UniS University of Surrey

EPSRC

TIWO – Television in Words

EPSRC GR/R67194/01, January 2002-5

Report on Workpackage 1

Data Models and Knowledge Representation for Video Annotation

15 November 2004

FINAL VERSION

Andrew Salway, Yan Xu and Andrew Vassiliou
Department of Computing, University of Surrey

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

This report firstly provides an overview of video data compression formats, metadata standards, video data models and the use of knowledge representation formalisms in systems that process video data. Secondly, this report gives an overview of the work done in the TIWO project towards modelling video data, specifically films, in terms of their narrative content: the work comprises a series of UML models of narrative based on a theory of narratology (Chatman 1978), and the representation of two feature films using a knowledge representation formalism called Plot Units (Lehnert 1981).

Pointers are given to sources of further information including websites, TIWO project documents and academic publications.

2 Description of TIWO Workpackage 1 from Project Proposal Document

A review of video content representation schemes will consider standards such as MPEG-4, MPEG-7 and the use of knowledge representation formalisms from artificial intelligence, especially Schank's scripts, plans and goals. Parallel to this will be a review of video data models that structure video data in terms of intervals, objects and the spatial, temporal and other relationships between them; including data models that link video data with text data. Domain modelling exercises to adapt the knowledge representation formalisms and video data models will take place during the two Round Table meetings in the period, and during meetings with individual Round Table members at their organisations.
Milestone 1: Data model / knowledge representation for video annotation with audio descriptions

3 Motivation

Compression formats, data models, metadata schemes and knowledge representation formalisms for video data all impact on the functionality of multimedia applications, and hence impact on the way users can access and interact with moving images like films and television programmes, and other associated media, such as audio description scripts.

Our work in TIWO has two particular requirements for storing, accessing and interacting with video data (films and television programmes) and audio description, which are:

(i) The need to integrate video data and text data.

(ii) The need to deal with the narrative content of video data.

One way to describe narrative is that it is a sequence of connected events. Events may be connected by causal-effect relationships, either physically or mentally. For example, jealousy may cause a murder, or pushing someone may cause his/her injury. Different forms of media can express narrative: film, novels, drama, poems, etc. Thus, narrative links text and video in

a way that they all express the same narration. For example, "The English Patient" is manifested in different ways such as novel, drama, and film. They all express the same narrative. A video data model may benefit from narrative and be able to provide more semantic content for information retrieval.

4 Overview of Video Data Formats, Models and Knowledge Representation

4.1 Standards for Video Data

Starting with compression formats for video data, there is a fundamental difference between 'traditional' standards such as MPEG-1 and MPEG-2, that conceive of the moving images as a sequence of frames, and the emerging MPEG-4 standard that conceives of moving images as sets of audio-visual objects and sets of instructions (including user interaction) that governs their playback. Thus, there is the potential for a new generation of video editing, retrieval and browsing systems. For example, consider the difference between editing video data by 'cutting and pasting' sequences of frames, and 'cutting and pasting' objects and backgrounds.

Access to multimedia archives relies on metadata, i.e. 'data about data', in order for users' requests for information (queries) to be matched successfully against the multimedia artefacts. There are metadata standards that are intended to be generic and extensible, such as Dublin Core, which specifies 15 core elements of metadata, and MPEG-7, which specifies a wide range of low-level and high-level features for describing multimedia content. Other metadata sets are driven by the needs of more specialist user groups such as organisations in the broadcasting industry. For example the work of the European Broadcasting Union (EBU), the Society of Motion Picture and Television Engineers (SMPTE) and the pro-MPEG group on the Advanced Authoring Format (AAF) for media content creation and the Multimedia Exchange Format (MXF) for interoperability between television production applications.

Websites:

- The MPEG Homepage: www.chiariglione.org/mpeg/
- The Dublin Core Metadata Initiative homepage: http://dublincore.org/
- For an overview of AAF and MXF, see: www.snellwilcox.com/knowledgecenter/mxf_aaf.html

4.2 Generic Video Data Models

A video data model supports the storage of data about video content. Put another way, it is a formal description of what comprises video content, where video content is typically understood from a viewer's perspective, e.g. what is seen happening on-screen. To support

querying-retrieval of video data in general purpose video databases, a number of generic models have been proposed to describe video content in terms of things/objects/entities alongside happenings/actions/events organised in space and time (Agius and Angelides, 2001, Cheng 2002, Dumas 2002). When dealing with a particular kind of video data it is sometimes appropriate to use a more specific data model; for example one video data model for a film browsing application allows for the storage of data about actors, shots, scenes, edit effects, etc. (Corridoni et al. 1996).

Increasingly, collateral text is being seen as a crucial resource for video indexing in a wide variety of applications, to complement the audio-visual features that can be extracted directly from video data. Video data may be annotated with textual descriptions manually, but in many cases extant collateral text can be exploited, such as speech from audio data and/or closed captions, exploited by systems such as *Informedia* to index, retrieve and summarise news broadcasts and documentary programmes (Wactlar et al 2000). Another example integrating collateral text with video data is a video data model called *VideoText*, which is based on free text annotation and mapped with logical video segments and a corresponding query language (Jiang, Montesi and Elmagarmid, 1999). The data model contains information about the relationships between text and video data.

Existing data models do capture video content at quite a high level; where low-level is close to the bit-stream and pixels, and high-level is close to what a human viewer understands by watching the video. The generic conception of video content as things and happenings organised in space and time is complete when considering a common-sense view of what can be seen on screen, however it does not capture the higher-level interpretations of a story that viewers might form from watching the unfolding sequence of action, i.e. narrative aspects of video content.

For a more extensive review please see Yan Xu's MPhil-PhD transfer report 'Retrieving and Browsing Story Elements' (2003) which is available on the TIWO website.

References:

Agius, H. W. and Angelides, M.C., 2001. Modelling Content for Semantic-Level Querying of Multimedia. *Multimedia Tools and Applications*, 15, pp. 5-37.

Cheng, T. S. 2002. The Event Matching Language for Querying Temporal Data, *IEEE Transactions on knowledge and data engineering*, vol. 14, no. 5.

Corridoni, J. M., Alberto D. B., Dario L. and He W., 1996., Multi-perspective Navigation of Movies, *Journal of Visual Languages and Computing*. 7, no. 4, pp. 445-466.

Dumas, M. 2002. A Sequence – based Object-Oriented Model for Video Databases. *Multimedia Tools and Applications*, 18, pp. 249-277.

Jiang, H., Montesi, D. and Elmagarmid, A.K. 1999, Integrated Video and Text for Content-based Access to Video Databases. *Multimedia Tools and Applications*, 9, pp. 227-249.

Wactlar, H., Olligschlaeger, A., Hauptmann, A., and Christel, M., 2000. 'Complementary Video and Audio Analysis for Broadcast News Archives'. *Communications of the ACM*, February, 43(2), pp.42-47.

4.3 Knowledge Representation for Multimedia Systems

Story understanding systems have been the subject of research in artificial intelligence for several decades, though the stories are typically verbal and relatively short and simple when compared with the kinds of films we are trying to deal with. A common feature of these systems is the need to represent common sense knowledge, alongside a representation of the emerging story. A useful survey of these systems, and the knowledge representation systems on which they are based, is given by Erik Mueller (Mueller 2003).

A variety of knowledge representation formalisms have been used in multimedia applications in order to process the video data more in terms of its meaning. By combining *concept ontology* and a *semantic network*, objects, events and actions in the video are represented by nodes and their relationships are represented by links (Roth, 1999). This system can retrieve any of those concepts or related events. Although it can infer related events, the kind of relationships between events is not specified. Another intelligent multimedia information retrieval (IMIR) system, called *SmartVideoText*, combines keywords processing techniques with *conceptual graphs*, can infer some concepts-similar objects, or other objects as the agent or patient of the current object (Kokkoras et al. 2002). A system design of browsing the structure of multimedia stories is applied into one simple story to capture its theme and plot and one complex story, a film, to capture only its theme, based on their *story thread* representation (Allen and Acheson, 2000). However, there is no classification of types of causalities. Characters' development is excluded, as well as relationships among people. By using operators to represent some attributes of narrative, a system is ideal to process texts into knowledge base of characters, spatial relationships of characters, props, locations and state (as new actor or existing one) (Callaway and Lester, 2002). Common sense, such as how a person hammers a screw into a piece of wood as a physical action, can be simulated by conceptual graphs (Parkes, 1989). Thus, an IMIR system can retrieve objects, actions and their temporal relationships. When context-free knowledge unites context-dependent knowledge, a system can infer context-related events physically or mentally at a simple level (Tanaka, Ariki and Uehara, 1999).

For a more extensive review please see Yan Xu's MPhil-PhD transfer report 'Retrieving and Browsing Story Elements' (2003) which is available on the TIWO website.

References:

Allen, R. B.; and Acheson, J. 2000. Browsing the Structure of Multimedia Stories. *In Proceedings of the 5th ACM Conference on Digital libraries,*. New York: ACM Press, pp. 11-18.

Callaway, C. B., and Lester, J. C. 2002. Narrative Prose Generation, *Artificial Intelligence* 139, pp. 213-252.

Kokkoras, F., Jiang, H., Vlahavas, I., Elmagarmid A. K., Houstis, E. N., and Aref, W. G., 2002. Smart VideoText: a Video Data Model based on Conceptual Graphs, *Multimedia Systems.* 8, pp. 328-338.

Mueller, Erik T., 2003, Story understanding through multi-representation model construction. In Graeme Hirst & Sergei Nirenburg (Eds.), *Text Meaning: Proceedings of the HLT-NAACL 2003 Workshop*. East Stroudsburg, PA: Association for Computational Linguistics. pp. 46-53.

Parkes, A. P. 1989, The Prototype CLORIS System: Describing, Retrieving and Discussing Videodisc Stills and Sequences. *Information Processing & Management.* Vol. 25, No. 2, pp. 171-186.

Roth, V. 1999. Content-Based Retrieval from Digital Video. *Image and Vision Computing,* 17, pp. 531-540.

Tanaka, K., Ariki, Y., and Uehara, K. 1999 Organization and Retrieval of Video Data. *IEICE Transactions on Information and systems* E82-D (1), pp. 34-44.

5 Modelling Narrative Content of Films in TIWO

5.1 A Data Model of Narrative?

Given the prior existence of a video-text data model, domain modelling to deal with integrating video data and audio description seemed less important, but we felt a clear need to capture high-level semantic video content, i.e. narrative and its constituent elements: events, states, characters, locations etc. Thus we used Chatman (1978), as a domain expert, specifically his book 'Story and Discourse: Narrative Structure in Fiction and Film', to try and model some of the elements of narrative. To do this we used the Unified Modelling Language (UML); UML is used to specify, visualise, construct and document the artefacts of software systems. It has also been extended to other fields such as business modelling. In this case we are used it to model elements of narrative as artefacts of a software system or framework.

The first chapter of Chatman's book was chosen for modelling, as it was a good overview of the book's content. Each section of Chapter 1 was considered separately with respect to whether the subject being discussed or defined could be modelled with the UML. Key sentences were taken, quite literally in some cases, and turned into UML diagrams.

As an example let us consider the concept of a story as described by Chatman. Chatman states that a story (histoire) is the content of the narrative or chain of events plus elements of existents such as characters and settings. The story is, "[t]he set of events tied together which are communicated to us in the course of the work," [p. 20]. "Story, in one sense, is the continuum of events presupposing the total set of all conceivable details…" [p. 28]. To model 'Story' we should consider what other elements of narrative are associated with the concept, i.e. events and existents. "Events… are either acts or actions in which an existent is the agent of the event, or happenings, where the existent is the patient." [p. 32]. "An existent is either a

character or an element of setting (based on whether or not it performs a plot-significant action.)" [p. 32]. Although these are not the only narrative elements associated with 'story' by Chatman (others include *Discourse* and *Cultural codes*) it was enough to build up a UML *class* diagram for 'Story'.

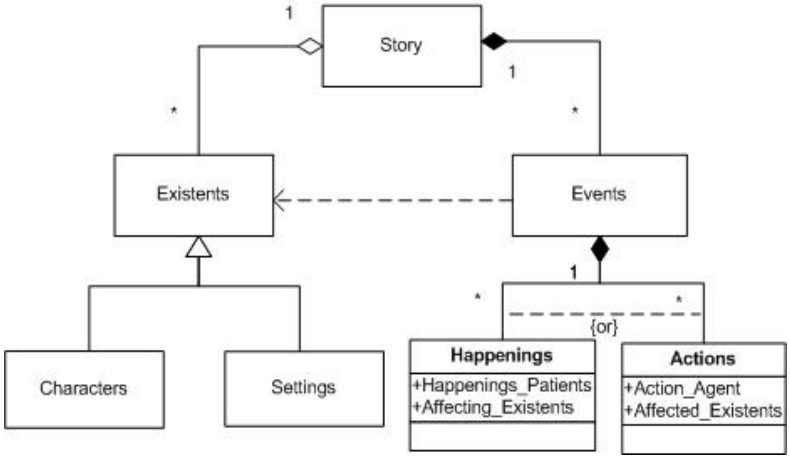


Figure 1. An example of modelling narrative elements using the UML's Class diagrams.

In Figure 1 we consider it impossible for a story to exist without events; we are saying that events are not merely parts of a story but are the essence of a story. Thus, we model the 'Story' – 'Events' association as a *composition*, i.e. stories are *made up of* events. With respect to existents, a story has existents but is *not composed of them* and so the association is modelled as an *aggregation*.

Events seem to be dependent on existents: an action cannot be carried out without an *actor* and a happening must affect some existent(s). Here we are saying that events are made up of actions or happenings. Thus a composition association is used to describe this relationship. Actions and happenings, as we are told, have agents and patients respectively and the consequences for Actions and happenings also affect existents in the story. Therefore, these components have been included as attributes of the classes 'Actions' and 'Happenings'.

A generalisation association was chosen to illustrate how existents can be split into subtypes. Although not directly specified by Chatman, it seems reasonable that an existent can be a character, a setting or another symbolic object.

This is just one example of one narrative concept, story, and not all of the associated narrative elements have been modelled here. This model is linked to other more extensive concepts of narrative such as 'Discourse' to form a more in-depth, *'big-picture'* of narrative. The complete models produced from this exercise are reproduced in Appendix 1 of this report.

For more details of this work please see Andrew Vassiliou's MPhil-PhD transfer report 'Representing Narrative in Multimedia Systems' (2004) which is available on the TIWO website.

References:

Chatman, S., *Story and Discourse: narrative structure in fiction and film*, Ithaca: Cornell University Press, 1978.

5.2 Representing Video Data with Plot Units

From our review of knowledge representation formalisms we were particularly interested in the idea of Plot Units proposed by Wendy Lehnert (1981). Plot Units, a kind of conceptual structure, are part of a simulation of how the human mind summarises a story by highlighting its central concepts. A narrative might be represented partly by plot units to show how emotion may affect what events are going to happen, or how events may affect characters' emotions. Perhaps it is also possible to find a kind of pattern that stories may share, even across multiple characters. Plot units have been adopted by Allen and Acheson (2000) into their concept of a *story thread,* as mentioned in section 4.3, to capture theme and plot of multimedia stories.

A simple plot unit is made up of two *affect states* and one *causal link*. Lehnert considers the central part of a story as characters' emotional reactions to events and states of affect, which are classified into three affect states: *Positive Event* (+), *Negative Event* (-) and *Mental State* (M). Diagonal links between linear state affects are invented to make causality explicit. Thus, she classifies four types of causal links: *motivation link* (m-link), *actualisation link* (a-link), *termination link* (t-link) and *equivalence link* (e-link). With three types of affect states and four type of causal links, there are 15 pair wise configurations which are called *primitive plot units* and they are named such as success, failure, loss etc. Connected primitive plot units become certain *complex plot units* for a single character. Using plot units to describe configuration of multiple characters needs some causal links between them, which are called cross links which do not have types as intra-character causal links do, so they can be linked between any pair of affected states between two different characters. For example, the mini 'story' of "how John gets his car started by Paul's help" can be represented as in Figure 2. This pattern, a specific combination of several plot units, is called an *honoured request*.

An example of the Plot Units for 7 scenes of the film *Harry Potter and the Philosopher's Stone* (2001) is given in Appendix 2 with causal link types shown. We have now represented two complete films with plot units – *The Matrix Revolutions* (2003) and *The Pelican Brief* (1993). Each film took approximately 35 hours to represent: some details about the number of plot units and types of link in each film are given in Table 1. These two films have a-links as the most common and e-links as the least common. The difference is that *The Pelican*

Brief has more m-links and cross-links than *The Matrix Revolutions*, which may be due to the genre of the film. In crime-thriller film characters tend to have more motivations and they interact with each other more often.

> *When John tried to start his car this morning, it wouldn't turn over. He asked his neighbour Paul for help. Paul did something to the carburettor and got it going. John thanked Paul and drove to work.*

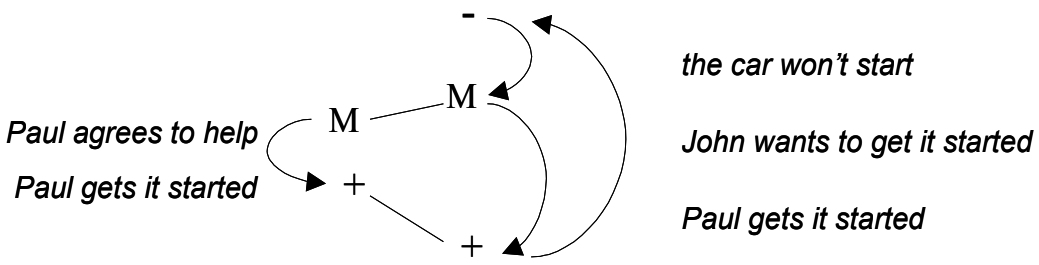


Figure 2. Plot Units of character John getting his car started, based on Lehnert (1981).

Table 1 – Statistics of plot units in two films.

Film	Number of Affect States	Number of Plot Units	a-links	m-links	t-links	e-links	Cross link
The Matrix Revolutions (124 minutes)	151	98	54	19	16	9	67
The Pelican Brief (134 minutes)	201	137	65	38	25	9	108

For more details of 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 in early 2005 and a paper will be submitted to IEEE International Conference on Multimedia & Expo, ICME 2005.

References:

Lehnert, W. G. 1981. Plot Units and Narrative Summarization. *Cognitive Science* 4, pp. 293-331.

Appendix 1: UML Models of Narrative Based on Chatman (1978)

Throughout his book Story and Discourse (1978), Chatman describes narrative structure and how narrative elements complement each other. The class diagram in Figure 3 was developed from Chatman's descriptions of the narrative elements. Many of the links, relationships and associations were explicitly mentioned in the book and the ones that were inferred were also taken from ideas and definitions in the book. The model largely parallels ideas from chapter 1 of Chatman's book, which gives an overview of his view of narrative structure. The split of narrative into story and discourse can be seen clearly in the model. The basic elements of a narrative such as events, states, existents and plot units have been captured in a way as to show their possible relationships and dependencies to one another. Although this model is by no means complete, it can serve as a framework for further modelling narrative and perhaps help in the development of systems that can process, navigate and extract information about narrative.

Appendix 1: UML Models of Narrative Based on Chatman (1978)

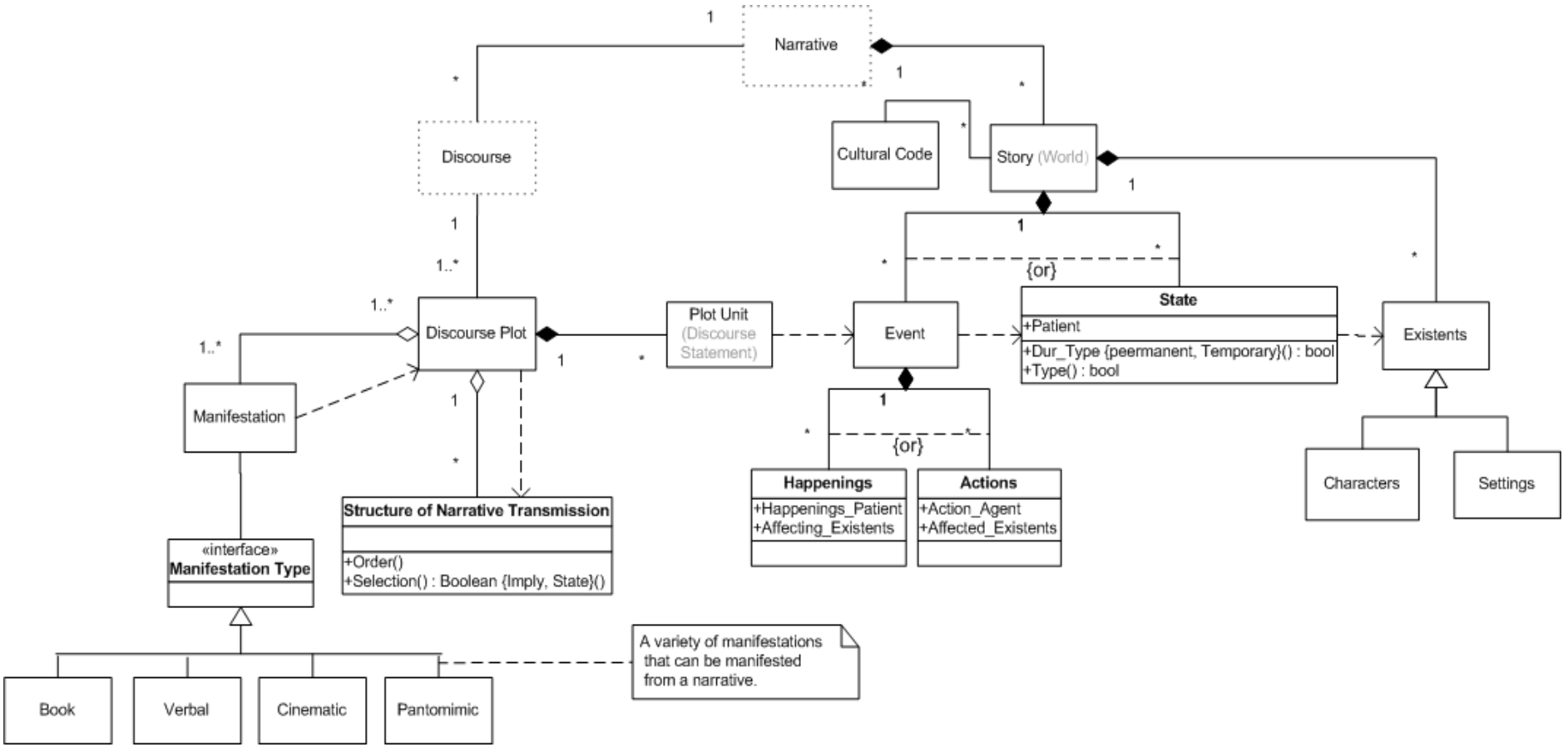
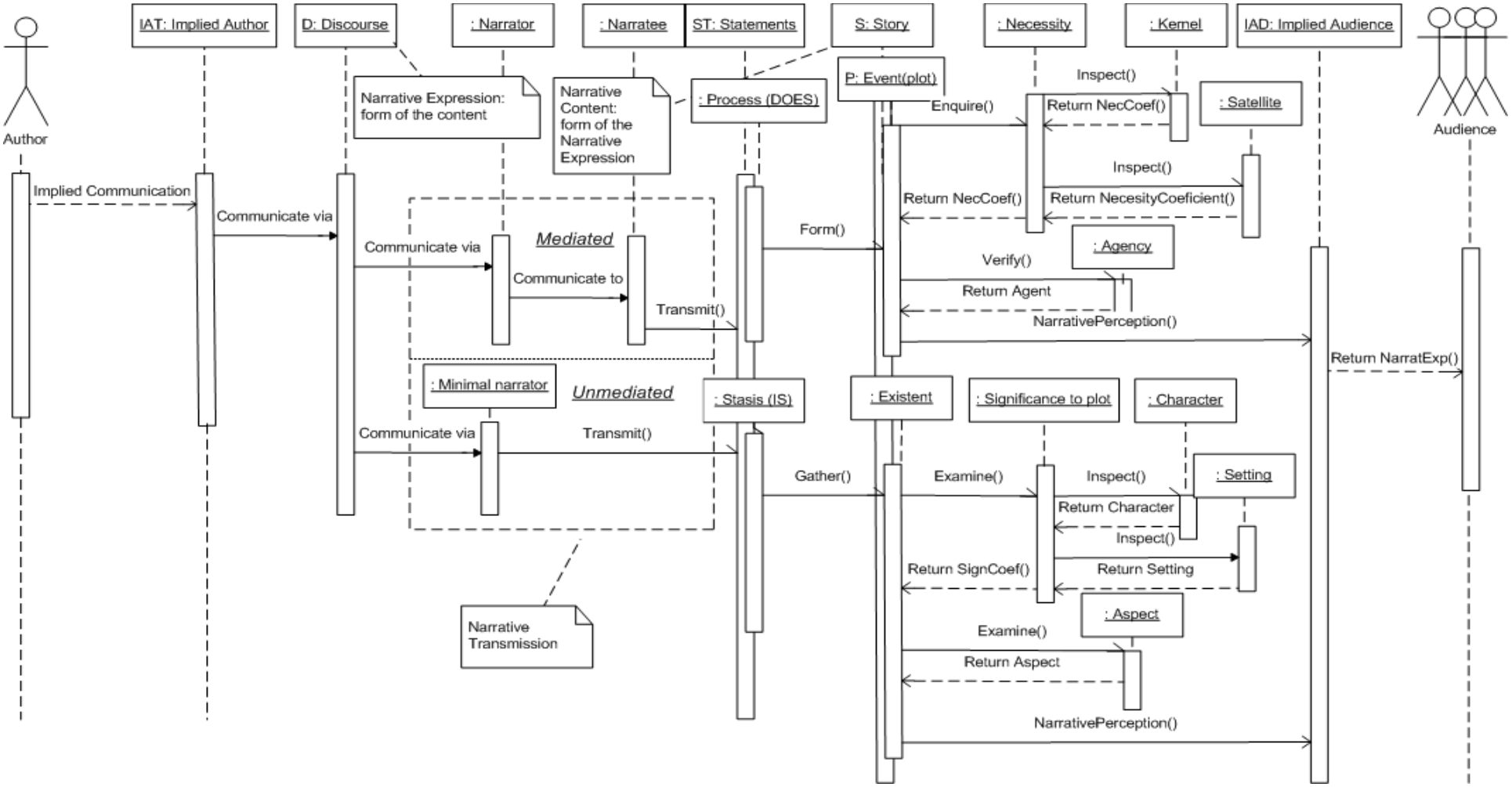


Figure 3. A possible model for narrative, and its elements, in the form of a UML class diagram.

Appendix 1: UML Models of Narrative Based on Chatman (1978)

Chatman describes an overall concept of narrative communication; the possible processes and flow of communication of narrative from the Author to the Audience. Figure 4, shows a UML sequence diagram that tries to capture in more detail these processes and the overall flow of communication. As well as this, there is a sequential element, i.e. the diagram tries to capture the idea of parallel processing in the audiences' minds and the sequence in which that happens (flow from left to right, top to bottom). This diagram serves to show the complexity of the human mind with respect to 'understanding' a narrative and how difficult it would be to model these processes, as well as their order of function.

Appendix 1: UML Models of Narrative Based on Chatman (1978)



Appendix 1: UML Models of Narrative Based on Chatman (1978)

This scene contains 3 events: ‘wake Harry up’, ‘in the kitchen’ and ‘before taking off to the Zoo’. These events are represented by plot units, as seen in the following event explanations.

1) Mrs. Dursley’s first mental state is that she wants Harry to cook breakfast for them, and the second mental state is that she wants to wake him up first; she actualised her desires by waking him up, and asking him to cook. Both events were negative events from the perspective of Harry. Dudley wanted to wake Harry too so he shouted and stepped on the stair. This event helped his mum to wake up Harry.
2) Dudley checked the numbers of gifts and feels unhappy with the amount. He requested more gifts from his father, Mr. Dursley who offered to give him more gifts. His desire is resolved by Mr. Dursley’s offer.
3) Mr. Dursley warns Harry not to have any “funny” stuff happen during the day. It seemed a problem for Harry. Later on, his magic made the snake come out from the cage and Dudley is locked inside instead. These things made Mr. Dursley very angry. His threatening failed in that sense, which seems to be positive for Harry.

Appendix 1: UML Models of Narrative Based on Chatman (1978)

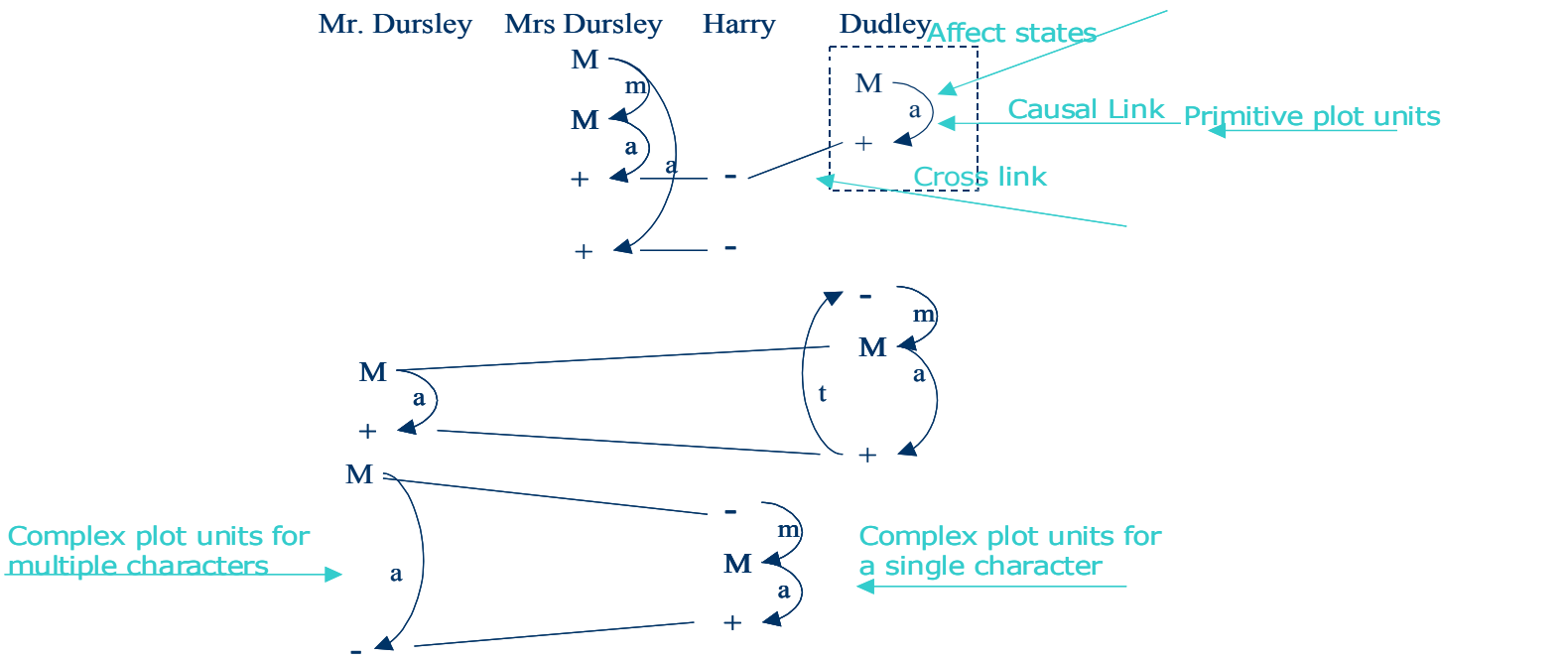


Figure 5. Example Plot Units from Harry Potter and the Philosopher's Stone, Scene 2.

Appendix 1: UML Models of Narrative Based on Chatman (1978)