Generating Multimedia Narrative For Virtual Humans

Steven Wark and Marcin Nowina-Krowicki

Joint and Operations Analysis Division

Defence Science & Technology Group

PO Box 1500, Edinburgh SA 5111, South Australia

{firstname.lastname}@dsto.defence.gov.au

ABSTRACT

In this paper we describe a prototype capability we have developed for creating and presenting multimedia presentations using virtual humans to provide narrative. The multimedia presentations are structured using Rhetorical Structure Theory to describe the narrative relations between each part of the presentation, so that different versions of the presentation can be automatically presented to suit given constraints such as time available or the prior knowledge of the audience. This approach allows reuse of the presentation material while ensuring that these different versions remain coherent. To simplify authoring of these presentations we have developed a web-based presentation editor and preview capability.

Categories and Subject Descriptors

- Information systems~Multimedia content creation
- Information systems~Document structure
- Information systems~Web interfaces
- Information systems~Multimedia and multimodal retrieval
- Human-centered computing~Interaction paradigms
- Human-centered computing~Graphical user interfaces
- Computing methodologies~Natural language processing

Keywords

Multimedia presentations; Virtual humans; Situational awareness.

1. INTRODUCTION

As more and more information becomes available in defence and national information systems through the proliferation of sensor and other data sources, the challenge for analysts and decision makers is not only to find relevant information in a timely manner, but to also integrate it with related information and display it in a contextually appropriate form. Meeting this 'Big Data' challenge will require increasing reliance on automation to assist with the discovery, retrieval, and fusion of contextually relevant information, and also increasing reliance on automation to assemble and present it in an appropriate and coherent manner. One approach to meeting this challenge uses a multimedia narrative paradigm modelled on an interactive news service capability [1]. This paradigm is based around animated virtual human characters [2] that provide narrative, and if appropriate, commentary on multimedia content. This multimedia content can include text, images, videos, graphs, diagrams, 2D/3D animations,

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

Copyright is held by the owner/author(s).

ACSW '16 Multiconference, February 02-05, 2016, Canberra, Australia
ACM 978-1-4503-4042-7/16/02.

<http://dx.doi.org/10.1145/2843043.2843368>

or geospatial scenes that help illustrate the narrative. The animated characters in this case, through character selection and the use of gestural and emotional cues, help establish, maintain, and change context so as to facilitate the users' understanding – they convey 'the story behind the data'. While this can currently be achieved using human narration in a briefing role, this system provides a capability that can potentially be accessed 'on demand' tailored to particular user roles and needs, using the most up-to-date information available.

Figure 1: Multimedia briefing on the fictitious 'Military Strikes in Atlantis' scenario.

Our current work is aimed at making use of human creativity to provide human-authored multimedia content that is then automatically assembled to suit different presentation constraints, such as the time available, the display environment, and audience's information requirements. This capability allows content to be generated once and re-used many times in many different situations, which would be useful for diverse briefing and training roles. This is achieved using an approach based on Rhetorical Structure Theory (RST) [3] to specify the rhetorical relationships between multimedia content within a presentation. Narrative coherence can be maintained as long as every multimedia element within the presentation bears a rhetorical relationship with some other element. Content is then chosen at 'run-time' based on these relationships so as to meet the presentation constraints while maintaining narrative coherence.

One of the challenges with this approach is to provide an intuitive authoring interface that abstracts away the complexity of assigning rhetorical relations to the content, and the rendering system needed to animate the virtual human characters, allowing the author to focus on the desired communicative goal. To achieve this we have developed a prototype web application, using the Meteor JavaScript app platform [4]. This system provides a graphical authoring interface, a media repository, and a presentation preview capability. It allows an author to create and assemble multimedia 'clips' and their RST relationships to other clips within a multimedia presentation, and generate an XML presentation document that can be rendered in different display

environments given different constraints. Our goal is to extend this system to support the automatic assembly of multimedia clips from a corpus to achieve a desired communicative goal.

2. SITUATIONAL AWARENESS

Situational awareness is a key requirement for decision makers and analysts. In the normal course of their roles this is achieved, in part, by exploration and manipulation of the data space in order to produce the products needed to support their analysis and decisions. This helps establish the context, and determine what is known, what is not known, and what is important and what is not important to a particular situation. When automation is introduced to handle large data sets, this pathway to situational awareness is largely lost. However, storytelling is widely regarded as an effective mechanism for experientially engaging an audience [5], and thus can establish the context needed to achieve situational awareness in these circumstances. Beyond this, narration associated with multimedia content can help explain a graphic, animation, or scene, and point out the significance of what is being displayed. Therefore, we have been exploring the use of automated multimedia narrative as a means of providing users with the elements needed to achieve situational awareness when automation is used for data analysis and fusion. In particular, interactive virtual humans have the potential to provide an engaging narrative that goes beyond a simple ‘voice-over’, through the display of appropriate emotions and the use of appropriate gestures. Modelling of the user’s cognitive and emotional state, and that of the virtual human ‘storytellers’, are important to achieving this engagement [6]. With the increasing shift towards autonomous systems with ‘humans on the loop’ rather than ‘humans in the loop’, there will be an increasing need for such engaging interfaces. Automation can introduce a variety of human factors issues, such as automation complacency and automation bias [7]. Automation complacency, where the user becomes disengaged from the task (trusting to the automation) and loses situational awareness, may be mitigated by an engaging and experiential interface such as provided by multimedia narrative. Automation bias, where the user bestows unwarranted trust in the automation, may be mitigated by ‘humanising’ the interface using animated characters so that human fallibility is associated with the information presented [8].

3. VIRTUAL ADVISERS

Virtual human storytellers are one of the key elements for our approach to multimedia narrative. While we have developed a generic multimedia narrative capability that supports a variety of different media and modes and that does not require a virtual human component, we see the involvement of a narrator as key to providing the benefits of multimedia narrative discussed above. We have developed real-time photo-realistic animated characters, known as ‘Virtual Advisers’ for this purpose [9]. Content is provided to Virtual Advisers as marked-up text in the ‘Talking Head Markup Language’ (THML) we have developed, that describes what the Virtual Adviser is to say, and how to say it. This can be generated programmatically, or automatically translated from other communication protocols. The Extensible Natural Gesture Animation Generation Engine (ENGAGE) provides the Virtual Adviser with automatic generation of character behaviour based on sentence syntax [2]. ENGAGE takes care of the intricacies of character animation.

We have used a modular architecture that supports an OpenSceneGraph [10] and a Java rendering engine. The OpenSceneGraph implementation provides a configurable virtual world that can display images, videos, and rolling captions akin to

a virtual news studio. The Virtual Adviser system coordinates the display of multimedia content within this virtual world with the character gestures and utterances. The Java implementation allows Virtual Advisers to be rendered in web applications and web pages with other multimedia content, and coordinated content through techniques such as AJAX.

Figure 2: Architecture used for Virtual Advisers supports different rendering engines.

4. IMMP

Our approach for the generation of multimedia narrative is based on the research done by many researchers over the last few decades on Intelligent Multi-Media Presentation (IMMP) systems [11]. IMMP systems automate the selection, design and presentation of multimedia content. Multimedia generation poses some unique challenges, such as how to tailor and coordinate text and graphics to enhance each other, but there is also much similarity to the challenges posed by natural language generation. Consequently, the techniques used for automated multimedia presentation have drawn heavily on the lessons learned during the development of systems for the automated natural language generation for textual discourse. Generalising the approaches taken with natural language generation, the generation of multimedia presentations has been treated as a goal-directed activity by many researchers [11]. A communicative goal is used to build a multimedia presentation, structured as a hierarchy of communicative acts, each supporting a specific sub-goal. For example, a presenter may point to an illustration or animation while providing a commentary, to achieve a specific sub-goal within the presentation that contributes to the intent (goal) of the whole presentation.

Creation and presentation of multimedia material, or re-using content in another context, can both be considered as communicative acts within different types of multimedia presentations:

1. Multimedia content is generated and used at the same time – for example, when a presenter draws on a blackboard and provides commentary.
2. Multimedia content is created and re-used at a later time by the same person – for example, when someone prepares in advance the content for a presentation.
3. Multimedia content is created and re-used by different people – for example, when someone reuses material from another source.

Thus, the communicative goals underlying the production of the material may be quite different from the goals to be achieved by

presenting it. An IMMP system needs to support these different interpretations.

The standard reference architecture for IMMP systems proposed by Bordegoni et al [12] provides a layered architecture that determines the appropriate content needed to achieve the communicative goals, designs the presentation layout, generates and assembles the multimedia content needed, and coordinates the presentation of that content. In this model, application domain knowledge and data is provided by knowledge services that inform the various layers. Our work takes a simplified approach to these knowledge services by representing available multimedia content, and the contextual relationships within an application domain, as a source document or presentation that is dynamically modified to suit different audience requirements. This allows domain experts to simply generate the source content without any particular expertise (or even awareness) of IMMP techniques.

Alternative approaches to IMMP systems seek to dynamically discover and interpret domain knowledge from available resources. For example, the News at Seven system [13] mines domain knowledge from available online news services, and audience sentiment and commentary, to dynamically generate ‘news stories’ and present them using virtual characters. Unfortunately, such resources are not often available for content in specialised domains (such as military and national security), and other approaches are necessary.

5. RHETORICAL STRUCTURE THEORY

Rhetorical Structure Theory was originally developed to represent the structure of a textual discourse, to guide computer generation of text. It has since been applied to the generation of multimedia presentations [11,14-15], as it provides a general framework for achieving coherence within a presentation by ensuring that every element has a role described by a rhetorical relation. In RST, one multimedia element known as the *satellite* is related to another ‘core’ element, known as the *nucleus*, by a rhetorical relation. A nucleus may have more than one satellite, and *multinuclear* relations also exist where no element is more central than any other. Multimedia elements may themselves be composed of sub-elements that have their own internal relationships, allowing a hierarchy to be formed describing how each part of a presentation is related to the other.

RST includes some 32 rhetorical relations: 25 satellite relations and 7 multinuclear relations. For our initial work, we wanted to simplify the set of relations to facilitate appropriate mark-up of a multimedia presentation by an author, while focusing on those relationships that will support content re-use in different presentation conditions. We thus chose a simplified set of 6 satellite relations:

- *Preparation (P)*: establishes the narrative context.
- *Elaboration (E)*: provides more information.
- *Background (B)*: provides situational context.
- *Conclusion (C)*: provides a restatement and narrative context.
- *Continuation (J)*: represents a relation where multiple elements need to be presented together in sequence. This may also be used to relate multinuclear elements that may be presented simultaneously. This relation effectively subsumes all of the multinuclear RST relations and represents them as a satellite relation.

- *Initialisation (I)*: a specialized case of preparation that *must always* be presented with its nucleus to ensure that the content displayed is independent of whatever has been shown beforehand. This was introduced to support re-use of multimedia content in different presentation situations.

In addition, we have adopted a structural relationship:

- *Nucleus (N)*: identifies that this is the core communicative element of a particular presentation (or clip). This is not an RST relationship per se, but serves to label the edge in the graph that can be constructed to represent the connection to the root node of a presentation (for example, see Figure 3)

RST relations do not, by themselves, necessarily dictate the sequencing of the multimedia elements shown in a presentation. However, for the purposes of simplicity we attached an implicit sequencing of elements based on the RST relations between them:

$$I < P < B < N < J < E < C$$

We anticipated that this may be overly restrictive when dealing with a reduced set of relations, but were curious to explore how suitable this would be for presentation generation. To date this simplification has not been found to be particularly restrictive, in part because the structure is flexible enough to provide simple work-arounds.

An example of a multimedia presentation that can be represented in this way is shown in Figure 3, showing a simplified graph representing a background briefing on a situation from the fictitious ‘Military Strikes in Atlantis’ scenario [16]. In this example, multimedia elements that are always presented together (simultaneously) are combined into a single multimedia node of the diagram. The element labelled ‘a’ represents the core element (*nucleus*) for briefing, while ‘d’ provides some *background* material and ‘e’ provides further *elaboration*. The elements ‘b’ and ‘c’ together provide *preparation* signaling the narrative structure of the presentation. In this case the element ‘b’ is presented first, followed by ‘c’ as the *continuation* that provides a narrative component. Further *elaboration* of elements ‘d’ and ‘g’ are provided by ‘f’ and ‘g’ respectively, and ‘f’ is elaborated still further by ‘h’.

Sequencing of this example based on the partial ordering rules above gives, for the full presentation:

$$b < c < d < f < h < a < e < g,$$

where content is ordered from the leaves of the tree first.

Inspection of this example illustrates how coherence is maintained in the presentation if multimedia elements are removed, provided that a path remains between each remaining element and the nucleus ‘a’.

6. MULTIMEDIA PRESENTATIONS

The current focus of our work is to provide an IMMP capability that can deliver multimedia narrative using animated characters. While a fully automated system is desirable, a semi-automated system that can re-use human-generated content is also of value for our application domain, allowing content re-use in different training and briefing situations. Such a semi-automated system takes advantage of the rich storytelling capabilities and intrinsic knowledge of human authors in how to compose and make use of existing multimedia content. Given this, we have adopted the following roadmap as one development path for an IMMP capability:

Figure 3: Rhetorical relations assigned to multimedia elements within the ‘Military Strikes in Atlantis’ scenario.

1. Re-use of human-authored presentations to suit different audiences, time constraints, and display environments.
2. Re-use of human-authored multimedia content that is automatically assembled into presentations at run-time.
3. Automatic generation and assembly of multimedia content, following defined schemas, at run-time.

Our initial work is aimed at addressing the first of these milestones, while providing a capability that supports the second. We are developing a simple graphical multimedia authoring capability that uses the rhetorical relations described above to annotate the content. A generated multimedia presentation can be exported as an XML document, which can be processed and rendered by an IMMP system as and when desired to suit the current presentation constraints (see Figure 4). The intent is that this editor can also be used to create, save, and retrieve self-contained collections of multimedia elements, or *clips*, which group content on a particular topic, and which can be assembled by an IMMP system and re-used in other presentations.

The previous example shown in Figure 3 represents such a multimedia clip, which describes the background to the ‘Military Strikes in Atlantis’ scenario. This clip could potentially be re-used in a variety of different presentations, with more or less detail displayed depending on the requirements for a particular presentation.

Note that this approach, while it is suitable for multimedia presentations using a virtual human narrator, does not require such a narrator – it is up to the IMMP system how any particular content is rendered.

6.1 Multimedia Presentation Structure

The generated multimedia presentation includes the textual components, references (URI) to other media, and the rhetorical relationships between them, to provide portability and rendering

with different layouts. The requirements identified for the multimedia presentation were:

- Contain information about the content of a presentation without requiring specification of the final presentation layout
- Allow multimedia elements to be collected into re-usable multimedia *clips* that convey particular concepts
- Allow tagging of multimedia elements with rhetorical relations to allow different coherent presentations to be produced under different presentation constraints
- Allow a hierarchy of multimedia elements to be constructed, linked by rhetorical relations, which provides scope for complex rhetorical structures
- Allow multimedia clips to be tagged with semantic information so they can be re-used in other presentations

A number of approaches were considered, including some existing multimedia standards, but none were found to have the features required. For this reason, a bespoke document structure was developed. Much of the rationale behind the design of this structure came from earlier work done to provide a multimedia presentation system [17].

There are two important dimensions defined in the multimedia presentation document: the multimedia content within the presentation and the rhetorical relations between them; and the way it is to be rendered. While in a purely automated system the latter would be determined automatically, for a scheme intended for human authoring it was considered important to allow the author to provide some level of design input, as well as simplifying the task of previewing the presentations created. In the format used, the author specifies an abstract rendering channel for each multimedia element. How this information is ultimately interpreted by an IMMP system determines how this influences the presentation design and realisation.

Figure 4: Concept for a graphical editor for re-usable multimedia presentations

6.2 Content Selection

The multimedia presentation thus described essentially provides the source material for our IMMP system. Content is selected and assembled from this presentation based on three criteria:

1. **topic:** Only those clips associated with a nominated topic (and their associated rhetorical relations) are selected
2. **audience requirements:** By favouring the selection of content related by particular rhetorical relations over others, the presentation can be slanted towards a particular style that suits different audiences. For example, favouring background over elaboration may be appropriate for an audience unfamiliar with the context of the topic presented, while favouring elaboration over background may be more appropriate for an audience familiar with the topic and requiring a more detailed exposition
3. **duration:** The more multimedia elements within a presentation, the longer it will generally take to present. For a time-constrained presentation, multimedia elements may need to be pruned to allow the presentation to fit within a nominated duration. The tree structure adopted in our approach allows content to be pruned from the presentation while retaining coherence. The strategy applied to prune the presentation will determine what rhetorical relations are favoured over others.

There are a number of approaches that can be applied to determine what content should be included to provide a coherent presentation. We chose, at least initially, to test a simple quantitative approach based on assigning different numerical weights, $w_{segment}(R)$ and $w_{clip}(R)$, in the range [0,1] to the rhetorical relations R assigned to a segment or clip respectively. To allow for different potential roles in the presentation structure, different weights are applied depending on whether the rhetorical relations apply to the segments of a clip, or the clips of a sequence. To evaluate how important any particular multimedia element is to the communicative goal of the presentation, a score is assigned by multiplying the weights of the segments traversed

to reach the nucleus of the clip, and the clips traversed to reach the nucleus of the sequence. The greater the score, the greater the assigned importance of the element to the communicative goal of the presentation, and hence its selection priority.

In this scheme, different assignments to the weightings will determine what rhetorical relations are deemed more important to the communicative goal, and the overall structure of the presentation produced. We chose two selection strategies based on the presentation structure initially, although we plan to implement more strategies in the future:

1. **Focussed:** Assigning $w_{clip}(R) < w_{segment}(R)$ for satellite relations, the content of the nucleus of the sequence will be favoured over the content of the satellite clips, tending towards a focus on the core content of the presentation.
2. **Overview:** Assigning $w_{clip}(R) > w_{segment}(R)$ for satellite relations, the nucleus of each clip will always be deemed to be more significant than the satellite segments of any clip. In this case the overall structure of the presentation will be favoured over detailed content within a clip.

In both of these approaches we use:

$$I = w_e(N) = w_e(J) = w_e(I) \geq w_e(P) \geq w_e(C) > w_e(B) > w_e(E) > 0$$

where $e \in \{clip, segment\}$. Different relative weightings can be used to refine these strategies or provide additional selection strategies.

By estimating the time required to render each multimedia segment, we are able to estimate the overall presentation duration for each of these strategies, allowing us to produce presentations within a nominated maximum duration for each of these weighting strategies.

Figure 5 shows examples of presentations generated from the 'Military Strikes in Atlantis' scenario with a maximum duration of 60 seconds, using the focussed and overview weighting strategies respectively.

Figure 5: Example 60 second presentations showing the difference between 'focussed' and 'overview' generation strategies.

The structure of the two presentations generated is clearly different, although features that have been assessed as important in both remain common. It is interesting to note that, without any constraints both of these strategies will generate identical presentations including all of the available content, and that the tightest constraints (shortest duration) will also produce the same presentation – consisting in this case of the nucleus of the Intelligence Update clip only. The strategies chosen will only determine the intermediate states between these two conditions.

7. THE PROTOTYPE SYSTEM

The multimedia presentation structure, and the rhetorical relations represented within it, forms the basis for our initial work on an IMMP system that supports content re-use. This was initially implemented as an XML document, which could be generated manually in a text editor or using any of the available XML editors, to explore the feasibility of this approach. The simple selection process described above was implemented as a groovy script to provide the run-time processing pipeline shown in Figure 4. However, this is hardly a feasible solution for a subject matter expert composing training material or a briefing, so a graphical IMMP editor was developed as a web application to simplify content creation. Some of the features of the IMMP system that support previewing of the generated presentations were also incorporated into this application.

7.1 Meteor

The Meteor web application platform [4] was selected as the framework for the prototype system due to its support for a reactive and flexible, rapid prototyping workflow. After investigating and comparing potential frameworks, Meteor's feature set stood out over the traditional Model-View-Controller based frameworks such as Rails, Grails, web2py and Django. These traditional web frameworks tend to enforce a strict application structure on the developer and require a great deal of boilerplate code to add functionality, such as automated distributed data synchronisation, that is readily available in Meteor.

The key Meteor features that led us to choose it as the web framework for the editor were:

- **Rapid prototyping:** application code changes are automatically pushed to the client while preserving session state, allowing changes to be immediately available in all connected clients.
- **Reactive programming:** provides an intuitive paradigm for User Interfaces (UI). When data changes the UI updates automatically, reacting to the change, without explicit event handling code. This helps avoid many of the traditional 'call-back hell' problems when writing UI code logic. This leads to far simpler UI logic, less coding and in turn fewer bugs.
- **Flexible app structure and workflow:** allows the developer to focus on program logic rather than adhering to the predefined framework structure, as less boilerplate code is required.
- **Automatic data synchronization:** communication between client and server is via data-on-the-wire rather than html-on-the-wire, minimising the bandwidth needed and providing greater presentation flexibility. It supports in-built latency compensation, when there is a lag between client and server, and conflict resolution between multiple clients. Data updates propagate reliably to all affected clients, using push rather than pull semantics.
- **Common JavaScript API on both client and server:** improves programming output by reducing context switching between development languages.

7.2 Architecture

The prototype system uses a distributed architecture consisting of the main IMMP Editor web application interacting with various web services (see: Figure 6). As discussed in the previous section the Meteor web application platform provides the framework for the IMMP Editor web application. For simplicity, presentation media, clips, and data model are currently stored in Meteor's native MongoDB databases and are synchronized between clients and server using Meteor's Distributed Data Protocol (DDP). All data changes in the system reliably propagate to all affected clients.

An IMMP web service is implemented using a REST interface to provide content selection. This service takes an XML

Figure 6: Web Editor Architecture

representation of the multimedia presentation from a client, in this case the editor, and returns a tailored presentation based on the constraints supplied.

The Virtual Adviser web service is used for preview and playback of the presentation. This provides a 3D Applet and AJAX JavaScript API that allows the IMMP Editor to play synchronized real-time presentations narrated by a Virtual Adviser, by controlling and interacting with the page's HTML content, including multimedia such as monitor images and captions.

A PowerPoint conversion service allows existing Microsoft PowerPoint presentation content to be reused. This service converts slides to images and extracts the notes section to provide the narration. Currently the integration between the PowerPoint conversion service and the Web Editor is in its initial stages, but the aim is to provide a system that allows presentation reuse through a semi-automated assignment of the appropriate RST relationships to the presentation content. In this case, slides are mapped to clips with a fixed rhetorical structure between segments, and initially the author would be prompted to assign RST relations between the clips during the import process.

7.3 Features

The Web Editor offers three main modes of operation: a Simple Editor, an Advanced Editor and a Preview/Playback capability.

7.3.1 Simple Editor

The Simple Editor provides a template based approach for quick presentation generation (see Figure 7). Users fill in the blanks in a presentation template and the system generates all the underlying presentation structure. A presentation template specifies the channels available for the presentation. Currently the rendering channels provided by each template are not user editable. For example, the "Simple VA Presentation" template provides a channel for narration, image/video monitor, text caption and classification banner. Future versions of the Simple Editor will allow the author to specify the channels available in a presentation template.

The Simple Editor has been designed for authors that may have some familiarity with presentation programs such as Microsoft's PowerPoint or Apple's Keynote, but no familiarity with Rhetorical Structure Theory or the underlying concepts in our IMMP system. The presentation template provides the user with a predefined presentation structure consisting of placeholders for clips representing the nucleus of the presentation, and satellite clips for potential preparation, background, continuation, elaboration and conclusion. These placeholders appear in the order they will be rendered. The author simply selects a placeholder and fills in the content for the required clips. Each added satellite clip can have satellite relations further specified, allowing for a flexible presentation structure.

The system uses a fixed RST structure between segments in each clip. Each clip consists of an initialisation segment that sets the media (e.g. monitor, caption and banner) for the clip and a nucleus consisting of the narration for the clip. A simpler approach would have been to simply associate a single nucleus segment with each clip, but this scheme gives the author the flexibility to expand the clip structure using the Advanced Editor while ensuring that clip initialisation is preserved.

7.3.2 Advanced Editor

The Advanced Editor provides finer grained control of the creation, editing and management of presentation workflow. The Advanced Editor allows explicit assignment of RST relationships at both the clip and segment level as well as control of other underlying presentation structures and associated attributes. The Advanced Editor provides alternative visualisations of the core presentation structures. These visualisations consist of a:

- **Flow view:** this shows the sequencing of content in each channel and their relative duration (estimated).
- **Cell view:** shows the structure of the presentation by showing what presentation elements are contained within each element.

Figure 7: Simple Editor

- **Tree view:** shows the presentation hierarchy as a tree.
- **Graph view:** shows the graph structure of the presentation and the relationships between elements
- **Storyboard view:** shows the presentation storyboard containing all of the multimedia content and its sequencing.

7.3.3 Preview/Playback

The preview capability allows the author to preview presentations, with the intention to extend this in future to support playback of the presentation to an audience. This is facilitated by an embedded 3D Virtual Adviser that provides narration and HTML5 page control to provide a monitor, caption and banner capability using an AJAX based presentation system. The presentation system allows the Virtual Adviser and page components to interact in a synchronised fashion. The user can control the desired presentation length and select the content selection strategy (e.g. focused or overview). The IMMP Web Service is used to generate a presentation based on the user specified constraints. The presentation is then delivered in real-time in the user's browser with synchronized narration, monitor and captioning. Currently this capability is constrained to a fixed layout for the rendering channels. Future versions will allow the user to specify the channel layout and rendering options to suit different audience preferences.

8. DISCUSSION

The prototype capability we have developed provides a modular framework that allows us to explore the issues related to development of a fully automated multimedia presentation system, and provide some useful capability along the way. As

such, there is scope for further work in each of the components developed, guided by user evaluation studies and feedback. We will discuss some of the issues and potential next steps in the following sections.

8.1 IMMP development

The use of rhetorical relations to link existing multimedia content, whether it is generated clips or segments, or more primitive multimedia elements, is a powerful method of retaining coherence when generating a multimedia presentation. This approach not only facilitates re-use of existing content, but also provides a template that could be used to guide automatic generation of content and/or narrative – which could indeed be re-used by a human presenter. As such, further refinement of our approach is planned. In particular, as discussed earlier, we plan to refine the existing weighting strategies, and add additional ones to suit emerging audience requirements. We have several degrees of freedom available in how to approach this: do we tune the weightings applied to the rhetorical relations between clips, or between the content within a clip? We believe that this is best answered by exploring how authors structure content using the IMMP editor. For example, a heavy reliance on the Simple Editor mode would make the weightings on the relationships between clips more important. On the other hand, a heavy reliance on re-use of existing multimedia clips would suggest that the weightings of the relationships between segments within a clip are more important.

We have deliberately simplified the set of RST relations we use in order to simplify the process of tagging content with these relations. It may prove necessary to extend this set in future to meet emerging presentation requirements. Two examples that see common usage in written works are:

- **adjunct:** secondary content, not essential to the primary communicative goal, such as in a printed sidebar. This could potentially be treated as an additional satellite relationship with lower weighting and precedence than elaboration.
- **summary:** a short restatement of previously presented content. This may apply, for example, as an alternative to a chain of elaborations, but not to any individual elaboration on its own. This case is not amenable to the simple weighting scheme applied so far, but would require additional rules to identify when it needs to be included. A summary provided as a recap of previously presented material could otherwise be provided as a conclusion.

Furthermore, as we move away from human generated content we could expect to be able to make use of the full set of RST relations, automatically generated to guide content selection.

Another potential refinement of our approach relates to how content is managed within a multimedia segment. Additional tags could be applied to the multimedia elements to identify which, for example, is critical to the communicative goal of the segment, or which is a re-statement to reinforce the point being made. This would provide cues to the IMMP system that could, for example, use the virtual human narrator to direct the attention of the audience to a particular element, or decide whether a segment should be presented at all (and hence any satellites of this segment) if an appropriate rendering channel is not available in the current display environment for that element.

Finally, while this approach uses the rhetorical relationships to maintain coherence in the generated presentations, it relies on the creativity of the human author to ensure that the content is presented in an interesting way – i.e. it leaves the storytelling art to the author. This means that, unless care is taken, a generated presentation may be coherent but not interesting (despite the best efforts of the virtual human narrator). One possible approach to address this may be to use a formal semantics for narrative to tag the clips in the presentation, and use a formal model for narrative generation (see, for example, Bui et al [18]) to ensure that the appropriate narrative elements are used to provide an engaging presentation. This may be too much to expect from a novice content author, but may be appropriate for an automated or semi-automated generation system. In this case, a more complete set of RST relations, which includes relations such as *evidence* and *contrast*, could help guide the selection of content.

8.2 Editor Development

Currently the system is heavily user driven allowing authors to create, edit, delete and preview presentations. Presentation templates provide the main mechanism of abstracting the underlying system from the author. In the next phase of development we plan to start semi-automating parts of the workflow. Intelligent hints will be used to guide the author's RST assignment. For example, the system may suggest to the author that the current clip looks like a summary of another clip. We also want to improve the re-use provided by the system at both the clip and multimedia (e.g. images, videos, captions, etc.) level. Topic-based retrieval and re-use of clips will allow the selection of clips based on the current topic and ontology.

An important capability for situational awareness in many domains is the ability to display geospatial information and relationships. We currently achieve this using a 3D geospatial display system developed using OpenSceneGraph [10]. To support a web-based playback capability future work will integrate the Cesium WebGL Virtual Globe and Map Engine [19]

with the IMMP Editor Preview/Playback capability, to provide a synchronized display for rendering geospatial channels.

The current web-based implementation for the Virtual Adviser as a Java plugin is being deprecated. A WebGL implementation of the Virtual Adviser is being explored as a replacement. This approach provides native browser support and allows compositing of the 3D adviser with HTML elements. This enables new rendering possibilities, for example: using HTML HUD style overlays on top of the adviser or using HTML components in the scene background.

Future work will explore technologies for semi-automating content construction and tagging of presentation elements. The final phase will see the system move from a semi-automated system to an automated Intelligent Multimedia Presentation System.

8.3 Multimedia Web Services

The IMMP Editor currently stores media, clips, and presentations within Meteor's native database. Future work will look into how we can store this within a graph database to support more complex queries of this content based on the relationships between them. Additionally, it would be useful to provide a web service interface to this media repository so that other applications could discover and use its content.

We currently rely on the authors to tag media as it is added to the IMMP Editor so that it can be re-used. A more robust approach could be to use automated techniques to classify and tag the media instead. It is important to ensure that media is not re-used inappropriately, and hence provide misleading or confusing information. In the current system this depends on the judgement of the presentation author. For an automated or semi-automated system, this needs to be handled in some other way. One approach to resolving this could be to ensure that any captions associated with (or generated for) a graphic, image or video, which provide context, are included with any material generated from it. In this case, automatic revision of the caption may be necessary to ensure that it correctly relates the material to the current narrative context.

8.4 Automated News Services

Our ultimate goal is to produce an automated news service that can provide up-to-the-minute situational awareness on demand. We do not see this as precluding human generated content in specialised 'expert' domains in the foreseeable future, and so the techniques we have described here will still have a role when automation can be used routinely for other content. As the capabilities of automated systems develop towards this goal, we see that there will still be a role for the IMMP Editor described here to allow human editorial oversight of the generated content.

9. CONCLUSION

In this paper we have described work done towards providing a multimedia presentation capability using virtual human characters to provide the narration, which allows dynamic re-use of content to suit different presentation and audience requirements. This work provides initial steps towards providing an automated multimedia narrative capability suitable for comprehension of complex and dynamic 'big data', and interaction with autonomous systems.

Our approach uses a simplified set of rhetorical relations from Rhetorical Structure Theory (RST) to link multimedia content into re-useable multimedia clips, which can be similarly linked to produce a multimedia presentation. The RST relations are

analysed by an Intelligent Multi-Media Presentation (IMMP) system to select appropriate content to provide a presentation tailored to suit a particular audience and time constraints, which still retains coherence. A simple weighting scheme for our simplified set of rhetorical relations has proven surprising effective for this purpose.

The exploitation of human storytelling capabilities is a key element to achieving multimedia narrative with the current system. A web-based IMMP Editor has been developed using the Meteor web framework to allow a human author to assemble multimedia clips into an engaging narrative. The RST relations provide cues that help preserve the narrative structure of the presentation, although it relies on the author to structure the presentation to facilitate this. A variety of visualisations are provided to the author to assist them with this. The IMMP Editor also includes a simple preview/playback capability to support presentation authoring.

Development of the IMMP Editor is ongoing, to include features to increase its flexibility and provide aids to simplify content development and re-use. Further refinement of the approach to content selection is anticipated to support a wider span of audience requirements. Finally, we anticipate that this approach will provide useful insights towards development of an automated multimedia narrative capability that can automatically discover, create, assemble, and deliver content through engaging multimedia storytelling.

10. ACKNOWLEDGMENTS

Our thanks to Dr Ian Dall and Dr Jason Scholz for encouraging this work, and the Chief Joint and Operations Analysis Division for supporting it.

11. REFERENCES

- [1] Wark, S., and Lambert, D. A. 2007. Presenting the Story Behind the Data: Enhancing Situational Awareness using Multimedia Narrative. In *Proceedings of the Military Communications Conference (MILCOM) 2007*. (Orlando, Florida, 29-31 October 2007), IEEE, DOI=10.1109/MILCOM.2007.4455058
- [2] Nowina-Krowicki, M., Zschorn, A., Pilling, M., and Wark, S. 2011, ENGAGE: Automated Gestures for Animated Characters. In *Proceedings of the Australasian Language Technology Association Workshop (ALTA) 2011*, (Canberra, Australia, December 2011), 166-174
- [3] Mann, W. C., and Thompson, S. A. 1988. Rhetorical Structure Theory: Toward a Functional Theory of Text Organization, *Text*, 8 (3), 243-281
- [4] Coleman, T., and Greif, S. 2015. *Discover Meteor: Building Real-Time Javascript Web Apps*. (Version 1.9 updated 15 April 2015), <https://www.discovermeteor.com/>
- [5] Hsu, J. 2008. The Secrets Of Storytelling: Our Love For Telling Tales Reveals The Workings Of The Mind, *Scientific American Mind*, 19 (4), Aug/Sep 2008 46-51
- [6] Marsella, S., and Gratch, J. 2014. Computationally Modeling Human Emotion, *Communications of the ACM*, 57 (12), Dec 2014, 56-67, DOI=10.1145/2631912
- [7] Carr, N. 2014. *The Glass Cage: Where Automation Is Taking Us*. The Bodley Head, London, 2014.
- [8] Smith, G., Dry, M., and Lee, M. 2005. Trust and Confidence: Presenting Information Over a Network Using Text Or A Speaking Head, *TTCP HUM TP7 Workshop on Aerospace Human Factors Issues in Network-Centric Warfare* (Salisbury, UK 2005)
- [9] Lambert, D.A. 1999. Advisers With Attitude for Situation Awareness, *Proceedings of the 10th International Command and Control Research and Technology Symposium (ICCRTS) 1999* (La Salle IL, USA)
- [10] OpenSceneGraph Project Website: <http://www.openscenegraph.org/>
- [11] Andre, E. 2000. The Generation of Multimedia Presentations. in Dale, R., Moist, H. and Somers, H. eds. *Handbook of Natural Language Processing*, Marcel Dekker Inc., New York NY, 314-338
- [12] Bordegoni, M., Faconti, G., Feiner, S., Maybury, M.T., Rist, T., Ruggieri, S., Trahanias, P., and Wilson, M. 1997, A Standard Reference Model For Intelligent Multimedia Presentation Systems, *Computer Standards & Interfaces*, 18, 477-496.
- [13] Nichols, N., and Hammond, K., 2009, Machine-Generated Multimedia Content, *Proceedings of the 2nd International Conferences on Advances in Computer-Human Interactions (ACHI) 2009* (Cancun, Mexico)
- [14] Paris, C., Colineau, N., and Estival, D. 2004, Intelligent Multi Media Presentation Of Information In A Semi-immersive Command and Control Environment, in *Proceedings of the Australasian Language Technology Workshop (ALTA) 2004* (Sydney Australia, December 2004)
- [15] Andrews, T., Broughton M., and Estival D. 2006. Implementing An Intelligent Multimedia Presentation Planner Using An Agent Based Architecture, in *Proceedings of the International Conference on Intelligent User Interfaces* (Sydney Australia, January 2006)
- [16] Lambert, D. A., Wark, S., Bossé, É., Pigeon, L., Blackman, C. and Hinman, M. 2012. A Coalition Approach to High-Level Information Fusion, in Blasch, E., Bossé, É. and Lambert, D. A. eds, *High-Level Information Fusion Management and Systems Design*, Artech House, Boston MA, 251-276
- [17] Wark, S., Zschorn, A., Broughton, M., and Lambert, D. 2004. FOCAL: A Collaborative Multimodal Multimedia Display Environment. In *Proceedings of SimTecT 2004 - Simulation - Better Than Reality?* (Canberra Australia, May 2004)
- [18] Bui, V., Abbass, H., and Bender, A. 2010. Evolving Stories: Grammar Evolution For Automatic Plot Generation, in *Proceedings of the 2010 IEEE Congress on Evolutionary Computing (CEC)*, (Barcelona, July 2010), DOI=10.1109/CEC.2010.5585934
- [19] Cesium – WebGL Virtual Globe and Map Engine website: <http://cesiumjs.org/>