# Deciphering visual gist and its implications for video retrieval and interface design

**Meng Yang**
Open Video Project
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360
yangm@ils.unc.edu

**Gary Marchionini**
Open Video Project
University of North Carolina at Chapel Hill
Chapel Hill, NC 27599-3360
march@ils.unc.edu

**ABSTRACT**
How do people make sense of a video based on viewing a few frames of that video? What elements constitute the "visual gist" in their minds? Answers to these questions will give implications to both content-based video retrieval and the interface design (e.g., key-frame selection) of digital video libraries. A preliminary study was conducted to unravel the issues and 45 subjects participated in the study. After viewing a fast forward surrogate, the subjects were asked to choose pictures which they thought would "belong to" the video. And they were also asked to think aloud during their selection processes. Nine visual gist attributes (e.g., people, objects and actions) were generated using the grounded theory method and their frequencies were also compared and analyzed.

**Author Keywords**
Visual gist understanding, video retrieval, video surrogate, fast forward, user studies.

**ACM Classification Keywords**
H5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

**INTRODUCTION**
How do people make sense of a video based on viewing a few frames of that video? What do people remember after they view a single picture or other surrogate? In particular, what attributes constitute the "visual gist" in their minds? Answers to these questions will influence the interface design for digital video libraries.

Designing effective browsing techniques is a key issue for effective video retrieval [11]. Empirical evidence shows that video information needs sometimes are hard to express in words, but are easily clarified when the picture/video clips are seen and also users feel that browsing often required less effort and time than formulating a refined query [8]. Among those browsing techniques, video surrogation plays an important role, and varied types of video surrogates such as storyboards (arrays of frames), slide shows and fast forwards have been created and their use studied (author cite here). However, important questions remain: how to select a "good" keyframe, or a "good" key clip to represent the whole videos? What are the "salient" scenes in the videos to be chosen? There are no firm rules to follow during these filtering processes. User-centered design suggests that understanding human needs, behavior, and expectations is the best way to drive design. To this end, we have conducted several studies of how people use video surrogates to make sense of video. In this paper, we present the qualitative results of a study that focused on the effectiveness of fast forward surrogates. We operationalized this investigation by focusing on the features people identified as useful, specifically, we aimed to determining what would be the mostly frequently utilized attributes to aid in determining visual gist after watching a fast forward surrogate.


*CHI 2005*, April 02-07, 2005, Portland, Oregon, USA.
ACM 1-58113-454-1/02/0004.

**RELATED LITERATURE**
Empirical evidences have shown that people have superior memory for pictures over words [6, 12]. It was thus concluded by [10] that the visual code is qualitatively superior to the verbal code as a mediator of recall. Then the following question becomes: What makes the pictures so memorable? In practice, the details of the image are not well remembered [7]. A phenomenon that has been referred as "change blindness" [13] illustrates how badly people can recognize differences between two versions of the same scene. This is also explored in movies [5], where cuts between views render subjects insensitive to changes in clothing, props or even the identity of actors. [11] explains this phenomenon by noting that "observers do not remember the scene per se. Rather, they remember the gist of the scene". However, it is apparent that there is no "consistent" gist understanding between different people, since people might remember quite different things from the same picture.

If people can remember the gist of the scene, what does the gist include? First, the objects in the scene should at least be a part of the gist [15]. The selection of objects is partly governed by attention. [1] found that participants often paid more attention to keyframes selected from videos with one of the following features: text in pictures, interaction information, symbols, novelty, emotion and people. In addition, [4] discovered that object, people, social status, color, body part, location, specific detail, and activity were the most frequently mentioned attributes when she asked students to write down individual descriptions of six color images. However, studies also show that "a gist is more than a list" [15]. Another component coded in the visual gist could be the relationships between objects: for instance, the object relations and spatial layout. Additionally, there could be also some information coded in visual gist that is not available in the pictures: for instance, people's imaginations, or associated ideas [7]. Similar gist understanding in videos has also been explored. [14] consider visual gist as a combination of topicality, narrative structure, and visual style. While this concept needs additional clarification, participant comments clearly indicated that they formed a more holistic view of gist, beyond topic and narrative.

## METHODOLOGY

A visual gist comprehension task (see Figure 1) was embedded in a study conducted in Spring 2002 [15], which aimed to study the optimal speed of fast forward video surrogate. The participants first watched a fast forward surrogate and then were asked to complete six tasks: object recognition (textual), object recognition (graphical), action recognition, linguistic gist comprehension (full text), linguistic gist comprehension (multiple choice), and visual gist comprehension (For more details about these six measures and the study performance, please see [15]). In the visual gist comprehension task, they were asked to select pictures that "belonged" in the video represented by the surrogate. No title, abstract or other types of linguistic cues were presented to the participants, only the fast forward surrogates. The frames presented in this task did not appear in the surrogate they watched before, thus this visual gist comprehension task was different from another task also used in this study ---- the object recognition task, which asked the subject to select pictures they had seen in the surrogate. In the study, the participants did indicate that they could differentiate between the object recognition task and the visual gist selection task. The stimuli used in this visual gist comprehension task were 12 still images (see Figure 1 for an example). Six of them were selected from the target video (but had not been seen in the fast forward surrogate). Of the remaining six keyframes, three were selected from a different video of a similar style and three were selected from a different video of a different style. The participants were also asked to "think aloud" during the visual gist comprehension task and the process was videotaped for further analysis.

More details about the videos and surrogates used in this study are as follows: four fast forward speeds (1:32, 1:64, 1:128 and 1:256) were examined for four video clips. Each of the 45 participants in this study interacted with four video surrogates. The four videos and four fast forward rates were counterbalanced so that each video/surrogate speed combination was approximately equally represented. The four videos used in this study were selected from the Open Video Project repository (www.open-video.org). Two factors were considered in the video selection: color and video structure, with two documentaries and two narrative videos, two color and two black and white videos.

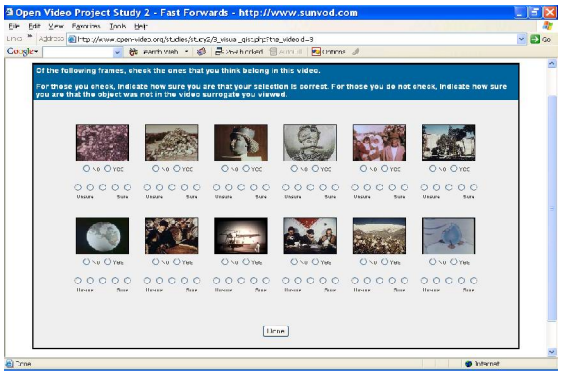


**Figure 1. Visual gist comprehension task**

## RESULTS

Although the participants often complained that the fast forward surrogates were very fast, their performance on the visual gist task was generally good (70%). Using the grounded theory method, two researchers coded participant utterances from the visual gist comprehension task and identified nine visual gist attributes. Brief explanations and two examples of user statements follow for each attribute.

### *Object*

The objects contained in the fast forward surrogates such as cars or bridges were the most frequently mentioned criteria the participants used to judge the "relevance" of the pictures. Examples of the participants' citations include: "This could be there because there were a lot of stones, rocks and groves"; "These are the old fashion cars they were showing".

### *People*

In her studies, [1] identified that people were one of the key visual-attention elements participants reported in sense making, which is consistent with the results from this study. People were the second most frequently mentioned attributes, often using some specific

characteristic such as age, gender, dress, or emotion as selection criteria: “These people seem a little too dressed up”; “This little boy was too unhappy to be in there”

*Setting/environment*

After watching the fast forward surrogate, in spite of the fast forward rate, participants typically got a general impression about what the context or environment was in the video, such as indoor vs. outdoor scenes, big city vs. rural area scenes, night vs. daytime scenes. Settings were among the top most frequently mentioned criteria. “It didn't show any airport, airplanes, just downtown, business day sort of environments.”; “I would say this belongs because it is about a beach and there is water there.”

*Action/activities/events*

Compared to still image surrogates such as storyboards, the fast forward surrogate has the advantage of giving users more information about the actions or events in the original video. Participants often used those actions or events related to the characters to make judgments, which mostly occurred in the two narrative videos. “Looks like she was changing the baby's diaper, I don't remember seeing that”; “Sitting at a dinner, I don't remember anybody eating.”

*Theme/topic*

Sometimes, the participants made judgments by more general topical or theme issues. Sometimes they might not know exactly what the theme of the video was, but they could get some general impressions, for instance, about Middle East, about history, about recreation, and about courtship. Examples of the partcipants’ citations are: “These people are water skiing. The video was like recreation or leisure-time related”; “It seemed to be no alive thing, all was about history”;

*Time/period*

The participants also considered the time period of the videos or of the objects in the videos as a criterion for relevance judgments. “This one (image) could be (there) but only because of the time period of the car”; “This (picture) is too modern for that period”.

*Geographical location*

This criterion mostly occurred in the documentary videos. “Yes, I think this one looks like an Egyptian environment”. “This looks like dolls in American, not there”.

*Plot*

In the narrative videos, the participants would infer a plot to determine whether some object or person was present. “…what was going on in the plot, what made sense in these shots, the characters involved, the places they were, just comparing them to the plots that I know that I did see”; “I don't think there was any hands-on work. It didn't fit into the plot. I am pretty sure about that”.

*Visual perception*

Visual perception such as color, video type, and video qualities were often mentioned. “I distinctly saw a green car”; “This one, the video quality looks good, but I don't think that one belongs”;

## DISCUSSION

Nine primary visual gist cues were found in this study. Among them, “people” and “object” were the two most frequently cited criteria by the participants. The “people-orientation” fact was actually consistent with the findings in [1, 9]. The action and activities related with those

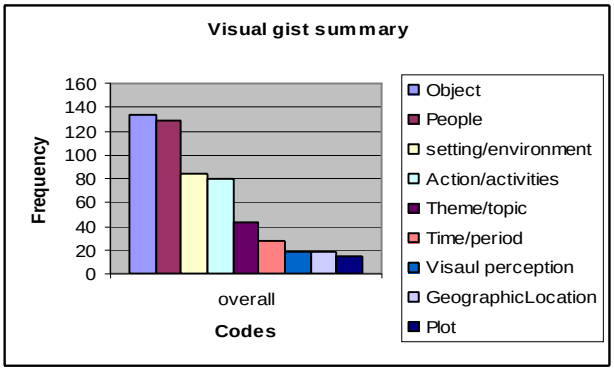


**Figure 2. Attribute frequency**

characters were also commonly recognized by the participants, especially in the narrative videos. Perhaps the characters were more consistent in the video and thus their actions/activities were critical to the comprehension of the plots or the themes. Clearly, human form and face detection techniques and interfaces that support users in specifying such queries are important.

The objects contained in the video were also important visual gist cues. In particular, there were usually some “stand-out” objects which were recognized by most of the participants, such as cars in the video “On the Run” and the “Ferris wheel” in the video “Coney Island”. Frequency of occurrence of the objects may be an important factor, in which case techniques for information retrieval that leverage term frequencies may be useful for video retrieval interfaces. Alternatively, it may be the case that some objects draw more attention (e.g., the Ferris wheel). For these cases, manually identifying the most salient “stand-out” objects may be required.

Another important attribute was settings/environments, which provide the context for the characters and their activities. More importantly, the participants constantly used settings to infer the existence of the people and objects, or the people’s activities, etc. Themes/topics were also sometimes mentioned by the participants to judge the

relevance of the pictures. Since no linguistic metadata were provided in this study, it might be hard for the participants to infer the exact theme of the videos, but they did get some impressions about the topic, such as about the Middle East, about history, about recreation, and about courtship.

### Implications for user interfaces

Many video surrogates are generated by selecting "key" frames or "key" segment to represent the original videos. Results from this preliminary study will provide guidelines and criteria for keyframe or key segment selection. Current automatic keyframe generation techniques (e.g., [2]) are good at selecting dissimilar keyframes to represent the whole video and thus reduce redundancies. However, few user studies have been conducted to test the applicability or the appropriateness of these methods. These methods focus on the physical attributes of the video content, not on users' attention and understanding.

The results here suggest that user interfaces include a range of visual surrogates that afford people several cues to build visual gist. If storyboards are used, designers might choose a few human figures and objects, and one each of the other attributes if they appear in the video. This work also demonstrates that people can be highly effective at using visual surrogates to make sense of a video and designers should augment textual metadata with visual cues in video retrieval interfaces.

## FUTURE WORK

Nine visual gist attributes were generated in this preliminary study and they provide empirical implications for user interfaces for digital video libraries. However, due to the exploratory nature of this study, more work is required to evaluate and develop the visual gist attribute set. Visual gist understanding can be affected by many factors such as video genre, video surrogate types and user tasks. In this study, only documentary and narrative videos were used, and the visual gist cues may be different for other types of videos such as news and sports videos. Additionally, the participants were not given any situated task when they watched the fast forward surrogates and their visual gist understanding might differ if they had been assigned some search tasks or used some other types of video surrogates. Future studies will address these issues.

## ACKNOWLEDGMENTS

This work was partially supported by NSF IIS 0099638. We also thank our study participants.

## REFERENCES

1. Ding, W., Marchionini, G., Soergel, D. Multimodal Surrogates for Video Browsing. In: *Proc. of Digital Libraries '99:* 85-93
2. Dufaux, F. Key frame selection to represent a video. *ICIP 2000. Vol, II.* p. 275-278
3. Grodal, T. Emotions, Cognitions, and Narrative Patterns in Film. In *Passionate views : film, cognition, and emotion*, edited by Plantinga, C. & Smith, G. M. 1999, 127-145
4. Jorgensen, C. Image attributes in describing tasks: an investigation. *Information Processing & Management, 34(2/3),* 1998, 161-174
5. Levin, D. T. & Simons, D. J. Failure to detect changes to attended objects in motion pictures. *Psychological Bulletin, 4, 1997,* 501-506
6. Lieberman, L. R., & Culpepper, J. T. Words versus objects: comparison of free verbal recall. *Pscychol. Rep. 17,1965,* 983-988
7. Mandler, J & Ritchey G. H. Long term memory for pictures. *Journal of Experimental Psychology [Human learning and memory], 3,* 1977. 386-396
8. Markkula, M. and Sormumen, E. Searching for photos: journalists' practices in pictorial IR, *The Challenge of Image Retrieval Research Workshop,* 1998.
9. Massey, M.; Bender, W. Salient stills: process and practice. *IBM Systems Journal*, 35(3-4), 1996, 557-573.
10. Paivio, A. & Csapo, K. Picture superiority in free recall: Imagery or dual coding? *Cognitive Psychology, 5*, 1973, 176-206
11. Ponceleon, D., Srinivasan, S., Amir, A., Petkovic, D., & Diklic, D. Key to effective video retrieval: Effective cataloguing and browsing. In *Proc. ACM Multimedia*, 1998, 99-107.
12. Shepard, R. N. Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, 6, 1967, 156-163
13. Simons, D. J. & Levin, D. T., Change blindness. *Trends Cognitive Science, 1,* 1997, 261-267
14. Wildemuth, B. M., Marchionini, G., Wilkens, T., Yang, M., Geisler, G., Fowler, B., Hughes, A., & Mu, X. Alternative surrogates for video objects in a digital library: users' perspectives on their relative usability. *Proc., the European Conference on Digital Libraries (ECDL), 2002, 493-507*
15. Wildemuth, B. M., Marchionini, G., Yang, M., Geisler, G., Wilkens, T., Hughes, A., & Gruss, R. How fast is too fast? Evaluating fast forward surrogates for digital video. *Proc., Joint Conference on Digital Libraries, 2003,* 221-230
16. Wolfe, J.M. Visual memory: what do you know about what you saw? *Current Biology, 8(9),* 1998, 303-304.