ADVANCING CONTENT-BASED RETRIEVAL EFFECTIVENESS WITH CLUSTER-TEMPORAL BROWSING IN MULTILINGUAL VIDEO DATABASES

Mika Rautiainen, Tapio Seppänen, Timo Ojala
MediaTeam Oulu, University of Oulu, P.O.BOX 4500, FIN-90014 University of Oulu
{firstname.lastname@ee.oulu.fi}

ABSTRACT

Interactive experiments on video retrieval systems need to address the problem of internal validity, i.e. how much the test users' experience affects the retrieval effectiveness. This paper compares the semantic retrieval performance of novice users and expert system developers. The test system utilizes cluster-temporal browsing, which combines chronological video structure and computation of similarities into single interface. Interactive experiments with eight test users were carried out in a database of ~80 hours of multilingual news video from TRECVID 2005 benchmark. A cluster-temporal browser was found to improve the retrieval effectiveness by 12% with novice system users. Expert users were able to achieve 18% better performance than the novice users. Additionally, manual search experiments demonstrated that search performance can be improved by 19-25% when a plain text search is supplemented with content-based features.

1. INTRODUCTION

Conventional content-based retrieval paradigm consists of a query with attributes and examples describing the needed content and a mechanism to compute relevance using the actual content data. Related work can be found from [1][2][3].

Baseline performance for a content-based video search uses typically automatic speech transcripts. However, automatic transcripts have language domain constraints, are dependent on the quality of the audio source and do not contain all of the existing semantic information. Transcript information can be augmented by recognizing additional concept terms from the video and by computing similarities between content samples [4]. Relevance feedback improves the content-based retrieval performance by incorporating information from prior relevance judgments [5].

In [6] cluster-temporal browsing was introduced as an interactive navigation tool for search guided browsing in video databases. The novelty lay in combining the video time-line and content-based clusters into a dynamic view. Previous experiments have found cluster-temporal browser to improve retrieval effectiveness of novice users by 22% over sequential search with relevance feedback [7].

One of the open questions with the interactive search experiments is the internal validity, i.e. how much the professional experience of a user affects the retrieval effectiveness. This study measures such bias with interactive experiments on a large multilingual video database. Sections 2 and 3 describe cluster-temporal browsing and the test system. Section 4 presents the experimental results. Section 5 gives conclusions.

2. CLUSTER-TEMPORAL BROWSING

Moving on a video timeline is a classical example of video browsing. It is intuitive but time consuming search strategy with regard to large video collections. Content-based search is supposed to retrieve relevant content from the videos but it can not guarantee complete relevance due to semantic gap. Content-based navigation and browsing can be useful tools to alleviate this problem [8][10][11][19][20][20].

Rodden and Wood's [9] user tests suggest that the most desirable features for a photo archive browser are chronological navigation and visual previews with large number of images. In cluster-temporal video browsing both of them are considered: temporally adjacent shots are used to concurrently retrieve and display a number of nearest neighbors from content-based feature clusters.

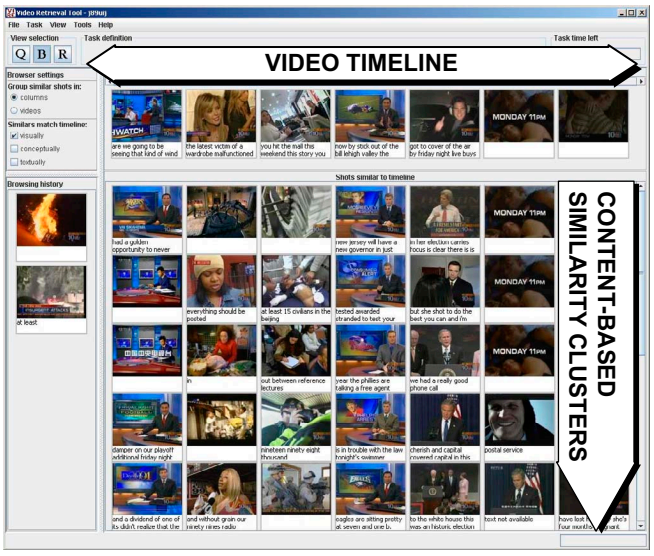

Figure 1. Cluster-temporal browsing interface

The arrows in Figure 1 illustrate how cluster-temporal browsing combines content-based clusters and temporal order to maximize information in a single view. The row beneath the horizontal arrow displays the current video of interest, with chronological key frame sequence. The vertical arrow is in a panel that contains similar shots from the rest of the database. Each vertical column represents a retrieved cluster where the results are ordered with downward decreasing similarity with the query shot atop of the column. The columns form a matrix of similar shots created from the entire database, from which the user can open a new relevant video to the timeline. When the new video is selected, similarity view is immediately updated. The user can also navigate to different temporal location in the selected video. Similarity matrix is then recomputed using the currently visible timeline segment.

3. EXPERIMENTAL VIDEO RETRIEVAL SYSTEM

The experimental video retrieval system consists of a query server, a content-based search interface, a result container with relevance feedback and a cluster-temporal browser. The server has visual *v*, concept *s*, and text *l* indexes. It creates a relevance ranking based on feature similarities and fusion of result sets. The search interface is employed for manual query definition. The result container collects selected shots and uses them as relevance feedback by creating new content-based example queries. The browser acts as a navigational tool where relevant shots are collected using direct interaction with the system.

3.1. Content-based feature indexes

Visual feature indexes are based on color and structure of a video shot. Used low-level features are described in [12][13]. The computed Color Correlogram (CC) and Gradient Correlogram (GC) features describe statistical co-occurrences of colors and edges in a single key frame. Dissimilarities of the individual features are based on city-block distance. CC and GC queries generate two separate rank-ordered lists of search results, which are combined using sum of ranks [4] that has been found effective when the dimensions of the features vary.

The semantic concept index is constructed of detected concept confidences. We have developed two types of detectors, SVM classifiers [14] and propagated labeling based on positive examples [13]. The following concepts are implemented using SVM: entertainment, faces, newsroom, outdoor, desert, natural-disaster, and snow. The detectors with propagated labeling: fire-explosion-smoke, maps-charts, meeting-footage, nature-footage, sports, water, and weather.

A text index is constructed from the available automatic speech recognition (ASR) and machine translation (MT) transcripts. The words are first pre-processed with stop word removal and stemming and then indexed into a database. Grouping words into speaker segments improves contextual organization for the index and whip up the text search. The textual similarity of shots is computed using prioritized ranking combined with weighed term frequency score [14]. We have also constructed an example-based text search engine using text from the example shot as the query to allow cluster-temporal browsing using text similarity. [14]

Queries require retrieval from any or all of the three described indexes. The rank-ordered sub-results can be considered as votes from the 'experts'. Fusion of the feature lists is created using a variant of Borda [15] count voting:

$$f^t(n) = sum\left(\frac{w_v \cdot v^t(n)}{V^t_{\max}}, \frac{w_s \cdot s^t(n)}{S^t_{\max}}, \frac{w_l \cdot l^t(n)}{L^t_{\max}}\right) \quad (1)$$

$$F^t = \left[sort\{f^t(1),...,f^t(N)\} \right]_X \quad (2)$$

where $f^t(n)$ = overall rank of a result shot *n* to the search topic definition *t*

$v^t(n), s^t(n), l^t(n)$ = rank of a result *n* by independent search engines

w_v, w_s, w_l = weights of the search engines: 1, 1, 2 (*v,s,l* respectively)

$V^t_{\max}, S^t_{\max}, L^t_{\max}$ = last rank of the independent result lists.

3.2. Search interface and result container with relevance feedback

Our test system incorporates traditional content-based tools for constructing queries and providing relevance feedback. The search interface allows users to define queries manually: First, the user can select examples for visual search. Second, user can configure a semantic query from the list of given semantic concepts. Third, text query is created by typing words to a text box. From the retrieval results, the user can pick relevant shots to the result container or select any shot as a start point for browsing with the cluster-temporal browser.

The result container collects every selected relevant item into a list. Selected shots are considered as positive examples and used in a relevance feedback query, which is directed to visual and text search engines. The results are displayed under the selected shots. Each time a new relevant shot is added, query is regenerated. Found results are displayed as an additional resource in order to help finding more relevant shots from the other parts of the database.

3.3. Cluster-temporal browser interface

In addition to the traditional search tools, our test system provides cluster-temporal browsing as an alternative search strategy. User can select any shot from the other interfaces and open the related video in the browser interface. User can pick relevant shots either from the video timeline or the

similarity view and add them to the result container. The browser allows the configuration of similarity parameters (text, visual or both) and tracks browsing history by showing a list of latest shots that the user has accessed.

4. EXPERIMENTS AND RESULTS

Our experiments focused on measuring improvement in retrieval effectiveness with the cluster-temporal browsing over the traditional content-based search techniques, namely sequential queries with relevance feedback. We also tested one-time manual search performance with different feature configurations.

Our test system was developed and evaluated on 80 hours of English, Arabic and Chinese video data from TRECVID 2005, which is a U.S. National Institute of Standards and Technology (NIST) led retrieval benchmark providing common framework for research groups to test their content-based video retrieval systems [16]. Training and testing of the system were performed in separate databases. NIST provided 24 search topics that were used in the experiment. The search results were sent to NIST for evaluation. A search topic contained one or more example clips of video or images and textual topic description to aid the query definition. Video database was initially segmented into shots by Fraunhofer HHI [17]. ASR and MT transcripts were provided by NIST and Carnegie Mellon University.

The interactive experiments were carried out with a group of eight test users: four were novices with the system (S5-S8) and the other four were involved in the retrieval system development (S1-S4) but had not seen the test search topics or any content from the test database. The users were mainly information engineering undergraduate students having good skills in using computers and searching the web but had little experience in searching video databases. Test was organized into latin square configuration to minimize the effect of 'random' proficiency for certain search topics and system configurations. See Table 1.

Table 1. Interactive test configuration

Run ID	Searcher ID [topic set IDs]			
I1Q	S1[TG1]	S3[TG2]	S2[TG3]	S4[TG4]
I2B	S2[TG1]	S4[TG2]	S1[TG3]	S3[TG4]
I3Q	S7[TG1]	S5[TG2]	S6[TG3]	S8[TG4]
I4B	S8[TG1]	S6[TG2]	S5[TG3]	S7[TG4]

In practise, the four sets of six topics (TG1-TG4) together with two search system variants (IxQ: sequential queries with relevance feedback and IxB: the same system augmented with cluster-temporal browser), were distributed between the eight users. After six search topics, at halfway of the experiments, users were given a break with refreshments to dispel the effect of fatigue. Search time for a topic was limited to 12 minutes totalling in an approximately three hour experiment with 12 topics per user. Novice users received 30 minutes of training with the search system before the actual experiments.

Average precisions for the four different search configurations are shown in Table 2. MAP shows the mean value of the average precisions for 24 search topics. Although the MAP values seem low, on average the runs obtained 8.25 correct results within top 10 results which can be considered very high for versatile semantic search tasks. The same value for 'hits at depth 30' was 20.36 which is approximately two times higher than our formerly reported results [18]. Novice users achieve 12% improvement in retrieval effectiveness by using the cluster-temporal browser. This result is in line with our previous findings with the novice users (22% improvement) [7]. Novice user logs showed that 49% of the relevant shots originated from the cluster-temporal browser, 33% from the sequential searches and 18% from the relevance feedback. This demonstrates high browser utilization during the experiments.

On average the expert users (system developers) were able to achieve 18% improvement over the novice user performance. Overall, the group of experts did not benefit from using the cluster-temporal browser. This can be explained with their level of expertise; two of the experts were responsible for developing and training semantic concept detectors whereas the other two were system and interface developers. Knowing the classification performance of the individual concept detectors helped the expert users to construct efficient semantic queries. This can be seen from the topics that created the largest increases in average precision for the benefit of run I1Q: Find shots of 'Mahmoud Abbas', 'Hu Jintao', 'Omar Karami' and 'people shaking hands'. These topics were successfully retrieved using text search and concepts 'meeting-footage' and 'faces'. Due to this specialist knowledge of the system experts, we believe that testing with novice users results in better external validity.

Table 2. MAP and total relevant for search runs

Search Run ID	MAP
I1Q (expert users)	0.264
I2B (expert users)	0.242
I3Q (novice users)	0.202
I4B (novice users)	0.226
Mean (interactive)	0.218
Max (interactive)	0.414
M5T (txt search baseline)	0.081
M6TS (txt+semantic)	0.097
M7TE (txt+examples)	0.102
Mean (manual)	0.067
Max (manual)	0.169

Table 2 also shows the results for manual search runs, which disclose one-time retrieval effectiveness without interaction loop between the system and the user. We tested three manual search configurations: plain text search for baseline run (M5T), text search combined with semantic feature query (M6TS) and text search combined with visual example clips (M7TE). The results show that baseline text search performance can be improved as much as 19-25% by combining content-based features to the search. Visual example clips were found to be more significant than semantic concept queries for boosting retrieval effectiveness.

Overall the performances of our interactive runs were around the mean of all TRECVID interactive runs. Our search runs returned the highest number of unique relevant shots among the TRECVID search participants, which indicates that our system facilitates locating novelty information. [22]

5. CONCLUSIONS

The experiments in large multilingual video collection show that cluster-temporal browsing amplifies the retrieval effectiveness over the conventional content-based retrieval techniques for novice users. Developers as expert users know the system's strengths and weaknesses, for example the classification rates for concept classifiers, and are therefore able to select more efficient query configurations for different types of topics. Due to this, search tests using developers has lower external validity than with novices.

Manual search results indicate that content-based features can improve traditional text search on transcripts. Visual examples were found to contribute slightly more than concept queries.

6. REFERENCES

[1] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, D. Steele and P. Yanker, "Query by Image and Video Content: The QBIC System," *IEEE Computer Magazine* 28(9), 1995, pp. 23-32.

[2] H.D. Wactlar, T. Kanade, M.A. Smith, and S.M. Stevens, "Intelligent Access to Digital Video: Informedia Project," *IEEE Computer*, 29 (5), pp. 46-52, May 1996. See also http://www.informedia.cs.cmu.edu/.

[3] H. Lee, A. Smeaton, C. O'Toole, N. Murphy, S. Marlow, and N. O'Connor, "The Físchlár Digital Video Recording, Analysis, and Browsing System," *Proc. of RIAO 2000*, Paris, France, pp. 1390-1399, April 12-14 2000.

[4] M. Rautiainen, T, Ojala, and T. Seppänen, "Analysing the performance of visual, concept and text features in content-based video retrieval," *6th ACM SIGMM International Workshop on Multimedia Information Retrieval MIR 2004*, New York NY, pp. 197-205, 2004.

[5] G. Salton and C. Buckley, "Improving retrieval performance by relevance feedback," *Journal of the American Society for Information Science*, vol. 41, pp. 288-297, 1990.

[6] M. Rautiainen, T. Ojala, and T. Seppänen, "Cluster-temporal video browsing with semantic filtering", *Proc. of Advanced Concepts for Intelligent Vision Systems*, Ghent, Belgium, pp.116-123, 2003.

[7] M. Rautiainen, T. Ojala, and T. Seppänen, "Content-based browsing in large news video databases," *Proc. of 5th IASTED International Conference on Visualization, Imaging and Image Processing*, Benidorm, Spain, 2005.

[8] D. Heesch and S.M. Rüger, "NN^k Networks for Content-Based Image Retrieval," *Proc. of 26th European Conference on IR Research,* Sunderland, UK, pp. 253-266, April 5-7, 2004.

[9] K. Rodden & K.R. Wood, "Searching and organizing: How do people manage their digital photographs?" *Proc. of the Conference on Human Factors in Computing Systems*, pp. 409 – 416, April 2003.

[10] M. Yeung, B.L. Yeo, W. Wolf, and B. Liu, "Video browsing using clustering and scene transitions on compressed sequences", *Proc. of Multimedia Computing and Networking*, pp. 399-413, February 1995.

[11] H.J. Zhang, J. Wu, D. Zhong, and S.W. Smoliar, "An integrated system for content-based video retrieval and browsing", *Pattern Recognition*, Vol. 30(4), pp. 643-658, Apr 1997.

[12] M. Rautiainen and D. Doermann, "Temporal color correlograms for video retrieval", *Proc. of 16th International Conference on Pattern Recognition*, Canada, pp. 267-270, 2002.

[13] M. Rautiainen, T. Seppänen, J. Penttilä and J. Peltola, "Detecting semantic concepts from video using temporal gradients and audio classification", *Proc. of Int. Conference on Image and Video Retrieval*, Urbana, IL, pp. 260-270, 2003.

[14] M. Rautiainen, M. Varanka, I. Hanski, M. Hosio, A. Pramila, J. Liu, T. Ojala, T. Seppänen, "TRECVID 2005 Experiments at MediaTeam Oulu", *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID 2005, Gaithersburg, MD, 2005. http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/oulu.pdf

[15] T. Ho, J. Hull, and S. Srihari, "Decision combination in multiple classifier systems," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* 16(1), pp. 66–75, 1994.

[16] *TREC Video Retrieval Evaluation Home Page*, URL: http://www-nlpir.nist.gov/projects/trecvid/.

[17] C. Petersohn, "Fraunhofer HHI at TRECVID 2004: Shot Boundary Detection System," *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID 2004, URL: www-nlpir.nist.gov/projects/tvpubs/tvpapers04/fraunhofer.pdf

[18] M. Rautiainen, T. Ojala, and T. Seppänen, "Cluster-temporal browsing of large news video databases," *Proc. of 2004 IEEE International Conference on Multimedia and Expo*, Taipei, Taiwan, pp. 751-754, 2004.

[19] K. Cox, "Information retrieval by browsing," In Proc. of the 5th International Conference on New Information Technology, Hongkong, 1992.

[20] I. Campbell and K. van Rijsbergen, "The ostensive model of developing information needs," *In Proc. of CoLIS 2*, Danmarks Biblioteksskole, Copenhagen, pp. 251-268, 1996.

[21] S. Santini, and R. Jain, "Integrated Browsing and Querying for Image Database," *IEEE Multimedia*, Vol. 7, No.3, pp.26-39, 2000.

[22] A. Smeaton and T. Ianeva, "Slides: Search Task Overview", *TREC Video Retrieval Evaluation Online Proceedings*, TRECVID 2005, Gaithersburg, MD, 2005. http://www-nlpir.nist.gov/projects/tvpubs/tv5.papers/tv5.search.slides.final.pdf