Web on the Wall: Insights from a Multimodal Interaction Elicitation Study

Meredith Ringel Morris
Microsoft Research
Redmond, WA, USA
merrie@microsoft.com

ABSTRACT

New sensing technologies like Microsoft's Kinect provide a low-cost way to add interactivity to large display surfaces, such as TVs. In this paper, we interview 25 participants to learn about scenarios in which they would like to use a web browser on their living room TV. We then conduct an interaction-elicitation study in which users suggested speech and gesture interactions for fifteen common web browser functions. We present the most popular suggested interactions, and supplement these findings with observational analyses of common gesture and speech conventions adopted by our participants. We also reflect on the design of multimodal, multi-user interaction-elicitation studies, and introduce new metrics for interpreting user-elicitation study findings.

Author Keywords

Gestures, speech, multimodal input, user-defined gestures, participatory design, interactive walls, web browsers.

ACM Classification Keywords

H.5.m. Information interfaces and presentation (e.g., HCI): Miscellaneous.

INTRODUCTION

Marc Weiser's vision of ubiquitous computing [28] asserted that users would have access to an ecology of interactive devices at varying scales: "tabs" (small devices, a niche now filled by mobile phones), "pads" (mid-sized devices, a niche now filled by tablets and laptops), and "boards" (very large interactive surfaces). This last class of devices remains relatively rare, with large touch-screens such as those made by Perceptive Pixel [perceptivepixel.com], SMART [smarttech.com], and Microsoft [microsoft.com/en-us/pixelsense] costing several thousands of dollars, placing them out of reach of typical consumers. However, recent consumer technologies such as Microsoft's Kinect gaming device [kinect.com] can be used to cheaply add rich interactivity to user's televisions (the Kinect was introduced in 2010, costs about $150, and contains an array microphone and depth-camera). Unlike specialty multi-touch surfaces, TVs are quite common (96.7% of U.S. households had a TV as of 2011 [17]), and large-screen TVs (40'' diagonal and greater) are increasingly popular and cheap [25].

Figure 1. Living room laboratory used in our study. Participants were seated on the couch facing the large-screen TV, which has a Kinect mounted on top.

In this paper, we employ a user-centered approach to investigate how the speech and gesture sensing afforded by devices like Kinect might support turning the TV from an entertainment device into one that enables casual interaction with more task-oriented applications. In particular, we focus on the use of a web browser, since the web is a rich and general platform whose uses span the gamut from leisure to productivity [9] and which is often used in a collaborative or social fashion [15], thereby making it suitable for multi-user settings like living rooms.

We conducted a study with 25 participants, many of whom described scenarios that have motivated them to desire the ability to interact with the internet on their TV from the comfort of their couch. We then describe an elicitation study in which these participants proposed speech and gesture interactions for fifteen common web browser functions (the *referents* [30]). We introduce two new metrics (max-consensus and consensus-distinct ratio) for analyzing user-elicitation studies, and present the most popular commands suggested by our participants. We also discuss common conventions and biases observed among our participants when designing and performing gesture and speech interactions.

ITS'12, November 11–14, 2012, Cambridge, Massachusetts, USA.

RELATED WORK

We build upon prior work on web browser interaction techniques and on end-user interaction elicitation studies.

Browser Interactions

Researchers have considered various ways of updating the traditional web browser experience for large form-factor screens. For example, the WeSearch [14] and WebSurface [26] projects both reimagine the task of web search for a multi-user, multi-touch tabletop computer. The FourBySix system [12] illustrated how multiple mice and keyboards, combined with co-located projection when placed atop an interactive tabletop, could facilitate novel web search interactions. As in these projects, we are interested in the experience of web browsing and search in large-display, multi-user scenarios; however, we focus on interaction from a distance with a vertically-oriented television screen, rather than interaction with a tangible- or touch-sensitive horizontal surface.

WebGlance [19] was a prototype public electronic bulletin board (such as might be found in a train station), that enabled web browser interactions through the use of a cell phone – links were numbered so as to be selectable by a phone's keypad, enabling commuters to opportunistically engage with a browser on a public display. CoSearch [2] also allows a group of users to share control of a web browser on a single monitor (such as a public computer terminal in a library) through use of personal mobile phones. In contrast, we focus on using speech and gesture to control the browser, rather than mediation through a remote control or phone, and focus on a home-based rather than public use scenario.

Microsoft released a developer SDK for the Kinect in 2012 [kinectforwindows.org]. The DepthJS project [depthjs.media.mit.edu] is a JavaScript toolkit that augments the Apple Safari and Google Chrome browsers with the ability to recognize some Kinect inputs, such as gesture-based tab switching or button pressing. A 2011 competition by PrimeSense (developer of the Kinect depth camera technology) encouraged people to develop a "natural interaction based web browser app." The winning entry, SwimBrowser [swimbrowser.tumblr.com] allows users to use "swimming" gestures to perform functions such as forward/back navigation, zooming, and selecting bookmarks. In June 2012, shortly after we conducted our study, Microsoft announced the addition of web browsing functionality to its Xbox gaming system [27], controlled by a combination of voice commands and by the use of Windows phones and tablets as remote controls ("SmartGlass"). In this work, we consider how both the speech and gesture-recognition capabilities of a Kinect might be used to facilitate internet access on one's TV; however, we adopt a user-centered approach of interaction elicitation to gain insight into users' concerns and mental models surrounding this use case, as well as to identify interactions that are guessable [29, 30] by end users.

End-User Elicitation Studies

Participatory design [24] incorporates end-users into the design process. User-elicitation studies are a specific type of participatory design in which users are shown *referents* (an action's effects) and are asked to supply the corresponding *signs* (interactions that result in the given referent) [8, 30]. User-elicitation approaches are able to produce non-conflicting gesture sets [30], and to produce gestures that end-users find preferable to those designed by HCI professionals [16].

User-elicitation methodologies have been used to design a variety of gesture interface types, including multi-touch gestures [4, 5, 10, 18, 30], foot gestures [1], unistroke gestures [29], pen gestures [7], and mobile phone motion gestures [23]. In this work, we use this technique to design multimodal interactions enabled by a Kinect sensor, including speech and free-space gestures.

Some studies have used user-elicitation to generate vocabularies for speech-based interfaces, including [11, 21, and 22]. Mignot et al.'s study of a multimodal (speech + gesture) system for furniture layout [11] found that users preferred speech for more abstract commands and gesture for more straightforward ones. In this work, we examine when users choose to use speech, gesture, or both for the task of browsing the web on a TV in their living room, and identify common speech and gesture interactions for a set of fifteen browser functions, as well as discussing users' perceptions of the benefits and drawbacks of each modality.

STUDY DESIGN

To better understand users' current practices, desires, and concerns regarding the use of their television to browse the Web, we conducted a lab-based user study in April 2012. Because television-viewing is often a group activity, and because the appeal of using a large display for internet browsing (rather than a personal device such as a smart phone or tablet computer) is enhanced in multi-user settings [2, 15], we recruited groups of users to participate jointly.

The study took place in a lab set up to look like a living room, with a Kinect perched atop a wall-mounted 63'' TV. The TV was connected to a PC, and displayed output at a resolution of 1600 x 1200 pixels. Participants were seated on a sofa located 11 feet, 8 inches (3.56 meters) in front of the TV screen (Figure 1).

We recruited 12 groups of participants (11 pairs and 1 triad, for 25 total participants). Participants were external to our organization, and were paid for their time. Group members knew each other prior to the study, and played Kinect-based video games together at least once per month. Five groups consisted of spouses/romantic partners, four consisted of parents with their children, and three consisted of roommates/friends. Sample occupations included student, homemaker, network engineer, property manager, emergency medical technician, audio engineer, director of sales, and grocery cashier.

Ages ranged from 12 to 49 years old (mean of 25). 72% of participants were male. All participants had prior experience using Kinect (64% used it at least once per week; only 16% were self-described novices). All participants used a web browser at least once per day, and 92% used a search engine daily.

The study consisted of three parts. First, participants answered a series of interview questions about their current experiences and desires surrounding internet use while watching TV or playing Kinect. Next, participants participated in an elicitation exercise, in which they proposed speech and/or gesture commands for controlling a web browser on their television. Finally, participants individually completed post-study questionnaires.

For the elicitation portion of the study, participants were told that they would use their web browser on their television to plan a shared weekend activity. Participants were told that they could use any combination of speech and/or gestures that they felt a Kinect should be able to recognize in such a scenario, and that the experimenter would act as the "Wizard of Oz" and ensure that the system "reacted" properly to their envisioned commands.

The experimenter then walked the participants through a sequence of fifteen common browser functions, henceforth referred to as *referents* [30], in an order that would be plausible for the given scenario of weekend activity planning (Table 3). As in the methodology presented by Wobbrock et al. [30], for each command the experimenter stated the command name and demonstrated the *effect* of the command, then prompted the participants to suggest interactions that would *cause* this command to be executed. When participants demonstrated an interaction, the Wizard used the mouse and/or keyboard to produce the result of the command in the Web browser.

Participants were able to suggest more than one interaction, if desired, and were able to consult with each other during the elicitation process. Groups typically brainstormed several interactions together, and then identified which of the proposals they actually thought were "good" (these were the ones we recorded as proposed interactions). Note that group members did not always agree on what made a good interaction, and our methodology included noting which and how many members of each group "voted" for each proposal.

Participants, who had been seated on the couch for the interview portion of the study, were told that they were free to sit, stand, or otherwise move about, depending on what seemed most comfortable and realistic to them.

RESULTS

In this section, we present qualitative and quantitative results from our interviews with participants, the interaction-elicitation exercise, and the post-study questionnaire.

Use Scenarios

Participants were asked about scenarios in which they might find it beneficial to interact with a web browser on their TV both in an interview at the beginning of the study and on the post-study questionnaire.

Interview

During the pre-elicitation exercise interview, we asked participants whether they had ever had the experience of sitting together around their TV and wanting to use the web. Eight of the twelve groups (66.7%) indicated that this was a familiar scenario (and three indicated that they already attempted to use smartphones and/or laptops in their living room to fulfill such desires). These groups then described the information needs they had in this setting.

Gaming was a theme motivating living room search needs for four of the groups. This class of information need included looking up reviews of games the user was considering purchasing from their device's online game or app store, finding descriptions of new or upcoming games, and finding hints on how to beat current game challenges.

Television and movies was another common theme motivating a desire for living room search; this theme was also mentioned by four groups. Example tasks groups recalled wanting to complete included looking up a movie's show time in a local theatre, finding movies or TV shows the group might be interested in watching, looking up trivia about actors in programs currently being viewed, and accessing fantasy sports league websites and information while viewing live sports on TV.

A desire to use the internet to look up *general trivia* facts while in the living room was also relatively commonplace, being mentioned by three groups. Three groups also mentioned a desire to access *social media*, including social network sites and e-mail, as a supplement to their TV viewing or video-gaming experience.

Several groups also recalled more unique information needs that had prompted them to wish they could use their television as a platform for web browsing and search. One group mentioned that they often wished they could perform *online shopping* activities, such as further researching products seen in commercials during TV shows. Another described a desire to use their TV as a mechanism to *order food* for delivery while playing video games. One group mentioned a desire to use their TV to search for *music* (on services like Pandora or YouTube) in order to provide background atmosphere for activities like video-gaming.

Post-Study Questionnaire

On the concluding questionnaire, participants used a five-point Likert scale to indicate how likely they would be to use a gesture + speech enabled browser on their TV for each of several scenarios. Table 1 summarizes these ratings, which suggest that participants overall were enthusiastic about using the web on their TV. "Choosing a movie" was the most popular scenario, followed by "trivia," whereas

Scenario	Median	Mean
choosing a movie to see	5	4.5
looking up facts and trivia	4	4.2
using social networking applications	4	4.2
finding and viewing photos online	4	4.1
selecting a restaurant	4	4.0
shopping online	4	3.8
research for work or school projects	4	3.6

Table 1. Participants indicated enthusiasm for using a Kinect-enabled web browser for a variety of tasks on the post-study questionnaire (using a five-point Likert scale).

"research for work or school projects" garnered the least enthusiasm. A Friedman test to evaluate differences in medians was significant, $\chi 2(6, N=25) = 28.68$, $p<.001$. Follow-up pairwise comparisons using Wilcoxon tests confirms that the research/school scenario was significantly less appealing than choosing a movie ($p = .003$) or looking up trivia ($p = .028$), perhaps because the form factor and setting do not lend themselves well to longer, more in-depth, or less playful activities.

User-Defined Multimodal Browser Control Techniques

Participants proposed a total of 357 interactions for the fifteen referents, 170 of which were distinct (considering distinctness on a per-referent basis). Table 2 shows the breakdown of proposed interactions by modality – speech interactions were the most common, and multimodal interactions were quite rare. Gesture interactions tended to be more agreed upon, as reflected in their lower representation among distinct gestures than total gestures.

Each referent had a mean of 23.8 proposed interactions (11.3 distinct interactions). On average, 3.7 distinct gestures, 6.9 distinct speech commands, and 0.7 distinct multimodal commands were suggested for each referent.

Wobbrock et al. [30] introduced a formula for an *agreement score* for a given referent, where P_r is the set of proposed interactions for referent r, and P_i is the subset of identical proposed interactions for that referent (Equation 1):

$$A_r = \sum_{P_i \subseteq P_r} \left(\frac{|P_i|}{|P_r|} \right)^2 \quad (1)$$

However, this notion of agreement was formulated for studies in which each participant was constrained to suggest exactly one interaction per referent. Since participants in our study suggested as many interaction synonyms for a referent as they felt were appropriate, comparing agreement scores across referents is not meaningful. Consider, for example, the case of the *open link in separate tab* referent. This referent received three distinct proposals for multimodal interactions, each of which were suggested by a single participant, resulting in a score $A_{multimodal} = 0.33$. This referent also received 8 distinct gesture interaction proposals, 7 of which were suggested by a single user and one of which was suggested by three separate users, resulting in a score $A_{gesture} = .16$. This lower score for agreement in the gesture modality as compared to the gesture + speech modality conflicts with our intuitive understanding of "agreement" – we would expect a higher score in the gesture modality in this case, since there was a proposed gesture that exhibited consensus from three users, whereas all of the multimodal suggestions were singletons. This example illustrates how the agreement score does not correctly account for situations in which the size of P_r varies for different referents (or modalities within a referent, if that is the desired comparison). We therefore introduce metrics that better capture the notion of "agreement" for elicitation studies in which participants can propose an arbitrary number of interaction synonyms for each referent: *max-consensus* and *consensus-distinct ratio*.

	Total Proposed Interactions	Distinct Proposed Interactions
Gesture	40.9%	32.9%
Speech	56.0%	60.6%
Multimodal	3.1%	6.5%

Table 2. Proportion of the 357 total and 170 distinct proposed interactions employing each modality. Speech interactions were suggested most often, but gesture interactions exhibited greater consensus.

The *max-consensus* metric is equal to the percent of participants suggesting the most popular proposed interaction for a given referent (or referent/modality combination). The *consensus-distinct ratio* metric represents the percent of the distinct interactions proposed for a given referent (or referent/modality combination) that achieved a given *consensus threshold* among participants; the default assumption is a consensus threshold of two, meaning at least two participants proposed the same interaction. These two metrics together convey a sense of both the peak and spread of agreement, either (or both) of which may be of interest depending on the target goal of an elicitation exercise. For example, if the goal is to design a system with a single, highly guessable command per referent, then max-consensus may be more important, whereas if the goal is to understand diversity of opinion surrounding a referent, or conceptual complexity of a referent [30], consensus-distinct ratio may be more helpful. More generally, referents achieving high values for both *max-consensus* and *consensus-distinct ratio* can be considered highly suitable for representation using user-elicited interaction, as such scores would be indicative of strong agreement on a primary interaction with few other contender interactions. Raising the consensus threshold above two may be applicable in scenarios where very high agreement among users is desired in a final command set.

Table 3 provides max-consensus and consensus-distinct ratios for each referent in our study, as well as breaking these metrics down by modality on a per-referent basis.

Referent	Max-Consensus	Consensus-Distinct Ratio	Max-Consensus (Gesture)	Consensus-Distinct Ratio (Gesture)	Max-Consensus (Speech)	Consensus-Distinct Ratio (Speech)
Open Browser	32%	0.471	32%	1.000	20%	0.462
Search Engine Query	24%	0.500	**4%**	**0.000**	24%	0.750
Click Link	52%	0.375	52%	0.500	12%	0.333
Go Back	28%	0.444	28%	0.600	28%	0.250
Go Forward	24%	0.500	20%	0.600	24%	0.400
Open Link in Separate Tab	**12%**	**0.059**	12%	0.125	**4%**	**0.000**
Switch Tab	28%	0.385	28%	0.500	16%	0.333
Find In Page	16%	0.308	12%	1.000	16%	0.300
Select Region	24%	0.333	24%	0.667	4%	**0.000**
Open New Tab	24%	0.333	24%	0.250	20%	0.400
Enter URL	28%	0.300	20%	0.333	28%	0.286
Reload Page	36%	0.667	12%	0.500	36%	0.750
Bookmark Page	28%	0.500	28%	0.500	20%	0.500
Close Tab	20%	0.455	16%	0.250	20%	0.571
Close Browser	24%	0.353	24%	0.250	12%	0.385

Table 3. The fifteen referents used for the input elicitation exercise. The sequence was chosen to create a logical narrative for the task of planning a weekend activity. The overall max-consensus and consensus-distinct ratio are shown for each referent (using a consensus-threshold of 2), as well as being broken down within each referent according to modality. The highest-scoring referent(s) for each metric are indicated with grey shading, and the lowest-scoring are indicated with a bold font. Values for multimodal interactions are not shown, since these were rarely proposed and never achieved any consensus.

Across all referents and modalities, mean max-consensus was 26.7% and mean consensus-distinct ratio was .399 (using a consensus threshold of two).

Gestures tended to exhibit greater commonalities among participants than speech interactions. Gestures had a mean max-consensus of 22.4%, whereas speech interactions had a mean max-consensus of 18.9%. Gestures also had a higher mean consensus-distinct ratio (.472) than speech interactions (.381). However, these differences between metrics for gesture and speech were not statistically significant. For individual referents, however, the difference in our agreement metrics varied quite a bit across modalities, suggesting that some referents may be best mapped to certain modalities. For instance, the referent *select region* appears better suited to gesture rather than speech interactions, as it achieved no consensus among speech commands, yet had a max-consensus score of 24% for gesture interactions and a consensus-distinct ratio of .667 for proposed gestures.

Table 4 shows the proposed interactions from our study having a consensus threshold of three. Note that this is a list of popular interaction suggestions, but it is not a conflict-free interaction set – for example, conflicting directions for the flick motion for the "go back" and "go forward" commands would need to be resolved; the proper way to do this is not necessarily straightforward, due to strong individual preferences for each underlying metaphor, as discussed in the next section.

Themes in Elicited Interactions

Our observations during the elicitation exercise identified several common themes in participants' actions and comments. These observations relate to the perceived benefits and drawbacks of the available interaction modality, conventions commonly adopted by participants (and sources of bias that may have unconsciously influenced these), and areas in which participants' preferences diverged sharply.

Modality

Overall, very few participants proposed multimodal interactions, instead choosing one modality or the other for a given command (only 11 of the 357 proposed interactions were multimodal). However, though most did not create multimodal interactions, users often proposed *multi-modal synonyms*, i.e., creating both a speech interaction and a gesture interaction as alternatives for the same command. P18 noted, "it's nice to have both options." P22 echoed that, saying "it would be really nice if there were multiple commands to use," and P6 said, "I would actually want a few options [to issue this command]." This desire for multi-modal synonyms seemed to stem from users' perceptions that both speech and gesture had modality-specific

Referent	Interaction	#
Open Browser	hand-as-mouse to select browser icon	8
	"open browser"	5
	"internet"	3
	"<browser name>" (e.g., "Internet Explorer," "Firefox," "Chrome")	3
Search Engine Query	"<query>"	6
	"search <query>"	5
Click Link	hand-as-mouse to select link	13
	"<link #>" *(assumes all links have a number assigned to them)*	3
Go Back	"back"	7
	flick hand from right to left	7
	hand-as-mouse to select back button	5
	flick hand from left to right	4
Go Forward	"forward"	6
	flick hand from right to left	5
	flick hand from left to right	5
	hand-as-mouse to select forward button	3
Open Link in Separate Tab	hand-as-mouse hovers on link until context menu appears, then hand-as-mouse to select menu option	3
Switch Tab	hand-as-mouse selects tab	7
	"next tab"	4
	"tab <#>" *(assumes all tabs have a number assigned to them)*	3
	flick hand	3
Find in Page	"find <query>"	4
	hand-as-mouse to select a find button, then type on virtual keyboard	3
Select Region	hand-as-mouse sweeps out diagonal of bounding box	6
	hand-as-mouse acts as highlighter, sweeping over each item to be included in region	3
Open New Tab	hand-as-mouse to select new tab button	6
	"new tab"	5
	"open new tab"	5
Enter URL	"<url>" (e.g., "its2012conf.org")	7
	type on virtual keyboard	5
	"go to <url>"	3
Reload Page	"refresh"	9
	"refresh page"	9
	move finger in spiral motion	3
Bookmark Page	hand-as-mouse selects bookmark button	7
	"bookmark page"	5
Close Tab	"close tab"	5
	hand-as-mouse to select close button on tab	4
	"close tab <#>" *(assumes all tabs have a number assigned to them)*	3
Close Browser	hand-as-mouse to select close button on browser	6
	"close browser"	3
	"exit"	3
	"exit all"	3

Table 4. User-elicited gesture and speech commands for a web browser on a living room TV. Only interactions with a consensus-threshold of three or higher are included due to space constraints; the number of participants suggesting each is shown in the "#" column. Gesture interactions that use "hand-as-mouse" may be referring to either "palm-as-mouse" or "finger-as-mouse" postures, and may use either dwell, click, or push as the selection technique, as described in the "Common Conventions" section.

drawbacks that might make them unsuitable under particular circumstances.

The primary perceived drawbacks of gesture input were related to ergonomics and exertion. For instance, P20 noted "I feel like my arm would get tired," and "I'm a big fan of the voice commands, because it takes less arm movement." P1 noted that for switching tabs "the speech command would be easiest because you don't have to move your arm." P8 noted that for closing the browser, speech was preferable "because it's easier." The use of two hands seemed to be perceived as particularly burdensome, with P13 noting that for clicking a link "I wouldn't want to use two hands," and P14 commenting, "I wouldn't want to use two hands for browsing, either… I would want to be doing things as simply as possible with one hand." This echoes prior findings in the domain of surface computing, in which users preferred simpler, one-handed interactions [16]. The

fact that all users chose to remain seated on the couch during the entire elicitation exercise despite a reminder from the experimenter that they could move about if they wished, also suggests that users view this scenario as one in which exertion is undesirable. Another drawback of gesture input for this use scenario was the possibility of large gestures to invade companions' personal space – several partners had to scoot apart from each other on the couch in order to avoid accidentally hitting each other while performing gestures.

Speech input had perceived drawbacks as well. In light of the living room scenario, in which multiple users are often present and conversing, many users were concerned about the ability of a speech system to detect whether an utterance was intended as an interaction or as part of a conversation. For example, P14 noted a preference for gestures since they "can't get misconstrued by the machine." Some users prepended signifiers like "command" or "xbox" to their speech interactions in order to make them distinct from casual conversation.

Users were also concerned that speech input might be disruptive to other members of their households, such as in the evening if other family members were sleeping. For example, P12 created a gesture synonym for the "bookmark" command, noting that they wanted a gesture alternative so that they "could just do it quietly." Other groups had the opposite concern, that other members of the household would be so noisy that a system would not be able to distinguish speech commands from the ambient noise. For example, P16 noted, "our house is always loud… it would never work."

Sources of Bias

As in past gesture-elicitation studies that found users' ability to suggest gestures for new form-factors was heavily influenced by their experience with WIMP interfaces [4, 30], we saw that several types of past experiences influenced the types of gesture and speech commands users proposed. Despite the Kinect being a relatively new device (launched in November 2010, about 18 months before our study was conducted), a series of conventions around its use have already emerged in the gaming community, and their influence on our users' creativity was apparent, such as preceding speech commands with the signifier "xbox," using the term "bing" to signify a desire to perform a web search, performing a dwell action to select items on-screen, and waving a hand around to get the device's "attention" before performing a gesture. Several participants also mentioned that their impression of speech interfaces was based on their experience using speech to operate their mobile phones (such as for hands-free driving scenarios), and that poor recognition in such settings biased them towards creating gestures rather than speech commands for this new scenario.

Biases from traditional desktop computing scenarios also influenced participants' behavior, despite the living room setting. The concept of a "cursor" was prevalent, with most participants assuming that a traditional mouse cursor would track their hands movements on-screen. P12 hesitated when inventing a bimanual gesture, expressing concern that the system "would probably have to have two cursors or something." Similarly, people often treated their hand as a virtual mouse by tapping and clicking; one participant in group 3 kept saying "the mouse" when referring to his hand. Several users also assumed virtual keyboards would appear on the TV, and their invented gestures consisted of typing on them by hunting and pecking with a single finger (rather than simulating keyboard-free touch-typing, e.g., [6]). Some gestures were metaphors based on imagined keyboards, such as when P8 pushed his hand forward after saying a search term out loud, explaining, "I'm trying to simulate hitting enter on the keyboard."

The influence of traditional desktop computing was so strong that several groups invented voice commands based on keyboard shortcuts. For example, P22 suggested the speech interaction "keyboard control F" for *find in page*, noting "if things could be as close to Windows as possible, it would make things a whole lot easier for everybody." P24 also suggested the voice command "control F" for the find in page function. For *reload page*, both P6 and P14 suggested the speech command "F5." P14 also used the spoken command "alt F4" for *close browser*.

Despite being influenced by prior systems, many participants seemed to recognize that legacy interactions could be modified to better suit this new scenario. For example, participants suggested that icons and buttons might be made larger to be more suitable as touch targets, that certain regions of the large-screen TV (such as corners or edges) might be set aside as dedicated menu areas, and that the need to exit applications might not be applicable, but rather that switching among a limited set of always-running applications might be more suitable in an entertainment scenario.

Common (and Divergent) Conventions

Several groups proposed the convention of assigning numbers to various components of the user interface, such as links within a web page, results returned from a search engine, browser tabs, and autocomplete or spell correction suggestions. This numbering system then formed the basis for simple speech commands to navigate the interface, e.g., "tab 4" or "link 2."

There were several cases in which participants displayed common but divergent patterns in their proposed interactions. Wobbrock et al. noticed this type of phenomenon in their study of user-elicited surface gestures, when different participants used either one, two, or three fingers to perform "single finger" pointing gestures [30]. In our study, diverging conventions were evident in the ways in which users mimicked the mouse, performed selections, specified regions, and performed flicking gestures.

Pretending that one's hand was a proxy for the mouse was a common metaphor in the gestures users invented. However, there were two distinct ways in which users mimicked the mouse with their hand, the *palm-as-mouse* gesture (in which the user held their hand with the palm parallel to the TV screen) or the *finger-as-mouse* gesture (in which users held their hand in a "gun" pose, with the index finger pointing at the screen).

When mimicking a mouse with their hand, users tended to adopt one of three conventions for selecting an item under the imagined projection of the mouse cursor: *dwell*, *click*, and *push*, with *dwell* and *click* being the most common (for example, for the *click link* referent, six users employed dwell and seven employed click). These three selection methods were compatible with either the *palm-as-mouse* or *finger-as-mouse* posture, and differed in the action performed once the imagined projection of the cursor was positioned over the desired on-screen target: the *dwell* gesture involved holding the "mouse" hand still for a few seconds, the *click* gesture involved rotating the wrist about the x-axis by 45- to 90-degrees, and the *push* gesture involved moving the hand along the z-axis toward the TV.

The *select region* referent (in which a user selected a two-dimensional on-screen region, such as a paragraph of text they wished to place on the clipboard) resulted in two different metaphors that users tended to adopt – *highlighter* and *bounding box*. The seven users employing the *highlighter* metaphor imagined that their "mouse" hand was selecting any item that they swept over individually, whereas the nine users employing the *bounding box* metaphor imagined that anything within a region whose outer limits they defined would be included in a selection.

Referents that tended to evoke "flick" gestures (particularly the *go back* and *go forward* referents) were the most problematic in terms of individual differences, in that the two most prominent metaphors users employed were in direct conflict with each other. The first flick metaphor users employed was the *book metaphor*, in which a flick from the right to the left would map to the forward direction (this is the motion one would use to turn a page when reading a book). The alternative metaphor was the *arrow metaphor*, in which a flick from right to left would map to the backward direction (since this motion would correspond to drawing an arrow pointing backward). Some users switched back and forth between which metaphor they employed, creating not only cross-user but also within-user conflicts for this gesture. For instance, P14, pondering both flick directions as possibilities for the *go back* command, wondered, "I don't know which way you'd want to go, which way makes sense." Similarly, P12 attempted to create *go back* and *go forward* commands that were inverses of each other, but switched metaphors in between, resulting in identical right-to-left flicks for both commands. The preference for these two metaphors was roughly equal, with five users employing a book metaphor for *go forward* and five employing an arrow metaphor.

Statement	Median	Mean
I would enjoy using a browser in this manner.	5	4.4
Gesture commands seem like an effective way to control the browser in this setting.	4	4.4
Speech commands seem like an effective way to control a browser in this setting.	5	4.3
I had fun operating the browser in this manner.	5	4.3
I cooperated with my partner(s) when operating the browser.	4	4.2
The interactions with the browser felt natural.	4	4.1
My partner(s) and I took on different roles in operating the browser.	4	3.6
The text in the web browser on the TV was difficult to read.	2	2.7
I felt tired after performing the activities.	2	2.6
My partner(s) and I got in each other's way when operating the browser.	2	2.3
I felt physically uncomfortable performing the activities.	1	1.6

Table 5. Participants' ratings of the usability of various aspects of the envisioned TV-web-browsing scenario (on a five-point Likert scale).

User-Suggested Referents

In addition to the fifteen referents that we presented to each group for the elicitation exercise (Table 3), several groups spontaneously proposed interactions for other functions that they felt would be important to their living room internet experience.

Three groups proposed a *scroll* referent. P10 and P15 accomplished this by extending an arm out to one's side or in front of one's body and moving it up and down; P22 used a spoken command, "scroll down."

Two groups proposed *zoom*, in order to make the text on the TV screen more readable. P16 accomplished this by moving two palms along the diagonals of an imagined rectangle, and P19 by spreading two palms out horizontally or using the voice command "zoom."

P13 observed that the large, widescreen TV form-factor would support tiling two tabs onto the two halves of the screen to support comparisons, and accomplished this by holding two palms together and then moving them horizontally in opposite directions.

P4 proposed *undo* by pulling a hand away from the screen along the z-axis. P16 used a similar backwards motion, combined with a finger-pinch of the moving hand, to represent *summon menu*.

Post-Study Questionnaire

The post-study questionnaire asked participants to reflect on the simulated TV-browser experience by using a five-

point Likert scale to rate their level of agreement with a series of statements, shown in Table 5.

Participants' responses indicated that, overall, they found the experience of using a web browser on a large-screen TV with Kinect control to be enjoyable. Participants disagreed with statements suggesting that gesture control of the browser was physically uncomfortable or tiring (a concern sometimes expressed in other studies of gesture interactions, e.g. [16]), nor did users' age correlate with their ratings of fatigue. Users also slightly disagreed that it was difficult to read the text displayed in the browser from their seated position on the couch; however, participants were generally not reading pages in depth. Additionally, two groups spontaneously invented a "zoom" gesture (which would increase font size), suggesting some desire for changes in web page readability for this form factor.

Participants rated both speech and gesture commands as effective ways to control a web browser in their living room; a Wilcoxon signed ranks test found that the difference in the overall ratings for speech and gesture was not statistically significant.

DISCUSSION

Our findings suggest that users would be highly receptive to the ability to use their TVs for web browsing. Two-thirds of participant groups readily identified scenarios when they had desired such interactions in the past, and all groups indicated enthusiasm for the scenarios proposed on the post-study questionnaire. However, the scenarios participants described and those they rated most highly suggest that living room internet use is likely appropriate for short-duration, casual tasks, more similar to the tasks labeled "lean-back internet use" by Lindley et al. [9] than to the multi-user productivity tasks often associated with collaborative web search (e.g., [2, 12, 14, 26]). Such casual, short-lived tasks seem particularly suited to computer-vision and speech-recognition based interactions, since these methods tend to be more lightweight than finding a dedicated interaction device, starting an appropriate application, etc. The recent addition of seated skeleton recognition by the Kinect technology [3] makes the gestures proposed by our participants (which they all chose to perform while seated) feasible to recognize. The difficulty of reading long articles from a distance, even on large-screen displays, and potential fatigue from performing gesture interactions for long periods [16] are usability challenges that would need to be overcome to make in-depth internet use tasks suitable for this form factor; offering multi-modal synonyms for interaction (i.e., speech alternatives to gesture commands) is one step toward expanding the suitability of this interaction platform to a broader array of scenarios.

Participants viewed neither gesture nor speech as ideal interaction methods, but rather saw pros and cons of each method depending on both situational factors (e.g., their current energy levels, or the locations and activities of other members of their household) and application-specific factors (e.g., the target referent). Wobbrock et al. [30] identified the benefit of including synonyms in a user-defined gesture set to increase guessability and coverage of proposed gestures; we found that a multi-modal elicitation study offers a related benefit of creating *multimodal synonyms*, which can support users' expressed desires to access the same functionality with different modalities in different circumstances. Our findings also provide some insight for designers as to which modalities may be more suitable for which referents – for example, the conflicting nature of the two popular metaphors for choosing flick gesture direction for the *go forward* and *go back* referents might suggest that a speech command would be less confusing to users.

Designing our study such that users could offer any number of interaction suggestions, using either (or both) modalities, offered challenges and benefits. Because this study design resulted in different numbers of proposed interactions for different referents, Wobbrock et al.'s agreement metric [30] was not applicable; however, we found that *max-consensus* and *consensus-distinct ratio* were useful alternative metrics for gaining insight into our data. Our less constrained study design allowed us to gain insight into which referents might be more suited to either speech or gesture control by examining differences in the total interactions proposed, max-consensus, and consensus-distinct ratio for each referent in each modality.

Our multi-user study design, in which groups of two or three previously-acquainted users simultaneously participated in the elicitation activity, had benefits and drawbacks. We chose this design to add ecological validity, since the target use scenario (interacting with a TV in the living room) is one in which multiple users are often present. We envisioned that pairs/triads participating together might propose cooperative interactions, such as cooperative gestures [13] or role-based collaborative web search actions [20], although users in our study did not do so (perhaps a fact which is noteworthy in its own right). Nonetheless, the multi-user design seemed to help shy users participate more actively by enabling them to build off of their partner's suggestions. Additionally, this configuration enabled relevant concerns to surface, such as the difficulty of performing large gestures without accidentally striking a companion seated nearby, or the likelihood of conversation being conflated with commands. However, it is possible that users were more likely to converge on suggestions in this arrangement; formally comparing single-user and multi-user elicitation methodologies to quantify tradeoffs between the two techniques is an area for future work.

CONCLUSION

In this paper, we explored appropriate gesture and speech interactions for enabling users to control a web browser on their television. Our contributions include interview and questionnaire results that illustrate the situations in which

users wish to use the internet on their TV, user-elicitation results that identify popular suggestions for speech and gesture commands to support such interaction, and observational insights that identify common conventions and biases surrounding the use of speech and gesture interactions in this setting. We also offer a methodological contribution by introducing two new metrics for analyzing the results of user-elicitation studies, and reflecting on the benefits and challenges of multimodal and multi-user elicitation study designs.

REFERENCES

1. Alexander, J., Han, T., Judd, W., Irani, P., and Subramanian, S. Putting Your Best Foot Forward: Investigating Real-World Mappings for Foot-Based Gestures. *Proceedings of CHI 2012*.
2. Amershi, S. and Morris, M.R. CoSearch: A System for Co-located Collaborative Web Search. *Proceedings of CHI 2008*.
3. Cooper, D. Kinect for Windows SDK reaches v1.5, now works when you're sitting down. *Engadget*, 5/21/2012.
4. Epps, J., Lichman, S., and Wu, M. A study of hand shape use in tabletop gesture interaction. *Extended Abstracts of CHI 2006*.
5. Findlater, L., Lee, B., and Wobbrock, J.O. Beyond QWERTY: Augmenting Touch Screen Keyboards with Multi-Touch Gestures for Non-Alphanumeric Input. *Proceedings of CHI 2012*.
6. Findlater, L., Wobbrock, J., and Wigdor, D. Typing on Flat Glass: Examining Ten-Finger Expert Typing Patterns on Touch Surfaces. *Proceedings of CHI 2011*.
7. Frisch, M., Heydekorn, J., and Dachselt, R. Investigating Multi-Touch and Pen Gestures for Diagram Editing on Interactive Surfaces. *Proceedings of ITS 2009*.
8. Good, M.D., Whiteside, J.A., Wixon,D.R., and Jones, S.J. (1984) Building a User-Derived Interface. *Communications of the ACM*, 27(10), 1032-1043.
9. Lindley, S., Meek, S., Sellen, A., and Harper, R. "It's simply integral to what I do": Enquiries into how the web is weaved into everyday life. *Proceedings of WWW 2012*.
10. Micire, M., Desai, M., Courtemanche, A., and Yanco, H.A. Analysis of Natural Gestures for Controlling Robot Teams on Multi-Touch Tabletop Surfaces. *Proceedings of ITS 2009*.
11. Mignot, C., Valot, C., and Carbonell, N. An experimental study of future 'natural' multimodal human-computer action. *Conference Companion INTERCHI 1993*, 67-68.
12. Morris, M.R., Fisher, D., and Wigdor, D. Search on Surfaces: Exploring the Potential of Interactive Tabletops for Collaborative Search Tasks. *Information Processing and Management*, 2010.
13. Morris, M.R., Huang, A., Paepcke, A., and Winograd, T. Cooperative Gestures: Multi-User Gestural Interactions for Co-located Groupware. *CHI 2006*.
14. Morris, M.R., Lombardo, J., and Wigdor, D. WeSearch: Supporting Collaborative Search and Sensemaking on a Tabletop Display. *Proceedings of CSCW 2010*, 401-410.
15. Morris, M.R. and Teevan, J. Collaborative Web Search: Who, What, Where, When, and Why. *Morgan & Claypool*, 2010.
16. Morris, M.R., Wobbrock, J.O., and Wilson, A.D. Understanding Users' Preferences for Surface Gestures. *Proceedings of Graphics Interface 2010*.
17. NielsenWire. "Nielsen Estimates Number of U.S. Television Homes to be 114.7 Million." May 3, 2011.
18. Nielsen, M., Störring, M., Moeslund, T.B., and Granum, E. A procedure for developing intuitive and ergonomic gesture interfaces for HCI. *International Gesture Workshop 2003*.
19. Paek, T., Agrawala, A., Basu, S., Drucker, S., Kristjansson, T., Logan, R., Toyoma, K., and Wilson, A. Toward Universal Mobile Interaction for Shared Displays. *Proceedings of CSCW 2004*, 266-269.
20. Pickens, J., Golovchinksy, G., Shah, C., Qvarford, P., and Back, M. Algorithmic Mediation for Collaborative Exploratory Search. *Proceedings of SIGIR 2008*.
21. Robbe-Reiter, S., Carbonell, N. and Dauchy, P. Expression constraints in multimodal human-computer interaction. *Proceedings of IUI 2000*, 225-228.
22. Robbe, S. An empirical study of speech and gesture interaction: Toward the definition of ergonomic design guidelines. *Conference Summary CHI 1998*, 349-350.
23. Ruiz, J., Li, Y., and Lank, E. User-Defined Motion Gestures for Mobile Interaction. *Proc. of CHI 2011*.
24. Schuler, D. and Namioka, A. (1993) Participatory Design: Principles and Practices. Hillsdale, NJ: Lawrence Erlbaum.
25. Tarr, G. Large-Screen TV Sales Grow in Q4, FY 2010. *Twice*, Feb. 15, 2011.
26. Tuddenham, P., Davies, I., and Robinson, P. WebSurface: An Interface for Co-located Collaborative Information Gathering. *Proceedings of ITS 2009*.
27. Warren, C. Internet Explorer Coming to Xbox. *Mashable*. June 4, 2012.
28. Weiser, M. The Computer for the 21st Century. *Scientific American*, September 1991, 66-75.
29. Wobbrock, J.O., Aung, H.H., Rothrock, B., and Myers, B.A. Maximizing the guessability of symbolic input. *Extended Abstracts of CHI 2005*, 1869-1872.
30. Wobbrock, J.O., Morris, M.R., and Wilson, D. User-Defined Gestures for Surface Computing. *Proceedings of CHI 2009*.