Andrew Salway and Eleftheria Tomadaki, “Temporal Information in Collateral Texts for Indexing Video”.

Procs. LREC 2002 Workshop on Annotation Standards for Temporal Information in Natural Language, eds. Andrea Setzer and Robert Gaizauskas, pp. 36-43.

Temporal Information in Collateral Texts for Indexing Movies

Andrew Salway and Eleftheria Tomadaki

Department of Computing
University of Surrey
Guildford, Surrey
GU2 7XH
United Kingdom
a.salway@surrey.ac.uk

Abstract

This paper suggests that video indexing is an interesting and important natural language application for which it is crucial to identify temporal information in collateral text that articulates the semantic content of moving images. Recently a rich source of information about the content of films and television programmes has become available in the form of audio description scripts. The analysis of the expression of temporal information in a corpus of audio description scripts leads to a discussion of some consequences for schemes to annotate such information in a video indexing application.

1. Introduction

The further development of digital libraries to retrieve, manipulate, browse and generate complex multimedia artefacts depends upon the machine-based representation of those artefacts, and in particular their 'semantic content'. An image can be understood at different levels of meaning: an image sequence, like a movie, can also tell a story by depicting a sequence of events. A crucial part of a film's semantic content is the narrative that it relates. As the story unfolds, the viewer constructs their understanding of the story guided by the director's careful sequencing of scenes and editing of shots. A machine-level representation of a film should maintain its rich structure and detail the entities, events and themes depicted; but how can a representation be instantiated for a given film?

One general approach to video indexing is based on the association of moving images and *collateral text* so that keywords, and potentially richer representations, are extracted from text fragments. Consider, for example, the speech of news and documentary presenters, sports commentaries and even newspaper film reviews. The challenge is to explicate the relationship between the moving image and the text. This involves dealing with temporal information in various ways; for example it is necessary to associate text fragments with the video intervals for which they are true; temporal relationships between the events depicted in the moving image must be extracted from the text; and, the time at which the action takes place must be ascertained.

A newspaper film review gives an incomplete and temporally re-ordered account of the events in a film. The speech of a newsreader is temporally aligned with the moving image but does not always refer to the visual information directly – much of a news broadcast comprises head and shoulders shots of newsreaders or stock video footage used to illustrate a story, thus keywords are more likely to be indicative of general story content than refer directly to what can be seen. By contrast, an audio description is a kind of 'narrative monologue' that gives a detailed account of what can be seen on screen in which the text order tends to follow the order of events in a programme or film. Audio description enhances the enjoyment of most kinds of films and television programmes for visually impaired viewers. In the gaps between existing speech the audio description gives key information about scenes, people's appearances and on-screen actions so that in effect the story conveyed by the moving image is retold in words.

We are interested in applying information extraction technology to generate machine-level representations of video content from audio description. It is hoped that the enhanced representation of video content could facilitate more complex querying ("find all clips showing X happening" – where X is a detailed description of events) and perhaps also contribute further to systems for video generation and maybe even question answering about what happened in a movie and why. As a first step towards information extraction we are considering the annotation of a corpus of audio description scripts to explicate what and how information is conveyed. At the moment priority is being given to temporal information because it seems to be crucial for the proper integration of moving images and collateral text.

Audio description is scripted before it is recorded. An audio description script is thus a text which is 'written to be spoken' and includes time-codes to indicate when each utterance is to be spoken. The task of processing audio description scripts is constrained because audio describers follow guidelines that restrict the language they use, i.e. normally the present tense, simple sentences and few pronominal references. This restricted language, the presence of time-codes and the relatively straightforward chronological order of the texts make audio description scripts a good starting point for extracting information for video indexing.

Though it is straightforward to associate a time-coded text fragment approximately with a video interval, a more precise association requires consideration of tense and aspect. For example, consider how the following fragments relate to intervals in the moving image: they are from audio description for *The English Patient* – time codes are in the format [minute:second][1].

[1] This sample is reproduced from *The English Patient.* Please note that further examples in the paper are fictitious but based closely on actual audio description and maintain grammatical structure (i.e. only names and events have been changed).

[11:43] Hanna passes Jan some banknotes – a near instantaneous event in the present tense, so the fragment relates to a short video interval at the time of speaking;

[11:55] Laughing, Jan falls back into her seat – the present participle indicates that 'laughing' is ongoing and so relates to a longer video interval that includes the instantaneous 'falls back';

[12:01] An explosion on the road ahead – use of nominalisation to refer to an event;

[12:08] The jeep has hit a mine – the present perfect indicates that the event is completed and the video interval that the text relates to must have start and end points before the time-code of the text fragment (general knowledge tells us that this event occurred before the explosion and was its cause).

Once text fragments have been associated with video intervals the events depicted in the steady flow of the moving image must be related to each other according to a different time-line – that of the diegetic world depicted by the movie. For example the 'hit mine' event happens immediately before the 'explosion' described above and it might be appropriate to label the relationship with causality. There are also examples of simultaneous and included events, such as – *he prevents her from leaving, holding her firmly*. Events in a movie are grouped in scenes where each scene has a (normally) unique combination of time and location. In audio description an explicit time reference might be used to introduce a new scene, e.g. *October 1944*; *later* is also used to introduce scenes and indicate story progression.

This paper suggests that video indexing is an interesting and important natural language application for which it is crucial to identify and analyse temporal information in collateral texts that articulate the semantic content of moving images. A review of video retrieval systems shows that the use of collateral text is important, but in order to extend the approach to more kinds of video material and collateral text it will be essential to process temporal information. The conceptualisation of time and events with respect to semantic content in digital video systems is outlined, particularly for films (Section 2).

We attempt to formalise the challenge of integrating video data and collateral text by describing three tasks that would contribute to the use of collateral text for video indexing. These tasks guided the analysis of an audio description script corpus (70,856 words): prominent expressions of temporal information are quantified and exemplified (Section 3). The results begin to give a basis for discussing what would be required of a scheme to annotate temporal information in this scenario: existing annotation schemes are reviewed and some tentative extensions are proposed (Section 4). The paper closes by considering further directions for this work (Section 5).

2. Digital Video Systems

Video data can be indexed with visual features based on the distribution of pixels, e.g. colour, texture, shape and motion: however a 'semantic gap' appears between video databases and users who often conceive their information needs in terms of the relationships between entities, events and themes to be depicted in the video sequence of interest. Indexing could be achieved by attaching keywords and other descriptors manually to either whole video data files or intervals and regions within them. A cheaper alternative is to use language technology to process 'collateral text'; Srihari introduced this term to refer to textual information associated with visual information, specifically photo captions (1995). Video data sometimes includes an *integral* textual component in the form of speech and closed-captions. Other *external* textual information arises in the production and distribution of video material, e.g. scripts and production notes, and now audio description (legislation in a number of countries makes the provision of audio description mandatory for an increasing amount of digital TV and film output).

In news broadcasts and documentary programmes much of the information content is carried by the spoken words of the presenters, and the subjects on which they are speaking will reflect, albeit to varying degrees, the entities, events and themes shown in the accompanying moving images. The *Informedia* system indexes news broadcasts and documentary programs by keywords that are extracted from speech and closed captions: since the speech is time-aligned with the moving images the keywords can be associated with specific video intervals (Wactlar et al., 1999). This approach has been extended into a multi-lingual context in the Pop-Eye and Olive projects, and to deal with sports footage in the current MUMIS project (de Jong et al., 2000). Other researchers have applied text segmentation techniques to the speech stream of video data in order to segment video sequences (Mani et al., 1997; Takeshita, Inoue and Tanaka, 1997). The transcribed speech of news presenters has been exploited in a system for browsing through news broadcasts by following hypertext links between terms and viewing associated video sequences (Shahraray, 1999). Research and systems focused on accessing broadcast news, including further tasks like multi-stream segmentation, combined name/face recognition and multimedia summarisation are collected in Maybury (2000).

There are moving images that do not contain 'integral' text, but that can be indexed with text that was produced specifically to elucidate the video's content. The *WebSEEK* system, which has indexed hundreds of thousands of images and videos on the WWW, selects keywords from the text of hyperlinks to images and videos on WWW-pages (Smith and Chang, 1997); note that this system only indexes whole videos rather than intervals. Another system, developed at the Japan Broadcasting Corporation, parses the notes kept in the production of wildlife documentary programs that describe the entities and events in the recorded footage and are time-coded. Queries for video intervals can be made in terms of the relationships between entities and actions (Kim and Shibata, 1996).

More recently there have been developments to combine visual and textual features for the classification of video sequences. For example, visual features may indicate the location of a scene (indoors/outdoors) and whether there is one or many people in the shot, and textual features may indicate the nature of the spoken words (a news report / a political speech): taken together these features suggest whether a video sequence depicts a political rally, an outside news broadcast, or a political party's conference (Satoh, Nakamura and Kanade, 1999).

Textual information from TV sit-com scripts has been combined with visual features, through a process involving user interaction, so a system can locate scenes containing a particular character (Wachman and Picard, 2001).

Collateral text could potentially be used for extracting information about other kinds of video, including those with rich semantic content like films and dance sequences. In specialist domains, like dance, there is an extensive range of collateral texts available (dance programmes, newspaper reviews, textbooks, choreographer's notes, biographies, etc.) and spoken commentaries can be elicited from experts asked to 'describe' and to 'interpret' sequences. The KAB system was developed to index fixed-length intervals of dance videos with keywords from such commentaries: this work also specified requirements for a general system to process diverse collateral texts (Salway and Ahmad, 1998; Salway, 1999). A key requirement is a video data model and representation scheme that captures semantic video content, including temporal organisation, at an appropriate level of detail to facilitate complex queries, browsing and even video generation; the link with collateral text also needs to be modelled.

In the video database literature semantic content is usually treated as comprising the objects and events depicted by a moving image and the spatio-temporal relationships that hold between them; for a survey see Chen, Kashyap and Ghafoor (2000). Descriptions of objects and events (as keywords and propositions, for example) are associated with intervals of video data which can be modelled either as a hierarchy of discrete intervals (Weiss, Duda and Gifford, 1995) or as multi-layered overlapping intervals (Davenport, Aguierre Smith and Pincever, 1991).

Allen's (1983) temporal logic has been applied widely in video data models to facilitate reasoning about video content and more complex queries: the number of the 13 possible interval relationships that are used varies between applications. A hierarchical model is appropriate for dealing with film in terms of scenes and shots (see Corridoni et al., 1996). However, to capture more detail about the events within a shot it might be necessary to allow for overlapping video intervals and more description of the relationships between events.

Knowledge representation schemes aim to provide unambiguous representations of meaning and to facilitate inferencing: a number of proposals have been made to use such schemes for semantic video content. Regarding the composition of events/sub-events in moving images, particularly in stereotypical situations, a framework was developed based on Schank and Abelson's scripts (1977), see (Parkes, 1989; Nack and Parkes, 1997). Semantic networks have been used in a video browsing system to elaborate the description of events, for example to specify participants and causal relationships between events (Roth, 1999). The use of conceptual dependency graphs and story grammars has also been discussed (Tanaka, Ariki and Uehara, 1999). Independent to this, but sharing some similar goals, researchers in computer vision have proposed levels to deal with complex visual information at stages from raw visual input to final representation, e.g. 'change – event – verb – history' (Nagel, 1988), and specifically for human motion 'movement – activity – action' (Bobick, 1997).

3. Temporal Information in Collateral Text

This section characterises the expression of temporal information in a corpus of audio description scripts with respect to three tasks we consider important for video indexing with collateral text. First though, in order to extend the use of collateral text to index films it is necessary to explicate how a linear text relates to a film with multi-faceted content. The discussion here is limited to film content that is conveyed visually, and hence accessible through audio description - we are not currently considering dialogue and sound effects. The focus is on films and accompanying audio description but much of what is discussed could be relevant to other kinds of video and other collateral text types.

3.1. Integrating Moving Images and Text

In order to integrate audio description text with film at a semantic level it is necessary to deal with film in terms of the shots and scenes by which it is structured. It is also important to recognise two timelines: (i) film time, i.e. the time it takes to watch the film; and, (ii) story time, i.e. the time in which the events depicted take place. Figure 1 shows how a film (stored as a video data file) can be modelled in terms of shots which are defined as continuous pieces of filming, and scenes which are characterised by each having a unique combination of location and time. The story timeline is shown in parallel with layers of events taking place. Of course the relative position of events may differ between the two timelines, e.g. the film may depict events in a different order than they happen in the story, and events that are happening at the same time but in different locations will be depicted in different scenes. For video retrieval purposes it is important to maintain temporal relationships between events; different sub-sets of Allen's relationships will be required for different applications.

The structure of film provides some useful constraints when dealing with temporal information. It is reasonable to assume that all events depicted within a scene take place close together in the story timeline, and are likely to form larger events (information about scene boundaries may be available from sources like film scripts and automatic video analysis). When considering how events are depicted at the shot level it is important to note film-making techniques that are used to convey that an event is taking place, or has taken place, without showing it in its entirety; a director may choose to portray only the end result of an event and allow the viewer to infer that the event took place.

The collateral text is shown as a series of time points that indicate the time at which the speaker starts the utterance (assuming a temporally aligned collateral text, like an audio description). The three tasks outlined next relate to the extraction of temporal information from collateral text to: (i) associate an utterance with the video interval for which it is true, be it a shot, scene or some other interval; (ii) specify event-event relationships – here we only consider relationships holding within a scene (in film time); (iii) establish the time at which scene is set (in story time).

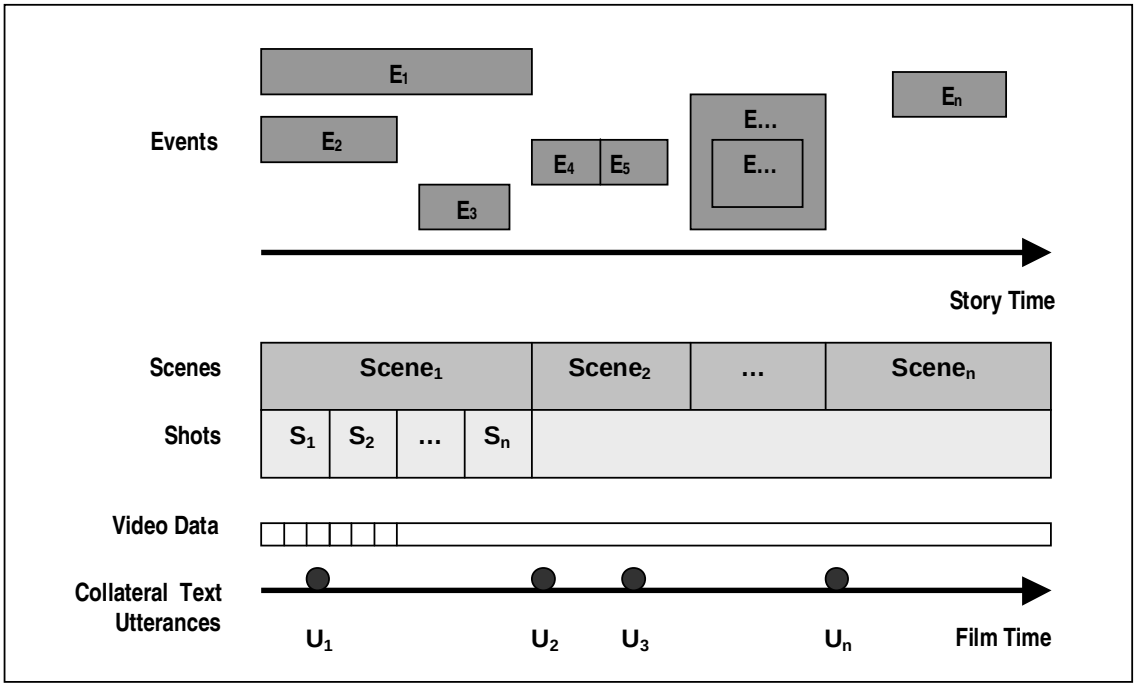

Figure 1. The organisation of a film's content in terms of shots and scenes (which relate to film time) and the events that comprise the semantic video content (which relate to story time). Collateral text such as audio description is temporally aligned with the video data in film time.

3.1.1. Task 1: Associate an audio description fragment with the interval in film time for which it is true.

Given a time coded text fragment it is relatively straightforward to associate it approximately with the video interval for which it is true, i.e. the interval in which the event it refers to is taking place; the time-code plus and minus an arbitrary number of seconds works as a crude approximation of start and end times. However, it is desirable to be more precise about at least one of: start time, end time or duration. (A greater challenge, not addressed here, is to ascertain whether the event is depicted on-screen throughout the duration). As well as events, it is also appropriate to deal with states if they change significantly during the movie, e.g. to indicate scenes in which a character is a child or grown-up.

The problem can be gauged to some extent by considering an earlier feasibility study for indexing moving images with audio description (Turner, 1998). A small sample of video material with accompanying audio description was analysed (including a film and various kinds of television programme). Results showed that overall about 50% of shots were described but only about 40% of the audio description referred directly to the shot on-screen at the time of speaking.

To ascertain appropriate video intervals it may help to consider some of the aspectual features of events classified by Comrie (1976). Whether an event has internal structure (*punctual* / *durative*) gives some information about its duration; this may be an inherent characteristic of a verb but may be modified grammatically, e.g. with the progressive. Knowing about an event's end result, if it has one (*telic* / *atelic*), gives information about its completion (and in audio description may be all that is referred to).

3.1.2. Task 2: Event-event relationships in story time (within the same scene).

Moving images can depict many events at the same time, and in the case of film the temporal organisation of events and relationships such as event / sub-event and causality are crucial to a viewer's understanding. As discussed previously some or all of Allen's 13 temporal relationships might be needed, though whether they can all be extracted from collateral text remains to be seen. In a narrative dialogue by default events are mentioned in the order in which they occur – however, events may occur simultaneously, or there may be stylistic reasons to mention them out of order.

Ascertaining basic temporal relationships, like before / after / overlapping, may be possible just from the collateral text. However, to construct rich representations of composite events within scenes perhaps relies more on prior 'world knowledge' than it does on information immediately available in a narrative monologue (cf. the use of Schank's scripts to deal with semantic video content); the problem becomes harder still when event-event relationships across different scenes are considered. A lexical resource, like WordNet, might help as a first step to relate events, in light of the entailment relations for verbs described by Fellbaum (1998). When considering temporal inclusion some sets of verbs are *co-extensive*, e.g. 'march and walk', 'whisper and talk'; whereas other share a relationship of *proper inclusion*, e.g. 'sleep and snore'. Having access to these relationships may help to

associate descriptions of the same event and to establish sub-event relationships. Other relationships allow events to be associated according to *backward presupposition*, e.g. 'forget and know', and on the basis of *causality*, e.g. 'show and see'.

3.1.3. Task 3: Establish the time a scene takes place (in story time).

A viewer's appreciation of a film requires knowing when it is set, and if it is set over a long time period then the time of each scene must be known – thus information needs to be extracted to give each scene a time. Unless a film is based upon true-life events then it is normally set within a time period without specific dates being implied. Similarly, within the course of a day in story time exact times are usually less important than whether the viewer knows it is morning, afternoon, evening or night (of course exact times will be crucial for some plots). Unless otherwise indicated the assumption is that scenes are ordered sequentially according to the story timeline, but for some movies the use of flashback will have to be dealt with.

3.2. Temporal Information in Collateral Text: a case study with audio description scripts

The intention of this case study is to quantify and exemplify prominent expressions of temporal information in audio description scripts: the analysis is organised around the three tasks for video indexing described in the previous section. The corpus comprises audio description scripts for 12 movies, covering a range of movie genres, and written by six different describers. It currently totals 70,856 words – this will be expanded to around 500,000 words in coming months.

When carrying out the analysis we considered the variety of ways temporal information can be expressed in English as outlined by Quirk et al. (1985), i.e. by using tense, aspect, adverbials, prepositional phrases, subordinate clauses, nouns and proper nouns. Of course in narrative monologues text order is highly informative about the order in which events take place. Our 'conceptualisation of time' is guided by approaches to video data modelling, i.e. the association of event and state descriptions with video intervals, the specification of interval relationships following Allen (1983), and the organisation of complex events using knowledge representation schemes. Theoretical perspectives on events, such as Comrie's classification of aspect (1976) and Fellbaum's entailment relations (1998) were also considered.

Based on the 50 most frequent verbs in the corpus it appears that the majority of events are material processes (84%), with some mental processes (10%), a few relational processes (4%) and a few behavioural processes (2%), following Halliday (1994).

3.2.1. Information to Associate Text Fragments with Video Intervals

The present tense proliferates in the audio description corpus. It is even used to describe events that are about to happen, for example to describe speech acts which cannot be described at the time they occur – *the doctor questions Tom*. The occasional use of the present perfect is important to describe events after they happened (possibly because there was not an opportunity to describe them at the time they occurred, or because only the end result is depicted on screen) – *the cake has been eaten*. Past events are also sometimes referred to in relative clauses used to identify unnamed characters – *the woman who visited Paul is walking down the street*. In order to be more precise about the start, end or durations of events it seems that a variety of aspectual information is important, especially aspectual verbs and the inherent aspect of verbs.

In the audio description corpus the verbs *start, stop, begin* and *finish* occur relatively far more frequently than they do in the general language British National Corpus (BNC), Table 1 shows just 3rd person singular forms. These verbs almost always appeared in the present tense to refer to another event so it would be straightforward to use them to compute their arguments' start and end times.

Verb	Abs. Freq	Ratio with BNC
stops	105	65.79
starts	60	25.13
begins	19	4.67
finishes	3	25.65

Table 1: Showing prominent aspectual verbs in Surrey's audio description corpus (only for 3rd person singular). The third column is calculated by dividing relative frequency in the audio description corpus with relative frequency in the British National Corpus (BNC)

Regarding the duration of events the adverb *still* is frequently used with durative events that have not finished at the time of speaking (85 occurrences of *still* in the corpus; 62 of these are in the time sense). There was little sign of time expressions giving information about exact durations but relatively short periods were frequently indicated with *for a moment* (29 occurrences). Other frequent, adverbs like *slowly* (111 occurrences) and *quickly* (20) might make a small contribution to understanding the duration of an event.

The grammatical marking of progressive aspect does not appear to be significant for the task of associating text fragments with video intervals. In a narrative monologue we learn nothing about the duration of the event from the distinction between 'he runs' and 'he is running'; the simple present and the progressive are used interchangeably in the corpus. In fact it is probably an event's inherent aspect that is most important to determine, at least approximately, its duration in film time. In general language this will be problematic given that the many senses of common verbs often have different aspectual characteristics, however in specialist domains it may be possible to store default durations for events, like dance movements.

3.2.2. Information to Specify Event-Event Relationships in Film Time

The most frequent conjunction to indicate events happening at the same time in the audio description corpus was *as* (350 occurrences in the time sense) – *the children*

play as the crowd moves away; sometimes *as* indicates more of a connection between events – *she continues to hide as the monster approaches*. Used only to indicate simultaneity (without implying further connection between events) *while* occurred 37 times. Both these conjunctions indicate some degree of overlap between events but further information is required to know whether the events are strictly simultaneous, whether one is included within the other, or if they simply overlap. Non-finite verbs with sub-ordinate clauses tend to indicate that the second event is included in the first – *Coughing, Mary gives the medicine to Tom.* When linking events *and* was 'ambiguous' as to whether the events occurred serially or in parallel.

Occurring only in its time sense, *then* (173 occurrences) was still relatively more frequent in the audio description corpus than the BNC. Though it is redundant as far as indicating sequence is concerned (that is already conveyed by text order) it does seem to imply the completion of the first event before the start of the next one – *Sarah chops the tomatoes then fries an egg*. Furthermore in many examples it suggests that the events meet in time, i.e. the endpoint of one equals the start of the other. (This kind of information could be useful in relation to our Task 1). The less frequent *when* (29 occurrences) and *until* (20) were used respectively to indicate the start and end of events in relation to other events or states, often suggesting that the first event led to the second – *she walks through the forest until she finds the house*.

Like *then*, *now* (40 occurrences) adds little or nothing by way of basic temporal information in these narrative monologues, however it does seem to indicate a change or contrast between two events across a passage of audio description – *Jane is dancing with George … Now she is dancing with her cousin*. Perhaps surprisingly, *after* and *before* occur relatively infrequently in the corpus (compared with the BNC) and when they are used they only serve to emphasise the sequence of events already conveyed by text order, i.e. we find 'after E1, E2' but not 'E2 after E1', and 'E1 before E2' but not 'before E2, E1'.

The adverb *again* is prominent in the corpus (141 occurrences – 2.5 times relatively more often than in the BNC). It generally indicates that an event is happening for a second time within a scene – for video retrieval purposes it might be useful to relate the two instances.

3.2.3. Information to Specify When Scenes Take Place in Story Time

The most frequent time expressions used to locate a scene on the story timeline relate to non-specific times of day: *night* (37 occurrences), *morning* (19), *evening* (11), *dusk* (6), *dawn* (3). Less frequent were expressions for non-specific times of year, i.e. months (without years), seasons and festival days (17 occurrences in total). This probably reflects the fact that the progression of time during a film is more often at the granularity of days. The relative paucity of specific times and dates (there were only a few examples) is explained in part by the fact that for many films the viewer need only understand a general time period. This may be conveyed by costumes, props and, for times of day, lighting: these will all be referred to by audio description.

Scenes are sometimes introduced with one of the time expressions mentioned above – indicating a change of time is a shortcut to indicate a new scene. Quite often *later* (32 occurrences) is used for this purpose and as such may be a useful cue for scene segmentation.

4. An Annotation Scheme for Temporal Information in Collateral Text?

Based on the preceding analysis this section discusses some tentative requirements of an annotation scheme for temporal information in collateral text: such a scheme would be a step in applying information extraction technology to the task of video indexing. The extent to which existing schemes would cater for these requirements is reviewed. A number of factors suggest that some extensions to existing schemes will be required: (i) there seems to be a need to maintain two timelines; (ii) if practical in terms of time and inter-annotator agreement, it would be desirable to record aspectual information regarding the internal structure of events and end-states; (iii) also subject to practicality, it is important for film to specify sub-event and causal relationships.

A canonicalized representation of times was proposed as part of a set of guidelines for annotating temporal expressions by Mani et al. (2001), who targeted a variety of text genres such as both print and broadcast news, and meeting scheduling dialogues. The emphasis of their approach was on detailing different classes of time expressions like points in time (when), durations (how long) and frequencies (how often) and handling context-dependent expressions. It also addressed fuzzy temporal boundaries that arise from the use of phrases that refer to times of year and times of the day, e.g. *summer* and *morning*, and addressed non-specific times, such as *a sunny day in April* (not a specific day, nor a specific year).

Of the time expressions dealt with it is points in time that seem to apply most directly in our scenario in order to locate events (at the granularity of scenes) on the story timeline (our Task 3). Though duration and frequency relate to the kinds of aspectual characteristics that we would like to describe for events, they annotate only words and phrases that express this information directly; though there were some frequent phrases in the audio description corpus for which it might be applicable – *for a moment*.

Another scheme that has been proposed is more concerned with associating temporal information with events, and annotating the temporal relationships between events (Setzer and Gaizauskas, 2000; Setzer, 2001); this scheme was developed initially for newswire texts but is extensible. Annotations are attached to the heads of finite verb groups as representatives of events, as well as to temporal expressions. It is possible to specify the type of event (Occurrence / Perception / Reporting / Aspectual) as well as the tense and grammatical aspect of the verb. The annotations have attributes to specify five event-event and event-time relationships: 'before', 'after', 'includes', 'included' and 'simultaneous'. The features of the scheme that have been summarised here are exemplified in Appendix A which shows the annotation of 9 utterances of audio description.

The annotation of event-event relationships within a scene (our Task 2) would be dealt with quite

comprehensively by Setzer's and Gaizauskas' scheme: although as many as 13 temporal relationships (from Allen) are discussed in the video retrieval literature the five used in this scheme would probably serve most purposes. Being able to annotate aspectual events, i.e. to indicate the start and end time of occurrence events, is certainly important given their frequency in the audio description corpus - cf. our Task 1. For other parts of Task 1 it might be necessary to extend the scheme to specify the start and end of events when there is no explicit time expression, or to do it relative to the time-code in the text; a further minor extension would be to allow for the annotation of states as well as events. It certainly would be desirable to be able to specify causal and sub-event relationships between events as these are crucial to the narrative structure of movies, however this would depend upon annotators' ability to apply them consistently.

5. Closing Remarks

Dealing with temporal information is an important first step towards generating machine-level representations of video content from collateral text, especially when dealing with a complex multimedia artefact, like film, and richly informative collateral text, like audio description. This work is in its early stages but the three tasks outlined here begin to give us a handle on some of the challenges involved in integrating moving images and narrative monologues. The corpus analysis showed an extensive range of temporal information that needs to be dealt with in respect to these tasks. Progress will be made by more extensive application of existing annotation schemes leading to decisions about exactly what is required by way of extensions. Such decisions need to be informed by considerations of any new scheme's practicality (is it simple enough to be applied consistently and quickly) and the extent to which it captures important information (the criteria for which will vary between video applications and users). The final test would perhaps be a comparison of video retrieval using: (i) unannotated audio description (i.e. relying on time codes and text order alone); (ii) annotated audio description (with no further processing); and, (iii) machine-based representations generated from annotated audio description.

6. Acknowledgements

This research was carried out as part of the Television in Words project (TIWO) supported by an Engineering and Physical Sciences Research Council (EPSRC) grant, GR/R67194/01. The authors would like to thank the members of the TIWO Round Table for sharing their knowledge of audio description and providing samples for our corpus. Finally, we are very grateful for the comments of two anonymous reviewers and have tried to take heed.

7. References

Allen, J.F., 1983. Maintaining Knowledge About Temporal Intervals. *Communications of the ACM* 26 (11):832-843.

Bobick, Aaron F., 1997. Movement, Activity, and Action: The Role of Knowledge in the Perception of Motion. *Philosophical Transactions of the Royal Society of London Series B – Biological Sciences* 352 (1358):1257-1265.

Chen, S.-C., R.L. Kashyap, and A. Ghafoor, 2000. *Semantic models for Multimedia Database Searching and Browsing*. Kluwer Academic Publishers.

Comrie, B., 1976. *Aspect: an introduction to the study of verbal aspect and related problems*. Cambridge University Press.

Corridoni, J.M., A. Del Bimbo, D. Lucarella, and H. Wenxue, 1996. Multi-perspective Navigation of Movies. *Journal of Visual Languages and Computing*, 7:445-466.

Davenport, G., T. Aguierre Smith, and N. Pincever, 1991. Cinematic Primitives for Multimedia. *IEEE Computer Graphics and Applications* July:67-74.

de Jong, F., J.-L. Gauvain, D. Hiemstra, and K. Netter, 2000. Language-Based Multimedia Information Retrieval. *Proceedings RIAO 2000 Content-Based Multimedia Information Access, Paris, April 2000*, 713-722.

Fellbaum, C., 1998. A Semantic Network of English Verbs. In Fellbaum (eds.), *WordNet: an electronic lexical database*. Cambridge MA: The MIT Press.

Halliday, M. A. K., 1994. *An Introduction to Functional Grammar*. London: Edward Arnold, 2nd edition.

Kim Y.-B., and M. Shibata, 1996. Content-Based Video Indexing and Retrieval – A Natural Language Approach. *IEICE Transactions on Information and Systems* E79-D(6):695-705.

Mani I., D. House, M.T. Maybury, and M. Green, 1997. Towards Content-Based Browsing of Broadcast News Video. In M. Maybury (ed.), *Intelligent Multimedia Information Retrieval*. Menlo Park CA / Cambridge MA: AAAI Press / MIT Press.

Mani, I., G. Wilson, L. Ferro, and B. Sundheim, 2001. Guidelines for Annotating Temporal Information. *Procs. HLT 2001, First International Conference on Human Language Technology Research*, San Francisco: Morgan Kaufmann.

Maybury, Mark, 2000 (ed.). Special Issue – 'News on Demand'. *Communications of the ACM*, 43(2).

Nack, F., and A. Parkes, 1997. Toward the Automated Editing of Theme-Oriented Video Sequences. *Applied Artificial Intelligence*, 11:331-366.

Nagel, H.-H., 1988. From image sequences towards conceptual descriptions. *Image and Vision Computing*, 6(2):59-74.

Parkes, A. P., 1989. The Prototype CLORIS System: Describing, Retrieving and Discussing Videodisc Stills and Sequences. *Info. Proc. and Management*, 25(2):171-186.

Quirk, R., S. Greenbaum, G. Leech, and J. Svartvik, 1985. *A Comprehensive Grammar of the English Language*. London and New York: Longman.

Roth, V., 1999. Content-based retrieval from digital video. *Image and Vision Computing*, 17:531-540.

Salway, A., 1999. *Video Annotation: the role of specialist text*. Ph.D. thesis, University of Surrey.

Salway, A., and K. Ahmad, 1998. Talking Pictures: Indexing and Representing Video with Collateral Texts. *Procs. 14th Workshop on Language Technology* -

Language Technology for Multimedia Information Retrieval, 85-94.
Satoh, S., Y. Nakamura, and T. Kanade, 1999. Name-it: Naming and detecting faces in news videos. *IEEE Multimedia*, 6 (1):22-35.
Schank, R. C. and R. P. Abelson, 1977. *Scripts, Plans, Goals and Understanding: an inquiry into human knowledge structures*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Setzer, A., 2001. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*. Ph.D. thesis, University of Sheffield.
Setzer, A., and R. Gaizauskas, 2000. Annotating Events and Temporal Information in Newswire Texts. *Procs. LREC 2000, 2nd Int. Conf. On Language Resources and Evaluation*, 1287-1293.
Shahraray, B., 1999. Multimedia Information Retrieval Using Pictorial Transcripts. In B. Furht (ed.), *Handbook of Multimedia Computing*. Florida: CRC Press.
Smith J.R., and S.-F. Chang, 1997. Visually Searching the Web for Content. *IEEE Multimedia* July-Sept:12-20.
Srihari, R.K., 1995. Computational Models for Integrating Linguistic and Visual Information: A Survey. *Artificial Intelligence Review*, 8(5-6):349-369.
Takeshita, A., T. Inoue, and K. Tanaka, 1997. Topic-based Multimedia Structuring. In M. Maybury (ed.), *Intelligent Multimedia Information Retrieval*. Menlo Park CA / Cambridge MA: AAAI Press / MIT Press.
Tanaka, K., Y. Ariki, and K. Uehara, 1999. Organization and Retrieval of Video Data. *IEICE Trans. on Information and Systems*, E82-D(1):34-44.
Turner, J.M., 1998. Some Characteristics of Audio Description and the Corresponding Moving Image. *ASIS Annual Meeting*, 35:108-117.
Wachman, J. S., and R.W. Picard, 2001. Tools for Browsing a TV Situation Comedy Based on Content Specific Attributes. *Multimedia Tools and Applications*, 13(3):255-284.
Wactlar, H.D., M.G. Christel, and Y. Gong, and A.G. Hauptmann, 1999. Lessons Learned from Building a Terabyte Digital Video Library. *Computer*, Feb:66-73.
Weiss, R., A. Duda, and D.K. Gifford, 1995. Composition and Search with a Video Algebra. *IEEE Multimedia*, Spring 1995:12-25.

Appendix A

Annotation of an audio description script following Setzer's scheme

The following passage of audio description (from *The English Patient*) has been annotated following the scheme and guidelines given by Setzer (2001). The sample here exemplifies how: (i) tense and aspect features can be associated with an event's verb; (ii) how the class of an event can be noted; (iii) how relationships between events can be specified. The sequence of events inherent in the text order has not been annotated, though it could have been – only exceptions to the 'default' have been annotated, e.g. simultaneous events, and events that are mentioned in a different order to which they occur.

[11.43] Hanna <event eid=1 tense=present class=occurrence> passes </event> Jan some banknotes.
[11.55] <event eid=2 tense=present class=occurrence aspect=progressive relatedToEvent=3 eventRelType=includes> Laughing </event>, Jan <event eid=3 tense=present class=occurrence relatedToEvent=4 eventRelType=simultaneous signalID=1> falls </event> back into her seat <signal sid=1> as </signal> the jeep <event eid=4 tense=present class=occurrence> overtakes </event> the line of the lorries.
[12.01] An <event eid=5 tense=present class=occurrence relatedToEvent=6 eventRelType=after> explosion </event> on the road ahead.
[12.08] The jeep has <event eid=6 tense=present class=occurrence aspect=perfective > hit </event> a mine.
[12.09] Hanna <event eid=7 tense=present class=occurrence> jumps </event> from the lorry.
[12.20] Desperately she <event eid=8 tense=present class=occurrence> runs </event> towards the mangled jeep.
[12.27] Soldiers <event eid=9 tense=present class=occurrence> try </event> to stop her.
[12.31] She <event eid=10 tense=present class=occurrence> struggles </event> with the soldier who <event eid=11 tense=present class=occurrence> grabs </event> hold of her firmly.
[12.35] He <event eid=12 tense=present class=occurrence> lifts </event> her bodily from the ground <event eid=13 tense=present class=occurrence aspect=progressive relatedToEvent=12 eventRelType=simultaneous> holding </event> her tightly in his arms.