Towards a Gesture-Based Story Authoring System: Design Implications from Feature Analysis of Iconic Gestures During Storytelling[⋆]

Sarah Anne Brown[1][0000−0002−3103−927X], Sharon Lynn Chu[1][0000−0003−4790−2020], Francis Quek[2], Pomaikai Canaday[3][0000−0001−5089−7184], Qing Li[4][0000−0002−2015−1733], Trystan Loustau[5][0000−0002−8687−6446], Sindy Wu[1][0000−0003−3904−3449], and Lina Zhang[2][0000−0002−9847−7338]

[1] University of Florida, Gainesville FL 32611, USA
[2] Texas AM University, College Station TX 77843, USA
[3] Georgetown University, Washington DC 20057, USA
[4] Santa Fe College, Gainesville FL 32606, USA
[5] Florida State University, Tallahassee FL 32306, USA

Abstract. Current systems that use gestures to enable storytelling tend to mostly rely on a pre-scripted set of gestures or the use of manipulative gestures with respect to tangibles. Our research aims to inform the design of gesture recognition systems for storytelling with implications derived from a feature-based analysis of iconic gestures that occur during naturalistic oral storytelling. We collected story retellings of a collection of cartoon stimuli from 20 study participants, and a gesture analysis was performed on videos of the story retellings focusing on iconic gestures. Iconic gestures are a type of representational gesture that provides information about objects such as their shape, location, or movement. The form features of the iconic gestures were analyzed with respect to the concepts that they portrayed. Patterns between the two were identified and used to create recommendations for patterns in gesture form a system could be primed to recognize.

Keywords: Gesture Analysis · Gesture-Based Storytelling Systems · Iconic Gestures · Machine Learning · Human-Computer Interaction.

1 Introduction

Today's gesture recognition systems are limited in the kinds of applications they can be applied to. When applied to storytelling, gesture systems have been mostly limited to the use of gestures to provide commands to the system (e.g., [8]) or to manipulate tangible objects related to the story (e.g., [11]). Following Quek's taxonomy [17], the former are *semaphoric gestures* that define sets of

[⋆] Research supported by National Science Foundation Grant #1736225 *To Enact, To Tell, To Write: A Bridge to Expressive Writing through Digital Enactment.*

pre-defined whole static or dynamic hand poses, and the latter are *manipulative* uses of hand movement whose purpose is typically to generate a control signal. This paper is concerned with conversational gestures that are performed in conjunction with speech. More specifically, we are interested in how gesture systems can be designed to support story authoring through the feature analysis of free gestures produced in naturalistic conversational/storytelling contexts. Such gestures are termed *gesticulation*: gestures that are constructed, typically unwittingly, at the moment of speech [16]. Gesticulation is creatively produced, and is normally impermeable to 'whole gesture' recognition techniques typically used in machine learning approaches that recognize exact repeated performances. For example, one does not always produce the stylized 'turning steering wheel' gesture when one says the word 'car'. Research has shown that gesticulation carry their meaning in gestural aspects or features that carry the mental image of the multimodal discourse utterance (e.g., [18]). With the availability of technologies that can capture and detect the movements of the hand in relatively high fidelity such as the Leap Motion and Kinect systems [14], there are increasing opportunities to enable the creation of story products (e.g., a comic) to be driven by storytelling gesticulations and speech.

Our research in this paper addresses specifically iconic gestures, which provide representational information about objects, such as their shape, location, or movement. This initial focus is because iconic gestures are critical to storytelling: of the existing gesture types, iconic gestures are frequently used to aid in visual depictions of concrete objects [16] . We conducted a study whereby 20 participants (including 3 pilot participants) were video recorded while retelling stories from various cartoon stimuli. Our analysis involved a feature-based analysis of iconic gestures extracted from these recordings.

2 Background and Related Work

2.1 Gesture-based Storytelling Systems

The use of gestures in storytelling applications can be classified into three categories. In the first category, users are asked to perform certain gestures during the consumption of stories so as to increase their engagement in the experience. For example, in Kistler et al.'s work [8], players are asked to perform gestures indicated on screen during quick time events (QTE) during their engagement in the choose-your-own-adventure story 'Sugarcane Island'. In the second category, users use gestures to control a tangible object related to the story being told or created. For instance, in Liang et al.'s *Puppet Narrator* system [11,10], children use specific hand gestures to control a puppet avatar to perform basic movements like 'move right' or 'stretch'.

In the third category, users can perform free gestures to contribute to storytelling scenarios. For example, Kistler et al. [7] conducted a study asking participants explicitly to perform full-body gestures for a set of given in-game actions (e.g., ask permission, approach supervisor, sit at a bar) in a story-based scenario. They performed an analysis of the participants' full-body gestures using a high-level scheme, coding for 'form', 'gesture type', and '(involved) body parts'. An example of their results is that most participants gestured "talking to a supervisor" using a metaphoric gesture, while an iconic gesture was mostly used to indicate "sitting at a bar and waiting". They subsequently built a recognizer for the specific gestures identified in their study. Our study, in contrast, focuses on gesticulations produced in naturalistic storytelling contexts.

2.2 Gesture Recognition Systems and Approaches

As outlined in Rautaray and Agrawal [20] and Al-Shamaylehk et al. [1], hand gesture recognition, typically accomplished through machine learning algorithms [21], consists of three steps - **Detection:** Detecting the hands and extracting necessary features from them for recognition and/or tracking (tracking only necessary when the application is dynamic in nature, as opposed to static gesture recognition); **Tracking:** Maintaining detection of the hands from frame to frame; and **Recognition:** The final interpretation of what the hands semantically express in the context of a given application.

The features typically used in gesture machine learning algorithms include specific pixel values, whole three-dimensional hand models, or two-dimensional hand shapes [20]. A main focus in the literature has been on the segmentation of whole gestures for interpretation [4,6,12]. However, as can be observed during natural discourse, gestures flow in and out of each other near seamlessly at times. Our work looks at gesture features at a much lower level by using grounding from psycholinguistic research [16,9] with a goal to inform the development of gesture recognition systems for storytelling.

We note as well that few gesture recognition systems currently deal with completely naturalistic settings. The closest application of gesture recognition systems for natural, conversational gestures is to recognize sign language gestures. In that case, the gestures can be argued to be conversational, but not necessarily natural, as they stem from a predefined vocabulary.

2.3 The Semantics of Gestures and their Features

While variations exist between different gesture taxonomies proposed in the literature [13,20], iconic gestures tend to be a common category across many of them. As defined by McNeill, iconic gestures "bear a close formal relationship to the semantic content of speech" [16], furnishing imagistic information about their referents such as their shape, location, or movement. McNeill emphasizes the need to interpret the meaning of an iconic gesture in combination with its associated spoken utterance. An example of an iconic gesture is a speaker spreading his hands out wide while describing a tree. In this case, the iconic gesture is representing the width of the tree. As such, iconic gestures aid speech by depicting a visual representation in the mind of the speaker from which both gesture and speech proceeds.

At this point, it is useful to note that the information provided by iconic gestures may either be redundant (e.g., "she is really tall [gesture with raised hand with palm facing down indicating height]), or complementary (e.g., "he went through the entrance [gesture with two inward facing hands close together signifying that the entrance is narrow]).

Kopp, Tepper, and Cassell [9] conducted a study that analyzed the features of iconic gestures by detailing their shapes and spatial properties/relationships (that they called *image description features*) and their morphologies or forms. They hypothesized that there exists an observable relationship between the physical forms of an iconic gesture's features and the image-describing meanings that can be derived from them. Their research primarily resulted in the framework with which they propose to continue their analysis, and a limited variety of gestures their conversational agent can utilize in to providing users with directions. Thus, their gesture research was not done for storytelling. They studied discourse in the context of direction-giving, with the intent of informing conversational agents that could describe an environment or give travel directions to a user.

A critical reason for why the analysis of gesture features is important to inform gesture-based storytelling systems is that within a discourse, gesture features tend to repeat themselves to convey similar meanings. Quek et al. [19] called such recurrences *catchments*. A catchment is "recognized when gesture features recur in two or more (not necessarily consecutive) gestures". For example, within a discourse session, a speaker may always perform a gesture of moving the right hand, palm faced down, to indicate horizontal surfaces. Research on the Catchment Feature Model has showed that addressing gestures as a whole often limits gesture recognition to a set vocabulary [18]. However, Quek et al. studied recurrent gesture features as they occur within individual speakers. We hypothesize that it is possible to find gesture patterns across speakers if the feature analysis is taken to a level that is abstracted enough.

3 Data Collection

The goal of our research was to find the commonalities between iconic gestures that are produced during naturalistic storytelling in terms of form features and the concepts they portray, such that design implications for gesture-based storytelling systems can be derived. We had a total of 17 adult participants in our final study, 15 male and 2 female, between the ages of 18 and 34. Before the final study, we also conducted a pilot study with 1 male and 2 female participants within that same age range to test our protocol. Participant recruitment took place both via e-mail and via an online recruitment system which offered course credit to enrolled students.

Our study protocol was similar to that used in previous gesture studies by McNeill [16]. After presenting participants with a cartoon stimulus, we asked them to retell the story of the cartoon from start to finish to the researcher in a conversational context. The researcher primarily remained as a listener. We video recorded the exchange from two angles - one close and one angled from slightly farther away. Participants were not informed that we were looking at gestures specifically, but rather that we were investigating how people tell stories to prevent them from being self-aware of their gestures during the retelling tasks.

Our study stimuli consisted of a combination of short scenes and full short films: 2 full cartoon shorts (a 5-minute Loony Tunes cartoon, *Box Office Bunny* [22] produced by Warner Bros. Animation, a 8-minute short film titled *Alike* produced by Pepe School Land [15]), and 5 cartoon scenes, each under a minute in length, extracted from 2 additional shorts (Pixar's *La Luna* [3] and *Alarm*, produced by MESAI [5]). The cartoon shorts were selected for being non-abstract in nature (having concrete objects and environments, even if they are fantastical or stylized), and having a clear sequence of events. Each participant watched and retold the same set of cartoons, enabling us to compare gestures across participants. In total, including the pilot study, 37 retellings of full cartoon shorts were collected, and 85 retellings of cartoon scenes.

4 Data Analysis

The analysis in this paper focuses on the retellings of only the Loony Tune's short *Box Office Bunny* [22]. The coding was done by 6 coders, who first underwent basic training sessions in gesture analysis.

Extracting Iconic Gestures Since the identification of iconic gestures can sometimes be subjective, 2 to 3 coders were assigned to code the same retellings. Coders were also asked to give a 'confidence of iconicity' score (on a scale of 1-5) with each iconic gesture identified, as suggested by the coding scheme outlined in McNeill [16]. Once an iconic gesture was identified, a gesture was coded as describing an Object, Action, or Position (or some combination of the three, in certain cases). After the lists of iconic gestures across all coders were synthesized, a total of 161 iconic gestures were found across all the cartoon retellings. The confidence scores were averaged for each gesture across its coders. This resulted in 65 iconic gestures with confidence averages above the median (2.667). These made up the final dataset that we used for further analysis.

Identifying Gesture Concepts From the first round of analysis, 43 iconic gesture instances were classified as referencing an Action, 15 an Object, 8 a Position. A second round of analysis was done classifying the gesture referents:

Actions were divided into: *Movement* (32 gesture instances across 11 participants): the movement of an object(s) or character(s) from one location to another; *Character Action* (10 gesture instances across 6 participants): an action performed by a character(s); and *Object Action* (1 gesture across 1 participant): an action in relation to an object(s), e.g., bombs exploding.

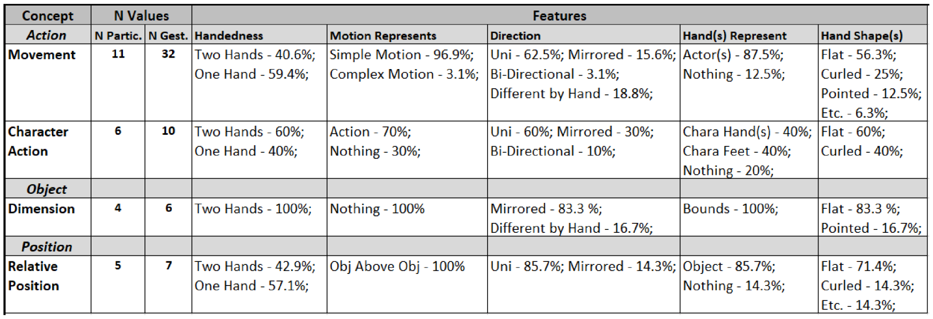
Objects were divided into: *Dimension* (6 gestures across 4 participants): defining the dimensions of an object, e.g., width or height; *Explicit Shape* (4 gestures across 3 participants): describing the shape of an object; Followed by *Volume* (3 gestures across 2 participants): defining the general volume in which an object would exist. Ex - cupping hands around a loose area regardless of object shape; and *Implicit Shape* (2 gestures across 2 participants): describing an implied feature of the shape of an object through the gesture, e.g., forming flat palms to imply the ground is flat.

Positions were divided into: *Relative* (7 gestures across 5 participants): position of an object from an external perspective, in relation to another object, e.g., the movie theater was on top of Bugs Bunny's home; and *Internal* (1 gestures across 1 participants): position of an object from the perspective of oneself or a character, e.g., pointing at the top of one's head to show where Bugs Bunny's ears would be.

Coding Gesture Form Features We used the coding scheme proposed by Church and recommended by McNeill [16]. The scheme involves the coding of hand shape, handedness, space in which the gesture was performed, view, and motion direction, and meaning of the gesture (what the motion and the hands represent). For coding hand shape, the proposed method was to match the shape to an existing table of American Sign Language signs [2]. Ho wever, we found that the iconic gestures aligned with only a limited number of signs in our dataset. We thus reduced the coding scheme to three sets of form features: (i) Flat palms and fingers; (ii) Curled fingers; and (iii) One or more Pointed fingers. Coding of the other aspects of the gesture was relatively unchanged from the original coding scheme.

5 Results

We considered only concepts that had a minimum of 4 gesture instances across at least 4 participants in our results. Furthermore, we excluded entirely divergent and entirely convergent gesture features since both provide little discriminatory potential. The relevant results are shown in Figure 1, and clear visual examples of each analyzed concept can be found in Figure 2.



Concept	**N Values**		**Features**				
Action	**N Partic.**	**N Gest.**	**Handedness**	**Motion Represents**	**Direction**	**Hand(s) Represent**	**Hand Shape(s)**
Movement	**11**	**32**	Two Hands - 40.6%; One Hand - 59.4%;	Simple Motion - 96.9%; Complex Motion - 3.1%;	Uni - 62.5%; Mirrored - 15.6%; Bi-Directional - 3.1%; Different by Hand - 18.8%;	Actor(s) - 87.5%; Nothing - 12.5%;	Flat - 56.3%; Curled - 25%; Pointed - 12.5%; Etc. - 6.3%;
Character Action	**6**	**10**	Two Hands - 60%; One Hand - 40%;	Action - 70%; Nothing - 30%;	Uni - 60%; Mirrored - 30%; Bi-Directional - 10%;	Chara Hand(s) - 40%; Chara Feet - 40%; Nothing - 20%;	Flat - 60%; Curled - 40%;
Object							
Dimension	**4**	**6**	Two Hands - 100%;	Nothing - 100%	Mirrored - 83.3 %; Different by Hand - 16.7%;	Bounds - 100%;	Flat - 83.3 %; Pointed - 16.7%;
Position							
Relative Position	**5**	**7**	Two Hands - 42.9%; One Hand - 57.1%;	Obj Above Obj - 100%	Uni - 85.7%; Mirrored - 14.3%;	Object - 85.7%; Nothing - 14.3%;	Flat - 71.4%; Curled - 14.3%; Etc. - 14.3%;

Fig. 1. Results (%s reflect counts of each code within a given sub-concept)

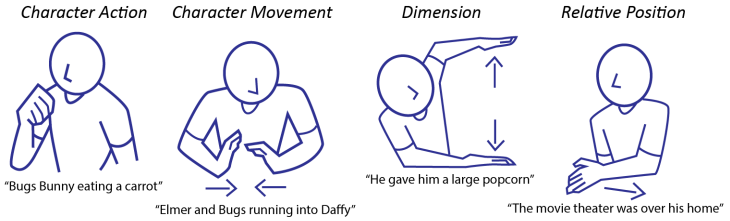

Character Action
Character Movement
Dimension
Relative Position
"Bugs Bunny eating a carrot"
"Elmer and Bugs running into Daffy"
"He gave him a large popcorn"
"The movie theater was over his home"

Fig. 2. Visual Examples of Concepts as Gestures (representative, but not directly reflective of participant gestures)

6 Discussion and Design Implications

As described in our Background section, gestures echo or add to information presented in speech. From the uncovered gesture feature patterns, we discuss below implications of how a gesture-based system can take advantage of storytellers' gesticulations to adjust visual story output.

For gestures referencing *Movement*, motions of the hand always depicted the imagined movement of the object referenced - these movements were split into the two codes under the feature depicting what "Motion Represents" (see Figure 1). Within this feature, "Simple Motion" detailed movements of an object(s) or character(s) from one location to another, and "Complex Motion" detailed those that by comparison included a secondary motion, e.g., characters spinning together when they collide in a cartoon-ish fashion, as oppose to one character pursuing another with no further motions described. Thus, detecting hand motions in a gesture-based storytelling system could hypothetically reliably help to trace out objects' movements.

Gestures representing *Character Actions* provide a unique challenge, as their features appear to be highly tied to the specific actions being referenced, e.g., Bugs Bunny eating a carrot in our given cartoon represented all of gestures where the "Hand(s) Represent" the hand(s) of the character, Bugs Bunny. We are limited in a true analysis of this sub-concept, because the only two actions were described with any regularity from the given cartoon was the aforementioned Bugs Bunny eating a carrot, or the starring characters dancing on a gum-covered carpet (4 of either, for a total of 8 of the 10 identified *Character Actions*). However, much like *Movement*, there was a majority of hand motion directions were uni-directional, and motions of the hand tended to represent the action (comparable to the same feature, "Motion Represents" coded solely as different types of movement in *Movement*).

Gestures representing *Dimension* was one of the most promising concepts with distinguishable form features - the hands consistently (83.3% or above) represented two separate bounds of a given dimension (under the feature "Hand(s) Represent"), moved in mirrored directions (under "Directions") and, and with the exception of one example, were always both flat ("Hand Shape(s)"). These are set features a system could be built to detect so as to adjust a visualized object's dimension (with the object being determined from speech), and is reliable in that the other concepts do not replicate this set of form features.

Finally, gestures for *Relative Position* consisted largely of uni-directional motions (with larger percentages than in both *Movement* and *Character Action*), paired with a flat hand shape. We are limited by our chosen stimulus, *Box Office Bunny* [22], in that gestures produced for this sub-concept solely described objects that existed above or on top of another object. A common example was descriptions of the movie theater sitting upon Bugs Bunny's home. It is hard to say without a broader range of relative relationships, but it could be that other potential relationships share the same commonalities in form features. In which case, a system would look for a uni-directional motion paired with a flat hand to determine a positional relationship between one object and another. And perhaps, in the absence of a flat hand shape, the system could move on to explore the possibility that a *Movement* or *Character Action* is being portrayed, as they were the next concepts to have majorities in uni-directional hand motions, going down by highest percentages within those concepts. Thus, what we are describing is a system going through potential concepts by the probability that the detected form features match trends discovered through this analysis; filtering-down through an emerging taxonomy to determine in the end, what concept is likely being gestured, if any.

7 Conclusion

In this paper we presented an analysis of iconic gestures during naturalistic storytelling as investigated through specific concepts they portray and their form features. We coded iconic gestures as extracted from retellings of a cartoon stimulus, and our findings suggest that across gestured concepts, patterns can be found for specific form features. Though certain sub-concepts from our initial overarching concepts provided mixed results, there were notable patterns within the sub-concepts of Dimension and Relative Position. Our results provide a starting point to develop gesture-based systems that can recognize free gestures during naturalistic storytelling to produce concrete story outputs such as a cartoon animation or a comic.

As a limitation of our analysis, the sample size for each category of gesture concepts are unequal because in naturalistic contexts, we had no control over what content participants decided to include in their story retelling and what the gestured about. Moreover, we were limited to just one cartoon stimulus in the work presented. Many more storytelling gesture stimuli need to be analyzed.

References

1. Al-Shamayleh, A.S., Ahmad, R., Abushariah, M.A., Alam, K.A., Jomhari, N.: A systematic literature review on vision based gesture recognition techniques. Multimedia Tools and Applications **77**(21), 28121–28184 (2018)

2. Baker, C., Friedman, L.: On the other hand: New perspectives on american sign language. In: Language, Thought and Culture, pp. 215–236. Academic Press New York (1977)
3. Casarosa, E.: La luna (2011)
4. Chambers, G.S., Venkatesh, S., West, G.A.W., Bui, H.H.: Segmentation of intentional human gestures for sports video annotation. In: 10th International Multimedia Modelling Conference, 2004. Proceedings. pp. 124–129 (Jan 2004). https://doi.org/10.1109/MULMM.2004.1264976
5. Jang, M.H.: Alarm (2009)
6. Joshi, A., Monnier, C., Betke, M., Sclaroff, S.: A random forest approach to segmenting and classifying gestures. In: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). vol. 1, pp. 1–7 (May 2015). https://doi.org/10.1109/FG.2015.7163126
7. Kistler, F., André, E.: User-defined body gestures for an interactive storytelling scenario. In: IFIP Conference on Human-Computer Interaction. pp. 264–281. Springer (2013)
8. Kistler, F., Sollfrank, D., Bee, N., André, E.: Full body gestures enhancing a game book for interactive story telling. In: International Conference on Interactive Digital Storytelling. pp. 207–218. Springer (2011)
9. Kopp, S., Tepper, P., Cassell, J.: Towards integrated microplanning of language and iconic gesture for multimodal output. In: Proceedings of the 6th international conference on Multimodal interfaces. pp. 97–104. ACM (2004)
10. Liang, H., Chang, J., Kazmi, I.K., Zhang, J.J., Jiao, P.: Hand gesture-based interactive puppetry system to assist storytelling for children. The Visual Computer **33**(4), 517–531 (2017)
11. Liang, H., Chang, J., Kazmi, I.K., Zhang, J.J., Jiao, P.: Puppet narrator: utilizing motion sensing technology in storytelling for young children. In: 2015 7th International Conference on Games and Virtual Worlds for Serious Applications (VS-Games). pp. 1–8. IEEE (2015)
12. Madeo, R.C.B., Lima, C.A.M., Peres, S.M.: Gesture unit segmentation using support vector machines: Segmenting gestures from rest positions. In: Proceedings of the 28th Annual ACM Symposium on Applied Computing. pp. 46–52. SAC '13, ACM, New York, NY, USA (2013). https://doi.org/10.1145/2480362.2480373
13. Maricchiolo, F., Gnisci, A., Bonaiuto, M.: Coding hand gestures: A reliable taxonomy and a multi-media support. In: Esposito, A., Esposito, A.M., Vinciarelli, A., Hoffmann, R., Müller, V.C. (eds.) Cognitive Behavioural Systems. pp. 405–416. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
14. Marin, G., Dominio, F., Zanuttigh, P.: Hand gesture recognition with leap motion and kinect devices. In: 2014 IEEE International Conference on Image Processing (ICIP). pp. 1565–1569 (Oct 2014). https://doi.org/10.1109/ICIP.2014.7025313
15. Martnez Lara, D., Cano Mendz, R.: Alike (2015)
16. McNeill, D.: Hand and mind: What gestures reveal about thought. University of Chicago press (1992)
17. Quek, F.: Gesture and interaction,. Encyclopedia of Human-Computer Interaction Vol 1 **54**, 288 – 292 (2004)
18. Quek, F.: The catchment feature model for multimodal language analysis. In: null. p. 540. IEEE (2003)
19. Quek, F., McNeill, D., Bryll, R., Duncan, S., Ma, X.F., Kirbas, C., McCullough, K.E., Ansari, R.: Multimodal human discourse: Gesture and speech. ACM Trans. Comput.-Hum. Interact. **9**(3), 171–193 (Sep 2002). https://doi.org/10.1145/568513.568514

20. Rautaray, S.S., Agrawal, A.: Vision based hand gesture recognition for human computer interaction: a survey. Artificial intelligence review **43**(1), 1–54 (2015)
21. Trigueiros, P., Ribeiro, F., Reis, L.P.: A comparison of machine learning algorithms applied to hand gesture recognition. In: 7th Iberian Conference on Information Systems and Technologies (CISTI 2012). pp. 1–6 (June 2012)
22. Van Citters, D.: Box office bunny (1991)