IEICE TRANS. INF. & SYST., VOL.E101–D, NO.7 JULY 2018

PAPER

Feedback Control Model of a Gesture-Based Pointing Interface for a Large Display

Kazuaki KONDO[†a)], ***Member***, **Genki MIZUNO**[†], ***Nonmember***, ***and*** **Yuichi NAKAMURA**[†], ***Member***

SUMMARY This study proposes a mathematical model of a gesture-based pointing interface system for simulating pointing behaviors in various situations. We assume an interaction between a pointing interface and a user as a human-in-the-loop system and describe it using feedback control theory. The model is formulated as a hybrid of a target value follow-up component and a disturbance compensation one. These are induced from the same feedback loop but with different parameter sets to describe human pointing characteristics well. The two optimal parameter sets were determined individually to represent actual pointing behaviors accurately for step input signals and random walk disturbance sequences, respectively. The calibrated model is used to simulate pointing behaviors for arbitrary input signals expected in practical situations. Through experimental evaluations, we quantitatively analyzed the performance of the proposed hybrid model regarding how accurately it can simulate actual pointing behaviors and also discuss the advantage regarding the basic non-hybrid model. Model refinements for further accuracy are also suggested based on the evaluation results.

key words: *pointing interface, human-in-the-loop system, control theory*

1. Introduction

Large display devices are becoming less expensive, thereby making it easier to communicate using visual contents shown on large displays, such as digital signage and slide presentation on a wide screen. Figure 1 shows an example of such open communication styles. In those environments, an easy and intuitive remote interaction scheme rather than an oscillatory interface such as a touch panel is required. A gesture-based pointing interface is one of those schemes requiring no additional equipment for users. It recognizes an indicator's pointing posture, estimates a pointed location on the screen, and shows a pointer there. However, the present interface design is not always best for users, as a result of inaccurate sensing, ambiguity of pointing postures, and variation in how they indicate. An advanced scheme will be necessary to construct a more comfortable interface.

The purpose of this study is to build a mathematical model of a gesture-based pointing interface system and analyze its performance. The model enables simulation-based performance evaluation and interface design. Once a mathematical model of a target pointing interface system is established, its behaviors in various pointing situations can be simulated and its usability can be evaluated. It can aid in designing an interface by using a trial-and-error strategy without experimental evaluation. This advantage is extremely important for human inclusive systems because conducting real evaluations with participants requires immense effort and involves difficulties in ensuring fair experimental conditions. Furthermore, the repeatability of human behavior is low. Each participant must constantly repeat pointing under the same conditions to collect sufficiently many samples from which obtaining general and essential analyses/evaluations.

Fig. 1 Open communication using pointing gesture.

In this paper, we assume a basic interface that simply draws a small circular pointer at an estimated location. This will help to clarify existing problems and provide a fundamental framework for modeling improved interfaces with advanced design. A key idea of the modeling is to assume that a pointing interface and its user construct a human-in-the-loop system, because an indicator observes relationship between how he or she wants to indicate and a shown pointer to change one's own behavior. This feedback structure is written as a control model. Additionally, we describe general pointing behaviors as a superposition of two typical feedback control loops, a target value follow-up component and a disturbance compensation one. This hybrid structure comes from the assumption that behaviors for reaching a pointing target and those for maintaining a pointer on the target have different features and are independent.

2. Related Work

Using a laser pointer is a typical way through which images projected on wide screens can be pointed at from a distance. That is quite intuitive and easy, but all of the control depends on the users and reflects bad influences present in human habits, such as hand vibrations and tendency to flick the pointer. Some previous studies have proposed the use

Manuscript received September 19, 2017.
Manuscript revised February 17, 2018.
Manuscript publicized April 4, 2018.
[†]The authors are with Academic Center for Comupting and Media Studies, Kyoto Universiy, Kyoto-shi, 606–8501 Japan.
a) E-mail: kondo@ccm.media.kyoto-u.ac.jp
DOI: 10.1587/transinf.2017EDP7298

of remote pointing interfaces [1]–[4] that use visual sensing technology and special devices to estimate the target locations on screens. This indirect way can modify a pointer location and its form adaptively, which suppresses bad influences present when directly using a laser pointer. This study also focuses on a similar type of indirect pointing interface that estimates an on-screen target location from the hand gestures of the user. Since this research does not need an additional device for pointing, users can come into and get out from open communication at any time.

A particular problem of the current gesture-based pointing interfaces is the difficulty in estimating where an indicator wants to indicate. Thus we often have considerable effort to show a pointer at the intended location. One reason for this is the accuracy of the pointing pose measurement. Significant measurement accuracy is required when using a remote pointing interface system, because even extremely small errors in body coordinates are amplified on a screen. Previous studies have proposed the use of markerless pose estimation methods that do not interfere with the indicator's behavior [5]–[7]. However, their performance do not satisfy the stringent requirements because visual sensing is weak to self-occlusion and non-rigid body deformation. Another flaw in gesture-based pointing interface is the ambiguity of pointing postures, which corresponds to the inconsistency between the 3D pose structure and the desired target location in the indicator's mind. Although previous studies [8]–[11] have investigated their relationship, the exact nature is still uncertain and is heavily influenced by existing conditions and the user's individual traits. The aforementioned studies assumed accurate measurement based on the special devices and failed to consider the influence of the pointing posture ambiguity. With these shortcomings in mind, one of the purposes in this current work is to create a mathematical model that can predict how measurement errors affect a user's pointing behavior.

Another issue is the purpose of the pointing. Most of the conventional research, including [12] and [4], assume the pointing gesture as input media for a computer like a mouse control, whereas this work assumes it as a way to make an audience aware of the target location. Since the former requires fast and accurate selection of a target, movement time (MT) for the selection [13] and the spatial error noted at that time [2] have been used to evaluate interface usability. However, pointing for acquiring attention requires additional evaluation from audience viewpoint. A rapidly moving pointer is often lost from audiences' perception. For similar reason, the pointer should remain at the target location for a duration of time so as to afford the audience the chance to recognize it. Vibrating pointers are oftentimes distracting, drawing attention away from the target visual contents. These problems highlight the importance of evaluating transient pointing behaviors and the pointer stability before and after convergence, respectively

The aforementioned issues motivated us to build a mathematical model based on the control theory in order to find a better way of handling various disturbances that may occur in such a system, including measurement errors. Since predicting a participant's pointing behaviors is a factor of temporal trajectories, we can extract features from this theory to evaluate its usability and applicability. Any system of this kind would be a good tool for analyzing particular issues associated with using a gesture-based pointing interface. Although Kondo et al. proposed the use of a feedback control model for a gesture-based pointing interface system [14], the work was only focused on the user's traits when approaching to the target location without or sufficiently small measurement errors. Our proposed model aims to extrapolate both the approaching and adjusting behavior of the user with certain individual traits to describe a general mode for superposition of the two.

3. Feedback Control Model for a Gesture-Based Pointing Interface System

3.1 Pointing Situation

In this paper, a pointing situation as shown in Fig. 2 is assumed. One of the users standing in front of a wide screen points at a particular spot and the others look at that. This situation is a typical style in communication via large scale visual contents. Our work is focused on how measurement errors can affects the indicator's pointing behaviors. Thus, we assume that pointing gesture with the arm being straight. Pointing with a bent arm and/or with the user standing at side of the screen often occur but are excluded in this study, because there will be ambiguity in the indicator's pointing postures, even under the same experimental conditions.

A sequence of pointing behavior proceeds as follows. All of the location values are defined on a screen coordinate.

1. A pointing target arises in an indicator's intention. Let its position on the screen be p_t. He or she begins a change in posture to point at it.
2. The computer estimates the pointed location as p_e using visual sensing through the camera. The pointer is shown at the location p_c determined by the visualization filter with p_e used as its input.

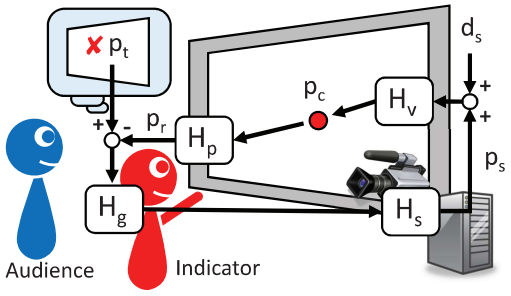

Fig. 2 The relationship between the indicator and the gesture-based pointing interface. Vision sensors measure his or her pointing posture and show a pointer at the estimated pointing location. The modules H_g, H_p, H_s, and H_v describe the indicator's body dynamics, visual perception, estimation of the pointed location, and pointer visualization. Additionally, d_s refers to disturbance given to the pointed location estimation.

3. The indicator recognizes the pointer location as p_r and adjusts the pointing posture to move it closer to p_t.
4. Steps (2)-(4) continue until $p_t - p_r$ becomes zero.

3.2 Basic Model

The above procedure can be modeled as a feedback control loop. It is formulated with four transfer functions H_g, H_p, H_s, and H_v to denote the indicator's body dynamics, the visual perception catching a pointer location, the computer estimating the indicated location, and the visualization filter, respectively [14]. Each component in the loop is formulated as follows.

Pointing Interface Side :

The estimation of the pointed location involves visual sensing using cameras and pointing pose estimation based on the measurement. However, its accuracy is not high and considerable estimation errors remain as noted in the previous section. To formulate this feature, we assume that the interface correctly estimates the indicated location with a specific latency, and that the estimation error is described as an additional disturbance d_s. We formulate H_s as

$$H_s(s) = e^{-\tau_s s} \tag{1}$$

where τ_s denotes the latency. The formulation of d_s depends on which pose estimation method is assumed. A vision-based method usually searches optimal values in a neighborhood of those in the previous frame. This results in estimation error being accumulated. Thus, in this study we assume a random walk sequence as the disturbance d_s and formulate it as

$$d_s(t) = \begin{cases} 0 & t = 0 \\ d_s(t-1) + s_f x & t > 0 \end{cases} \tag{2}$$

where x and s_f denote an update factor that follows an independent and identical standard normal distribution $\mathcal{N}(0, 1)$ and its scaling, respectively. The formulation of H_v expresses how the pointer location is determined from the estimated location. Since this study assumes simple drawing without any modification, H_v is formulated simply as

$$H_v(s) = e^{-\tau_v s} \tag{3}$$

with a latency for visualization τ_v.

Indicator Side :

Considering natural pointing, once an indicator configures a pointing target, it usually remains for some duration. This results in a formulation of the target value input p_t as a step signal

$$p_t(t) = \begin{cases} 0 & t = 0 \\ T_r & t > 0 \end{cases} \tag{4}$$

where T_r denotes the distance from an initial location. The characteristics of human pointing behavior can still be uncertain, especially from an analytic viewpoint. We configured a formulation of H_g through the investigation of actual pointing trajectories. In the preliminary experiments, we observed responses for step inputs p_T without drawing a pointer, i.e., no feedback information is provided to an indicator to introduce pure step responses of H_g. We confirmed that the step response trajectories converge after some overshooting and then formulated H_g as a second-order lag element :

$$H_g(s) = e^{-\tau_g s} \frac{K_g}{T_g^2 s^2 + 2\zeta T_g s + 1} \tag{5}$$

where K_g, τ_g, T_g and ζ denote a body movement gain, a dead time to begin a pointing action, a parameter determining pointing speed, and a damping coefficient of pointing fluctuation, respectively. H_p is also difficult to formulate explicitly, because p_r is an internal value and cannot be measured well. Furthermore pointer's size, shape, and clarity also affect human visual perception of its location. In this paper, we simply consider H_p to be a first-order lag element

$$H_p(s) = \frac{K_p}{T_p s + 1} \tag{6}$$

that guarantees a step response of the total system converging to a target value. The parameters K_p and T_p have meanings similar to those in H_g. The latency of perception is sufficiently smaller than other latencies and is ignored.

Merge into a total system :

Let the output of H_s be a control value p_s, because an indicator can not know the amount of disturbance and thus tries to change p_s to be closer to p_t. Given the four transfer functions and input signals $p_t(t)$,$d_s(t)$, the control value $p_s(t)$ is described as

$$p_s(s) = p_t(s)G_{st}(s) + d_s(s)G_{sd}(s) \tag{7}$$

where $p_s(s)$, $p_t(s)$, and $d_s(s)$ correspond to frequency domain descriptions of $p_s(t)$, $p_t(t)$, and $d_s(t)$, respectively, with Laplace transform. $G_{st}(s)$ and $G_{sd}(s)$ are described as

$$\begin{aligned} G_{st}(s) &= \frac{H_g(s)H_s(s)}{1+H_g(s)H_s(s)H_v(s)H_p(s)} \\ G_{sd}(s) &= -\frac{H_g(s)H_s(s)H_v(s)H_p(s)}{1+H_g(s)H_s(s)H_v(s)H_p(s)} \end{aligned} \tag{8}$$

based on a closed loop theorem.

3.3 Hybrid Model

In addition to $G_{sd} = -H_p G_{st}$ introduced from Eq. (8), the

frequency characteristic of H_p is determined by only one parameter T_p, G_{sd} and G_{st} are strongly dependent. This results in a considerable trade-off between the approximation performance shown by G_{sd} and G_{st}. An advanced model with more DOF would be necessary to approximate actual pointing behaviors accurately, and it must also follow the proposed control model shown in Fig. 2. Hence, we propose that both G_{sd} and G_{st} consist of the same formulations of H_g, H_s, H_v, and H_p while they are characterized by different parameter sets $\phi^T = [K_g^T, T_g^T, \zeta^T, \tau_g^T, K_p^T, T_p^T, \tau_s, \tau_v]$ for G_{st} and $\phi^D = [K_g^D, T_g^D, \zeta^D, \tau_g^D, K_p^D, T_p^D, \tau_s, \tau_v]$ for G_{sd}. We refer to G_{st} with ϕ^T and G_{sd} with ϕ^D in this hybrid model as a "target value follow-up component" and a "disturbance compensation component", respectively.

4. Experimental Environment

A pointing interface was implemented as shown in Fig. 3 for experimental evaluation of the proposed model. A short focal length projector RICOH PJ WX4141 projects approximately $2.2m \times 1.4m$ visual contents on a screen (white wall in the figure). Participants stand at a distance of approximately $2.5m$ from the screen and are asked to straighten their arms while pointing. The pointed location is estimated as being where the indicating vector connects the reference points in the head and the finger tip to the screen [8]. The screen served as an apparent FOV for the a participant at approximately 47×31 degrees, which is an intermediate size when considering that of the effective visual field of 30×20 degrees and the stable fixation field of 60×45 degrees to 90×70 degrees. Humans can detect, identify, and discriminate visual stimuli in the former FOV with natural eye movements, whereas the latter requires additional head movement. Our configuration allows slight head movement, under the assumption that it does not affect so much to the pointing reference location in the head. With this in mind, the aforementioned pointing model adequately approximates the locations indicated by the participants.

Figure 4 shows the pointing conditions that comprise the pointing start location (SL) and the pointing length (PL) corresponding to the step input size (Tr). For our study, these parameters were configured so as to analyze the model's prediction performance under various experimental conditions. The magnetic field-based 3D pose sensor POLHEMUS Liberty is used to measure pointing postures accurately . Instead of measurement errors expected with a vision-based sensor, simulated error sequences following Eq. (2) are provided. This implementation is intended to acquire $p_s(t)$ and $d_s(t)$, individually. Although it does not perfectly reflect the system behaviors in practical use, it is acceptable for analyzing them. The magnetic sensors are attached on finger tips of the participants' right hands and temples in order to acquire the reference point location as their midpoint. No sensor is attached directly between the participants' eyebrows so as to avoid obstruction of the natural field of vision. A circular pointer with a diameter of approximately 1 cm (assumed as a "point" for the participants) was drawn at the estimated location on the screen.

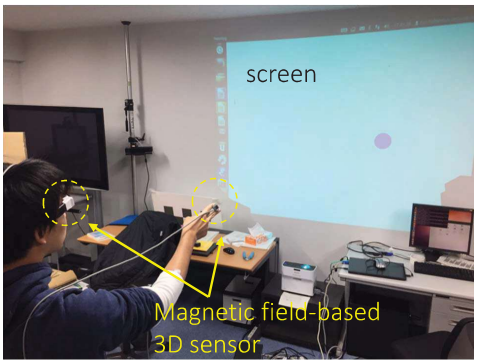

Fig. 3 Experimental environment

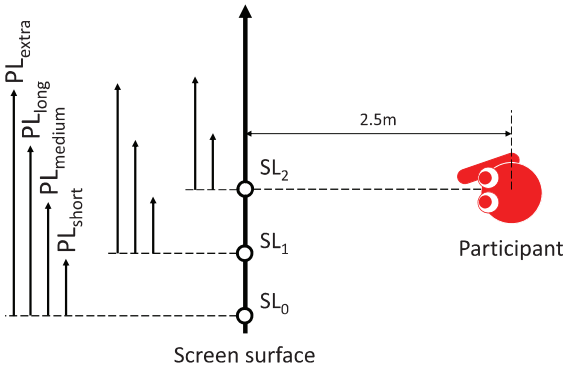

Fig. 4 The experimental conditions for the pointing start location (SL) and the pointing length (PL). SL_2 is located at the front of the participant. SL_1 and SL_0 are 50 and 100 cm left from the participant's side, respectively. The configurations of PL are $PL_{short}, PL_{medium}, PL_{long}, PL_{extra}$ = 40, 80, 120, 160 cm, respectively.

5. Model Calibration

5.1 Methodology

The parameter sets ϕ^T and ϕ^D must be configured to build a mathematical model of a target pointing interface system. Typically, these are determined to explain sample behaviors well during actual pointing. This procedure is referred to as model calibration. The latency parameters τ_s and τ_v are assumed to be inherent and stable features. Thus they are configured as durations for the estimation of the pointed location and the visualization of the pointer, respectively. The remaining parameters are estimated using a non-linear optimization that minimizes residuals about $p_s(t)$, namely between values simulated by the model and the actually measured one.

Here we explain a detailed procedure followed when calibrating G_{st} using response behaviors for step inputs. A similar scheme is applied to calibrating G_{sd} using those for disturbance inputs. We assume that they are independent and do not change even when both inputs are provided simultaneously. We formulate the non-linear optimization for the calibration as minimizing sum of squared errors :

$$\phi^T_{ind} = argmin \sum_t \left(p_s(t) - \mathcal{L}^{-1}\left(\frac{T_r}{s} G_{st}(s, \phi^T)\right)\right)^2. \quad (9)$$

A trust region non-linear optimization method is applied to solve Eq. (9). To avoid convergence to a local minimum, multiple optimizations begin with various initial values to enable at least few of them to reach the global minimum. The optimal parameter set is taken to be the best result of all of them.

To acquire general pointing feature of a target participant, the above optimization is conducted individually for multiple pointing trial $i = 1, 2, \cdots, N$ to acquire $\phi^T_{ind}(i)$. Then, the final parameter set is estimated as their mean values $\phi^T_{fin} = \frac{1}{N}\sum_i \phi^T_{ind}(i)$. A direct optimization of the common parameter set for all trials is simpler than the above two step calibration, but it works as averaging dispersed pointing trajectories in a real domain, though not in a frequency domain. In contrast, two-step calibration averages the system behaviors in both the real and frequency domains and thus appears to maintain essential characteristics.

5.2 Target Value Follow-Up Component

The experimental procedure for providing a step input signal p_t into the participant's intention and to induce its response p_s is configured as follows.

1. The measurement begins when a participant indicates an initial target at SL_1 visualized on the screen and the pointer remains in that location.
2. The initial target suddenly disappears. Simultaneously, a new target pops up at $PL_{medium} = 80$ cm right to the initial target. The participant changes his or her posture to move the pointer onto the new target.
3. The measurement stops when the participant calls the finish of the pointing action.

A sufficiently small noise was expected because of the measurement accuracy of the magnetic sensor and the straight arm condition. Thus we could use Eq. (7) with $d_s = 0$ for calibration. The participants are $M = 5$ university students. Each of them conducted $N = 20$ trials under the same condition.

Although the original parameter set ϕ^T contains six variables $K^T_g, T^T_g, \zeta^T, \tau^T_g, K^T_p$, and T^T_p, its actual DOF is reduced with a constraint for convergence of response trajectories. Without disturbance, a pointing trajectory typically converges to a target location after sufficient time passes. This can be described as the mathematical constraint that the stationary error $e(t) = p_t(t) - p_s(t)$ at $t = \infty$ must be zero under the condition of $p_t(t)$ being a step signal and $d_s = 0$. It is formulated as

$$\frac{p_s(t)}{p_t(t)}|_{t=\infty} = \lim_{s\to 0} s \cdot \frac{1}{s} G_{st}(s) = 1 \quad (10)$$

based on the final-value theorem and results in

$$K^T_g = \frac{1}{1 - K^T_p}. \quad (11)$$

This means that either of the variable K^T_g or K^T_p is sufficient for calibration (with the other being determined automatically). This study selects K^T_p because of its finite range $0 < K^T_p < 1$ while K^T_g has only the lower limit $0 < K^T_g$. The ranges for the other parameters are configured as $0 \le T^T_g \le 1000$ ms, $0 < \zeta^T \le 1.5$, $0 \le \tau^T_g \le 1000$ ms, and $0 \le T^T_p \le 300$ ms based on features of the measured trajectories such as rising times, amount of overshoots, and degrees of convergence. Three values uniformly distributed in each range are used to configure various initial parameter sets with their combinations. Totally non-linear optimizations beginning from $3^5 = 243$ patterns are validated in each trial.

The final results of the optimal parameter sets are shown in Table 1. The estimated feedback gains K_p were quite small and reached to the lower limit. There was almost no feedback effect observed in the participants' pointing behaviors.

Figure 5 shows the measured trajectories and the simulated trajectories on the screen coordinate using the individually calibrated parameter set ϕ^T_{ind} and the average ϕ^T_{fin}, respectively, whereas Fig. 6 shows the mean simulation error in N trials used to analyze error transitions in accordance with the elapsed time. Note the model's accuracy (represented in green curves) as given by individually calibrated models. The absolute errors are within several centimeters totally of each other, and those values observed in the stationary periods are less than 1 cm. The remaining error is thought to be due to indelible waggles due to dynamic nature of the human body. Although the second-lag order element seems to be insufficient, the development of a vastly improved model with a slightly higher DOF will eliminate these errors.

Next, we compared individually calibrated models and the model with their mean values, shown in green and red, respectively. The model's performance becomes worth as expected, because it is well known that human pointing behaviors will have extremely large diversity even under the same conditions. But its influence is out of our expectation. Simulation errors tend to increase, especially during the points noted at the first overshoot and the subsequent counter overshoot. This increase might arise from variations in the pointing speed intentions in a participant's mind. When they place priority on reaching a target location earlier, then large overshoot and damping appear, resulting in errors, whereas accurately moving a pointer can minimize the occurrences of these errors. The mismatch in the point-

Table 1 Optimized parameter set ϕ^T_{fin} for each participant. Each K^T_g is calculated from the corresponding K^T_p based on Eq. (11).

Participant	K^T_g	T^T_g	ζ^T	τ^T_g	K^T_p	T^T_p
#1	1.04	149	0.800	416	0.0384	0.00124
#2	1.08	108	0.873	406	0.0772	0.000956
#3	1.08	123	0.836	340	0.0776	0.000702
#4	1.05	155	0.850	419	0.0512	0.00210
#5	1.06	96.0	0.769	435	0.0587	0.00131

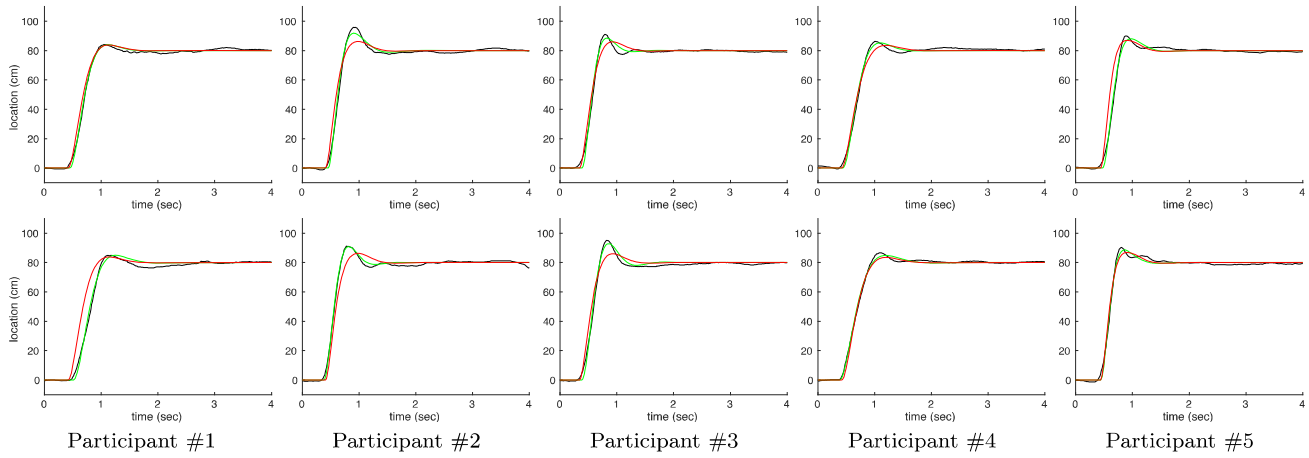

Fig. 5 Reproduction accuracy by G_{st} on pointing trajectory space. The results for the representative two trials of three participants are shown in the figures. The horizontal and vertical axes denote the elapsed time from when the new pointing target appears and the distance from the initial location, respectively. (black) $p_s^{step}(t)$: the measured step response trajectories as ground truth. (green) $\overline{p_{s,ind}^{step}}(t)$: the trajectories simulated with the parameter sets ϕ_{ind}^T optimized for the individual trajectory. (red) $\overline{p_{s,fin}^{step}}(t)$: those with the final parameter sets ϕ_{fin}^T shown in Table 1.

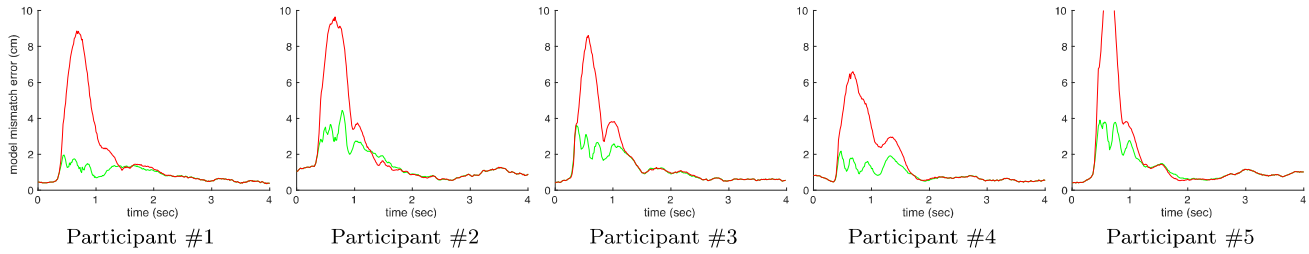

Fig. 6 Reproduction accuracy by G_{st} in difference from ground truth trajectory. The transitions in the figures are average values in N trials for acquiring general feature. (green) $diff_{ind}^{step}(t) = \frac{1}{N}\sum_i |p_s^{step}(t,i) - \overline{p_{s,ind}^{step}}(t,i)|$ for the individually calibrated model. (red) $diff_{fin}^{step}(t) = \frac{1}{N}\sum_i |p_s^{step}(t,i) - \overline{p_{s,fin}^{step}}(t,i)|$ for the model with the averaged parameter.

ing beginning time indicates that there is diversity in the dead time between visual stimulus and body movement. It is quite natural to assume that the time needed to detect the disappearance of the initial target changes over the course of the trials. Variables, which can be used to describe the degree of the pointing speed intention and the dead time experienced, might be helpful in accounting for diversities seen in the participants' pointing behavior.

5.3 Disturbance Compensation Component

Similarly, G_{sd} is calibrated using pointer location stabilizing behaviors against disturbance inputs d_s. The experimental procedure for insert a disturbance d_s into the system and inducing its response p_s is configured as follows. A step input is not provided in order to acquire a pure response to a given disturbance sequence.

1. The measurement begins when a participant indicates a stable target visualized on the screen and the pointer stays at that location.
2. A disturbance sequence begins being added to an estimated location and then the pointer also start vibrating. The participant changes his or her own pointing posture to maintain the pointer on the stable target.
3. The measurement stops after a specified experimental duration if 4.0 seconds.

The participants and the number of trials are the same as those in the previous section. A different disturbance sequence generated based on Eq. (2) with $s_f = 2.0$ cm is used in each trial. In this calibration, we have no reasonable constraint among the parameters, while K_g^D and K_p^D appear in only their product form $K_g^D K_p^D$ in G_{sd}. The actual DOF of the non-linear optimization becomes five. The configured ranges are $0.5 \le K_g^D \le 7.0$, $0 \le T_g^D \le 1000$ ms, $0 < \zeta^D \le 15$, $0 \le \tau_g^D \le 1000$ ms, $0.5 \le K_p^D \le 7.0$, and $0 \le T_p^D \le 300$ ms. The manner of constructing various initial parameter sets for avoiding convergence to a local minimum and to acquire the final result from $N = 20$ trials is the same as those in the calibration of the target value follow-up

component.

The final results of the optimal parameter sets ϕ^D_{fin} are shown in Table 2. As we expected, estimated parameters, especially the feedback gain $K^D_p K^D_g$, were quite different from K^T_p and K^T_g for the target value follow-up component. This result supports our hypothesis that G_{st} and G_{sd} should be characterized using different parameter sets (namely ϕ^T and ϕ^D), respectively, rather than using one common parameter set. The simulated trajectories and their absolute accuracy are shown in Fig. 7, and Fig. 8, respectively. The bouncing signals presented in Fig. 7 should be also evaluated by using the similarity of the trajectory curves. Thus, as shown in Fig. 9, we quantified the similarity as a factor of coherence between the measured and the simulated trajectories. As noted by the simulated results, shown as green curves, the individually estimated models describe the approximated trajectories of actual pointing behaviors of the participants (Fig. 7) within several centimeters (Fig. 8). However, the simulated curves are slightly smoother than the actual pointing behaviors and are without the typical bumps seen in them. The low coherence values at higher

Table 2 Optimized parameter set ϕ^D_{fin} for each participant. K^D_g and K^D_p are shown in their product form.

Participant	$K^D_g K^D_p$	T^D_g	ζ^D	τ^D_g	T^T_p
#1	16.7	427	8.07	166	34.8
#2	24.6	433	8.6	161	42.5
#3	28.0	515	8.85	146	64.8
#4	22.8	518	7.95	165	74.9
#5	24.5	319	8.93	147	16.6

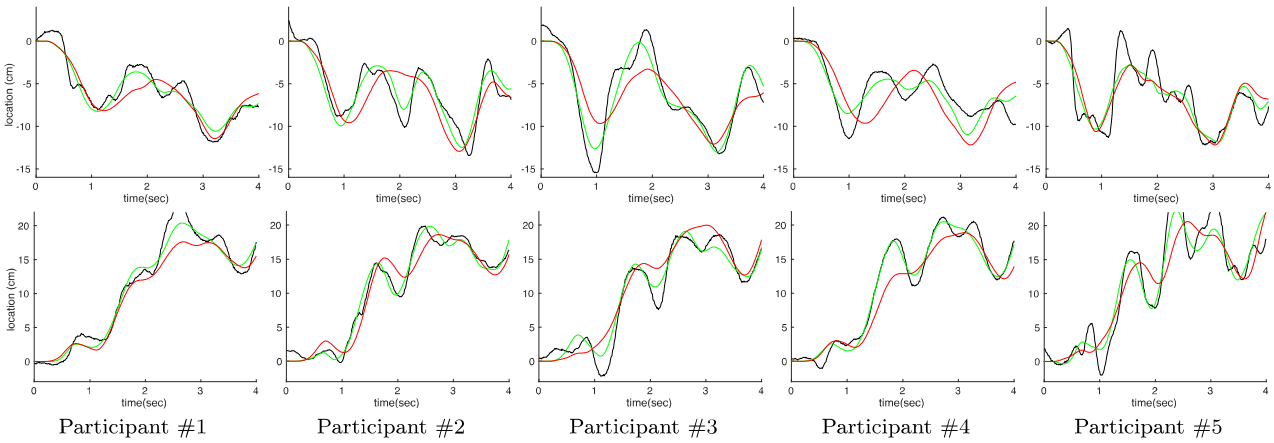

Fig. 7 Reproduction accuracy by G_{sd} on the pointing trajectory space. The results for the representative two trials of three participants are shown in the figures.
(black) $p^{dist}_s(t)$: measured disturbance compensation trajectories as ground truth. (green) $\overline{p^{dist}_{s,ind}}(t)$: trajectories simulated with the parameter sets ϕ^D_{ind}. (red) $\overline{p^{dist}_{s,fin}}(t)$: those with the final parameter sets ϕ^D_{fin}, shown in Table 2.

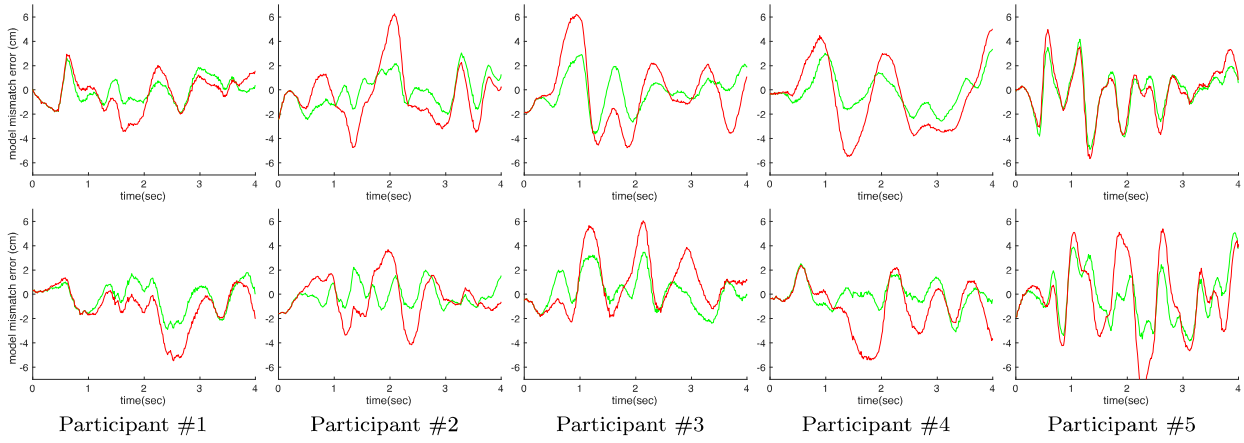

Fig. 8 Reproduction accuracy by G_{sd} in the viewpoint of difference from ground truth trajectory. Each figure corresponds to that at the same cell in Fig. 7. Note that the transitions are not averaged values in N trials to see their feature, individually. (green) $diff^{dist}_{ind}(t) = p^{dist}_s(t) - \overline{p^{dist}_{s,ind}}(t)$ for the individually calibrated model. (red) $diff^{dist}_{fin}(t) = p^{dist}_s(t) - \overline{p^{dist}_{s,fin}}(t)$ for the model with the averaged parameter.

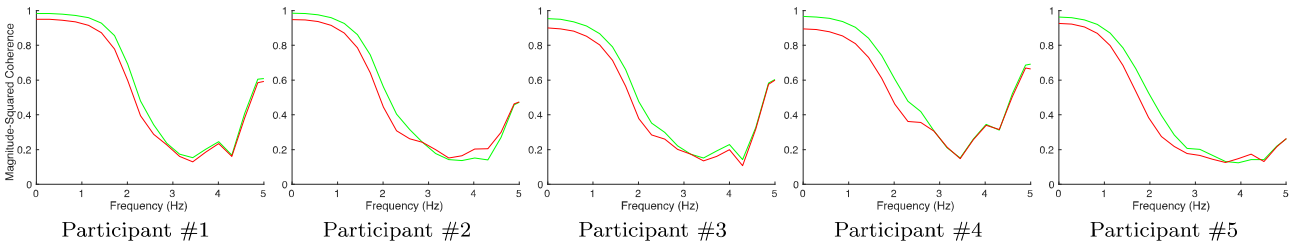

Fig. 9 The reproduction accuracy by G_{sd} in the viewpoint of signal coherence, namely, signal similarities in the frequency domain. The coherence for less than 5 Hz are shown, because we confirmed that the power spectrum on more than 4 Hz are quite small and not dominant in the pointing behaviors. This feature explains well the upper limit of human behavior frequency. (green) the coherence between $p_s^{dist}(t)$ and $\overline{p_{s,ind}^{dist}}(t)$. (red) those between $p_s^{dist}(t)$ and $\overline{p_{s,fin}^{dist}}(t)$. The transitions in the figures are average values in N trials for acquiring general feature.

frequency domain also indicate that compensating behaviors for frequently bumping disturbance are not described well. Possible reasons for this are (1) the participants' body might have been shaking as a result of the frequent movements of their pointing gestures and (2) the sequence of the pointing behaviors comprised multiple compensation movements that may have different response features. Presumably the intended speed of the pointing movement for the compensation are not stable, because disturbances are often unpredictable. Thus there would be considerable diversity and it results in the model mismatch errors.

When the estimated parameters for each pointing trajectory are integrated, the influence of the diversity in a sequence is accumulated. This ultimately decreases the accuracy of final parameter results. A possible way to overcome this mismatch is to optimize the parameters in the frequency domain while our method has applications in the real domain. Since frequency responses are fundamental for compensating for this type of disturbance, the diversity observed in the frequency domain might be small.

6. Model Evaluation

The proposed hybrid model consists of G_{st} and G_{sd} calibrated by using the responses for a single type of input signal, namely step input only or disturbance input only. The purpose of this experiment is to evaluate the model's performance in other situations when both of them are given simultaneously. This corresponds to a typical case of practical pointing situations. The evaluation method is to analyze how accurately the model simulates actual behaviors, comparing simulated p_s using Eq. (7) and measured one. Responses to simultaneous inputs of p_t and d_s were induced and measured in a manner similar to that explained in the calibration section. The participants were also the same as in the calibration. Their own calibrated models were used when simulating their behaviors. To evaluate the model's prediction performance in various situations, each participant conducted seven trials for various combinations of SL and PL as shown in Fig. 4 corresponding to the pointing start location and the size of screen, respectively. These conditions also induce standing position relative to the screen. A different disturbance sequence, of course different from those used for the calibration, was used in each trial while the same disturbances in the same order were used for each participant.

The non-hybrid basic model in which a common parameter set used for both of G_{st} and G_{sd} is assumed as a comparison reference. To calibrate it under conditions similar to those of the proposed hybrid model, the input signals and the measured trajectory data corresponding to SL_1 and PL_{medium} conditions are used. One considerable issue is data segmentation. When evaluating the k-th trial in each target value category, the remain six trials with $l \neq k$ in the SL_1 and PL_{medium} condition are used for the calibration. This method is a kind of cross validation. The parameter ranges are the same as those configured in Sect. 5.3. The manner of constructing various initial parameter sets and acquiring the final calibration result are also same as those in the calibration of the proposed model except for the number of trials.

The simulated and the actual pointing trajectories of the all participants are shown in Fig. 10. In most cases, the trajectories simulated by the proposed hybrid model show more similarities with the actual trajectories than those simulated by the reference model. That is remarkable considering the disturbance compensation durations, which were seen after the pointer reached to the target. An improvement in the overall performance can be also found in the mean values of all trials, for example the small absolute errors after an elapsed time of 1 s and the larger coherence at a frequency higher than $2Hz$ as shown in Fig. 11. These results support our hypothesis that the proposed hybrid model is a much better way to approximate human pointing behaviors despite the presence of considerable disturbance.

With relation to transient duration, the prediction performance of the two methods was almost even. When analyzing this result, we found an important report by Woodworth et al. [15]. This report proposes that switching from the target value follow-up phase to the disturbance compensation phase is better, whereas our proposed model assumes significant responses to the disturbance even in the transient durations and simply superposes these two phases. Applying the switching or a weighted sum formulation is a possible way to solve this problem. Comparisons of the coher-

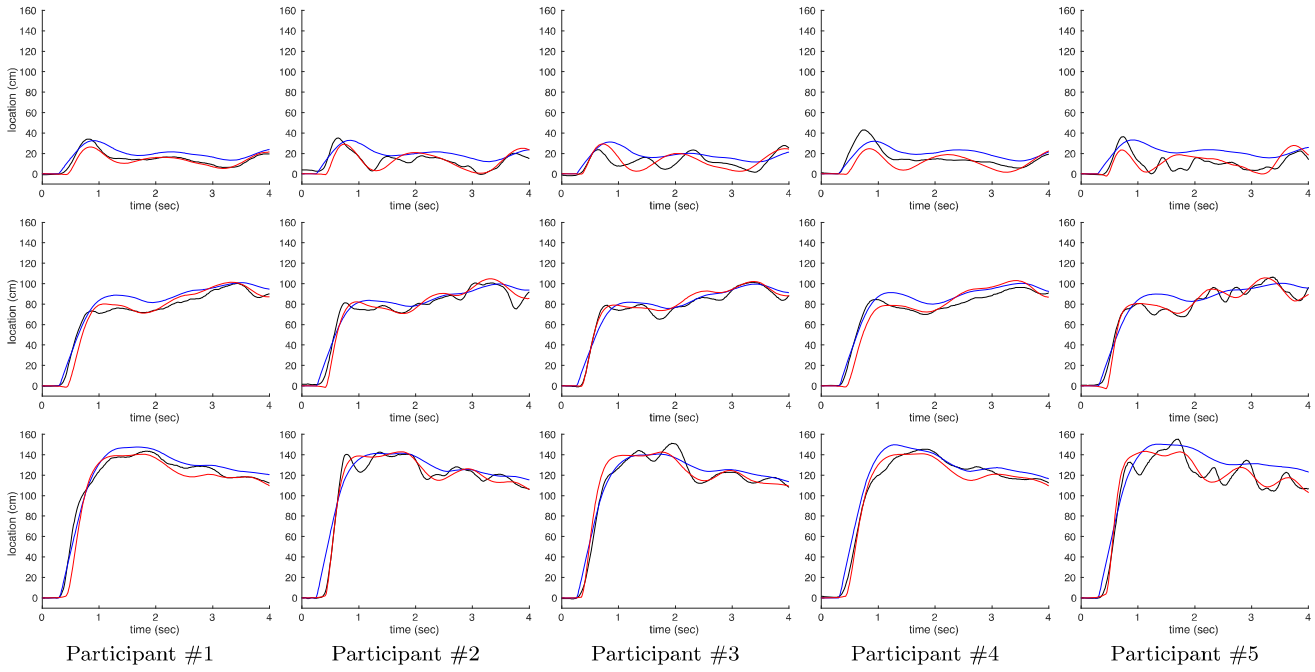

Fig. 10 The performance comparison for the general situations when step signals and disturbance sequences are provided simultaneously. The figures in each row correspond to representative trials under the pointing lengths $PL_{short}, PL_{medium}, PL_{long}$ with the initial target SL_1. (black) $p_s^{gen}(t)$: actual pointing behavior trajectories. (blue) $\overline{p_{s,normal}^{gen}}(t)$: trajectories simulated by the non-hybrid model using a common parameter set for both of G_{st} and G_{sd}. (red) $\overline{p_{s,hybrid}^{gen}}(t)$: those by the proposed hybrid model using ϕ_{fin}^{T} for G_{st} and ϕ_{fin}^{D} for G_{sd} shown in Table 1 and Table 2, respectively.

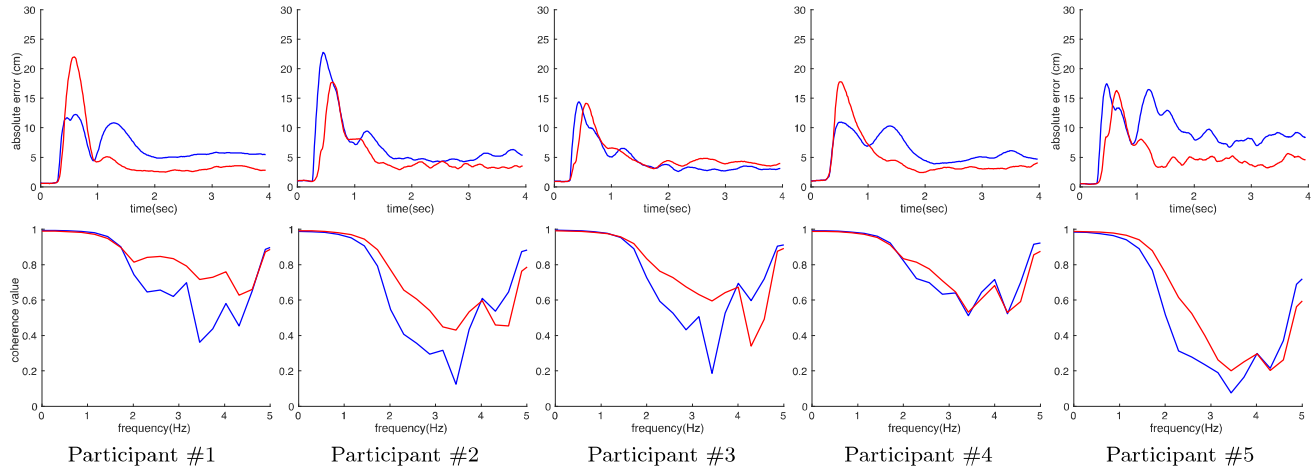

Fig. 11 The statistic evaluation of the model's prediction performance as mean values with related to all trials of each participant. (top-row) the absolute error. (bottom-row) the coherence value. (red) performance of the proposed hybrid model. (blue) that of the reference non-hybrid model.

ence values within the meaningful frequency range of less than 4 Hz indicate that the proposed model is advantageous in all cases with the exception of participant #4. As shown in Fig. 10, his pointing behavior tends to be smoother than those of other participants, even under the same experimental conditions. It indicates that his response to the high frequency disturbance becomes small when a pointing target at a distance and disturbance are given simultaneously. Thus improvements made by the proposed hybrid model in predicting their influence are not comparatively significant.

Herein, we focus on the relationship between the pointing start location SL and the model's prediction performance. Their comparison in terms of the absolute error values and the coherence values are shown in Fig. 12. The absolute errors observed after convergence do not change much for the various pointing start locations, whereas those

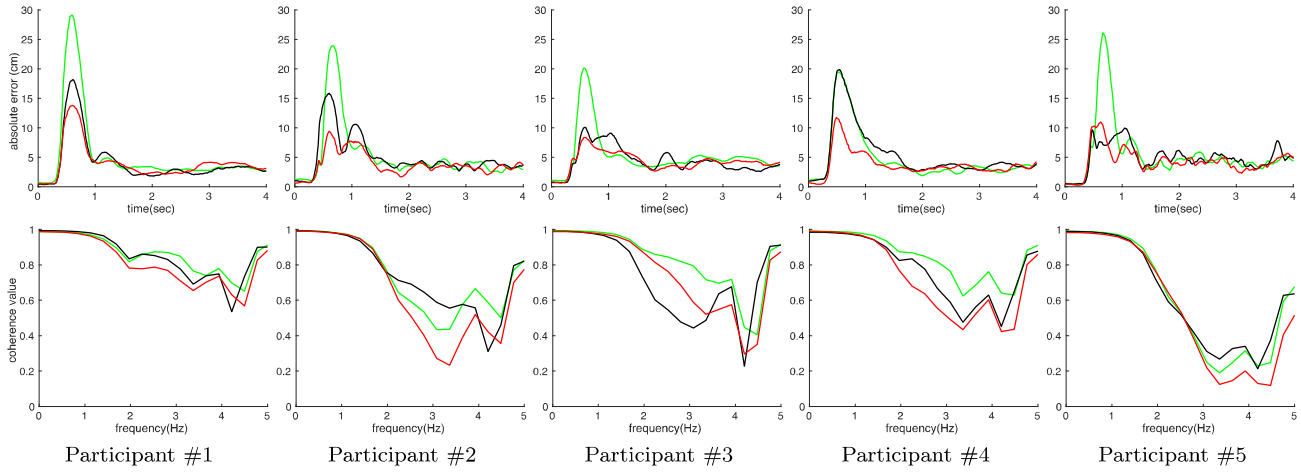

Fig. 12 The comparison of the model's prediction performance at various pointing start positions. The green, black, and red curves correspond to SL_0, SL_1 and SL_2 conditions, respectively.

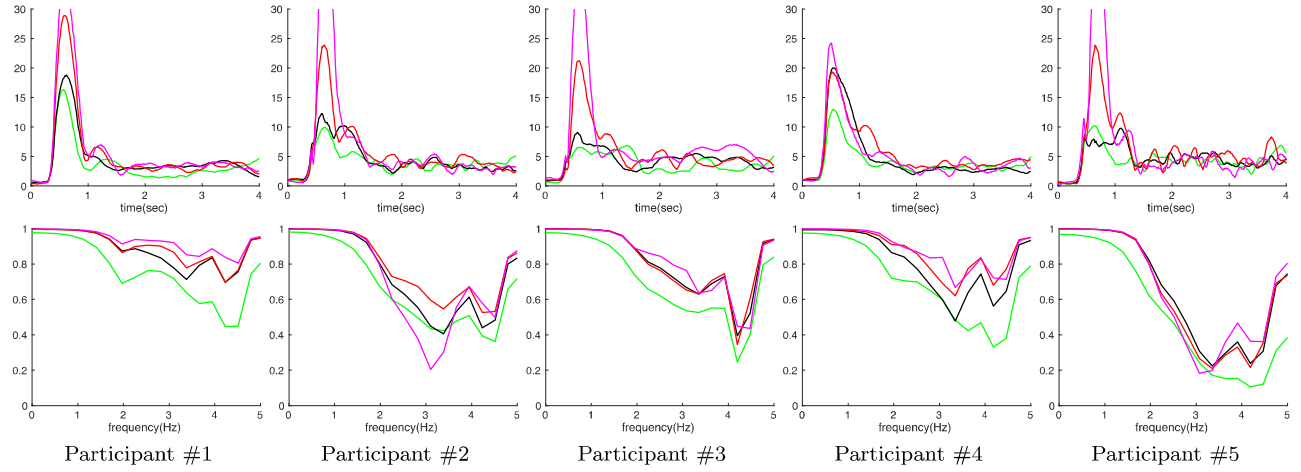

Fig. 13 The comparison of the model's prediction performance at various pointing lengths. The green, black, red, and magenta curves correspond to PL_{short}, PL_{medium}, PL_{long} and PL_{extra} conditions, respectively.

in the transient durations tend to increase in the order of SL_2, SL_1, and SL_0, i.e., with a change in the distance from the front of the participants. The pointing behaviors that start from diagonal locations become unstable and varied when compared to those initiated from directly in front of the participants. These results are accumulated as part of the absolute error calculations. Comparison of the coherence values indicates that SL does not affect to the model's prediction performance in low frequency domains of approximately less than 2 Hz, whereas its influence in higher frequency domains strongly depends on the individuality of the participants. Interestingly, the fluctuating trend observed with the increase in frequency is common for all the SL conditions and is worth further investigation.

The relationship between the PL and the model's prediction performance are shown in Fig. 13. Comparisons of stationary periods reveal that there are similarities in the two aforementioned cases. The absolute errors seen during transient duration tend to become large in relation to the PL. This is not surprising, considering that the model's prediction values rise when there is a large increase in the input signal, which ultimately corresponds to a longer PL. However, the value of the prediction errors does not match the amplitude of PLs. Human field of vision is a probable factor for the reason of the mismatch. When participants could see both the initial and the destination target locations without the need for additional head movements, their gestures tended to be more stable, resulting in smaller errors for PL_{short} and PL_{medium}. In contrast, when the pointing lengths were longer, such as in the case of PL_{long} and PL_{extra}, the participants were required to rotate their heads to recognize the destination target locations. Doing so induces diversity in the participants' pointing behaviors and results in numerous errors. When focusing on the coherence comparison, the PL_{short} condition gives the worst prediction performance contrary to the absolute error aspect. Presumably, this also arises from the issue of how the target value follow-up and the disturbance compensation components should be inte-

grated. For PL_{short}, the amplitude of disturbances compared to the pointing length is considerable even in the transient periods. Therefore disturbance compensation behavior was dominant and reaching behavior to the target location did not appear, which does not match the assumption of the proposed hybrid model that receives the same amount of the effect from both components every time.

7. Conclusion

This study proposed a feedback control model of a pointing interface system for constructing a more comfortable pointing environment. The model is formulated as a superposition of a target value follow-up component and a disturbance compensation one induced from the same feedback loop but with different parameter sets to describe human pointing features well. The two optimal parameter sets were determined individually to reproduce actual pointing behaviors accurately for step input signals and random walk disturbance sequences. The calibrated models can be used to simulate pointing behaviors for arbitrary input signals expected in practical pointing situations. The evaluation results indicated that absolute errors from actual trajectories at the beginning of the pointing often exceed 10 centimeters while those after the convergence are within several centimeters. Through the signal coherence analysis, the trajectories simulated by the proposed hybrid model had higher similarity than those by the non-hybrid model in the frequency domain of approximately more than 2 Hz.

Most of the model prediction errors could be justified using the following arguments: (1) the proposed hybrid model formulated a simple superposition of the target value follow-up and the disturbance compensation components and, (2) diversity in the participants' pointing behaviors, even those which occur under the same conditions, should not be ignored. A possible solution for the former issue is to expand the superposition to include a weighted sum in which the weights would change based on the elapsed time and the distance to the target location. The latter issue requires further investigation in an additional model parameter that would describes pointing behavior diversity and build a probabilistic prediction model.

In this study, the pointing posture, the pointing direction from the initial location to the destination, the standing distance from the screen, and the body direction in relation to the screen were all fixed. The performance under various configurations should be evaluated, including when the pointing gesture began from the position of the arms hanging limply at the user's sides. The current proposed model assumes a simple pointing interface that draw a small circular pointer. This must be expanded to advanced pointer visualizations for interface improvement. Now we assume a larger burred pointer at smoothed location to indicate not a point but a region with less effort.

References

[1] R. Kopper, D.A. Bowman, M.G. Silva, and R.P. McMahan, "A human motor behavior model for distal pointing tasks," Int. J. Human-Computer Studies, vol.68, no.10, pp.603–615, 2010.
[2] A. Cockburn, P. Quinn, C. Gutwin, G. Ramos, and J. Looser, "Air pointing: Design and evaluation of spatial target acquisition with and without visual feedback," Int. J. Human-Computer Studies, vol.69, no.6, pp.401–414, 2011.
[3] M. Nancel, E. Pietriga, O. Chapuis, and M. Beaudouin-Lafon, "Mid-air pointing on ultra-walls," ACM Trans. Computer-Human Interaction, vol.22, no.5, pp.1–62, 2015.
[4] D. Vogel and R. Balakrishnan, "Distant freehand pointing and clicking on very large, high resolution displays," Proc. 18th Annual ACM Symposium on User Interface Software and Technology, pp.33–42, 2005.
[5] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," Proc. 2011 IEEE Conference on Comput. Vis. Pattern Recognit., CVPR '11, Washington, DC, USA, pp.1297–1304, IEEE Computer Society, 2011.
[6] M. Yoshimoto and Y. Nakamura, "Cooperative gesture recognition: Learning characteristics of classifiers and navigating user to ideal situation," Proc. 4th IEEE International Conference on Pattern Recognit. Applications and Methods, pp.210–218, 2015.
[7] K. Nickel and R. Stiefelhagen, "Pointing gesture recognition based on 3d-tracking of face, hands and head orientation," Proc. IEEE Int. Conf. Multimodel Interfaces, pp.140–146, 2003.
[8] M. Fukumoto, Y. Suenaga, and K. Mase, "Finger-Pointer": Pointing interface by image processing," Computers & Graphics, vol.18, no.5, pp.633–642, Sept. 1994.
[9] D.Y.P. Henriques and J.D. Crawford, "Role of eye, head, and shoulder geometry in the planning of accurate arm movements," J. Neurophysiology, vol.87, no.4, pp.1677–1685, 2002.
[10] K. Kondo, G. Mizuno, and Y. Nakamura, "Analysis of human pointing behavior in vision-based pointing interface system: Difference of two typical pointing styles," Proc. 13th IFAC/IFIP/IFORS/IEA Symposium on Analysis, Design, and Evaluation of Human-Machine Systems, HMS 2016, vol.49, pp.367–372, IFAC-PapersOnLine, 2016.
[11] S. Ueno, S. Naito, and T. Chen, "An efficient method for human pointing estimation for robot interaction," Proc. IEEE Int. Conf. Image Processing, pp.1545–1549, 2014.
[12] R. Jota, M.A. Nacenta, J.A. Jorge, S. Carpendale, and S. Greenberg, "A comparison of ray pointing techniques for very large displays," Proc. Graphics Interface, pp.269–276, 2010.
[13] P.M. Fitts, "The information capacity of the human motor system in controlling the amplitude of movement," J. Experimental Psychology, vol.47, no.6, pp.381–391, 1954.
[14] K. Kondo, Y. Nakamura, K. Yasuzawa, H. Yoshimoto, and T. Koizumi, "Human pointing modeling for improving visual pointing system design," Proc. Int. Symp. Socially and Technically Symbiotic Systems, pp.393–400, 2015.
[15] R.S. Woodworth, "The accuracy of voluntary movement," Phychological Review Monograph Sullplement, vol.3, no.13, pp.1–119, 1899.

Kazuaki Kondo He received B.E. and M.E. degrees in Engineering Science, Ph.D. degree in Information Science and Technology, from Osaka university, in 2002, 2004, and 2007, respectively. He was a research associate in the Institute of Scientific and Industrial Research in Osaka University, from 2007 to 2009. He has been an senior lecturer at Kyoto University, Japan, from 2015. He has been working on the research of computer vision, human-computer interaction, and media computing. Current his major work is to analyze human behaviors in various daily environments for designing human-friendly interfaces or supports, such as pointing, manipulating equipment's, and doing several tasks simultaneously.

Genki Mizuno He received B.E degree in electrical engineering from Kyoto University, in 2015. He is now a graduate student in master's course of Kyoto University. He has been working on analysis of pointing behavior.

Yuichi Nakamura He received B.E, M.E, and Ph.D degrees in electrical engineering from Kyoto University, in 1985, 1987, and 1992, respectively. From 1990 to 1993, he worked as an instructor at the Department of Electrical Engineering of Kyoto University. From 1993 to 2004, he worked for Institute of Information Sciences and Electronics of University of Tsukuba, Institute of Engineering Mechanics and Systems of University of Tsukuba, as an assistant professor and an associate professor, respectively. Since 2004, he has been a professor of Academic Center of Computing and Media Studies, Kyoto University. His research interests are on computer vision, multimedia, human-computer and human-human interaction including distance communication, and multimedia contents production.