Interactive Storytelling via Video Content Recombination

Julie Porteous
Teesside University, UK
j.porteous@tees.ac.uk

Sergio Benini
University of Brescia, Italy
sergio.benini@ing.unibs.it

Luca Canini
University of Brescia, Italy
luca.canini@ing.unibs.it

Fred Charles
Teesside University, UK
f.charles@tees.ac.uk

Marc Cavazza
Teesside University, UK
m.o.cavazza@tees.ac.uk

Riccardo Leonardi
University of Brescia, Italy
riccardo.leonardi@ing.unibs.it

ABSTRACT

In the paper we present a prototype of video-based storytelling that is able to generate multiple story variants from a baseline video. The video content for the system is generated by an adaptation of forefront video summarisation techniques that decompose the video into a number of Logical Story Units (LSU) representing sequences of contiguous and interconnected shots sharing a common semantic thread. Alternative storylines are generated using AI Planning techniques and these are used to direct the combination of elementary LSU for output. We report early results from experiments with the prototype in which the reordering of video shots on the basis of their high-level semantics produces trailers giving the illusion of different storylines.

Categories and Subject Descriptors

H.5.1 [**Multimedia Information Systems**]: Video (e.g., tape, disk, DVI)

General Terms

Algorithms

Keywords

Video Summarisation, LSU, Interactive Storytelling

1. INTRODUCTION

Interactivity and content adaptation are two of the most important trends for the development of new media. Until now, the former has been mostly associated with graphics-based media such as computer games, whilst multimedia research has embraced the challenge of video content personalisation and adaptation. Early research in interactive movies led to widespread feeling that the video medium was incompatible with the sophisticated combinatorics of story generation required to fully implement the concept, because of the inability to generate video content in real-time, unlike the real-time generation of computer graphics. For that reason, most work in Interactive Storytelling (IS) is now based on the generation of 3D graphics [7, 10], and IS is conceived of as the future of computer games rather than film.

In this paper, we seek to challenge the status quo and re-introduce video as the medium for IS. Our working hypothesis is that an interactive video-based storytelling system, realised using a combination of state-of-the-art video summarisation and interactive storytelling techniques, will be able to generate output videos that display the degree of variability that has till now only been achieved with real-time generated graphics. To test this hypothesis we have developed a prototype video-based storytelling system which uses Michael Radford's 2004 screen adaptation of Shakespeare's Merchant of Venice [9]. During development of the prototype the video content is identified via automatic LSU identification (discussed in section 2) and mapping of LSU to narrative actions (section 3). This video content is then used in the real-time system, in which different narrative variants are generated (section 4), and accompanying video content is re-combined and output to the user (section 5).

2. LSU IDENTIFICATION

A video can be segmented into a hierarchy of partitions. At the highest level, a video can be completely and disjointly segmented into a sequence of scenes, where each scene conveys a high-level concept of a story, which we refer to as Logical Story Units (LSU) [5]. On a lower level, scenes can be segmented into a sequence of basic video segments named shots, which are the longest continuous frame sequences from a single camera take. Shots sharing common perceptual low-level characteristics can be clustered together into higher entities called groups (or clusters) of shots. Finally, at the lowest level of the hierarchy, one or more key-frames can be extracted from shots as static significant examples of the shots visual content.

The first step of the segmentation process used in this work is the extraction of the shots from the film [3]. For each shot a key-frame is extracted, divided into squared blocks and analysed in the LUV colour space to feed a Tree-Structured Vector Quantization algorithm [4] which outputs a code-book, i.e. a pool of representative colours used as a signature for the corresponding key-frame.

Then shots are grouped using a hierarchical clustering algorithm which computes a measure of similarity between shots on the basis of the code-books extracted. At the beginning of the hierarchical clustering each shot belongs to a different cluster, then the algorithm iteratively merges the

MM'10, October 25–29, 2010, Firenze, Italy.

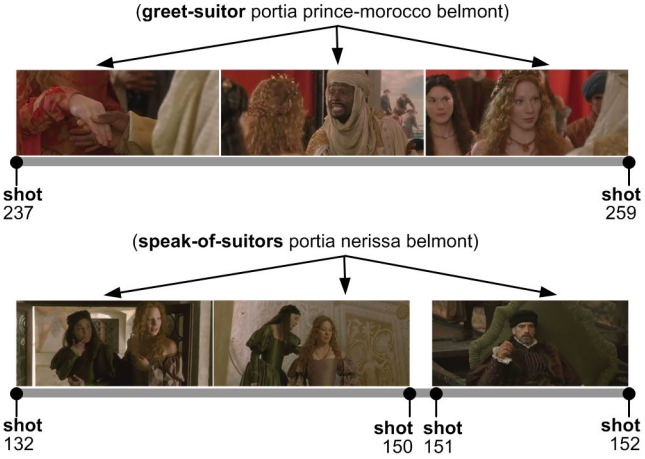

Figure 1: Mapping LSU to Narrative Actions.

two most similar clusters, where similarity between clusters is given by the average similarity between all shots belonging to the two clusters. The final hierarchical grouping is obtained by exploiting the properties of the associated dendrogram tree and the codebook distortions related to different branches of the dendrogram (see [1] for details).

Finally, LSU segmentation is computed on the basis of the obtained shot grouping, by reconstructing temporal relations between adjacent shots. Following [12], we represent the video using a Scene Transition Graph (STG), where nodes are clusters of visually similar and temporally close shots, and edges represent transitions between subsequent shots. After the removal of cut-edges, each well connected subgraph represents an LSU. Cut-edges constitute reliable LSU boundaries since they are one-way transitions from one set of highly connected clusters to another set that feature completely new visual-content.

3. MAP LSU TO NARRATIVE ACTIONS

Our video-based storytelling prototype uses an AI Planning approach to narrative generation that requires a model of the story domain as input. This model includes key actions of the main characters along with character attributes represented as predicates and is specified during a domain analysis phase. As part of the domain analysis the automatically identified LSU are mapped to high level concepts corresponding to narrative actions. These mappings are then used at run-time to identify appropriate LSU and enable the collating of required video content for output.

To map LSU to narrative actions we identified segments in the film where narrative actions occur and then associated with each segment a single LSU that best overlapped the concept expressed by the corresponding narrative action. For our initial analysis of the Merchant of Venice we identified 21 narrative actions (with analysis at approximately the level of scenes in the original play). For 14 of these narrative actions the correspondence with a single LSU was almost perfect, while for the remaining 7 it was still possible to identify a representative LSU that captured the crucial part of the action.

Some sample LSU mappings are shown in figure 1. At the top of the figure there is an exact correspondence between the LSU (shot 237 to 259) and the narrative action, where Portia greets her suitor the Prince of Morocco. However, in the example at the bottom of the figure there is over 90% correspondence between the LSU and the narrative action where Portia speaks of her suitors: the segment from shot 132 to 150 depicts this topic but the final 10% of the LSU (shot 151 to 152) isn't as relevant since it shows Antonio.

4. NARRATIVE GENERATION

Generativity is central to IS since it underlies all forms of story variation. We have adopted an AI planning approach to narrative generation since previous work has demonstrated that the use of planning is sufficient to enable the generation of different story variants. Within this approach the narrative generator is an AI planner that is input a description of an initial state of the story world, a description of the desired goal state and a set of actions that change the state of the story world; and the narrative generation "problem" is to generate a sequence of actions that lead from the initial state to a state in which the goal is true (the story variant).

The narrative generator in our video-based storytelling prototype is an AI planner we have developed that is targeted at the requirements of narrative generation for IS [8]. It also includes features directed at further enhancing the power of a planning based approach to generate story variants. One feature is the use of character Point of View, a concept we introduced in [8], to describes a character's perspective on an overall plot through which a story can be told. It is an important concept that can be a source of variation which also preserves genre and "semantic" consistency by generating narrative variants that don't disrupt the story genre. Another feature is the inclusion of "wildcard" actions in the domain model: actions that can result in different contextual interpretations depending on their relative position within a causally chained sequence of actions. In the context of video-based storytelling this means that these wildcards can be used to introduce different semantics to output narratives and hence contribute to the generation of different story variants. This is more than an editing or presentation effect since the automatic re-ordering changes the semantics and contents of subsequent portions of the narrative.

An example wildcard, shown highlighted in figure 2, is *(enjoy-lavish-lifestyle bassanio venice)*. Our system is able to automatically generate variants in which this wildcard appears both *after* and *before* Bassanio has secured Antonio's financial help, sequence (a) and (b) respectively. Different contextual interpretations result from the wildcard position: in (a) Bassanio prepares for his pursuit of Portia in a lavish style that is in keeping with his status; whereas in sequence (b) Bassanio is shown as a profligate, squandering money and exploiting his friends affection for financial gain.

5. VIDEO CONTENT RECOMBINATION

This part of the system handles the recombination of suitable segments of video content from the LSU corresponding to each narrative action for display to the user. Rather than merely outputting segments of video for the LSU corresponding to the narrative action, the approach taken is to produce an informative skim of the LSU's in an innovative approach to generation of video content on-the-fly.

Our approach is to model each LSU using a Hidden Markov

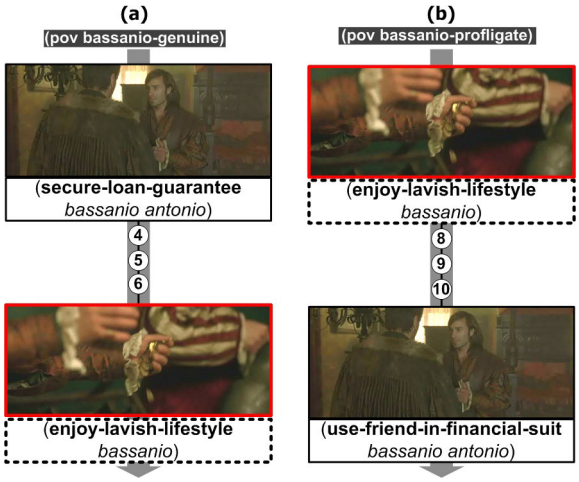

Figure 2: Wildcard position changes contextual interpretations: (a) Bassanio prepares to pursue Portia in lavish style; (b) Bassanio is a profligate who squanders money and financially exploits Antonio.

Model (HMM) [2]. In the model the HMM states representing concepts correspond to distinct clusters of visually similar shots, the state transition probability distribution captures the shot pattern structure of the LSU, and shots constitute the observation set. The use of HMM's is important because it provides a means to generate different informative skims of a given LSU. The method is: for each LSU, a skimmed version can be generated as an observation sequence of the associated HMM, that is: $O_1 O_2 \cdots$, where each observation O_j is one of the symbols of the observation set: a shot of the original video. This is important because the generation of LSU skims provides a mechanism to introduce variation, whilst at the same time preserving genre and "semantic" consistency in terms of structure informativeness.

This is an innovative approach to the generation of narrative variations. Given a story variant obtained by recombination of narrative actions mapped on LSU, a HMM can be built on each LSU, resulting in the possibility of having a reinforcement of a particular concept or aspect of the story (e.g. selection of observation shots for states of the model could be driven by specific paradigms such as the presence of a character or level of character motion).

6. EVALUATION

The first part of the evaluation is to assess the quality of the output videos generated by our prototype. For this we compare "LSU-based PoV" video where content corresponds to actions in narrative variants produced by the generator with "ground-truth PoV" video where the content is the video that would be output if the mapping between narrative actions and LSU was perfect. The quality of the mapping between LSU and narrative actions is evaluated using the criteria for LSU evaluation in [11], with slight modifications to the measures of *coverage* and *overflow*. For comparison of LSU-based PoV video and ground-truth PoV video, the measures are defined as follows: *coverage* $\mathcal{C} \in [0,1]$ is the fraction of shots in the LSU-based PoV video that overlap with the ground-truth PoV video; and *overflow* $\mathcal{O} \in [0,1]$ is the fraction of the LSU-based PoV video overlapping shots that do not belong to the ground-truth PoV video.

Our experiments with LSU-based PoV video generation were carried out using the PoV for three different characters: Shylock, Antonio and Portia. The measured values of *coverage* and *overflow* are shown in the table below, where the low values of *overflow* and the high scores of *coverage* reveal the good performance of the proposed mapping mechanism on the narrative variants.

Point of View	Shylock	Antonio	Portia
Coverage	0.708	0.693	0.702
Overflow	0.023	0.001	0.000

The second part of the evaluation is to assess the ability of the system to generate different video variants from a single baseline video. To illustrate this, figure 3 shows two very different narrative variants, one generated from the PoV of Portia a loyal daughter and the other from a PoV of Bassanio an honest friend and suitor. For Portia, this variant focuses solely on one sub-plot of the play, the "caskets" sub-plot, while the variant for Bassanio also introduces elements of the "pound-of-flesh" sub-plot (following the analysis in [6]). In the video from Portia's PoV the initial actions follow the caskets sub-plot up to the arrival of Bassanio but in this variant Portia doesn't follow Bassanio when he returns to Venice and culminates in Portia despairing over her father's will and rejection by Bassanio. For Bassanio's PoV the initial phase of the narrative centres on his involvement in the pound-of-flesh sub-plot and the bond between Antonio and Shylock. It isn't until after Bassanio has won Portia's hand in marriage that he returns to Venice with the remainder of the narrative following the order of events as they unfold in the play: once Antonio is freed by the court the exchange of rings triggers the "rings" sub-plot which is only resolved at the end of the play when Bassanio and Portia are re-united.

This re-combination around a baseline plot makes it possible to generate very different variants that give different meaning to aspects of the plot and open the possibility of different story lines. The generated videos are qualitatively different, much more than could be achieved via editing.

7. CONCLUSIONS

We have shown that video recombination techniques could replace computer graphics in the dynamic generation of narratives and have proposed a complete framework to achieve it. Whilst the generative potential of graphics-based interactive storytelling systems may remain higher, this has to be moderated by the fact that most graphics-based systems do not generate individual characters' realistic behaviour and instead rely on pre-defined animation sequences for narrative actions, making it de facto closer in philosophy to our approach. If we also consider that dynamically generated graphics are of a lesser quality than both offline-rendered animations and video, the potential for a video-based approach becomes convincing. The level of combinatorics supported by our approach exceeds by at least one order of magnitude what can be achieved with tree-based video branching, whilst also being more economical in terms of content production and more sophisticated in the underlying narrative backbone (planning can enforce causality over the entire action sequence, not just locally).

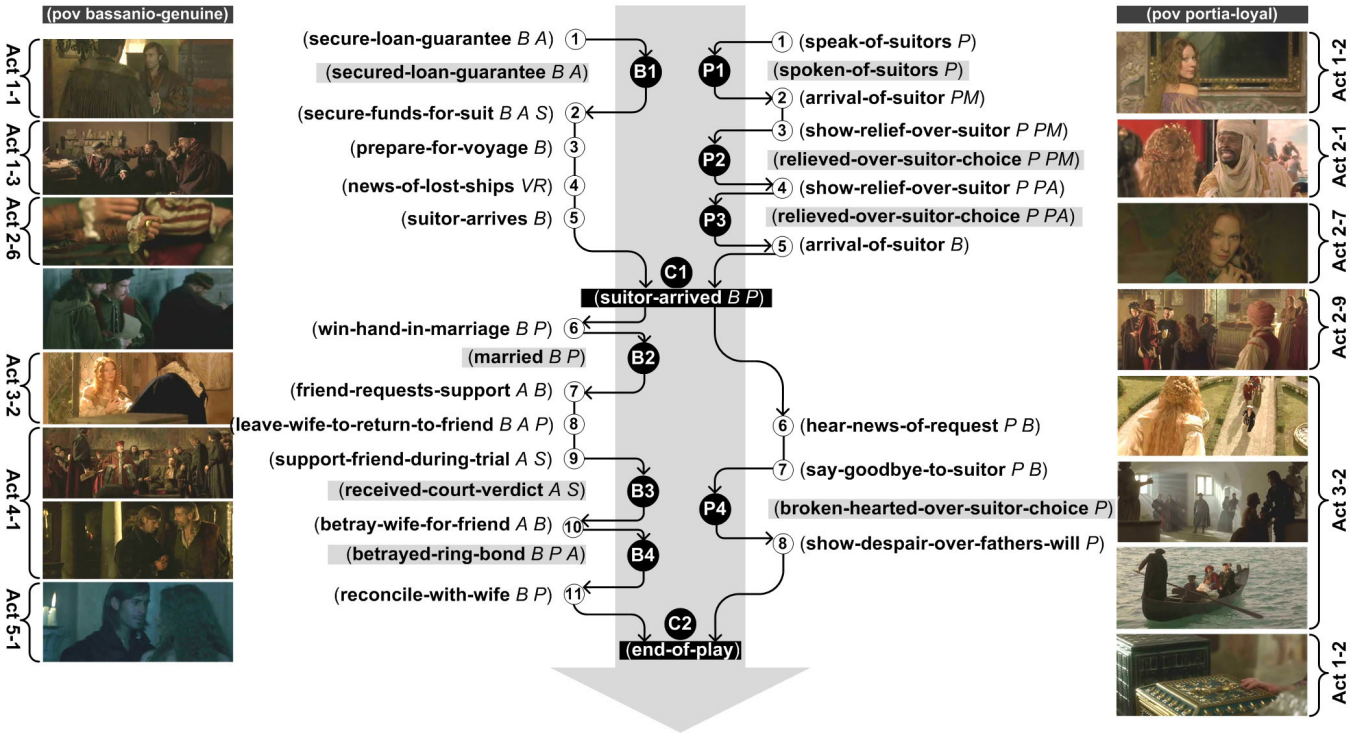

Figure 3: Comparison of video variants for Bassanio and Portia. For Bassanio this variant ends happily after resolution of the key story sub-plots. However, Portia's variant ends unresolved in sadness and despair.

A central contribution of this work has been to show that narrative causality can be manipulated by non-linear video recombination, unlike most adaptive video techniques that preserve the linear ordering of scenes (they mostly operate through scene deletions). This has been achieved by unifying the representational philosophy of video-based LSU with the narrative action formalisation of planning operator. Our current approach uses soundtrack only to determine the correspondence between original narrative actions and the film. In future we intend to explore how best to integrate soundtrack into generated videos where soundtrack fit is poor (e.g. when wildcards have been inserted out of sequence). One approach is to use a planner capable of reasoning about durative actions, to reason about soundtrack and generate "voice over" where required (a technique employed by Radford [9]).

8. ACKNOWLEDGMENTS

This work has been funded (in part) by the European Commission under grant agreement IRIS (FP7-ICT-231824).

9. REFERENCES

[1] S. Benini, P. Migliorati, and R. Leonardi. Hierarchical structuring of video previews by leading cluster analysis. *Signal, Image and Video Processing*, 2010.

[2] S. Benini, P. Migliorati, and R. Leonardi. Statistical skimming of feature films. *International Journal of Digital Multimedia Broadcasting*, 2010.

[3] J. S. Boreczky and L. A. Rowe. Comparison of video shot boundary detection techniques. In *Storage and Retrieval for Still Image and Video Databases IV*, Los Angeles, California, January 1996.

[4] A. Gersho and R. M. Gray. *Vector Quantization and Signal Compression*. Kluwer Academic, 1992.

[5] A. Hanjalic, R. L. Lagendijk, and J. Biemond. Automated high-level movie segmentation for advanced video retrieval systems. *IEEE Trans. on CSVT*, 9(4), 1999.

[6] J. L. Hinely. Bond Priorities in The Merchant of Venice. *Studies in English Literature, 1500-1900*, 20(2):217–239, 1980.

[7] M. Mateas and A. Stern. Structuring Content in the Façade Interactive Drama Architecture. In *Proc. of the 1st Conf. on Artificial Intelligence and Interactive Digital Entertainment (AIIDE-05)*, 2005.

[8] J. Porteous, M. Cavazza, and F. Charles. Narrative generation through characters' point of view. In *Proc. of 9th Int. Conf. on Autonomous Agents and MultiAgent Systems (AAMAS 2010)*, 2010.

[9] M. Radford. The Merchant of Venice. MGM, 2004. Copyrighted images reproduced under "fair use" policy.

[10] M. Riedl and A. Stern. Believable Agents and Intelligent Story Adaptation for Interactive Storytelling. In *Proc. 3rd Int. Conf. on Technologies for Interactive Digital Entertainment (TIDSE)*, 2006.

[11] J. Vendrig and M. Worring. Systematic evaluation of logical story unit segmentation. *IEEE Trans. on Multimedia*, 4(4):492–499, December 2002.

[12] M. M. Yeung and B.-L. Yeo. Time-constrained clustering for segmentation of video into story units. In *Proc. of Int. Conf. on Pattern Recognition*, 1996.