Augmenting EMBR Virtual Human Animation System with MPEG-4 Controls for Producing ASL Facial Expressions

Matt Huenerfauth
Rochester Institute of Technology (RIT)
Golisano College of Computing and Information Sciences
20 Lomb Memorial Drive, Rochester, NY 14623
matt.huenerfauth@rit.edu

Hernisa Kacorri
The Graduate Center, CUNY
Computer Science Ph.D. Program
365 Fifth Ave, New York, NY 10016
hkacorri@gc.cuny.edu

1. Motivations

Our laboratory is investigating technology for automating the synthesis of animations of American Sign Language (ASL) that are linguistically accurate and support comprehension of information content. A major goal of this research is to make it easier for companies or organizations to add ASL content to websites and media. Currently, website owners must generally use videos of humans if they wish to provide ASL content, but videos are expensive to update when information must be modified. Further, the message cannot be generated automatically based on a user-query, which is needed for some applications. Having the ability to generate animations semi-automatically, from a script representation of sign-language sentence glosses, could increase information accessibility for many people who are deaf by making it more likely that sign language content would be provided online. Further, synthesis technology is an important final step in producing animations from the output of sign language machine translation systems, e.g. [1].

Synthesis software must make many choices when converting a plan for an ASL sentence into a final animation, including details of speed, timing, and transitional movements between signs. Specifically, in recent work, our laboratory has investigated the synthesis of syntactic ASL facial expressions, which co-occur with the signs performed on the hands. These types of facial expressions are used to convey whether a sentence: is a question, is negated in meaning, has a topic phrase at the beginning, etc. In fact, linguists have described how a sequence of signs performed on the hands can have different meanings, depending on the syntactic facial expression that performed [8]. For instance, an ASL sentence like "MARY VISIT PARIS" (English: Mary is visiting Paris.) can be negated in meaning with the addition of a Negation facial expression during the final verb phrase. As another example, it can be converted into a Yes/No question (English: Is Mary visiting Paris?) with the performance of a Yes-No-Question facial expression during the sentence.

The timing, intensity, and other variations in the performance of ASL facial expressions depend upon the length of the phrase when it co-occurs (the sequence of signs), the location of particular words during the sentence (e.g., intensity of Negation facial expression peaks during the sign NOT), and other factors [8]. Thus, it is insufficient for a synthesis system to merely play a fixed facial recording during all sentences of a particular syntactic type. So, we are studying methods for planning the timing and intensity of facial expressions, based upon the specific words in the sentence. As surveyed in [7], several SLTAT community researchers have conducted research on facial expression synthesis, e.g., interrogative questions with co-occurrence of affect [11], using clustering techniques to produce facial expressions during specific words [10], the use of motion-capture data for face animation [2], among others.

2. Features

To study facial expressions for animation synthesis, we needed an animation platform with a specific set of features:

1. The platform should provide a user-interface for specifying the movements of the character so that new signs and facial expressions can be constructed by fluent ASL signers on our research team. These animations become part of our project's lexicon, enabling us to produce example sentences so that our animations can be tested in studies with ASL signers.
2. The virtual human platform must include the ability to specify animations of hand, arm, and body movements (ideally, with inverse kinematics and timing controls), so that we can rapidly produce those elements of the animation.
3. The platform must provide sufficiently detailed face controls so that we can create subtle variations in the face and head pose, to enable us to experiment with variations in the movement and timing of elements of the face.
4. The platform should allow the face to be controlled using a standard parameterization technique so that we can use face movement data from human signers to animate the character.

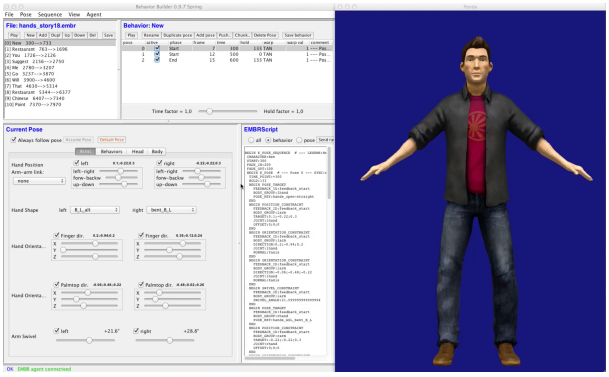

Fig. 1: EMBR user-interface for controlling virtual human

The open-source EMBR animation platform [3] already supported features 1 and 2, listed above. To provide features 3 and 4, we selected the character "Max" from this platform, and we enhanced the system with a set of face and head movement controls, following the MPEG-4 Facial Action Parameter standard [5]. Specifically, we added controls for the nose, eyes, and eyebrows of the character, which are portions of the face that are used extensively in syntactic facial expressions in ASL. The MPEG-4 standard was chosen because it is a well-defined face control formalism (thereby making any models we investigate of face animation more readily applicable to other virtual human animation research), and there are various software libraries available, e.g., [9], for automatically analyzing human face movements in a video, to produce a stream of MPEG-4 parameter values representing the detailed movements of the face over time.

Matt Huenerfauth, Hernisa Kacorri. 2015. Augmenting EMBR Virtual Human Animation System with MPEG-4 Controls for Producing ASL Facial Expressions. The Fifth International Workshop on Sign Language Translation and Avatar Technology (SLTAT'15), Paris, France, April 9-10, 2015.

Specifically, our laboratory implemented the following:

- We added facial morphs to the system for each MPEG-4 facial action parameter: Each of these parameters specifies vertical or horizontal displacements of landmark points on the human face, normalized by the facial proportions of the individual person's face. Thus, the morph controls for the Max character's face had to be calibrated to ensure that they numerically followed the MPEG-4 standard.
- Prior researchers have described how wrinkles that form on the forehead of a virtual human are essential to the perception of eyebrow raising in ASL animations [11]. So, we increased the granularity of the wireframe mesh of the character's face where natural wrinkles should appear, and wrinkle formation was incorporated into the facial morphs.
- To aid in the perception of wrinkles and face movements, a lighting scheme was designed for the character (see Fig. 2).
- Our laboratory implemented software to adapt MPEG-4 recordings of a human face movement to EMBRscript, the script language supported by the EMBR platform. In this way, our laboratory can directly drive the movements of our virtual human from recordings of human ASL signers. The MPEG-4 face recordings can be produced by a variety of commercial or research face-tracking software; our laboratory has used the Visage Face Tracker [9].

a.  b. 

Fig 2: (a) forehead with eyebrows raised before the addition of MPEG-4 controls, facial mesh with wrinkling, and lighting enhancements, (b) eyebrows raised in our current system

At the workshop, we will demonstrate our system for constructing facial expressions using MPEG4 controls in EMBR; we will also show animation examples synthesized by the system.

3. Science

The primary goal of our implementation work has been to support our scientific agenda: to investigate models of the timing and intensity of syntactic facial expressions in ASL. As part of this work, it will be necessary for us to periodically conduct user-based studies with native ASL signers evaluating the quality of animations that we have synthesized.

As an initial test of our ability to synthesize animations of ASL with facial expressions using this new animation platform, we conducted a pilot test with 18 native ASL signers who viewed animations that were generated by our new system: full details appear in [6]. The animations displayed in the study consisted of short stories with Yes-No Question, WH-Question, and Negation facial expressions, based upon stimuli that we released to the research community in [4]. The participants answered scalar-response questions about the animation quality and comprehension questions about their information content.

In this pilot study, the participants saw animations that were driven by the recording of a human; we previously released this MPEG-4 data recording of a human ASL signer performing syntactic facial expressions in [4]. The hand movements were synthesized based on our project's animation dictionary, which native signers in the lab have been constructing using the EMBR user-interface tool. Because the data-driven animations contained facial expressions and head movement, they utilized the skin-wrinkling, lighting design, and MPEG-4 controls of our new animation system. As compared to animations without facial expression shown as a lower baseline, participants reported that they noticed the facial expressions in the data-driven animations, and their comprehension scores were higher [6].

While this pilot study was just an initial test of the system, these results suggested that our laboratory will be able to use this augmented animation system for evaluating our on-going research on designing new methods for automatically synthesizing syntactic facial expressions of ASL. In future work, we intend to produce models for synthesizing facial expressions, instead of simply replaying human recordings.

4. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under awards 1506786, 1462280, & 1065009. We thank Andy Cocksey and Alexis Heloir for their assistance.

5. REFERENCES

[1] Ebling, S., Way, A., Volk, M., Naskar, S.K. (2011). Combining Semantic and Syntactic Generalization in Example-Based Machine Translation. In: Mikel L. Forcada, Heidi Depraetere, Vincent Vandeghinste (eds.), Proceedings of the 15th Conference of the European Association for Machine Translation, Leuven, Belgium, p. 209-216.

[2] Gibet, S., Courty, N., Duarte, K., Naour, T.L. 2011. The SignCom system for data-driven animation of interactive virtual signers: methodology and evaluation. *ACM Transactions on Interactive Intelligent Sys (TiiS)*, 1(1), 6.

[3] Heloir. A, Nguyen, Q., and Kipp, M. 2011. Signing Avatars: a Feasibility Study. *2nd Int'l Workshop on Sign Language Translation and Avatar Technology (SLTAT)*.

[4] Huenerfauth, M., Kacorri, H. 2014. Release of experimental stimuli and questions for evaluating facial expressions in animations of American Sign Language. *Workshop on the Representation & Processing of Signed Languages, LREC'14.*

[5] ISO/IECIS14496-2Visual, 1999.

[6] Kacorri, H., Huenerfauth, M. 2015. Comparison of Finite-Repertoire and Data-Driven Facial Expressions for Sign Language Avatars. *Universal Access in Human-Computer Interaction*. Lecture Notes in Computer Science. Switzerland: Springer International Publishing.

[7] Kacorri, H. 2015. TR-2015001: A Survey and Critique of Facial Expression Synthesis in Sign Language Animation. Computer Science Technical Reports. Paper 403.

[8] Neidle, C., D. Kegl, D. MacLaughlin, B. Bahan, and R.G. Lee. 2000. *The syntax of ASL: functional categories and hierarchical structure*. Cambridge: MIT Press.

[9] Pejsa, T., and Pandzic, I. S. 2009. Architecture of an animation system for human characters. In. *10th Int'l Conf on Telecommunications (ConTEL)* (pp. 171-176). IEEE.

[10] Schmidt, C., Koller, O., Ney, H., Hoyoux, T., and Piater, J. 2013. Enhancing Gloss-Based Corpora with Facial Features Using Active Appearance Models. *3rd Int'l Symposium on Sign Language Translation and Avatar Technology (SLTAT).*

[11] Wolfe, R., Cook, P., McDonald, J. C., and Schnepp, J. 2011. Linguistics as structure in computer animation: Toward a more effective synthesis of brow motion in American Sign Language. *Sign Language & Linguistics*, 14(1), 179-199.