A Touch-panel based User Interface and Utilization of User's Memories for Known-item Search (KIS) Task in TRECVID 2011

Yuan Zhou, and Takashi Yukawa
Nagaoka University of Technology, Nagaoka-shi, Niigata 940-2188 Japan

Abstract

This year the KSLab-NUT team participated in the interactive know-item search task of TRECVID 2011. The keywords of the run of which runid is I_B_YES_KSLab-NUT_1 are iPad-as-UI, salient word refining, and video-length verification. The authors' approach this year is to develop an easy-to-use and intuitive system, therefore iPad is chosen as user interface of the client. As the description of the know-item search task, the authors believe that user has some clues or memories relative to the search query in mind. Thus, salient word refining and video-length verification are suggested separately during the search process.

1. Introduction

This year the team from Knowledge Systems Lab at Nagaoka University of Technology participated in the interactive know-item search (KIS) task (KSLab-NUT). This is the first participation in TRECVID and the authors developed most of the components of the system from the ground-up. As in the interactive search, the authors' approach of this year is to develop simple, easy-to-use and user-centered components of the search engine and the authors are looking forward to extending current system with more functions and technologies in the coming years.

Since interactive KIS task is evaluated by user satisfaction, mean elapsed time, and mean average precision, the emphasis of this paper is three-fold. First, iPad is chosen as an easy-to-use and user-friendly user interface to facilitate search process and the browsing of returned videos. This component is supposed to the contribution of the user satisfaction of our system. Second, as the description of the KIS task, it is assumed that user remembered some key information or clues of the video in their mind as well as the "key visual cues" in the query. Therefore, two memory-aid functions, salient-word refining and video-length verification, are implemented to the system during the search process due to the psychological studies of human memory [18][16]. These two components are postulated to the contribution of the mean elapsed time and mean average precision.

2. Related Work

The system built by C. Foley et al. [4] used iPad last year as the user interface and it is believed to improve the user's experience and satisfaction greatly. L. Chaisorn et al. [10] proposed a search system using several search techniques including "salient term/phrase selection", "custom stemming and inflectional normalization", etc. and reached the highest average precision in TRECVID 2010 conference. However, researches concerning the manipulation of user's memory in video retrieval, not only in TRECVID since 2009 KIS is set as a main task of the conference, but in the entire research field of video retrieval as well are comparatively inadequate though it is believed that memory aid designed system is helpful and effective [1][13].

3. Interactive Known-Item Search System using iPad and Salient Word Refining

Figure 1 provides an overview of our video search system. The user interface for the system is built as an application running on the iPad. Meanwhile, on the server side, a Tomcat-based Java web service [9] is running on a separate PC and HTTP protocol (post call) is used as the communication between the client and server. For indexing and retrieval, Apache Lucene is used.

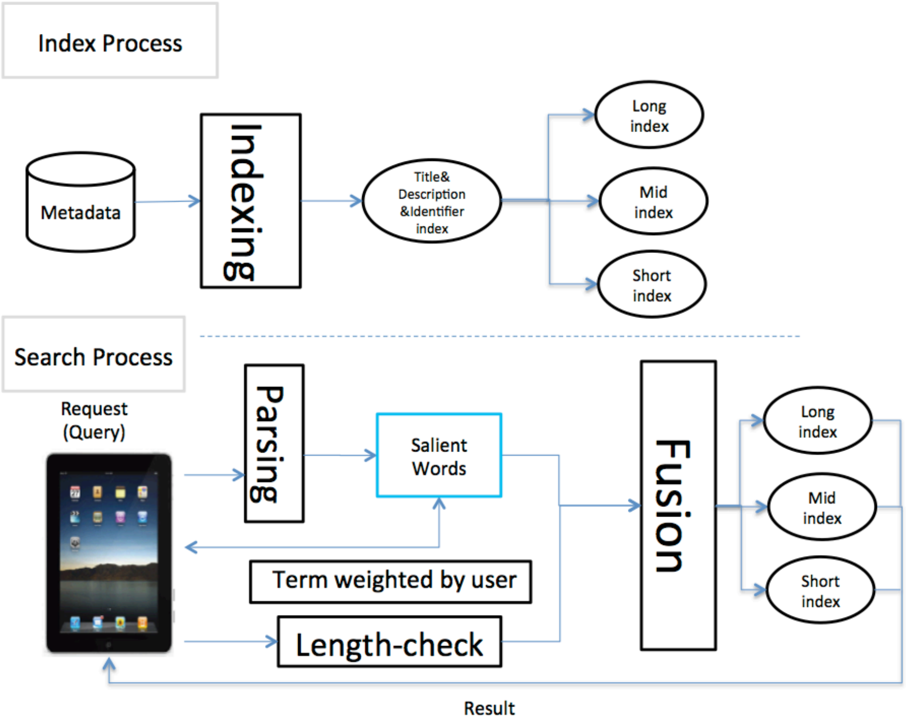


Figure 1: System Overview

3.1 Indexing Process

During the indexing process, the authors simply choose the metadata provided along with each video data. As it is mentioned before, Apache Lucene is used to build the index of our system and Lucene is a free/open source informational retrieval library that uses the inverted index and tf-idf weighting scheme. The indexing process creates an instance of each document within the directory and populates it with Fields that consists of name and value pairs [11]. In this paper, considering the effectiveness of the information, the author selected three fields of each metadata to index: identifier, title, and description.

3.2 Searching Process

Figure 2 is the searching process in detail and there are three main characteristics in this process: iPad as user interface, salient word refining, and video length verification.

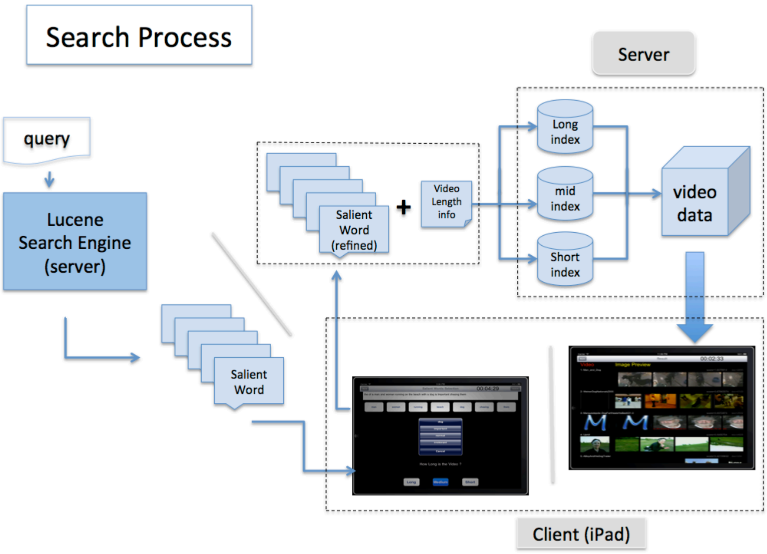


Figure 2: Search Process

3.2.1 iPad as User Interface

Because the user satisfaction is one of the three evaluation methods of KIS task, the authors determined to seek an easy-to-use and intuitive user interface. The iPad, which is released on Apr. 2010, has changed the view of usage of the traditional PC a lot until now and with the time elapses, it is believed to influence the habit of the computer usage of human being. Users can interact it with just touching on the screen by their figures. In a word, it is believed that using iPad as user interface of our system would raise the user satisfaction and perhaps also shorten the mean elapsed time.

Upon starting the application, the user is required to input the query first (as shown in Figure 3).

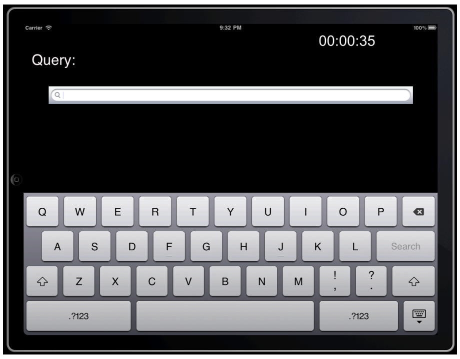
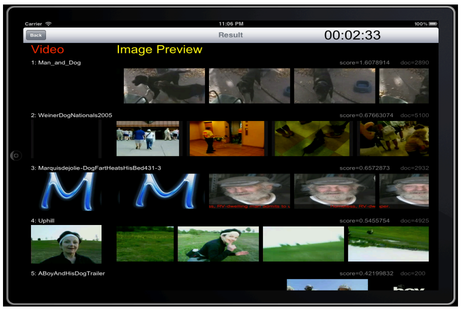


Figure 3: Start of the Search

After the query input, several salient words will be shown on the upper-side of the screen and each word can be touched and its importance (or salient word's weight) could also be changed by user. As shown in Figure 4, there will be a preview of each video result returned in ranked order, which is up to 10 possible video results, with a maximum of 10 representative keyframes being displayed in the "image preview" section on its right side.



1: Man_and_Dog score=1.6078914 doc=2890

2: WeinerDogNationals2005 score=0.67663074 doc=5100

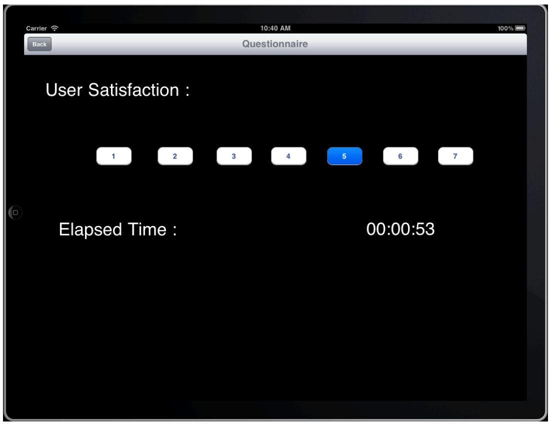
3: Marquisdejolie-DogFartHeatsHisBed431-3 score=0.6572873 doc=2932

4: Uphill score=0.5455754 doc=4925

5: ABoyAndHisDogTrailer score=0.42199832 doc=200

Figure 4: Preview of the video result

The top ranked video result is supposed to have the highest score calculated by Lucene search engine, which means the most possibly correct answer to the user's query. User can play the video by simply tapping on it in the "video" section. Once the user finds what they believe is the correct answer to their query, they can tap "found it" button to end this search. Finally, as shown in Figure 5, there will be a user satisfaction investigation and elapsed time of the whole search process shown on the screen.



User Satisfaction :

1 2 3 4 5 6 7

Elapsed Time : 00:00:53

Figure 5: User Satisfaction Investigation

3.2.2 Salient Word Refining

Since the author attended the interactive KIS task, concentrating on the user's consideration is always on the first rank while designing an interactive video search system. It is postulated that refining the salient words processed by the search system automatically once again based on the user's memory would be effective and helpful, and not time-consuming to the search process.

3.2.2.1 Studies concerning memory

In psychology, memory is the ability of an organism to store, retain, and recall information and experiences.

There are three kinds of memory types defined by how long it passes when one wants to recall something: sensory memory, which corresponds approximately to the initial 200-500 milliseconds after an item is perceived; short-term memory, which allows one to recall for a period of several seconds to a minute without rehearsal; and long-term memory, which can store large quantities of information for potentially unlimited duration (sometimes a whole life span). Due to the description of the KIS task, the author believes that considering the time period until one wants to find a certain previous watched video, it is highly possible that this time period belongs to long-term memory.

Declarative memory (sometimes referred to as explicit memory) [17] is one of two types of long-term memory and refers to memories which can be can be consciously recalled such as facts and knowledge. Furthermore, declarative memory can be divided into two categories: episodic memory that stores specific personal experiences and semantic memory that refers to the memory of meanings, understandings, and other concept-based knowledge unrelated to specific experiences such as generalized knowledge that does not involve memory of a specific event.

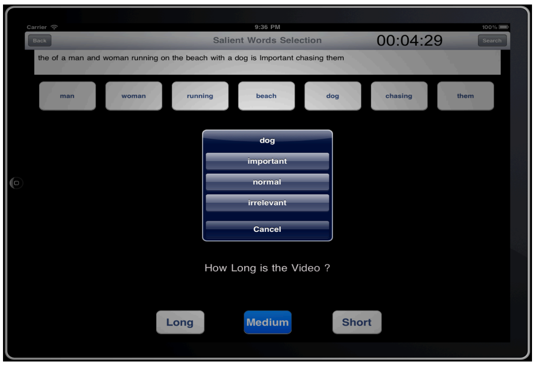
Episodic memory is the memory of autobiographical events (times, places, associated emotions, and other contextual knowledge) that can be explicitly stated. This implies that a detail from a past even in which the user was involved might be difficult to recall, the name of a document, for example. But the context of the event can be easier to remember [13]. For example, one may be able to recall: the place where the document was received, the people present when it was handed over, or the task being carried out at the time.

Researchers in psychology have developed theories about episodic memory for a period [16][7] and observed that human beings naturally organize memories for past events into episodes [2], and the location of the episode, who was there, what was going on, and what happened before or after, are all strong cues for recall [16][14]. Studies by Eldridge et al. [5][6] have confirmed these findings, and moreover, have led author to believe the following scenario: user seems to remember the important words or some representative words that can describe a key concept, situation or a scene in the previously watched video.

3.2.2.2 Salient word refining function

In order to acquire the salient words of each query, when user completes inputting request, it will be directly sent to the server and processed by Lucene search engine. In this research, the authors didn't develop our specific analyzer class for Lucene search engine but simply choose its original analyzer named

“StandardAnalyzer”. User can adjust the importance as “important”, which weighs the word, “normal”, which keeps it as the same as the default, or “irrelevant”, which deletes the word from the salient word list, of each salient word intuitively or based on their memories. Figure 6 illustrates this scenario.



the of a man and woman running on the beach with a dog is Important chasing them

How Long is the Video ?

Long [✓ Medium] Short

Figure 6: Salient Word Refining

3.2.3 Video Length Verification

One can see colors, hear sounds and feel textures. Also, one can perceive time [15]. Human beings seem to have a special faculty, distinct from the five senses, for detecting time. It is conceded that the perception of temporal duration is crucially bound up with memory. It is some feature of our memory of the event (and perhaps specifically one’s memory of the beginning and end of the event) that allows human beings to form a belief about its duration. It seems likely that it is intimately connected with what W. Friedman calls ‘time memory’: that is, memory of when some particular event occurred. That there is a close connection here is entailed by the plausible suggestion that we infer (albeit subconsciously) the duration of an event, once it has ceased, from information about how long ago the beginning of that event occurred [8].

In the real world, although storing everything may seem ideal, it creates a problem akin to the way humans store their experiences. In psychological studies of human memory, it has been suggested that the phenomenon of “forgetting” is basically a retrieval problem. The information is stored, but it has barely a right way to retrieve it by humans. In order to help users with this difficult task, we use memory aids.

Human memory retrieval process is triggered by retrieval cues: stimuli related to an experience that facilitate the recall of other information related to the same experience. A person wishing to remember details of an event often uses cues --- when he saw something, where he put it, who sent it, for what event --- to find the information. Cues useful for retrieval include events, time, people [3], and activities (what and where) [12].

There has been a growing interest in the investigation of everyday memory [18] in recent years. However, little work has been done in studying what people remember (or forget) about a previous watched video. In a previous research concerning on the meeting video retrieval [1], the author invented a new approach to meeting video retrieval based on the concept of using memory aids to remember important information from meeting the users have attended.

One study was conducted to comprehend the types of things people remember (or forget) about the meetings.

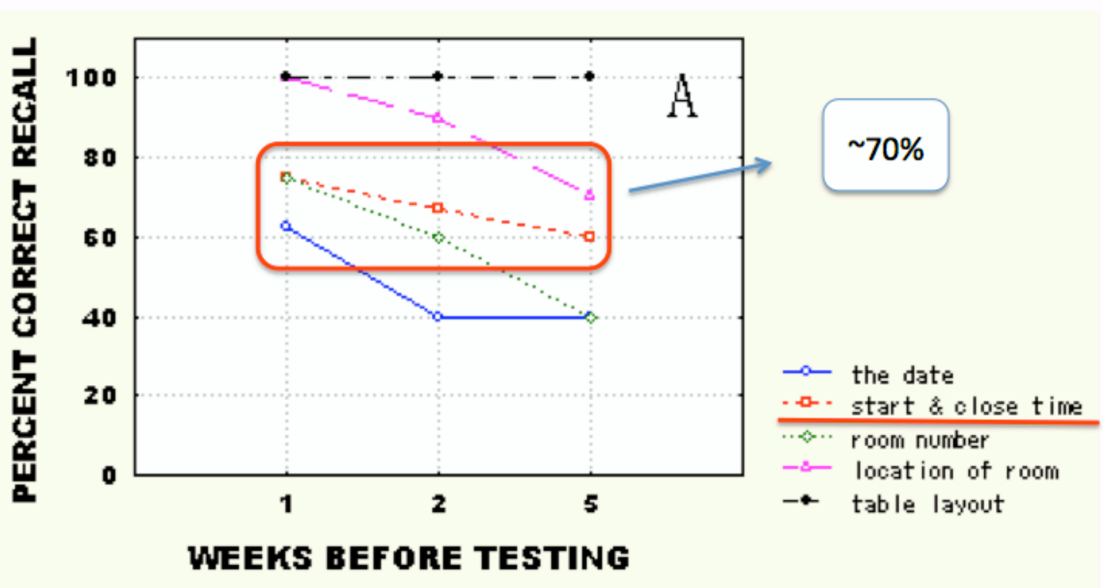


Figure 7: Result of Memory Test

Figure 7 shows that nearly 70% of the tester remembered the start and close time of the meeting in average.

On the other side, from the description of the KIS task, user is supposed to have seen the video before. Therefore, it is assumed that user may have the impression of the length of the video he/she wants to find roughly. Based on this assumption, the authors divided the video data (also the index) into three parts: "short", which stands for 10 to 20 seconds, "medium", which stands for 21 to 60 seconds, and "long", which stands for 61 seconds to 4.1 minutes.

On the bottom of the same screen of figure 6, there will be three buttons representing "long", "medium", and "short" appeared for being chosen during the refining process of the salient words.

4. Experimental Evaluation

Table 1 shows the results of the experiment run on our system to test its performance. In this experiment, 25 test topics are chosen exclusively to interactive KIS task. Although interactive runs will be scored in terms of the video found or not, it is believed that mean inverted rank of each run may not be underestimated. For example, there could be a situation that user might not have enough time (5 minutes limited) to browse and find the correct video which is not comparative upper in the rank.

Table 1: Result of the Overview Performance of the System

Index & Methods combination	25 Test Topics		
	Mean Precision	Mean Elapsed Time	Mean User Satisfaction
XML	0.320	3.544	4.000
+ Salient Word refining & Length Verification	0.400		

In table 1, it is counted by 1 if the correct video ID is on the preview list (maximum 10 possible videos, ranked 1-10) no matter which rank it is on when the authors calculate the Mean Precision while scoring 1/rank, which is according to the rank of the correct answer on the list when calculating the Mean Inverted Rank.

From the baseline (only XML), our system achieves MP of 0.320, which means it found 8 correct videos while MIR of 0.176 due to the different ranks of each video. After salient word refining and video length verification function being implemented, the performance rises up to 0.400 in MP, which means 10 videos are found correctly while 0.373 in MIR, which suggests a huge increase of each video's rank. For example, as for NO.0519 test topic, the correct answer would not be shown on the preview list by the baseline at first. Then after setting the video length as "short", user could find the correct answer shown at 5th rank on the preview list. Moreover, if the user chose salient word "moon" as important, it would be ranked at 3rd.

Moreover, according to the table 1, mean elapsed time and mean user satisfaction stay the same as 3.544 and 4.000 because the authors believe that they have very little effect on both evaluations.

Through the experiment, the authors found that the reason why the performance (MP) of our system is comparatively low is because only metadata has been used to build the index of our system. Since over half of the metadata of the correct answer of the topics (actually 13 out of 25) doesn't contain any useful and key information of that video, there is no way for the system to search for those queries.

5. Summary and Future Work

In this year the authors presented our experiment in only Interactive search at this year's TRECVID workshop.

The interactive search system is based on using Apache Lucene during indexing and searching process. This year only metadata provided by each video has been used as the index. iPad is chosen as the user interface of the system to achieve better user satisfaction and implement salient word refining as well as video length verification to acquire better MP and mean elapsed time. As the result, it is assumed that the system has achieved good performance.

For the future work, it is considered that there is necessary to expand the index with OCR and/or ASR data because the authors found that there are several topics include the query concerning on the visual level or the acoustic level. Second, as iPad being used as the user interface, it is supposed that it might be a good idea to implement voice control function using Artificial Intelligence, such as Siri, which is supplied by iPhone4S, to improve the user experience and achieve a better user satisfaction.

References

[1] A. Jaimes, K. Omura, T. Nagamine, and K. Hirata. Memory Cues for Meeting Video Retrieval. FXPal Japan, Corporate Research Group, Fuji Xerox Co., Ltd., Japan. 2004.

[2] Barsalou, L. W. (1988). The content and organization of autobiographical memories. In U. Neisser & E. Winograd (Eds.), Remembering reconsidered: Ecological and traditional approaches to the study of memory. Pp. 193-243. Cambridge: Cambridge University Press.

[3] C. Ahlberg, B. Shneiderman. Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays. Proc. Of the SIGCHI conference on Human factors in computing systems: celebrating interdependence, pp. 313-317, Boston, MA, 1994.

[4] C. Foley, J. Guo, D. Scott, P. Ferguson, P. Wilkins, K. M. Cusker, E. S. Diaz, C. Gurrin, A. F. Smeaton, X. Giro-i-Nieto, F. Marques, K. McGuinness, N. E. O'Connor. TRECVid 2010 Experiments at Dublin City University. In TRECVid 2010 – Text Retrieval Conference TRECVID Workshop, MD, USA, 2010. National Institute of Standards and Technology.

[5] Eldridge, M., Barnard, P. & Bekerian, D. (1994). Autobiographical memory and daily schemas at work. Memory.

[6] Eldridge, M., Lamming, M. & Flynn, M. (1992). Does a video diary help recall? In A. Monk, D. Diaper, & M. D. Harrison (Ed.), People and Computers VII., VII pp.257-269. York: Cambridge University Press.

[7] E. Tulving. Episodic and semantic memory. In Organization of memory, ed. E Tulving, W. Donaldson, 1972. pp.381-403. New York: Academic.

[8] F. William. About Time: Inventing the Fourth Dimension. The MIT Press. 1990.

[9] J. Murach, A. Steelman, “Murach’s Java Servlets and JSP, 2nd Edition”, Mike Murach & Associates, INC. 2008.

[10] L. Chaisorn, K. W. Wan, Y. T. Zheng, Y. Zhu, T. S. Kok, H. L. Tan, Z. X. Fu, and S. Bolling. TRECVID 2010 Know-item Search (KIS) Task by I2R. In TRECVid 2010 – Text Retrieval Conference TRECVID Workshop, MD, USA, 2010. National Institute of Standards and Technology.

[11] M. McCandless, E. Hatcher and O. Gospodnetic, “Lucene in Action”, Manning Publication, 2010.

[12] M. Lamming, and D. Bhom. SPECS: Another Approach to Human Context and Activity Sensing Research, using Tiny Peer-to-peer Wireless Computers. In proc. UbiComp 2003, Seattle, Washington, 2003.

[13] M. Lamming and M. Flynn, 1994. “Forget-me-not” Intimate Computing in Support of Human Memory. In Proceedings of FRIEND21, ’94 International Symposium on Next Generation Human Interface, Meguro Gajoen, Japan.

[14] Saywitz, K., Bornstein, G. & Geiselman, E. (1992). Effects of cognitive interviewing and practice on children’s recall performance. Journal of Applied Psychology, 77(5), 3-15.

[15] Stanford Encyclopedia of Philosophy. The Experience and Perception of Time. 2000: http://plato.stanford.edu/entries/time-experience/

[16] Tulving, E. (1983). Elements of Episodic Memory. Oxford University Press.

[17] Ullman MT. contributions of memory circuits to language: the declarative/procedural model. Cognition 2004; 92: 231-70.

[18] U. Neisser. Memory Observed: Remembering in Natural Context. W.H. Freeman and Company, New York, NY, 1982.