8
Application and Evaluation

Two prototype interactive narratives were produced to validate the proposed approach to video-based interactive storytelling: “*The Game of Love*” and “*Modern Little Red Riding Hood*”.

The Game of Love pertains to a romantic drama genre and tells the story of a young boy named Peter, who falls in love with an unknown girl and tries to do anything to get closer to her. The main characters of the narrative are: the young lover, Peter; the unknown girl, Anne; Anne’s best friend, Carol; and two imaginary creatures, a little angel and a little devil. The story takes place in six main locations: a university, Peter’s house, Anne’s house, a party, a beach, and the city square. In the main storyline, Peter falls in love with Anne at the university and tries to know more about her by hacking her Facebook page. After getting some information, Peter manages to go out with Anne on a date, but she find out that he invaded her social network account. Users are able to influence the decisions made by the main characters and change the future of the young couple. Figure 8.1 shows some scene from “*The Game of Love*”.

PUC-Rio - Certificação Digital Nº 1021793/CA

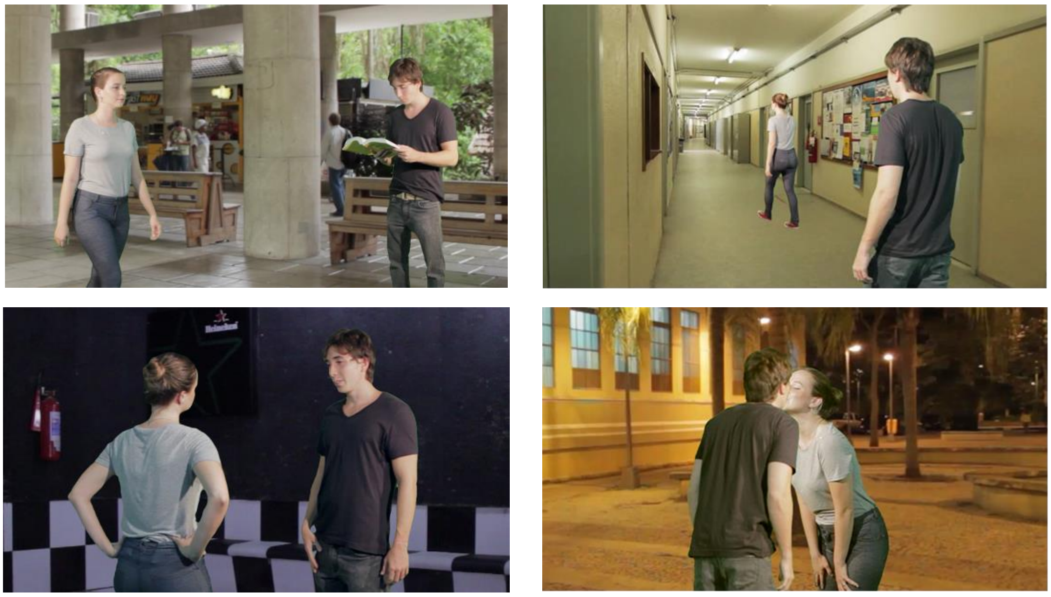


Figure 8.1: Scenes from “The Game of Love”.

The second prototype video-based interactive narrative developed, "*Modern Little Red Riding Hood*", is an adaptation of the famous Little Red Riding Hood fairy tale. It tells a modern and comic version 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, the Big Bad Wolf, and the woodcutter. The story takes place in three main locations: the Little Red Riding Hood house, the forest, and the grandmother's house. The prototype is able to generate a considerable number of diversified stories to comply with the users desires. In the more conventional stories, the narrative evolves following the traditional fairy tale plot with the Big Bad Wolf tricking the Little Red Riding Hood and getting to her grandmother's house first, eating the grandmother and attacking Little Red Riding Hood when she finds out what happened. In stories with a more unconventional 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 more comic outcome, the Big Bad Wolf eats both Little Red Riding Hood and her grandmother, and then gets a stomach ache. Figure 8.2 shows some scene from "*Modern Little Red Riding Hood*".

PUC-Rio - Certificação Digital Nº 1021793/CA

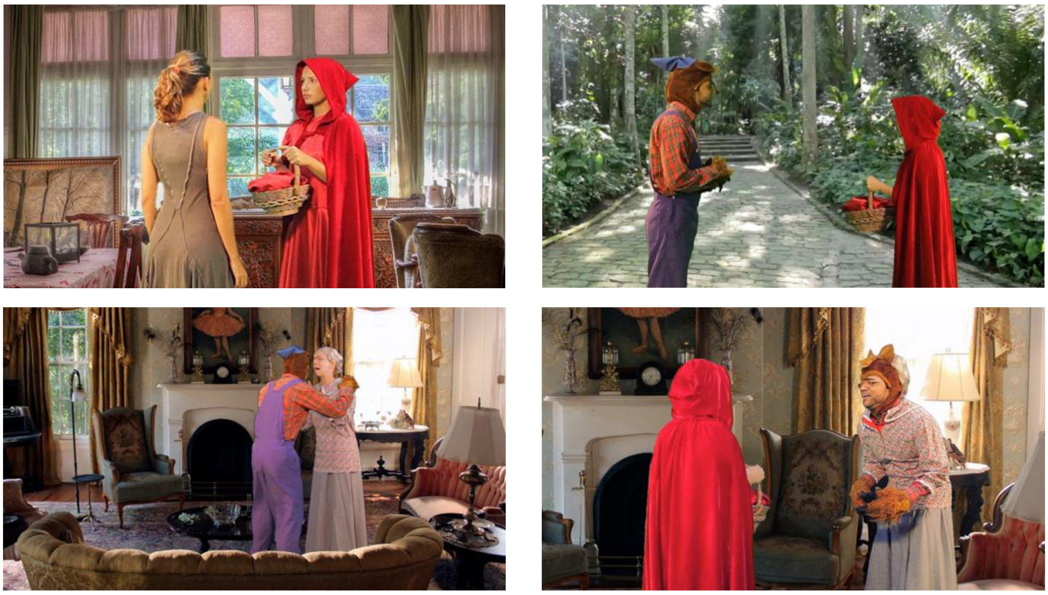


Figure 8.2: Scenes from "*Modern Little Red Riding Hood*".

In order to evaluate the proposed methods for video-based interactive storytelling from a technical point of view, we performed two tests: a performance and accuracy test to validate the methods of dramatization and user interaction, and a visual evaluation test to compare the compositing results automatically produced by the proposed system with the results manually produced by human filmmaking professionals. The following sections describe these tests.

8.1. Technical Evaluation

The technical evaluation concerns the accuracy and the real-time performance of the video compositing and user interaction methods used in the video-based interactive storytelling system. The evaluation was mainly focused on the methods that are based on image processing and machine learning algorithms, which are the most time-consuming processes and require a validation of accuracy. Each method was evaluated individually and the results are presented in the next sub-sections. The computer used to run the experiments was an Intel Xeon E5620, 2.40 GHZ CPU and 24 GB of RAM.

PUC-Rio - Certificação Digital Nº 1021793/CA

8.1.1. Video Editing

The video-based interactive storytelling system implements two video editing techniques that are used in real-time to automatically select the best shots to compose the scenes and the most adequate scene transitions to join two different shots.

8.1.1.1. Shot Selection

In order to validate the shot selection method proposed in this thesis, we performed two tests: (1) a recognition rate test to check the accuracy of the predicted shots; and (2) a performance test to check the necessary time to select a new shot.

As presented in Chapter 6, the shot selection method uses two neural networks for each type of scene. The first one is trained to classify the best camera angle for the shot, and the second is trained to select the best type of shot for the selected camera angle. In order to evaluate the accuracy of the shot selection method, for each type of scene implemented in the prototype application (total of

PUC-Rio - Certificação Digital Nº 1021793/CA

8 types), we created 5 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 to 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. Table 8.1 and Figure 8.3 show the computed results of this test with the training set size ranging from 10 to 50 samples. The presented percentages of accuracy correspond to the average of the results obtained for the neural networks used in the different types of scenes.

Number of Samples	**10**	**20**	**30**	**40**	**50**
Camera Angle Accuracy	82.2%	89.6%	92.5%	95.2%	98.4%
Shot Type Accuracy	76.5%	85.3%	89.4%	93.3%	97.5%
Shot Selection Accuracy	72.3%	81.7%	87.6%	92.2%	96.2%

Table 8.1: Recognition rate of the shot selection method with training sets ranging from 10 to 50 samples.

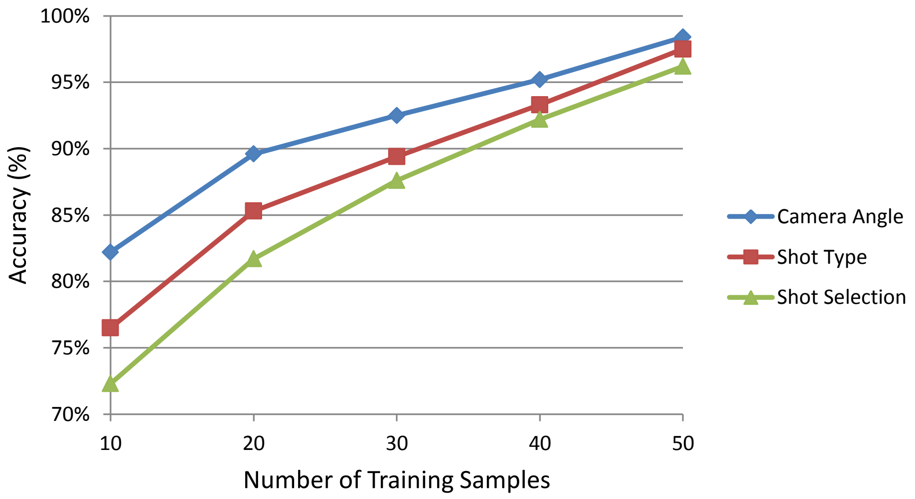


Figure 8.3: Recognition rate of the shot selection method with training sets ranging from 10 to 50 samples.

In order to evaluate the performance of the proposed solution, we used our shot selection method to predict the shots for a sequence of 5 scenes, with a total of 30 different shots. For each shot, we calculated the time necessary to extract the features used as input for the neural networks and to perform the classification process to select the best shot. As a result we got the average time of 18.3 milliseconds (standard deviation of 6.2 milliseconds), which indicates the capacity of the proposed method to selected the best shots in real-time.

The results of the recognition rate test indicate the capacity of the proposed method to learn and replicate the editing style of a human editor. The good recognition rates achieved with small training sets indicate that a human editor can train the neural networks without having to choose the best shots for too many scenes. 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 another human editor that has a different editing style, the recognition rates will probably be lower. This, however, was expected because every editor has its own style and preferences. The proposed technique is capable of learning this personal editing style and replicating it during the video compositing process, which keeps the signature of the human artist in the computer generated content.

PUC-Rio - Certificação Digital Nº 1021793/CA

8.1.1.2. Transition Selection

The capacity of the proposed method to select the most adequate transitions for video segments was evaluated by comparing the results of the proposed method with the decisions made by human editors of well-known movies.

Firstly, we analyzed the initial scene of the movie The Lord of the Rings: The Return of the King (New Line Cinema 2003). The scene starts with *Déagol* falling into the river *Anduin* and finding the ring, and ends with *Frodo* and *Sam* following *Gollum* through the *Vale of Morgul*. The test sequence had approximately 8 minutes and a total of 94 shots manually separated into individual video files (where the frames that contain transition effects were eliminated by hand). This sequence was chosen because the scene starts in the past and gradually progresses to the present story time, enabling the evaluation of

PUC-Rio - Certificação Digital Nº 1021793/CA

different temporal transitions. Given the ordered sequence of shots of the movie, we computed (for each consecutive shot) the transition between the shots (see Figure 8.4) and then compared the result with the original transition used in the movie. In this case, there is no need of a story planner or video compositing algorithms, because the movie is linear and all scenes are pre-recorded. In the video-based interactive storytelling system, the temporal and spatial information used by the algorithm is automatically calculated, but in this test we manually annotated this information in each shot. The result of the test is shown in Table 8.2.

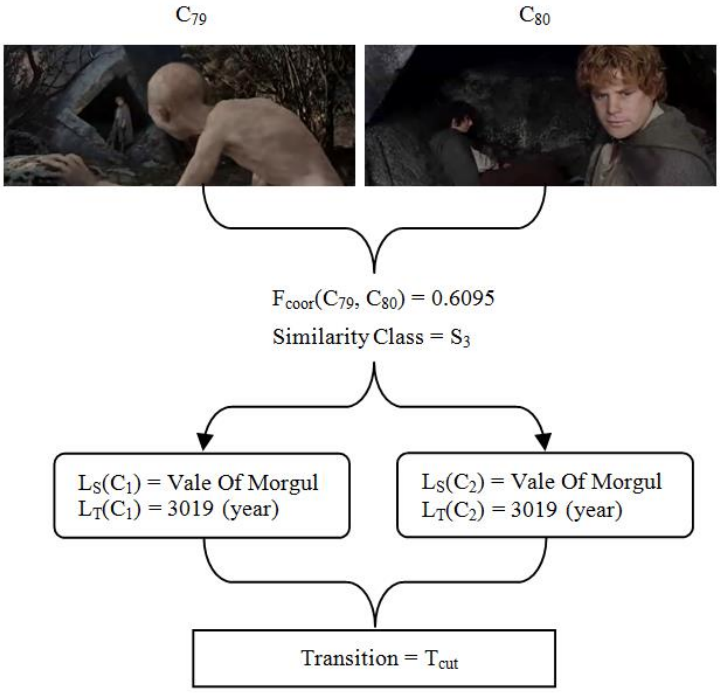


Figure 8.4: Example of a transition computation between two shots (C_{79}, C_{80}) of The Lord of the Rings: The Return of the King. Copyrighted images reproduced under "fair use" policy.

Transitions	Cut	Dissolve	Wipe	Fade
Original	86	6	0	2
Our Method	84	7	0	2
Hits Rate	97.6%	85.7%	100%	100%

Table 8.2: Comparison between the original transitions in the Lord of the Rings: The Return of the King with the transitions selected by our method.

PUC-Rio - Certificação Digital Nº 1021793/CA

We found two transitions that do not match the original transitions. The first one is a cut classified by the proposed method as a dissolve. Analyzing the video segments it is difficult to justify why the actual editor chose a cut, because it is clear that the shots occur in different times and the cut causes some disorientation in the audience. This disorientation does not occur using a dissolve transition. In the other mismatch, our method classified the transition as a jump cut. Indeed, visually analyzing the film we see that there is a jump cut in a very short fighting scene. Actually we cannot affirm that the human editor is wrong in these cases, because each editor has his/her own style.

As a second evaluation test, we selected a movie by another director in a different film genre: the classic Psycho, directed by Alfred Hitchcock (Universal Studios 1960). We analyzed the last scene of the film. The scene starts with *Lila* going to investigate *Mrs. Bates'* house and finishes in the end of the film, when *Mary's* car is pulled out of the swamp. The test sequence had approximately 14 minutes and a total of 153 shots manually separated in individual video files. The results of this test are shown in Table 8.3.

Transitions	**Cut**	**Dissolve**	**Wipe**	**Fade**
Original	150	2	0	1
Our Method	150	2	0	1
Hits Rate	100%	100%	100%	100%

Table 8.3: Comparison between the original transitions in Psycho with the transitions selected by our method.

The automated transition selection method does not involve complex computing tasks; however, its computing complexity grows according to the resolution of the analyzed video segments. In order to check the performance of the proposed method, we tested its execution in the most common video resolutions. For each resolution, we calculated the necessary time to compute the histograms, calculate the histogram correlation and classify the transition. This process was executed in a sequence of 40 video segments and the average time was computed for each resolution. The results of the performance tests are shown in Table 8.4.

Resolution	**704x480**	**1280x720**	**1920x1080**
Time(*ms*)	8.43	16.81	22.74

Table 8.4: Performance results of the transition selection method with different video resolutions.

8.1.2. Photography and Music

The validation of the methods used to select the best visual effects and soundtracks for the narratives was based on two tests: (1) a recognition rate test to check the accuracy of the predicted visual effects and soundtracks; and (2) a performance test to check the necessary time to perform the prediction process.

In order to evaluate the accuracy of the neural networks used to select the best visual effects and soundtracks for the narrative we simulated several scenes varying the type of scene, number of actors, emotional states and actions, and then, for each scene, we asked to a human director of photography and a human music director to make the selection of the visual and audio profile that best represent the scene emotion. Each decision generates one sample for both neural networks, which includes all the features used as input for the neural network, together with the identification of the selected visual effect and soundtrack for the simulated scene. Based on the samples collected, we created 5 training sets with a different number of samples and, for each one, a testing set with half the size of the corresponding training set. The training sets were used to train the neural networks and the samples of the current test set were then predicted. Table 8.5 and Figure 8.5 show the computed results of this test with the training set size ranging from 10 to 50 samples.

Number of Samples	**10**	**20**	**30**	**40**	**50**
Visual Effects Accuracy	89.7%	92.4%	95.9%	98.1%	98.8%
Music Accuracy	90.2%	93.1%	94.8%	97.5%	98.4%

Table 8.5: Recognition rate of the visual effects and music selection method with training sets ranging from 10 to 50 samples.

PUC-Rio - Certificação Digital Nº 1021793/CA

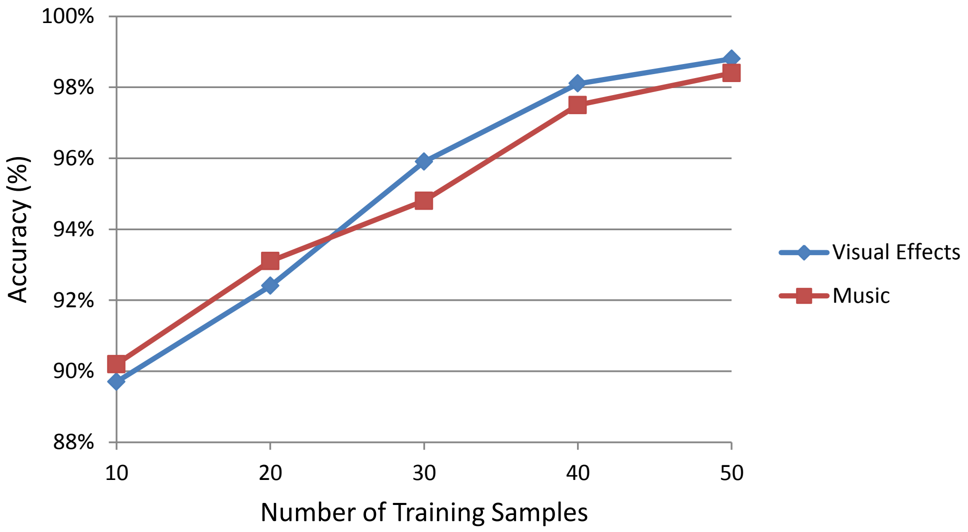


Figure 8.5: Recognition rate of the visual effects and music selection method with training sets ranging from 10 to 50 samples.

In order to evaluate the performance of the proposed solution, we used our method to select the visual effects and soundtracks for a sequence of 20 scenes and, for each one, we calculated the time necessary to extract the features used as input for the neural networks and to perform the classification process. As result we get the average time of 8.4 milliseconds (standard deviation of 4.6 milliseconds) to select the best visual effects and an average time of 7.9 milliseconds (standard deviation of 5.3 milliseconds) to select the soundtracks to the scenes, which indicates the capacity of the proposed method to perform its task in real-time.

The results of the recognition rate test are similar to the ones obtained by the neural networks trained to select the best shots for the scenes and also indicate the capacity of the proposed method to learn and replicate the decisions made by a human director of photography and music director using small training sets. This approach allows the system to learn the personal style of human filmmakers and replicate it during the dramatization of a video-based interactive narrative.

PUC-Rio - Certificação Digital Nº 1021793/CA

8.1.3. Frame Compositing

The frame compositing process is the most time-consuming task and must be performed in real-time to allow the system to generate video frames while the narrative is being exhibited. The algorithm comprises several image processing methods and its complexity grows according to the number of scene elements that have to be composed in the frame. In order to evaluate the performance of the parallel architecture of the proposed frame compositing process, we conducted a performance test to check the average frame rate of the proposed system with the number of threads ranging from 1 to 8. Five sequences of 4 basic actions with an increasing number of actors were simulated and dramatized by the system, generating a total average of 600 frames per sequence. The results of the performance tests of the parallel composing architecture are shown in Figure 8.6.

The results of the performance experiments show 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 system can compensate the cost of the frame compositing task by dividing the work among multiple CPU cores.

PUC-Rio - Certificação Digital Nº 1021793/CA

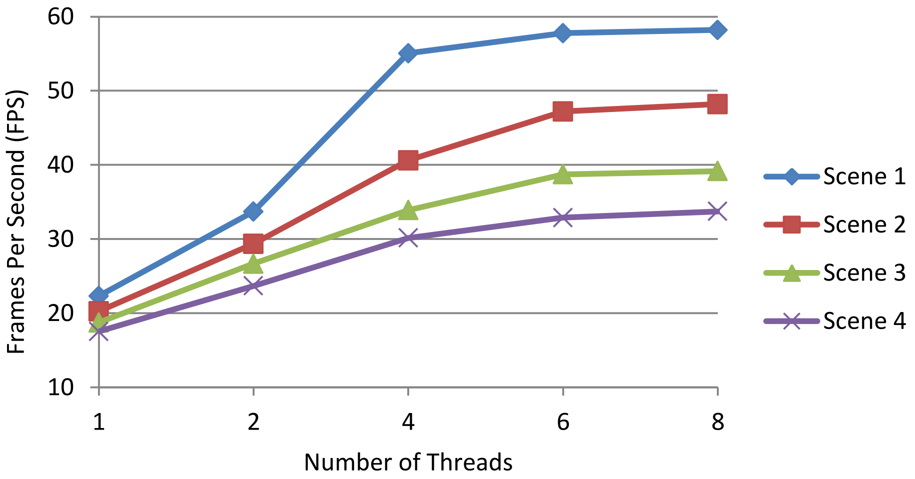


Figure 8.6: Performance results of the parallel composing architecture with the number of actors in the frame ranging from 1 to 4 and with the number of compositing threads ranging from 1 to 8.

8.1.4. Natural Language Interface

The user interaction methods adopted in the video-based interactive storytelling system are mostly based on natural language. In order to evaluate the natural language processing algorithms implemented in the user interaction module, we performed two experiments: (1) the recognition rate test, to check the accuracy of the predicted suggestions; and (2) the performance test, to check the time needed to process the input suggestions and recognize them as first-order logic sentences. For both tests, we used a set of 107 text suggestions collected from users that were testing the interaction system. After a manual analysis of the suggestions, we found 81 suggestions that were manually classified as valid suggestions.

For the recognition rate test, we used our method to extract valid story suggestions from the text suggestions and then compared the results with the results obtained by the manual classification. As a result we got a recognition rate of 90.6%, with only 10 valid suggestions being incorrectly classified as invalid suggestions. The main reason for the incorrect classifications was the occurrence of spelling mistakes.

To evaluate the performance of our method, we again utilized the 107 suggestions collected from users, and calculated the average time necessary to recognize them as first-order logic sentences. As a result we got the average time of 2.7 milliseconds to process an input suggestion and recognize them as first-order logic sentences (standard deviation of 1.3 milliseconds).

Similarly, we evaluated the method utilized to recognize user satisfaction. During the tests of the system, we collected a set of 43 text comments expressing user satisfaction. Then, we used our simplistic method of sentiment analysis to classify the comments as positive and negative comments then compared the results with the results obtained through a manual classification. As result we get a recognition rate of 97.6%, with only 1 positive comment incorrectly classified as negative. The time consumed by the algorithm is almost insignificant (less than 0.001 milliseconds).

In the experiments, the multi-user natural language interface produced good results. However, natural language processing is not a trivial task. It is possible that our parser will not correctly recognize every possible valid sentence, but we believe that it will be able to recognize the sentences in the most part of the cases without the audience being aware of mistakes.

PUC-Rio - Certificação Digital Nº 1021793/CA

PUC-Rio - Certificação Digital Nº 1021793/CA

8.2. Visual Evaluation

The visual evaluation 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 proposed 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 human 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 human and system had available the same video resources to compose the frames. Table 8.6 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.

	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.*"

Table 8.6: Description of the selected basic actions used in the visual evaluation test.

In order to perform the task, the human subjects decided to use the Adobe After Effects CS6. The results of the visual evaluation test comparing the initial frames of the scenes composed by the human professionals and the initial frames automatically generated by the proposed video-based dramatization system for the three selected basic actions are shown in Table 8.7.

During the experiment, we also recorded the time both human and system spent to complete the tasks. Table 8.8 shows a comparison of the time spent by the subjects to complete the composition of a single frame of each basic action.

PUC-Rio - Certificação Digital Nº 1021793/CA

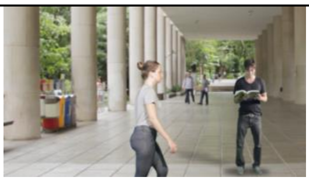
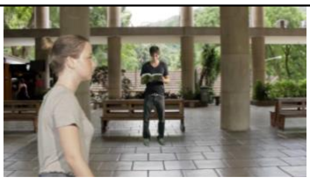
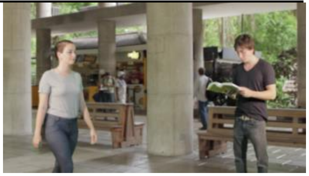
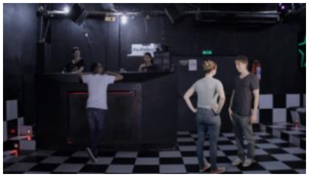
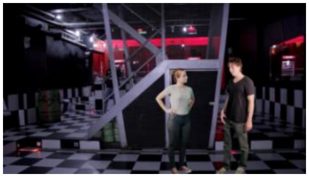
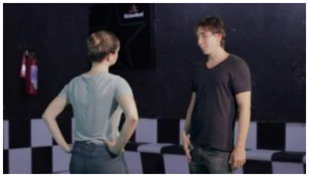
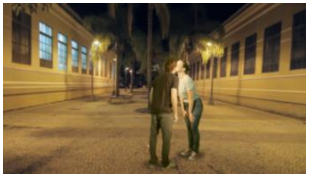
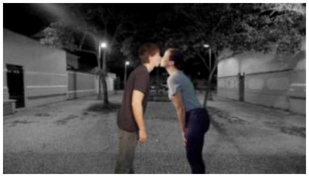
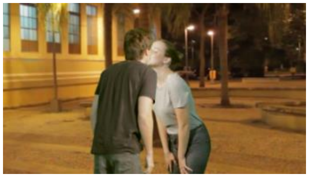
	Team 01	Team 02	System
Action 1			
Action 2			
Action 3			

Table 8.7: Visual comparison between the selected frames of the scenes composed by the human subjects and the corresponding frames automatically generated by the proposed video-based dramatization system for 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*)

Table 8.8: Comparison between the times spent by the human professionals and the system to compose the scenes representing the three basic actions.

Although the scenes automatically generated by the system have different angles and distances from those specified by the human teams, the results are similar in quality in a way that it would be difficult to a human to identify which ones was generated by a computer software. These similarities indicate the capacity of the proposed automatic video compositing methods to generate frames similarly as video compositing professionals do. In addition, the system is capable of generating multiple frames instantaneously, while the human professional takes several minutes to composite the frames.