Understanding Mid-Air Hand Gestures: A Study of Human Preferences in Usage of Gesture Types for HCI

Roland Aigner[1], Daniel Wigdor[2], Hrvoje Benko[3], Michael Haller[4], David Lindlbauer[4], Alexandra Ion[4], Shengdong Zhao[5], Jeffrey Tzu Kwan Valino Koh[6]

[1]Ars Electronica Futurelab
Ars Electronica GmbH
Linz, Austria
roland.aigner@aec.at

[2]Department of Computer Science
University of Toronto
Toronto, ON
dwigdor@dgp.toronto.edu

[3]Microsoft Research
Microsoft Corporation
Redmond, WA
benko@microsoft.com

[4]Media Interaction Lab
University of Applied Sciences Upper Austria, Hagenberg, Austria
mi-lab@fh-hagenberg.at

[5]Department of Computer Science
National University of Singapore
Singapore
zhaosd@comp.nus.edu.sg

[6]Interactive and Digital Media Institute
National University of Singapore
Singapore
j.koh@nus.edu.sg

ABSTRACT

In this paper we present the results of a study of human preferences in using mid-air gestures for directing other humans. Rather than contributing a specific set of gestures, we contribute a set of *gesture types*, which together make a set of the core actions needed to complete any of our six chosen tasks in the domain of human-to-human gestural communication without the speech channel. We observed 12 participants, cooperating to accomplish different tasks only using hand gestures to communicate. We analyzed 5,500 gestures in terms of hand usage and gesture type, using a novel classification scheme which combines three existing taxonomies in order to better capture this interaction space. Our findings indicate that, depending on the meaning of the gesture, there is preference in the usage of gesture types, such as pointing, pantomimic acting, direct manipulation, semaphoric, or iconic gestures. These results can be used as guidelines to design purely gesture driven interfaces for interactive environments and surfaces.

INTRODUCTION

Recent advances in computer vision, particularly in real-time hand [26] and body tracking [23] are empowering computers and intelligent environments to recognize human gestures from a distance. These novel technologies reduce barriers to interaction and increase the input bandwidth between the user and the computer, without requiring the user to wear or acquire a tracked object. Since surface and tabletop systems have expanded to include proximity sensing [13], there is potential for computer vision based mid-air interaction to become a part of tabletop computing [12] and large-scale displays [3]. The recognition of mid-air gestures enables designers to create interfaces that enable explicit control of such systems.

As is true of any new input modality, there is a need for extensive design of the physical primitives which make-up the interaction [1]. Since the WIMP GUI is still a prevalent metaphor in computer interfaces, a simple approach is to adopt its "point and click" and related primitives. For example, Vogel and Balakrishnan [25] built a pointing-and-click interface for large screen displays which simulates mouse input. They identified issues with precision in pointing, ambiguity of finger movements, and lack of physical feedback. These issues are critical for the control of a point-and-click, mouse-based UI, but offer little guidance to designers wishing to develop new, possibly more suitable input primitives. In particular such transformation may reduce much of the richness of the new input modality, and sacrifice available bandwidth [8].

Figure 1. Two participants in our study: one was gesturing to direct the actions of the other. We excluded the audio channel in order to force and observe common gestural behavior.

To create interfaces that leverage a gestural input paradigm, there is a need to build gestural interfaces from scratch – including the design of new physical primitives. One common approach has been to elicit gestures applying the user defined interfaces (UDI) methodology. This technique is useful in determining a novice's "first guess". However, researchers have found that there is little consensus among users in association between gestures and their expected effect [24, 28]. In contrast, designed languages can be formed to avoid conflicts, but also to be ergonomic and high-bandwidth. In view of this, we believe that gesture languages need to be designed, rather than observed. A key

Microsoft Research Technical Report MSR-TR-2012-111.

limitation, however, is the lack of guidelines to aid designers in creating those languages.

To overcome this issue, we seek to bridge the gap between those two methodologies by investigating what linguists term *gesture types* [19]. Although gesticulation humans perform during speech is quite different in its nature, we build on this traditional linguistic practice for a more systematic investigation of gestures in HCI. Rather than seeking to construct a dictionary of particular gestures for a given set of functions, classifying physical actions enables us to create middle ground between elicited and designed gestures. While a UDI methodology might provide viable but inconsistent results, our approach enables for a thorough *design* of a gesture language by providing guidance that, for example, a gesture for a particular action should be chosen to involve unimanual static hand postures or bimanual movements, or that it should use pointing, descriptions of the objects, or semaphoric codes.

In order to produce our gesture types, we conducted a study. Our methodology was similar to that of UDI, in that participants were given a 'before' and 'after' situation, where user-selected gestures were used to make the transition between those two states [28]. Unlike previous work, our study provided a *recognizer* – a second participant, who was observing, via a muted video feed, the actions of the first participant. The first participant was tasked with guiding the observer in completing various tasks, using only gestures shown over a video feed (see Figure 1), thus lacking speech. The 'human' recognizer enabled participants in our study to perform gestural interaction within a closed feedback loop, enabling refinement over time. The choice of using a human instead of an artificial gesture recognizer was also taken to prevent from biased behavior and to provide an unconstrained environment to the directing participant. While this study setup is similar to previous works in CSCW [10, 17], the goals and information collected were different. In our methodology the actor was meant to communicate only using mid-air gestures, in order to accomplish a task in a 'virtual' environment, just like in controlling an interactive computing system.. Except for visual feedback showing the result of the given input, the communication was intentionally one-way, as with a unimodal gesture interface. Additionally, the study was designed not to address high-level human communication tasks, but low-level actions such as manipulating objects, guiding avatars, and describing simple shapes, in order to collect gesture data for each of them for further analysis. As we will discuss, the refinement process saw the introduction of new types of gestures, ad hoc formalisms were created, and whole gesture languages were observed.

During this process, we collected approximately 5,500 gestures from 12 participants during 6 different scenarios. We analyzed those in terms of *expected effect* and *gesture type*, using our own classification scheme. We also collected data about the number of hands used, by splitting into unimanual and bimanual gestures. The results indicate that for each expected effect of a gesture, there are preferences in both type and number of hands. These preferences, reported in detail in this paper, represent two contributions. First, they provide a baseline measurement of conventional human behavior (after a short grounding process) which goes beyond the 'first guess' provided by UDI. Second, the results, in and of themselves, form the basis for the development of guidelines which designers can use to design their own gesture languages, taking into account not only user preference, but all the factors relevant to the context of the language [27].

In this paper, we present the design of the study, our classification system, and our findings about gesture type usage, including user choices for uni- vs. bimanual gestures. Our contribution is our classification scheme, and knowledge of the suitability of gesture types for the chosen gesture effects.

RELATED WORK

Gesture Classification

Many different classifications have been proposed for speech-related hand movements [5, 19]. Many of these are modifications of Efron's [6] pioneering work about conversational behavior between Jewish and Italian immigrants in New York City. For example, Kendon [16] arranges gestures along a continuum, which reflects the gesture's relation to accompanying speech. He proposed classes of *gesticulation* (beat, cohesives), *language-like* (iconic), *pantomimes*, *emblems* (deictic), and *sign language* (symbolic), with the necessity of accompanying speech to facilitate communication declining in this order.

Cadoz [5] classifies gestures according to their function into *semiotics*, *ergotic* and *epistemic* gestures. McNeill [19] separates them into the classes of *iconics*, *metaphorics*, *beats*, *cohesives*, and *deictics*. These authors investigate the multidisciplinary research field of human gesturing. As a result, their classifications are created for gestures accompanied by speech, such as in narration, and are therefore not appropriate for our domain. In contrast, Karam and schraefel [15] created a more extensive taxonomy, especially tailored for HCI, and serves as the basis for our classification. For what they call 'gesture styles', they propose the classes of *deictic*, *gesticulation*, *manipulation*, *semaphores*, and *sign language*. In this model, *gesticulation* is also equated with *iconic* or *depictive* gestures, which are used to depict physical shapes and forms referred to by speech.

Our work adopts elements of where appropriate to build our classification system, primarily by extending Karam and schraefel's model. We also build on other work in HCI.

Gestures in HCI

Many novel interfaces, such as those for in-air gesture detection [2, 3], multitouch [22], tangible [14], and organic [18] user interfaces try to overcome limitations of WIMP interaction. Interfaces combining hand-gesture and speech try to provide means of control by mimicking human-to-human communication [4, 24].

In the present work, we are focusing on purely gesture-driven human computer interfaces, without coverbal gesticulation. A notable demonstration of unimodal mid-air gesture usage is found in the demos for the g-speak platform (oblong.com). The designers show several sample applications for interacting with virtual environments, including drawing, image and video editing, and navigation in 3D space. They use bimanual interaction extensively, primarily to establish reference frames. This work extends earlier work in HCI, such as Charade [2], as well as work specifically dedicated to the creation of gesture languages.

Creating Gesture Languages

The study of the construction of gesture languages can be roughly divided into two camps: one viewpoint is that gesture languages must be *elicited*, through the study of potential users in their natural environments, or through user definition [20, 21, 28]. This approach prioritizes the probability that a new users' first guess of a gesture will yield the desired effect. The alternative camp proposes that gesture languages must be *designed* through an examination of the unique physical and psychological implications of a given input device and context of its use, and then taught to the user [1, 9, 27]. Their view is that 'guessability' may be only one of several factors in the design of a language.

With our work we attempt to bridge the gap between these camps. By building on UDI methodology, we are seeking for typical user behaviors in terms of hand usage for constrained communication. These behaviors, however, are in the form of *types*, rather than particular gestures. Through the definition of these classes, we seek to provide generally useful guidance on the issue of 'guessability' to those practitioners and researchers who are considering this and several other factors in *designing* gestural languages.

In seeking gesture types, we follow the general approach proposed by Nielsen et al. [21]. They suggest that in order to build a gesture interface it is necessary to first identify the functions that will be evoked (or 'effects' of the gestures). Subsequently, the most appropriate gesture for each of those functions has to be found (in our case, gesture *types*). In our methodology, we select 10 common functions ('effects') required in many interactive systems, and elicit and classify gesture performed by our participants into gesture types.

TARGET GESTURE EFFECTS

A difficulty in defining a classification of gesture types is the need to develop generic and representative gesture effects (system or application functions). Foley et al. [7] identified six generic tasks, reflecting the user's intentions: *select position* and *orient* an object, *ink* (draw lines), *enter text*, and *specify scalar values*. This work is somewhat limited for our purpose, because it assumes mouse-like primitives. For example, it assumes that possible effects such as *accepting* or *refusing* will be composed of other primitives, such as selection of a menu item, by selecting (activating) a button, or by entering text.

To achieve coverage of frequently desired gesture effects, we attempted to find as many effects as possible: during a brainstorming session with four researchers, we gathered a large set of tasks and actions for *content creation*, *manipulation*, *management*, and *navigation* in the physical world. We do not claim the resulting list to be complete, for example it excludes tasks for heavy data input, such as text input. However it should cover most actions one would like to perform with a gesture interface. Next, we split high-level actions into atomic physical acts, in order to find a basic set of effects. For example, the task of arranging several blocks by their color can be decomposed into several acts of moving an object, which in turn can be split into grasping, translating and releasing. This is similar to the logical primitives approach taken by Foley et al., except that our primitives are based on physical acts, rather than an assumed interaction model.

From this process, we formed a list of 10 fundamental, non-overlapping effects:

- *Select:* in contrast to identification, selecting refers to the actual action of *picking* or *grasping* of physical entities.
- *Release:* this effect is the complement to selecting, thus it expresses the *depositing* of previously selected entities.
- *Accept:* this represents *agreeing* in binary queries and includes similar meanings such as *approving* and *affirming*, as well as *continue* in the sense of "you are doing it right, go on".
- *Refuse:* the complement effect to accepting is often used for answering "no" in binary queries, but it also includes "wrong", "stop", or "undo".
- *Remove:* in contrast to releasing, removing entities means to *dismiss* them, to *erase* them, or even to move them off the table.
- *Cancel:* this effect represents *pausing* or *stopping* of a currently ongoing process, or even *terminating* the whole overall task.
- *Navigate:* in contrast to translation, navigation refers to the *guiding* (both movement and rotation) of a representation of the actor's hand or tool, such as a pen. This representation is considered to be similar to a mouse cursor.
- *Identify*: this effect includes *identification*, *indication* or *description* of an entity or location.

- *Translate*: other than navigation, translation refers to the movement or the dragging of a physical entity.
- *Rotate:* this effect represents the action of rotation of physical entities.

Having defined the atomic effects of gestures we wished to evaluate, we turn our attention now to the design of our experiment, which was meant to answer the question: *what* types *of gestures do users perform in order to achieve these effects?* We extended the UDI methodology in order to learn not only 'first guess' behavior, but to record the results of refinement of techniques over time – given the presence of a closed feedback loop.

EXPERIMENT FOR COLLECTING GESTURE DATA

The goal of this study was to determine mappings of user-defined gesture types and number of hands used to each of the effects. To do this, we constructed tasks which, in order to accomplish them, would include the 10 effects we identified. The experiment was done in pairs of two: one was designated the *actor*, who performed gestures, and the *wizard*, who observed the gestures and carried out the effects that they understood. The wizard was provided with the tools needed to complete each task, but was not made aware of the goals, or which of the tools would be required to accomplish them. There was no *a priori* communication about the tasks between the wizard and actor, nor were any guidelines provided to shape their communications. The wizard was not given aid to understand what task would be performed, beyond the placement of the necessary tools within their reach.

Participants

Twelve participants (8 male), with ages ranging from 23 to 54 years (M = 30.92, SD = 10.21) were recruited from a Central European community. To avoid technical bias, we chose candidates with affiliations ranging from school teacher, lawyer, and process manager to IT technician and software developers. Mean self-reported daily computer usage was 6.07 hours (SD = 3.08). 7 participants reported frequent use of gesture based touch interfaces for a duration ranging from 4 months to 4 years (M = 2.17, SD = 1.59), such as smartphones and tablets, 6 of those also had used mid-air gesture interfaces (3 of them frequently) prior to the study, such as Microsoft Kinect, WiiMote and PlayStation Move. The remaining 5 did not have mentionable experience with gesture interfaces whatsoever. The participants were divided into 6 groups of 2 strangers. To eliminate biasing by gesture vocabularies established during the earlier trials, the task order was counterbalanced using a 6×6 Latin Square.

Design

Both actor and wizard were placed in separate rooms to prevent them speaking or using other forms of auditory communication. Both participants were connected by a video chat system (Skype), with the audio channel disabled, so the only reasonable high-bandwidth channel left was gestures. In our experiment gesture communication was meant to be one-way, thus from actor to wizard, while the desired closed feedback loop was established by showing the wizard's actions to the actor. Both participants were placed in front of a camera/screen setup, as shown in Figure 2. The actor was only able to see the table, all tools, and the wizard's hands; the wizard was shown both the upper body and the head of the actor. In order to help the wizards to locate their work, the borders of the area visible to the actor were marked on the table.

For each task, the actor was provided with the goal of the results to be achieved. Depending on the task, this was either a drawing or a photograph. The image was placed next to the actor's screen throughout the task.

Participants had to finish each task within 10 minutes; after that time, the task was suspended. Since we intended to collect as many gestures as possible, we designed the tasks in a way that prevented their accomplishment during the allotted time. All participants were told not to hurry, to communicate at a convenient speed, and simply try to complete as much of each task as possible.

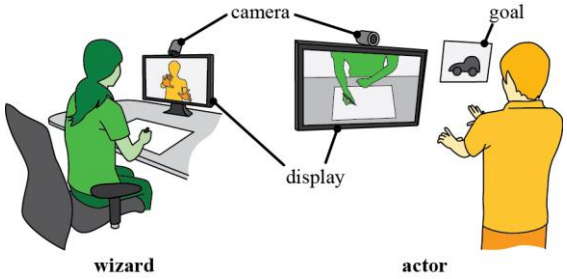

Figure 2. In our study, 2 participants, the *actor* and the *wizard*, were separated spatially, and were only allowed visual communication using video conferencing.

After each task, actor and wizard switched roles and moved on to the next task. This means that each participant acted as both actor and wizard, providing greater variety of gesture types for our analysis.

Tasks

We chose the 6 tasks to be completed during the study. Since our focus is on gesture interfaces for HCI they were not chosen to be particular meaningful everyday human-human communication tasks but to cover a majority of low-level action, thus our 10 chosen effects. For example building a figure out of toy blocks would at least require (i) identification of an object, or alternatively navigation of the wizard's hand, (ii) grasping, (iii) translation and rotation of the object, and (iv) releasing. Samples of the goal images for each of the 6 types of tasks are shown in Figure 3.

Pilot studies aided in refinement of the tasks to fine-tune their difficulty and diversity. The selected six tasks were:

- *Blocks:* the objective was to reconstruct a given figure out of DUPLO blocks (Figure 3a), and to dismantle the construction again after 9 minutes.

- *Drawing:* the goal was to draw a number of figures onto a blank piece of paper (Figure 3b). After 6 minutes, the actor had to instruct the wizard to memorize everything drawn so far and to redraw it on a new sheet of paper. This was included to find gestures for the effect *save*.

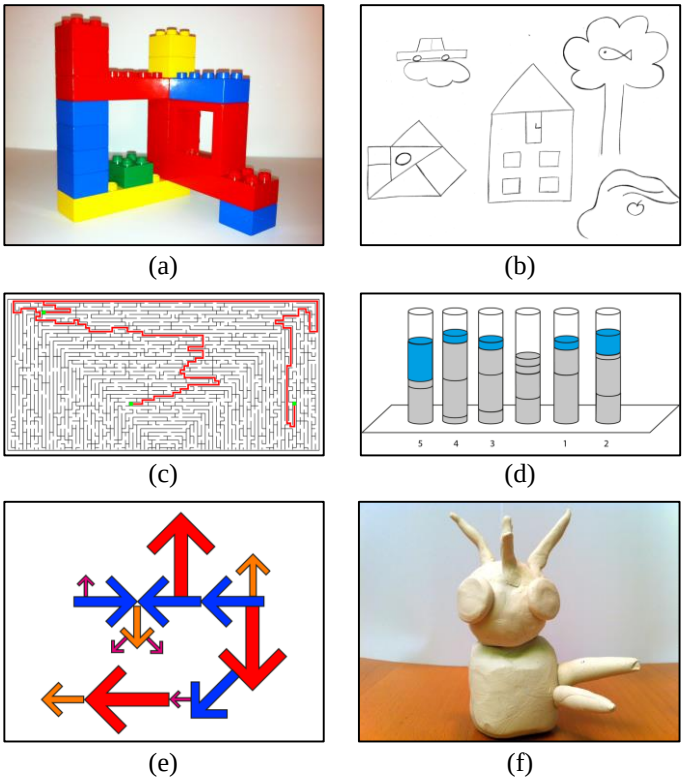

Figure 3. Samples of the goal images provided to each *actor*, and hidden from each *wizard*.

- *Maze:* the actor had to make the wizard draw a series of preselected paths through a maze (Figure 3c).
- *Meter:* six glasses with liquid were to be filled to specified levels and in a specified order. There were several intermediate states to accomplish (Figure 3d).
- *Arrange:* the objective was to recreate a specific arrangement of colored cutout arrow shapes, as depicted on a printout presented to the actor (Figure 3e).
- *Sculpt:* wizard was to recreate an abstract model out of clay, as depicted in a set of photographs provided to the actor (Figure 3f).

To avoid agency by the wizard, we incorporated aspects into each task which were unlikely to be anticipated, such as moving backwards in the maze task, drawing a fish onto a tree, and dismantling the incomplete block figure.

We recorded both the actor's and the wizard's video stream and used them for our analysis. Twelve participants directed three tasks of ten minutes each, which resulted in six hours of video data. In order to classify the gesture types, we developed the following classification scheme.

GESTURE TYPE CLASSIFICATION SCHEME

In a pilot study, we used the taxonomy of Karam and schraefel [15]. As we conducted the coding of the video, we found the need to modify Karam & schraefel's scheme. In particular, it was clear that their *gesticulation* class failed to capture the distinction between *iconic* and *pantomimic* gestures, and so was split into those two categories. We further observed iconic gestures, like semaphoric in the earlier scheme, could be further divided into *static* and *dynamic*, though *stroke* did not apply.

As formal sign languages have to be learned just like any spoken language, we did not expect them to appear during the study, so we omitted them in our classification scheme. Since they are especially designed to be versatile and high-bandwidth, they are rather complicated and inconvenient to use and learn [21]. As a result, we find them to be unsuitable for guessable and intuitive gestures.

Additionally, we replaced the term *deictic* by *pointing*, since our application did not use speech communication, and thus there can be no deixis per se. Our classification scheme, which is a modified version of the previous work, is shown in Figure 4. We now explain each of the gesture types in greater detail, and then present the results of the coding exercise.

Pointing

Pointing is used to indicate objects and directions, which does not necessarily involve a stretched index finger. It may also be performed with multiple fingers, the thumb, a flat palm, etc. Note that gestures which resemble pointing but which are intended to mean "yes, you understand" were labeled as semaphoric, given their loaded meaning.

Semaphoric

Semaphoric gestures are hand postures and movements, which are used to convey specific meanings. Mostly gesture and meaning are completely unrelated and strictly learned. Therefore, we consider semaphorics to be the gestures most dependent on the actor's background and experience.

Static semaphorics are identified by a specific hand posture. Examples would be a thumbs-up, meaning "okay", or a flat palm facing from the actor, meaning "stop".

Dynamic semaphorics convey information through their temporal aspects. A circular hand motion meaning "rotate" is of this type, as well as a *repeatedly* flicking or waving of the hand sideward, meaning "no".

Semaphoric *strokes* represent hand flicks and are similar to dynamics, since they are identified only by hand motion, but which are single, stroke-like movements. They may be compared to the familiar touch and stylus gestures on iOS and Windows. An example would be a *single*, dedicated sideward flick of the hand, meaning "dismiss this object".

Pantomimic

Pantomimic gestures are used to demonstrate a specific task to be performed or imitated, which mostly involves motion and particular hand postures. They usually performed by an

actor without any objects actually being present, such as filling an imaginary glass with water, by tilting an imaginary bucket. They often consist of multiple low-level gestures, e.g., (i) grabbing an object, (ii) moving it, and (iii) releasing it again. We code these as a single pantomimic gesture.

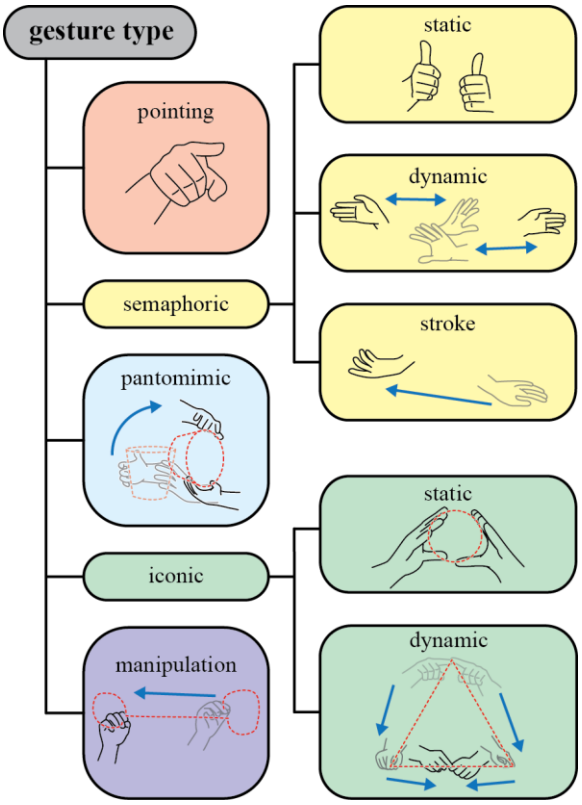

Figure 4. The classification we used to analyze gestures in this research, including examples for each of the gesture types.

Iconic

Iconic gestures are used to communicate information about objects or entities, such as specific sizes, shapes, and motion paths.

Static iconics are performed by static hand postures. In contrast to static semaphorics they do not rely on a commonly known vocabulary, instead they are rather spontaneous, such as forming an "O" with index finger and thumb, meaning "circle".

Dynamic iconics are often used to describe paths or shapes, such as moving the hand in circles, meaning "the circle". Compared to concatenated flicks (which would be semaphoric strokes), the motions are usually performed more slowly. Another difference is that in strokes, the actual range of the movement does not hold information about the action, however in dynamic iconics it does.

Manipulation Gestures

Manipulation gestures are used to guide movement in a short feedback loop. Thus, they feature a tight relationship between the movements of the actor and the movements of the object to be manipulated. The criterion for distinguishing them from pantomimic and dynamic iconic gestures is the presence of the feedback loop. In the case of manipulation gestures, the actor waits for the entity to "follow" before continuing, instead of performing *beforehand*, only causing a reaction *subsequently*.

RESULTS

To review the results of our study, we begin with classifying each of the observed gestures used for each effect.

Gesture Effects

Three researchers collected approximately 5,500 gestures and categorized them using our classification scheme, constantly consulting each other in order to prevent from biasing and diverging interpretations. We also differentiated by unimanual and bimanual gestures. The overall results are depicted in Figure 5. This represents our primary contribution: the types of gesture chosen for each of the desired effects.

Note that some gestures may include elements of more than one type, as actors displayed diverse creativity in composing gestures from different types. This is especially apparent in the case for bimanual gestures, since gestures might be combined, such as expressing "move the round block" by forming a round static-iconic gesture and then pantomiming a movement with that shape. Also, some movements were not gestures to convey meaning, such as when the actors hesitated or when they were irritated. As a result, the sums for any given effect may not add up to 100%.

As predicted, actors used a wide variety of gestures to accomplish the same effect. However, the *type* of gesture that they used was often consistent across time, and participant. Thus, while a classification of particular gestures might find a high degree of variance [28], our results suggest that classifying by type reveals a much greater degree of consistency.

Select

Selection was most often indicated with pantomimic gestures, primarily in the form of "grasping". Bimanual gestures were rarely present for selection.

Release

Pantomimics, semaphoric strokes and iconic dynamics showed high proportions of bimanual acting. Pantomimic gestures were "releasing hand gestures", thus the counterpart of the grasping gestures used for selection. Semaphoric strokes were usually flicks downwards, mostly with stretched palms facing down, indicating placement of the object down onto the table.

Accept

All gestures were semaphoric static, either by showing thumbs-up hand poses, okay-signs (foming an "o" with index finger and thumb), or the previously described pointing-like semaphoric gesture.

Refuse

97% were semaphoric dynamic gestures, either by waving sideward with one or two hands and palms facing down,

while index finger or palm were outstretched, or by waving sideward with the index finger pointing up.

Remove

Semaphoric strokes were sideward flicks. Bimanual strokes mainly showed both hands flicking into the same direction. Unimanual pantomimics were mostly *throw-away*-gestures, e.g., showing the actor throwing an imaginary object backwards, over the shoulder. The high bimanual portion was mainly made up of *put-aside*-gestures, i.e. grasping an invisible object with both hands and releasing it offside.

Cancel

Strokes were generally flicks sideward, with outstretched palms facing down and both hands moved into opposite directions. Semaphoric statics mostly showed hands with outstretched palms facing away from the actor ("stop" gesture). Okay-signs and thumbs-up hand poses were sometimes observed to indicate "Alright, good job, stop now." Semaphoric dynamics were variations of the strokes mentioned above, but with the hands repeatedly waving, instead of flicking once.

Navigate

Navigation was mainly done by pointing into the desired direction, which was mostly accompanied with a rhythmic movement. Manipulation gestures guided the wizard's hand as if it were a remote representation of the actor's hand, comparable to a mouse cursor. Semaphoric strokes were used to indicate the direction to move towards. Bimanual gestures were rarely present.

Identify

While pointing seems to be obvious for identification, iconic dynamic gestures were used to describe the desired object by drawing its shape in mid-air. Semaphoric statics were used to identify objects by showing a number of fingers. Identification like this requires the prior establishment of a code; actors would occasionally index objects by numbering them at the beginning of the task. Iconic static gestures were used to refer to objects by showing their size.

Translate

Semaphoric dynamics were mostly repetitious waving towards the respective direction, with one or both hands. Pointing gestures were mostly rhythmic pointing into the desired directions; this was generally done with one hand. About 10% out of those included self-references, thus using one hand or arm, or even other parts of the body to establish a reference frame. Pantomimic gestures were demonstrations of the translations, including hand gestures mimicking grasping and releasing. In manipulation gestures, actors guided the wizard using slow movements, while the wizard followed accordingly

Rotate

Iconic dynamics were performed by according rotations of the hand, mostly with an outstretched palm, while the hand was considered a representation of the object. In the *arrange* task the tip of an arrow was sometimes indicated by an outstretched index finger or thumb. Pantomimic gestures simply showed the rotations to be performed by a combination of grasping, rotating, and releasing. Most of the semaphoric dynamic gestures were circular motions of the hand, with the index finger or even the palm outstretched. The bimanual portion represents flip-gestures, with both arms moving in a half circle, ending up in crossed arms, referring to a rotation of 180°.

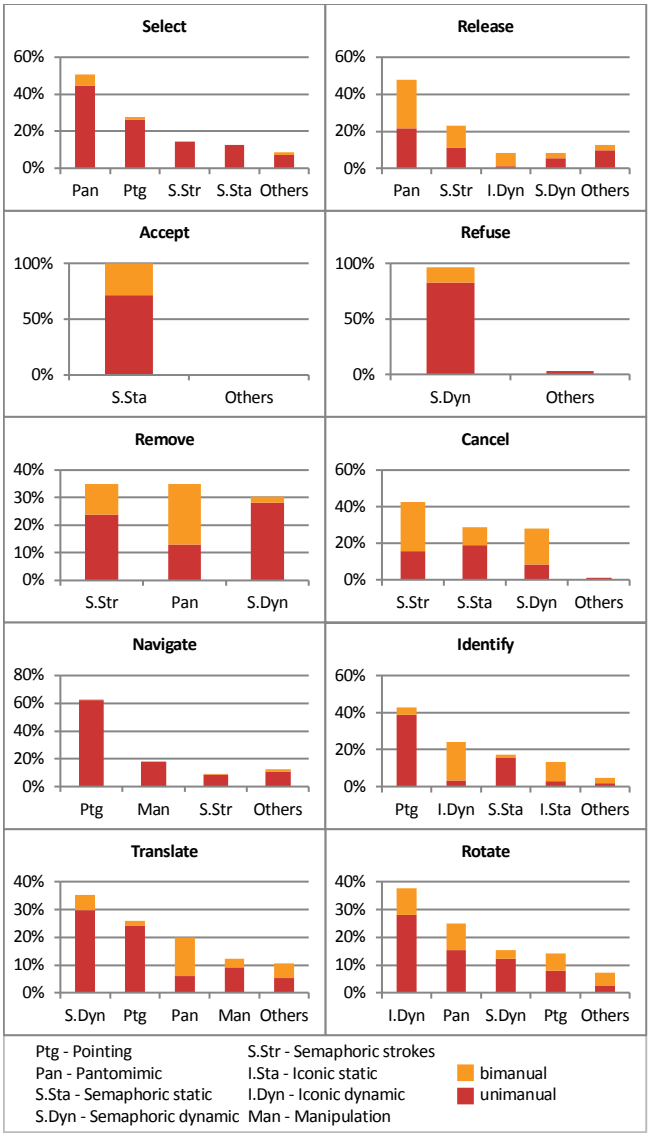

Figure 5. The type histograms for each of the effects.

As intended by the study designers, gesture communication was one-way for most the time. Gestures performed by the wizards, e.g. due to confusions, were insignificant, although the participants were not explicitly advised to refrain. Feedback was mostly given by the effect, visible on screen.

Influence of Participant's Experience

For select, remove, translate, and rotate, significant differences between groups were observed depending on

the participant's prior experience with gesture interfaces. For the rest of the effects, there are only minor differences: there was a relatively low change within type usage (average 4.27%, *SD* = 0.0372), while for each effect, the types with the highest portion were the same for both novices and experts, and the second most prominent types only changed in *navigate* and *deactivate*.

Reference Frames

The usage of reference frames, such as self-references, is shown in Figure 6. They are especially present in iconic static, iconic dynamic and pantomimic gestures, with a notably bimanual dominance, validating previous design efforts [11]. Reference frames using one hand usually involve prior established contexts, such as drawing a rectangle into mid-air with the outstretched finger, representing the drawing canvas, followed by a circular movement in the upper left corner, meaning "draw a circle of this size into this corner of the sheet."

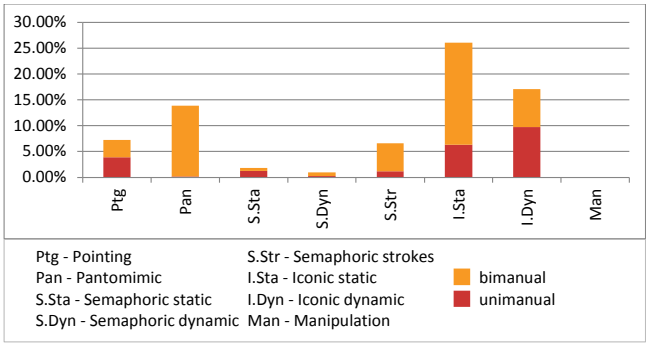

Figure 6. Reference frame usage for each of the types.

It was our goal to record results in the presence of a feedback loop, and without the introduction of a recognizer which would artificially affect results. Nonetheless, that the wizard was a human observer has an obvious impact on our measurements: vocabularies which were easily understood by the counterpart became the most present overall. Some effects saw gesture types change after a short period, as the actor adapted to the preference of the wizard. This was most prominent for *translate*, *rotate*, *remove*, and *cancel*, whereas *accept* and *refuse* did not show significant changes over time. While this grounding process did not introduce significant error into our data, since failing gestures were dropped instantly, it does have an impact on how to interpret our results. Thus, they do not rely on the most *spontaneous* gestures for each action, but on the most *common* instead. It remains an open question as to whether the grounding process would be similar when communicating with an artificial gesture learning system. The degree to which this shift occurred varied across desired effect - the remaining six effects were less affected by this shift. In particular, *accept* and *refuse* did not show significant changes of gesture type over time.

DISCUSSION AND DESIGN IMPLICATIONS

In general, participants displayed a lot of unanticipated behavior. For example, we expected manipulation gestures to be much more present, especially for translation, rotation and navigation. Instead, participants tended to iconically describe motion paths and object shapes, and to demonstrate rotations beforehand. In latter case, a pointing finger seems to be a good way of clarifying the exact angle of rotation whenever the rotated object had a clear front and back side, just like the arrows in the *arrange*-task. If the object did not have such clear distinctions of front and back sides, a pointing finger might still be useful as an "anchor". These results have significant implications for the design of alternative implementations of "direct manipulation".

Some of the results were highly consistent, especially those for *accept* and *refuse*. Other results showed much more variety, which suggests that there is more flexibility in gesture language design for those cases. For example, for *remove*, the three types of semaphoric strokes, pantomimics and semaphoric dynamics were performed at similar rates. This suggests that those types are equally suitable.

The fact that the types for *accept* and *refuse* were distinctly different is particularly interesting, because these two effects are dichotomous. The static hand postures for accepting are completely opposed to the dynamic hand motions used for refusing. This is highly valuable since confusing dichotomous effects would be particularly severe, yet static and dynamic motions can easily be distinguished by a gesture recognition system.

Another set of dichotomous effects is *select* and *release*. While both were mostly accomplished by pantomimic gestures, bimanual acting was highly present in the latter, which could be valuable for distinguishing between them. In contrast to accept/refuse, it would even be possible to distinguish between them by context in many applications.

For identification of objects, pointing is often ambiguous. In applications with many objects located close to each other, it might be valuable to describe objects iconically instead. Alternatively, pointing can also be used to preselect a number of possible candidates (the ones close to the spot the user is pointing to). Those could be augmented, with indices, so the user is able to select the desired object by showing the corresponding number with a semaphoric static gesture, for example by stretching out three fingers to for "object #3". It is important to note that this gesture type was employed despite each participant also discovering the cursor-like 'manipulation' gestures, suggesting a potential preference vs. traditional 'cursor' pointing.

Although there is little consensus between users regarding association between specific gestures and effects, there seem to be some exceptions. One gesture was particularly prevalent throughout all participants: all actors pointed one or both index fingers to their head meaning, "memorize the drawing" in the drawing task.

The processes of translation and navigation seem to be very similar at first sight since both effects refer to movements: those of an object to be manipulated in the former, and those of an avatar or cursor in the latter. By comparing the results, however, we can see that there are significant differences in type. While translation is mostly accomplished by semaphoric dynamic gestures, navigation is generally done by pointing. This indicates that there is a substantial difference in how to instruct movements, depending on the relation between actor and entity.

Actions for *object manipulation* (translations and rotations) are mostly accomplished using hand motion, thus by dynamic gestures. However, translation is done by semaphorics, e.g., by repetitious waving, which has to be explicitly terminated when the target is reached. In contrast, rotation already shows one concluded movement, e.g., by rotating the palm around 45°, which already contains all the information required for performing the action from initialization of the action to its conclusion.

Pointing is highly present in both navigation and identification types. While this might indicate danger of confusion on first sight, identification can mostly be distinguished from navigation by watching the hand pose: while the actor points directly at the objects (thus on the screen) in the former, s/he points sideward (left, right, up, down) in the latter. Also the context (e.g., an application's state) might be an indicator in many cases.

Semaphoric strokes are mainly used for removing and cancelling. While there are other types in removing, which are almost as frequent, the result for cancelling is much more distinct, and is highly bimanual.

Overall, the extent of bimanual gesticulation is clearly noticeable. This implies that it is appropriate to extensively incorporate it into HCI interfaces. Bimanual gestures are highly popular in pantomimic and iconic gesture types, with the exception of those used for the effect of navigation and selection.

Limitations of this Work

A potential limitation of this work is possible dependence on culture. So far, we did not compare actors across different countries and cultures. While particular gesture selections may vary across cultures, we hypothesize that only minor differences in gesture *type* would be observed. E.g, our central-European actors frequently made 'stop' gestures by facing the palm of a single hand towards the wizard. The Japanese equivalent for that effect is crossing both arms in front of the upper body. In both cases, the type is semaphoric static.

While using a human recognizer instead of a gesture recognition system may have led to a more unconstrained behavior of the actors, it is not clear yet to what extent this introduced characteristics of human-human dialogue and how they may affect our findings. While we argue that it would be reasonable to model these in a designed gesture language, more research has to be done in order to judge their respective applicability in HCI.

Another limitation is the availability of additional communication channels to the wizards, especially facial expressions and body postures. Although they already were part of the communication during the pilot study, we observed the overwhelming majority of them to be used only for emphasis, rather than to convey independent or additional meaning. As a result, we found them to be completely redundant information channels, so we did not intend to hide them from the wizards.

Although we designed our tasks to reduce the opportunity for agency of the wizard to interfere, some did occur. However, we did not find that asymmetrically benefited any one gesture effect or type.

CONCLUSION & FUTURE WORK

In this paper, we presented the design of a user study, which extended UID, a new classification system for mid-air gestures, and observations regarding gesture type usage for ten different gesture effects.

It is interesting to note that, as of today, many systems are built upon the metaphor of direct manipulation while our results point out that the interfaces may better map to users' expectations if they were able to recognize other types of gestures for each action. Our results can serve as a guide, for both designers and researchers, to the type of gestures which are appropriate for various effects.

In particular, our results show that users tend *not* to use manipulation gestures to translate or rotate physical objects, while selecting and releasing is primarily done pantomimicly. There are huge differences in terms of motion for accepting and refusing. Bimanual acting is highly present, although only for certain combinations of gesture types and effects.

The most direct application of these results will be the development of alternative gesture languages, taking into account fundamentals of gestural communication. It is our hope that designers will begin to design gesture languages based not only on those mappings which are most 'guessable' by a novice, but instead taking into account the many factors of context which affect the definition of gestures. Our results can serve as a guide to those designers, since it frees them from UDI's simplistic 1:1 mapping of particular gestures to particular effects, and instead provides categories of gesture types from which they can draw.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge that the work for this paper was funded by the FFG on behalf of the Austrian Government as part of the research project CADET.

REFERENCES

1. Apitz, G., and Guimbretière, F., CrossY: a crossing-based drawing application. In *Proc. UIST 2004*, ACM Press (2004), 3-12.
2. Baudel, T., and Beaudouin-Lafon, M., Charade: remote control of objects using free-hand gestures. In *Commun. ACM 36*, 7, ACM Press (1993), 28-35.
3. Benko, H., and Wilson, A.D., Multi-point interaction with immersive omnidirectional visualizations in a dome. In *Proc. ITS 2010*, ACM Press (2010), 19-28.
4. Bolt, R., "Put-that-there": Voice and gesture at the graphics interface. In Proc. *7th Annual Conference on Computer Graphics*, ACM Press (1980), 262-270.
5. Cadoz, C., *Les réalités virtuelles*. Flammarion, Paris, 1994. ISBN: 2-08-035142-7.
6. Efron, D., *Gesture and environment*. The Hague: Mouton, 1972.
7. Foley, J.D., Wallace, V.L., and Chan, P., The human factors of computer graphics interaction techniques. In *IEEE computer graphics and applications 4*, 11, IEEE Computer Society Press (1984), 13-48.
8. Forlines, C., Wigdor, D., Shen, C., and Balakrishnan, R. 2007. Direct-touch vs. mouse input for Tabletop displays. In *Proc. of CHI '07*, 647-656.
9. Freeman, D., Benko, H., Morris, M.R., and Wigdor,D., ShadowGuides: visualizations for in-situ learning of multi-touch and whole-hand gestures. In *Proc. ITS 2009*, ACM Press (2009), 165-172.
10. Fussell, S.R., Setlock, L.D., Yang, J., Ou, J., Mauer, E., Gestures over video streams to support remote collaboration on physical tasks. In *Human-Computer Interaction 19*, L. Erlbaum Associates Inc. (2004), 273-309.
11. Gustafson, S., Bierwirth, D., and Baudisch, P. 2010. Imaginary interfaces: spatial interaction with empty hands and without visual feedback. In *Proc. of UIST '10*. 3-12.
12. Hilliges, O., Izadi, S., Wilson, A.D., Hodges, S., Garcia-Mendoza, A., and Butz, A., Interactions in the Air: Adding Further Depth to Interactive Tabletops. In *Proc. UIST '09*, ACM Press (2009), 139-148.
13. Izadi, S., Hodges, S., Taylor, S., Rosenfeld, D., Villar, N., Butler, A., and Westhues, J. Going beyond the display: a surface technology with an electronically switchable diffuser. In *Proc. UIST '08*. ACM Press (2008), 269-278.
14. Jordà, S., Geiger, G., Alonso, M., and Kaltenbrunner, M., The reacTable: exploring the synergy between live music performance and tabletop tangible interfaces. In *Proc. TEI 2007*, ACM Press (2007), 139-146.
15. Karam, M. and schraefel, m.c. (2005) A Taxonomy of Gestures in Human Computer Interactions. Technical Report ECSTR-IAM05-009, Electronics and Computer Science, University of Southampton.
16. Kendon, A., How gestures can become like words. In F. Poyatos (Ed.), *Crosscultural perspectives in nonverbal communication*. Hogrefe, 1988, 131-141.
17. Kirk, D., Crabtree, A., Rodden, T., Ways of the hands. In *Proc. ECSCW '05*, Springer-Verlag New York, Inc. (2005), 1-21.
18. Koh, J.T.K.V., Karunanayaka, K., Sepulveda, J., Tharakan, M.J., Krishnan, M., and Cheok, A.D., Liquid interface: a malleable, transient, direct-touch interface. In *Comput. Entertain. 9*, 2, ACM Press (2011), 7:1-7:8.
19. McNeill, D., *Hand and Mind: What Gestures Reveal About Thought.* The University of Chicago Press, Chicago, London, 1995. ISBN: 0-226-56134-8.
20. Morris, M.R., Wobbrock, J., and Wilson, A. Understanding Users' Preferences for Surface Gestures. Proceedings of Graphics Interface 2010, 261-268.
21. Nielsen, M., Störring, M., Moeslund, T.B., and Granum, E., A procedure for developing intuitive and ergonomic gesture interfaces for HCI. In: Camurri, A., and Volpe, G., (Eds.) *Gesture Workshop 2003. LNCS (LNAI) 2915*. Springer, Heidelberg, 2004, 409-420.
22. Reisman, J.L., Davidson, P.L., Han, J.Y., A screen-space formulation for 2D and 3D direct manipulation. In *Proc. UIST 2009*, ACM Press (2009), 69-78.
23. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A., Real-time human pose recognition in parts from single depth images. In *Proc. CVPR 2011*, IEEE (2011), 1297-1304.
24. Stern, H.I., Wachs, J.P., and Edan, Y., Designing hand gesture vocabularies for natural interaction by combining psycho-physiological and recognition factors. In *Int. J. Semantic Computing 2*, 1 (2008), 137-160.
25. Vogel, D., and Balakrishnan, R., Distant freehand pointing and clicking on very large, high resolution displays. In *Proc. UIST 2005*, ACM Press (2005), 33-42.
26. Wang, R.Y., and Popović, J., Real-time hand tracking with a color glove. In *ACM Trans. Graph. 28*, 3, ACM Press (2009), 63:1-63:8.
27. Wigdor, D., and Wixon, D., *Brave NUI World: Designing Natural User Interfaces for Touch and Gesture*. Morgan Kaufmann, 2011. ISBN: 978-0-12-382231-4.
28. Wobbrock, J.O., Morris, M.R. and Wilson, A.D., User-defined gestures for surface computing. In *Proc. CHI 2009*, ACM Press (2009), 1083-1092.