Measuring Time Gestures with the Microsoft Kinect

Daniel Lenzen (dlenzen@ucsd.edu)
University of California – San Diego
Department of Cognitive Science, 9500 Gilman Drive
La Jolla, CA 92093

Abstract

Gestures related to time can reveal implicit representations of the TIME is SPACE metaphor (Núñez & Sweetser, 2006). While past research has shown that gestures illustrate the direction of future and past on timelines, no detailed analysis of timelines has been possible. Using the Kinect depth camera and body tracking technology, we tracked participants' co-speech gestures while explaining time-related concepts. We present data collected with novel, relatively unsupervised Kinect-based methods that offer evidence similar to traditional gesture-coding methods and could provide the opportunity for novel theoretical findings.

Keywords: Gesture; Motion Capture; Time; Metaphor

Introduction

Space-Time Metaphors

Human cultures frequently use orientational metaphors to reason and communicate about abstract concepts such as time and number. That is, numbers and temporal concepts are mapped onto spatial locations (Lakoff & Johnson, 1980). These metaphors are readily available in the language cultures use. In particular, English speakers use phrases grounded in the sagittal (front-to-back) axis when discussing time – e.g. "Thanksgiving is behind us". While some cultural artifacts such as calendars and timelines use the lateral (left-to-right) or vertical (up-down) axes to visually represent time, these do not typically appear in English expressions. Additionally, those artifacts (as well as writing direction) tend to vary more frequently across cultures than the front-is-future, past-is-behind representation of time.

These spatial metaphors are not just linguistic epiphenomena, but have been shown to be embodied in physical experience. Reaction times to the presentation of time words (visual or auditory) is significantly faster when the direction of reaction movements (e.g. button press, slider) is congruent with the participant's time-space metaphor (Sell & Kashak, 2011).

While these sensorimotor tasks offer behavioral support to time-space mappings, they constrain responses to single dimensions. Co-speech gesture, on the other hand, offers a three-dimensional canvas for conveying information.

Time & Gesture

Gesture can be a channel into understanding spatial or imagistic features of cognitive representations (McNeill, 1992), making it a ripe domain for investigating abstract concepts understood metaphorically with spatial concepts. Studying gestures about time, for example, has revealed features of conceptualizations of time (see Cooperrider, Núñez & Sweetser, 2014, for a review). For example, when discussing time, gesturers use the sagittal (front-to-back) and horizontal (left-to-right) axes in varying ways depending on cultural factors. That is, members of a culture that has a dominant language that is written left to right tend to gesture about sequential events with the past to the left and the future to the right, while the opposite is true for users of languages that are written right to left.

This directional difference is also found for the slightly different, sagittal axis (Núñez & Sweetser, 2006). The Aymara tribe in the Andes mountains speak and gesture about the past being in front of them and the future behind. These studies show different cultural conceptualizations of time in relation to oneself exist, and gesture can support and expand on linguistic evidence of these conceptualizations.

While English has many expressions about time that rely on the sagittal axis but not the lateral axis (i.e. an upcoming event is not "to the right"), English speakers tend to gesture along the lateral axis when discussing time naturally (Núñez & Cooperrider, 2013 for a review). When directly instructed to gesture about time, however, participants switched to primarily using the sagittal axis (Casasanto & Jasmin, 2012). This intentionally communicative setting mirrors the use of the sagittal axis when discussing time in American Sign Language

(ASL). ASL strictly uses the sagittal axis for deictic references to time and the lateral axis for sequencing events (Emmorey, 2001).

While the axis and direction in which people gesture for different temporal events has been documented (i.e. forward, backward, left, right), there has been almost no work on the details of *how* people gesture in these directions. Recent work has suggested that co-speech gestures sometimes blend the sagittal and lateral timelines (Walker & Cooperrider, in press), but there has been no detail in the profiles of these gestures. We will examine the three-dimensional location, distance and velocity of these gestures.

Time gestures could be simple binary movements – forward or backward, right or left – as has been coded in previous research. That is, when talking about different times in the past, co-speech gestures might move backward towards the same (or a random) location and at the same (or random) speed no matter how distant the time concept may be. The information in these gestures would be redundant with the speech and only reinforce when the event happened. In this case, we would expect to find only a difference of direction of gesture between future and past words.

On the other hand, if the timeline metaphors are strongly embodied and spontaneously displayed, gestures along the lateral and sagittal axes related to time concepts could encode the perceived distance from now (some front-center location on the body) of those concepts. A gesture about the deictic concept "recently" could have a different profile from a gesture about "previously", despite their common feature of being related to the past. In this case, we would expect to find a difference in speed or trajectory of gesture based on perceptual distance of the concept.

New developments in body-tracking technology from contact-free sensors can add a level of detail to coarser measures of gesture. Depth cameras can contribute yet another level – information in the z-axis – overcoming the 2-dimensionality of normal RGB video.

Body Tracking

Most cognitive gesture research has been conducted with video cameras and human coding, but there have been many recent advancements tracking human motion with computer vision techniques on RGB video (Song, Dimirdjian, & Davis, 2012) or measured directly with physical motion capture sensors placed on the body (Lu, & Huenerfauth, 2010). Computer vision techniques for tracking gesture are a promising avenue, but are computationally intense and currently less accurate than computer vision algorithms using the depth data offered by the Kinect (Han, Shao, Xu, & Shotton, 2013). Motion capture systems offer precise tracking of body movements in a fixed laboratory setting (for exception see Glowinski et al., 2013) at a greater cost of equipment, space, and setup time.

The Microsoft Kinect offers a non-invasive, cheap, mobile alternative to traditional motion capture at a cost of tracking accuracy. Previous research has used the Kinect to study laughter (Mancini, Varni, Niewiadomski Volpe, & Camurri, 2014)

With basic data techniques and integrative software, the Kinect can offer researchers new measurements of spontaneous and intentional gesture. This study uses the Kinect to add three-dimensional, quantitative detail to our understanding of time-related gestures.

Method

Participants

24 UCSD undergraduates participated in the gesture study. All participants learned English before the age of 4. Three participants were excluded from further analysis because of significant exposure to a sign language. Three participants were excluded for having body tracking data for less than 80% of frames, leaving 18 participants (13 female, 5 male; 1 left-handed).

Thirty participants from the same population as the gesture study participated in the online norming study.

Materials

Twelve time-related words were selected from those used in Lakens, Semin, & Garrido (2011). Words were classified as either past, present, or future-related (e.g. "Past", "Today", "The day after tomorrow) and either deictic referential or sequential (e.g. "Recently", "Earlier").

Additionally, 10 spatial and 14 abstract concepts were used as filler words (e.g. "Front", "Hero").

Norming Study

In order to determine how far in the past or future the stimulus words were perceived to exist, an independent group of 12 participants were given an online survey and asked to mark

the position on a horizontal line with "Present" marking the midpoint and no endpoints (Lakens et al., 2011). Future words were rated as significantly to the right of the midpoint (t-test $p < .001$; mean = 34.8/100; s.d. = 28.5). Past words were rated as significantly to the left of the midpoint (t-test $p < .001$; mean = -40.2/100; s.d. = 30.6).

Procedure

During the gesture tasks, participants sat on a stool facing a Kinect v1.0 placed 1.7m away at eye-level. Stimuli were presented using Microsoft Powerpoint on a laptop placed below the Kinect. Stimuli automatically progressed, and participants were not interrupted when the next word appeared – resulting in some variability in the length of each trial. An experimenter was present and sat next to the Kinect and laptop. Audio, RGB frames, depth frames, and 10 upper-body joint estimations (30 fps) were recorded continuously during the gesture tasks using ChronoSense software (Weibel et al., 2015).

Gesture Study

Participants viewed time, space or abstract words and were asked to explain the concept to a person who did not know what they meant. There was no mention of speech or gesture. Participants had 15s to respond.

Responses were segmented in ChronoViz (Fause et al., 2011), a software tool designed to integrate Kinect data and allow for data annotation. Spontaneous gesture trials were segmented based on the speech produced – ending when the participant started describing the next stimulus.

Results

Data Cleaning

Occasionally, mostly due to occlusion, the Kinect is unable to track joints. In our study, wrist estimates were present for significantly more time points than hand estimates (89.8% present vs. 51.1% present), so the wrist is used as a proxy for hand location. Data segments in which both wrists were not tracked (10.2% of total data) were excluded from analysis.

Wrist data was smoothed using a moving average filter (span of 10). Note that this smoothing limits the influence of outliers – useful for erroneous data points, but also underestimates the peaks of gestures, working against an effect we might find.

To normalize across participants, peak locations were calculated as z-scores (standard deviations) from the mean locations across all responses (including spatial and filler terms).

Responses in which there was little movement were excluded from analyses. Lack of movement was defined as less than 1.8m of total movement – an average of test data segments in which we observed no noticeable, meaningful gesture.

Average Peak Analysis

During spontaneous gesture responses, the average rightward (participant's perspective) wrist maximum was further to the right for the 5 future-related phrases than the 5 past-related words (-2.10 vs. -1.50 standard deviations from the participant's mean lateral position; $p < 0.01$).

The average leftward maximum was further to the left for the 5 past-related words than for the 5 future related words (2.65 vs. 2.06 standard deviations left of the mean lateral position; $p < 0.05$).

Along the sagittal axis, the average forward wrist maximum was not significantly further forward for the 5 future-related phrases than the 5 past-related words (-1.91 vs. -1.75 standard deviations from the mean sagittal position; $p = 0.3$).

The average backward maximum was not significantly further back for the 5 past-related words than for the 5 future related words (1.46 vs. 1.35 standard deviations from the mean sagittal position; $p = 0.4$).

We used gesture peaks as a proxy for activity on a particular axis to compare responses to deictic and sequential words. Right hand responses to sequential words had forward maxima significantly less far forward than those for sequential words (-2.00 vs. -2.62 standard deviations forward of the mean wrist position; $p < 0.01$). No other significant differences between deictic and sequential words occurred.

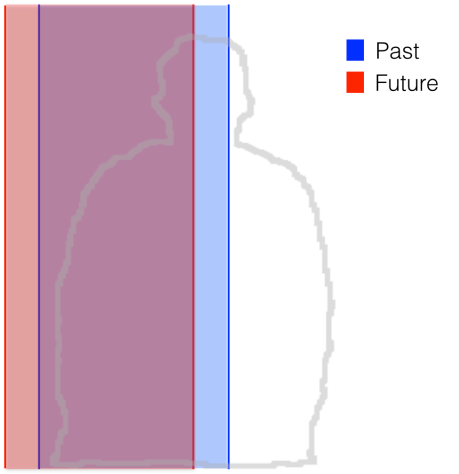

Fig 1. Average lateral rightward and leftward maxima for future and past responses.

Metaphor Blends

Based on previous research, we predicted a relationship between forward gestures and rightward gestures, and backward and leftward gestures. Right hand maxima further to the participants' right were significantly positively correlated ($r = 0.25$; $p < 0.05$) with maxima further in front of the participant and significantly negatively correlated ($r = -0.25$; $p<0.05$) with maxima further behind (or less far in front of) the participant. This suggests responses with a gesture further to the right tend to have a gesture further forward and tend to not have a gesture further back.

Right hand maxima further to the participants' left are significantly negatively correlated with maxima further in front of the participant ($r = -0.37$; $p < 0.01$) and not significantly correlated with maxima further behind (or less far in front of) the participant. This suggests responses with right hand gestures further to the left tend to have gestures less far forward, but not further back from the mean position.

Total Distance

As an indicator of relative activity between hands, we calculated cumulative distance traveled during the response windows. Across all spontaneous gesture responses, the right wrist travelled significantly further on average than the left wrist (3.34m vs. 2.56m; $p < 0.01$). The right wrist did not travel significantly more for future words than for past words (3.07m vs. 3.20m; $p > 0.5$), and the left wrist did not further for past words than for future words (2.54m vs. 2.28m; $p > 0.25$).

Discussion

Our new methods of measuring co-speech, spontaneous gesture supports previous hand-coded findings while requiring less human supervision. More nuanced analyses and more human supervision could offer new insight on well-known gesture findings.

We have replicated the findings of Casasanto and Jasmin (2012) that there are more significant differences in spontaneous, co-speech gesture along the lateral axis than the sagittal axis. Similar to Walker and Cooperrider (in press), we found a relationship between more forward gestures and more rightward gestures, and a negative relationship between rightward gestures and further back (less far forward) gestures. This relationship does not exist for the left hand

Further analysis could examine the distance traveled along the particular axes of interest and help understand how gesturers use the two axes in greater detail. We have only reported distance in general, and used peaks as a proxy for activity along an axis – a coarse measure. We can analyze intentional, elicited gesture in the vein of previous research to provide additional detail. In the future, these methods could be co-opted to provide novel findings for theoretical issues.

A future study will compare these gestures to more developed systems of manual movements – sign languages. Sign languages have grammatical rules for locations and movement trajectories, and previous research has found the use of timelines in various sign languages (Engberg-Pedersen, 1993). Few attempts have been made to understand exactly how signers use the sagittal timeline.

Emmorey (2001) states that in ASL, points to locations behind the signer can only have spatial, and not temporal meaning. With these methods we can examine if signs related to more distant past or future events subtly encode that distance in location or trajectory as a "gestural" component of sign, despite the grammatical rule.

Acknowledgments

Thanks to Seana Coulson and Ben Bergen for project design and analysis suggestions. Thanks to Jim Hollan, So-One Hwang, Carol Padden and the UCSD Chancellor's Interdisciplinary Collaboratories Fellowship for support.

References

Casasanto, D., & Jasmin, K. (2012). The hands of time: Temporal gestures in english speakers. *Cognitive Linguistics*, 23(4), 643-674.

Cooperrider, K., Núñez, R. & Sweetser, E. (*2014*). The conceptualization of time in gesture. In Müller, C., Cienki, A., Fricke, E., Ladewig, S.H., McNeill, D., & Tessendorf, S.(Eds.), *Body-Language-Communication (vol. 2)*. New York: Mouton de Gruyter.

Emmorey, K. (2001). *Language, cognition, and the brain: Insights from sign language research*. Mahwah: Lawrence Erlbaum Associates, Publishers.

Engberg-Pedersen, E. (1993). *Space in Danish Sign Language: The semantics and morphosyntax of the use of space in a visual language*. Hamburg, Germany: Signum-Verlag.

Fouse, A., Weibel, N., Hutchins, E., & Hollan, J.D. (2011). ChronoViz: A System for Supporting Navigation of Time-Coded Data. *Proceedings of ACM Conference on Human Factors in Computing Systems,* 299-304.

Glowinski, D., Mancini, M., Cowie, R., Camurri, A., Chiorri, C., & Doherty, C. (2013). The movements made by performers in a skilled quartet: a distinctive pattern, and the function that it serves. *Frontiers in psychology*, *4*.

Han, J., Shao, L., Xu, D., & Shotton, J. (2013). Enhanced Computer Vision with Microsoft Kinect Sensor: A Review. *IEEE transactions on cybernetics*, 43, 5.

Huenerfauth, M., & Lu, P. (2010). Accurate and Accessible Motion-Capture Glove Calibration for Sign Language Data Collection. *ACM Transactions on Accessible Computing*, 3, 1, 2.

Lakens, D., Semin, G.R., & Garrido, M.V. (2011). The sound of time: Cross-modal convergence in the spatial structuring of time. *Consciousness and Cognition*, 20, 437-443.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago: University of Chicago Press.

Mancini, M., Varni, G., Niewiadomski, R., Volpe, G., & Camurri, A. (2014, April). How is your laugh today?. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems* (pp. 1855-1860). ACM.

McNeill, D. (1992). *Hand and mind: What gestures reveal about thought.* Chicago: University of Chicago Press.

Núñez, R., Cooperrider, K., (2013). The tangle of space and time in human cognition. *Trends in Cognitive Sciences*, 17(5), 220-229.

Núñez, R., & Sweetser, E. (2006). With the future behind them: Convergent evidence from Aymara language and gesture in the crosslinguistic comparison of spatial construals of time. *Cognitive Science*, 30, 1-49.

Sell, A.J., & Kashak, M.P. (2011). Processing time shifts affects the execution of motor responses. *Brain and Language*, 117, 39-44.

Song, Y., Demirdjian, D., & Davis, R. (2012). Continuous body and hand gesture recognition for natural human-computer interaction. *ACM Trans. Interact. Intell. Syst.* 2, 1.

Walker, E., & Cooperrider, K. (in press). The continuity of metaphor: Evidence from temporal gestures.

Weibel, N., Rick, S., Emmenegger, C., Ashfaq, S., Calvitti, A., & Agha, Z. (2015). LAB-IN-A-BOX: semi-automatic tracking of activity in the medical office. *Personal Ubiquitous Computing,* 19, 317–334.