Físchlár-TRECVid-2004: combined text- and image-based searching of video archives

Noel E. O'Connor, Hyowon Lee, Alan F. Smeaton, Gareth J. F. Jones
Edward Cooke, Hervé Le Borgne, Cathal Gurrin
Centre for Digital Video Processing and Adaptive Information Cluster, Dublin City University, Ireland
Email: oconnorn@eeng.dcu.ie

Abstract— **The Físchlár-TRECVid-2004 system was developed for Dublin City University's participation in the 2004 TRECVid video information retrieval benchmarking activity. The system allows search and retrieval of video shots from over 60 hours of content. The shot retrieval engine employed is based on a combination of query text matched against spoken dialogue combined with image-image matching where a still image (sourced externally), or a keyframe (from within the video archive itself), is matched against all keyframes in the video archive. Three separate text retrieval engines are employed for closed caption text, automatic speech recognition and video OCR. Visual shot matching is primarily based on MPEG-7 low-level descriptors. The system supports relevance feedback at the shot level enabling augmentation and refinement using relevant shots located by the user. Two variants of the system were developed, one that supports both text- and image-based searching and one that supports image only search. A user evaluation experiment compared the use of the two systems. Results show that while the system combining text- and image-based searching achieves greater retrieval effectiveness, users make more varied and extensive queries with the image only based searching version.**

I. Introduction

This paper describes the Físchlár-TRECVid-2004 system developed for Dublin City University's participation in the TRECVid 2004 content-based information retrieval benchmarking exercise [1]. This paper focuses on our participation in the TRECVid interactive search task. This task can be summarized as follows: given the specified video test data, a set of query topics, and the provided common shot boundary reference (supplied by CLIPS-IMAG), the system should return a ranked list of shots which best satisfy the user's information need as expressed in each of the query topics. The search task test collection consisted of 64 hours of (MPEG-1) content from CNN Headline News and ABC World News Tonight broadcasts recorded during the second half of 1998. In addition to this, our system also used the Automatic Speech Recognition (ASR) transcripts supplied by LIMSI [2], Closed Caption (CC) transcripts from the broadcast and Optical Character Recognition (OCR) results on the video images and motion and face feature extraction results that were donated by Carnegie Mellon University (CMU).

The Físchlár-TRECVid-2004 system is a search/browse system based on the Físchlár Digital Video System [3][4]. Two variations of the system were developed for our experiments. System A provides text querying functionality and image-based relevance feedback, whereby the user initiates a query by typing some text and/or adding video/image examples that come with the topic. During the search, the user can include any keyframes determined as relevant from the video into subsequent queries. System B relied solely on searching based on keyframe images without any text-based querying. Thus the only way the user can initiate a search in System B is by including video/image examples in a search query panel. Thereafter, keyframes can be added to or removed from the query in an attempt to improve the search result. The system was used to conduct an interactive search experiment for the 25 topics provided for the 2004 TRECVid interactive search task. For our experiments, 16 experienced users carried out this search task each working under a time constraint of 15 minutes per topic.

The remainder of this paper is organised as follows. Section II presents an overview of the system, including descriptions of the text and image search engines and how the results of these are combined (sections II-A, II-B and II-D, respectively). The relevance feedback approach employed is described in section II-C. The user interface design is discussed in section II-E, whilst user searching is described in section II-F. The experimental set-up used and results obtained are discussed in sections III and IV respectively, whilst conclusions are presented in section V.

II. System Overview

The Físchlár-TRECVid-2004 system supports web-based remote access and has an XML-based architecture that uses MPEG-7 compliant video description schemes. The system builds on our previous work for TRECVid-2003 [5], however the 2004 system described here provides for much finer user control of individual visual features in searching. A key objective of our 2004 system being to enable us to study users' ability to make use of individual low-level visual features in interactive searching. A system overview is presented in figure 1 and the main processing blocks are described in the following subsections.

A. Text Search

To support text searching we used three sources of data, ASR, CC and OCR, and a separate search engine for each. All indexed text was processed to remove stopwords and stemmed using the Porter algorithm. Document terms are weighted using the standard Okapi BM25 algorithm [6], and a matching

0-7803-9390-2/06/$20.00 ©2006 IEEE

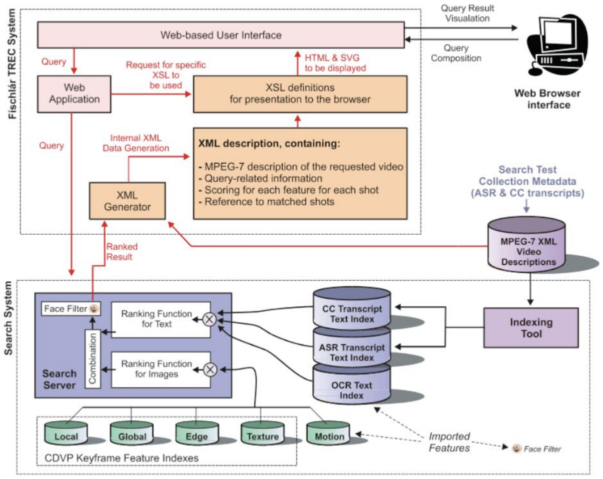

Físchlár TREC System
Web-based User Interface
Query Result Visualation
Query Composition
Web Browser interface
Query
Request for specific XSL to be used
HTML & SVG to be displayed
Web Application
XSL definitions for presentation to the browser
Query
Internal XML Data Generation
XML description, containing:
- MPEG-7 description of the requested video
- Query-related information
- Scoring for each feature for each shot
- Reference to matched shots
XML Generator
Search Test Collection Metadata (ASR & CC transcripts)
MPEG-7 XML Video Descriptions
Ranked Result
Search System
Face Filter
Ranking Function for Text
Combination
Ranking Function for Images
Search Server
CC Transcript Text Index
ASR Transcript Text Index
OCR Text Index
Indexing Tool
Local
Global
Edge
Texture
Motion
CDVP Keyframe Feature Indexes
Imported Features
Face Filter

Fig. 1. Físchlár-TRECVid-2004 System Overview

score is computed by summing matching terms. The document scores from each engine were summed to produce a final score for each document. Since many shot documents are short, the matching score is smoothed using the scores of neighbouring documents. 15% or 10% of the score of the first and second preceding and following shots are added respectively. The final outcome of the text search engine is a ranked list of 1,000 shots. Depending on the user requirements as expressed in the query, these were either used to rank shots directly before they are presented to the user, or combined with the results of an image search result to generate a final ranked list.

B. Image Search

MPEG-7 image features are automatically extracted from all keyframes to support image matching. The following four features are extracted – detailed descriptions can be found in [7] – using the aceToolbox [8], a toolbox of audiovisual analysis tools based on a cut-down implementation of the official MPEG-7 eXperimentation Model [9]:

- Colour Layout (CLD) – a compact and resolution-invariant representation of local colour in an image based on quantized DCT coefficients of small image blocks;
- Scalable Colour (SCD) – measures colour distribution over an entire image based on applying a Haar transform to a HSV histogram;
- Edge Histogram Descriptor (EHD) – measures the spatial distribution of edges by categorising the different kinds of edges found in image blocks;
- Homogenous Texture Descriptor (HDT) describes directionality, coarseness, and regularity of patterns in images by means of a bank of orientation and scale sensitive (Gabor) filters.

In addition, two other features donated by CMU were used, corresponding to shot motion and face detection. The latter enables the user to filter retrieved shots based on the presence/absence of a human face.

The similarity between images was estimated by the L2 Minkowsky (Euclidean) distance for each of the features. At query time, the user could select which (or all) of the five features were important via the interface and each of these features were combined together to produce a final feature ranked list. Separate ranked lists of the top 500 shots were generated for each feature and combined by summing the individual shot rank to avoid incompatibility issues associated with dissimilar score distributions in lists from different features.

C. Relevance Feedback

Relevance feedback is available for both text and image search results. For text, queries were expanded based on ASR and CC transcripts (with stopwords removed and stemmed) of the shots added to the query panel. The top 10 available expansion terms are selected using Robertson's offer weight [6] and added to the query. For images, the keyframe of the associated shot is simply used as another keyframe for the image matching engine to process.

D. Combining Image and Text Search

Depending on the user's query and the system used, image and text ranked lists were combined based on rank position to generate the final list of 200 top ranked shots that was presented to the user. In System B, this corresponded to the top ranked image results alone.

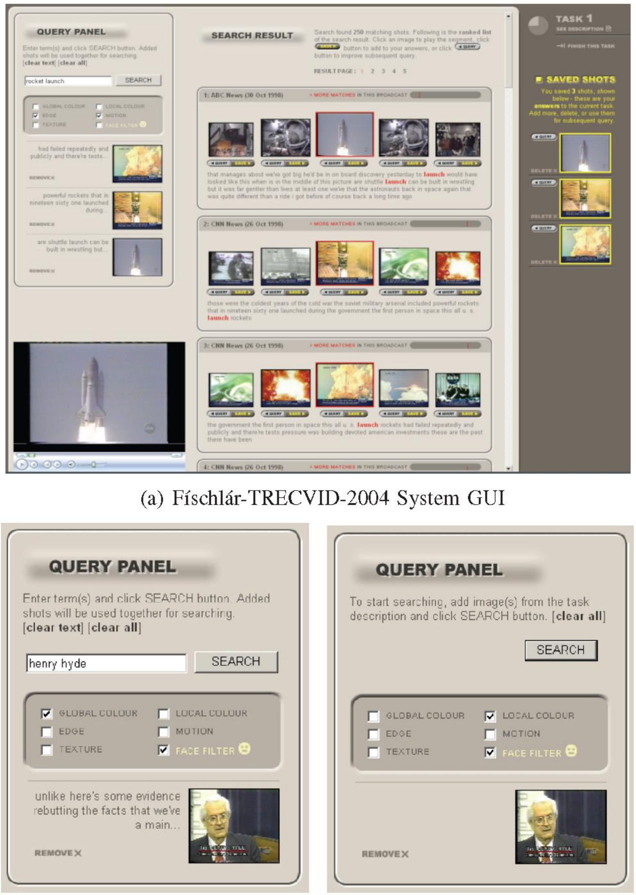
E. User Interface Design

Figure 2(a) shows the overall system interface for System A (the interface for System B is a subset of this, and is described below). Given the nature of the search task (searching for video shots within a limited time period) it was necessary to provide support in the interface for both administrative user interaction (i.e. task number, task description, time remaining, saved shots) and search/browse functionality (i.e. query panel, search results, playback). Thus, screen real estate is divided into two areas: administrative area and work area.

The work area is divided into a query panel area and a search result area. A consistent look and feel is employed throughout to enhance usability and user satisfaction. This includes 3-D buttons and bevelled lines employed to imply the possibility of interaction. As the user finds shots that satisfy the topic, they add them to the saved shot list on the right of the interface. Shots from this list can be incorporated in the query at any time, in addition to those in the current ranked list. The administrative area also contains a timer that indicates time remaining for this search task and the possibility to redisplay the topic description.

F. Search Procedure

In operation, the user is first presented with a topic description along with example videos/images. In System B, there is no text box (see figure 2(b) vs. 2(c)) and thus when the example videos/images are added to the query the user is required to check at least one of the checkboxes corresponding to the 6 visual features used for visual retrieval. Search results



Enter term(s) and click SEARCH button. Added shots will be used together for searching. [clear text] [clear all]

rocket launch

□ GLOBAL COLOUR ☑ EDGE □ TEXTURE □ LOCAL COLOUR ☑ MOTION □ FACE FILTER

had failed repeatedly and publicly and there're tests...

powerful rockets that in nineteen sixty one launched during...

are shuttle launch can be built in wrestling but...

Search found 250 matching shots. Following is the ranked list of the search result. Click an image to play the segment, click SAVE button to add to your answers, or click QUERY button to improve subsequent query.

1: ABC News (30 Oct 1998)

that manages about we've got big he'd be in on board discovery yesterday to launch would have looked like this when is in the middle of this picture are shuttle launch can be built in wrestling but it was far gentler than lives at least one we've that the astronauts back in space again that was quite different than a ride i got before of course back a long time ago

2: CNN News (26 Oct 1998)

those were the coldest years of the cold war the soviet military arsenal included powerful rockets that in nineteen sixty one launched during the government the first person in space this all u. s. launch rockets

3: CNN News (26 Oct 1998)

the government the first person in space this all u. s. launch rockets had failed repeatedly and publicly and there're tests pressure was building devoted american investments these are the past there have been

4: CNN News (26 Oct 1998)

You saved 3 shots, shown below - these are your answers to the current task. Add more, delete, or use them for subsequent query.

(a) Físchlár-TRECVID-2004 System GUI

Enter term(s) and click SEARCH button. Added shots will be used together for searching. [clear text] [clear all]

henry hyde

☑ GLOBAL COLOUR □ EDGE □ TEXTURE □ LOCAL COLOUR □ MOTION ☑ FACE FILTER

unlike here's some evidence rebutting the facts that we've a main...

To start searching, add image(s) from the task description and click SEARCH button. [clear all]

□ GLOBAL COLOUR □ EDGE □ TEXTURE ☑ LOCAL COLOUR □ MOTION ☑ FACE FILTER

(b) System A

(c) System B

Fig. 2. Interface Design

are presented in the middle column of the search area. The keyframe of the retrieved shot is shown in the centre with a highlighting red bounding box. The preceding and following shots immediately neighbouring it are also displayed at smaller sizes to show its context. The text in the ASR transcript associated with the shots is shown below the keyframes with words matching the query highlighted. Each entry displays the name and date of the broadcast, and also the approximate location of the matched shot within the broadcast. Given a relevant shot in a particular broadcast, it is quite likely that there will be another one nearby. For this reason, the system provides functionality for the user to view all matches within a particular broadcast. This corresponds to a compromise between full browsing of a broadcast (which complicates the user's task) and just browsing search results (which may lead to relevant shots being missed). At all levels of browsing, the user has the possibility to add to or remove a shot from the query or the saved shot list.

III. Experimental Set-up

For our experiments, 16 test users were recruited from within the University, excluding the system developers. All users had high levels of computer experience, but varying levels of experience with information retrieval systems. Prior to the experiments, participants were provided with remote and unsupervised access to the system in order to familiarise themselves with both system variants. Furthermore, each user was provided with the opportunity to complete training searches before the actual experiment. Thus, we can classify our participants as experienced users. Each user completed 12 query searches with 15 minutes allocated for the completion of each one. Users were assigned to a randomly sequenced set of topics. For the first six queries they used one system and the other six used the other system, with another user being assigned the same topic groups in random order using the systems in reverse order. Users completed pre- and post-search and post-experiment questionnaires. After each topic search was completed, all saved shots were collated and submitted to the TRECVid coordinators for manual judgement of the relevance of each saved shot.

IV. Results and Discussion

A general description of the TRECVid experiments, including an indication of how the performance of participants' systems compared to each other is available in [10]. In this section, we focus on a detailed comparison of our two system variants.

Retrieval results are presented in Table I, whilst selected user interaction log data is documented in Table II. Overall, System A outperformed System B and this was to be expected. The average mean average precision (MAP) for System A was over twice that of System B. The total number of shots our users found as relevant (see Table II) whether correct or not was almost twice higher with System A than with System B, supporting the recall figures of Table I. However the average deviation about the mean for System A at 0.023 (not shown) is over twice that of System B at 0.011 (not shown), suggesting that System A emphasizes variances in user ability more than System B. This could also be due to the fact that System B retrieves fewer relevant shots since if we examine MAP distribution as a function of the number of relevant shots found then the differences between the average deviations disappear. Average recall clearly illustrates that not using text reduces the recall of System B to 47% of System A (taking the median as opposed to the average gives a value of 46% which suggests that outliers do not affect this). For 3 of the 24 topics, System B outperformed System A (topics 140, 142 and 144). This can be considered to imply that searching for visually striking shots is thus more amenable to image only based retrieval. System A significantly outperformed System B (average MAP <25% that of System A) on 7 topics that typically correspond to searching for shots with no distinctive visual features.

The number of searches and refinements for each query was higher for System B than System A in the given 15-minute period, partly indicating the quicker and more experimental nature of interaction with System B. This is also indicated the considerably higher number of query images added and deleted with System B compared to System A ('Interaction with query panel' in Table II). Users took more time at each search iteration with System A, spending longer in formulating

TABLE I

RETRIEVAL RESULTS FOR SYSTEM A (TEXT AND IMAGE) VS. SYSTEM B (IMAGE ONLY)

	System A				System B			
Run	MAP	MP@rel	P@10	Recall	MAP	MP@rel	P@10	Recall
1	0.203	0.532	0.683	0.222	0.066	0.182	0.506	0.082
2	0.190	0.481	0.665	0.181	0.094	0.250	0.504	0.097
3	0.191	0.506	0.652	0.180	0.088	0.215	0.395	0.083
4	0.133	0.397	0.564	0.139	0.074	0.195	0.314	0.077
AVG	0.179	0.479	0.641	0.181	0.081	0.211	0.430	0.085

TABLE II

INTERACTION STATISTICS

	System A (#)	System B (#)
Finding relevant shots		
Saved shots	2681	1483
(added)	2733	1532
(removed)	303	614
Searching		
Total searches	1112	1254
Searches with text	714	N/A
Interaction with query panel		
Topic example image added	469	791
Shot keyframe added	81	45
Image removed from query panel	303	614
Image feature usage		
Local Colour	501	798
Edge	482	838
Global Colour	355	606
Texture	274	447
Motion	261	312
Face	163	242
Browsing		
Within broadcast results viewed	244	266
Full broadcast viewed	131	140
'Next 20 results' requested	765	1418

the text query and looking at the text transcript search results. In System B, query formulation in the form of feature checkboxes was a quicker and a more frequent process. The number of features used as part of a query ('Image feature usage' in Table II) shows considerable difference between the two systems, with 56% more use on average in System B. Trying various combinations of these features was observed as the major component of the querying behaviour in System B. The number of times the users viewed the next pages of a search result was considerably higher in System B than System A, indicating the less precise results of feature-based searches and our users' subsequent wishes to see more results. In System B, in the absence of accurate text-based matching, image-based querying followed by more comprehensive browsing was required. As a result, users conducted more frequent within-broadcast browsing as well as much more search result paging (to next 20 results, next 20 results, etc.).

V. CONCLUSION

In this paper we described the system built for our participation in the 2004 world-wide TRECVid benchmarking activity for video shot retrieval and how we evaluated content-based information retrieval using two system variants – one that supports combined text- and image-based search and the other image-only search. Whilst the result of the evaluation of the two system variants may appear self-evident, i.e. combined image- and text-based searching is significantly better than image-based searching alone, this experiment has allowed us to quantify how significant this difference actually is. In addition, it has provided us with an insight into the more experimental manner in which users formulate queries using image features only.This will be important in the context of developing video shot retrieval for video data where ASR, CC and OCR text is not available, such as CCTV or home movies.

ACKNOWLEDGMENT

Part of this work was based on work supported by Science Foundation Ireland under grant number 03/IN.3/I361. The support of the Informatics Directorate of Enterprise Ireland is also gratefully acknowledged. Part of this work was also supported by the European Commission under contract FP6-001765 aceMedia (URL: http://www.acemedia.org).

REFERENCES

[1] P. Over. (2004) TRECVID2004 guidelines. [Online]. Available: http://www-nlpir.nist.gov/projects/tv2004/tv2004.html

[2] J. L. Gauvain, L. Lamel, and G. Adda, "The LIMSI broadcast news transcription system," *Speech Communication*, vol. 37, pp. 89–108, May 2002.

[3] H. Lee and A. F. Smeaton, "Designing the user-interface for the Físchlár digital video library," *Journal of Digital Information*, vol. 2, no. 4, May 2002.

[4] A. F. Smeaton *et al.*, "The Físchlár-news-stories system: Personalised access to an archive of tv news," in *Proc. RIAO 2004 - Coupling Approaches, Coupling Media and Coupling Languages for Information Retrieval*, Avignon, France, Apr. 2004.

[5] C. Gurrin, H. Lee, and A. F. Smeaton, "Físchlár TRECVID2003: System description," in *Proc of the* 12^{th} *ACM International Conference on Multimedia*, New York, NY, Oct. 2004, pp. 938–939.

[6] S. E. Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford, "Okapi at TREC-3," in *Proceedings of TREC-3*, D. K.Harman, Ed., 1995, pp. 109–126.

[7] B. S. Manjunath, J. R. Ohm, V. Vasudevan, and A. Yamada, "Color and texture description," *IEEE trans. on circuits and systems for video technology*, vol. 11, no. 6, June 2001.

[8] N. E. OConnor *et al.*, "The aceToolbox: low-level audiovisual feature extraction for retrieval and classification," in *Proc of the* 2^{nd} *European Workshop on the Integration of Knowledge, Semantic and Digital Media Technologies (EWIMT 04)*, London, U.K., Nov. 2005.

[9] *Visual eXperimentation Model (XM) version 10.0*, MPEG-7 Std. ISO/IEC/JTC1/SC29/WG11, N4062, 2001.

[10] W. Kraaij, A. F. Smeaton, P. Over, and J. Arlandis. (2004) TRECVID2004 - an overview. [Online]. Available: http://www-nlpir.nist.gov/projects/tvpubs/tvpapers04/tv4overview.pdf