Semantic Visual Features in Content-based Video Retrieval

Xiangming Mu

School of Information Studies, University of Wisconsin-Milwaukee P.O. Box 413, Milwaukee, WI 53201. Tel: (414) 229-4707; Fax: (414) 229-6699

A new semantic visual features (e.g., car, mountain, and fire) navigation technology is proposed to improve the effectiveness of video retrieval. Traditional temporal neighbor browsing technology allows users to navigate temporal neighbors of a selected sample frame to find additional matches, while semantic visual feature browsing enables users to navigate keyframes that have similar features to the selected sample frame. A pilot evaluation was conducted to compare the effectiveness of three video retrieval designs that support 1) temporal neighbor browsing; 2) semantic visual feature browsing; and 3) fused browsing which is a combination of both temporal neighbor and semantic visual feature browsing. Two types of searching tasks: visual centric and non-visual centric tasks were applied. Initial results indicated that the semantic visual feature browsing system was more efficient for non-visual centric tasks.

Introduction

Access to digital video from news sources such as CNN, MSNBC, or ABC has become commonplace. To make digital multimedia resource discovery and search more convenient, multimedia digital libraries are being developed for research and education. Increasingly, students or instructors are consulting video collections in search of video shots within larger video “documents” to be used in their projects or lectures. Viewing all videos in full length to find the desired video shots may be feasible for a small collection, but can be very time intensive for a large collection. The ability to search within individual videos, much in the same way that full text searching allows users to search for content instead of their bibliographic surrogates, would greatly increase access to video content.

Most of current video retrieval systems are still text-based. Video collections in these systems are usually searched in the same way as searching for text documents. Video

related text descriptions such as title, author, abstracts, etc. are indexed to facilitate these types of retrieval. For instance, the Vanderbilt Television News Archive database includes over 705,000 records of indexed video summaries for video searching (http://tvnews.vanderbilt.edu/tvn-database-info.pl). One disadvantage for this approach, however, is that fast-growing volumes of digital collections volumes make this manually indexing very labor intensive.

Content-based video retrieval focuses on visual descriptions about the video. For instance,

> A student looks for "a shot showing a lion chasing a goat" that can be used in his final project
> A instructor looks for "a shot showing the path of hurricane Katrina" that can be illustrated in his lecture
> "Shots of George W. Bush entering or leaving a vehicle (e.g., car, van, airplane, helicopter, etc) (he and vehicle both visible at the same time)"- –Topic 159 from TRECVID 2005

Recent research on content-based video retrieval indicated that initially performing a text-based query and subsequently proceeding with neighbor or visual similarity browsing proved to be an effective retrieval strategy (Wildemuth et al., 2003; Heesch et al., 2004; Mezaris et al., 2004 ; Amir et al., 2005). Human beings are usually good at pattern recognition through navigation. A retrieval system supporting navigation functions would provide users additional means for content related searching tasks.

In this paper we propose a new video content browsing technique: semantic visual feature browsing. Our purpose is to evaluate its effectiveness as compared to traditional temporal neighbor browsing technique for two types of retrieval tasks: visual centric tasks and non-visual centric tasks. After the introduction of related research, a description of the semantic visual feature browsing algorithm will be given. The user interface of a prototype web-based video retrieval system that supports semantic visual feature browsing will be then illustrated. Finally, the methodology of a pilot user study and some initial results from the study will be presented, followed by a brief discussion.

Related Research

Video retrieval in the context of a digital library has only recently begun to be studied from a research perspective. Gauch et al.'s study (1994) was among the earliest to address the challenges of digital video retrieval in a digital library environment, particularly for content-based retrieval. More recently, Marchionini and Geisler (2002) described the structure and content of the Open Video Digital Library (http://www.open-video.org/)

project to study different issues that affect the design, development, and use of digital video and to serve as a test bed for future video retrieval research. There is clearly interest in user access to digital video content. For example, Chen and Choi (2005) reported that more than 80% of participants in a study of analog video usage by college students would be interested in accessing videos online if they were available.

Video can be an audio-and video-centric genre (Li et al 2000). Automatic Speech Recognition (ASR) technology has been developed to turn audio into text (Christel et al., 1998) and to provide textual description of the video content. Even though the quality of the ASR transcript is usually not as good as the human generated video description, they are still the primary data resource for shot level video retrieval systems (Mezaris et al., 2005; Wildemuth et al., 2004; Amir et al., 2004; Heesch et al., 2004; Cooke et al., 2004).

In video retrieval, various browsing technologies are widely supported to augment text based query search, in particular when exact queries are hard to form (Carmel et al 1992). This may be because human beings are good at rapidly finding patterns, recognizing objects, generalizing or inferring information from limited data, and making relevance decisions (Helander, 1998; Shneiderman, 1998). Amir et al. (2005) indicated that a search can be defined as a global operation over an entire collection, while browsing, which usually follows a search operation, operates on the results of the search to pinpoint the correct matches.

Houten et al. (2000) described five types of general browsing/navigation models in general: *Model navigation* (e.g., storyboard video surrogates) presents all video information on one-page; *Hierarchical navigation* (Zhang et al., 1995) structures video surrogates in a hierarchical tree; *Relational navigation* groups small video segments into clusters; *Sequential navigation* (or neighbor browsing) allows navigation in the temporal space; and *Conceptual navigation* usually supports metaphor to aid browsing.

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 (Heesch et al., 2004; Wildemuth et al., 2003). 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 (Christel, et al., 1998).

Mezaris et al. (2004) 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 (Houten et al., 2000). As a result, a function like “finding similar shots like this”

can be supported. In this project, we refer this function as “visual similarity browsing”. Visualization technology such as visual networks can be used to enhance visual similarity browsing (Heesch et al., 2004), but the effectiveness needs to be further verified.

Research also revealed that the performances of video retrieval are dependent on search tasks (Carmel et al., 1992). Heesch et al. (2004) found that network browsing was particularly helpful for relatively hard queries where low-level physical features (e.g., color, texture, and layout) were less informative. Yang et al. (2004) demonstrated that the retrieval recall and precision between two system designs, consisting of transcript based systems (transcript only and transcript + high-level features) and feature-based system (high-level feature only), were directly linked to the search tasks. Two types of tasks were defined by the author in their study, generic topic related tasks (e.g., a kind of person, object, event, action, and geographic location) and specific topic related tasks (e.g., named person, object, event, action, and geographic location).

Semantic Visual Feature Browsing

Video’s semantic visual feature is defined as a high level semantic description of video content, as opposed to low-level physical description features such as camera motion (pan, tilt, and zoom). Examples of semantic visual features can be indoor/outdoor, people, car, and explosion. Naphade et al. (2005) proposed a 39-feature light weight 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, Setting/Scene/Site, People, Objects, Activities, Events, and Graphics; the second layer contains a number of sub-categories to provide further classification and specification. 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 that show the face of “Condoleeza Rice”, a list of other shots that have or partly have the features of “politics, face, person, government leader, police/private security personnel”, which are features of a picked Condoleezza video shot, can be utilized for looking for more matches. These “similar” shot keyframes can be presented in filmstrip to facilitate browsing (see Figure 1, area D below).

The selection of the shot keyframes are based on the level of “feature similarity”. In study, each keyframe Fi has a 39 dimension Boolean feature vector $F_i = (f_{i1}, f_{i2}, f_{i3},..., f_{i39})$ based on the ontology proposed by Naphade et al. (2005), and

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

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

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

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, m is 6.

Content-based video browsing System

A novel content-based video retrieval and browsing system was developed as a research platform to examine the effectiveness of various video browsing technologies under different search tasks in the context of multimedia digital libraries. The system supports two types of browsing: temporal neighbor browsing and semantic visual feature browsing. Figure 1 is the starting web page and user can give keywords in the textfield.

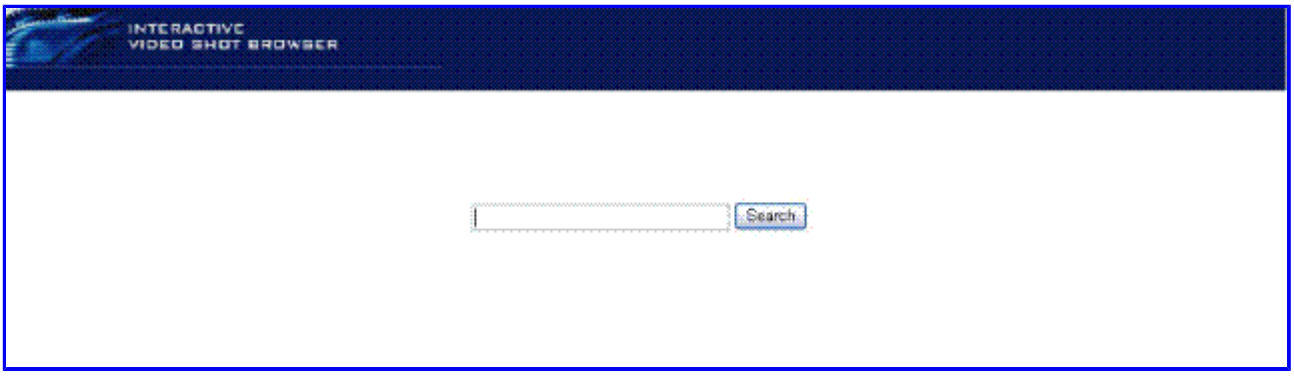

Figure 1: Starting page of the video retrieval and browsing system

After user's clicking on the "search" button, the main system interface will be presented (see Figure 2)

The main interface includes three panels: a search (Part A), result (Part B and C), and browsing panel (Part D). On the top part (part A) a traditional text input field is provided for text-based query. Text metadata includes a video's title, abstract/description, closed caption, and audio transcript. In our study the videos' transcripts were utilized for text-based retrieval.

In the middle (part B and C) is the result panel. Part B is on the left to display transcript of a selected sample keyframe from Part C, which displays the search results in storyboard style. In Part C the keyframe of the most matched shot is placed at the top-left corner, while the least matched is placed at the bottom-right corner. In Part B a dropping box with a "Selection" tag allows users to add or remove desired shot keyframes for their tasks.

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 adapted 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. Once user selects a sample keyframe from the storyboard panel in Part C, a list of the sample's neighboring shots or shots with similar visual features will be displayed as a filmstrip with the sample highlighted in the center. All the neighboring or similar frames will be displayed in the same size as the sample.

Figure 2: User interface of the content-based video browsing and retrieval system

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): focusing on visual features of a keyframe, for instance, “find shots of Tony Blair”.
- Non-visual centric tasks(NCT): focusing on non-visual features of a keyframe, for instance, “find shots about bird flu”

These tasks represent two complementary facets of shot retrieval. Visual centric tasks tend to emphasize visual information delivered through the visual channel. Non-visual centric tasks, on the other hand, emphasize semantic information that can be delivered through both the visual and audio channel.

Video Collection

The user study used data selected from the TRECVID 2005 data collection, including about 86 hours of news videos (137 segments with average duration of about half an hour) donated from MSNBC, CNN, NTDTV, CCTV, and LBC and associated metadata. Metadata included shot boundaries, shot key frames, shot durations, Automatic Speech Recognition, and semantic high level visual features for each keyframe. The semantic visual features were collaboratively created by some TRECVID 2005 project participants.

Experiment systems

Two shot browsing methods was studied: temporal neighbor and semantic visual feature browsing. Accordingly, three types of retrieval systems were compared:

- Temporal neighbor browsing system (TN): After the initial text query search, users were allowed to use the temporal neighbor browsing function to aid retrieval.
- Semantic visual feature browsing system (SF): After the initial text query search, users were allowed to use the semantic visual feature browsing function to aid retrieval.
- Fused browsing system (FU): After the initial text query search, users were allowed to use the temporal neighbor and semantic visual feature browsing functions to aid retrieval.

All the above three systems were modified from the prototype we previously introduced (figure 2, which represnted the FU system). For instance, for TN system, there were no tabs in part D and only temporal neighbor browsing function was supported.

Evaluation schema

Our scenario was to ask participants to look for five video shots that could be used in a student's project or a lecturer's presentation slides. Instead of using the recall and precision, our performance evaluation schema included the following two dimensions:

- Effectiveness: whether users can complete the required tasks and feel comfortable with their choices (levels of confidence will be asked).
- Efficiency: whether users can complete the required tasks quickly.

Goals

Our goals in this pilot user study were to

- Compare the performance of three video content retrieval and browsing systems
- Achieve a better understanding of the relationships between browsing methods and video seeking tasks,
- Provide recommendations for the improvement of design and development of content based video retrieval systems

Participants

Seven volunteer participants (five students, one professor, and one academic staffs) from multiple majors and programs participated in the study and six of them completed all the tasks (one student quitted due to personal reasons).

Tasks

Six tasks were modified from the TRECVID 2005 project tasks to represent two types of retrieval: visual and non-visual centric retrieval. In particular, users were asked to find shots of:

Visual Centric Tasks(VCT)	
Task 1:	face of George W. Bush
Task 3:	something (e.g., vehicle, aircraft, building, etc) on fire with flames and smoke visible
Task 5:	a tall building (with more than 5 floors above the ground)
Non-visual Centric Tasks (NCT)	
Task 2:	George W. Bush's foreign policy
Task 4:	damages caused by a fire (e.g., physical, psychological and social)
Task 6:	information about downtown development

Experiment D esign

Within-subject design was adapted and each participant was asked to complete these six search tasks (using all three systems). A Latin Square design was adapted to attempt to balance the learning effect. It took about half an hour for one participant to complete all the runs.

	TN	SF	FU
VCT	2	2	2
NCT			

Experiment Procedure

The study was conducted between Feb. 4th and 8th on the campus of UWM. A laptop with wireless network connection was provided in the study for students who did not have the Internet access. After a brief description of the intention, a list of tasks printed on paper was given to the participant. For each task, the time used to complete an individual search and users' confidence level (0-4 and 4 denotes the highest confidence) on their selections for that task were recorded. The task completion time was used as an index for efficiency evaluation and the confidence score for effectiveness index. Following completion of all the search tasks, participants were interviewed for about five minutes to give comments.

Initial result and analysis

In general, the participants were able to perform the tasks successfully without any extra training. The only case that a participant quitted the study was due to personal reasons. We did not find any difficulty for users to give confidence level scores for their selections.

The results showed that the average time to complete the six tasks was 98 seconds and the average confidence score was about 2.9 (0-4 scale and 0 for no confidence while 4 for highest confidence, see Table 1). The semantic visual feature browsing system (SF) was the most efficient one in terms of time spent for completing the tasks. From confidence scores we found that semantic visual browsing (2.7) is almost as effective as the temporal neighbor browsing (2.8), while the fused system performed the best (3.3).

Table 1: Efficiency (Time) and effectiveness (confidence score) for three systems

	TN	SF	FU	Average

Time (seconds)	98.8	80.2	115.7	98.2
Confidence score (0-4 and 4 is most confident)	2.8	2.7	3.3	2.9

Further analysis of the performances on visual centric tasks versus non-visual centric tasks, we found that users were more confident on the traditional temporal neighbor browsing (2.5) and the fused system (2.8) than that of the semantic visual feature browsing system (2.1). One explanation based on the post-experiment interviews might be that the semantic visual browsing only provided "visual" similar shots and this visual similarity was not their expected visual similarity. For instance,

"I am looking for President Bush but I got a number of (keyframes for) anchorman", commented by one participant.

Table 2: Efficiency (Time) and effectiveness (confidence score) for visual centric tasks

	TN	SF	FU	Average
Time (seconds)	128.7	119.3	152.7	132.5
Confidence score (0-4 and 4 is most confident)	2.5	2.1	2.8	2.5

Compared to visual centric tasks, the non-visual centric tasks were completed much faster (see Table 3). The average time was about 62.3 seconds, which was only half of the 132.5 seconds for that of the visual centric tasks. In addition, users felt more confident for their selections (3.5) than they did for the visual centric tasks (2.5). In general, for non-visual centric tasks, semantic visual feature browsing was more effective (3.5) than temporal neighbor browsing. On the other hand, semantic visual feature browsing was also the most efficient system (only 40.7 seconds in average) in this category of retrieval tasks.

The finding that the semantic visual browsing system performed well in non-visual centric tasks surprised us initially. Further analysis of the interview data revealed that users' more relying on transcript rather than the "visual similarity" in their confidence judgments could lead to a "shorter" browsing time. In addition, we did not provide transcripts for the neighbor shot keyframes in the filmstrip (Part D, Figure2), and consequently users might feel less confident in their selections when using the temporal neighbor browsing system.

Table 3: Efficiency (Time) and effectiveness (confidence score) for non-visual centric tasks

	TN	SF	FU	Average
Time (seconds)	69.0	40.7	77.3	62.3
Confidence score (0-4 and 4 is most confident)	3.1	3.5	3.8	3.5

Participants' performances on each of the tasks were given in Table 4. We found that when users spent more time on a specific tasks (e.g., task 3 and task 5), their confidence level were accordingly low (e.g., 1.9 and 1.7 for task 3 and 5 respectively).

Table 4: Efficiency (Time) and effectiveness (confidence score) for individual task

	T1	T2	T3	T4	T5	T6	Average
Time (seconds)	39.0	41.7	175.0	20.3	186.7	125.4	98.2
Confidence score (0-4 and 4 is most confident)	3.8	3.9	1.9	4.0	1.7	2.1	2.9

Conclusions and future work

In general we found that the fused system that supports both semantic visual browsing and temporal neighbor browsing was the most effective system, even though not the most efficient one. The semantic visual similarity browsing performed efficient in non-visual centric tasks and had a similar level of effectiveness as the temporal neighbor browsing.

Findings from the study provided valuable recommendations for the improvement of the system design, such as 1) highlighting the keywords in the transcript 2).providing a larger browsing scope for both the filmstrip (Part D in Figure 2) and the storyboard (Part C in Figure 2), and 3) providing transcripts for the keyframes in the filmstrip.

A large-scale usability study based on an improved browsing system is planed to further explore the relationships between the browsing technologies and video search tasks. Other factors such as the video genre could also be added in the future study.

Acknowledgement

Thanks to the TRECVID2005 project for providing videos and metadata. My colleagues, Dr. Dietmar Wolfram and Dr. Wooseob Jeong provided invaluable helps in the user study methodology and system user interface design; Rebecca Hall helped to implement the web interface of the new browsing system.

References

Amir,A., Argillander,O.J., Berg, M., Chang,S.F., Franz, M., Hsu,W., Iyengar, G., et al. (2004) IBM Research TRECVID-2004 video retrieval system *TRECVID2004*

Carmel, E., Crawford, S., & Chen, H. (1992) Browsing in hypertext: a cognitive study *IEEE*

Transactions on Systems , Man, and Cybernetics 22, 865-884

Chen, H-L., & Choi, G. (2005) Construction of a digital video library: a socio-technical pilot study on college students' attitudes *Journal of Academic Librarianship* 31(5), 469-476

Christel G. M., Smith, A. M., Taylor, R. C., & Winkler, B. D. (1998) *Evolving video skims into useful multimedia abstractions*

Cooke, E., Ferguson,P., Gaughan, G., Gurrin, C., Jones, J.F. G., Borgne, L.H., Lee,H., et al. (2004) TRECVID2004 experiments in Dublin City University *TRECVID2004*

Evans, J. (2003) The future of video indexing in the BBC *TRECVID* workshop

Gauch, S., Aust, R., Evans, J., Gauch, J., Miden, G., Niehaus, D., & Roberts, J. (1994) The Digital Video Library System: Vision and Design *Proceedings of The First Annual Conference on the Theory and Practice of Digital Libraries* Retrieved January 18, 2006 from http://www.csdl.tamu.edu/DL94/paper/gauch.html

Heesch, D., Howarth, P., Magalhaes, J., May, A., Pickering, M., Yavlinsky, A., and ruger, S. (2004) Video retrieval using search and browsing *TRECVID2004*

Helander, M. (Ed.). (1988) *Handbook of human-computer interaction* Amsterdam, Netherlands: North-Holland

Houten, V., Setten, V., and Enschede, E. O. (2000) Video browsing & summarization: user interaction techniques for video browsing & summarization *TI/RS/2000/163* Retreived 11/15/2005 from https://doc.telin.nl/dscgi/ds.py/Get/File-12409/user_interaction.pdf

Li, F.C., Gupta, A., Sanocki, E., He, L., and Rui, Y. (2000) Browsing digital video *Proceedings of CHI 2000* p. 169-176

Marchionini, G., Geisler, G. (2002) The Open Video Digital Library *D-Lib Magazine* 8(12). Retrieved January 18, 2006 from http://www.dlib.org/dlib/december02/marchionini/12marchionini.html

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 *TRECVID2004* program

Murdock, V. and Croft, W. B. (2002) Task-orientation in question answering *Proceedings of the 25th International ACM SIGIR Conference, Tampere, Finland, August*

Naphade, R. M., Kennedy, L., Kender, R. J. , chang, S., Smith, R. J. , Over, P., and

Hauptmann, A. (2005) A light scale concept ontology for multimedia understanding for TRECVID 2005 Retrieved 11/15/2005 from http://www-nlpir.nist.gov/projects/tv2005/LSCOMlite_NKKCSOH.pdf

Rose, B., and Rosin, L. (2001) The Internet & Streaming: What consumers want next Arbitron/Edison Media Research Internet VIIRetrieved Nov. 20, 2005 from http://www. edisonresearch.com

Salampasis, M., tait, J., & Bloor, C. (1998) Evaluation of information-seeking performance in hypermedia digital libraries *Interacting with computers* 10, 269-284

Shatford, S., (1986) Analyzing the subject of a picture: a theoretical approach *Cataloguing & Classification Quarterly* 5(3), 39-61

Shneiderman, B. (1996) The eyes have it: a task by data type taxonomy for information visualizations *Proceedings of the IEEE conference on visual languages* pp. 336-343

Tse, T., Vegh, S., Marchionini, G., & Shneiderman, B. (1999) An exploratory study of video browsing user interface designs and research methodologies: Effectiveness in information seeking tasks *Proceedings of the 62nd Annual Meeting of the American Society for Information Science* p. 681-692

Wildemuth, M. B., Yang, M., Hughes, A., Gruss, R., Geisler, G., & Marchionini, G. (2003) Access via features versus access via transcripts: user performance and satisfaction *TRECVID2004*

Yang, M. Wildemuth, M. B., Marchionini, G., (2004) The relative effectiveness of concept-based versus content -based video retrieval *Proceedings of the 12th annual ACM international conference on Multimedia* pp.368-271

Zhang, H., J., Low, C.Y., Smoliar, S., W., & Wu, J.H. (1995) Video parsing, retrieval and browsing: an integrated and content-based solution *Proceedings of the third ACM international conference on Multimedia* 15-24