Making Movies from Make-Believe Games

Adela Barbulescu, Antoine Begault, Laurence Boissieux, Marie-Paule Cani, Maxime Garcia, Maxime Portaz, Alexis Viand, Pierre Heinish, Romain Dulery, Rémi Ronfard, et al.

Eurographics Workshop on Intelligent Cinematography and Editing (2017), pp. 1–5
W. Bares, V. Gandhi, Q. Galvane, and R. Ronfard (Editors)

Making Movies from Make-Believe Games

Adela Barbulescu[1], Antoine Begault[1], Laurence Boissieux[1 2], Marie Paule Cani[1],

Maxime Garcia[1], Maxime Portaz[2], Alexis Viand[3], Pierre Heinish[3], Romain Dulery[3], Remi Ronfard[1], Dominique Vaufreydaz[2]

[1]Univ. Grenoble Alpes, Inria, LJK, F-38000 Grenoble France
[2]Univ. Grenoble Alpes, Inria, LIG, F-38000 Grenoble France
[3]Grenoble-INP

Abstract

Pretend play is a storytelling technique, naturally used from very young ages, which relies on object substitution to represent the characters of the imagined story. We propose "Make-believe", a system for making movies from pretend play by using 3D printed figurines as props. We capture the rigid motions of the figurines and the gestures and facial expressions of the storyteller using Kinect cameras and IMU sensors and transfer them to the virtual story-world. As a proof-of-concept, we demonstrate our system with an improvised story involving a prince and a witch, which was successfully recorded and transferred into 3D animation.

Categories and Subject Descriptors (according to ACM CCS): I.3.7 [Computer Graphics]: Three-Dimensional Graphics and Realism—Animation H.5.2 [Information Interfaces and Presentation]: User Interfaces—Interaction styles

1. Introduction

There has been increasing interest in recent years in providing virtual storytelling tools that can be used to teach narrative skills to young children [GPS10]. While early work has investigated the use of digital puppets [HRvG97] and interactive spaces [BID*99], many researchers have noted that tangible interaction with actual physical puppets is usually more engaging to children even in the digital age. According to Walton [Wal94], props such as puppets and figurines play a central role in make-believe games, where they serve as *prompters* for imagination, but also as *objects* of the imaginary worlds created in the game, and as generators of the *fictional truths* that make up the story.

In this work, our goal is to create movies showing the fictional truths created by the storyteller in simple games of make-believe, using puppets and figurines as props. Our system allows the storyteller to improvise a scene with two figurines in hand by interpreting the lines of the two characters in front of two Kinect cameras (see Figure 1). The front camera records the storyteller's facial expressions and head movements. The top camera records the movements of the figurines. The figurines are also equipped with inertial measurement units (IMU) recording their linear accelerations and angular positions.

Our system decomposes the story into alternating speaking turns where the storyteller is playing the part of one character or the other. Then the head movements and facial expressions of the storyteller are transferred to the corresponding character. Head movements and facial expressions of the non-speaking character are automatically computed so that they smoothly integrate with those transferred during speech turns. In addition, the relative movements of the two figurines are transferred to the corresponding animated characters, with suitable adaptations.

2. Related work

Using tangible interaction for creating stories is not a new idea. A variety of puppetry interfaces that can be used to control character animation in real time were reviewed by Sturman [Stu98]. The digital marionette [OTH02] is a physical device equipped with sensors, which can be used to create relatively complex character animation using multi-tracking. The i-theatre [MCP09] investigates the creation of puppet-like tangible interfaces augmented with embedded sensors (e.g. RFIDs and accelerometers) and multimedia capabilities integrating the different elements of a traditional puppets-theater (e.g. music, lighting, etc.) into a virtual representation.

The 3D puppetry system by Held et al. [HGCA12] uses a combination of image-feature matching and 3D shape matching to identify and track physical puppets with a kinect camera and then render the corresponding 3D models into a virtual set. PuppetX by Gupta et al. [GJR14] is a construction kit for building articulated puppets

that can be manipulated and virtualized in real-time using a combination of finger and body gestures. [CMB15] combine multi-touch interaction on a tabletop (NIKVision) with tangible interaction with physical props for playful experiments with kindergarten children.

Most of the above previous work relies on real-time interaction with a magic mirror metaphor showing the story in the virtual world as it is being played in the real world. While this may be attractive, it also causes problems of divided attention between those two mental work spaces. In contrast, we describe a system where the storyteller is allowed to improvise freely with sensor-equipped objects in front of Kinect cameras, and watch the story he was imagining as a separate step.

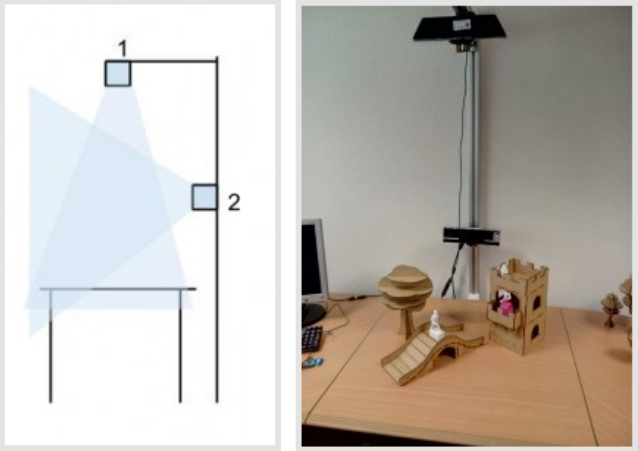

Figure 1: *Acquisition setup. There are two Kinect devices: one looking down to follow figurines, one following narrators. The stage is the table in the middle.*

3. Recording system

Recently, several storytelling systems such as PuppetX [GJR14] or 3D puppetry [HGCA12] have used RGB-D sensors. A problem with these vision-based systems is that they are very sensitive to occlusion. The storyteller must be very careful to keep the puppet visible for the camera at all times while acting. As this may interfere with the narration, we reduce occlusion problems as much as possible by using Inertial Motion Units (IMU) sensors. We further enrich the recordings with body pose, facial expression and voice of the storytellers. Therefore, we use one kinect at the top of the stage combined with a set Inertial Motion Units (IMU) in order to track the puppets. This configuration was set to address the tasks of puppet identification and localization in 4D (3D space + time). In addition the IMU sensors provide the angular position at each time for each puppets making the system recover the 6D path for each figurine. The second kinect records the storyteller's face features using the commercial system FaceShift. This markerless motion capture system returns accurate head rotation and translation (30fps, absolute values with camera placed at 0), gaze direction and facial expressions (represented by a total of 48 blendshapes), along with synchronized voice signal. Using this software we are able to transfer the face features (blendshapes) to the corresponding 3D model.

4. Animation system

The storyteller uses his voice, gestures and expressions to impersonate two characters. Our animation system uses the data recorded with the systems described in Section 3 in order to generate separate animations for the two characters. This section describes the algorithms implemented for the decomposition of the storyteller's recorded data into two audiovisual data streams that correspond to the characters' performances (see Figure 2).

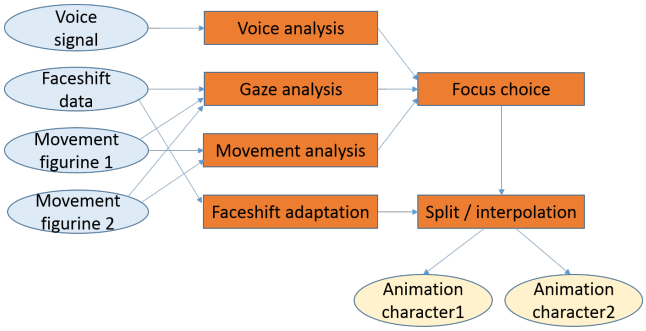

Figure 2: *System workflow: the voice signal, motion recorded with Faceshift and figurine movements are analyzed in order to determine the focus turns of the characters to which the adapted motion is transferred. Next, the separate animation data for the two characters is obtained by splitting the storyteller's recorded data and interpolating the missing motion from the non-focused parts.*

4.1. Animation layers

We can safely assume that the characters take talking turns that do not overlap as the storyteller can only interpret one character at a time. Also, we consider that the gestures and expressions deployed during storytelling present two main components: *expressive*, which is intended for the actual interpretation of the character (for example: lip movements while interpreting a character that is talking) and *focusing*, which appears as a consequence of manipulating the figurines and helps indicate which figurine is currently interpreted (for example: the storyteller may lean his head and gaze towards the focused figurine while performing). Determining which figurine is being focused at each moment is an essential task for our system.

Considering these components of a storyteller's performance, we propose that the movements of the virtual characters are obtained using several layers of animation:

- Rigid body motion is transferred directly from the tracked movements of the corresponding figurine.
- Head rotation, gaze, facial expressions and voice are transferred directly from the storyteller if the corresponding figurine is focused, or are automatically generated otherwise.
- Body and head rotations and facial expressions may be readapted in order to edit out the focusing component of the transferred data (for example, to raise the head and change the gaze direction) or to satisfy certain cinematographic constraints (for example, rotating the bodies such that the face is visible).

4.2. Story analysis

Since we want to generate dialog stories, an important task is extracting the sentences and then assigning them to the corresponding figurines i.e. the ones that are focused during speech. We first recorded a performance, in which a female storyteller interprets a male and a female character. For this performance we annotated the sentence intervals with the characters to which they correspond. Next we computed a set of parameters for each type of data as an optimization problem for assigning the right figurine with its corresponding annotated sentence. These parameters can then be used for a new storytelling performance.

Voice analysis. First of all voice informs whether one character is talking. A strategy in impersonating characters is changing the voice pitch, intensity or rhythm. Particularly, female characters are delivered with high-pitched voices and monsters or other negative characters with low-pitched or slow-paced voices. We use the Praat software to compute the voice pitch and intensity. However, tests show that the storytellers have difficulty maintaining distinctive pitch strategies for the two characters throughout an entire story. For this reason, we only use the intensity of the voice signal to extract sentences. The sentences are extracted by separating the silent frames from speech frames using an intensity threshold, then by concatenating the successive speech frames.The best result is obtained for an intensity threshold of 50 dB. The same parameter can be used in extracting sentences for a new performance if the recording is done with the same microphone and under similar audio conditions.

Gaze analysis. Gaze direction is an important cue because storytellers tend to look at the figurine that they are currently interpreting. Therefore, determining which figurine is focused depends on the angle of gaze direction and the current positions of the figurines. The simplest solution is determining whether the gaze is oriented towards the left or the right relative to an imaginary vertical plane which passes through the center of the storyteller and equally divides the space between the figurines.

Movement analysis. Storytellers also tend to move more the figurine which is currently focused. In order to determine the focused figurine during speech, we compare the amount of variation of translation and rotation for a set of frames for the two figurines. We compute the sum of 1-norm between the current position and orientation and the one at a previous frame, for the last N frames. N is determined by varying this number such that the best matches are obtained with the annotated frames. The number of frames obtained is 78, for a framerate of 30 fps.

Focus choice. After analyzing the voice and extracting sentences, we use gaze and figurine movements to determine which figurine is focused. The problem of choosing the figurine is also solved as an optimization problem where we assigned weight coefficients for the two methods and varied their values. We found that the best result is obtained when the weight attributed to the movement analysis method is 1 and for the gaze analysis is 0. This shows that the figurine movement is a better cue for assigning the focused character at the level of the sentence.

Facial motion adaptation. Once the focus is obtained, we can directly transfer the expressions and head motion of the storyteller to the assigned characters. A few adaptations are first needed in order to correct the storyteller's motions which are caused by manipulating the figurines: head and gaze oriented downwards. Head and gaze pitch values are scaled such that the head and gaze direction are raised. The blendshapes corresponding to eyelid opening are also scaled in order to correspond to the adapted gaze direction.

Splitting/interpolation. At this point we have determined the focus and the expressions to transfer to the corresponding characters. In order to generate a separate animation stream for each character, we need to generate head motion and facial expressions for the intervals where a character is not focused.

The first step is to split the adapted motion data according to the focus intervals. For the non-focus intervals i.e. when the character is silent, generated motion data should look natural and not distract the attention from the speaking character. For this reason we opt for neutral facial expressions. We carry linear interpolation for head motion during all non-focused intervals between the last and first pose of the respective neighboring focused intervals. To ensure smooth transitions between the transferred and the automatically generated neutral expressions, we also carry linear interpolation of all blendshapes during intervals of 250 ms at the beginning and end of the non-focused interval.

5. Experimental results

We used a set of predefined characters that are recurring in folktales: the prince, the princess, the witch, the ogre. A professional 3D artist created stylized body and head models for these characters. The heads of the virtual characters are rigged as blendshape models which correspond to the ones obtained with Faceshift. The body models are 3D printed to obtain the figurines. To each figurine, a plinth is added to incorporate an IMU sensor.

Recording. In order to evaluate our system, we recorded a male storyteller delivering a short story with two characters: the prince and the witch. The system imposes two constraints: figurines must be used within the surface of the table (80x80 cm) and the storyteller should perform at a distance between 60 and 120 cm from the front Kinect. The text of the story was prepared beforehand but the interpretation and manipulation of the figurines were deployed freely. We consider that the story starts when the storyteller moves the first figurine and it ends when both figurines are released. The entire story lasts 1 minute and 25 seconds and consists of 12 sentences alternating between the two characters. Figure 3 illustrates different steps in the generation of the animated story.

Animation. We analyzed the story using the algorithms described in Section 4 with the previously computed parameters. The voice analysis algorithm extracts sentences with a total of 84% speech frames correctly identified. Using the gaze analysis algorithm leads to correctly assigning 85% of the speech frames to the focused characters, while using the movement analysis obtains 89% recognition rate. The focus choice algorithm leads to a final correct assigning of 89% speech frames since we only use the figurine movement to choose the focused figurines. We notice that gaze analysis is less reliable than figurine movement especially because the storyteller tends to switch the gaze direction to the other figurine before the sentence ends.

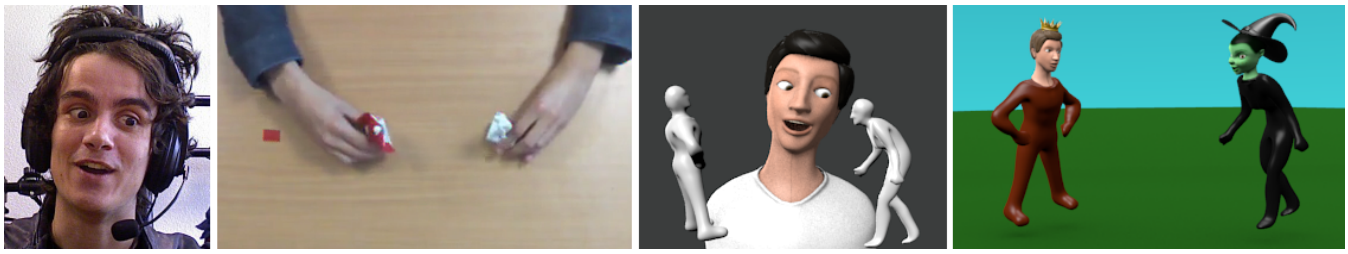

Figure 3: *Corresponding frame for: (a) the front Kinect video data displaying the storyteller's expressions, (b) the top Kinect video data with the current orientations and positions of the figurines, (c) a reconstructed scene including a virtual storyteller with directly transferred expressions and virtual figurines with directly transferred movements and (d) the final animated version of the story.*

Besides the adaptation of motion recorded by the storyteller, we introduced new modifications for satisfying cinematographic rules in the virtual scene. We carried automatic body orientation according to the camera position in the virtual scene by rotating the bodies with 25 degrees towards the camera. We did this because we want to see the characters' faces in generated animation, while still maintaining the illusion of dialogue.

6. Limitations and future work

This system is presented as a tool for making virtual movie scenes from pretend-play for children such that the animated result can be viewed after the performance takes place. In future work, it may be useful to provide real-time feedback to storytellers, allowing them to add and control scene components such as cameras, lights and surrounding environments. The acquisition system described in this paper has intrinsic limitations, i.e. only one storyteller and two figurines can be recorded at a time. In addition, the Faceshift software we use for capturing the narrator's blendshapes is no longer available. However several alternative software such as Mixamo (Unity plug-in) or Kinect Face HD can be considered.

Our system is currently limited to facial animation. In future work, we would like to infer hand gestures and full body animation intended by the storyteller and transfer them to the virtual characters in a similar fashion. We are currently studying the strategies deployed by different storytellers to display the desired actions using a combination of explicit voice commands and implicit motion cues.

7. Conclusion

We presented a system for creating movies from improvisational games of make-believe, where figurines and their movements are transfered into imaginary characters. Gestures and expressions of the storyteller are tracked and analyzed such that an animated version of the dialog story is generated. Our system enables the generation of separate animation streams for each character, using the performance of one storyteller. This is done by first extracting the sentences using the voice signal and then choosing which character performs each sentence. This choice is made by analyzing the storyteller's strategies towards "focusing" i.e. attributing a performance to a certain figurine by either looking towards that figurine or moving it more. Our experiments indicate that a combination of voice prosody, figurine motion and storyteller gaze is sufficient for determining speaking turns between the two characters. This result needs to be confirmed with more extensive testing involving children as well as professionally-trained puppeteers.

In future work, we would like to develop intelligent cinematography and film editing techniques suitable to the display of the imaginary story worlds created in make-believe games.

Acknowledgment

The work was supported by the European Research Council advanced grant EXPRESSIVE (ERC-2011-ADG 20110209) and the PERSYVAL-Lab (ANR-11-LABX-0025-01) Labex.

References

[BID*99] BOBICK A. F., INTILLE S. S., DAVIS J. W., BAIRD F., PINHANEZ C. S., CAMPBELL L. W., IVANOV Y. A., SCHÜTTE A., WILSON A.: The KidsRoom: A Perceptually-Based Interactive and Immersive Story Environment. *Presence: Teleoperators and Virtual Environments 8*, 4 (1999), 369–393. 1

[CMB15] CEREZO E., MARCO J., BALDASSARRI S.: Hybrid games: Designing tangible interfaces for very young children and children with special needs. In *More Playful User Interfaces: Interfaces that Invite Social and Physical Interaction* (Singapore, 2015), Nijholt A., (Ed.), Springer, pp. 17–48. 2

[GJR14] GUPTA S., JANG S., RAMANI K.: Puppetx: A framework for gestural interactions with user constructed playthings. In *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces* (New York, NY, USA, 2014), AVI '14, ACM, pp. 73–80. 1, 2

[GPS10] GARZOTTO F., PAOLINI P., SABIESCU A.: Interactive storytelling for children. *Proceedings of the 9th International Conference on Interaction Design and Children IDC 10 2*, 1 (2010), 356. 1

[HGCA12] HELD R., GUPTA A., CURLESS B., AGRAWALA M.: 3d puppetry: A kinect-based interface for 3d animation. *Proceedings of UIST 2012* (2012), 423–433. 1, 2

[HRvG97] HAYES-ROTH B., VAN GENT R.: Story-Making with Improvisational Puppets. In *Proceedings of the first international conference on Autonomous agents (AGENTS '97)* (1997), pp. 1–7. 1

[MCP09] MAYORA O., COSTA C., PAPLIATSEYEU A.: itheater puppets: Tangible interactions for storytelling. In *International Conference on Intelligent Technologies for Interactive Entertainment* (2009), Springer, pp. 110–118. 1

[OTH02] OORE S., TERZOPOULOS D., HINTON G.: A desktop input device and interface for interactive 3d character animation. In *In Proc. Graphics Interface* (2002), pp. 133–140. 1

[Stu98] STURMAN D. J.: Computer puppetry. *Computer Graphics and Applications, IEEE 18*, 1 (1998), 38–45. 1

[Wal94] WALTON K. L.: *Mimesis as Make-Believe. On the Foundations of the Representational Arts.* Harvard University Press, 1994. 1