Associating Characters with Events in Films

Andrew Salway
Burton Bradstock Research Labs
Dorset, DT6 4QR
United Kingdom
andrew@bbrel.co.uk

Bart Lehane
Centre For Digital Video Processing
Dublin City University
Ireland
lehaneb@eeng.dcu.ie

Noel E. O'Connor
The Adaptive Information Cluster
Dublin City University
Ireland
oconnorn@eeng.dcu.ie

ABSTRACT

The work presented here combines the analysis of a film's audiovisual features with the analysis of an accompanying audio description. Specifically, we describe a technique for semantic-based indexing of feature films that associates character names with meaningful events. The technique fuses the results of event detection based on audiovisual features with the inferred on-screen presence of characters, based on an analysis of an audio description script. In an evaluation with 215 events from 11 films, the technique performed the character detection task with $Precision = 93\%$ and $Recall = 71\%$. We then go on to show how novel access modes to film content are enabled by our analysis. The specific examples illustrated include video retrieval via a combination of event-type and character name and our first steps towards visualization of narrative and character interplay based on characters occurrence and co-occurrence in events.

Keywords

Film, Video Retrieval, Video Visualization, Audio Description.

1. INTRODUCTION

As the size and availability of digital film archives increase, so too does the need for tools that assist those who need to access them. This could include film fans in the general public, students and scholars of the medium or editors and other broadcast/film professionals. Given the significant volume and data-rich nature of the content to be accessed, the efficacy of such tools critically depends on their ability to perform automated semantic analysis and description of film content, be it for video retrieval, browsing, summarization or visualization [1]. In an ideal world, an appropriate indexing of film content (as espoused in [2]), would mirror that of a book whereby access is facilitated via either a linear *Table of Contents* or a non-linear *Index* listing the occurrences of important semantic concepts. Much research to date has focused on the former, as evidenced by the array of works on shot-level and scene-level analysis of content in recent years [3, 4, 5, 6]. For the latter, it is not fully understood what constitutes such an *Index* for film content, but we know it would typically include the film's characters and the events in which they are involved, perhaps along with the range of cinematic techniques used by directors to convey a story. In this paper, we discuss our initial work towards creating a non-linear Index of important objects (in this case characters) and events

The significant challenges posed by semantic analysis of video content, usually referred to as the 'semantic gap' [7], requires the fusion of several multimodal information streams. A meaningful and useful description of 'who, what, where, when and why' as depicted in a film, typically cannot be generated by video analysis alone. Rather, multiple additional sources to the moving image itself, such as the soundtrack and the various kinds of text that are associated with a film, should also be considered in any analysis.

For the first time, the work presented here combines the analysis of a film's audiovisual features with the analysis of an accompanying *audio description*, in order to automatically produce a semantic description of a film in terms of the characters and events depicted. We choose this text source to complement our audiovisual analysis, since compared with a film script, an audio description gives a more accurate account of who and what can be seen on-screen. Moreover, this account is tightly time-aligned with the film. Our analysis of audiovisual features leads to structuring the film into three kinds of event: *Dialogue*, *Exciting* and *Musical*. Our analysis of audio description identifies the on-screen presence of the film's main characters who can then be associated with these detected events. The resulting description of a film in terms of events and associated characters enables innovative modes of information access to the film content.

Section 2 provides the necessary background for our work. It motivates and describes the events that we consider important and provides more information on audio description in general. It also reviews the work that we consider most related to ours in the current state of the art in the analysis of film content. Our method for associating characters and events is described in detail in section 3 and evaluated in section 4. Section 5 then describes novel kinds of access to film content that are enabled by our analysis – video retrieval via a combination of event-type and character name and

our first steps towards visualization of plot and character interplay based on character occurrence and co-occurrence within events. Finally, we draw some conclusions and outline directions for future research in section 6.

2. BACKGROUND

2.1 Event-Based Indexing

The basic atomic unit of a film is generally considered to be the shot and the task of video segmentation at this level is now considered straightforward. However, a shot-level segmentation is not always an appropriate access mechanism for a user. For example, a single shot of a car chase carries little meaning when viewed independently, it may not even be possible to deduce that a car chase is taking place from a single shot. However, when viewed in the context of the surrounding shots it forms a meaningful *event* that the viewer can recognise. In our work, we consider an event to be an interval in the film which viewers recognise and remember as a semantic unit. A conversation between a group of characters, for example, would be remembered as a semantic unit ahead of a single shot of a person talking in the conversation. Similarly, a car chase would be remembered as 'a car chase', not as 50 single shots of moving cars.

At a higher level of abstraction, a film is made up of scenes and various techniques for scene-based segmentation based on the analysis of audiovisual features have been proposed [4, 5, 6, 8, 9, 10]. However, we argue that a scene-based representation fails to achieve the granularity required for most users. A scene may contain a conversation, followed by a fist/gun fight, which would be recognised and remembered as two distinct events by the audience and so should be modelled as such. Furthermore, it is often problematic for human observers to recognise scene boundaries, especially in a medium as diverse and creative as films where occurrences of dramatic devices such as flashbacks / flash-forwards, dream sequences, cross-cutting etc. are common. Our previous studies [11] indicate, on the other hand, that users can strongly relate to an event-based structure imposed on a film when applied in a retrieval scenario.

We work with three event types of major events – Dialogue, Exciting and Musical events – detected using the techniques presented in [12] and reviewed briefly in section 3.1. These three types of event typically account for $> 90\%$ of a film, whilst being intuitive for a user to understand. Dialogue constitutes a significant part of any film, and the viewer usually obtains much of the information about the plot, story, background etc. of the film from the dialogue. Dialogue events are not necessarily constrained to a set number of characters so a conversation between any number of characters is classified as a dialogue event in our approach. Exciting events typically occur less frequently than Dialogue events, but are central to many films. Examples of exciting events include fights, car chases and battles. The Musical event type is distinguished by the presence of music, and often the absence of dialogue, and includes montage sequences, emotional scenes (such as somebody crying), and diagetic music, i.e. music being played within the film. Previous work on event-based indexing has typically concentrated only on the detection of dialogue events [13][14, 15] or dialogue and a restricted set of specific kinds of action events [16].

2.2 Audio Description

There are many text sources that provide information about a film's semantic content. Film scripts give information about on-screen action, but are not time-coded, and tend to be wordy and use colloquial language, which can make them difficult to analyse. Furthermore, they do not necessarily match well to the final edit of the film (except for post production scripts). A higher-level account of a film's story can be found in a plot summary, which concentrates on the characters' goals and some of the major film events, although not normally on the ending. Subtitles or closed captions are an important text resource for semantic analysis. However, unlike news and sports video, in the case of film, the relationship between what is transcribed in subtitles and what can be seen on-screen is often disjunct. Subtitles exist so that the deaf and hard of hearing can read what they cannot hear: an audio description, on the other hand, allows blind and visually impaired film and TV audiences to hear a description of what they cannot see. In the interval between existing dialogue, a describer relates essential details of the on-screen action via an additional soundtrack that is delivered via a radio link to headphones in cinemas and also via DVD and digital television broadcasts.

In the UK, over 200 cinemas provide audio description and it is available for most major new films. A professional audio describer must ensure that blind and visually-impaired audiences have enough information about on-screen action in order to follow the story being told, yet this must be sufficiently succinct to fit between the existing dialogue. An audio description is scripted before it is recorded and includes precise timecodes to indicate at which point of the film a particular line of audio description is to be spoken. An example of an excerpt from an audio description is presented in figure 1. It should be clear from this example that the audio description is informative about many aspects of the on-screen action: who is present, what they look like, what they are wearing, their facial expressions, and of course, what they are doing. Clearly, audio description can be an important resource, not only for the purpose for which it was originally intended, but also for video indexing.

Evidence that audio description uses a simplified language, making it amenable to information extraction, is presented in [17]; information about characters' emotions was extracted in [18]. A set of commonly occurring actions in audio description and film scripts was identified in [19], and a set of common dialogue events in subtitles was identified in [20]. Film scripts and plot summaries are contrasted with audio description in [21] and [22], respectively. In the work reported here, the extraction of information from an audio description is restricted to data about characters' on-screen presence because this can be achieved with a very high-level of *Precision*.

2.3 Related Work

Previous work [23, 24] has shown the benefit of analysing the occurrence of characters on-screen. It was shown how the analysis of a character's appearance on-screen and subsequent disappearance gives a rhythm that can be used for topic segmentation and film classification, and for mining other semantic structures. However, in that work the annotation of character presence was done manually. The au-

00:45:25 Danny looks up and follows Russ across the warehouse.

00:45:44 They hover in the doorway. Danny paces.

00:46:03 Incredulous Russ looks away. He runs his fingers over his mouth.

00:46:23 They stare steadily at each other. A smile creeps across Danny's face. Russ turns away, folding his arms.

00:46:32 The two men nod gently at each other.

00:46:37 In the Bellagio Museum, Tess, wearing an oriental style suit, with a high collar stands serenely, gazing at a picture. The painting, in a rectangle frame, is Woman with Guitar by Pablo Picasso. Greys, browns and blues mingle in a cubist style. A short bald man chatters away beside Tess. A taller man joins them. They turn their heads as Benedict saunters in. Tess introduces him to the taller man, then Benedict, hands behind his back, studies the painting.

Figure 1: An excerpt from the audio description script of the film Ocean's Eleven (Dir. Steven Soderbergh). Note here the timecodes are in HH:MM:SS format

tomation of such annotation is a key contribution of our work. Other complementary information about semantic film content has been obtained by the analysis of audiovisual features alone, for example, models of affective content [25, 26] and information about key points in a film's story coinciding with changes in its tempo - a function of motion within the frame and shot length [27]. The task of associating names with faces detected in frames of television programmes and films was recently addressed with promising results in [28]. Their technique aligns scripts (containing information about who is saying what) with subtitles (containing information about what is being said when) in order to cue a face detector as to which characters were likely to be seen talking in which frames. This technique was limited to frontal faces and of course relies on a character to be speaking in order to be detected. This technique will not work for characters not facing the camera, nor for scenes without dialogue. We believe that the analysis of audio description will give a much more complete set of character occurrences than scripts and subtitles, though there will of course be cases when these can provide complementary information.

3. PROPOSED APPROACH

As illustrated in figure 2, our technique follows three steps: (i) Event detection; (ii) Character detection; (iii) Matching characters to events. In fact, we believe that potentially an iterative approach, as indicated by the dashed line, is warranted. That is, it is feasible that rather than a straightforward matching process, the results of event detection could inform the analysis of the audio description and vice versa. However, this is not considered in the work reported here, but targeted for future work. In this section, we describe the three steps in detail.

3.1 Event Detection

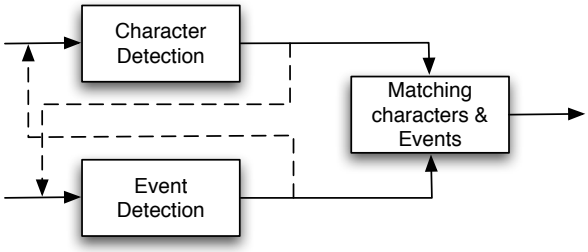


Figure 2: Overall Approach

The event segmentation and classification technique used here is described in detail and evaluated in [29, 30, 12]. It exploits a number of observations about film creation principles. For example, when filming a dialogue event, the director ensures that the audience can clearly interpret the words being spoken. Thus, a relaxed filming style is used with a low amount of camera movement, few long shots, high amounts of shot repetition and clearly audible speech [31]. Conversely, when shooting an exciting part of a film, the director aims to stimulate the audience. The common techniques for creating excitement involve fast-paced editing, combined with high amounts of movement and widely varying camera positioning [32]. Finally, musical events are shot with a strong musical soundtrack, usually combined with slower paced editing and low amounts of camera movement [31, 11].

We extract a set of audiovisual features to detect the presence of these film-making principles. Firstly, *shot boundaries* are detected using a standard colour histogram technique. Once the shot boundaries are known, the *editing pace* can be inferred. Two motion features are extracted per shot, MPEG-7 *motion intensity* [33] and a measure of *camera motion* [29]. A support vector machine based classifier is used to classify audio into one of: *speech*, *music*, *silence* and *other audio*. Finally, a measure of *shot repetition* is implemented [12]. A set of Finite State Machines (FSMs) is used in order to detect parts of a film where particular features are prominent. For example, in order to detect dialogue events, FSMs are used in order to detect areas which contain various combinations of speech shots, still cameras and repeating shots. The output of the FSMs is filtered, and a list of dialogue events is created. A similar process is repeated for the exciting events (where fast-paced editing and camera movement are sought) and for musical events (where, among others, shots with silence and music are sought). An evaluation with ten films of very different genres and origins [12], found that 95% of Dialogue events, 94% of Exciting events and 90% of Musical events were detected by the system, which indicates the reliability of the event detection system. It was also found that 91% of all shots in a film were categorised into one of the event classes by the event detection system.

3.2 Character Detection

This step involves the identification of who the main characters are in the film, how they are referred to in the audio description, and their gender for pronoun resolution. It

results in a log of the time-code of every utterance that contains each character's name, or a pronoun referring to them. Preprocessing was required to tidy-up some of the raw audio description scripts and to get them into a standard format. In order to identify the characters in a film we generated a list of capitalized words occurring > 4 times in the audio description script, with common stopwords removed. This threshold removed most incidental characters. Also, the set of capitalized words occurring $<= 4$ times contained a much higher percentage of non-characters than the set of capitalized words occurring > 4 times. Even so, some manual filtering of the resulting list was required in order to remove the names of commonly occurring locations which were capitalized. Manual intervention was also required in rare cases when a character was referred to in different ways in different parts of a film, including mis-spellings. In the film Enigma, the character Hester Wallace is at first described as 'Miss Wallace' whilst she is on formal terms with the main character, and then as 'Hester'. In the case of The English Patient, though we may believe the patient and the character Count Almasy to be the same person we chose to keep them as distinct characters, but we had to enter the non-capitalized instance of 'patient' manually.

For characters with common names it was possible to disambiguate their gender automatically, e.g. by looking up lists of male and female names generated by the US Census [34]. However, in about half the films that we dealt with many of the characters were referred to by their surnames, or had names that would not be found in census data of any country, e.g. 'Shrek' and 'Stitch', so manual intervention was required. Given a list of the main characters and their genders, pronoun resolution was achieved by looking back from each occurrence of / he | she | he's | she's / to the first character name of matching gender. This relatively simple method proved to be effective because of the nature of audio description; describers try hard to avoid ambiguous pronouns because it would distract the audience. We did not change the automatically generated output for our pronoun resolution.

Because we wanted a complete list of characters and their genders, with no false entries, we had to accept some manual intervention during this step. We tried a system for automatic named entity recognition [35] and whilst it performed well, it was not able to give the complete list of character names that we wanted, nor did it give genders. We also considered alternative information sources such as film scripts and the Internet Movie Database. The main problem was that different sources can refer to the same character in quite different ways, e.g. by first name or second name, or not by name at all.

3.3 Linking Characters and Events

The previous steps give us a set of events, with start and end times, and a set of Timecode:Character pairs. Initially, we made the assumption that a character is present in an event if his/her timecode is between the start and end time of the event. However, some preliminary evaluation and subsequent investigation of the audio description scripts suggested this assumption had to be modified. Audio describers must fit their descriptions around existing dialogue, and other important sounds, so what is being said in a piece of description does not always relate exactly to what is happening on-screen at that moment. In particular, if a new event starts with a long stretch of dialogue then the audio description for that event might be positioned at the end of the previous event. This observation motivated the use of an offset to integrate the event and character occurrence data. After trial and error with some small-scale evaluation we settled on an offset such that a character was associated with an event if:

$$EventStart - 10s < CharacterTimecode < EventEnd - 3s$$

By including audio description from 10s before the start of an event this offset should capture all advance mentions of characters, and hence improve *Recall* albeit at the expense of *Precision*. By ignoring audio description from the last 3s of an event *Precision* is improved with little or no reduction of *Recall* because the last 3s are likely to include an advance description of the next event, and it is unlikely that a character's only mention in an event will come in the last 3s.

4. EVALUATION

This section describes our initial objective evaluation of the performance of the technique proposed in section 3. The task evaluated here is this: given a set of film events, detect which characters can be seen on-screen at some point in the duration of each event. A sample of 20% of all events automatically detected from eleven films was selected for the evaluation; the films represented a broad range of styles and genres. The sample totalled 215 events which were sampled evenly across the films, i.e. from the first film Events 1, 6, 11... were used, from the second film Events 2, 7, 12..., etc. For each selected event, all the main characters that could be seen on-screen in the duration of the event were manually logged. In order to keep the evaluation objective we included any main character however briefly they appeared on-screen, and even if they did not seem to play any significant part in the event. By 'main character' here we mean a character included in the character list derived as detailed in section 3.2. Based on this, we could then measure *Precision* and *Recall* for our character detection technique:

$$Precision = \frac{CD}{TCR}$$

$$Recall = \frac{CD}{CSOS}$$

where $CSOS$ is the number of characters seen on-screen, CD is the number of characters detected who were on-screen and TCR is the total number of characters returned.

Evaluation results are shown in table 1 both for individual films, and collectively for all 215 events (the 'Total' row). The high *Precision* values are not surprising since an audio description tends to mention a character at around the time they are seen on-screen. Where *Precision* did suffer, however, this was typically due to one of three reasons. First, some audio description utterances are long, e.g. they take

Film Name	*Precision (%)*	*Recall (%)*
American Beauty	88	71
Chocolat	100	66
Enigma	94	82
High Fidelity	93	68
Lilo & Stitch	97	83
Ocean's Eleven	91	63
The Pelican Brief	96	92
Shrek	100	64
The English Patient	90	65
The Road to Perdition	82	73
The Royal Tenenbaums	96	61
Total	**93.4%**	**71.3%**

Table 1: Results of character detection in events compared with manual annotation

10s or more to speak. A character mentioned towards the end of a long utterance is still associated with the timecode at which the utterance started, i.e. the timecode is too early for that character. This problem could be alleviated by splitting long utterances and estimating timecodes for their segments, based on an expected rate of speech. Second, occasionally a character's name is mentioned in phrases like 'Mary goes into John's room' and 'Mary picks up the book that John had' - where John is not present on-screen. Third, our simple method for pronoun resolution did not distinguish between Subject and Object in utterances, so in 'Mary gives the book to Beth. Later, she gives the gun to John', 'she' is resolved to 'Beth' rather than to 'Mary'. A deeper linguistic analysis of the audio description script could help to alleviate the second and third problems, but it is not clear to us whether the extra effort involved would result in a significant improvement in *Precision*.

Low *Recall* figures are due to the fact that the target set for the evaluation included all characters that could be seen on-screen in the duration of the event, however briefly, and even if they did not play any role in the on-screen action. However, an audio description will not always have time to mention all characters and will concentrate on those who are most important in the scene. We assume then that the *Recall* for important characters is higher than the values provided in table 1. However, there are some reasons why audio description will never give 100% *Recall*. If a character can be recognised by their distinctive voice then they will not always be mentioned in the audio description; likewise if a character is mentioned explicitly in the dialogue. Near the start of a film, if a character's name is unknown to the sighted audience then the audio description will not refer to them by name, but rather as, for example 'a tall man in a brown suit'. Resolving such a phrase with a character's name is beyond the state-of-the-art in natural language processing.

5. ILLUSTRATIVE APPLICATIONS

In this section we present evidence of the potential usefulness of our approach in terms of the different modes of information access to feature film content that it supports. It should be noted, that the presented applications of the proposed technology are illustrative in nature. Our future work will target prototype system development based on these initial investigations, complete with a formal evaluation of their effectiveness.

5.1 Retrieval by Character and Event-Type

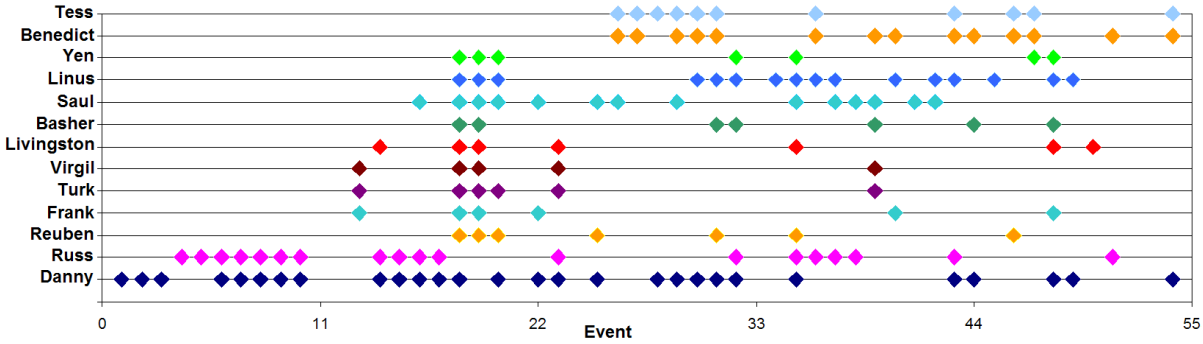
As a result of the character to event matching process it is possible to retrieve events containing particular characters. For example, it is quite straightforward to implement a retrieval system that can answer the query *Find me exciting parts of the film 'The Road To Perdition' that contain the character 'Michael'*. Similarly, it is possible to query a film and retrieve events in which specific characters are interacting. If a database with character names assigned to actors/actresses were to be leveraged, such as the aforementioned Internet Movie Database, it would be possible to support queries for similar events in which a particular actor/actress appears *across* films e.g. *Find me all dialogue events featuring Humphrey Bogart and Ingrid Bergman.*

5.2 Character Occurrence and Plot

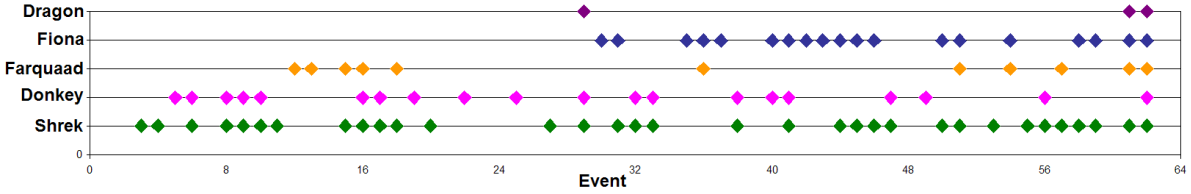
Based on the indexing performed, it is possible to garner a high level picture of the evolution of a film's plot and characters relation therein. For example, figures 3(a) and 3(b) shows a timeline for the appearance of characters in sequential dialogue events in the films Ocean's Eleven and Shrek, respectively. Ocean's Eleven is a heist film that tells the story of a casino robbery planned by the characters Danny and Russ, and involves a large number of accomplices. As can be seen from figure 3(a), in the early part of the film most of the dialogue events involve Danny and Russ together, however, around a third of the way through, the rest of the gang begin to assist in the planing of, and ultimately undertake, the heist as can be seen from the increase in characters present. Also, around half way through the film, Danny's love interest, Tess, and her husband, Benedict, become heavily involved in the film and thus start to feature prominently in the timeline. Thus, the film's narrative can be visualised albeit from quite a high level. Similarly, the film Shrek primarily involves two characters, Donkey and Shrek. The various characters temporal relation to the plot of the film is clearly reflected in Figure 3(b) as Shrek and Donkey initially trek to Prince Farquaad's palace, who sends them off to rescue Fiona, who is being held captive by the Dragon. Once she is rescued they begin to make their way back to Farquaad where all of the characters then appear together for the final climactic scenes of the film.

5.3 Character Relationships

By examining the events in which characters appear together it is possible to build up a picture of the character interactions in the film. To this end, the character co-occurrence for character X and Y, $C(X,Y)$, is declared as the total number of events in which characters X and Y appear together, divided by the number of events in total in which character X appears. This can be represented as $C(X,Y) = \frac{X \cup Y}{X}$. Note that $C(X,Y) \neq C(Y,X)$ as $C(Y,X) = \frac{Y \cup X}{Y}$. Tables 2 and 3 contain the co-occurrence figures for characters in the film American Beauty. Each row contains the co-occurrence figures for a single character. So, taking the character Lester as an example, he appears with himself 100% of the time, with Carolyn in 40% of the events that he is in, with Jane in 43% of events, and so on.



(a) Ocean's Eleven



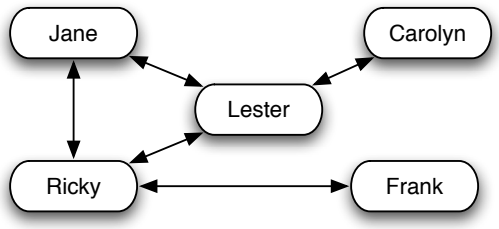
(b) Shrek

Figure 3: Character appearance in dialogue events

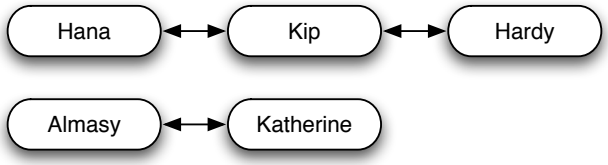
In table 2, the character pairs with the highest mutual co-occurrence are highlighted. This table can give insights as to the various character interactions. For example, Lester appears in most events with his daughter Jane. Carolyn is present in all of the events that Buddy appears in and thus $C(Buddy, Carolyn)$ is given a value of 1.00. However, $C(Carolyn, Buddy)$ is only given a value of 0.23. This suggests that Buddy's involvement in the film is dependant on Carolyn (Carolyn, one of the main characters, has a brief affair with Buddy). While it may be difficult for a machine to interpret these results in a meaningful manner, we believe that they can certainly help a human observer in gaining insight into the relationships implicit in the film.

By finding the relationships in which both $C(X, Y)$ and $C(Y, X)$ have high values, it is possible to detect the characters whose relationships are important to the film, assuming that important relationships are given an increased amount of screen time – this is not always true, but is a good rule of thumb, at least for our initial work. Table 3 highlights via different colours the relationships in which both co-occurrence figures are above a predefined threshold (in this case the threshold was chosen as the average co-occurrence figure, 0.379). For example, $C(Lester, Carolyn)$ and $C(Carolyn, Lester)$ are both above the threshold, and so it can be deduced that their relationship is particularly meaningful to the film, which is indeed the case.

In total there are five co-occurrences that are higher than average dual co-occurrence interactions. By analysing the high dual co-occurrence figures, it is possible to automatically create graphs of the character interactions of a film. Figures 4(a) and 4(b) illustrate two such examples. Figure 4(a) shows all of the high dual co-occurrences for American Beauty (as highlighted in table 3). Note how Lester has a lot of interactions with the various main characters. This



(a) American Beauty



(b) The English Patient

Figure 4: Main character interactions

reflects the nature of the film, where Lester is the central character and much of the film revolves around him and the plot is narrated from his perspective. Figure 4(b) shows the main interactions in the film The English Patient. Note that this reflects the fact that the film consists of relatively unrelated parallel stories rather than a single predominant one, as illustrated by the two sets of relationships which are not dependent on each other.

6. CONCLUSION

	Lester	Carolyn	Jane	Buddy	Angela	Ricky	Frank
Lester	1.00	0.40	**0.43**	0.09	0.36	0.38	0.21
Carolyn	**0.61**	1.00	0.45	0.23	0.23	0.35	0.13
Jane	0.49	0.34	1.00	0.02	0.34	**0.56**	0.12
Buddy	0.57	**1.00**	0.14	1.00	0.00	0.14	0.14
Angela	**0.71**	0.29	0.58	0.00	1.00	0.54	0.29
Ricky	0.46	0.28	**0.59**	0.03	0.33	1.00	0.41
Frank	0.56	0.22	0.28	0.06	0.39	**0.89**	1.00

Table 2: Character co-occurrence – highest co-occurrence for each character highlighted

	Lester	Carolyn	Jane	Buddy	Angela	Ricky	Frank
Lester	1.00	**0.40**	**0.43**	0.09	0.36	**0.38**	0.21
Carolyn	**0.61**	1.00	0.45	0.23	0.23	0.35	0.13
Jane	**0.49**	0.34	1.00	0.02	0.34	**0.56**	0.12
Buddy	0.57	**1.00**	0.14	1.00	0.00	0.14	0.14
Angela	**0.71**	0.29	0.58	0.00	1.00	0.54	0.29
Ricky	**0.46**	0.28	**0.59**	0.03	0.33	1.00	**0.41**
Frank	0.56	0.22	0.28	0.06	0.39	**0.89**	1.00

Table 3: Dual co-occurrence highlighted by different colours

In this paper, we demonstrated the benefits of combining the results of audiovisual event-based indexing with an analysis of an accompanying audio description. We showed how in this manner it is possible to automatically associate film characters with the memorable events in which they occur. We explained why event-based indexing is preferable to shot-based and scene-based alternatives. We also note here that event-based indexing is better than indexing the occurrence of characters by single timestamps, which we could have achieved using audio description scripts without audiovisual analysis. Event-based indexing means that we can retrieve relevant surrounding material in response to users queries for characters. It also means that we can produce less cluttered and more meaningful visualisations which show clearly which characters appear together in events. We then have a good basis for character co-occurrence analysis, i.e. it would not make sense to analyse character co-occurrence based on how close their timestamps were. All in all, this further establishes the need for combining audiovisual analysis with information extraction from text. Text segmentation techniques applied to audio description could not access the film making principles embodied in the audio and visual data streams of film data. Conversely, audiovisual analysis is not able to describe semantic film content at the high level that audio description does.

The approach was evaluated and initial results are encouraging. Further, the potential usefulness of the technique in terms of novel way of accessing film content was demonstrated. A number of directions for future work have already been identified. We plan to further develop the individual analyses, specifically to leverage the other's results in an iterative manner – the feedback loop depicted in figure 2. Future techniques will add different sources of text information, such as scripts or subtitles, in order to provide as rich a description as possible. We also plan to build a number of real applications based on the possibilities outlined in section 5. The actual effectiveness of the approach will then be evaluated in the context of these applications using a variety of user groups – general fans of feature films and film and media students at the University's School of Communication Studies have already been identified as target groups. Further investigation may facilitate character classification based on the types of events that they appear in, e.g. a character that appears in an unusually high amount of exciting events could be a villain. Additional inferences about a films' story may also be possible with further analysis. For example, it may be possible to detect of key points in a film by examining the first and last occurrences and co-occurrences of characters and their activities at those times.

7. ACKNOWLEDGEMENTS

We are grateful to BBC, RNIB and itfc for providing the audio description scripts used here. AS thanks Elaine White for her work on the evaluation data. The support of the Irish Research Council for Science, Engineering and Technology is gratefully acknowledged. This material is based on works supported by Science Foundation Ireland under Grant No. 03/IN.3/I361.

8. REFERENCES

[1] N. Dimitrova, H.-J. Zhang, B. Shahrary, I. Sezan, T. Huang, and A. Zakhor, "Applications of video-content analysis and retrieval," in *IEEE Multimedia*, July-Sept. 2002.

[2] P. Salembier, "Overview of the mpeg-7 standard and of future challenges for visual information analysis," in *Eurasip Journal on Applied Signal Processing*, 2002.

[3] C. Cotsaces, N. Nikolaidis, and I. Pitas, "Signal processing magazine, ieee, pages 28-37," in *Video shot detection and condensed representation. a review*, vol. 23, 2006.

[4] M. Yeung and B.-L. Yeo, "Video visualisation for compact presentation and fast browsing of pictorial content," in *IEEE Transactions on Circuits and Systems for Video Technology*, 1997, pp. 771–785.

[5] T. Liu and J. R. Kender, "Proceedings of the ieee workshop on content-based access on image and video libraries," in *A Hidden Markov Approach to the*

Structure of Documentaries, 2000.

[6] Y. Cao, W. Tavanapong, K. Kim, and J. Oh, "Audio-assisted scene segmentation for story browsing," in *Proceedings of the International Conference on Image and Video Retrieval*, 2003.

[7] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval: the end of the early years," in *IEEE Transactions Pattern Analysis and Machine Intelligence*, 2000.

[8] M. Yeung and B.-L. Yeo, "Time constrained clustering for segmentation of video into story units," in *Proceedings of International Conference on Pattern Recognition*, 1996.

[9] Y. Rui, T. S. Huang, and S. Mehrotra, "Constructing table-of-content for video," in *ACM Journal of Multimedia Systema*, 1998, pp. 359–368.

[10] J. Zhou and W. Tavanapong, "Shotweave: A shot clustering technique for story browsing for large video databases," in *Proceedings of the Workshops XMLDM, MDDE, and YRWS on XML-Based Date Management and Multimedia Engineering-Revised Papers*, 2002.

[11] B. Lehane, N. O'Connor, A. Smeaton, and H. Lee, "A system for event-based film browsing," in *3rd International Conference on Technologies for Interactive Digital Storytelling and Entertainment, Lecture Notes in Computer Science (LNSC) Vol 4326*, 2006.

[12] B. Lehane and N. O'Connor, "Movie indexing via event detection," in *7th International Workshop on Image Analysis for Multimedia Interactive Services, Incheon, Korea, 19-21 April*, 2006.

[13] R. Leinhart, S. Pfeiffer, and W. Effelsberg, "Scene determination based on video and audio features," in *In proceedings of IEEE Conference on Multimedia Computing and Systems*, 1999, pp. 685–690.

[14] Y. Li and C.-C. J. Kou, *Video Content Analysis using Multimodal Information*. Kluwer Academic Publishers, 2003.

[15] Y. Li and C.-J. Kou, "Movie event detection by using audiovisual information," in *Proceedings of the Second IEEE Pacific Rim Conferences on Multimedia: Advances in Multimedia Information PRocessing*, 2001.

[16] L. Chen, S. J. Rizvi, and M. Ötzu, "Incorporating audio cues into dialog and action scene detection," in *Proceedings of SPIE Conference on Storage and Retrieval for Media Databases*, 2003, pp. 252–264.

[17] A. Salway, "A corpus-based analysis of the language of audio description," in *Selected Proceedings of Media for All*, 2007.

[18] A. Salway and M. Graham, "Extracting information about emotions in films," in *Procs. ACM Multimedia, pp 199-302*, 2003.

[19] A. Salway, A. Vassiliou, and K. Ahmad, "What happens in films?" in *Procs. IEEE ICME*, 2005.

[20] V. Lingabavan and A. Salway, "What are they talking about? information extraction from film dialogue," in *Dept. of Computing Technical Report CS-06-07, University of Surrey*, 2006.

[21] A. Vassiliou, "Film content analysis: a text-based approach," in *PhD dissertation, Dept. of Computing, University of Surrey*, 2006.

[22] E. Tomadaki, "Cross-document coreference between different types of collateral texts for films," in *PhD dissertation, Dept. of Computing, University of Surrey*, 2006.

[23] K. Shirahama, K. Iwamoto, and K. Uehara, "Video data mining: Rhythms in a movie," in *Procs. IEEE Int. Conf. Multimedia and Expo, ICME*, 2004.

[24] K. Shirahama, Y. Matsuo, and K. Uehara, "Mining semantic structures in movies," in *LNCS 3392, 2005, pp. 116-133.*, 2005.

[25] A. Hanjalic and L. Xu, "Affective video content representation and modeling," in *IEEE Trans. Multimedia 7(1), 2005, pp. 143-154*, 2005.

[26] C. Chan and G. Jones, "Affect-based indexing and retrieval of films," in *Procs. ACM Multimedia 2005, 427-430*, 2005.

[27] B. Adams, C. Dorai, and S. Venkatesh, "Towards automatic extraction of expressive elements for motion pictures tempo," in *IEEE Trans. Multimedia 4 (4), 2002, pp. 472-481.*, 2002.

[28] M. Everingham, J. Sivic, and A. Zisserman, "Hello! my name is... buffy' - automatic naming of characters in tv video," in *Proceedings of the 17th British Machine Vision Conference (BMVC2006)*, September 2006.

[29] B. Lehane, N. O'Connor, and N. Murphy, "Dialogue scene detection in movies," in *International Conference on Image and Video Retrieval (CIVR), Singapore, 20-22 July 2005*, 2005, pp. 286–296.

[30] B. Lehane and N. O'Connor, "Action sequence detection in motion pictures," in *The international Workshop on Multidisciplinary Image, Video, and Audio Retrieval and Mining*, 2004.

[31] D. Bordwell and K. Thompson, *Film Art: An Introduction*. McGraw-Hill, 1997.

[32] K. Dancyger, *The Technique of Film and Video Editing. History, Theory and Practice*. Focal Press, 2002.

[33] B. Manjunath, P. Salember, and T. Sikora, *Introduction to MPEG-7, Multimedia content description language*. John Wiley and Sons ltd, 2002.

[34] (2005) Frequently occurring first names and surnames from the 1990 census. [Online]. Available: http://www.census.gov/genealogy/www/freqnames.html

[35] (2005) Named entity tagging demonstration. [Online]. Available: http://l2r.cs.uiuc.edu/ cogcomp/eoh/nedemo.html