Modeling and Synthesizing Spatially Inflected Verbs for American Sign Language Animations

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

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

ABSTRACT
Animations of American Sign Language (ASL) have accessibility benefits for many signers with lower levels of written language literacy. This paper introduces a novel method for modeling and synthesizing ASL animations based on movement data collected from native signers. This technique allows for the synthesis of animations of signs (in particular, inflecting verbs, which are frequent in ASL) whose performance is affected by the arrangement of locations in 3D space that represent entities under discussion. Mathematical models of hand movement are trained on examples of signs produced by a human animator. Animations of ASL synthesized from the model were judged to be of similar quality to animations produced by a human animator, and these animations led to higher comprehension scores (than baseline approaches limited to selecting signs from a finite dictionary) in an evaluation study conducted with 18 native signers. This novel technique is applicable to ASL or other sign languages. It can significantly increase the repertoire of generation systems and can partially automate the work of humans using scripting systems.

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
American Sign Language (ASL) is a full natural language with a distinct syntax, word order, and vocabulary from English; it is the primary means of communication for over 500,000 people in the U.S. [14]. Due to various educational factors and levels of exposure to language, a majority of deaf high school graduates in the U.S. have a fourth-grade (age 10) English reading level or below [21]. Many deaf adults have difficulty reading English text on computers, 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]. Animated avatars are more advantageous than video when content is often modified, content is generated or translated automatically, or signers wish to preserve anonymity.

As surveyed in [6], there are two major groups of ASL computer animation research: scripting software (e.g., [1, 22]) or generation software (e.g., [2, 4, 11, 23]). Scripting software allows a human who knows ASL to arrange signs on a timeline to produce animations of ASL sentences without having to manually control all the joints of a virtual human character's body; the software synthesizes an animation from the sentence timeline created by the human. Generation software plans an ASL sentence based on an English input sentence or other information. Unfortunately, modern scripting and generation technologies do not yet address how to produce many types of ASL signs whose motion path is affected by the context in which they appear. Many modulations to sign performance are grammatically governed and essential to understanding an ASL sentence. In fact, many ASL verbs change their motion path to indicate 3D locations in the surrounding space where their subject and/or object have been associated.

This paper presents a novel technique for automatically synthesizing animations of ASL verb signs that undergo such spatial-modulation. Section 6 describes our method, in which we collect multiple examples of the performance of a sign and then fit mathematical models of its movement (based on different settings of the linguistic parameters that affect its performance). Section 7 presents the results of several forms of evaluation of our approach, including an experiment with 18 native ASL signers. The ultimate goal of our work is to construct an animation lexicon of ASL verbs that are spatially parameterized on the 3D location of their subject and/or object (so that a specific instance can be synthesized as needed by ASL scripting or generation software).

2. SPATIALLY INFLECTED ASL VERBS
During a conversation (or during a single-signer multi-sentence discourse), ASL signers frequently associate the entities under discussion (people, things, concepts, etc.) with 3D locations around their bodies [9, 12, 13]. For instance, after mentioning an

ASSETS'10, October 25–27, 2010, Orlando, Florida, USA.

entity for the first time, a signer will typically point at a 3D location in the surrounding space. For the remainder of the conversation, if the signer wants to refer to this entity, he/she will perform a pronoun sign (in which he/she will point to the 3D location). Both participants in a conversation will share the 3D points in space to refer to these entities being discussed, and it is expected that both conversational partners will remember how they have been arranged (since the name of the entities may not be said again during the conversation – their locations are used).

ASL linguists have proposed various models for representing locations where entities are associated [9, 12, 13]. Some believe signers use locations on a semi-circular arc floating at chest height in front of their torso (e.g., see Fig. 2 in section 5 for an idea of what this "arc" might look like) [12, 13]. Other ASL linguists believe that signers use arbitrary 3D locations in the surrounding signing space (at different heights and at different distances from the signer) to represent entities [9]. In both views, signers are not limited to picking from a finite set of locations around their body; they can choose from continuous locations in 3D space.

Many ASL verbs change how they are performed based on the locations of their subject and/or object [9, 15, 16]. While all verbs have a *citation-form* (a standard way they are performed that you might see in a dictionary), many verbs can be *inflected* (grammatically modified to match the linguistic context in which they appear). Verbs typically change their hand path or orientation so that the motion of their citation-form is *deflected from* the 3D location of their subject and *toward* their object; the resulting path is a combination of the verb's citation-form and the subject/object locations. Verbs can be divided into classes as to whether their motion is modified based on: (1) subject only, (2) object only, (3) both subject & object, or (4) neither [9, 16].[1] Fig. 1 shows the ASL verb BLAME, a verb of type "(3)." In many ASL sentences, the subject or object is not overtly expressed: it is the verb's inflection that reveals the identity of its subject and object. If a signer does explicitly mention the subject and object of every verb, then it is legal to use uninflected verbs, but the resulting sentences tend to appear less fluent. Signers who view ASL animations find those that include inflected verbs easier to understand (than those with uninflected verbs) [5].

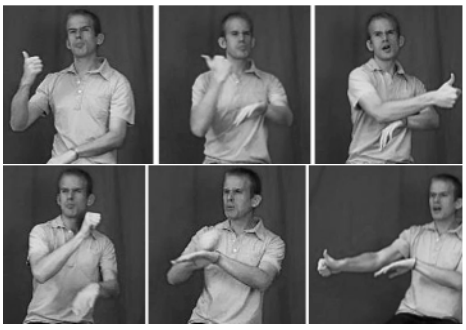

Fig. 1. Two inflected versions of the ASL verb BLAME: on the top row, the subject has been positioned on the left side and the object on the right; on the bottom row, the subject has been positioned on the right side and the object on the left.

[1] Verbs can also indicate adverbial information (often via facial expression) or temporal aspect (e.g., continuous, distributed, or repeated action) by other modifications to a verb's motion path; these types of verb modifications are not the focus of this paper.

3. LIMITS OF CURRENT ASL SYSTEMS

ASL signers can establish infinitely many arrangements of entities in 3D locations in space, and an inflected verb is a combination of the citation-form movement and the arrangement of space. Thus, it is not possible to include all possible performances of such verbs in the dictionary of ASL scripting or generation systems. Most ASL generation systems merely store a single *uninflected* version of each verb in their dictionary; so, the quality of the ASL animations they produce is limited. To measure how much this limitation impacts the ease-of-comprehension of ASL animations, we conducted a study (reported in [5]). We asked 8 native ASL signers to evaluate ASL animations of two versions: (1) a version that included association of entities with locations in space and spatially inflected verbs (carefully produced by a human animator) and (2) a version without the spatial-associations or inflected verbs. The use of spatially inflected ASL verbs led to a significant improvement in user performance on comprehension questions about the animations (the scores doubled) [5]. If ASL animation technology could produce spatially inflected verb forms, then we anticipate significant benefits for deaf users.

If an ASL scripting system were to restrict the user to selecting only signs in its dictionary, then the user who wants to build a sentence that uses a specific inflected version of a verb may become frustrated when it is not in the dictionary. Thus, many scripting systems permit the user to create "custom" signs. For instance, VCom3D Sign Smith Studio [22] is a commercially available ASL animation scripting system. If a user wants to insert a sign into a sentence that does not appear in the standard dictionary, then the user can use accompanying software (called Gesture Builder) to animate a detailed movement of a virtual human character's body. The user uses a GUI to drag and orient the hands of the character to produce a set of animation keyframes that specify a movement of the body for a single sign. This new sign can be saved as an XML file and imported into Sign Smith Studio. This is a somewhat time-consuming process, but it does enable the user to add a wide variety of signs (that weren't included in the software's original dictionary) into sentences. Of course, animating all instances of spatially inflected verbs in this way would be impractically time-consuming. Thus, users of ASL scripting systems tend to use *uninflected* versions of verbs. Another challenge is that some artistic 3D animation skill is needed to produce realistic and understandable signs in this way. Users of scripting software are already required to know ASL grammar (in order to produce correct ASL sentences), but if they are also trying to add "custom" signs, then they also need some skill in animating the 3D movements of the virtual character.

The ultimate goal of our research is to construct computational models of ASL verbs that could be used to partially automate the work of human authors using scripting software or to underlie generation/translation systems. We model ASL verbs whose paths and orientations depend on the locations in 3D space where their subject/object have been established. Such signs are time-consuming for users of scripting software to produce, and they are not in the repertoire of most modern ASL generation software.

4. RELATED WORK

Marshall and Safar's [11] British Sign Language generator could associate entities with a finite number of locations in the signing space (approximately 6), and the system produced a few verbs whose subject/object were positioned at these locations. However,

the verbs discussed in [11] involve simple motion paths for the hands from subject to object locations, and the system did not allow for the arrangement of spatial reference locations at arbitrary locations in the signing space (human signers use a wide variety of locations, not just a fixed, finite set). In sections 5 and 6, we will describe a more general approach that can synthesize verbs whose subjects/objects are not restricted to a finite set of locations and whose motion paths are more complex.

Aside from [11], prior sign language animation researchers have not studied how to model spatial inflection of verbs. However, some have explored how novel varieties of a sign can be synthesized at run time based on details of the sentence in which it appears. For example, in their English-to-ASL translation work, Zhao et al. [23] explored how a sign can be modulated based on parameters that control the "energy" or "effort" of the movement, to produce subtle adverbial modifications to the sign's meaning.

Using 3D movement data of human performances of French Sign Language, Segout and Braffort [19] study coarticulation (how hand/finger positions of surrounding signs affect the current sign). They seek a model that would allow them to synthesize novel sign movements based on the 3D position of the hands before/after a sign. The similarity to our research is that they want to represent the movement of a sign in a parameterized manner such that a novel form can be synthesized as needed for an animation. However, in our research, it is the arrangement of subject/object in the surrounding signing space that affects the verb performance (not the body positions for the previous/subsequent signs); thus, linguistic research on coarticulation is less relevant to our work.

Animation researchers (not studying sign language) have also synthesized novel human animations from sample animations produced by humans (or from data collected via motion-capture) using interpolation techniques that "blend" human movement data [17, 18, 20]. Rose et al. [18] generated novel animations from a small number of recorded motion-capture examples of an action; e.g., for "reaching" actions parameterized on the 3D location of the target object being reached for. Their motion-capture data had to be pre-processed by humans to: (1) mark the 3D location of the target object and (2) identify key time points in the movement that correspond across examples of the action. Next, they performed B-spline approximation of all of the human body's joint angles over time and interpolated their data using low-order polynomials and radial basis functions [18]. Section 6 describes our methodology for modeling ASL inflected verbs, which has been influenced by this prior research. Because of the linguistic regularity of ASL sign movements, we have been able to adapt and simplify techniques designed by prior researchers.

5. GOAL OF OUR RESEARCH

The goal of our research is to produce a parameterized animation lexicon of ASL inflecting verbs. Given a 3D location of where in the signing space the subject and object of a verb is placed, we want to produce an *instance* of the verb that has been properly inflected – i.e., its motion-path has been modified to reflect the subject and object locations in 3D space. We would like to train our model on instances of ASL verbs collected from native signers. Because it is not possible to collect infinitely many signs (for all possible subject/object locations), our model must be able to synthesize previously-unseen instances of a verb for novel arrangement of subject/object locations. In our current research, we focus on five verbs (listed in Table 1), but we intend for our methodology to be generalizable to other ASL inflecting verbs and other sign languages (see section 8). The five verbs represent a variety of inflection patterns: some inflect based on their subject and object location (shown as "Subj+Obj" in Table 1), and some inflect based on their object location only. The verbs represent a mixture of one- and two-handed signs. Animations of signs appear on our lab website: http://latlab.cs.qc.cuny.edu/assets2010/. Table 1 describes the movement of MEET as symmetrical: i.e., if a signer associates one entity on the left side of the signing space (e.g. "John") and one on the right (e.g. "Mary"), then the sign MEET in the sentence "John MEET Mary" would look identical to the performance of MEET in the sentence "Mary MEET John."

Table 1: Five ASL Inflecting Verbs Examined in This Paper

Verb	*Inflection*	*#Hands*	*Description*
ASK1h	Subj+Obj	1	'ask a question': a bending index finger moves from Subj ('asker') to Obj ('askee')
GIVE2h	Subj+Obj	2	'give to someone': hands move as a pair from the Subj ('giver') to Obj ('recipient')
MEET	Subj+Obj	2	'two people meet': hands move from Subj and Obj toward each other symmetrically
SCOLD	Obj only	1	'scold/reprimand': extended index finger wags at the Obj ('person being scolded')
TELL	Obj only	1	'tell someone': index finger moves from signer's mouth to Obj ('person being told')

Because of the capabilities of modern virtual human animation software, we can make several simplifications for our research:

- *Solving Body Joint Angles through Inverse Kinematics.* While [18] interpolated models for all the joints of the body, we do not need all these joints to specify a sign. ASL signs are *linguistic* and can be represented by a smaller number of parameters: handshape, hand location, and hand orientation [10]. Thus, if we can build a model that predicts hand location and orientation, we can synthesize an animation of most ASL signs. Specifically, inverse kinematics can be used to successfully calculate shoulder, elbow, and wrist angles that will get the hand to a desired location/orientation. In prior research, native signers judged ASL animations produced by a system using inverse kinematics to be understandable [7].
- *Motion Interpolation through Keyframes.* Modern animation software can also interpolate a motion path for an object (like a hand) given a list of "keyframes" (locations/orientations for the hand at specific moments on a timeline). Thus, a complex sign movement can be reduced into a list of static keyframes – this further simplifies our modeling task. In fact, [18] used a similar simplification; the motion-capture data they used to build their models of an action had to be pre-processed by a human to identify its basic keyframes. Several ASL animation systems (e.g. [7, 22]) use a keyframe-based approach, with good-quality, understandable results [5, 7]. For each verb we modeled, we have defined a list of keyframes, which generally correspond to the apexes of movement curves.

Other linguistic aspects of ASL allow us to make some further simplifying assumptions for our verb-modeling research:

- *Handshape.* The handshape of an inflecting ASL verb generally does not change based on how the subject/object are positioned [9]. Thus, we hard-code a single handshape for each keyframe of the verb. This approach allows us to model a verb whose handshape changes during its performance (i.e., keyframes can

have different handshapes), but the handshape itself is not affected by the subject/object location. In future work, we may lift this assumption: in some verbs in which the finger points to a subject/object location, the finger joint angle changes slightly.

- *Subject/Object Locations on an Arc.* Section 2 discussed how ASL linguists disagree whether human signers associate entities under discussion with arbitrary locations throughout the 3D space or use locations on an "arc" around their body. For our current research, we assume all locations are on an arc. While this still leaves open the possibility of infinitely many locations at which combinations of subjects and objects could be placed, it does reduce the dimensionality of our verb-modeling task. Verbs like "ASK1h" whose movement path is affected by both subject and object can be thought of as being parameterized on two values: the arc-position of the subject and the arc-position of the object. Fig. 2 displays the specific arc location and numbering scheme used in our research. In future work, we may also relax this assumption: in that case, the modeling task becomes more complex. Verbs like "ASK1h" would now be parameterized on *six* values: location of both the subject (*x,y,z*) and the object (*x,y,z*) in the 3D signing space.

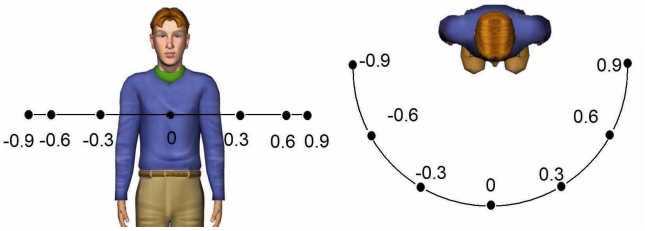

Fig. 2. Front & top view of arc-positions around the signer.

Thus, after considering the assumptions outlined above, the specific goal of our research is as follows: based on samples of verb performances collected from a native signer, we seek to build a mathematical model of the location and orientation of the hand during the keyframes of the performance of ASL inflecting verbs. This model will be parameterized on the location of the subject and/or object of the verb – each specified by a real number that represents the position on an arc around the signer. Location of the hand will be encoded as (*x,y,z*) coordinates and the 3D orientation of the hand will be encoded as a quaternion, which is a quadruple of four real numbers: the first 3 numbers represent a vector in 3D space and the 4th number is an angle of rotation on the axis represented by this vector [3]. ("Euler angles" are another common way to represent 3D orientation, but quaternions are more mathematically well-behaved during 3D interpolation).

6. METHODOLOGY

Our overall methodology is to collect samples of instances of ASL inflecting verbs (for a variety of subject and object locations) and fit low-order polynomial models to the data so that we can predict hand location/orientation (and thereby synthesize animations of novel verb instances, as needed) for given subject and object locations. After constructing these models for the five example ASL verbs discussed in this paper, we conducted several forms of evaluation – including a study with 18 native ASL signers.

6.1 Collecting Samples of Inflected Verbs

To collect instances of verbs, we asked a native ASL signer to use VCom3D Gesture Builder [22], which enables the creation of new signs (to add to the VCom3D dictionary). The GUI allows the user to drag the hand and arms of a virtual human character to produce static poses for keyframes on a timeline. The user can press a button to "play" the animation they have produced to see what the sign would look like as it is performed. After first practicing with the software (to produce a few dozen "practice" signs), the signer was asked to produce instances of verbs – for given locations of subject and object on the arc around the signer. A clear plastic sheet was overlaid on the computer monitor with a scale drawing of an arc with angles on the arc labeled (similar to Fig. 2). The signer was given a list of verbs to produce, e.g., GIVE1h with the subject at arc-position 0.3 and the object at arc-position 0.9. The signer was also told how many keyframes to use for each verb (and time index for each keyframe); e.g., all of the collected instances of the verb GIVE1h for different subject and object locations would use the same number of keyframes (with the hands at different locations/orientations for each instance).

For each instance of a verb we collected, we needed to record:

- Location of the subject of this instance of the verb represented as a real-number specifying a location on the "arc",
- Location of the object of this instance of the verb,
- For each keyframe of this instance of the verb:
  - Location of the hand represented as (*x,y,z*) coordinates,
  - Orientation of the hand represented as a quaternion.

The XML file produced by Gesture Builder stores keyframes with hand locations and orientations (specified as quaternions); so, it was easy to extract this information above for each instance of a verb we collected. To collect a variety of instances of each verb, we divided the "arc" around the signer into seven discrete locations (Fig. 2). For verbs that inflect for object location only, we collected seven instances (one for the object at each of these locations). For verbs that inflect for both subject and object location, we collected 42 instances for each non-reflexive combination of subject/object locations on the arc. (ASL signers tend to express reflexive verbs such as "John asks himself" with an uninflected form of a verb.) Because of the symmetry in how the verb MEET is performed, only 21 instances of this verb were collected. (MEET with subject at arc-position 0.3 and object at 0.9 is identical to MEET with subject at 0.9 and object at 0.3).

6.2 Building Models of ASL Verb Inflection

For each keyframe of a verb instance, there are 7 values to be fit: 3 parameters of the location (*x,y,z*) and 4 parameters of the orientation quaternion (q_0,q_1,q_2,q_3). We decided to fit 3rd-order polynomial models for each of these 7 independent parameters. For the verbs whose performance is affected by the arc-position of the object only, the model can be formalized as a function parameterized on one value. Let's assume that o is the arc-position of the verb's object and that m is one of the seven parameter values for an instance of a verb: x, y, z, q_0, q_1, q_2, q_3. In this case, the function for each parameter has the form:

$$m = f(o) \quad (1)$$

For verbs whose performance is affected by the arc-positions of both the subject and the object, the model can be formalized as a function parameterized on two values. Assume that s and o are the arc-positions of the subject and object respectively and that m is one of the 7 parameters for an instance of a verb: x, y, z, q_0, q_1, q_2, q_3. In this case, the function for each parameter has the form:

$$m = f(s, o) \quad (2)$$

Thus, for a one-handed verb with two keyframes, 14 functions would be required: to specify all 7 values of the hand for each keyframe. For a two-handed verb with two keyframes, 28 functions would be required. (All five of the verbs that we modeled had two keyframes.) 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.

To obtain the functions, we used MATLAB code to identify the coefficients for a polynomial of degree 3 that best fit the training data – in a least-squares sense. (The overall solution of least-squares polynomial-fitting minimizes the sum of the squares of the errors between each value in the training data and the model's prediction for that value.) We trained our model on the instances of ASL verbs we collected from a native signer (as described in section 6.1). Functions parameterized on object arc-position only, as shown in equation (1), contained terms up to o^3. Functions parameterized on both subject and object arc position, as shown in equation (2), contained terms up to s^3 and o^3 and all possible cross-product terms $s^a o^b$ where $a \leq 3$, $b \leq 3$, $a+b \leq 3$. Thus, at the end of the training, we had a function (a 3rd-order polynomial) that predicts the value of each of the 7 verb parameters for each keyframe, given subject and object arc-positions.

We selected *polynomials* to model each of our verb parameters as a compromise between two competing design issues: (1) our desire to accurately model the verb parameters and (2) our desire to use a simple modeling approach that would be easy for other sign language researchers to replicate for their own use. We selected to use *3rd-order* polynomials because fitting to high order polynomials is numerically sensitive and requires more data. Thus, we first tried to use 1st-order and 2nd-order models before we eventually decided upon 3rd-order models. To illustrate how lower order polynomials did not accurately fit some training data, Fig. 3 shows 3D plots of the *x* location of the right hand in keyframe 1 of "GIVE2h." In Fig. 3, the *subject* axis is the arc-position value of the verb's subject and the *object* axis is the arc-position of the object. The vertical axis (also indicated by color-coding) shows the *x* coordinate value of the hand. The "dots" on the plot show the value for the hand's *x*-coordinate from the data collected from the human animator, and the "open-circles" show the hand's *x*-coordinate as predicted by our model. (The trend in the dots indicates that the subject arc-position is the major influence on the hand's x-coordinate at the beginning of the performance of the verb GIVE1h.) In Fig. 3, the 3rd-order model does a better job of fitting the data collected (as indicated by the proximity of the dots to the open-circles in the plot on the right).

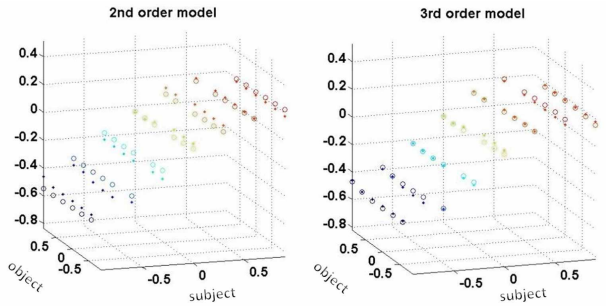

Fig. 3. Comparison of 2nd-order and 3rd-order models for the *x* coordinate of the right hand for keyframe 1 of GIVE1h.

6.3 Synthesizing Animations from the Models

The output of our verb model is a list of the locations (*x*,*y*,*z*) and orientations (as quaternions) for each keyframe of an instance of a verb (for a given subject and object location). For each keyframe, we know its time index on a timeline for that verb and the handshape. (These values are assumed to be constant across all instances of a verb; this assumption may be relaxed in future work.) 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. 4 shows keyframes of the verbs "ASK1h" and "GIVE2h" for subject at arc-position -0.6 and object at 0.3; our lab website includes example animations of the other verbs we produced: http://latlab.cs.qc.cuny.edu/assets2010/.

Fig. 4. The top row shows keyframes 1 and 2 of ASK1h produced by our model; the bottom row shows GIVE2h.

7. EVALUATION OF OUR MODELS

This section presents two types of evaluation studies we conducted to measure the quality of our verb inflection animation models. Section 6.2 described how we used the verb instances collected from a native signer to fit 3rd-order polynomial models for each of the seven verb parameters for each hand for each keyframe. To build a good model that would be used in a future ASL animation system, we would make use of as much training data as possible when fitting our models. However, *for evaluation purposes only*, to produce the animations of ASL verbs used throughout the evaluation studies presented in sections 7.1 and 7.2 below, we wanted to be more rigorous. To make the task of our model a little more difficult, we followed a "leave-one-out" strategy, as follows: For example, when we want to produce an animation of an instance of "GIVE2h" with subject at arc-position 0.3 and object at arc-position 0.9, we trained models using all instances of "GIVE2h" except the human animator's instance of "GIVE2h" with subject at arc-position 0.3 and object at arc-position 0.9. In this way, the specific instance that we were producing did not appear in the training data. We believe that this more difficult form of evaluation is a better way of measuring the

quality of our model – in actual use, the model might be asked to synthesize verb instances that did not appear in its training data.

7.1 Comparing Model to Human Data

To evaluate how well the signs produced by our model match signs produced by a human animator, we decided to focus on a specific instance of each of our five verbs: the instance of the verb with the subject at arc-position -0.6 and the object at arc-position 0.3. (Of course, for some of the verbs, the arc-position of the subject is irrelevant.) We used our model to predict the 7 parameters $\{x, y, z, q_0, q_1, q_2, q_3\}$ for each hand for each keyframe of the five verbs. Next, we asked the human animator to return to the lab on three different days. Each day, he used Gesture Builder to produce an instance of each of the five verbs – assuming that the subject is at arc-position -0.6 and the object is at arc-position 0.3. Despite being asked to assume identical subject and object locations, the animations he produced on different days varied slightly. Finally, we compared the instance of each of the five verbs predicted by our model to the three collected samples of each verb from the human animator. Fig. 5 shows the close-up view of the finger differences between the model and the human animator's versions for the first keyframe of the verb "ASK1h".

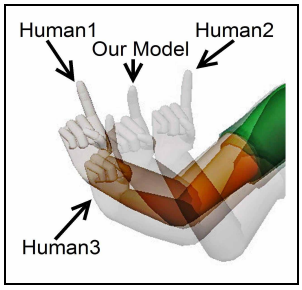

Fig. 5. Close-up view of differences between keyframe 1 of ASK1h produced by our model or by the human animator.

We calculated the average of the differences between our model and each of the three human-produced versions. For the (x,y,z) location coordinates, we used Euclidean distance. Since quaternion representations of the orientation of an object have a geometric meaning, there is a natural distance metric for comparing two quaternions [3]. If we assume that "object 1" has an orientation of $q=(q_0,q_1,q_2,q_3)$ and object 2 has an orientation of $p=(p_0,p_1,p_2,p_3)$, then the magnitude of the angle of rotation required to re-orient object 1 to match the orientation of object 2 is the value *theta* calculated by the pair of equations below [3]:

$$d(p, q) = |q \bullet p| = |q_0p_0 + q_1p_1 + q_2p_2 + q_3p_3| \quad (3)$$

$$theta(q, p) = arccos(d(q, p)) \quad (4)$$

Because the instance of the verbs produced by the human animator varied on each of the three days of data collection, we also calculated the average of the pairwise differences between the three human-produced versions to estimate the variance in human-produced signs. Fig. 6 compares "model vs. human" to the "human vs. human" pairwise differences. To give the reader a sense of scale for the "location differences" in Fig. 6, the width of the virtual human's shoulders is about 0.35 units. The values shown for "orientation differences" are based on equations (3) and (4). The amount that our model differed from each of human samples was similar to the amount that the human-produced samples differed from each other. There were no significant differences between the bars shown in Fig. 6 ($p>0.05$, t-test).

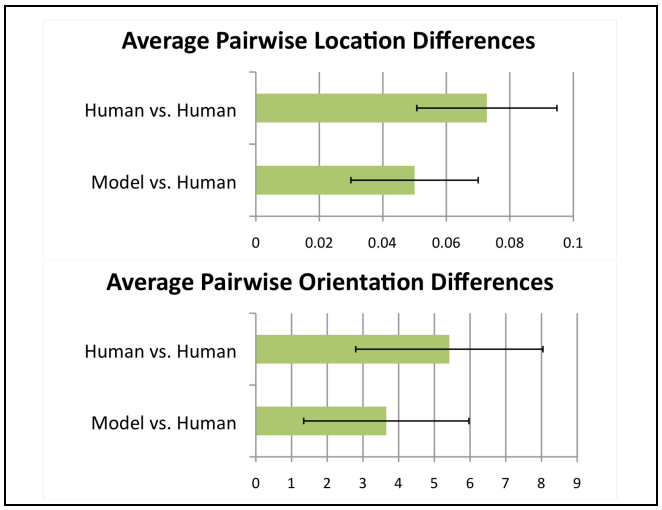

Fig. 6. Comparison of pairwise differences between model vs. human signs and pairwise differences among human signs.

7.2 Evaluation Study with Deaf Users

To evaluate how well target users would understand and enjoy the inflected verb animations produced by our model, we conducted an evaluation study in which 18 native ASL signers evaluated ASL computer animations of three types: (1) with inflected verbs synthesized using our model, (2) with inflected verbs produced by a human animator (native signer), and (3) with uninflected verbs (standard dictionary versions of each verb). Sign Smith Studio was used to produce the ASL animations for this study; we have used this software in prior ASL animation research [5]. In prior studies, we have developed best-practices to ensure that responses given by participants are as ASL-accurate as possible [7]. We've discussed how participants should be native ASL signers, how to ask questions to screen for such participants, and how the study environment should be ASL-focused with little English influence. For the current study, all instructions and interactions were conducted in ASL by a native signer. Ads posted on Deaf community websites in New York City asked potential participants if they had grown up using ASL at home or had attended an ASL-based school as a young child. Of the 18 participants, 12 participants learned ASL prior to age 5, and 4 participants attended residential schools using ASL since early childhood. The remaining 2 participants have been using ASL for over 15 years, learned ASL as adolescents, attended a university with classroom instruction in ASL, and use ASL on a daily basis to communicate with a significant other or family member. There were 12 men and 6 women of ages 20-56 (median age 30.5).

In phase 1 of the study, participants viewed animations of a virtual human character telling a short story (in ASL) that included instances of the five verbs we are studying (appearing 5-6 times in each story). Stories were an average of 55 signs in length and included 3-5 main characters, each of which was associated with a different location on the arc around the signer. Thus, many verb instances were produced in which people associated with locations on the arc served as the subject or object of the inflecting verbs. The 9 stories were produced in 3 versions: some with verbs synthesized by our model, some with verbs produced by a native signer using Gesture Builder, and some using uninflected dictionary form of each verb. A fully-factorial

within-subjects design was used such that: (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 one time, the participants answered a set of four multiple-choice comprehension questions, which focused on information conveyed by the verbs. Questions focused on whether they understood and remembered the subject and object of each verb. Details of the methodology we have used for similar ASL animation evaluation studies are described here [5, 7].

In phase 2 of the study, participants viewed three animations side-by-side, and they could re-play each animation as many times as they wished. The animations shown in this phase consisted of three versions of a single ASL sentence (shown side-by-side), e.g. "John point-to-position-0.3 ASK1h Mary point-to-position-0.9." The only difference between the three versions was whether the verb: was produced by our model, was created by a human animator using Gesture Builder, or was the uninflected version of the verb in the VCom3D dictionary. A variety of arc-positions for subject and object were used throughout the study (the three versions shown at one time all used the same arc-positions), and the arrangement of the three animations on the screen was randomized (sometimes our model's version was leftmost, sometimes the human animator's version was, etc.). Participants were asked to focus on the verb and consider its grammaticality, understandability, and naturalness in each of the 3 versions of the sentence. They were asked to assign a 1-to-10 Likert-scale score to each of the three versions of the animation.

Fig. 7 shows the results of the side-by-side comparison and comprehension question studies. Error bars indicate the *standard error of the mean* for each value; significant pairwise differences are marked with a star. Statistical tests were planned prior to data collection. To check for significant differences between "phase 2" Likert-scale scores for each version of the animations, a Kruskal-Wallis test was performed (pairwise Mann-Whitney U-tests with Bonferroni-corrected p-values). A non-parametric test was used because the scalar response data was not normally distributed. An ANOVA was used to look for significant differences between comprehension question scores in "phase 1."

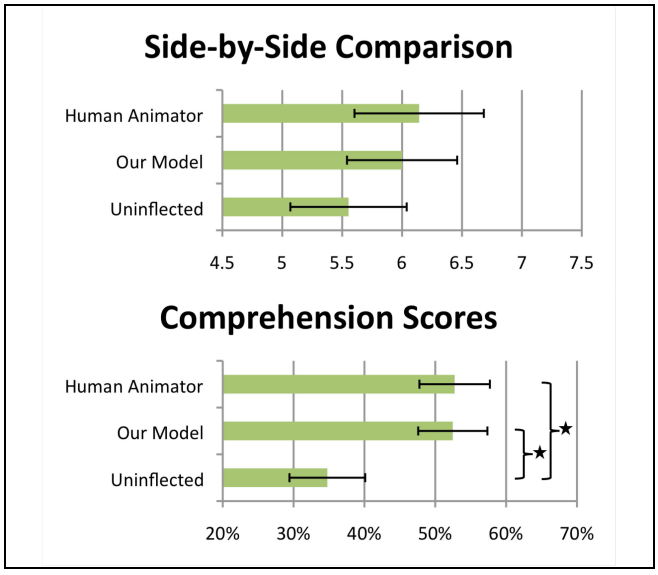

Fig. 7. Results of evaluation study with native ASL signers.

In both the Likert-scale data and the comprehension-score data, our model tended to have similar performance to the inflected verb animation produced by the human animator using Gesture Builder; this result suggests that our model was producing an ASL sign of similar quality to the human animator. For comprehension question scores, both our model and the human animator's verb scored significantly higher than the uninflected verb animations.

8. CONCLUSION

This paper has presented and evaluated a novel approach to synthesizing animations of ASL signs whose performance is based on the arrangement of entities under discussion in the signing space. Specifically, an approach to modeling the location and orientation of the hands during the performance of ASL inflecting verbs was described that enables an animation system to produce *infinitely many* versions of a verb – based on the values of input parameters that specify the location of the subject and object of a verb. Prior ASL animation systems have included 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.

The key advantage of our methodology is that it 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 animator. Thus, the model can produce instances of a sign that were never collected. Further, we have found that human animators tend to vary in the way that they produce an instance of a sign on different occasions (as described in section 7.1). Our methodology would also allow for the use of data consisting of multiple copies of an instance of a sign for the same subject and object position (but with different hand location and orientation values). Our modeling approach can "average" across these multiple examples of a verb performance in a principled manner.

While this paper demonstrated the approach for five verbs, the methodology (collecting samples of verb instances and training models) can be applied to many more ASL verbs – and to other ASL signs (e.g., pronoun signs, in which the signer points to a 3D location). Further, the methodology is also applicable to other sign languages, many of which include signs whose movements or orientations are affected by how the signer has established entities in the space around his/her body. Further, many sign languages (including ASL) have regional dialects; this data-driven model-creation approach allows for the creation of models of signs for specific sub-varieties of a sign language (if appropriate samples of signs are collected as training data for each dialect).

Prior research has shown that the use of inflected verb forms improves signer's comprehension of information presented in an ASL animation [5], and this was confirmed by the results presented in section 7.2. By following the methodology described in this paper, creators of sign-language generation software can enable an infinite variety of inflecting verb instances to be included in the repertoire of their generation software. Creators of sign-language scripting software can also use this technique to construct models of signs for their dictionaries that allow the user to easily include a specific instance of an inflecting verb in a sentence (without needing to perform the time-consuming creation of a "custom" sign, as described in section 3). Incorporating parameterized models of signs (that can be modulated to produce different versions as needed) should make sign-language scripting software easier and more efficient to use.

9. FUTURE WORK

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 or some in which the hands move in close proximity to each other. Such verbs may pose a greater challenge for the polynomial models presented in this paper; we may experiment with new modeling techniques. Further, section 5 discussed several simplifying assumptions for the current version of our model; we will explore relaxing these assumptions as needed to accommodate the larger set of ASL verbs we model. For instance, we may begin to model how handshape is affected by subject/object location, and we may model subject/object location as 3D points in space (instead of positions on the arc around the signer). We may also allow for varied timing of keyframes so that verb instances with longer movement paths produce animations that use more time than instances in which the hands travel a shorter distance. Overall, our approach has been to use the simplest model possible (based on the set of verbs we are modeling) and then to add complexity to the model only as necessary (to accommodate a wider variety of signs).

We have been using Sign Smith Studio [22] to synthesize virtual humans; we will later implement our models in our own virtual human animation codebase (under development), which includes inverse kinematics and motion interpolation capabilities.

We have been using Gesture Builder [22] to collect training data, but there are challenges to this approach: it can be difficult for some native ASL signers to produce accurate and realistic signs using an animation tool. The way they move the virtual human character may not accurately reflect how they would move their own body when signing. If the collected training data is not natural, then the resulting model is lower quality. Our lab is conducting a multi-year project to use motion-capture (body suits, sensors, gloves) to collect ASL performances [8]. In future work, we intend to use this data to train our models. While possibly more natural, motion-capture data poses new challenges: identifying keyframes in the performance, cleaning up "noise" in the motion data, extracting appropriate location and orientation parameters, etc. As we shift from using human-animator data to motion-capture data in future work, we will address these issues.

10. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0746556. This work was supported by PSC-CUNY Research Award Program, 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.

11. REFERENCES

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

[2] 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.

[3] Hanson, A.J. 2006. *Visualizing Quaternions.* San Francisco: Morgan-Kaufmann/Elsevier, ISBN 978-0-12-088400-1.

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

[5] 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.

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

[10] Liddell, S., Johnson, R. 1989. American Sign Language: The phonological base. *Sign Language Studies* 64, 195-277.

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

[12] 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.

[13] 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.

[14] 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.

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

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

[17] Park, S.I., Shin, H.J., Shin, S.Y. 2002. On-line locomotion generation based on motion blending. In *Proc. SIGGRAPH/ Eurographics Sympos. on Computer Animation*, 105-111.

[18] Rose, C., Cohen, M.F., Bodenheimer, B. 1998. Verbs and adverbs: multidimensional motion interpolation. In *IEEE Computer Graphics and Applications* 18(5):32-40.

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

[20] Sloan, P.P., Rose, C., Cohen, M.F. 2001. Shape by example. In *Proc. Symposium on Interactive 3D Graphics*, 135-144.

[21] 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.

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

[23] 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.