HAL
open science

Figurines, a multimodal framework for tangible storytelling Author version

Maxime Portaz, Maxime Garcia, Adela Barbulescu, Antoine Begault, Laurence Boissieux, Marie-Paule Cani, Rémi Ronfard, Dominique Vaufreydaz

To cite this version:

Maxime Portaz, Maxime Garcia, Adela Barbulescu, Antoine Begault, Laurence Boissieux, et al.. Figurines, a multimodal framework for tangible storytelling Author version. WOCCI 2017 - 6th Workshop on Child Computer Interaction at ICMI 2017 - 19th ACM International Conference on Multi-modal Interaction, Nov 2017, Glasgow, United Kingdom. hal-01595775v1

HAL Id: hal-01595775

https://inria.hal.science/hal-01595775v1

Submitted on 27 Sep 2017 (v1), last revised 2 Oct 2017 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Figurines, a multimodal framework for tangible storytelling

Maxime Portaz[1], Maxime Garcia[2], Adela Barbulescu[2], Antoine Begault[2], Laurence.Boissieux[1], Marie-Paule Cani[2], Rémi Ronfard[2], Dominique Vaufreydaz[1]

[1] Univ. Grenoble Alpes, CNRS, Inria, Grenoble INP*, LIG, 38000 Grenoble, France

[2] Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP*, LJK, 38000 Grenoble, France

Dominique.Vaufreydaz@inria.fr

Author version

Abstract

This paper presents *Figurines*, an offline framework for narrative creation with tangible objects, designed to record storytelling sessions with children, teenagers or adults. This framework uses tangible diegetic objects to record a free narrative from up to two storytellers and construct a fully annotated representation of the story. This representation is composed of the 3D position and orientation of the figurines, the position of decor elements and interpretation of the storytellers' actions (facial expression, gestures and voice). While maintaining the playful dimension of the storytelling session, the system must tackle the challenge of recovering the free-form motion of the figurines and the storytellers in uncontrolled environments. To do so, we record the storytelling session using a hybrid setup with two RGB-D sensors and figurines augmented with IMU sensors. The first RGB-D sensor completes IMU information in order to identify figurines and tracks them as well as decor elements. It also tracks the storytellers jointly with the second RGB-D sensor. The framework has been used to record preliminary experiments to validate interest of our approach. These experiments evaluate figurine following and combination of motion and storyteller's voice, gesture and facial expressions. In a make-believe game, this story representation was re-targeted on virtual characters to produce an animated version of the story. The final goal of the *Figurines* framework is to enhance our understanding of the creative processes at work during immersive storytelling.

Index Terms:Puppetry, Storytelling, Multimodal data fusion, RGB-D sensor, IMU sensor

1 Introduction

Storytelling is the art of creating or sharing a narrative. As famously emphasized by French structuralist Roland Barthes, forms of narrative are numerous and diverse around the world, and they always played an important role in human society [2].

* Institute of Engineering Univ. Grenoble Alpes

Figure 1: View from a storytelling session with a 6-year old child using the *Figurines* framework. At left, the frontal view with body and face detections (eyes, ears, nose, neck, shoulders, arms, hands) and facial landmarks of the storyteller using OpenPose [1]. At right, top view with body and hand detection, and decor tracking (blue points). Tracking of figurines is done using a hybrid IMU/RGB-D approach.

Narrative is widely considered to be a fundamental part of human cognition and understanding [3]. Today, storytelling is increasingly performed digitally and it is becoming easier to create stories using available digital technologies [4]. Storytelling is especially important for children education by supporting and enhancing creative expression and learning, as well as encouraging team work and sharing of personal experience [5, 6, 7].

Several authoring tools exist to help a narrator develop and construct his story [8]. In the film and game industries, motion capture is often used to create the 3D representation of movements or object manipulations. Existing motion capture systems generally require expensive hardware setup and/or attached marker [9]. They are not usable by the general public, especially by young children. The emergence of RGB-D sensors (like the Kinect) and of Fab-Labs with their new widespread technologies (3D printing, laser cutting ...) lets people envision developing lighter and cheaper systems. In parallel, the increasing performance of human perception algorithms thanks to Deep Learning [1, 10, 11, 12] tends to improve human analysis capabilities of interaction systems. Even if interaction systems take advantages of these recent progresses, there are still challenging problems to address. A first challenge concerns the underlying algorithms for object tracking and occlusion. These problems, still under investigation in many laboratories [13], are addressed in our system by a hybrid IMU/RGB-D approach. This challenge is not in the main scope of this paper and will not be detailed. A second challenge is to develop these tools while allowing rich narrative creation, with free character expressions and motion. The intrusiveness of the acquisition system must be reduced to a minimum, so that it does not disrupt the narrative flow of the storyteller. This challenge is the core of our system design.

This paper presents *Figurines* , a narrative capture/playback system for adults and children. Its first goal is to record one or two narrators while they are telling and playing a story. A story is composed of animated characters, evolving with rigid decor elements, in a given scene. In order to provide a natural interaction and to help the storytellers to be immersed in the story world, we propose to use tangible interaction with *diegetic* objects like figurines and decor elements. Diegetic objects are part of the story and they are present (i.e. exist) in the story, and not just tools or symbols artificially added to represent the story components. The second goal of the framework is to pro-

duce a complete synchronous representation of the story: 3D position and orientation of figurines, position of the decor elements and acting from the storytellers. One future purpose of this information is to analyze playing sessions to understand storytelling mechanisms. In our first experiments, this information was used to produce a 3D animated movie from the storytelling session. The final goal of the *Figurines* framework is, analyzing these session data, to enhance our understanding of the creative processes at work during immersive storytelling.

In Section 2, we present existing systems and related work. In Section 3, we explain our design and implementation. Finally, in Section 5, we present preliminary experimental results and evaluations made to test our system. Several storytelling sessions with one or two child narrators are analyzed. Limitations and lessons learned are discussed and conclusions are drawn.

2 Related Work

During the past years, many systems have been developed. Harley et al. [8] propose a survey and a classification of these systems. Different tangible interfaces exist to help children to experiment storytelling. Sylla et al. [14] and Chu et al. [15] present storytelling tangible interfaces for children. These systems use tangible objects as story components. They are suitable for children but they limit possible interaction and story structure. Character animation and orientation have been investigated with fully augmented puppets [16, 17] or even using the body of the puppeteer [18]. These interfaces allow precise and accurate control of the character. Tangible interfaces also exist with a simple webcam to track moves of a toy robot [19]. A large amount of work is needed to create the puppet and make it available to children. *ShadowStory* and *iTheater* [20, 21] let children animate puppets using accelerometer enabled devices.

Recently, several storytelling systems have taken advantage of RGB-D sensors. *PuppetX* [22] uses skeleton or hand/finger motions acquired from an RGB-D sensor. It retargets animation on a specific articulated puppet with servo-motors using manually defined rules to match between moves and puppet degrees of freedom. *3D puppetry* [23] takes advantage of the depth and associated color data. Using rigid colorized 3D models, it tracks poses of so-called puppets (cars, boats, etc.) to compute their 3D positions in the scene. *MotionMontage* [24] uses a similar approach to animate a virtual object along a path using a physical rigid object. One problem with purely vision-based systems is that their performance degrades rapidly in cases of occlusion. The storyteller must be very careful to keep the puppet visible from the camera at all times while acting. One way to do so is to provide feedback to the storyteller but this may interfere with the narration fluidity, moreover for children. Narrators may look to the feedback screen regardless of their narrative goals. To tackle this problem in the *Figurines* framework, we decided to combine IMU (Inertial Motion Unit) sensors with an RGB-D sensor in order to reduce occlusion problems as much as possible. This choice leads us to address the drifting problem of IMU path reconstruction with improvement over existing algorithms (see section 4.1). Similarly to *PuppetX* [22] and to *i-marionette* [18], we also want to benefit from storyteller body language to increase relevance of the resulting animation. Therefore, we include a second RGB-D camera to track the storytellers and record their body poses, facial expressions and voices.

3 Design and Implementation

In this section, we first define the recording scenario, prior of the design of our system. Then, the acquisition setup is presented. Last, design of the narrative elements and gathered data about figurines and storytellers are described.

3.1 Recording scenario

To increase playfulness, the system must be non intrusive and must not impose narrative schemes. The storytellers can use any object of any shape to play their story. The first type of objects is called *figurine* in the framework. A figurine is an object of importance in the story (prince, animal, car...). The second type of objects is *decor element*. As far as they fit into the playground (see 3.2), any object can be a decor element.

The recording scenario of the narrative session is quite simple. The narrators place themselves in front of the recording table. Several decor elements and different figurines are available:

1. storytellers choose among decor items and figurines to play with;
2. they start to organize the stage at their convenience to tell the story;
3. they can freely play their story as long as they want.

To increase the recreational dimension of the storytelling session, all calibrations are done off-line without the narrators. No specific action is mandatory from the narrator to help the system. This property is obviously even more important for children.

3.2 Acquisition setup

The acquisition system records everything from the augmented figurines, the storytellers and the decors. In its current configuration, due to space constraint and camera view angle, the system handles at maximum two simultaneous narrators and a 70cm x 70cm playing area. The full setup is shown on figure 2. It tracks figurines on the stage and the storytellers with several devices. An overhead Kinect and a set of Inertial Motion Units (IMU) track the figurines. The top RGB-D device is accurate enough for tracking moves and distance of figurines, and decor elements on the stage. The frontal Kinect complete the top one to record the storytellers and their behaviors (see fig. 2 and 1).

As stated before, the main difficulty to address is occlusion problems. Occlusions can be caused by the storyteller himself, by another figurine, or by a decor element. Figurines can disappear from the overhead camera. That's why we included IMUs in our setup (see 3.3). This brings additional advantages over previous works: it is not mandatory to scan the figurines before tracking, and most importantly, we can use non-rigid objects (articulated, dressed and/or soft puppets for instance).

The acquisition setup records lots of raw data. All the streams from the RGB-D devices are recorded synchronously in uncompressed format at full frame rate: RGB, depth, infrared, skeletons, faces and audio streams. All IMUs information is synchronously stored in the system. Due to technical reason, data processing has to be done off-line after the recording. Data from the IMUs cannot be synchronized on-line, due to the low power Bluetooth 4.0 that could not transfer data when the acquisition

Figure 2: Acquisition setup. There are two Kinect devices: one looking down to follow figurines, one following narrators. The playground area is the table in the middle. Its size in our experiment is 70cm x 70cm.

frame rate is 200Hz. Moreover, real time processing would cause a limited choice for computer vision algorithms used in the framework. As on-line processing is not mandatory for the storyteller, and would not improve the session, it is not a constraint in the framework.

3.3 Figurine and decor design

In the last years, new widespread and affordable digital fabrication technologies are available for researchers but also, within Fab-Labs, for the general public. We decided to take benefit from these technologies to build personalized figurines and decor elements. As can be seen, we created a specific box to serve as basement for 3D printed figurine (figures 3 and 4) or as IMU container for enhanced usual puppet (figure 5). One benefit of this process is that we can create upon request almost any figurine. Another benefit arises when we produce a 3D movie from the storytelling session (see figure 6). The printing models can be reused for the 3D rendering.

As said, figurines are of importance for the story. This justifies the need for a fine tracking. Figurines are thus augmented with an IMU to improve its monitoring (figures 3 and 4). Any puppet or toy that can be enhanced with an IMU can be integrated in a story. Even decor elements, if they need to be active part of the story can be equipped and become a figurine. For instance, Figure 5 in Appendix shows a handmade car build by a child with bricks. Preliminary experiments with different consumer

Figure 3: At left, 3D models used to print an enhanced princess figurine. At right, the result dressed figurine.

Figure 4: Figurine examples. Left, the IMU in the basement of the soldier figurine. Center and right, the soldier and prince dressed figurines respectively.

IMUs demonstrated that even a 50Hz frame rate is not reliable enough to reconstruct 3D path because of acceleration information sparsity and drift. Among professional available IMUs, we selected the 10-degrees-of-freedom Hikob Fox IMU[1]. This choice was driven by several technical aspects. These sensors are very light (~20 gr with their embedded battery, memory card and printed basement). They are able to record gyroscope, accelerometer and magnetometer data at high frame rate (up to 200Hz) on their memory card. Their storage and their rechargeable battery allow an autonomy of several hours. Last crucial point, using wireless synchronization, all IMUs share a common time reference. For decor elements, everything that fit the playing area can be used and tracked by the framework. In our prototype, we designed decor objects with a laser cutter (see figure 2). These decor elements fit perfectly with the figurines, as their scale has been chosen accordingly.

[1] http://www.hikob.com/wp-content/uploads/2015/06/HIKOB_FOX_ProductSheet_EN.pdf (last seen 07/2017)

4 Storytelling data

This section describes data computed on the figurines and the decor elements on stage, and about the storytellers. All the processing described here are offline after the recording of the narrative session. In this article, we do not detail underlying mathematics and algorithms used in the framework but present the general principle.

4.1 Stage tracking

Using the overhead RGB-D device, it is possible to track decor elements. The first step is to use the depth data to do their automatic detection over the playground. This automatic detection can be corrected at any time while processing for mis-detected objects. In a second step, the system tracks moves from decors elements using a dense optical flow. In the current implementation, decor tracking follows center of objects in 2D position (x,y), that is, the system tracks them but does not provide neither their orientation nor their height ($z = 0$). Once again, if this information is mandatory for the storyteller (a magic tree for instance), the decor element can be transformed into a figurine.

One can say that the figurine tracking is a hybrid IMU/RGB-D algorithm. The main tasks to address are figurine identification, tracking in 3D space and computing orientation over time. For the identification task, when a new mobile object is detected in the depth data, it is compare to synchronous IMUs data of actual moving figurines. If one figurine matches the current motion, its label is tagged over the mobile object. For the tracking aspect, as far as a figurine can be seen in the Kinect view, standard tracking paradigm using a Kalman filter is applied. When a figurine disappeared (under a decor element, hidden by the narrator hand, out of the camera view, ...), the only available information are gathered from IMUs. The IMUs give us the acceleration and the orientation, so we can compute the position from the last known position integrating twice the acceleration. A drifting problem appears fast with this method. Our implementation corrects the drift using an improved version of Neto's algorithm [25] and let us reconstruct the 3D path of the figurine until it is identifiable again. Finally, using magnetometer and gyroscope information, the figurine orientation is computed using Madgwick's method [26].

4.2 Storytellers' information

The tracking of the storytellers is done using both RGB-D devices. As seen on figure 1, the recordings include body tracking (frontal and from top), face tracking and sound. Body tracking is limited to upper-body tracking (head, shoulders, torso, arms, wrist and hands). The body tracking algorithm uses a modified version of the *Realtime Multi-Person Pose Estimation* algorithm [10]. Faces and hands are tracked using OpenPose [1, 10, 11, 12]. Using this software, we are able to compute facial landmark, head pose, eyes and facial Action Units. The recorded sound is tagged into voice segments. Optionally, speaker diarization can be applied to partition the audio stream according to the storyteller identity [27]. Using such a system could improve information gathered by the system and the speech slot association with figurines within the storytelling session.

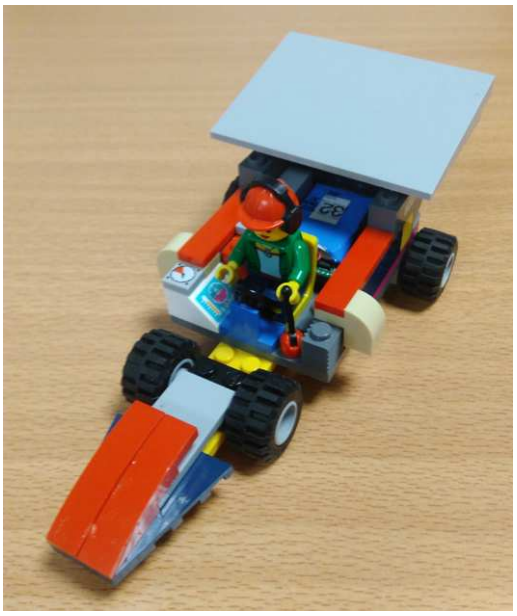

Figure 5: Child handmade Lego© car with an IMU (behind the driver's seat).

5 Evaluation of narrative session recordings

Preliminary evaluations were conducted in a dedicated room. According to the scenario (see section 3.1), we filmed narrative sessions with several users to check the usability of the system. We conducted experiments with children to verify simplicity and usability of the system: two children play the storyteller role, individually then together. We also wanted to check that narrative process is not disturbed by the recording system. Several sessions with adults have been made during the development phase. As the room does not have one-way mirror, the examiner remained in the room with the narrators for the entire session. Observations about these sessions are discussed in the following paragraphs.

In this first experiment, the storytellers did not model their own characters. We tried to provide them with enough different figurines recurring in folktales: the prince, the princess, the witch, the soldier/knight, the dragon, the horse and the wolf. We provided standard decor elements like several different trees, bridge, house, tower and etc.

5.1 Children sessions

Three narrative sessions with two children were recorded. Each child records one session alone (6 minutes for the 6-year old boy, 20 minutes for the 7-year old girl). They perform also an 11 minutes narrative session together.

The system was able to reconstruct the figurines movements and to track the decor elements. The main problems come from the storyteller tracking. The children faces were often lost on the video, due to three reasons:

1. To take advantage of the whole scene, children have to stretch their arm and to come closer to the table. Doing so, they are hidden by their arm (see fig 1) or too close to the RGB-D device to get depth data.

2. The children are sometimes hidden by the decor (the tower in our experiment).

3. In one session, the child tells his story looking at the experimenter, thus his face in profile was not detected.

For these children sessions, we recovered the upper body information 94.03% of time. The percentage of face detection is also 74.36% over time (only 11.48% using the standard Kinect2 face detection). There is no significant difference with sessions with one or two children.

Regarding the storytelling aspects, children seem not to be disturbed by the acquisition system. With one child, the presence of the examiner was a discomfort. As said, the child looked at the examiner and tended to explain the story to him, instead of playing each character role. In further experiments, we will equip the room with a one-way mirror to solve that problem.

5.2 Adult sessions

Four adults played with all the figurines and the scenery we built for a total of 13 minutes. The system was able to track each figurine, the stage and the storyteller. The face was detected and tracked with more accuracy than with children (98.02% of the time). Adults have longer arm and do not need to get closer to the table. The upper body is tracked 100% of the time. Contrary to children, in our experiments, adults have much more difficulties to imagine stories with imposed characters. As we did not let them print or construct their own figurines for these preliminary experiments, some of the participants expressed discomfort using the provided figurines. Partly for this reason, all adult storytelling sessions are shorter.

5.3 Make-believe games

We also used our framework to record imaginary dialogues for a make-believe game [28]. The storyteller's voice, gestures and facial expressions combined with movements of instrumented figurines were transferred to virtual characters in order to obtain an animated version of the dialogue. A rendering example is presented in figure 6 and in this video https://hal.inria.fr/hal-01518981v2/file/wicedcrc.mp4.

6 Lessons learned

Figurines our storytelling framework, benefits from RGB-D sensors and IMU technologies. However, it has intrinsic limitations. Due to the setup (figure 2) and the Kinect angle of view, a maximum of 2 storytellers can be recorded at a time. The number of figurines has been limited to 4 in all our experiments. This is not a strict limitation but one can figure out that increasing the number of figurines may lead to less accurate figurine identification. For the RGB-D acquisition system, we learned some lessons. The frontal Kinect must be higher and further away from the stage. It will overhang the decor elements and prevent from occlusions. This will improve perception of child storytellers.

The 3D reconstruction is efficient in our context but it is not perfect. Even small collisions of the figurines with solids have a huge effect on the measured IMU accelerations. These perturbations cannot trivially be filtered. 3D path reconstruction is actually under investigation. Manual corrections may be needed to improve quality of

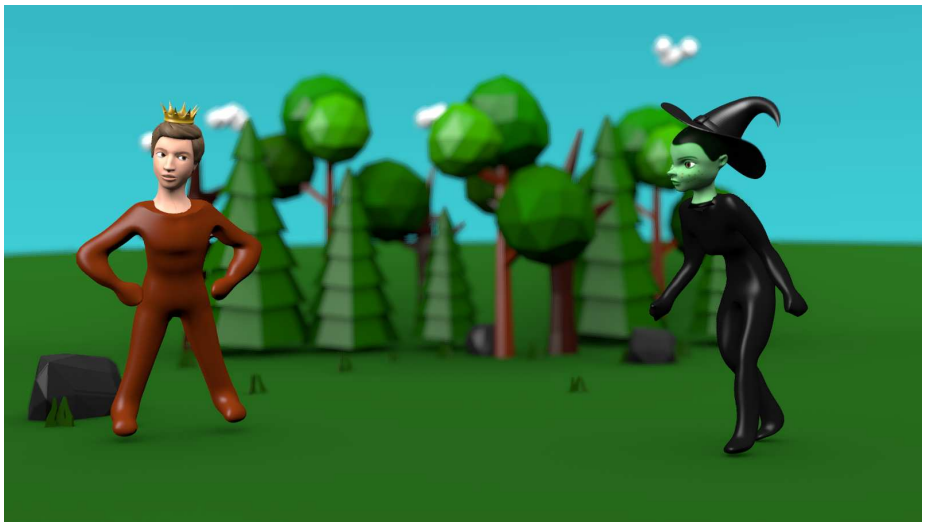

Figure 6: 3D rendering example generated from a recorded session of a make-believe game.

the 3D path reconstruction from the recordings. In our case it is not an issue, as artistic corrections can also be done to improve expressiveness in the final rendering like accentuating some moves or expressions for instance.

7 Conclusion

We have presented *Figurines* a hybrid multimodal framework for recording and playback of storytelling sessions involving tangible interaction with objects. In this framework, tangible objects are figurines enhanced with IMUs and decor elements. Figurines can be articulated, dressed and/or soft puppets. The system records the storytelling sessions using two RGB-D sensors. The overhead RGB-D sensor follows the figurines on the stage using computer vision algorithms in combination with IMUs. Up to two storytellers are monitored with a frontal RGB-D sensor. The system records their facial expressions, upper body motion and voice activity. Output of the system is a synchronous representation of the story: 3D position and orientation of figurines, position of the decor elements and interpretation of the storytellers. This information can be used as input for a 3D rendering system to produce a video animation of the story. Our framework can already be used to create multimodal recording of make-believe games, and we hope this will enhance our understanding of the creative processes at work during immersive storytelling.

In future work, we would like to let users, both children and adults, freely design and print their own story worlds, including sets, props and characters [29] and turn their stories into movies using intelligent tools for 3D animation [30] and cinematography [31]. A variation of the *Figurines* framework has also been used in an experiment to monitor players while solving Chess problem [32]. In this setup, the gathered data are used to infer mental state and chess level of the players.

8 Acknowledgments

The authors would like to thank the experimentation participants who contributed by sharing their stories. We want to thank the SED team for their technical support. Prototyping was done using the Amiqual4Home facilities (ANR-11-EQPX-0002). This work was partly funded by the PERSYVAL-Lab (ANR-11-LABX-0025-01) Labex.

References

[1] (2017) Openpose: A real-time multi-person keypoint detection and multi-threading c++ library. [Online]. Available: https://github.com/CMU-Perceptual-Computing-Lab/openpose

[2] R. Barthes, "An introduction to the structural analysis of narrative," *New literary history*, pp. 237–272, 1975.

[3] K. Newman, "The case for the narrative brain," in *Proceedings of the second Australasian conference on Interactive entertainment*. Creativity & Cognition Studios Press, 2005, pp. 145–149.

[4] J. Van Dijck, "Users like you? theorizing agency in user-generated content," *Media, culture, and society*, vol. 31, no. 1, p. 41, 2009.

[5] B. Nojavanasghari, T. Baltrušaitis, C. E. Hughes, and L.-P. Morency, "The future belongs to the curious: Towards automatic understanding and recognition of curiosity in children," in *Workshop on Child Computer Interaction*, 2016, pp. 16–22.

[6] J. A. Fails, A. Druin, and M. L. Guha, "Interactive storytelling: interacting with people, environment, and technology," *International Journal of Arts and Technology*, vol. 7, no. 1, pp. 112–124, 2014.

[7] S. Benford, B. B. Bederson, K.-P. Åkesson, V. Bayon, A. Druin, P. Hansson, J. P. Hourcade, R. Ingram, H. Neale, C. O'Malley *et al.*, "Designing storytelling technologies to encouraging collaboration between young children," in *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 2000, pp. 556–563.

[8] D. Harley, J. H. Chu, J. Kwan, and A. Mazalek, "Towards a framework for tangible narratives," in *Proceedings of the TEI'16: Tenth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, 2016, pp. 62–69.

[9] D. J. Sturman, "Computer puppetry," *Computer Graphics and Applications, IEEE*, vol. 18, no. 1, pp. 38–45, 1998.

[10] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," in *CVPR*, 2017.

[11] T. Simon, H. Joo, I. Matthews, and Y. Sheikh, "Hand keypoint detection in single images using multiview bootstrapping," in *CVPR*, 2017.

[12] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *CVPR*, 2016.

[13] M. Camplani, S. Hannuna, M. Mirmehdi, D. Damen, A. Paiement, L. Tao, and T. Burghardt, "Real-time rgb-d tracking with depth scaling kernelised correlation filters and occlusion handling," in *Proceedings of the British Machine Vision Conference (BMVC). pp*, 2015, pp. 145–1.

[14] C. Sylla, S. Gonçalves, P. Brito, P. Branco, and C. Coutinho, "A tangible platform for mixing and remixing narratives," in *Advances in Computer Entertainment*. Springer, 2013, pp. 630–633.

[15] J. H. Chu, P. Clifton, D. Harley, J. Pavao, and A. Mazalek, "Mapping place: Supporting cultural learning through a lukasa-inspired tangible tabletop museum exhibit," in *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, 2015, pp. 261–268.

[16] W. Yoshizaki, Y. Sugiura, A. C. Chiou, S. Hashimoto, M. Inami, T. Igarashi, Y. Akazawa, K. Kawachi, S. Kagami, and M. Mochimaru, "An actuated physical puppet as an input device for controlling a digital manikin," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 637–646.

[17] A. Mazalek and M. Nitsche, "Tangible interfaces for real-time 3d virtual environments," in *Proceedings of the international conference on Advances in computer entertainment technology*. ACM, 2007, pp. 155–162.

[18] S.-Y. Lin, C.-K. Shie, S.-C. Chen, and Y.-P. Hung, "Action recognition for human-marionette interaction," in *Proceedings of the 20th ACM international conference on Multimedia*. ACM, 2012, pp. 39–48.

[19] R. Slyper, G. Hoffman, and A. Shamir, "Mirror puppeteering: Animating toy robots in front of a webcam," in *Proceedings of the Ninth International Conference on Tangible, Embedded, and Embodied Interaction*. ACM, 2015, pp. 241–248.

[20] F. Lu, F. Tian, Y. Jiang, X. Cao, W. Luo, G. Li, X. Zhang, G. Dai, and H. Wang, "Shadowstory: creative and collaborative digital storytelling inspired by cultural heritage," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2011, pp. 1919–1928.

[21] O. Mayora, C. Costa, and A. Papliatseyeu, "itheater puppets tangible interactions for storytelling," in *International Conference on Intelligent Technologies for Interactive Entertainment*. Springer, 2009, pp. 110–118.

[22] S. Gupta, S. Jang, and K. Ramani, "Puppetx: a framework for gestural interactions with user constructed playthings," in *Proceedings of the 2014 International Working Conference on Advanced Visual Interfaces*. ACM, 2014, pp. 73–80.

[23] R. Held, A. Gupta, B. Curless, and M. Agrawala, "3d puppetry: a kinect-based interface for 3d animation." in *UIST*. Citeseer, 2012, pp. 423–434.

[24] A. Gupta, M. Agrawala, B. Curless, and M. Cohen, "Motionmontage: A system to annotate and combine motion takes for 3d animations," in *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems*, ser. CHI '14. New York, NY, USA: ACM, 2014, pp. 2017–2026. [Online]. Available: http://doi.acm.org/10.1145/2556288.2557218

[25] P. Neto, J. N. Pires, and A. P. Moreira, "3-d position estimation from inertial sensing: minimizing the error from the process of double integration of accelerations," *CoRR*, vol. abs/1311.4572, 2013. [Online]. Available: http://arxiv.org/abs/1311.4572

[26] S. Madgwick, "An efficient orientation filter for inertial and inertial/magnetic sensor arrays," *Report x-io and University of Bristol (UK)*, 2010.

[27] M. Najafian and J. H. Hansen, "Speaker independent diarization for child language environment analysis using deep neural networks," in *Spoken Language Technology Workshop (SLT), 2016 IEEE*. IEEE, 2016, pp. 114–120.

[28] A. Barbulescu and M. Garcia and A. Begault and M. P. Cani and M. Portaz and A. Viand and R. Dulery and L. Boissieux and P. Heinish and R. Ronfard and D. Vaufreydaz, "A system for creating virtual reality content from make-believe games," in *2017 IEEE Virtual Reality (VR)*, March 2017, pp. 207–208.

[29] M. Skouras, B. Thomaszewski, S. Coros, B. Bickel, and M. Gross, "Computational design of actuated deformable characters," *ACM Transactions on Graphics (TOG)*, vol. 32, no. 4, p. 82, 2013.

[30] J. Chai and J. K. Hodgins, "Performance animation from low-dimensional control signals," *ACM transactions on Graphics, Proceedings of SIGGRAPH*, pp. 686–696, 2005.

[31] Q. Galvane, R. Ronfard, M. Christie, and N. Szilas, "Narrative-driven camera control for cinematic replay of computer games," in *Proceedings of the Seventh International Conference on Motion in Games*, ser. MIG '14. ACM, 2014, pp. 109–117.

[32] T. Guntz, D. Vaufreydaz, R. Balzarini, and J. Crowley, "Multimodal observation and interpretation of subjects engaged in problem solving," in *1st Behavior, Emotion and Representation: Building Blocks of Interaction Workshop at 5th International Conference on Human-Agent Interaction*. ACM, 2017.