Designing a Theoretically Motivated Animation System for Environment Aware Virtual Humans

Antonio Origlia[1,2,*], Maria Di Maro[1,2]

[1]Dept. of Electrical Engineering and Information Technology, University of Naples Federico II
[2]URBAN/ECO Research Center, University of Naples Federico II

Abstract

As Virtual Humans are getting more and more explored, as a natural interface towards different services, the way these are used for Cultural Heritage applications also needs to be updated accordingly. This paper describes the new version of the animation control system for the Maya Virtual agent, which was previously used to implement Cultural Heritage presentations in virtual environments. The work-in-progress described here aims at obtaining a stronger animation system from the theoretical point of view, matching work that is being done from the dialogue management point of view. The design of this new animation system includes multiple techniques for animation generation and a priority mechanism for gestures categories to fall into, depending on the background theory that will be adopted.

Keywords

Virtual Humans, Animation system, Cultural Heritage Presentations

1. Introduction

Language is inherently grounded in reality, intertwined with the physical and social contexts in which it operates. One crucial aspect of this grounding is the incorporation of gestures into the communication system. Gestures, alongside spoken words, form a symbiotic relationship, collectively referencing and interacting with the world. In this context, semiotics plays a crucial role in understanding nonverbal communication. Through semiotics, gestures are seen as signs, representing ideas, emotions, or actions, which are interpreted by individuals or groups [1]. This perspective on a grounded language underscores the necessity of integrating gestures as a fundamental component of a communication system for Real-Time Interactive 3D (RTI3D). As specific machine-interpretable languages exist to specify with different degrees of precision the desired output even from machine learning models, techniques for animating characters within RTI3D engines should be considered, emphasising the importance of aligning animation systems with communication theories. The choice of the appropriate technique for animation systems should, therefore, reflect the theory, as it is linked to specific communicative functions. Moreover, the organisation of these modules into a complete system integrated with other systems dedicated to communication, primarily the speech control system, can help develop and test computational theories of natural communication. In general, all animation techniques

Workshop on Advanced Visual Interfaces and Interactions in Cultural Heritage (AVICH 2024), June 4th, 2024, co-located with the 17th ACM Conference on Advanced Visual Interfaces, Arenzano (Genoa), Italy.

[*]Corresponding author.

antonio.origlia@unina.it (A. Origlia); maria.dimaro2@unina.it (M. Di Maro)

animate a hierarchy of *bones* (skeletons), which in turn deform the 3D meshes they are linked to. **Parametric** animations are the oldest way to implement movement in characters. Whether controlled using manually set interpolated keyframes or frame-by-frame generated parameters to control skeletal meshes, parametric animations are a good choice to implement *precision* movements. These typically depend on the relationship of the gesture with the environment (like pointing or touching) but may also be used to implement specific hand shapes or movements in conjunction with arm dynamics. The use of Inverse Kinematics, consisting of computing the orientation of parent bones chains, allows to generate coherent movements and limb configurations to implement gestures. **Motion capture** has become increasingly accessible, offering advantages over parametric animations by accurately capturing the subtle nuances of natural movement. This has been a typical issue in the digital games industry, where artists would often struggle to create realistic characters. However, suspension of disbelief may help reduce people expectancy while Virtual Humans movement performance may, instead, conflict with their increasingly realistic appearance and with the expectations set by the entertainment industry as "Entertainment gaming experiences color players' interactions with other digital media, setting expectations for user experience and interactivity" [2]. Motion capture animations are better used to implement gestures that require a less strict implementation than parametric gestures and can be paired with small, randomly generated variations in speed or arms involved to increase naturalness. The large quantity of motion data made available by recent motion capture technology has also enabled the development of **machine learning** techniques using Deep Neural Networks to generate animation curves. The most notable example, in this sense, is provided by NVidia as part of its Omniverse framework as the audio2face[1] (lipsync and facial expressions) and audio2gesture[2] (body movement) tools. Both are based on neural networks generating animation curves from speech data which can be streamed to RTI3D engines.

2. The Maya animation system

The architecture presented in this work is based on a RTI3D engine, the Unreal Engine 5, and it is designed to interact with the FANTASIA [3, 4] plugin[3] for Embodied Social Interactive Agents. A FANTASIA Conversational AI follows these main principles: **Behaviour Trees** (BT) [5] are used to organise and prioritise dialogue moves; **Graph Databases** (i.e., Neo4j [6]) are used for knowledge representation and dialogue state tracking; **Probabilistic Graphical Models (PGM)**, implemented using the aGRuM library [7], are used for decision making; LLMs are used to verbalise the decisions taken by PGMs. While a FANTASIA Virtual Human (Maya) using pointing gestures was used in the Cultural Heritage presentations setting[8, 9], the system described in this paper generalises the principles used in the previous work. A FANTASIA Virtual Human is equipped with an Interaction Manager implemented as a BT. This generates high-level instructions as activated tasks that are then executed by lower-level modules, like the navigation system for moving around the environment or the FANTASIA components providing access to PGMs and Graph Databases. Once the high-level instructions

[1] https://www.nvidia.com/en-us/ai-data-science/audio2face/

[2] https://docs.omniverse.nvidia.com/extensions/latest/ext_audio2gesture.html

[3] https://github.com/antori82/FANTASIA

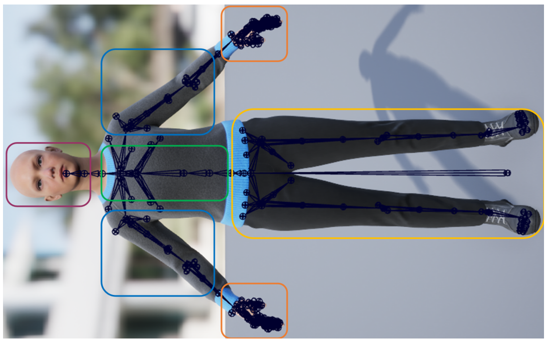

Figure 1: The animation system controlling different parts of the skeleton (body). **Lower body (yellow)**: the locomotion system controls this section, matching legs movement with character speed and orientation (warping). It uses parametrically altered mocap data; **Upper body (green)**: partly controlled by the locomotion system to match navigation speed and orientation. It mainly uses parametrically altered mocap data; **Left/Right Arm (blue)**: each arm has its own controller, implementing requests from the Interaction Manager using either parametric data or parametrically altered mocap data. Falls back to ML-generated data; **Left/Right Hand (orange)**: each hand has its own controller, implementing requests from the Interaction Manager using either parametric data or parametrically altered mocap data. Falls back to ML-generated data; **Head**: manages head orientation and combines it with facial expressions and head movements to match ML-generated lip-sync and nods.

are generated, the virtual representation of the character can implement the action by speaking, walking, and using gestures. Synchronising these activities and making them coherent with the virtual environment is primarily handled by the animation system, consisting of an *event graph*, receiving instructions from the Interaction Manager, and by an animation graph, selecting and combining animations (blending). It is always possible for the Interaction Manager to independently activate specific animations on specific parts or on the full skeleton, overriding, even temporarily, the animation system. The animation system consists of different animation controllers which results are combined to produce, at each frame, the appropriate pose of the character, summarised as shown in Figure 1. Figure 2 shows the steps used to create animations in the presented architecture. Early decisions are considered to be low priority, as they may be overwritten by later steps. Machine learning has low priority for body movements as it provides only a general estimate of coherent movements without actual communicative intentions. In presence of more strict necessities, like in the case of pointing or emphasis gestures, parametric or motion capture data can be used. A similar strategy is adopted for the lower body, concerning walking animations. After computing the final pose, given the animation system setup, the Interaction Manager can override parts of the pose by activating slot animations.

3. Pointing gestures: an example

The pointing system used in [8, 9] has been reimplemented as part of this generalised architecture. Differences from the previous system, which had an ad-hoc design for the specific experiments it was needed for, consist in two main points. First of all, the previous system relied on a fixed environment so that semantically annotated concepts could be assumed. The new

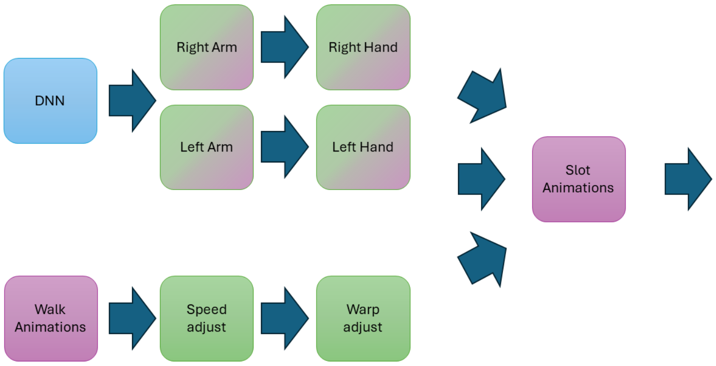

Figure 2: The animation system workflow. Moving from left to right, animations are generated using either machine learning (cyan) or motion capture (purple) data. Then, parametric or motion capture data can be used to entirely or partially overwrite movement on specific parts of the skeleton. After computing the final pose, this can be further modified using slot animations.

system provides a general interface for FANTASIA characters to query the environment and obtain available semantic labels on 3D models [10]. As in the previous system, the Interaction Manager of a FANTASIA character will consider the centroid of vertex points that are relevant for the pronounced concept to obtain the parameters needed to orient the appropriate arm and to request the index-pointing hand shape on the coherent hand. Secondly, the previous system made use of concatenative Text-to-Speech (TTS) technology, for which better support to the Speech Synthesis Markup Language (SSML) was available. To generate pointing gestures, labelled items were marked according to the SSML syntax (*mark* tag) so that a speech synthesizer would be able to provide the time offset at which each labelled item in the text was actually pronounced by the synthetic voice. Neural TTS, at present time, does not yet support the *mark* tag so the produced speech is now passed through a forced alignment module producing timing information *ex-post*. Differently from the previous architecture, where only pointing gestures could be produced by the avatar, these gestures are now framed within the priority system described in the previous Section. An example of the pointing gesture is shown in Figure 3.

4. Conclusions

We have presented a work-in-progress for the improved animation system of the Maya Virtual Human, designed to support visits in semantically annotated virtual environments. While previous system iterations supported pointing gestures only, the system is now generalised to support a computational theory of multimodal communication. Both the modules composing this system and the way they are organised leverage on modern methods for combining multiple sources of motion data. Their organisation also defines a priority mechanism for gestures, which can be used to organise a decision process coherent with models coming from humanities research. This topic is currently in course of investigation and will further develop the proposed

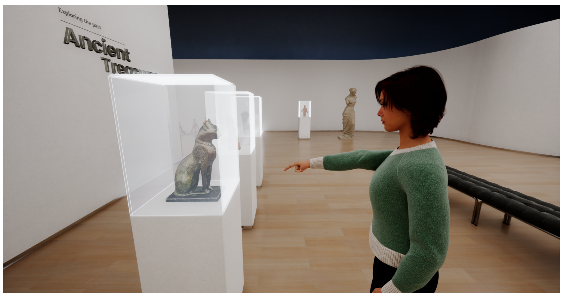

Figure 3: A pointing gesture: the Virtual Human Maya pointing at a target object of interest.

animation system to be fully integrated with the FANTASIA approach to dialogue management.

References

[1] C. K. Ogden, I. A. Richards, The meaning of meaning, New York & London: Harcourt Brace Jovanovich (1923).

[2] K. Squire, H. Jenkins, W. Holland, H. Miller, O. Alice, K. Philip Tan, K. Todd, Design principles of next-generation digital gaming for education, Educational Technology (2003) 17–23.

[3] A. Origlia, F. Cutugno, A. Rodà, P. Cosi, C. Zmarich, Fantasia: a framework for advanced natural tools and applications in social, interactive approaches, Multimedia Tools and Applications 78 (2019) 13613–13648.

[4] A. Origlia, M. Di Bratto, M. Di Maro, S. Mennella, A multi-source graph representation of the movie domain for recommendation dialogues analysis, in: Proceedings of the Thirteenth Language Resources and Evaluation Conference, 2022, pp. 1297–1306.

[5] G. Flórez-Puga, M. A. Gomez-Martin, P. P. Gomez-Martin, B. Díaz-Agudo, P. A. Gonzalez-Calero, Query-enabled behavior trees, IEEE Transactions on Computational Intelligence and AI in Games 1 (2009) 298–308.

[6] J. Webber, A programmatic introduction to neo4j, in: Proceedings of the 3rd annual conference on Systems, programming, and applications: software for humanity, 2012, pp. 217–218.

[7] G. Ducamp, C. Gonzales, P.-H. Wuillemin, aGrUM/pyAgrum : a Toolbox to Build Models and Algorithms for Probabilistic Graphical Models in Python, in: 10th International Conference on Probabilistic Graphical Models, volume 138 of *Proceedings of Machine Learning Research*, Skørping, Denmark, 2020, pp. 609–612. URL: https://hal.archives-ouvertes.fr/hal-03135721.

[8] A. Origlia, M. Grazioso, M. L. Chiacchio, F. Cutugno, 3d avatars and semantic models annotations for introductory cultural heritage presentations., in: AVI^{2}CH, 2022.

[9] A. Origlia, M. L. Chiacchio, M. Grazioso, F. Cutugno, Increasing visitors attention with introductory portal technology to complex cultural sites, International Journal of Human-Computer Studies 180 (2023) 103135.

[10] V. Cera, A. Origlia, F. Cutugno, M. Campi, et al., Semantically annotated 3d material

supporting the design of natural user interfaces for architectural heritage., in: AVI* CH, 2018.