Synthesizing and Evaluating Animations of American Sign Language Verbs Modeled from Motion-Capture Data

Matt Huenerfauth [1], *Pengfei Lu* [2], *Hernisa Kacorri* [2]

[1] Rochester Institute of Technology, Golisano College of Computing and Information Sciences
[2] The Graduate Center, CUNY, Doctoral Program in Computer Science
matt.huenerfauth@rit.edu, pengfeilv@gmail.com, hkacorri@gradcenter.cuny.edu

Abstract

Animations of American Sign Language (ASL) can make information accessible for many signers with lower levels of English literacy. Automatically synthesizing such animations is challenging because the movements of ASL signs often depend on the context in which they appear, e.g., many ASL verb movements depend on locations in the signing space the signer has associated with the verb's subject and object. This paper presents several techniques for automatically synthesizing novel instances of ASL verbs whose motion-path and hand-orientation must accurately reflect the subject and object locations in 3D space, including enhancements to to prior state-of-the-art models. Using these models, animation generation software could produce an infinite variety of indicating verb instances. Using a corpus of motion-capture recordings of multiple performances of eight ASL indicating verbs, we modeled the signer's hand locations and orientations during each verb, dependent upon the location in the signing space where the subject and object were positioned. In a user study, ASL signers watched animations that included verbs synthesized from these models, and we found that they had similar quality to those produced by a human animator.

Index Terms: American Sign Language, accessibility for people who are deaf, animation, natural language generation

1. Introduction

This paper describes technologies for automating the creation of animations of American Sign Language (ASL), which is a natural language that consists of movements of the hands, body, head, and face. ASL is the primary means of communication for over 500,000 people in the United States [18]. ASL is a natural language, and the grammar, word-order, and vocabulary of the language is distinct from English. For various language-exposure and educational reasons, many deaf adults have lower literacy levels. In fact, standardized testing suggests that the majority of deaf high school graduates in the U.S. (typically students age 18) have a fourth-grade reading level or below (typically students age 10) [22].

Given these literacy trends, when English text is presented online, the text may sometimes be too difficult for many of these users. Conveying information content through videos or animations of ASL could make information more accessible. As discussed in [14], because a human signer must be re-filmed, videos are ill-suited to contexts where the information: is often modified, might require later corrections, is generated automatically in response to a query, or is produced by automatic translation technologies. Animations of sign language that are produced automatically from an easy-to-update script can overcome these limitations and make it easier to incorporate ASL content on websites or other media. A challenge is that ASL signs must be "customized" so that they are performed in a specific manner that matches how the signer has set up locations around their body to represent entities under discussion. This paper focuses on a ubiquitous class of ASL verbs, called "indicating verbs," and it describes research on technologies to automatically produce understandable animations of these verbs for use in ASL animations, with an ultimate goal of increasing information accessibility for users who are deaf.

1.1. Spatial Reference Points, Indicating Verbs

In ASL, entities under discussion (concepts, people, etc.) can be associated with locations in the 3D space around the signer's body [10, 16, 17]. For example, the signer may point to a location in space immediately after an entity is mentioned for the first time; when signers want to refer to the entity again, they do not repeat its name. Instead, they point to the location in space. Some linguists (e.g., [16, 17]) have proposed that ASL signers can be thought of as using a semi-circular arc (at chest height around their torso) as the range of possible locations where entities may be established. The location of the spatial reference point for an entity could be represented as an angle on this arc. Individual signs may vary in how they are performed in a particular sentence, based on a variety of linguistic factors. For instance, temporal aspect, manner, or spatial depiction can be conveyed through modifications to the performance of an ASL verb [4, 5]. However, the focus of this paper is a class of ASL verbs referred to as "indicating verbs" by [10] (also known as "inflecting verbs" [19] or "agreeing verbs" [1].) The movement path and hand orientation of these verbs is affected by the spatial reference points for the verb's subject and/or object [10, 19].

When a signer is asked to perform an indicating verb in isolation or when such a verb is listed in a dictionary, the prototypical verb performance that is seen is typically referred to as a "citation form" or "uninflected form," which has not been modified to indicate locations in the signing space for its subject or object. When an indicating verb is performed in a sentence, the signer will modify the hand locations and orientations used to perform the verb, often tracing a unique motion-path through the signing space, which indicates the locations of the spatial reference points for the verb's subject and/or object. In fact, in ASL sentences that include an indicating verb, the subject or object is often not overtly expressed. That is, the signer does not need to point to the spatial reference locations for the subject or object as long as the verb's motion-path and orientation reveals the identity of its subject and object. If a signer does choose to explicitly mention the subject and object of the verb, then it is legal for

SLPAT 2015, 6th Workshop on Speech and Language Processing for Assistive Technologies, pages 22–28,
Dresden, Germany, 11 September, 2015.

the signer to simply use the uninflected form of the verb, but the resulting sentences may appear less fluent. Signers who view ASL animations find those that include citation forms of indicating verbs more difficult to understand (as compared to versions of animations in which indicating verbs indicate the locations of the subject and object) [7].

Generally, the motion path of indicating verbs moves away from the subject and toward the object, but the verb performance is actually a complex interaction of: (a) the verb's citation-form motion-path and hand orientation, (b) the location of the subject's spatial reference point, and (c) the location of the object's spatial reference point. ASL verbs can be partitioned into multiple classes, based on whether their motion is modified based on: (1) subject only, (2) object only, (3) both, or (4) neither [10, 19]. Figure 1 shows the verb EMAIL, which is a verb of type (3).

Figure 1: Verb EMAIL with: (a) subject on the left and object on right or (b) with opposite arrangement.

This paper describes our research on automating the creation of animations of ASL indicating verbs. Section 2 briefly summarizes some prior work on modeling ASL indicating verbs. Section 3 describes new techniques for automatically synthesizing animations of ASL verb signs. Section 4 presents an experiment with 18 native ASL signers who evaluate animations resulting from our modeling techniques. Finally, section 5 presents conclusions and avenues for future work.

2. Prior Work on ASL Verbs

Researchers have investigated methods to speed the creation of sign language animations. *Scripting* systems, e.g., [23], allow a human who is knowledgeable of ASL to assemble sentences by drawing upon pre-built words in a dictionary to create a timeline for a performance. A common limitation is that the user may not find the exact sign (or version of a sign) that is needed for a particular sentence, e.g., most systems include only the citation form of verb signs because it is not practical to include hundreds of versions of each verb for various possible arrangements of the verb's subject and object in the signing space. As discussed in [6], other researchers have focused on building *generation* systems, which further automate the production of animation, e.g. research on machine translation of written text into sign language animation. In order for the machine translation output to include indicating verbs, some method is needed for automatically predicting how the motion-path and orientation of a verb would be affected by the locations of the verb's subject and object in space. Sections 2.1 and 3 describe research on automatically synthesizing novel performances of ASL verb signs for any desired combination of subject and object arrangement in the signing space: Such software would be useful in both scripting and generation systems, thereby making it easier to add indicating verbs to animations.

Marshall and Safar [15] designed an animation generator that could associate entities with up to six locations in the signing space and produce British Sign Language verbs whose subject/object were positioned at these locations. However, the verbs involved simple motion paths, and the system did not allow for the arrangement of subjects and objects at arbitrary locations in the signing space (a small number were enabled).

Some researchers have studied videos of performances of ASL verbs to design algorithms for specifying how the arms should move for specific arrangements of subject and object [21]. While the results were promising, a human animation programmer was needed to design the necessary algorithms. By contrast, our research is based on the idea that the only input should be a set of examples of how an ASL verb is performed for various given arrangements of subject and object, with the software automatically learning a model of how a signer's hands should move, given where the subject and object is located in space.

Other researchers have collected motion-capture recordings of signing and used this data to synthesize novel verb signs: Duarte and Gibet [2] collected French Sign Language data via motion capture, and they reassembled elements of the recordings to synthesize novel animations. They used several "channels" to represent their recorded signs, e.g., channels of eye, head, spine, and arms, and they mixed information from the channels of different recordings to produce new animations. For a small number of verb signs, these researchers played the recording of the verb in reverse (from the original recording) to produce a version of the verb with the subject and object in opposite locations. For example, they recorded several indicating verbs with a few combinations of subject/object, e.g., "I invite-you" and "you-invite-I." However, they did not try to build a model of how to synthesize novel inflections for verbs for any arrangement of subject or object in the signing space (the focus of this paper).

2.1. Earlier Work on ASL Indicating Verb Modeling

In earlier work, researchers have designed data-driven methods for synthesizing animations of ASL indicating verbs, for any desired arrangement of the subject and object on an arc around the signer. However, there were significant limitations in that prior work, which we address with some novel modeling approaches described and evaluated in this paper.

As described in [11], verb performances were collected from human ASL signers to create a training data set for animation modeling research. The data included the location (x, y, z) and orientation (*roll*, *pitch*, *yaw*) for the hands, torso, and head of the signer. The native ASL signer performed ASL verbs signs, for given arrangements of the subject and object in the signing space. Targets were positioned around the laboratory at precise angles, relative to where the signer was seated, corresponding to positions on an arc around the signer's body. The signer was asked to perform ASL verbs, e.g., EMAIL, with one target as subject and another as object. In this way, 42 examples of verb forms were recorded for each verb, for various combinations of subject and object locations. Because the verbs considered contained relatively straight motion paths for the hands, they were modeling using two keyframes (one at the beginning of each hand's motion path and one at the end).

Thus, the location (x, y, z) and orientation (*roll*, *pitch*, *yaw*) for the hands were extracted at each keyframe. (For signs with more complex paths, additional keyframes might be required.) This data was used to learn a model to predict the motion-path of a signer's hands for that verb, for novel arrangements of the subject and object on the arc around the signer. In prior work [8, 12, 13], two major types of modeling approaches were created for ASL indicating verbs:

Point-Based Modeling: This model [8, 12] predicted a starting location and the ending location of the hands for the verb, as distinct points in the 3D signing space; the virtual human animation software interpolated between these location points. Based on the position on the arc around the signer where the subject and object of the verb were located, the coefficients of six polynomial models were "fit" from training data for each for each hand (x_{start}, y_{start}, z_{start}, x_{end}, y_{end}, z_{end}), and, at run time, the models were used to estimate these values to synthesize a particular verb instance that was needed for an animation [8].

Vector-Based Modeling: The "point" model was not ideal: When different human signers perform a verb (e.g., EMAIL with subject at arc position on the left and object at arc position on the right), not all of the humans select exactly the same 3D point for their hands to start and stop. What is common across the performances is the *direction* that the hands move through space. Thus, in [13], a new modeling approach was proposed, called "vector" based modeling. Each verb was modeled as a tuple of values: the difference between the x-, the y-, and the z-axis values for the starting and ending location of the hand. Using this model, researchers followed a similar polynomial fitting technique summarized in [8], except that the model used fewer parameters. The "vector" model used only three values per hand ($delta_x$, $delta_y$, $delta_z$), instead of six per hand in the prior "point" model, which represent start and end location of the hand as (x_{start}, y_{start}, z_{start}, x_{end}, y_{end}, z_{end}). Of course, knowing the direction that the hands should move is insufficient: to create an animation, the starting and ending locations for the hands must be selected. At run time, a Gaussian mixture model of hand location likelihood (that had been trained for each ASL verb) was used to select the starting position for each hand (to identify a path that travels through a maximally-likely region of space) to synthesize a particular verb instance for an animation [13].

3. Novel Modeling Approaches

Limitations of prior work ASL verb modeling included:

- While hand orientation (*roll, pitch, yaw*) was modeled using artificially produced testing-data from a human animator in [8], researchers never attempted to model the orientation (*roll, pitch, yaw*) of the hands, based on a training set of motion-capture data *from humans*. Since hand orientation must be selected when producing an ASL animation, this was a major limitation of prior work.
- The "vector" model in prior work treated the left and right hands of the signer as completely independent motion vectors that needed to be selected. Section 3.1 will discuss how this led to low quality animation results for some verbs, and it will address this limitation.

Researchers had never before conducted a user-based evaluation (with native ASL signers viewing animations and answering questions) to compare the point-based and vector-based modeling approaches for synthesizing verbs. This paper presents the first user-based comparison of the quality and understandability of verbs synthesized by those two verb models, trained on motion-capture data from human signers. In addition to the conduct of the user-based study (section 4), another novel aspect of this paper is that we have enhanced and modified the Vector-Based Model, that was first described by [13], in several new ways, as described below.

3.1. Relative Hand Location Modeling

Some ASL verbs involve a movement in which the two hands come into close proximity or interact in a specific spatial orientation. For example, when performing the verb EMAIL as seen in Figure 1, the right hand must pass through the "C" handshape of the left hand. This close-proximity articulation of the two hands is essential for this verb's understandability. As another example, the ASL verb COPY requires the signer's two hands to come into close proximity at the beginning of the performance, as seen in Figure 2 and Figure 3.

Figure 2: Inflected version of ASL verb COPY with the subject on right and the object on left.

Figure 3: Inflected version of ASL verb COPY with subject on left and object on right.

There is a limitation in the original vector-based model, defined in [13]: That model did not explicitly represent the *relative* location between the left and right hands. It represented the direction of each hand's movement, with the starting location of each hand selected *independently,* based on the Gaussian model of hand location likelihood for that ASL verb. Applying such a technique to several examples of verbs such as EMAIL and COPY with specific hand proximity requirements, it was apparent that independently modeling the direction of both the left and right hands led to animations in which the relative positions of the two hands were not correctly preserved during the performance of the verb. For instance, for a verb like EMAIL, the right hand did not always move precisely through the opening produced by the left hand.

Therefore, we re-implemented and modified that original vector-based model, as follows: we model the left hand position *relative to* the right hand's position at each keyframe of the verb. At run time, we used our model to predict a hand movement direction vector for the right hand only. When we needed to synthesize a specific verb instance, we first selected a right hand starting location based on the Gaussian model. Then, we used our model of left hand relative-location to select a left hand location for each key-frame, *relative* to the right hand. Our new vector-based model, for verbs with two keyframes, would model nine values ($delta_x$, $delta_y$, $delta_z$) for

the right hand and ($relative_x$, $relative_y$, $relative_z$) for the left hand for each keyframe of the verb. In the prior "point" model, for a two-keyframe verb, there would be a total of twelve values modeled, the start and end location of both hands as (x_{right}, y_{right}, z_{right}, x_{left}, y_{left}, z_{left}). Given this new vector-based model (with the left-hand locations represented as relative to the right hand locations), we trained our enhanced vector-based models on the motion-capture data of ASL verbs that had been recorded by prior ASL animation researchers [14].

3.2. Modeling Hand Orientation

In prior work, researchers had not modeled *orientation* of the hands for ASL verbs using motion-capture data collected from human signers [13]. In this section, we present a novel method for modeling hand orientation (and an evaluation in section 4). Because there are various popular methods of representing the orientation of 3D objects (e.g., Euler angles, axis-angle, or 3x3 rotation matrices), we had to select an approach that was well-suited to representing hand orientation for modeling ASL verbs. Almost all orientation representations are actually representations of the 3D *rotation* of an object from a starting orientation; they all assume that a 3D object enters the universe with some initial orientation. They differ as follows:

- *Euler angles* represent a sequence of three rotations about the local axes of an object in 3D space. For instance, a first rotation about the object's z-axis by an angle α, a second rotation about its x-axis (which might have been affected by the first rotation) by an angle β, and another rotation about the object's z-axis, by an angle γ [3].
- The *axis-angle* representation is a rotation representation that consists of a unit vector <*x*, *y*, *z*> indicating an axis of rotation in a three-dimensional space and an angle *theta* indicating the magnitude of the rotation [3].
- A *rotation matrix* is another way to represent orientations of 3D objects; in this case, a 3x3 matrix can be used to represent a rotation. To rotate a point in three-dimensional space (represented as column vectors), you can multiply it by the 3x3 rotation matrix [3].

Since there are methods for converting between various orientation representations, we were free to select whichever representation for our modeling of hand orientation of ASL verbs. We wanted to select an approach with desirable mathematical properties. Specifically, we prefer methods of modeling orientation that avoid gimbal lock (described below) and were well suited to interpolation (meaning that when you numerically average the numbers that represent the orientation, the resulting 3D orientation of the object looks realistic). Techniques for computing representative orientations from measured 3D data have been described by several researchers, e.g., [5, 24], and the relative tradeoffs of many of these techniques have also been investigated, e.g., [25]. Some relevant considerations are summarized below:

- If we had used Euler angles, we may have encountered problems due to gimbal lock, a phenomena in which the first Euler rotation causes the axes of the system to align in such a way that a degree of freedom is lost [3].
- If we had used axis-angle representations, we may have encountered problems because axis-angle representations are not a unique representation of orientation (meaning that there are multiple possible ways to represent the same resulting final orientation of an object). Thus, there is no guarantee that simple interpolation of the numbers of the orientation representation will result in a realistic-looking 3D orientation for the final object (because the resulting orientation produced through interpolation may not be on the shortest path on the great arc between the two original orientations).
- If we had used 3x3 rotation matrices to represent orientation for modeling, this would have made our modeling more complex because this representation uses a large number of parameters (specifically, nine) to represent orientation.

For these reasons, we selected a less common method of representing orientations: Simultaneous Orthogonal Rotation Angles (SORA). SORA represents a rotation as a vector of three values (φ_x, φ_y, φ_z) that represent three *simultaneous* rotations around the coordinate system axes. (Euler angles represent *sequential* rotations.) SORA has been used in the areas of real-time angular velocities estimation [20]. The simplicity of SORA makes it possible for our orientation modeled in a single step, and avoids several of the problems with other approaches, outlined above. There are also standard ways to convert between SORA and other orientation representations [11, 20]. While [25] identify some limitations of SORA (similar to discontinuities encountered with axis-angle), we have found SORA to be an effective modeling approach for ASL verb orientation (as shown in Section 4.)

We performed our modeling as follows: First, we converted the motion-capture data into SORA format. Then, we trained the orientation models for all eight verbs (TELL, SCOLD, GIVE, MEET, ASK, EMAIL, SEND, and COPY). Since the rotation component for each axis can be isolated when using SORA, we consider the axes independently when we fit 3rd order polynomials to predict each component of SORA. Figure 4 outlines the procedure. At run-time, given some *s* and *o* values (i.e., subject and object location on the arc around the signer), we independently predict each of the values of φ_x, φ_y, and φ_z. After modeling each SORA value, we converted this back to axis-angle to synthesize a verb animation.

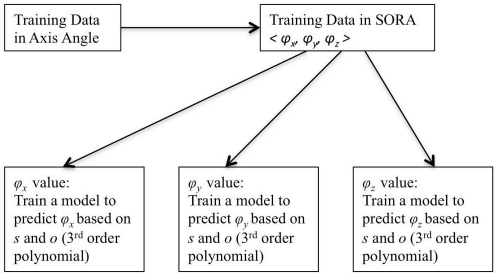

Figure 4: Training verb orientation data using SORA.

4. USER-BASED EVALUATION STUDY

A user study was conducted to evaluate animations synthesized by our point-based model and by our vector-based model, trained on the recorded data of the eight ASL verbs. The overall methodology of this study, including the recruiting practices, format of comprehension questions, and other details follows the general approach used in prior ASL evaluation research, e.g., [8]. Of the 24 participants, 13 had used ASL since infancy, 6 participants had learned ASL before age 8, and 2 participants began using ASL at a school with primary instruction in ASL since age 10. The remaining 3 participants identified as deaf, attended schools and

university with instruction in ASL, and had spouses or partners with whom they used ASL on a daily basis. There were 17 men and 7 women of ages 24-58 (median age 33).

The experiment consisted of two phases: In phase 1 of the study, we used a set of 12 ASL stories and comprehension questions that we designed and produced as stimuli. The stories and questions were adapted from those used in [8] for use in this current study; the stories were edited so that they included the eight ASL verbs listed in Table 1. The animations 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 12 stories and their questions were designed so that the questions related to information conveyed by a specific verb in the story. The comprehension questions were difficult to answer because of the stories' complexity, because participants saw the story before seeing the questions, and because they could only view the story one time. Each story was produced in four different versions, based on the form of the verb used in the animation:

- PointModel: inflected verb using our point-based model
- VectorModel: inflected verb using vector-based model
- Animator: inflected verb produced by a human animator
- Uninflected: uninflected citation-form of the verb

It is important to note that all of the animations presented were ***grammatical***, including the Uninflected stimuli. As described in section 1.1, verbs in ASL do not require spatial inflection during sentences, so long as the identity of the subject and object is otherwise indicated in the sentence. The animations presented in this study included in this information in the form of noun phrases or pointing pronouns in each sentence, identifying the subject and object. So, there were no non-grammatical sentences shown to participants in the study.

Section 3.2 mentions how the orientation model of the vector-based model is identical to the orientation model of the point-based model, so, the hand orientations in these two types of animation are identical – only the locations of the hands differ.

In this within-subjects study design:

- No participant saw the same story twice.
- The order of presentation of each story was randomized.
- Each participant saw 3 animations of each version.

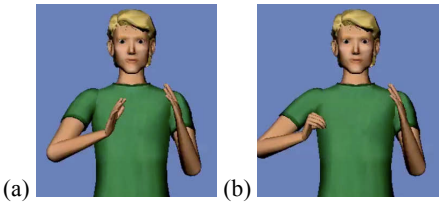

Figure 5: Example of ASL verb COPY produced by the vector model, as it appears in the study

Figure 5 shows example images for the verb COPY, produced by the vector model, as it appeared in a story during the study. In this example, the animated signer described a story in which several students (set up at locations in the signing space) were working on homework, and one student copied another student's homework. One of the comprehension questions for this story asked which of the students copied the homework.

Animation examples from this study may be accessed here: http://latlab.ist.rit.edu/slpat2015/

Table 1: Verbs collected in the training data set and which appear in the stimuli in study in section 4.

Verb	***Indicates***	***1- or 2-handed***	***Description of Movement***
ASK	Subject and Object	1	'ask a question': a bending index finger moves from Subj ('asker') to Obj ('askee')
GIVE	Subject and Object	2	'give to someone': hands move as a pair from the Subj ('giver') to Obj ('recipient')
MEET	Subject and Object	2	'two people meet': hands move from Subj and Obj toward each other, and meet somewhere in-between.
SCOLD	Object only	1	'scold/reprimand': extended index finger wags at the Obj ('person being scolded')
TELL	Object only	1	'tell someone': index finger moves from signer's mouth to Obj ('person being told')
COPY	Subject and Object	2	'copy from someone': right flat hand against left flat hand near Obj ('someone') moves toward Subj ('copier').
EMAIL	Subject and Object	2	'email to someone': right hand (bent-flat) passed through the cavity of the left hand (C shape) from Subj to Obj.
SEND	Subject and Object	2	'send to someone': a "B" hand with fingertips' quickly slide over the back of other hand, moving from Subj to Obj.

After watching each story once, participants answered 4 multiple-choice comprehension questions that focused on information conveyed by the indicating verbs. This study followed the methodological details of prior ASL animation research studies, as described in [8, 9, 11]. Figure 6 shows the comprehension question accuracy scores. A Kruskal-Wallis test (alpha=0.05) was run to check for significant differences between comprehension scores for each version of the animations. Only one pair of values had a significant difference (marked with a star in the Figure).

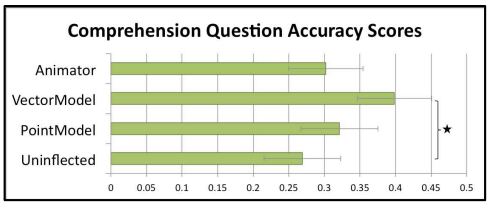

Figure 6: Comprehension question scores in phase 1.

In phase 2, participants viewed four animations of the same sentence side-by-side; e.g., "John point_to_arc_position_0.9 ASK Mary point_to_arc_position_-0.6." (Arc position 0.9 is on the signer's far right side, and arc position -0.6 is on the signer's left side.) The only difference between the four versions that were displayed on the screen was whether the *verb* in the sentence was: (a) synthesized from our point-based model, (b) synthesized from our vector-based model, (c) created by a human animator, or (d) 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 four versions shown at one time all used the same arc-positions). Participants answered 1-to-10 Likert-scale questions about the quality of the verb in each of the 3 versions of the sentence. Figure 7 shows the results. To check for significant differences between Likert-scale scores for each version, a Kruskal-Wallis test (alpha=0.05) was performed; significant pairwise differences are marked with a star.

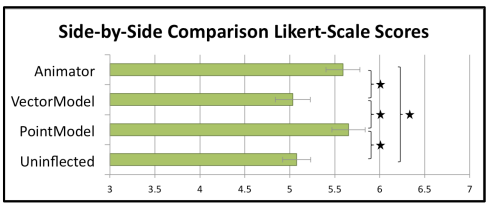

Figure 7: Subjective Likert-scale scores in phase 2.

4.1. Discussion of Results

For the comprehension question scores collected in phase 1 of the study, the vector-based model had significantly higher scores than the stories with the uninflected version of the verbs. This is a positive result because it indicates that the vector-based modeling approach led to more understandable stories. Prior work [9] has shown that comprehension-question based evaluation of animations is necessary to accurately measure the understandability of ASL animations.

For the subjective scores of animation quality collected during the side-by-side comparisons in phase 2 of the study, the animations containing verbs produced by the human animator received significantly higher scores than the uninflected animations. This was an expected result: the Animator animations were hand-crafted by a native ASL signer with proper ASL verb inflection movements, whereas the Uninflected animations were considered our lower baseline.

Similar to the Animator animations, our PointModel animations received higher subjective evaluation scores than the Uninflected animations. Verbs produced using this modeling technique received higher scores from native ASL signers. Uninflected verb animations are still used in many sign language animation systems; so, this indicates that our modeling technique is superior to that lower baseline. Because the 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), it was a positive result that the PointModel achieves this high score.

It is also notable that the PointModel received statistically higher subjective scores than the VectorModel, and the VectorModel did not receive statistically higher scores than the Uninflected animations. This result may indicate that there were problems with some animations produced using the VectorModel in this study. Figure 8 shows per-verb results from phase 2. It is important to note that none of the differences in Figure 8 were statistically significant; however, looking at this figure, we *speculate* that the VectorModel may have performed poorly for TELL and SCOLD. Among the verbs in this study, these two verbs are special, in that they inflect for object position only. (Their movement path is not modified based on where the subject of the verb is positioned on an arc around the signer.) Further, when human signers perform these verbs, their motion path is oriented away from the signer's chin (in the case of TELL) or heart (in the case of SCOLD). Since the VectorModel does not explicitly model the starting location of a verb (the location is selected based on a search through the Gaussian mixture model representing hand location probability), the VectorModel may lead to verb animations in which the starting location is somewhat inaccurate. For some ASL verbs, this may not have a significant impact on the perceived quality of the verb, if the overall direction of the verb movement is correct. However, for TELL and SCOLD, it may be the case that the beginning location of these verbs is very important for the correct production of the sign. For this reason, the vector model may not be appropriate for verbs of this type. Investigating the suitability of the vector model for different classes of ASL verbs, that have particular constraints on their starting locations, is an open area of future research.

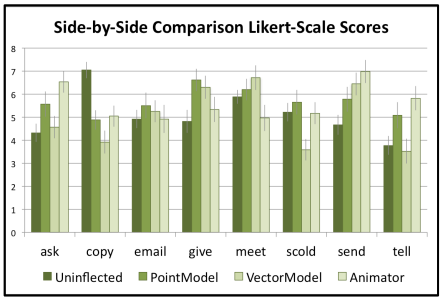

Figure 8: Per-verb results from phase 2 of the study.

5. Conclusions, Future Work

This paper has described our modeling methods and construction of a parameterized lexicon of ASL verb signs, whose motion path depends on the location in the signing space associated with the verb's subject and object. Specifically, we have described enhancements (representing hand orientation and relative location of the hands) to two prior state-of-the-art methods for generating ASL indicating verb animations (i.e., the point-based model and vector-based model of [8, 11, 13]). We have used motion capture data of sign language performances from native signers as a training data set for learning our models. In a user-based evaluation with 24 participants, we evaluated whether these models were able to produce more understandable ASL verb animations.

In future work, we intend to collect a larger set of recordings of ASL indicating verbs, including some with more complex movements of the hands, to evaluate whether the modeling techniques perform well for an even larger variety of signs. We may also explore how subject/object locations affect the signer's handshape during a verb signs: handshape was not affected by subject/object location in our current modeling approaches. We will study how the speed or timing of verb movements varies with the location of subject/object in the signing space. While our current work has focused on verb signs, we believe these modeling techniques may also be applicable to ASL pronouns and other signs whose movements are affected by the arrangement of spatial reference points in the signing space. Further, while this paper focused on ASL, we expect that researchers studying other sign languages internationally may wish to replicate the data-collection and verb-modeling techniques to produce models for signs that are affected by spatial locations.

6. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 0746556 and 1506786. Jonathan Lamberton and Miriam Morrow recruited participants and collected response-data during the user-based evaluation study.

7. REFERENCES

[1] Cormier, K. 2002. Grammaticalization of indexic signs: How American Sign Language expresses numerosity. Ph.D. Dissertation, University of Texas at Austin.

[2] Duarte, K., Gibet, S. 2011. Presentation of the SignCom project. In Proc 1st Int'l Workshop on Sign Language Translation and Avatar Technology, Berlin, Germany, 10-11.

[3] Dunn, F., Parberry, I. 2002. 3D Math Primer for Graphics and Game Development. A K Peters/CRC Press. 2nd edition.

[4] Emmorey, K., Bellugi, U., Friederici, A., Horn, P. 1995. Effects of age of acquisition on grammatical sensitivity: Evidence from on-line and off-line tasks. Applied Psycholinguistics, Cambridge University Press. 16(1):1-23

[5] Gramkow, C. 2001. On Averaging Rotations. Journal of Mathematical Imaging and Vision 15: 7–16, 2001. Netherlands: Kluwer Academic Publishers.

[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., Lu, P. 2012. Effect of spatial reference and verb inflection on the usability of sign language. Universal Access in the Information Society 11(2):169-184.

[8] Huenerfauth, M., Lu, P., 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.

[9] Huenerfauth, M., Zhao, L., Gu, E., Allbeck, J. 2007. Evaluating American Sign Language generation through the participation of native ASL signers. In Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility (Assets '07). ACM, New York, 211-218.

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

[11] Lu, P. 2013. Data-driven synthesis of animations of spatially inflected American Sign Language verbs using human data. Ph.D. dissertation, City University of New York, NY, USA.

[12] Lu, P., Huenerfauth, M. 2011. Synthesizing American Sign Language spatially inflected verbs from motion-capture data. In Proceedings of the 2nd International Workshop on Sign Language Translation and Avatar Technology (SLTAT), in conjunction with ASSETS 2011, Dundee, Scotland.

[13] Lu, P., Huenerfauth, M. 2012. Learning a vector-based model of American Sign Language inflecting verbs from motion-capture data. In Proceedings of the 3rd Workshop on Speech and Language Processing for Assistive Technologies (SLPAT '12). ACL, Stroudsburg, PA, USA, 66-74.

[14] Lu, P., Huenerfauth, M. 2014. Collecting and evaluating the CUNY ASL corpus for research on American Sign Language animation. Computer Speech & Language 28(3):812–831.

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

[16] 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.

[17] Meier, R. 1990. Person deixis in American Sign Language. In S. Fischer, P. Siple (eds.) Theoretical issues in sign language research. Chicago: U. Chicago Press, 175-190.

[18] 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.

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

[20] Stančin, S., Tomažič, S. 2011. Angle estimation of Simultaneous Orthogonal Rotations from 3D gyroscope measurements. Sensors 2011, 11, 8536-8549.

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

[22] 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.

[23] VCom3D. 2014. Homepage. http://www.vcom3d.com/

[24] Pennec, X., FIllard, P., Ayache, N. 2006. A Riemannian Framework for Tensor Computing. International Journal of Computer Vision 66(1):41–66, January 2006, Springer.

[25] Allgeuer, P., Behnke, S, 2014. Fused Angels for Body Orientation Representation. In proceedings of the 9th Workshop on Humanoid Soccer Robots, IEEE-RAS International Conference on Humanoid Robots (Humanoids), Madrid, Spain, 2014.