An examination of automatic video retrieval technology on access to the contents of an historical video archive

PETRELLI, Daniela <http://orcid.org/0000-0003-4103-3565> and AULD, Dan

This paper has been published in

Program: Electronic library and information systems

42 (2), May 2008

An Examination of Automatic Video Retrieval Technology on Access to the Contents of an Historical Video Archive

Daniela Petrelli, Dan Auld

Department of Information Studies, University of Sheffield, Sheffield S1 4DP, UK

d.petrelli@shef.ac.uk

Research Paper

Purpose
To provide a first understanding on the constraints historical video collections pose to video retrieval technology and the potential that an online access offers to both archive and users.

Design/methodology/approach
A small and unique collection of videos on customs and folklore was used as case study. Multiple methods were used to investigate the effectiveness of technology and the modality of user access. Automatic keyframe extraction was tested on the visual content while the audio stream was used to automatic classification of speech and music clips. The user access (search vs. browse) was assessed in a controlled user evaluation. A focus group and a survey provided insight on the actual use of the analogue archive. The results of these many studies was then compared and integrated (triangulation).

Findings
The amateur material challenged automatic techniques for video and audio indexing thus suggesting that the technology must be tested against the material before deciding on a digitization strategy. Two user interaction modalities, browsing vs. searching, were tested in a user evaluation. Results show users preferred searching but browsing becomes essential when the search engine fails in matching query and indexed words. Browsing was also valued for serendipitous discovery; however the organization of the archive was judged cryptic and therefore of limited use. This indicates that the categorization of an online archive has to be thought in terms of users who might no understand the current classification. The focus group and the survey showed clearly the advantage of an online access even when the quality of the video surrogate is poor. The evidence gathered suggests that the creation of a digital version of a video archive requires a re-thinking of the collection in terms of the new medium: a new archive should be specially designed to exploit the potential the digital medium offers. Similarly user's needs have to be considered before designing the digital library interface as needs are likely to be different than those imagined.

Research limitations/implications
The study is limited to a single archive; other institutions provided the technology used on which we had no influence.

Practical implications
The guidelines drawn as result of this study could impact on the strategy and policy of digitization projectsthat include video archives.

Originality/value
This paper is the first attempt to understand the advantages offered and limitations hold by video retrieval technology for small video archives like those often found in special collections.

1 Introduction

The digitisation of material within libraries and archives has become commonplace as technology can now offer better quality of image reproduction. The primary reasons for an institution to start a digitization project are to preserve material and widen access. However, potentials and risks are still a matter of debate.

Digitisation has been extolled as an effective and economic conservation measure (Weber 1997). Digital files are resistant to decay and help to guard against the complete loss of information held on volatile and delicate media, but attention is needed to migrate files to the next technology generation, to keep the digital form accessible and cut down redundancy (Owen et al 2000). The need to keep up with technology advancements raises concern about digital information being lost in the migration and the possibility that files will eventually become unusable (Cain 2003).

A digital copy can increase the access to rare and delicate artefacts as it can be used by a wider public remote from the actual object (Bouche 1999). However, digital copies may lose some of the visual quality of the originals during compression (Cain 2003), though interested parties could be satisfied by a high quality digital surrogate (Bouche 1999).

When dealing with video material the problems and advantages outlined above are amplified. Many within the film archive sector, particularly those at the executive and organisational level, think that digital video technology holds great potential and will enlarge access to film archives for teaching and learning purposes (Enser and Sandom 2002). In contrast, film archivists showed a negative attitude toward video retrieval and scepticism on how useful such technology will be (Sandom and Enser 2002). The main concern is that people looking for information will not be able to find what they want using such systems: The current technology is not sophisticated enough to automatically annotate and index pictures to include enough details to satisfy complex queries (Enser and Sandom 2002). Archivists claim video retrieval would not cut down on their workload as they would have to manually annotate every video in a digital library to provide a sufficient service.

Although based on some truth, this view does not consider how computers could help in providing a better service. If film archives offered digital access, the number of enquiries would decrease as the surrogate material would be likely to satisfy the needs of all but the most demanding researchers (Owen et al 2000). Moreover, a rich online catalogue would not only allow users to autonomously search and identify individual resources, but would also be a valuable mean for the promotion of the collection (Owen et al 2000). Benefits are likely to be higher for small collections or highly-focussed parts of bigger video archives as rarely seen material, often unique, can become more accessible through the digital medium. However the potential of a digitization project on small collections has not been investigated so far. Research on digital video libraries has been carried out mainly on extensive projects that had required substantial effort of engineers and computer scientists, e.g. Open Video Project (Marchionini and Geisler 2002), Físchlár (Smeaton et al. 2004), Informedia (Wactlar et al 1999). These projects have significantly contributed to the advancement of sophisticated techniques, e.g., video analysis of automatic scene segmentation[1] (Lienhart 2001), speech recognition for automatic transcription of the spoken audio (Brown et al. 2001), summarization and visualization of video content (Hughes et al 2003), (Christel et al 1998). However small archives have specific needs different from those of huge ones and therefore should be considered separately.

First and foremost, small archives have to face budget constraints. A small budget means limited resources available to buy expensive devices or acquire the best technology; hiring specialized personnel is not an option and the limited working hours of an archivist or the voluntary effort of enthusiasts is a realistic estimation of the resources available for a digitization project. However, having only limited time available may not be a big issue: as the archive is small, the time needed to look after it is likely to have an impact on its accessibility whereas the same effort is unlikely to have a perceivable effect on bigger archives.

---

[1] An extensive computer vision bibliography is maintained by Keith Price at http://iris.usc.edu/Vision-Notes/bibliography/applicat818.html

Secondly, small video archives (or specialized archives of a bigger collection) may encompass historical or cultural material that does not conform to the studio quality that has driven technological development so far. Indeed it is not unusual for small collections to originate from personal donations, therefore the quality of the material is often amateurish.

Finally the users of specialized small archives have to be considered before a digitization project starts as they may have specific needs and expectations that should be met by the final system.

This paper explores constraints and possibilities for a small video archive to be digitized and made accessible online to a wider audience. A real collection of amateur and historical videos currently used by scholars and students is used as a case study. Part of this collection has been digitized and the effectiveness of techniques for automatic indexing has been assessed. The online access to the digital collection was then investigated in a controlled user evaluation and explored with actual archive users in a focus group.

The different perspectives considered in the study (the archive, the user, the technology) provide a better understanding of the potentials and pitfalls and offer suggestions on the steps a small digital library projects should undertake before more formal planning takes place (Shen et al 2005, Gonçalves et al 2004).

The paper is organized as follow. A description of the collection used follows in section 2. Section 3 provides an overview of digital video retrieval. The digitization step and the technology assessment are described in section 4. The user evaluation and focus group follows in section 5. Section 7 summarises the results of the study before the conclusions in section 8.

2 The NATCECT Archives

Academic libraries often hold special collections of multimedia material. Generally the legacy of scholars, sometimes those archives develop into a collective heritage that needs to be preserved but also made accessible and available to the widest public. The National Centre for English Culture and Tradition NATCECT[2] at the University of Sheffield is an example.

The archive was established in 1968 as a repository for material collected since 1964 by Professor J.D.A.Widdowson through the Sheffield Survey of Language and Folklore. The NATCECT today houses material on all aspects of English folklore and language. The collection includes text (fieldwork notebooks, newspaper cuttings, filled questionnaires), printed ephemera, images (transparencies and photographs), audio (on tapes), and video. The archive is partially catalogued and open a few hours a week; users are academics, scholars, university students, and enthusiasts. The audio-visual collection spans from the 1920's to the present day and encompasses children's language, calendar and social customs, rites of passage, folk belief, legends, anecdotes, jokes, regional and social dialects, slang and colloquialisms, occupational vocabulary, proverbs, sayings, traditional drama, music, and dance.

[2] http://www.shef.ac.uk/natcect/archive

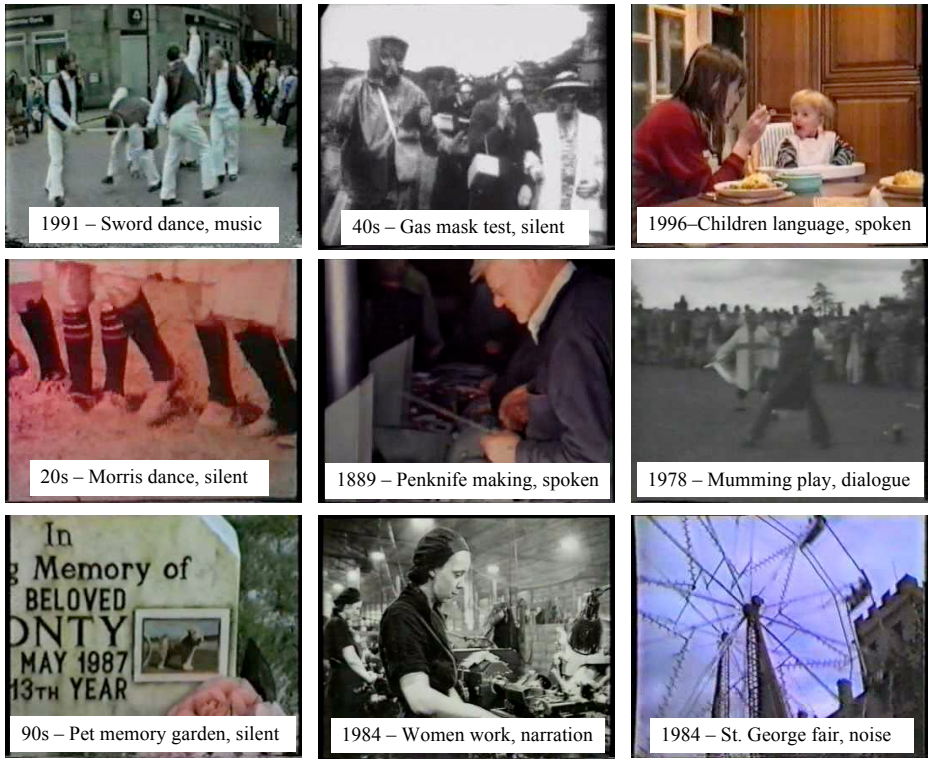

Figure 1. Examples of the material in the NATCECT collection: the date of recording, the topic, and the type of audio are indicated.

There are approximately 230 videos stored on different media (VHS tapes, Umatic tapes, and high density tape reels), most have been recorded by amateurs and donated to the Centre, examples are in Figure 1. Videos have been classified respect to their generic content and using a specific organization; for example "childlore" includes children games – skipping or clapping hands; "custom fixed date" lists traditions related, for example, to the 25th of December, Christmas, while "custom movable date" is used for traditions not linked to a fixed date like Shrove Tuesday. For the broad categories, e.g. dance, sub-categories are used, e.g. coconut, morris, sward.
Videos in the collection are, in many cases, the only type of footage of its kind (e.g. footage of male Morris dancers in the 1920s). For its nature the NATCECT would benefit from a digitization project that makes its material accessible on the Web.

3 A Brief Overview of Video Retrieval

Early digital video libraries allowed access to material via searching the associated textual descriptions. Metadata such as title, actors, and date was also used. More recently techniques that directly operate on video content (i.e. moving images, audio, and occasionally overlaid text) have been explored and incorporated into operating video libraries. This section briefly describes both approaches and discusses those in the perspective of a small and video collection like the NATCECT one.

3.1 Metadata-Based Video Retrieval

Metadata describes the characteristics of information in encoded structures that can help to identify, discover, manage and assess material. Metadata can be seen as being an access provider and asset protector with its specification of rich contextual information bringing order to a potentially shapeless resource, e.g. images, video, audio (Christel and Wactlar 2002). It should provide the crucial details needed for aiding retrieval at the level needed by users when searching for information and judging its relevance.
Metadata within digital video libraries can be of two types (Jain and Harumpapur 1994): *content dependent* metadata relates directly to the visual and aural content of the video; *content*

independent metadata refers to information pertaining to the videos but that does not have anything to do with the actual content, e.g. producer, actors, recording location.
Content dependent metadata (CDM) is important to aid resource location: exhaustive shot lists that detail the minutest piece of action taking place in every single shot are often produced (Sandom and Enser 2002). CDM satisfies specific requests looking for shots like "aeroplanes taking off", but would fail in satisfying requests like "the assassination of John F. Kennedy". For those a higher, semantic level of metadata is needed (Sandom and Enser 2002). To help the writing of CDM synchronized with the video sequence, tools such as IBM VideoAnnEx[3] have been developed. As the manual annotation can be a strain, the possibility of automatically detecting features that can be part of CDM has been researched since 2002 by the TRECVID community[4]. Results are encouraging, but not yet robust enough to exit research laboratories.
Content independent metadata encapsulates information that is not directly about the video but is needed by the archive, e.g. categorization, shelf location, provenance, or property. Although this metadata is not used for searching by content it can provide a base for clustering the material and/or filter the search result. For its nature content independent metadata have to me handwritten but are essential as they reflect the cataloguing of the material.

3.2 Content-Based Video Retrieval

Content Based Video Retrieval techniques (CBVR) seek to employ a range of indexing and retrieval methods taken from the analysis of the visual and aural content of a video segment. Those include speech recognition and transcription of the commentary, scene and boundary detection in moving images, shape and object detection in single key frames (Smeaton 2004).
Automatic speech recognition can index the spoken audio elements of video: a time aligned spoken word transcript can be generated and enriched by temporally aligned captions derived from overlaid text (Witbrock and Hauptmann 1997). Natural language processing tools are used to form the recognised speech into paragraphs from which textual summaries and titles of videos are generated (Witbrock and Hauptmann 1997). Textual summaries derived from the transcription can be used as surrogates to describe the video content to the user.
The video stream can be segmented through shot boundary detection, a technique to divide a video sequence into manageable segments by applying image processing techniques (see (Smeaton 2004) for details). During shot boundary detection a keyframe image that represents the segment is automatically extracted; the sequence of keyframes is used to represent the visual content of the whole video. This visual surrogate is analogous to the index of a book that is used to determine which parts look interesting or worthy of further consultation (Yeo and Yeung 1997). Shot boundary detection is a quite robust technique when applied to high quality standard videos (i.e. film or TV programs) but it can fail and generate an abnormal amount of keyframes when the source does not conform to shot rules.
More sophisticated techniques could be applied to the video stream to distinguish between, for example, indoor or outdoor scenes, with people, with faces (Gros et al. 2005), and summaries can be created by analysing the transcription of the commentary (Sadlier et al. 2002) or the moving images (Assfalg et al. 2002).
Much of the advancement in this area derives from the collective effort of researchers around the world that participate in the TRECVID initiative. Although huge progress has been made, at present CBVR still faces significant challenges before any meaning can be assigned to a sequence (Petkovic and Jonker 2004).

4 Digitization, Quality Assessment, and Technical Feasibility

Only part of the whole NACTECT video collection was digitized; the choice was driven by the video quality (only the best were used), practicality (formats that could be immediately digitized, i.e. VHS), budget issues (completion with the available hardware), and time constraints. The

[3] IBM VideoAnnEx Annotation Tool uses MPEG-7 metadata layer to record "scene descriptions, key object descriptions, event descriptions, and other lexicon sets" http://www.alphaworks.ibm.com/tech/videoannex (accessed 1.8.2007).
[4] http://www-nlpir.nist.gov/projects/t01v/ (accessed 1.8.2007)

digitization was done in-house using commercial hardware and software[5]. The result does not compare with professional digitization; however our interest here was not on preserving the archive content (for which the best possible quality is advisable), but on enlarging its use via on-line access. A "good enough" result would suffice in this context: assessing if the video quality was indeed good enough was a point of investigation in the user evaluation.
The digital videos were then processed for indexing: keyframe extraction was applied to the visual channel while the audio was used to test automatic classification. The analysis of the results is reported in this section.

4.1 Visual channel

In all, 53 cassettes were digitized, totalling 37 hours. The length of the resulting video varied greatly from a few minutes to over 2 hours and covers 80 years, from the 20s to more recent times, with most of the recordings done in the 80s and early 90s (see Table 1).

Era	20s to 50s	70s	80-84	85-89	90-94	95-99	00-03
Number	3	7	14	10	11	6	2

Table 1. Distribution of the videos by recording year in the digitized collection.

A wide proportion (47%) of the video was black and white (B&W); this includes some of those recorded in the early 80s and all the older ones[6]. Colour is one of the main features used during visual indexing, therefore having half of the collection in B&W can be a real limitation.
The image quality of the original analogue material in the NATCECT collection is often low due to the initial recording (old, hand held recording devices) and subsequent multiple format transfer, e.g. from Super8 to VHS (likely a further VHS copy). As a result the quality of the digital material is poor with respect to broadcasting standards (e.g. TRECvid material): NATCECT B&W videos often have blurred images (as in the Mumming play image in Figure 1) whereas colours spread out of the proper border (as in St.George fair in Figure1). The poor quality negatively affected the keyframe extraction during the shot boundary detection process[7]. The performance of the algorithm varied greatly from video to video: the average was 1 frame per 67 seconds, with a minimum of 1 frame per 536 sec. (9 minutes) and a maximum of 4 frames per 1 second (resulting in an abnormal 1880 keyframes for a video of less then 8 minutes). An excessive number of keyframes does not properly represent the video content and can actually become, as in the extreme case reported, an impediment when the keyframes are used to access the video. A small number of frames (e.g. 1 every 9 minutes) is not effective either as it still requires the user to fast forward and rewind the clip to find the point of interest.

Another feature used to index video is the change of scene that derives from cutting and rearranging shots in post production editing. It is not unusual in the NATCECT material that the camera is set in the same position for the whole recording and the scene changes because characters enter or exit the scene, as for example in a Mumming[8] play. Interestingly while this is a good indicator for theatrical settings this is often not the case for street recording, e.g. a group of sword dancers performing in a public square: passers by are not the focus, on the contrary they are purely background noise.
An analysis of the 53 videos indicates that the scene change is often due to stop and start recording; as in the case for the compilation of dances performed by the same group on different occasions or by different groups during the same event.
For just a limited set (5%) the video has been edited with proper cutting and closed-captioning[9] providing enough material for a multimodal indexing (Snoek and Worring 2005).

---

[5] A camcorder was connected to a VHS player in input and to a standard PC in output. The camcorder converted the analogue signal into digital; commercial software was used to capture the digitized video on the PC and to compress it into MPEG-1 format.
[6] An interesting exception is a final fragment of the Morris dancers recorded in the 20s that is in colour, likely one of the earliest experiments of the kind.
[7] The shot boundary detection process was run at UCD using the software developed for the Fischlār project.
[8] A Mumming is a traditional English play that represents the fight of good against evil.
[9] Text added to overlap the image, common in TV broadcasting, e.g. news programs.

4.2 Audio channel

As mentioned in section 2.2, the content of the audio channel can be used to index the video. The NATCECT collection was problematic for the audio component as it was for the visual one. The audio of each video was manually assessed and associated to one or more of the following categories:

- Silent: 4 videos (7%) did not have any audio component;
- Street noise : 15 videos (26%) had an audio component but useless in term of meaning;
- Voice:
    - Non related speech: the audio contains people talking but the actual content does not relate to the video content (e.g. street recording of dancing captured voices of spectators speech as they were next to the camera); this occurred in 5 videos (8%);
    - Speech: 8 videos (14%) had a clearly identifiable speaker; the recording however might have occurred in noisy places, e.g. penknife making recorded in the workshop in Fig. 1;
    - Narration : associated to edited videos account for 6 items (10%);
    - Play: 13 videos (22%) were theatrical representation with multiple speakers acting, e.g. mumming play;
    - Nursery rhymes: these 5 videos (8%) include skipping games and clapping hands; many children are saying the same rhymes while performing some actions;
- Music : a consistent part of the collection 15 (26%) contains music registered outside or indoor; the recording is always live, i.e. never added in production; Often the music accompanies dancing;
- Songs : 8 videos (14%) contain recitals of traditional songs.

The classification shows that only 50% of the video has some form of speech that could index the video content. However a high percentage of this is problematic for automatic speech recognition as the recording conditions (e.g. in a noisy environment, multiple voices) impact on its accuracy measured at about 70% [10]. A further complexity is the challenging content (e.g. children telling jokes, dialects) for which it is reasonable to expect a further decrease in performance.

In light of these considerations, speech recognition was considered unlikely to be of easy and effective use in indexing the NATCECT archive. Instead as 40% of the audio was musical (music or song) a test was performed to see if the video could be automatically classified as musical or speech. The Marsyas[11] system was used to compute the audio properties; the audio was split up into 5 second segments (using a 3 second overlap) and a nearest neighbour classification algorithm was applied to each segment. The speech-music classification resulted correct in about 60% of the cases, against an accuracy of 90% reported in literature (Pinquier et al. 2003). This 30% reduction in the classification performance is due to the nature and quality of the material. Indeed, the distinction between music and speech is not always neat: songs might not have a musical accompaniment or speech might have been accidentally recorded during a music playing (e.g. street recording of nearby talk).

4.3 Discussion

The technical assessment showed that automatic technology are challenged when the material is not of studio quality and can indeed fail dramatically as in the extreme case of the keyframe extraction. However, even with limitations and failures, automatic techniques can be valuable in creating a video digital library, as in the following examples.
Often the VHS cassettes in the NATCECT collection contain many recordings on the same tape: the digitization creates then a long video of heterogeneous (though related) clips. For example a tape with the gas mask training in the 40s included also a rescue from a burning building,

[10] Data as reported by Wekipedia – Speech recognition page on the 1.8.2007 http://en.wikipedia.org/wiki/Automatic_speech_recognition
[11] http://opihi.cs.uvic.ca/marsyas/

munitions assembly, bike rides, man-to-man combat, and a mass. Similarly a tape recorded at a dance festival collates many dances of different kinds and performed by different groups.
Automatic scene detection can be applied to segment longer video files into smaller and homogeneous items, e.g. split the dance file into single performances. This division would allow a finer grained retrieval but it is necessary to maintain a connection among clips coming from the same tape as, for example, the dances were performed at the same event.
Keyframe extraction is an effective way to visually index the content, as discussed in the next section devoted to user interaction. To overcome the problem of an excessive number of keyframe the archivist should be given the possibility of discarding the duplicates and keep only those that represent the content and are meaningful entry points in the clip.

As for the video component, the automatic analysis of the audio channel can be a valuable tool for classifying the clips onto brad categories. Given the imprecision of the technology when applied to street recording, the supervision of the archivist is needed. However this step should be faster that manually classify the whole archive as more than half of the clips should have been classified correctly.
Similarly the use of speech recognition can be useful when the full indexing of the content of a clip is needed, for example the transcription of a play would allow the retrieval of the individual lines of speech. As at most it can be expected to reach a quality of 70% checking and correction are essential. Although the checking is likely to be more efficient than transcribing, its cost in term of time could be high and consideration should be spent to select only those clips for which a transcription would provide a real advantage.

5 Assessing the User Experience

5.1 Browsing vs. Searching

Indexing the visual content is essential to let the user judge at a glance what a video is about: the keyframes extracted during the boundary detection process were used as a visual surrogate. A cleaning step was initially done: redundant or noisy images were manually removed before storing the keyframes in the database. This process was not a cherry-picking of the best quality and most representative keyframes but rather a cleaning step that preserved the essence of the automatic shot boundary detection phase and did not required too much time.

Archive number	VC0007
File number	0017.mpg
Title	Britannia Coconut Dancers
Author	Unknown
Location	Bacup (Lancashire)
Participants	Unknown
Producer name	John Smith
Era	1980-1989
Date	1981
Colour	Black and white
Duration	42 min 1 secs

Category	Dance
Detailed category	Coconut Dancing
Synopsis	Footage of the Britannia Coconut Dancers performing in various outdoor and indoor locations. They skip down the streets followed by children who mimic their dances. Their dance involves a lot of hand movements. They have things on their knees that make a clacking sound when hit. They also dance in formation with flowery hoops. Their faces are blackened.

Figure 2. Example of metadata created for the archive.

An XML file was compiled for each video and was enriched with manually written metadata. Figure 2 shows an example. The metadata included all the information the archive kept of each video; A short textual description of the video content (the synopsis) summarising the most significant actions taking place was added.

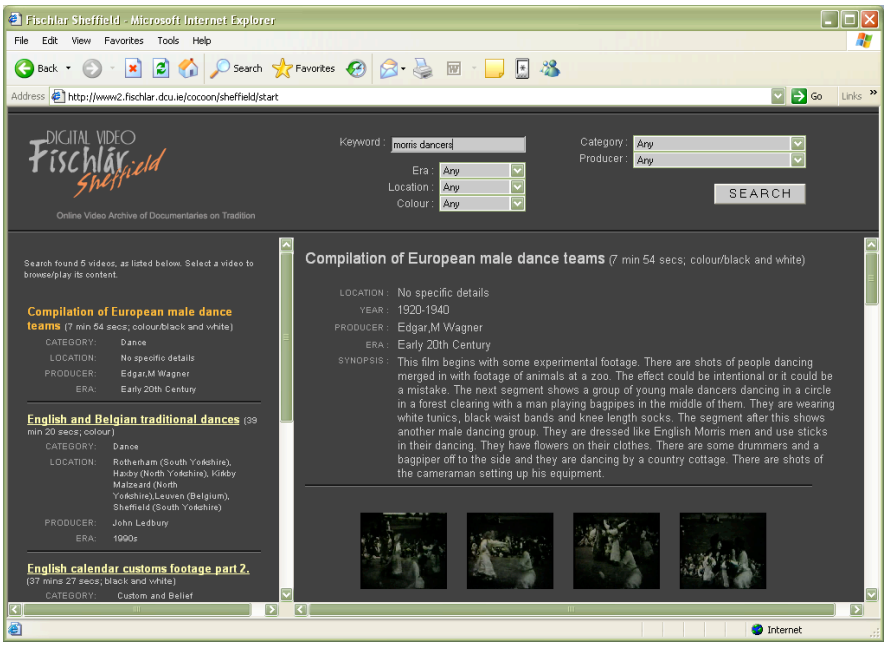

Figure 3. The search interface: search options on the top, results on the left.

From the whole metadata set, five attributes were selected to be used on the interface as filters (era, location, colour, category, producer) and could be selected by the user on the search interface (top right in Figure 3). The synopsis was used to index the videos for free text retrieval. The annotation of every single keyframe was considered too time consuming and not necessarily worth the effort. The retrieval acted then at the video level.

Both the digitization and the metadata creation steps kept the structure of the archive as it was, that is to say what was contained in a tape was digitized as a single video clip and the archive information was added in the metadata.

A system for online interactive video retrieval was set up at the Centre for Digital Video Processing in Dublin building upon much previous work (Smeaton et al 2004). The interface was via a conventional web browser: the user could search video content by free text queries (on the indexed synopsis) as well as by selecting one or more of the five attributes offered as filters (fig.3, top right). The search mechanism was very basic, via a keyword exact match. Filters and free text could be used singularly or combined.

The list of retrieved videos was displayed on the left side of the screen. If no video was retrieved a message was displayed in place of the list suggesting further actions. If the user had selected a video, its details would be displayed on the right, metadata and synopsis on top, keyframes below (Figure 3).

A second interface (fig.4) was created as a baseline for the evaluation. Videos were grouped with respect to their category; titles were listed on the left, the selected video on the right. The two interfaces (Figure 3 and 4) differ in their search mechanisms only. The rationale for comparing a

video search system against a simplistic list is due to the size of the NATCECT collection and its use. A collection of a few tens of videos (up to a couple of hundred) accessed regularly (even if rarely) could be easily browsed as after a little while the content becomes familiar to its user. This interaction mode is of course not an option for bigger collections, but if a browsing interface is good enough for small archives, then the cost of building a digital video archive can be vastly reduced.

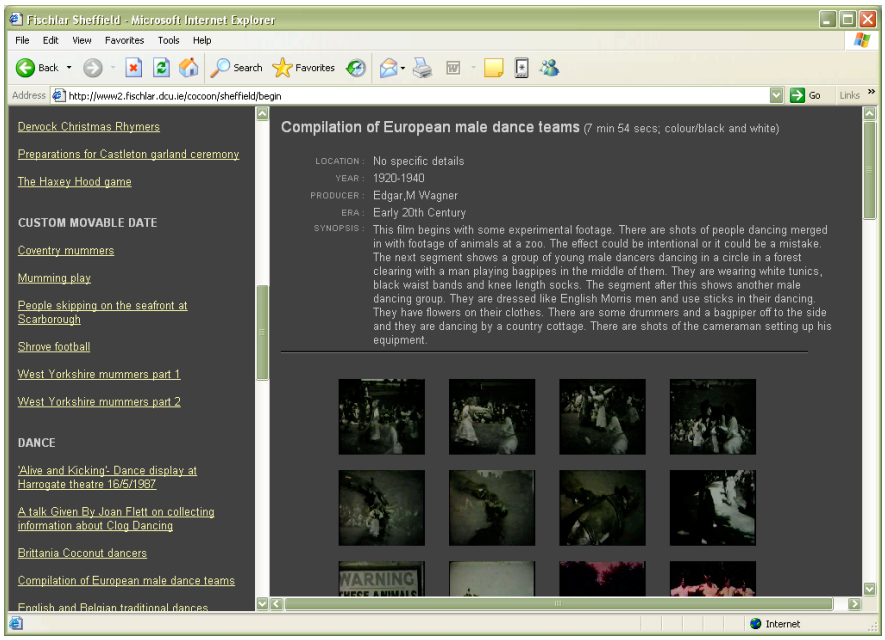

Figure 4. The browse interface: videos grouped by category are on the left.

5.2 The User Evaluation

The comparative evaluation involved 24 participants; each used the two interfaces in a within subject experiment. Participants were academics or students at the University of Sheffield; a few had used the analogue NATCECT archive before. Participants were required to accomplish 4 tasks, 2 using each system. Tasks were of 2 types: a) to retrieve (a minimum number of) videos relevant for a written scenario representing realistic uses (Borlund 2000) (e.g. finding video material to show in a lecture), and b) to locate the only video containing a given keyframe (2 keyframes per task). Keyframe location tasks carried the potential of showing the variety of query terms users would generate while searching. Participants had 10 minutes for each task but were free to stop when they felt they had found enough material. A 5 minutes self-training with the search system was scheduled at the beginning. Participants were invited to verbalize their thoughts while accomplishing the task.
The interaction was recorded to later extract objective data, i.e. queries, task accomplishment rate and time. Actions (e.g. scrolling, video play) and comments were captured in a video. Questionnaires were used to collect subjective data. They comprised: personal profile, familiarity with the topic, user satisfaction with each system (derived from QUIS (Chin et al. 1998)), and systems comparison. An interview was held at the end to discuss interesting behaviours observed and collect participants' feedback.

5.2.1 Efficiency and Effectiveness

The search system was the most efficient and effective for completing tasks: an average retrieval time of 92 sec. per video was calculated for the search system that retrieved 175 relevant clips; the browsing average time was 103 sec. for 162 relevant clips. The difference decreases sensibly if only the successful cases are considered: 74 sec. for the search, 76 sec. for the browse. Indeed the failure rate of the browsing system was higher (13 failures over 48 tasks, 27%) then the searching one (11 failures on 48 tasks, 22%), but the vast majority of the browsing failures were due to the time expiring (10 out of 13) while for the search system it was the user who gave up most of the time (9 out of 11). An analysis of the logs shows that of the total of 332 queries submitted only

41% retrieved something. Of those 61% were text only, 24% were mixed text and filters, and 11% were filters only. The most successful retrieving mode was through filters only (81%) while text only was successful in 47% of cases. The least successful was the mixed modality in just 21% because filters were used to focus the result of the search, but on a small scale collection this easily results in empty sets. For example the highest number of videos retrieved for a text query was 11 for 'mumming' (a popular play that represents the fighting of good against evil) but those videos belonged to 5 different categories and applying a filter would reduce the result to 5 at best, or to 0 at worst.
Similarly, the exact match criteria used by the searching mechanism showed its limitation inside a small collection, e.g. 'mumming' retrieved 11 while 'mummers' retrieved 2 and 'mummer' not one. This system feature forced users to try many similar terms before finding the only successful one, if found at all. In the evaluation participants entered as many as 27 queries in order to find the requested video, however this is unlikely to happen in real life as after a while users would assume the desired content is not available.
As discussed above, search and filters apply to different texts (synopsis and metadata); as the two may not be in line, inconsistent behaviour was at times displayed, e.g. 'drama' as a category retrieved 5 videos but 'drama' as free text retrieved just 1 (which belonged to the 'dance' category). This clearly shows that all attributes (synopsis, title and metadata) should be indexed and used as sources of evidence for the free text search.

5.2.2 User Satisfaction and Behavior

Despite the high rate of query failure, user satisfaction questionnaires showed a strong preference for the search interaction (67%) over the browsing one (20%), with 13% neutral. Even questions on the same interface features (usefulness of the synopsis or keyframes) were rated higher for the search system than for the browse system. In the interview participants motivated their choice as "being used to searching", "don't like to scroll up and down" and more in general "feeling in control" with respect to browsing through someone else's organization. In a few cases the preference for the searching could have been a dislike for the browsing as proposed rather then for the modality in itself. Critical comments such as "unclear categories", "flat hierarchy", and "strange titles" supported this interpretation.
The 5 participants who preferred the browsing system motivated their choice upon such factors as exhaustiveness of the browsing ("with keywords you never know if you have actually retrieved everything") and easier access to the material ("it is easier to pull out what there is"). Those users valued the transparency of browsing against the opacity of searching. A participant who did not express a preference for either system said "It was nice to browse through and see what was in there. I would never have known that there were dance videos in the collection if I'd been using the text searchable system". Other participants showed a similar interest in the content they encountered while browsing: a few watched some clips; most said they would have watched if they were not in need of finishing the task. This shows the importance of serendipitous search and casual discovery particularly for material used for study and research purposes.

Positive comments were collected on the contents of the archive despite the sometimes poor quality of the videos. As one participant pointed out, the amateur videos were superior as they were "a more genuine record and there's less of a chance that it has been staged" and "it tells you something about the equipment used […] a little super 8 camera". The small size of the video play area was rated just sufficient, though good enough for the task in hand: "the size of the viewing screen could be a little larger but the current size is perfectly usable". The advantage in the data exploration offered by the digital format is evident: "to jump in on a clip and start the video at any point is brilliant", "to get a very good idea of what's on the system just by skimming through". One comment pointed out the importance of bookmarking the retrieved material "to give someone else a reference to a particular keyframe".
A final point to consider is how users interacted with the video surrogates. Behaviours observed were of 3 types: *text driven* (15%) read carefully the text and ignored the keyframes, *image driven* (15%) skimmed through the images and ignored the text, and *holistic* (70%) used both text and images to judge video relevance, thus confirming (Hughs et al. 2003). Except for those few who just read the text (15%), all the other participants (85%) exhibited frustration with the number of keyframes. Many, when scrolling through several screens of images, gave up and looked at other material instead thus missing relevant video shots.

5.3 Further Insights from a Focus Group

A focus group was held with the current users of the NATCECT video collection, i.e. potential users of the digital archive. Eight people participated, all were new to the system: 2 academics, a researcher, the archivist, and 4 masters students. An introduction to the system was given, comments and questions were solicited. Then the group was left to interact freely with the system; their behaviour was observed. A questionnaire was finally distributed.
The behaviour of this group was fairly different from that observed in the evaluation: these people were familiar with the archive terminology and had no problem with titles and categories. When using the search they wished to find content at clip level; they were looking for highly conceptual clips, e.g. the doctor giving the medicine to the dead St. George in a mumming play. Only expert-added metadata could offer the needed level of content to make the search effective, a task they said they were willing to do to improve the system retrieval.
The visual surrogate (keyframes) was used much more than the text one (synopsis), partially because the content was known and the text description did not add new information. Images were used to jump into the video at the desired point, time was spent watching the videos and commenting on the content. There was a generic complaint about the number of keyframes, their quality, and their similarity. The videos were also deemed to be too small, but when the file size was discussed the current format was considered to be good enough for research tasks thus supporting (Bouche 1999) that most users would accept a surrogate that is not as good in quality as the original.
Comments on the usefulness of such a system were enthusiastic; the most common remarks were centred around enthusiasm for having the collection more accessible and better organized. The possibility of accessing the material from outside of the archive by using the web was greatly appreciated.
The videos could be played only if they were stored locally; Web access would just display the keyframes. When questioned about the possibility of searching the archive from everywhere but being limited to scan through the images all participants agreed that this would be very useful anyway as many of them do not travel to Sheffield regularly. The possibility of searching the archive would improve their research and watching the selected relevant material could be postponed to the next visit.

6 Digitization as Archive Assessment

One of the key aspects of the digital medium is its high flexibility: movies do not need to be seen in sequence, but viewers can jump wherever they like. Even though users could jump into the middle of a video when it was retrieved, by keeping the archive format the non-sequential mode was not exploited. The digitization step should include a reassessment of the current archive organization in order to take advantage of the digital medium. The first step would be to decide which grain better represents the material.
A revision of the information collected into the metadata is also appropriate. This revision step should encompass the assessment of the terminology used when the material is presented to the general public as the archive categorization, fully clear to accustomed users, can be problematic and can prevent effective use by newcomers.
The physical archive should also be assessed to guarantee consistency across the whole collection by checking, for example, that similar material belongs to the same category or has a similar description. Consistency is fundamental for effective retrieval and becomes crucial when offering on-line access since the help and mediation of the archivist is not available to the user.

7 User-Archive Interaction Design

The user evaluations and the focus group clearly showed that searching is the preferred modality for interacting with a video archive. However this fact should not lead to the overlooking of other possibilities, e.g. designing an effective browsing system. Archive aims like promoting the discovery of the collection or limiting the costs, should influence any informed design as much as the user's preference. Indeed improving familiarity with the collection, an important factor when

the use is regular as it is likely to be for scholarly use, should push the designers to consider browsing as the main interaction modality.
If the collection consists of a few hundred videos it is possible to fully explore it via simple browsing in a reasonable time frame.
A further reason for considering browsing as the main interaction mode is the potential for serendipitous information discovery. Watching videos is highly entertaining and is likely to occur often in natural settings (i.e. not under time constraints as in the evaluation), as happened during the focus group. The possibility of casual discovery is particularly relevant with the NATCECT collection and alike where unique pieces exist: missing or finding that single one makes a difference.

The use of faceted metadata as a way of organizing, accessing, and navigating a collection has proved to be very effective with images (Yee et al. 2003); it is therefore likely the effect would be replicated with other visual material (i.e. video). This requires a better understanding of archive users and uses, but would not need special technical skills or technology, as would be the case with an advanced information retrieval system.

The exact match approach used for the retrieval proved to be very cheap to implement but too poor to be of any actual use. As the browsing was negatively affected by the archive terminology, the suggestion is to concentrate on providing a good filtering system. Indeed setting up a good browsing facility is likely to involve less technical expertise than setting up a good search system.
If a search functionality is implemented, then a more open indexing and retrieval is needed to capture as much content as possible. The indexing of all the textual components of the metadata would be mandatory to enlarge the retrieval as much as possible. Stemming should be included as well as query expansion via thesaurus and synonyms.
In our study, free text search and filters did not work in a synergistic way but as an antagonist with the result of often reporting an empty set. If both search and browse are available on the interface the layout should discourage the use of mixed queries, by, for example, clearly separating text and filter options.

8 Conclusions

A small digital library from an analogue collection of homogeneous amateur videos was created and current automatic technologies to index audio and video channels were tested. Results were much lower than those reported in literature when similar techniques are applied to broadcasting quality video. An online video archive was set up and two interactions (searching and browsing) were evaluated with users. In essence, our minimalistic approach offered a good enough service: the appreciation of a group of potential users for the result of the low-budget digitization project showed the high potential for the promotion and accessibility of material in rare video archives. The possibility of autonomously browsing through the collection and viewing the material online greatly overcame reservations about the videos quality.
However to take full advantage of the digital medium, time and effort have to be spent to rearrange and annotate the digital collection. The organization of the analogue archive proved to be too rigid and coarse grained to take full advantage of the flexibility offered by the digital medium.

A semi-automatic tool could be created to facilitate an archivist in creating the digital collection, annotate the material and offer a more fine grained access to video clips. A first attempt in this direction has been done by Mu and Marchionini (2003) who propose a graphical interface for capturing and annotating educational material. Conversely from their approach, we envisage a tool that incorporates CBVR technology as much as possible but leaves the human with the final decision of what is worth including depending on the quality of the output of the automatic step. Indeed the visual surrogate (i.e. the sequence of keyframes) proved to be the weakest point as the sometimes huge number of images hampered the correct identification of the desired clip.
The role of the user could also change to that of editor of the archive. High profile users (researchers, scholars) can supplement more precise descriptions associated with the videos thus enriching the text that is used for retrieving the video. Our study shows such expert and reliable users are willing to do so in order to improve the quality of the archive and the future retrieval. The digital archive would then become a collective resource used, annotated and shared by the user community.

Acknowledgements

Part of this work was submitted by Dan Auld in partial fulfilment of the Masters degree in Librarianship. The authors are grateful to Alan Smeaton and his group at the Digital Video Centre in Dublin for the time and energy spent in setting up the system for the user evaluation; and to Simon Trucker for performing the audio analysis and classification. We thank NATCECT's director and archivist for their support and all the people who participated in the study.

References

1. Assfalg, J., Bertini, M., Colombo, C. and Del Bimbo, A. (2002), "Semantic Annotation of Sports Videos", *IEEE Multimedia*, Vol. 9 No. 2, pp. 52-60.
2. Borlund, P. (2000), "Experimental components for the evaluation of interactive information retrieval systems", *Journal of Documentation*, Vol. 56 No. 1, pp. 71-90.
3. Bouche, N. (1999), "Digitization for scholarly use: The Boswell Papers Project at the Beinecke Rare Book and Manuscript Library", Council on Library and Information Resources. http://www.clir.org/pubs/reports/strategies.html (accessed 7 November 2007)
4. Brown, E. W., Sriniyasan, S, Coden, A., Ponceleon, D., Cooper, J. W. and Amir A. (2001), "Toward Speech as Knowledge Resource", *IBM Systems Journal*, Special issue on Knowledge Management, Vol. 40 No. 4, available at http://www.research.ibm.com/journal/sj/404/tocpdf.html (accessed 7 November 2007)
5. Cain, M. (2003), "Being a library of record in a digital age", *Journal of Academic Librarianship,* Vol. 29 No. 6, pp. 405-410.
6. Chin, J. P., Diehl, V. A. and Norman, K. L. (1998), "Development of an instrument measuring user satisfaction of the human-computer interface", CHI '98. ACM Press, pp. 213-218.
7. Christel, M.G., Smith, M.A., Taylor, C.R. and Winkler, D.B. (1998), Evolving video skims into useful multimedia abstractions", CHI '98 ACM Press, pp. 171-178.
8. Christel, M.G. and Wactlar,H.D. (2002), "Digital video Archives: Managing through metadata", In: Building a national strategy for digital preservation: Issues in digital media archiving. http://www.clir.org/pubs/reports/pub106/video.html (accessed 8 November 2007).
9. Enser, P. and Sandom, C. (2002), "Retrieval of archival moving imagery - CBIR outside the frame?" Proc. CIVR 2002. Springer Verlag (Vol. 2383) 206-214.
10. Film Archive Forum "Moving History: Towards a policy for the UK moving image archives", London: BUFVC. http://www.bufvc.ac.uk/faf/mh.pdf (accessed 8 November 2007)
11. Goncalves, M. A., Fox, E., Watson, L. T. and Kipp, N.A. (2004) "Streams, Structures, Spaces, Scenarios, Societies (5S): A Formal Model for Digital Libraries", *ACM Transactions on Information Systems*, Vol. 22 No. 2, pp. 270–312.
12. Gros, P., Delakis, M. and Gravier, G. (2005) "Multimedia Indexing: The Multimedia Challenge", ERCIM News No. 62, July 2005, 11-12.
13. Hare, J. S., Lewis, P. H., Enser, P. G. B. and Sandom, C. J. (2006) Mind the Gap: Another look at the problem of the semantic gap in image retrieval. In *Proceedings of Multimedia Content Analysis, Management and Retrieval 2006,* SPIE Vol. 6073.
14. Hughes, A., Marchionini, G., Wildemuth, B. and Wilkins,T. (2003) "Text or pictures ? An eyetracking study of how people view digital video surrogates", Proc. 2nd Int. Conf. on Image and Video Retrieval CIVR 2003. (Vol. 2728) Springer Verlag, pp. 271-280.
15. Jain R. and Harumpapur A. (1994) "Metadata in video databases", *Sigmod Record*, Vol. 23. No. 4, pp. 27-33.
16. Lienhart, R. (2001) "Reliable transition detection in videos: A survey and practitioner's guide", *International Journal of Image and Graphics (IJIG),* Vol. 1, No. 3, pp. 469-486.
17. Marchionini, G. and Geisler, G. (2002) "The Open Video Digital Library", *D-Lib Magazine*, Vol. 8, No. 12, available at http://www.dlib.org/dlib/december02/marchionini/12marchionini.html (accessed 11 November 2007)
18. Mu X. and Marchionini G. (2003) "Enriching Video Semantic Metadata: Authorization, Integration, and Presentation", ASIST Annual Meeting 2003, pp. 316-322.
19. NATCECT http://www.shef.ac.uk/natcect/ (accessed 11 November 2007)
20. Open Video Project [The] http://www.open-video.org/index.php (accessed 4.3.2005)
21. Owen, C, Pearson, T., Arnold, S. (2000) Meeting the challenge of film research in the electronic age. D-Lib Magazine. 6(3) http://www.dlib.org/dlib/march00/owen/03owen.html (accessed 23.9.2004)
22. Petkovic, M. & Jonker, W. (2004) Content based video retrieval: A database perspective. Dordrecht: Kluwer ac. press.
23. Pinquier, J., Rouas, J-L., Andre-Obdrecht, R. (2003) Robust Speech/Music Classification in Audio Documents. In Proceedings of the International Conference on Acoustics, Speech, and Signal Processing (ICASSP 03), Vol. 2.

24. Preece, J, Rogers, Y. & Sharp, H. (2002) Interaction design: Beyond human-computer interaction. Wiley.
25. Sadlier D, Marlow S, O'Connor N and Murphy N. (2002) MPEG Audio Bitstream Processing Towards the Automatic Generation of Sports Programme Summaries. ICME 2002 - IEEE International Conference on Multimedia and Expo.
26. Sandom, C., Enser, P. (2002) VIRAMI - Visual Information Retrieval for Archival Moving Imagery. Library and information report 129. London: The council for museums, archives and libraries.
27. Shen1, R., Gonçalves, M. A., Fan, W. and Fox E. (2005) Requirements Gathering and Modeling of Domain-Specific Digital Libraries with the 5S Framework: An Archaeological Case Study with ETANA. Proceedings of European Conference on Digital Libraries ECDL 2005, Springer LNCS 3652, 1 – 12
28. Smeaton A.F. (2004) Indexing, Browsing and Searching of Digital Video. ARIST - Annual Review of Information Science and Technology, 38(8) 371-407.
29. Smeaton A.F., Gurrin C, Lee H, Mc Donald K, Murphy N, O'Connor N, O'Sullivan D, Smyth B and Wilson D. (2004) The Físchlár-News-Stories System: Personalised Access to an Archive of TV News. RIAO 2004, Avignon, France, 26-28 April 2004.
30. Snoek, C., Worring, M. (2005) Multimodal Video Indexing: A Review of the State-of-the-art. Multimedia Tools and Applications, 25, 5–35.
31. UK Audio Visual Archive Strategy Steering Group. (2004) Hidden Treasures. The UK audiovisual archive strategic framework. London: Chameleon Press
32. Wactlar H., Christel M., Gong Y., Hauptmann A. (1999) Lessons Learned from the Creation and Development of a Terabyte Digital Video Library. IEEE Computer, 32(2), 66-73.
33. Weber, H. (1997) Digitisation as a method of preservation? Commission on Preservation & Access. http://www.clir.org/pubs/reports/digpres/digpres.html.
34. Witbrock, M., Hauptmann, A.G. (1997) Speech Recognition and Information Retrieval: Experiments in Retrieving Spoken Documents. Proc. of the 1997 DARPA Speech Recognition Workshop http://www.nist.gov/speech/publications/darpa97/ (accessed 19.2.2005)
35. Yee, K.P., Swearingen, K., Li, K, Hearst, M. (2003) Faceted Metadata for image Search and Browsing. Proc. of CHI 2003 401-408.
36. Yeo, B.L. & Yeung, M. (1997) Retrieving and visualizing video. Communications of the ACM 40 (12) 43-52.