Learning to Generate Understandable Animations of American Sign Language

Matt Huenerfauth
matt.huenerfauth@rit.edu

Department of Computer Science, Queens College
Computer Science and Linguistics, The Graduate Center
The City University of New York (CUNY)
New York, NY, USA

Department of Information Sciences and Technologies
Golisano College of Computing & Information Sciences
Rochester Institute of Technology (RIT)
Rochester, NY, USA

Abstract—Standardized testing has revealed that many deaf adults in the U.S. have lower levels of written English literacy; providing American Sign Language (ASL) on websites can make information and services more accessible. Unfortunately, video recordings of human signers are difficult to update when information changes, and there is no way to support just-in-time generation of web content from a query. Software is needed that can automatically synthesize understandable animations of a virtual human performing ASL, based on an easy-to-update script as input. The challenge is for this software to select the details of such animations so that they are linguistically accurate, understandable, and acceptable to users. Our research seeks models that underlie the accurate and natural movements of virtual human characters performing ASL, using the following methodology: experimental evaluation studies with native ASL signers, motion-capture data collection from signers, linguistic analysis of this data, statistical modeling techniques, and animation synthesis.

Keywords—American Sign Language; animation; accessibility; technology for people who are deaf or hard-of-hearing

I. INTRODUCTION

This paper provides an overview of our research on automatically synthesizing animations of American Sign Language (ASL). In particular, it is meant to serve as an introduction and survey of some of the major publications produced by our laboratory during the years 2006 to 2014.

Section II explains the accessibility motivations for conducting research in this area, and Section III describes the major goals and applications of our work. Section IV highlights several major technical accomplishments of our research, with a focus on identifying the key contribution in each area. This section provides several citations to additional publications from our laboratory where the reader can find more substantial technical background on the work that is described briefly in this overview paper. Finally, Section V identifies several new avenues of future research.

II. MOTIVATIONS

Unable to hear spoken language during the critical language-acquisition years of childhood, only half of deaf high school graduates in the United States have a fourth-grade reading level in English [34]. However, many of these adults have sophisticated fluency in ASL, which is a primary means of communication for one-half million people in the U.S. [29]. ASL is not a direct transliteration of English; it is a distinct natural language, with its own unique word-order, grammar, and vocabulary. Many people are fluent in ASL but have only limited fluency in English. In fact, ASL contains phenomena without a direct parallel in written/spoken languages: During a conversation, an ASL signer can set up various locations around them in space to represent people, things, or concepts under discussion [19, 27, 28]. To refer to these items again in a conversation, the signer will point or aim their torso, eyes, or head at these locations. Given these literacy and linguistic issues, software/websites must present information in the form of ASL to make it accessible to many people who are deaf.

A. Why Can’t We Simply Use Videos of Human Signers?

While it is possible to post videos of real human signers on websites, animated avatars are better if the information content is often updated or is synthesized on demand (e.g., result of a website search or database query) [26]. It is prohibitively expensive to continually re-film a human performing ASL whenever information must be updated, and it is impossible to use video if information is synthesized on demand. Because ASL signers use the space around their body to set up the items they are discussing, it is not possible to “cut and paste” videos of different sentences together to produce an understandable result. Assembling video clips of individual signs together to produce sentences leads to poor results, due to discontinuities in the video blending and because of the need to modify the movements of signs based on the surrounding words. Websites with videos would be likely to update their English text information more frequently than their ASL videos, leading to out-of-date information for people who are deaf.
B. Why Automatically Synthesize ASL from Sparse Input?

One way to produce animations of ASL would be for a skilled animator (fluent in ASL) to create a virtual human that moves in the correct manner using general-purpose 3D animation software. Unfortunately, this approach is too time-consuming (with an approximately 1:100 ratio between length of the video produced and the animator time needed), and this approach depends too much on the skill of the 3D animator. In studies at our lab with ASL signers evaluating animations of ASL, we have determined that the understandability of ASL animations depends on subtle use of speed/timing, particular motion paths/orientations of the hands, and other complex movement constraints. It is too difficult for a human animator to specify all of the body joint angles for a virtual human to produce high quality ASL animation. For this reason, we conduct research on methods to synthesize animations of ASL from sparse input specifications.

III. RESEARCH GOALS AND APPLICATIONS

A. Goals of Research on ASL Animation Synthesis

Given a high-level plan for the sentence to be produced (e.g., that the ASL sentence should contain a given list of words), the goal of our research program is to design intelligent software that will automatically and accurately:

- Select which items under discussion should be set up in locations in space around the signer and where in the 3D space they should be positioned (to match typical locations that human ASL signers would select).
- Select how the motion path of verbs must change based on how locations have been set-up in space for the subject and the object of the verb. The motion paths and hand orientation of many ASL verbs undergo complex modifications to indicate the 3D location in space of the verb’s subject and object [1, 19, 32].
- Select how the motion paths and handshapes of ASL signs blend together when one sign follows another – or other types of linguistic “coarticulation” effects.
- Select linguistically accurate velocities/accelerations of the hands and appropriate duration of pauses during signing, based on linguistic factors.
- Select the most appropriate location for the eye gaze of the signer, which is governed by linguistic rules, discourse-level factors, and other constraints.
- Select accurate movements of the facial muscles and head motion/orientation to perform facial expressions during ASL signs, which communicate linguistically essential information about the meaning of sentences. The same sequence of movements on the hands can have widely different meanings, depending on the facial expressions, which can indicate negation, questions, conditional-clauses, topic-phrases, etc. [30]
- Provide a lexicon of ASL signs that can be used to rapidly synthesize sentences or longer passages. Such a resource is needed for scripting technologies (below).

B. Applications of This Technology

There are multiple applications for software that can automate the synthesis of animations of ASL [10]. For example, there are several immediate and long-term uses:

- To add ASL information to a website such that it can be updated easily or dynamically synthesized on-demand, the website designer can encode a “script” for the ASL sentences in the website, which is synthesized into animations as needed. If the author of this script were responsible for determining all of the movements for every sign or making all of the “selections” listed above, it would be too difficult a task. Our software could synthesize an animation, given a “sparse” script of the ASL sentence that is needed. There are several other research groups internationally who are investigating technologies for scripting sign language animations, e.g., [3, 35].
- Many researchers are investigating automatic written-language-to-sign-language machine translation technology, e.g., [2, 4, 33]. While this is a very difficult problem (the state-of-the-art is still limited), our ASL synthesis software would be a necessary final step in the pipeline of any translation system. Our software would convert a symbolic specification of a sign language sentence into a full animation to be displayed.

IV. TECHNICAL CONTRIBUTIONS IN ASL ANIMATION SYNTHESIS AND EVALUATION

A. HCI Experimental Research to Measure the Understandability of ASL Animations

We have conducted iterative design and evaluation of linguistic and assistive technologies for over a decade, including hundreds of hours of studies with ASL signers to evaluate ASL animation technology [8, 14, 15, 18, 21]. It has been essential for our research that there are many ASL signers and members of the deaf community who are members of our research team. This has enabled our research team to build relationships with the local Deaf community when recruiting research subjects, to synthesize ASL linguistics research when designing ASL software, and to collaborate with researchers who are deaf or hard-of-hearing.

From 2008 to 2013, our laboratory at CUNY organized a summer research program for high school students who are deaf or hard-of-hearing in New York City; the program encouraged students to pursue higher education and research careers in the sciences while also forming stronger ties between the lab and the local community.

During experiments with signers, we have quantified how variations and enhancements to ASL animations result in benefits for people who are deaf. Specifically, we have found that the understandability of ASL animations is affected by: modulations in the speed [8], insertion of pauses at linguistically appropriate locations [7], the use of space around the signer to represent entities under discussion [14], and the “inflection” of ASL verbs [13, 14].
These experiments guide our lab’s research on animation-synthesis technologies: allowing us to prioritize what aspects should receive attention and allowing us to evaluate the quality of ASL animations that result from our computational linguistic improvements. Evaluation studies of animations synthesized by our ASL animation software have demonstrated that our lab has significantly advanced the state-of-the-art for synthesizing animations of ASL verb signs [13, 23, 24], speed and timing relationships in ASL [14], and accurate facial expressions [16, 17, 18].

B. HCI Methodological Research on How to Best Evaluate ASL Animations

Prior research on ASL animation technology has lacked rigorous methodological research to understand how experimental design choices affect the outcome of studies, and our lab has provided guidance for this maturing field. We have designed several new experimental methodologies for user-based studies with ASL signers interacting with linguistic technology. Our lab has measured the effect of showing different baselines for comparison during evaluation studies – thereby enabling, for the first time, comparison of the results of studies that had used different experiment designs [18].

We have also designed new recruitment and screening protocols to effectively identify participants with specific levels of ASL skill [15], linguistic stimuli for use in studies evaluating ASL technology [6, 17], and new question-types and modalities for studies with ASL signers [8, 11, 17, 18]. In recent work, we have studied the use of eye-tracking technology with ASL signers to measure where they look on ASL videos or animations; we have learned how to automatically detect the quality of ASL animations based on eye metrics [16].

C. NLP and Animation Research on Automatically Synthesizing ASL Animations

Our lab has designed new algorithms for selecting movement details for ASL animations, given a script with sparse information about the sentence. Our lab has identified ideal speed/timing for ASL animations [7, 8], placement of pauses during sentences [7, 8], planning of complex sentences called “classifier predicates” [6, 9, 15], temporal coordination during animations [5], and the linguistically accurate selection of the motion-path and hand orientation of ASL verbs to indicate locations in 3D space around the signer where entities under discussion have been established [21, 23, 24]. This final work on ASL verbs has been data-driven, making use of statistical models trained from examples of human movement collected in our ASL motion-capture corpus (see below).

D. Collecting and Annotating a Corpus of ASL Multi-Sentence Passages using Motion-Capture Equipment

In NSF-funded research, our lab has created the first corpus of ASL containing motion-capture data of multi-sentence passages and linguistic annotation [12, 20, 22, 25, 26], which enables new data-driven techniques for producing ASL animations, e.g., [24]. In addition to the corpus creation itself, we have published about our custom configuration of motion-capture gloves, eye-tracker, head-motion-tracker, and bodysuit, and we have developed novel calibration protocols for our equipment. We have developed recruitment protocols, annotation manuals, and other necessary methodologies for corpora creation and annotation. We have also investigated various prompting strategies to encourage signers to perform multi-sentence passages with specific linguistic properties that we want to collect [12, 22, 25], thereby avoiding using scripts, which can lead to unnatural linguistic results.

Fig. 2. Images of the motion-capture equipment used to collect the corpus of American Sign Language at our laboratory: (a) H6 head-mounted eye-tracker from Applied Science Laboratories, (b) Interense IS-900 sensor used to record head position and orientation, (c) Immersion 22-sensor Cyberglove used to record handshape, (d) IS-900 overhead ultrasonic speaker array that is used in conjunction with the sensor shown in (b).
Fig. 3. Images from videos recorded during data collection, showing front view, face close-up view, and side view. The motion-capture body suit worn by signers is visible in the image, as well as other equipment shown in Fig. 2.

Fig. 4. Screenshot of tool used to annotate the corpus, developed by [30].

Fig. 5. Examples of facial expressions generated for our animated signer.

At our laboratory, our linguistic researchers and graduate students are assisted by high school and undergraduate student researchers who are ASL signers. They “annotate” the data we collect – marking the time-span of individual signs and other linguistic phenomena in the ASL recordings. In addition to creating a valuable research experience for students who are deaf or hard-of-hearing, the data we are gathering benefits our ASL animation research – and is useful for other researchers studying ASL linguistics or recognition of ASL from video.

E. Collaborative Research on Synthesizing Facial Expressions for ASL Animations

Facial expressions convey essential grammatical information about ASL sentences, and they can dramatically affect the meaning of sentences (e.g., changing sentences into questions or negating the meaning of a verb phrase). In NSF-funded collaborative research with computer vision scientists at Rutgers University and ASL linguists at Boston University, our laboratory is designing new statistical models that govern the movements of a signer’s face to accurately produce understandable facial expressions.

The team at Boston has linguistically annotated a corpus of videos with many facial expressions, and the Rutgers team is tracking facial landmarks in the video [31].

Our team is designing models that link the linguistic phenomena with specific facial movements to produce animations, and we are building an infrastructure for synthesizing animations using MPEG-4 facial animation parameters. This is a challenging task due to the necessary temporal coordination of the face, synchronization of these movements with words in the sentence, and blending of one facial expression with another (simultaneously or sequentially). Our lab is conducting periodic evaluations with native ASL signers to evaluate animations with facial expressions synthesized based on our computational linguistic models [16, 17, 18].

V. Future Research Agenda

After the relocation of our laboratory to the Rochester Institute of Technology in the summer of 2014, we intend to continue our research on automatic ASL animation synthesis along the following research avenues:

- In addition to facial expressions that communicate syntactic information about phrases and sentences, which are the focus of our current work, we would like to investigate two additional types of common facial expressions in ASL: lexically-specific facial expressions that must co-occur with specific words and affective/emotional facial expressions that convey the signer’s mood/tone. We plan to research this wider variety of facial expressions. The challenge is that these additional types of facial expressions interact with the syntactic facial expressions in complex ways, and thus may require more sophisticated modeling.

- Given our lab’s recent release of the large motion-capture corpus of ASL, there are many unexplored avenues of research in which this data resource could be mined to train statistical models of a variety of ASL linguistic phenomena – which could, in turn, be used to synthesize high-quality ASL animations. While we have already begun to successfully model some types of verb modifications, we would like to study the speed/pausing of signers in the corpus, their selection of points in space for items under discussion, their use of torso movement (based on linguistic structures), and their eye-gaze directions. Prior to the existence of our
corpus, none of these topics could be easily investigated in an empirical manner for ASL animation research, and we are excited to see how data-driven approaches can advance the field.

- Given the ASL technologies our lab has developed, we want to investigate how they can be combined into a usable and effective ASL animation scripting infrastructure to allow users to compose, edit, and update ASL sentences for providing information on websites. Such software could revolutionize how signers create presentations and messages in ASL, and it could have a substantial impact on the way in which students learning ASL practice and develop their skills.

In addition to this research on ASL animation synthesis, our laboratory is interested in investigating issues that relate the accessibility of captioning technologies for people who are deaf and hard-of-hearing and the creation of tools to benefit students who are learning American Sign Language.

ACKNOWLEDGMENT
Our team consists of many researchers, current and former students, and visiting scholars who have contributed to the research surveyed in this paper, including: Hernisa Kacorri, Pengfei Lu, Allen Harper, Jonathan Lambert, Miriam Morrow, Molly Sachs, Meredith Turtletaub, Priscilla Diaz, Jennifer Marfino, Fang Zhou Yang, Wesley Clarke, Amanda Krieger, Evans Seraphin, Christine Singh, Kaushik Pillapakkam, Fatimah Mohammed, Giovanni Moriarty, Kenya Bryant, Raymond Ramirez, Kelsey Gallagher, Aaron Pagan, and Jaime Penzella. In addition, we thank our collaborators at other universities, including Carol Neidle at Boston University, Dimitris Metaxas at Rutgers University, YingLi Tian at the City College of New York, and Elaine Gale at Hunter College.

REFERENCES


