Kinect in the Kitchen: Testing depth camera interactions in practical home environments

Galen Panger
School of Information
University of California, Berkeley
gpanger@berkeley.edu

Abstract
Depth cameras have become a fixture of millions of living rooms thanks to the success of the Microsoft Kinect as a gaming device. Yet to be seen is whether depth camera interaction can succeed as widely in other areas of the home, or in the workplace. This research takes the Kinect into the kitchen, where touchless, in-the-air interaction techniques could be a boon for messy hands, but where interactions with the device must be interleaved with cooking, cleaning and socializing. We implement a recipe navigator, timer and music player, which users navigate through one-dimensional menus and “push” gestures. More experimentally, we test alternate-limb navigation, where users may hold out their head or a foot to navigate when their hands are full. We evaluated our system with five subjects who baked a chocolate chip cookie recipe in their own kitchens. Subjects liked the simplicity of navigating in the air with their hands but did not find navigating with other body parts useful, and in all cases placement of the Kinect and display unit was a challenge.
Keywords
Depth camera, gestures, push gesture, kitchen, cooking, recipes, joint selection, Kinect

ACM Classification Keywords
H.5.2 [Information interfaces and presentation]: UserInterfaces – Evaluation/methodology, theory and methods, user-centered design.

General Terms
Design, human factors.

Introduction
The release of the depth camera-based Microsoft Kinect in November 2010 was a consumer success, setting a record for the fastest-selling consumer electronics device over a period of 60 days [23]. The Kinect has also been a success among hackers, tinkerers, designers and other non-academic communities who began experimenting with the device's capabilities even before the first beta release of the Kinect software development kit (SDK) in June of 2011 [38]. Implementations span the breadth of gesture-based arcade games like Tetris and Pong [22] to gesture recognition toolkits [8], forays into posture recognition [2], and experiments with interactive shop windows [17]. Inevitable attempts to mimic the interface of the film Minority Report were made and became popular [18].

Microsoft also shares a broad vision for depth camera interactions in daily life. The company released a video on the Kinect's one-year anniversary envisioning the device's use in scenarios that go beyond its initial application to the gaming world and include areas such as medicine, home entertainment, classroom instruction, physical therapy, remote bomb diffusion, musical instrument simulation, and more [29].

The Kinect and the Kinect SDK should enable research that tests these visions of depth cameras in everyday life, though publications in this vein are still relatively scarce. However, this research, like our current project, will build on a long history of visions and experiments with gesture-based interaction methods.

Related Work
Gesture and depth camera research
Perhaps the seminal work in this space is Myron Kreuger’s mid-1970s Videoplace, where users’ full-body silhouettes could interact with projected graphical objects in real-time [25]. The DigitalDesk is another relatively early experiment that utilized gestures to interact with digital objects on a work desk [43]. Closer to the current effort is Freeman and Weissman’s 1998 study of television control through hand gestures [12]. Past consumer electronics precedents for gesture-based interaction include Intel's Me2Cam and the MagicEye for the Sony PlayStation [28].

Work by Watson [41], Kammerer and Maggioni [20], and Acharya and Mitra [3] provide helpful surveys of the field of gesture recognition, highlighting the ways we use our hands to communicate fundamental ideas, such as 'stop,' 'over there,' or 'come closer' [41], and the ways in which gestural computing is an attempt to model our interactions with the physical world [3]. Relevant to the research in our current study is Kammerer and Maggioni's note that gestural control can be an improvement over other input methods "wherever an awkward physical environment hampers the operation of complex systems." For example,
“gloves or oily hands make using a keyboard or touch screen tricky” [20].

Recent studies on gestural control include “data miming,” where objects are recognized based on a user’s gestural description [16]; a gesture keyboard [31]; teaching HCI through experiential interaction with depth cameras [40]; and “PyGmI,” which is a wearable gestural control system using cameras and projectors, and color markers worn on the fingers [35]. This last example departs from much of the work in gesture recognition in that it requires users to wear a physical marker for the purpose of tracking their active hand. Much of the promise of depth camera interaction, however, is that users do not need to wear anything special, especially in situations like kitchens where the physical marker could obstruct the user's task or become dirtied.

While much gesture recognition and depth camera research explores the promise of the interaction space, the technology has advanced enough to be tested in practical environments. If depth cameras are to be more widely utilized in the home and workplace, they need to serve unmet needs, tolerate task interleaving, where we need not devote ourselves entirely to the depth camera interaction in order to use it successfully, and be usable. A usability study by Jakob Nielsen and Donald Norman was very pessimistic about gestural control of smartphone touchscreens (iOS and Android) due to visibility and consistency concerns [33], but a first-impressions usability review of the Kinect by Jakob Nielsen was positive, saying that Microsoft’s extensive user testing paid off, even though visibility and consistency concerns were not entirely reduced [32].

Experiments in the kitchen of the future

Much like gesture-based interactions, the kitchen has long been a focal point for futurists, who hope that digital technology could be tailored to enhance cooking and nutrition, and solve common problems [1]. The Honeywell Kitchen Computer of 1969 offered to help women generate a menu around a desired entree; it cost $10,000 and required extensive training [30]. General Motors and Monsanto also offered interesting ideas for the future, near the mid-20th century. More recently, Microsoft’s Home of the Future offers recipe suggestions based on ingredients and available appliances and will project recipes onto the kitchen counter.

These efforts, not surprisingly, have led to criticism as well as exhortations to focus on simplicity. Reviewing the Microsoft home, the SF Chronicle wrote that the smart house is still “more nag than household helper” [11]. A lengthy “kitchen manifesto” in 2002 proclaimed the need for technologists to focus on the intimate rituals and cultural context of cooking, which meant focusing more on simplicity than on multiplying functionality [4]. Echoing this sentiment is Martha Stewart, who in a 2008 interview said her vision was to design “silence” into the home of the future. “I don’t want my refrigerator talking to me,” she said, “I do have a brain. Functionality has to be good, but it doesn’t have to be invasive” [21].

Notwithstanding this articulated desire for simple integration of digital technology, a number of efforts have been extended to turn the kitchen into a multi-sensor array [5, 6, 36]. MIT’s CounterIntelligence program used sensors and multiple projected displays to tell users about the contents of their refrigerator and...
how to follow recipes [5, 6]; unfortunately, however, the design was information-dense and required that the kitchen be dark so that projections were visible. Other enhanced kitchen ideas such as CounterActive and KitchenSense assume that foods of the future will come with embedded RFID tags that will allow them to be monitored [9, 19, 26], though that is a dubious assertion.

Recipes, cooking assistance, and nutrition management are another active area of exploration [37, 27]. McEwan [37] explores the use of the “kitchen of the future” mentioned in Siio, Hamada, and Mima [36] to take photo snapshots of food as it is prepared to automatically generate recipes. Cooking Navi, for example, tests foot pedals versus waterproof touch pens for recipe navigation and finds users preferred foot pedals because their hands were frequently dirty or wet [20], mirroring our current approach. eyeCook uses the user’s gaze as well as speech recognition to focus on elements of recipes in order to help define and explain unknown terms or steps [7]. Speech recognition is perhaps the best hands-free alternative to the depth camera, and in fact it is built into the Kinect [24], though usability may be limited when the kitchen gets noisy. Finally, Taichi et al. uses weight to estimate calories in a food, and employs visual recognition and other techniques to disambiguate between food and cooking utensils during food preparation [10]. Wackier efforts include robots and virtual dining [42, 34]. For the current project, cooking Navi and eyeCook are most relevant.

**Implementation**

The design of our system is motivated by two complementary goals. First, we set out to build a no-frills system that would allow us to cheaply gather data on the suitability of depth camera interaction in the kitchen through testing in real users’ homes. Second, we sought to reflect the concerns mentioned above for simplicity, when it comes to introducing a new piece of digital technology into the kitchen environment, and for visibility and consistency in interaction design. Though our prototype consists of an open laptop connected to a Kinect and a portable speaker, it’s not hard to imagine future consumer devices with significantly less bulk, such as a depth camera dock for a tablet computer or a wall-mounted computer display with a depth camera built in.

A third goal was to explore the use of other body parts for navigation, aside from the hands. While this added complexity, we wanted to enable users to navigate even when their hands were full.

For our implementation, we used C# and the Microsoft Kinect SDK Beta 2, which includes skeleton tracking for determining the location of 20 joints, or moving body parts, like the hands, feet, head, elbows and knees.

We developed three interfaces: a recipe navigator, kitchen timer and music player [FIGURES]. The recipe navigator allows the user to step through a recipe’s ingredients and instructions. The music player allows the user to choose from a number of songs, or listen to National Public Radio through the organization’s API, and offers the standard song display and playback controls. The timer can be set in hours, minutes and seconds and the normal start, stop and reset controls apply. When the user navigates away from the music player or kitchen timer, the current song or time remaining is displayed when applicable in the lower-
right-hand corner of the display. Also, due to the Kinect’s requirement that users stand several feet away from the device, all of our interfaces use large type.

The lower-left-hand corner of the display features three orienting indicators [FIGURE]. The first is the RGB video stream from the Kinect, which is intended to help users understand where in the frame they’re located and how much of their bodies are in the frame. The second is a display of circles indicating where the system thinks each joint available for navigation is located according to skeleton tracking; the green circle is the joint currently navigating. Third, next to the mini-skeleton is a larger green circle with a text label that shows the joint currently navigating so that it is clear to the user.

Our interface tracks the right hand by default, but also allows for navigation with the left hand, head, either foot, and either knee (each joint’s movements are scaled so that the user can reach all controls within the joint’s assumed range of motion). To switch to another joint, the user holds the joint out toward the Kinect sensor for two seconds, whereupon the system updates to reflect the new active joint. Navigation and control of functionality across the recipe navigator, kitchen timer and music player is accomplished through a horizontal bar of buttons [FIGURE]. To press a button, the user performs a “push” gesture, whereby they rapidly move their active joint toward the Kinect, like they are pushing the button. Some menus such as the ingredients list and song list allow for the rapid previewing, or scrubbing, of their contents without forcing the user to perform a push gesture to view each one [FIGURE].

We took this approach to our interface because it is simple—users need only worry about positioning themselves along the x-axis and “pushing” along the z-axis (toward the Kinect). This eliminates the need for a two-dimensional cursor and also reduces the need for movement along the y-axis which is difficult for joints other than the hands. Because all controls are displayed in one place, available functions are highly visible and the overall presentation is consistent, helping to address the main concerns about depth camera interactions raised by Nielsen [32]. Furthermore, because this is a depth camera, the user need not wear, hold or touch anything physical in order to navigate the system.

Finally, our implementation addresses the reality that users will interleave their interactions with our system with their cooking, cleaning or social activities in the kitchen. We chose our joint selection and button pressing gestures because we felt that accidentally holding a joint out for two seconds (joint selection) or pushing the active joint (button pressing) would be unlikely relative to the alternatives (such reversing the two gestures). In order to further cut down on accidental activations and to facilitate task interleaving,
a lock button appears in all menus, much like the lock buttons on smartphones.

**Evaluation**

The user study attempted to answer the question of whether our system allows people to comfortably and successfully navigate recipes, manage the timer and listen to music while cooking in the kitchen, which is intended to shine light on the feasibility and desirability of depth camera interactions in key home environments such as the kitchen. Study subjects were required to bake a chocolate chip cookie recipe using the system and their success and enjoyment were assessed to help answer our research question. Chocolate chip cookies were selected for the recipe because they're quick, everyone loves them and the process of mixing and separating the dough onto the cookie sheet is guaranteed to get hands messy. All ingredients and utensils (if needed by the subject) were supplied.

Five graduate students were recruited from the Berkeley computer science and information systems departments via email solicitation and direct requests. Because the purpose of the study was to venture into real home environments, we recruited subjects with kitchens in their homes. To facilitate the placement of our system, the Kinect, laptop, speaker and cables were placed on a rolling cart [PHOTO].

Subjects first familiarized themselves with the system by performing a set of given navigational tasks, which allowed them to attempt navigation with each of the joints. At this stage, data was collected on the success of direct, focused interactions. Then, once the subject completed the initial tasks, they followed the recipe and prepared the chocolate chip cookies, up to the point where the timer was set and the cookies were placed in the oven. Each test lasted about an hour from setup to takedown.

Data was collected on the rate of gesture recognition failure (false positives and negatives) and time to task completion, as well as the subjective experiences of the subjects in terms of their enjoyment, feeling of success and level of frustration. Subjects were also asked to comment on whether or not they attempted to navigate while their hands were messy, as well as how they felt 'in the air' interactions compare to other systems they might use in the kitchen to navigate recipes, music or the radio. Descriptive statistics and data plotting are most illustrative here; we do not have a 'control' or other system to compare our implementation to (except from the subjects' memories of other systems they use in the kitchen). Our current approach is more about feasibility.

**Results**

- For initial, determined tasks:
  - Figure: False negatives (system does not respond to user intention) and false positives (system mistakenly responds) across subjects 1-5 and average for joint selection gesture
  - Figure: Same for push gesture
  - Figure: Total navigation time across subjects for initial tasks: 1-5 and average

- For cookie baking task:
  - Figure: Number of intentional joint selections with system across subjects 1-5 and average
  - Figure: Number of intentional push and "scrub" gestures across subjects 1-5 and average
Figure: Number of total false positives, false negatives or other errors divided by total number of intentional interactions (joint selections, push and scrub gestures) across subjects 1-5 and average

Qualitative results:
- Figure: Likert ratings for enjoyment, feelings of success and frustration across subjects 1-5 and average
- Descriptive analysis of all comments, including comments on messy hands and comparisons to other systems

Discussion
Report on success of user studies and evaluate results:
- H1: Subjects will struggle with joint navigation using anything but their hands.
- H2: Kinect placement will be difficult, especially given 6-foot distance requirement.
- H3: System will fail frequently.
- H4: Users will express enjoyment but also frustration with errors.

Speculate on future directions:
- Kinect rotation and (if not implemented) tilt to follow user
- Accommodation of multiple users
- Microsoft says future Kinect hardware reported to perform better closer to the user.
- Depth cameras can succeed in the kitchen and future devices will be less bulky and intrusive on the user's kitchen space.

References
[31] Murase, Taichi et al. ”Gesture Keyboard Requiring Only One Camera.” UIST. 2011
[38] Stevens, Tim. ”Kinect for Windows SDK beta launches, wants PC users to get a move on.” Engadget. 2011.