Gesture Interfaces: Minor Change in Effort, Major Impact on Appeal

Xiaoxing Liu
Mechanical and Industrial Engineering
University of Iowa
Iowa City, USA
xiaoxing-liu@uiowa.edu

Geb W. Thomas
Mechanical and Industrial Engineering
University of Iowa
Iowa City, USA
geb-thomas@uiowa.edu

ABSTRACT
Making gestures easy for imaging systems to reliably recognize often comes at the expense of user effort. But what is the impact of increasing a gesture's effort, even slightly, on user preference? We investigate physical effort, system reliability, and user satisfaction in two experiments. The first explores eight basic command gestures. Participants preferred the less effortful gestures in two of the three easy-difficult gesture pairs when they perceived the difference in effort to be significantly different. The second experiment explores two separate three-dimensional pointing and selection conditions that differ only in the movement distance required to finish the task. In both experiments, there is a significant negative correlation between a gesture's effort and its appeal. The results show the great impact that effort has on a user's willingness to utilize the system. The findings provide evidence that the trade-off between user effort and system reliability must be carefully considered to build an effective gesture interface.

Author Keywords
Gesture interfaces; usability evaluation; user experience; effort-based measurement.

ACM Classification Keywords
H.5.2. User Interfaces: evaluation/methodology.

INTRODUCTION
Despite their natural contribution to human communication [7,9] and decades of research in many fields [13,19,21], mid-air 3D gesture interfaces have been widely adopted in only a few niche application areas. Thanks to the recent low-cost body-tracking devices like Microsoft Kinect [30] and Leap Motion [31], the pace of research has increased, and gesture systems are more broadly distributed. Nevertheless, there is no consensus on how to design effective gesture interfaces or how to measure their success. This lack of understanding has limited opportunities to improve their application.

CHI 2017, May 06-11, 2017, Denver, CO, USA
© 2017 ACM. ISBN 978-1-4503-4655-9/17/05 $15.00
DOI: http://dx.doi.org/10.1145/3025453.3025513

A gesture system's recognition reliability is a concern among researchers. Various algorithms have been explored to achieve higher reliability [12,16,23,29]. However, higher recognition reliability alone does not always lead to a more usable system. It is likely that designers select specific gestures that are more easily distinguished by the technology being employed in the recognition. These decisions often sacrifice the utility, memorability, comfort and naturalness of the gestures from the user's viewpoint, resulting in a less favorable system. Meanwhile, researchers have been relying on user-defined gesture sets to develop user-friendly gesture lexicons [5,18,27]. Still, user-developed gestures, often extracted without an actual gesture recognition system, may not be reliably interpreted by gesture systems. Wright et al. [28] report an experiment in which some of the most preferred gestures had low system performance, and gestures with great system performance were sometimes disliked.

Researchers have noted that effort is a factor in gesture preferences and selection [4,17] and it must be taken into account when assessing the usability of a gesture interface [2,20]. However, effort remains a subordinate design consideration. The trade-off between the needs of the system and the desires of the user remains unbalanced.

To some extent, people's limb trajectories follow predictable patterns. Gestures appear to minimize some "cost," either to minimize metabolic energy costs [1], movement time, distance, peak velocity, jerk (rate of change of acceleration) [8], or peak acceleration [6,15,25]. Stern et al. [24] proposed a gesture vocabulary design approach with evaluation of both intuitiveness and comfort. They found that users naturally filtered out difficult gestures in the intuitive experiment when selecting gestures. This suggests that effort could be a fundamental factor that affects the attractiveness of the gestures from the user's viewpoint.

Perhaps evaluating system performance is easier with current usability testing methods than evaluating effort, which is more difficult to measure. The role effort plays for an effective gesture interface, and more importantly, the trade-off between user effort and system reliability, have not received much attention, or are often poorly understood. This trade-off needs to be explicitly considered in designing a user interface that is appealing and easy to use. Therefore, we present an evident investigation of the relationship between the perceived physical effort and user satisfaction for several gesture interfaces. We hypothesize that physical effort is fundamental to satisfaction and people prefer even slightly less effortful gestures that are able to communicate clearly. This relationship is first evaluated for 8 basic hand gestures in experiment 1, then corroborated in experiment 2 by using a three-dimensional pointing and selection task.

EXPERIMENT 1

Experiment 1 measures the effort and user preference for basic gesture elements independent of any specific task. It explores whether users prefer less effortful gestures and the effect of recognition reliability on perceived effort and appeal by asking participants to repeat each gesture multiple times, then rate the physical effort and appeal of the gesture.

Participants

Sixteen female and fourteen male adults, with a combined mean age of 25.6 years, were recruited from the University of Iowa community and participated in this experiment. Twenty-one participants had experience with gesture interfaces including in-game motion. The participants were divided into two cohorts balanced by gender and gesture interface experience. The cohorts completed the gestures in reverse sequence to offset any effect introduced by the order of the procedures, including the effect of fatigue.

Apparatus

The Leap Motion controller was used to capture hand gestures. The experiment, programmed in JavaScript as a webpage, was individually conducted on a PC computer with a 12.5-inch IPS display at a resolution of 1366 x 768 pixels in a room with soft light. The experiment was presented in the Chrome web browser in full-screen mode. The display was adjusted to each participant's eye level.

The Gestures

Experiment 1 considered four pairs of basic gestures which could be either used as commands themselves or as components of more complex gestures. Each gesture pair shared a similar movement, but one was slightly larger to make it more physically effortful, or "difficult" (Table 1). A pinch, the act of squeezing the thumb and index finger, is paired with a grab, closing all five fingers from an open position to make a fist. A finger key tap, an up-and-down tapping movement with one finger, is paired with a palm key tap, which involves maintaining the forearm in the air while using the whole palm or forearm to tap as if bouncing a ball. Swiping, a long, linear movement of the hand and fingers, is paired with a sideways wrist motion with an upward motion of the forearm. Finally, the circle gesture pair, using a finger to draw a circle in the air, differs only in the diameter of the circle, either less than 9 cm or greater than 12cm.

Pair	Easy Gesture	Difficult Gesture
1	Pinch	Grab
2	Key Tap – Finger	Key Tap - Palm
3	Swipe Left	Swipe Up
4	Circle Small	Circle Large

Table 1. Eight gestures tested in experiment 1

Procedure

A pilot test indicated that system recognition reliability strongly influenced participant ratings of effort and satisfaction. When the system repeatedly failed to recognize a participant's attempt to perform a gesture, the participant had to perform the same gesture many times. As a result, participants became frustrated and increased their subjective rating of the gesture effort and lowered their subjective rating of preference. Thus, the experiment was designed to eliminate the confounding effect of gesture recognition reliability. For the first round of testing, each participant performed each gesture 20 times for the experimenter, simulating a completely reliable gesture recognition system (the Wizard of Oz experiment approach [14]). After completing each gesture pair, participants rated the gestures by answering the following questions on a scale of 0-10:

Q1. How physically tiring is this gesture?

Q2. How appealing is this gesture?

In the second round of testing, participants repeated each gesture until the system successfully recognized it 20 times, causing the participants to experience the effects of occasionally unrecognized performance attempts. This allowed us to measure the effect of reliability on the gesture attempts. In the second round, participants were paced at the speed of one gesture attempt per second. After completing each pair, participants responded to the following three survey questions on a scale of 0 – 10:

Q3. How physically tiring is this gesture?

Q4. How well does the system recognize this gesture?

Q5. How appealing is this gesture?

Participants always performed the gesture in pairs, but the order of easy / difficult gesture was balanced. They also followed the same condition order in the two test rounds.

Results

Figure 1 shows the mean perceived physical effort ratings (Q1) of the eight gestures (all error bars show 95% CIs). Because the responses were not normally distributed, four separate Wilcoxon signed-rank tests [26] measured the difference between easy and difficult gestures in each pair. The differences between three gesture pairs were significant: the pinch / grab pair ($Z = -0.36$, $p = 0.02$), the tap pair ($Z = -$

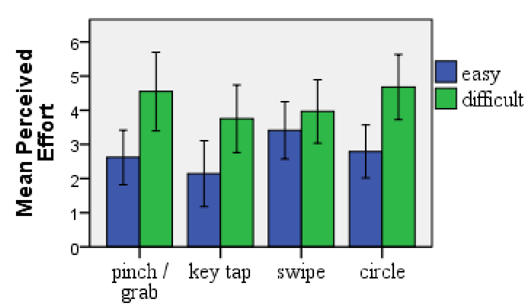

Figure 1. Mean perceived effort of 8 gestures. Error bars indicate 95% confidence interval.

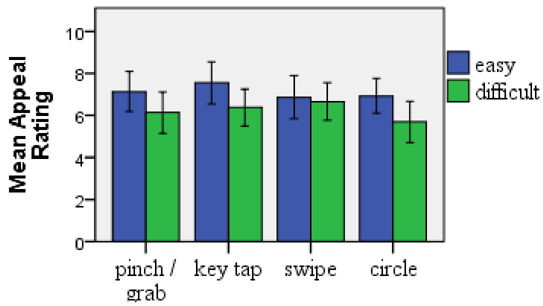

Figure 2. Mean appeal (Q2) rating of 8 gestures. Error bars indicate 95% confidence interval.

2.98, p = 0.003), and the circle pair (Z = -3.18, p = 0.001). The difference between swipe gestures was not significant (Z = -1.96, p = 0.051).

Figure 2 shows the mean appeal ratings (Q2) for the eight gestures (error bars show 95% CIs). Four separate Wilcoxon signed-rank tests revealed that the appeal ratings of two pairs were significantly different: the tap pair (Z = -2.22, p = 0.027) and the circle pair (Z = -2.44, p = 0.015). The differences between pinch / grab pair (Z = -1.84, p = 0.066), and the swipe pair (Z = -1.15, p = 0.251) did not reach significance.

The correlation between effort and appeal ratings across all eight gestures was measured with a Spearman's rank-order correlation [22]. There was a significant, negative correlation between appeal and effort rating, $r_s(230) = -0.54$, $p < 0.001$.

Swipe left received the lowest rating in perceived recognition reliability (M = 6.66, SD = 2.1), while grab was perceived as most reliable (M = 7.83, SD = 2.1). Gestures in between, ranked from lowest to highest, were circle large (M = 6.72, SD = 2.6), tap small (M = 6.86, SD = 2.3), circle small (M = 7.07, SD = 2.4), tap large (M = 7.07, SD = 2.4), swipe up (M = 7.59, SD = 2.2), and pinch (M = 7.72, SD = 2.4).

The relationship between changes in reliability and changes in appeal was measured with a Spearman's rank-order correlation between the participants' perception of system reliability (Q4) and the difference in appeal between the simulated, fully reliable system and the actual system (Q2 – Q5). Figure 3 shows the scatterplot of change in appeal, Q2

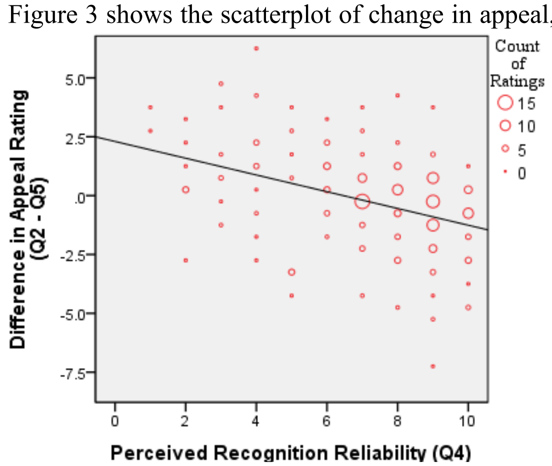

Figure 3. Regression fit plot of projected difference in appeal rating as a function of perceived reliability.

- Q5, as a function of the perceived recognition reliability of the actual system, Q4. There was a negative correlation at a statistically significant level, $r_s(230) = -0.39$, $p < 0.001$.

Finally, we conducted a similar correlation between changes in reliability (Q4) and changes in perceived effort (Q1 – Q3). There was a significantly positive correlation, $r_s(230) = 0.28$, $p < 0.001$, indicating the lower recognition reliability of the actual system caused an increased effort rating.

Discussion

In three of the four gesture pairs, participants were able to perceive that one gesture required more effort than the other. Participants did not report that the paired swipe gestures differed in effort. Perhaps the difference was too subtle. As predicted, participants preferred the less effortful tap and circle gestures. The pinch / grab pair effect was in the expected direction, but did not achieve significance. Also as hypothesized, the negative correlation between each gesture's appeal and perceived effort is indicative of the fact that the more effortful gestures were less appealing.

Finally, user satisfaction is negatively affected when the system does not consistently and reliably recognize user input. The negative correlation between perceived recognition reliability (Q4) and the difference in appeal caused by the change in reliability between the ideal and actual systems (Q2 – Q5) indicated that: 1) the more reliably the system recognized a gesture, the smaller the difference between Q2 and Q5; and 2) when gestures were perfectly recognized by the system (that is, when Q4 = 10), participants tended to find the system more appealing than when they simply performed the gesture for the experimenter (Q2 – Q5 < 0).

EXPERIMENT 2

Experiment 2 explores the relationship between physical effort and user preference, using a three-dimensional, pointing and tapping-to-select gesture.

Participants

The 30 participants from experiment 1 also completed this experiment, with the second cohort performing the procedures in reverse order as before.

Apparatus

The hardware and setup was the same as in experiment 1. The three-dimensional interaction was programmed using the Three.js JavaScript library.

Procedure

The test environment displayed 16 white, spherical targets that were regularly spaced on a rectangular grid (Figure 4). A wire frame hand model was animated by the participant's hand movements as perceived by the Leap Motion system. A round, red selection cursor was rendered just beyond the fingertip of the index finger. Shadows of the cursor and targets were projected onto two walls and the ground to provide navigational cues in the three-dimensional space. Each participant was asked to move his or her hand to control the cursor and use the tap gesture to select all spheres. The

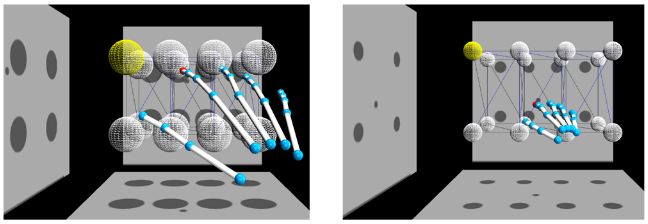

Figure 4. Experiment 2 condition A (left) and B (right).

current target sphere appeared in yellow. After a successful selection, the current target was removed from the environment and the next target sphere appeared in yellow. Target order was randomized. Visual feedback was given when the cursor intersected with the targets.

Each participant finished the selection tasks in two conditions: A and B. The movement distance required to finish the task in condition B was 2.2 times longer than in A. Because moving a shorter distance requires less energy, condition A was expected to be less effortful. Despite the difference in physical effort, the two conditions were designed to have constant *selection reliability*. That is, although the targets in condition A appeared bigger than in condition B, they were equally easy to be precisely tapped. Equalizing selection reliability requires the hand-to-target size ratio to be the same in the two conditions. In this experiment, condition A only used the central area of Leap's working space while condition B used a broader range. Because the performance of the sensor is inconsistent across its working space [10], we increased the hand-to-target size ratio slightly in condition A (A:B = 1.1:1) to offset this performance inconsistency.

To verify that the selection reliability of the two conditions was comparable, each participant's total recognized tapping attempts was counted. When participants finished the selection task in both conditions, they were asked to answer the following survey questions on a scale of 0 – 10:

Q1. How physically tiring is this task?

Q2. Do you experience difficulties in finishing the tasks precisely? How difficult / ambiguous is the interaction?

Q3. How appealing is the interaction for the selection task?

Results

The total selection attempts and mean participant response to survey question Q1 through Q3 are reported in Table 2.

The selection reliability was calculated by dividing the number of targets (16) by the total selection attempts. A paired t-test revealed that the difference between conditions A and B in selection reliability was not significantly different at the $p = 0.05$ level ($t = -1.82$, $p = 0.079$).

	Attempts	Q1	Q2	Q3
A	21.1 (4.8)	3.73 (2.20)	3.57 (2.21)	7.63 (1.43)
B	19.9 (3.6)	6.17 (2.44)	5.30 (2.48)	5.57 (2.56)

Table 2. Mean (and standard deviation) of participants' response to Q1 through Q3, and total attempts

Three separate Wilcoxon signed-rank tests measured the differences between the two conditions. The differences for all three questions were significant: the perceived effort (Q1), $Z = -4.35$, $p < 0.001$, the perceived accuracy (Q2), $Z = -3.49$, $p < 0.001$, and the appeal (Q3), $Z = -3.95$, $p < 0.001$.

To test the hypothesis that people prefer interactions that are less effortful, a Spearman's rank-order correlation with samples from both conditions was calculated to determine the relationship between the participants' perceived effort (Q1) and a condition's appeal (Q3). There was a significant, negative correlation between the two variables, $r_s(58) = -0.71$, $p < 0.001$, indicating that less effort was preferred.

Discussion

Experiment 2 explored user preferences for two separate three-dimensional pointing interfaces with similar selection reliability but different levels of gesture effort. Specifically, completing the two conditions in a comparable number of attempts, participants perceived the condition that required shorter movement distance (condition A) to be less effortful. This condition also enjoyed a significantly higher appeal rating. Moreover, the strong correlation between the perceived effort and appeal supported our hypothesis that people prefer less effortful interactions.

It is worth noting that participants also perceived condition B as significantly more difficult to complete, despite finishing it with fewer attempts. Reasons for this might be (or the combination of): 1) the targets at the corners in condition B reached the extreme of the Leap Motion's working space. The occasional loss of tracking at the border might have made condition B feel "difficult" to finish. The selection reliability, meanwhile, would not be affected by unrecognized gestures. 2) The participants may not have completely understood the difference between effort and accuracy. Thus, the longer movement distance in condition B may have been interpreted as more difficult to complete.

CONCLUSION AND FUTURE WORK

The results of both experiments demonstrate the importance of gesture effort, even when there are only small differences, such as tapping with a finger versus tapping with a palm, or making a small circle versus a big circle.

One limitation of this study is the lack of an objective measure of physical effort because it is difficult to define an empirical effort measure for such small movements. None of the standard questionnaires for system usability seem relevant, but it may be worth pursuing some such as NASA TLX [11] and Borg CR10 scale [3] in the future.

Effort is critical to the appeal of a gesture interface and has been overlooked. Designers too often sacrifice effort in order to achieve higher recognition reliability. The experiments here show that physical effort greatly affects a gesture's or a system's attractiveness, for both command tasks and pointing tasks. Thus, the trade-off between user effort and system reliability must be carefully considered in order to make gesture interfaces more usable in the future.

REFERENCES

1. R. McN Alexander. 1997. A minimum energy cost hypothesis for human arm trajectories. *Biological cybernetics* 76, 2: 97–105.
2. Kayne Barclay and Danny Wei. 2011. A quantitative quality model for gesture based user interfaces. In *Proceedings of the 23rd Australian Computer-Human Interaction Conference* (OzCHI '11), 31–39. https://doi.org/10.1145/2071536.2071540
3. Gunnar Borg. 1998. *Borg's perceived exertion and pain scales*. Human Kinetics, Champaign, IL, US.
4. Marcio C. Cabral, Carlos H. Morimoto, and Marcelo K. Zuffo. 2005. On the usability of gesture interfaces in virtual reality environments. In *Proceedings of the 2005 Latin American conference on Human-computer interaction* (CLIHC '05), 100–108. https://doi.org/10.1145/1111360.1111370
5. Li-Chieh Chen, Po-Ying Chu, and Yun-Maw Cheng. 2016. Exploring the Ergonomic Issues of User-Defined Mid-Air Gestures for Interactive Product Exhibition. In *International Conference on Distributed, Ambient, and Pervasive Interactions*, 180–190.
6. Sascha E. Engelbrecht. 2001. Minimum principles in motor control. *Journal of Mathematical Psychology* 45, 3: 497–542.
7. Louie J. Fant. 1972. *Ameslan: An introduction to American Sign Language*. National Association of the Deaf.
8. Tamar Flash and Neville Hogan. 1985. The coordination of arm movements: an experimentally confirmed mathematical model. *The journal of Neuroscience* 5, 7: 1688–1703.
9. Susan Goldin-Meadow and Marolyn Morford. 1990. Gesture in early child language. In *From gesture to language in hearing and deaf children*. Springer, 249–262.
10. Jože Guna, Grega Jakus, Matevž Pogačnik, Sašo Tomažič, and Jaka Sodnik. 2014. An Analysis of the Precision and Reliability of the Leap Motion Sensor and Its Suitability for Static and Dynamic Tracking. *Sensors* 14, 2: 3702–3720. https://doi.org/10.3390/s140203702
11. Sandra G. Hart. 2006. Nasa-Task Load Index (NASA-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 50, 9: 904–908. https://doi.org/10.1177/154193120605000909
12. N. A. Ibraheem and R. Z. Khan. 2012. Vision Based Gesture Recognition using Neural Networks Approaches: A Review. *International Journal of human Computer Interaction (IJHCI)* 3, 1: 1–14.
13. H. Jiang, B. S. Duerstock, and J. P. Wachs. 2016. User-Centered and Analytic-Based Approaches to Generate Usable Gestures for Individuals With Quadriplegia. *IEEE Transactions on Human-Machine Systems* 46, 3: 460–466. https://doi.org/10.1109/THMS.2015.2497346
14. J. F. Kelley. 1984. An Iterative Design Methodology for User-friendly Natural Language Office Information Applications. *ACM Trans. Inf. Syst.* 2, 1: 26–41. https://doi.org/10.1145/357417.357420
15. Winston L. Nelson. 1983. Physical principles for economies of skilled movements. *Biological cybernetics* 46, 2: 135–147.
16. Dung Duc Nguyen, Thien Cong Pham, and Jae Wook Jeon. 2009. Fingertip detection with morphology and geometric calculation. In *Intelligent Robots and Systems, 2009. IROS 2009. IEEE/RSJ International Conference on*, 1460–1465.
17. Michael Nielsen, Moritz Störring, Thomas B. Moeslund, and Erik Granum. 2004. A procedure for developing intuitive and ergonomic gesture interfaces for HCI. In *Gesture-Based Communication in Human-Computer Interaction*. Springer, 409–420.
18. Anna Pereira, Juan P. Wachs, Kunwoo Park, and David Rempel. 2015. A user-developed 3-D hand gesture set for human–computer interaction. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57, 4: 607–621.
19. Julie Rico and Stephen Brewster. 2010. Usable gestures for mobile interfaces: evaluating social acceptability. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 887–896.
20. Matthias Schwaller and Denis Lalanne. 2013. Pointing in the air: Measuring the effect of hand selection strategies on performance and effort. In *Human Factors in Computing and Informatics*. Springer, 732–747.
21. Fernando Winckler Simor, Manoela Rogofski Brum, Jaison Dairon Ebertz Schmidt, Rafael Rieder, and Ana Carolina Bertoletti De Marchi. 2016. Usability Evaluation Methods for Gesture-Based Games: A Systematic Review. *JMIR Serious Games* 4, 2. https://doi.org/10.2196/games.5860
22. C. Spearman. 1904. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology* 15, 1: 72–101. https://doi.org/10.2307/1412159
23. T. Starner and A. Pentland. 1995. Real-time american sign language recognition from video using hidden markov models. In *Computer Vision, 1995. Proceedings., International Symposium on*, 265–270.
24. H.I. Stern, J.P. Wachs, and Y. Edan. 2006. Human Factors for Design of Hand Gesture Human - Machine Interaction. In *IEEE International Conference on Systems, Man and Cybernetics, 2006. SMC '06*, 4052–4056. https://doi.org/10.1109/ICSMC.2006.384767
25. Yoji Uno, M. Kawato, and R. Suzuki. 1989. Formation and control of optimal trajectory in human multijoint arm movement. *Biological cybernetics* 61, 2: 89–101.
26. Frank Wilcoxon. 1945. Individual Comparisons by Ranking Methods. *Biometrics Bulletin* 1, 6: 80–83. https://doi.org/10.2307/3001968

27. Jacob O. Wobbrock, Meredith Ringel Morris, and Andrew D. Wilson. 2009. User-defined gestures for surface computing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1083–1092.
28. Michael Wright, Chun-Jung Lin, Eamonn O'Neill, Darren Cosker, and Peter Johnson. 2011. 3D Gesture Recognition: An Evaluation of User and System Performance. In *Pervasive Computing*, Kent Lyons, Jeffrey Hightower and Elaine M. Huang (eds.). Springer Berlin Heidelberg, 294–313.
29. Cheoljong Yang, Yujeong Jang, Jounghoon Beh, D. Han, and Hanseok Ko. 2012. Gesture recognition using depth-based hand tracking for contactless controller application. In *2012 IEEE International Conference on Consumer Electronics (ICCE)*, 297–298. https://doi.org/10.1109/ICCE.2012.6161876
30. Microsoft Kinect. Retrieved December 23, 2016 from https://developer.microsoft.com/en-us/windows/kinect
31. Leap Motion. Retrieved April 16, 2015 from https://www.leapmotion.com/product