Aesop: A Visual Storytelling Platform for Conversational AI

Tim Meo, Aswin Raghavan, David A. Salter, Alex Tozzo, Amir Tamrakar and Mohamed R. Amer
SRI International
201 Washington Rd
Princeton, NJ08540, USA
firstName.lastName@sri.com

Abstract

We present a new collaborative visual storytelling platform, Aesop, for direction and animation. Aesop consists of a language parser, human gesture monitoring, composition graphs, dialogue state manager, and an interactive 3D animation software. Aesop thus enables 3D spatial and temporal reasoning which are both essential for storytelling. Our key innovation is to enable conversational AI using both verbal and non-verbal communication, which enables research in language, vision, and planning.

1 Introduction

Storytelling, a vital part of both ancient and modern cultures, critically relies on the ability to convey a specific narrative (a complex task in communication). We define narrative as the description of a sequence of events, similar to illustrated books for children, with the intent of informing, educating, or entertaining. Aesop, our storytelling platform, will enable research on communication with Artificial Intelligence with the goal of collaboratively visualizing such a narrative.

Currently, the interaction between humans and machines focuses on handling directives given by the human. We believe that the future of AI will be a mixed-initiative collaboration with human partners. An AI agent's ability to communicate a goal to a human collaborator, and operate as an equal in a process, opens the door for many potential applications which lie outside current focuses of the AI community.

The goal of Aesop is to enable communication with machines to understand and relate both parties' perception of the world. This is a well-established problem in AI - specifically, the intersection of computer vision and natural language processing. Seamless communication with AI is critical in the performance of complex and creative tasks; in a mixed-initiative system, a human and a machine operate as a team, in which both collaborators perform the tasks at which they respectively excel. The more effectively the collaborators are able to identify and communicate about these tasks, the more efficiently the task can be accomplished.

Aesop will help address open research problems such as: conversational AI; joint language and gesture reasoning; planning and dialogue management; and narrative.

2 Related Work

SHRDLU [Winograd, 1972] attempted to ground language in a physical world. It consisted of hard coding of rules, physical and linguistic, and lacked 3D representation, focusing instead on symbolic abstraction. Wubble World [Hewlett *et al.*, 2007], which was introduced to advance the research initiated by SHRDLU, enabled learning in a physical world by interactive game-play with a simplistic 3D environment that evolved with time. More recently, [Perera *et al.*, 2017; Kim *et al.*, 2018] presented a blocks world platform for building structures. Their system combines natural language understanding, planning, and dialogue management. It supports communication about structures where the goals are shared between the computer and the human using natural language. The aforementioned platforms, however, are all limited to simplistic and rigid environments.

New platforms, such as the Minecraft-based Project Malmo [Johnson *et al.*, 2016], focus on end-to-end learning by solving various tasks in the 3D environments. These tasks range from navigation to collaborative problem solving using language. Similarly, the Quake III Arena-based DeepMind Lab [Beattie *et al.*, 2016] focuses on a maze navigation task and has been extended to incorporate language and learning, via an end-to-end approach which combines reinforcement and unsupervised learning [Hermann *et al.*, 2017]. While these platforms are more advanced than [Hewlett *et al.*, 2007] and [Perera *et al.*, 2017; Kim *et al.*, 2018], they are limited by the rigidity caused by boxy, inexpressive worlds and narrowly defined tasks assigned to the agents, offering little in the way of interaction with human users.

3 System Architecture

We describe our system illustrated in Fig. 1. Aesop consists of: a language parser, which takes input text and outputs parses; a gesture tracking module, which takes inputs depth and color information and outputs deictic coordinates; a composition graph module, which takes input parses and outputs a spatio-temporal event graph; a dialogue manager, which takes inputs actions, tracks their execution, adds them to the graph, and outputs communicative actions and generated language; various movie making agents, which take input composition graphs and output API calls; and finally, the Muvizu animation software, which takes inputs API calls and deictic coordinates, and outputs animations.

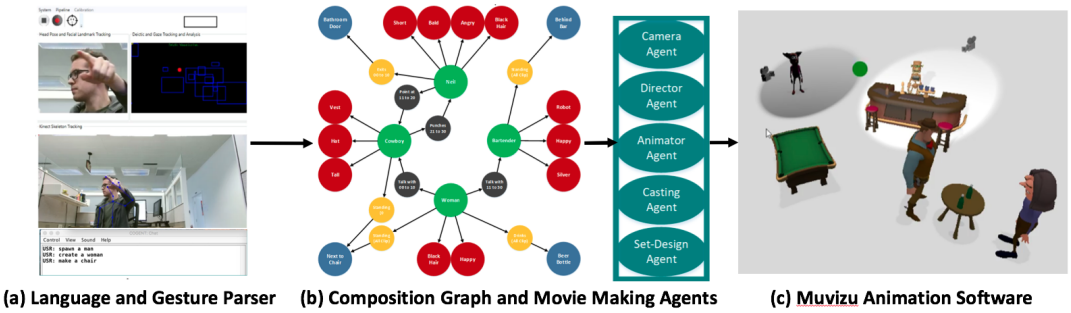

Figure 1: Block diagram of our system (Best viewed in color).

Language and Gesture Parser: Aesop uses a symbolic language parser called TRIPS [Allen *et al.*, 2008], which generates a semantic representation structured around events. TRIPS uses a general language-level ontology, augmented with domain-specific knowledge about the visual domain. The output "logical form" is a directed acyclic graph of ontology tokens representing entities, events, and their spatial, temporal, and lexico-semantic relationships. The logical form is used to determine an appropriate problem-solving act that represents a common goal between the human and computer. Nonverbal behaviors, meanwhile, are tracked by our own Multimodal Behavior Analytics system, which utilizes a Microsoft Kinect as its sole sensor. This sensor provides a high resolution RGB video stream, a 3D depth video stream of the person, and a high quality audio stream via an onboard microphone. A fully-articulated 3D skeleton of the person is tracked via the sensor's 3D-depth data. This skeleton is used by our gesture recognition analytics to detect hundreds of unique gestures made by the individual. Figure 1(a) illustrates our language and gestures components.

Composition Graphs: these graphs are used for reasoning about spatio-temporal and object-attribute relationships within the animation. The composition graph serves as an intermediate representation between natural language and the visual domain. In Fig. 1(b), actors (green nodes) are associated with their attributes (red nodes), spatial relationships with props (blue nodes), and temporal relationships such as interactions with other actors (gray nodes) and actions (yellow nodes). Aesop implements the composition graph as a probabilistic graphical model using the *Edward* [Tran *et al.*, 2016] library. Users can incrementally build the graph using natural language. Aesop extracts relationships from the output of the language parser and adds, removes or modify nodes in the graph with appropriate prior probability distributions. Parameters of the underlying probabilistic model to be learned from the MovieGraphs dataset [Vicol *et al.*, 2015].

Dialog Manager: given a graph, Aesop instantiates a possible animation corresponding to the state of the 3D engine in Muvizu [Muvizu, 2018], using movie-making agents which issue environmental and communicative acts. Our agents are managed using Markov Decision Process (MDP) scheduling the different actions given to each agent, as illustrated in Fig. 1(b). We use the notion of a collaborative state to represent the exchange between the user and the system, as they each express their own goals. Changes in collaborative problem state occur via Collaborative Problem Solving Acts [Ferguson and Allen, 2007] which refer to state-change acts applied to goals, or domain-specific actions. Maintaining these states as a graph enables the system to return to the relevant plan or action once subproblems in the collaboration process have been resolved. Aesop can spawn actors, change their attributes and locations, animate various motions with emotional modifiers, manipulate multiple cameras and capture video. When the graph consists of missing information, Aesop then decides whether to prompt the user for a clarification versus sampling from the probabilistic model initiating a communicative act.

Animation Software: we utilize Muvizu [Muvizu, 2018] as our animation software for visual story telling. Muvizu is designed to feel like an actual movie production pipeline, cutting out many of the barriers associated with the creation of an animated film. Muvizu provides a large library of built-in assets to rapidly assemble a scene, and enables users to select characters, customize their appearance, position cameras and lights around the set to prepare for the scene's key shots, issue directions to the actors, and record on-screen action in real time. Muvizu's internal library of pre-generated assets includes 80 characters and 600 props (with the ability to import own 3D object file), 1000 character accessories, 6 types of lights, 900 pre-animated character actions with mood-based modifiers (pointing angrily vs. with fright), and 19 visual effects for cameras. Shots are layered with visual and audio effects, voice tracks, and music, and finally exported as a video file. A user can direct character eye and head movement, and automatically lip-sync characters with audio tracks. Fig. 1(c) illustrates an animation created in Muvizu. All of the program's internal assets are directly exposed via an API, enabling interaction with external applications.

4 Conclusion

We presented Aesop, a novel visual storytelling system, for directing and animating. It exploits multimodal communication with the computer. It enables research on conversational AI using verbal and non-verbal communication, dialogue management, and gesture and language understanding.

Acknowledgements

This work is funded by DARPA W911NF-15-C-0246. The views, opinions, and/or conclusions contained in this paper are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied of the DARPA or the DoD.

References

[Allen *et al.*, 2008] James F. Allen, Mary Swift, and Will de Beaumont. Deep semantic analysis of text. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP 08. Association for Computational Linguistics, 2008.

[Beattie *et al.*, 2016] Charles Beattie, Joel Z. Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, Julian Schrittwieser, Keith Anderson, Sarah York, Max Cant, Adam Cain, Adrian Bolton, Stephen Gaffney, Helen King, Demis Hassabis, Shane Legg, and Stig Petersen. Deepmind lab. *CoRR*, abs/1612.03801, 2016.

[Ferguson and Allen, 2007] George Ferguson and James Allen. Mixed-initiative systems for collaborative problem solving. *AI magazine*, 28(2):23, 2007.

[Hermann *et al.*, 2017] Karl Moritz Hermann, Felix Hill, Simon Green, Fumin Wang, Ryan Faulkner, Hubert Soyer, David Szepesvari, Wojciech Marian Czarnecki, Max Jaderberg, Denis Teplyashin, Marcus Wainwright, Chris Apps, Demis Hassabis, and Phil Blunsom. Grounded language learning in a simulated 3d world. *CoRR*, abs/1706.06551, 2017.

[Hewlett *et al.*, 2007] Daniel Hewlett, Shane Hoversten, Wesley Kerr, Paul R Cohen, and Yu-Han Chang. Wubble world. In *AIIDE*, pages 20–24, 2007.

[Johnson *et al.*, 2016] M. Johnson, K. Hofmann, T. Hutton, and D. Bignell. The malmo platform for artificial intelligence experimentation. In *International Joint Conference on Artificial Intelligence*, 2016.

[Kim *et al.*, 2018] Sujeong Kim, David Salter, Luke DeLuccia, Kilho Son, Mohamed R. Amer, and Amir Tamrakar. Smilee: Symmetric multi-modal interactions with language-gesture enabled (ai) embodiment. In *North American Association for Computational Linguists*, 2018.

[Muvizu, 2018] Muvizu. Muvizu animation software. `http://www.muvizu.com/`, 2018.

[Perera *et al.*, 2017] Ian E. Perera, James F. Allen, Lucian Galescu, Choh Man Teng, Mark H. Burstein, Scott E. Friedman, David D. McDonald, and Jeffrey M. Rye. Natural language dialogue for building and learning models and structures. In *AAAI*, 2017.

[Tran *et al.*, 2016] Dustin Tran, Alp Kucukelbir, Adji B. Dieng, Maja Rudolph, Dawen Liang, and David M. Blei. Edward: a library for probabilistic modeling, inference, and criticism. *arXiv preprint arXiv:1610.09787*, 2016.

[Vicol *et al.*, 2015] Paul Vicol, Makarand Tapaswi, Lluis Castrejon, and Sanja Fidler. Moviegraphs: Towards understanding human-centric situations from videos. *CoRR*, abs/1712.06761, 2015.

[Winograd, 1972] Terry Winograd. Understanding natural language. *Cognitive Psychology*, 3(1):1 – 191, 1972.