Joint Virtual Reality Conference of EuroVR - EGVE - VEC (2010)
T. Kuhlen, S. Coquillart, and V. Interrante (Editors)

Short Paper: Virtual Storyteller in Immersive Virtual Environments Using Fairy Tales Annotated for Emotion States

I. V. Alexandrova [1,2†], E. P. Volkova [2], U. Kloos [1], H. H. Bülthoff [2,3] & B. J. Mohler [2]

[1] Reutlingen University
[2] Max Planck Institute for Biological Cybernetics
[3] Department of Brain and Cognitive Engineering, Korea University, Seoul, 136-713 Korea

Abstract

This paper describes the implementation of an automatically generated virtual storyteller from fairy tale texts which were previously annotated for emotion. In order to gain insight into the effectiveness of our virtual storyteller we recorded face, body and voice of an amateur actor and created an actor animation video of one of the fairy tales. We also got the actor's annotation of the fairy tale text and used this to create a virtual storyteller video. With these two videos, the virtual storyteller and the actor animation, we conducted a user study to determine the effectiveness of our virtual storyteller at conveying the intended emotions of the actor. Encouragingly, participants performed best (when compared to the intended emotions of the actor) when they marked the emotions of the virtual storyteller. Interestingly, the actor himself was not able to annotate the animated actor video with high accuracy as compared to his annotated text. This argues that for future work we must have our actors also annotate their body and facial expressions, not just the text, in order to further investigate the effectiveness of our virtual storyteller. This research is a first step towards using our virtual storyteller in real-time immersive virtual environments.

Keywords: *Augmented Reality and Virtual Reality, Virtual Humans, Storytelling*

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Virtual reality

1. Introduction and Background

Naturally animated virtual humans are used for many applications in training, interactive storytelling and populating immersive virtual environments (VEs). Virtual storytellers are an example of virtual humans that are currently a topic of investigation for many scientists [BC04, KHG*07, MTK09]. The storyteller should be able to present the story in a way that the audience understands, interprets, and perceives the emotions that the story should convey. In addition, the storyteller should engage the listener in the story and make the audience feel the presence of the emotional expressions. This could be done not only with motions, but also with voice modulation.

Therefore the realistic emotions expression of virtual humans is one of the most important factors for providing the illusion of personality and human-like behavior [MTK09, KHG*07]. According to [MTK09] interactive virtual humans should be able to make decision based on memory of past events or relationships with other characters. Thus, they will be able to convey their personality to the human users. Other researchers are working on the development of a virtual human that is able not only to express realistic emotions but also to sense the gestures and facial expressions of the trainee or the audience and act accordingly to the sensed emotions [KHG*07]. In addition, several studies have shown

† ivelina.alexandrova@tuebingen.mpg.de The authors gratefully acknowledge the support of the Max Planck Society and the WCU (World Class University) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (R31-2008-000-10008-0).

that motion is also important for expressing emotions and interacting with the human user [MJM*09, KHG*07]. However, realistic expression of emotions is not enough to fully engage the audience in the story. Similar to the human storytellers the virtual ones should be able to modulate their voice to emphasize the dramatic parts of the text and distinguish between the characters in the story and change their voice accordingly [BC04].

Virtual storytelling is a process that should not always have only a storyteller and a listener or an audience. Immersive VEs give more power to the user and enable him/her to experience the story by taking part in the story itself. [EBBA09] present research, which describes an interactive drama storytelling system for multiple users. The system enables the users to experience the same story from different perspectives depending on their role.

As a first step towards building our virtual storyteller we decided to concentrate on the realism of his body and facial expressions. Our goal is to create a virtual storyteller, which can be used to study the expression and perception of emotions in real-time immersive virtual environments. Therefore we present our approach for creating a virtual storyteller by morphing body and facial emotional states (ESs) based on previously annotated texts. To verify the realism of the presentation of the storyteller we performed a user study, in which we compared our virtual storyteller to an animation of an actor, telling a fairy tale.

2. Approach

In this section we present the setup of the virtual storyteller, which uses natural body and facial ESs and morphs between them based on previously annotated texts. Then we describe an actor animation used as a baseline to test the realism of our automatically generated virtual storyteller.

2.1. Setup of the automatically Generated Virtual Storyteller

For creating the body animations of the virtual storyteller we have used 15 basic body ESs (neutral, relief, disturbance, joy, sadness, hope, despair, interest, disgust, compassion, hatred, surprise, fear, approval, anger: see Figure 1). The 15 recorded animations were short dynamic clips of a person's motions captured using Moven, an inertial motion capture suit from Xsens. The dynamic clips ranged from 2 to 6 seconds in length. The avatar used for the storytelling was a virtual character (Rocketbox Studios). The captured animations were mapped to the character in Autodesk Motion Builder 2009. The animated character was then exported to Dassault Systemes 3DVIA Virtools 5.0. For smoother animation transitions and greater programmer flexibility each body ES used for the storyteller begins and ends with a pose close to neutral.

Another approach, BlendShapes, was used for generating

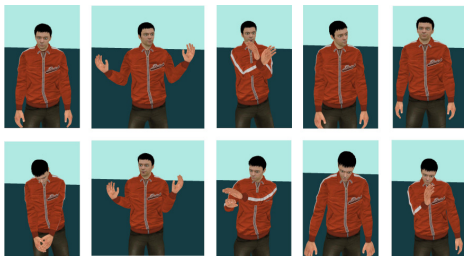

Figure 1: *Body expressions from top left to bottom right: neutral, joy, fear, disturbance, approval, sadness, surprise, disgust, anger, despair.*

the different facial ESs of the virtual storyteller. The process of blending shapes requires several meshes of a face. It blends from one mesh to the other by changing the weights of the different meshes. In this case each mesh expresses a particular motion of the face (smile, mouth open, eyes open, eyebrows raised, see Figure 2). For expressing the different emotions of the virtual storyteller we predetermined the weights of the meshes for each of the ESs. Also, to make the storyteller's facial ES more natural, a randomized blinking (1 to 4 seconds) was used. Since, his eyes are separate from the face meshes, we are also able to move the eyes and direct the storyteller's gaze, which we hope will make him appear more realistic.

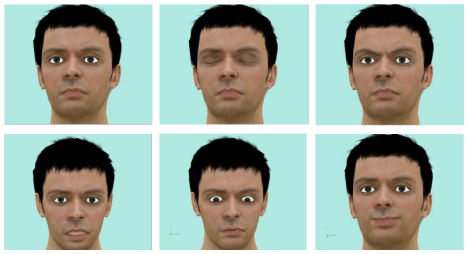

Figure 2: *Facial motions from top left to bottom right: neutral expression, eyes closed, eyebrows down, mouth open, eyes moved, smile.*

The text that was used as input for our virtual storyteller first consisted of 8 short Brothers Grimm's fairy tales ranging from 1200 to 1400 words (consistent with [VMM*10]). They were written in standard German. More recently the number and the length of the fairy tales has been increased - 79 longer fairy tales with average length of 2500 words have been added. The texts of the fairy tales were manually annotated for ESs by native speakers. The annotators used the 15 ESs described earlier. The speech of the storyteller was automatically generated using Natural Reader 9.0 from the files of the annotated texts. We also automatically parsed the timing of each word, and therefore automatically created a list of the ESs with their subsequent timing [Rap95]. Therefore we were able to increase the naturalness by fully synchronizing the annotations and the expressions.

Thus, when the system has an annotated fairy tale as an input it produces the speech generated from the annotated text, produces a file with the ESs and the timing, and finally automatically creates an animation of the body and face by morphing and blending between previously recorded/determined ES animations.

2.2. Animated Actor Animation

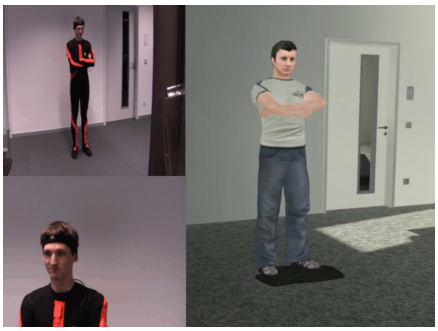

Figure 3: *Left: Amateur Actor in Moven Suit with close up face video capture. Right: Animation of Storytelling in our Virtual Environment Setup using full body motion capture information.*

To test the realism of the virtual storyteller, an actor animation was setup. For making the actor animation we have used a motion capture data of a whole fairy tale told by a German amateur actor. The amateur actor was asked to read a German fairy tale, called *Godfather Death (Gevatter Tod)*, and annotate its text according to his emotional perception of the text using the 15 ESs (see Figure 1). For annotating the whole text he has used 12 unique emotions and divided the fairy tale into 38 chunks with average length of 12.61 seconds (in time spoken in presentation) and with standard deviation 10.30 seconds. Then the amateur actor was asked to tell the fairy tale which he previously annotated with ESs. The duration of the storytelling was 479.8 seconds. His body motions were captured with a Moven inertial tracking suit. In addition, we used two cameras to record videos of him, while he was telling the fairy tale (see Figure 3). The first camera captured the full body of the amateur actor and the second one captured only his face. The video of the latter was used in the post-processing for the facial expressions, while the video of the first one was for validation of the motion capture data.

After the recordings, the motion capture data was mapped to another virtual character (Rocketbox Studios). For recording the facial expressions of the amateur actor and transferring them to the character we have used faceAPI, a face tracking technology developed by Seeing Machines. The software tracks the facial features such as eyebrows, lips, nose, etc. and can be used in both real-time applications and post-processing. However, in the user study we have used post-processing for mapping the facial expressions of the amateur actor to the actor animation. The sound used for the voice of the actor animation was extracted from the video recordings of the human actor. Thus, we made the actor animation while still retaining as much information from the actor as possible (face and body motions, voice).

3. User Study

To test our virtual storyteller we have used the video of the animated actor animation as a baseline. To make a fair comparison between our baseline video and the video of our virtual storyteller, the fairy tale annotated by the amateur actor was used as an input for the virtual storyteller. The output of the system was recorded as a video of the automatically generated virtual storyteller. Thus the virtual storyteller was telling the same fairy tale as in the actor animation and expressing the annotated amateur actor's emotions by morphing between the ESs. Since we used two different people for recording the animations of the actor animation and the virtual storyteller, we prevented the use of motions typical for a particular person in the animation of both storytellers.

Finally, the presentation of the actor animation was compared to the virtual storyteller. For this we only used the animations and not the sound, since potentially the sound likely conveys rich information about the emotional expression and we are primarily concerned with whether the automatically generated animations also convey the same emotional meaning as the motion capture animation. To analyze how people perceive the emotions expressed by our virtual storyteller we performed a preliminary analysis.

Five participants (3 male and 2 female; average age - 26.8) were asked to watch a video without sound of the actor animation and a video without sound of the virtual storyteller. After watching the first video (the one with the actor animation) they had to watch it again and try to annotate 10 to 20 different emotions and mark the time of the emotion peak. Then they had to do the same for the second video. They did not know that both videos present the same fairy tale.

4. Results

The emotional meaning conveyed by the virtual storyteller was compared to the actor animation. The participants reported that it was rather challenging to locate the exact emotions of the actor animation, while it was much easier to tell the emotions expressed by the virtual storyteller. This was probably due to the fact that the emotions expressed by the virtual storyteller are sort segments of 15 possible animations and it was easy to tell when a short animation started and stopped. Furthermore the problems that participants had when locating the emotions in the video of the actor animation were most probably due to the fact that humans often use complex body motions to express a certain emotion.

The results of this user study showed that for the video of

Results from User Study		
Participant: Video	Intended	Swapping
Novice: Virtual Storyteller	39.4%	36.4%
Novice: Actor Animation	18.2%	20.45%
Actor: Actor Animation	15%	17%

Table 1: *Intended refers to the percent correctly annotated in the video as compared to the annotated text of the actor. Swapping refers to the percentage of emotions where surprise and happiness and sadness and anger were swapped. Percentages are in overall percent of the annotations provided by the user or actor.*

the actor animation one participant labeled a sequence of 19 emotions, two - a sequence of 12 emotions and the remaining two - a sequence of 11 emotions. For the video with virtual storyteller four participants labeled a sequence of 20 and one participant labeled a sequence of 17 emotional expressions. 39.4% of the emotions of the video with virtual storyteller were perceived as the amateur actor intended. 36.4% of the wrong annotations were due to the swapping of happiness with surprise and sadness with anger and vice versa. The emotional expression in the actor animation was very hard to determine and did not appear to line up with the annotated text of the actor - the matched annotations were only 18.2%, while the wrong annotations due to swapping of emotions 20.45%.

We therefore had the actor himself annotate the video so we could determine if our annotations of the video lined up with this annotation better than the text. Interestingly, although before telling the story the amateur actor annotated the text with a sequence of 38 emotions, he marked 64 emotion instances when watching the video of the actor animation. Surprisingly, only 15% of the marked emotions matched with the initial annotation and 17% were wrong due to swapping of the emotions. In future recordings of actors we will therefore have our actors also annotate their body and facial expressions, not just the text, in order to further investigate the contribution and timing of the body/facial motion to emotional expression of stories (see Table 1).

5. Future Work

We believe it is important that the storyteller presents our annotated stories with a variety of expressive emotions to the audience. Therefore we want to collect several different body motions of each emotion in order to make our virtual storyteller more believable and human-like. Furthermore, for more realistic and expressive facial emotions we are going to use a different face mesh for each ES, instead of the individual predefined motions of the eyes, the eyebrows or the mouth used so far. To make the storyteller more believable it is also necessary to integrate and synchronize the opening and the closing of the mouth to the spoken text.

We are collaborating with scientists from the field of computational linguistics to gain automatic ESs in annotations of texts using natural language processing and machine learning. This will enable us to automatically animate a large collection of fairy tales and scenarios we need for our further work.

6. Conclusion

Our program is a tool which enables the user to create a virtual storyteller based on annotated emotional comprehension of the text. This gives us the power to manipulate the body and facial expressions of the virtual storyteller and observe their impact on the participants when perceiving emotions and learning in real-time immersive virtual environments. In this paper we have presented an approach for animating a virtual storyteller by using morphing of facial and body ESs based on previously annotated texts. This system runs in real-time (120fps), and therefore can be used for creating a virtual storyteller in immersive VEs. It is a first step towards being able to fully manipulate and control the perception of emotions in learning experiments in VEs. Since, emotions are a crucial factor for conveying personality and perceiving information in conversations and storytelling ([MTK09, KHG*07, AS05]), we believe that our approach is a promising tool for manipulation of animations of emotions for real-time experiments in immersive VEs.

References

[AS05] ALM C. O., SPROAT R.: Perceptions of emotions in expressive storytelling. *Interspeech 2005* (Dec 2005), 533–536.

[BC04] BICKMORE T., CASSELL J.: Social dialogue with embodied conversational agents. *In Natural, Intelligent and Effective Interaction with Multimodal Dialogue Systems, New York: Kluwer Academic* (Dec 2004).

[EBBA09] ENDRASS B., BOEGLER M., BEE N., ANDRÉ E.: What would you do in their shoes? experiencing different perspectives in an interactive drama for multiple users. In *ICIDS '09: Proceedings of the 2nd Joint International Conference on Interactive Digital Storytelling* (Berlin, Heidelberg, 2009), Springer-Verlag, pp. 258–268.

[KHG*07] KENNY P., HARTHOLT A., GRATCH J., SWARTOUT W., TRAUM D., MARSELLA S., PIEPOL D.: Building interactive virtual humans for training environments. *I/ITSEC* (2007).

[MJM*09] MCDONNELL R., JÖRG S., MCHUGH J., NEWELL F. N., O'SULLIVAN C.: Investigating the role of body shape on the perception of emotion. *ACM Trans. Appl. Percept. 6*, 3 (2009), 1–11.

[MTK09] MAGNENAT-THALMANN N., KASAP Z.: Modelling socially intelligent virtual humans. In *VRCAI '09: Proceedings of the 8th International Conference on Virtual Reality Continuum and its Applications in Industry* (New York, NY, USA, 2009), ACM, pp. 9–9.

[Rap95] RAPP S.: Automatic phonemic transcription and linguistic annotation from known text with Hidden Markov Models. In *Proceedings of ELSNET Goes East and IMACS Workshop* (1995), Citeseer.

[VMM*10] VOLKOVA E. P., MOHLER B. J., MEURERS D., GERDEMANN D., BÜLTHOFF H. H.: Emotional perception of fairy tales: Achieving agreement in emotion annotation of text. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies* (June 2010).