Multimodal, Multi-User and Adaptive Interaction for Interactive Storytelling Applications

Edirlei Soares de Lima[1] Bruno Feijó[1] Simone Barbosa[1] Fabio Guilherme da Silva[1]
Antonio L. Furtado[1] Cesar T. Pozzer[2] Angelo E. M. Ciarlini[3]

[1]PUC-Rio, Departamento de Informática, Brasil
[2]UFSM, Departamento de Eletrônica e Computação, Brasil
[3]UNIRIO, Departamento de Informática Aplicada, Brasil

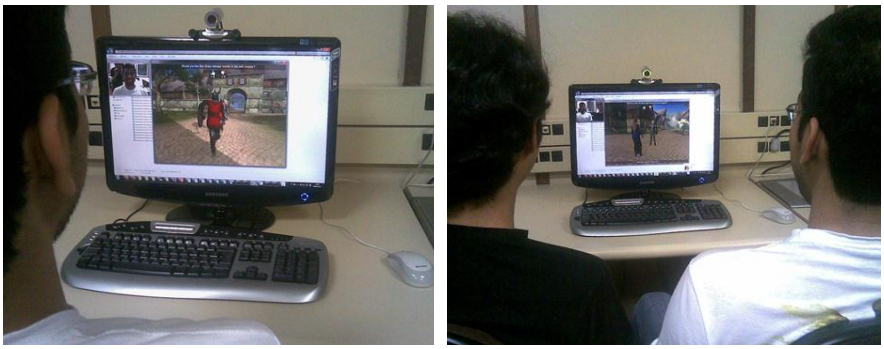

Figure 1: Users interacting with the multimodal, multi-user, and adaptive interaction system for interactive storytelling.

Abstract

The ability that users have to interact and change stories according to their own desires is what differentiates interactive narratives from conventional films. Moreover, this ability expands the boundaries of computer games towards new forms of digital entertainment. However, designing an interaction model for an interactive storytelling system involves several challenges, from the need for natural interaction interfaces to adequate multi-user settings. In this paper we present the development and evaluation of a multimodal, multi-user, and adaptive interaction system for an interactive storytelling application.

Keywords: Interactive Storytelling, Multimodal Interaction, Adaptive Interaction.

Authors' contact:
{elima, bfeijo, simone, faraujo, furtado }@inf.puc-rio.br,
pozzer@inf.ufsm.br,
angelo.ciarlini@uniriotec.br

1. Introduction

Interactive storytelling is a form of digital entertainment where authors, public, and virtual agents participate in a collaborative experience. Crawford [2004] defines interactive storytelling as a form of interactive entertainment in which the player plays the role of the protagonist in a dramatically rich environment. The experience offered to the public by an interactive story differs substantially from a linear story. An interactive story offers a universe of dramatic possibilities to the spectator. In this form of entertainment, the audience can explore an entire set of storylines, make their own decisions, and change the course of the narrative.

Typically, the way viewers interact with a storytelling system is directly linked to the story generation model: character or plot-based model. Character-based approaches [Cavazza et al. 2002][Young 2001][Aylett et al. 2006] give to the system great freedom of interaction. Usually, the story is generated based on the interactions between the viewer and the virtual characters. In some cases, the viewer can acts as an active character in the story. In plot-based approaches [Grasbon and Braun 2001][Ciarlini et al. 2005], the interaction options are quite limited. The users can perform only subtle interferences to guide the progress of the narrative plot.

The level of interaction in storytelling must be careful planned. Viewers should keep their attention on the narrative content and should not be distracted by the interaction interface. Another important aspect that must be considered during the design of an interaction model for an interactive storytelling system is the need of a multi-user interface. As in conventional TV and cinema, there may be more than one viewer watching the story at the same time. An interaction model must offer equal possibilities of interaction to all viewers. Another aspect that must be observed by an interaction system is the existence of several stereotypes of viewers. Some viewers like to interact actively with the story, others prefer to opine only on key points, while some prefer just to watch the story. The interaction system should adapt itself to the different types of viewers.

In this paper we present the development and evaluation of a multimodal, multi-user, and adaptive interaction system for an interactive storytelling application. The paper is organized as follows. Section 2 presents related works on multimodal interfaces, applications of interaction models, and the use of user stereotypes in interactive storytelling systems. Section 3 presents the architecture of our storytelling system. Section 4 presents the proposed multi-user and multimodal interaction system. A evaluation of the system is described in section 5. Finally, in section 6, we present the concluding remarks.

2. Related Work

Many works have already been done with the objective of using multimodal interfaces as a means of human-computer interaction. Cohen et al. [1989] show how the use of natural language together with gesture can overcome the limitations of each modality alone. The combination of speech and gesture provides a highly proficient communicative behavior to interact with applications in a more transparent experience than traditional GUI interfaces. In the field of virtual environments, Weimer and Ganapathy [1989] develop a virtual environment with speech and hand gesture input. These authors use a DataGlove for hand tracking where the thumb gestures are used to initiate a pick and the index fingertip is used like a stylus. Koons and Sparrell [1994] present an interface that let users interact with 3D objects in a virtual environment with speech and gestures using DataGloves. Lucente et al. [1998] presents a multimodal user interface where 3D objects shown on a wall-sized display were controlled by speech and natural gestures.

There are also some related works that use multimodal interfaces to human-computer interaction in the field of interactive storytelling applications. Dow et al. [2006] present an augmented reality version of the desktop based interactive drama Façade [Mateas 2002]. The players can move through a physical apartment and interact with two autonomous characters using gestures and speech. In a similar approach, Cavazza et al. [2004] present an interactive storytelling application that captures the user's video image and insert him in a virtual world populated by virtual actors. In the interaction system the users are able to interact with virtual actors using body gestures and natural speech. In other similar work, Cavazza et al. [2007] present an immersive interactive storytelling system where the users can interact with the virtual world using multimodal interaction. The verbal interaction is based on predefined speech sentences and the non-verbal interaction uses the user body orientation and distance from the virtual character to detect his attitude.

The use of a collaborative multimodal interaction model in an interactive entertainment application is explored by Tse et al. [2007]. These authors present a gaming interaction system based on a multi-touch table. These authors attempted to create a system that allows overlapping speech and gesture acts, as in the following example: "Put that" <points to an object> "there" <points to a place>. Maybes-Aminzade et al. [2002] present a set of techniques to enable members of an audience to participate, either cooperatively or competitively, in shared entertainment experiences. The techniques allow theater audiences to control onscreen activities by leaning left and right in the seats, batting a beach ball while its shadow is used as a pointing device, and pointing laser pointers at the screen. Carbini et al. [2006] present a cooperative storytelling application where user speech and gesture actions are interpreted in order to cooperatively build a story with another user. Kuka et al. [2009] present DEEP SPACE, a multi-user interactive storytelling system where the users can interact with the story drawing 2D objects that are transferred to the story as 3D objects and characters. Kurdyukova et al. [2009] present the evaluation of a multi-user interactive storytelling system where users are able to interact with virtual characters using various forms of interactions (cell phones, dance pads, Wiimotes and radio-frequency identifications (RFID)). The main problem revealed by their study was the disorientation caused by the multiple interaction options, the users frequently forgot the correct way of using the interaction devices.

The main difference between the interaction model presented in this work and the above-mentioned ones is the combination of a multimodal interface in a multi-user environment that automatically adapts the interaction options according to the user's preferences. Moreover, we also consider a simpler interaction interface that does not distract the viewers from the narratives and offers equal possibilities of interaction for several users at the same time.

3. System Architecture

The Logtell [Ciarlini et al. 2005] is an interactive storytelling system that focuses the logical coherence on its strategy of generating narratives. It is a plot-based system, but it uses some features of the character-based approach by using rules of inference of goals. These inference rules provide objectives to be achieved by the characters when certain situations are observed. The system has a client/server architecture (Figure 2) which supports multiple users sharing and interacting in the same or different stories. The client-side is responsible for user interaction and dramatization of stories. At the application server side there is a pool of servers sharing the responsibility of creating and controlling multiple stories, which are presented in different clients.

The idea behind Logtell is to capture the logics of a genre through a temporal logic model and then verify what kind of stories can be generated by simulation

combined with user intervention. In this way, Logtell does not simply focus on different ways of telling stories but on the dynamic creation of plots. The temporal logic model is composed of typical events and goal-inference rules. Plots are generated by multiple cycles of goal-inference, planning, and user intervention.

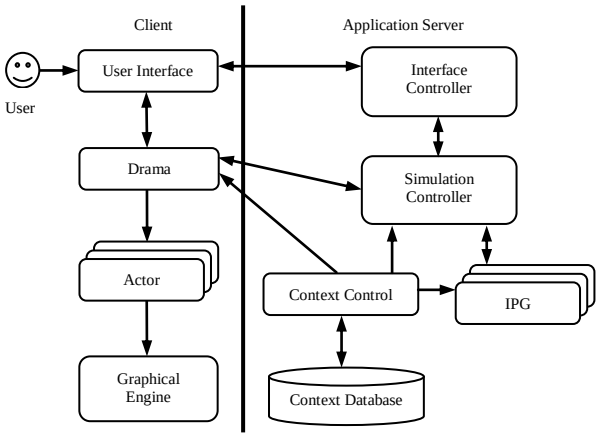

Figure 2: Logtell Architecture

The users can interact with the system suggesting events to story chapters. If the suggestion is considered valid according to the temporal logic model, the suggestion is allocated in the story plot. The plots are composed by ordered sequences of events generated in real-time by the planning algorithm [Furtado and Ciarlini 1999] based on the user intervention. Each plot event is modeled as a nondeterministic automaton [Doria et al. 2008], where situations observed in the world are associated to states, and micro-actions that virtual actors can perform are associated to the transitions. In general, there is always a set of states that can be reached after the execution of an action, the selection of which transition must occur could be a user choice or randomly chosen according to weights associated to the transitions. The user interaction in the micro-action level is considered a minor intervention, since this interaction does not necessarily change the story plot, except in some cases where the micro-action intervention leads to different final states in the automaton. Figure 3 shows an example of automaton created to represent the possibilities for the dramatization of an event where a villain kidnaps a victim.

The dramatization system represents the stories generated by the planning system in a 3D environment. The characters are represented through 3D models and their actions through animations. The system provides a set of parameterized actions that can be used to visually represent (dramatize) the generated stories. The dramatization system has the goal of emphasizing the dramatic content of the scenes and presents them in the most attractive and engaging way to the viewers. The architecture of the system is composed by a set of cinematography-inspired autonomous agents that controls the dramatization, actors, cameras, lights and music. The agents use emotional information of the actors and environment to emphasize the emotion of the scenes using cinematography techniques and concepts. Our cinematography-inspired dramatization system is described with more details on our previous works [Lima et al. 2009][Lima et al. 2010].

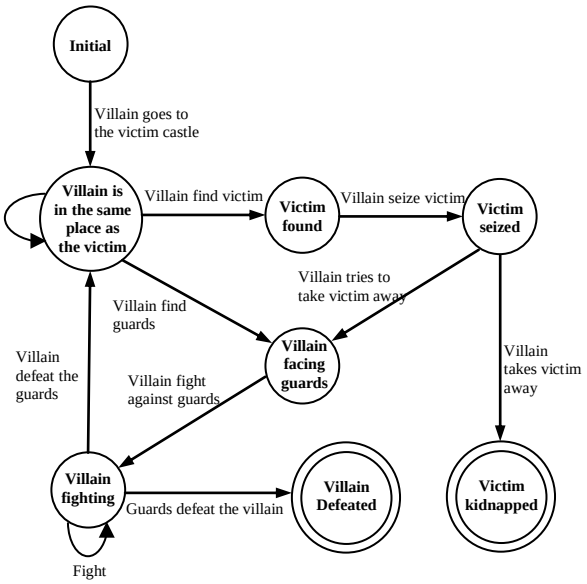

Figure 3: Example of automaton representing the possibilities for the dramatization of a kidnap event.

4. Multimodal, Multi-User and Adaptive Interaction Model

To design our multimodal, multi-user, and adaptive interaction model we follow some requisites for the design of multimodal interfaces described by Reeves et al. [2004]. Adapting some of these concepts to the interactive storytelling domain, we define the following requisites to our interaction system:

- **Natural Interaction:** The multimodal interaction must be natural. The viewers must feel comfortable interacting with the system;
- **Adaptable Interaction:** The multimodal interface must adapt itself to the needs and abilities of different viewers;
- **Consistent Interaction:** The result of an input shared by different interaction modalities must be the same;
- **Error Handling:** The system must prevent and handle possible mistakes in the interaction, as well allowing the viewers to easily undo their actions;
- **Feedback:** The system always must give a feedback to the viewers when some action resultant from a multimodal interaction be executed;
- **Equal Interaction:** In a multi-user scenario, the interaction system must offer equal possibilities of interaction to all viewers**.**

In our model, the multimodal interaction interface is based on gestures and speech. The choice of these interaction modalities was made due to the need of natural interaction modalities in a multi-user setting. Gestures and speech provide a natural interaction interface and allow the interaction of several users by using computer vision and speech recognition techniques. The viewers are free to use both interaction modalities.

The architecture of the interaction system presented in this paper is shown on Figure 4. The system uses a conventional camera and a microphone to capture the input of the system. The viewers are located on the video input by the Haar Classifier algorithm, and the viewer's speech is recognized by the SVM Classifier based on the audio input. The Interaction Interpreter module analyses and interprets the viewer's gestures and speech commands. Next, the Eigenfaces Classifier and the SVM GMM Classifier identify the viewer based on the profile of the viewers (which is stored in the Viewers Profiles Database). Each interaction is then recorded in the appropriated viewer's profile. The Profile Manager updates the viewer's profile based on the viewer's interactions and the atmospheric traits associated to the events as modeled in the Atmosphere Database. Before the viewer's interaction affects the system, the Interaction Validator module checks if the viewer is not interacting for the second time in the same option (for example, to avoid a viewer voting more than one time in the same option). Finally, the user interaction is sent to the Story Suggestion System.

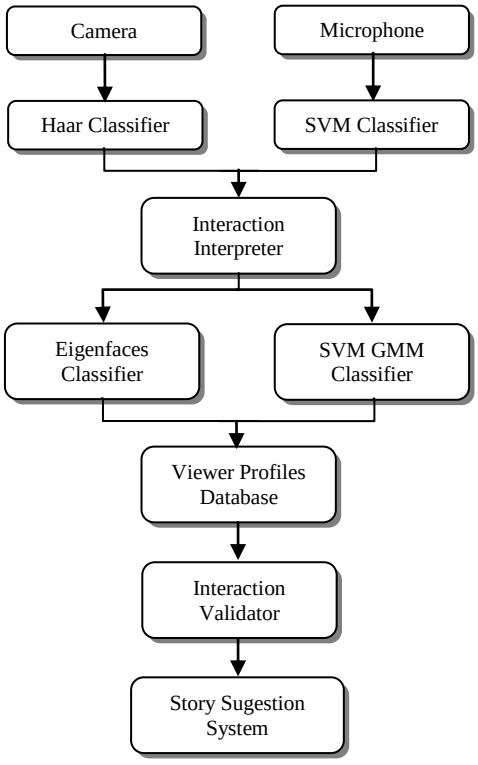

Figure 4: Multimodal, multiuser and adaptive interaction architecture.

4.1 Gesture Recognition

Gesture recognition provides a more natural and powerful way for human-computer interaction. Several studies have been done with the objective of creating systems able of understanding human gestures [Cohen et al. 1989][Weimer and Ganapathy 1989][Koons and Sparrell 1994][Lucente et al. 1998]. Currently, the entertainment industry is making the most significant use of such technology. Examples of applications include several games for the PlayStation Move [PlayStationMove 2010] and the Xbox Kinect [Kinect 2010] (previous called project natal). These new interactions forms are changing the way players interact with games.

Interacting with a game through natural body gestures involve movements like kicking, punching, jumping. These actions (in most part of the cases) are related to the game, where the player controls the virtual avatar using her/his own body. However, in an interactive storytelling application the users are treated more like viewers, that is, they usually watch the story from a third person point of view. This is even more visible in the Logtell, where the users interact with the story just giving suggestions about what should happen in the story chapters. The reason of such limited interaction is related to the nature of this form of entertainment: an interactive storytelling system focuses on the dramatic content of the story and not on creating challenges to the players.

Choosing appropriated and natural body gestures to interact with a multimodal interactive application is an important phase of the development process. Considering the interaction possibilities offered by Logtell system (where the viewers choose what should happens in the story from a set of suggestions) together with the need of natural body gestures, we decide to transform the set of suggestions in a set of single questions that ask if such event should happen in the next chapters. The viewers just need to approve or disapprove a suggestion. If they disapprove it, the next suggestion is asked. When the spectators approve a suggestion, the choice is sent to the story generator system. In this way, there is the need of only two body movements, one to approve and another to disapprove suggestions. The most natural body movements to approve/disapprove something are the head nodding and shaking. Head nods and shakes are very simple in the sense that they only provide yes/no, understanding/ misunderstanding, approval/disapproval meanings. However, their importance must not be underestimated because they meaning is almost universal [Mehrabian and Ferris 1967].

The interaction system uses a webcam to capture the environment and the spectators. The camera can be placed in any location where it can capture the face of the viewers (usually on the top of the TV). To recognize the head nods and shakes, the interaction system first needs to find the spectators on the camera

image. This is executed by the system using a machine learning approach for visual object detection. The Haar Classifier method is used to detect the spectator's faces on the camera frames. This method, proposed by Viola and Jones [2001] and improved by Lienhart and Maydt [2002], consists of a classification method created to detect rigid objects on images. The core basis for Haar classifier object detection is the Haar-like features. These features, rather than using the intensity values of a pixel, use the change in contrast values between adjacent rectangular groups of pixels. The contrast variances between the pixel groups are used to determine relative light and dark areas. Two or three adjacent groups with a relative contrast variance form a Haar-like feature. Haar features can easily be scaled by increasing or decreasing the size of the pixel group being examined. This allows features to be used to detect objects of various sizes. The great advantage of this classifier is that it quickly rejects regions where the probability of finding an object is low. About 70% to 80% of images that do not contain the searched object are rejected in the first two interactions. Moreover, according to Bradski and Kaehler [2008], the method has high recognition rates with few false positives and few false negatives.

The Haar Classifier algorithm recognizes and estimates the position and size of all faces in the analyzed images (as illustrated on Figure 5). The interaction system uses the Haar Classifier algorithm in each frame captured by the camera. Each face found in the image can be considered a spectator. However the Haar Classifier is not able to identify the spectator, the algorithm just detects human faces. So, the Haar Classifier does not solve the problem completely. The interaction system must know the spectators to adapt itself to the spectator's stereotype. To solve this problem, the interaction system incorporates another statistical classifier to identify the spectators. The principal component analysis method (PCA), also called Eigenface [Turk and Pentland 1991] is the most common algorithm used for face identification. The method is based on statistical data extracted from train images. PCA is considered one of the techniques that provide the best performance [Zhang et al. 1997]. The main idea of the PCA is to obtain a set of orthogonal vectors (Eigenfaces) that optimally represent the distribution of the pixel intensities. Once the corresponding Eigenfaces are computed, they are used to represent the training dataset and are used to identify the same faces in other images.

The spectator identification process consists of two steps. First, the Haar Classifier algorithm finds the viewer's faces in the camera image, and then the Eigenfaces algorithm is used to identify these viewers. The training process of the Eigenfaces classifier is done in real time. At the beginning of a new story, the system identifies possible new viewers and asks for their names. Then the system captures some training samples of the new users' faces and save the Eigenface data in the training dataset.

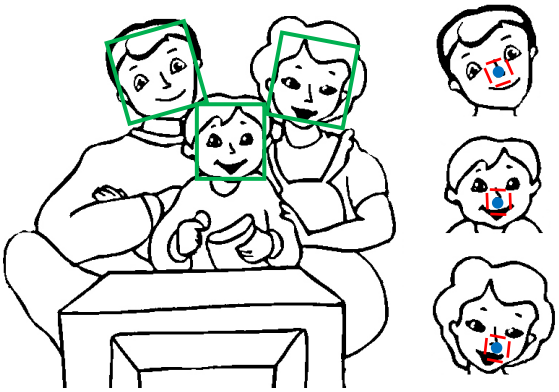

Figure 5: Illustration of the position and size of all viewers' faces recognized by the Haar Classifier algorithm

To recognize the head gestures, the interaction system uses dynamic interaction points. These points consist of four planes (up, down, left and right) created around the center of the spectators head (Figure 6). The position of the dynamic interaction points is updated when the spectator head stay approximately in the same position for more than one second. The dynamic interaction points are used to detect the head shaking and nodding. When the center point of the head hits the left and the right interaction points in a short period of time, the head shake is detected. When the center point hits the up and the down interaction points, the head nod is detected.

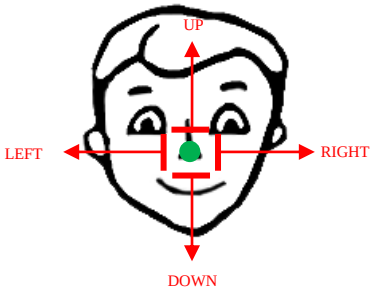

Figure 6: Illustration of the dynamic interaction points used by the system to recognize the viewer gestures.

The interaction system analyses and interprets the gesture of all viewers in real time. The viewer identification allows the system to ensure equal interaction possibilities for all users. For example, the same user cannot vote two times for the same story suggestion, and if the user chooses another suggestion, the system can change the viewer vote. Moreover, the viewer identification also allows the system to analyze and register all the interaction choices for each viewer; in this way, the system is able to learn the viewer's preferences.

4.2 Speech Recognition

Speech is a powerful human communication skill and also can be used as a natural form of human-computer interaction. There are several studies in the field of speech recognition applied as a form of human-

computer interaction [Cohen et al. 1989][Weimer and Ganapathy 1987][Koons and Sparrell 1994][Lucente et al. 1998]. In those studies is possible to distinguish three areas of research: 1) isolated word recognition, where words are separated by distinct pauses; 2) continuous speech recognition, where sentences are produced continuously in a natural manner; and 3) speech understanding, where the aim is not on the transcription but understanding of the speech.

As mentioned earlier, we decide to transform the Logtell suggestions mechanism of story events in a set of questions that ask if an event should happen in the next chapters. The spectators just need to approve or disapprove a suggestion. This simplified interaction is also useful for the speech recognition system, the system just needs to understand two words: "yes" and "no". These two words are quite distinct, which ensures high correct recognition rates. The recognition and understanding of isolated words is one of the most basic approaches to speech recognition; however it is totally adequate for the context of this work.

The approach to speech recognition adopted by the present work is based on a supervised machine learning method used to classify and recognize the input speech based on a set of features extracted from the audio signal. More specifically, we train a support vector machine (SVM) classifier with audio signal features and use its classification to recognize unknown input signals.

SVM, proposed by Vapnik [1995], is a method for pattern recognition of general purpose and it's indicated for small sample sets. It consists of a supervised learning method that tries to find the biggest margin to separate different classes of data. Kernel functions are employed to efficiently map input data, which may not be linearly separable, to a high dimensional feature space where linear methods can then be applied. The essence of the SVM method is the construction of an optimal hyperplane, which can separate data from opposite classes using the biggest possible margin.

To use the SVM classifier in the speech recognition system is necessary to follow two steps. The first phase is the supervised training process, where features of words samples are provided to the classifier. The second step is the prediction process, where the knowledge acquired through the training process is used to classify an unknown input.

The first step in the training process is the pre-processing phase; in this phase, the audio is normalized to eliminate large variations in the signal amplitude. The audio samples are also discretized to reduce the number of samples along the signal. After the pre-processing phase, the feature extraction process is performed. The audio samples are divided into 49 frames and, for each frame and for the entire signal, the amplitude average and its standard deviation are calculated. The total of 100 features is extracted from the audio signal. The vector of features extracted from the signal constitutes one training sample. To create a training database several training samples are extracted from several input signals. In our case, the input audio signals are composed by several recordings of different people saying "yes" and "no".

The SVM classifier is able to recognize the words "yes" and "no", but it is no able to identify the speaking viewer. As occurs with the gesture recognition, the speech recognition system must recognize the viewers that are interacting with the system to adapt the interaction to the user's preferences. To identity the viewer's speaking we adopt an approach also based on a support vector machine classifier, but using a Gaussian Mixture Model (GMM) supervector [Reynolds et al. 2000]. The GMM model provides a text-independent speaker identification methodology with high recognition rates [Campbell et al. 2006]. The training process of the speaker identification classifier is done in real-time together with Eigenfaces training process. At the beginning of a new story, the system identifies possible new viewers by their faces using the Eigenfaces algorithm. While the system is capturing some training samples for the Eigenfaces algorithm, the speech recognition system asks some questions (that must be answered using the microphone) to the new viewers. The system extracts some features from the viewer voice and stores them in a training dataset. This training dataset is associated to the same user profile used by the gesture recognition system.

With the first SVM trained to recognize the words "yes" and "no" and the second SVM trained to identify the viewers by their voice, the speech recognition system is able to be used by the viewers to interact with the Logtell stories. The speech capture and recognition is performed in real time through a microphone. To prevent the processing of noise from the environment, the system continuously analyzes the audio input and only performs the classification of signals whose amplitudes are larger than the sound of the environment. When a large variation in the audio amplitude is detected, the system starts recording the input signal. When the amplitude of the input signal returns to the environment level, the system pre-processes the signal and performs the feature extraction. The extracted features are then sent to the classifier to determine the class of the input signal.

4.3 Adaptive Interaction

The goal of an adaptive system is to adapt the application interface to a specific user based on the user's behavior, goals, preferences, and actions [Jaimes and Sebe 2007]. The objective of our adaptive interaction model is to adapt interaction options to viewers' preferences. As mentioned earlier, the public who watches an interactive story is diversified. Simplifying this idea, it is possible to distinguish three

types of viewers clearly: (1) the viewers that like to interact actively with the story; (2) the viewers that prefer to opine only on key points; and (3) the viewers that just watch the story. The interaction possibilities offered by our interactive storytelling system also can be dived in two types: (1) the plot level suggestions, which are allocated directly in the story plot; and (2) the micro-action level suggestions, which influence the scene events. The viewers are asked frequently to interact with the micro-action suggestions, while the plot level suggestions occur only in key points. Considering these options, it's possible to model three stereotypes of viewers:

1) **Active viewers:** Viewers that like to interact actively with the story. These viewers are asked for both plot level and micro-action level suggestions.
2) **Conventional Viewers:** Viewers that like to interact only in key points. These viewers are asked only for plot level suggestions.
3) **Passive viewers:** Viewers that like only to watch the stories without interact with them. These viewers are not asked to interact with the story.

The system uses the history of interactions to classify the viewers within adequate stereotypes. As mentioned early, the system is able to identify the viewers by their faces using computer vision techniques. In this way, the system associates each viewer to a unique user profile. All interactions of the viewers are recorded in the user profiles. The history of interactions is stored by the system for all dramatizations sessions. All the viewers who watch a story will have a profile. Analyzing the viewer's history of interactions it's possible to determine their preferences of interactions and classify each viewer in one of the viewer's stereotypes. Viewers that interact in more than 10% of the micro-action level suggestions are classified in the "active viewers" stereotype. Viewers that have 10% or less of interventions in the micro-action level suggestions, and still have more that 10% of interventions in the plot level suggestions, are classified in the "conventional viewers" stereotype. The viewers that have less than 10% of interventions in the micro-action level suggestions and less that 10% of interventions in the plot level suggestions are classified as a "passive viewer" stereotype.

The viewer's stereotype is updated during the dramatization; this means that the stereotypes can change during the dramatization sessions. Active viewers can become conventional viewers if they completely stop interacting with the micro-action level suggestions. Likewise, conventional viewers can become passive viewers if they completely stop interacting with the plot level suggestion. When the viewer watching the story is classified as a conventional viewer, the system sometimes stochastically suggests a micro-action level suggestion. If the viewer interacts with the micro-action suggestion, the system continues suggesting micro-actions until the viewer stops interacting with the micro-action suggestions. This allows the system to adapt itself if a conventional viewer suddenly becomes an active viewer. The same is done with the passive viewers; sometimes the system creates plot level or micro-action suggestions to check if the viewers really don't want to interact with the story.

When a group of viewers is interacting with the story the system adopts the stereotype of the most active viewer. In this way, the active viewer acts as a leader to the group, encouraging the other viewers to interact with the story.

5. Evaluation

To evaluate the interaction system presented on this paper, we performed two tests: a technical test to check the performance and accuracy of the system, and then a user evaluation test to check the system's usability from a Human-Computer Interaction (HCI) perspective. The following sections describe these tests.

5.1 Technical Evaluation

Although the user evaluation is the most significant form of evaluating an interaction system, a technical evaluation can't be discarded. It is important to evaluate the overall computational performance of the system, especially when the system is based on computer vision and signal processing techniques that do not ensures correct classifications all the time. Moreover, some of these techniques require high computational processing that must be done in real time. In this section we present some tests that we performed to technically evaluate the multimodal, multi-user, and adaptive interaction model presented in this paper.

To evaluate our gesture recognition system, we simulate three scenarios with different number of viewers and in different environments. In each simulation the viewer's perform the interaction gestures for approximately 2 minutes. Each simulation was recorded in a video and used as the input to the gesture recognition system. Then we computed the correct and wrong results of the gesture recognition algorithms. False positives and false negatives are also considered wrong classifications and are included in the overall recognition rate. The result of this test is shown in table 1.

To evaluate the computational performance of our gesture recognition system we use the same video simulations of the previous experiment. However instead of calculating the recognition rate, we compute the average time necessary to process the classification algorithms. The test was ran in an Intel Core i7 2.66 GHZ CPU, 8 GB of RAM using a single core to process the algorithms. The result of this test is shown on table 2.

Table 1. Gesture Recognition Rate

	Number of Viewers	Haar C. Recognition Rate	Eigenfaces Recognition Rate	Gestures Recognition Rate
Video A	1	100 %	100%	100%
Video B	2	97.8%	98,3%	98,4%
Video C	3	96.2%	92,9%	99,2%

Table 2. Gesture Recognition Performance

	Number of Viewers	Haar C. Recognition Time (*ms*)	Eigenfaces Recognition Time (*ms*)	Total Time (*ms*)
Video A	1	150.99	6.86	157,85
Video B	2	158.28	8.42	166,7
Video C	3	164.63	8.87	173,5

Similar tests were applied to evaluate the speech recognition system. We use audio from the previous scenarios as the input to the speech recognition system and then compute the correct and wrong results. The result of the recognition rate test is shown on table 3. We also computed the average time necessary to process the classification algorithms. The result of computational performance of the speech recognition system is shown on table 4.

Table 3. Speech Recognition Rate

	Number of Viewers	SVM Recognition Rate	SVM/GMM Recognition Rate
Video A	1	96.1%	100%
Video B	2	95.7%	92.4%
Video C	3	94.8%	83.9%

Table 4. Speech Recognition Performance

	Number of Viewers	SVM Recognition Time (*ms*)	SVM/GMM Recognition Time (*ms*)	Total Time (*ms*)
Video A	1	18.23	42.83	61.06
Video B	2	17.89	44.02	61,91
Video C	3	18.44	43.74	62,18

The high accuracy in the predicted gestures and voice commands indicates that the system (in most cases) executes correctly the viewer's interactions. The lower processing time allows the interaction system to be executed in real time.

5.2 User Evaluation

To effectively evaluate our interaction system, we have conducted a preliminary user evaluation with nine participants, six male and three female, all between 20 and 28 years old, with diverse backgrounds: two cinema professional, four graduate students and two undergraduate students in Computer Science, and a graduate student in Fine Arts. To evaluate the interaction system in multi-user settings, they were divided in three groups of three participants.

We asked the groups of participants to interact with two versions of our interactive storytelling system, one based on a traditional GUI interface and the other using the multimodal interface. In order to reduce learning effects, two of the groups used the traditional GUI interface first, and the other one used the multimodal interface first.

After using each version, the participants filled out a questionnaire with 10 questions about their motivation to interact with the story, their understanding of how to do so, the effort to do so, and how they influenced the story. After having interacted with both systems, the participants were interviewed about their preferences and experiences using the systems.

Figure 8 summarizes the results of the questionnaire. As can be seen, in this preliminary evaluation, the multimodal interface has shown that it requires more efforts by the participants, but also increased their motivation to interact with the story. As for the interviews, all participants stated they preferred to interact with the multimodal version, because it was more interesting, attractive and allows the groups to have more freedom to interact with the stories, despite the slightly increased effort, mostly due to some limitations of the speech recognition algorithms. They also pointed out the "sense of competition" that emerges when some participants are trying to guide the story to specific ending and the others want a different ending.

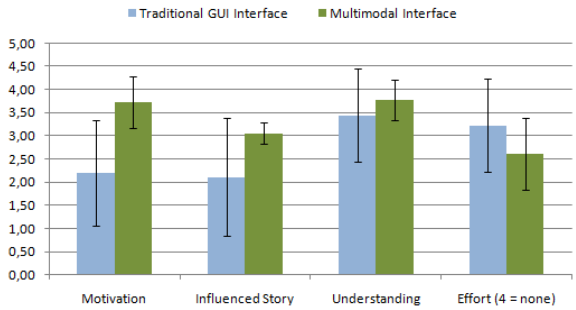

Figure 7: Averages and standard deviation of questionnaire topics in both versions of the system.

Although the quantitative results are inconclusive due to the small number of participants, the increased motivation and interest in influencing the story, especially expressed in the interviews, indicate that this is a promising direction of research.

6. Conclusion

In this paper we present the development and evaluation of a multimodal, multi-user, and adaptive interaction system for an interactive storytelling application. In our tests, the proposed model showed good results and fulfilled the interaction requisites of our interactive storytelling system. We believe that the main contribution of this work is the interaction model that combines the simple and natural multimodal interaction interface in a multi-user setting that automatically adapts the interaction options according to the user's preferences.

The multimodal interface allows the free interaction of several users simultaneously. However some limitations must be pointed: first, the number of users interacting with the system using gestures is limited by the field of view of the camera. A similar limitation occurs with the speech interaction; there is always a maximal distance where the microphone can capture the viewer's voice. Another limitation that can be observed is that the speech recognition system does not recognize correctly the voice interactions when more than one user is trying to interact at the same time. The recognition of speech from several users at same time is an open field of research in the area of signal processing. The identification of the viewers based on their voice also has some limitations. Humans can easily change the tone of their voices, and when the change is significant, the system will not identify the viewer. This trick can be used by the viewers to generate more votes in the desired choices. In future works we intend to improve our interaction system to overcome some of these limitations and conduct a more expressive user studies to effectively evaluate the usability of the system.

Acknowledgements

This work was supported by CAPES through the project RH-TV-Digital 133/2008, and CNPq.

References

AYLETT, R., DIAS, J., PAIVA, A., 2006. An affectively driven planner for synthetic characters. *In: Proceedings of 16th International Conference on Automated Planning and Scheduling*, pp. 2-10.

BRADSKI, G., KAEHLER, A., 2008. Learning OpenCV: Computer Vision with the OpenCV Library, O'Reilly.

CAMPBELL, W.M., STURIM, D.E., REYNOLDS, D.A., 2006. Support vector machines using GMM supervectors for speaker verification. Signal Processing Letters, IEEE, Vol. 13, No. 5. pp. 308-311.

CARBINI, S., DELPHIN-POULAT, L., PERRON, L., VIALLET., J.E., 2006. From a Wizard of Oz experiment to a real time speech and gesture multimodal interface. Signal Processing, (12), 3559-3577.

CAVAZZA, M., CHARLES, F., MEAD, S. J., MARTIN, O., MARICHAL, X. AND NANDI A., 2004. Multimodal acting in mixed reality interactive storytelling. IEEE Multimedia, pp. 30-39.

CAVAZZA, M., CHARLES, F., MEAD, S., 2002. Character-based interactive storytelling. IEEE Intelligent Systems. Volume 17, Issue 4, pp. 17-24.

CAVAZZA, M., LUGRIN, J.L., PIZZI, D., CHARLES, F., 2007. Madame Bovary on the Holodeck: Immersive Interactive Storytelling. *In: ACM Multimedia 2007*, pp. 651-660.

CIARLINI, A.E.M., POZZER, C.T., FURTADO, A.L., 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, p. 133-140.

COHEN, P. R., DALRYMPLE, M., MORAN, D. B., PEREIRA, F. C., SULLIVAN, J. W., JR, R. A. G., SCHLOSSBERT, J., AND TYLER, S. W., 1989. Synergistic use of direct manipulation and natural language. *In: Proceedings of the SIGCHI conference on Human factors in computing systems: Wings for the mind*, pp. 227-233.

CRAWFORD, C., 2004. Chris Crawford on Interactive Storytelling. Berkeley, Estados Unidos, New Riders.

DORIA, T.R., CIARLINI, A.E.M., ANDREATTA, A., 2008. A Nondeterministic Model for Controlling the Dramatization of Interactive Stories. *In: Proceedings of the 2nd ACM Workshop on Story Representation, Mechanism and Context.* Vancouver, Canada, pp. 21-26.

DOW, S., MEHTA, M., LAUSIER, A., MACINTYRE, B., AND MATEAS, M., 2006. Initial Lessons from AR-Façade, An Interactive Augmented Reality Drama. *In: ACM SIGCHI Conference on Advances in Computer Entertainment (ACE'06),* Los Angeles, CA.

FURTADO, A., CIARLINI, A., 1999. Operational Characterization of Genre in Literary and Real-Life Domains. *In: Proceedings ER'99 Conceptual Modeling Conference*, pp. 460-474.

GRASBON, D., BRAUN, N., 2001. A morphological approach to interactive storytelling. *In: Proceedings of the 2001 Living in Mixed Realities*, pp. 337-340.

JAIMES, A., SEBE, N., 2007. Multimodal Human Computer Interaction: A Survey, *In: Proceedings of 11th IEEE International Workshop on Human Computer Interaction (HCI)*, pp. 116-134.

KINECT, 2010. Xbox. Available from: http://www.xbox.com/en-US/kinect [Accessed in 26 nov. 2010].

KOONS, D. AND SPARRELL, C., 1994. ICONIC: speech and depictive gestures at the human-machine interface. *In: Proceedings of CHI '94: Conference companion on Human factors in computing systems*, pp. 453-454.

KUKA, D., ELIAS, O., MARTINS, R., LINDINGER, C., PRAMBÖCK, A., JALSOVEC, A., MARESCH, P., HÖRTNER, H. BRANDL, P., 2009. DEEP SPACE: High Resolution VR Platform for Multi-user Interactive Narratives. *In: Proceedings of the 2nd Joint International Conference on Interactive Digital Storytelling: Interactive Storytelling*, pp. 185-196.

KURDYUKOVA, E. ANDRÉ, E. LEICHTENSTERN, K., 2009. Introducing Multiple Interaction Devices to Interactive Storytelling: Experiences from Practice. *In: Proceedings of the 2nd International Conference on Interactive Digital Storytelling (ICIDS)*, pp. 134-139.

LIENHART, R., MAYDT, J., 2002. An Extended Set of Haar-like Features for Rapid Object Detection, EEE ICIP, vol. 1, pp. 900-903.

LIMA, E.S., FEIJÓ, B., FURTADO, A.L., POZZER, C.T., CIARLINI, A., 2010. Director of Photography and Music Director for Interactive Storytelling. *In: Proceedings of IX Brazilian Symposium on Games and Digital Entertainment*, pp. 129-137.

LIMA, E.S., POZZER, C., ORNELLAS, M., CIARLINI, A., FEIJÓ, B., FURTADO, A., 2009. Virtual Cinematography Director for Interactive Storytelling. *In: Proceedings of the International Conference on Advances in Computer Entertainment Technology*. Greece.

LUCENTE, M., ZWART, G. J., AND GEORGE, A. D., 1998. Visualization Space: A Testbed for Deviceless Multimodal User Interface. *In: Proceedings of AAAI Spring Symposium on Intelligent Environments*, pp. 98-02.

MATEAS, M., 2002. Interactive Drama, Art, and Artificial Intelligence. Ph.D. Thesis - School of Computer Science, Carnegie Mellon University, Pittsburgh, United States.

MAYNES-AMINZADE, D., PAUSCH, R. AND SEITZ, S. 2002. Techniques for Interactive Audience Participation. *In: Proceedings IEEE International Conference on Multimodal Interfaces (ICMI).*

MEHRABIAN, A., FERRIS, S.R., 1967. Inference of attitude from nonverbal communication in two channels. Journal of Counseling Psychology 31(3), 248-252.

PLAYSTATIONMOVE, 2010. Motion Controller. Available from: http://us.playstation.com/ps3/playstation-move/ [Accessed in 26 nov. 2010].

RAUSCHERT, I., AGRAWAL, P., SHARMA, R., FUHRMANN, S., BREWER, I., MACEACHREN, A., WANG, H., AND CAI, G., 2002. Designing a human-centered, multimodal GIS interface to support emergency management. *In: Proceedings of the 10th ACM international symposium on Advances in geographic information systems.* McLean, USA, pp. 119-124.

REEVES, L.M., MARTIN, J.C., MCTEAR, M., RAMAN, T., STANNEY, K., SU, H., WANG, Q., LAI, J., LARSON, J., OVIATT, S., BALAJI, T., BUISINE, S., COLLINGS, P., COHEN, P., KRAAL, B., 2004. Guidelines for multimodal user interface design, Communications of the ACM 47 (1), pp. 57-69.

REYNOLDS, D.A., QUATIERI, T.F., DUNN, R.B., 2000. Speaker Verification Using Adapted Gaussian Mixture Models, Digital Signal Processing, Volume 10, Issues 1-3, pp. 19-41.

SOWA, T. AND WACHSMUTH, I., 2003. Coverbal Iconic Gestures for Object Descriptions in Virtual Environments: An Empirical Study. *In: Proceedings of Gestures. Meaning and Use*, pp. 365-376.

TSE, E., GREENBERG, S., SHEN, C., 2006. GSI demo: multiuser gesture/speech interaction over digital tables by wrapping single user applications. *In: Proceedings of the 8th International Conference on Multimodal Interfaces.* pp. 76-83.

TSE, E., GREENBERG, S., SHEN, C., FORLINES, C., 2006. Multimodal Multiplayer Tabletop Gaming. *In: Proceedings Third International Workshop on Pervasive Gaming Applications (PerGames'06).*

TURK, M.A., PENTLAND, A.P., 1991. Face recognition using eigenfaces. *In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Patter Recognition*, pp. 586-591.

VAPNIK, V., 1995. The Nature of Statistical Learning Theory. New York, Estados Unidos: Springer.

VIOLA, P., JONES, P., 2001. Rapid object detection using a boosted cascade of simple features. *In: Proceedings of the CVPR01*, pp. 511–518.

WEIMER, D. AND GANAPATHY, S. K., 1989. A synthetic visual environment with hand gesturing and voice input. *In: Proceedings of the SIGCHI conference on Human factors in computing systems: Wings for the mind*, pp. 35-240.

YOUNG, M., 2001. An Overview of the Mimesis Architecture: Integrating Intelligent Narrative Control into an Existing Gaming Environment. *In: Working Notes of the AAAI Spring Symposium on Artificial Intelligence and Interactive Entertainment.*

ZHANG, J., YAN, Y., LADES, M., 1997. Face recognition: eigenface, elastic matching, and neural nets. *In: Proceedings of the IEEE*, Vol. 85, No. 9, pp. 1423-1435.