Content-based Video Retrieval: Does Video's Semantic Visual Feature Matter?

Xiangming Mu
University of Wisconsin-Milwaukee
3210 N. Maryland Ave., Milwaukee, WI53211
(414)-229-6039, mux@uwm.edu

ABSTRACT
A new shot level video browsing method based on semantic visual features (e.g., car, mountain, and fire) is proposed to facilitate content-based retrieval. The video's binary semantic feature vector is utilized to calculate the score of similarity between two shot keyframes. The score is then used to browse the "similar" keyframes in terms of semantic visual features. A pilot user study was conducted to better understand users' behaviors in video retrieval context. Three video retrieval and browsing systems are compared: temporal neighbor, semantic visual feature, and fused browsing system. The initial results indicated that the semantic visual feature browsing was effective and efficient for Visual Centric tasks, but not for Non-visual Centric tasks.

Categories and Subject Descriptors
H.3.3 [**Information Search and Retrieval**]: Relevance feedback

General Terms
Algorithms, Design, Experimentation,

Keywords
Content-based, video browsing, video retrieval, user interface

1. INTRODUCTION
Browsing technologies are supported in video retrieval to augment text based query search, particularly when exact queries are hard to form. Browsing usually follows a search operation to pinpoint the correct matches.

For shot level content-based retrieval (where a shot represents a series of consecutive frames with no sudden transition), temporal neighbor browsing is the most common navigation method [2]. Temporal neighbor browsing allows users to navigate around the selected sample shot keyframe (a single frame that is representative of the content of a shot) from a text query returns. Potential relevant shots may appear just before or after the sample one due to the asynchronous of the visual content and its related transcript. Mezaris et al.[3] noted that a visual similarity re-search using a sample picked keyframe is a good design for retrieval. Various visual features including color histograms, text, camera movement, face detection, and moving objects can be utilized to define the similarity. As a result, a function like "finding similar shots like this" can be supported. In this project, we refer to this function as "visual similarity browsing." Visualization technology such as visual networks can be used to enhance visual similarity browsing [2], but the effectiveness needs to be further verified.

One limitation of these technologies is that no functions are provided to support users to look for specific visual objects, such as people, car, map, etc., even though automatic semantic visual feature extraction technologies [1] have been developed to make these metadata easily obtainable. Targeting this problem, a new semantic visual feature browsing technology is developed in this paper.

SIGIR'06, August 6–11, 2006, Seattle, Washington, USA.
ACM 1-59593-369-7/06/0008.

2. SEMANTIC VISUAL FEATURE
Video's semantic visual feature is defined as a high level semantic description of video content, such as indoor/outdoor, people, car, and explosion. Naphade et al. [4] proposed a 39-feature lightweight ontology for TRECVID project (which was also used in our study). This ontology has a two-layer structure: the top layer includes seven categories: Program category, People, Objects, Setting/Scene/Site, Activities, Events, and Graphics; the second layer contains sub-categories for further classification. For instance, under the top layer category vehicle, the sub-categories include airplane, car, bus, truck, and boat/ship.

Semantic visual feature browsing allows users to navigate around shots that have similar visual features of a selected sample shot. For instance, to search for shots with the face of "Condoleezza Rice," a list of "similar" shots that share or partly share the features of "politics, face, person, government leader, police/private security personnel" are retrieved for more matches.

The selection of the similar shot keyframes are based on the score of "feature similarity". In the study, each keyframe F_i has a 39 dimension binary feature vector $F_i = (f_{i1,} f_{i2,} f_{i3,...,} f_{i39})$ based on the ontology proposed by Naphade et al. [4], and

$$f_{ij} = \begin{cases} 0 & does\,not\,have\,feature j \\ 1 & has\,feature j \end{cases} \qquad j=1,2,\ldots 39$$

For a selected keyframe $F_{s,}$ the feature similarity d_{sj} between F_i and another keyframe F_j is

$$d_{sj} = \left| F_s \bullet F_j \right| = \sqrt{\sum_k (f_{sk} \bullet f_{jk})^2} \qquad k=1,2,\ldots,39$$

As a result, a full list of semantic visual feature similarity index for a selected keyframe F_s will be

$$D_s = (d_{s1,} d_{s2,} d_{s3,...,} d_{sm})$$

where m is the total number of keyframes in the collection.

In practice, usually only top n elements (or none zero elements) in the index will be utilized to support semantic visual feature browsing. In our study, n was defined as six.

3. SYSTEM USER INTERFACE

A new content-based video retrieval and browsing system was developed as a research platform to examine the effectiveness video browsing technologies. The system supports two types of browsing: temporal neighbor browsing and semantic visual feature browsing.

Figure 1 is the main interface of the system. On the top part (part A) a traditional text input field is provided for text-based query. In our study the videos' transcripts were utilized for text-based retrieval.

Figure 1: User interface of the video browsing system

In the middle (part B and C) is the result panel. Video transcript of a selected shot is displayed in Part B, while Part C shows the matched shot keyframes in storyboard style.

At the bottom of the interface (Part D) is a browsing panel where two browsing methods are supported. After performing a text query, users can subsequently proceed with further navigation in this area to find more matches. A tabbed layout is adopted to facilitate users switching among browsing methods. "TEMPORAL" tag will lead to temporal neighbor browsing and "FEATURE" tag will go to the semantic visual feature browsing. All the neighboring or similar frames will be displayed in the same size as the sample, which is highlighted in the middle of the filmstrip.

4. PILOT USER STUDY

A pilot user study was conducted to evaluate the effectiveness of the semantic visual browsing technology. Two types of video searching tasks were selected: *Visual centric tasks* (VCT) focus on visual features of a keyframe and *Non-visual centric tasks* (NCT) focus on non-visual features of a keyframe. The data was obtained from the TRECVID 2005 data collection, including about 86 hours of news videos (137 segments with average duration of about half an hour). The semantic visual features were collaboratively created by TRECVID 2005 project participants. Three types of retrieval systems were compared: *Temporal neighbor* (TN), *Semantic visual feature* (SF), and *Fused* (FU) browsing system. The Fused browsing system allows users to use both the temporal neighbor and semantic visual feature browsing functions to aid retrieval. Figure 1 is the screenshot of the Fused system interface (Temporal and Feature tabs are located on the left top of the browsing filmstrip in area D).

5. RESULTS AND FUTURE WORK

Six volunteer participants from multiple majors and programs of campus participated in the study. Initial data for the effectiveness (indicated by the users themselves) and efficiency (average time spent on tasks) of the three systems are listed below.

Table 1: Average effectiveness (1-5 scale with 5 very effective)

Tasks	TN	SF	FU
VCT	4.7	**4.7**	4.3
NCT	4.7	**2.3**	4.0

Table 2: Average efficiency (seconds used to complete the task)

Tasks	TN	SF	FU
VCT	81.7	**57.7**	90
NCT	126.7	**336.7**	55.0

We found that Semantic Visual Feature (SF) was very effective and efficient for Visual Centric tasks (VCT), but not for Non-Visual Centric Tasks (NCT). The participants also described the system as easy to learn and manipulate. One user, however, complained about the slow response of the system.

In the future a large-scale usability study based on an improved browsing system and evaluation plan will be conducted to validate and further explore the relationships between the browsing technologies and video search tasks. In addition, we will consider evaluating the system with different video genres.

6. ACKNOWLEDGMENTS

I am grateful to Dr. Dietmar Wolfram and Dr. Wooseob Jeong for their valuable comments about the pilot user study design.

7. REFERENCES

[1] Chang,S.-F., Hsu, W., Kennedy, L., Xie, L. , Yanagawa, A. , Zavesky, E., and Zhang, D.-Q., 2005, Columbia University TRECVID-2005 Video Search and High-Level Feature Extraction, in Proceedings of TRECVID2005

[2] Heesch, D., Howarth, P., Magalhaes, J., May, A., Pickering, M., Yavlinsky, A., and ruger, S. (2004). Video retrieval using search and browsing. In proceedings of TRECVID2004.

[3] Mezaris, Y., Doulaverakis, H., Herrmann, S., Lehane, B., O'Connor, N., Kompatsiaris, I., and Strintzis, G. M. (2004). Combining textual and visual information processing for interactive video retrieval: SCHEMA's participation to TRECVID2004. In proceedings of TRECVID2004 program.

[4] Naphade, R. M., Kennedy, L., Kender, R. J. , chang, S., Smith, R. J. , Over, P., and Hauptmann, A. (2005). Alight scale concept ontology for multimedia understanding for TRECVID 2005.