Interactive Film Recombination

FABRIZIO GUERRINI, NICOLA ADAMI, SERGIO BENINI, and ALBERTO PIACENZA,
Department of Information Engineering, University of Brescia
JULIE PORTEOUS, School of Computing, Teesside University Digital Futures Institute
MARC CAVAZZA, School of Engineering and Digital Arts, University of Kent
RICCARDO LEONARDI, Department of Information Engineering, University of Brescia

In this article, we discuss an innovative media entertainment application called Interactive Movietelling. As an offspring of Interactive Storytelling applied to movies, we propose to integrate narrative generation through artificial intelligence (AI) planning with video processing and modeling to construct filmic variants starting from the baseline content. The integration is possible thanks to content description using semantic attributes pertaining to intermediate-level concepts shared between video processing and planning levels. The output is a recombination of segments taken from the input movie performed so as to convey an alternative plot. User tests on the prototype proved how promising Interactive Movietelling might be, even if it was designed at a proof of concept level. Possible improvements that are suggested here lead to many challenging research issues.

CCS Concepts: • **Computing methodologies** → *Scene understanding*;

Additional Key Words and Phrases: Interactive storytelling, logical story unit, semantic description, Markov chains, narrative modeling

ACM Reference format:
Fabrizio Guerrini, Nicola Adami, Sergio Benini, Alberto Piacenza, Julie Porteous, Marc Cavazza, and Riccardo Leonardi. 2017. Interactive Film Recombination. *ACM Trans. Multimedia Comput. Commun. Appl.* 13, 4, Article 52 (August 2017), 22 pages.
https://doi.org/10.1145/3103241

1 INTRODUCTION

This article deals with a recent cutting-edge research direction in the broader field of Interactive Storytelling [13], one of the foremost technologies currently being researched and developed for the creation and diffusion of new media entertainment systems, by applying it to movies—an application referred to as Interactive Movietelling. A well-established field, Interactive Storytelling aims at creating new media content for the presentation of a narrative, in which the evolution of

Authors' addresses: F. Guerrini, N. Adami, S. Benini, A. Piacenza, and R. Leonardi, Department of Information Engineering, University of Brescia, Via Branze 38, 25123 Brescia, Italy; emails: {fabrizio.guerrini, nicola.adami, sergio.benini, alberto.piacenza, riccardo.leonardi}@unibs.it; J. Porteous, School of Computing, Teesside University Digital Futures Institute, Campus Heart, Southfield Rd, Tees Valley, Middlesbrough TS1 3BX, UK; email: j.porteous@tees.ac.uk; M. Cavazza, School of Engineering and Digital Arts, University of Kent, Jennison Building, Canterbury, Kent, CT2 7NT, UK; email: m.o.cavazza@kent.ac.uk.

https://doi.org/10.1145/3103241

the story is made dynamic, that is to say it can be modified and/or influenced by the user in real time. Interactive Storytelling specifically refers to the ability to change the story, underpinning the content independently from the visual medium used to present the narrative, which could range from text, audio, video, all the way up to computer graphics and virtual reality rendering systems. As it is easily discernible, Interactive Storytelling is a strongly interactive discipline, bringing together humanities (psychology, drama theory, etc.) and many technical fields such as computer science and multimedia signal processing. The latter includes, among the others, some form of automated reasoning mechanism (e.g., artificial intelligence) for the narrative generation engine, which is considered a key enabler for Interactive Storytelling, human-computer interaction, to allow the user to intervene in a variety of ways (that could be either direct like keyboard inputs, menu selections, and speech commands, or indirect like physiological inputs), content analysis, computer vision, and computer graphics (if new content needs to be created and rendered).

As such, Interactive Storytelling and its ramifications represent a potential revolution in the way media entertainment is experienced by users because it brings interactivity into current and future generation digital media content, as well as into more established mediums. Ultimately representing the bridge between computer games (in which engaging, open-ended plots are acquiring more and more importance from the user's perspective for the commercial success of the overall product) and traditional narrative experiences such as movies (where the narrative quality of the scripted content is, in the end, their most important asset and, unlike games, do not require any interaction from a user to progress), it should not come as a surprise that Interactive Storytelling is attracting huge interest from both traditional broadcasters and computer game producers. Following a number of academic projects in the mid to late 1990s [24, 32], an ever-growing number of research efforts have been dedicated to this endeavor. In particular, in Europe, it culminated in the Integrated Research in Interactive Storytelling (IRIS) Network of Excellence project [17].

The following section recaps the pertinent literature on the subject of this article. Then, Section 3 describes the proposed architecture of a whole system, without delving into technical details, to give a flavor of the motivation underlying each design choice. After that, the prototype implementation of the system is analysed in Section 4, showing how it reflects the proposed workflow. Technical details on that particular implementation are postponed till Section 5. Experimental results in the form of user tests and analysis of the video recombination process performance are reported in Section 6, and conclusions are drawn in Section 7.

2 RELATED WORK

Interactive Storytelling was initially considered widely incompatible with the video medium because of its inherent inability to generate new content on the fly, and thus to fully leverage state-of-the-art narrative engines' powerful combinatorics. In fact, early attempts to develop interactive movies relied on branching narratives [12], following the well-known gamebook scheme, but the huge costs associated with multiple video shooting schedules and the fact that the necessary interaction at fixed points was perceived as too cumbersome by the users prevented these approaches from attaining any degree of popularity with producers and audiences alike. Therefore, until recently, research in Interactive Storytelling has been mostly associated with computer graphics [21] that allow real-time content generation. However, despite the rapid progress that the 3D-rendering field has enjoyed lately, the visual quality of the interactive stories generated by graphics engines are still nowhere close to that achievable with shot video. Therefore, while there is significant agreement on the interest and impact of Interactive Storytelling in general, the use of video has been generally seen as being too challenging to be considered.

As it turns out, recent advancements in video personalization techniques, for example, video summarization [5], has made possible a rekindling of interest in Interactive Storytelling based on

video. The development of specific techniques for the semantic representation of video [14] provided a means of interfacing the video semantic content with the narrative. Most of the recent work on video-based Interactive Storytelling is about the so-called emergent storytelling paradigm, or storyfication [35], that uses temporal and semantic relationships within the video content to attribute meaning to a sequence of events, for example, forming a life narrative from personal videos [8, 7, 37] or constructing documentaries from user-contributed content [15, 35, 36].

Global plot properties are not considered in these works, which instead follow a bottom-up approach by attaching semantic information to static, basic units of content (e.g., shots), usually through manual tagging, and then achieving a discourse-based output operating only with local constraints. Therefore, these systems do not enforce global narrative constraints but use only the elementary components of planning actions. For example, the New Media for a New Millennium (NM2) project [37] improved branching narrative techniques but did not make use of any reasoning engine, thus it does not maintain global causal consistency. Other works such as IDIC [32] and AUTEUR [24] use some planning concepts, but only to describe individual actions that are then concatenated to obtain short output videos. Local narrative properties are also the backbone of the work of Jung et al. [18], but the need for editorial relations to support the narrative is acknowledged.

The approach followed in all those works, however, takes a very different perspective on narrative with respect to the one associated to movies. Applying Interactive Storytelling to movies is actually about constituting different variants of the original story, i.e., alternative courses of action that still preserve the global narrative properties and dramatic nature of the medium, while the previously cited works aim to generate a coherent narrative using content lacking any sophisticated form of original structure.

To deal with the challenge of generating alternative stories from the same baseline movie, the process of simple reordering of short sequences according to their fixed semantic and temporal relationship as done using the storyfication paradigm is not sufficient. Instead, it is necessary to leverage the combinatorial properties of individual video segments by capitalizing on the Kuleshov Effect [22], which explains how a viewer could attach very different interpretations to the same segment of video content depending on its context, for example, based on what content immediately follows or precedes the considered segment. Since state-of-the-art narrative generation techniques are able to take into account contextual phenomena, the same video content can be flexibly adapted to the alternative narrative, provided that the semantics of individual segments of video is made compatible with the global logic of the artificial intelligence (AI)-based narrative generation.

Indeed, recognizing the potential of this approach, a specific part of the IRIS project focused on Interactive Movietelling, since it specifically dealt with movies. Following a first analysis on the potential of video recombination for conveying different narratives [29], by the end of the project, a working prototype system had been demonstrated [28]. The requirements induced by the global aspects of a plot output by a reasoning engine are considered as an essential key to improve the narrative experience. Its underlying idea is closer to that of Interactive Storytelling systems using top-down, plan-based narrative generation with 3D graphics [30], integrated with the use of video to preserve the aesthetic quality of the generated media. Video processing and the reasoning engine are integrated through a shared semantic representation of the content to obtain a consistent alternative story. In the end, the system is able to concatenate some of the original shots of the baseline movie conveying an alternative narrative by exploiting their semantic description and the global plot aware reasoning engine. In the rearrangement of original video segments, a particular attention is paid to ensure that the shooting stage remains as consistent as possible and that the temporal structure of the overall scene is preserved.

This article enumerates the desirable features of an Interactive Movietelling system using the prototype system as a baseline for illustration, and at the same time, collects all the research insights gathered so far to point out future evolutions of this promising digital entertainment application.

3 INTERACTIVE MOVIETELLING WORKFLOW

In this section, we describe the workflow that is deemed necessary to enable Interactive Movietelling. It is thus essential that the components described in what follows should be included in the design of an entertainment system having the features proposed in this work.

To apply Interactive Storytelling to movies, and in light of what we have discussed in Section 2, it is necessary to integrate a video processing module with a reasoning engine. In fact, the synergy between these two technologies would allow to compensate for each other's limitations and, therefore, to improve on what state-of-the-art video-based storytelling systems would achieve if directly applied to movies. In particular, the global narrative properties used by the reasoning engine cannot be directly mapped to the pure video features and, at the same time, the reasoning engine does not know the details about the available content and its flow characteristics. Given these considerations, an Interactive Movietelling system should rely on the interposition of a *semantic integration layer* between the video processing and the reasoning engine. A possible solution for the implementation of such a layer is the construction of a shared semantic representation to enable communication between content that can be automatically identified by video processing tools and the model of the narrative domain used for narrative generation. The semantic integration layer, which is at the core of this design, allows to link the low-level semantic description provided by video analysis with the high-level perspective taken by the narrative engine, in particular, the plot representation through a sequence of logical actions. Such representation is required to define the backbone of a narrative structure, even in non-video–based narrative generation systems [30].

Therefore, the proposed Interactive Movietelling workflow deals with two different levels of granularity when analyzed from the point of view of the video processing or reasoning engine, respectively. The latter constructs an alternative story with respect to the original one by concatenating what we refer to as *narrative actions*. A narrative action can be seen as a representation of a high-level interaction involving a set of characters, such as "character A travels to location L" or "character A welcomes character B." The task of the video processing unit, on the other hand, is to assemble a video clip for such narrative actions from existing video segments of an original movie. The best way to do this is to represent every narrative action by a sequence of a few shots taken from the available baseline movie using an appropriate semantic description. For example, "A welcomes B" may be represented by an outdoor establishing shot followed by a sequence of indoor, close-up shots, one with character A, one with character B, and one with them both happily chatting.

In the end, following this design, a new output video can be generated providing a meaningful recombination of the original movie shots conveying the new alternative story. The IRIS prototype follows this conceptual architecture closely, as depicted in Figure 1. The top part of the figure deals with the video processing part, the bottom deals with the reasoning engine, and between them lies the semantic integration layer. It is certainly desirable to pre-process available data before running the reasoning engine. The architecture of the IRIS prototype exemplifies this very well. In particular, the preliminary data analysis workflow, on the left of Figure 1, is done offline while the runtime core, on the right, runs in real time, as is detailed in the next section that briefly describes how these tasks are performed in the prototype.

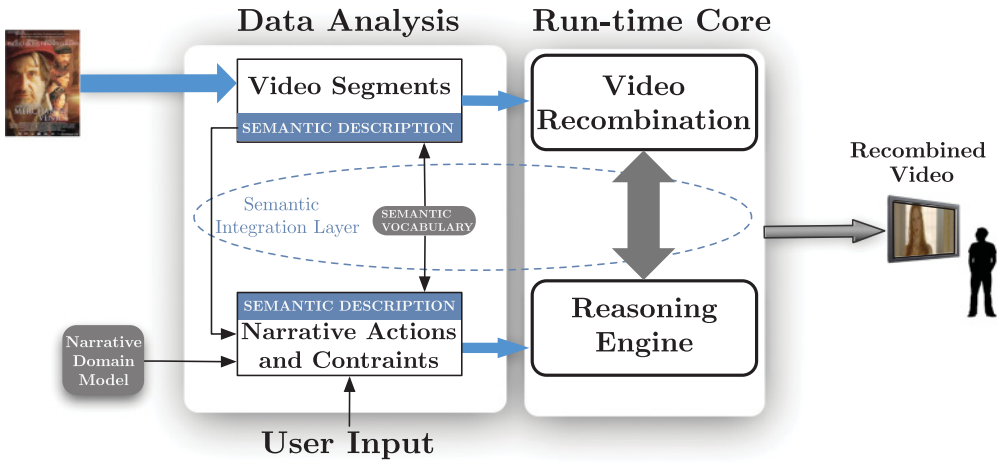


Fig. 1. Architecture of the Interactive Movietelling prototype, adhering to the proposed workflow: the baseline video is segmented into shots and their semantic description is created and then sent to the reasoning engine. The semantic integration layer handles communication between the video processing unit, which performs more semantic modeling of the video content, and the narrative generation module in charge of the plot construction to automatically produce novel filmic variants through video recombination.

4 ANALYSIS OF THE INTERACTIVE MOVIETELLING PROTOTYPE

In this section, we briefly discuss how the Interactive Movietelling prototype implements the proposed workflow, breaking it down into the off-line data analysis, the runtime core and the user's experience, which these choices entail. The section concludes with a description of the role of the author and a recapping of the data flow in the prototype.

4.1 Data Analysis

The input baseline movie is first segmented by the video processing unit into shots. A *semantic vocabulary* specifying all the semantic fundamentals needed to describe the video content is preemptively prepared and shared between the video processing and the reasoning engine subsystems (see Table 1 in Section 5.2). The shots are then semantically described according to this vocabulary, that is, each shot is tagged manually or (semi-) automatically to form a particular set of *semantic attributes*, illustrated in Section 5.2. The description is then made available to the reasoning engine. Therefore, shots constitute a basic unit of consistent semantic content. From a practical point of view, each shot is a separate video file at system disposal (which will then prepare the appropriate playlist to convey the narrative) and they are associated to an XML file containing the semantic description.

Each possible instantiation of semantic attributes is called a *semantic point*. In the reasoning engine domain model, each high-level narrative action is mapped into a *semantic set*, which is a list of semantic points needed to reproduce a certain action, i.e., to accurately convey a conceptual meaning when associated with a specific verbal interaction.

Returning to the previous example, suppose the narrative action "A welcomes B" needs to be represented. A possible mapping of this narrative action to a particular semantic set, as specified by the author in the narrative model, could be the following: a semantic point whose semantic attributes specify that no character is present for the establishing shot, a semantic point involving character A, another involving character B, and a fourth one involving both characters, all points

sharing similar environmental attributes. The task of the video processing unit is then to assemble the appropriate shots, matching the needed semantic points stated previously, through a runtime process of *video recombination*, explained in Section 4.2.

To let the video recombination process take advantage of the already existent narrative structure of the baseline movie, the video processing unit aggregates groups of adjacent shots using the shots' low-level features and their temporal relations, forming so called *Logical Story Units* (or LSUs) [4], which model the baseline movie scenes (see Section 5.1). By joining the LSU segmentation information and the semantic description of the shots therein, as described in Section 5.2, a new set of models is obtained that are referred to as *Semantic Story Units* (or SSUs) [28], which are basically Markov chains. The nodes of the Markov chains are semantic clusters (*SCs*) that group shots in the same LSU, which share a same semantic description. The SSUs embody both the baseline movie scenes' temporal structure and the semantic content of the constituent shots and are to be walked through at the time of the video recombination to generate new narrative actions as required.

The SSUs constitute the semantic integration layer needed to let the video-based recombination system and the narrative engine cooperate. The semantic clusters and the associated transitions inherent in the statistical model that has been built through the previously stated process represent a self-consistent semantic story unit. Thus, the succession of shots, taken from the suitable semantic clusters and chosen by performing a random walk on the Markov Chain, realizes a coherent instance of the required narrative action, as long as the number of shots is at least comparable with that of the movie portion that concurred in the construction of the SSU. In general, the probability associated to the transitions drives the succession of shots without introducing a deterministic pattern while still allowing to retain the structure of the constituent movie scene, a fact that has been explored in related contexts as well, for example, movie summarization [3].

The reasoning engine narrative model is also enriched by a process called *SSU fusion*, performed by the video processing module to propose new narrative actions to be added in the narrative domain, taking advantage of the SSU modeling. This process is explained in Section 5.4.

4.2 Runtime Core

When the SSUs are constructed, the preliminary data analysis phase is concluded and the user can begin interacting with the system. At the start of a user session, he/she specifies through a simple interface the user input for the alternative plot, that is the plot goals and the characters involved. The user input can be seen as a series of constraints on the alternative plot and, therefore, on the sequence of narrative actions that the reasoning engine may construct: more details follow in Section 4.3. The story variant output by the narrative generation module is constructed in a way that preserves the global narrative properties, while at the same time exploiting local causality and consistency as guaranteed by the video content modeling process—this is, in a nutshell, the key advantage of the semantically integrated approach.

Video recombination, as detailed in Section 5.3, is mostly performed at runtime by part of the video processing module using the SSUs obtained in the data analysis stage. In this case, specific requests for each narrative action are issued by the reasoning engine and served by sequencing appropriate shots carrying the needed semantic information. The sequence of shots is chosen using the SSUs by either deleting or substituting semantic clusters of available SSUs to reflect the needed semantic content and then extracting shots from the resulting models.

4.3 User Experience

This section describes the user's perspective, that is, how the user can drive the flow of the system and what output is expected. All pertinent details are to be found in subsequent sections,

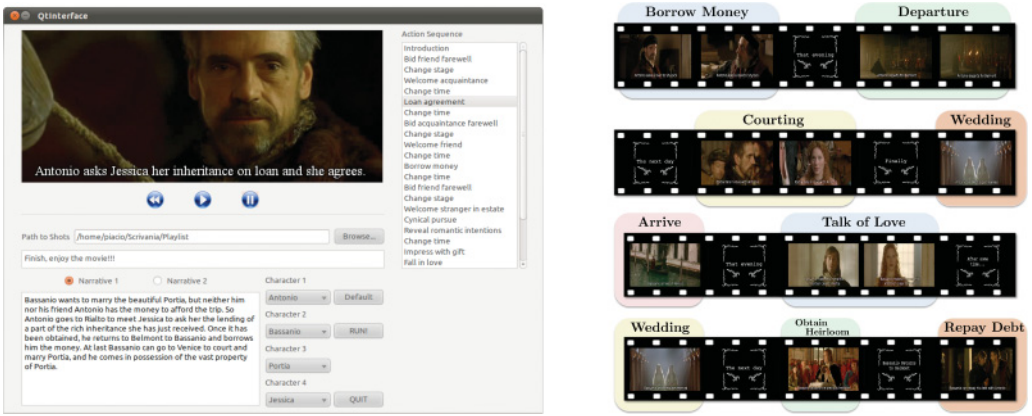


Fig. 2. Left: the user interface containing the video player (top left), the output's narrative action list (top right), and the user configurable settings for the narrative (bottom). Right: a representation of output video clips for two different narratives, which highlights the separation between narrative actions and the roles of subtitles and text panels.

accordingly. The illustrations and results presented in this article are based on Michael Radford's 2004 movie adaptation of the Shakespeare Play, *The Merchant of Venice* [31].

The IRIS Movietelling prototype system has aroused a great deal of interest as a working proof of concept where it has been demonstrated at international conferences, both from multimedia researchers and industrial practitioners. The prototype system interface is shown in the left part of Figure 2. A video clip demonstration of the prototype can be played back and it is provided as supplementary material of this article. The output is a system generated filmic variant of the baseline input movie. It can potentially represent a completely new story, while still using the original locations and physical actors since no new content is generated. The new content can be played back to the user after a few seconds of computation. The video player comes with the usual rewind, play, and pause command buttons, placed in the top left corner. At the beginning of a standard session, the user is presented with a list of choices that allows him/her to influence the generation of the filmic variants.

This process is as follows: first, the user chooses a different narrative with respect to the original one among those generated by the system, using the authored model of the narrative domain. In particular this model defines: (i) the initial narrative state, (ii) the goal narrative state, (iii) the set of narrative actions that can form part of generated story lines by modifying the current narrative state, and (iv) the previously mentioned crucial association between each narrative action and the semantic characteristics (involved characters, shooting scene settings, structure of the interaction, and so on) that must be present to correctly depict it. Moreover, the user can swap the roles of the original cast. In particular, the user can select the characters he/she wants to feature in the output narrative, chosen among the main characters of the original movie, using the sliding menus labeled "Character 1" and so on in the bottom part of the interface (see Figure 2).

In the prototype implementation, only two alternative plots have been envisioned, so the user can choose which one simply with a pair of radio buttons: another sliding menu could be included when more alternatives are available. A plot synopsis is reported in the bottom text window, with the selected characters in the appropriate role. Not all character combinations are possible for a given alternative plot, and in those cases, the text appears in red. Obviously, the same character cannot be cast in more than one role. Additionally, video content resources may be insufficient

to correctly represent key narrative actions with certain characters in them: see Section 5.5. The "default" setting proposes a default character combination for the chosen plot goal.

When the "run" button is pressed, the plot selected by the user forces the narrative generation module to build a narrative path that satisfies the input requirements, and then the user is presented with the actual output of the system, i.e., a recombination of video segments taken from the baseline video. The plot itself is rendered as a sequence of narrative actions, reported in the right column of the interface. They provide a glimpse of the overall narrative structure, and, in addition, clicking on one of them jumps the playback to that point to ease playback.

Audio information is currently discarded because the original soundtrack could only exceptionally preserve consistency with respect to the recombined video content. Instead, the meaning of what is played back relies at present on subtitles that describe the semantics of the scene. Any change of context, such as a character traveling to another location, is highlighted by a transition panel in the style of old silent films.

A schematic representation of an example of output video for each narrative is reported in the right part of Figure 2. For this purpose, each story has been compressed to just a handful of narrative actions. Each image represents the central frame of any given segment. The colored clouds grouping them are titled with the corresponding narrative action. The action itself is described through the subtitles and the black panels separate two actions whenever a significant change of context, i.e., a temporal jump, is present in the narrative.

4.4 The Role of the Author

According to the system framework that we have described so far, a human operator is needed to set up the system, which is referred to as the *author*. The issue of the respective roles of authors and users in this context has been discussed in the interactive narrative literature described in Section 2. Accordingly, the role of the author is distinct from that of the final user who enjoys the Interactive Movietelling system as described in Section 4.3, though they may not necessarily be different persons. In particular, the author is responsible for all those data analysis processes that we have described in Section 4.1 and that can be summarized as follows.

First, the movie shots need to be semantically described offline, a process which is thoroughly detailed in Section 5.2. In brief, the author has to describe the shots for each movie just once, using the provided semantic vocabulary. As a matter of fact, the author could also change the semantic vocabulary itself, this way changing the narrative modeling of the actions altogether, but usually this is not needed as the semantic attributes already present in the current implementation are very general. This way, they also can be (at least in principle) extracted with fair accuracy by automatic systems. However, since the precision required by the system is very high given that a single mis-described shot can cause great harm to the narrative action being rendered, the author's manual intervention is necessary to supervise the shot's semantic description. The description can be performed on just a central keyframe for each shot, and it approximately takes about 30 minutes for a 2-hour movie.

In addition, the author has to formalize narrative actions, largely through their Planning Domain Description Language (PDDL) components, so as to allow the dynamic generation of narrative variants according to user preferences. In practical terms, this process sets up the alternative plot goals that are to be proposed to the final user (see Section 4.3 and Figure 2), in terms of narrative states that are to be reached by the engine during runtime execution. Also, the authors have to prepare the narrative domain, namely the association between the narrative actions used by the narrative engine to advance the plot and the semantically described shots needed to render them. Both processes are described in Section 5.5. An example of such mapping has been previously cited

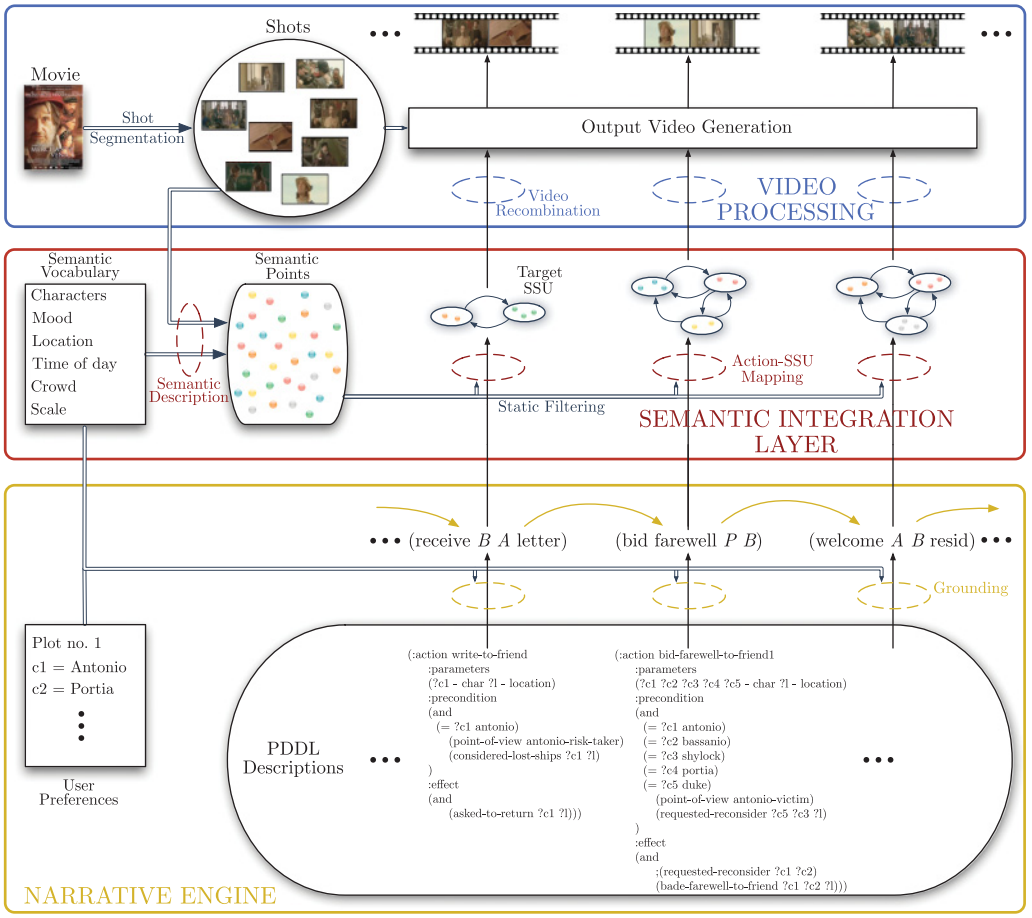


Fig. 3. The architecture of Figure 1, revisited from a data flow perspective. In the flowchart, solid black lines represent the path taken by high-level narrative actions as they flow toward their low-level, video representation, i.e., a sequence of shots. Thicker, grey lines show additional information flows. Yellow arrows in the narrative engine represent the narrative evolution. Rough separations between the narrative engine, the video processing unit, and the shared semantic integration layer are illustrated by the colored boxes.

when we introduced the "welcoming" action. The author is tasked with providing the narrative engine with a set of such mappings, one for each narrative action.

A last optional task for the author concerns the enrichment of the narrative domain suggested offline by the video recombination subsystem during the SSU fusion process, described in Section 5.4. This process mirrors the narrative actions mapping, but instead of relying on the author to build such mapping, the video recombination subsystem suggests new mappings starting from the structure of the original content. If the author chooses to do so, his/her only task is just to validate the proposed mappings according to their perceived quality.

4.5 Data Flow

The previous sections have described the implementation that the IRIS prototype proposes for the conceptual architecture depicted in Figure 1. Accordingly, it is beneficial to revisit Figure 1 from the perspective of the data flow, illustrated in Figure 3, before delving into the system details. The

plot generation takes place in the narrative engine (bottom), and each narrative action follows a separate vertical path toward its video medium representation as it is generated. First, the narrative action, starting from its PDDL description, is grounded using the user preferences. The variables being grounded are, in the end, semantic attributes among those present in the semantic vocabulary. The action is then mapped to a suitable target SSU using the provided narrative domain. The SSU are composed of semantic points obtained through the description of the movie shots. The video processing unit then takes over and tries to build the required SSU through the video recombination process. If this operation is successful, the output video segment depicting the intended narrative action can be finally obtained.

To keep the flow illustration simple, some of the processes touched upon in the rest of the article are not depicted in Figure 3. For example, the fail condition branching described in Section 5.3, which would be represented with a feedback channel from the output video generation subsystem all the way down to the narrative engine, is absent. Also omitted are the off-line SSU fusion process of Section 5.4 and the subtitling process. Nevertheless, Figure 3 is a useful streamlined representation recapping the data communication taking place in the IRIS prototype.

5 DETAILED SYSTEM DESCRIPTION

The baseline video is first segmented along the temporal dimension: this process is briefly described in Section 5.1. Then, the content is described through a set of intermediate-level semantic attributes. Structured models based on both the temporal segmentation and the proposed semantic description can be further derived, as explained in Section 5.2. At last, the video content is recombined to form the alternative plot using the structural models: Section 5.3 recaps the video recombination process. Then, the details on SSU fusion are given in Section 5.4 and finally Section 5.5 describes the narrative generation engine.

5.1 Video Segmentation

At the atomic level, the baseline video is first decomposed into shots using a traditional shot-cut detector [11], which typically works by analyzing the variations of the statistical color intensity distributions of the video frames. Sequences of shots conveying a common concept in the context of the story are then grouped into LSUs [4]. To do so, the shots are first clustered into nodes using both a measure of visual similarity and temporal distance. In particular, visual clustering is obtained using a Tree-Structured Vector Quantization algorithm run over the CIELUV color space values of the square 8x8 blocks describing the shot keyframe content. Visual clusters can be determined through a process of hierarchical clustering. In the end, the video is represented by a Scene Transition Graph (STG), where the nodes represent the visual clusters and the edges correspond to shot transitions. In the case of movies, it can be shown that the STG can be decomposed into cyclic subgraphs, each representing a distinct LSU, separated by cut-edges. An instance of an LSU segmentation process is shown in Figure 4.

5.2 Semantic Description and Modeling

A semantic vocabulary has been designed to define intermediate level concepts with which each shot can be tagged, i.e., the semantic attributes listed in Table 1. The selection of the vocabulary has been made to be sufficiently expressive to include all necessary attributes for an acceptable rendering of the semantic interplay that will be used for narrative generation. The more precise the semantic representation of each shot, the simpler the video recombination is. On the other hand, having too many or too detailed tags makes the narrative generation too convoluted.

We adhered to the following principles when selecting which intermediate attributes to include in the vocabulary. The first of the attributes is a list of the characters present in the shot,

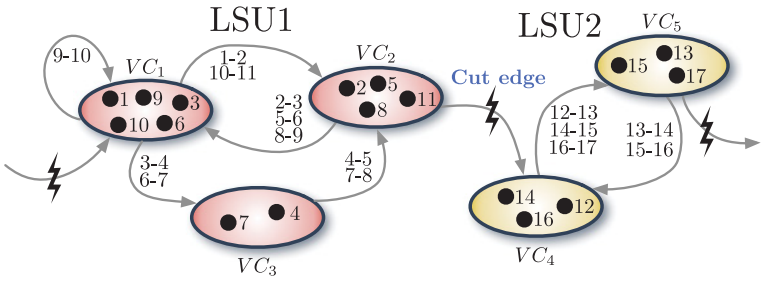


Fig. 4. LSU segmentation using visual clusters of shots and temporal transitions. The visual clusters, VC_1–VC_5, are obtained through hierarchical clustering; black points inside clusters represent individual shots. The numbers on the links refer to shot transitions. Cut-edges are also highlighted.

Table 1. Semantic Vocabulary: The List of Intermediate-level Attributes to Associate to Each Shot and Their Possible Values

Descriptor	Values
Characters	Anonymous tags
*Mood**	Ternary: "positive," "negative," or "neutral"
Location	Binary: "indoor" or "outdoor"
Time of day	Binary: "daytime" or "nighttime"
Crowd	Binary: "crowded" or "not crowded"
Scale	Ternary: "wide", "medium" or "close-up"

*one for each *characters* tag.

specified by an anonymous tag such as "A" or "B," and so on. The default association between these anonymous tags and the movie actors is provided offline by the author. Of course, the characters present in a narrative action are necessary because they are what primarily drive the narrative forward [23]. Attached to each character, there is also a mood indicator taking three possible values: "positive," "negative," or "neutral." These tags help the coherence between the intended narrative action and the actual content being produced, and the mood of the characters is also important to convey the intended narrative action. Both tasks can be performed with satisfying precision by current video processing technologies, i.e., character recognition (for example, relying on face recognition, as we explored in [27]) and facial expression to extract the mood of the characters [34].

However, since the considered shot is being repurposed to represent a part of the narrative action that is, in general, different from the one represented in the original scene from which it is taken, it is not necessary to describe, in depth, the characters' emotions in the shot. Instead, it is sufficient to "cluster" the emotions into three classes ("positive," "negative," and "neutral") and then let the context induced by adjacent shots and the other attributes to convincingly carry the intended meaning of the narrative action. Moreover, facial expressions may also change very rapidly, even within the same shot, while a ternary description such as the proposed one is a much more stable descriptor to associate to a character.

Next, a ternary valued tag indicating the shot scale is used to enhance conveying of appropriate feelings or emphasis, avoiding abrupt and uncomfortable jumps in the framing in the reconstructed video. Shot scale, intended as the distance between the camera and the main subjects of the considered take, is an important tool to effectively describe the semantic content of the scene, as well [1]. In particular, it can be argued that the more distant the camera, the more detachment exists between the viewer emotional response and the scene content [6]. For this reason, wide shot scales

are appropriate for establishing shots or transition shots, and close-up shots are good to convey the character's mood. Generally, even if it is in principle a continuously valued feature, the shot scale is generally describable by three scales as we proposed and can be reliably estimated by automatic algorithms [19, 20].

Last, for the purpose of keeping consistency within the set of shots representing a given narrative action, three additional binary tags specify the "environment" of the shot: time of the day, location ("indoor"/"outdoor"), and unnamed crowd presence or not. These attributes have an important role in the semantic description of the scene [25]. They are inter-dependent, for example, the time of day attribute can be set to a wildcard value in case the location attribute is indoor. Some narrative actions may accept various sets of values for these descriptors. For example, the "welcoming" narrative action is more or less indifferent to time of day, location, and crowd presence. Hence, the only important aspect is to guarantee that each shot chosen for the representation of this narrative action has consistent attributes regardless of which they are, otherwise, there would be coherency problems such as night and day transitions every other shot. Other narrative actions may instead require a fixed environmental attribute; for example, a "traveling" action needs outdoor attributes. This set of binary descriptors can be automatically extracted as well (e.g., see the work of Chan et al. [10] and Serrano et al. [33]).

As the semantic point is defined as a possible combination of semantic attributes, it follows that more than one shot can be associated to a single semantic point if they all share the same set of description attributes. For example, the semantic point containing the following attributes: no characters, outdoor, daytime, crowded and wide depth of field, may describe more than one shot in the baseline movie, and they may also be temporally distant.

A semantic modeling of the LSU pattern can also be constructed. To recap, each LSU forms a subgraph of a STG without cut-edges (see Figure 4). The original LSU clusters are composed of visually similar shots. To relate instead to a more realistic semantic context, shots belonging to each LSU (which are by definition temporally adjacent) are reclustered on the basis of their semantic description. As such, shots within an LSU that are associated to one semantic point can be merged to form a same SC. At the end of such semantic clustering process, a possibly different subgraph may result, called SSU.

The SSU is, in fact, a Markov chain like the LSUs, as shown in Benini et al. [5]. However, with respect to the LSU where the edges represent the shot transitions, the SSU enables the construction of a statistical model that possesses a transition probability matrix P. The matrix P is fitted with values obtained through maximum likelihood estimation using actual temporal transitions; that is, the probability p_{ij} is defined as the number of temporal transitions existing between shots belonging to the semantic cluster SC_i and shots belonging to the semantic cluster SC_j, divided by the total number of shots in SC_i.

According to the correspondence between visual clusters and semantic clusters, various scenarios are possible—they are depicted in Figure 5. A perfect correspondence between visual clusters (VC) and SC may exist at times (Figure 5(a)). Sometimes, this does not happen since visually similar shots may be associated to different semantic points (Figure 5(b)). Due to this non-perfect mapping, in the resulting SSU, an additional cut-edge may exist with respect to the original LSU, as in Figure 5(c). In such a case, if SC_k is a sink node, then $p_{kk} = 1$.

To summarize, the temporal segmentation into LSUs can be associated to original movie scenes. SSUs are instead necessary for the video recombination process. The combination of the two models captures the structural semantic behavior of the baseline movie scenes. Once an SSU is associated to any given LSU, the system can perform video recombination by manipulating SSUs and selecting shots associated to a particular SSU, as described in what follows.

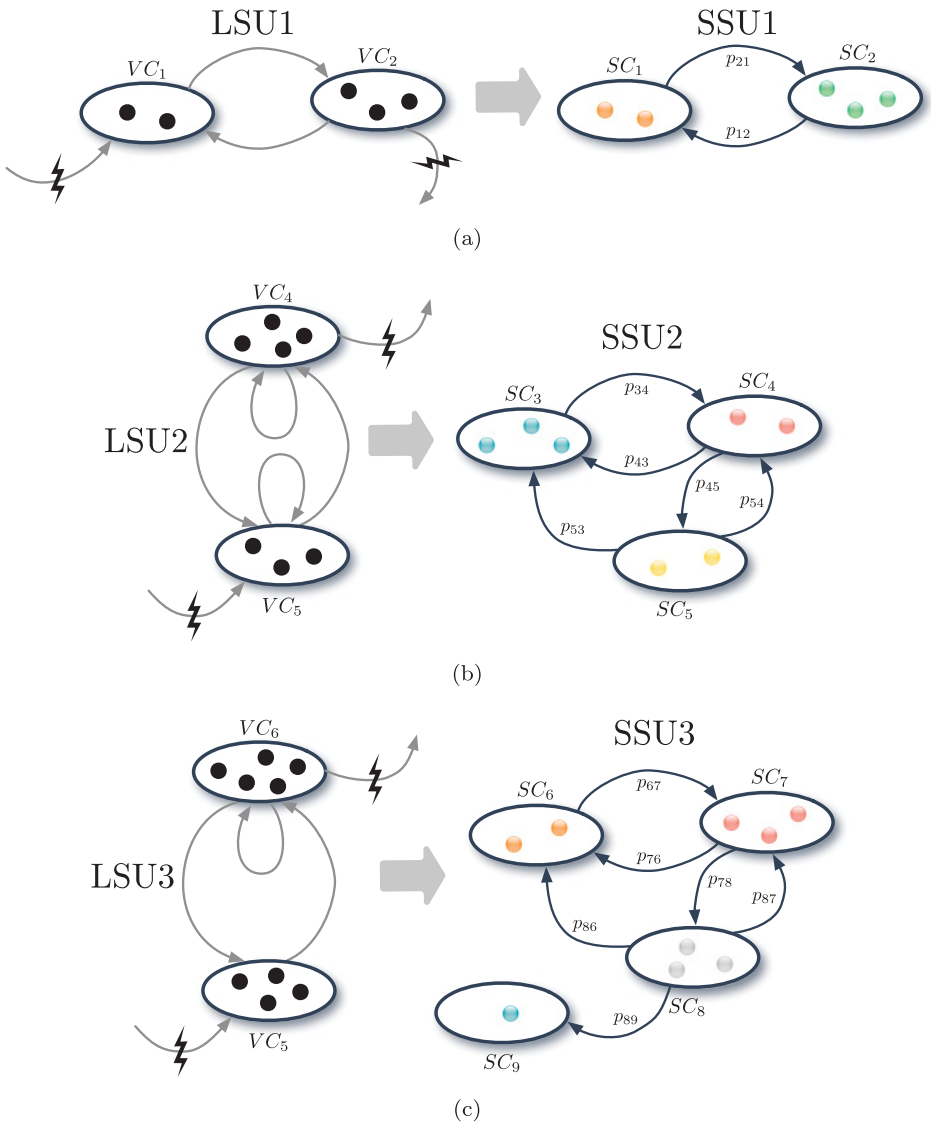


Fig. 5. SSU generation in various *VC-SC* correspondence cases. In the diagrams on the right, colored points correspond to semantic points and darker arrows represent Markov chain edges. In (a), the visual clusters and the semantic clusters are perfectly matched; in (b), one of the visual cluster has spawned two different semantic clusters; and in (c), an additional cut-edge has been added, SC_9 being a sink node.

5.3 Video Recombination

Video recombination is performed at runtime to answer to specific requests from the reasoning engine for its next narrative action to fit the plot objective. At the time of domain definition, the narrative action has been associated into a specific semantic set. The anonymous character tags within it are set according to the actual characters involved and all this information is then shared with the video processing unit. The latter now has the task to choose some shots from the baseline

video consistent with the requested semantic set, which means that all the associated semantic points may be instantiated as needed from the original movie footage, possibly after some shot manipulation. For this purpose, it uses the content semantic modeling provided by the SSUs. As said previously, each SSU is a Markov chain where the nodes represent different semantic points and are populated by those shots described by such semantic points. Since each SSU is temporally confined by a mother LSU, it is possible to have semantic clusters belonging to different SSUs described by the same semantic point. There are three possible scenarios: there is an exact match between the requested semantic set and an available SSU; an SSU can be successfully manipulated to provide for the requested semantic set; or a fail condition is reported to the narrative engine.

5.3.1 Exact Match. In the first scenario, the semantic set is constituted by exactly those semantic points belonging to an available SSU. In such a case, a random walk across the associated Markov chain can be immediately constructed until the needed shots are extracted. Each shot is chosen from its semantic cluster with the only constraint of local causality, i.e., not reversing the temporal order of the shots within a same semantic cluster.

The first scenario is likely to happen when the reasoning engine requests a narrative action already present in the baseline movie or at least a one similar to it in terms of semantic attributes. When, instead, the reasoning engine requests a narrative action whose semantic set is not present in the available SSUs, a second scenario is obtained.

5.3.2 SSU Successful Manipulation. In this case, two processes called *cluster substitution* and *cluster deletion* are performed so as to change the baseline SSUs to construct a target SSU satisfying the requested semantic set, that is an SSU with a one-to-one correspondence between its semantic points and the requested ones. For this purpose, the system identifies at least a candidate SSU that contains at least a matching semantic point with respect to the requested set and has no less semantic clusters than the target SSU. The candidate SSUs are then sorted by the number of semantic points satisfying the request, with the tie-breaking criteria first being the fewest number of shots unrelated to the needed semantic clusters, and then, average visual similarity (already computed for the visual clustering that took place for the construction of the LSU segmentation). Now, the best candidate SSU is processed to substitute and/or delete some of its clusters to match the required semantic set.

Figure 6 illustrates this process with an example. Suppose that the required semantic set is constituted by three semantic points. The best candidate SSU (top left) has SC_1 and SC_3 matching two of the needed semantic points, but neither SC_2 nor SC_4 are consistent for the third set of semantic points. The video processing unit then identifies another SSU (top right) with the needed semantic point, in this case SC_5. Then, SC_2 is substituted by a subset of SC_5 (the number of shots in the original SSU is preserved to minimize the perturbation to the original Markov chain), while SC_4 is deleted. In the end, the target SSU is constructed (bottom). Again, in this whole process the tie-breaking criteria is related to the number of involved shots followed by visual similarity. If more than one semantic cluster needs to be substituted and/or deleted, the process is iterated.

The process of cluster substitution is so designed that the underlying structure of the candidate SSU, i.e., its transition matrix, which is well formed because it is present in the baseline movie, should be perturbed in the least possible way. Special care is given to avoid problems with sink nodes (where the random walk would be trapped indefinitely during the shot extraction), both originally present in the candidate SSU (see Figure 5(c)) and formed by the substitution/deletion process because modifying the edges of the graph can isolate a semantic cluster. If a sink node SC_k is present, its unitary transition probability p_{kk} is redistributed uniformly among all the semantic clusters of the target SSU effectively, eliminating the sink node problem.

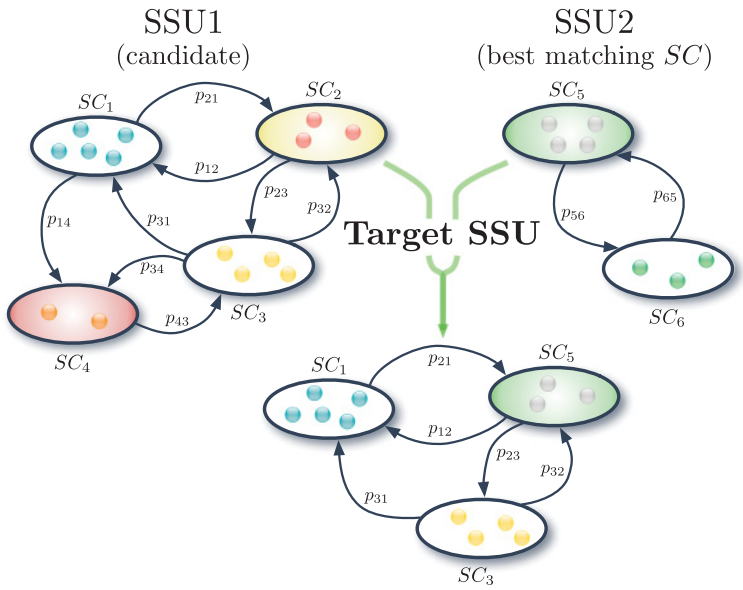


Fig. 6. Semantic cluster substitution and deletion: some of the shots in SC_5 (green) substitute those in SC_2 (yellow) while SC_4 (red) is deleted.

5.3.3 Fail Condition. If the original SSU structure has been modified to a large extent to match the requested semantic set, the output clip associated to the considered narrative action could be of poor quality. For this reason, in a third possible scenario, the video processing unit could report a fail to the reasoning engine, meaning that it must rewind its engine and compute another narrative path avoiding the failed narrative action. The fail condition is evaluated by a heuristic cost, which takes into account both the number of deleted and substituted shots that were needed to form the target SSU. In particular, a cost C is computed as follows:

$$C = 2n_d + n_c, \tag{1}$$

where n_d is the total number of shots (i.e., semantic points) that is necessary to delete from one or more semantic clusters to form the new SSU and n_c is the number of substituted shots: of course, the former process is more damaging for the SSU structure. For example, for Figure 6 $n_d = 2$ (for SC_4) and $n_c = 3$ (for SC_2 becoming SC_5), giving $C = 7$. If C exceeds a threshold set by the author, heuristically estimated by watching some of the constructed narrative actions, the fail condition is set. In our experiments, we set $C = 5$.

5.4 Semantic Story Units Fusion

Runtime video recombination is not the only type of processing that is applied on the SSUs. During the formation of the baseline movie SSUs, the LSU segmentation allowed for the exploitation of the well-formed temporal structure of the baseline movie. For the same reason, to further enrich the narrative domain model before the user session begins, it is reasonable to try to fuse SSUs sharing some semantic information (i.e., both contain one or more *SCs* with the same semantic description), constructing larger SSUs with elements of both, constituting original SSUs. In this way, possible new narrative actions could be identified even before the reasoning engine is started. To do this, video clips corresponding to these possible narrative actions are generated as independent output and validated by a human author. If any of them is deemed appropriate, in the sense that the author

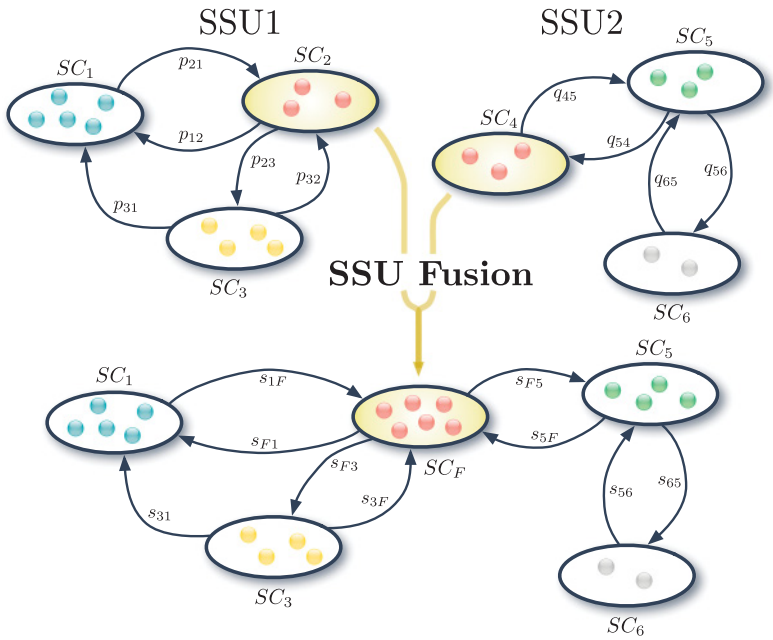


Fig. 7. SSU fusion process: the fused SSU (bottom) has aggregated SC_2 from SSU1 and SC_4 from SSU2, which represent the same semantic point, into the fused cluster SC_F.

thinks that the output clip can reasonably represent a new narrative action, the latter and the association with the semantic features of the corresponding fused SSU are added to the narrative domain model and thus made available as an additional action for the construction of even richer alternative plots. Of course, before performing the SSU fusion proper, the author could first add to the reasoning engine domain model those SSUs directly lifted from the original movie likely associated to original movie scenes and thus to meaningful narrative actions.

To obtain a coherent output, only pairs of SSUs with at least a matching semantic cluster should be considered. The fused SSU possesses the shots of both the constituent SSUs, but the shots belonging to those semantic clusters with matching semantic points are grouped together. Figure 7 illustrates the case where two semantic clusters, SC_2 and SC_4, belonging to different SSUs, share the same semantic point, and as such, in the fused SSU, they are merged in SC_F. The transition probabilities of the resulting Markov chain model are inherited by the original SSUs where possible, that is, in the parts of the chain unaffected by the fusion: for example, in Figure 7, $p_{31} = s_{31}$. For the transitions involving the fused cluster, the resulting transition probabilities are a weighted mix of the original probabilities, dependent on the number of shots contributed by each SSU. More details on how the transition probabilities are handled during the fusion process can be found in the work of Piacenza et al. [26].

5.5 Narrative Generation

The reasoning engine works in conjunction with the video processing unit to construct a consistent story by preserving the global narrative properties of the alternative plot. This approach transposes to the video medium the philosophy of narrative generation, which has been successfully demonstrated in previous narrative generation systems based on 3D animation [30]. In particular, a forward-chaining state-based planner is employed. Narrative generation based on AI Planning

techniques preserve both local and global consistency. As a matter of fact, the resulting complete decoupling between the high-level representation of the plot and the baseline movie content description is a remarkable result that allows the construction of different filmic variants using all the video content resources available in a flexible way. This can be differentiated from other video recombination methods, for instance, those used in summarization, which cannot always guarantee the logical consistency of action presentation or characters involvement, unlike methods based on Planning.

The narrative states refer to categories of actions and character attributes, both generic and specific to the baseline plot, and constitute the domain model of the reasoning engine, which is formalized using the PDDL language (Figure 3 depicts two PDDL descriptions). Of course, PDDL descriptions are always specific to a given interactive narrative, the payoff being in the number of story variants that can be produced from this single formalization. An important reasoning engine task is to properly construct its domain model by identifying the main actors, actions as well as logical constraints for actions, and situations. This logical domain can be mapped onto any smaller set of predicates corresponding to semantic labels obtained from video analysis/clustering [5]. The narrative actions are mapped to a semantic set (ensemble of semantic points) according to a predetermined pattern assessed by a human interpreter during system development.

These mappings are included in the domain model of the reasoning engine. At this stage, since the narrative actions represent the interaction between at most two characters, a maximum of four different semantic points is usually sufficient. Using more complex narrative actions would, of course, require more complex, and perhaps more flexible mappings.

The association between the narrative action and the semantic features is done as to correctly convey the action meaning, as illustrated by the previous example on which and how many characters have to be present and so on for a "welcoming" action (see also Figure 3). A narrative action can be produced, i.e., staged, by different sets of shots, as long as their semantic description is consistent with the action decomposition into semantic features, therefore, the presentation of each individual action is not limited to a fixed segment of the baseline movie. This important feature introduces combinatorial properties for video segments on a principled basis, something that had been the preserve of 3D graphics-based narrative generation. During runtime, generic attributes in the narrative actions are resolved using the specific semantic description needed by the plot, so as to enforce consistency.

Moreover, additional constraints on the narrative generation process are imposed by the original content present in the baseline video. In fact, since the semantic description of the shots is shared with the reasoning engine, the domain model is updated so that the reasoning engine can avoid those narrative actions that would be translated into unavailable semantic content, i.e., the needed required semantic points for those specific semantic sets that are never instantiated. This operation can be seen as a *static action filtering* process, which guarantees that the video recombination could always be performed as the requested semantic set does not contain a semantic point not available in the baseline movie.

Put in another form, the narrative construction procedure is constrained by the available video content in the sense that it must be capable of adapting the generation process to avoid areas of the narrative space for which video data is not available. These adaptations are twofold:

- —static modifications that can be applied to any narrative generated for the given domain and video data;
- —dynamic modifications to recover from unexpected presentation failure.

The first refers to the static action filtering already discussed: this is achieved by filtering the variables in each narrative action as they are ground (i.e., substituted with specific character names

and so on), applying the action-semantics mapping and accepting only those actions that map to semantics appearing in the video, thus avoiding those for which no representative shot exists in the video data. The dynamic plan modifications occur when the video recombination process reports a fail due to the excessive manipulation of the existing SSUs as the video processing unit attempts to identify or find a plausible set of video segments for the requested semantic set.

Overall, while most of the organization of video content into meaningful units is the result of video processing and analysis, narrative generation is in charge of preserving the consistency of the story "backbone" and, in doing so, optimizes the management of semantic resources by generating appropriate contexts that leverage on the semantic interpretations available, as well as ensuring that narrative generation will not require semantic units that are unavailable in context.

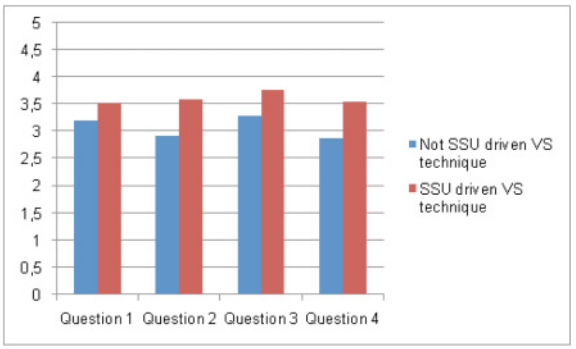
6 EXPERIMENTAL RESULTS

In this section, we thoroughly report obtained results using the prototype Interactive Movietelling system and point out possible improvements both in the evaluation framework to more effectively collect information on the system workflow, and in the system engine as indicated by the user tests themselves.

The most practical way to evaluate an Interactive Movietelling system is, of course, through extensive user tests. For the prototype, we explored how comprehensible the narratives' output by the system were to users. To avoid the users actually grading the accompanying subtitles instead of the overarching narrative, we asked them to focus on the video content and to use the subtitles as a recap on what is said in the scene. The QUEST model [16] represents narratives as a conceptual graph that provides measures that are able to rate the relative quality of comprehension questions. Asking users to assign goodness of answer (GOA) values to question-answer pairs and assessing their correlation with QUEST-predicted quality has proven a useful technique for measuring presentation effect on comprehension in Interactive Storytelling applications such as the one presented here. An additional benefit of this approach over free-form questionnaires is that it eliminates the need for qualitative assessment of user responses.

Correlation between the QUEST model expected quality of question-answer pairs and user ratings would provide strong evidence that the Interactive Movietelling system produces easily comprehended narratives. To determine if this is so, a narrative and video instantiation was produced for each of the three example initial states. Four questions and four answers were randomly selected from the QUEST model of each of our three narratives. This gave 16 question-answer pairs for each narrative, which were presented to 10 participants for a total of 480 evaluations. Participants were asked to watch the video for a narrative and rate the goodness of each answer for each question with a value from 1 (very bad) to 5 (very good).

User responses were compared against measures of reachability and arc distance in the QUEST graph for each narrative. We set expected values for the GOA with 5 (very good) for those with arc distance 1, 4 for those with arc distance 2, and so on, with 1 (very bad) expected for question-answer pairs that are unreachable in the QUEST graph. The mean difference between these expected values and those of the participants was 1.07—significantly lower than the 1.6 mean that would result from random selection. This was significant with $p < 0.01$ by a two-tailed single sample T-test. Furthermore, the correlation between user GOA and the arc distance measure was 0.49 by Pearson product-moment coefficient, which can be interpreted as somewhere between a medium and large correlation. Given that no normalization between participants' results was performed and that the relationship between our arc distance measure and GOA is not necessarily linear, this level of correlation is strong evidence that the video-based presentation of stories has not compromised comprehensibility.



	SSU driven		*Not SSU driven*	
	Mean	Conf. Int.	Mean	Conf. Int.
Question 1	3.19	0.34	2.88	0.31
Question 2	3.23	0.35	2.50	0.41
Question 3	3.50	0.27	3.00	0.29
Question 4	3.23	0.31	2.57	0.31

Fig. 8. Visual Quality Test Results: users were questioned about shot adequacy, coherency, transitions, and enjoyability of videos generated with and without our SSU techniques (see text for detail).

Table 2. Users' Acceptance and Agreement for Both Output Clips Sets

Clips set	*Accept (averaged on 33 clips)*	*Agreement (ratio)*
Nearest SSU Case	~18	0.63
2nd-Nearest SSU Case	~14	0.7

Subjective tests were also run on the quality of video content by generating recombined video clips, of about 4 minutes, relating to two alternative plots ("SSU driven" in Figure 8). For comparison, the same plots were used to generate video that didn't exploit the SSU techniques from Section 5.3; instead, output video was formed by taking shots satisfying the semantic patterns guaranteeing only causality of the shots in the same narrative action ("Not SSU driven" in Figure 8). These four videos were shown in random order to users, who were asked the following questions for each video. First, does the pace of the shots seem right to the user, not too frenetic nor too slow; second, is the shot's visual content coherent with the subtitle's meaning, conveying the narration; third, is the transition between consecutive narrative actions smooth or does it appear artificial; and fourth, is the recombined video pleasant, with an emphasis on perception rather than understanding. Again, the answers were integer grades ranging from 1 (low quality) to 5 (high quality). Figure 8 reports the Mean Opinion Score (MOS) of the answers, along with the 95% confidence interval. From the grades given to the content quality provided by the Interactive Movietelling system, it can be concluded that the users were generally satisfied with the experience, although there is still room for improvement. Also, Figure 8 highlights that the shot recombination process benefits from the SSUs underlying structure inherited by the LSUs, as the user grades for the "SSU driven" clips are clearly better than the "Not SSU driven" ones.

Also worth considering as part of evaluating video recombination is to look at the fusion process. We asked some interviewees to play the authoring role and assess the content generated from the fused SSUs by watching output clips obtained by performing a random walk through the shots of the fused SSU and evaluating if some kind of meaning could be attached to the resulting scene. Of the 37 non-trivial (i.e., with more than a single semantic cluster) SSUs present in the baseline movie, just 33 had at least another SSU with one or more matching clusters, that is, the other 4 had no matching clusters among all the others SSUs and, therefore, are not eligible for the fusion process. Two sets of output clips were obtained by considering, in addition to pairs of SSUs having the best (lowest) associated distance, also those pairs having the second-best associated distance. The results pertaining to these two sets are shown in the rows of Table 2, which report the average clip acceptance and the user agreement ratio, expressed as the ratio between the overlap between

users' acceptance decisions and the total size of the accepted set. As expected, it can be observed that the accepted SSUs in the second row are less than those in the first; therefore, confining the analysis only to the nearest SSU in the fusion process is good, since, as SSUs with higher visual distance are fused, the resulting output clips could be more confusing for the user and, therefore, it is more difficult to give a global meaning to the generated narrative scene. Also, the obtained results in the first row show, by employing the proposed method in the movietelling framework, that fusing SSUs is a viable solution to expanding the narrative model: in fact, the results show that among the 33 proposed new scenes, about 18 can be considered with an acceptable meaning.

In the end, it is clear that, with respect to mainstream Interactive Storytelling systems based on graphics created and rendered on the fly, introducing interactivity in the context of Interactive Movietelling is much more challenging. In the existing prototype, interactivity is limited to an initial selection of the alternative story from a fixed number of narratives, which, in practice, amounts to fix the narrative goal, and the characters' role that instead influences the narrative construction. Starting from these inputs, the planner computes the narrative path that cannot be modified anymore. To improve interactivity, the user could also be able to influence the narrative goal through both direct and indirect interaction. Indirect interaction refers to user monitoring. The user viewing the content can be captured on camera with the intent of detecting his/her gaze and perceived emotions. In addition, the user could be allowed to interact directly with the system, for example, through a simple user interface, to change the story preferences during its playback, of course, according to certain limits dictated by the past and current narrative states, that is, what has already happened in the alternative story. Both these kinds of information would influence the narrative evolution by acting as input to the narrative engine.

Also, user tests showed the semantic description of video content is absolutely necessary to act as a common communication ground between the video processing and the AI planner. This description relies on intermediate-level semantic attributes that are in between the low-level features that can be extracted from raw material and high-level narrative actions with which the planner reasons to construct the alternative narrative. Personal attributes are the most important to convey the narrative, but a second type of information to be extracted from the video content, collectively referred to as the scene attributes set, is essential, too. In fact, their role is to describe the scene context to allow retaining consistency when the video segments are ultimately recombined by the video processing unit. A further attribute that, if incorporated, would possibly improve the semantic description and, in turn, could help in deriving more powerful narrative action models is a description of the emotional state of the scene. This attribute would express which emotional state the director was trying to convey using cinematographic techniques such as lighting or pacing (see, e.g., Benini et al. [2] and Canini et al. [9]).

7 CONCLUSIONS

This article overviews Interactive Movietelling, an innovative application that combines video processing and construction with a reasoning engine to form a filmic variant of a baseline movie according to user input through a simple interface. The integration between high-level concepts pertaining to narrative actions forming the basic building blocks of the plot in the reasoning engine side and video analysis and processing is achieved using a shared vocabulary of intermediate-level semantic attributes. Among the prominent features of the system, there is also narrative generation constrained by the available video resources and novel Markov models manipulation techniques as an accessory to the video recombination process. User tests conducted on the prototype, completed with the complex SSU construction subsystem, show encouraging results in terms of enjoyability and comprehensibility of the output filmic variants. Furthermore, they highlight the importance of the underlying idea of leveraging the pre-existent logical structure of the baseline content.

Many ideas from recent advancements need to being experimented upon to further the application. Automatic semantic description using state-of-the-art tools modified to fit and leverage the particular application they will be used into is clearly a priority—so that more movies can be experimented upon as a result of the reduction in time involved in the current manual description. The usage of the audio portion, both for better semantic content rendering and more enjoyable output, is also being considered. In addition, more flexible and multi-layered semantic modeling processes need to be investigated to both expand the narrative domain and improve the video output quality.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Jonathan Teutenberg for significant contributions provided in the realization and testing of the prototype.

REFERENCES

[1] K. Bálint, T. Klausch, and T. Pólya. 2016. Watching closely. *Journal of Media Psychology* 1 (2016), 1–10.
[2] S. Benini, L. Canini, and R. Leonardi. 2011. A connotative space for supporting movie affective recommendation. *IEEE Trans. on Multimedia* 13, 6 (2011), 1356–1370.
[3] S. Benini, P. Migliorati, and R. Leonardi. 2007. A statistical framework for video skimming based on logical story units and motion activity. In *Proc. Int. Workshop on Content-Based Multimedia Indexing*. IEEE, Bordeaux, France, 152–156.
[4] S. Benini, P. Migliorati, and R. Leonardi. 2010. Hierarchical structuring of video previews by leading-cluster-analysis. *Signal, Image and Video Processing* 4, 4 (2010), 435–450.
[5] S. Benini, P. Migliorati, and R. Leonardi. 2010. Statistical skimming of feature films. *Int. Journal of Digital Multimedia Broadcasting* 2010 (2010), 1–11.
[6] S. Benini, M. Svanera, N. Adami, R. Leonardi, and A. Kovács. 2016. Shot scale distribution in art films. *Multimedia Tools and Applications* 75, 23 (2016), 16499–16527.
[7] S. Bocconi, F. Nack, and L. Hardman. 2008. Automatic generation of matter-of-opinion video documentaries. *Web Semantics: Science, Services and Agents on the WWW* 6, 2 (2008), 139–150.
[8] K. Brooks. 1997. Do story agents use rocking chairs? The theory and implementation of one model for computational narrative. In *Proc. ACM Int. Conf. on Multimedia*. 317–328.
[9] L. Canini, S. Benini, and R. Leonardi. 2013. Affective recommendation of movies based on selected connotative features. *IEEE Trans. on Circuits and Systems for Video Technology* 23, 4 (2013), 636–647.
[10] A. Chan, Z.-S. Liang, and N. Vasconcelos. 2008. Privacy preserving crowd monitoring: Counting people without people models or tracking. In *Proc. Int. Conf. on Computer Vision and Pattern Recognition*. IEEE, 1–7.
[11] C. Cotsaces, N. Nikolaidis, and I. Pitas. 2006. Video shot detection and condensed representation. A review. *IEEE Signal Processing Magazine* 23, 2 (2006), 28–37.
[12] Chris Crawford. 2003. *Chris Crawford on game design*. New Riders.
[13] Chris Crawford. 2004. *Chris Crawford on Interactive Storytelling*. New Riders.
[14] M. Davis. 1994. Knowledge representation for video. In *Proc. Nat. Conf. on Artificial Intelligence (AAAI)*. 120–127.
[15] P. Garcia, D. Bulterman, and L. Soares. 2008. Human-centered television: Directions in interactive television research. *ACM Trans. on Multimedia Computing* 4, 4 (2008), 1–7.
[16] A. Graesser and D. Hemphill. 1991. Question answering in the context of scientific mechanisms. *Journal of Memory and Language* 30, 2 (1991), 186–209.
[17] IRIS. 2011. Integrating Research in Interactive Storytelling (IRIS) Network of Excellence. http://iris.scm.tees.ac.uk. (2011). [Online].
[18] B. Jung, J. Song, and Y. Lee. 2007. A narrative-based abstraction framework for story-oriented video. *ACM Trans. on Multimedia Computing, Communications, and Applications* 3, 2 (2007), 11.
[19] K. Karsch, C. Liu, and S. Kang. 2014. Depth transfer: Depth extraction from video using non-parametric sampling. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 36, 11 (2014), 2144–2158.
[20] A. Kovács. 2014. Shot scale distribution: An authorial fingerprint or a cognitive pattern? *Projections* 8, 2 (2014), 50–70.
[21] M. Mateas and A. Stern. 2005. Structuring content in the façade interactive drama architecture. In *Proc. Conf. on AI and Interactive Digital Entertainment (AIIDE)*. 93–98.
[22] D. Mobbs. 2007. The Kuleshov Effect: The influence of contextual framing on emotional attributions. *Social Cognitive and Affective NeuroScience* 1, 2 (2007), 95–106.
[23] J. Monaco. 2000. *How to Read a Film: The World of Movies, Media, and Multimedia: Language, History, Theory*. Oxford University Press.

[24] F. Nack. 1996. *AUTEUR: The Application of Video Semantics and Theme Representation for Automated Film Editing*. University of Lancaster.
[25] Y. Nakamura and T. Kanade. 1997. Semantic analysis for video contents extraction—spotting by association in news video. In *Proc. ACM Int. Conf. on Multimedia*. ACM, 393–401.
[26] A. Piacenza, F. Guerrini, N. Adami, and R. Leonardi. 2012. Markov chains fusion for video scenes generation. In *Proc. Eur. Signal Processing Conf.* 405–409.
[27] A. Piacenza, F. Guerrini, N. Adami, and R. Leonardi. 2013. Tracking characters in movies within logical story units. In *Proc. Int. Workshop on Multimedia Signal Processing*. IEEE, 183–188.
[28] A. Piacenza, F. Guerrini, N. Adami, R. Leonardi, J. Porteous, J. Teutenberg, and M. Cavazza. 2011. Generating story variants with constrained video recombination. In *Proc. ACM Int. Conf. on Multimedia*. 223–232.
[29] J. Porteous, S. Benini, L. Canini, F. Charles, M. Cavazza, and R. Leonardi. 2010. Interactive storytelling via video content recombination. In *Proc. ACM Int. Conf. on Multimedia*. 1715–1718.
[30] J. Porteous, M. Cavazza, and F. Charles. 2010. Applying planning to interactive storytelling: Narrative control using state constraints. *ACM Trans. on Intelligent Systems and Technology* 1, 2 (2010), 10.
[31] M. Radford. 2004. The Merchant of Venice (film adaptation). (2004).
[32] W. Sack and M. Davis. 1994. IDIC: Assembling video sequences from story plans and content annotations. In *Proc. ICMCS*. 30–36.
[33] N. Serrano, A. Savakis, and A. Luo. 2002. A computationally efficient approach to indoor/outdoor scene classification. In *Proc. Int. Conf. on Pattern Recognition*, Vol. 4. IEEE, 146–149.
[34] C. Shan, S. Gong, and P. McOwan. 2009. Facial expression recognition based on local binary patterns: A comprehensive study. *Image and Vision Computing* 27, 6 (2009), 803–816.
[35] E. Yu-Te Shen, H. Lieberman, and G. Davenport. 2009. What's next?: Emergent storytelling from video collection. In *Proc. SIGCHI Conf. on Human Factors in Computing Systems*. 809–818.
[36] T. Smith and G. Davenport. 1992. The stratification system a design environment for random access video. In *Proc. Int. Workshop on Network and Operating System Support for Digital Audio and Video*. 250–261.
[37] V. Zsombori, M. Ursu, J. Wyver, I. Kegel, and D. Williams. 2008. ShapeShifting documentary: A golden age. In *Proc. Eur. Conf. on Interactive Television*. 40–50.

Received February 2017; revised May 2017; accepted May 2017