Synthesizing American Sign Language Spatially Inflected Verbs from Motion-Capture Data

Pengfei Lu
The City University of New York (CUNY)
CUNY Graduate Center
Doctoral Program in Computer Science
365 Fifth Ave, New York, NY 10016
+1-212-817-8190
pengfei.lu@qc.cuny.edu

Matt Huenerfauth
The City University of New York (CUNY)
Queens College and Graduate Center
Computer Science and Linguistics
65-30 Kissena Blvd, Flushing, NY 11367
+1-718-997-3264
matt@cs.qc.cuny.edu

ABSTRACT
People who are deaf or hard-of-hearing who have lower levels of written-language literacy can benefit from computer-synthesized animations of sign language, which present information in a more accessible form. This paper introduces a novel method for modeling and synthesizing American Sign Language (ASL) animations based on motion-capture data collected from native signers. This technique allows for the synthesis of animations of verb signs whose performance is affected by the arrangement of locations in 3D space that represent entities under discussion. Mathematical models of hand location were trained on motion-capture recordings of a human producing inflected verb signs. In an evaluation study with 12 native signers, the ASL animations synthesized from the model were judged to be of similar quality to animations produced by a human animator. This animation technique is applicable to other ASL signs and other sign languages used internationally – to increase the repertoire of sign language animation generation systems or to partially automate the work of humans using sign language animation scripting tools.

Categories and Subject Descriptors
I.2.7 [**Artificial Intelligence**]: Natural Language Processing – *language generation, machine translation*; K.4.2 [**Computers and Society**]: Social Issues – *assistive technologies for persons with disabilities*.

General Terms
Design, Experimentation, Human Factors, Measurement.

Keywords
American Sign Language, Accessibility Technology for People who are Deaf, Animation, Natural Language Generation.

1. INTRODUCTION
Animations of a virtual human character performing sign language can increase the accessibility of information for many people who are deaf or hard-of-hearing, many of whom may have difficulty reading complex written-language texts. This paper focuses on American Sign Language (ASL) and producing accessible sign language animations for people who are deaf in the U.S., but many of the linguistic issues, literacy rates, and animation technologies discussed within are also applicable to other sign languages used internationally. ASL is a natural language used as a primary means of communication for over one-half million people in the U.S. [17]. ASL and English are distinct languages, with different syntax, word order, and vocabularies (without word-for-word translations), and thus it is possible to be fluent in ASL yet have significant difficulty reading English texts. In fact, due to various educational and language-exposure factors, a majority of deaf high school graduates (student age 18 or older) in the U.S. have a fourth-grade (age 10) English reading level or below [23]. This lower rate of written-language literacy poses key accessibility challenges for deaf adults who must obtain information from English text on computers, video captions, or other sources. Technologies for automatically generating computer animations of ASL can make information and services accessible to deaf people with lower English literacy [6]. While videos of sign language are feasible to produce in some contexts, animated avatars are more advantageous than video when the information content is often modified, the content is generated or translated automatically, or signers scripting a message in ASL wish to preserve anonymity.

Section 2 of this paper discusses several complex ways in which signs vary in how they are performed based on the context of the sentence in which they are used. Section 3 surveys the current state of the art in sign language animation technologies in regard to the production of inflecting verb signs. Section 4 summarizes how, in prior research, we introduced a novel technique for automatically synthesizing animations of such verb signs [9]. In this paper, we apply this modeling technique to motion-capture data collected from human ASL signers, and we evaluate whether the resulting animations are understandable and accurate. Section 5 describes our method, in which we collect multiple examples of the performance of a sign using motion-capture equipment and then fit mathematical models to this data. Section 6 presents an evaluation study we conducted with 12 native ASL signers. Section 7 discusses related work on synthesizing animations of sign language verbs. Finally, section 8 discusses our conclusions and future work – our ultimate goal is to create a lexicon of ASL verbs that are parameterized on the 3D location of their subject and/or object (so that a specific verb performance can be synthesized as needed by ASL animation software).

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SLTAT 2011, 23 October 2011, Dundee, UK.
Copyright 2011 the Authors.

2. USE OF SPACE, INFLECTED VERBS

ASL signers often associate people, things, places, or concepts under discussion with arbitrary 3D locations in space around their bodies [12, 13, 15, 16]. After an entity is mentioned, a signer may point to a 3D location in space around his/her body. Later, to refer to this entity again, the signer (or his/her conversational partner) can point to this 3D location in order to refer to the entity. Various researchers have studied this pronominal use of space [11, 12, 15, 16], and some believe that signers tend to pick 3D locations on a semi-circular arc floating at chest height in front of their torso [15, 16]. Whether or not signers limit their selection of 3D locations to this arc around their body (or whether they pick 3D locations at different heights and distances from their body, as argued by [12]), there are an infinite number of locations in 3D space where entities may be associated for future pronominal reference.

The locations selected by the signer to associate with entities under discussion have an effect on how later signs in the conversation are performed. While ASL verbs have a standard, prototypical way of being performed, some verbs can be inflected to indicate the 3D location in space at which their subject and/or object have been associated [12, 18, 19]. When such verbs are performed in a sentence, their prototypical motion path may be deflected such that the movement or orientation goes from the 3D location of their subject and toward the 3D location of their object. The resulting performance is a synthesis of the verb's prototypical motion path and the locations associated with the subject and object in the signing space. Sometimes called "inflecting" [19], "indicating" [12], or "agreeing" [1] verbs, these verbs are the focus of our research and shall be referred to as "inflecting verbs" in this paper. Because the verb sign itself reflects the 3D locations in space of its subject and/or object, the names of the subject and object may not be otherwise expressed in the sentence. If the signer chooses to explicitly mention the subject and object of a verb, then it is legal for him to use the prototypical (uninflected) version of the verb, but the resulting sentences tend to appear less fluent (see study in [10]).

Fig. 1. Two inflected versions of the ASL verb BLAME: on the top row, the subject has been associated with a location on the left side and the object on the right; on the bottom row, the subject, on the right side, and the object, on the left.

Fig. 1 shows an example of an ASL verb BLAME that changes its performance based on the 3D locations in space around the signer where the verb's subject and object have been previously associated. While this example verb is affected by the location of both its subject and object, broadly, linguists divide ASL verbs into classes based on whether their motion is inflected based on: (1) subject only, (2) object only, (3) both subject & object, or (4) neither [12, 19]. Further, there are other categories of ASL verbs (e.g., "depicting," "locative," or "classifier" verbs) whose movements convey complex spatial information (e.g., a specific 3D motion path of an object or a manner of movement of a person being discussed); such verbs are not the focus of this paper.

3. SIGN LANGUAGE ANIMATION

In prior publications (e.g., [6]), we have surveyed recent technologies and research systems for synthesizing animations of sign language. In summary, there are two major groups of ASL computer animation research: scripting software (e.g., [3, 23]) or generation software (e.g., [4, 5, 14, 24]).

- Software for "scripting" sign language can be thought of as a "word processing" tool for creating animations of sign language. It allows a human who knows ASL to arrange signs on a timeline to produce animations of ASL sentences; this approach is more efficient than requiring the human to manually specify all the joints of a virtual human character's body. The software automates much of this process and synthesizes an animation of a virtual human from the sentence timeline created by the human. The scripting system makes linguistic and human-movement choices about how specific signs should appear when used in a particular sentence, how the human's body should move from the end of one sign until the beginning of the next, and various other detailed animation issues unspecified by the human user when creating a timeline of the sentence to be performed.
- The second major type of sign language animation system is called a "generation" system; such software plans a sign language sentence based on some source of information. A human user does not manually specify all of the signs to be performed by the system on a timeline; the generation software does this automatically. Various researchers have studied the "machine translation" of written-language sentences into sign language animations automatically; such research also falls into the category of sign language "generation" – in this case, the input information source used by the software is a text in a written language.

The way in which ASL signers assign entities under discussion to locations in space around their bodies not only affects where they point in space during later pronominal reference, but it also affects inflecting verbs' motion path, based on the 3D location assigned to their subject and/or object in the surrounding signing space. These linguistic aspects of ASL pose challenges for animation systems; it is insufficient to store a single version of each sign in the system's dictionary; for pointing signs, inflecting verb signs, and other space-influenced signs, the system would need to synthesize a specific instance of the sign based on how space was arranged. Because of the infinite ways in which locations could be assigned to entities under discussion, it is not possible to pre-store all the possible combinations of all the signs the system may need.

Because of the challenge in producing animations that use space for pronominal reference and correctly synthesize signs influenced by the spatial arrangement of entities, it is natural to consider whether we could make a simplification: Would animations of sign language that do not use spatial layout of entities or inflection of verbs be understandable to deaf users? To evaluate this question, in prior research, we conducted

experimental studies in which native ASL signers viewed animations of ASL stories that varied in whether or not entities were associated with locations in space and whether or not verbs were spatially inflected based on these locations [10]. We found that animations which lack this use of space for pronominal reference and which lack verb inflection are less understandable to deaf ASL signers, as measured by comprehension questions and Likert-scale subjective evaluation scores. Thus, producing animations with proper space use and verb inflection is important for producing easily understandable and useful ASL animations.

Unfortunately, current generation and machine translation systems for producing animations of sign language generally do not include the ability to produce inflected versions of verb signs. They also typically do not make extensive use of spatial locations to represent entities under discussion. (Thus, the output of these systems looks much like the animations without space use and without verb inflection that we evaluated in [10].) One of the most sophisticated uses of verb inflection in a generation system is the British Sign Language animation generator produced by Marshall and Safar [14], which could associate entities under discussion with a finite number of locations in the signing space (approximately 6 locations). Further, the system's repertoire included a few verbs whose subject/object were positioned at these locations. A limitation of this work is that most of the verbs handled by their system involved relatively simple motion paths for the hands from subject to object locations. Also, their system did not allow for the arrangement of pronominal reference points at arbitrary 3D locations in the signing space.

Section 7 discusses the ASL synthesis research of [21, 22] focused on verb inflection; this comparative discussion is reserved until after details of our methods are described in sections 5 and 6.

Sign language scripting technologies also currently do not include the capability of setting up arbitrary locations in the surrounding signing space around the character associated with entities under discussion. For instance, Sign Smith Studio, a commercially available sign language scripting system for ASL, allows users to ask the character to point to a location on the left or the right, but not infinitely many possible locations [24]. Further, the system contains a single uninflected version of most of the ASL verbs in its dictionary. While the company provides companion animation software that enables users to precisely pose the virtual human character to produce novel sign performances as needed, this is a time-consuming process. Carefully animating the movements of a virtual character for each inflected verb form would significantly slow down the process of scripting an ASL animation. For this reason, the users of most scripting software tend to use uninflected verb forms and limited use of space for pronominal reference, thereby producing animations of sign language that are less fluent and likely less understandable for deaf users.

4. OUR PRIOR MODELING RESEARCH

The goal of our research is to construct computational models of ASL verbs that can partially automate the work of human users of scripting software or be used within generation systems. We want to create a parameterized dictionary of ASL verb signs such that: Given the name of the verb, the location in space associated with of the verb's subject, and the location in space associated with the verb's object, our software should be able to produce a movement of the virtual human character that is a linguistically-accurate spatially-inflected instance of the verb. In prior work [9], we proposed a lexicon creation technique that could use samples of ASL verb signs produced by human animators, fit a mathematical model to the data samples, and then use this model to synthesize novel ASL verb sign instances (properly inflected for different locations of subject and object – including combinations that were not present in the training data used to build the model). This prior work is the foundation for the research presented in this paper, and so, our prior research methodology is summarized below.

Our technique is a data-driven approach that is based on the collection of samples of sign language performance from human signers. In [9], we describe how we use a commercially available sign language animation tool called VCom3D Gesture Builder [24] to gather samples of ASL verbs for our research. This software allows a human user who is knowledgeable of sign language to animate the movements of a virtual human character with an easy-to-use GUI to produce a novel ASL sign by dragging and moving the hands of a signer and arranging keyframes for the animation on a timeline. The new sign is saved as an XML file, and it can be imported into the VCom3D's Sign Smith Studio scripting tool (or used in other animation systems). Using the Gesture Builder tool, native ASL signers were asked to produce dozens of examples of a set of several ASL verbs (including: ASK, GIVE, TELL, SCOLD) for various combinations of subject and object locations in the surrounding signing space. For our research, we assume that signers tend to place entities in space on an arc-like area of space around their bodies, as shown in Fig. 2. The Gesture Builder software (and our animation approach) is keyframe-based, and it uses inverse kinematics and motion interpolation to synthesize a full animation from a list of hand location targets for specific keyframe times during the animation.

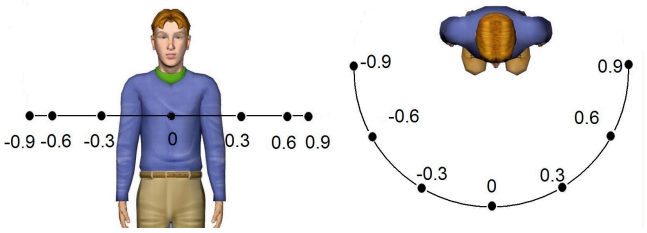

Fig. 2. Front & top view of arc positions around the signer.

For data collection, we identified seven equally-spaced locations on the arc, as shown in Fig. 2. We asked the signer to produce all possible combinations of each of the five verbs for these seven locations. Thus, for verbs like ASK and GIVE whose movement path is affected by the location of both their subject and object in space, the native signer produced 42 examples of each verb for all non-reflexive combinations of the seven arc positions. For verbs like TELL and SCOLD whose movement path is affected by their object location only (not affected by their subject location), the native signer produced seven examples of each verb (for all seven possible object locations around the arc). Table 1 lists the four ASL verbs studied in section 5 that were also studied in our prior research [9] and includes a brief description of each. While we focus on these four verbs as examples of our lexicon-building methodology, we intend for our work to be generalizable to other ASL verbs and other sign languages used internationally. The website of our laboratory includes example animations of each of these verbs we studied: http://latlab.cs.qc.cuny.edu/sltat2011/.

We extracted hand locations (x,y,z) from the XML files for each keyframe of each verb; each verb animation produced by the signer consisted of two keyframes. Thus, for a two-handed verb

(e.g., GIVE) that is affected by both subject and object positions, we collected 504 location values per verb: 42 examples x 2 keyframes x 2 hands x 3 (x,y,z) values.[1] To later synthesize a verb animation for a given arrangement of subject and object arc positions around the signer, we needed a mathematical model of each location parameter (x,y,z) for each hand for each keyframe for each verb. We modeled each parameter using a third-order polynomial as discussed in [9], and we set the coefficients of each model using a least-squares fit to the verb samples collected from the native signer using the Gesture Builder software. For verbs that inflect for object only, these functions were parameterized on object arc position only. For verbs that inflect for both subject and object, these functions were parameterized on both subject and object arc positions. For example, the model for the x value of the right hand for the first keyframe of the verb GIVE had the following form: for subject arc position s and object arc position o, the model contained terms up to s^3 and o^3 and all possible cross-product terms $s^a o^b$ where $a \leq 3$, $b \leq 3$, $a+b \leq 3$. After setting the coefficients, this function predicts the right hand's x location at the beginning of the verb GIVE, given the subject's and the object's arc positions around the signer.

Table 1: Four ASL Inflecting Verbs Examined in This Paper

Verb	*Inflection Type*	*Description*
ASK	Subject & Object	The signer moves an extended index finger from the "asker" (subject) to the "person being asked" (object). During the movement, the finger bends into a hooked shape. (ASL "1" to "X" handshape.)
GIVE	Subject & Object	In this two-handed version of the sign, the signer moves two hands as a pair from the "giver" (subject) toward the "recipient" (object). (Both hands have an ASL "flat-O" handshape.)
SCOLD	Object Only	The signer "wags" (bounces up and down while pointing) an extended index finger at the "person being scolded" (object). (ASL "1" handshape.)
TELL	Object Only	The signer moves an extended index finger from the mouth/chin toward the "person being told" (object). (ASL "1" handshape.)

Fig. 3. Keyframes 1 and 2 of GIVE produced by our model.

To synthesize a novel verb sign animation, we select the verb and specify the subject and object arc positions. Given the models we had fit on the verb sample data, this information is sufficient to produce an XML file representing the sign, which can be imported into VCom3D Sign Smith Studio. This software allows the user to script sentences of ASL and allows "custom" signs (e.g., inflected verbs produced by our model or by a human animator) to be imported into the sentence. Given the XML file, the software handles the keyframe interpolation and inverse kinematics to synthesize an animation of a human character. Fig. 3 shows keyframes of the verb "GIVE" for subject at arc position -0.6 and object at 0.3 – as synthesized by our model. After constructing models for a set of example ASL verbs, we conducted an evaluation study with native ASL signers [9].

[1] In [9], we also model hand orientation, but we focus only on hand location in this current paper. [9] also explains how we assume that the handshapes for a verb sign are consistent in how they're performed across different inflections for subject and object arc positions; thus, we do not model how handshape is affected by different subject and object arc positions around the signer (but this could be done in future work).

5. CURRENT WORK: MOTION CAPTURE

Our previous methodology (section 4) had been to collect samples of instances of ASL inflecting verbs (for a variety of subject and object locations) by asking a native ASL signer with animation experience to produce these verbs using the Gesture Builder sign creation software [24]. Then, we extracted hand position information from the keyframes for each verb, and this data was used to fit third-order polynomial models for hand position parameterized on the subject's and object's position on the arc. The problem with this approach was that it may be difficult for some signers to produce accurate and natural ASL signs using an animation tool; the way the signer actually moves when signing may look different than how they *think* they move when producing an ASL sign using an animation tool. If unrealistic sign examples collected in this manner are used as training data for our model, then the quality of the resulting animation may suffer. In other projects at our lab, we are collecting a large motion-capture corpus of ASL [8], and so, we were curious whether we could use motion-capture equipment to record ASL verb performances -– instead of asking a signer to create verb animations using Gesture Builder. After videotaping and motion-capture recording a signer performing a variety of ASL verb signs, we extract hand position information from keyframes in the motion-capture data stream, and we use this information as our source data for fitting polynomial models. Thus, the mathematical models and the way in which the models are used to synthesize an animation are identical to our prior work, what is novel is that we are now using motion-capture data as our source of verb examples. While using motion capture has the potential to yield more natural movement data, there are challenges of including motion capture in this verb lexicon-building approach. We now must identify timecodes in the motion-capture data stream that correspond to the beginning and ending keyframes of each verb recorded, we must clean up noise in the data, and we may find that the variation in movement of actual humans is wider than the variation we observed in the ASL verb signs produced by signers using the Gesture Builder tool – thereby making our modeling work more difficult.

A native ASL signer was recruited to perform the ASL verb signs while being recorded via motion-capture equipment. To avoid coarticulation effects from adjacent signs, the signer was asked to perform single examples of each inflected form of the verb for given locations in the surrounding signing space where the subject and object of that verb were associated. Fig. 4. shows how the laboratory was arranged during the data collection. Seven 10cm colored paper squares were attached to the walls of the laboratory in such a way that they corresponded to the angles of the seven points shown on the arc in Fig. 2. Specifically, the two squares visible in Fig. 4 correspond to arc positions 0.9 and 0.6 in Fig. 2. These squares served as "targets" for the signer to use as "subject" and "object" when performing various inflected verb forms; the use of color-coded squares was found to be less error-prone than the use of numbered or labeled target locations positioned around the room. Another native ASL signer sitting behind the video camera prompted the performer to produce each inflected verb

form by pointing to the colored squares for the subject and the object of the verb sample that we wanted to record (Fig. 5). At the beginning of the session, the signer was asked to make several large arm movements with some sudden direction changes to facilitate the later synchronization of the motion-capture stream with the video data (Fig. 6). Occasionally during the recording session (and whenever the signer made a mistake and needed to repeat a sign), the signer was asked to sign the sequence number of the verb example being recorded (Fig. 7); this facilitated later analysis of the video. An Intersense IS-900 motion-capture system with an overhead ultrasonic speaker array and hand, head, and torso mounted sensors with directional microphones and gyroscope were used to record location and orientation data of the hands, torso, and head of the signer during the study. After the recording session was completed, the motion-capture data values were transformed to match the coordinate system of the VCom3D animation system XML files. The various arm positions in Fig. 6 (and an arms-straight-up pose) were used to scale the data from the recorded human to the body size of the VCom3D avatar.

After the recording session, a native ASL signer viewed the video (Fig. 8) to identify the time index (video frame number) that corresponded to the start and end movement of each verb sign that we recorded. (If we had modeled signs with more complex motion paths, we might have needed more than two keyframes.) These time codes were used to extract hand location (x,y,z) data from the motion-capture stream for each hand for each keyframe for each verb example that was recorded. With this information, we were able to use the modeling technique summarized in section 4; we fit new polynomial models to this data. These new models could then be used to synthesize animations of ASL verb signs – the difference from our prior work is that these new models are based on actual human movements from our motion-capture session.

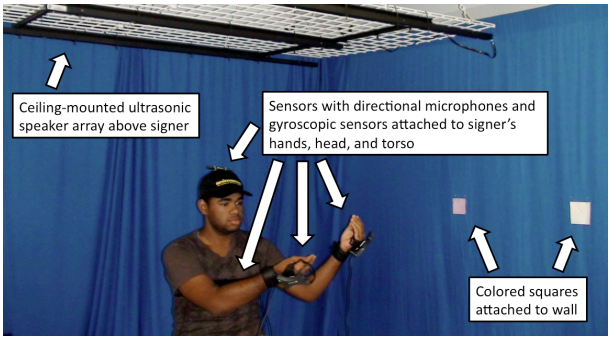

Fig. 4. This three-quarter view illustrates the layout of the laboratory during the motion-capture data collection; the signer is facing a camera (off-screen to the right). Sitting behind the camera is another signer conversing with him.

1.	GREEN	GIVE	YELLOW
2.	BLUE	GIVE	RED
3.	WHITE	GIVE	YELLOW
4.	"YOU"	GIVE	RED
5.	GREEN	GIVE	PURPLE

Fig. 5. The signer behind the camera prompted the recorded signer to perform each inflected verb by pointing to the appropriate colored squares on the walls; this is an excerpt of the verbs-to-collect list that the prompter viewed as a guide.

Fig. 6. These photographs extracted from the video recorded during the study demonstrate some of the arm movements the signer was asked to perform at the beginning of the session to facilitate later calibration of the collected motion-capture data.

Fig. 7. The photo on the left illustrates how the signer was asked to say the number that corresponded to the sentence being performed during the session; this facilitated later analysis of the video recording. The photo on the right shows a close-up view of the hand-mounted sensor used in the study.

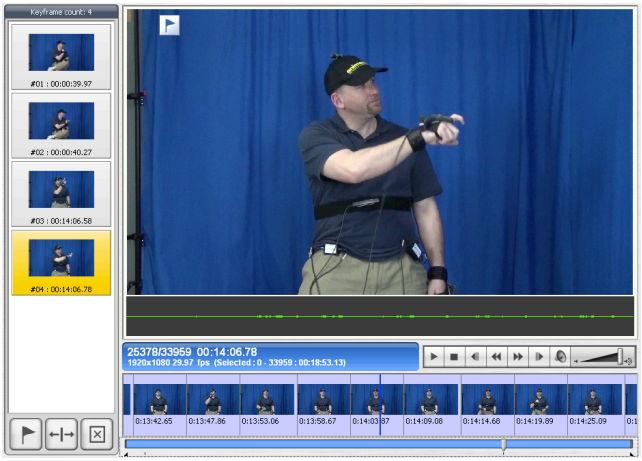

Fig. 8. This screen-capture demonstrates the use of the TMPGEnc video analysis software to identify the times of keyframes in the video that correspond to the beginning and ending hand positions for each verb sign sample collected.

6. EVALUATION OF CURRENT WORK

In order to determine whether our ASL verb lexicon-building approach worked well with the motion-capture training data, we conducted an evaluation study of the understandability and naturalness of animations synthesized using our new models. The overall methodology of this evaluation study, including the recruiting practices, genre of ASL stories used as stimuli, format of comprehension questions, and other details follows the general approach of our prior evaluation research reported in [9, 10]. In this study, 12 native ASL signers evaluated ASL animations of three types: (1) with inflected verbs synthesized using our new model (based on the motion-capture data), (2) with inflected verbs produced by a human animator (a native signer used the Gesture Builder program to produce each verb example), and (3) with

uninflected verbs (prototypical dictionary versions of each verb that do not reflect the subject/object locations in space). Electronic advertisements were used to recruit 12 participants for this study; a screening questionnaire [7] helped determine if potential participants were native signers. Of the 12 participants, 10 had used ASL since infancy, and 2 participants had learned ASL at ages 6 and 10 through attendance at a residential school with instruction in ASL. These final 2 participants had been using ASL for over 20 years, attended schools or university with classroom instruction in ASL, and used ASL on a daily basis to communicate with a significant other or family member. There were 8 men and 4 women of ages 21-47 (median age 32).

Sign Smith Studio and Gesture Builder [24] were used to create the ASL animations for this evaluation study, using hand location and orientation data based on our models or other sources. In this paper, we are currently focused only on the modeling of hand location from the motion-capture data; we intend to use the orientation data from our recording session in future work. However, in order to synthesize verb animations for an evaluation study, we needed some source of hand orientation data for our inflected verb forms. Thus, for this evaluation study, we used the hand orientation information from the Gesture Builder verb examples produced by a native signer and used as training data in our prior work [9]. Thus, the verb animations shown to participants in this evaluation study are produced from hand location data determined by our model trained on motion-capture and hand orientation values that were set by a human ASL signer.

In future applications of our research in which someone wants to synthesize animations of sign language from a model trained on verb examples, it would be best to use as much training data as possible when fitting the model coefficients. However, in an evaluation study, it would be better to be more rigorous: to make the model's task more difficult, we use a "leave one out" strategy: To produce an animation of an instance of a verb (e.g., "ASK" with subject at arc position 0.9 and object at arc position 0.3), we trained models using all instances of "ASK" except the instance being synthesized for the evaluation study stimuli (in this case, we omitted the example of "ASK" with subject at 0.9 and object at 0.3 from our training data, then trained the model coefficients, and then synthesized the example of this verb for use in the study).

The experiment consisted of two phases: In phase 1, participants viewed animations of short ASL stories and answered comprehension questions after viewing the animation only once. In phase 2, participants saw three side-by-side animations of the same ASL sentence – with three different versions of the verb used in each sentence: inflected, human animator, or uninflected.

In phase 1, we used a set of nine ASL stories and comprehension questions that we had originally produced as stimuli for [9]. The stories and questions were adapted for use in this study to exclude any ASL inflected verbs that were not the four listed in Table 1. The animation consisted of a single onscreen virtual human character who tells a story about 3-4 characters, who are associated with different arc positions in the signing space surrounding the virtual signer. The stories were an average of 55 signs in length, and the comprehension questions were difficult to answer because of the stories' complexity, because participants saw the stories before seeing the questions, and because they could only view the story once time. Each story was produced in three versions: (1) with inflected verbs from our model, (2) with inflected verbs produced by the human animator, and (3) with uninflected versions of the verbs. In this within-subjects study design: (1) no participant saw the same story twice, (2) order of presentation was randomized, and (3) each participant saw 3 animations of each version. After watching each story once, participants answered 4 multiple-choice comprehension questions that focused on information conveyed by the inflecting verbs. Further methodological details of similar studies we have conducted may be found in [7, 9, 10]. Fig. 9 shows the comprehension question accuracy scores with error bars indicating the standard error of the mean. An ANOVA was run to check for significant differences between comprehension question scores for each version of the animations; no significant pairwise differences were observed between the versions for any of the verbs individually – nor from the data from all verbs combined.

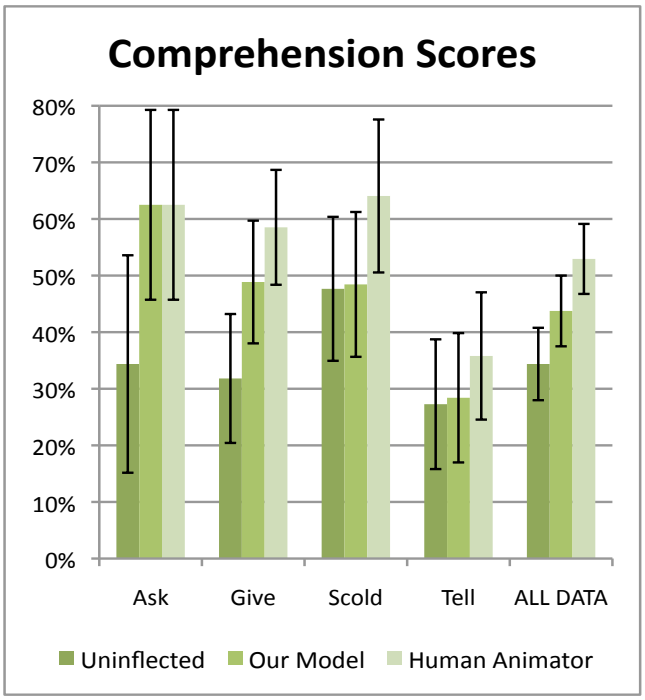

Fig. 9. Results of comprehension questions; no significant pairwise differences (ANOVA, alpha=0.05).

In phase 2, participants viewed three animations of the same sentence side-by-side; for example, we used the ASL sentence "John point-to-position-0.9 ASK Mary point-to-position-0.3." The only difference between the three versions was whether the verb was synthesized from our model, created by a human animator using Gesture Builder, or an uninflected version of the verb. Participants could re-play the animations multiple times, and a variety of arc positions were used in the animations (the three versions shown at one time all used the same arc positions). Participants answered 1-to-10 Likert-scale questions about the grammaticality, understandability, and naturalness of the verb in each of the 3 versions of the sentence. Fig. 10 shows the results. To check for significant differences between Likert-scale scores for each version, a Kruskal-Wallis test was performed (for scalar data that is not normally distributed, a non-parametric test is best); significant pairwise differences are marked with a star in Fig. 10.

In both the Likert-scale data and the comprehension-score data, our model tended to have performance between the inflected verb animation produced by the human animator using Gesture Builder and the uninflected version of the verb; this result suggests that our model was producing an ASL sign of better quality than

uninflected forms. For side-by-side comparison scores, both our model and the human animator's verb scored significantly higher than the uninflected verb animations. Because the human animator version of the verbs was considered our upper baseline for this study (since it reflects the careful creation of an inflected verb form during a time-consuming process), achieving scores that are similar to that of the human animator is a positive result.

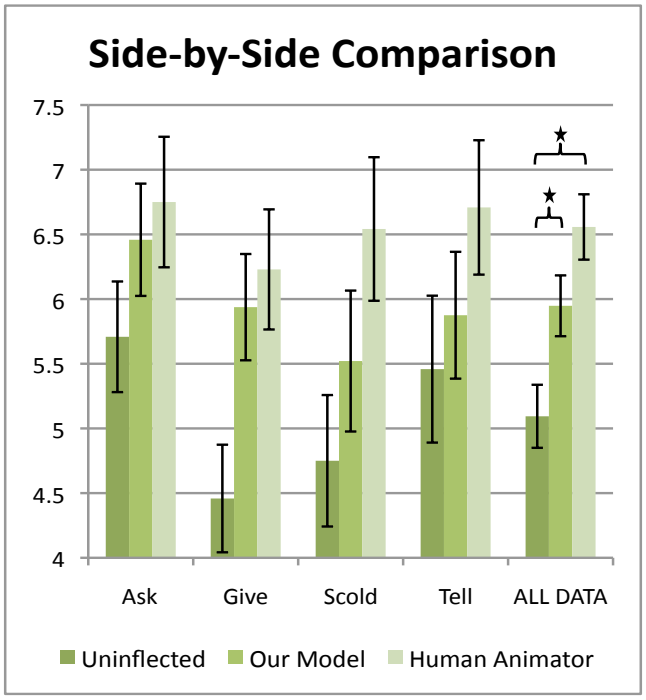

Fig. 10. Results of side-by-side comparison Likert-scale questions; significant pairwise differences marked with stars (Kruskall-Wallis test, alpha=0.05).

7. RELATED WORK

The most related work is [21, 22], who collected video examples of inflected verbs from human ASL signers and analyzed these videos to note the locations of the hands in the image for different verbs. Based on these observations, they invented algorithms for planning motion paths for the hands for each verb. Our research differs in several ways: Instead of analyzing videos to identify hand locations, we collect 3D coordinates of the hands using *motion-capture sensors* worn by the signer. Further, our signer performed several dozen inflected forms of each verb (that we specifically requested) based on colored *targets* arranged on an arc around the signer in the recording studio; the targets served as subject/object locations for verbs (details in [9]). Further, we *automatically* build mathematical models of hand movement for each verb by fitting our model's parameters to collected motion-capture data. Thus, with additional data collection for new verbs, new animation models can be created without verb-specific animation work. While [21, 22] asked signers to judge whether an animation of a verb they synthesized was a 1st-, 2nd-, or 3rd-person singular/plural form, we conducted an evaluation with native ASL signers in an environment minimizing English influences to determine whether they could understand the specific meaning of inflected verb forms used in an animated passage – measured via comprehension questions. In particular, the stories in our study included multiple entities simultaneously associated with several 3D locations around the animated signer; participants had to disambiguate which entity was the subject and object of each inflected verb in the animation in order to understand the meaning and answer the comprehension questions correctly. A further distinction is that [21, 22] modeled the incorporation of repetitive movements indicating variations in temporal aspect (repetition, regularity, etc.) for verbs with plural objects; we do not model this optional temporal aspect inflection in our work. Finally, [21, 22] also model how to synthesize forms of their verbs with subject and object positioned at different heights around the signer (possible in narrative contexts in which signers are emphasizing interactions between adults and children or short humans); our use of an arc-model for entity-placement does not allow this possibility, but this restriction could be relaxed in future work.

Aside from [21, 22] and the British Sign Language generator [14] discussed in section 3, few sign language animation researchers have studied the spatial inflection of verbs. However, some researchers have explored how variations in the movement or performance of a sign language animation can be synthesized based on linguistic details of the sentence in which it appears. For example, Zhao et al. [25] studied how to build an English-to-ASL machine translation system, and they studied how a sign can be varied based on parameters that control the "energy" or "effort" of the movement. They used these parameters to synthesize alternate versions of signs that conveyed adverbial modifications to the sign's meaning. The similarity to our research is that they sought to produce a lexicon of signs that were parameterized on a small number of parameters and could be produced in a specific instance when needed for an animation.

Other research on French Sign Language animation [20] has examined how to produce linguistically accurate coarticulation movements between signs based on an analysis of movements of humans in video data. These researchers digitally analyzed the movements of human signers in video to determine mathematical models of their movements, and then these models could be used during animation synthesis. The similarity to our research is that they were using data from human signers to develop models of signing movements that could be used during synthesis to produce specific instances of sign language movements.

8. CONCLUSION AND FUTURE WORK

This paper has presented and evaluated a novel approach to using motion capture data of sign language performances to construct an animation lexicon of signs whose specific movements vary depending on the context in which they are used. The models created in this paper can be used to synthesize animations of ASL signs whose performance is based on the arrangement of entities under discussion in the signing space; the particular focus of this paper has been modeling the location of the hands during ASL inflecting verbs – based on the values of input parameters that specify the location of the subject and object of a verb. Prior ASL animation systems typically include only a single uninflected version of each verb in their dictionary or only produced a finite variety of verb performances based on a few arrangements of subject and object in the signing space. In prior research [9], we had developed the novel mathematical modeling and animation synthesis approach used in this paper, and the novel contribution of this current work is that we attempted to fit the coefficients of our models to a new source training data – motion capture recordings of human signers performing ASL inflected verbs. In this way, we were able to analyze the suitability and robustness of

our modeling technique to the potentially more noisy (yet more realistic) movement data that is obtained through motion capture.

While prior sign language animation researchers have used motion-capture data to build lexicons [2], our approach allows for the synthesis of an infinite variety of instances of a sign – based on the collection of a finite number of instances from a human performer. The model can produce instances of a sign that were never collected. Using this technique, generation software could include flexible lexicons that can be used to synthesize an infinite variety of inflecting verb instances, and scripting software could more easily enable users to include inflecting verbs in a sentence (without requiring the user to create a custom body movement for each inflected verb sign). While this paper demonstrates our method on four ASL verbs, this technique should be applicable to more ASL verbs, more ASL signs parameterized on spatial locations, and signs in other sign languages used internationally.

In future work, we will collect samples of and model a larger set of ASL inflecting verbs, including some with more complex movements of the hands, and we will experiment with more sophisticated modeling techniques (than the simple polynomial model in this paper). We will also use hand orientation data from our motion-capture sessions to synthesize hand orientation for sign animations. We also plan to experiment with representing subject/object location as 3D points in space (instead of positions on the arc around the signer), and we may model how the timing of verb animation keyframes varies with subject/object position. We will also record and analyze data from additional signers to determine whether our modeling approach is effective when you include multiple copies of signs for the same subject and object position (if you record a human signer on multiple occasions or blend data from multiple human performers); we want to determine if our approach could be used to "average" across these multiple examples of a verb performance in a principled manner.

9. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0746556 and 1065009. This work was supported by a PSC-CUNY Research Award, Siemens A&D UGS PLM Software (Go PLM Grant Program), and a free academic license for character animation software from Visage Technologies AB. Jonathan Lamberton recruited participants and collected response-data during the user-based evaluation study.

10. REFERENCES

[1] Cormier, K. 2002. Grammaticalization of Indexic Signs: How American Sign Language Expresses Numerosity. Ph.D. Dissertation, University of Texas at Austin.

[2] Cox, S., M. Lincoln, J. Tryggvason, M. Nakisa, M. Wells, M. Tutt, S. Abbott. 2002. Tessa, a system to aid communication with deaf people. In *Proceedings of Assets '02*, 205-212.

[3] Elliott, R., Glauert, J., Kennaway, J., Marshall, I., Safar, E. 2008. Linguistic modelling and language-processing technologies for avatar-based sign language presentation. *Univ Access Inf Soc* 6(4), 375-391. Berlin: Springer.

[4] Fotinea, S.E., E. Efthimiou, G. Caridakis, K. Karpouzis. 2008. A knowledge-based sign synthesis architecture. *Univ Access Inf Soc* 6(4):405-418. Berlin: Springer.

[5] Huenerfauth, M. 2006. *Generating American Sign Language classifier predicates for English-to-ASL machine translation*, dissertation, U. of Pennsylvania.

[6] Huenerfauth, M., Hanson, V. 2009. Sign language in the interface: access for deaf signers. In C. Stephanidis (ed.), *Universal Access Handbook*. NJ: Erlbaum. 38.1-38.18.

[7] Huenerfauth, M., L. Zhao, E. Gu, J. Allbeck. 2008. Evaluation of American sign language generation by native ASL signers. *ACM Trans Access Comput* 1(1):1-27.

[8] Huenerfauth, M., P. Lu. 2010. Annotating spatial reference in a motion-capture corpus of American Sign Language discourse. In *Proc. LREC 2010 workshop on representation & processing of sign languages.*

[9] Huenerfauth, M., P. Lu. 2010. Modeling and synthesizing spatially inflected verbs for American sign language animations. In *Proceedings of the 12th international ACM SIGACCESS conference on Computers and accessibility (ASSETS '10).* ACM, New York, NY, USA, 99-106.

[10] Huenerfauth, M, P. Lu. (in press). Effect of spatial reference and verb inflection on the usability of American sign language animation. In *Univ Access Inf Soc*. Berlin: Springer.

[11] Klima, E., U. Bellugi. 1979. *The Signs of Language.* Harvard University Press, Cambridge, MA.

[12] Liddell, S. 2003. *Grammar, Gesture, and Meaning in American Sign Language.* UK: Cambridge U. Press.

[13] Lillo-Martin, D. 1991. *Universal Grammar and American Sign Language: Setting the Null Argument Parameters*. Kluwer Academic Publishers, Dordrecht.

[14] Marshall, I., E. Safar. 2005. Grammar development for sign language avatar-based synthesis. In *Proc. UAHCI'05*.

[15] McBurney, S.L. 2002. Pronominal reference in signed and spoken language. In R.P. Meier, K. Cormier, D. Quinto-Pozos (eds.) *Modality and Structure in Signed and Spoken Languages*. UK: Cambridge U. Press, 329-369.

[16] Meier, R. 1990. Person deixis in American sign language. In S. Fischer, P. Siple (eds.) *Theoretical issues in sign language research*. Chicago: University of Chicago Press, 175-190.

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

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

[19] Padden, C. 1988. *Interaction of morphology & syntax in American Sign Language.* New York: Garland Press.

[20] Segouat, J., A. Braffort. 2009. Toward the study of sign language coarticulation: methodology proposal. In *Proc. Advances in Computer-Human Interactions*, 369-374.

[21] Toro, J. 2004. Automated 3D animation system to inflect agreement verbs. *Proc. 6th High Desert Linguistics Conf.*

[22] *Toro, J. 2005. Automatic verb agreement in computer synthesized depictions of American Sign Language.* Ph.D. dissertation, Depaul University, Chicago, IL.

[23] Traxler, C. 2000. The Stanford achievement test, 9th edition: national norming and performance standards for deaf & hard-of-hearing students. *J Deaf Stud & Deaf Educ* 5(4):337-348.

[24] VCom3D. 2010. Homepage. http://www.vcom3d.com/

[25] Zhao L., K. Kipper, W. Schuler, C. Vogler, N. Badler, M. Palmer. 2000. A machine translation system from English to American Sign Language. In *Proc. AMTA'00*, 293-300.