Rewarding the Viuser: A Human-Televisual Data Interface Application

Keir Smith
*iCinema Centre for Interactive Cinema Research,
University of New South Wales,
P.O. Box 259 Paddington NSW, 2021, Australia
keirs@cse.unsw.edu.au*

abstract: Interaction between any visual information user (viuser) and televisual data will be significantly enhanced by the application of the technologies and techniques presented in this paper. The current temporally linear, "streamed", one-channel-at-a-time televisual content display is replaced by a two part process of deconstruction and reconstruction. Multiple incoming streams of data are disaggregated, down to the most basic unit - the shot. Using purpose built software each shot is analysed, tagged and indexed according to its attributes, creating a database of shots. This database of shots can be searched by a number of different parameters including; keywords, image similarity, colours, patterns, setting, camera movement and sound. The selected shots are then interactively aggregated, according to user preference. Control over the resultant content and display is in the hands of the viuser. The need for this application became clear in the context of the *T_Visionarium* project currently being explored at the iCinema Centre, but is separate from that project in form and intent. The application has implications for future discussion, development and study of human-televisual data interaction.

1. Introduction

The Televisual Habitus

Television is ubiquitous in the Western world where it can be found in almost every house and sometimes in every room. The dominant mode of interaction with the TV monitor is through the equally commonplace remote control. Pointing and clicking can take you up or down a channel. If you know the number, or location, of the channel that you want to see, you can usually direct your receiver to it. You then view this televisual stream the same way the content provider intended: continuous, in a "full screen" mode. There are other modes of interaction, such as watching the content as chosen by another viewer (a family member perhaps) or viewing a summary channel like those offered by most cable TV providers or PC TV solutions. There have been many technological advances, from the ability to preview another channel in a small off center window [1] through to the much-touted HDTV (High Definition TV) or DTV (Digital TV)[2]. But even if you tape a program and watch it later, or you choose to view a player's statistics as you watch a game, your TV viewing is always temporally linear, its parts are presented in the same order that its producers intended.

General Description

Instead of trawling through your TV options by using a printed TV guide or an automatically generated channel by channel summary, imagine searching through every shot in the last 24 hours to find just those that had fast camera movement, loud noises or both and then having the capacity to combine these in fresh ways. This paper describes such a technology. The underlying infrastructure supplies an analysed, tagged and indexed database of every shot that has been segmented from data recorded over a selected period. At iCinema we are actively researching many alternative modes of interaction. One such mode is *T_Visionarium*, an interactive immersive virtual environment by Dennis Del Favero, Neil Brown, Jeffrey Shaw and Peter Weibel [10], which inspired the application presented in this paper and is discussed below.

Everyday TV offers a near totally passive mode of consumption, the only activity being the channel selection from a list of pre-ordered options. This paper does not attempt to propose a replacement to TV viewing habits. It proposes an alternative, active and potentially creative engagement with a familiar data set, one which may lead to new, unexpected and, often, unintended synchronicities. The goal is not to search through pre-digested information, for more information of the same type. This technology is not a Google for TV.

Currently data, of all descriptions, is being accrued and stored faster than we can access it. Greg Papadopolous [15] notes that processors are doubling performance every eighteen months while personal and commercial data storage is doubling every nine to twelve months. Jim Gray continues, adding that disk access times are only improving by ten percent a year, "the fundamental problem is that we are building a larger reservoir with more or less the same diameter pipe coming out of the reservoir"[14].

This technology is linked, but not limited, to incoming television signals. It could be used to explore the vast stores of audiovisual information currently available. In the future, as more descriptive video encoding schemes (such as MPEG 7) are implemented and supported, this application should be one of a large number of applications that attempt to make use of this ever growing reservoir.

2. Current Related Developments

MPEG 7/21

Although relevant and comprehensive, it is hard to tell which parts of the MPEG 7 or 21 specifications will be implemented by the main industry partners. If fully implemented, MPEG 7 offers a "standard for description and search of audio and visual content" [3].

MPEG 7 intends a secondary time-coded, descriptive stream, in a format such as XML, to run concurrently with audiovisual stream. This descriptive stream can also be stored or parsed separately from the audiovisual content [4]. If MPEG 7 is supported to the fine level of granularity required to categorise, delineate and describe each shot or scene, this standard could be a significant advance for audiovisual applications that use both live and stored data in alternative and creative ways.

Automated Analysis, Indexing and Tagging

There are a number of research institutes committed to video indexing and cataloguing. Some of the most successful of these are the Multimedia Knowledge Management group at the Imperial College London [5], the Centre for Digital Video Processing at Dublin City University [6] and the Informedia-II research project at Carnegie Mellon University [7]. They have been presenting their research work and technological advances at conferences and competitions such as Video Retrieval Evaluation workshop at TREC [8]. Techniques like real-time or near real-time shot segmentation have been perfected. Computer vision, image analysis, speech recognition, pattern matching, colour spectral analysis, face recognition and sound analysis technologies have all been applied to video data indexing efforts [12,13,16,17].

3. A Human-Televisual Data Interface Application

Through a graphical user interface (GUI), that is either augmented/superimposed on the output display or supplied on an auxiliary device, the viuser is able to select any number of query parameters to be applied to the database. The parameters can be chosen from any of the supplied categories and applied at distinct levels of specificity. The viuser can add further search criteria to the current configuration or abandon it and begin a new combination of queries. As the user continually updates his or her selection to incorporate the more complex options available, the original baggage of conventional habits of viewing are forgotten.

The Application

The user interface (UI) for this application is a wireless, colour screen, personal digital assistant (PDA), the Zaurus SL-C700 from Sharp. This Zaurus was chosen from a number of PDAs of similar size and weight, for its screen resolution and its Linux development compatibility and support. The Zaurus SL-C700 weighs about 225 grams and has a 74 by 53 millimetre front-lit screen with a 640 x 480 pixels resolution [9]. The SL-C700 screen is touch sensitive and the PDA comes with a stylus and built-in keyboard. The keyboard will be hidden, and the whole PDA encased in a protective rubber covering with only the screen showing, while being used for this project.

The screen of the PDA contains 4 buttons, ‘change selection’, ‘new search’, ‘save’ and ‘load’. If the viuser chooses the ‘new search’ option the screen will display a series of buttons arranged in a row, one for each category. The categories are *keywords, image similarity, colours, patterns, source channel, setting, camera movement* and *sound*. At the bottom of the screen is a docking area where you can place your parameters as you choose them and at the far right of this is a ‘display selection button’. To the left of the docking area is a ‘return’ button.

The *patterns* category contains buttons with stylised representations of *vertical lines, horizontal lines, a face* and *a ball*. The *sound* category contains buttons for *loud, crying, laughter, music* and *silence*. The *camera movement* category *contains pan left, pan right, pan up, pan down, zoom in, zoom out, chaotic scene* and *static scene*. The *colour* category contains a colour wheel or a set of RGB levels, allowing accurate colour selection. The *setting* category contains *horizon, night, outside* and *inside* as options. The *keywords* section contains *machines, children, water, greetings, touch, food, text* and *dialogue* among others. *Source channel* offers a text list of the channels that were recorded to form the shot database. The *query images* category offers a series of small images including *an old mans' face, a red ball, a bird-eating spider, a pair of shoes* and *a coastline*, there are more query images if the viuser scrolls to the left or the right.

These catagories, and the elements within them, represent the first generation of possible selection criteria. An empirical usability study is being designed to inform the next iteration of the user interface and the selection criteria that it contains.

After selecting 'new search', one could then enter the *camera movement* section by selecting the appropriate button. If the *pan right* button is selected with a finger or the Zaurus stylus, a small iconic version of the button would appear in the search parameter dock at the bottom of the PDA's display. Any number of other parameters could be added or removed before the viuser commits to their selection criteria by clicking the 'display selection' button.

User selections are wirelessly communicated to the visualisation device, the visual component displays on a commonplace CRT monitor and the audio output comes from a speaker system, alternately a projector and a public address (PA) system might be used.

Technical Specification

For this deployment we recorded 24 hours of TV data, one hour each from 24 satellite TV channels from the Australasian region. At 25 frames per second at PAL resolution, this amounts to about 1.8 Terabytes of uncompressed data. The data was then segmented into shots and a representative key frame for each shot was extracted using the purpose built software. A test of relevant TV programming suggested that the average length of a shot would be about four seconds, or 100 frames at 25 frames per second. Although the average time varies greatly due to the type of program (MTV for example is closer to one second per shot) for a wide range of programming this is a good estimate. This means that there will be roughly 22,000 images in our keyframe database. Using software developed in-house at the UNSW Computer Science and Engineering school, these images were then analysed in two separate ways. Initially a large image texture, shape and colour feature vector was created. Analysis of the feature vector reveals shape, pattern and colour histographic information and allows us to calculate image similarity. The second analysis extracts audio density, in-frame movement and panning/zooming camera motion. It also offers us a semi-automated manual tagging environment.

This project seeks to predict the future of possible television interaction, as such it is not limited by the automatic analysis that is possible using current technologies. In addition to the automated techniques we manually tagged attributes that current technology cannot supply. It is hoped that in the future all currently used parameters, and many we've not thought of, are generated in real-time for the home viuser.

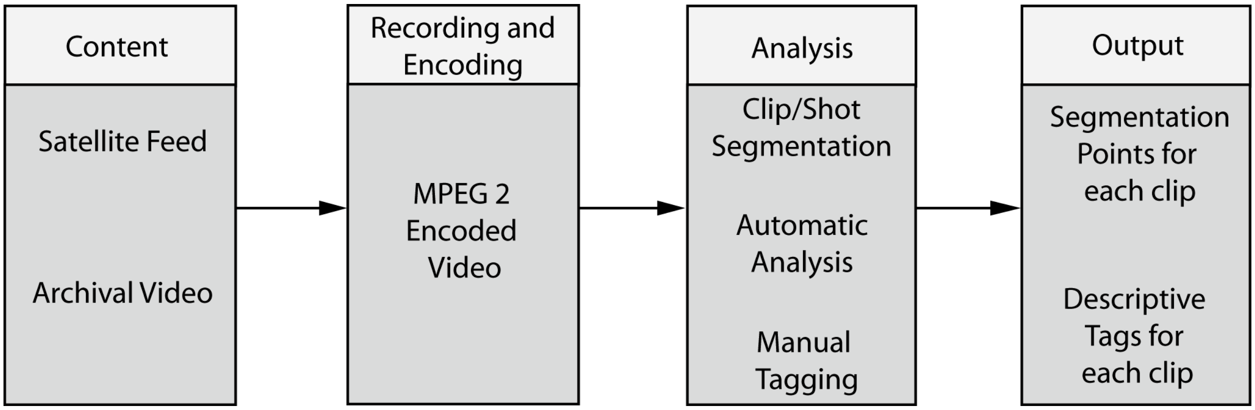


Figure 2. The data preparation process.

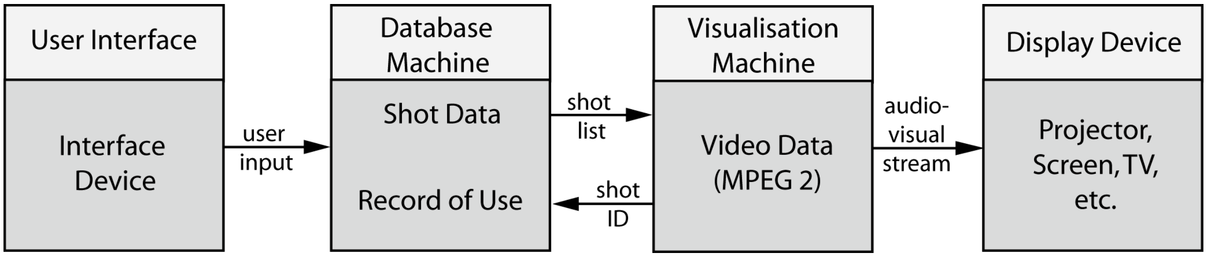


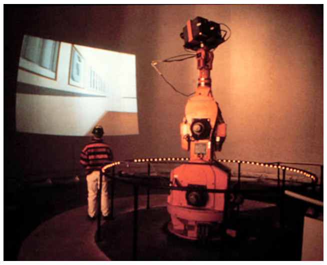
Figure 3. The interactive display system.

All the automated and manually entered information is represented in a purpose built postgreSQL database, which is indexed. Matches for all the supplied query images are calculated, allowing us to remove the feature vector from the search area. It is this final reduced indexed database that the interface device formulates queries for, based upon sometimes simplistic, often complicated, user selections.

The derivation of the display can be thought of conceptually as three separate processes, which may or may not be running on different computers: the visualisation machine (VM), the database (DB) and the user interface (UI). The UI supplies the DB with the selection criteria, the DB then uses this to create a shot list, an ordered list of matching shots. In the case of *T_Visionarium* a shot list for each channel is supplied, in this paper's example only one list is supplied with shots ordered from the most to the least relevant. Shots that do not match are not referenced at all. This shot list is supplied to the VM, which begins their ordered playback. The VM also sends the ID of each shot as it is played, to the DB for analysis, logging and for additional interaction, such as making a own movie out of the shots previously displayed or for a replay functionality.

If the selection is so specific that only two shots are retrieved from the whole database, which can happen especially when using the channel filter, then those shots are simply looped. The same system can be easily adapted to show hundreds of channels at any one time, in which each display stream is the output of a different channel filter being applied. The technology being developed is easily scalable to accommodate any number of additional channels.

T_Visionarium



The ideas presented here began to emerge in a very specific context, as part of my research within a team of digital artists and computer scientists working on the *T_Visionarium* project; this work is greatly indebted to them. This project is funded by the Australian Research Council, for the iCinema Centre for Interactive Cinema Research at the University of New South Wales.

T_Visionarium: An Extended Virtual Environment, is an “interactive immersive virtual environment set within a dome, 12 meters in diameter and 9 meters high, made of inflatable fabric and articulated to a satellite reception, recording and database system. It allows viewers to spatially navigate a televisual database and apply a recombinatory search matrix to create emergent narratives from the database's network of digital streams.” [10]

Upon entering the dome the viewer places a position-tracking head-mounted device, with cableless stereo headphones. After a short personal calibration, the system will, with the feedback from the position tracker, tell the visualisation machine at which point on the dome the viewer is facing. As the viewer looks around the dome, the projector, which is mounted on a pan and tilt mechanism, displays the image where their face is pointing. All the channels remain virtually placed in the same area of the dome, for example CNN will always be above the entrance and BBC will remain 25 degrees to its right. The viewer will be able to see only a small section of the dome as they move their head.

Once activated by a small tactile switch, a semi-transparent menu based selection system, similar to that of some VCRs, is superimposed over the current display. As described above the viewer can change selection criteria or begin a new combination of queries. With each user action, the shots selected for every channel, in all parts of the dome, will be updated. For example, if the viewer makes a selection specifying only shots with faces and loud noise, the space could be transformed into a sea of singing shouting faces, each one revealed as the user looks around the dome.

In *T_Visionarium*, after a user chooses a set of selection parameters, the system "extracts and distributes all the corresponding broadcast embodiments of the parameter over the entire projection surface of the dome. The viewer, by moving his or her head in different directions repositions the projected image, shifts from one channel's embodiment of the selected parameter to the next" [10].

The output of *T_Visionarium* is such that only a few hundred people a day can witness it. This raises the issue of limited compared to (in principle) unlimited user access. This paper provides a solution to this problem based on an accessible screen-based framework, where many interaction paradigms can be explored and studied. In contrast to *T_Visionarium* this accessible system shows just one shot at a time allowing the user to see which shots, in a full 24 hours of television, most closely match their selection and then to interactively add other selection criteria.

4. Future Possibilities

At iCinema we are currently working on a number of modes of interaction, both from a user evaluation standpoint and as a test bed for artistic exploration. These include:

1. A wide screen, touch-screen version. In this version all the possible channels are on the screen at all times. The centre of the screen is filled by the active channel, which supplies the sound track. Small, low-resolution versions of the other channels are presented either as if floating in a sea or as a rigid matrix, surrounding the main display. The user can select any one of the other, low-resolution, channels to replace the current active channel. A superimposed UI can be brought up and combinative queries can be formulated and applied to all visible channels.
2. A performance version. In this application an auxiliary process, with access to the same database, is used to display the UI and small low-resolution version of the current output for the content author/artist. Another process is providing the output visualisation on a screen or projector, for audience view. The artist can use a mouse and keyboard to interact with the UI.
3. The utopian home user version[1]. Imagine a couch, with built in PDA, in front of a large TV display augmented by eye-tracking technology that can offer very accurate information about where on the screen the viewer is looking, for example [11]. This screen position information could be used as a form of channel selection. The display

[1] I can imagine some people would view this as some sort of nightmare.

would look the same as outlined in point 1. In this case however if a viewer looks for a prolonged period at a clip playing in a low-resolution window surrounding the main display, this clip will begin to increase in size. If the viewer maintains their focus on this expanding window, it will replace the current main display in the centre of the screen and begin to provide the audio track. When the user looks down at the UI on the PDA to change their shot selection criteria, the system will know that the eyes are not looking at the screen and will maintain the current output until the eyes' focus returns.

It is intended that a function be added that allows the user to interactively order and record a list of shots into their own "show". The raw ingredients would be the most recent shot list, or a previously saved shot list. Although the resolution constraints of the PDA predicates such functionality for the version outlined in this paper, this functionality would be possible for the larger resolution touch-screen and performance versions.

Content producers and providers, as envisaged by MPEG 7, could generate and supply a description stream meeting the requirements of a system such as the one outlined in this paper. This would allow a screen-based installation in the home, by removing the time and computationally intensive segmentation, analysis and tagging elements of the system.

Choosing to split the video data into shots is an artifice of this particular solution. An alternate solution could include details of scene change (although scenes are much harder to accurately distinguish, either computationally or by hand) so that the output would be presented in a scene-by-scene structure. This sort of application would be more beneficial in a learning or teaching environment.

A serious question raised by this, and any other work of its kind, relates to digital rights management (DRM). Although DRM systems are "aimed at increasing the kinds and/or scope of control that rights-holders can assert over their intellectual property assets" [18], it is intended that this work would fall under "fair use" protection from copyright law. But Fred von Lohmann, senior intellectual property attorney for the Electronic Frontier Foundation, argues that under the planed changes to copyright law, DRM will fail to protect the full range of future fair uses made possible by new technological advances, such as DTV [18]. Whether or not this work would survive a court challenge is another matter entirely.

In the utopian world described above, all content producers, including advertisers, would have full knowledge of the technology and could cater their imagery to meet as much of the tagging criteria as possible. But in the utopian world of televisual interaction envisaged by the author, the system would automatically remove all the advertisements unless the *retain advertisements* filter was enacted.

Other future work efforts include: incorporation of live query images and sounds, reduction and eventual elimination of manual input requirements and close monitoring of the relevant MPEG standards.

5. Conclusion

By taking commonplace data, like television quality video, and reducing it to its constituent component, the shot, a new way of interfacing with the data is made possible. In this application segmented shots are tagged, catalogued and indexed for interactive, real-time retrieval. The (re)combination and display of the shots is controlled by the individual user's selection criteria.

The next stage in the development of this application is an empirical usability study of the various techniques proposed and the current selection criteria, aided by the logging apparatus being incorporated as part of the software implementation. By utilising an iterative, user-centred interface/interaction design process, we hope to create a usable interface to a massive data reserve, that will allow a (any) viuser to become their own content provider. The more interesting the source material and inventive the query, the more rewarding the final visualisation.

List of figures

References

[1] The Toshiba CN36H97 TV is one example of the channel within a window technology. It contains 2 turners allowing you to watch 2 channels simultaneously. http://www.supremevideo.com/television/toshiba/cinemaseries/cn36h97.htm

[2] Advanced Television Systems Committee houses the technical standards for digital television. http://www.atsc.org/

[3] Motion picture experts group official website. http://mpeg.telecomitalialab.com/

[4] MPEG 7 specification. http://mpeg.telecomitalialab.com/standards/MPEG-7/MPEG-7.htm

[5] Multimedia Knowledge Management**,** Department of Computing at the Imperial College, London, has made some remarkable advances. http://km.doc.ic.ac.uk/

[6] The Centre for Digital Video Processing is a cross-disciplinary research centre are a collaboration between the School of Computing and the School of Electronic Engineering at Dublin City University. http://www.cdvp.dcu.ie/

[7] CMU's Informedia II Digital Video Library research project covers auto summarisation and visualisation across multiple video documents and libraries http://www.informedia.cs.cmu.edu/dli2/

[8] The TREC[2] video retrieval evaluation workshop. http://www-nlpir.nist.gov/projects/trecvid/

[9] Technical specification from: http://www.dynamism.com/zaurus7xx/index.shtml

[10] The iCinema *T_Visionarium* project website. http://www.icinema.unsw.edu.au/projects/prj_tvis.html

[11] Seeing Machines, Canberra Australia, create vision based human machine interfaces. http://www.seeingmachines.com

[12] M. Brown, J. Foote, G. Jones, K. Sparck Jones, S. Young, Automatic Content-Based Retrieval of Broadcast News, *ACM Multimedia*, 1995.

[13] R. Cendrillon, B. Lovell, Real-Time Face Recognition using Eigenfaces, *Proc. of the SPIE International Conference on Visual Communications and Image Processing*, 2000.

[14] J. Gray, A Conversation with Jim Gray, *ACM Queue* vol. 1, no. 4, 2003.

[15] G. Papadopolous, The future of computing. Unpublished talk at *NOW* workshop, Lake Tahoe, USA, July 1997.

[16] C. Snoek, M. Worring, Multimodal Video Indexing: A Review of the State-of-the-art, *Multimedia Tools and Applications*, 2002.

[17] M. Witbrock, A. Hauptmann, Speech Recognition for a Digital Video Library, *Journal of the American Society of Information Science*, 1998.

[18] F. von Lohmann, Fair Use and Digital Rights Management: Preliminary Thoughts on the (Irreconcilable?) Tension between Them, *Computers, Freedom & Privacy*, 2002.[3]

[2] The TREC (Text REtrival Conference) series is sponsored by the National Institute of Standards and Technology in the USA http://trec.nist.gov

[3] http://www.eff.org/IP/DRM/fair_use_and_drm.html