VIDEO INFORMATION RETRIEVAL USING OBJECTS AND OSTENSIVE RELEVANCE FEEDBACK

Paul Browne and Alan F. Smeaton
Centre for Digital Video Processing
Dublin City University
Glasnevin, Dublin 9, Ireland
{ Paul.Browne, Alan.Smeaton } @computing.dcu.ie

ABSTRACT

In this paper, we present a brief overview of current approaches to video information retrieval (IR) and we highlight its limitations and drawbacks in terms of satisfying user needs. We then describe a method of incorporating object-based relevance feedback into video IR which we believe opens up new possibilities for helping users find information in video archives. Following this we describe our own work on shot retrieval from video archives which uses object detection, object-based relevance feedback and a variation of relevance feedback called ostensive RF which is particularly appropriate for this type of retrieval.

1. Introduction To Video IR

Two very important areas for video information retrieval (IR) research are visual feature extraction and retrieval evaluation.

In the area of feature extraction, current approaches are still not very accurate for many medium and high level feature types. Features can be broadly divided into low level and high level:

*Low-level feature*s examine low level information from video content like colour, edge and motion. These contain a small amount of semantic information and *individually* provide poor discriminatory power, such as for example, a colour histogram.

High-level features indicate semantic information and have good discriminatory power. A low-level feature could say that video segment α contains a human face whereas a high-level feature would say the human face in α belongs to Mr T.

The main high-level features currently in use by video IR systems which are not restricted to a specific genre are speech transcripts and these can be generated automatically from spoken audio or from closed caption information. A number of genre-specific features like object detection can also be extracted for domain specific video content like television news, sports and cartoons.

SAC' 04, March 14th, 2004, Nicosia, Cyprus.

For the last 3 years the Text Retrieval Conference (TREC) has evaluated video IR tests and. during the past 3 years of TREC the number of groups participating in the video track (TRECVID) has grown considerably as has the scale and challenges for each of the groups. Current evaluated tasks include shot boundary detection, news story detection, automatic and real user query topic search. TRECVID is important for evaluation of video IR and we shall return to it later in this paper.

2. The Role of Video Retrieval

There have been almost no studies of the kinds of topics or queries which are submitted to current operational video information retrieval systems. What information available anecdotally from video libraries in TV archives and national depositories shows user topics varying from precise to vague, and from the abstract to the specific. The VIRAMI (Visual Information Retrieval for Archival Moving Imagery) study showed that specifically named persons, places, objects and events were the common needs of the clients of eleven film archives [Enser & Sandom 2002], similar to those of still-image archives a number of which have been studied previously [Markkula & Sormunen 1999] [Armitage & Enser 1997. From these results the early conclusion has been that the current approaches to content-based retrieval have very limited real-world application [Enser 2000]. An early study on the RTE Broadcast archive by ourselves showed that among 32 user searches analysed which had been requested to the broadcast news archive, requests for persons (15) and locations (16) and activities (16) were the most common whereas time (3), camera effects (1) and low-level visual features (0) were not.

This variety of user information needs has been used as the catalyst for the development of topics in each of the three years of TRECVID and as a result we have TRECVID topics like "Rockets taking off" or "city rooftops" or "shots of one or more groups of people " or "the Golden Gate Bridge". It is possible to develop video analysis techniques which can detect features such as indoor, outdoor, urban setting, female speech, text overlay or fire and these can be used individually or in combination to help narrow down a search for some query types such as *rockets taking* off (outdoor+fire) or *city rooftops* (outdoor+cityscape). However, if the video information retrieval research community uses these broad requirements of current video information retrieval as the driver for developing video IR research systems then we have a very long way to go if we want to automatically search for

abstract concepts like "government plan to improve reading standards" or "bullying at school".[1]

The few indicators of video IR needs that we do have to act as drivers for our research, do come from video archive environments but with the growth in availability of digital video, searching TV archives will be only one of many video databases. The availability of digital camcorders in the mass market, and mass market access to digital forms of TV (TiVo boxes) and movies (via DVDs) will create new user demands and we simply don't know what kinds of video navigation demands are or will be. What we do know is that the user needs of the mass market will not be like the user needs of the TV archives searchers who currently search through meticulously indexed content with closed indexing vocabularies and structured, home-grown ontologies.

Real people in the real world, doing real information seeking and in a hurry, use web search engines and give 2-word queries to be run against billions of web pages. We expect, and get, sub-second response time and we complain when there are no relevant web pages in the top 10 presented to us. The (text) information retrieval community (including ourselves) have developed sophisticated techniques for query expansion, relevance feedback, post-retrieval clustering, visualisation of results, passage retrieval and many others which have all shown to yield more effective retrieval when used but unless a technique fits into the simple model of *2-word query as input and document list as output*, it won't be used in mainstream web searching. Some other (text) information retrieval research such as links based retrieval, pseudo relevance feedback, document length normalisation and other term weighting strategies are used in mainstream web searching, but only because they fit the simple interaction model.

So what does this tell us about video information retrieval; it tells us that the level of interaction from users in video IR must be low and we must develop video IR techniques to match this.

Current video IR research systems are able to perform automatic structuring of video into shots with keyframes [Boreczky *et al*. 2000], [Yeo & Yeung 1997], or higher-level domain-specific units such as stories in TV News [Pickering *et al*. 2003] [Ide *et al*. 2003]. Current video IR systems use manual metadata about their video contents (date, location, actors, etc.) to help users search. Current video IR systems can use automatic speech recognition (ASR) or closed captions taken from video and text search based on this can be successfully combined with keyframe browsing [Christel & Warmack 2001] [Smeaton *et al*. 2002]. Video IR systems can also use features automatically extracted from video such as the TRECVID features and combine these with text search through ASR or closed captions, and keyframe browsing [Smeaton & Over 2003] [Browne *et al*. 2002]. While all these developments are significant advances and represent the state of development of this field, they all place the burden of interpretation of the information need upon the user who is required to turn a need for shots about *bullying at school* into a text search for "school children bully yard playtime harass" against captions and where the desired shots also have the features OUTDOOR, BUILDINGS, PEOPLE. .

[1] examples of real search requests submitted to the BBC archive.

In all our work on video IR, and as far as we know the work of others, one of the problems we've not been able to address is capturing the user's real information needs and especially as this evolves during a search session. It is a well-accepted fact in information seeking that users' information needs will evolve as their search session progresses [Ingwersen, 1992]. This is partially due to the searcher learning more about the area as they view material and partially due to them clarifying in their own mind what they are looking for. The technique used to redress this in text IR is relevance feedback, but in video IR where the searcher is already burdened with having to do so much work and interpretation as well as query formulation, how can this be incorporated in a way which is simple ? In other words, how can an already-stressed user indicate what it is about a video clip which makes it relevant in as easy a fashion as possible ? When viewing a video clip on a computer, the most natural way to feedback relevant facets of a video clip is to point and click. Point and click on an entire keyframe is too generic, so point and click on an *object* is what is needed.

In this paper we report on work we are doing which provides video information retrieval in a way not dissimilar to other contemporary video IR research systems but we address some of the previously untackled problems that video searchers have by including a type of relevance feedback called ostensive relevance feedback which models shifting information needs. We also incorporate a facility whereby a user can indicate what explicitly within a shot makes it relevant by clicking on the relevant *object*.

3. The Simpson's and the User's Information Need

The real information need of users normally extends well beyond the capabilities of traditional visual retrieval which uses colour and edge matching and into the area of object based retrieval. In many cases the user will be searching for specific objects. So why not automatically index all the objects in the video content ? The task of automatic object based detection is extremely difficult in the narrow sense and computationally infeasible in the broad sense, and there is as of yet no method of object detection across wide domains. Current object based retrieval needs to focus on narrow tasks with a reasonable level of system training in order to be feasible. In order to meet real users information needs in object-based video retrieval it is necessary to select a narrow domain. For the purposes of our experiments it was decided to use the animated cartoon series of the Simpsons as our video content.

Why select animated content ? Cartoon animation is a good choice for object-based detection as the character objects generally have very distinct outlines making detection, identification and tracking easier. Edges can be identified with more success in animated content than natural, and this is vital for object identification. Motion is also less of a problem here as there is a smaller amount of camera movement (panning, zooming, etc.) in comparison with natural video content.

"The Simpsons" is one of the most popular TV programmes and has a loyal following all over the world transcending social, religious and political differences. It has been running for just over fourteen years with over 300 23-minute episodes produced. This total content comes to about 115 hours.

There are currently three seasons of Simpsons content available on DVD (each season is 22 episodes) and a number of special themed releases. Closed captions can be extracted automatically from the DVD content. For the purposes of our experiments we extract each individual episode from the DVD, transcode into MPEG-1 and extract the closed caption information. Shot boundary detection was run on the content and each I-frame (JPEG Image) was extracted.

The object detection we developed is based on computing a similarity score between any two shapes and when used with predefined character templates for the Simpsons characters it makes object identification possible. For retrieval we have created a number of predefined templates for the ten main Simpsons characters in their various poses. Object detection works by comparing the edges in a candidate image with those templates.

In developing a search system for Simpsons content which incorporates the techniques we have described earlier, it is important to realise that the search task is to find a shot which the user has most likely not previously seen. A search task of finding a previously viewed shot would be a simple case of recall of previously viewed material and that search scenario would be satisfied using straightforward keyframe browsing and metadata searching. In this search task a user expects to have to browse and navigate a lot for each search as the ideal shot is what is needed so users are encouraged to take their time, get high precision, and the searches mostly involve finding characters or character combinations in particular settings. One example of a search would be for "Marge and Homer falling out or fighting" and a relevant shot might or might not have one or both of these characters present.

4. Ostensive RF for Shot Retrieval

One technique in text-based IR which has consistently proved to yield improved retrieval effectiveness is relevance feedback [Belkin et al, 1996]. Allowing a user to make relevance judgements on video can help compensate for the lack of features' discriminatory powers especially in the visual domain where high-level feature extraction remains very difficult and works only in narrow areas. Shot retrieval with relevance feedback requires an initial relevance judgement before results are provided but once a judgement is made, shots can be re-ranked based on their similarity to the judged shot.

In relevance feedback, a user's query can be continuously refined. As *each* shot is given a relevance judgement, remaining results are re-ranked automatically. With *each* feedback iteration the user will view results and ask the question "Do the results answer my information need?" and the user can select a shot as being relevant or non-relevant. The following shows this process:

(1) Query is submitted

(2) Shots are ranked and displayed to the user.

(3) User marks a shot as relevant / non-relevant.

(4) The browsing view changes based on query and included relevance judgements. Are more shots needed?

IF (YES) <GOTO 2>

ELSE (NO) <EXIT>

The idea behind ostensive relevance feedback (O-RF) which differentiates it from the more commonly used traditional relevance feedback, is that it recognises and incorporates the fact that as people search for information over time, their initial information requirement also changes and evolves. Searchers normally start out with an unrefined or vague information need which becomes more sharply focused as their search continues and exposure to information changes their information need. Previous research on O-RF reported in [Campbell, 2000] used textual content from a French test corpus of images with retrieval based on manually annotated image descriptors.

O-RF requires a method for weighting the decay corresponding to the degree of importance of previously viewed relevant items as searching progresses. This is because in O-RF the ranking of unseen objects at any point in time is a function of the original query plus all the viewed and relevant objects seen up to that point in the search. The contribution that a previously viewed and relevant object makes to the computation of the ranking of other objects requires a decay function (OD) to weight the similarity between that viewed object, and each of the objects to be ranked. Provided that a non-zero similarity value exists between the ranked object and the viewed object, the weight must also take into account the relative point in the search at which the relevance judgement was made.

5. Shot Comparison

In tailoring O-RF to video IR and retrieval of video shots, we need an accurate method of computing similarity between entire shots and this presents us with a number of challenges

1. Each shot in video consists of a number of still images and it is these *sets* of images that need to be compared.
2. Each still image can be indexed by a number of visual features such as colour or the presence of specific objects, which need to be combined.
3. All shots with relevance judgements made need to be compared against each *individual* candidate shot from the corpus. Thus if a user has judged 10 shots and 10,000 shots remain to be ranked, that is 10*10,000 shot comparisons.

To compare shots with a varying number of images there are a number of methods that one can employ, and basically the aim is to reduce numerous still images to a representative form to reduce comparison complexity. The following are some possibilities:

- Aggregate and average the feature information for each image from both shots and compare using the averages.
- Aggregate and average the feature information for each image from both shots and find the still image closest to the average for both shots, and compare these directly.
- Take each image from the candidate shot and compare against each image from the shot with the relevance judgement, aggregate and average comparison values.
- Compare each image from both shots as before but remove some such as the highest and lowest 10% before averaging or use only the top 10% of image comparison scores.
- Take the middle image from each shot as the visual representation.

Having evaluated these alternatives on the queries described in section seven it appears that the final option offers higher performance and thus is the method used in our work.

Apart from the fact that each shot contains many still images there are also numerous other visual features in video which can be extracted and used in shot comparison. Normalizing *the differing* feature values can be accomplished fairly crudely by finding an average comparison score for each of the required features. This works in practice by dividing Feature1's average score by some value X, multiplying Feature2's average score by some value Y and thus we bring all the Feature scores averages to 'roughly' the same level. Feature comparisons can be normalized using this method.

We have an additional requirement that *each* candidate shot in the corpus is compared against *all* shots with relevance judgements, only after this can overall comparison similarity totals be computed ($Total^{Text}$ and $Total^{Visual}$). Calculation requires that for each candidate shot a comparison score is obtained for each relevant (X_{rel}) and non-relevant ($X_{non\text{-}rel}$) shot judgement with the result placed in one of two totals; added to the relevant or non-relevant total. When all shot judgements have been compared against the candidate shot the two totals are averaged with the overall total made up of the non-relevant total subtracted from the relevant total. The following formula shows how $Total^{Visual}$ is calculated using a shot comparison score() function:

C : Index shot Corpus , Cj , Ci $\notin X_{rel}$, Ci $\notin X_{non\text{-}rel}$

X_{rel} : Relevant Judgement Set, $X_i^{rel} \in X_{rel}$

k : Relevant Judgement number

$X_{non\text{-}rel}$: Non-Relevant Judgement Set, $X_i^{non-rel} \in X_{non\text{-}rel}$

L : Non-Relevant Judgement number

$$Total^{Visual}(C_j) = \left(\frac{\sum_{i=1}^{k} \frac{score(X_i^{rel}, C_j)}{OD_i}}{k} \right) - \left(\frac{\sum_{i=1}^{l} \frac{score(X_i^{non-rel}, C_j)}{OD_i}}{l} \right)$$

A text and visual shot comparison rank is done by the following:

$$Rank(C_j) = W_1 * Total^{Text}(C_j) + W_2 * Total^{Visual}(C_j)$$

The remaining issue that needs to be considered is the ostensive nature of relevance judgements. In section four we discussed how a users information need changes over time and this needs to be reflected in the generation of the shot(s) to shot comparisons.

This is accomplished by weighting relevance shot judgements based on the order (time) in which they are made. The latest shot judgement will have the highest weighting while the earliest judgement will have the smallest. Shot comparisons involving each judgement will now be weighted based on their order.

There are a number methods in which this ostensive shot weight decay can be obtained from logarithmic based to linear and these will require real user evaluation in order to find an optimal solution. The following are two examples of shot decay weighting.

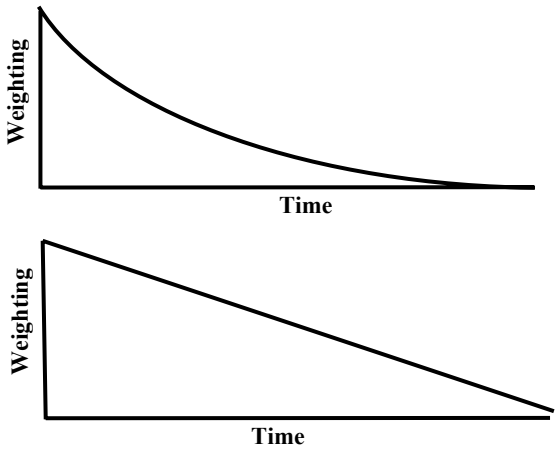

Figure 5.1 Sample Shot Weight Decay Graphs

6. Fischlár-Simpsons

To evaluate ostensive relevance feedback and object retrieval a video IR system has been designed and built for Simpsons based shot retrieval. The general visual comparison features of this system can be used for retrieval any content type with an additional Simpsons specific feature of object based retrieval.

The current system is broken into three sections, Search & Options, Browse Results and Browse Shot Context. The first section (Figure 1(a)) of the system facilitates initial query input or query expansion. Textual terms can be added to search the indexed closed captions while the drawing query section allows the user to sketch the content they are interested in.

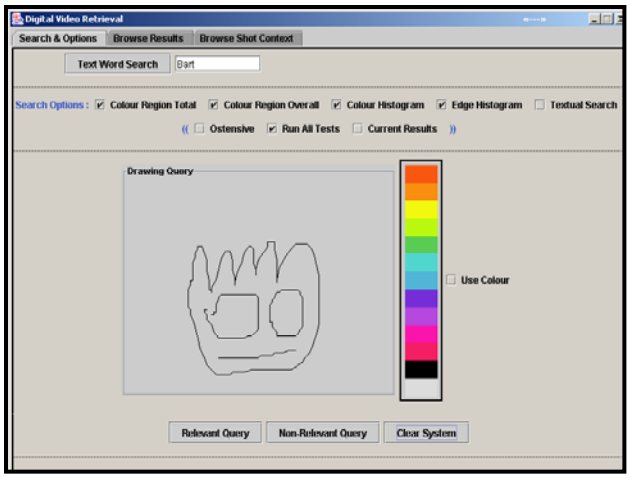

Fig 1(a): Search & Options Pane

When the user modifies the query the "Browse Results" section is automatically updated (Fig 1(b)). From here the user can select a shot as relevant or non-relevant to the query with a left or right mouse click. Clicking the middle mouse button allows the user to browse the context of a shot result (Fig 1(c)).

Fig 1(b): Browse Results Pane

From the "Browse Shot Context" pane shown below, the previous four shots and the next six are shown. The current shot is displayed at full size with object information displayed where available. The user can select object(s) displayed with the white borders in that full size current shot to be used to expand the query as in Fig 1(c).

Fig 1(c): Browse Shot Control

There are a variety of possible methods for generating and/or expanding a user query with the system:

1. User does a sketch of their query need (see Fig 1a).
2. As above with the addition of colour to the sketch query
3. Region Colour is used solely as the query input
4. The user types in text keywords
5. The content is browsed and complete shot(s) are selected as relevant or non-relevant
6. The context of a shot is browsed and an object within a highlighted shot is selected as part of the search criterion (see Fig 1c).

The following is a sample worked query scenario where the user's topic is to find content *of the Simpsons character Homer who is outdoors on a cloudy day and with a scared facial expression:*

- The user makes an initial query by selecting the colour blue.
- User presses the "submit query" button.
- Results are ranked.
- The user views the first 20 shots.
- The user moves onto the next 20 shots by clicking next.
- A shot of Homer is visible outdoors so the user selects the complete shot as relevant with a left mouse click.
- Results are re-ranked
- A shot of clouds is one of the shots shown to the user
- The user selects this shot as relevant.
- The results are again re-ranked
- One shot shows Homer on a skateboard looking scared.
- Select this as relevant.
- The results are re-ranked again
- The user browses the shot context of the first shot presented
- When viewing the shot context, the user selects the Homer object displayed as relevant
- The results are re-ranked again……..

7. System Evaluation

There are many aspects to evaluation of the work we have described in this paper. A full, thorough evaluation must incorporate everything from the accuracy of object identification and low level feature extraction during the indexing phase, to eventual end-user satisfaction. Information retrieval evaluation is dominated by the TREC paradigm which pays importance to precision and recall during retrieval and incorporates indexing accuracy as part of that. In our work we perform a more step-wise evaluation, and measure the effectiveness of different components as well as performance of the whole.

In order to evaluate the performance of the system during development, 10 narrow query topics were chosen for retrieval of Simpsons content using the following four low level features:

- Regional Average Colour (9 Region * 3 Colour RGB)
- Regional Largest Colour (9 Region * 3 Colour RGB)
- Regional Hue Histogram (4 Region * 18 Colour bin)
- Regional Edge Histogram (4 Region * 16 Edge bins)

The reason for selection of narrow topics is that some might have valid results in over half the corpus of shots (a search for Homer for example). This part of the evaluation was developed to test system integration and the application of low-level visual feature extraction rather than object retrieval. A baseline of three hours (4868 shots) was searched for shots matching each of the 10 query topics listed below. As matching shots were found they were added to a list of valid shot numbers for each of the topics

creating a topic baseline. The following are the 10 query topics that were used:

1. Find Content containing Mr Smithers
2. Dr Hibbert
3. Grandpa Abe Simpson
4. Itchy and/or Scratchy (Cartoon Cat and Mouse)
5. Milhouse (Bart's friend)
6. Krusty the clown
7. Skateboards
8. Selma and/or Patty (Marge's sisters)
9. Cars (trucks and vans are also acceptable)
10. Newscaster Kent Brockman

As query input for *each* of the 10 topics *four* valid or relevant shots were found. These four shots were selected based on realistic relevance feedback searches using the system during an early construction phase and they were represented using their keyframes. In order to give an idea of the system RF performance, the 10 query topics represented as the associated four shot judgements were each fed into the system and the results compared with the marked up ground truth. The relevance feedback approach described in section five was used without ostensive shot weighting applied.

So how does this work in practice? For *each* topic, each keyframe of the four query input shots is compared against the keyframe from each shot in the full corpus. At each of the four iterations (4 query judgements) the dissimilarity score for all of the 4868 shots was increased by some amount by adding the new score to the total. When complete, the scores were ranked in ascending order and the position of shots compared with the topic baseline.

For the purposes of shot keyframe comparison, four visual features were used, 4 * Regional Colour Histograms, 9 * Regional Largest & Average Colours and 9 * Regional Edge Histograms. While a keyframe from the middle of a shot was used during comparison, we did try alternatives such as shot frame averaging but found that the middle keyframe performed best.

As described in section 5 visual feature normalization was done crudely by finding an average comparison score for each feature, dividing the Colour average Histogram score by some value X, multiplying the average Edge Histogram score by some value Y and thus we bring all the score averages to 'roughly' the same level. For future comparisons using the four combined visual features each Colour Histogram score is divided by X while the Edge Histogram score multiplied by Y.

As changes are made to the system to vary the decay function for O-RF for example, one can get an idea of the performance 'benefit' or 'reduction' in terms of retrieval. Future work will include performance evaluation of ostensive relevance feedback by taking inputs from real user searches, applying ostensive shot weighting and comparing a non-ostensive weighting scheme and comparing performance results.

Looking at the results for each topic in this baseline as presented in Table 1, there are three main sections. The first is the recall and average precision for the first 100 ranked results, the second is the recall and precision for the first 1000 ranked results while the final section is the recall and precision over all documents (4868 shots). These results are based on standard relevance feedback using the four low-level visual features, four pre-selected shot judgements and 10 query topics described earlier.

Topic	Doc at 100		Doc at 1000		Over all Docs	
	Recall	Avg Prec	Recall	Avg Prec	Recall	Avg Prec
1	16	0.187	49	0.061	74	0.041
2	7	0.714	31	0.161	54	0.093
3	17	0.353	39	0.154	52	0.115
4	22	0.454	28	0.357	58	0.172
5	8	0.375	17	0.176	42	0.071
6	9	0.889	21	0.381	34	0.235
7	13	0.385	24	0.208	46	0.109
8	23	0.391	41	0.219	56	0.161
9	22	0.818	70	0.257	124	0.145
10	16	1.000	20	0.8	28	0.571
AVG	15.3	.557	34	.276	56.8	.171

Table 1: Ten Topic Results

At present we are about to run evaluation experiments of the whole system using a larger set of user topics than the sample topics described in this paper. We have build object detectors for Simpsons characters and run these against the test database and are evaluating its accuracy and we will use this as the basis for automatic indexing of the content. In the user experiments we will use real users to perform interactive searching using object-based relevance feedback, with, and without, ostensive relevance feedback on user generated queries.

8. Conclusions

One valid criticism of the work we have presented here is that the application domain is somewhat contrived, in that there are not many people who really want to search through the Simpsons archival material. This is true, but this work is meant to be a stepping-stone towards the kind of video information retrieval systems we want to build which incorporate object-based interaction between user and video, and which utilise a more thorough cognitive model of user behaviour and needs than is currently used in mainstream (text) information retrieval model.

The approach suggested here could be applied on any content domain provided that the relevant object information can be extracted. Object extraction is easier for animated material and is naturally the best domain type to evaluate hence our selection of 'The Simpsons'.

At the time of writing we are about to commence evaluation of interactive user searching and we expect to have results to present by the time of the SAC workshop. Our expectations are that the use of point-and-click relevance feedback will yield faster searching (namely users getting to a point of satisfaction with the

shots that they have found much quicker as measured in elapsed time). We also believe that ostensive relevance feedback will result in higher quality retrieval, though not necessarily faster retrieval. Using O-RF should help users to browse through sections of the archive which will contain more query-relevant material, resulting in more relevant shots being found, though we expect this benefit to occur only when O-RF is combined with object-based relevance feedback.

9. Acknowledgements

The support of the Informatics Directorate of Enterprise Ireland is gratefully acknowledged. Part of this material is based upon work supported by the IST programme of the EU in the project IST-2000-32795 SCHEMA.

This material is based on works supported by Science Foundation Ireland under Grant No. 03/IN.3/I361

10. REFERENCES

[1] [Armitage & Enser , 1997] Armitage L & Enser P.G.B. Analysis of user need in image archives. Journal of Information Science 23(4), 1997, 287-299.

[2] [Belkin *et al*. 1996] Belkin N.J *et al*, Using relevance feedback and ranking in interactive searching. Procs. 4th Text Retrieval Conference, NIST Special Publication 500-236. Pp.181-210. 1996.

[3] [Boreczky *et al*. 2000] Boreczky, J., Girgensohn, A., Golovchinsky, G., and Uchihashi, S. An interactive comic book presentation for exploring video". Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI 2000), The Hague, The Netherlands, April, 2000, 185-192

[4] [Browne *et al*. 2002] Browne, P., Czirjek, C., Gurrin, C., Jarina, R., Lee, H., Marlow, S., Mc Donald, K., Murphy, N., O'Connor, N., Smeaton, A. and Ye, J. Dublin City University Video Track Experiments for TREC 2002. TREC 2002 - Text REtrieval Conference, Gaithersburg, Maryland, 19-22 November, 2002.

[5] [Campbell, 2000] Campbell I, Interactive evaluation of the Ostensive Model using a new test collection of images with multiple relevance assessments, *J. of Information Retrieval*, 1, 2000. pp 85-112.

[6] [Christel & Warmack, 2001] Christel, M. & Warmack, A. The Effect of Text in Storyboards for Video Navigation. IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), Salt Lake City, UT, May 7-11, 2001.

[7] [Enser & Sandom, 2002] Enser, P.G.B. & Sandom, C.J., Retrieval of archival moving imagery - CBIR outside the frame? In: Lew, M.S., Sebe, N. & Eakins, J.P. Image and video retrieval; International Conference, CIVR 2002, London, UK, July 18-19, 2002 Proceedings (Lecture Notes in Computer Science, 2383) Berlin Springer-Verlag, 2002, 206-214.

[8] [Enser, 2000] Enser, P.G.B. Visual image retrieval: seeking the alliance of concept-based and content-based paradigm. Journal of Information Science, 26(4), 2000, 199-210.

[9] [Ide *et al*. 2003] Ide, I., Mo, H., Katayama, N. and Satoh, S. Topic-based structuring of a very large-scale news video corpus. AAAI Spring Symposium in Intelligent Multimedia Knowledge Management, 2003.

[10] [Ingwersen, 1992] Ingwersen P. 1992 *Information Retrieval Interaction.* Pub. Taylor Graham, London.

[11] [Markkula & Sormunen 1999] Markkula, M. & Sormunen, E. End-user searching challenges indexing practices in the digital newspapaer photo archive. Information Retrieval, 1(4), 1999, 259-286.

[12] [Pickering *et al*. 2003] Pickering, M., Wong, L. and Ruger, S. ANSES: summarisation of news video. Content-based Image and Video Retrieval (CIVR2003), 2003.

[13] [Smeaton & Over 2003] Smeaton, A. and Over, P. TRECVID: Benchmarking the Effectiveness of Information Retrieval Tasks on Digital Video. CIVR 2003 - International Conference on Image and Video Retrieval, Urbana, IL, USA, 24-25 July 2003.

[14] [Smeaton *et al*. 2002] Smeaton, A. Challenges for Content-Based Navigation of Digital Video in the Físchlár Digital Library. In: Lew, M.S., Sebe, N. & Eakins, J.P. Image and video retrieval; International Conference, CIVR 2002, London, UK, July 18-19, 2002 Proceedings (Lecture Notes in Computer Science, 2383) Berlin Springer-Verlag, 2002, pp. 215-224.

[15] [Yeo & Yeung 1997] Yeo, B-L. and Yeung, M. Retrieving and visualizing video. Communications of the ACM, 40(12), 1997, 43-52.