Kinect in the Kitchen: Testing Depth Camera Interactions in Practical Home Environments

Galen Panger
University of California, Berkeley
School of Information
102 South Hall #4600
Berkeley, CA 94720 USA
gpanger@berkeley.edu

CHI'12, May 5–10, 2012, Austin, Texas, USA.
ACM 978-1-4503-1016-1/12/05.

Abstract
Depth cameras have become a fixture of millions of living rooms thanks to the Microsoft Kinect. Yet to be seen is whether they can succeed as widely in other areas of the home. This research takes the Kinect into real-life kitchens, where touchless gestural control could be a boon for messy hands, but where commands are interspersed with the movements of cooking. We implement a recipe navigator, timer and music player and, experimentally, allow users to change the control scheme at runtime and navigate with other limbs when their hands are full. We tested our system with five subjects who baked a cookie recipe in their own kitchens, and found that placing the Kinect was simple and that subjects felt successful. However, testing in real kitchens underscored the challenge of preventing accidental commands in tasks with sporadic input.

Author Keywords
Depth camera; Kinect; gestures; push gesture; kitchen; cooking; recipes; home; joint selection

ACM Classification Keywords
H.5.2 [Information Interfaces and Presentation]: User Interfaces – Interaction Styles, User-Centered Design, Evaluation/Methodology.

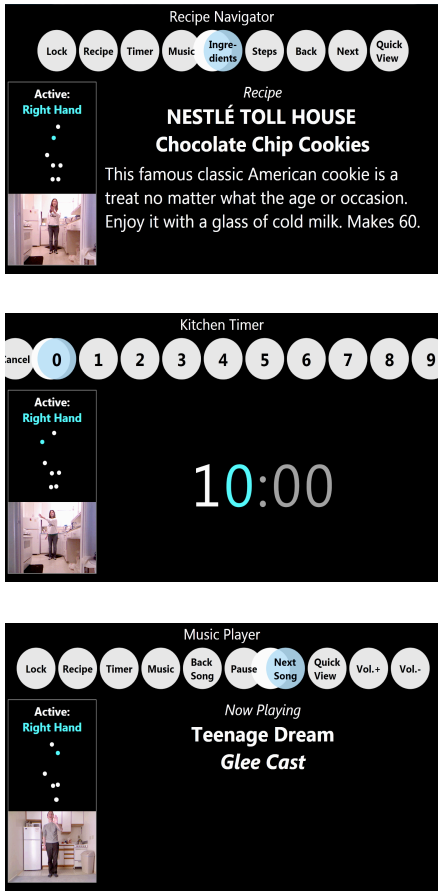

Figure 1. The three implemented applications of Kinect in the Kitchen. On the top is the Recipe Navigator main menu; in the middle the user is setting the Kitchen Timer for 10 minutes; on the bottom is the Music Player main menu.

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 [11]. Depth cameras can track body movements in 3-D space and thus allow for computer input through full-body, touchless, in-the-air gestures. They are especially consumer-friendly because they do not require users to hold physical controllers or wear physical markers. But while depth camera interactions are a proven success in gaming, we are interested in how they might succeed, in the near-term, outside the living room in other areas of the home, especially the kitchen. In order to be successful beyond the living room, depth camera interactions should provide a competitive advantage beyond being fun. Furthermore, depth camera interactions need to support sporadic input, so that users may intersperse system commands with their cooking and other tasks while in view of the depth camera.

Related Work

Depth cameras have recently been the focus of a variety of non-gaming experiments on the part of enthusiasts and researchers. Ideas from researchers include "data miming," where objects are recognized based on a user's gestural description [7], and tabletop interfaces that recognize gestures and objects performed or held above the table surface [6].

A survey of the field of gestural control by Kammerer and Maggioni points to the potential of depth camera interactions to succeed in the kitchen. The authors note that gestural control can be helpful "wherever an awkward physical environment hampers the operation of complex systems," such as when "gloves or oily hands make using a keyboard or touch screen tricky" [9]. Oily, messy, oven-gloved or full hands are common to kitchen tasks and thus gestural control could be a natural fit. Depth cameras provide a further advantage in the kitchen, however, because they do not require the user to hold or wear anything special, which is not the case for all in-the-air gesture systems.

A number of past efforts have brought futuristic though somewhat impractical interaction paradigms to the kitchen. MIT's CounterIntelligence program, for example, used sensors and multiple projected displays to tell users about the contents of their refrigerator and how to follow recipes [2], but it was information-dense and required that the kitchen be dark so that projections were visible. Other ideas such as CounterActive and KitchenSense assume that foods of the future will come embedded with RFID tags [8, 4], though this is doubtful especially for fresh foods.

Other examples from the literature on digital interactions in the kitchen focus more on near-term practical solutions. Two systems, Cooking Navi and eyeCook, relate closely to our current effort. Cooking Navi tests foot pedals against waterproof touch pens for recipe navigation and finds users prefer foot pedals because of dirty hands [5]. eyeCook employs the user's gaze as well as speech recognition to focus on elements of recipes that can be defined or explained [3]. Speech recognition and foot pedals represent good hands-free alternatives or supplements to the depth camera, though both have limitations. Here, we narrow our approach to depth cameras in order to flesh out their capabilities in the kitchen.

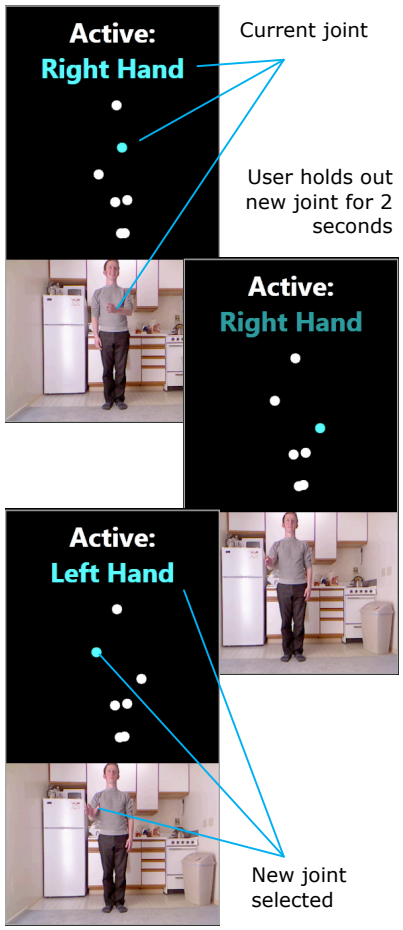

Figure 2. The body positioning area and the joint selection gesture. To switch from navigating with one joint (at top, the right hand) to another, the user holds out his new joint for two seconds (at center, the left hand). The system then updates to the new joint (at bottom).

While designing our interface, we kept in mind Jakob Nielsen's initial review of the Kinect, where he noted that many Kinect games suffer from consistency and visibility challenges. Users struggle to remember the right gesture to perform because they vary from game to game and because they are not presented on the screen to prompt the user [12].

Similarly, we also kept in mind lessons from cooking specialists. Bell and Kaye's 2002 "kitchen manifesto" proclaims the need for technologists to focus on the intimate rituals of cooking, which means emphasizing simplicity over multiplying functionality [1]. 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. "Functionality has to be good, but it doesn't have to be invasive" [10].

Design

With this background in mind, we focused on three goals for the design of our system. First, we set out to build a no-frills prototype to cheaply gather data on the feasibility of depth cameras in the kitchen through testing in real users' homes. Second, we sought to reflect the concerns mentioned above for simplicity, visibility and consistency. Third, we explored the use of other body parts, or joints, for navigation aside from the hands. While this added complexity, we wanted to enable users to navigate when their hands were full.

We developed three interfaces: a recipe navigator, kitchen timer and music player (Figure 1). 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 pre-populated songs. The timer can be set in minutes and seconds, and when it elapses, an alarm sounds. Due to the Kinect's requirement that users stand several feet away from the device, all of our interfaces use large type.

On the left side of the display is a column of orienting indicators (Figure 2). On the bottom of the column is the RGB video stream from the Kinect, which is intended to help users understand how much of their bodies are in the frame. In the middle is a display of circles indicating where the system thinks each joint available for navigation is located. On the top is a label indicating which body part is currently navigating.

Our interface tracks the right hand by default, but also allows for navigation with the left hand, head, either foot, or either knee. Joint movements are scaled to help users reach controls on both sides of the screen, though scaling means joints move more quickly, which makes it harder to point precisely. To switch to another joint, the user holds the joint out toward the Kinect sensor past a threshold for two seconds (Figure 2). Though the threshold is invisible, the active joint label dims as soon as the user reaches it. Navigation across the system is accomplished through a horizontal bar of large buttons, behind which floats a button-sized white cursor that helps users hit buttons accurately (Figure 3, top). To press a button, the user performs a "push" gesture, whereby they move their active joint toward the Kinect like they are pushing the button. In addition to stepping individually through songs and recipe instructions, users can also push a "Quick View" button to sweep through the lists by hovering over the item number (Figure 3, bottom).

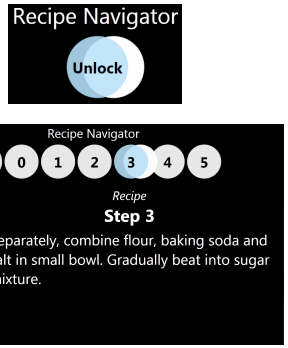

Figure 3. At top, the white cursor highlighting the "unlock" button. At bottom, the user is in Quick View mode, which allows them to quickly skim through recipe steps or songs simply by hovering over their corresponding number.

We took this approach to our interface because it is fairly simple. Users need only worry about positioning their active joint along the x-axis and reaching and pushing along the z-axis toward the Kinect. This eliminates the need for a two-dimensional cursor and also reduces y-axis movement, which is difficult for joints other than the hands. Because the body is mirrored for the user and all controls are displayed in one place, the body and available functions are visible rather than hidden to the user and the overall presentation is consistent, helping to address the concerns about visibility and consistency raised by Nielsen [12], noted above. Furthermore, because this is a depth camera, the user can but need not wear or hold anything physical in order to navigate.

Finally, our implementation attempts to address the reality that users will intersperse their interactions with our system with their cooking, cleaning and social activities in the kitchen. We chose our gestures because we felt that, with the right optimizations, holding a joint out to select it or pushing the active joint to press a button would be unlikely accidental triggers relative to alternatives. For example, the "hover" gesture would be problematic for our interface given that in some menus all x-axis positions map to a button, and thus users are always hovering over a button. In addition, to cut down on accidental activations and to facilitate task interleaving, a lock button appears in most menus, which hides buttons in the current menu and replaces them with a single "unlock" button (Figure 3, top).

Implementation

For our implementation, we used C# and the Microsoft Kinect software development kit (SDK) Beta 2, which provides skeleton tracking for determining the location of 20 joints. Scaling the movements of our joints was accomplished using the Coding4Fun Kinect API.

Limitations of the depth camera technology and the early stage of Microsoft's Kinect SDK provided some challenges. Libraries are limited such that no standard gestures or mappings to UI events are provided. In addition, joints end in single points, meaning that gestures like opening or closing the hand cannot be implemented using the SDK, though they might be valuable. Depth cameras also generate a significant amount of static, enough that Microsoft provides a "smoothing" function for joint tracking, though this causes it to feel less responsive. We use the smoothing function to reduce the jerkiness of joint movements.

Our push gesture was implemented by sampling the z-axis velocity and triggering when the active joint velocity was at a certain threshold toward the Kinect. Ceilings on active joint x- and y-axis velocities and on average non-active joint z-axis velocity were placed to limit accidental activations by non-push movements. In addition, a small wait time after a button is highlighted and before it is pushable was implemented to reduce accidental activations when sweeping the hand across the screen. In practice, it was difficult to find a balance of these parameters. In a future iteration, we might set a distance threshold in addition to a velocity threshold, and we might average a sample of several frame velocities, rather than trigger on a single frame.

Our joint selection algorithm was based on the z-axis distance of the active joint-to-be from the average of the other joint distances from the Kinect. When the user hit our distance threshold and held for 2 seconds, the system switched to navigating with that joint. An

Figure 4. The laptop and portable speaker (on the top shelf) and Kinect sensor (on the second shelf) were placed on a rolling cart to facilitate placement of the system in kitchens.

additional caveat was added to the algorithm so that the hands had to be a certain distance from one another, to avoid accidentally switching between them when holding something with both hands. In practice, this worked well and accidental switches were rare.

Evaluation

The user study attempted to answer the question of whether our system allows people to comfortably and successfully navigate recipes, manage a timer and listen to music while cooking. Five students were recruited from a graduate Berkeley computer science course. Subjects were required to bake a chocolate chip cookie recipe in their own kitchens using the system. Chocolate chip cookies were selected for the recipe because the process of mixing and separating the dough onto the cookie sheet tends to get hands messy. All ingredients were supplied, as were utensils if needed. To facilitate the placement of our system, the Kinect, laptop, speaker and cables were placed on a rolling cart (Figure 4). Tests took about an hour.

Subjects first performed a set of tasks that allowed them to attempt navigation with each joint and test the three applications and lock button. Then subjects followed the recipe in the system and prepared the cookies, setting the timer while baking and listening to music. While subjects were cooking, observations were made on the frequency of gesture errors as well as how well users understood the interface. After the baking was finished, subjects were directed to an online survey which they completed after the experimenter left.

Results and Discussion

Subjects in the survey reported feeling successful using the system, and reported high levels of ease and pleasure, and low levels of frustration. They also felt the current implementation, provided it were able to load other recipes and music, was nearly as helpful as they could imagine the interaction style being generally (Figure 5). All subjects reported navigating while their hands were messy and comments about this were enthusiastic.

Our observations were not quite as favorable. Accidental button pushes were too common. During focused interaction, accidental pushes occurred while sweeping the hand across the screen, especially when changing directions. Pushes also occurred when subjects were focused elsewhere. All users to a lesser extent also suffered from system failures to recognize their pushes, which often appeared to be due to their pushing too quickly (a limitation likely due to smoothing by the Kinect SDK).

Lock buttons on the screen were appreciated by subjects but used rarely. Two subjects thought the lock was automatic, though locking in those cases resulted from accidental pushes. In the future, locking should be automated when the user turns sideways (and thus x-axis joint positions collapse inward) to their side counters or on the way to turning to face counters behind them. Unlocking should be a two-step rather than one-step process to prevent accidental unlocking.

There were significant successes, however, including the surprising ease of positioning the Kinect cart, which was done by the experimenter. In all but one case, the camera was positioned so that the subject was always in the frame. The distance requirement meant that the cart was placed generally outside of the kitchen and out of the way, which one subject noted freed up counter

Figure 5. In surveys, subjects rated themselves an average of:

5.6
out of 7 on how *successful* they felt using the system. 1 meant "very unsuccessful" and 7 meant "very successful."

5.4
out of 7 on how *helpful* they see this style of interaction being in the kitchen, generally. 1 meant "very unhelpful" and 7 meant "very helpful."

4.8
out of 7 on how *helpful* the current prototype was to them. 1 meant "very unhelpful" and 7 meant "very helpful."

2.2
out of 5 on how *frustrated* they felt using the system. 1 meant "no frustration" and 5 meant "extreme frustration."

4.2
out of 5 on how much *ease or pleasure* they felt using the system. 1 meant "no pleasure" and 5 meant "extreme pleasure."

space over a recipe book. Subjects took advantage of the body positioning area to keep themselves in the frame, though a future iteration would do more to show subjects when they step out of the frame. An apron was worn by one subject and worked fine.

One-dimensional menu navigation was also successful, and pointing errors were rare because users lined up the white cursor with the buttons before pushing. But menus should be improved to make accidental activations less costly. Before resetting the timer, for example, a confirmation should be required. And subjects appreciated being able to rapidly sweep through recipe steps and songs in Quick View, though selections should also be two steps to reduce errors.

Alternate-limb navigation was ultimately a success only in the case of the head and even then it was limited because only one subject ever used it for a significant amount of time. Observing subjects, however, it was clear that using the head, while socially awkward, was relatively easy. Users were adept at switching to the head and using it to position the cursor and push buttons. Legs posed balance issues, and knees were especially hampered by their limited range of motion.

Overall, we think depth camera interactions can be successful in the kitchen in the near-term with more work, particularly, on accidental activations. Automatically locking the screen when the user turns away would help, as would optimizing our gesture recognition. It's important in the future to support or at minimum tolerate multiple users in the kitchen as well.

Ultimately, it's not difficult to assemble a laptop, Kinect and cart with the given software. Dedicated devices are possible for the future, too. However, it's clear that in-the-air gestural control remains a foreign concept to users and that in order to feel comfortable with the interaction style they need persistent reminders about the gestures available to them as well as feedback on their performance.

References

[1] Bell, Genevieve and Kaye, Joseph. Designing Technology for Domestic Spaces: A Kitchen Manifesto. *Gastronomic: The Journal of Food and Culture*. 2002.

[2] Bonanni, Lee and Selker. CounterIntelligence: Augmented Reality Kitchen. CHI 2005.

[3] Bradbury, Shell and Knowles. eyeCook: Hands On Cooking - Towards an Attentive Kitchen. CHI 2003.

[4] Chen, Chang, Chi, Chu. A Smart Kitchen to Promote Healthy Cooking. UbiComp 2006.

[5] Hamada, Okabe, and Ide. Cooking Navi: Assistant for Daily Cooking in Kitchen. Multimedia 2005.

[6] Hilliges, Izadi and Wilson. Interactions in the Air: Adding Further Depth to Interaction. UIST 2009.

[7] Holz, Christian and Wilson, Andrew. Data Miming: Inferring Spatial Object Descriptions from Human Gesture. CHI 2011.

[8] Ju, Hurwitz, Judd and Lee. CounterActive: An Interactive Cookbook for the Kitchen Counter. Evaluation 2001.

[9] Kammerer, B. and Maggioni, C. GestureComputer - History, Design and Applications. In *Computer vision for human-machine interaction*. 1998.

[10] Kelly, Kevin. I Do Have a Brain. Wired. 2008.

[11] Kinect Confirmed As Fastest-Selling Consumer Electronics Device. Guinness World Records. 2011.

[12] Nielsen, Jakob. Kinect Gestural UI: First Impressions. Jakob Nielsen's Alertbox. 2010.