UniS University of Surrey

EPSRC

TIWO – Television in Words

EPSRC GR/R67194/01, January 2002-5

Report on Workpackage 3

AuDesc System Specification and Prototypes

15 November 2004

FINAL VERSION

Compiled by:
Andrew Salway
Department of Computing, University of Surrey

With input from:
TIWO Round Table
Marie Campbell (BBC), Garry Duguid (ITFC), Denise Evans (RNIB), Claude le Guyader (ITFC), James O'Hara (ITFC), Jane Westrop (BBC), Russ Wood (Softel)

PhD Students at Surrey: Elia Tomadaki, Andrew Vassiliou, Yan Xu
MSc Student at Surrey: Mike Graham
BSc Students at Surrey: Elizabeth Jones, Pete Sessions

Contact: a.salway@surrey.ac.uk

TIWO Project web-site:

www.computing.surrey.ac.uk/personal/pg/A.Salway/tiwo/TIWO.htm

Contents

1 Executive Summary

As part of the TIWO project we are specifying, designing and prototyping a system, called AuDesc, to assist in the production and repurposing of audio description. Section 3 summarises the requirements and desired functionality of such a system as discussed by TIWO Round Table members in meetings from December 2001 to September 2003. Section 4 comments on the feasibility of implementing such functionality whilst considering available commercial systems and the state-of-the-art as described in the literature on intelligent multimedia information systems. Some the challenges raised by the desired functionality are aligned with TIWO's research aim, which is to develop a computational account of narrative in multimedia systems. The five prototype systems reported in Section 5 are being implemented and evaluated in the context of the desired functionality from the round table and in the context of this research aim.

2 Description of TIWO Workpackage 3 from Project Proposal Document

An audio description system will be specified, designed and prototyped following standard software engineering techniques. It will incorporate the deliverables of WP1 and WP2: continuous prototyping will facilitate quick evaluation of the results from these workpackages with feedback from the Round Table.

Requirements gathering, testing and evaluation will be carried out in Round Table meetings and during visits to members. The emphasis will be on the repurposing of existing systems, including systems developed at Surrey that integrate visual and textual data, and commercial multimedia database systems like Informix. System development will also take account of existing software products used for audio description, like Softel's ADePT, and off-the-shelf speech recognition and video analysis packages.

The system will be evaluated by the user group for the tasks of style checking audio descriptions, customising them and indexing-retrieving video content. The evaluation will be designed to gauge how efficiently the system generates representations of video content from audio descriptions and how well the representations capture the 'meaningful' aspects of video content for the users.

Milestone 3a - month 12: AuDesc system specification
Milestone 3b - month 27: Delivery of prototype AuDesc system to Round Table
Milestone 3c - month 30: Evaluation of prototype AuDesc

3 Requirements Analysis and Desired Functionality

Ideas for the requirements and the functionality of the AuDesc system were discussed by the Round Table during meetings in December 2001, April 2002, September 2002 and April 2003. This section attempts to summarise the major points arising from these discussions. First some general requirements relating to the production of audio description are listed (Section 3.1). The desired functionality elicited from the Round Table can be split into that which is to do with supporting the work of audio describers in the production of audio description (Section 3.2), and that which is to do with facilitating the repurposing of audio description, e.g. for video indexing, retrieval and browsing (Section 3.3). There is some cross-over in these requirements and the technologies that may support them, e.g. the indexing-retrieval of video material may help an audio describer in their professional practice as well as: help a film fan access their favourite films.

3.1 Requirements Relating to the Production of Audio Description

Audio description helps visually impaired viewers watch most kinds of films and television programmes, including dramas, situation comedies, soap operas, documentaries, and sometimes live news broadcasts. In the gaps between existing speech, audio description gives key information about scenes, events, people's appearances, actions, gestures, expressions and body language so that in effect the story conveyed by the moving image is retold in words. There is potential for audio description to be used by the whole television and film audience, for example to 'watch' a film on an audio CD or on a device with a small visual display. Audio description also supports access to theatre, art galleries, museums, sporting events and multimedia content but these domains are beyond TIWO's scope. The following factors were considered by the Round Table to be relevant when specifying requirements for technology to support the production of audio description.

General Factors for the Production of Audio Description

- Audio description is delivered via digital television broadcasts, in some cinemas and on some DVD/VHS film releases.
- Audio description is normally scripted before it is recorded.
- Audio describers must take care to ensure that audio description interacts with the existing dialogue and sound effects.
- They are also careful to strike the right balance between frustrating the audience with insufficient information to follow the story, and patronizing them by spelling out obvious inferences. The appropriate amount of description may vary between a film watched in a cinema and on television.
- Describers should maintain their 'descriptive parlance' throughout an audio description. Appropriate describers and description styles are selected for programme / film.

~ To produce an 8000-word audio description for a 2-hour film may take 60 person hours, with many viewings and more than one describer: however a 30 minute soap opera which is almost full of dialogue and has familiar scenes and characters may take only 90 minutes to describe.

~ Audio description has to be tightly synchronised with programme / film material, to within a frame, especially when there are only short gaps in dialogue. Software is available to assist in the preparation of audio description scripts, the recording of audio description and its synchronization with programme and film material. Current computer systems to assist the production of audio description present video data on screen alongside a window that marks time-coded speech-free intervals into which the describer can type their descriptions. Once the script has been completed and reviewed it is spoken, recorded and synchronized with the video data by time-codes.

~ Legislation and regulation mean that audio description is becoming increasingly available in countries like the UK, US, Canada, Germany, France, Spain, Italy and Japan.

~ The legislation and regulation tends to focus on the quantity of audio description that broadcasters must provide, but in time it should also address the quality of audio description.

~ Some audio description professionals follow guidelines produced by government agencies, e.g. the UK's Office of Communications. These include guidance about what information to include and the style of language to use. There are differences in styles, e.g. between British and American audio description.

~ Some organisations develop such guidelines into in-house procedures for the production of audio description, based around the use of software systems, from initial scripting through to editing and then recording.

~ New delivery mechanisms / technology give further scope for audio description. Background information can be given either before a film on VHS, and more so on DVD, or TiVO or during the opening titles of a TV programme.

SUMMARY

- Audio Description is a professional process with a strong concern for guidelines and quality control.
- The production of audio description can be a collaborative process with teams of describers working on the description of a single film or television series.
- Once produced, there is an increasing need to 'repurpose' audio description:
    - For different languages.
    - For different audiences, e.g. cinema vs. television, and young vs. old.
    - When programme / film material is edited and reused.

Related Information

- For the latest information about audio description from a global perspective see the website of Audio Description International: www.adinternational.org/

- The W3 Consortium are developing "an interoperable timed text format" which may have implications for the production and delivery of audio description in the future, see: www.w3.org/AudioVideo/TT/

- Conferences dealing with theoretical and practical issues for audio description in a global context include:

- In So Many Words: language transfer on the screen, 2004: www.surrey.ac.uk/LIS/CTS/insomanywords.htm

- International Conference on Audiovisual Translation, 2005: www.fti.uab.es/transmedia/

- Languages and the Media: www.languages-media.com

3.2 Functionality to Support the Production of Audio Description

The following are ideas for functions to be incorporated into existing software and procedures for audio description production. They were elicited and discussed during meetings of the TIWO Round Table. These functions broadly address the wish to make the production of audio description more effective, by saving time whilst maintaining, if not improving, the quality of the description. The functions are ordered here according to our judgements about how feasible it would be to implement them given current technology, and the current state-of-the-art in research and development of multimedia information processing technologies.

3.2.1 On-line Access to Lexical resources

Give audio describers easy, and possibly proactive, access to lexical resources.

- Thesauri – to prevent repetition in descriptions;
- Specialist terminologies – e.g. when describing documentaries;
- Concordances of existing audio descriptions for 'ways to describe X' to prevent repetition, e.g. different ways to describe the sea;
- Concordances of earlier descriptions from within the current film or TV series which could be useful to ensure consistent description, especially when there is more than one audio describer involved.

3.2.2 On-line Access to Related Textual Information

Give audio describers easy, and possibly proactive, access to online texts like film scripts and plot summaries, or web-pages about the subject matter of a documentary.

- There is a danger of being given too much information when the need is to concentrate on what is to be described.
- There is a need to analyse what kinds of information are available in the different kinds of collateral texts associated with television programmes and films, such as scripts – spotting lists, dialogue lists, shooting scripts, post-production scripts, original books, screenplays, other audio descriptions, film reviews, directors' commentaries, etc.

3.2.3 Browsing and Searching Video Material and Audio Description Scripts

Give audio describers alternative ways to browse and search through the video material and the audio description script that they are working on, e.g. to work out which scenes need further description, or to see how they described a character previously.

- Currently audio describers navigate the video material and audio description using conventional video controls and the time-codes of audio description. An interface could present key-frames for scenes in the video to facilitate video skimming.
- A describer should be able to associate personal notes with scenes and search over these, as well as the current audio description script, and subtitles, to locate scenes containing certain characters and objects. Some processing will be necessary to extract a list of names and to resolve personal pronouns.

3.2.4 'Live' or nearer real-time audio description

Give audio describers tools to support 'live' audio description, either for live television broadcasts when scripting is not possible, or for producing a first draft audio description script by speaking while viewing material for the first time.

Tools might include speech, audio and video processing for:
- speech recognition;
- automatic detection of gaps in dialogue / quiet (where audio description can go);
- automatic scene detection – to ensure some description of every scene
- identifying key-frames that are high priority for description

Once the describer's verbalisation has been transcribed through speech recognition, further processing could perhaps automatically edit the description to fit in with dialogue gaps, and scene boundaries, and perhaps bolster it with information from related textual information, and other sources.

There is a need for observations of what happens when audio describers speak while watching material for the first time.

3.2.5 Automatic Style-checking

Give audio describers tools that check audio description scripts against an organisation's in-house style, or against national guidelines.

- These tools could help to train novice audio describers
- They might also be useful for quality control purposes, e.g. to help senior audio describers in editing others' scripts
- Quality stands to be an increasingly important issue when broadcasters are required to provide audio description in increasing quantities, and so possibly resort to using lower-skilled describers.

Automatic checks could be made for features such as:
- Confusing use of pronominals
- Overuse of proper nouns, and repetition of other words
- Use of inappropriate tenses
- Vocabulary appropriate for audience age-group, and genre/period of film

3.2.6 Customisation and Translation

Give audio describers tools that help them customise, or translate, existing audio description scripts for different audience groups. (Or tools that do this under viewers' control at the point of viewing).

- Some audience groups may prefer more/less detail or interpretation in the audio description: starting with a maximal description perhaps some could be automatically filtered (by recognising elaborative, or interpretative text fragments).
- Different audience groups (within the same language community) might prefer different kinds of vocabulary, e.g. youth vs. adult, or regional dialects: maybe dictionaries could be used to swap automatically between vocabularies.
- On some delivery platforms, like DVD and TiVO, there is potential to 'layer' audio descriptions which do not have to fit in with the realtime playout of the video material, i.e. the viewer can stop the video in order to listen to extended audio description.
- In an evermore global media market then translation of media assets, including audio description in a key issue for media producers and suppliers. The relatively simple nature of the language used in audio description (simple that is say compared to a novel), may mean automatic translation systems fair better than usual. However current automatic translation technology is unlikely to satisfy viewers when top quality audio description is required (cf. current poor quality automatic subtitle translation).
- If not automatic translation, then machine tools can be used to assist the translation of audio description, e.g. to identify key events that need translation (and check that they are covered by translation), and to check that the emotional tone of an audio description is maintained in a translation.

3.2.7 Subsequent editing of programme / film material

Give audio describers tools that help them edit audio description when a programme or film is edited for a subsequent viewing.

Much programme and film material is edited for subsequent viewings, either to fit into television schedules, or in accordance with what can be shown at certain times of day. Without producing a new audio description from scratch, this raises problems to do with making smooth and coherent edits:

- The speaker's voice – the original speaker might not be available
- The spaces available for audio description change
- The visual content to be described changes

3.3 Functionality for the Repurposing of Audio Description

Subtitles and closed captions have been repurposed in various ways, such as people watching television when someone in the same room is reading, and by video retrieval systems like *Informedia* as a source of keywords for video indexing. Such repurposing of audio description could hasten its uptake.

The fact that audio description is produced specifically to be informative about the characters, objects and events depicted by moving images, the fact that it uses a relatively simple subset of natural language, and the fact that audio description fragments are linked to video data by timecodes, all mean that audio description is a ripe source of machine-processable representations of video content. The extent to which machine-processable representations of video capture a user's understanding of video content determines how intuitively video databases can be accessed via retrieval, browsing and summarization applications.

Various kinds of representations are required for numerous tasks that researchers envisage intelligent multimedia information systems performing:

- **Video Retrieval** – locating scenes in films containing certain characters, or finding films that have a similar storyline to one that you have previously enjoyed.
- **Hypervideo Browsing** – browsing video files in a non-linear, but non-arbitrary, fashion, e.g. to explore cause-effect links between different events depicted in a film (cf. possibilities with DVD players).
- **Video Summarisation** – presenting viewers with an abridged version of a film / programme, or summarising the action that they have missed if joining a television series midway (cf. possibilities with TiVO boxes).
- **Question Answering** – "why did that just happen?"
- **Proactive links to related information** – e.g. webpages when watching a documentary (cf. possibilities for making a film the centre of an educational experience for children – an idea suggested by Prof. Fabio Ciravenga, University of Sheffield).
- **Enhancing the viewing experience** – e.g. by controlling the viewing environment to suit the mood of the film.

4 Currently Available Technology

This section lists some systems that in various ways do, or could, provide some of the functionality to support the production and the repurposing of audio description.

- Softel design and manufacture innovative media products including Interactive TV broadcast servers & Authoring Solutions, MPEG stream analysers, Teletext systems, and Subtitling/Closed Captioning workstations & transmission solutions. Including, Softel ADePT – to prepare audio description scripts and to synchronise recording: www.softel.co.uk/l4_products_subtitling_adept_main.php

- For USB devices and PCI cards that enable digital TV reception (including audio description) on PCs: www.nebula-electronics.com

- For an overview of hardware to facilitate the delivery and reception of audio description in different contexts see: www.adinternational.org/ADItech.html

- For organisation-level multimedia archiving, retrieval and delivery, including automated speech and video processing: www.virage.com (incorporating Dremedia), and for speech search www.softsound.com, both part of www.autonomy.com/

- Also, Aurix Media Content Management Suite, including 20/20 speech technology for speech detection and speaker / speech recognition, e.g. for assisted subtitling: www.aurix.com/dynamic/about/

- Informix video datablade (part of IBM): www-306.ibm.com/software/data/informix/blades/video/

- Technologies to support teletext, subtitling and interactive TV: www.sysmedia.com

- Informedia II Digital Video Library: synthesises audio, video and text processing for indexing video data, and a variety of retrieval and browsing strategies: www.informedia.cs.cmu.edu/ [Research at Carnegie Mellon University]

- GATE (Generalised Architecture for Text Engineering) – freely available system for language processing tasks, especially information extraction: gate.ac.uk

5 Prototyping AuDesc

In TIWO we have developed models and algorithms for generating machine-executable representations of semantic video content from different kinds of text that describe the moving image. Previously, systems dealing with semantic video content have treated it as an inventory of events and existents, organised in space and time, but have not dealt the narrative aspects of moving images. Video retrieval systems tend to use visual features, or information extracted from one kind of text – typically subtitles, or closed captions. We focussed on films where dealing with semantic content involves modelling and generating representations of a film's narrative, i.e. a sequence of events connected by cause-effect relationships where the agents of cause-effect are often characters with mental states, goals, beliefs and desires. Our approach is to extract and integrate information from different kinds of texts associated with films, including film scripts, plot summaries and audio description. Results will be applied to assist audio description professionals and film viewers in retrieving and navigating digital film libraries. Progress has been made with respect to three main challenges: cross-document co-reference; extraction of information about characters' emotions; and, novel kinds of video browsing.

Five systems have been prototyped after consideration of: the requirements from Section 3 that cannot be met with current technology; the current state-of-the-art in international research and development of multimedia information processing systems; and, most importantly, the overarching aim of the TIWO project which is to "develop a computational understanding of narrative in multimedia systems".

A paper written early in the TIWO project gives the background to these developments:

Salway, Graham, Tomadaki and Xu (2003), 'Linking Video and Text via Representations of Narrative', AAAI Spring Symposium on Intelligent Multimedia Knowledge Management, Palo Alto, 24-26 March 2003. **Available from TIWO website**

5.1 Keyword Video Retrieval

A basic video retrieval task is to provide results for queries of the form 'Show me pieces of video of X' where X is a description of something or something happening. Whilst computer vision is not sufficiently developed to support such retrieval from general domain video sources, like films, there seems to be an opportunity to reuse audio description. After all, audio description is specifically produced to put important visual information into words. Because audio description scripts are timecoded the video retrieval task can be tackled by: (i) matching the keywords in the query with words in the audio description; (ii) identifying the timecodes of audio description fragments with matching words; and, (iii) returning the user video clips around those timecodes.

Of course this approach is liable to problems faced by most text information retrieval systems, such as synonymy and ambiguity. The other limit on its success is the extent to which the audio description mentions things of interest to the user of the video retrieval system, and perhaps the degree to which the timing of the audio description is aligned with the on-screen action.

As part of the TIWO project a system was developed to implement this idea of keyword based video retrieval using audio description scripts. To evaluate this idea we considered one of the tasks used as part of VideoTREC. About 30 minutes of film was broken into 30-second intervals and each was marked as either relevant or irrelevant for each of the VideoTREC queries. Then the audio description for the film was analysed to determine how many relevant clips could be potentially matched on the basis of keywords in the audio description: by 'potentially' we mean allowing for some query expansion. The results were quite encouraging for some kinds of queries.

As well as helping with video retrieval, the alignment of keywords from audio description with intervals of video data may also be useful for training video feature detectors, i.e. systems that learn associations between the audio-visual features of video data and words (labels of what is in the video). To pursue this idea might require the stricter alignment of keywords with the relevant video intervals in which case temporal information in the audio description might need to be processed in order to annotate temporal relationships between the text fragments and the video data.

Further Information

The system implementation and evaluation against VideoTREC queries was carried out by Pete Sessions for his final year project in his BSc Computing and Information Technology at University of Surrey. His report and software are held by the Department of Computing.

5.2 Extraction of Information About Characters' Emotions in Films

In order to deal with the narrative content of video data we have to deal with more than descriptions of what can be seen happening on screen. We have found that one way to access information about a film's narrative is to concentrate on *characters' emotional states*. A character's emotional state can be considered as their reaction to events unfolding around them, and their reaction is determined by how they think those events impact on their goals. Thus information about characters' emotional states can be revealing of a film's narrative, modelled as a sequence of events connected by cause-effect relationships. We have developed a method for extracting information about characters' emotions from time-aligned texts such as audio description and film scripts. This information appears to be useful for video retrieval by story similarity, and for reasoning about a film's narrative.

Put briefly, emotion tokens found in audio description scripts are mapped to emotion types, e.g. 'afraid' → FEAR, and that instance of the emotion type is given the timecode of the audio description fragment containing the emotion token. This method gives output like the 'emotion graph' for the film *Captain Correli's Mandolin*, shown in Figure 1. In the audio description for the film *Captain Corelli's Mandolin* there were 52 tokens of 8 emotion types. The story of this film concerns a love triangle between an Italian officer (Corelli), a Greek woman (Pelagia), and a Greek partisan (Madras) on the occupied Greek island of Cephallonia during World War II. A high density of positive emotion tokens appear 15-20 minutes into the film, e.g. JOY and LIKE, corresponding to Pelagia's betrothal to Madras. The negative emotion tokens which immediately follow are associated with the invasion of the island. The cluster of positive emotions between 68-74 minutes occurs during scenes in which the growing relationship between Pelagia and Corelli becomes explicit. The group of FEAR, DISTRESS and SELF-REPROACH tokens between 92-95 minutes maps to a scene in which German soldiers are disarming their former Italian allies, during which a number of Italians are gunned down. The clusters of emotion tokens appear to identify many of the dramatically important sequences in the film.

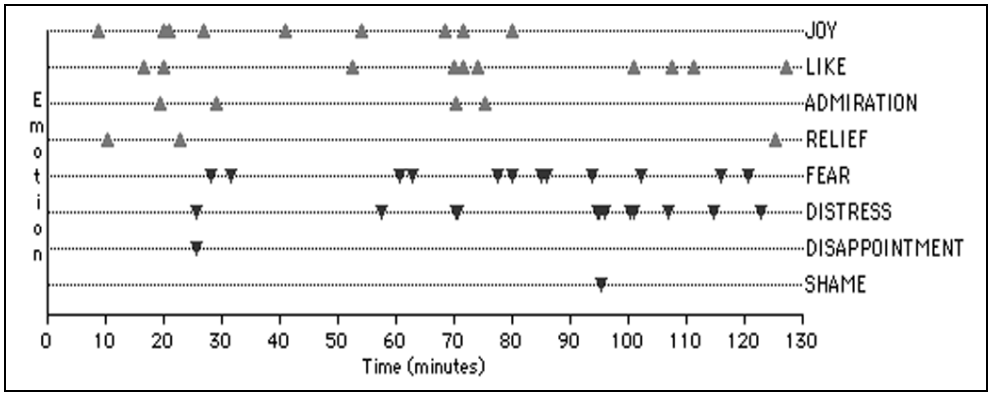

Figure 1. A plot of emotion tokens found in audio description for *Captain Corelli's Mandolin*; emotions only included if > 0.

For evaluation of this method, and a discussion of how it might be applied to video retrieval by story similarity, video summarization and video browsing, see (Salway and Graham 2003). The method was refined to associate each instance of an emotion type with one of the film's characters. The audio description scripts were passed through a part-of-speech tagger and rules based on parts-of-speech determined which character the emotion should be associated with. The results were passed automatically to Microsoft Excel for display, Figure 2. This strand of the project was extended further by testing the method on another type of text (screenplays) and by developing a metric to measure the similarity of two graphs.

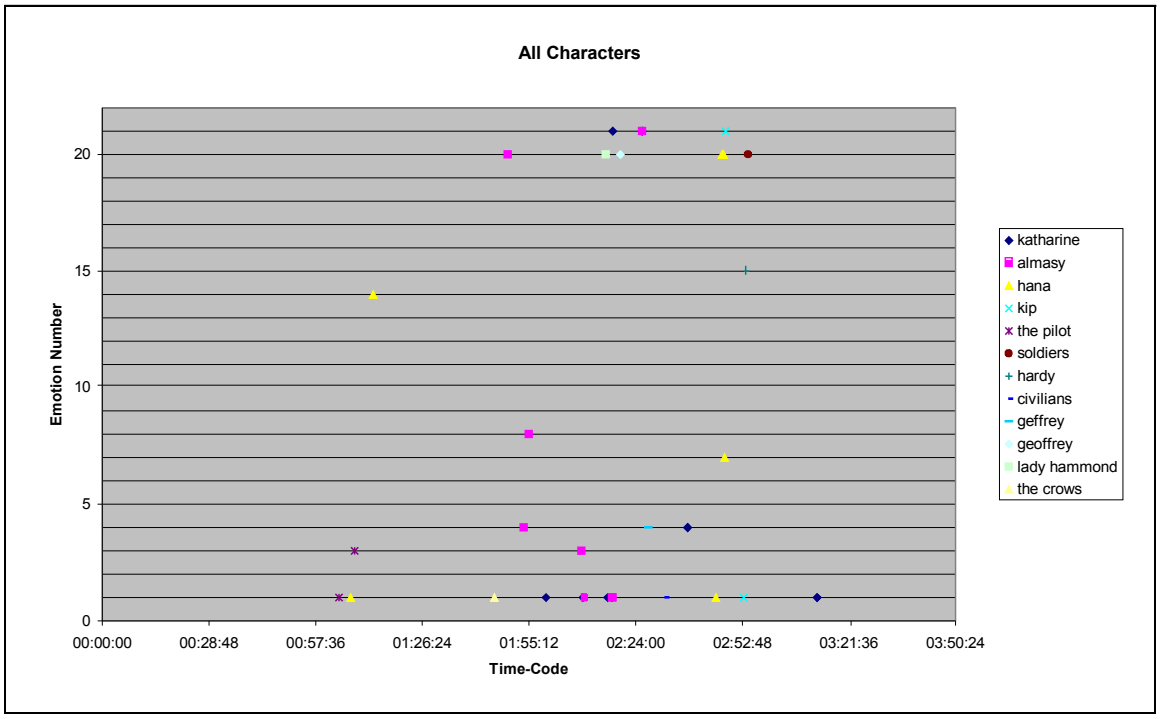

Figure 2. Output of a system that automatically associates the occurrence of an emotion with one of the film's characters. Note the 22 emotions are referred to by numbers in this display.

Further Information

This work has been reported in two conference papers.

Vassiliou, Salway and Pitt (2004), 'Formalising Stories: sequences of events and state changes', IEEE Conference on Multimedia and Expo, ICME 2004.

Salway and Graham (2003), 'Extracting Information about Emotions in Films', ACM Multimedia 2003, Berkeley, 2-8 November 2003.

The initial analysis and first implementation was carried out by Mike Graham as part of his 2003 MSc dissertation at the Department of Computing, University of Surrey. His dissertation is held by the University library. Subsequent work to assign emotions to characters, and the second implementation was carried out by Elizabeth Jones for her final year project in her BSc Computing and Information Technology at University of Surrey. Her report and software are held by the Department of Computing. Application of the method to another type of text (screenplays) and the innovation of a metric to compare two graphs for similarity is due to Andrew Vassiliou as part of his PhD research. For more information, please see his 2004 MPhil-PhD transfer report 'Representing Narrative in Multimedia Systems' – available on the TIWO website.

5.3 Integrating Information from Audio Description with Information from Other Texts related to Films

There is potential to combine information from multiple sources in order to create richer and more complete metadata for video data, e.g. for a video of a football match metadata can be extracted from the television and radio commentaries and a variety of newspaper reports. In the case of films, there is the potential to supplement information from audio description with information from other texts such as screenplays and plot summaries.

A first step in integrating information from different texts is to identify *cross-document co-reference*, i.e. fragments of different texts that refer to the same entity or event. Most previous work has concentrated on information about entities extracted from different texts of the same type, e.g. news stories. We are working on information about events in two very different text types – plot summaries (typically about 200 words long, referring to about 10 major events in a film) and audio description (typically about 5000-8000 words long, describing the on-screen action for the visually impaired). Our method is to select keywords for each event in the first text (plot summary) and do an IR-like search in the second text (audio description). Selecting and matching verbs directly is not possible, for example a 'murder' event mentioned in a plot summary is described as a sequence of smaller actions in the audio description. Selecting and matching the participants of events, and their grammatical roles, achieves about 50-60% precision and recall. Ongoing work concerns 'query expansion' of verbs and we are evaluating existing schemes for event decomposition and knowledge representation for this task. An example of output from our algorithms is shown in Figure 3.

Plot summary
A young, shell-shocked war nurse (Hana) remains behind to tend her doomed patient

Audio description
[23:54] Hana makes her patient comfortable

Audio description
[45:09] Gently she washes the tender skin on the patient's chest.

Figure 3. Automatically computed cross-document co-reference between a plot summary and the audio description for the film *The English Patient*.

For more information about this work please see Elia Tomadaki's MPhil-PhD transfer report 'Integrating Information from Collateral Media' (2003) which is available on the TIWO website. Her PhD thesis is due for completion in 2005, and a paper is being prepared for submission to Computational Linguistics.

5.4 Browsing Video by Cause-Effect Links

One motivation for generating machine-processable representations of a film's semantic content is to facilitate novel kinds of video browsing. We are developing a video browsing system based on representations of characters' affect states and goals. At any point in the film the user is shown key-frames from other scenes that are related to the current scene. The system is being evaluated in terms of how it helps users find answers to questions they have about a film, particularly of the kind 'Why did X do Y', and in terms of how they think it could be applied to DVD players in the future.

An example of our system in use is shown in Figure 4. Two Supreme Court Justices were killed in scene 4 (the first screenshot). When the user clicks "Highlight Scenes" button when they are watching scene 4, the user sees three highlighted scenes, which have related events by causal links and cross-links (the second screenshot). If the user clicks scene 2, then the events associated with: 'assassinator getting his mission' are represented by a-link and cross-link (Event 5 and Event 6).

For more information about this work please see Yan Xu's MPhil-PhD transfer report 'Retrieving and Browsing Story Elements' (2003) which is available on the TIWO website. Her PhD thesis is due for completion in 2005, and a paper is being prepared for submission to IEEE International Conference on Multimedia & Expo, ICME 2005.

The NAFI System is freely available for use by other researchers.

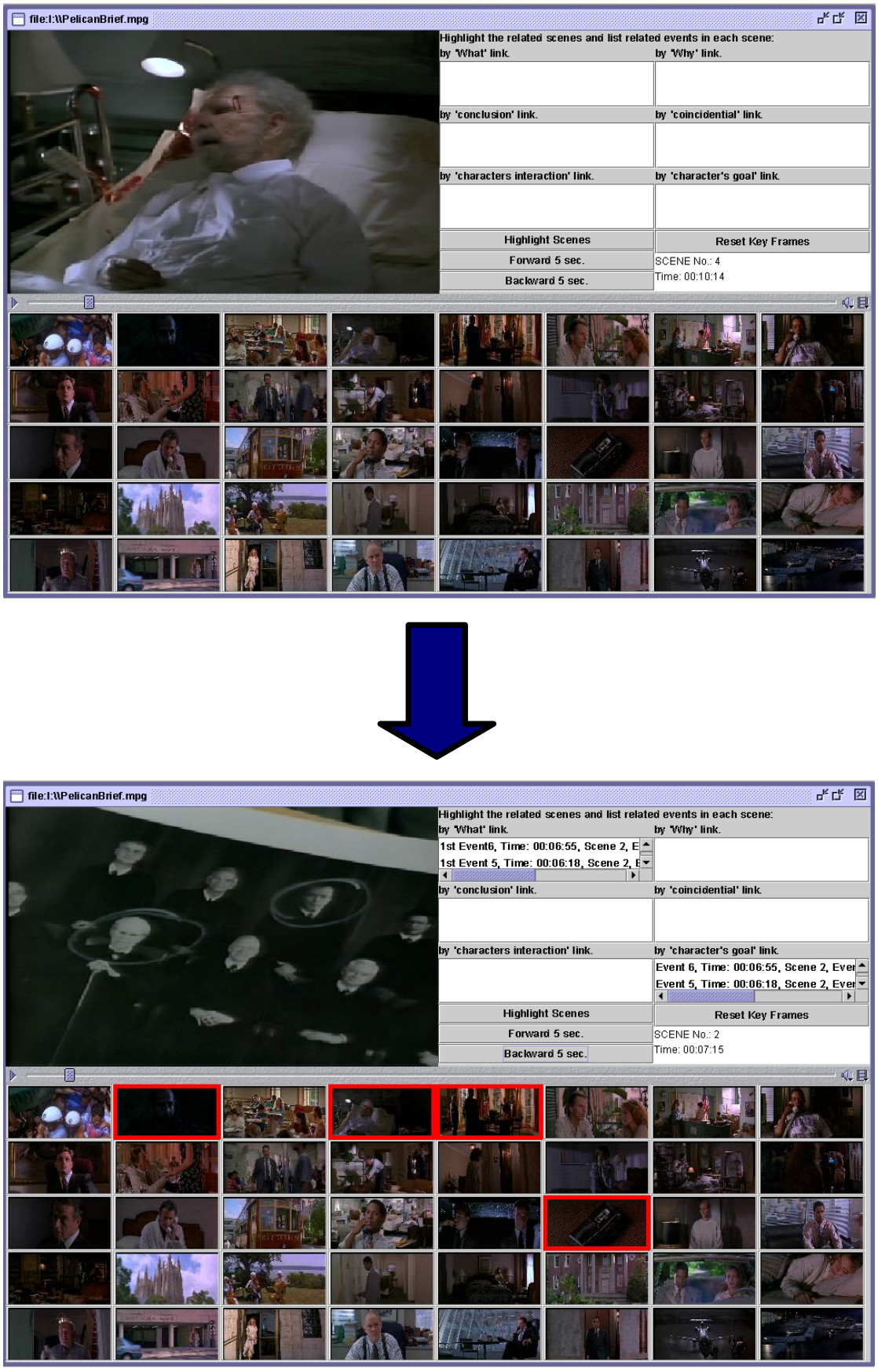

Figure 4. Browsing between two scenes in the film *The Pelican Brief* using the NAFI system.

5.5 Frequent Cues for Extracting Narrative-related Information from Audio Description?

Recent analysis of the TIWO Audio Description Corpus has highlighted a number of phrases that tend to occur frequently in the audio description and the screenplay for most films, Table 1. We believe that these phrases might be useful cues for extracting narrative-related information because they commonly occur around characters, settings, locations, motion words (verbs) and other frequent words. We believe that these phrases may help us extract information about time, the spatial arrangements of characters, characters' behaviours and goals and even plot information. We believe that these phrases can help us automatically instantiate a model for narrative.

Table 1 – Average occurrences of lexical units (phrases) in films.

Phrase	Average Occurrences in Audio Description of Films	Average Occurrences in Screenplays
"looks at"	22.4	41.6
"turns to"	10.9	15.8
"opens door"	4.6	5.8
"smiles at"	3.6	2.3

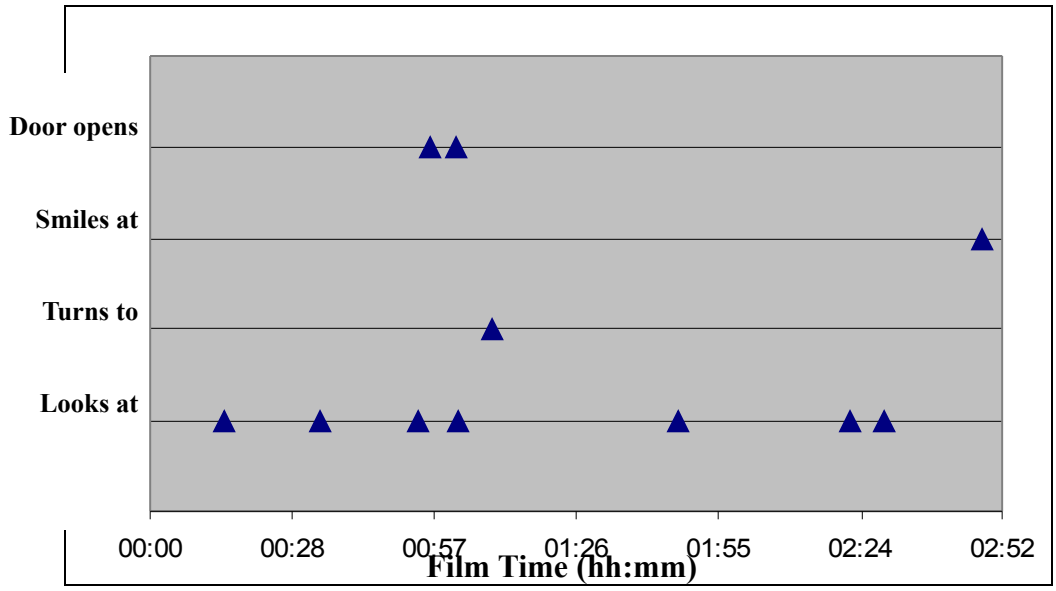

Figure 5. The occurrences of the lexical units "look at", "turns to", "smiles at" and "opens door" in the Audio Description for *The English Patient*.

For a more information about this work please see Andrew Vassiliou's MPhil-PhD transfer report 'Representing Narrative in Multimedia Systems' (2003) which is available on the TIWO website. His PhD thesis is due for completion in Autumn 2005, and a paper is being prepared for submission to IEEE Transactions on Multimedia

5.6 Evaluation from TIWO Round Table

The systems reported here have all been evaluated using standard techniques of multimedia information systems, such as Precision and Recall, and usability metrics such as time-to-complete-task. These metrics have been used to ask the questions:

- how well does the system generate representations of video content from audio description?
- how well the representations capture the meaningful aspects of video content for the users?

In the final two Round Table meetings (November 2004 and February 2005) this evaluation will be supplemented with feedback and evaluation from Round Table members. In particular they will be asked to consider the potential applicability of these systems to address some of the requirements previously elicited from the Round Table.