Implementation and Evaluation of Animation Controls Sufficient for Conveying ASL Facial Expressions

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

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

ABSTRACT

Technology to automatically synthesize linguistically accurate and natural-looking animations of American Sign Language (ASL) from an easy-to-update script would make it easier to add ASL content to websites and media, thereby increasing information accessibility for many people who are deaf. We are investigating the synthesis of ASL facial expressions, which are grammatically required and essential to the meaning of sentences. To support this research, we have enhanced a virtual human character with face controls following the MPEG-4 Facial Action Parameter standard. In a user-study, we determined that these controls were sufficient for conveying understandable animations of facial expressions.

Categories and Subject Descriptors

H.5.2 [**Information Interfaces and Presentation**] User Interfaces – *evaluation/methodology*; K.4.2 [**Computers and Society**]: Social Issues – *assistive technologies for persons with disabilities*.

General Terms

Design, Experimentation, Human Factors, Measurement.

Keywords

Accessibility Technology for People who are Deaf, MPEG-4, Facial Expression, American Sign Language, Animation.

1. INTRODUCTION

There are approximately 500,000 users of ASL in the U.S. [4], and it is possible for users to have fluency in ASL but difficulty with written English because the two languages are distinct. Many signers prefer to receive information in the form of ASL. One simple method of presenting ASL content online would be to display video recordings of human signers on websites, but this approach is not ideal: the recordings are difficult to update, and there is no way to support just-in-time generation of content. Software is needed that can automatically synthesize understandable animations of a virtual human performing ASL, based on an easy-to-update script as input. This software must select the details of the movements of the virtual human character so that the animations are understandable and acceptable to users. Facial expressions and head movements are essential to the fluent performance of ASL, conveying: emotion, variations in word meaning, and grammatical information during entire syntactic phrases. This paper focuses on this third use, which is necessary for expressing questions or negation. In fact, a sequence of signs performed on the hands can have different meanings, depending on the syntactic facial expression that co-occurs [5]. E.g., a declarative sentence (ASL: "JOHN LIKE PIZZA" / English: "John likes pizza.") can become a Yes-No question (English: "Does John like pizza?"), with the addition of a Yes-No Question facial expression (eyebrows raised, head tilted forward). Similarly, the addition of a Negation facial expression (left and right headshaking with some brow furrowing) during the verb phrase "LIKE PIZZA" can change the meaning of the sentence to "John doesn't like pizza." The word NOT is optional, but the facial expression is required. For interrogative questions (with a "WH" word like what, who, where), a WH-Question facial expression (head tilted forward, eyebrows furrowed) is required during the sentence, e.g., "JOHN LIKE WHAT."

There is variation in how these facial expressions are performed during a given sentence, based on the length of the phrase when the facial expression occurs, the location of particular words during the phrase (e.g., NOT, WHAT), the facial expressions that precede or follow, the overall speed of signing, and other factors. Thus, for an animation synthesis system, it is insufficient to simply play a single pre-recorded version of this facial expression whenever it is needed. For this reason, we are researching how to model the performance of facial expressions in various contexts. Other researchers are also studying synthesis of facial expressions for sign language animation, e.g., interrogative questions with co-occurrence of affect [8], using clustering techniques to produce facial expressions during specific words [7], etc.

2. IMPLEMENTATION & EVALUATION

To support this research, we had to parameterize the face of our virtual human character so that we can control it by specifying a vector of numbers. Then, a full performance is a stream of such vectors. We needed a parameterization with some properties:

- Values should be invariant across signers with different face proportions who are performing an identical facial expression so we could use recordings from multiple humans in our work.
- The parameterization must be sufficient for controlling the face of a character and should be invariant across animated characters with different facial proportions. This property would allow us to use a variety of characters in our work.
- The parameterization should be a well-documented, standard method of producing and analyzing facial movements. This property would enable our research to be useful for other researchers, using other animation platforms.

The MPEG-4 standard [3] defines a 3D model-based coding for face animation and has all the above properties. In short, a face is

ASSETS '14, Oct 20-22 2014, Rochester, NY, USA
ACM 978-1-4503-2720-6/14/10.
http://dx.doi.org/10.1145/2661334.2661387

controlled by setting values for 68 Facial Action Parameters (FAPs), which are displacements of points shown in Fig. 1a with the displacements normalized according to scaling factors based on the proportions of the character's face. This normalization allows for a set of 68 FAPs to produce equivalent facial expression on faces of different sizes or proportions.

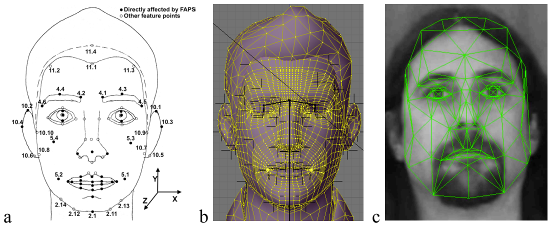

Figure 1: (a) Some MPEG-4 feature points, (b) wireframe and feature points in Max, (c) Visage tracker adaptive mask.

Our lab extended the character named Max from the open source animation platform EMBR [1] (Fig. 1b) with MPEG-4 FAPs for the upper face controlling the eyes, eyebrows, and nose. EMBR allows for head and torso movements, enables blinking as a background behavior, and has been used for creating sign language animations. As part of our enhancements to EMBR, a professional artist modified the surface mesh and constraints to cause the skin on the face to wrinkle automatically as the face controls are modified. The artist also assisted in the design of a lighting scheme for the character to highlight these wrinkles, which are essential to perception of ASL facial movements [8].

We conducted a user study, where 14 native ASL signers viewed animations of short stories and then answered comprehension questions and scalar-response questions as to whether they noticed the correct facial expression. The 18 stories included Yes-No Question, WH Question, or Negation (6 of each type), and the comprehension questions were engineered so that the correct answer depended on understanding the facial expression. We publically released these stimuli and evaluation questions for evaluating facial expression animations; details appear in [2]. In a between-subjects design, we compared two types of animations with identical hand movements but differed in their face, head, and torso movements: (a) ***driven*** by a recording of a human performing that type of facial expression or (b) face, head, and torso movements are static and ***neutral*** throughout the story. The type "b" animations therefore did not reveal any of the capabilities of the new MPEG-4 controls or skin-wrinkling of our character. Face and head movements for the ***driven*** animations were created using Visage Face Tracker, automatic software [6] that provides MPEG-4 compatible output. Fig. 1c illustrates the 3D mask in the tracking system that is fitted to a native signer's face. We implemented software to convert MPEG-4 data to EMBRscript, the script language supported by the EMBR platform. Example shown in Fig. 2 and at: http://latlab.cs.qc.cuny.edu/2014assets/.

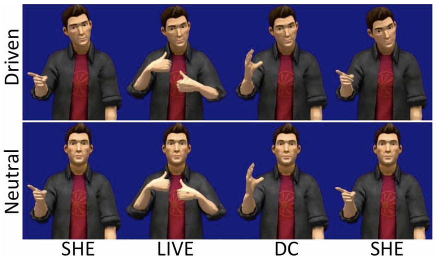

Figure 2: Screenshots from a human-recording-driven and neutral version of a Yes-No Question stimulus in the study.

Fig. 3 displays the scores of the comprehension questions and the question that asked if participants noticed the correct facial expression. Medians are shown above each boxplot. There was a significant difference in the Notice scores (Mann-Whitney test used since the data was not normally distributed, p<0.00014). There was also a significant difference in the comprehension question scores (t-test, p<0.000001). Note that comprehension scores depend on the difficulty of the questions asked; so, such scores are meaningful only for comparison within a single study.

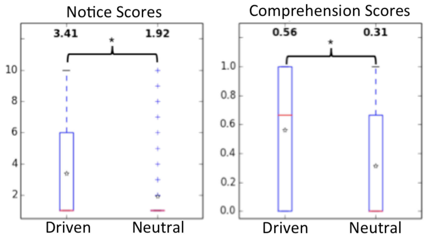

Figure 3: Notice and Comprehension scores for animations with facial expressions (Driven) and without (Neutral).

These results indicate that our new animation system is a useful platform for evaluating our on-going research on designing new methods for automatically synthesizing facial expressions of ASL. This finding is significant because it allows for research on ASL facial expression to take advantage of prior tools and research on facial animation with MPEG-4. In order to evaluate the expressivity of our character, we used human recordings in this study; however, in future work, we will be investigating learning-based models for *automatic* synthesis of ASL facial expressions.

3. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under award number 0746556 and 1065009. We thank Andy Cocksey and Alexis Heloir for their assistance.

4. REFERENCES

[1] 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).*

[2] 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.*

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

[4] Mitchell, R., Young, T., Bachleda, B., and Karchmer, M. 2006. How many people use ASL in the United States? Why estimates need updating. *Sign Lang Studies*, 6(3):306-335.

[5] 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.

[6] 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.

[7] 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).*

[8] 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.