

To Move or To Remove? A Human-Centric Approach to Understanding Gesture Interpretation

Sukeshini A. Grandhi, Gina Joue, Irene Mittelberg

Human Technology Center, Natural Media and Engineering, RWTH Aachen University, Germany
{grandhi, joue, mittelberg}@humtec.rwth-aachen.de

ABSTRACT

This paper explores how naïve observers recognize and interpret transitive actions (actions involving manipulation of objects) without accompanying speech, in order to derive guidelines for the design of gesture interpretation systems. Semi-structured interviews with 11 observers, interpreting 106 video clips of transitive actions elicited unstaged from 16 participants, reveal that people are generally able to interpret the transitive action as well as characteristics of the object manipulated despite individual variations in how people naturally gesture. In particular, people focus primarily on hand movement and hand shape to correctly interpret object characteristics, and on manner of movement of arms and/or final location of hands to interpret the goal of the transitive action (e.g., arrange objects vs. clear objects). These findings provide insights on aspects of gestures one can focus on to inform and guide the design of gesture interpretation models for interfaces that allow for individual variations in natural gesture production.

Author Keywords

Gestures, interpretation, recognition, naturalness, human centric, user, interaction design, cognitive principles, models

ACM Classification Keywords

H5.2. User interfaces: User-centered design, Theory and Methods

General Terms

Design, Human Factors, Experimentation

INTRODUCTION

The intuitive, natural and pervasive use of hand gestures in human-human communication has set HCI researchers and practitioners in the pursuit of the design of touchless gestural interfaces that inherit these qualities [2]. “Intuitive” actions do not require excessive deliberation; likewise, “natural” actions come spontaneously [10]. However, translating these qualities from gestural communication into touchless gestural interfaces requires not only understanding what constitutes a gesture vocabulary that is intuitive and natural for users to

interact with the computer, but also an accurate gesture recognition and interpretation system for the computers to respond appropriately.

Most work on improving gesture interface systems focuses on improving recognition via statistical pattern recognition and classification. These systems work best with simple gesture vocabularies (such as static hand poses and simple hand movement) as these facilitate distinctive recognition and segmentation [21]. However, a truly powerful gesture interface would be one that could also cope with the individual variations that might come with more natural and intuitive gestures produced by users. Some of these gestures sometimes appear similar yet do not convey the same meaning, which poses an additional complexity to gestural interfaces. For example, if we proposed a computer interface that allows users to perform the hand gesture of cleaning a whiteboard to convey that the screen be erased and a goodbye wave to turn off the screen, because these were considered the most intuitive for users, we need to ensure that the system be able to differentiate them. These two gestures are an example of how certain gestures have predominantly similar kinetic and visuo-spatial characteristics, in this case a raised forearm and side-to-side movement of the hand, and yet can convey different meanings, possibly conveyed in details such as finer hand shape characteristics or context.

Although people sometimes also need context to identify transitive actions [26], they still appear to be able to identify the goal of the gestures despite individual differences in an action, which can be stumbling factors for recognition accuracy in automatic gesture interpretation. This suggests that there exist certain factors or characteristics that humans reliably use to interpret and derive meaning from the gestures they see. In this paper we investigate what aspects humans focus on and what strategies they use to identify, interpret and differentiate one gesture from another, and how we can use these factors and strategies to inform and guide gesture interpretation for gestural interfaces. In particular, we focus on gestures illustrating transitive actions, namely, actions involving the manipulation of objects, for example, peeling an orange, erasing a white board or arranging folders on a shelf. Such gestures are of relevance for gesture-based serious (e.g. movement rehabilitation) or entertainment games that often require selection and manipulation of objects.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

DIS 2012, June 11-15, 2012, Newcastle, UK.

Copyright 2012 ACM 978-1-4503-1210-3/12/06...\$10.00.

BACKGROUND

Cognitive scientists including linguists, semioticians, psycholinguists and neuroscientists have extensively studied gestures in the last three decades. While these studies predominantly focus on human-human communication, they lay the foundation for understanding gestures for interaction with computer systems.

Many gesture classifications exist [4,5,6,10,15,23,24]. In terms of some of these classifications, transitive actions can be seen as pantomimes [1], which are iconic [23], ergotic [4] in the sense that they are pantomimes of the manipulation of objects, and also communicative [4] in the sense that the gestures have to be communicated to a receiving system.

Gesture Description and Analysis

The primary approach to gesture analysis to date has been manual annotation based on high-level gesture classifications and/or lower fine-grained level descriptions based on spatio-temporal aspects of the gesturing body part. These fine-grained annotations are often free-form verbal descriptions. Stokoe [31] developed one of the first notational systems to describe sign language, in terms of hand shape, position, orientation and movement. Kendon's unpublished annotation scheme suggests fine-grained descriptions of spatial positions and changes in arms, hands, torso and head of the gesturer, and was implemented computationally [22]. Based on work by Laban [19], in the performing arts, Laban Movement Analysis is used to analyze body movements in terms of *body*, *space*, *effort/dynamics*, *shape* and *relationship*. Laban notation [9] allows physical and structural annotation of movements, such as *direction of movement*, *place and position* of the body parts involved and the duration of their movement. Partially developed on Labanian analysis and experimental gesture studies, Neuroges [20] offers an annotation scheme using combined kinematic and functional characteristics.

Gesture analysis also involves segmenting a gesture into different phases: *preparation*, the *stroke* that contains the primary meaning of the gesture, and the *retraction* [6], and how these phases can be organized into meaningful chunks, or gesture phrases. Gesture phrases, in turn, are grouped into kinematic chunks, or gesture units [1]. Many of these analysis-heavy approaches focus on describing gestures in a spatio-temporal manner that are valuable in guiding gesture recognition, but they often do not specify how a finite and complete set of these characteristics can map onto action goals in a computationally tractable manner.

Role of naïve observers in gesture interpretation for HCI

Most automatic gesture recognition is based on the recognition of either static poses or simple predefined motions. People perform a limited set of gestures in constrained domains [13], but increased application complexity entails increased gesture vocabulary and hence greater gesture recognition difficulty. A small handful of research has turned to the intuitions of naïve observers to tackle this issue. Most of this prior work has focused on

iconic gestures involved in object description [17,18,29] or object manipulation tasks [12]. While some researchers argue that, generally, gestures without accompanying speech cannot be unambiguously or fully interpreted by naïve observers [8,12], other researchers suggest that at least iconic gestures may be sufficiently identified based on visuo-spatial features [17,29]. Kopp [17] found that people relied on hand shapes, movements and hand orientations to extract visuo-spatial information. To develop a model for an iconic gesture interpretation system, Sowa and Waschmuth [29] analyzed a corpus of coverbal iconic gestures produced by 37 subjects while describing five objects. Based on this corpus, they developed a set of defining gesture features: linearity of movement, distance between palms in two handed gestures, hand aperture, palm orientation, index finger direction and curvature of hand shape.

In a Wizard of Oz study, Hauptmann [13] found that gesture production varied in terms of number of hands and fingers used in performing spatial manipulations but does not report the accuracy of the Wizard's interpretations. In an exploratory study on the feasibility of task-specific intuitive human computer interaction using gestures, Hummels and Stappers [14] found that accurate interpretations of objects (as sketched by an artist in a Wizard of Oz role) could be made from gestures performed by 12 product designers with no accompanying speech. This study did not report on the degree of consistency and similarity in people's gestures, but it established that certain core features exist to accurately identify object features from gestures of different people. Though this study looked at interpretations of gestures, it was based on a single individual and might not have been reflective of how people in general interpret these gestures.

RESEARCH QUESTIONS

In this paper, we go beyond the question of whether observers can accurately interpret gestures. We investigate what aspects untrained observers focus on and the strategies they use to recognize and interpret unstaged (hence natural) gestures without accompanying speech, in order to cull guiding principles for the design of gesture interpretation systems. In other words, we use naïve observers, as robust but flexible pattern matchers and categorizers, to interpret gestures performed by naïve performers. Specifically we ask

1. How do people interpret transitive action gestures without accompanying speech and what aspects of the gestures do they focus on to identify/interpret what they see? In particular,
 - a. What is perceived as being common across gestures representing a particular transitive action (recognition and interpretation by common characteristics)?
 - b. What is perceived as being different between gestures representing different transitive actions (recognition and interpretation by differentiation)?
2. How can we use this understanding of human perception/interpretation of transitive action gestures to

provide insights for gesture interpretation models in human computer interaction?

USER STUDY

In order to address the above research questions we adopted a human-centric approach [30,35]. We conducted a study where participants, informed that the study was about people’s intuitive understanding of gestures of transitive actions, watched video clips (3-5 seconds each) of others performing such gestures. The transitive actions were chosen to explore the differences in gesture interpretation with respect to

1. manipulation of different objects (size and shape) within the same conceptual type of task: clearing coins vs. crumbs vs. toys
2. manipulation of the same object in different tasks: clearing coins vs. arranging coins in a pile
3. object manipulation in different planes (horizontal vs. vertical) within the same conceptual type of task: clearing coins vs. clearing whiteboard and arranging coins in a pile vs. arranging folders in a row on a shelf.

Stimuli

The video stimuli were data from a previous study [10] where 16 adults (8 female) were asked to gesture (and verbally explain) the action needed to go from “before” to “after” pictures, presented as pairs (Figure 1). These scenarios reflect simple computer tasks (e.g. erase, arrange) that were camouflaged as everyday non-computer scenarios to minimize the influence of conceptual models of these tasks on pre-existing input devices. Of the 912 videos of 19 tasks, 108 videos of gestures lasting 3-5 seconds each were used as stimuli for the present study. This included 96 videos, or 6



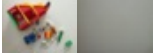
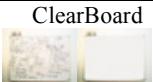
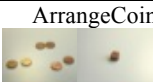
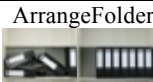
ClearCrumbs 	ClearCoins 	ClearToys 
ClearBoard 	ArrangeCoins 	ArrangeFolders 

Figure 1: The six transitive actions used in the study

transitive actions performed by 16 different people, plus 12 distraction videos. All videos were presented muted. (Video file uploaded with the paper presents sample video stimuli of the 6 transitive actions used).

Subjects

Eleven (7 female) native or near-native English speakers were recruited through public advertisements. Participants were a mix of full-time and part-time working adults and students from various fields such as aviation, marketing and communication, chemistry, English literature, and engineering. All participants were strongly right-handed based on the Edinburgh Handedness Survey and between 20-40 years old, except one who was between 41-50 years old.

Procedure

The study was conducted in three parts.

In the first part, participants were asked 1) to view videos in each given group 2) to remove any videos that did not belong to the group, and 3) to name the transitive action that they believed was being depicted in each group. Participants had no prior knowledge of the transitive actions shown in the gesture videos.

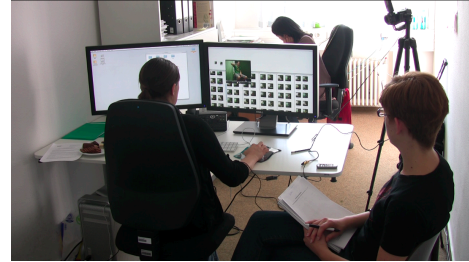


Figure 2: Study setup

In the second part, participants were asked to verbally describe and provide reasons for how they identified each transitive action, to experimenter 1, who was seated in front of them in a manner where she could not see the videos or the participants (Figure 2). Experimenter 2 sat behind participants to take notes on participants’ sorting behavior as well as to facilitate various parts of the study. This setup was to encourage participants to verbalize their thoughts rather than pointing to the screen. A semi-structured interview style was adopted to understand and confirm the reasoning behind participants’ gesture interpretation. Participants were systematically questioned to explain what characteristics they perceived to be common to gestures belonging to the same group and what differentiated gestures between groups.

In the third part, participants had to match the transitive action in each gesture group 1) to the best “before” picture of objects that were presented to them, as well as 2) to the best before-and-after picture pair presented to them.

All sessions were recorded using two high-speed video cameras and a DV camera for audio and two visual perspectives. Participants viewed one of two orders of the six gesture groups arranged on the screen, and the video clips of gestures were randomly arranged within each group to minimize order effects. At the end of the study, participants were questioned on their overall gesture interpretation strategy used in general for the study.

RESULTS

Part 1: Task-specific gesture interpretation strategy

In Part 1, participants spent an average of 24 minutes viewing the videos in the six categories and performing the task. They correctly removed at least one distractor from each category, or implicitly ignored them while interpreting the transitive actions. Most participants accurately identified the transitive actions depicted in all the six categories in a generic manner (clearing or collecting something small) rather than specifically identifying the object (clearing or collecting coins) (see Table 1 and Table 2).

Category	Action	Object
Clear Crumbs	gathering (8/11); spreading (3/11)	small, powdery, non-graspable
Clear Coins	collecting (10/11); distributing (1/11)	small, sweepable, pickable
Clear Toys	removing (10/11); grouping (1/11)	numerous, light, small-medium
Clear Board	wiping (10/11); painting (1/11)	flat surface, large, vertical
Arrange Coins	collecting/placing in one location (11/11)	small
Arrange Folders	placing objects at eye level (1/11); catching balloons (1/11).	block-like, rectangular, not heavy

Table 1: Participants' interpretation of gestures, by task

All participants were also successful in identifying differences and similarities between transitive actions with respect to object size, interaction plane/space, and primary goal of the action. However, what they focused on in expressing these similarities and variation varied, as outlined in Table 2.

Categories	Similarity	Difference
ClearToys vs. Coins vs. Crumbs	Arm movements horizontal (9/11)	Objects in Toys larger than in Coins and Crumbs (8/11)
ClearCoins vs. Board	Cleaning or getting rid of something (7/11)	Plane of interaction horizontal (Coins) vs. vertical (Board) (10/11); manner of movement forward-backward vs. side-to-side (8/11)
ClearCoins vs. ArrangeCoins	Object size (9/11)	Object in Arrange picked vs. object in Clear brushed (9/11)
ClearBoard vs. ArrangeFolders	Action at eye level (10/11)	Manner of movement with one object in Board vs. several in Folders (10/11)
ArrangeFolders vs. Coins	Repetitive movements; Rearrange (7/11)	Objects in Shelf larger than objects in Coins (10/11)

Table 2: Perceived similarities and differences in gestures

Part 2: General gesture interpretation strategies

In part two, we recorded 574 minutes of semi-structured interviews, which were transcribed. We used both open and axial coding [33] of the transcriptions to develop initial categories of how object characteristics and goals of transitive actions were described using hand shapes, hand orientation, number of hands, body part location, movement direction, movement type and manner/quality of movement. From our analysis of these categories, the following themes emerged on how naïve observers seem to derive object characteristics and the goals of the transitive actions: the influence of hand shapes (T1), number of hands (T2) and manner of movement (T5) on object interpretation; the influence of arms on interpreting interaction space (T4); the influence of non-dominant hand use (T3) and manner of movement (T6) in action interpretation; and the combination of shape and movement information (T7). These themes are illustrated through representative quotes below and are partially compiled and modeled as strategies in Figure 3.

T1. Influence of hand shapes on object interpretation

Participants used hand shapes to derive object characteristics such as size, shape and number.

Type of handgrip gave an indication of object size: spread fingers were interpreted as handling larger objects while hand grips with touching fingertips or precision grip indicated smaller objects were being handled. For instance, participants S4 and S5 describe how the object in the ArrangeCoins category is small as revealed by the tighter fingertip positions in the handgrip.

S4: I think it's small...because of the pinching motion they used to pick it up.

S5: ...That something small being picked up....basically the way that the fingers are placed very close to each other to make a sort of... pincer, tweezer-type thing...

Similarly, S2 describes object in category ArrangeFolders.

S2: they could be rearranging objects (in ArrangeFolders) that are larger than a grain of salt because their...thumb does not touch the other fingers. (in ArrangeCoins) they're pinching, possibly salt...because...their thumb...(and) their index finger (are) picking it up...connect(ing) and then releasing it.

While handgrips were explicitly used to interpret object size, the absence of a handgrip was also used to identify the size and number of objects. S2, S4 and S9 describing objects for the ClearCrumbs category, where no grip indicated very small objects or numerous objects, was typical.

S9: crumbs...they are not gripping anything, the palm is open....suggest that it is possible to drag them all together in one small place, I mean which is their left palm

S4: In (ArrangeCoins) the action is much more discrete - they are picking things up one at a time and collecting. So it made me think that there is not so many. Whereas in

(ClearCoins) they are collecting it but...sweeping which implies to me that there is much more.

S2: The hand doesn't really move... it is an indication that it's not something that you would pick up because when you pick things up you need to move your hand, but when you sweep, you just want to cover the largest possible area that you can...So...your arm and hands stretch out and then just do the sweeping motion

In a similar vein several participants interpreted a changing hand shape as handling multiple objects in succession while a constant hand shape meant handling a single object.

S4: ...one case (ClearBoard) the hand is just basically holding the same the whole time whereas with (ArrangeFolders) the hand keeps...holding and letting go

S8: ...in the (ArrangeFolders) category, the hand picks up a book or push it...whereas in category (ClearBoard), it's more or less the same...the hand is fixed. In category (ArrangeFolders), they seem to be dealing with lots of relatively smaller individual objects, what I think are the books – it's sort of more dynamic because of that. And then, in category (ClearBoard), they're just dealing with a static big object which is this board.

Handgrip was also used to interpret the shape of an object:

S4: the hand...is...holding in sort of a rectangular way...so that the four fingers are all together and the thumb on the other side, and giving it room to... a rectangular object...

S9: the fingers are together and the thumb far away which seems to suggest it's a flat, block kind of thing.

A flat hand shape was perceived as indicating a flat surface (ClearBoard):

S3: They seem to be pulling their hands towards themselves at the apex when their arm would be at 90-degree angle to their body...So it looks to be a flat object...

S8: ...it (the hand) was sorta being kept kind of quite rigid as if it was following a flat surface

S11: The hand always stayed flat against an imaginary wall

T2. Influence of number of hands used in object interpretation

Participants interpreted object size or weight from the number of hands used.

S1 and S2's description of objects in the ArrangeFolders category illustrate how using a single hand implied a lighter and smaller object.

S1: It is light enough for them to pick up with one hand

S2: The objects are relatively small...using only one hand per object implies against the size and the mass.

S10: ...they only need one arm to hold it and it doesn't seem to be that heavy, because they can move with such ease.

T3. Influence of non-dominant hand in action interpretation

When two hands were used, participants interpreted the use of the non-dominant hand as a tool to collect or gather small objects:

S1: I am very sure in this case (ClearCoins) that they are collecting because they all have their less dominant hand like a cup, to collect all the things together.

T4. Influence of arms in interpreting interaction space

Participants interpreted the space where the object interacted based on how the gesturer's arms moved. In describing the interaction space for the ClearBoard and ArrangeFolders category:

S3: So both of them seem to have...the upper arm and the forearms are perpendicular...they seem to be about the same height.

S7: They are both on a vertical surface...because of the...lifting of their arms high up in front of them....it's two different things they are doing though

In describing the difference in interaction space between ClearCoins and ClearBoard categories, the following accounts were typical.

S3: Category (ClearCoins) is again at the table level with the upper arms...the forearms parallel to the floor. In category (ClearBoard) it is the opposite. It's at eye level with upper arm parallel to the ground and the forearm perpendicular.

S9: ...in category (ClearCoins) objects are kept in front of them on a horizontal surface or a flat surface...Holding out the upper arm, the way the lower arm moves gives away this difference. So, in category (ClearBoard), the upper arm is vertical, it's perpendicular to the ground and in category (ClearCoins), it is parallel to the ground.

T5. Influence of manner of movement in object interpretation

Participants had various interpretations of object size, quality and object value/worth given the manner of hand and arm movement. Participants sensed the object to be of lesser value, undesirable or of low interest to the gesturer based on less-precise movements or what seemed like routine movements that did not require much attention. For instance,

S7: It (ClearToys) was not really of value (the way) that they were moving. Because they were not really looking at it, it wasn't something very valuable...it was just shoved to the side.

The following accounts from S5 and S10 illustrate how the difference in manner of movement aids interpretation of the objects' worth and quality.

S5: Category (ArrangeCoins) is much more precise than category (ClearCoins) is. It doesn't really matter if any of the objects fall or...just want to get many in one full scoop sort of thing.

S10: The (ClearCrumbs) category seems to be slightly

more delicate because they're using their hands more and they are moving slower than in (ClearCoins and ClearToys).... The hand (reveals this is delicate) because it's...more determined...or...more delivered...

Participants also used manner of movement to gauge the size of the objects on which an action is being performed. S2 interpreted the object in ClearToys to be small based on manner of movement.

S2: I see something very small. Can be easily brushed to one side... because no one appeared to be using that much force in their sweeping movements.

Size of movements was also used to interpret the size and shape of the object as illustrated for ClearBoard category:

S2: a lot of it has to do with how large the movement is....everyone seemed to go pretty wide and it sort of communicated that they were working with something larger than themselves....The upper arm in the way it was moving.

S4: I was thinking either a window for washing or like a wall...something flat...there wasn't much motion in the direction like I guess backwards and forwards. Mostly side to side...at a relatively constant distance from what they are interacting with...

T6. Influence of manner of movement in action interpretation

Participants varied in their interpretation of repetitive movements. Repetitive movements from one location to different locations were seen as stacking or arranging while movements from different locations to the same location was seen as gathering. S10 and S11 describe how they interpret the manner of repetitive movement as arranging in the ArrangeFolders category.

S10: Their forearm and their hands do this repeated action of up down...and to me that suggests putting one thing into another area and then repeating that action.

S11: There was a definite repetition aspect to it except... I guess it is like the reverse of stacking... You start at the same place but ended up at different place.

S9 describes the reverse for the ArrangeCoins category.

S9: they are individually picking small objects and putting it in a bottle or a jar... this reference point. Some of them are using their ...left palm as this object as...this jar and with the right...hand they individually pointing to objects and bringing it back to this reference

Manner of movement also provided clues to direction of movement. The following accounts of S1 and S2 in describing ClearCoins and ClearBoard, in particular, highlight the difference between movement from left to right or side-to-side as compared to movement that pulls and pushes or goes back and forth.

S1: they both involve rotating back and forth at the elbow but in one of them (ClearBoard) your forearm direction

is... to left and right of your palm...And the other one (ClearCoins) is moving as if it was from and into your palm.... like if you were to you push forth and pull back; left and right of your palm would be as if to wave right and wave left.

S2: Category (ClearBoard) their hands are moving back and forth as if they're painting something or cleaning something.... the main part is how the forearm and the rest of the arm interact. Elbows help to back and forth to paint. ...I think about painting a door or a wall. The table (in ClearCoins) is horizontal and they're sitting vertically ...next to it...The action is sort of a sweeping movement....

T7. Combining shape and movement

At the end of the study session, participants were asked about their overall strategy in interpreting the gestures shown to them. They all reported to have used a combination of hand shapes and motion to identify and differentiate between transitive actions as witnessed in their interpretation style throughout the study. For instance S5 and S11 described how they differentiated between whether an object was being gathered or pushed away based on the palm direction even though the arm motion was similar.

S11: But the difference for me...seems pushing (ClearToys) away whereas in the other two (ClearCoins and ClearCrumbs) seems like it's pushing towards them...the direction (of) the arms when they stop the movement and ... their palms are facing outwards away from them when they're pushing it away from themselves and when they are gathering or collecting it. The palms also end up pointing towards them.

S5: ...the hand cupping to collect or gather whatever it is...(suggests) that the object is being brought towards the person, although in the same category there's someone who seem(s) to be doing the opposite, pushing it away. But it's again the same sort of movement with the hand in the opposite direction. Sort of closing the hand towards you (vs.) opening the hand out away from you

When hand and arm movements were similar, participants noted differences in the final destination of the hand to identify where the objects were being deposited. S4 illustrates this in describing how objects were cast aside as compared to being collected towards self.

S4: I think in some sense they all perform the sweeping action....the way they do it initially is pretty similar but for the (ClearToys) it's more like sweeping it to the side, past kind of the center...past whatever is in front of them as opposed to (ClearCrumbs and ClearCoins) it's more like that they are sweeping...it towards...their hand, or the center. Basically more like in front of them.

Participants also used arm orientation to differentiate between actions when the forearm motions seemed similar. S11 described how the transitive action in ClearBoard was perceived to be on a vertical plane while the transitive action

in CoinsClear was perceived to be on the horizontal plane, based on the arm orientation and hand position.

S11: ...a swiping motion I suppose is a similarity but, one (ClearBoard) is more vertical. It's different swiping so your palm is facing outwards and it feels like you are almost waving to someone whereas in category (ClearCoins) here (you are) collecting something and your hands moving towards you and your palm never really faces the outside.

Part 3: Gesture matching

When presented with pictures of objects before manipulation (“before” pictures), all except one participant identified the Board and the Folders as the only possible object(s) that can be manipulated for the ClearBoard and ArrangeFolder

	Level of information	% times correctly matched (n=11)	% times (also) matched with other objects / tasks (n=11)
Clear Crumbs	Object	73%	27% with Coins, Toys
	Action	82%	18% with ClearCoins, ClearToys
Clear Coins	Object	64%	36% with Crumbs, Toys, Folders
	Action	73%	27% with ClearCrumbs, ArrangeCoins, ClearToys
Clear Toys	Object	55%	45% with Crumbs
	Action	55%	45% with ClearCrumbs, ClearCoins
Clear Board	Object	100%	None
	Action	100%	None
Arrange Coins	Object	64%	36% with Toys, Crumbs
	Action	82%	18% with ClearCoins
Arrange Folders	Object	100%	None
	Action	100%	None

Table 3: Gesture identification performance

categories respectively. Some level of uncertainty with respect to the exact objects manipulated was displayed by the participants for ClearToys, ClearCoins, ClearCrumbs and ArrangeCoins as shown in Table 3. However, except for

ClearToys, this possibility reduced considerably when participants were shown the before-and-after pictures, from which they could also infer the underlying transitive action for the categories.

DISCUSSION

Findings from this qualitative study provide insights on what characteristics of gestures people focus on to interpret unstaged gestures and how they interpret these characteristics in identifying gestures and differentiating between them. These insights have implications for how we can design gesture interpretation models and gesture vocabulary for applications that employ transitive action gestures.

We found that people adopt a consistent but limited set of strategies in identifying gestures of transitive actions even though they are faced with various gestural variations due to individual differences in natural gesture production. While there is no single set of hand or arm features that clearly and uniquely maps onto a given gesture of transitive action as seen in the performance of the gesture matching task (Table 3), people predominantly focus on hand shapes, hand orientation and manner of hand movements to deduce characteristics of the object being manipulated and combine these characteristics with arm movements to deduce the overall goal of the transitive action. Often, participants focused on the hand in isolation to gather meaning. This is in support of the observation that hand shape rarely changes meaningfully when hands are in motion [28]. These strategies are iteratively used to construct the overall meaning of the gesture performing the transitive action. One participant described his strategy *“(It) changed...sometimes the movement, sometimes it was the ways the hands were shaped...kind of a little bit like, playing charades or something when you’re trying to just figure out. It’s like a storyline, and then we’re trying to figure out what the action is. So, like...telling story, trying to explain something...looking at them as sort of actors.”* Overall, these strategies are perhaps successful because of the 1) iconic relations between hand shapes and an object’s geometric properties (e.g. flat surface as depicted by a flat hand) and/or 2) affordances of an object (e.g. being graspable only by a tweezer). These are both based on people’s embodied experience, previous interactions with such objects and experience performing these daily tasks as well as the human body’s own anatomical/bodily constraints. Our hypothesis is reinforced by participants often drawing from their life experiences to make sense of the gesture, whether it is *“from years of eating at a table and cleaning these crumbs”* or *“maybe experience comes in thinking what would you try and arrange something in that way? A book?”*

We have schematized these strategies in Figure 3 in a pseudo-network to show how some of the features that emerged in our data can be used to infer transitive actions and adapted to more formal or computational implementations. Each node in the network indicates a discriminating feature category (e.g. “Arm.Mvmt”) and some

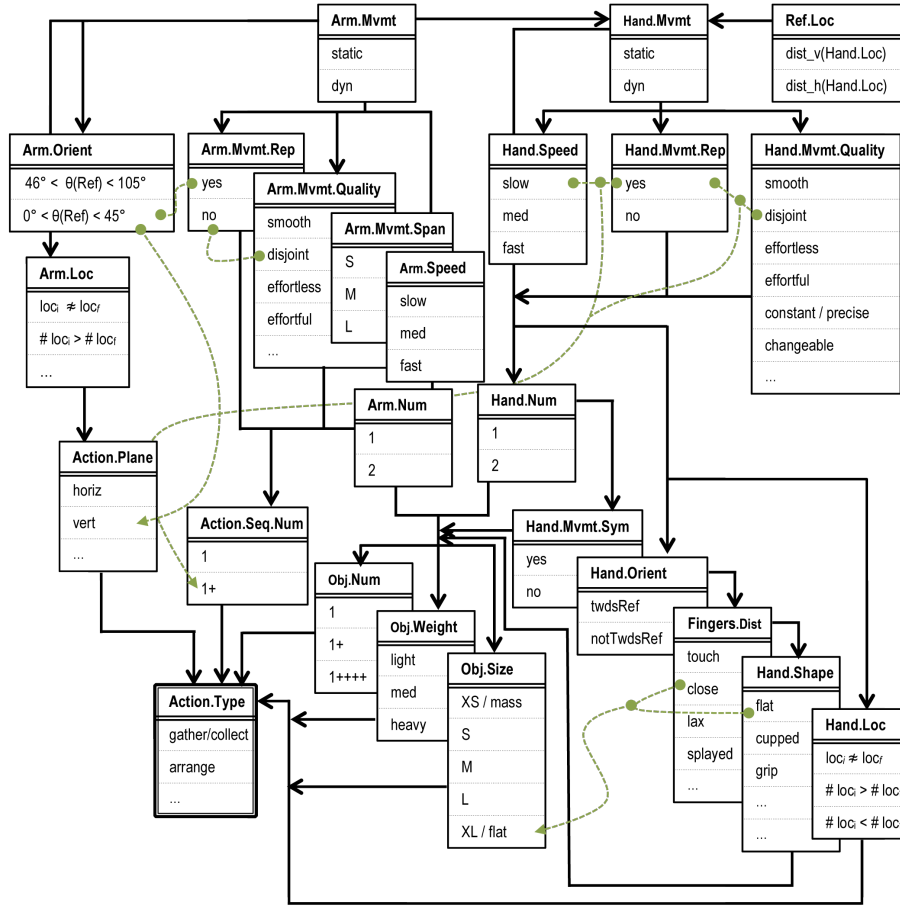


Figure 3: Gesture interpretation strategy model with a partial illustration of ClearBoard (dotted green lines)

of its possible values (“static”, “dynamic”) that could be used in gesture recognition models. Arrows indicate the transitions through the nodes of the recognition network. Note that there can be multiple observation nodes at any given time point.

Implications for gesture interpretation models

In developing interpretation models for gesture recognition systems, previous work has suggested that a gesture module should capture position data of user’s hand shape, orientation, location and movement characteristics, which can then be interpreted [16,17,29]. Findings from this study suggest that people use a consistent set of interpretation strategies, as schematized in Figure 3, which can provide insights into the parameters necessary to aid in reducing uncertainty and in inferencing in gesture recognition and interpretation systems. To draw on Bobick’s [3] distinctions of automatic motion recognition domains, we argue that action interpretation is only possible based on the integrality of movement (motion primitive with the clearest machine interpretability), activity (sequence of movements), and action (interpretation of higher-level motions often requiring context). Another important finding of this study is that there seems to be a limited possible set of interpretations for any partially ambiguous gesture. This implies that there are key decisive features or sets of features, which can provide the

maximal discriminatory power. A more comprehensively mapped out set of key features and strategies could provide a multi-attribute decision making model. For example, one such key subset of features could be hand orientation and arm movement direction in determining action goal: 1) if the palm is facing the user and hand is moving away from the user, then the object is being reached for; 2) if the palm is facing a particular direction but moving towards that direction, it may suggest that the object is being moved towards that direction.

While interpretation is context-dependent, the implications of this finding is that within an application domain where the designer is aware of the nature of the objects and possible manipulations required, a limited set of gesture variations and interpretations can be easily extracted for an interpretation module by 1) obtaining user-defined gestures and 2) naïve observer-driven interpretations of gesture features.

Applications

Since the gestures we explored were transitive actions (manipulation of objects), this study directly demonstrates the methodological utility of using human-based strategies to aid gesture interpretation in serious gaming, such as R.O.G.E.R. [24], or entertainment games, which both often require selection and manipulation of objects. Translated to rehabilitation game application, our approach would involve

collecting user-centric gestures to identify patient vs. healthy group differences and individual healthy variations in a given set of actions. Aspects that are found to be different between the two groups' actions would be more reflective of "true" errors, which the gesture interpretation system could then use to adjust the levels of game difficulty and even provide the player specific feedback or situations which would train those aspects further.

Our findings have implications not only for gesture interpretation modeling, but also for gesture vocabulary design. Given that the principles of gesture interpretation in this study are drawn from natural unstaged gestures, user-defined gestures that allow for intuitive and natural gesture production can be implemented more successfully. This in turn can enhance user experience and increase guessability [37] for first-time or occasional users.

Limitations and Future Directions

Humans are statistical processors. Of course, even though the study was designed to draw on what visual primitives and low-level strategies people may be tapping into in order to interpret the transitive action gestures presented, it is never certain that these are the actual strategies and that people are not availing of other unspoken but crucial inferencing. One could also argue against the validity of using classifiers based on human interpretation in the first place. Like the development of automatic speech recognition systems, we expect purely stochastic models to outperform top-down-heavy models, but we also expect stochastic models to plateau and lack the flexibility of hybrid models, which are essentially what humans are. As one of the most flexible pattern recognizers, we humans are a resource to tap into. The validity of this approach is a topic of debate in modeling and system performance assessment but has its strong supporters. For example, using people's behavior and intuitions to benchmark models and systems is argued for in [1] who used human ratings of the acceptability of signing variability, and the discrepancies of these ratings from computer ratings, to benchmark the representability of test data and test systems.

The substantial level of consistency across people in interpreting the gestures used in this study warrants serious consideration of people's interpretation strategies and approaches in building gesture interpretation models. Future studies are required to test the accuracy of the gesture interpretation strategies found here on different corpora, to also extend beyond transitive actions. These additional studies can aid in further developing the preliminary gesture interpretation strategy model proposed in Figure 3 into a Bayesian model or a multi-attribute decision making model in order to compare with other gesture recognition algorithms.

CONCLUSION

In this study, we explored the value of bootstrapping onto how people recognize and interpret unstaged gestures of

transitive actions. We have shown that people are good at interpreting unstaged gestures given a specific context, despite individual differences in how these gestures are performed. While this work was done with no specific application in mind, it aimed to 1) put together a set of universal and fundamental gesture interpretation strategies that can be used to improve high-level interpretation models and 2) propose and provide validation of a method/approach to building gesture interpretation models based on how humans decode gestures. Our findings achieved both these aims. Being able to have a systematic set of strategies to interpret unstaged gestures implies that we can exploit user-generated gesture vocabularies for interaction with systems that will allow us to retain naturalness and intuitiveness, as well as handle individual differences. It advocates an interface design approach for touchless gesture vocabularies which is based on what is most intuitively performed with little cognitive effort, and thus moves away from ad-hoc gesture choices. Thus, the implications of this work for design are 1) at the backend feasibility level (gesture interpretation system), 2) in a holistic design approach, and 3) at the interface, namely the gesture vocabulary design itself (what types of gestures can be considered that balance practical implementation yet maximize user experience).

ACKNOWLEDGMENTS

This work has been funded by the Excellence Initiative of the German Federal and State Governments.

REFERENCES

1. Arendsen, J., Lichtenauer, J.F., ten Holt, G., van Doorn, A. J. and Hendriks, E. A. Acceptability ratings by humans and automatic gesture recognition for variations in sign productions. In *Proc. of 8th IEEE Int. Conf. on Automatic Face & Gesture Recognition*, (2008), 1-6.
2. Bolt, R.A. and Herranz, E. Two-Handed Gesture in Multi-Modal Natural Dialog. *Proc. UIST '92*. ACM Press, 7-14.
3. Bobick, A. Movement, Activity, and Action: The Role of Knowledge in the Perception of Motion. *Phil. Trans. Royal Society London B*, 352, pp.1257-1265, 1997.
4. Bressem, J. Notating gestures – Proposal for a form based notation system of co-verbal gestures. Unpublished manuscript, 2008.
5. Cadoz, C. Le geste canal de communication homme/machine: La communication instrumentale. *Techniques et Sciences Informatiques 13* (1994), 32-61.
6. Efron, D. *Gesture and Environment*. New York: Kings Crown Press, 1941.
7. Ekman, P. and Friesen, W. The repertoire of nonverbal behavior: Origins, usage and coding. *Semiotica 1* (1969), 49-98.
8. Feyereisen, P., Van de Wiele, M., and Dubois, F. The meaning of gestures: what can be understood without speech? *Cahiers de Psychologie Cognitive. European Bulletin of Cognitive Psychology* 8, 1 (1988), 3-25.

9. Freedman, N. The analysis of movement behavior during the clinical interview. In Siegman, A. W. and Pope, B. (Eds.), *Studies in dyadic communication*. Pergamon, New York, 1972.
10. Grandhi, S.A., Joue, G., Mittelberg, I. Understanding Naturalness and Intuitiveness in Gesture Production: Insights for Touchless Gestural Interfaces. In *Proc. of the 2011 Annual Conference on Human factors in computing systems*, ACM, New York, NY (2011), 821-824.
11. Guest, A.H. *Labanotation: the system of analyzing and recording movement*. 4th ed. Routledge, New York, 2005.
12. Hadar, U. and Pinchas-Zamir, L. The Semantic Specificity of Gesture: Implications for Gesture Classification and Function. *Journal of Language and Social Psychology* 23, 2, 204-214.
13. Hauptmann, A.G. Speech and Gestures for Graphic Image Manipulation. In *Proc. CHI '89*. ACM Press (1989), 241-245.
14. Hummels, C., Stappers and P.J. Meaningful Gestures for Human Computer Interaction: Beyond Hand Postures. In *Proc. FG'98. IEEE Computer Society Press*, Los Alamitos (1998), 591-596.
15. Kendon, A. How Gestures can become like words. In Poyatos, F. (Ed.), *Crosscultural Perspectives in Nonverbal Communication*, Toronto: C.J. Hogrefe (1988), 131-141.
16. Koons, D.B., Sparrell, C.J. and Thorisson, K.R. Integrating simultaneous input from speech, gaze and hand gestures. In M. T. Maybury (Ed.), *Intelligent multimedia interfaces*. MIT Press (1993), 257-276.
17. Kopp, S. The spatial specificity of iconic gestures. In *Proc. KogWis05*, Schwabe (2005), 112 - 117.
18. Krauss, R.M., Dushay, R.A., Chen, Y. and Rauscher, F. The Communicative Value of Conversational Hand Gestures. *Journal of Experimental Social Psychology* 31, (1995), 533-552.
19. Laban, R. von. *The mastery of movement on the stage*. MacDonald & Evans, London, 1950.
20. Lausberg, H. and Sloetjes, H. (2009). Coding gestural behavior with the NEUROGES-ELAN system. *Behavior Research Methods* 41, 3 (2009), 841-849.
21. Lee, J. C. In search of a natural gesture. *XRDS* 16, 4 (June 2010), 9-12.
22. Martell, C. *FORM: An experiment in the annotation of the kinematics of gesture*. Unpublished Doctoral Dissertation (2005). University of Pennsylvania.
23. McNeill, D. *Gesture and Thought*. University of Chicago Press (2005).
24. Microsoft TechDays 2011. <http://capecalm.tv/2011/03/01/microsoft-supports-roger-1st-medical-serious-game-on-kinect-enmicrosoft-techdays-2011/>
25. Müller, C. *Redebegleitende Gesten. Kulturgeschichte – Theorie – Sprachvergleich*. Berlin Verlag (1998).
26. Osiurak, F., Jarry, C., Baltenneck, N., Boudin, B., and Le Gall, D. Make a gesture and I will tell you what you are miming. Pantomime recognition in healthy subjects. *Cortex* (2001), doi:10.1016/j.cortex.2011.01.007.
27. Poggi, I. From a Typology of Gestures to a Procedure for Gesture Production. In Wachsmuth, I. and Sowa, T. (Eds.) *Gesture and Sign Language in Human-Computer Interaction*. Springer (2002), 158-168.
28. Quek, F. Hand gesture interface for human-machine interaction. In *Proc. of the Virtual Reality'93 Conference*. New York (1993), 13-19.
29. Sowa, T. and Wachsmuth, I. Interpretation of Shape-Related Iconic Gestures in Virtual Environments. In Wachsmuth, I. and Sowa, T. (Eds.) *Gesture and Sign Language in Human-Computer Interaction*. Springer (2002), 21-33.
30. Sowa, T. The Recognition and Comprehension of Hand Gestures – A Review and Research Agenda. In Wachsmuth, I. and Knoblich, G. (Eds.) *Modeling Communication*, Springer (2008), 38-56.
31. Stokoe, W. *Sign Language Structure*. University of Buffalo Press (1960).
32. Streeck, J. *Gesturecraft. The Manufacture of Meaning*. John Benjamins, Amsterdam, 2009.
33. Strauss, A. and Corbin, J. 1990. Basics of Qualitative Research: Grounded Theory Procedures and Techniques. Sage, Newbury Park, CA.
34. Wachs, J.P., Kölsch, M., Stern, H. and Edan, Y. Vision-based hand-gesture applications. *Communications of the ACM* 54 (February 11), ACM Press, 60-71.
35. Wexelblat, A. Analysis of Natural Gestures at the User Interface. *ACM Transactions on Computer-Human Interaction (ToCHI)*, September, 1995.
36. Wilkins, D. Why pointing with the index finger is not a universal (in sociocultural & semiotic terms). In Kita, S. (Ed.), *Pointing: where language, culture, and cognition meet*. Lawrence Erlbaum Associates, Mahwah, NJ (2003), 171-215.
37. Wobbrock, J.O., Morris, M.R. and Wilson, A.D. User-defined gestures for surface computing. *Proc. of the 27th Int. Conf. on Human Factors in Computing Systems*, ACM, New York, NY (2009), 1083-1092.