Why is the Creation of a Virtual Signer Challenging Computer Animation ?

Nicolas Courty, Sylvie Gibet

Why is the Creation of a Virtual Signer Challenging Computer Animation ?

Nicolas Courty and Sylvie Gibet

Université de Bretagne Sud, Laboratoire VALORIA, Bâtiment Yves Coppens, F-56017 Vannes, FRANCE

Abstract. Virtual signers communicating in signed languages are a very interesting tool to serve as means of communication with deaf people and improve their access to services and information. We discuss in this paper important factors of the design of virtual signers in regard to the animation problems. We notably show that some aspects of these signed languages are challenging for up-to-date animation methods, and present possible future research directions that could also benefit more widely the animation of virtual characters.

1 Introduction

Signed languages (SL), defined as visual languages, were initially intended to be a mean of communication between deaf people. They are entirely based on motions and have no written equivalent. They constitute full natural languages, driven by their own linguistic structure. Accounting for the difficulties of deaf to read text or subtitles on computers or personal devices, computer animations of sign language improve the accessibility of those media to these users [27, 5, 20, 9]. The use of avatars to this purpose allows to go further the restrictions of videos, mostly because the possibilities of content creation with avatars are far more advanced, and because avatars can be personalized along with the user's will. They also allow the anonymity of the interlocutor.

However, animating virtual signers has revealed to be a tedious task [17], mostly for two reasons: *i)* our comprehension of the linguistic mechanisms of signed languages are still not fully achieved, and computational linguistic software may sometimes fail in modeling particular aspects of SL *ii)* animation methodologies are challenged by the complex nature of gestures involved in signed communication. This paper focuses on this second class of problems, even though we admit that in some sense those two aspects are indissociable.

In fact, signs differ sensibly from other non-linguistic gestures, as they are by essence multichannel. Each channel of a single sign (those being the gestures of the two arms and the two hands, the signer's facial expressions and gaze direction) conveys meaningful information from the phonological level to the discourse level. Moreover, signs exhibit a highly spatial and temporal variability that can serve as syntactic modifiers of aspect, participants, etc. Then, the combination in space and time of two or more signs is also possible and sometimes mandatory

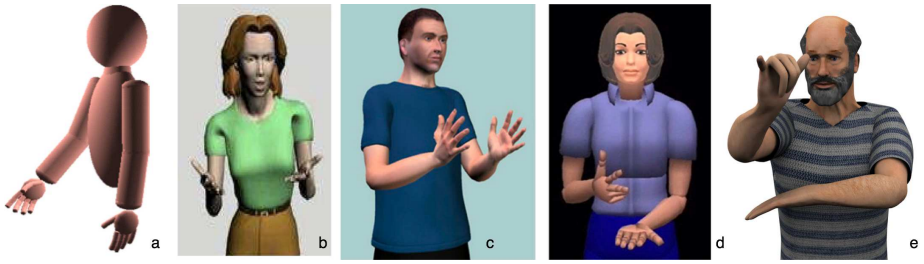

Fig. 1. Some virtual signers classified in chronological order: (a) the GESSYCA system [10] (b) Elsi [8] (c) Guido from the eSign european project [20] (d) the virtual signer of the City University of New-Yord [17] (e) Gerard [2]

to express concisely ideas or concepts. This intricate nature is difficult to handle with classical animation methods, that most of the time focus on particular types of motions (walk, kicks, etc.) that do not exhibit a comparable variability and subtleties.

The remainder of the paper is organized as follows: a brief state-of-the-art presents some existing virtual signers and the two aspects of sign generation: procedural and data-driven methods (Section 2), then challenges in the production of signs are exposed (Section 3) and finally a collection of unresolved virtual character animation problems are presented (Section 4).

2 Existing Virtual Signers

We first begin by reviewing some of the technologies used to animate virtual signers. Figure 1 presents in chronological order some existing virtual signers.

2.1 Descriptive and generative methods

Several gesture taxonomies have already been proposed in [19] and [28], some of which rely on the identification of specific phases that appear in co-verbal gestures and sign language signs [22]. Recent studies dedicated to expressive gesture rely on the segmentation and annotation of gestures to characterize the spatial structure of a sign sequence, and on transcribing and modeling gestures with the goal of later re-synthesis [21].

Studies on sign languages formed early description/transcription systems, such as [33] or [31]. More recently, at the intersection of linguistics and computation, gestures have been described with methods ranging from formalized scripts to a dedicated gestural language. The BEAT system [4], as one of the first systems to describe the desired behaviors of virtual agents, uses textual input to build linguistic features of gestures to be generated and then synchronized with speech. Gibet et al. [10] propose a gesture synthesis system based on a quantified description of the space around the signer; using the HamNoSys [31]

Fig. 2. Photo of the motion capture settings in the Signcom project

sign language notation system as a base, the eSign project has further designed a motion specification language called SigML [7]. Other *XML*-based description languages have been developed to describe various multimodal behaviors, some of these languages are dedicated to conversational agents behaviors, as for example MURML [24], or describe style variations in gesturing and speech [30], or expressive gestures [11]. More recently, a unified framework, containing several abstraction levels has been defined and has led to the XML-based language called BML [36], which interprets a planned multimodal behavior into a realized behavior, and may integrate different planning and control systems.

Passing from the specification of gestures to their generation has given rise to a few works. Largely, they desire to translate a gestural description, expressed in any of the above-mentioned formalisms, into a sequence of gestural commands that can be directly interpreted by a real-time animation engine. Most of these works concern pure synthesis methods, for instance by computing postures from specification of goals in the 3D-space, using inverse kinematics techniques, such as in [10], [35], [23]. Another approach uses annotated videos of human behaviors to synchronize speech and gestures and a statistical model to extract specific gestural profiles; from a textual input, a generation process then produces a gestural script which is interpreted by a motion simulation engine [29].

Alternatively, data-driven animation methods can be substituted for these pure synthesis methods. In this case the motions of a real signer are captured with different combinations of motion capture techniques. Since it is not possible to record every possible sentences, new strategies are to be devised in order to produce new utterances, The next paragraph presents an example of a fully data-driven approach.

2.2 An example of a full data-driven approach: the Signcom project

An example of a full data-driven virtual signer is given by the Signcom project, which aims at improving the quality of the real-time interaction between real humans and avatars, by exploiting natural communication modalities such as

Fig. 3. Screenshots of the virtual signer "Sally" from the Signcom project

gestures, facial expressions and gaze direction. Based on French Sign Language (FSL) gestures, the real human and the virtual character produce statements towards their interlocutor through a dialog model. The final objective of the project consists in elaborating new ways of communication by recognizing FSL utterances, and synthesizing adequate responses with a 3D avatar. The motion capture system uses Vicon MX infrared camera technology to capture the movements of our LSF informants at frame rates of 100 Hz. The setup was as follows: 12 motion capture cameras, 43 facial markers, 43 body markers, and 12 hand markers. In order to replay a complete animation, several post operations are necessary. First, the fingers' motion were reconstructed by inverse kinematics, since only the fingers' end positions were recorded. In order to animate the face, cross-mapping of facial motion capture data and blendshapes parameters was performed [6]. This technique allows to animate directly the face from the raw motion capture data once a mapping has been learned. Finally, since no eye gazes were recorded during the informants performance, an automatic eye gazing systems was designed. Figure 3 gives some illustrations of the final virtual signer "sally" replaying captured motions. A corpus annotation was also conducted. Annotations expand on the mocap data by identifying each sign type with a unique gloss, so that each token of a single type can be easily compared. Other annotations include grammatical and phonological descriptions.

From recorded FSL sequences, multichannel data are retrieved from a dual-representation indexed database (annotation and mocap data), and used to generate new FSL utterances [2], in a way similar to [1]. At that time, the final system is currently under evaluation with native LSF signers.

3 Challenges in Sign production

Though data-driven animation methods significantly improve the quality and credibility of animations, there are nonetheless several challenges to the reuse of motion capture data in the production of sign languages. Some of them are presented in the following.

Spatialization of the content As sign languages are by nature spatial languages, forming sign strings requires a signer to understand a set of highly spatial and temporal grammatical rules and inflection processes unique to a sign language. We can separate plain signs that do not use space semantically (like the American Sign Language sign HAVE which does not make any notable use of space other than which is necessary for any sign) from signs that incorporate depiction. This second group of signs includes the strongly iconic signs known as depicting verbs (or classifiers), which mimic spatial movements, as well as size-and-shape specifiers, which concern static spatial descriptions.

Moreover, indicating signs like indicating verbs and deictic expressions require the signer to interface with targets in the signing space by effecting pointing-like movements towards these targets. Indicating verbs include such signs as the LSF sign INVITER, in which the hand moves from the area around the invited party toward the entity who did the inviting . Depending on the intended subject and object, the initial and final placements of the hand vary greatly within the signing space. Deixis, such as pronouns, locatives, and other indexical signs are often formed with a pointed index finger moving toward a specific referent, though other hand configurations have been reported in sign languages, such as American Sign Language.

Small variations can make big semantic differences Sign languages require precision and rapidity in their execution, but at the same times imperfection in the realization of the signs or bad synchronization can change the semantic content of the sentence. We give here some challenging elements in the execution of signs:

- **Motion precision.** The understandability of signs require accuracy in the realization of the gestures. In particular in finger spelling the degree of openness of a fingers leads to different letters. Some of fhe different hand shapes used in FSL only differ by the positions of one finger or by the absence or not of a contact. This calls for a great accuracy in the capture and animation processes.
- **spatio-temporal aspects of the gestures.** The sign language being a language with highly spatio-temporal components, the question of timing and dynamics of gesture is crucial. In fact, three elements are of interest for a sign: first, the spatial trajectory of the hands are rather important. They do not only constitute transitions in space between two key positions, but may be constituent of the sign. This raises the problem of the coding of this trajectory. Second, synchronization of the two hands is a major component,

and usually hands do not have to this regard a symmetric role, In the case of PAS D'ACCORD (not agree), the index start from from the forehead and meets the other index in front of the signer. The motion of the second hand is clearly synchronized on the first hand. Third, the dynamics of the gesture (acceleration profile along time) allows the distinction between two significations. An example is the difference between the signs CHAISE (chair) and S'ASSEOIR (to sit), which have the same hands configurations, the same trajectories in space, but different dynamics. Let us finally note that the dynamics of contacts between the hand and the body (gently touching or striking)is also relevant.

– **facial expressions and non manual elements.** While most of the description focus on the hands configuration and their motions, important non manual elements should also be taken into account, like shoulder motions, head swinging, changes in gazes or facial mimics. For example, the gaze can be used either to recall a particular object of the signing space, or either directed by the dominating hand (like in the sign LIRE, to read, where the eyes follow the motion of fingers). In the case of facial mimics, some facial expressions may serve as adjectives (for instance inflated cheeks will make an object big, while wrinkled eyes would make it thin) or indicate wether the sentence is a question (raised eyebrows) or an affirmation (frowning). It is therefore very important to preserve these informations in the facial animation.

4 Unresolved animation problems

Regarding the different requirements exposed in the previous Section, several unresolved computer animation problems are presented here. Those problems are not particularly exclusive to the animation of virtual signers, and can address more widely general virtual character animation problem.

High frequency full body and facial motion capture. Signs are by nature very dexterous and quick gestures, that involve at the same time several modalities (arms, hands, body, gaze and facial expressions). Capturing accurately all these channels with an appropriate frequency (> 100 Mhz) actually pushes motion capture equipment to their very limits. It could be argued that splicing methods such as [26] would allow to capture independently the different modalities, and then combine them during a post process phase. However, the temporal synchronization issues raised by this method seem hard to alleviate. Moreover, asking the signer to perform alone the facial expressions corresponding to given sentences is also out of reach, since most of the facial mimics are generally done unconsciously. A parallel could be drawn with non-verbal communication: could we ask someone to perform accompanying gestures of an unspoken discourse ? Finally, new technologies such as surface capture [32], that captures simultaneously geometry and animation, are very attractive, but yet the resolution is not sufficient to capture the body and the face with an adequate precision, and only

very few methods exist to manipulate this complex data in order to produce new animations.

Expressivity filtering. As seen in the previous Section, the spatio-temporal variability of signs can be used as as adjectives, or in a more general way, to inflect the nature of a sentence and enhance the global expressivity of the virtual signer. It has been shown [15] that temporal alignment methods [13] can be efficiently used to change the style and expressivity of a captured sentence. Nevertheless, big variations in style are can not only obtained by changing the timing of gestures, but most often by the change of spatial trajectories, and sometimes may inflect the entire sentence. Most of existing methods that build statistical models [37] of gestures may fail for this purpose, mostly because the style transfer is encoded by higher level linguistic rules, and because pure signal approaches are insufficient to model this variability.

Advanced motion retargeting. Most of the actual motion retargeting techniques focus on the adaptation of motion to changing the physical conditions of the motion [34] or more frequently kinematic constraints [18, 25, 12] through the use of inverse kinematic techniques. In the case of sign language the spatial relations between the fingers and the arms or the head are key elements for the comprehension of the discourse and should be preserved in the retargeting process. To this end, the recent work of Ho and colleagues [16] is really attractive, provided that the important relation between limbs could be preserved by their methods. Yet Its application to sign language synthesis remains to be explored. Whereas interaction with the floor or objects in the environment lead to hard constraints which lead to difficult optimization problems and procedures, constraints in sign language may be more diffuse or expressed qualitatively (e.g. "the thumb should touch the palm of the hand"). Algorithms dealing with such fuzzy or high level constraints could be extremely interesting, both numerically (more degrees of freedom while optimizing) and from a usability point of view. Finally, since arms motions are involved, a planing phase may also be required to avoid self collisions. Combined inverse kinematics and planing algorithms could be used [3], as well as more recent hybrid approaches [38]. Yet, real time algorithms for this class of problems remain to be found.

Multichannel combinations. As exposed in [2], the possibility of building new signed utterances by composing selectively pre-exisiting elements of a corpus data is possible. In this option, not only the spatial coherency should be preserved, but as well the channel's temporal synchronization:

- *spatial coherency.* Sign language allows to combine different gesture with different meanings at the same time, thus providing several information in a minimum of gestures. This combination differs from the classical blending approaches which mix motions together to produce new ones [1], as far as topological constraints should be preserved in the composition process. An example is given in Figure 4, where the same pose indicates at the same

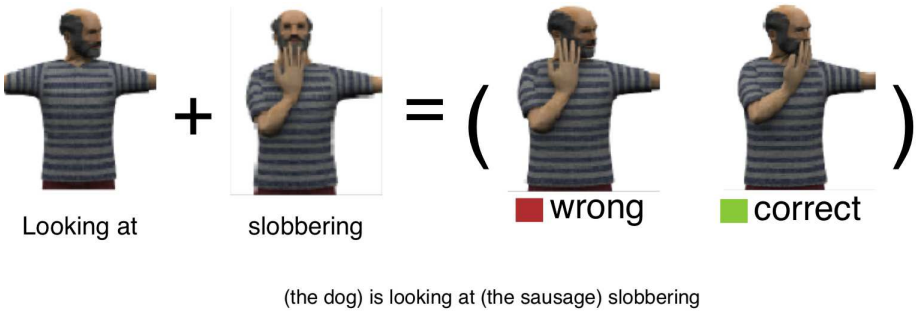

Fig. 4. Combination of two signs ("looking" and "slobbering")

time that a dog is looking at (first sign) something while slobbering (sign 2). If both signs were to be recorded independently, a naive blending operation would fail because the hand would not anymore be located in front of the mouth. Moreover, as exposed in the previous Section, every spatialized gestures should be retargeted with respect to the current signing space. This brings us back to the problem of advanced motion retargeting, but also clearly reveals that the combination process should be driven by more abstract definition, possibly of linguistic nature.

– *temporal synchronization.* It is likely that the different motion elements have not the same duration. The consequent problem is twofold: *i)* a common timeline has to be found, eventually as the result of a combinatorial optimization, or driven by linguistic rules. Up to our knowledge though, no existing model of sign language describe such temporal rules or model the synchronization of the different channels *ii)* once a correct time plan has been devised, the temporal length of the motion chunks has to be adapted, while preserving the dynamic of the motions. To this end, time warping techniques can be used [13]. However, inter channels synchronizations may exist (for example between the hand and the arm motions [14]). Those synchronization schema can be extracted from analysis, but the proper way to introduce this empirical knowledge in the synthesis process has not been explored yet.

5 Conclusion

We examined in this article the different challenges posed by the animation of virtual agent communicating in sign language. While data-driven animation techniques clearly lead to the best natural results, a lot of improvements are still mandatory to fulfill the requirements of sign languages. Among others, capture techniques and retargeting algorithms are severely challenged by the complex spatial and temporal schemas involved in signs. In parallel, those improvements should accompany progresses in the modeling of sign language, which is in itself a critical issue. In a second step, the usability and acceptability of virtual signers to the community of deaf people should also be evaluated thoroughly, notably

through the help of native signers. Though those issues have recently attracted the attention of several research groups, a lot remain to be done before signing avatars can be used in our everyday environments.

References

1. O. Arikan, D. Forsyth, and J. O'Brien. Motion synthesis from annotations. *ACM Trans. on Graphics*, 22(3):402–408, July 2003.
2. C. Awad, N. Courty, K. Duarte, T. Le Naour, and S. Gibet. A combined semantic and motion capture database for real-time sign language synthesis. In *Proc of IVA*, volume 5773 of *LNAI*, pages 432–38. Springer-Verlag, Berlin, Heidelberg, 2009.
3. D. Bertram, J. Kuffner, R. Dillmann, and T. Asfour. An integrated approach to inverse kinematics and path planning for redundant manipulators. In *ICRA 2006: Int. Conf. on Robotic and Automation*, pages 1874–1879, 2006.
4. Justine Cassell, Joseph Sullivan, Scott Prevost, and Elizabeth F. Churchill. *Embodied Conversational Agents*. The MIT Press, 2000.
5. Y.H. Chiu, C.H. Wu, H.Y. Su, and C.J. Cheng. Joint optimization of word alignment and epenthesis generation for chinese to taiwanese sign synthesis. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(1):28–39, January 2007.
6. Z. Deng, P.i-Y. Chiang, P. Fox, and U. Newmann. Animating blendshape faces by cross-mapping motion capture data. In *Proc. of the 2006 symp. on Interactive 3D graphics and games*, pages 43–48, Redwood City, California, March 2006.
7. R. Elliott, J. Glauert, V. Jennings, and J. Kennaway. An overview of the sigml notation and sigml signing software system. In *Workshop on the Representation and Processing of Signed Languages, 4th Int'l Conf. on Language Resources and Evaluation*, 2004.
8. M. Filhol, A. Braffort, and L. Bolot. Signing avatar: Say hello to elsi!,. In *Proc. of Gesture Workshop 2007*, LNCS, Lisbon, Portugal, June 2007.
9. S. Fotinea, E. Efthimiou, G. Caridakis, and K. Karpouzis. A knowledge-based sign synthesis architecture. *Universal Access in the Information Society*, 6(4):405–418, 2008.
10. S. Gibet, T. Lebourque, and P.F. Marteau. High level specification and animation of communicative gestures. *Journal of Visual Languages and Computing*, 12:657–687, 2001.
11. B. Hartmann, M. Mancini, and C. Pelachaud. Implementing expressive gesture synthesis for embodied conversational agents. *Gesture in Human-Computer Interaction and Simulation*, 3881:188–199, 2006.
12. C. Hecker, B. Raabe, R. Enslow, J. DeWeese, J. Maynard, and K. van Prooijen. Real-time motion retargeting to highly varied user-created morphologies. *ACM Trans. on Graphics*, 27(3):1–11, 2008.
13. A. Héloir, N. Courty, S. Gibet, and F. Multon. Temporal alignment of communicative gesture sequences. *Computer Animation and Virtual Worlds*, 17:347–357, July 2006.
14. A. Héloir and S. Gibet. A qualitative and quantitative characterisation of style in sign language gestures. In *Gesture in Human-Computer Interaction and Simulation, GW 2007, Lecture Notes in Artificial Intelligence, LNAI*, Lisboa, Portugal, 2009. Springer Verlag.
15. A. Héloir, M. Kipp, S. Gibet, and N. Courty. Evaluating data-driven style transformation for gesturing embodied agents. In *Intelligent Virtual Agent (IVA 2008)*, volume 5208 of *LNCS*, pages 215–222, Tokyo, Japan, September 2008.

16. E. Ho, T. Komura, and C.-L. Tai. Spatial relationship preserving character motion adaptation. *ACM Trans. on Graphics*, 29(4):1–8, 2010.
17. M. Huenerfauth, L. Zhao, E. Gu, and J. Allbeck. Evaluation of american sign language generation by native asl signers. *ACM Trans. Access. Comput.*, 1(1):1–27, 2008.
18. K. j. Choi and H. s. Ko. On-line motion retargetting. *Journal of Visualization and Computer Animation*, 11:223–235, 2000.
19. A. Kendon. *Tools, Language and Cognition*, chapter Human gesture, pages 43–62. Cambridge University Press, 1993.
20. J. R. Kennaway, J. R. W. Glauert, and I. Zwitserlood. Providing signed content on the internet by synthesized animation. *ACM Trans. Comput.-Hum. Interact.*, 14(3):15, 2007.
21. M. Kipp, M. Neff, K. Kipp, and I. Albrecht. Toward natural gesture synthesis: Evaluating gesture units in a data-driven approach. In *Intelligent Virtual Agents (IVA'07)*, pages 15–28, 2007.
22. S. Kita, I. van Gijn, and H. van der Hulst. Movement phase in signs and co-speech gestures, and their transcriptions by human coders. In *Proc. of the Int. Gesture Workshop*, volume 1371 of *LNCS*, pages 23–35. Springer-Verlag, London, 1997.
23. S. Kopp and I. Wachsmuth. Synthesizing multimodal utterances for conversational agents. *Journal Computer Animation and Virtual Worlds*, 15(1):39–52, 2004.
24. A. Kranstedt, S. Kopp, and I. Wachsmuth. MURML: A Multimodal Utterance Representation Markup Language for Conversational Agents. In *Proceedings of the AAMAS02 Workshop on Embodied Conversational Agents - let's specify and evaluate them*, Bologna, Italy, July 2002.
25. R. Kulpa, F. Multon, and B. Arnaldi. Morphology-independent representation of motions for interactive human-like animation. *Comput. Graph. Forum*, 24(3):343–352, 2005.
26. A. Majkowska, V. B. Zordan, and P. Faloutsos. Automatic splicing for hand and body animations. In *SCA '06: Proc. of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 309–316, 2006.
27. I. Marshall and E. Safar. Grammar development for sign language avatar-based synthesis. In *In Proc. of the 3rd Int. Conf. on Universal Access in Human-Computer Interaction (UAHCI 2005)*, 2005.
28. D. McNeill. *Hand and Mind - What Gestures Reveal about Thought*. The University of Chicago Press, Chicago, IL, 1992.
29. M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel. Gesture modeling and animation based on a probabilistic re-creation of speaker style. *ACM Transactions on Graphics*, 27(1):233–51, March 2008.
30. Han Noot and Zsófia Ruttkay. Variations in gesturing and speech by gestyle. *Int. J. Hum.-Comput. Stud.*, 62(2):211–229, 2005.
31. S. Prillwitz, R. Leven, H. Zienert, T. Hanke, and J. Henning. *Hamburg Notation System for Sign Languages - An Introductory Guide*. University of Hamburg Press, 1989.
32. J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Computer Graphics and Applications*, 27(3):21–31, 2007.
33. William C. Stokoe. *Semiotics and Human Sign Language*. Walter de Gruyter Inc., 1972.
34. S. Tak and H.-S. Ko. A physically-based motion retargeting filter. *ACM Tra. On Graphics*, 24(1):98–117, 2005.
35. D. Tolani, A. Goswami, and N. Badler. Real-time inverse kinematics techniques for anthropomorphic limbs. *Graphical Models*, 62(5):353–388, 2000.

36. H. Vilhalmsson, N. Cantelmo, J. Cassell, N.E. Chafai, M. Kipp, S. Kopp, M. Mancini, S. Marsella, A.N. Marshall, C. Pelachaud, Z. Ruttkay, K. Thorisson, H. van Welbergen, and R.J. van der Werf. The behavior markup language: Recent developments and challenges. In *IVA 2007*, 2007.
37. J. M. Wang, D. J. Fleet, and A. Hertzmann. A. multifactor gaussian process models for style-content separation. In *Proc. of int. conf. on Machine Learning (ICML)*, June 2007.
38. L. Zhang, M. C. Lin, D. Manocha, and J. Pan. A hybrid approach for simulating human motion in constrained environments. *Computer Animation and Virtual Worlds*, 21(3–4):137–149, 2010.