**	mean difference = .055	p < 0.001
U1 > **C2**	mean difference = .055	p < 0.001
U1 > **C3**	mean difference = .109	p < 0.001
U2 > **C3**	mean difference = .078	p < 0.001
U3 *<* ***C1***	*mean difference = -.051*	*p = 0.007*
U3 *<* ***C2***	*mean difference = -.051*	*p = 0.005*
	Confidence	
U1 > **C1**	mean difference = .240	p = 0.021
U1 > **C2**	mean difference = .271	p = 0.008
U1 > **C3**	mean difference = .313	p = 0.006
U3 *<* ***C1***	*mean difference = -.250*	*p = 0.028*
	T_Time	
U1 < **C2**	mean difference = -2.141	p = 0.007
U3 < **C1**	mean difference = -2.260	p = 0.006
U3 < **C2**	mean difference = -4.120	p < 0.001

Note: Pairwise comparisons which were against the hypothesis were emphasized in *Italic*.

pling V) than with the coordinated surrogates C3 in terms of accuracy, but they did not appear to be very confident about their keyframe determination performance with U2.

Namely, the automatically-generated uncoordinated surrogates were uniformly more helpful than the coordinated ones, yielding higher accuracy, higher confidence ratings, and less task completion time. However, no significant difference was found between the two manually generated surrogates C3 (manual A + V) and U3 (manual A + manual V).

Although the task completion time for the keyframe determination task was statistically significantly shorter with the uncoordinated surrogate U3 (Manual A + Manual V) than with the coordinated surrogates C1 (Systematic subsampling A + V) and C2 (Magic A + V), U3 did not lead to higher accuracy or higher confidence ratings than the coordinated surrogates C1 and C2.

4.5.2 Manually-generated vs. Automatically-generated Surrogates

Table 4.17 summarizes the mean differences and p values for the Automatic vs. Manual pairwise comparisons which are significant at the 0.05 level.

We hypothesized that people would perform the task better with the manually-generated surrogates than the automatically-generated surrogates. However, the manually-generated surrogates C3 and U3 yielded lower accuracy means and higher confidence ratings than the automatically-generated ones, which was surprising. Participants did complete the keyframe determination task more quickly with most of the manually-generated surrogates than with the automatically-generated ones, which was expected.

Table 4.17: Summary of Pairwise Comparisons (Automatically vs. Manually) Significant at 0.05 Level for Task 3 Accuracy, Confidence, and Task Completion Time.

Accuracy		
C1 > C3	*mean difference = .054*	*p = 0.002*
C2 > C3	*mean difference = .054*	*p = 0.002*
U1 > C3	*mean difference = .109*	*p < 0.001*
U2 > C3	*mean difference = .078*	*p < 0.001*
C1 > U3	*mean difference = .051*	*p = 0.007*
C2 > U3	*mean difference = .051*	*p = 0.005*
U1 > U3	*mean difference = .106*	*p < 0.001*
U2 > U3	*mean difference = .074*	*p < 0.001*
Confidence		
U1 > C3	*mean difference = .313*	*p = 0.006*
C1 > U3	*mean difference = .250*	*p = 0.028*
U1 > U3	*mean difference = .490*	*p < 0.001*
***T*_Time**		
C1 > **C3**	mean difference = 1.490	p = 0.005
C2 > **C3**	mean difference = 3.349	p < 0.001
C1 > **U3**	mean difference = 2.260	p = 0.006
C2 > **U3**	mean difference = 4.120	p < 0.001
U1 > **U3**	mean difference = 1.979	p = 0.001
U2 > **U3**	mean difference = 2.500	p < 0.001

Note: Pairwise comparisons which were against the hypothesis were emphasized in *Italic*.

The manually-generated surrogates C3 and U3 were high-quality video summaries created by human judges after viewing the video, hence were very helpful for the free-text gist written task. However, as with the keyword determination task, the number of choices and fine granularity of the keyframes made the keyframe determination task more of a measure of specific token recognition, instead of purely a measure of inference. Hence, although C3 and U3 were the most helpful surrogates for the free-gist written task, they were the least helpful surrogates for the keyframe determination task.

4.5.3 Task 3 Summary

The participants had the third lowest accuracy means and the lowest mean confidence ratings on keyframe determination task among all the tasks, which can be explained by the novelty and difficulty of this task.

In general, the participants performed the keyframe determination task with higher accuracy and more quickly with the uncoordinated surrogates than with the coordinated ones, though the participants were not always more confident in their task responses with the uncoordinated surrogates.

Similar to the keyword determination task, the expectation that people would do better on the task with the manually-generated surrogate than the automatically-generated surrogate was not borne out either.

Overall, the uncoordinated surrogates were more helpful than the coordinated surrogates for the keyframe determination task with respect to accuracy, confidence ratings, and task completion time, which was consistent with the hypothesis. Also as hypothesized, the participants performed the keyframe determination task more quickly with the manually-generated surrogates than the automatically-generated ones. However, the manually-generated surrogates yielded lower accuracy and confidence ratings than the automatically-generated ones, which was surprising.

4.6 Task 4: Visual Excerpt Determination Task

Table 4.18 presents the means and standard deviations for Task 4 (Visual excerpt determination task) for each of the six surrogates.

Table 4.18: Task 4. Visual Excerpt Determination Task (N = 48, df = 5, 43)

Surrogate	**Accuracy** (Max: 4.0)		**Confidence** (Max: 5.0)		***T*_Time*** (sec)	
	Mean	SD	Mean	SD	Mean	SD
C1	0.914	0.108	3.568	0.813	34.036	12.218
C2	0.904	0.091	3.807	0.661	33.536	11.719
U1	0.922	0.089	3.891	0.857	32.333	10.870
U2	0.914	0.086	3.625	1.024	31.807	11.228
C3	0.911	0.136	3.813	0.943	29.188	11.998
U3	0.859	0.168	3.698	0.966	26.948	11.005

We hypothesized that people would perform the task more accurately, more confidently, and more quickly with the uncoordinated surrogates than the coordinated surrogates. Likewise, we hypothesized that people would perform the task more accurately, more confidently, and more quickly with the manually-generated surrogates than the automatically-generated surrogates.

As shown in Table 4.18, participants did well (i.e., accuracy mean > 0.85) in the visual excerpt determination task with all surrogates. However, participants were not very confident in their responses – the confidence mean on this task was even lower than the mean on the free-gist written task. Since people are not often asked to determine if a visual segment is extracted from a particular video except in the case of re-finding a video, there may have been a novelty effect here as people are not used to this kind of task.

Figures 4.10, 4.11, and 4.12 present the accuracy, confidence ratings, and task completion time with the six surrogates for the visual excerpt determination task.

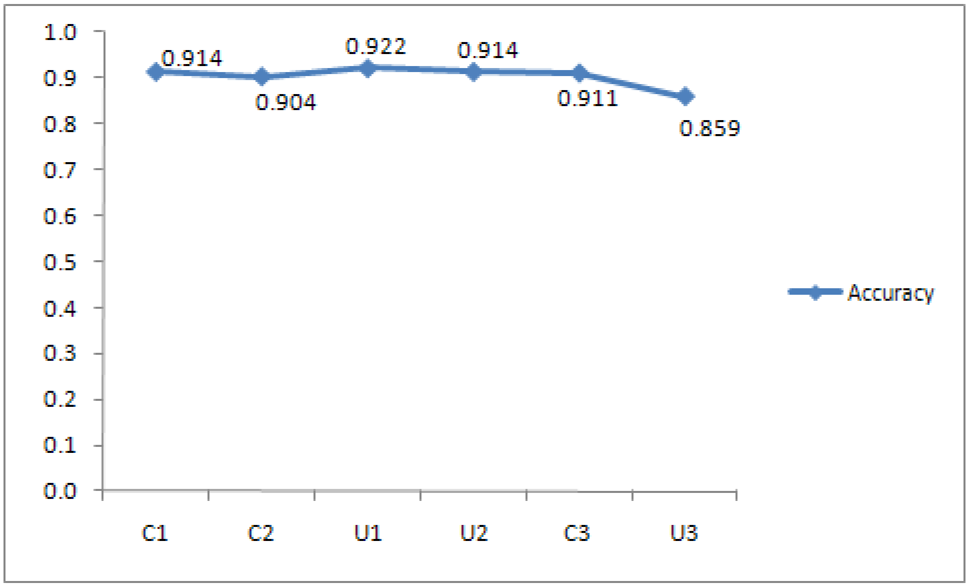


Figure 4.10: Mean Accuracy for Task 4 (Visual Excerpt Determination Task) by Surrogate Condition. Possible data value range: 0.0 - 1.0

One-way repeated-measures ANOVA showed no statistically significant effects of different surrogates for the visual excerpt determination task at the .05 level in terms of accuracy (F(5, 43) = 1.535, p = 0.199), and participants' confidence levels (F(5, 43) = 2.355, p = .056). The main effect of different surrogates on the task completion time was significant (F(5, 43) = 11.560, p < .001).

Next we present experimental results organized by the two research questions.

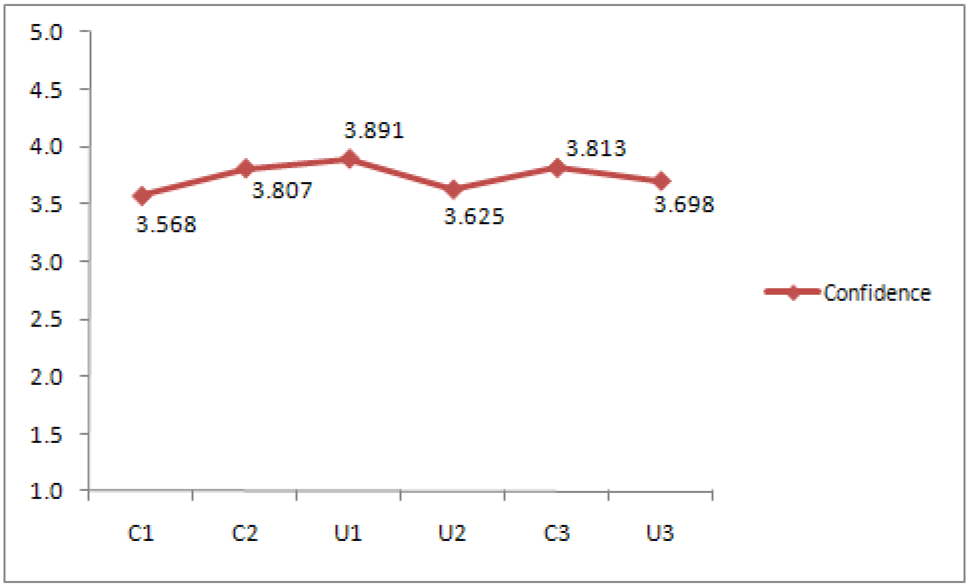


Figure 4.11: Participants' Mean Confidence for Task 4 (Visual Excerpt Determination Task) by Surrogate Condition. Possible data value range: 1 - 5.

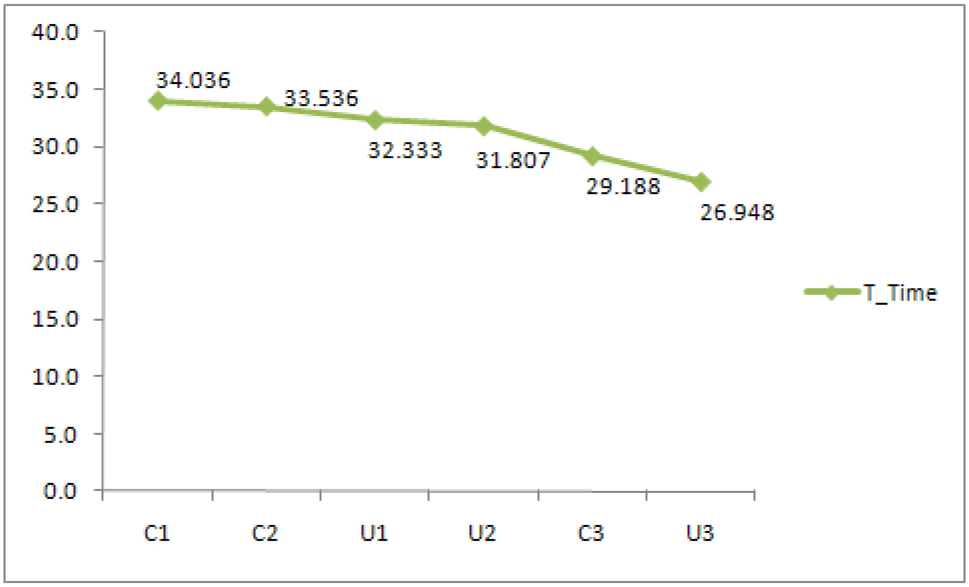


Figure 4.12: Mean Task Completion Time (in seconds) for Task 4 (Visual Excerpt Determination Task) by Surrogate Condition.

4.6.1 Coordinated vs. Uncoordinated Surrogates

Table 4.19 summarizes the mean differences and p values for the Coordinated vs. Uncoordinated pairwise comparisons on this task which are significant at the 0.05 level. As mentioned above, no significant differences were found among different surrogates on accuracy and confidence levels for the visual excerpt determination task.

Table 4.19: Summary of Pairwise Comparisons (Coordinated vs. Uncoordinated) Significant at 0.05 Level for Task 4 Completion Time

	***T*_Time**	
U3 < **C1**	mean difference = -7.089	p = 0.004
U3 < **C2**	mean difference = -6.589	p = 0.004

Though we hypothesized that people would perform the visual excerpt determination task more accurately, more confidently, and more quickly with the uncoordinated surrogates than the coordinated surrogates, no statistically significant differences were found among different surrogates on accuracy and confidence levels for the task. Results on the task completion time were consistent with our hypothesis.

Participants spent statistically reliably less time in completing the task with uncoordinated surrogate U3 (Manual A + Manual V) than with coordinated surrogates C1 (Systematic subsampling A + V) and C2 (Magic A + V). No other differences in Task 4 completion times between coordinated and uncoordinated surrogates were significant.

4.6.2 Manually-generated vs. Automatically-generated Surrogates

The mean differences and p values for the pairwise comparisons on Task 4 completion time significant at the critical value 0.05 are shown in Table 4.20. No statistically significant effects were found on the type of surrogate on Task 4 accuracy and confidence ratings.

Table 4.20: Summary of Pairwise Comparisons (Automatically vs. Manually) Significant at 0.05 Level for Task 4 Completion Time

T_Time		
C3 < **C1**	mean difference = -4.849	p < 0.001
C3 < **C2**	mean difference = -4.349	p < 0.001
U3 < **C1**	mean difference = -7.089	p = 0.004
U3 < **C2**	mean difference = -6.589	p = 0.004
U3 < **U1**	mean difference = -5.385	p < 0.001
U3 < **U2**	mean difference = -4.859	p < 0.001

As we expected, participants completed the visual excerpt determination task more quickly with the manually generated surrogates than with the automatically-generated ones. Participants spent statistically reliably less time in completing the task with manually-generated surrogate C3 (Manual A + V) than with automatically-generated surrogates C1 (systematic subsampling A + V) and C2 (Magic A + V).

Interestingly, although the participants were the least accurate and moderately confident in the visual excerpt determination task with the manually-generated surrogate U3, they spent statistically less time in completing the task with U3 than with any of the automatically-generated surrogates C1 (systematic subsampling A + V), C2 (Magic A + V), U1 (Magic A + Storyboard V), and U2 (Magic A + Systematic subsampling V).

4.6.3 Task 4 Summary

To summarize, the participants did well on Task 4 (Visual excerpt determination task) with all surrogate conditions, and no statistically reliable differences were observed in Task 4 accuracy for the six different surrogates.

Because people are not used to the task (i.e., we are not often asked to determine if a segment is extracted from a particular video), participants' confidence ratings in their responses to this task were not very high compared to other tasks in this study, despite their good performance. Similar to the accuracy result, the differences in Task 4 confidence among the different surrogates were not statistically significant.

Surrogate condition had a significant effect on the completion time for the visual excerpt determination task. In general, as we expected, the participants performed the task more quickly with one of the uncoordinated surrogates than with some of the coordinated ones, and more quickly with the manually-generated surrogates than with the automatically-generated ones.

4.7 Task 5: Audio Excerpt Determination Task

Table 4.21: Task 5. Audio Excerpt Determination Task (N = 48, df = 5, 43)

Surrogate	Accuracy		Confidence*		T_Time*	
	(Max: 4.0)		(Max: 5.0)		(sec)	
	Mean	SD	Mean	SD	Mean	SD
C1	0.908	0.106	3.776	0.901	32.708	9.786
C2	0.898	0.121	4.229	0.618	31.073	9.165
U1	0.896	0.100	4.109	0.836	31.323	8.641
U2	0.902	0.102	4.005	0.913	32.151	8.193
C3	0.919	0.114	4.219	0.967	29.781	8.861
U3	0.919	0.135	4.219	0.875	28.000	9.790

Table 4.21 presents the means and standard deviations for Task 5 (Audio excerpt

determination task) for each of the six surrogates. Strong effects of different surrogates were observed for the audio excerpt determination task in participants' confidence ratings (F(5, 43)=4.394, p=.003) and task completion time (F(5, 43)=4.483, p=.002), but not in accuracy (F(5, 43) =.467, p=0.799). Figures 4.13, 4.14, and 4.15 plot the accuracy, confidence ratings, and task completion time with the six surrogates for the audio excerpt determination task.

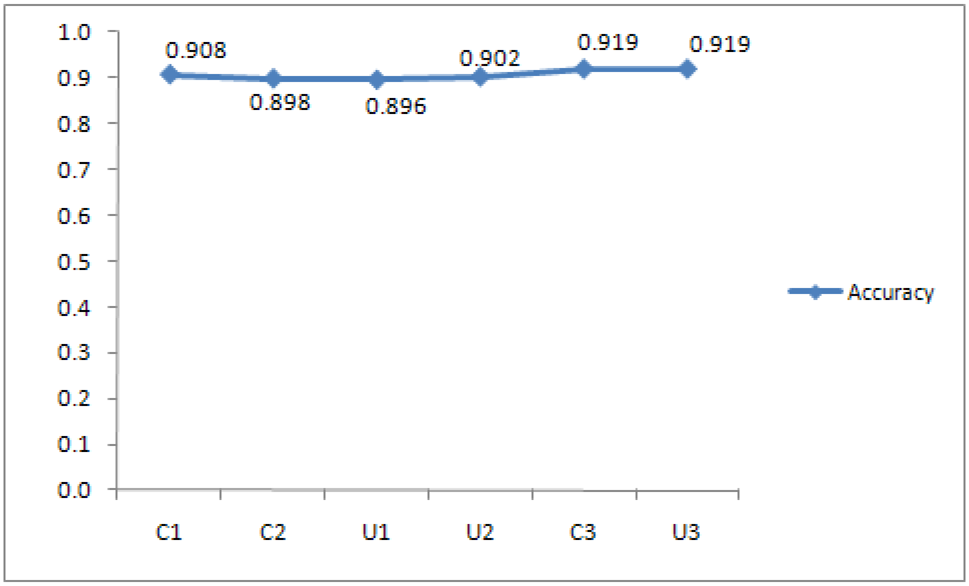


Figure 4.13: Mean Accuracy for Task 5 (Audio Excerpt Determination Task) by Surrogate Condition. Possible data value range: 0.0 - 1.0

In general, participants did very well (i.e., accuracy mean > 0.89) in the audio excerpt determination task with all surrogates, and they were quite confident in their responses – the confidence ratings for the audio excerpt determination task were the second highest among all tasks, only slight lower than Task 6 (Verbal gist determination task). As we discussed earlier, we observed a novelty effect in Task 4 (Visual excerpt determination task) because people were not used to this type of task. Although people may not be used to the task of determining if an

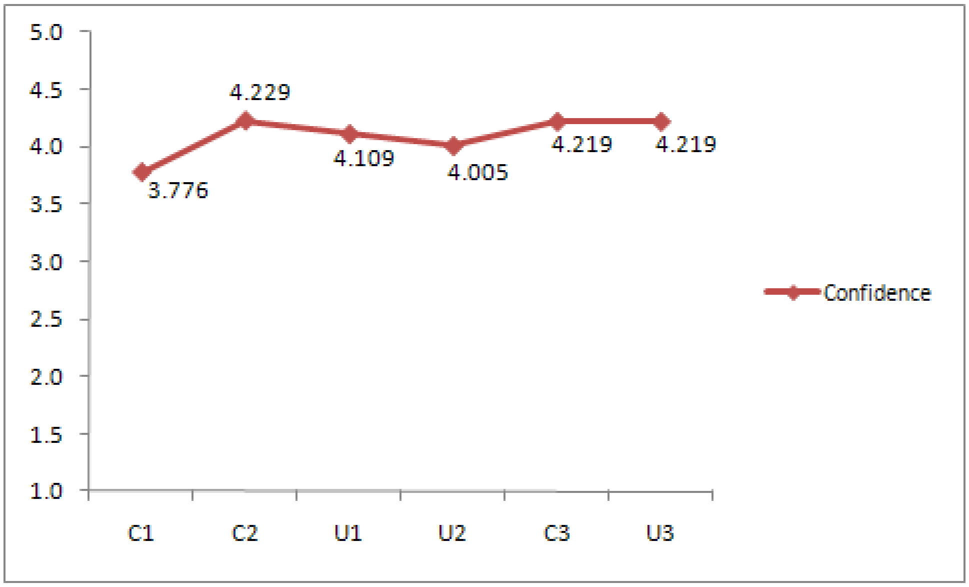


Figure 4.14: Participants' Mean Confidence for Task 5 (Audio Excerpt Determination Task) by Surrogate Condition. Possible data value range: 1 - 5.

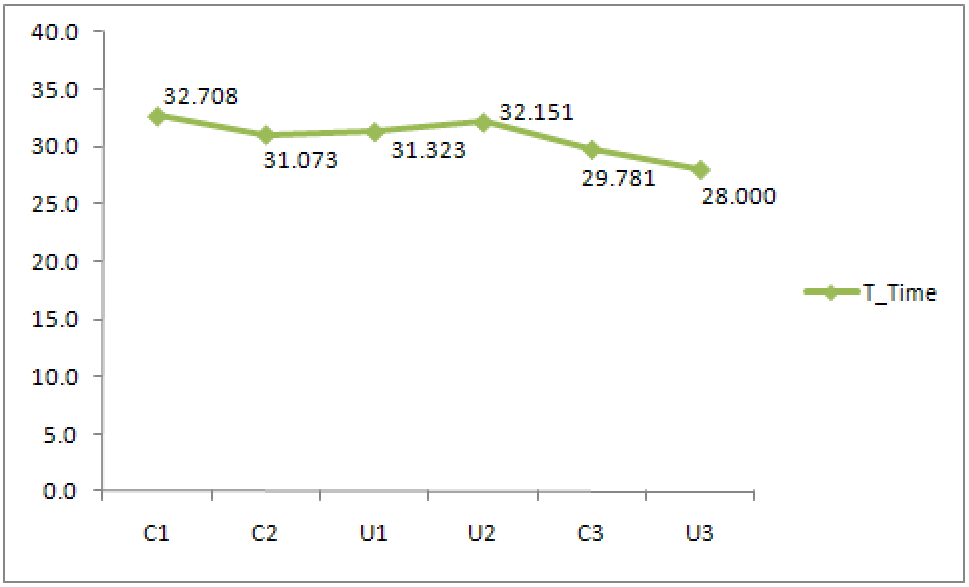


Figure 4.15: Mean Task Completion Time (in seconds) for Task 5 (Audio Excerpt Determination Task) by Surrogate Condition.

audio segment is extracted from a particular video, there did not seem to be a novelty effect here. Perhaps they had gained some similar experience with the visual excerpt determination task, or there is some inherent naturalness to aural expression of words.

Next, we will organize the results by the two research questions as follows.

4.7.1 Coordinated vs. Uncoordinated Surrogates

Table 4.22 summarizes the mean differences and p values between the coordinated and uncoordinated surrogates in terms of confidence and task completion time on the audio excerpt determination task for those that are significant at the 0.05 level.

Table 4.22: Summary of Pairwise Comparisons (Coordinated vs. Uncoordinated) Significant at 0.05 Level for Task 5 Confidence and Task Completion Time

	Confidence	
U1 > **C1**	mean difference = .333	p = 0.002
U3 > **C1**	mean difference = .443	p < 0.001
	***T*_Time**	
U3 < **C1**	mean difference = -4.708	p = 0.011

We hypothesized that people would perform the audio excerpt determination task more accurately with the uncoordinated surrogates than with the coordinated surrogates, but there were no statistically significant differences in the accuracy result.

For the audio excerpt determination task, the participants were statistically reliably more confident in their task responses with the uncoordinated surrogates U1 (Magic A + Storyboard V) and U3 (Manual A + Manual V) than with the coordinated surrogate C1 (manual A + V), which was expected.

We hypothesized that people would perform the audio excerpt determination task more quickly with the uncoordinated surrogates than with the coordinated surrogates. In this study, participants completed the audio excerpt determination task most quickly with the uncoordinated surrogate U3 and most slowly with the coordinated surrogate C1, and the task completion time difference between U3 and C1 was statistically reliable (p=0.011). None of the task completion time differences in other coordinated and uncoordinated surrogate pairs were statistically reliable at the .05 level.

4.7.2 Manually-generated vs. Automatically-generated Surrogates

Table 4.23 summarizes the mean differences and p values between the manually-generated surrogates and the automatically-generated surrogates in confidence and task completion time on the task that are significant at the 0.05 level.

Table 4.23: Summary of Pairwise Comparisons (Automatically vs. Manually) Significant at 0.05 Level for Task 5 Confidence and Task Completion Time

	Confidence	
C1 < C3	mean difference = -.443	p < 0.001
C1 < U3	mean difference = -.443	p < 0.001
U2 < U3	mean difference = -.214	p = 0.032
	***T*_Time**	
C1 > C3	mean difference = 1.490	p = 0.005
C1 > U3	mean difference = 3.349	p < 0.001
U1 > U3	mean difference = 2.260	p = 0.006
U2 > U3	mean difference = 4.120	p < 0.001

No statistically significant differences were observed in the accuracy result across different surrogates for the audio excerpt determination task. The manually-generated surrogates led to almost uniformly better performance than the automatically-generated surrogates in terms of confidence ratings and task completion time, which was consistent with the hypothesis.

As shown in Table 4.23, participants were statistically reliably more confident in their task responses with the manually-generated surrogate C3 (Manual A + V) than with the automatically-generated surrogate C1 (Systematic subsampling A + V). They were also statistically reliably more confident in their task responses with the manually-generated surrogate U3 (Manual A + Manual V) than with the automatically-generated surrogates C1 (Systematic subsampling A + V) and U2 (Magic A + Systematic subsampling V).

The task completion time with manually-generated C3 (Manual A + V) was statistically reliably less than the task completion time with automatically-generated C1 (Systematic subsampling A + V), and the task completion time with manually-generated U3 (Manual A + Manual V) was statistically reliably less than the task completion time with automatically-generated C1 (Systematic subsampling A + V), U1 (Magic A + Storyboard V), and U2 (Magic A + Systematic subsampling V). Other pairwise comparisons in task completion time between manually- and automatically- generated surrogates were not significant at the .05 level.

4.7.3 Task 5 Summary

To summarize, the participants did very well on the audio excerpt determination task with all surrogate conditions, and no statistically reliable differences were observed in Task 5 accuracy for the six difference surrogates. Participants were also very confident in their task responses. Though the audio excerpt determination task was to some extent similar to the visual excerpt determination task, no novelty effect was observed here. Maybe they had gained some experience with this kind of task when performing the visual excerpt determination task.

We hypothesized that people would perform better on the task with the uncoordinated surrogates than the coordinated surrogates, and would perform better with the manually-generated surrogates than the automatically-generated surrogates. In fact, there were no significant differences across the six surrogates on task 5 accuracy. Nevertheless, participants were generally more quick in producing their responses to the task and were more confident in their responses using the manually-generated surrogates than using the automatically-generated surrogates, which was consistent with the hypothesis. But there were no statistically reliable differences in the task completion time between coordinated and uncoordinated surrogates except in one pair: it took participants longer to completed task 5 with C1 than with U3 (mean difference = 4.708, p = .011).

4.8 Task 6: Verbal Gist Determination Task

Table 4.24: Task 6. Verbal Gist Determination Task (N = 48, df = 5, 43)

Surrogate	**Accuracy*** (Max: 4.0)		**Confidence*** (Max: 5.0)		***T*_Time*** (sec)	
	Mean	SD	Mean	SD	Mean	SD
C1	0.938	0.121	4.193	0.791	14.766	6.378
C2	0.974	0.077	4.464	0.455	14.776	6.577
U1	0.974	0.093	4.401	0.714	13.813	4.849
U2	0.953	0.111	4.203	0.865	13.380	4.079
C3	0.979	0.101	4.490	0.664	14.375	8.117
U3	0.969	0.122	4.479	0.699	11.146	4.434

Table 4.24 presents the means and standard deviations for accuracy, confidence, and task completion time for each of the six surrogates for Task 6 (Verbal gist

determination task). Strong effects of different surrogates were observed for this task in accuracy (F(5, 43) = 2.586, p = 0.039), participants' confidence ratings (F(5, 43) = 2.641, p = 0.036), and task completion time (F(5, 43) = 3.253, p = 0.014).

Figures 4.16, 4.17, and 4.18 plot the accuracy, confidence ratings, and task completion time on the verbal gist determination task across the six surrogates.

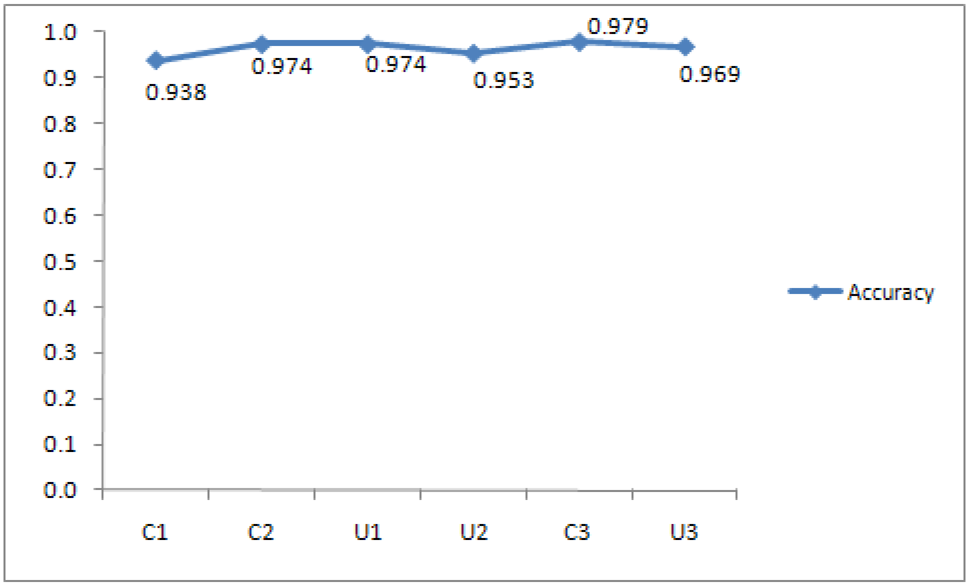


Figure 4.16: Mean Accuracy for Task 6 (Verbal Gist Determination Task) by Surrogate Condition. Possible data value range: 0.0 - 1.0

Participants did very well (i.e., accuracy means > 0.93) in the verbal gist determination task with all surrogates, and they were very confident in their responses (i.e., confidence means > 4.1). The task completion time for the task was also the shortest among all six tasks.

The verbal gist determination task was the last task in our study. Though we carefully ordered the six tasks so as to minimize potential knowledge gain from previous tasks, participants might still collect new information from the previous tasks

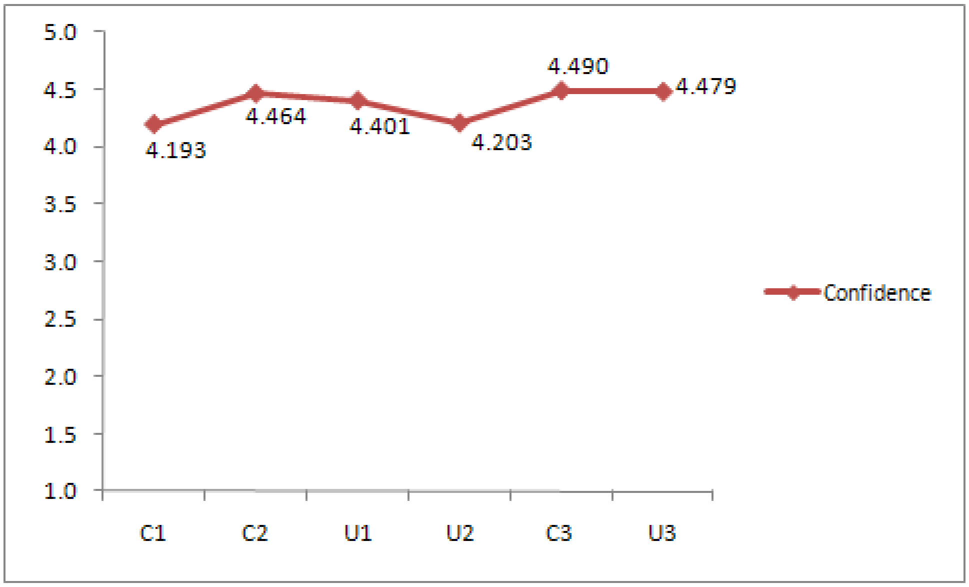


Figure 4.17: Participants' Mean Confidence for Task 6 (Verbal Gist Determination Task) by Surrogate Condition. Possible data value range: 1 - 5.

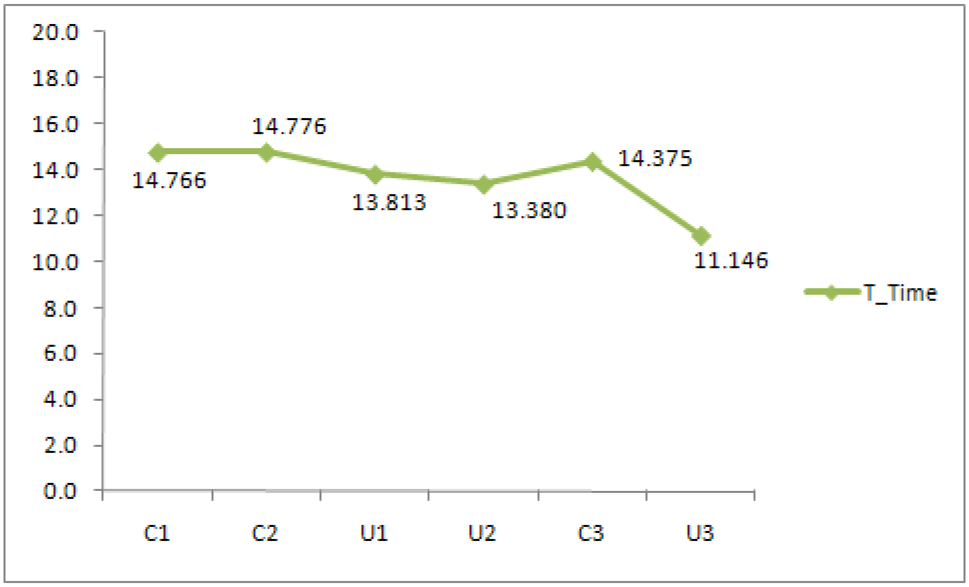


Figure 4.18: Mean Task Completion Time (in seconds) for Task 6 (Verbal Gist Determination Task) by Surrogate Condition.

which would help them perform the last task better. Also, as with the visual excerpt determination task and the audio excerpt determination task, the verbal gist determination task asked the participants to pick one correct answer from the 4 choices. It is relatively easy compared to the free-gist written task where participants were asked to compose text summaries based only on the brief surrogate they experienced, and the keyword determination task and the keyframe determination task where participants were asked to pick an unknown number of correct answers from a total of 10 choices. All in all, the verbal gist determination task was the easiest task in this study. It is not surprising that participants performed this task with the highest accuracy, the highest confidence ratings, and the least task completion time.

As with the other 5 tasks, we hypothesized that people would perform the verbal gist determination task better with the uncoordinated surrogates than with the coordinated surrogates, and better with the manually-generated surrogates than with the automatically-generated surrogates. In the following paragraphs, experimental results related to the two research questions are presented.

4.8.1 Coordinated vs. Uncoordinated Surrogates

Table 4.25 summarizes the mean differences and p values between the coordinated and uncoordinated surrogates in terms of confidence and task completion time on Task 6 (Verbal gist determination task) for those that were significant at the 0.05 level. No statistically reliable difference in the accuracy was observed on this task between the coordinated and uncoordinated surrogates.

As expected, participants' confidence ratings in their task 6 responses were statistically reliably higher with the uncoordinated surrogate U3 than with the co-

Table 4.25: Summary of Pairwise Comparisons (Coordinated vs. Uncoordinated) Significant at 0.05 Level for Task 6 Confidence and Task Completion Time

	Confidence	
U3 > C1	mean difference = .286	p = 0.032
U2 < C2	*mean difference = -.260*	*p = 0.045*
U2 < C3	*mean difference = -.286*	*p = 0.011*
	***T*_Time**	
U3 < C1	mean difference = -3.620	p = 0.006
U3 < C2	mean difference = -3.630	p = 0.007
U3 < C3	mean difference = -3.229	p = 0.039

Note: Pairwise comparisons which were against the hypothesis were emphasized in *Italic*.

ordinated surrogate C1 (Systematic subsampling A + V). We also expected that participants would have higher confidence in their task responses with the uncoordinated surrogates than with the coordinated surrogates. However, participants' confidence ratings were statistically reliably higher with the coordinated surrogates C2 (Magic A + V) and C3 (Manual A + V) than with the uncoordinated surrogate U2 (Magic A + Systematic subsampling V).

Participants completed the task statistically reliably more quickly with uncoordinated surrogate U3 (Manual A + Manual V) than with all coordinated surrogates: C1 (Systematic subsampling A + V), C2 (magic A + V), and C3 (Manual A + V).

4.8.2 Manually-generated vs. Automatically-generated Surrogates

Table 4.26 summarizes the mean differences and p values (< .05 critical value) between the manually-generated surrogates and the automatically-generated sur-

Table 4.26: Summary of Pairwise Comparisons (Automatically vs. Manually) Significant at 0.05 Level for Task 6 Accuracy, Confidence, and Task Completion Time.

	Accuracy	
C1 < C3	mean difference = -.042	p = 0.010
	Confidence	
C1 < C3	mean difference = -.297	p = 0.019
C1 < U3	mean difference = -.286	p = 0.032
U2 < C3	mean difference = -.286	p = 0.011
U2 < U3	mean difference = -.276	p = 0.005
	***T*_Time**	
C1 > U3	mean difference = 3.620	p = 0.006
C2 > U3	mean difference = 3.630	p = 0.007
U1 > U3	mean difference = 2.667	p = 0.002
U2 > U3	mean difference = 2.234	p = 0.002

rogates in terms of accuracy, confidence ratings, and task completion time on the verbal gist determination task.

Participants performed the verbal gist determination task statistically significantly more accurately with the manually-generated surrogate C3 (Manual A + V) than with the automatically-generated surrogate C1 (Systematic subsampling A + V). Other pairwise comparisons of accuracy between the manually-generated and the automatically-generated surrogates were not statistically significant.

Results in participants' confidence ratings were consistent with our hypothesis. Participants were statistically significantly more confident in their responses to this task with the manually-generated surrogates C3 (Manual A + V) and U3 (Manual A + Manual V) than with automatically-generated surrogate C1 (Systematic subsampling A + V) and U2 (Magic A + Systematic subsampling V). No statistically reliable difference in participants' confidence ratings on this task was observed between the two manually-generated surrogates C3 (Manual A + V) and U3 (Manual

A + Manual V).

Participants completed the task statistically significantly more quickly with the manually-generated surrogate U3 (Manual A + Manual V) than with all four automatically-generated surrogates: C1 (Systematic subsampling A + V), C2 (Magic A + V), U1 (Magic A + Storyboard V), and U2 (Magic A + Systematic subsampling V), while the task completion time with manually-generated surrogate C3 (Manual A + V) was not statistically reliably different from any of the automatically-generated surrogates.

4.8.3 Task 6 Summary

To summarize, the participants did extraordinarily well on the verbal gist determination task with all surrogate conditions. The accuracy mean and participants' confidence ratings were the highest on this task, and participants completed this task most quickly, among all six tasks.

Surrogate condition had significant effects on accuracy, confidence ratings, and task completion time for the verbal gist determination task. Generally, the participants performed the task with higher accuracy, higher confidence ratings, and less time with the manually-generated surrogates than with the automatically-generated ones. Although participants performed the verbal gist determination task more quickly with the uncoordinated surrogates than the coordinated ones, the expectation that people would perform the task more accurately and more confidently with the uncoordinated surrogates than with the coordinated surrogates was not borne out.

4.9 Subjective Measures

After completing the tasks for all test videos in each surrogate condition, participants were asked to rate their experience with the surrogate on the four subjective scales including **usefulness**, **usability**, **enjoyment**, and **engagement**. Participants first completed twelve 5-point Likert scale questions (Davis, 1989) on usefulness and usability, and then completed eight 7-point semantic differential scales (Ghani et al., 1991) on engagement and enjoyment.

One-way repeated-measures ANOVAs were conducted across the six surrogates on each of four subjective measures: usefulness, usability, engagement and enjoyment. Table 4.27 summarizes the means and the standard deviations for participants' reports of subjective measures for each surrogate.

Table 4.27: Participants' Subjective Ratings on the Surrogates

Surrogate	**Usefulness***		**Usability***		**Enjoyment***		**Engagement**	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
C1	2.935	0.917	3.573	0.939	3.385	1.494	3.922	1.503
C2	3.175	0.919	3.656	1.006	3.729	1.544	4.094	1.542
U1	2.946	1.090	3.573	0.962	4.063	1.453	4.099	1.488
U2	2.908	0.967	3.356	1.007	3.677	1.416	3.990	1.678
C3	3.449	1.107	3.872	1.149	4.026	1.597	4.278	1.692
U3	3.062	1.088	3.581	1.080	3.630	1.476	3.927	1.354

Note: Usefulness and Usability were rated 1-5. Enjoyment and Engagement are rated 1-7.

Strong effects of different surrogates were observed for three out of the four subjective measures: usefulness (F(5, 43) = 3.419, p = .011), usability (F(5, 43) = 3.533, p = .009), and engagement (F(5, 43) = 4.957, p = .001). The type of surrogates did not have strong effects on participants' ratings of enjoyment in using the surrogates (F(5, 43) = 1.714, p = .152). Figure 4.19 (a) and (b) plot the means of the four subjective scales.

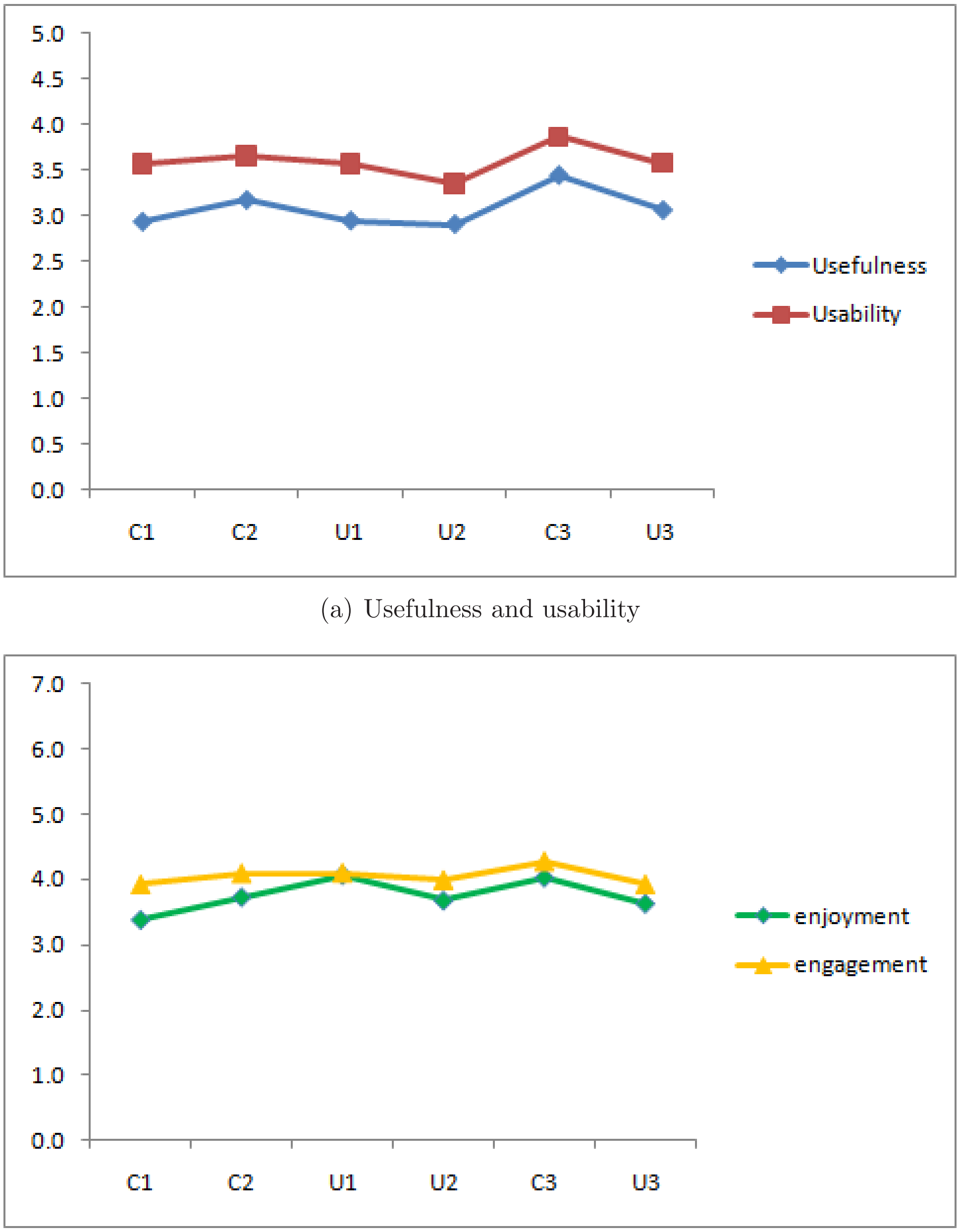
Participants' subjective ratings demonstrated their strong preferences for the manually-generated coordinated surrogate C3 (Manual A + V), which was rated the highest among all surrogates for both usefulness and usability. Automatically-generated uncoordinated surrogate U2 (Magic A + Systematic subsampling V) was rated the lowest for usefulness and usability, though the difference was only statistically significant in comparison with C3.

In terms of enjoyment and engagement, the manually-generated coordinated surrogate C3 (Manual A + V) were rated the highest for engagement and the second highest for enjoyment among the six surrogates. Interestingly, the automatically-generated uncoordinated surrogate U1 (Magic A + Storyboard V) was rated most enjoyable by the participants in this study, which was later on reinforced by participants' comments in favorable of the storyboard with audio surrogate in the post-session questionnaire. The automatically-generated coordinated surrogate C1 (Systematic subsampling A + V) was rated the lowest for both enjoyment and engagement.

4.9.1 Coordinated vs. Uncoordinated Surrogates

Table 4.29 summarizes the mean differences and p values of the pairwise comparisons between the coordinated and uncoordinated surrogates significant at the .05 level in three of the four subjective scales: `usefulness`, `usability`, and `enjoyment`. No significant differences were observed in participants' ratings of `engagement` in any pair of the coordinated and uncoordinated surrogates.

We expected that the uncoordinated surrogates would be more useful yet less usable than the coordinated ones for the purpose of video sense-making. On average,



(a) Usefulness and usability

(b) Enjoyment and engagement

Figure 4.19: Subjective Measures: (a) Usefulness and usability, (b) Enjoyment and engagement.

Table 4.28: Pairwise Comparisons (Coordinated vs. Uncoordinated) of the Subjective Measures Significant at 0.05 Level

Usefulness		
U1 < C3	*mean difference = -.503*	*p = 0.007*
U2 < C3	*mean difference = -.542*	*p = 0.002*
U3 < C3	*mean difference = -.387*	*p = 0.004*
Usability		
U1 < C3	mean difference = -.299	p = 0.006
U2 < C3	mean difference = -.515	p = 0.001
U3 < C3	mean difference = -.290	p = 0.003
Enjoyment		
U1 > C1	*mean difference = .677*	*p = 0.001*
U3 < C3	mean difference = -.396	p = 0.017

Note: Pairwise comparisons which were against the hypothesis were emphasized in *Italic*.

participants' ratings on usability were expected, but their ratings on on usefulness was surprising. In this study, the coordinated surrogates were found both more useful and more usable than the uncoordinated ones.

In fact, the coordinated surrogate C3 (Manual A + V) was rated statistically reliably more useful (which was surprising) and statistically reliably more usable (which was expected) than the uncoordinated surrogate U1 (Magic A + Storyboard V), U2 (Magic A + Systematic subsampling V), and U3 (Manual A + Manual V), which parallels the the accuracy and confidence results for two linguist gist tasks: Task 1 (the free-text gist written task) and Task 6 (the verbal gist determination task).

4.9.2 Manually-generated vs. Automatically-generated Surrogates

Table 4.29 summarizes the mean differences and p values of the pairwise comparisons significant at < .05 critical value between the manually-generated surrogates and the automatically-generated surrogates in the four subjective scales: usefulness, usability, engagement and enjoyment.

Table 4.29: Pairwise Comparisons (Automatically vs. Manually) of the Subjective Measures Significant at 0.05 Level.

	Usefulness	
C1 < **C3**	mean difference = -.515	p < 0.001
U1 < **C3**	mean difference = -.503	p = 0.007
U2 < **C3**	mean difference = -.542	p = 0.002
	Usability	
C1 < **C3**	mean difference = -.299	p = 0.003
C2 < **C3**	mean difference = -.215	p = 0.013
U1 < **C3**	mean difference = -.299	p = 0.006
U2 < **C3**	mean difference = -.515	p = 0.001
	Enjoyment	
C1 < **C3**	mean difference = -.641	p < 0.001
C2 < **C3**	mean difference = -.297	p = 0.030
U1 > ***U3***	*mean difference = -.432*	*p = 0.027*
	Engagement	
C1 < **C3**	mean difference = -.356	p = 0.016

Note: Pairwise comparisons which were against the hypothesis were emphasized in *Italic*.

The manually-generated surrogate C3 (Manual A + V) was rated the highest by the participants among all surrogates in terms of usefulness and usability. Pairwise comparisons found C3 significantly reliably more usable than all automatically-generated surrogates C1 (Systematic subsampling A + V), C2 (magic A + V), U1 (Magic A + Storyboard V), and U2 (Magic A + Systematic subsampling V). Additionally, C3 was also significantly reliably more useful than all automatically-generated surrogates but C2.

Furthermore, the manually-generated surrogate C3 (Manual A + V) was rated statistically significantly higher than the automatically-generated surrogates C1 (Systematic subsampling A + V) and C2 (magic A + V) for enjoyment, and was rated higher than the automatically-generated surrogate C1 for engagement. Also, the manually-generated surrogate U3 (Manual A + V) was rated statistically significantly higher than the automatically-generated surrogate U1 (Magic A + Storyboard V) for enjoyment.

4.9.3 Subjective Measures Summary

In sum, the manually-generated coordinated surrogate C3 (Manual A + V) was rated statistically reliably more usable than all other surrogates, and was rated statistically reliably more useful than all other surrogates but C2 (Magic A + V).

The automatically-generated uncoordinated surrogate U1 (Magic A + Storyboard V) was the most enjoyable surrogate in this study, but its high enjoyment ratings were most likely to be related to the fact that U1 was uncoordinated: it was fun to listen to the well-summarized audio extracts when looking at the storyboard visual.

4.10 Post-session Questionnaire

In spite of a few anomalies on the performance measures, the quantitative data clearly demonstrate that uncoordinated surrogates were effective in video sense-making and do not penalize efficiency – it took the participants less time to complete the gist tasks with the uncoordinated surrogates, and there was no difference in the consumption time of different surrogates. These results were strongly reinforced by participants' responses to the post-session questionnaire and suggest that people are able to integrate two distinct sets of surrogates that use different sensory channels but are not temporally coordinated, though they may not like the uncoordinated channels as much as the coordinated ones.

Which type of surrogates did you find easier to learn to use?

The participants were asked which surrogates they found easier to learn to use. Note that participants were only given 4 options to choose from – "Synchronous Audio and Visual", "Asynchronous Audio and Visual", "Audio and Storyboard", and "No difference" – because the surrogates were not labeled with specific names in the study such that the participants may not easily differ C2 from C3, and U2 from U3.

Thirty-four of the 48 participants voted for "Synchronous Audio and Visual", 6 voted for "Audio and Storyboard", 1 voted for "Asynchronous Audio and Visual", and 7 voted for 'No difference'".

Which type of surrogates did you find easier to use?

When asked what type of surrogates they found easier to use, thirty-seven of the 48 participants voted for "Synchronous Audio and Visual", 9 voted for "Audio and Storyboard", none voted for "Asynchronous Audio and Visual", and 2 voted for 'No difference'".

Which type of surrogates did you like the best overall?

The participants were asked which surrogates they liked most. The votes were actually quite similar to the votes to the two questions "Which type of surrogates did you find easier to learn to use?" and "Which type of surrogates did you find easier to use?".

Thirty-five of the 48 participants selected “Synchronous Audio and Visual” , with 9 selecting “Audio and Storyboard”, 2 selecting “Asynchronous Audio and Visual”, and 2 selecting ‘No difference‘”.

Participants were then asked to comment on some open-ended questions in the post-session questionnaires. The following summarizes the participants’ comments on different surrogates.

What did you like about each of the surrogate conditions?

The participants were asked to comment what they liked about each of the surrogates. Some participants liked the coordinated surrogates (i.e., C1, C2, and C3) for their familiarity and user friendliness. They commented that the coordinated surrogates were “very user friendly”, were what they were “used to”, and “gave a good synthesis of visual and audio input”. Participants liked the coordinated surrogates because they “were easier to use”, and “allowed you to focus”.

Many participants said that the coordinated surrogates helped them “understand what was going on in the video” and intrigued them to watch the full video. They said that the coordinated surrogates “mimicked the trailer style of popular films”, “were more informative and made me more curious about what was in the rest”.

More specifically, participants liked the coordinated surrogates when the snippets had a natural flow, and did not like when the snippets “jumped around” without transition. One participant commented: “The videos with the smoothest transitions were the best videos”, whereas surrogates created by systematically subsampling the video “were disjointed” and “were difficult to follow”.

In spite of the uncoordinated audio and visual presentation, some participants actually liked storyboards with audio (i.e., U1) because they were easier to follow than the surrogates with uncoordinated audio and moving images (i.e., U2 and

U3). One participant commented that “Still pictures accompanying audio feels more natural than moving video [where] you expect to have sound, but having sound that doesn’t match. It was easier to concentrate with the storyboard than with the unsynchronized.” Another participant noted, “[it] was easy to peruse the images while listening to the audio and make my own connections.” One said, “The storyboard condition was very simple, and allowed the small bit of information provided to be more wholly taken in.” These positive user comments on U1 help explain the results on the subjective ratings. As we discussed in Section 4.9.3, participants had statistically significantly higher subjective ratings of enjoyment on U1 than on 3 of the other 5 surrogates: C1 (Systematic Subsampling A + V), U2 (Magic A + Systematic Subsampling V), and U3 (Manual A + Manual V).

Participants’ positive comments on the uncoordinated surrogates with moving audio and moving visual (i.e., U2 and U3) varied. Several participants felt that the uncoordinated surrogate provided a greater amount of information about the video in a short time than the coordinated surrogates did. They said the uncoordinated surrogates “provided the most complete data”, “gave you more information”, and “provided a great breadth of material from the original film.”

Some participants thought the uncoordinated surrogates could become more usable as people got used to them. One participant noted, “it was difficult to process, but more manageable once I got the hang of it.” One participant said he actually liked “the surrealism” of the uncoordinated.

What did you dislike about each of the surrogate conditions?

Participants were also asked to comment on what they disliked about each of the surrogates. Some commented that the coordinated surrogate C1 (Systematic subsampling A + V) was “choppy” and “annoying”. One participant said C1 was “too choppy to find the common flow. It was hard to follow the brief snippets jumping around the topic". Another said that C1 "did not adequately give me enough context to know what the video would be about." These comments paralleled participants' relatively low subjective ratings of enjoyment on C1.

A number of participants noted that "the storyboard condition did not provide enough information". One commented that "the visuals were mostly useless". Some participants made comments on the quality and size of the image. One said, "Sometimes it was hard to figure out what the images were because they were so small." One said, "the audio and storyboard didn't keep my attention because I wasn't so worried about looking at the screen while listening to the audio." Some participants liked the storyboard with audio in general, but commented on the visual presentation of the storyboard. One said "it feels very odd to be listening but not seeing motion." Another said, " I didn't like how the storyboard didn't have a lot of pictures (it had more audio information than images - if it was an equal mix I would really like it."

Some participants pointed out the problems with the multimodal surrogates with uncoordinated audio and visual channels. Most participants did not like the uncoordinated surrogates with moving audio and moving visual, such as U2 (Magic A + Systematic subsampling V) and U3 (Manual A + Manual V). The motion of the uncoordinated audio and video made the surrogates "disturbing","distracting", and "hard to process". One noted, "The more out of sync the audio and video were, the more it was difficult to obtain information about the video." Another stated, "The unsynchronized [surrogates] were hard to understand and frustrated me." Some participants found the uncoordinated audio and visual "confused my brain", and they felt "disoriented while trying to listen to mismatched audio". One participant joked, "Don't let the military get hold of this one because they will start using it as a new interrogation method in place of waterboarding."

4.11 Summary

One of the goals of creating surrogates for the video is to allow users to make sense of the video or make decisions about their relevance quickly based on the surrogates (Song and Marchionini, 2007). High compaction rates (i.e., the ratio of time to view the full video to the time to view a surrogate) are generally desirable as long as people are able to make sense of the full information object. The compaction rates (i.e., inversely proportional to the surrogate consumption time) for the six surrogates investigated were **not** significantly different from each other, ranging from **29.96** (i.e., 28.5 minutes × 60 ÷ 57.073 seconds) for C3 (Manual A + Manual V) to **30.7** (i.e., 28.5 minutes × 60 ÷ 51.781 seconds) for U1 (Magic A + Storyboard V). Thus, all surrogates had very good compaction rates.

In addition, the different surrogates were evaluated based on participants' task completion accuracy, participants' confidence ratings in their task responses, and the task completion time, for each of the six tasks: Task 1 (Free-text gist written task), Task 2 (Keyword determination task), Task 3 (Keyframe determination task), Task 4 (Visual excerpt determination task), Task 5 (Audio excerpt determination task), and Task 6 (Verbal gist determination task).

Results demonstrated that the type of the surrogates had statistically significant effects on accuracy for four out of the six tasks: Task 1 (Free-text gist written task), Task 2 (Keyword determination task), Task 3 (Keyframe determination task), and Task 6 (Verbal gist determination task).

Statistically significant effects of surrogates were observed on participants' confidence ratings for all tasks except Task 2 (Keyword determination task) and Task 4 (Visual excerpt determination task), and statistically significant effects of surrogates were observed on task completion time for all six tasks.

For the majority of the tasks, the uncoordinated surrogates were more helpful than the coordinated surrogates in terms of overall accuracy, participants' confidence ratings, and task completion time, but the manually-generated surrogates were only more helpful than the automatically-generated surrogates in the task completion time.

Among the four automatically-generated surrogates, the uncoordinated surrogates U1 (Magic A + Storyboard V) and U2 (Magic A + Systematic subsampling V) were on average more helpful than the coordinated surrogates C1 (Systematic subsampling A + V) and C2 (Magic A + V) on the six tasks.

Participants' subjective ratings were more favorable for the coordinated surrogate C2 (Magic A + V) and the uncoordinated surrogate U1 (Magic A + Storyboard V) in terms of usefulness, usability, enjoyment, and engagement. As commented by the participants' in the exit questionnaire, the coordinated surrogate C1 (Systematic subsampling A + V) which "chop[s] off mid word or sentence" was found "annoying" and "difficult to follow". On the contrary, the coordinated surrogate C2 (Magic A + V) which had "longer sections, with more consecutive speech" was what people were used to and "allowed the greatest level of focus and understanding of the material presented". They were "easy to use" and people "liked the familiarity of the synchronized surrogates".

Consistent with the result of our previous study (Song and Marchionini, 2007), many participants liked the uncoordinated surrogate U1 (Magic A + Storyboard V) because it was "easy to peruse the images while listening to the audio " and "associate the correct image to the audio". Most participants did not like the other uncoordinated surrogate U2 (Magic A + Systematic subsampling V). Though the surrogate "showed a lot of different aspects and ideas at once so you got a lot of information in a short time", the motion of the uncoordinated audio and video made it “distracting” and “really difficult to comprehend”.

Among the two manually-generated surrogates, the coordinated surrogate C3 (Manual A + V) was uniformly more helpful than the uncoordinated surrogates U3 (Manual A + V) in all aspects. Participants had higher accuracy and higher confidence ratings with C3 in almost all tasks. They also rated C3 as more useful, more usable, more enjoyable, and more engaging than U3.

Tables 4.30 through 4.33 summarize the mean differences of all pairwise comparisons among the six surrogates, as well as the number and the percentage of surrogate pairs with statistically significant mean differences, with respect to task completion accuracy, confidence ratings, task completion time, and subjective measures, respectively. The asterisks in the tables denote mean differences which are significant at the .05 level.

For task completion accuracy, 14 out of the 30 pairwise comparisons were found statistically significant at .05 level for Task 1. For Task 2 and Task 3, there were 16 and 22 pairs respectively for which the accuracy mean differences were statistically significant. For Task 4, Task 5, and Task 6, pairwise comparisons in accuracy means were found statistically significant only for 6, 0, and 4 pairs respectively.

For confidence ratings, 10 to 12 pairwise comparisons out of the 30 were statistically significant at .05 level for Task 1, Task 3, Task 5, and Task 6. No statistically significant difference was found in any surrogate pairs for Task 2, and only 6 pairs were found to have statistically significant differences for Task 4.

For task completion time, the number of of statistically significant pairwise comparisons ranged from 8 (26.7%) to 16 (53.3%) for the six tasks, with Task 3 having the most significant pairs and Task 5 having the fewest significant pairs.

Table 4.30: Pairwise Comparisons of Task Completion Accuracy among Surrogates

Surrogate		**Mean Difference** (I - II)					
(I)	(II)	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**	**Task 6**
C1	C2	-.188(*)	0.021	-0.001	0.010	0.009	-.036(*)
	C3	-.576(*)	.082(*)	.054(*)	0.003	-0.012	-.042(*)
	U1	-0.151	0.018	-.055(*)	-0.008	0.012	-0.036
	U2	-.188(*)	0.013	-0.024	0.000	0.005	-0.016
	U3	-.346(*)	.083(*)	.051(*)	.055(*)	-0.012	-0.031
C2	C1	.188(*)	-0.021	0.001	-0.010	-0.009	.036(*)
	C3	-.388(*)	.061(*)	.054(*)	-0.008	-0.021	-0.005
	U1	0.036	-0.003	-.055(*)	-0.018	0.003	0.000
	U2	0.000	-0.008	-0.023	-0.010	-0.004	0.021
	U3	-0.159	.062(*)	.051(*)	0.044	-0.021	0.005
C3	C1	.576(*)	-.082(*)	-.054(*)	-0.003	0.012	.042(*)
	C2	.388(*)	-.061(*)	-.054(*)	0.008	0.021	0.005
	U1	.424(*)	-.064(*)	-.109(*)	-0.010	0.023	0.005
	U2	.388(*)	-.069(*)	-.078(*)	-0.003	0.017	0.026
	U3	0.229	0.001	-0.003	0.052	0.000	0.010
U1	C1	0.151	-0.018	.055(*)	0.008	-0.012	0.036
	C2	-0.036	0.003	.055(*)	0.018	-0.003	0.000
	C3	-.424(*)	.064(*)	.109(*)	0.010	-0.023	-0.005
	U2	-0.036	-0.005	.031(*)	0.008	-0.007	0.021
	U3	-0.195	.065(*)	.106(*)	.063(*)	-0.023	0.005
U2	C1	.188(*)	-0.013	0.024	0.000	-0.005	0.016
	C2	0.000	0.008	0.023	0.010	0.004	-0.021
	C3	-.388(*)	.069(*)	.078(*)	0.003	-0.017	-0.026
	U1	0.036	0.005	-.031(*)	-0.008	0.007	-0.021
	U3	-0.159	.070(*)	.074(*)	.055(*)	-0.017	-0.016
U3	C1	.346(*)	-.083(*)	-.051(*)	-.055(*)	0.012	0.031
	C2	0.159	-.062(*)	-.051(*)	-0.044	0.021	-0.005
	C3	-0.229	-0.001	0.003	-0.052	0.000	-0.010
	U1	0.195	-.065(*)	-.106(*)	-.063(*)	0.023	-0.005
	U2	0.159	-.070(*)	-.074(*)	-.055(*)	0.017	0.016
Sig. pair num.		14	16	22	6	0	4
	(%)	(46.7%)	(53.3%)	(73.3%)	(20.0%)	(0.0%)	(13.3%)

* - The mean difference is significant at the .05 level.

Table 4.31: Pairwise Comparisons of Participants' Confidence Ratings in their Task Responses among Surrogates

Surrogate		**Mean Difference** (I - II)					
(I)	(II)	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**	**Task 6**
C1	C2	-.391(*)	-0.057	0.031	-.240(*)	-.453(*)	-.271(*)
	C3	-.474(*)	0.000	0.073	-0.245	-.443(*)	-.297(*)
	U1	-.302(*)	-0.057	-.240(*)	-.323(*)	-.333(*)	-0.208
	U2	-0.125	0.026	0.078	-0.057	-0.229	-0.010
	U3	-0.193	0.042	.250(*)	-0.130	-.443(*)	-.286(*)
C2	C1	.391(*)	0.057	-0.031	.240(*)	.453(*)	.271(*)
	C3	-0.083	0.057	0.042	-0.005	0.010	-0.026
	U1	0.089	0.000	-.271(*)	-0.083	0.120	0.063
	U2	0.266	0.083	0.047	0.182	0.224	.260(*)
	U3	0.198	0.099	0.219	0.109	0.010	-0.016
C3	C1	.474(*)	0.000	-0.073	0.245	.443(*)	.297(*)
	C2	0.083	-0.057	-0.042	0.005	-0.010	0.026
	U1	0.172	-0.057	-.313(*)	-0.078	0.109	0.089
	U2	.349(*)	0.026	0.005	0.188	0.214	.286(*)
	U3	.281(*)	0.042	0.177	0.115	0.000	0.010
U1	C1	.302(*)	0.057	.240(*)	.323(*)	.333(*)	0.208
	C2	-0.089	0.000	.271(*)	0.083	-0.120	-0.063
	C3	-0.172	0.057	.313(*)	0.078	-0.109	-0.089
	U2	0.177	0.083	.318(*)	.266(*)	0.104	0.198
	U3	0.109	0.099	.490(*)	0.193	-0.109	-0.078
U2	C1	0.125	-0.026	-0.078	0.057	0.229	0.010
	C2	-0.266	-0.083	-0.047	-0.182	-0.224	-.260(*)
	C3	-.349(*)	-0.026	-0.005	-0.188	-0.214	-.286(*)
	U1	-0.177	-0.083	-.318(*)	-.266(*)	-0.104	-0.198
	U3	-0.068	0.016	0.172	-0.073	-.214(*)	-.276(*)
U3	C1	0.193	-0.042	-.250(*)	0.130	.443(*)	.286(*)
	C2	-0.198	-0.099	-0.219	-0.109	-0.010	0.016
	C3	-.281(*)	-0.042	-0.177	-0.115	0.000	-0.010
	U1	-0.109	-0.099	-.490(*)	-0.193	0.109	0.078
	U2	0.068	-0.016	-0.172	0.073	.214(*)	.276(*)
Sig. pair num.		10	0	12	6	10	12
	(%)	(33.3%)	(0.0%)	(40.0%)	(20.0%)	(33.3%)	(40.0%)

* - The mean difference is significant at the .05 level.

Table 4.32: Pairwise Comparisons of Task Completion Time among Surrogates

Surrogate		**Mean Difference** (I - II)					
(I)	(II)	**Task 1**	**Task 2**	**Task 3**	**Task 4**	**Task 5**	**Task 6**
C1	C2	-2.052	-0.318	-1.859(*)	0.500	1.635	-0.010
	C3	4.885	0.906	1.490(*)	4.849(*)	2.927(*)	0.391
	U1	0.833	0.198	0.281	1.703	1.385	0.953
	U2	2.375	0.547	-0.240	2.229	0.557	1.385
	U3	8.135(*)	1.948(*)	2.260(*)	7.089(*)	4.708(*)	3.620(*)
C2	C1	2.052	0.318	1.859(*)	-0.500	-1.635	0.010
	C3	6.938(*)	1.224(*)	3.349(*)	4.349(*)	1.292	0.401
	U1	2.885	0.516	2.141(*)	1.203	-0.250	0.964
	U2	4.427	0.865	1.620	1.729	-1.078	1.396
	U3	10.188(*)	2.266(*)	4.120(*)	6.589(*)	3.073	3.630(*)
C3	C1	-4.885	-0.906	-1.490(*)	-4.849(*)	-2.927(*)	-0.391
	C2	-6.938(*)	-1.224(*)	-3.349(*)	-4.349(*)	-1.292	-0.401
	U1	-4.052	-0.708	-1.208	-3.146	-1.542	0.563
	U2	-2.510	-0.359	-1.729	-2.620	-2.370	0.995
	U3	3.250	1.042	0.771	2.240	1.781	3.229(*)
U1	C1	-0.833	-0.198	-0.281	-1.703	-1.385	-0.953
	C2	-2.885	-0.516	-2.141(*)	-1.203	0.250	-0.964
	C3	4.052	0.708	1.208	3.146	1.542	-0.563
	U2	1.542	0.349	-0.521	0.526	-0.828	0.432
	U3	7.302(*)	1.750(*)	1.979(*)	5.385(*)	3.323(*)	2.667(*)
U2	C1	-2.375	-0.547	0.240	-2.229	-0.557	-1.385
	C2	-4.427	-0.865	-1.620	-1.729	1.078	-1.396
	C3	2.510	0.359	1.729	2.620	2.370	-0.995
	U1	-1.542	-0.349	0.521	-0.526	0.828	-0.432
	U3	5.760(*)	1.401(*)	2.500(*)	4.859(*)	4.151(*)	2.234(*)
U3	C1	-8.135(*)	-1.948(*)	-2.260(*)	-7.089(*)	-4.708(*)	-3.620(*)
	C2	-10.188(*)	-2.266(*)	-4.120(*)	-6.589(*)	-3.073	-3.630(*)
	C3	-3.250	-1.042	-0.771	-2.240	-1.781	-3.229(*)
	U1	-7.302(*)	-1.750(*)	-1.979(*)	-5.385(*)	-3.323(*)	-2.667(*)
	U2	-5.760(*)	-1.401(*)	-2.500(*)	-4.859(*)	-4.151(*)	-2.234(*)
Sig. pair num.		10	10	16	12	8	10
	(%)	(33.3%)	(33.3%)	(53.3%)	(40.0%)	(26.7%)	(33.3%)

* - The mean difference is significant at the .05 level.

Table 4.33: Pairwise Comparisons of Subjective Measures among Surrogates

Surrogate		**Mean Difference** (I - II)			
(I)	(II)	**Usefulness**	**Usability**	**Enjoyment**	**Engagement**
C1	C2	-0.240	-0.083	-.344(*)	-0.172
	C3	-.515(*)	-.299(*)	-.641(*)	-.356(*)
	U1	-0.012	0.000	-.677(*)	-0.177
	U2	0.027	0.217	-0.292	-0.068
	U3	-0.128	-0.008	-0.245	-0.005
C2	C1	0.240	0.083	.344(*)	0.172
	C3	-0.275	-.215(*)	-.297(*)	-0.184
	U1	0.228	0.083	-0.333	-0.005
	U2	0.267	0.300	0.052	0.104
	U3	0.112	0.075	0.099	0.167
C3	C1	.515(*)	.299(*)	.641(*)	.356(*)
	C2	0.275	.215(*)	.297(*)	0.184
	U1	.503(*)	.299(*)	-0.036	0.179
	U2	.542(*)	.515(*)	0.349	0.288
	U3	.387(*)	.290(*)	.396(*)	0.351
U1	C1	0.012	0.000	.677(*)	0.177
	C2	-0.228	-0.083	0.333	0.005
	C3	-.503(*)	-.299(*)	0.036	-0.179
	U2	0.039	0.217	.385(*)	0.109
	U3	-0.116	-0.008	.432(*)	0.172
U2	C1	-0.027	-0.217	0.292	0.068
	C2	-0.267	-0.300	-0.052	-0.104
	C3	-.542(*)	-.515(*)	-0.349	-0.288
	U1	-0.039	-0.217	-.385(*)	-0.109
	U3	-0.155	-0.225	0.047	0.063
U3	C1	0.128	0.008	0.245	0.005
	C2	-0.112	-0.075	-0.099	-0.167
	C3	-.387(*)	-.290(*)	-.396(*)	-0.351
	U1	0.116	0.008	-.432(*)	-0.172
	U2	0.155	0.225	-0.047	-0.063
Sig. pair num.		8	10	14	2
	(%)	(26.7%)	(33.3%)	(46.7%)	(6.7%)

* - The mean difference is significant at the .05 level.

For subjective measures, 8 out of the 30 surrogate pairs were found to have statistically significantly differences in usefulness ratings, 10 pairs were found to have statistically significantly differences in usability ratings, 14 pairs were found to have statistically significantly differences in enjoyment ratings, and only 2 pairs were found to have statistically significantly differences in engagement ratings.

Chapter 5

DISCUSSION

5.1 Surrogate Interfaces

In general, usability guidelines suggest that people want to have control over their user interfaces. For all surrogates examined in this study, the participants were given control to replay or stop the surrogates. Though the number of replays and the number of stops were not large (i.e., 279 times and 49 times out of 960 plays, respectively), participants did choose to replay and stop the surrogates while consuming the surrogates. The control over the surrogate interface could make the participants more accurate and more confident in doing the gisting tasks; thus it is strongly suggested that real-world video retrieval systems should grant users control over the interfaces.

In our previous studies (Song and Marchionini, 2007; Marchionini et al., 2009), the text-to-speech synthesizer was used to create the audio surrogates, which sounded too mechanical to be easily understood. In this study, the audio came directly from the video clips, thus the audio quality was not an issue in general. However, for the coordinated surrogate C1 (Systematic Subsampling A + V), the visual and audio channels were extracted as 5 seconds out of every 120 second interval, which led to complaints by many participants that it was too "choppy" and "annoying" to find the common flow because the sound often chopped off mid words or sentences.

5.2 Task Performance Correlations

Performance (accuracy) scores on some of the video gisting tasks were correlated. Table 5.1 summarizes the correlations between task accuracy scores which were significant at the .05 level.

As stated earlier, task 2 (the keyword determination task) and task 3 (the keyframe determination task) were more measures of specific token recognition than measures of gist inference, while the other four tasks were gist inference tasks.

Performance on the free-text gist written task had moderate but statistically significant correlations with performance on the other 3 gist inference tasks: visual excerpt determination task ($r = .313$, $p < .001$), audio excerpt determination task ($r = .349$, $p < .001$), and verbal gist determination task ($r = .294$, $p < .001$). A reasonable explanation is that a person's ability to recognize the visual clip or audio clip that "belongs" to a video and a person's ability to select a correct description of a video's gist are related to the person's understanding of the entire video.

No statistically significant correlations were found between performance scores on the free-text gist written task and performance scores on either the keyword determination task or the keyframe determination task. One possible explanation is that people's ability to recognize small pieces of information (i.e., keywords and keyframes) in a video is not closely related to their general understanding of the entire video.

Performance on the keyword determination task was strongly correlated to per-

Table 5.1: Statistically Significant Correlations on Task Accuracy (p < .05)

Surrogate (I)	**Surrogate (II)**	**Correlation**	**Sig.**
Task 1	Task 4	.313	.000
	Task 5	.349	.000
	Task 6	.294	.000
Task 2	Task 3	.529	.000
	Task 4	.145	.014
	Task 6	.216	.000
Task 3	Task 2	.529	.000
	Task 4	.124	.035
	Task 6	.221	.000
Task 4	Task 1	.313	.000
	Task 2	.145	.014
	Task 3	.124	.035
	Task 5	.537	.000
	Task 6	.307	.000
Task 5	Task 1	.349	.000
	Task 4	.537	.000
	Task 6	.256	.000
Task 6	Task 1	.294	.000
	Task 2	.216	.000
	Task 3	.221	.000
	Task 4	.307	.000
	Task 5	.256	.000

Task 1 - Free-text gist written task.
Task 2 - Keyword determination task.
Task 3 - Keyframe determination task.
Task 4 - Visual excerpt determination task.
Task 5 - Audio excerpt determination task.
Task 6 - Verbal gist determination task.

formance on the keyframe determination task (r = .529), and the correlation was statistically significant (p < .001). The strong correlation between the two was not surprising: the two tasks resembled each other – they had similar formats, and they both had plausible distractors selected from other videos in the same video collection. Furthermore, performance on the keyword determination task was also weakly but statistically significantly correlated to performance on the visual excerpt determination task (r = .145, p = .014) and performance on the verbal gist determination task (r = .216, p < .001).

Likewise, besides its strong correlation with the keyword determination task, performance on the keyframe determination task was weakly correlated to the performance on the visual excerpt determination task (r = .124, p = .035) and performance on the verbal gist determination task (r = .221, p < .001). Although the correlations were weak, they were statistically significant. A possible explanation is that the recognition and inference required to select the keywords or keyframes that "belong" to a video do not rely on the inference required for an accurate construction of the video's topical gist (i.e., expressing the video's gist on their own, as in the free-text gist written task), whereas people's ability to recognize the correct keywords or keyframes for a video is (weakly) related to their ability to recognize the correct visual clip or verbal gist statement for the video (or vice versa). The lack of relationship with the audio excerpt determination task suggests that selecting an audio clip that "belongs" to a video is not relying on the same gisting processes as the keyword or keyframe determination task.

Scores on the visual excerpt determination task were highly correlated to scores on the audio excerpt determination task (r = .537, p < .001), and moderately correlated to the scores on the the verbal gist determination task (r = .307, p < .001). Scores on the audio excerpt determination task were weakly correlated to scores on the verbal gist determination task (r = .256, p < .001). It is reasonable that a person's ability to recognize the correct audio or visual clip that "belongs" to a video is related to their ability to recognize the correct verbal gist statement for the video (or vice versa).

In summary, participants' performance on the four gist inference tasks was statistically significantly correlated with each other. Participants' performance scores on the two token recognition tasks were strongly correlated with each other, and were also statistically significantly correlated with the visual excerpt determination task and performance on the verbal gist determination task. Some of the tasks relied on similar inferential processes, and some did not.

5.3 Coordinated vs. Uncoordinated Surrogates

We hypothesized that the uncoordinated surrogates would yield higher accuracy, and higher confidence ratings than the coordinated surrogates on the video gisting tasks, but the results were not always as expected. For example, the coordinated surrogate C2 (Magic A + V) led to comparable accuracy and confidence performance with the uncoordinated ones on the two verbal gist tasks: the free-text gist written task and the verbal gist determination task.

Participants performed the free-text gist written task more accurately with the uncoordinated surrogates than with the coordinated surrogates, but they were not accordingly confident in their responses.

For the keyword determination task, the results were mixed. Participants had statistically reliably higher accuracy with the uncoordinated surrogates U1 (magic A + storyboard V) and U2 (magic A + systematic subsampling V) than the coordinated surrogate C3 (manual A + V), but had statistically reliably higher accuracy with the coordinated surrogates C1 (Systematic subsampling A + V) and C2 (Magic A + V) than with the uncoordinated surrogate U3 (manual A + manual V).

The uncoordinated surrogates were more helpful than the coordinated surrogates for the keyframe determination task with respect to accuracy and confidence ratings, which was as we expected.

Participants did well on the visual excerpt determination task and the audio excerpt determination task with all surrogates. No statistically reliable differences were observed among the surrogates on the accuracy of any of the two tasks. These tasks were easy to perform, and did not tell us much about the differences in the effectiveness of different surrogates.

Results on the accuracy and confidence ratings on the verbal gist determination task were mixed too. The uncoordinated surrogates U1 (Magic A + Storyboard V) was very helpful in selecting the correct gist description for a video, yet the coordinated surrogate C2 (Magic A + V) was very helpful too.

Participants performed almost uniformly more quickly with the uncoordinated surrogates than the coordinated surrogates for all tasks, which was consistent with our hypothesis on the task completion time.

Participants' subjective ratings on the surrogates showed that the participants felt the coordinated surrogates were more useful and more usable than the uncoordinated ones regardless of their actual performance. No significant differences were observed in participants' ratings of engagement, but participants experienced statistically significantly more enjoyment using the uncoordinated surrogate U1 (Magic A + Storyboard V) than using the coordinated surrogate C1 (Systematic subsampling A + V), which was reinforced by their open-ended comments on what they liked and disliked about each surrogate. To put it simply, it was easy to peruse the images while listening to the audio and make connects between the images and audio, whereas it was hard to follow the flow of the surrogate when it was choppy and jumping around.

Though most participants did not like the uncoordinated surrogates with audio and moving visual, such as U2 (Magic A + Systematic subsampling V) and U3 (Manual A + Manual V), they did acknowledge the great breath of information carried in the uncoordinated surrogates. In addition, some participants felt the uncoordinated multimodal presentation was "a bit shocking at first" because they were "used to something else" and "didn't expect that type of stimulus". However, some participants thought the uncoordinated surrogates could become more usable as people got used to them: "It was difficult to process, but more manageable once I got the hang of it." It is possible that the uncoordinated presentations will be more usable and gradually liked by users when they see more of them in real-life video retrieval and browsing systems.

tasks. The automatically-generated uncoordinated surrogate U2 (Magic A + Systematic subsampling V) led to reasonable performance results, but was commented as "distracting" by many participants. The coordinated surrogates C2 (Magic A + V) was generated automatically using the MAGIC system, had reasonable performance results, and was liked by the participants for its user friendliness, smooth transitions, and good quality summaries. Therefore, we recommend adding U1 (Magic A + Storyboard V) and C2 (Magic A + V) in video retrieval and browsing systems (as they both can be generated automatically), and giving users control over the surrogate interfaces to select the appropriate surrogates to use.

5.4 Manually-generated vs. Automatically-generated Surrogates

Automated video summarization is a difficult challenge. As expected, the manually-generated surrogates in general led to higher accuracy than the automatically-generated ones on the video gisting tasks, with only two exceptions: the keyword determination task and the keyframe determination task.

Both of these tasks were more measures of specific token recognition rather than measures of gist inference. Thus the most effective surrogates for the gist inference tasks may not be equally helpful for the token recognition tasks, and vice versa. The automatically-generated surrogates may not be the most helpful in constructing a story of the video, yet they provided pieces of information throughout the video with fine granularity. Thus they were found to be very effective and even superior to the manually-generated surrogates on the token recognition tasks.

When the accuracy, confidence ratings, and task completion times were aggregated over all six tasks, manually-generated surrogates were only more helpful than the automatically-generated surrogates in task completion time.

Taking all factors into account (i.e., creation cost, compaction rate, accuracy, confidence, task completion time, and subjective measures), the automatically-generated uncoordinated surrogate U1 (Magic A + Storyboard V) was the most effective surrogate for the video gist tasks. The automatically-generated uncoordinated surrogate U2 (Magic A + Systematic subsampling V) led to reasonable performance results, but was seen as "distracting" by many participants. The coordinated surrogate C2 (Magic A + V) was generated automatically using the MAGIC system, had reasonable performance results, and was liked by the participants for its user friendliness, smooth transitions, and good quality summaries. Therefore, we recommend adding U1 (Magic A + Storyboard V) and C2 (Magic A + V) in video retrieval and browsing systems (as they both can be generated automatically), and giving users control over the surrogate interfaces to select the appropriate surrogates to use.

5.5 Learning Effect

We were also interested in whether there was any learning effect with the gist tasks. In the study, we provided participants with training videos for 4 out of the 6 surrogate conditions, and the result on the training videos were excluded from the data analysis. For the two manually-generated surrogate conditions, no training videos were provided due to the limited number of manually-generated surrogates which were very expensive to create. However, the manually-generated coordinated surrogate C3 was always tested after coordinated surrogate C2, and the manually-generated uncoordinated surrogate U3 was always tested after uncoordinated surrogate U2, such that the participants could get some training with similar types of surrogates before working with the manually-generated ones.

Table 5.2 presents the task accuracy means on each of the six tasks for each of the 20 test videos.

If there were learning effects with the tasks, the accuracy means for each task would increase as participants work with more test videos in each two- or four-video block (i.e., there were 2 or 4 training videos for each surrogate condition), as delimited by horizontal lines in Table 5.2. However, no learning effects were shown on any of the six tasks. Therefore, the accuracy differences among different test videos were due to video differences rather than learning effect.

Table 5.2: Average Task Accuracy by Video.

Video id	**Average Accuracy**					
	Task1	**Task2**	**Task3**	**Task4**	**Task5**	**Task6**
2	1.292	0.856	0.829	0.818	0.870	0.979
3	2.333	0.781	0.775	0.896	0.974	0.979
4	1.094	0.792	0.729	0.859	0.635	0.938
5	0.948	0.717	0.763	0.875	0.953	0.938
7	1.990	0.819	0.742	0.979	0.948	0.938
8	1.292	0.865	0.777	0.974	0.979	0.979
9	1.708	0.806	0.746	0.896	0.922	0.938
10	2.156	0.819	0.894	0.927	0.938	0.979
11	2.000	0.704	0.765	0.823	0.865	1.000
12	1.302	0.767	0.754	0.870	0.891	0.938
14	1.365	0.681	0.896	0.964	0.974	1.000
15	1.417	0.638	0.790	0.964	0.964	0.958
16	1.760	0.731	0.838	0.964	0.995	1.000
17	2.281	0.856	0.717	0.823	0.844	0.938
19	1.438	0.685	0.781	0.932	0.833	0.896
20	1.906	0.765	0.702	0.844	0.688	0.958
21	1.615	0.863	0.792	0.948	0.938	0.958
22	0.719	0.696	0.725	0.953	0.964	0.979
23	1.990	0.640	0.671	0.880	0.953	0.979
24	2.354	0.704	0.646	0.969	0.969	0.979

5.6 Limitations of the study

There are some limitations for this user study. Due to the limited number of video surrogates and limited types of video gisting tasks covered in this study, the results from this study might not be able to be generalized to other types of video gisting or searching tasks. Nevertheless, the findings do provide implications to interface design for video retrieval systems.

To simplify the study design and reduce individual video differences, a collection of instructional videos of similar structures and similar conceptual levels were used as test videos in this study. Results found for instructional videos may not be generalized to other video genres where the videos have different structures. Different types of video surrogates may be effective and useful for retrieving and making sense of videos from other genres.

The study was a within-subjects user study of 48 participants. As mentioned in Section 3.3.3.4, it was impossible to completely counterbalance the order of all 6 surrogate conditions with a group of 48 participants. To completely counterbalance all 6 conditions, we would need at least $6! = 720$ participants. Thus, the surrogate conditions were only partially counterbalanced by first grouping the surrogates into coordinated and uncoordinated, and then counterbalancing the groups as well as the conditions within each group. In addition, the manually-generated surrogates were always the last conditions in the coordinated and uncoordinated groups. Given this compromised experimental design, some ordering effects of surrogates may still exist in favor of the manually-generated surrogates.

Task 4 (Visual excerpt determination task) and Task 5 (Audio excerpt determination task) were easy tasks. Participants did very well in these two tasks with all surrogate conditions and thus there was little variance to detect in accuracy across the six surrogates. The surrogates did differ significantly in terms of task completion time for Task 4 and Task 5. In future studies, if we are more interested in accuracy and less interested in confidence ratings or task completion time, we may consider removing these two tasks from study procedures as they do not tell us much in evaluating the effectiveness of different video surrogates for video gisting and sense-making.

Previous studies concluded that performance on some tasks may be affected by the video with which the participant was interacting (Wildemuth et al., 2003). Differences in the video may be related to the participants' ability to making sense of the video using the surrogates. As the manually-generated surrogates were only created for 4 videos out of the 24, a possible video characteristics effect may exist in the results when comparing the automatically-generated and the manually-generated surrogates. In future studies, all surrogates to be examined should be created for all test videos in the study to eliminate or reduce the video characteristics effect.

In this study, we conducted one-way repeated-measures ANOVA and post hoc pairwise comparisons to investigate the effectiveness of different surrogate conditions and to address the research questions. Alternatively, two-way repeated-measures ANOVA could be used if we have a balanced experimental design – two coordinated surrogate conditions (one manually generated, and one automatically generated), and two uncoordinated surrogate conditions (one manually generated, and one automatically generated) – with the ordering of the surrogate conditions fully counter-balanced.

A two-way repeated-measures (within-subjects) ANOVA can be conducted to evaluate the effect of coordination between the audio and visual channels of the surrogates and automation of surrogate creation. The within-subjects factors will be coordination between the surrogate channels with two levels (coordinated and uncoordinated) and automation of surrogate creation with two levels (automated and manual). The overall hypotheses can then be evaluated for both the main effects and interaction effect between the two factors – coordination and automation.

However, due to the limitation of the design of this study, we had six surrogate conditions and a not fully balanced experimental design. Therefore, one-way repeated-measures ANOVA was a more appropriate analysis approach for this study.

Moreover, a significant F value in a one-way ANOVA indicates that there are statistically reliable differences in the means, but does not tell you where those differences are. Thus, multiple comparisons of group means should be conducted to follow the main effect if it is significant. Various methods have been developed for doing the multiple comparisons of group means. In this study, we used the Least Significant Difference (LSD) t test for the multiple comparisons of group means, which does not control the overall probability of rejecting the hypotheses that some pairs are different and makes no adjustment for the number of comparisons (if there is more than one). With the LSD test, if we do enough comparisons, some comparisons may show up as significant just by chance.

Methods such as the Bonferroni adjustment, the Sidak adjustment, and the Scheffe test, can be used to offer adjustment to compensate for the fact that the chance of rejecting the null hypothesis increases with each additional pairwise comparison when using the LSD test. The adjustment methods reduce the overall chance of falsely rejecting each hypothesis; however, the adjustment methods also increase the chance that we do not reject the null hypothesis when we should indeed reject it. It is possible that we will not find any pairs to have significant differences when the adjustment methods are used for the multiple comparisons, while the LSD test will suggest some significant differences. Considering the pros and cons of the adjustment methods, we stuck with the LSD method which makes no adjustment.

Chapter 6

CONCLUSION

Good surrogates allow users to quickly derive the gist of the video without having to take time to watch the full video; hence they are crucial to video retrieval and browsing systems. This research evaluated the effectiveness of coordinated and uncoordinated multimodal surrogates by conducting a within-subjects user study with 48 participants. The study investigated the effects of manually and automatically generated surrogates with coordinated (i.e., pre-processed integration) and uncoordinated (i.e., user-centered integration) audio and visual channels, on six recognition and inference tasks, and users' perspective of the surrogates.

We hypothesized that participants would spend more time in consuming the uncoordinated surrogates than consuming the coordinated surrogates, and spend more time in consuming the manually-generated surrogates than consuming the automatically-generated surrogates. However, there were no statistically significant differences in the surrogate consumption time among the six surrogates.

We hypothesized that people would perform the video gisting tasks better with the uncoordinated surrogates than with the coordinated surrogates, and that people would perform the video gisting tasks better with the manually-generated surrogates than with the automatically-generated surrogates, in terms of task accuracy, confidence ratings, and task completion time.

Results demonstrated that the type of the surrogates had statistically significant effects on accuracy for four out of the six tasks: free-text gist written task, keyword determination task, keyframe determination task, and verbal gist determination task. No statistically significant main effects of surrogates were found on task accuracy for the other two tasks: visual excerpt determination task and audio excerpt determination task.

Although the uncoordinated surrogates did not yield statistically reliably higher accuracy than the coordinated surrogates in all tasks, the uncoordinated surrogate U1 (Magic A + Storyboard V) was in general the most cost-effective surrogate for video gisting, considering creation cost, accuracy, confidence, task completion time, and subjective measures. Moreover, the multimodal presentation of U1 was liked by many participants. As noted by a participant, "The storyboard condition was very simple, and allowed the small bit of information provided to be more wholly taken in."

Furthermore, statistically significant main effects of surrogates were also found on participants' confidence ratings on four of the six tasks: free-text gist written task, keyframe determination task, audio excerpt determination task, and verbal gist determination task.

Although no statistically reliable differences were shown on the time participants spent on consuming the surrogates, the time participants spent on completing the tasks was statistically reliably different among the surrogates for all six tasks. In general, the participants performed the tasks more quickly with the uncoordinated surrogates than with the coordinated ones, and more quickly with the manually-generated surrogates than with the automatically-generated ones.

Moreover, the manually-generated surrogates led to higher accuracy than the automatically-generated ones on the free-text gist written task, visual excerpt determination task, audio excerpt determination task, and verbal gist determination task, but not in the keyword determination task and the keyframe determination task, which were both measures of specific token recognition rather than measures of gist.

Despite the high accuracy performance on the uncoordinated surrogates, participants were not more confident in their responses with the uncoordinated surrogates than with the coordinated ones, and most participants still preferred the coordinated surrogates rather than the uncoordinated ones, according to their comments in the post-session questionnaires.

The results and findings of this study have implications for user interface design of future digital video retrieval systems. There was a lot of variability in the performance results: none of the surrogates was found the most effective for all gisting tasks. However, the uncoordinated surrogate U1 (Magic A + Storyboard V) and the coordinated surrogate C2 (Magic A + V) were in general very effective surrogates for all video gisting tasks considering compaction rate, accuracy, confidence, task completion time, and subjective measures. U1 and C2 are also cost-effective surrogates because they can both be generated automatically, thus have relatively low creation cost. Therefore, we recommend adding these two surrogates in video retrieval and browsing systems. We also recommend that real-world video retrieval systems grant users control over the interfaces, and put the users in control of selecting the appropriate surrogates to use when searching and making sense of videos. Video retrieval systems can also use personal profiles of the users to determine or recommend the default video surrogate(s) to be shown to the users (e.g., coordinated or uncoordinated).

The user study described in this dissertation was an extrinsic evaluation of different types of video surrogates based on some specific video gisting tasks. The gisting tasks evaluated the participants' ability to infer an overall understanding of the video and construct a story of the video, recognize small pieces of information (i.e., keywords and keyframes) that occurred in the video, and recognize audio clip, visual clip, or verbal gist that belonged in the video based on the video surrogates viewed. The tasks were associated with the users' needs to select video frames or clips for re-use and the users' needs to select videos from a collection for particular purposes. These tasks had been employed in several past studies, and were found practical for differentiating the effectiveness of different surrogates. These tasks can be further refined in future studies in video retrieval research, and we're also interested in developing new tasks which are more realistic and reflect real-world user needs and video retrieval tasks. For example, given a specific topic, a person may be asked to search and browse a video repository and retrieve a few relevant videos based on the surrogates available for the videos without watching the full videos. Furthermore, the tasks should be performed by real-world users who will use or who will be interested in using the surrogates to search for videos in real video retrieval and browsing systems.

In addition to the subjective measures, more intrinsic evaluation may be added in future studies to judge the quality of the surrogates directly rather than using the gisting tasks. For example, the surrogates may be judged based on the coverage of important or interesting events in the source videos or their similarity to the ground-truth summaries generated by human judges.

With the rapid and mass adoption of powerful mobile devices such as iPhones and iPads, more and more people have started watching online video on their mobile devices. Therefore, surrogates usable and helpful for quickly making sense of video on mobile devices are needed. Do humans behave differently when using video surrogates on computers and on mobile devices? Are the video surrogates designed for computers equally useful and helpful for mobile devices? What are the limitations for video surrogates on mobile devices? In this study, the screen sizes of the surrogates were determined according to past study experiences. Although past study experiences demonstrated that these surrogate sizes were adequate for the users, one participant in this study noted that “the images ... were so small.” For mobile devices, the screen real estate is limited; hence, it is important that surrogates designed to be used on mobile devices only require minimal screen space (e.g., audio surrogates). Future studies may address these issues through carefully designed user studies.

Appendix: Metadata for the Test Videos

6000 — NASA Connect - Ahead Above The Clouds. NASA Connect Video containing five segments as described below. NASA Connect Segment exploring new and future technology to help meteorologists predict hurricanes and other severe weather. The video explores GIFTS, or geostationary satellites, and other developing technologies at NASA. NASA Connect Segment explaining what hurricane hunters do and how they do it. The video explores the instruments they use to collect data from a hurricane and the types of data collected such as temperature, moisture, air pressure and wind. NASA Connect Segment explaining software tools and products that use interactivity to network NASA research data. The video describes dynamic websites that use visualization, simulation, and remote sensing tools to help students study hurricanes. NASA Connect Segment explaining the fundamentals of hurricanes and how meteorologists predict hurricanes. The video also features a meteorologists from The Weather Channel to explain how data is collected and how hurricanes are predicted. NASA Connect Segment involving students in an activity that uses a game called the Imperfect Storm. Students must track a hurricane, predict the probability of landfall, and issue watches and warnings.

6048 — NASA Connect - Virtual Earth. NASA Connect Video containing six segments as described below. NASA Connect segment explaining Earth System Science. The video also explores how modern technology studies the many different areas of Earth System Science. NASA Connect segment exploring NASA's three mission statements. The video explores NASA's purpose to achieve these missions for the planet Earth. NASA Connect segment explaining the mathematical standard of representation. The video gives examples and explores the purpose of representation. NASA Connect segment involving students in two web activities that teach about the lithosphere and hydrosphere. The video explores the two activities called Earthquake Hunters and Water World. NASA Connect segment involving students in an activity that is an introduction to systems. The video explores earth systems and the global water cycle. NASA Connect segment explaing basic facts about systems and subsystems. The video also introduces the study of Earth System Science.

6063 — NASA Connect - The Venus Transit. NASA Connect Video containing six segments as described below. NASA Connect segment explaining how scientists determined the distance between the earth and the sun. The video also explores the geometric technique called parallax. NASA Connect segment involving students in a classroom activity that uses graphing, measurement, and ratios to construct a scaled model of the Solar System. NASA Connect segment exploring what it means to scale and why scientists use scale models and drawings. The video also explores math terms that are associated with scale models and drawings. NASA Connect segment that explores how astronomers and scientists use astronomical units in measuring distances in the Solar System. NASA Connect segment that challenges students to participate in an activity to scale the universe. The video involves students in a proposal to determine a new baseline distance to use for an astronomical unit. NASA

Connect segment that explains the Venus Transit and compares it to a solar eclipse.

6089 — NASA Connect - Measurement, Ratios, and Graphing: Safety First. NASA Connect Video containing five segments as described below. NASA Connect Segment that explores the safety of airports. The video explains the Federal Aviation Administration's primarily responsibility is maintaining the safety of public aviation. NASA Connect Segment that explores instructional technologies relating to the show. These tools include a compact disc called Gate To Gate produced by NASA to introduce students to the air traffic control system. ASA Connect Segment involving students participating in an activity that explores the air traffic control system. Its objectives are to analyze aircraft coordinates, use tools to determine distance, and apply ratios to calculate air safety travel index. NASA Connect Segment that explores the air traffic control system. It also looks at NASA's program to study safer aviation techniques in the air and on the ground. NASA Connect Segment that explores the safety of air travel through new technologies. It also explains the math, science, and technology that NASA scientists use in their research.

6102 — NASAConnect - Plane Weather. NASA Connect video containing five segments as described below. NASA Connect Video that explains how meteorology, specifically icing, effects the ground operations of aircraft. Explores research being conducted to study the effects of icing by using refrigerated wind tunnels. NASA Connect Segment that explores the Joint Runway Friction Measurement Program that investigates aircraft losing traction on icy runways. NASA Connect Segment that explores meteorology and how it affects aviation safety throughout the National Airspace System. Explains the importance of daily forecasts and tools like satellites to understand complex processes and the fundamentals of weather. NASA Connect Segment involving students in an activity that investigates how surface conditions influence the coefficient of friction between two surfaces. NASA Connect Segment explaining the aviation weather channel and why pilots need to have a continual awareness of the changing nature of the atmosphere on their flight route.

6108 — NASAConnect - Recipes For the Future. NASA Connect Video containing five segments as described below. NASA Connect Video answering questions from emails and call-ins. Two experts sit in to answer questions about future space vehicles, composite materials, and daily uses for those materials. NASA Connect Segment involving students in an activity that investigates the strenth and deflection of composite material with and without reinforcement. It reviews vocabulary including polymer, fiber, stress cracks, and maximum deflection. NASA Connect Segment exploring composite material, what it is, and how it is made. Explains the goals of composites are to develop stronger, more durable, lighter weight materials for space vehicles. NASA Connect Segment explaining the process of testing new materials. It also explores the process for testing and analyzing structures for new space vehicles at room temperature and extreme temperatures. NASA Connect Segment that explores how scientists use recipes in chemistry to formulate new combinations and build new materials. Explains the difference between chemical and physical changes of substances.

6122 — NASAConnect - Tools of the Aeronautic Trade. NASA Connect Video containing five segments as described below. NASA Connect Video that explains aerodynamic forces that affect aircraft performance and how these forces relate to each other. NASA Connect Video involving students in an activity to create a wind tunnel to test the effect of drag while emphasizing data analysis. NASA Connect Segment explaining the development of the US standard system of measurement and the metric system and how the two systems differ. NASA Connect Segment exploring a SEMAA school targeting math, science, and technology. Students demonstrate interactive simulation software product called FoilSim. NASA Connect Segment explaining wind tunnels and how they are used as research tools. It also explores the SR-71 Blackbird and why it's used as an ideal research test plane.

6293 — NASA Connect - Quieting The Skies. NASA Connect Video containing four segments as described below. NASA Connect segment exploring the research and study efforts applied towards acoustics and noise, especially that related to aircraft. The segment also explains the study of psychological effects of noise on people. NASA Connect segment featuring a panel of two experts from NASA that answer students' questions by phone and email. The questions pertain to aircraft and noise reduction. NASA Connect segment exploring all the basics of sound including how it works and how it travels. The video also explains how the ear works. NASA Connect segment involving students in an activity called the Speed of Sound. The students investigate how sound waves travel at different speeds under various conditions.

6077 — NASA Connect - Festival of Flight. NASA Connect Video containing four segments as described below. NASA Connect Segment involving students in an activity to gather and graph statistical data and build mathematical models in a project involving rocket propulsion. NASA Connect Segment explaining how NASA uses computer simulation to design spacecraft, including the next reusable launch vehicle. NASA Connect Segment explaining how launch vehicles overcome the force of gravity through the force of thrust. NASA Connect Segment explaining how Reusable Launch Vehicles are designed and used by NASA for launch, space travel, and re-entry.

6082 — NASA Connect - Geometry of Exploration: Water Below the Surface of Mars. NASA Connect Video containing six segments as described below. NASA Connect Segment involving students participating in an activity to measure and calculate ellipses. The activity explains ellipses and their relation to Earth and Mars. NASA Connect Segment exploring ideas of water on Mars. It also explains the Mars Microprobe and its navigation on mars and how this relates to geometry. NASA Connect Segment explaining why we are exploring Mars. It also reveals tools and techniques used to explore Mars. NASA Connect Segment explores a National Arts, Sciences and Technology Education Initiative called the Mars Millenium Project. Allows students to participate in activity to design a community for Mars inhabitants in the year 2030. NASA Connect Segment that explores how NASA scientists use geometry to navigate spacecraft from Earth to Mars. It also explains the goals and accomplishments of the Viking Mission. NASA Connect Segment that explains who Pythagoras was and how he contributed towards geometry. Also it explains how geometry is used in everyday life.

6006 — NASA Connect - Better Health From Space To Earth. NASA Connect Video containing seven segments as described below. NASA Connect Video involving students in an activity that estimates average daily energy needs. The video also explains BMR and other vocabulary relating to energy. NASA Connect Segment exploring the mathematical concepts estimation and measurement. The video relates these concepts to daily activities and to health and nutrition. NASA Connect Segment involving students in a web activity. The video explains how students complete the Exercise Project and the Heart Plot Project. NASA Connect Segment exploring good nutrition and exercise. NASA Connect Segment explaing how astronauts exercise in space and how they endure long-duration space flights. The video also explores ways of measuring levels of fitness. NASA Connect Segment explaining the importance of good nutrition and specifically nutrients such as calcium. The video explores bones and effects on astronaut's bones. NASA Connect Segment involving students in an activity that applies estimation and measurement skills. The video explores estimations of serving sizes for different foods.

6014 — NASA Connect - Dancing In The Night Sky. NASA Connect Video containing five segments as described below. NASA Connect Segment exploring the Aurora Borealis or Northern Lights. This segment exlains this natural phenomena and its history. NASA Connect Segment involving students in an activity that investigates the Aurora Borealis. During the activities the students use geographic coordinates to find and plot locations on maps, draw conclusions using graphical data, and convert centimeters to kilometers. NASA Connect Segment exploring ground-based instruments and rockets used to analyze and research the auroras. The segment also explains the concepts of data analysis and measurement in scientific research. NASA Connect Segment explaining Earth oribiting satellites that record and analyze the causes of auroras. The segment explores the IMAGE satellite and other technology. NASA Connect Segment explaining what NASA is doing to explore auroras.

The segment also answers questions like what are the phases of the Aurora and how scientists use satellite images to monitor auroras.

6020 — NASA Connect - Geometry and Algebra - Glow With the Flow. NASA Connect Video containing six segments as described below. NASA Connect Segment explaining air flow. The video describes how drag, lift, and thrust work. NASA Connect Segment exploring drag and agebraic relationships. The video explains flow visualization and air flow and how engineers use algebra in their work. NASA Connect Segment explaining the new concept aircraft in development known as the blended wing body. The video explains how engineers and scientists uses geometry to help with development. NASA Connect Segment involving students in a classroom activity called What A Drag. The video explores how shape affects drag. NASA Connect Segment involving students in a classroom activity. The video explores how surface area affects drag. NASA Connect Segment exploring computer simulation tools for research on drag. The video features the Mars Airbourne Explorer simulation computer program.

6027 — NASA Connect - Geometry of Exploration - Eyes Over Mars. NASA Connect Video containing six segments as described below. NASA Connect Segment involving students in a classroom activity that measures shadows and uses geometry to determine sizes of angles. NASA Connect Segment explaining questions about Erastothenes, the Earth's circumference, parallel lines, angle relationships, and a transversal. NASA Connect Segment featuring an online activity to show students how to design a planetary observer like the Mars Global Surveyor. NASA Connect Segment explaining surveying and how surveyors use geometry. NASA Connect Segment exploring how the

Mars Global Surveyor works and how students survey Mars by using shadows, angles, and geometry. The video also explains how land formations are measured on Mars. NASA Connect Segment explaining how NASA scientists survey Mars with the Mars Global Surveyor. The video also explains aerobraking and how geometry influences this.

6034 — NASA Connect - Personal Satellite Assistant - The Astronaut's Helper. NASA Connect Video containing six segments as described below. NASA Connect Segment exploring the aspects of microgravity and how it affects objects in space. Explores object motion and friction and tests the PSA prototype in accordance with these forces. NASA Connect Segment exploring more aspects of the Personal Satellite Assistant. It explains motion and its relationship with the mass of objects in connection to the PSA. NASA Connect Segment explaining mechanical systems. It also compares and contrasts a mechanical system to the system of the International Space Station and Personal Satellite Assistants. NASA Connect Segment explaining the literary origins of robots. It also explores the development of the robot and how scientists use robots in research and technology. NASA Connect Segment exploring the different types of robots. It also explores robots such as the Mars Rover that scientists at NASA use to explore beyond the Earth. NASA Connect Segment involving students in an activity that investigates volume and surface area in two different cylinders. The video also explains basic mathematical functions to help answer the questions.

6041 — NASA Connect - Shapes of Flight. NASA Connect Video containing six segments as described below. NASA Connect segment exploring the first types of flights including kite flights. The video explores Kitty Hawk, North Carolina and experimental airplanes at a yearly festival. NASA Connect segment explainging the fundamentals of flight and the science behind it. NASA Connect segment involving students in an activity exploring glide ratio and surface area. NASA Connect segment featuring two NASA experts in a question and answer session. The video involves people calling in and emailing questions for the experts to answer. NASA Connect segment explaining how different forces affect aircraft. The video also explores team work and engineering for conducting research. NASA Connect segment explaining the process of modeling and testing model aircraft. The video features two experts who explain how wind tunnels work.

6267 — NASA Connect - The A-Train Express. NASA Connect Video containing six segments as described below. NASA Connect segment explaining aerosols and their affect on the changes of climate and weather. The segment also explores the lidar technique in the new CALIPSO satellite. NASA Connect segment involving French students in an activity called the Aerosols Protocol. The segment investigates how the sun's light is absorbed by particles in the atmosphere.? NASA Connect segment explaining the difference between weather and climate. The segment explores what factors determine weather and how climate is affected by the weather. NASA Connect segment exploring the GLOBE International science program. The segment explains how the program helps scientists collect environmental data from all over the world. NASA Connect segment explaining how scientists use satellites to predict weather. The segment explores the Afternoon Constellation, or the collection of satellites known as the 'A' Train as well as weather balloons, weather stations and local weather observers. NASA Connect segment involving students in an activity called Size Up the Clouds. The segment explores simulated cloud types to estimate precipitation content.

6274 — NASA Connect - The Future of Flight Equation. NASA Connect Video containing six segments as described below. NASA Connect segment involving students in a web activity that teaches how to use different shapes to design different aircraft. The segment also features an online tutorial for instruction in technology. NASA Connect segment exploring the current situation of commercial flight and what kinds of new technology is in place to help pilots today. NASA Connect segment explaining the tools, techniques, and requirements of designing an aircraft. The segment also explains the importance in wind tunnels and model planes. NASA Connect segment exploring the future of aircraft such as NASA's new experimental plane, the Hyper X with a scram jet. NASA Connect segment involving students in a web activity featuring the Plane Math Website to teach students about aeronautical principles, geometric and algebraic math concepts, and aircraft design. NASA Connect segment exploring the process of flight testing. The segment features the Hyper-X and answers questions pertaining to its test stage.

6298 — NASA Connect - Proportionality - Modeling the Future. NASA Connect Video containing five segments as described below. NASA Connect segment involving students in an online activity that features an Airplane Design Workshop that gives an example how artificial intelligence helps engineers in modeling and designing aircraft. NASA Connect segment involving students in an activity that explores the Fibonacci Sequence. The segment explores ratios, measurements, and proportionalities. NASA Connect segment explaining ratios and proportions. The segment describes how these math concepts helped the Wright Brothers to invent the first flying machine. NASA Connect segment explaining how the Fibonacci sequence and the Golden Ratio help NASA engineers research, design and develop airplanes. NASA Connect segment exploring transportation growth since the early 1900s and how the patterns of this growth are mathematical and are related to the Fibonacci sequence.

6304 — NASA Connect - Wired For Space. NASA Connect Video containing six segments as described below. NASA Connect segment exploring how algebra and arrays are used in NASA's activities. The segment also explains voltage, current, amp, and resistance. NASA Connect segment explaining how NASA is using electricity and magnetism to propell spacecraft into orbit. The segment also explains acceleration, mass, and force in an algebraic equation. NASA Connect segment involving students in an online activity that investigates a physics module on electricity and magnetism. The activity studies static charge, moving charge, voltage, resistance, and current. NASA Connect segment involving students in an activity called Make It Go which simulates NASA research. It uses an Electrodynamic Demonstration Unit to investigate electricity and magnetism. NASA Connect segment exploring how NASA is researching to design, build and test a new propulsion technology that uses magnetism, electricity, and tethers instead of rocket engines. NASA Connect segment explaining how NASA uses tethers to help propell spacecraft already in orbit. The segment also explores the NASA project called ProSEDS which is the first to experiment with a tether system.

Bibliography

Abracos, J. and Lopes, G. P. (1997). Statistical methods for retrieving most significant paragraphs in newspaper articles. In *ACL/EACL Workshop on Intelligent Scalable Text Summarization*. 33

Adjeroh, D. A. and Lee, M.-C. (1995). Mechanisms for automatic extraction of primary features for video indexing. In *ICSC '95: Proceedings of the Third International Computer Science Conference on Image Analysis Applications and Computer Graphics*, pages 493–494, London, UK. Springer-Verlag. 62

Agnihotri, L., Devera, K., McGee, T., and Dimitrove, N. (2001). Summarization of video programs based on closed captions. In *Proc. SPIE. Conf. Storage and Retrieval for Media Databases*, page 599C607, San Jose, CA. 42

Alatan, A. A., Akansu, A. N., and Wolf, W. (2001). Multi-modal dialog scene detection using hidden markov models for content-based multimedia indexing. *Multimedia Tools Appl.*, 14(2):137–151. 69

American Library Association (2000). *Task Force on Metadata: Final Report*. Committee on Cataloging: Description and Access. Association for Library Collections and Technical Services. http://www.libraries.psu.edu/tas/jca/ccda/tf-meta6.html. 60

Amir, A., Srinivasan, S., and Ponceleon, D. (2003). *Efficient Video Browsing Using Multiple Synchronized Views*, chapter 1. in Video Mining, pages 1–30. Kluwer Academic Publishers, Boston. 74, 110

Armitage, L. H. and Enser, P. G. (1997). Analysis of user need in image archives. *Journal of Information Science*, 23(4):287–299. xviii, 15, 16

Arons, B. (1994). Pitch-based emphasis detection for segmenting speech recordings. In *Recordings*, pages 1931–1934. 50

Arons, B. (1997). Speechskimmer: A system for interactively skimming recorded speech. *ACM Transactions on Computer Human Interaction*, 4:3–38. 35, 36, 39, 40

Babaguchi, N., Kawai, Y., and Kitahashi, T. (1999). Event based video indexing by intermodal collaboration. In *Proceedings of First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM&apos;99*, pages 1–9. 69

Begg, I. (1972). Recall of meaningful phrases. *Journal of Verbal Learning and Verbal Behavior*, 11:431–439. 27

Berger, A. A. (2004). *Media Analysis Techniques.* Sage Publications, Inc., 3rd edition edition. 14

Blecic, D. D., Bangalore, N. S., Dorsch, J. L., Henderson, C. L., Koenig, M. H., , and Weller, A. C. (1998). Using transaction log analysis to improve opac retrieval results. *College and Research Libraries*, 59(1):39C50. 114

Boekelheide, K., Brown, A., Fu, X., Marchionini, G., Oh, S., Rogers, G., Saelim, B., Song, Y., and Stutzman, F. (2006). Audio surrogation: for digital video a design framework. SILS Technical Report TR-2006-02, University of North Carolina at Chapel Hill, UNC. 31, 74, 76, 77

Boguraev, B. and Neff, M. (2000). Lexical cohesion, discourse segmentation and document summarization. In *Proceedings of RIAO-2000, content-based multimedia information access.* 140

Bordwell, D. (1989). A case for cognitivism. *IRIS Spring*, (9):11–41. 18

Branigan, E. (1992). *Narrative Comprehension and Film.* Routledge. 18

Cabasson, R. and Divakaran, A. (2003). 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. 41, 49

Carneiro, G., Chan, A. B., and Moreno, P. J. (2007). Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(3):394–410. Member-Vasconcelos,, Nuno. 62

Chen, F. R. and Withgott, M. (1992). 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. 33, 39, 40

Chen, J., Taskiran, C., Albiol, A., Delp, E., and Bouman, C. (2004). Vibe: A compressed video database structured for active browsing and search. *IEEE Transactions on Multimedia*, 6(1):103–118. 74

Christel, M. and Conescu, R. (2006). Mining novice user activity with trecvid interactive retrieval tasks. In *Image and Video Retrieval*, pages 21–30. 105

Christel, M., Stevens, S., Kanade, T., Mauldin, M., Reddy, R., and Wactlar, H. (1996). Techniques for the creation and exploration of digital video libraries. In *Multimedia Tools and Applications, B. Furht, Editor*. Kluwer Academic Publishers. 42

Christel, M. G. (2009). *Automated Metadata in Multimedia Information Systems: Creation, Refinement, Use in Surrogates, and Evaluation*. Morgan and Claypool Publishers. 105

Christel, M. G., Hauptmann, A. G., Warmack, A. S., and Crosby, S. A. (1999). Adjustable filmstrips and skims as abstractions for a digital video library. In *ADL '99: Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries*, page 98, Washington, DC, USA. IEEE Computer Society. 73

Christel, M. G., Smith, M. A., Taylor, C. R., and Winkler, D. B. (1998). 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. ACM Press/Addison-Wesley Publishing Co. 2, 4, 34, 38, 56, 76, 79, 87, 89, 99

Christel, M. G., Winkler, D. B., and Taylor, C. R. (1997). Improving access to a digital video library. In *INTERACT '97: Proceedings of the IFIP TC13 Interantional Conference on Human-Computer Interaction*, pages 524–531, London, UK, UK. Chapman & Hall, Ltd. 23, 107

Cole, M., Frankel, F., and Sharp, D. (1971). Development of free recall learning in children. 4:109–123. 30

Cooper, M. and Foote, J. (2002a). Automatic music summarization via similarity analysis. In *the 3rd International Conference on Music Information Retrieval (ISMIR '02)*, page 81C85, Paris, France. 39

Cooper, M. and Foote, J. (2002b). Summarizing video using non-negative similarity matrix factorization. In *Int. Workshop on Multimedia Signal Processing*, St. Thomas, U.S. Virgin Islands. Available: `http://citeseer.nj.nec.conV56859i.html`. 46

Damnjanovic, U., Fernandez, V., Izquierdo, E., and Martinez, J. M. (2008). Event detection and clustering for surveillance video summarization. In *WIAMIS '08: Proceedings of the 2008 Ninth International Workshop on Image Analysis for Multimedia Interactive Services*, pages 63–66, Washington, DC, USA. IEEE Computer Society. 71

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly*, 13(3):319–340. 84, 112, 156, 209 de Silva, G. C., Yamasaki, T., and Aizawa, K. (2005). Evaluation of video summarization for a large number of cameras in ubiquitous home. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 820–828, New York, NY, USA. ACM. 83

Dilley, M. and Paivio, A. (1968). Pictures and words as stimulus and response items in paired-associate learning of young children. 6:231–240. 30

Dimitrova, N. and Abdel-Mottaleb, M. (1997). Content-based video retrieval by example video clip. In *Storage and Retrieval for Image and Video Databases (SPIE)*, volume Vol. 3022, pages 59–70. 59, 67

Ding, W., Marchionini, G., and Soergel, D. (1999). Multimodal surrogates for video browsing. In *DL '99: Proceedings of the fourth ACM conference on Digital libraries*, pages 85–93, New York, NY, USA. ACM. 2, 22, 23, 24, 31, 79, 87, 108, 113

Ding, W., Marchionini, G., and Tse, T. (1997). Previewing video data: Browsing key frames at high rates using a video slide show interface. In *International Symposium on Research, Development and Practice in Digital Libraries*, pages 151–158, Tsukuba Science City, Japan. 99

Eakins, J. P. and Graham, M. E. (1999). Content-based image retrieval: A report to the jisc technology applications programme. Technical report, Institute for Image Data Research, University of Northumbria at Newcastle. 17, 18, 32

Ekin, A. and Tekalp, A. M. (2003). Automatic soccer video analysis and summarization. *IEEE Trans. on Image Processing*, 12:796–807. 49, 83

Enser, P. (1995). Pictorial information retrieval. *Journal of Documentation*, 51(2):126–170. 16

Ericsson, K. A. and Simon, H. A. (1984). *Protocol Analysis: Verbal Reports as Data.* MIT Press, Cambridge, MA. 113

Erol, B., Lee, D.-S., and Hull, J. (2003). 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. IEEE Computer Society. 54, 56

Farin, D., Effelsberg, W., and deWith, P. H. N. (2002). 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. 45

Fenstermacher, K. D. and Ginsburg, M. (2003). Client-side monitoring for web mining. *J. Am. Soc. Inf. Sci. Technol.*, 54(7):625–637. 114

Ferman, A. M. and Tekalp, A. M. (2003). Two-stage hierarchical video summary extraction to match low-level user browsing preferences. *IEEE Trans. Multimedia*, 5(2):244–256. 45, 46, 83

Florida, R. Z., Zwaan, R. A., and Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin*, 123:162–185. 12

Foote, J., Cooper, M., and Wilcox, L. (2000). Enhanced video browsing using automatically extracted audio excerpts. *IEEE*. 54, 80

Funk, M. and Reid, C. (1983). Indexing consistency in medline. *Bull Med Libr Assoc.*, 71(2):176–183. 137

Gan, C. and Donaldson, R. (1988). Adaptive silence deletion for speech storage and voice mail applications. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 36, no. 6:pp. 924–927. 36, 50

Gerber, S. and Wulfeck, B. (1977). The limiting effect of discard interval on time-compressed speech. *Language and Speech 20*, 2:108C115. 35

Ghani, J. A., Supnick, R., and Rooney, P. (1991). The experience of flow in computer-mediated and in face-to-face groups. In *ICIS '91: Proceedings of the twelfth international conference on Information systems*, pages 229–237, Minneapolis, MN, USA. University of Minnesota. 84, 112, 156, 209

Gong, Y. (2003). Summarizing audiovisual contents of a video program. *EURASIP J. Appl. Signal Process.*, 2003:160–169. 52, 53, 78

Gong, Y. and Liu, X. (2003). Video summarization and retrieval using singular value decomposition. *Multimedia Syst.*, 9(2):157–168. 34, 47, 56, 57

Gong, Y., Sin, L. T., Chuan, C. H., Zhang, H., and Sakauchi, M. (1995). Automatic parsing of tv soccer programs. *Multimedia Computing and Systems, International Conference on*, 0:0167. 67

Goodrum, A. and Spink, A. (2001). Image searching on the excite web search engine. *Information Processing & Management*, 37(2):295 – 311. 31

Goodrum, A. A. (2001). Multidimensional scaling of video surrogates. *J. Am. Soc. Inf. Sci. Technol.*, 52(2):174–182. 2, 23, 71, 85, 86, 87, 107

Greisdorf, H. and O'Connor, B. (2002). Modelling what users see when they look at images: a cognitive viewpoint. *Journal of Documentation*, 58(1):6–29. 17, 18, 32

Grodal, T. (1999). *Moving Pictures: A New Theory of Film Genres, Feelings, and Cognition*. Oxford University Press, USA. 19, 150

Gunther, R., Kazman, R., and MaccGregor, C. (2004). Using 3d sound as a navigational aid in virtual environments. 23(6):435–446. 31

Hampapur, A., Weymouth, T., and Jain, R. (1994). Digital video segmentation. In *MULTIMEDIA '94: Proceedings of the second ACM international conference on Multimedia*, pages 357–364, New York, NY, USA. ACM. 45

Hanjalic, A. and Zhang, H. (1999). An integrated scheme for automated video abstraction based on unsupervised cluster-validity analysis. *IEEE Trans. Circuits Syst. Video Technol*, 9(8):1280–1289. 45

Hanson, J. (1987). *Understanding Video: Applications, Impact, and Theory*, volume 19. Newbury Park, CA: Sage Publications. 11, 12, 14

Hauptmann, A., Christel, M., Lin, W., Maher, B., Yang, J., Baron, R., and Xiang, G. (2007a). Summarizing bbc rushes the informedia way. In *TVS '07: Proceedings of the international workshop on TRECVID video summarization*, New York, NY, USA. ACM. 48

Hauptmann, A., Yan, R., and Lin, W.-H. (2007b). How many high-level concepts will fill the semantic gap in news video retrieval? In *CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval*, pages 627–634, New York, NY, USA. ACM. 69

He, L., Sanocki, E., Gupta, A., and Grudin, J. (1999). Auto-summarization of audio-video presentations. In *Proceedings of the seventh ACM international conference on Multimedia*, pages 489–498, New York, NY, USA. ACM Press. 2, 50, 53, 71, 74, 84, 87, 89, 99

He, L., Sanocki, E., Guptra, A., and Grudin, J. (2000). Comparing presentation summaries: Slides vs. reading vs. listening. In *Proc. CHI'00*. 3

Heiman, G. W., Leighbody, G., and Bowler, K. (1986). Word intelligibility decrements and the comprehension of time-compressed speech. *Perception and Psychophysics*, vol. 40, no. 6:pp. 407–411. 35, 77

Herout, A., Beran, V., Hradis, M., Potúcek, I., Zemcík, P., and Chmelar, P. (2007). Trecvid 2007 by the brno group. In *TVS '07: Proceedings of the international workshop on TRECVID video summarization*, New York, NY, USA. ACM. 58

Huang, M., Mahajan, A. B., and DeMenthon, D. F. (2004). Automatic Performance Evaluation for Video Summarization. Technical Report LAMP-TR-114,CAR-TR-998,CS-TR-4605,UMIACS-TR-2004-47, University of Maryland, College Park. 91, 92

Hughes, A., Wilkens, T., Wildemuth, B. M., and Marchionini, G. (2003). Text or pictures? an eyetracking study of how people view digital video surrogates. In Bakker, E. M., Huang, T. S., Lew, M. S., Sebe, N., and Zhou, X. S., editors, *CIVR*, volume 2728 of *Lecture Notes in Computer Science*, pages 271–280. Springer. 3, 24, 25, 75, 89, 110

Hutchinson, B. D. (2004). An attention model of film viewing. *Journal of Moving Image Studies*. 20

Huurnink, B. and de Rijke, M. (2007). Exploiting redundancy in cross-channel video retrieval. In *MIR '07: Proceedings of the international workshop on Workshop on multimedia information retrieval*, pages 177–186, New York, NY, USA. ACM. 47

Ide, I., Yamamoto, K., Hamada, R., and Tanaka, H. (2001). An automatic video indexing method based on shot classification. In *In Systems and Computers in Japan*. 62

Iqbal, Q. and Aggarwal, J. K. (2002). Cires: A system for content-based retrieval in digital image libraries. In *in Invited Session on Content-based Image Retrieval: Techniques and Applications, 7 th International Conference on Control Automation, Robotics and Vision (ICARCV*, pages 205–210. 66, 106

Iqbal, Q. and Aggarwal, J. K. (2003). Feature integration, multi-image queries and relevance feedback in image retrieval. In *6th International Conference on Visual Information Systems (VISUAL 2003)*, pages 467–474. 106

Jaimes, A., Benitez, A. B., Jorgensen, C., and Chang, S.-F. (2000). Experiments in indexing multimedia data at multiple levels. *Idea Mart: Classification for User Support and Learning, ASIS SIG Classification Research Workshop*. 16, 17

Jaimes, A. and Chang, S. F. (2000). A conceptual framework for indexing visual information at multiple levels. In *SPIE Internet Imaging 2000*, pages 2–15. 17, 32

Jaimes, A., Jorgensen, C., Benitez, A. B., and Chang, S. F. (1999). Experiments for multiple level classification of visual descriptors contribution. *Contribution to ISO/IEC JTC1/SC29/WG11 MPEG99/M5593*. 16, 17, 63

JISC Digital Media (2006). *An Introduction to Metadata*. A JISC Advisory Servie. http://www.jiscdigitalmedia.ac.uk/crossmedia/advice/metadata-overview/. 60, 61, 62

Jones, S., Cunningham, S. J., Mcnab, R., and Boddie, S. (2000). A transaction log analysis of a digital library. *International Journal on Digital Libraries*, 3:152–169. 115

Kennedy, L. and Ellis, D. (2003). Pitch-based emphasis detection for characterization of meeting recordings. In *Proc. IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Virgin Islands. 33

Kim, J.-G., Chang, H. S., Kang, K., Kim, M., Kim, J., and Kim, H.-M. (2004). Summarization of news video and its description for content-based access. *International J. Imaging Syst. Technol*, 13(5):267–274. 57

Kulhavy, R., Stock, W., Woodard, K., and Haygood, R. (1993). Comparing elaboration and dual coding theories: The case of maps and text. Vol. 106:483–498. 31

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174. 139

Lass, N. and Leeper, H. (1977). Listening rate preference: Comparison of two time alteration techniques. *Perceptual and Motor Skills*, 44:1163–1168. 36

Lee, M. and Roskos-Ewoldsen, D. (2004). Subtitles, inferences, and movie comprehension: Predictions from the event index model. In *the Annual Meeting of the International Communication Association*, New Orleans, LA. 12, 20

Li, B., Pan, H., and Sezan, I. (2003). 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. 41, 48, 53

Li, J. and Wang, J. Z. (2006). Real-time computerized annotation of pictures. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 911–920, New York, NY, USA. ACM. 62, 65

Li, Y., Dorai, C., and Farrell, R. (2005). 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. ACM. 44, 125, 171

Li, Y., Zhu, Q., and Cao, Y. (2004). Automatic metadata generation based on neural network. In *InfoSecu '04: Proceedings of the 3rd international conference on Information security*, pages 192–197, New York, NY, USA. ACM. 33

Lie, W.-N. and Lai, C.-M. (2004). 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. 2, 52, 53, 56, 57, 79

Lienhart, R., Pfeiffer, S., and Effelsberg, W. (1997). Video abstracting. *Commun. ACM*, 40(12):54–62. 34, 53

Liu, Z., Zavesky, E., Gibbon, D., Shahraray, B., and Haffner, P. (2007). At&t research at trecvid 2007. In *TVS '07: Proceedings of the international workshop on TRECVID video summarization*, New York, NY, USA. ACM. 57

Lyman, B. and McDaniel, M. (1990). Memory for odors and odor names: Modalities of elaboration and imagery. 16:656–664. 29

Mani, I. and Maybury, M. (1999). *Advances in automatic text summarization*. Cambridge, MA: MIT Press. 174

Marchionini, G. (2006). Human performance measures for video retrieval. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 307–312, New York, NY, USA. ACM. 112

Marchionini, G., Geisler, G., and Brunk, B. (2000). Agileviews: A human-centered framework for interfaces to information spaces. In *Proceedings of the Annual Conference of the American Society for Information Science*, pages 271–280. 110

Marchionini, G., Song, Y., and Farrell, R. (2009). Multimedia surrogates for video retrieval: Toward combining spoken words and imagery. *Journal of Information Processing & Management*. http://dx.doi.org/10.1016/j.ipm.2009.05.007),. 7, 31, 75, 80, 87, 109, 110, 112, 119, 125, 127, 133, 140, 151, 174, 227

Martone, A. F., Taskiran, C. M., and Delp, E. J. (2004). Automated closed-captioning using text alignment. In *Proc. SPIE. Storage and Retrieval Methods and Applications for Multimedia Conference*, pages 108–116. 44

Massey, M. and Bender, W. (1996). Salient stills: process and practice. *IBM Syst. J.*, 35(3-4):557–573. 23

Mayer, R. and Moreno, R. (1998). A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory. 90(2):312–320. 31

Metadata Working Group (2009). *Guidelines For Handling Image Metadata*. lhttp://www.metadataworkinggroup.com/pdf/mwg_guidance.pdf. 61

Mihajlovic, V., Petkovic, M., Jonker, W., and Blanken, H. (2007). *Multimedia Retrieval*, chapter 10. Multimodal Content-based Video Retrieval, pages 271–294. Data-Centric Systems and Applications. Springer Berlin Heidelberg. 55

Minifie, F. (1974). *Time-Compressed Speech*, chapter Durational Aspects of Connected Speech Samples, pages 709–715. Scarecrow, Metuchen, NJ. 36

Mohamad Ali, N., Smeaton, A. F., Lee, H., and Brereton, P. (2009). Developing, deploying and assessing usage of a movie archive system among students of film studies. In *HCI International 2009 - 13th International Conference on Human-Computer Interaction*, San Diego, CA, USA. (in press). 74

Nagasaka, A. and Tanaka, Y. (1992). Automatic video indexing and full-video search for object appearances. In *Proceedings of the IFIP TC2/WG 2.6 Second Working Conference on Visual Database Systems II*, pages 113–127, Amsterdam, The Netherlands, The Netherlands. North-Holland Publishing Co. 45

Neuburg, E. (1978). Simple pitch-dependent algorithm for high quality speech rate changing. *Journal of the Acoustic Society of America*, 63(2):624–625. 36

Ngo, C.-w., Jiang, Y.-g., Wei, X., Wang, F., Zhao, W., Tan, H.-k., and Wu, X. (2007). Experimenting vireo-374: Bag-of-visual-words and visual-based ontology for semantic video indexing and search. In *TVS '07: Proceedings of the international workshop on TRECVID video summarization*, New York, NY, USA. ACM. 58

Nielsen, J. and Levy, J. (1994). Measuring usability: preference vs. performance. *Commun. ACM*, 37(4):66–75. 89

Omoigui, N., He, L., Gupta, A., Grudin, J., and Sanocki, E. (1999). Time-compression: systems concerns, usage, and benefits. In *CHI '99: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 136–143, New York, NY, USA. ACM. 38

Over, P., Kraaij, W., and Smeaton, A. (2005). Trecvid 2005: An introduction. In *Proc. TRECVID 2005*, pages 1–14. http://www.cdvp.dcu.ie/Papers/TRECVid2005_Overview.pdf. 3, 64

Over, P., Smeaton, A. F., and Kelly, P. (2007). 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. ACM. 57, 82, 99, 100

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Rinehart, and Winston. 26, 28

Paivio, A. (1972). Symbolic and sensory modalities of memory. In *M.E. Meyer (Ed.), The Third Western symposium on learning: Cognitive learning*. Bellingham, WA: Western Washington State College. 29

Paivio, A. (1975). Coding distinction and repetition effects in memory. In *G.H. Bower (Ed.), The Psychology of Learning and Motivation: Advances in Research and Theory*, volume 9. New York: Academic Press. 27

Paivio, A. (1986). *Mental representations: A dual coding approach.* Oxford: Oxford U. Press. 6, 26

Paivio, A. (2006). *Mind and Its Evolution: A Dual Coding Theoretical Approach.* Lawrence Erlbaum. 26, 28, 30

Paivio, A. and Begg, I. (1974). Pictures and words in visual search. *Memory & Cognition*, 3:515–521. 28

Paivio, A. and Csapo, K. (1973). Picture superiority in free recall : Imagery or dual coding? *Cognitive Psychology*, 5:176–206. 26

Panofsky, E. (1955). Meaning in the visual arts: meanings in and on art history. *Doubleday.* 13, 14, 17, 19, 32, 63

Panofsky, E. (1972). Studies in iconology: Humanistic themes in the art of the renaissance. *New York: Harper & Row.* 13

Peyrard, N. and Bouthemy, P. (2003). Motion-based selection of relevant video segments for video summarisation. *Multimedia and Expo, IEEE International Conference on*, 2:409–412. 45

Ponceleon, D., Amir, A., Srinivasan, S., Syeda-Mahmond, T., and Petkovic, D. (1999). Cuevideo: Automated multimedia indexing and retrieval. In *Proceedings of ACM MM '99.* 140

Ponceleon, D., Srinivasan, S., Amir, A., Petkovic, D., and Diklic, D. (1998). Key to effective video retrieval: effective cataloging and browsing. In *MULTIMEDIA '98: Proceedings of the sixth ACM international conference on Multimedia*, pages 99–107, New York, NY, USA. ACM. 65

Ponech, T. (1997). Visual perception and motion picture spectatorship. *Cinema Journal*, 37(1):85–100. 18, 19, 32

Pryluck, C. (1976). *Sources of Meaning in Motion Pictures and Television.* Amo Press. 24

Rasmussen, E. (1997). Indexing images. *Annual Review of Information Science and Technology*, 32:169–196. 62

Ratakonda, K., Sezan, I. M., and Crinon, R. J. (1999). Hierarchical video summarization. In *Proc. SPIE Conf. Visual Communications and Image Processing*, volume 3653, pages 1531–1541, San Jose, CA. 45

Sato, T., Kanade, T., Hughes, E. K., Smith, M. A., and Satoh, S. (1999). Video ocr: indexing digital new libraries by recognition of superimposed captions. *Multimedia Syst.*, 7(5):385–395. 66

Satoh, S., Nakamura, Y., and Kanade, T. (1999). Name-it: Naming and detecting faces in news videos. *IEEE MultiMedia*, 6(1):22–35. 69

Schooler, J. and Engstler-Schooler, T. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22(1):36–71. 28

Shatford, S. (1986). Analyzing the subject of a picture : a theoretical approach. *Cataloging & Classification Quarterly*, 6(3):39–62. 15, 17, 32, 63

Shepard, R. N. (1967). Recognition memory for words, sentences and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6:156–163. 28

Smeaton, A. F., Over, P., and Kraaij, W. (2004). Trecvid: evaluating the effectiveness of information retrieval tasks on digital video. In *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, pages 652–655, New York, NY, USA. ACM. 2, 23, 96

Smeaton, A. F., Over, P., and Kraaij, W. (2006). Evaluation campaigns and trecvid. In *MIR '06: Proceedings of the 8th ACM international workshop on Multimedia information retrieval*, pages 321–330, New York, NY, USA. ACM. 64, 105, 106

Smith, G. (2007). *Tagging: People-powered Metadata for the Social Web (Voices That Matter)*. New Riders Press, Thousand Oaks, CA, USA. 70

Smith, M. A. and Kanade, T. (1997). Video skimming and characterization through the combination of image and language understanding techniques. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 775, Washington, DC, USA. IEEE Computer Society. 55

Snoek, C. G. M. and Worring, M. (2005). Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25:5–35. 62, 70

Song, Y. and Marchionini, G. (2007). 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. ACM. 7, 23, 25, 31, 71, 75, 79, 80, 84, 87, 89, 110, 112, 119, 120, 151, 171, 183, 219, 220, 227

Spence, R. (2002). Rapid, serial and visual: A presentation technique with potential. *Information Visualization*, 1(1):13–19. 140

Taskiran, C., Cuneyt, M., Amir, A., Ponceleon, D., and Delp, E. (2002). Automated video summarization using speech transcripts. *Proc. SPIE Conference on Storage and Retrieval for Media Databases*, 4676:371–382. 39, 41, 42, 46, 89

Taskiran, C., Pizlo, Z., Amir, A., Ponceleon, D., and Delp, E. J. (2006). Automated video program summarization using speech transcripts. *IEEE Transactions on Multimedia*, 8(4):775–791. 42, 43, 44, 71, 82, 84, 87, 88, 90, 99

Taskiran, C. M. and Bentley, F. (2007). Automatic and user-centric approaches to video summary evaluation. In *Proceedings of the SPIE Conference on Multimedia Content Access: Algorithms and Systems*, volume 6506, San Jose, CA. 89, 92

Thompson, V. and Paivio, A. (1994). Memory for pictures and sounds: Independence of auditory and visual codes. *Canadian Journal of Experimental Psychology*, 48:380–398. 6, 29

Turner, J. M. (1995). Comparing user-assigned terms with indexer-assigned terms for storage and retrieval of moving images: research results. In *Proceedings of the 58th Annual Meeting of the American Society for Information Science*. 16

Uchihashi, S., Foote, J., Girgensohn, A., and Boreczky, J. (1999). Video manga: Generating semantically meaningful video summaries. In *ACM Multimedia'99*, pages 383–392. ACM Press. 34, 45

Valenzise, G., Gerosa, L., Tagliasacchi, M., Antonacci, F., and Sarti, A. (2007). Scream and gunshot detection and localization for audio-surveillance systems. In *Advanced Video and Signal Based Surveillance, 2007. AVSS 2007. IEEE Conference on*, pages 21–26. 67

Van Gemert, J. C., Snoek, C. G. M., Veenman, C. J., and Smeulders, A. W. M. (2006). The influence of cross-validation on video classification performance. In *MULTIMEDIA '06: Proceedings of the 14th annual ACM international conference on Multimedia*, pages 695–698, New York, NY, USA. ACM. 47

Wactlar, H. D. (2000). Informedia - search and summarization in the video medium. In *Proc. IMAGINA 2000 Conf.* 73

Wactlar, H. D., Kanade, T., Smith, M. A., and Stevens, S. M. (1996). Intelligent access to digital video: Informedia project. *Computer*, 29(5):46–52. 38

Wang, R., Naphade, M. R., and Huang, T. S. (2001). Video retrieval and relevance feedback in the context of a post-integration model. In *Proc. IEEE 4 th Workshop on Multimedia Signal Processing*, pages 33–38. 69

Weis, R. and Katter, R. (1967). Multidimensional scaling of documents and surrogatess. Report No. SP-2713. Santa Monica. 85

Wertheimer, M. (1961). *Experimental Studies on the Seeing of Motion. Classics in Psychology*. Newbury Park, CA: Sage Publications. 12

Wildemuth, B. M., Marchionini, G., Wilkens, T., Yang, M., Geisler, G., Fowler, B., Hughes, A., and Mu, X. (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. 2, 4, 22, 24, 31, 75, 89, 108, 110, 111, 112

Wildemuth, B. M., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., and Gruss, R. (2003). How fast is too fast?: evaluating fast forward surrogates for digital video. In *JCDL '03: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries*, pages 221–230, Washington, DC, USA. IEEE Computer Society. 2, 22, 37, 87, 99, 109, 111, 112, 238

Yahiaoui, I., Merialdo, B., and Huet, B. (2003). Comparison of multiepisode video summarization algorithms. *EURASIP J. Appl. Signal Process.*, pages 48–55. 34, 82, 88

Yang, J. and Hauptmann, A. G. (2006). 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. ACM. 47

Yang, M. (2005). *An exploration of users' video relevance criteria*. Ph.d. dissertation, The University of North Carolina at Chapel Hill, United States. Publication No. AAT 3190335. 18, 114

Yang, M., Wildemuth, B. M., Marchionini, G., Wilkens, T., Geisler, G., Hughes, A., Gruss, R., and Webster, C. (2003). Measures of user performance in video retrieval research. Sils technical report, University of North Carolina at Chapel Hill. 2, 22, 111, 112, 150, 151

Youngblood, G. (1970). *Expanded Cinema*. E.P. Dutton. 12

Zhang, H., Kankanhalli, A., and Smoliar, S. W. (1993). Automatic partitioning of full-motion video. *Multimedia Syst.*, 1(1):10–28. 45

Zhang, Y., Jansen, B., and Spink, A. (2008). Time series analysis of a web serach engine transaction log. *Information Processing and Management*, page 597C600. 115

Zhou, X. S., Zillner, S., Moeller, M., Sintek, M., Zhan, Y., Krishnan, A., and Gupta, A. (2008). Semantics and cbir: a medical imaging perspective. In *CIVR*

'08: Proceedings of the 2008 international conference on Content-based image and video retrieval, pages 571–580, New York, NY, USA. ACM. 65