Models for Rested Touchless Gestural Interaction

Darren Guinness
Dept of Computer Science Baylor University
Waco, TX 76798 USA
darren_guinness@baylor.edu

Alvin Jude
Ericsson Research
200 Holger Way
San Jose, CA, 95134
alvin.jude.hari.haran@ericsson.com

G. Michael Poor
Dept of Computer Science Baylor University
Waco, TX 76798 USA
michael_poor@baylor.edu

Ashley Dover
Dept of Computer Science Baylor University
Waco, TX 76798 USA
ashley_therriault@baylor.edu

ABSTRACT

Touchless mid-air gestural interaction has gained mainstream attention with the emergence of off-the-shelf commodity devices such as the Leap Motion and the Xbox Kinect. One of the issues with this form of interaction is fatigue, a problem colloquially known as the "Gorilla Arm Syndrome." However, by allowing interaction from a rested position, whereby the elbow is rested on a surface, this problem can be limited in its effect. In this paper we evaluate 3 possible methods for performing touchless mid-air gestural interaction from a rested position: a basic rested interaction, a simple calibrated interaction which models palm positions onto a hyperplane, and a more complex calibration which models the arm's interaction space using the angles of the forearm as input. The results of this work found that the two modeled interactions conform to Fitts's law and also demonstrated that implementing a simple model can improve interaction by improving performance and accuracy.

Keywords

Fitts; Pointing Device; Gestural Interaction; Fatigue;

CCS Concepts

•**Human-centered computing** → **Gestural input;** *Interaction paradigms;* Pointing devices;

1. INTRODUCTION

Touchless gestural interaction has achieved mainstream popularity over the last decade with the production of commercial off-the-shelf components like the Xbox Kinect and Leap Motion. Traditionally, gestural recognition has been the focus of research, while implementations have focused on coarse gestural recognition such as shape identification. However, commodity gestural recognition hardware has become increasingly more precise, with devices in some cases demonstrated to be accurate up to a sub-millimeter level [16] with low latency [6]. Due to these advances, researchers have started looking at gestures not simply for coarse actions, but for more fine use cases such as cursor navigation [20, 6, 32, 34].

The most common implementations of gestural input involve the user holding their arms out in mid-air. This mode of interaction results in arm and shoulder fatigue [38, 40, 17] commonly referred to as "Gorilla Arm Syndrome" [7]. This will need to be addressed before gestures can be accepted as a ubiquitous mode of interaction. Brown et al identified a simple way to address this problem by allowing the user to rest the elbow on a surface [6]. However, this simple solution may result in an interaction that is non-intuitive to the user, because it gives very little consideration to the actual mechanics of the human arm. In order to address this, Jude et al introduced a potential improvement called Personal Space [20] which used a calibration step to first map the user's input space.

In both aforementioned approaches, fatigue is addressed by performing gestural interaction from a rested position. However, they make separate but related claims: Brown et al [6] stated that fatigue is addressed simply by resting the elbow, while Jude et al [20] claimed that fatigue was addressed by modeling the user's input space from a rested position.

In this paper, we seek to investigate both approaches further, using standard evaluation methodologies provided by the ISO 9241-9 documentation. We aim to identify whether differences in performance exist between gestural interactions by using 3 different strategies: (1) a completely unmodeled approach, (2) the simple model introduced by Jude et al [20] which models palm positions onto a hyperplane, and (3) a more complex model introduced here, which models the interaction space with no loss of information, using the angles of the forearm as input. We also aim to identify whether learning is present in these gestural interfaces through the use of a longitudinal design, as current works that use only one session have not found learning [6, 34, 1].

For this study, the following two hypotheses were identified:

H1 Users will learn gestural interaction over time, allowing for an improvement in performance.

H2 An interaction with a model of the interaction space will perform better than an interaction which does not model the space.

The following metrics were collected to compare interactions: performance, accuracy, and subjective user feedback. Each of these metrics will be further explained within the following sections.

SUI '15 Aug 8–9, 2015, Los Angeles, CA, USA

ISBN 978-1-4503-3703-8/15/08…$15.00

DOI: `http://dx.doi.org/10.1145/2788940.2788948`

2. RELATED WORKS

2.1 Gestural Interaction

Gestural Interaction is a technique that leverages gestures from the body to interact with a computer. This type of interaction technique has been studied for over 3 decades since Richard Bolt's first implementation in "Put-That-There" [5]. Gestural interaction implementations are typically divided into two different types: (1) those that require the user to wear gloves, devices or specific markers and (2) touchless gestural interaction. The latter leverages the "Come As You Are" design principle [39], which states that users should not be required to wear devices or specific markers to interact with a system [40].

Recent devices such as the Xbox Kinect, Leap Motion, and Myo Armband have gained popularity amongst researchers, with some capable of sub-millimeter accuracy in static situations [16]. These devices have demonstrated the potential use of gestural interfaces in medical professions that require sterile environments [40, 29, 4], as an accessibility device for those with impairments [2, 15], and in mixed reality environments with head mounted displays [12]. These applications demonstrate the usefulness of gestural interaction, but more work is still needed before its ubiquitous adoption.

2.1.1 Gestural Fatigue

The standard method of using gestures involves users holding their arms up to the display for long periods of time. This has been known to cause a fatigue problem referred to as the "Gorilla Arm Syndrome" [42, 7, 40] and is considered to be a "known limitation" [38] of gestural interaction. Segen and Kumar stated that fatigue is one of the biggest issues with gestures after prolonged interaction [36].

A simple method of overcoming this issue is to allow users to rest their elbows on a chair armrest [14, 36]. Brown et al implemented this method to perform cursor navigation in an experiment with 2 possible modes of input: the whole hand and finger pointing. Jude et al implemented a very similar approach to the previous whereby cursor navigation was done with the whole hand, but the user's input space was first modeled during a calibration stage [20]. All these approaches make the same claim: performing gestural interaction from a rested position results in a more comfortable interaction and reduces the fatigue inherent to gestural interaction.

We accept the premise of the research above and attempt to further the knowledge in the area by investigating the effects of modeled versus unmodeled interactions.

2.1.2 Gestural Pointing

Recent touchless gestural pointing has been implemented in one of two general pointing methods. The first method, known as "ray pointing," is a popular approach used by many designers when implementing gestural interaction [6, 25, 3, 18]. Ray pointing uses ray casting to determine where the user is pointing [18]. This method has been demonstrated to be rapid but inaccurate [8].

Conversely, 'whole hand pointing', directly maps the 2D movements of the hand (by dropping 1 dimension) to the 2D cursor on screen [6]. The user can then move their hand or finger within this navigation space to move the cursor on screen. This method, implemented by [8, 6, 20], has been shown to be both rapid and accurate, even without visual feedback [8].

Our paper uses 'whole hand pointing' for the *Unmodeled* and *Hyperplanar* approaches. The *Spherical* approach combines both methods described above.

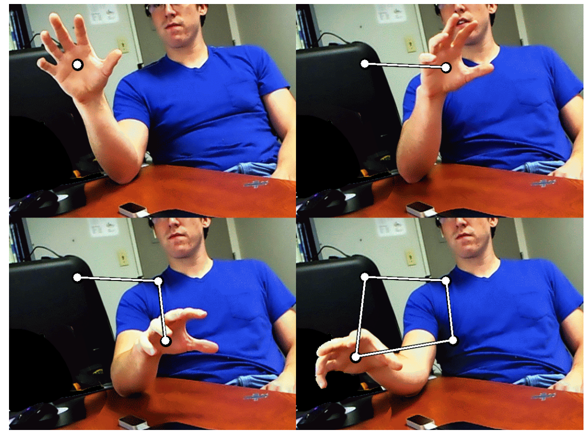

Figure 1: User calibrating their space taken from [20] with permission. This method of calibration was used for both the *Hyperplanar* and *Spherical* models.

2.1.3 Gestural Selection

Brown found that the finger tap gesture, which was considered the closest gesture to a mouse selection, performed inadequately for use in their experiment [6]. To combat this, [6] implemented a bimanual selection method, where users used their other hand to press the space bar when the cursor was over the target. Other researchers [17, 20, 34] used a 'hover-select' or 'dwell' method which required the user to hover over the target for between 250 and 1500 milliseconds. This was the method we chose to implement in our own experiment.

2.2 Learning Effects

Schmidt and Lee indicated that while we cannot directly observe learning, we can measure and report performance improvements, from which we can infer learning [35]. New pointing devices are expected to demonstrate performance improvements over time, making a one day study less descriptive of the performance of the device. To account for these improvements, researchers have run longitudinal studies, performing analysis on the data when no performance improvements were found [26]. This approach was used in our experiment in line with H1 over 3 days, based on results from our pilot which showed no significant improvements after day 3.

A simple method to report performance improvements would be to calculate the difference between means of throughput between both rounds. Researchers have shown that a better way to measure performance increase is with effect size [19] measured in Cohen's *d* [10], a practice which has recently been encouraged for use in the HCI community [21]. This metric represents the difference of the mean between 2 groups over the standard deviation:

$$\mathrm{EffectSize} = \frac{\overline{\mathrm{x_1}} - \overline{\mathrm{x_2}}}{\mathrm{s}} \tag{1}$$

In our experiment, we measure the difference in performance between days to indicate performance gained.

Researchers have indicated that random-order practice, otherwise known as Distributed Practice (e.g. A-B-C, B-C-A, C-A-B), generally benefits motor learning more than block order practice, known as massed practice (A-A-A, B-B-B, C-C-C) [24, 23]. A 3x3x3 balanced Latin cube design was used in our experiment, making the distributed practice approach trivial.

2.3 Pointing Device Evaluation

Since the focus of the paper is primarily gestural interaction, general pointing device evaluation related works will not be thoroughly reviewed. A brief explanation of the metrics is provided below, but we highly recommend the following works [13, 26, 37] for a more detailed read.

2.3.1 Fitts's Law

Fitts's law describes the relationship between Movement Time, distance, and accuracy for people in rapid motor tasks [37]. If an pointing interaction conforms to Fitts's law, it can be used as a predictive model used in user interface design [37]. To investigate whether an interface conforms to Fitts's law, a linear regression is performed over movement times (MT) on the corresponding effective Index of Difficulty conditions. The result is a regression equation which takes the form:

$$MT = a + b \times IDe \tag{2}$$

where a is the intercept term, b is the slope of the regression line and IDe is the index of difficulty.

The intercept a is desired to be near zero, and if positive is ideally below 400ms [37]. Once the regression has been performed, the *coefficient of determination (R^2)*, or "goodness of fit", is typically reported as it signifies the strength of the association between the IDe and MT [32]. R^2 is typically described as the degree to which the regression model explains the variation of the data, and the higher the R^2, the better the fit [27]. E.g., if $R^2 = .9$, this would mean that the regression line explains 90% of the variation.

The *linear correlation coefficient (R)* measures the strength and direction of the linear relationship between IDe and MT [32]. When there is no linear correlation or a weak linear correlation, R is close to 0. A correlation is described as strong if $R > 0.8$ and weak when $R < 0.5$ [32].

2.3.2 Performance

We measure Index of Difficulty (IDe) using an adjusted Shannon formulation:

$$ID_e = log_2\left(\frac{D_e}{W_e} + 1\right) \tag{3}$$

This formula has been adjusted to use the distance between the starting point of the cursor and the ending point, known as effective distance (D_e). Width has also been updated to use effective width (W_e), which utilizes the idea that the endpoints are normally distributed. We also have updated effective width (W_e) to use the standard deviation of both x and y. We use this to compute bivariate throughput, which has been demonstrated to have higher explanatory power [41]. Using the above formulation of *IDe*, we measure performance, or throughput, as:

$$\text{Throughput} = \frac{ID_e}{MT} \tag{4}$$

2.3.3 Accuracy Measures

Apart from performance measured in throughput, we also measure the accuracy of each interaction. The accuracy measures introduced by Mackenzie et al [26] include Target Re-entry (TRE), Task Axis Crossing (TAC), Movement Direction Change (MDC), Orthogonal Direction Change (ODC), Movement Variability (MV), Movement Error (ME), and Movement Offset (MO). Each accuracy metric characterizes a difference between the optimal path of the cursor and the actual path.

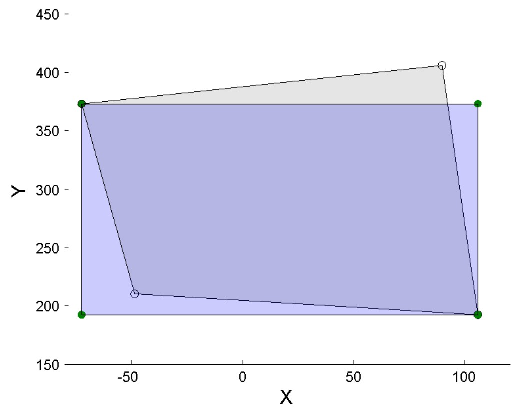

Figure 2: A front view of the Unmodeled (blue plane) and Hyperplanar (grey plane) interaction space. Both axis denotes interaction space measured in millimeters.

3. INTERACTION DESIGN

The interactions in this experiment were used specifically to test the hypotheses of this research. Therefore, we used an unmodeled interaction described by Brown et al [6], a simple model introduced by Jude et al [20], and a more complex model which we introduced for this experiment. We also built 3 angled stands $(30, 36, 44°)$ with Legos, to hold the Leap on a tilted incline as recommended by the previous authors [20].

3.1 Unmodeled Gestural Interaction

In this interaction, the space is constructed using two diagonal points from the calibration as shown in Figure 2. This model was replicated from [6], which drops the Z dimension and only uses X and Y for input into the source matrix. The points were chosen such that each screen boundary could be obtained without lifting the elbows from the table. We identify this approach as the *Unmodeled* interaction.

3.2 Rested & Calibrated Interaction – Hyperplanar Model

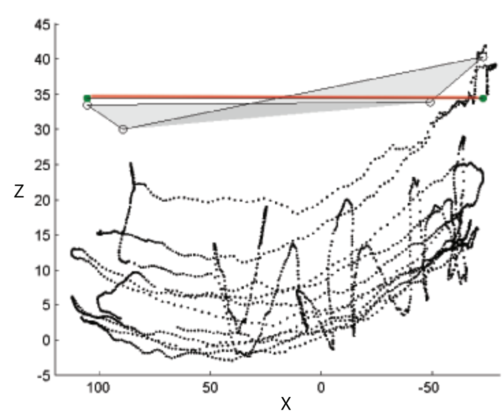

Figure 3: A top view of the Unmodeled (red plane) and Hyperplanar (grey hyperplane) with tracked hand movement (black points). Both axis denotes interaction space measured in millimeters.

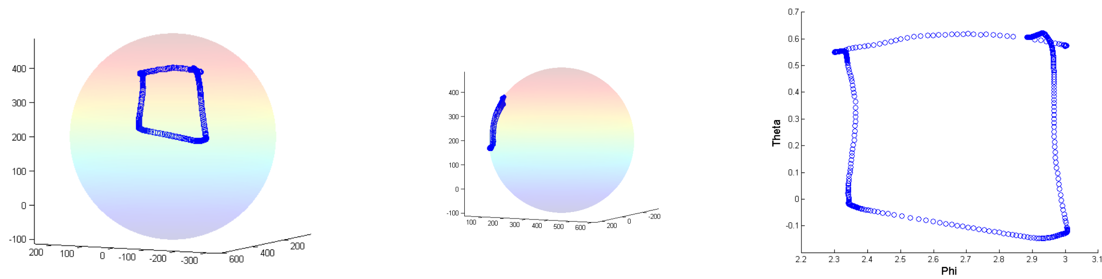

Figure 4: L-R: (1) Front view of *Spherical model*, (2) side view of the *Spherical* interaction, (3) plotting the angles (θ, ϕ) which is essentially a quadrangle.

The intuition behind this interaction is that the unmodeled interaction creates a space that does not map well to the interaction space. The example in figure 2 shows how it could be difficult for the user to hit the bottom left of the interaction space. By first modeling the interaction space, this interaction allows the users to be able to easily reach all corners of the screen from a rested position.

This interaction was originally referred to as a "planar" interaction. We however believe the term "hyperplanar" is more descriptive as as there is no guarantee that the 4 calibrated points in 3D lie on the same plane. This is better illustrated in figure 3.

We also note that despite the *Hyperplanar* approach being closer to the interaction space, it is not as accurate as it could be. The black dots in figure 3 show the actual interaction, which indicates some information loss, despite all 3 dimensions (X, Y, Z) in cartesian space used to build the model.

3.3 Rested & Calibrated Interaction – Spherical Model

The *Spherical* model introduced here is based on 2 intuitions. The first is that the movement of the hand from a rested position forms a part of a sphere, as shown in Figure 4. And second, that controlling an inherently 2-dimensional interface such as a monitor will be easier if the input itself is based on 2 dimensions. We built the *Spherical* model on both these intuitions. The interaction itself uses a spherical coordinate system with the azimuth (θ), mapped to medial and lateral shoulder rotation, and zenith (ϕ), mapped to elbow flexion and extension [28], of the forearm as input, making it 2-dimensional. This translation equates to a feature reduction from 3 features (X, Y, Z) to 2 (θ, ϕ) and a constant radius (r) with no loss of information. In contrast, other interactions that perform a reduction in dimensions generally do so by eliminating one dimension, generally Z or depth.

When performing the transformation from the input to the corresponding output, this feature reduction effectively removes one linearly independent column from the matrix, which causes a loss of precision as the matrix is now further from full rank [31]. To account for this we incorporate a plane-to-plane homography, otherwise known as a projective transformation with homogeneous estimation, which projects a 3rd dimension into a 2D image. We use this 3rd dimension to preserve rank during the transformation. The transformation also provides solutions to both determined and overly-determined systems of linear equations with a bounded error [11].

Although the input required is the azimuth (θ) and zenith (ϕ) of the forearm, these values were not directly obtainable from our input device, the Leap Motion controller. We did not consider using a different device that did have these values as it would be an unfair experiment, where there would be interaction with the input device. We therefore used the provided input by the Leap Motion controller, which is the X, Y, and Z positions of the palm in Cartesian coordinates, and transformed it to corresponding θ, ϕ and a static r. We measured the length of participants' forearms as radius r. The center of the interaction sphere is fixed and marked on the table surface, and the users are expected to position their elbow on this exact point throughout the experiment. Given these inputs, we were able to translate the coordinates in X, Y and Z to θ and ϕ. We then used θ and ϕ as input and the screen coordinates x and y as the intended output. We observed that using this coordinate system produces a model that closer represents the user's input, including being able to account for the curvature in the input, which was not achievable in the *Hyperplanar* model.

4. EXPERIMENTAL DESIGN

4.1 Participants

15 participants (M=9, F=6) between 19-27 years of age (mean = 20.6) took part in the experiment. All participants were students from a local university. Two participants self-reported as ambidextrous, but all elected to use their right hand for all interactions. All but 3 participants had used gestural interfaces before. Previously used gestural interfaces were limited to the Wii, Kinect, and/or PlayStation Move. Participants were compensated for their time.

4.2 Apparatus

All testing was conducted in a lab setting on a 30-inch Dell monitor set to 2560 x 1600 resolution. The Leap Motion Controller was used to recognize the hand position for the gestural navigation. The computer used an Intel i7-3820 CPU with 8 cores clocked at 3.6 GHz, with 32 GB RAM and ran Windows 7. The 36 degree stand was used in all cases except for 2 participants with longer forearms, for which the 30 degree stand was chosen.

4.3 Task

The ISO 9241-9 ring-of-circles task implementation from [41] was used to evaluate the performance of each interaction. This software was modified such that a 250 millisecond hover is used for selection. The task utilized 4 amplitudes {256, 512, 1024, 1408} and 3 target widths {64, 96, 128 } for 10 unique IDs ranging from 1.52-4.58 bits. Target amplitude and widths were identified from current literature [6] and extensive piloting. As we were using a large display, piloting revealed that a target width of 64px was the smallest target that was able to be selected by participants. The first 3 trials of each condition were taken as practice since this was the default in the software used.

4.4 Design

We used a 3x3x3 balanced Latin cube design with 3 interaction styles over 3 days and 3 orders. Each day included 1 session which lasted roughly 1 hour and was split into 2 rounds. In each round, participants would use all of the three interactions based on the Latin cube ordering. Participants were not told that they were using different interaction models.

4.5 Procedure

Participants were required to watch a video detailing the interaction and calibration method before they began trials on the first day. After the video, the calibration stage would begin. Once calibrated, participants were asked to test the interaction. A recalibration was allowed until they were pleased with the interaction. After which, participants were asked to watch a video detailing the ring-of-circles task. They then performed the task using the three gestural models. Participants were encouraged to take notes on the interaction they just used after each task, for ranking purposes. This was repeated until all 6 tasks (3 interactions $\times$ 2 rounds) were evaluated. Participants were then asked to rank the interactions from best to worst. These steps were repeated exactly every day of the experiment, with the videos only shown on day 1.

Only 1 calibration (see Figure 1) was performed each day at the beginning of the experiment to control for differing calibrations. Each model was then dynamically computed from the original source input points. All gestural interactions were performed with an off-the-shelf Leap Motion controller.

5. RESULTS

Analysis of our results was done to investigate 5 main aspects of each interaction: (1) Performance Improvement, (2) Conformance to the Fitts's Model, (3) Performance, (4) Accuracy, and (5) Subjective User Feedback. Metrics from (2), (3), and (4) were measured per trial across all participants per day. In each ring-of-circles task there were 23 trials (3 practice) in each of the 4 amplitude $\times$ 3 width conditions, which meant that there was a total of $20 \times 4 \times 3 = 240$ trials per task per participant. We incorporated 2 rounds with 15 participants, for a total of 240 $\times$ 2 rounds $\times 15$ participants $= 7200$ trials per interaction per day. In each analysis, the assumption of Sphericity was violated, thus a Greenhouse-Geisser (p_{GG}) epsilon correction was used to determine significance. Post-hoc tests were administered if $p_{GG} < .05$ using the MATLAB 'Bonferroni Method', which uses critical values from the t-distribution after an adjustment for multiple comparisons is made. We report the effect size of the Repeated Measures ANOVA as η_p^2 which is interpreted (0.01 = small, 0.06 = medium, 0.14 = large) as determined by Cohen [9, 33]. We report pairwise effect size, as measured by Cohen's *d* and fall back on Cohen's own guidelines for practical significance (0.2 = small, 0.5 = medium, 0.8 = large), as there are no domain-specific guidelines for pointing device evaluation.

5.1 Daily Improvement

We measure daily improvement as the difference in bivariate throughput between days in order to check for learning effects. These values are shown in Table 1, and all interactions were found to have statistical difference between days. The *Unmodeled* interaction showed a significant difference in performance between days ($F(2, 14370) = 3573$, $p_{GG} < .001$, $\eta_p^2 = 0.33$), as did the *Hyperplanar* interaction ($F(2, 14370) = 2786$, $p_{GG} < .001$, $\eta_p^2 = 0.28$) and the *Spherical* interaction ($F(2, 14370) = 2437$, $p_{GG} < .001$, $\eta_p^2 = 0.25$). The post-hoc test showed all interactions demonstrated significance between all days (*p <.001*).

	Unmodeled			Hyperplanar			Spherical		
	$\bar{x}$	σ	***d***	$\bar{x}$	σ	***d***	$\bar{x}$	σ	***d***
1	2.35	.38	-	2.44	.43	-	2.32	.40	-
2	2.69	.48	.77	2.74	.40	.74	2.61	.39	.74
3	2.79	.44	.23	2.82	.41	.18	2.67	.38	.15

Table 1: Mean ($\bar{x}$) and standard deviation (σ) of bivariate throughput for each device across days 1-3. The corresponding *d* values indicate difference in performance from the previous day measured with Cohen's d.

Due to these differences, we perform the full analysis of the interactions for performance, accuracy, and subjective feedback on data from day 3 only, while days 1 and 2 are considered practice. Therefore, all metrics reported in this paper is from day 3 of the experiment, unless stated otherwise.

5.2 Fitts's Regression

	Unmodeled	Hyperplanar	Spherical
Intercept	585.9	79.9	16.3
Slope	155.5	295.7	331.5
R^2	0.018	0.738	0.758
R	0.418	0.859	0.870

Table 2: Metrics for Fitts's regression of each interaction.

A Fitts's law model for each interaction was built by regressing the mean movement times (MT) on the corresponding effective Index of Difficulty conditions. The result is a regression equation in the form of Equation 2.

The results for each day and interaction are shown in Table 2, while a visualization of the data from day 3 can be seen in figure 5. The plots show that the modeled approaches are better explained by Fitts's law than the unmodeled approach. Additionally, Table 4 shows that the two modeled approaches consistently improve their fit between days, while the unmodeled approach does not.

5.3 Performance

The speed (specifically the Movement Time) of each interaction is considered a naive metric [37], but is reported nonetheless as we consider it a good description of each interaction. A better metric to use is throughput, and specifically bivariate throughput [41].

5.3.1 Movement Time

	Unmodeled		Hyperplanar		Spherical	
	Mean	**Stdev**	**Mean**	**Stdev**	**Mean**	**Stdev**
3	**1073.6**	495.39	**1063.4**	481.89	1110.4	538.24

Table 3: Mean and standard deviation of movement time per trial for each interaction in milliseconds. Smaller values are better.

A repeated measures ANOVA showed a significant difference in movement time, which was adjusted for dwell time, between interactions ($F(2, 14370) = 18.619$, $p_{GG} < .001$, $\eta_p^2 = 0.003$). The post-hoc test found that the *Unmodeled* interaction had significantly lower movement time than the *Spherical* interaction (*p <.001*, *d* = 0.07). The post-hoc test also showed that *Hyperplanar* interaction had significantly lower movement time than the *Spherical* (*p <.001 d* = 0.09).

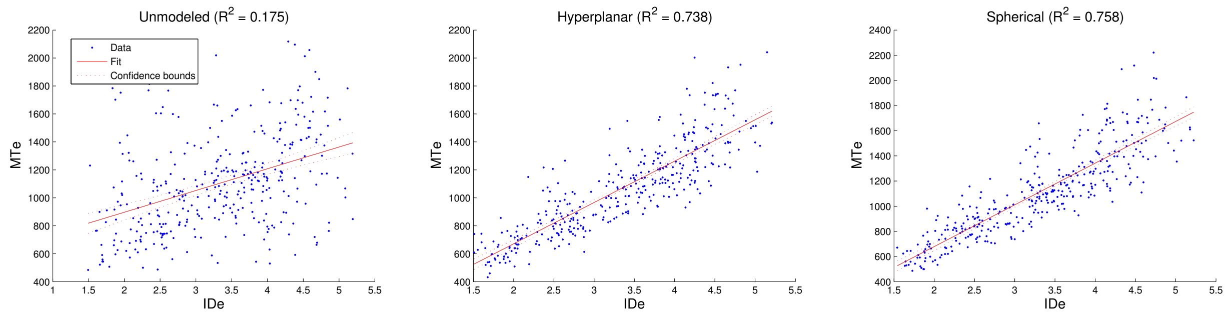

Figure 5: Mean movement time of all participants from day 3 as a function of effective Index of Difficulty (IDe), with computed Fitts's regression lines

	Unmodeled				Hyperplanar				Spherical			
	Intercept	**Slope**	R^2	R	**Intercept**	**Slope**	R^2	R	**Intercept**	**Slope**	R^2	R
1	453.05	248.18	0.31	0.56	113.29	334.08	0.63	0.79	18.97	381.91	0.65	0.81
2	785.04	108.17	0.07	0.26	59.47	308.46	0.73	0.86	38.62	332.73	0.71	0.84
3	585.86	155.53	0.18	0.42	79.86	295.67	0.74	0.86	16.30	331.50	0.76	0.87

Table 4: Intercept, slope, coefficient of determination (R^2) and correlation coefficient (R) of all 3 interactions over all 3 days.

5.3.2 *Bivariate Throughput*

A repeated measures ANOVA showed a significant difference in performance between interactions measured with bivariate throughput ($F(2, 14370) = 369.81$, $p_{GG} < .001$, $\eta_p^2 = 0.05$). A post-hoc test showed the *Hyperplanar* interaction had higher bivariate throughput than both the *Unmodeled* (p <*.005*, d = .05) and *Spherical* (p <*.001*, d = 0.35) interactions. The post-hoc test also showed that the *Unmodeled* interaction had higher bivariate throughput than the *Spherical* interaction (p <*.001*, d = 0.29).

5.4 Accuracy Measures

We report accuracy measures based on the metrics introduced by Mackenzie [26] consisting of Target Re-entry (TRE), Target Axis Crossing (TAC), Orthogonal Direction Change (ODC), Movement Variability (MV), Movement Error (ME) and Movement Offset (MO). These measures were taken directly from the *FittsStudy* software [41], except for TRE, as the software reports target entries (TE) instead, where TRE=TE-1. A lower number is better for all metrics except MO, where a closer distance to 0 is better.

	Unmodeled		Hyperplanar		Spherical	
	Mean	**Stdev**	**Mean**	**Stdev**	**Mean**	**Stdev**
TE	**1.27**	0.68	**1.28**	0.70	1.35	0.78
TAC	**1.45**	1.20	**1.49**	1.17	1.55	1.25
MDC	**2.77**	1.69	**2.78**	1.66	2.87	1.79
ODC	**1.08**	1.30	**1.05**	1.25	1.19	1.38
MV	23.50	20.73	**22.94**	18.85	23.83	20.06
ME	26.70	20.42	**25.72**	18.67	26.42	19.26
MO	-7.81	28.32	4.51	26.57	**2.04**	27.25

Table 5: Mean and standard deviation of accuracy measures metrics taken on day 3 of the experiment.

5.4.1 *Target Entries*

The *Unmodeled* interaction had the best TE score, followed by the *Hyperplanar* interaction. A repeated measures ANOVA showed a significant difference in the number of Target Entries between interactions ($F(2, 14370) = 27.20$, $p_{GG} < .001$, $\eta_p^2 = 0.004$). The post-hoc test identified a statistical significance between the *Unmodeled* and *Spherical* interactions (p <*.001*, d = 0.11), and between the *Hyperplanar* and *Spherical* interactions (p <*.001*, d = 0.10).

5.4.2 *Task Axis Crosses*

The *Unmodeled* interaction had the best TAC score, followed by the *Hyperplanar* interaction. A repeated measures ANOVA showed this to be statistically significant ($F(2, 14370) = 11.054$, $p_{GG} < .001$, $\eta_p^2 = 0.002$). The post-hoc test identified a significant difference between the *Unmodeled* and *Spherical* interactions (p <*.001*, d =0.08). The *Hyperplanar* interaction was also found to be significantly different from the *Spherical* interaction (p <*.05*, d = 0.04).

5.4.3 *Movement Direction Change*

The *Unmodeled* interaction had the best MDC score, followed by the *Hyperplanar* interaction. A repeated measures ANOVA showed this to be statistically significant ($F(2, 14370) = 8.262$, $p_{GG} < .001$, $\eta_p^2 = 0.001$). Post-hoc tests showed a significant difference between the *Unmodeled* and *Spherical* interactions (p <*.005*, d = 0.06). The post-hoc test also identified a significant difference between the *Hyperplanar* and *Spherical* interactions (p <*.005*, d =0.06).

5.4.4 *Orthogonal Direction Change*

The *Hyperplanar* interaction had the best ODC score, followed by the *Unmodeled* interaction. A repeated measures ANOVA found this to be statistically significant ($F(2, 14370) = 23.786$, $p_{GG} < .001$, $\eta_p^2 = 0.003$). A post-hoc test showed a significant difference between the *Unmodeled* and *Spherical* interactions (p <*.001*, d = 0.08), and between the *Hyperplanar* and *Spherical* interactions (p <*.001*, d =0.10).

5.4.5 *Movement Variability*

The *Hyperplanar* interaction had the best MV score, followed by the *Unmodeled* interaction. A repeated measures ANOVA found a significant difference between interactions ($F(2, 14370) = 3.918$, $p_{GG} < .05$, $\eta_p^2 = 0.0005$). A post-hoc tests showed a significant difference between the *Hyperplanar* and *Spherical* interactions (*p* <*.05*, *d* = 0.05).

5.4.6 Movement Error

The *Hyperplanar* interaction had the best ME score, followed by the *Spherical* interaction. A repeated measures ANOVA showed this to be statistically significant ($F(2, 14370) = 5.161$, $p_{GG} < .01$, $\eta_p^2 = 0.0007$). The post-hoc test identified a significant difference between the *Hyperplanar* and *Unmodeled* interactions (*p* <*.005*, *d* = 0.05).

5.4.7 Movement Offset

The *Spherical* interaction had the best MO score, followed by the *Hyperplanar* interaction. A repeated measures ANOVA showed this to be statistically significant ($F(2, 14370) = 424.97$, $p_{GG} < .001$, $\eta_p^2 = 0.06$). The post-hoc test identified a significant difference between the *Spherical* and *Hyperplanar* interactions (*p* <*.001*, *d* =0.09), and between the *Spherical* and *Unmodeled* interactions (*p* <*.001*, *d*= 0.36). The post-hoc test also showed that the *Hyperplanar* interaction was significantly different than the *Unmodeled* interaction (*p* <*.001*, *d* =0.45).

5.5 Subjective User Feedback

5.5.1 Preference

Participants were asked to rank the interactions by preference on a scale of 1-6, with 1 being the best interaction and 6 being the worst. We combined the rounds that used the same interaction. From these rankings we can order the interactions by most preferred to least as such: 1) *Hyperplanar*, 2) *Unmodeled*, 3) *Spherical*, with their respective scores of 2.57, 3.60, 4.33. A Friedman test showed that there was a statistically significant difference in preference rank between the interactions ($\chi^2(2) = 13.505, p = 0.0012$). Post hoc analysis with Wilcoxon signed-rank tests was conducted with a Bonferroni correction applied. The only statistically significance found was that the *Hyperplanar* interaction was significantly preferred over the *Spherical* ($Z = -3.524, p < .001$, $r = .23$).

5.5.2 Usability

	Unmodeled	Hyperplanar	Spherical
Operation smoothness	4.13	4.27	3.97
Operational effort	4.20	4.30	4.03
Accuracy	3.82	4.27	3.80
Operation speed	4.27	4.33	4.23
General comfort	4.27	4.40	4.07
Overall operation	4.17	4.33	4.03

Table 6: Mean reported usability & comfort metrics per interaction (1 = Most Negative, 5 = Most Positive).

Participants were asked to fill out 5-point independent rating Likert scale questions from Annex C. of the ISO 9241-9 which evaluated usability and comfort of the interaction immediately after using the interaction. Table 6 depicts the questions and the mean reported ratings about the usability the interaction. The reported ratings show reasonably high usability in all interactions. Friedman tests showed no statistical significance between interactions.

5.5.3 Fatigue

Participants filled out a 5-point independent rating Likert scale questions from ISO 9241-9 Annex C to collect data in regards to fatigue. Table 7 depicts the questions and the mean reported ratings per interaction. The reported ratings demonstrate a minor presence of fatigue in each interaction. Friedman tests showed no statistical significance between interactions.

	Unmodeled	Hyperplanar	Spherical
Finger fatigue	4.53	4.50	4.50
Wrist fatigue	4.37	4.37	4.17
Arm fatigue	4.33	4.33	4.33
Shoulder fatigue	4.57	4.53	4.40
Neck fatigue	4.87	4.87	4.87

Table 7: Mean reported fatigue per interaction (1 = Extreme, 5 = None)

5.5.4 Borg Scale

	Unmodeled	Hyperplanar	Spherical
Arm Effort	1.20	1.37	1.27
Shoulder Effort	0.95	1.02	1.05
Neck Effort	0.15	0.20	0.27

Table 8: Mean reported effort per interaction (0 = Nothing at all, 0.5 = very very weak (just noticable), 1 = very weak, ..., 10 very, very strong)

Previous work in gestures encouraged the use of the Borg Scale for arm, shoulder, and neck effort [17], and were therefore included in our subjective assessment. We used the Borg scale from ISO 9241-9 Annex C. Table 8 shows the questions and the mean reported ratings for each interaction. The table demonstrates that minimal effort was required when using the interactions. Friedman tests showed no statistical significance between interactions.

6. ANALYSIS

The goal of this study was to evaluate the hypotheses using the metrics from the previous section. We report our findings of the hypotheses by splitting hypothesis H2 to better illustrate the results based on each of the interactions.

H1 Users will learn gestural interaction over time, allowing for an improvement in performance.

We accept **H1** as the performance improvement analysis showed a significant improvement in all models between each day, which is indicative of learning.

H2a An interaction with a simple model of the interaction space will perform better than an interaction which does not model the space

We accept **H2a**. This decision was based on a comparison of performance, accuracy and Fitts's law conformance between the *Hyperplanar* and *Unmodeled* interaction.

In terms of performance, The *Hyperplanar* interaction obtained statistically better bivariate throughput than the *Unmodeled* interaction. However, this improvement was minor and not practically significant based on Cohen's interpretation.

With respect to accuracy, the *Hyperplanar* interaction obtained statistically better Movement Offset than the *Unmodeled* interaction, which was found to be a small-medium effect as defined by Cohen. It also had statistically better Movement Error, although the practical significance is minor.

In terms of Fitts' conformance, the *Hyperplanar* interaction demonstrated a strong linear correlation (R) between Movement Time and IDe, had an intercept that was within the ideal range, and a higher goodness of fit (R^2). The *Unmodeled* interaction, on the other hand, demonstrated a weak linear correlation (R), an intercept that was not within the ideal range, and a lower R^2. It can also be seen in Table 4 that the *Hyperplanar* interaction demonstrated an improvement over time with regards to Fitts's law, while the *Unmodeled* interaction did not.

Since the *Hyperplanar* interaction was found to be as good, or better, in the individual metrics of performance, accuracy, and predictability as measured by Fitts's conformance, we consider the *Hyperplanar* interaction as a whole to obtain better results than the *Unmodeled* interaction.

H2b An interaction which models the interaction space using a sphere will perform better than an interaction which uses a hyperplane.

We fail to reject the null hypothesis in the case of **H2b**. To evaluate this hypothesis we compared the *Hyperplanar* and *Spherical* interactions using each of the collected metrics: performance, accuracy, and Fitts' conformance.

In terms of performance, the *Hyperplanar* interaction obtained statistically better bivariate throughput than the *Spherical* interaction, which was found to be a small-medium effect as defined by Cohen.

In terms of accuracy, *Hyperplanar* interaction obtained statistically better accuracy than the *Spherical* interaction in 5 of the 7 accuracy metrics. None are considered to be practically significant. The *Spherical* interaction obtained significantly better Movement Offset than the *Hyperplanar* interaction, but was not practically significant.

Fitts' regressions demonstrated a strong linear correlation, an intercept within the ideal range, and a relatively high 'goodness of fit' in both the *Hyperplanar*, and *Spherical* interactions. However the *Spherical* interaction was found to have a higher correlation, goodness of fit, and an intercept that was closer to the ideal value when compared to the *Hyperplanar* interaction.

In addition to the aforementioned quantitative metrics, users also showed preference towards the *Hyperplanar* interaction which was shown to be significant. Taking all these into consideration, we cannot conclude that the *Spherical* interaction performs better than the *Hyperplanar*, thus we fail to reject the null hypothesis.

7. DISCUSSION AND FUTURE WORKS

Over the course of this work, we found performance improvements between all 3 days of the study which we believe is caused by participants learning the interaction. While this may seem trivially true, current literature has in some cases failed to identify learning in these gestural interfaces [6, 34, 1]. We believe this difference is due to our experimental design, which used a longitudinal study and random-order practice.

We found that fatigue and effort were reported to be minimal from the rested position, as demonstrated in Table 7 and Table 8. This finding further reinforces current literature that has stated that resting the elbow during interaction reduces fatigue [14, 36, 6, 20].

We also found that modeling the interaction space resulted in an interaction which conforms to Fitts's law. Conversely, we found that an interaction which does not model the input space was not well explained by Fitts's law, nor was it within the ideal intercept range in each of the 3 days. This means that the relationship between movement time and IDe is not consistently predictable in the unmodeled approach, and offers less value to those designing interfaces [37]. The approach used to detect these results was in line with the current standard's encouragement to regress the participant's mean movement time (MT) over their respective effective Index of Difficulty(IDe), as opposed to regressing over mean of means per Index of Difficulty [37].

We found that the interaction which used a simple model was significantly better than the unmodeled interaction in terms of both performance and accuracy, and only the latter was found to be practically significant as interpreted by Cohen. However, we also found that the interaction whose model was more complex in terms of the input space performed significantly worse than the other two models despite its mapping to the biomechanics of the body. This may be caused by the strict assumptions of the model. The two azimuthal (θ) and elevation (ϕ) angles were inferred from the previously measured elbow position and the palm position provided by the Leap Motion. Therefore, this model requires a higher tracking precision in order to perform optimally.

Upon review of the notes taken during the experiment, which included participants' feedback, we identified a few issues with the *Spherical* interaction. A recurring issue was with the static elbow placement. Participants had difficulty finding their calibrated elbow position even when a marker was present. We also found that during the pilot, participants mistakenly knocked the tracker off of its stand during their note-taking between rounds. This did not seem to significantly penalize the more simple models but caused large cursor jitters during the *Spherical* interaction. We attempted to solve this issue by building a new stand using Legos which allowed the tracker to tightly fit into place. While this increased stability, the stand itself was still movable when hit. We posit that a better interaction can be built using a device which directly tracks the arm in terms of relative angles. Another option would be to allow for dynamic elbow tracking in which both the hand and the elbow are dynamically tracked so that the elbow does not need to be in a fixed position. We believe that the better fit offered by the *Spherical* approach could yield positive results if these considerations are addressed.

While we provided overall effect size using η_p^2, and pairwise effect sizes for performance and accuracy metrics as measured by Cohen's d, interpreting these effect sizes however proved to be difficult. These interpretations are meant to be domain-specific [30], but no such guidelines exist within the domain of pointing device evaluation. We had to therefore fallback on the interpretations provided by Cohen, which are meant to be used as a last resort [22]. By providing the effect size of our study, we aim to provide better context for future research, and to contribute towards establishing guidelines for interpreting effect size within this domain.

8. CONCLUSION

In this study we used a longitudinal design to evaluate the two hypotheses. We learned that modeling the interaction space results in an interaction which can be explained by Fitts's law. Conversely, we learned that an unmodeled approach conforms weakly to Fitts's law. We also learned that a simple model of the user's interaction space resulted an interaction that was as fast and more accurate than an interaction which did not model the user's interaction space. Furthermore, we introduced a more complex model of the interaction space which maps the arm movement from a rested position to the 2D screen with no loss of information using the forearm angles as input. This more complex model did not exhibit better performance nor accuracy than the simpler model. We posit that this is due to the interaction having too many constraints and being unsuitable for use with an input device which uses the hand position as input. Finally, we showed that gestural interaction demonstrated performance improvements over multiple sessions, from which we infer learning.

9. REFERENCES

[1] Adhikarla, V. K., Sodnik, J., Szolgay, P., and Jakus, G. Exploring direct 3d interaction for full horizontal parallax light field displays using leap motion controller. *Sensors 15*, 4 (2015), 8642–8663.

[2] Bailly, G., Müller, J., Rohs, M., Wigdor, D., and Kratz, S. Shoesense: A new perspective on gestural interaction and wearable applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, ACM (New York, NY, USA, 2012), 1239–1248.

[3] Banerjee, A., Burstyn, J., Girouard, A., and Vertegaal, R. Pointable: an in-air pointing technique to manipulate out-of-reach targets on tabletops. In *Proceedings of the ACM International Conference on Interactive Tabletops and Surfaces*, ACM (2011), 11–20.

[4] Bigdelou, A., Schwarz, L., and Navab, N. An adaptive solution for intra-operative gesture-based human-machine interaction. In *Proceedings of the 2012 ACM International Conference on Intelligent User Interfaces*, IUI '12, ACM (New York, NY, USA, 2012), 75–84.

[5] Bolt, R. A. *"Put-that-there": Voice and gesture at the graphics interface*, vol. 14. ACM, 1980.

[6] Brown, M. A., Stuerzlinger, W., and Filho, E. J. M. The performance of un-instrumented in-air pointing. In *Proceedings of the 2014 Graphics Interface Conference*, GI '14, Canadian Information Processing Society (Toronto, Ont., Canada, Canada, 2014), 59–66.

[7] Carmody, T. Why 'gorilla arm syndrome' rules out multitouch notebook displays. *Wired, Oct 10* (2010).

[8] Cockburn, A., Quinn, P., Gutwin, C., Ramos, G., and Looser, J. Air pointing: Design and evaluation of spatial target acquisition with and without visual feedback. *International Journal of Human-Computer Studies 69*, 6 (2011), 401–414.

[9] Cohen, J. Statistical power analysis for the behavioral sciences.

[10] Cohen, J. A power primer. *Psychological bulletin 112*, 1 (1992), 155.

[11] Criminisi, A., Reid, I., and Zisserman, A. A plane measuring device. *Image and Vision Computing 17*, 8 (1999), 625–634.

[12] Ens, B., Hincapié-Ramos, J. D., and Irani, P. Ethereal planes: A design framework for 2d information space in 3d mixed reality environments. In *Proceedings of the 2nd ACM Symposium on Spatial User Interaction*, SUI '14, ACM (New York, NY, USA, 2014), 2–12.

[13] Fitts, P. M. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology 47*, 6 (1954), 381.

[14] Freeman, D., Vennelakanti, R., and Madhvanath, S. Freehand pose-based gestural interaction: Studies and implications for interface design. In *Intelligent Human Computer Interaction (IHCI), 2012 4th International Conference on*, IEEE (2012), 1–6.

[15] Guinness, D., Poor, G. M., and Jude, A. Gestures with speech for hand-impaired persons. In *Proceedings of the 16th International ACM SIGACCESS Conference on Computers & Accessibility*, ASSETS '14, ACM (New York, NY, USA, 2014), 259–260.

[16] Guna, J., Jakus, G., Pogačnik, M., Tomažič, S., and Sodnik, J. An analysis of the precision and reliability of the leap motion sensor and its suitability for static and dynamic tracking. *Sensors 14*, 2 (2014), 3702–3720.

[17] Hincapié-Ramos, J. D., Guo, X., Moghadasian, P., and Irani, P. Consumed endurance: A metric to quantify arm fatigue of mid-air interactions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, ACM (New York, NY, USA, 2014), 1063–1072.

[18] Jota, R., Nacenta, M. A., Jorge, J. A., Carpendale, S., and Greenberg, S. A comparison of ray pointing techniques for very large displays. In *Proceedings of Graphics Interface 2010*, GI '10, Canadian Information Processing Society (Toronto, Ont., Canada, Canada, 2010), 269–276.

[19] Jude, A., Poor, G. M., and Guinness, D. An evaluation of touchless hand gestural interaction for pointing tasks with preferred and non-preferred hands. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational*, ACM (2014), 668–676.

[20] Jude, A., Poor, G. M., and Guinness, D. Personal space: User defined gesture space for gui interaction. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, ACM (New York, NY, USA, 2014), 1615–1620.

[21] Kaptein, M., and Robertson, J. Rethinking statistical analysis methods for chi. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, ACM (New York, NY, USA, 2012), 1105–1114.

[22] Lakens, D. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology 4* (2013).

[23] Lee, T. D., and Genovese, E. D. Distribution of practice in motor skill acquisition: Learning and performance effects reconsidered. *Research Quarterly for Exercise and Sport 59*, 4 (1988), 277–287.

[24] Lin, C.-H. J., Sullivan, K. J., Wu, A. D., Kantak, S., and Winstein, C. J. Effect of task practice order on motor skill learning in adults with parkinson disease: a pilot study. *Physical therapy 87*, 9 (2007), 1120–1131.

[25] MacKenzie, I. S., and Jusoh, S. An evaluation of two input devices for remote pointing. In *Engineering for human-computer interaction*. Springer, 2001, 235–250.

[26] MacKenzie, I. S., Kauppinen, T., and Silfverberg, M. Accuracy measures for evaluating computer pointing devices. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, ACM (2001), 9–16.

[27] MacKenzie, I. S., and Teather, R. J. Fittstilt: The application of fitts' law to tilt-based interaction. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design*, ACM (2012), 568–577.

[28] McLester, J., and Pierre, P. *Applied Biomechanics: Concepts and Connections*. Cengage Learning, 2007.

[29] Mentis, H. M., O'Hara, K., Sellen, A., and Trivedi, R. Interaction proxemics and image use in neurosurgery. In *Proceedings of the SIGCHI Conference on human factors in computing systems*, ACM (2012), 927–936.

[30] Morris, P. E., and Fritz, C. O. Effect sizes in memory research. *Memory 21*, 7 (2013), 832–842.

[31] Nagy, J. G., Plemmons, R. J., and Torgersen, T. C. Iterative image restoration using approximate inverse preconditioning. *Image Processing, IEEE Transactions on 5*, 7 (1996), 1151–1162.

[32] Pino, A., Tzemis, E., Ioannou, N., and Kouroupetroglou, G. Using kinect for 2d and 3d pointing tasks: performance evaluation. In *Human-Computer Interaction. Interaction Modalities and Techniques*. Springer, 2013, 358–367.

[33] Richardson, J. T. Eta squared and partial eta squared as measures of effect size in educational research. *Educational Research Review 6*, 2 (2011), 135–147.

[34] Sambrooks, L., and Wilkinson, B. Comparison of gestural, touch, and mouse interaction with fitts' law. In *Proceedings of the 25th Australian Computer-Human Interaction Conference: Augmentation, Application, Innovation, Collaboration*, ACM (2013), 119–122.

[35] Schmidt, R. A., and Lee, T. *Motor Control and Learning, 5E*. Human kinetics, 1988.

[36] Segen, J., and Kumar, S. Look ma, no mouse! *Commun. ACM 43*, 7 (July 2000), 102–109.

[37] Soukoreff, R. W., and MacKenzie, I. S. Towards a standard for pointing device evaluation, perspectives on 27 years of fitts' law research in hci. *International Journal of Human-Computer Studies 61*, 6 (2004), 751–789.

[38] Teixeira, V. *Improving elderly access to audiovisual and social media, using a multimodal human-computer interface*. PhD thesis, Faculdade de Engenharia, Universidade do Porto, 2011.

[39] Triesch, J., and Von Der Malsburg, C. Robotic gesture recognition by cue combination. In *Informatik'98*. Springer, 1998, 223–232.

[40] Wachs, J. P., Kölsch, M., Stern, H., and Edan, Y. Vision-based hand-gesture applications. *Communications of the ACM 54*, 2 (2011), 60–71.

[41] Wobbrock, J. O., Shinohara, K., and Jansen, A. The effects of task dimensionality, endpoint deviation, throughput calculation, and experiment design on pointing measures and models. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, ACM (New York, NY, USA, 2011), 1639–1648.

[42] Yoo, J., Lee, S., and Ahn, C. Air hook: Data preloading user interface. In *ICT Convergence (ICTC), 2012 International Conference on*, IEEE (2012), 163–167.