2016 49th Hawaii International Conference on System Sciences

Hands that Speak: An Integrated Approach to Studying Complex Human Communicative Body Movements

Nadir Weibel, So-One Hwang, Steven Rick, Erfan Sayyari, Dan Lenzen, Jim Hollan
University of California San Diego, La Jolla, CA 92093, USA
{*weibel, soone, srick, esayyari, dlenzen, hollan*}*@ucsd.edu*

***Abstract*—Gestures, the visible body movements that are ubiquitous in human behavior, are key elements of natural communication. Understanding them is fundamental to designing computing applications with more natural forms of interaction. Both sign languages and everyday gestures reveal the rich signal capacity of this modality. However, although research is developing at fast pace, we still lack in-depth understanding of the elements that create the underlying symbolic signals. This is partly due to lack of tools for studying communicative movements in context. We introduce a novel approach to address this problem based on unobtrusive depth cameras and developed an infrastructure supporting naturalistic data collection. While we focus on sign language and gestures, the tools we developed are applicable for other types of body based research applications. We report on the quality of data collection, and we show how our approach can lead to novel insights and understanding of communicative movements.**

***Keywords*-Sign Language, Gestures, Depth cameras, Kinect, Visualization, ChronoViz**

I. Introduction

People of all ages and cultures use gestures to communicate. In the case of sign languages of deaf communities, hands and body form the primary motor channel for language, serving as a medium for transmitting complex messages. From simple pointing to the rapid sequence of words in signed sentences, humans display a broad continuum of visible communicative movements, which in turn correspond to different representational properties [1]. However, little is known about how these different representational properties arise in human communication and how they can be recognized at the signal level [2]. This is in part because we have not had the scientific tools for studying the physical form of human gestures in a detailed and replicable manner.

The most common practice today in gesture and sign language research is to watch video records of activity and code for features by hand. While quantitative analyses can be conducted on annotations [3], these rely on the language background and expertise of coders for making judgments, do not involve direct measurements of the signal, and are implemented with varying temporal resolution. Motion capture allows direct measurements of the signals but is not widely used because costs are prohibitively high for most researchers, multiple sensors must be placed on the body in a recording studio, preparation time is required for calibration, and the systems are often confined to laboratory settings.

The development of recording and analysis tools that are cheap, portable, easy to set up, and allow unencumbered movements is key to enable a more natural study of gesture and sign language. Natural and spontaneous productions could then be easily captured, communication could be studied in diverse situations through fieldwork, and access to data from a wide group of researchers would lead to fruitful comparisons, standardization of analysis methods, as well as faster progress at addressing scientific questions.

We believe that todays' technology has the potential to create new instruments and analysis tools for understanding the cognitive basis underlying humans' use of gestures [4]. Our contribution in this paper is a novel approach towards understanding and quantifying both sign language and gestural communication in naturalistic settings. We outline our approach based on a fusion of sensor data captured by Microsoft Kinect and LeapMotion,[1] and their integration within an infrastructure for multimodal data collection, exploratory analysis and interactive visualization.

In this paper we describe the steps we have taken to develop a system of tools and interfaces to study these movements, which will make it possible to pursue multiple avenues of scientific inquiry about human cognition and language, as well as human-computer interaction (HCI) applications. As we outline both our challenges and successes, our focus is on reporting accuracy and adequacy of depth-based sensors for studying sign language and gestures, and outlining initial results and future directions.

II. Background and Related Work

The recent introduction of depth cameras and their availability in commodity devices that capture audio, video, and 3D scene information in real-time, offer exciting potential for advancing research in visual forms of communication and cognition. Devices based on depth sensors that can extract 3D information, such as the Microsoft Kinect or the LeapMotion, provides increased robustness for applications like skeleton tracking, hand localization, finger tracking, and gesture recognition, due to the invariance of depth data as compared to clothing differences, lighting conditions, and background clutter. Although this technology was designed for other applications, it has recently started to be used for sign language recognition on small scales such as for vocabulary of German Sign Language [5], sentences in the restricted domain of a small American Sign Language (ASL) vocabulary [6], and ASL fingerspelling [7].

[1] http://kinectforwindows.com, http://leapmotion.com

1530-1605/16 $31.00 © 2016 IEEE
DOI 10.1109/HICSS.2016.82

Other groups in the fields of sign language, machine learning, computer vision, and HCI [8]–[11] have used other tools for building automatic sign language recognition systems. However, most existing sign language corpora consist only of video recordings that do not capture all the important characteristics of sign language expressions. These datasets are often based on video data only, typically relying on very controlled recording settings where light conditions are kept constant and signers sometimes even have to wear colored gloves to aid segmentation of the data [10]. Special motion capture gloves used in those settings [9], [12] restrict freedom of movement [12] and are expensive [11]. To improve the quality of the collected data often a combination of recordings from multiple sources, necessitating multiple camera setups [13]–[15] and light sources [16], are typically required if researchers want to be able to capture 3D information. Finally, databases are restricted to limited domains and small vocabularies, and sometimes contain data from only a few signers [11], [17]. This is mainly due to the time consuming and expensive nature of collecting and annotating these data.

We believe that a systematic approach towards enabling high-quality, unobtrusive and high-resolution collection of sign language and gesture data will enable better understanding of the structure of those communication systems, their movement primitives, as well as the relationship or differences, across different languages and gestures. While making progress towards achieving automatic sign/gesture recognition, the scientific investigation of features across a continuum of communication systems will further the understanding of human cognition.

III. Approach

Our overall goal is to use the data capture capabilities of depth-sensors to develop techniques and representations for signal analysis of human motion used for communication. In order to accompish these goals, we extended two existing tools, ChronoSense [18] and ChronoViz [19]. The multimodal multi-device management provided by ChronoSense's single interface was extended to enhance body tracking data captured through Kinect sensor and integrate LeapMotion for synchronized and simultaneous hand and digit tracking that the Kinect alone could not provide. ChronoViz and its analysis and interactive visualization infrastructure was extended to accommodate gesture and sign language data in order to allow for visualization, exploration, and preliminary analysis of motion data all in one interface. This empowers researchers from a diverse range of fields who do not necessarily have the signal processing or computing background required to directly work with the data.

While we acknowledge that the systems we summarize are not yet sufficient for automatic sign language recognition, we emphasize their potential for a range of scientific investigations. By describing both capabilities and limitations of the systems, we can aid sign language and gesture researchers in identifying studies that are feasible with current technology.

In the following sections, we describe the instruments and applications we used for data capture from natural productions by signers and nonsigners who were asked to gesture. We detail how our system was evaluated and the types of analyses we were able to conduct based on the kind and quality of captured data. We illustrate how depth and RGB data must play complementary roles in the investigation of the formational properties of gestures and signs. We also illustrate how a combination of linguistic annotations and movement data accommodates analysis of more complex communication.

IV. Data Collection, Visualization and Analysis

Our goal in data collection was to capture natural productions of symbolic, communicative movements from individuals who are skilled in a sign language as well as those who have no experience learning a sign language. The data that we report here include samples of 9 signers (6 signers of ASL, 1 signer of Japanese Sign Language, 1 signer of Al-Sayyid Bedouin Sign Language – an emerging village sign language found in Israel [20] – and 1 signer of German Sign Language, who performed the task in International Sign – a pidgin sign language that is often used to communicate with audiences of diverse sign language users [21]) and 9 nonsigning gesturers.

Since we were primarily interested in comparing these two samples and identify distinguishing features, while eliciting natural production, we also sought to minimize the risk that group differences would arise due to differences in the topics discussed rather than differences in experience of using hands and body for communication. To achieve this balance, we asked all the participants to watch and describe the same story. The participants watched a 6-minute animated film with three characters and no dialog and were then asked to narrate the events in the story as it was unfolding when played for a second time. We captured approximately 6 minutes of continuous data from each participant.

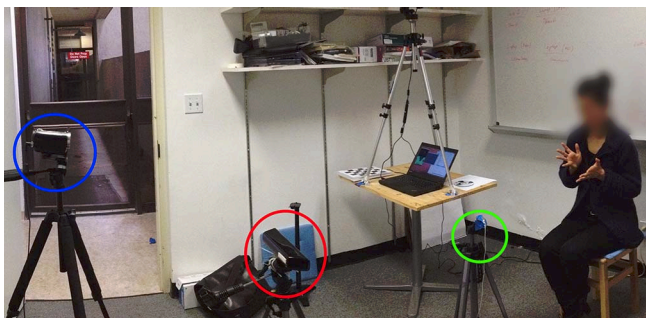

Figure 1. Experimental setup: In red on the bottom is the Kinect, recording body activity of our participant. In green, bottom-right is the Leap Motion device, recording hand activity. In blue on the left, the HD video camera.

Signers were asked to perform their task in their respective sign languages, and nonsigners were asked to improvise gestures to convey the story without speaking. All participants were instructed to stay seated during the task. In particular, this instruction limited the gesturers to using only their hands and upper body, similar to the articulation of sign languages. As described later in this section it was important to assess whether the quality of data recording differs between the two groups of participants. For example, the formational differences we expected to find may also result in less reliable data capture for one of the groups, potentially limiting the utility of instruments for studying communicative movements. We did not find differences in percent data captured.

In the remainder of this section we outline our approach for collecting high-resolution data from communicative movement of signers and nonsigners, including description of our recording instruments, evaluation of data collection performance across instruments, data processing and optimization, and data visualization for exploratory analysis.

A. Recording communicative movement

With the intent to capture as much information as possible about how our participants were using their bodies and hands to communicate naturally, we deployed multiple sensors in the environment. Our infrastructure included a Kinect for Windows (V1), a Leap Motion sensor, and an HD video camera. Figure 1 shows our setup as we are capturing sign language from an ASL participant. We placed Kinect cameras in front of the participants to capture their body motion. Additionally, we placed the Leap Motion sensor right in front of the participants' hands to get detailed movements of the fingers, while signing. The HD camera gave us a holistic view of the recorded experiments, an important element to validate the sign and gesture capture that we aimed to do through the depth sensing devices.

Data capture – We extended our ChronoSense application from earlier studies with Kinect in the wild [18]. By combining the existing Kinect SDK with the Leap Motion SDK in a single application we enable simultaneous and synchronized recording of body, hands and finger motion. The Kinect device projects a structured Infrared (IR) light array into the environment that is then assessed by the Kinect's IR camera to understand the 3D spatial positioning of everything in the scene. This enables understanding of objects and body movements, based on their distance (or depth) from the camera. Additionally, Kinect uses this IR depth data to analyze humans within view of the camera and construct real-time skeletal representations [22]. At the time of our experiments, we were using the Kinect for Windows V1 and it's associated v1.8 SDK. Even though this SDK reports 3D location data for 20 joints across the whole body, which ChronoSense records, our emphasis on human communicative body motion prompted us to narrow our focus to the upper body joints including the head, shoulders, elbows, wrists, and hands.

Our front-facing Kinect captured details related to larger scale body motion and data about hands movements relative to the body, but nothing related to the individual fingers. Given the importance of fingers for sign language we integrated the Leap Motion devices and its SDK. Integrating the Leap Motion directly into the ChronoSense tool enabled simultaneous tracking and recording of body, as well as palms and digits of the hands. Similar to the Kinect, this device uses IR light and two IR cameras to produce a 3D view of the scene above the device based on its cameras. The device's small size and limited IR range reduces the scope to only the detection and processing of hands. As such we are able to use the device to track and record the 3D position and orientation of the palm and up to five fingers of each hand. At the time of our experiments, we were using the Leap Motion v1.2 SDK.

To isolate the movement of our participants and reduce the chance of full-body pantomiming when tasked to silently communicate, we had all of our participants sit down prior to being presented with any tasks or stimulus. The Leap Motion device was attached to a tripod that was placed in front of the participant's legs and angled to point up towards the chest of each participant, the signing space. Along with the front facing Kinect, we were able to capture two different forms of bodily activity unobtrusively, at high resolution.

B. System Performance

To better understand the efficacy of our data collection methodology, we measured the performance of the single instruments in terms of captured data. We focus on percentage of body data (joints, hands, and fingers) captured through the different approaches outlined above. We developed a script that measured dropped data from each recorded body tracking and hand tracking, grouping the results by sensor type. Each row of each column of data was categorized as either having data present or not, and then a percentage of data presence versus lines read was generated.

Kinect – Over our 18 participants, we observed a 50%-55% capture rate for either hand, 79%-87% capture rate for either wrist, and 99% capture rate for a body being present as detected via the Kinect camera. This lower capture rate for the hands compared to the wrists and the upper arm is tied with the randomized decision forests implemented in the SDK [22] that procedurally classify joint detection down the limb starting from the shoulder. This prompted our decision to focus on gestural activity from wrist data points instead of hands, considering that these joints are so close in proximity.

Leap Motion – Next, we analyzed the same sample of experiments and participants with the Leap Motion device, observing a 48% rate of capture for one finger, 10% for two, 3% for three, 1.5% for four, 0.4% for all five and a 99% capture rate for a hand being present and detected at

all. So while hands were detected at a high rate, the natural gesturing and signing that our participants were performing did not allow for regular capture of more than one finger.

The dropout rate of hand joint from the Kinect camera is most directly connected to occlusion of the hands during complex signing or gestural movement, or depth segmentation issues that arise when hands are very close to or touching the body. For the Leap Motion device, the dropout rate for the individual finger tracking is most directly connected to the complex orientations of the hands and occlusion of fingers during signing and gesturing. The device operates best within a 2 meter dome from the center of the device,[4] and was developed to track hands oriented in a plane that is perpendicular to the device's cameras' directions, parallel with the surface of the device. Since natural signing and gesturing is not restricted to this space the device showed a reduced performance.

When adjusting for momentary gaps in the data, as explained in the Data Consolidation section below, the dropout rate decreased. Kinect improves to 60%-64% capture rate for hands and 86%-92% for wrists, Leap improves to 53% capture rate for one finger, 13% for two, 4% for three, 1.6% for four, 0.4% for all five fingers.

With these observations, we establish the degree to which communicative hand and body activity can be captured with our current array of technology. We also note that out of our assortment of sensors and data, gross body movements were most successfully being recorded as compared to hand and finger specific activity, especially when participants were left to use complex, natural, and unrestricted motions. With this in mind we began to explore how these data can lead to insights about how signers and nonsigners communicate through bodily motion.

C. Data Processing

Although the data collection performance outlined above provides enough data to allow researchers to perform analysis, the data recorded by the depth sensors often presents missing data, is incomplete, and is difficult to align across participants and experiments. In order to increase our analysis power and the flexibility and applicability of our approach, we process the data and perform three distinct operations: data recovery, de-noising, and data normalization.

Missing Data Recovery – Due to the naturalistic nature of the data collection, and the free movement of participants, when we look at the data frame by frame, the skeleton data generated by Kinect has many missing joint positions. This is more severe for the joints that are articulating the most, like signers' and nonsigners' hands. In order to enable both accuracy and enhanced precision at analysis, we aim to intelligently fill as many missing points as possible during data processing and include them as body joints in the data.

To do that we start by considering the concept of a *skeleton run*. We define a run as series of adjacent missing values or a series of adjacent known values. The number of consequent missing values, or known values is the *length* of the run. We calculated missing runs across all joints in our dataset and discovered that a large portion of missing values are part of runs of less than 0.5 seconds. Given the importance of the wrist for our specific analysis, we focus on wrist data here as an example. If we consider wrists while collecting body joints at a frame rate of 30Hz, and knowing that on average at most two signs might be performed within one second in natural production [23], 12 skeleton frames are less than one complete sign. From our experimental results (Fig. 2) we see that on average 61% of runs are of length less that 12. So choosing 12 as the maximum length of missing runs to get corrected, yields around 61% of the missing runs being part of a complete gesture. This means that most of our corrections would not span multiple gestures.

The realization that missing runs are mostly short, allows us to use an interpolation algorithm to fill in the missing values, without compromising the integrity of the dataset. We use the MATLAB function `Inpaintn` for this task [24]. This function minimizes the difference between the estimated values and the true values for known data. Smoothing penalty due to the processing of unknown missing data is achieved by utilizing a Laplace operator on top of the estimated values iteratively. This minimization was implemented using DCT (Discrete Cosine Transform) efficiently [25]. Mathematically the objective function, $G(\hat{X})$, is then written as:

$$G(\hat{X}) = ||W^{1/2} \circ (X - \hat{X})||^2 + s||\nabla^2 \hat{X}||^2 \qquad (1)$$

Where W is a matrix that indicates which element of X is missing, $\circ$ is an element-wise product operator, and ∇^2 is the Laplace operator.

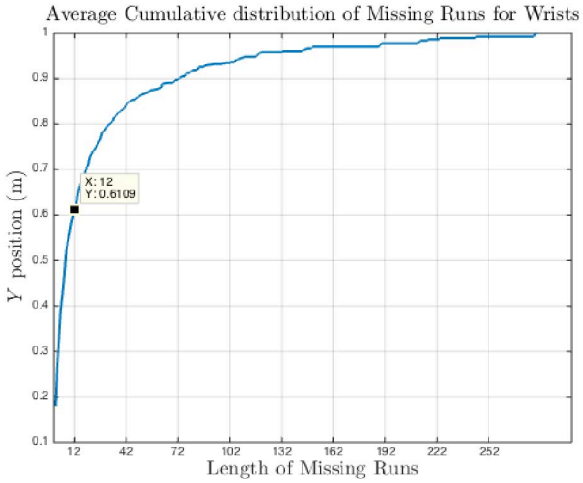

Figure 2. Average cumulative distribution for wrist missing runs. Around 61% of runs are less than 12 in length. We use this value for interpolation to balance effective correction and data completeness.

Best performance for this algorithm is achieved when the penalty and Laplace operator coefficient s is small [25]. We experimentally calculated the smallest possible value for s to be 10^{-3}. Algorithmically our data recovery process can be summarized as follows, where len is the maximum run length that could be retrieved by the algorithm. Based on our experimental results (see above) we set len as 12:

Algorithm 1 Data Recovery Algorithm

1: **procedure** FILL MISSING($Skeleton$, len)
2: **Set zero for missing and one for known values**
3: **Extract the runs and their length**
4: **Specify runs with length More than** len
5: **Run function** $Inpaintn$ **to interpolate missing values, the output is** $Skeleton_{new}$
6: **Set back Not a Numbers (**NaN**) for missing runs with length more than** len **to** $Skeleton_{new}$
7: **return** $Skeleton_{new}$
8: **end procedure**

Smoothing/Denoising – In addition to incidences of missing points, the skeleton generated by Kinect is also noisy [26]. To solve this problem we introduce a de-noising step that reduces the joint jitter. Although any conventional de-noising algorithm could be applied [26], a multivariate de-noising scheme was used because of the multi-dimensional nature of the skeleton data. In order to implement de-noising we use the MATLAB function `wmulden` [27], combining a generalization of univariate wavelet de-noising algorithm and Principle Component Analysis (PCA) [28], [29].

Our de-noising step is therefore defined as follows:

1- Assuming that the skeleton positions at each frame is $P-dimensional$, first perform the wavelet transform on each dimension of skeleton positions up to scale J.

2- Estimate noise properties and covariance matrix, $\hat{\Sigma}_{\varepsilon}$.

3- Compute Singular Value Decomposition (SVD) of the noise covariance matrix, $\hat{\Sigma}_{\varepsilon} = V\Lambda V^T$. Each (D_j) matrix is transformed with respect to the matrix V and then the $P-dimensional$ threshold is applied on top of (D_jV).

4- Find the PCA on the estimation coefficients A_J and select the $PJ+1$ as best principle components.

5- At last by inverting the transformation of the basis by V (multiplying by V^T) and wavelet transformation, reconstruct the de-noised version of the skeleton.

For de-noising, the wavelet transformation is computed up to $J = 10$. We use all of the PCA components for reconstruction. The J parameter helps avoiding washing skeleton information out, while decreasing jitters visually.

Combining consolidation and de-noising results in smoother overall joint data without interruptions that can then be further analyzed. Figure 3 shows the result of consolidation and de-noising on 20 seconds of the Y component of the right wrist of one of our participants. The original data is represented by a blue curve, the consolidated data is represented by the red curve, and the smoothed version is the green curve. An example for the resulting consolidated and de-noised skeleton is shown in Fig. 4.

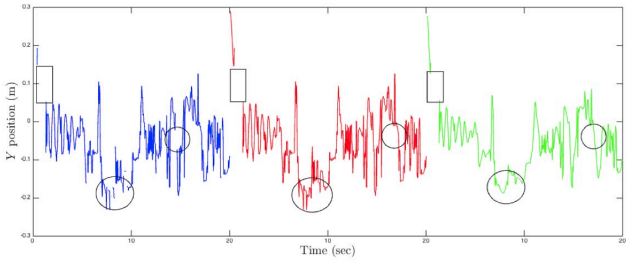

Figure 3. Combined consolidated and de-noised wrist data for Y component. Left: original data in blue. Center: consolidated data in red. Right: fully cleaned (i.e. consolidated and de-noised) in green. Outlined are areas where it is possible to see the effects of our data processing.

Normalization across Individuals – Even after data consolidation and de-noising, data issues related to the heterogeneous nature of the participants and their positioning with respect to the cameras remain. Individuals show variation in body size, and for each recording session, the center of the *world coordinate* (i.e. the center of the coordinate system used to measure joint positions) might be different as well. Because our recording environment allowed for flexible arrangements, the distance and angle to the camera was susceptible to variation, which in turn impacts the perceived body size and makes analysis more difficult. These are issues that need to be addressed when recording in natural environments and have high relevance for fieldwork conditions.

To solve this problem, we apply a process of normalization. The main step in normalization is to unify the center of the world coordinates. For this step, skeletons are shifted so that the coordinates of the *shoulder center* joints are positioned at the zero position for all participants. The next step is to represent skeletons as trees with the *shoulder centers* as the root. Since bone orientations, the interpolations between joints, are informative, a proper normalization algorithm should preserve bone angles and adapt their sizes. We apply the algorithm proposed in [30], that starts from

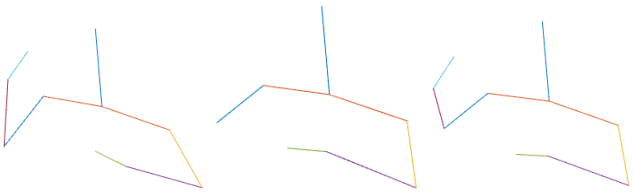

Figure 4. Result of consolidation and denoising process for a skeleton with a missing right hand. Left: complete skeleton, before the hand went missing. Center: subsequent frame, where a hand is cut out due to missing and noise data. Right: skeleton with reconstructed hand after cleaning process (same frame as the center skeleton).

the root of the tree and iteratively changes joint positions such that length of the bones be equal to average bone size. Mathematically, this is expressed as:

$$v_{init}^{(j-1,j)}(t) = \frac{S_{init}^{(j)}(t) - S_{init}^{(j-1)}(t)}{||S_{init}^{(j)}(t) - S_{init}^{(j-1)}(t)||} \quad (2)$$

$$S^{(j)}(t) = S^{(j-1)}(t) + v_{init}^{(j-1,j)}(t) \times b_{avg}^{(j-1,j)} - S_{init}^{(root)}(t) \quad (3)$$

where $S_{init}^{(j)}$ is the current joint, $S^{(j-1)}$ is its parent joint in tree, and $b_{avg}^{(j-1,j)}$ is the estimated average bone size.

D. Data Analysis and Visualization

Cleaning and normalizing data is a required step to prepare for analysis. However, tools to aid coding and analysis of movement data are also required. Those tools are key to understanding the dynamics and complexity of sign language and gesture. In this work we build on ChronoViz [19], a tool that already facilitates annotation, navigation, and analysis of multiple streams of video and other time-coded data, and we integrate visualization of body movements to facilitate analysis. These tools allowed us to explore the scientific application of understanding the differences in the dynamics and complexity of sign language and gesture.

The data we collected with our infrastructure provides joint estimation 30 times per second, from which we can recreate body position at any given moment. We are interested in the dynamics of communicative movements. Instead

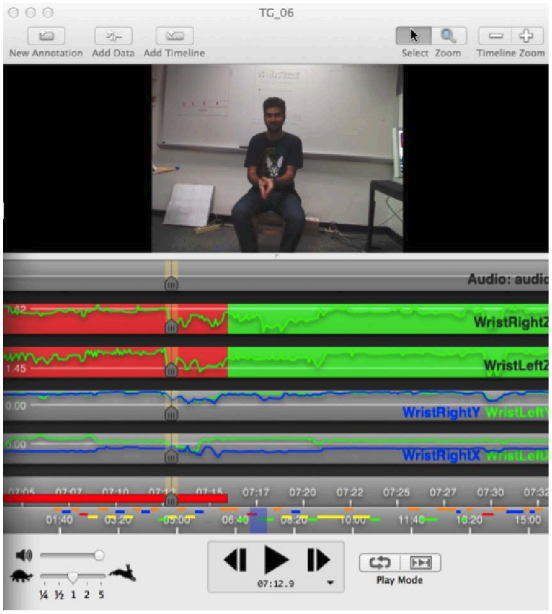

Figure 5. ChronoViz session showing RGB frames of a recorded gesture (top), and the single (X,Y,Z) components of the two wrists as line charts on dedicated timelines (top-down: Z component of right and left wrists, combined Y component of both wrists, combined X components of both wrists). Annotations show particular segments of interest.

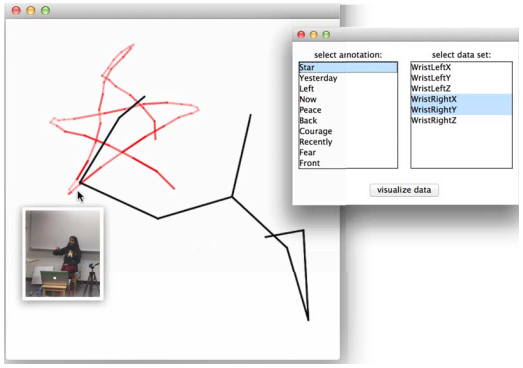

Figure 6. ChronoViz gesture/sign visualization plugin. Researchers can visualize movements of specific gestures over time. The selection box allows researchers to pick a specific annotation (e.g. the segment corresponding to the word *"Star"*) and plot the movement of particular joints of interest. Opacity of the line segment indicates speed and a tooltip displays the RGB frame of the current movement. Data can be superimposed on the real-time skeleton movements (as in this Figure) or on a silhouette representing the normalized participant's position (as in Fig. 7).

of analyzing joints one at a time, we visualize particular joints over the entire recording time and provide facilities for exploring particular dimensions of the movements as part of the ChronoViz exploratory data analysis tool. Figure 5 outlines gesture data in ChronoViz, while Fig. 6 shows our integrated visualization plugin.

The interactive visualization of body movements shown in Fig. 6 connects joint positions over time, with an averaged skeleton providing body context and highlighting the typology of the gestural response for that recording session. Although meaningful gestures occur only occasionally, when gestures are identified, this typically leads gesture researchers to focus on brief segments of motion for a long time, with the goal to understands the basic primitives behind the inspected motion. Our plugin allows users to zoom and filter onto meaningful segments of the collected data after these meaningful gestures have been identified. The plugin handles ChronoViz annotations directly, allowing researchers to identify segment of interest, create ChronoViz annotations, and then use those annotations to drive analysis. To support the larger researcher community, we also support time-based annotations imported from ELAN [3] a common annotation tool used by behavioral researchers.

Given our focus on sign language and gestures, we focus on hands and finger interactions, and highlight those details that are most important in this field [31] across three main parameters: (i) location of the hand, (ii) handshape, and (iii) movement. As outlined earlier we use wrist estimations as a proxy for hands movements. Figure 6, illustrates the location (or place of articulation) of the gestures, we plot the coordinates of the wrist collapsed across time, and connect them following the temporal order. Furthermore, we encode speed as the opacity of the path, which results in richer color for the fastest section of the gesture, typically considered to be the most meaningful portion of the movement. Finally,

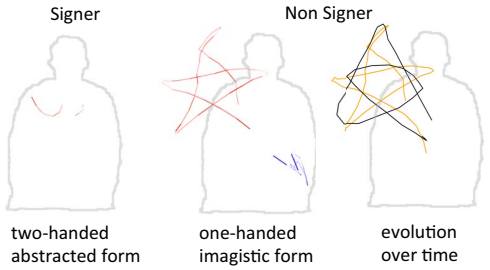

Figure 7. Comparing gestures and sign language. Left/Center: representations of a star across a signer (left) and a nonsigner (center); the movements of the two hands are indicated in blue and red respectively. Right: two gestures from the same nonsigner, both tracing the shape of a star, from different times in the session; the trace in black is a faster movement (1.3s vs. 1.9s) with a less defined shape; this illustrates for instance how gestures can evolve over a discourse session.

we link information about the handshape at different times of the gesture by providing a dedicated tooltip showing the RGB frame collected at the time-point nearest to the place the mouse crosses the path. Note that the depth information captured by our infrastructure provides Z-axis (towards/away from the camera) data for all of the joint estimates, allowing researchers to exploit our visualization facilities also on the sagittal axis. This is important since to date the difficulty of determining depth with traditional two-dimensional RGB video setups prevented gesture researchers from studying forward-backward aspects of gesture in detail. Our tool and visualizations can provide detailed information about this axis without additional cameras or sensors.

Finally, our visualization, allows researchers to compare gesture features across time or participants. Given our normalization step, collected data can be aligned independently of the data collection conditions. As shown in Fig. 7 (right), by overlaying two gestures from the same participant at different times researchers can visually compare the differences, and, when the meaning of the gesture is the same, compare how gesture evolves over time. Figure 7 (left/center) highlights how researchers can compare movements across participants.

V. Validation

In order to validate our approach we sought to find convergence between observational reports on gesture and sign language productions and the measurements we collected from the two groups. Before sign language research [32]–[34], it was widely assumed that sign languages are not natural human languages, but rather made up simulated actions and pictorial gestures. Demonstration of similarities of grammatical properties to spoken languages and differences from the gestures of nonsigners, along with evidence from neural processing and development, paved the way to the acceptance of sign languages as rich linguistic systems.

One approach to investigating the linguistic properties of a communication system is to identify the set of constraints and rules for how phonological features and words are combined. Unlike the relatively unconstrained use of the space around the body in pantomime/gesture production [32], the articulation space of the hands during signing is known to be smaller near the chest and face. Using ChronoViz, which allows for the customized selection and visualization of joint location over time, we were quickly able to validate that the measurements from our samples replicate such a pattern. Based on the pilot inspection that ChronoViz enables, we pursued quantitative analysis of spatial differences between signing and gesturing. One approach was to determine the proportion of time the hands/wrists fell in the shoulder-to-shoulder cubic space in front of the chest.

Beyond the categorical analysis of whether the hands/wrists fall within a pre-defined space, we also pursued richer descriptions of the spatial distribution of hand/wrist positions over the span of narrating a story. Figure 8 depicts a density plot showing the 3D position of the wrists of an ASL signer respective to their body. Density is represented by richness of the color. Figure 9, shows the same but for a nonsigner. As outlined in the two figures, the density of activity for a signer in the 3D space in front of their body is much larger than that of a non-signing participant. However, because nonsigners paused more in their narrations, they placed their hands/wrists at rest position on the lap. With our visualization tools it was easy to exclude those positions and check to see if the larger density of activity for nonsigner was an artifact of differences in articulation time.

We therefore limited the data analysis to the moments when participants were producing a meaningful segment, which were coded through time-aligned annotations. Even so, signers' articulations fell within a much more constrained space than nonsigners' gestures. While the spatial range of the movements are comparable, as constrained by physiological factors, graphs of distributions show differences in relationship of the hands in signers and nonsigners.

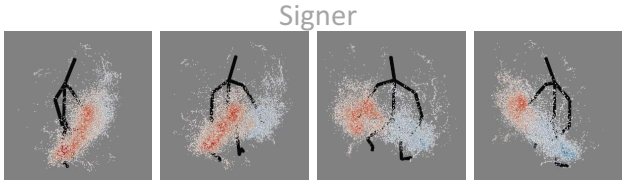

Figure 8. Density plot of ASL signer wrist movements during our experiments. The skeleton shown is the average position of all their body joints across a single recording session. The right wrist is represented by red dots, blue by the left wrist.

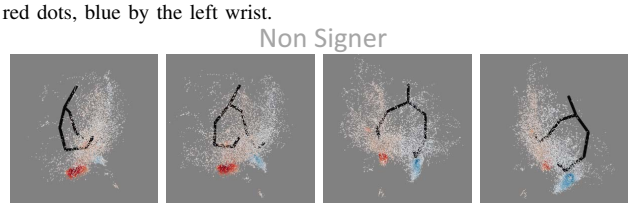

Figure 9. Density plot of nonsigner wrist movements. Note the much larger and less focused used of the space in front of the body needed to communicate and express concepts.

Figure 10 shows two plots of the same data with vertical and horizontal histograms breaking down spatial activity of the left and right wrist in two directions, for ASL signers and nonsigners respectively. The distributions suggest that there is a greater degree of lateral asymmetry in nonsigners than signers at the signal level. Analyzing sign language distribution in the three dimensions we see two equally unimodal and somewhat normal histograms representing activity for each wrist along the X axis. In comparison the same plots from a nonsigning subject show a strong single handedness as presented by the very uneven histograms in the X axis. The density of activity, represented by the brightness of the color, is much more widely distributed in front of the signer body with the nonsigner producing a smaller area of high density near the bottom of the plot.

Linguistic annotations on videos revealed that a significantly higher proportion of meaningful segments produced by signers were one-handed compared to nonsigners' productions. The spatial distribution of hand/wrist positions, collapsed for time, does not reflect the greater lateral asymmetry found in signers because even when the nondominant hand is not an active articulator in the production of a word, it is still kept within the primary articulation space. Nonsigners, in contrast, drop the inactive hand to resting position. Linguistic annotations validate the expectation that signers produce meaningful segments at much higher rates than gesturers [35]. At the signal level, we can see that this is correlated to the use of a more constrained articulation space that results in keeping both hands in this space to allow for the rapid transition between one- and two-handed signs. These results demonstrate the complementary roles of signal recordings and linguistic annotations for understanding communicative activities.

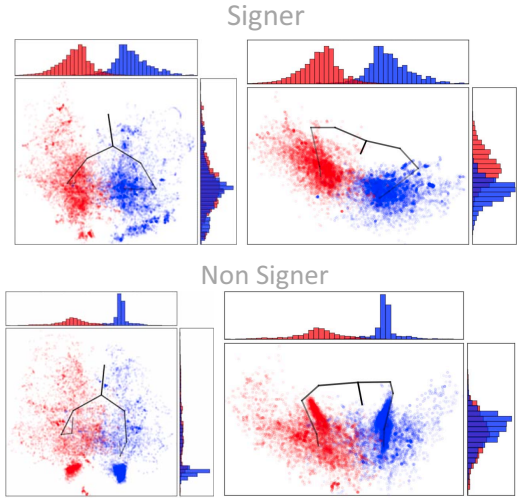

Figure 10. Spatial distribution of signers (top) and nonsigners (bottom) in the 3D space in front of them. The left plot shows the spatial distribution viewed from directly in front of the participant; the right plot shows the same data viewed from above looking down. Activity density is represented by the brightness of the color. Black lines represent the participant's average skeletal orientation over the course of our recording session.

Finally, in addition to gross-level analysis of the spatial data, we also validated system performance by examining the spatial trajectories associated with specific meaningful units. For example, stars are prominent in the animated film that the participants were asked to follow and describe. Nonsigners often represent a *star* by tracing the common five-pointed shape, whereas signers use more abstracted and less imagistic forms that are part of their language's lexical inventory. As shown in Figure 7 (left/center), our data shows trajectories that match what we can observe from videos. In this example, the nonsigner uses a one-handed imagistic form (with the left hand at rest), while the ASL signer produces the lexical sign, which is a two-handed form involving short, up-ward movements.

Using distributional data together with annotations, we plan to conduct more in-depth analysis to discover the set of location states for the hands while signing. These location states found among ASL signers are predicted to match descriptions of ASL phonology, where hand location is a contrastive phonological parameter, along with handshape, movement, and non-manual features. For example, articulations involving the same handshape and movement that are articulated at the forehead, nose, or chin, can result in three unrelated words [31]. By identifying specific signal segments that correspond to lexical sets using contrastive locations, we will be able to assess whether our system has sufficient spatial resolution for automatic recognition applications. In the future, we can take the continuous depth and annotated data together to identify signal features that correspond to boundaries between separate meaningful units, and compute the articulatory variability in producing the same meaningful unit within and across individuals. Machine learning techniques will then allow us to develop automatic segmentation and recognition.

VI. Discussion, Contribution and Future Work

In order to study the dynamics of natural human communicative movements and develop automatic sign/gesture recognition, systems for data capture should be designed to have the capability of tracking the multiple components that sign language researchers have identified as being critical for distinguishing meaning, while also being maximally portable. Portability is important not only for conducting fieldwork throughout all parts of the world where sign languages are found, but also for collecting large, natural data sets for machine learning applications. Additionally, analysis tools should be user friendly so that quality of data capture can be quickly validated and patterns in the signal can be intuitively ascertained even by researchers who do not have signal processing and computing backgrounds.

The example outlined in this paper demonstrate the capabilities that can be achieved with the integration of our tools as well as current limitations. Below, we summarize the approaches we have been taking to address these limitations.

We hope this information will aid not only sign language and gesture researchers but also other researchers in identifying studies that are feasible given the current technology.

Unlike previous approaches that collected limited sets of vocabulary items or sentences [5]–[7], we recorded natural productions of story narrations from both signers and nonsigners. The quality of data capture in these two groups was comparable. Although the rate of data capture using Leap Motion was not reliable enough to track hand configurations throughout, we were able to find distinctive features between fully linguistic communication from the gestures of nonsigners using skeleton tracking. Continuing this approach, making improvements on hand configuration detection, and adding features such as facial expression recognition will open the way to unprecedented analysis of communicative movements.

The richness of the data we collected allows researchers to investigate a variety of dimensions of the spatial and temporal dynamics of sign and gesture. While in this paper we only focus on one particular perspective, the tools we introduce are general and allow tackling a variety of other questions. The ability to automatically quantify and visualize movements, even the simple aspects depicted here, is an important first step for bootstrapping analysis. Findings from this project (1) confirm that the quantitative data collected within our infrastructure match the findings from separate annotation work we have done to analyze the communicative efficiency and representational differences between signing and gesturing and (2) converge with previous reports on sign language as being a constrained, rule-governed system that cannot be explained just by physiological limitations [32]. Our data show that sign productions are articulated in a more constrained space than pantomime gesture and that the coordination of the two hands differ as well. This initial outcome will lead us to explore questions about changes in signals that reflect linguistic organization, which in turn will inform theories on language evolution [36].

As evident from our initial findings, we envision the infrastructure, the methods and the tools that we present in this paper to be able to address, among others, fundamental questions around rhythmic features of communicative movements at multiple time-scales and investigate comparisons between sign language/gesture and speech, to ultimately lead to a better understanding of how these rhythms are coordinated and channeled through multiple avenues of the body for human language. Our goal here was to describe a methodology and a set of tools to aid researchers in better understanding communicative movement and designing systems to exploit this important form of interaction.

With the improvement of sensing technologies and software, we are actively improving and simplifying our capture system. We recently integrated the new Kinect for Windows V2 that provides a new higher resolution depth sensing camera that enable more robust segmentation and tracking. The new device can now track up to 26 joints, expanding to include thumb and fingers (as a single unit, not individual digits). The time-of-flight camera technology reduces interference from external sources of IR as compared to the structured IR system in V1. Additionally, the new Kinect V2 has been redeveloped at the hardware driver level from a single-application based system to a device that can provide the same raw data stream to multiple applications simultaneously. Kinect V2 also now has a HD camera with 1920x1080 color resolution. Previously the 640x480 resolution of the V1 Kinect color camera was much too low, requiring the use of an external HD video camera to capture gesturing and signing at a high resolution. These improvements provide new and enhanced capabilities while reducing the amount of hardware required to capture the same quality of data.

The Leap Motion device has also made strides forward through development of its SDK. The new V2.2 SDK now provides more reliable tracking, as well as useful information about the hands and fingers that are currently being tracked. Previously we could only capture information about a hand and its tracked fingers, now the SDK provides more meaningful information such as indication of left or right hand, as well as better labeling of the fingers being tracked as thumb, index, middle, etc. The new SDK has expanding tracking capability to better constrain hand shape to a model based on physical anatomy of the hands, which helps the device report more realistic positional data and allows for compensation of occlusion.

VII. Conclusion

In the history of science, new technologies for capturing and analyzing data have often led to significant scientific and practical advances. Depth cameras, like those used in the Microsoft Kinect and Leap Motion which can capture the 3D dynamics of human action in real-time, are changing the landscape of research. Coupling this technology with advances in computer vision and machine learning is creating exciting avenues for research and application not previously possible. However, to truly leverage these naturalistic modalities of human-computer interaction we need to understand how gesture is naturally used and design applications accordingly. Our contribution to this vision is a first step towards computational understanding of sign language and gesturing. We demonstrated how the fusion of multiple depth-based sensing devices promise to aid in the development of richer understanding surrounding gestural forms of interaction. Moreover, tools like ours which are sophisticated enough to accurately capture and represent sign language and other gestural signals will have broader applications for studying other types of embodied activity. Study of natural behavior in real-world settings has the incredible potential to yield a richer understanding of communication and cognition, and to better inform the design of new systems that seek to leverage such modalities.

REFERENCES

[1] S. Goldin-Meadow, "Talking and thinking with our hands," *Current Directions in Psychological Science*, vol. 15, no. 1, pp. 34–39, 2006.

[2] S. Goldin-Meadow, A. Shield, D. Lenzen, M. Herzig, and C. Padden, "The gestures ASL signers use tell us when they are ready to learn math," *Cognition*, vol. 123, no. 3, pp. 448–453, 2012.

[3] H. Lausberg and H. Sloetjes, "Coding gestural behavior with the neuroges-elan system," *Behavior research methods*, vol. 41, no. 3, pp. 841–849, 2009.

[4] S. Goldin-Meadow, "The role of gesture in communication and thinking," *Trends in cognitive sciences*, vol. 3, no. 11, pp. 419–429, 1999.

[5] S. Lang, M. Block, and R. Rojas, "Sign language recognition using kinect," in *Artificial Intelligence and Soft Computing*. Springer, 2012, pp. 394–402.

[6] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the kinect," in *Proc. ICMI '11*, 2011, pp. 279–286.

[7] N. Pugeault and R. Bowden, "Spelling it out: Real-time asl fingerspelling recognition," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, 2011, pp. 1114–1119.

[8] X. Chen, "http://www.dailymail.co.uk/sciencetech/article-2481515/microsoft-kinect-sensor-converts-sign-language-speech-text.html," in *Microsoft Faculty Summit*, 2013.

[9] C. Vogler and D. N. Metaxas, "Handshapes and movements: Multiple-channel american sign language recognition," in *Proc. of Gesture Workshop*, 2003.

[10] S. Ong and S. Ranganath, "Automatic sign language analysis: a survey and the future beyond lexical meaning," *IEEE Tran. on PAMI*, vol. 27, no. 6, pp. 873–891, 2005.

[11] H. Cooper, B. Holt, and R. Bowden, "Sign language recognition," in *Visual Analysis of Humans*. Springer, 2011, pp. 539–562.

[12] M. W. Kadous *et al.*, "Machine recognition of Auslan signs using PowerGloves: Towards large-lexicon recognition of sign language," in *Proceedings of the Workshop on the Integration of Gesture in Language and Speech*, 1996, pp. 165–174.

[13] T. Starner, J. Weaver, and A. Pentland, "Real-time american sign language recognition using desk and wearable computer based video," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1371–1375, 1998.

[14] R. Muñoz-Salinas, R. Medina-Carnicer, F. J. Madrid-Cuevas, and A. Carmona-Poyato, "Depth silhouettes for gesture recognition," *Pattern Recognition Letters*, vol. 29, no. 3, pp. 319–329, 2008.

[15] C. Vogler and D. Metaxas, "ASL recognition based on a coupling between HMMs and 3D motion analysis," in *Computer Vision, 1998. Sixth International Conference on*. IEEE, 1998, pp. 363–369.

[16] J. Segen and S. Kumar, "Shadow gestures: 3D hand pose estimation using a single camera," in *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, vol. 1. IEEE, 1999.

[17] P. Dreuw, J. Forster, Y. Gweth, D. Stein, H. Ney, G. Martinez, J. V. Llahi, O. Crasborn, E. Ormel, and W. s. o. Du, "Signspeak–understanding, recognition, and translation of sign languages," in *Proc. of 4th Workshop on the Representation and Processing of Sign Languages*, 2010, pp. 22–23.

[18] N. Weibel, S. Rick, C. Emmenegger, S. Ashfaq, A. Calvitti, and Z. Agha, "Lab-in-a-box: semi-automatic tracking of activity in the medical office," *Personal and Ubiquitous Computing*, vol. 19, no. 2, pp. 317–334, 2015.

[19] A. Fouse, N. Weibel, E. Hutchins, and J. D. Hollan, "Chronoviz: a system for supporting navigation of time-coded data," in *Proc. CHI '11*, 2011.

[20] W. Sandler, I. Meir, C. Padden, and M. Aronoff, "The emergence of grammar: Systematic structure in a new language," *PNAS*, vol. 102, no. 7, pp. 2661–2665, 2005.

[21] R. L. McKee and J. Napier, "Interpreting into international sign pidgin: An analysis," *Sign language & linguistics*, vol. 5, no. 1, pp. 27–54, 2002.

[22] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Commun. ACM*, vol. 56, no. 1, pp. 116–124, Jan. 2013.

[23] U. Bellugi and S. Fischer, "A comparison of sign language and spoken language," *Cognition*, vol. 1, no. 2, pp. 173–200, 1972.

[24] D. Garcia, "Inpaint over missing data in 1-D, 2-D, 3-D, ... N-D arrays. MATLAB Central File Exchange, url: http://www.mathworks.com/matlabcentral/fileexchange/27994-inpaint-over-missing-data-in-1-d–2-d–3-d–n-d-arrays m," 23 Jun 2010 (Updated 11 Nov 2013).

[25] G. Wang, D. Garcia, Y. Liu, R. De Jeu, and A. J. Dolman, "A three-dimensional gap filling method for large geophysical datasets: Application to global satellite soil moisture observations," *Environmental Modelling & Software*, vol. 30, pp. 139–142, 2012.

[26] M. Azimi, "Skeletal joint smoothing white paper," *Microsoft Developer Network. At URL: http://msdn. microsoft. com/en-us/library/jj131429. aspx last visited, 11th April*, 2014.

[27] "Wavelet Toolbox, wmulden, MATLAB version 8.4.0 (R2014b).The MathWorks Inc." Sep. 2014.

[28] M. Aminghafari, N. Cheze, and J.-M. Poggi, "Multivariate denoising using wavelets and principal component analysis," *Computational Statistics & Data Analysis*, vol. 50, no. 9, pp. 2381–2398, 2006.

[29] P. J. Rousseeuw and K. V. Driessen, "A fast algorithm for the minimum covariance determinant estimator," *Technometrics*, vol. 41, no. 3, pp. 212–223, 1999.

[30] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "Multi-scale deep learning for gesture detection and localization," in *ECCV Workshops*, 2014.

[31] W. C. Stokoe, *Sign language structure: The First Linguistic Analysis of American Sign Language*. Linstok Press, 1978.

[32] E. S. Klima and U. Bellugi, *The signs of language*. Harvard University Press, 1979.

[33] K. Emmorey, *Language, cognition, and the brain: Insights from sign language research*. Psychology Press, 2001.

[34] W. Sandler and D. Lillo-Martin, *Sign language and linguistic universals*. Cambridge University Press, 2006.

[35] W. Sandler, I. Meir, S. Dachkovsky, C. Padden, and M. Aronoff, "The emergence of complexity in prosody and syntax," *Lingua*, vol. 121, no. 13, pp. 2014–2033, 2011.

[36] W. Sandler, "Dedicated gestures and the emergence of sign language," *Gesture*, vol. 12, no. 3, pp. 265–307, 2012.