Video-based Interactive Storytelling Using Real-time Video Compositing Techniques

Edirlei Soares de Lima [a, *], Bruno Feijó [b], Antonio L. Furtado [b]

[a] *Rio de Janeiro State University (UERJ), Department of Computational Modeling, Nova Friburgo, RJ, Brazil*

[b] *Pontifical Catholic University of Rio de Janeiro, Department of Informatics, Rio de Janeiro, RJ, Brazil*

ABSTRACT

Interactive storytelling systems usually adopt computer graphics to represent virtual story worlds, which facilitates the dynamic generation of visual content. However, the quality of the images and motion produced by these systems is still inferior compared to the high quality experience found in live-action films. Interactive rates in photorealistic rendering for the film industry will not be possible for decades to come. A promising alternative is the replacement of 3D virtual characters with video sequences with real actors. In this paper, we propose a new method for video-based interactive narratives that uses video compositing algorithms that run at truly interactive frame rates. The proposed method is consistent with plots that are generated by nondeterministic planning algorithms. Moreover, we propose a system of artificial intelligent agents that perform the same roles played by filmmaking professionals. A user evaluation of the proposed method is presented. We believe that future improvements of the techniques proposed in this paper represent an important contribution to the quest for new and more immersive forms of interactive cinema.

Keywords: Interactive Storytelling, Video Compositing, Video-based Interactive Storytelling, Virtual Cinematography, Interactive Cinema.

[*] Corresponding author. Tel.: +55 22 2533-2332 (2106); Fax: +55 22 2533-2332.
E-mail address: edirlei.lima@uerj.br

1. Introduction

Since immemorial times, humans have been telling stories. Short stories that started out about hunts and tales of ancestors soon evolved into myths and legends. Over centuries, stories played an important role in human society and were used to teach, inspire, and entertain. With the advent of new technologies, new forms of storytelling were created, notably interactive narratives [1, 2, 3]. In an interactive storytelling system, authors, audience, and virtual agents engage in a collaborative experience. An interactive narrative transforms passive audiences into active users, allowing them to change the way the story unfolds.

The most robust forms of interactive narratives rely on artificial intelligence techniques, such as planning [37], to dynamically generate the sequence of narrative events rather than following predefined branching points. These forms are also known as non-branching interactive narratives. Usually, interactive storytelling systems use 2D/3D computer graphics for story visualization [1, 2, 3, 15]. Recently, films with interactive plots have been proposed as a new experience [4, 5, 6, 14, 13]. However, most of these experiences use prerecorded video scenes with plots based on the concept of branching narrative structures [36], which are known in the area of interactive storytelling to have several limitations such as authoring complexity and lack of story diversity.

Although artificial intelligence techniques can help improve the diversity of stories in interactive storytelling, they face the challenge of generating a visual representation for a story that is unknown beforehand in real-time. In branching narratives, all of the possible storylines are predefined by the author, and all scenes can be carefully planned and shot several times to achieve the best aesthetic quality. On the other hand, in non-branching interactive storytelling systems based on planning techniques, stories are created by the planning algorithm and guided to some extent by the user interactions, therefore, it is not easy to predict all of the possible storylines that can emerge. These unpredictable outcomes require an extra layer of artificial intelligence modules that are capable of representing the emergent narratives and produce the proper visualization of the story, which is even more challenging.

Usually, interactive storytelling systems try to handle the unpredictability of events by using 2D or 3D computer graphics to represent the virtual story worlds [1, 2, 3, 15], like a video game. This approach provides a visual medium that facilitates the dynamic generation of visual content, but even the most recent 2D/3D-based storytelling systems [15, 43, 44] are far from providing the image quality obtained from real camera shootings. Although animation is a powerful and popular storytelling medium, live-action films still attract more attention from the audience. A promising first step to bring interactive narratives closer to films is the replacement of 2D/3D virtual characters by video sequences with real actors.

The main problem of using videos to dramatize an interactive narrative is the lack of freedom caused by immutable prerecorded segments of videos, which reduce interactivity, limit story diversity, and increase production costs. In principle, every possible scene and variation of the story events has to be prerecorded. Consequently, the production costs of a traditional interactive film are multiplied by the number of storylines that can be generated by the system. Currently, there are no proposals for non-branching interactive films in the literature. Also, interactive storytelling technology is far from being

capable of producing digital contents with all the qualities and magic of a movie, such as: intense acting performances, dramatic scene composition, strong physical and psychological interactions among characters, and impeccable image and sound rendering. In this paper, we do not propose to fully cover the complex aspects of movie generation. Instead we focus on simple systems of interactive storytelling that consider non-branching interactive narratives and real actors. The goal of interactive storytelling systems is not to substitute traditional movies but to provide a new media for entertainment and learning. The present study addresses the first steps towards the development of this new type of multimedia production.

More specifically, we propose a new approach to video-based interactive narratives that uses video compositing techniques to dynamically create video sequences that represent the story events. In our method, actors are filmed by several cameras in front of a green screen in a variety of emotional states and situations that are compatible with the logical structure of the narrative. Afterwards, an automatic process controls the cinematography language, completes the required compositing scenes and chooses the view parameters (camera angle, zoom and movements). This approach allows the generation of more diversified stories and increases interactivity, while complying with the non-branching nature of the narrative structure.

The extensibility, but also the limitations, of the proposed approach must be made clear to the reader. As in any interactive storytelling system, we must handle three different types of limitation: types of genre, types of scene, and types of user interaction. First, in our model, any genre can be defined by rules and facts written in a temporal modal logic by a domain (i.e. genre) expert. In this paper, we present our experiments with the genre of folk tales: a variant of the well-known fairy tale Little Red Riding Hood. The extensibility of our model to other genres, such as detective stories and romance of chivalry, was successfully evaluated in some previous works [46, 47]. Secondly, as mentioned in a previous paragraph, our movie generation model has limitations in the type of scenes that it can generate, when compared to the flexibility and quality of a movie. However, our model is flexible enough to incorporate solutions that can mitigate the image quality problems above mentioned, while keeping a strong user's influence on the ongoing plot with surprising outcomes even for the original authors of the story – something not found in branching techniques. Thirdly, we do not have any particular user interaction limitation, mainly because our chapter-based structure can easily schedule the consequences of a complex interaction to take place in the next chapter.

Real time response is another important issue when we are using videos to visualize interactive narratives. Scenes with high quality images consume processing time and may delay the dramatization of the plot. Yet, delayed responses to user’s interactions seriously destroy the flows of movement and narrative. Therefore we need high-quality imagery produced at interactive frame rates. High quality images could be produced by photorealistic rendering techniques found in the modern film industry. However, these techniques cannot be used in interactive storytelling, because they are very time consuming (usually hours of processing time) and require constant human interventions. Algorithms that produce high quality images at nearly interactive rates (but not more than 10 frames per second) could be used, if a dramatization control system could generate a sequence of story events according to the required rendering time. However, this is a complex problem that is still under investigation (Doria et al.

[38]). Therefore, as an additional contribution to the area of interactive storytelling, our approach is based on a compositing algorithm that runs at really interactive frame rates, which guarantees fast responses and good-quality images.

As far as we are aware, no other study in the literature has yet proposed a method for interactive storytelling that uses video compositing techniques at truly interactive rates. The proposed method is consistent with plots that are generated by nondeterministic planning algorithms. Moreover, we propose a system of artificial intelligence agents that perform the same roles as filmmaking professionals. We believe that future improvements of the techniques proposed in this paper may represent an important contribution to the quest for new and more immersive forms of interactive cinema.

This paper is organized as follows. Section 2 examines previous works. Section 3 describes the architecture of our interactive storytelling system. Section 4 proposes a real-time video compositing method to create video-based interactive narratives. In section 5, we analyze the performance and the results produced by our method. Finally, in Section 6, we present concluding remarks.

2. Related Work

The first attempts to use prerecorded video segments to represent some form of dynamic narrative date back to the 1990s [7, 8, 9]. Since that time, several other experiments with interactive narratives using videos have been developed.

Terminal Time [10] is one of the early examples of an interactive narrative that uses videos to produce historical documentaries based on the audience's appreciation of ideological themes. In that system, video clips are selected from a multimedia database according to keywords associated with the documentary events and annotated video clips. In a similar approach, Bocconi [11] presents a system that generates video documentaries based on verbal annotations in the audio channel of the video segments. Chua and Ruan [7] designed a system to support the process of video information management, i.e., segmenting, logging, retrieving, and sequencing of the video data. This system semi-automatically detects and annotates shots for further retrieval based on a specified time constraint.

The idea of a generic framework for the production of interactive narratives is explored by Urso et al. [4]. The authors present the ShapeShifting Media, a system designed for the production and delivery of interactive screen-media narratives. The productions are made with prerecorded video segments and variations are achieved by the automatic selection and rearrangement of atomic elements of content into individual narrations. A similar approach is used by Shen et al. [12], wherein the authors present a video editing system that helps users compose sequences of scenes to tell stories by selecting video segments from a corpus of annotated clips. There are also some examples of video-based interactive narratives used in cinema. Last Call [14] is an interactive advert for the 13th Street TV Channel that has been experimentally exhibited in movie theaters. In Last Call, the audience interacts with the actress talking to her via cell phones. Based on the audience voice commands, the system selects a sequence of videos to be presented according to a fixed tree of prerecorded video segments.

In a more recent study, Porteous et al. [5] present a video-based storytelling system that generates multiple story variants from a baseline video. The video content is generated using an adaptation of video summarization techniques that decompose the baseline video into sequences of interconnected shots sharing a common semantic thread. The static video sequences are associated with story events and alternative storylines are generated using planning techniques. Piacenza et al. [6] present some improvements to these techniques using a shared semantic representation to facilitate the conceptual integration of video processing and narrative generation. Based on a similar approach, Liang et al. [40] propose a system to automatically produce new movies from existing videos in accordance with user created scripts. The system uses a database of semantically annotated video material to identify a group of optimal video segments to narrate the user designed story. Another recent study presented by Müller et al. [13] explores the use of videos in interactive storytelling. These authors describe a system for the production and delivery of interactive narratives, whose web-based client interface represents stories using short video snippets. Following a different approach, Lima et al. [20] present a real-time editing method for video-based interactive storytelling systems to automatically generate the most adequate shot transitions, so as to ensure the visual continuity of the film. However, like other previous works, Lima et al. [op. cit.] deal exclusively with static video segments, so many cinematography principles to ensure the consistency of the video stories have not been applied.

Most of these studies focus on the creation of stories by ordering video segments based on simple branching narrative structures and prerecorded video segments, without using powerful cinematography concepts. To bridge this gap, we propose a new approach to create video-based interactive narratives that uses video compositing techniques combined with intelligent algorithms that apply cinematography principles to create interactive narratives in real-time.

3. System Architecture

The present work is part of the Logtell Project [16]. Logtell is an interactive storytelling system based on temporal modal logic [15] and planning under nondeterminism [17]. It uses a hybrid planner that combines partial-order planning and task decomposition to efficiently address nondeterministic events, i.e. events that can have more than one outcome. Logtell conciliates plot-based and character-based approaches by logically modeling how goals can be brought about by previous situations and events. For each character role, there are goal-inference rules that provide objectives to be achieved by the characters when certain situations are observed.

As illustrated in Figure 1, the Logtell system is composed of three main modules: (1) Story Generator, which uses planning algorithms to create and update the story plot; (2) User Interaction, which manages user interactions and allows users to intervene in the narrative in a direct or indirect way; and (3) Story Dramatization, which represents the events of the story plot using videos. Each module integrates a dedicated controller in charge of handling the network communication between the components: a Planner Controller for the Story Generator, a Drama and an Interaction Controller for the Story Dramatization, and a Global and a Local Interaction Controller for the user Interaction module. Each controller is responsible for interpreting and managing the messages received from other modules. The

system adopts a client/server architecture, with the Story Generator and the User Interaction modules acting as servers, and the Story Dramatization module acting as a client interface. This architecture allows several instances of the Story Dramatization module to be connected with the Story Generator and the User interaction servers, allowing several users to watch and interact with the same or different stories. The communication between the modules is done through a TCP/IP network connection.

In the Story Generator module, stories are generated in chapters. In each chapter, goals to be achieved are specified either by the rules or by user interventions, and the planner tries to achieve them. The chapters are represented as contingency trees, where the nodes are nondeterministic events and the edges correspond to conditions that enable the execution of the next event. The nondeterministic events are executed by nondeterministic automata (NDA) composed of actions. The automata contain information about possible sequences of actions and are open to audience interventions. The basic actions correspond to the primitive actions that can be performed by the virtual characters during dramatization. Details on plot generation and NDA specification used by the Logtell system can be found in [15, 16, 17].

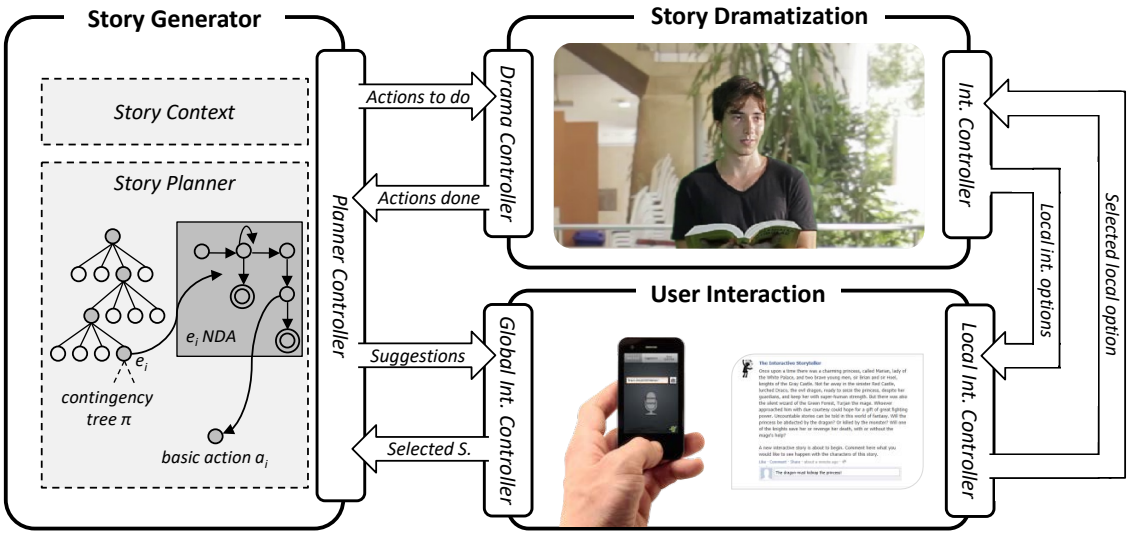


Fig. 1 Modules of the Logtell System.

The system offers two types of user interactions: global and local. In global user interactions, users are able to manipulate the characters' states and suggest events to next story chapters, directly interfering in the generation of the contingency trees for the chapters. Such interactions do not provide immediate feedback, but can directly affect the narrative plot. For example, in the context of the *Little Red Riding Hood* fairy tale, users may suggest that Little Red Riding Hood does not like her grandmother at the beginning of the story. This suggestion will influence the generation of the contingency tree for the chapter where Little Red Riding Hood finds out that the Big Bad Wolf has eaten her grandmother, making the little girl be thankful for the wolf's act. On other hand, local user interactions occur during the execution of the nondeterministic automaton and are usually more direct interventions, where users have to choose between the available options in a limited time. In this type of intervention, users can observe the results of their choices immediately, but such interventions only affect the story plot when the decision leads the execution of the nondeterministic automaton to a different final state. For example,

when Little Red Riding Hood meets Big Bad Wolf, users are directly asked to decide if the little girl should trust the wolf and follow his advices.

In the User Interaction module, local and global interactions are implemented through two interaction mechanisms: social networks and mobile devices. The first method is based on the idea of using social networks (such as Facebook, Twitter and Google+) as a user interface, allowing users to collaborate with the development of the stories in a social environment. The second interaction mechanism combines the use of mobile devices (such as smartphones and tablets) with natural language to allow users to freely interact with virtual characters by text or speech. For example, users can suggest that Little Red Riding Hood does not like her grandmother at any moment of the story by writing in a social network or speaking in the mobile application this fact (e.g.: "*Little Red Riding Hood shouldn't like her grandmother*", or "*I wish that Little Red Riding Hood didn't like her grandmother*"). Similarly, users can use this natural language interface during a local interaction to tell that Little Red Riding Hood should trust Big Bad Wolf (e.g.: "*You should trust Big Bad Wolf*", "*The wolf is trustful!*"). These user interaction methods are described in detail in [18, 42].

The main intended contribution of this paper is a new video-based dramatization architecture that focuses principally on the real-time video compositing task. The proposed architecture is inspired by cinematography theory, and the tasks of the system are assigned to agents that perform the same roles as filmmaking professionals. The cinematography-based agents share the responsibility of interpreting and presenting the narrative events using videos with live actors. Figure 2 illustrates the architecture. Scriptwriter is the agent responsible for receiving and interpreting the automata of the story events generated by the story planner. The Director agent is in charge of controlling the execution of the nondeterministic automata and the dramatization of basic actions, including the process of defining the location of the scenes, the actors and their roles. The Scene Composer agent, using real-time compositing techniques, transforms the scenes into a single piece of a motion picture. The Cameraman agent controls a virtual camera and suggests types of shots (e.g. close-up, medium shot, long shot) for the scenes. The Editor agent, using cinematography knowledge of video editing, selects the best shot for the scenes and controls the temporal and spatial continuity of the film. The communication between the Story Dramatization Client and the other modules of the system is handled by the Drama Controller and the Interaction Controller.

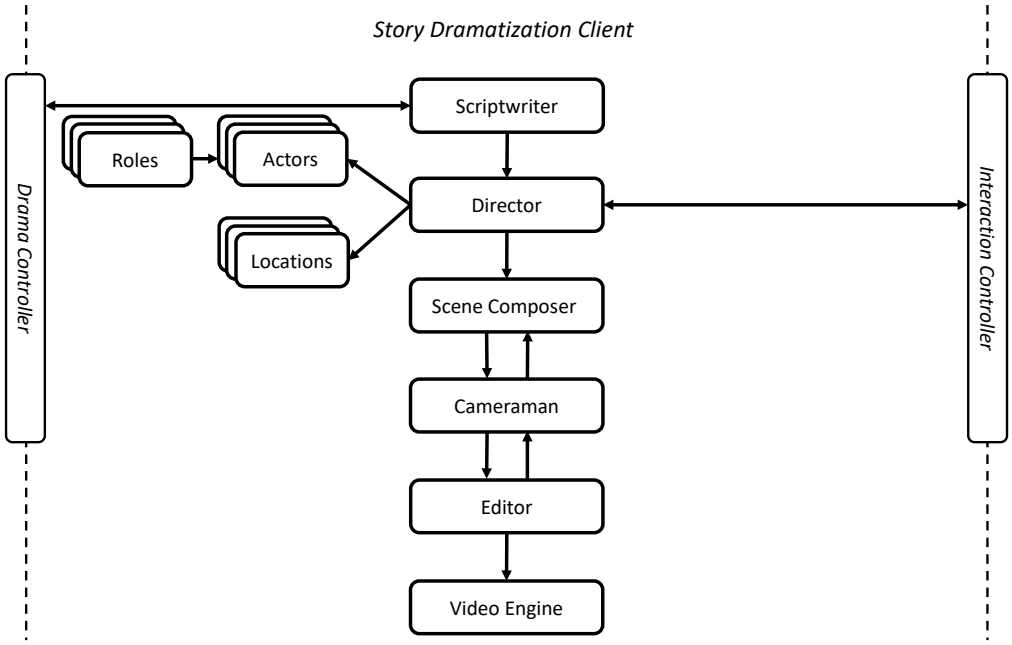


Fig. 2 The proposed video-based dramatization architecture

4. A New Method for Video-Based Interactive Storytelling

The proposed approach to video-based interactive narratives uses video compositing techniques that dynamically create video sequences to represent the story events generated by the planning algorithms. Video compositing is the process of assembling multiple visual elements from different sources into a single section of a motion picture. This process requires the application of a technique called matting, which is the process of extracting the visual elements from the background so they can be used by the compositing tasks. Chroma key (also referred to as green screen or blue screen) is the most common matting technique used in the film industry today [21]. Chroma key involves shooting the visual elements in front of a green or blue screen, and then executing an algorithm to remove the colored screen from the shot and replace it with the substitute background during the compositing process.

As in traditional filmmaking, a video-based interactive narrative must have a cinematic look and be composed of a variety of different shots, camera movements and transitions. In our system, to create such cinematic interactive narratives, actors and settings are both shot from 8 different angles at intervals of 45 degrees, which forms a circle around the subject, as illustrated in Figure 3. The multiple angles give the system the freedom to dramatize scenes by applying the basic cinematography concepts during the dramatization of the narrative.

The places where the story events may occur are also filmed from 8 angles at 45 degree intervals forming a circle around the stage. Each location is composed of a set of video or image layers representing the environment. Usually, outdoor locations are represented through videos, which are able to reflect the natural dynamism of the environment (e.g., leaves moving in the wind, people walking in the distance, birds flying around). Indoor locations that do not include dynamic elements are represented by static pictures. Figure 3 shows an example of a location composed of two layers, in which layer L_1

contains the image of the background of a restaurant and L_2 contains the image of a table with three chairs.

The locations also contain an undirected graph of waypoints that define the location's basic geometrical structure. The waypoints indicate where characters can be placed during the compositing process. There are three types of waypoints: (1) entrance/exit waypoints, which mark the points where characters can enter/leave the scene; (2) acting waypoints, which are used as a point of reference to place characters that are performing some actions in the scene; and (3) connection waypoints, which are used to create paths between the other waypoints. Each waypoint contains information about its specific position in the location and the angle that a character occupying that position must assume. In addition, each location contains the definition of a *front line* and *back line*, which delineates the region where characters can be placed during the compositing process. Both back and front lines include the description of the relative size that the characters must be when they are placed over the lines. Figure 3 shows an example of a location that contains a graph connecting five waypoints (W_1, W_2, W_3, W_4 and W_5) placed between the front and back lines (F_1 and F_2). In this figure, W_4 and W_5 are entrance waypoints, and W_1, W_2 and W_3 are acting waypoints.

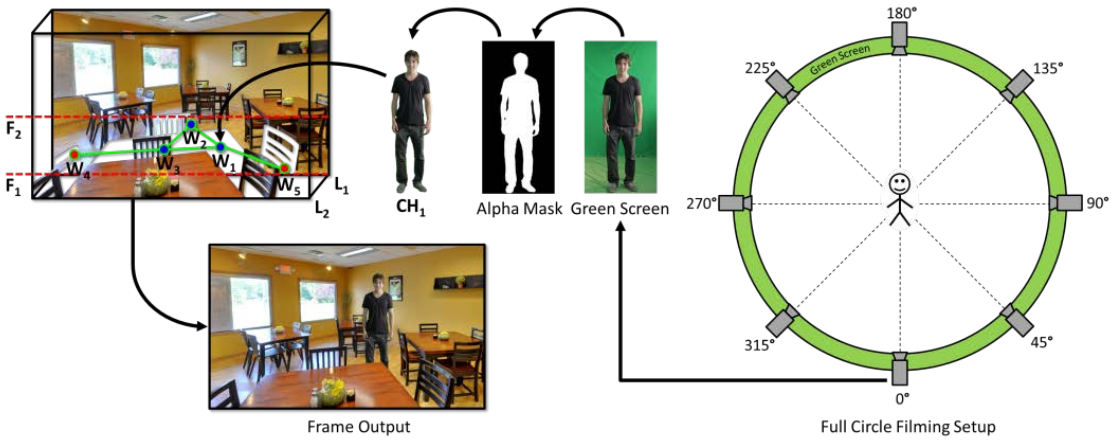


Fig. 3 Location with 2 layers (L_1 and L_2) e 5 waypoints, with the character CH_1 placed at waypoint W_1. The front line F_1 and the back line F_2 delineates the region for waypoints

The actors are filmed in a long shot (which shows their entire body), and locations are filmed in a wide shot to include the whole environment. After recording the videos of the actors performing their basic actions and dialogs, the raw material passes through a pre-processing phase, where the background of the video is removed using the chroma key matting technique, and alpha mask videos are created to define the clipping region that separates the actor from the background in the original video. Each frame of the alpha mask is a grey scale image in which black represents fully transparent pixels, white represents fully opaque pixels, and grey pixels represent a corresponding level of opacity (Figure 3).

Dialogs serve important functions within a traditional narrative film and are easily generated. In video-based interactive storytelling systems, however, we do not know satisfactory solutions for the dialog problems, especially lip sync problems. Therefore, in our method, we have the flexibility of recording essential pieces of dialog (in different emotional circumstances), which are designed to be reused in several compositing situations. This flexible approach, which we call *basic dialog approach*,

allows us to get rid of lip sync problems. There are two limitations of this approach: first, we need to shoot a great number of scenes; and secondly, some compositing scenes may appear unnatural if compared with real interactions between experienced actors. Nonetheless, these drawbacks can be minimized by a good design of basic dialogs, which can avoid excesses – such as long dialogs that can jeopardize the richness of user interaction, and a large number of short basic dialogs to be composed.

We are not proposing a general movie generation model, which can generate any type of scene automatically. The model is designed to have two types of users: film experts and regular spectators. While regular spectators only watch and interact with the system to create the final movie, film experts (with the help of genre experts) can interfere in the results of the system during the film production. For example, they specify default waypoints where actors are placed by the system during the compositing process for specific types of scenes. For scenes that include complex interactions among actors (e.g.: touching, fighting, and kissing) – which may not be adequately composed by the system in real-time – film experts can use methods and expedients that mitigate quality problems of image and compositing limitations. Such expedients may influence the way that basic actions are shot during the production process. For example, a kissing scene involving emotional touches and hugs may be simplified to kiss were both actors only touch their lips (Table 2 – Action 3). Scenes where the interaction between characters is indispensable can be represented as *linear scenes*. This type of scene consists of a prerecorded video of the entire scene, including characters, film set, camera movements, and different shots of the actions, as it occurs in a traditional film. When an event that is represented by a linear scene is generated by the planning algorithms, the dramatization system will exhibit the prerecorded video instead of compositing the event in real-time. We can interpret the combination of linear scenes and real-time scene compositing as a hybrid situation, which represents the integration of linear and nonlinear parts in interactive storytelling systems. This integration is an effective way to reduce the complexity of nonlinear narratives.

The process to create video-based interactive narratives is divided into two phases: (1) scene definition, where the logical description of the scene is configured; and (2) scene compositing, where the video frames representing the scene are generated by the system. Figure 4 offers an overview of the video compositing process, which is explained in the following sections.

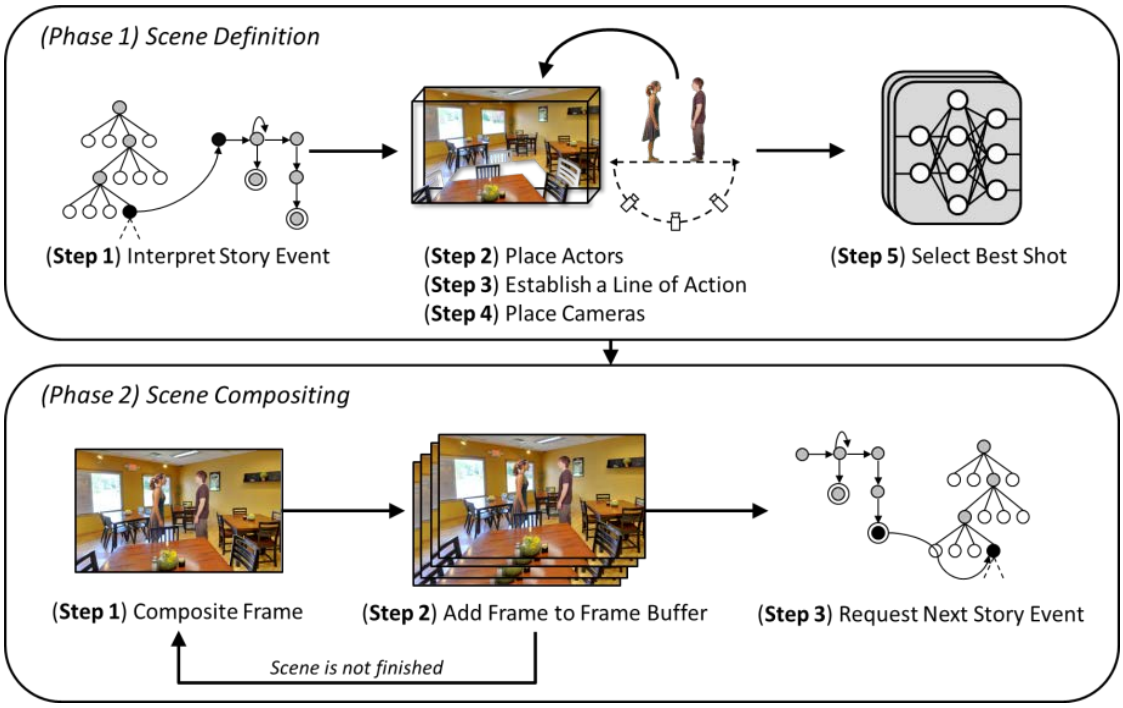


Fig. 4 An overview of the video compositing process

4.1. Scene Definition (Phase 1)

The first step of the video compositing process consists of interpreting the narrative events that must be dramatized. These events are automatically generated by the planning algorithms of our interactive storytelling system and are described in the form of ground first-order logic sentences (e.g. *lookAt*([*anne*], [*peter*], [*university*]). The narrative events are translated by the system into a logical scene structure, which comprises a list of the scene elements composing the scene that represents the event. There are three types of scene elements: (1) location, which defines the place where the event is happening and includes the video or image layers of 8 angles of the location, together with their respective waypoints and encoded information; (2) main characters, which include the videos and alpha mattes of the actively participating actors performing their current actions; and (3) supporting characters, which are the characters that are not directly participating in the action, but are in the same place where the action is happening.

The actors are placed on the scene according to the actions they are performing and the available waypoints of the location (Step 2, Phase 1, in Figure 4). The position and angle of the actors are defined according to the information provided by the waypoints, but their size must be automatically calculated by the system. As in the real world, the closer the actor is to the camera, the larger he/she must appear to be in relation to the rest of the scene. Accordingly, in our method, the width A_w and height A_h of an actor A in a location L are given by:

$$A_w(A,L) = A_v^w\left(\frac{\alpha(A,L)}{100}\right) \quad \text{and} \quad A_h(A,L) = A_v^h\left(\frac{\alpha(A,L)}{100}\right)$$

where A_v^w and A_v^h represent the original size (width and height, respectively) of the video of actor *A*, and the function $\alpha(A, L)$ computes the relative size of the actor through a linear interpolation between the front line L_{F1}^{pos} and back line L_{F2}^{pos} according to the actor's current position A_y and relative size on the front and back lines (L_{F1}^{size} and L_{F2}^{size}):

$$\alpha(A, L) = L_{F1}^{size}\big(1 - \gamma(A, L)\big) + L_{F2}^{size}\big(\gamma(A, L)\big)$$

where $\gamma(A, L)$ is a function that normalizes the current position A_y of the actor *A* in the interval [0,1]:

$$\gamma(A, L) = \frac{A_y - L_{F1}^{pos}}{L_{F2}^{pos} - L_{F1}^{pos}}$$

Although the actors are initially placed over the waypoints, their position may change during the dramatization of the action. Thus, every time the position of an actor is modified, its relative size is recalculated and updated. In addition, during dramatization, the system must maintain the relative size and position of the actors when the scene is viewed from different camera angles, especially when the actors are moving between waypoints.

After defining the basic configuration of the scene, the next steps of the compositing process are the establishment of a line of action and definition of the virtual cameras that can be used to film the scene (Steps 3 and 4, Phase 1 in Figure 4). Each shot requires placing the camera in the best position for viewing characters, setting, and action at a particular moment in the narrative. The approach employed to accomplish this task in the proposed system is based on the use of some standard arrangements for camera placement defined by cinematography theory [25]. In a scene of a dialog between two characters, for example, it is common to use the pattern known as the triangle system [23], whereby all possible shots for any subject are taken from three points forming a triangle within the currently chosen side of a "line of action", which is used to maintain the spatial continuity of the scenes [22]. To establish the virtual line of action in the scene, we adopt some common guidelines presented by Thompson and Bowen [24], which state that, in a scene with a single character, the line of action is usually given by the initial direction of the character. In scenes involving more characters, it is established by a line connecting the two most important characters in the scene. In this way, the virtual line of action is defined based on the position and orientation of the characters participating in the action.

After defining the cameras that can be utilized to film the scene, the system must select the "best" camera to be used (Step 5, Phase 1, in Figure 4). In actual filmmaking, directors have their own individual style and define the shots according to their knowledge and preferences. Our approach to this problem consists of using several artificial neural networks [26] trained to solve cinematography problems involving camera shot selection. Our model, purporting to represent the knowledge of a human film director, is illustrated in Figure 5. For each type of scene (*e.g.*, dialog scene, chasing scene, fighting scene), there are two artificial neural networks: the first one is trained to classify the best camera angle for the shot based on geometric information extracted from the scene; and the second network is trained to select the best type of shot based on the camera angle selected by the first neural network and on emotional information extracted from the characters participating in the scene. The emotional model

adopted in our system covers the six basic emotions proposed by Ekman and Friesen [27] and is simulated through a dynamic network of emotions and relationships [32].

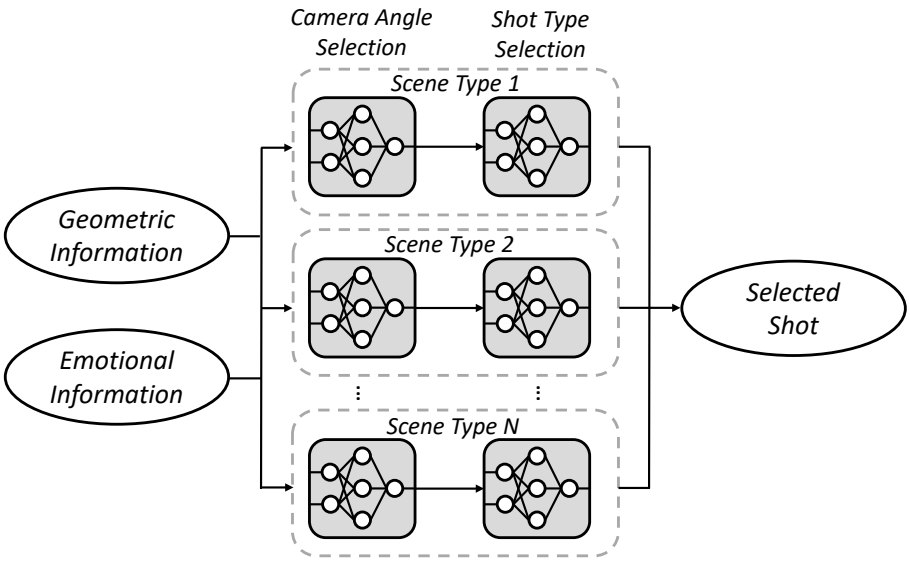


Fig. 5 Neural network system

Our method employs single hidden layer neural networks trained by a standard back-propagation learning algorithm using a sigmoidal activation function. The input of the neural network that selects the best camera angle comprises a set of geometric features that are extracted from the scene setup. It includes the angle and position of the characters participating in the action (X, Y and Z-index, relative to the center of the scene and arranged based on the order of importance of the characters in the scene), and the identification of the action performed by the main character. The number of input values depends on the type of scene and the number of characters involved in the action. For example, a dialog scene between two characters includes 9 input values and, consequently, for this type of scene, 9 nodes in the input layer of the neural network. The output of this neural network comprises the possible camera angles proposed by the Cameraman agent during the camera placement phase.

Once the camera angle has been selected by the first neural network, the next step is the selection of the type of shot. Usually, the decision of the best type of shot depends on the emotional content of the scenes [22]. More intimate shots, such as close-ups, are often employed when there is a substantial change in the emotions of characters, highlighting the facial expression of the subjects [24]. The input of the neural network used to select the best type of shot comprises a set of emotional features extracted from the dynamic network of emotions and relations. It includes the variation (relative to the previous shot) of the emotions and relations of characters participating in the action, together with the identification of the camera angle selected by the first neural network. The number of input values depends on the number of characters involved in the action; for instance, a dialog scene between two characters includes 15 input values (15 nodes in the input layer). The output of this neural network is composed of 5 nodes, which represent the five most common types of shots (i.e., close-up, medium close-up, medium shot, medium long shot and long shot). When the output is calculated, the activated neuron in the output layer indicates the selected type of shot.

The neural networks were implemented using the FANN library[1]. The networks were trained offline using training samples collected through a simulation process, during which we simulated 50 scenes and, for each one, the best shot (angle and shot type) was selected according to the opinion of a professional film editor. Each decision generates one training sample, which includes all of the features used as input for the neural networks, together with the selected camera angle and shot type for the simulated scene. Once the neural networks are trained, they can be used in real time to select the best cameras to film the scenes. An evaluation of the accuracy of the proposed method is presented in Section 5.1.

The use of neural networks trained by filmmaking professionals allows the system to learn the personal style of the human professionals and replicate it during the video compositing process. This feature endows the system with the ability to effectively apply cinematography rules and principles while keeping the signature of the human artist in the computer generated content.

4.2. Scene Compositing (Phase 2)

After defining the whole structure of a scene, the system starts the actual process of compositing video frames that represent the scene (Figure 4, Phase 2). This process, which is the most time-consuming task of the whole system, must be performed in real-time to generate at least 30 frames per second.

We adopt a parallel frame compositing architecture that is capable of managing and compositing multiple video frames (Figure 6). In this architecture, the Scene Compositing Control module manages the compositing process and execution of several threads that are responsible for compositing the video frames. Each frame of the scene is assigned to a thread and, when a thread finishes compositing a frame, it is added to the Frame Buffer, which is an ordered list of frames that are ready to be exhibited.

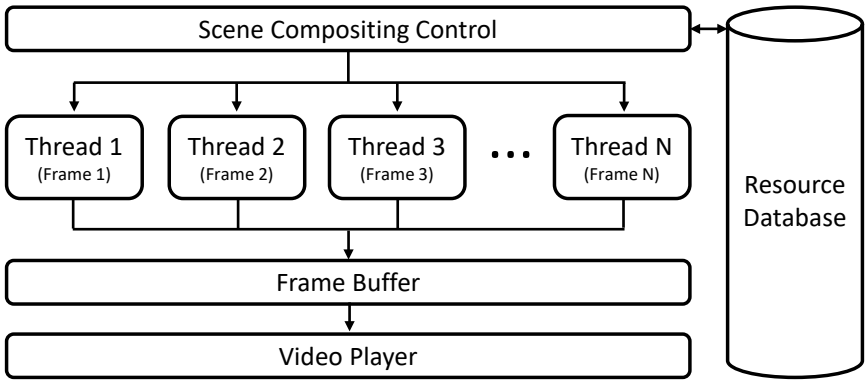


Fig. 6 Parallel video compositing architecture

Each compositing thread generates its assigned frame according to the information provided by the scene structure. The pseudocode of the compositing algorithm executed by each thread is shown in Figure 7. The algorithm receives the identification of the frame that has to be generated (`frame_id`) and a reference to the current scene structure (`scene_structure`). The compositing process starts by

[1] Fast Artificial Neural Network Library (FANN) - http://leenissen.dk/fann/

retrieving the background frame (`bg_frame`) of the current scene location, which is defined in the scene structure according to the angle chosen by the Editor agent to film the scene. It is important to notice that the frame is retrieved based on the identification of the frame that has been assigned to the thread (`frame_id`). Then, for each scene element present in the scene structure (actors and location layers), the algorithm retrieves its frame (`element_frame`) and alpha mask frame (`mask_frame`) based on the `frame_id` and orientation of the element in the scene. Next, the `element_frame` and `mask_frame` are combined to create an RGBA image (`alpha_frame`), which uses the `mask_frame` as the alpha channel of the image. Then, the `alpha_frame` is resized according to the width and height of the scene element that were defined in the previous steps of the compositing process. The next step consists of a clipping operation, which is performed over the `alpha_frame` in order to eliminate parts of the element that are not inside of the frame region defined by the angle and type of shot selected by the Editor agent. Before blending the `alpha_frame` with `bg_frame`, a color correction operation is performed to adjust the color of `alpha_frame` according to the color of its area in `bg_frame`. Then, an alpha GPU compositing operation is performed to blend the resulting `alpha_frame` with the `bg_frame`, which completes the compositing process of the scene element. The algorithm returns the composed frame (`bg_frame`).

```
1. function compose_frame(frame_id, scene_structure)
2.   get bg_frame of frame_id from the location defined in scene_structure
3.   for each scene_element in scene_structure do
4.     get element_frame and mask_frame of frame_id from scene_element
5.     combine element_frame and mask_frame to create an alpha_frame
6.     resize alpha_frame according to the size of scene_element
7.     perform clipping operation in alpha_frame
8.     correct the color of alpha_frame based on bg_frame
9.     perform an alpha GPU compositing operation blending alpha_frame
                                                              with bg_frame
10.  end
11.  return bg_frame
12.end
```

Fig. 7 Pseudocode of the compositing algorithm

Exposure is the amount of light collected by the camera sensor. If controlling exposure does not produce a bright enough image, then the signal gain of the camera can be adjusted to obtain a brighter image. In our system, the scene elements and background images are often captured at different levels of exposure or gain. Therefore, the compositing process requires an exposure/gain compensation to produce a better harmonized image compositing. This compensation is part of a more complete stage called color correction, in which more corrections are made to obtain a better exposure and balance of light or to adjust the image to match the color temperature to a predefined choice for each scene. In our algorithm, color correction means exposure/gain compensation only. Additionally, we do not use any technique of color grading, which is a process to further enhance an image or establish a new visual tone for the scenes.

The color correction algorithm used in our frame compositing process is based on the exposure compensation method proposed by Brown and Lowe [28] to correct color differences in panorama image stitching – which is the process of combining multiple images with overlapping fields of view to produce a segmented panorama or high-resolution image. Their method adjusts the intensity gain level of the

images by minimizing an error function, which is the sum of the gain-normalized intensity errors for all overlapping errors.

In our system, the exposure compensation method is used in the frame compositing process to adjust the exposure levels of the scene elements based on the background frame. The error function is:

$$e = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} N_{ij}\left(\left(g_i\bar{I}_{ij} - g_j\bar{I}_{ji}\right)^2/\sigma_N^2 + (1-g_i)^2/\sigma_g^2\right.$$

where I_i is the frame i with a scene element; I_j is the background frame j; g_i and g_j are the gains respectively; N_{ij} is the number of pixels in frame i that overlap frame j; and $\bar{I}_{ij}$ is the mean color value of the pixels in frame I_i that overlap frame I_j. The parameters σ_N and σ_g represent the standard deviation of the intensity errors and the gain standard deviation respectively, which have been empirically set to $\sigma_N = 10.0$ and $\sigma_g = 0.1$ [28]. Also we require $I \in \{0..255\}$. The criterion employed to determine the intensity gain level is the minimization of the error function e with respect to the gain g. After the step of gain correction, we do not apply any technique of multi-band blending as proposed by Brown and Lowe [*op. cit.*] although we strongly recommend it for future work.

The alpha compositing operation is applied to blend the alpha frame of the scene element with the background frame. The algorithm uses an "over" operator to blend together the color and alpha values of the images on a pixel-by-pixel basis [19]. The alpha compositing algorithm runs on a GPU and takes advantage of its parallel architecture to compute the color of several pixels simultaneously, which improves the performance of the process and allows the system to compose the interactive scenes in real-time.

5. Evaluation and Results

To evaluate the results produced by the proposed methods, we performed three tests: (1) a technical test to check the performance and precision of the video compositing methods; (2) a visual evaluation test to compare the compositing results automatically produced by our system with results manually produced by human video compositing professionals; and (3) a user evaluation test to check the user experience provided by our system from a Human-Computer Interaction (HCI) perspective. All video resources used in the evaluation tests were recorded at a resolution of 1080p and the computer used to run the experiments was an Intel Xeon E5620, 2.40 GHZ CPU with 24 GB of RAM.

5.1. Technical Evaluation

To evaluate the performance of the parallel architecture of the frame compositing process, we conducted a test to check the average frame rate of our system with the number of threads, ranging from 1 to 8. Four scene sequences comprising 4 basic actions were simulated and dramatized by the system, generating a total average of 600 frames per scene. The sequences were composed by the same basic actions with different number of actors per scene. The results of the performance tests of the parallel architecture are shown in Figure 8.

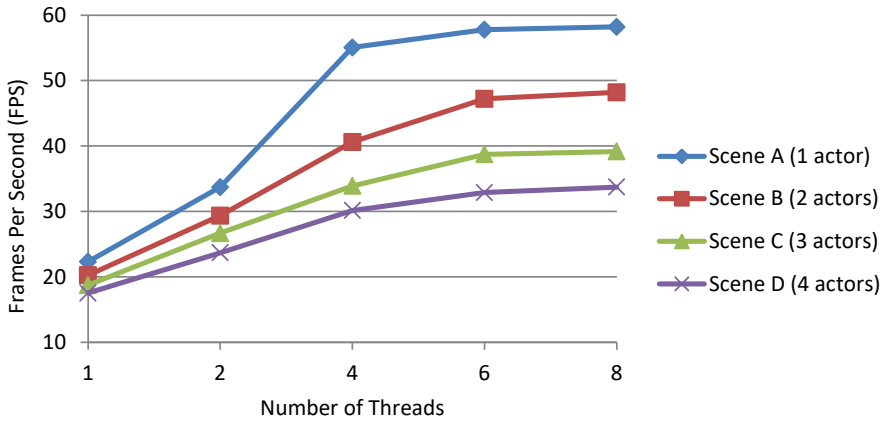


Fig. 8 Performance tests of the parallel architecture with the number of actors in the frame ranging from 1 to 4 and with the number of compositing threads ranging from 1 to 8

The results of the performance experiment showed that the process of compositing a frame becomes more expensive as more scene elements are added to the frame. However, the parallel architecture of the proposed system can compensate the cost of the frame compositing task by dividing the work among multiple CPU cores.

We also evaluated the accuracy of the proposed shot selection method. For each type of scene implemented in our prototype application (total of 12 types), we created 4 training sets with a different number of samples and, for each one, a test set with half the size of the corresponding training set. The samples were collected through a simulation process, where we created several scenes varying the type of scene, number of actors, emotional states and actions, and then, for each scene, we asked a human editor to make the selection of the best shot (angle and shot type) to film the scene. Each decision generates one sample, which includes all the features used as input for the neural networks, together with the selected camera angle and shot type for the simulated scene. The training sets were used to train the neural networks and the samples of the current test set were then predicted. Figure 9 shows the computed results of this test with the training set size ranging from 30 to 240 samples. The percentages of accuracy represented in the diagram correspond to the average of the results obtained via the neural networks used in the different types of scenes.

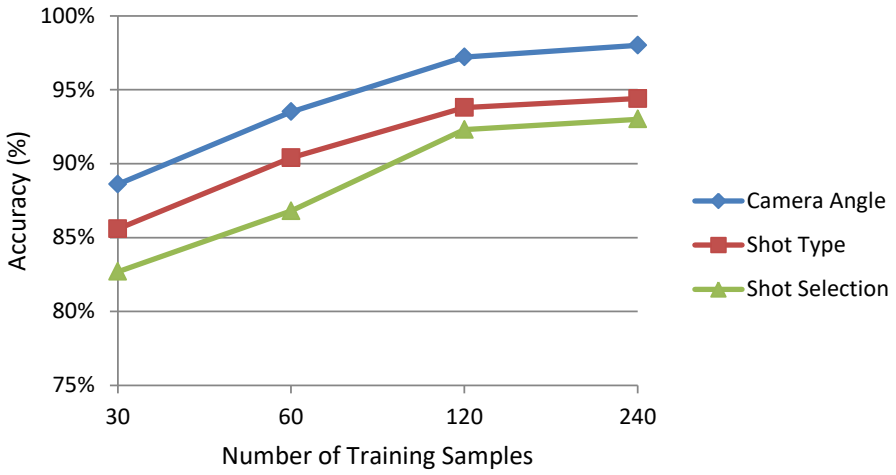


Fig. 9 Accuracy of the shot selection method with training sets ranging from 30 to 240 samples

The results of the accuracy test achieved by the shot selection method indicate the capacity of the method to learn and replicate the editing style of a human editor – as we mentioned before as one of our objectives. It is important to notice that we used training and testing samples generated by the same human editor. If we test the neural networks with samples generated by some other human editor, the accuracy will probably be lower, which is to be expected, since editors have their own individual style and preferences.

5.2. Visual Evaluation

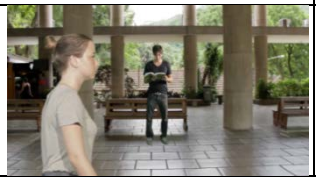
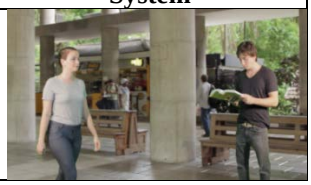
The second experiment to validate the proposed approach is a visual evaluation test, which concerns the overall aspects of the scenes composed by the system. In order to perform this test, we conducted an experiment comparing the results automatically produced by the system with the results manually produced by two teams of filmmaking professionals, where each team was composed of a film director and a video compositing professional. We selected a sequence of three basic actions and asked the two teams to compose the scene representing each of the basic actions. Then, we used our video-based dramatization system to generate the same sequence of basic actions. Both the teams and the system had available the same video resources to compose the frames. In order to perform the task, the human subjects decided to use the *Adobe After Effects CS6*. Table 1 shows the selected basic actions, including the logical description used by the dramatization system and the natural language description that was given to the human subjects.

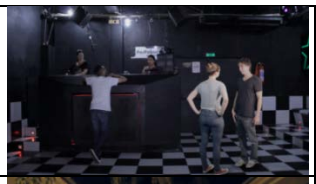
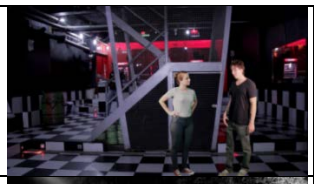
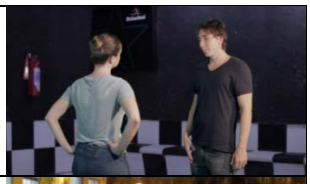
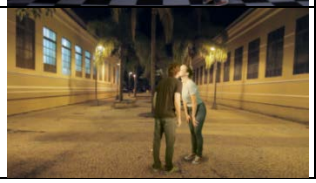
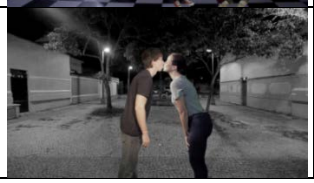
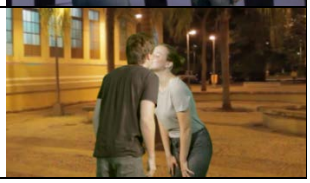
Table 1 Description of the selected basic actions used in the visual evaluation test

	Logical Description	**Natural Language Description**
Action 1	*GoIn*([*Anne*], [*University*])	"*Anne enters in the university where Peter is reading a book.*"
Action 2	*Tell*([*Peter*], [*S17*], [*Anne*], [*Nightclub*])	"*In the nightclub, Peter asks Anne if she likes to go out to parties.*"
Action 3	*Kiss*([*Peter*], [*Anne*], [*MainSquare*])	"*Peter kisses Anne in the Main Square.*"

The initial frames of the scenes composed by the human professionals and the initial frames automatically generated by the video-based dramatization system for the three selected basic actions are shown in Table 2.

Table 2 Visual comparison between the selected frames of the scenes composed by the human subjects and the corresponding frames automatically generated by the video-based dramatization system for the three basic actions

	Team 01	**Team 02**	**System**
Action 1			

Action 2			
Action 3			

During the experiment, we also recorded the time spent to complete the tasks. Table 3 shows a comparison of the time spent by the subjects to complete the composition of each basic action.

Table 3 Comparison between the time spent by the human professionals and the system to compose the scenes representing the three basic actions

	Team 01	**Team 02**	**System**
Action 1	36.20 (*min*)	29.16 (*min*)	3.55 (*sec*)
Action 2	50.47 (*min*)	26.22 (*min*)	2.42 (*sec*)
Action 3	20.16 (*min*)	39.17 (*min*)	2.04 (*sec*)

After completing the experiment, we conducted a simplified Turing-like Test to evaluate if human subjects would be able to differentiate the scenes created by the compositing algorithms and the scenes created by the filmmaking professionals. We asked 38 computer science students (29 male and 9 female, aged 18 to 26) to watch the video sequences of the three versions of each basic action (two created by the filmmaking professionals and one created by the system) and classify them according to whether they were created by humans or by a machine. The video sequences of each basic action were presented to the subjects in random order and without telling them how many sequences were created by a machine. Users were allowed to watch the videos how many times they wanted before classifying them. On average, each user spent 28.5 seconds (standard deviation of 8.4) to classify each video sequence.

Over the entire test set of 342 data points (38 users, each evaluating 9 videos), the “machine” version was correctly identified in 49 cases (out of a total of 114) and the “human” version was correctly identified in 121 cases (out of a total of 228), leading to an overall accuracy of 49.7%.

An ideal Turing Test is represented by the case where the computer and the human versions are indistinguishable, leading therefore to a random choice of 50% accuracy. The small difference between the achieved accuracy (49.7%) and the ideal Turing Test value (50%) suggests that the computer-generated and human-generated video sequences are hardly distinguishable, which is an indication of the capacity of the proposed automatic video compositing method to generate frames similarly to video compositing professionals. However, we must point out that, in this test, both filmmaking professionals and system had the same limited video resources available to compose the frames. This means that the filmmaking professionals were not allowed to shoot the video material and produce the scenes with the creativity and full liberty of choice that they would expect to enjoy in real film productions.

5.3. User Evaluation

To validate our approach, we developed a system prototype and used it to produce an experimental video-based, non-branching, interactive narrative called "*Modern Little Red Riding Hood*", which is an adaptation of the famous fairy tale *Little Red Riding Hood*. It is a modern, comic rendering of the original story, with funny and unexpected outcomes. The main characters of the narrative are: the girl called Little Red Riding Hood, her mother, her grandmother, and the Big Bad Wolf. The story takes place in three main locations: the Little Red Riding Hood's house, the forest, and the grandmother's house. The prototype is able to generate a fair number of diversified stories to comply with the desires of different users. In the more conventional stories, the narrative evolves following the traditional fairy tale plot, with the Big Bad Wolf tricking Little Red Riding Hood so as to be the first to arrive at her grandmother's house, eating the grandmother, and attacking the girl when she realizes what happened. In stories with a somewhat unconventional and awkward outcome, Little Red Riding Hood celebrates the death of her grandmother, and then shares her basket of goodies with the Big Bad Wolf. In stories with a sort of comic flavor, the Big Bad Wolf eats both Little Red Riding Hood and her grandmother, and then gets a stomach ache. Figure 10 shows a few scenes from our "*Modern Little Red Riding Hood*" and the interaction device based on text/voice commands used in the experiment[2]. Although our prototype is designed to offer additional ways to interact with the story through social networks, we did not consider these additional methods, because multimodal interaction is not in the scope of the present experiment.

Fig. 10 Scenes from "*Modern Little Red Riding Hood*" and the interaction device used in the experiment

For the production of "*Modern Little Red Riding Hood*", we produced a total of 648 film shots (referred here as "videos"), obtained by shooting films from 8 different angles: (1) 24 videos to represent the locations of the story; (2) 264 videos of Little Red Riding Hood's actions and dialogs; (3) 48 videos of the mother's actions and dialogs; (4) 56 videos of the grandmother's actions and dialogs; and (5) 256

[2] One of the story variants, with English subtitles, can be found as "Modern Little Red Riding Hood" in http://www.icad.puc-rio.br/~logtell/videos.php

videos of the Big Bad Wolf's actions and dialogs. Although this seems a huge amount of videos, it is important to notice that they are shot from multiple angles at the same time. The length of the finished videos totalizes 1 hour, 12 minutes and 30 seconds.

To assess the prototype, we conducted a user test based on the Schoenau-Fog method [33] to evaluate engagement in the interactive narratives. A total of 26 computer science students participated in the test. Twenty subjects were males and six females. Ages ranged from 18 to 25 years (mean of 20.8). Eighteen of them play video games at least weekly. None of them had previous experience with interactive storytelling systems.

We asked participants to watch and freely interact with the video-based interactive narrative. At the beginning of the story and after each chapter (total of 3), the dramatization was automatically paused and the participants were instructed by the system to answer a set of questions based on the Engagement Sample Questionnaire (ESQ) [33] with the addition of some questions regarding the user experience (curiosity, flow, and enjoyment). Each statement was followed by a seven-point Likert scale ranging from "strongly disagree" (-3) through "neutral" (0) to "strongly agree" (+3). After the last chapter, the final portion of the ESQ occurred, in which the participants were interviewed about their experience. On average, each session lasted 14.11 minutes (standard deviation of 4.32), and the average length of the stories generated by the participants is 3.28 minutes (standard deviation of 0.86).

When answering the question about the perceived time spent in the experience, the participants reported that they used an average of 8.4 minutes (standard deviation of 2.5). Of the total 26 participants, twenty-one (80.7%) wanted to do the experience again, mainly because they wanted to try out different choices and possibilities. Out of the five (19.3%) who did not want to try again, four (80%) thought the story was boring, mainly because they did not like the genre or found the story too childish. The other participant stated that the prototype had some technical problems, including random crashes that forced him to restart the story. We did not find necessary to separately indicate the average length of the stories generated by participants who found the story "boring", because it comes quite close to the general average. The possible variants of the story that can be generated by the system differ very little in terms of length. Thus, there seems to be no evidence that the "boring" label originated from stories either too short or too long.

The levels of continuation desire of three representative participants (P1, P2 and P3) and the average levels for all participants (Average) along the story chapters are shown on Figure 11.

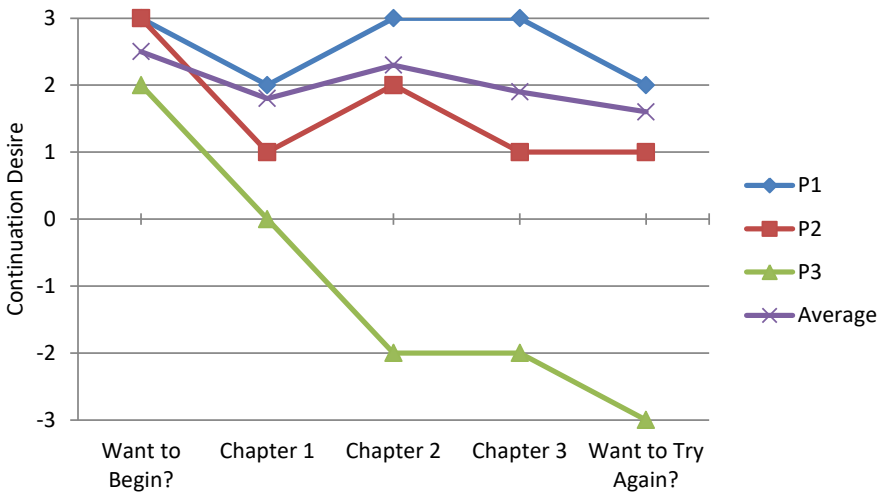


Fig. 11 Continuation desire levels

As far as the user experience is concerned, nineteen of the participants (73%) reported only positive feelings and affects during the experience (e.g. enjoyment, curiosity, excitement, pleasure). Seven (27%) mentioned some negative feelings (e.g. boredom, helplessness, frustration, confusion). Most of the negative feedback was in regards to users not liking the story and experiencing technical problems with the prototype.

During the interview, when questioned about the quality and realism of the video-based dramatization, fifteen participants (57.6%) indicated some lighting mismatches in the composition of some scenes with color differences between the actors and background. Twenty participants (76.9%) mentioned that they noticed the green background or a green tonality along the boundaries of the actors in some actions. Despite the alleged quality problems in the prototype, twenty-three participants (88.4%) declared that they would like to watch other interactive films, such as *Modern Little Red Riding Hood*.

Even though the results indicated that users were able to perceive some compositing problems during the dramatization, most of them expressed interest in continuing watching video-based narratives. This continuation desire is a fundamental element of the concept of play [41] and it is driven by the pleasure that emanates from the experience itself. This pleasure combined with the positive feelings reported by the majority of the participants is a direct indication of user engagement, which is a fundamental requisite for any successful interactive storytelling experience [33].

6. Concluding Remarks

Most of the previous works on interactive storytelling are about systems that either employ 2D/3D animation, or prerecorded video sequences that display branching narratives. Some of them explore more flexible ways of using video sequences [5, 6, 13, 20]. None of these studies, however, have proposed to create new scenes based on real-time video compositing and cinematography concepts.

In this paper, we present new algorithms and processes that combine expert knowledge on cinematography with a video compositing parallel architecture to guarantee real-time performance in video-based, non-branching, interactive storytelling. Our approach is not in competition with the

traditional sequential films, or even with branching interactive videos (such as [14]), especially with respect to professional-quality production process. What we purported to develop is a distinct form of storytelling, a novel type of multimedia content that provides a personalized way of enjoying stories, in which the audience is permitted to create their own, to some extent unexpected versions. Another interesting and scientifically novel part of the contribution of the present study is the use of neural networks to construct a cinematographic decision model. This method can also find economic success if future research expands it towards automated, real-time direction of CGI scenes.

The reason why the production process still remains rather limited, is that all basic actions and dialogs must be captured by camera shootings in advance. For our prototype application, we implemented 12 different types of scenes (dialoguing, walking, chasing, attacking, fighting, kissing, hugging, disguising, eating, releasing, celebrating, and escaping). Some of these scenes (e.g. the kissing scene in Table 2 (Action 3) and the hugging scene in Figure 10) involved complex interactions among the characters, which required the employment of compositing tricks (as usually done in the VFX industry [48]) and simplifications to allow the system to create them in real-time.

These basic actions can be combined with automatic shot transitions and automatic scene editing, which reduces the complexity of the process. More importantly, it should be noted that the underlying model of the system is extensible, in the sense that new types of scenes can be added, if so desired, and integrated to the current application. Furthermore, the extensibility of the scenes repertoire opens the possibility to try different genres (such as detective stories [46] and romances of chivalry [47]). As indicated in Section 4 – and as happened, in particular, with our Little Red Riding Hood fairy tale example – an application is delivered to the regular spectators (as we called the users that constitute the ultimate target audience of the system), only after the necessary basic actions are specified by film experts, who have the help of genre experts to cope with specific features of the genre involved.

The prototype application achieved its intended purpose of demonstrating the effectiveness of our approach (cf. Section 5), within the restrictions and limitations imposed by the available resources. With a larger budget for film production (i.e., by providing better lights and cameras, and by engaging more professionals to work in the post-production phase), some compositing problems could be avoided, such as the green tonality along the boundaries of the actors and lighting mismatches. In terms of efficiency, the process of editing and removing the background of the videos using a traditional chroma key matting technique requires a considerable amount of work in the post-production phase. To this end, the adoption of a faster matting process would be a better alternative. Despite these difficulties, the preliminary evaluation tests revealed a very encouraging result: the majority of the participants expressed interest in continuing watching this new form of video-based narratives while pointing out what should be improved.

Apart from the development and enhancement of prototypes, future work should mainly be directed toward problems related to the proposed approach itself. A major concern is the amount of narrative work for the production of the video material. For instance, in our prototype application, more than 1 hour of edited video material was necessary to generate narratives with the average length of 3.28 minutes. Although several actions can be reused to compose different scenes, filming them from 8 different angles

still generates a large number of video files, which grows in proportion to the number of different actions the characters can perform during the narrative. Restricting the branching storylines to conform to the conventions of the chosen genre may be a suitable way of controlling the amount of production work. In this regard, we may cite our recent studies of the recognized variants of Little Red Riding Hood, collected from oral storytellers [39][45]. The entire process can also be drastically improved if we use natural video matting (i.e., alpha matte computation from a video stream taken in a natural, uncontrolled environment) instead of blue-screen matting, for which we recommend the use of a camera array [34] or flash matting technique [35]. Moreover, natural video matting permits a much more comfortable environment for the actors, because the green screen is removed from the circle setup, as shown in Figure 3.

Another critical concern is the visual quality of the compositing results, which should come closer to the excellent quality of featured films. Improving the visual quality mostly depends on real-time techniques for realistic lighting, e.g., to illuminate the actors in harmony with the environment and to consider cast shadows and interreflections. This, however, remains an active field of research, and despite some promising results [29, 30, 31], interactive real-time rendering with complex illumination and materials is still an open issue.

While recognizing the production difficulties currently encountered by our experimental video-based implementation, for which we pointed out some promising solutions achievable with more ample resources, we claim that they do not undermine the positive feedback of our proposal. Indeed, the technique described in this paper still appears to be the most feasible way of extending non-branching interactive storytelling from 3D productions into interactive films. Overall, we believe that investing in video-based, non-branching, interactive narratives can significantly expand the present boundaries of traditional interactive films towards a more enjoyable form of digital entertainment.

Acknowledgements

We would like to thank CNPq (National Council for Scientific and Technological Development), FAPERJ (Carlos Chagas Filho Research Support Foundation of the State of Rio de Janeiro) and FINEP (Brazilian Innovation Agency), which belong to the Ministry of Science, Technology and Innovation, for the financial support. Also we are grateful to the anonymous reviewers for their valuable comments.

References

1. Mateas, M (2002) Interactive Drama, Art, and Artificial Intelligence. Doctoral Thesis. School of Computer Science, Carnegie Mellon University, Pittsburgh

2. Cavazza M, Charles F, Mead S (2002) Character-based interactive storytelling. IEEE Intelligent Systems, Special issue AI in Interactive Entertainment, 17 (4), pp 17-24

3. Pizzi D, Cavazza M (2007) Affective Storytelling Based on Characters' Feelings. In: AAAI Fall Symposium on Intelligent Narrative Technologies, Arlington, Virginia

4. Ursu MF, Kegel IC, Williams D, Thomas M, Mayer H, Zsombori V, Uomola ML, Larsson H, Wyver J (2008) ShapeShifting TV: Interactive screen media narratives, Multimedia Systems, 14 (2), pp 115-132

5. Porteous J, Benini S, Canini L, Charles F, Cavazza M, Leonardi R (2010) Interactive storytelling via video content recombination. In: Proceedings of the 18th ACM International Conference on Multimedia 2010, Firenze, Italy, pp 1715-1718

6. Piacenza A, Guerrini F, Adami N, Leonardi R, Porteous J, Teutenberg J, Cavazza M (2011) Generating Story Variants with Constrained Video Recombination. In: Proceedings of the 19th ACM International Conference on Multimedia 2011, Scottsdale, pp 223-232

7. Chua TS, Ruan LQ (1995) A Video Retrieval and Sequencing System. ACM Transactions on Information Systems, 13 (4), pp 373-407

8. Davenport, G., Murtaugh, M (1995) ConText Towards the Evolving Documentary. In: Proceedings of ACM Multimedia '95, San Francisco CA, 377-389

9. Ahanger G, Little TDC (1997) A System for Customized News Delivery from Video Archives. In: Proceedings of International Conference on Multimedia Computing and Systems, Ottawa, Canada, pp 526-533

10. Mateas M, Vanouse P, Domike S (2000) Generation of Ideologically-Biased Historical Documentaries. In: Proceedings of the Seventeenth National Conference on Artificial intelligence and Twelfth Conference on innovative Applications of Artificial intelligence. AAAI Press, pp 236-242

11. Bocconi S (2006) Vox Populi: generating video documentaries from semantically annotated media repositories. PhD Thesis, Technische Universiteit Eindhoven, Netherlands

12. Shen YT, Lieberman H, Davenport G (2009) What's Next? Emergent Storytelling from Video Collections. In: Proceeding of International Conference on Human factors in computing systems, ACM Press, pp 809-818

13. Müller W, Spierling U, Stockhausen C (2013) Production and Delivery of Interactive Narratives Based on Video Snippets. In: Proceedings of the 6th International Conference on Interactive Digital Storytelling, Istanbul, Turkey, pp 71-82

14. Jung Von Matt/Spree (2010) Last Call, Berlin. http://www.youtube.com/watch?v=qe9CiKnrS1w. Accessed 25 January 2016

15. Ciarlini AEM, Pozzer CT, Furtado AL, Feijó B (2005) A logic-based tool for interactive generation and dramatization of stories. In: Proceedings of the International Conference on Advances in Computer Entertainment Technology, Valencia, Spain, pp 133-140

16. Logtell (2016) Logtell Project Web Site. http://www.icad.puc-rio.br/~logtell/. Accessed 25 January 2016

17. Silva FGA, Ciarlini AEM, Siqueira SWM (2010) Nondeterministic Planning for Generating Interactive Plots. In: 12th Ibero-American Conference on AI, Bahía Blanca, Argentina, Springer, pp 133-143

18. Lima ES, Feijó B, Pozzer CT, Ciarlini AEM, Barbosa SDJ, Furtado AL, Silva FGA (2012) Social Interaction for Interactive Storytelling. In: Proceedings of the 11th International Conference on Entertainment Computing Bremen, Germany, pp. 1-15

19. Porter T, Duff T (1984) Compositing Digital Images, Computer Graphics, Vol. 18 (3), pp 253-259

20. Lima ES, Feijó B, Furtado AL, Pozzer C, Ciarlini A (2012) Automatic Video Editing For Video-Based Interactive Storytelling. In: Proceedings of the 2012 IEEE International Conference on Multimedia and Expo (ICME), Melbourne, Australia, pp 806-811

21. Foster J (2010) The Green Screen Handbook: Real-World Production Techniques, Sybex, Indianapolis, Indiana

22. Mascelli J (1965) The Five C's of Cinematography: Motion Picture Filming Techniques. Silman-James Press, Los Angeles

23. Brown B (2011) Cinematography: Theory and Practice: Image Making for Cinematographers and Directors. Focal Press, Waltham

24. Thompson R, Bowen C (2009) Grammar of the Shot, Focal Press, Burlington

25. Arijon, D (1976) Grammar of the Film Language. Silman-James Press, Los Angeles

26. Haykin SO (2008) Neural Networks and Learning Machines, 3rd Ed, Prentice Hall

27. Ekman P, Friesen WV (1971) Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17, pp 124-129

28. Brown M, Lowe D (2007) Automatic Panoramic Image Stitching Using Invariant Features. International Journal of Computer Vision, 74 (1), pp 59-77

29. Mehta S, Ramamoorthi R, Meyer M, Hery C (2012) Analytic Tangent Irradiance Environment Maps for Anisotropic Surfaces. Computer Graphics Forum, 31 (4), pp 1501-1508

30. Chabert CF, Einarsson P, Jones A, Lamond B, Ma WC, Sylwan S, Hawkins T, Debevec P (2006) Relighting human locomotion with flowed reflectance fields. In: ACM SIGGRAPH 2006 Sketches, New York

31. Ng R, Ramamoorthi R, Hanrahan P (2003) All-frequency shadows using non-linear wavelet lighting approximation. ACM Transactions on Graphics, 22 (3), pp 376-381

32. Lima ES, Feijó B, Barbosa SDJ, Furtado AL, Ciarlini AEM, Pozzer CT (2014) Draw Your Own Story: Paper and Pencil Interactive Storytelling. Entertainment Computing, 5 (1), pp 33-41

33. Schoenau-Fog H (2011) Hooked! – Evaluating Engagement as Continuation Desire in Interactive Narratives. In: Fourth International Conference on Interactive Digital Storytelling (ICIDS 2011), Vancouver, Canada, pp 219-230

34. Joshi N, Matusik W, Avidan S (2006) Natural video matting using camera arrays. ACM Transactions on Graphics, 25, pp 779-786

35. Sun J, Li Y, Kang SB, Shum H-Y (2006) Flash matting. ACM Transactions on Graphics, 25 (3), pp 772-778

36. Riedl MO, Young RM (2006) From Linear Story Generation to Branching Story Graphs. IEEE Computer Graphics and Applications, Vol. 26, No. 3.

37. Ghallab M, Nau D, Traverso P (2004) Automated Planning: Theory and Practice. Mor-gan Kaufmann Publishers, San Francisco, CA

38. Doria TR, Ciarlini AEM, Andreatta A (2008) A Nondeterministic Model for Controlling the Dramatization of Interactive Stories. In: Proceedings of the ACM Multimedia 2008 - 2nd ACM Workshop on Story Representation, Mechanism and Context - SRMC08, Vancouver, Canada, pp 21-26

39. Lima ES, Feijó B, Furtado AL (2015) Storytelling Variants: The Case of Little Red Riding Hood. Proceedings of the 14th International Conference on Entertainment Computing (ICEC 2015), Trondheim, Norway, pp 286-300

40. Liang C, Xu C, Cheng J, Min W, Lu H (2013) Script-to-Movie: A Computational Framework for Story Movie Composition. IEEE Transactions on Multimedia, 15(2), pp 401-414

41. Brown SL, Vaughan, CC (2009) Play: How it shapes the brain, opens the imagination, and invigorates the soul. Avery Publishing Group, NY

42. Lima ES, Feijó B, Furtado AL, Pozzer CT, Ciarlini AEM, Silva FG (2012) A Multi-User Natural Language Interface for Interactive Storytelling in TV and Cinema. In: Proceedings of the XI Brazilian Symposium on Computer Games and Digital Entertainment, Brasília, Brazil, pp. 154-161

43. Cavazza M, Lugrin J-L, Pizzi D, and Charles F (2007) Madame Bovary on the Holodeck: Immersive Interactive Storytelling. In: Proceedings of the 15th ACM International Conference on Multimedia, Augsburg, Germany, pp. 651-660

44. Mehlmann G, Endrass B, and André E (2011) Modeling and Interpretation of Multithreaded and Multimodal Dialogue. In: Proceedings of the 13th International Conference on Multimodal Interaction (ICMI 2011), pp. 385-392

45. Lima ES, Feijó B, Casanova MA, Furtado AL (2016) Storytelling variants based on semiotic relations. Entertainment Computing, 17, pp 31-44

46. Barbosa SDJ, Lima ES, Furtado AL, Feijó B (2014) Generation and Dramatization of Detective Stories. SBC Journal on 3D Interactive Systems, 5, pp 39-52

47. Barbosa SDJ, Silva FAG, Furtado AL, Casanova MA (2015) Plot Generation with Character-Based Decisions. Computers in Entertainment, 12, pp 1-21

48. Bredow R, Hastings A, Schaub D, Kramer D, Engle R (2015) From Mocap to Movie: The Polar Express. Course # 28, SIGGRAPH 2015, Los Angeles, CA