Liao, Yi-Chi; Mo, George B.; Dudley, John J.; Cheng, Chun Lien; Chan, Liwei; Kristensson, Per Ola; Oulasvirta, Antti

Practical approaches to group-level multi-objective Bayesian optimization in interaction technique design

Published in:
Collective Intelligence

DOI:
10.1177/26339137241241313

Published: 01/01/2024

Document Version
Publisher's PDF, also known as Version of record

Research Article

COLLECTIVE INTELLIGENCE

Collective Intelligence
Volume 3:1: 1–19

DOI: 10.1177/26339137241241313
journals.sagepub.com/home/col
Sage

Practical approaches to group-level multi-objective Bayesian optimization in interaction technique design

Yi-Chi Liao[1,*], George B Mo[2,*], John J Dudley[2], Chun-Lien Cheng[3], Liwei Chan[3], Per Ola Kristensson[2] and Antti Oulasvirta[1]

[1]Department of Information and Communications Engineering, Aalto University, Aalto, Finland

[2]Department of Engineering, University of Cambridge, Cambridge, UK

[3]Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu, Taiwan

Abstract

Designing interaction techniques for end-users often involves exploring vast design spaces while balancing many objectives. Bayesian optimization offers a principled human-in-the-loop method for selecting designs for evaluation to efficiently explore such design spaces. To date, the application of Bayesian optimization in a human-in-the-loop setting has largely been restricted to optimization, or *customization*, of interaction techniques for individual user needs. In practice, interaction techniques are typically designed for a target population or group of users, with the goal is to produce a design that works *well* for *most* users. To accommodate this common use case in interaction technique design, we introduce two practical approaches that facilitate multi-objective Bayesian optimization at the group level. Specifically, our approaches streamline the process of (1) deriving designs suitable for a group of users from data collected in individual user evaluations; and (2) deriving an initialization from group data to improve the efficiency of design optimization for new users. We demonstrate the advantages of these practical approaches in two multi-phase user studies involving the design of non-trivial interaction techniques.

Keywords

Human-in-the-loop optimization, interaction technique, interface design, optimization, design optimization, Bayesian optimization, pointing, haptics, input, touch

Introduction

Developing an interaction technique is hard. Technically, it involves setting the values of various design parameters that eventually shape the performance and experience of users. These configurable attributes implicitly define a multi-dimensional design space with theoretically infinite feasible operating points. A simple design task with three configurable design parameters, each with ten possible levels, already has 1,000 feasible operating points and most practical tasks face even larger design spaces. Unfortunately, the relationship between particular design

*The authors contribute equally to this paper.

Corresponding author:
Yi-Chi Liao, Department of Information and Communications Engineering, School of Electrical Engineering, Aalto University, Otakaari 1B, Espoo 02150, Finland.
Email: yichi.mdp@gmail.com

choices and their outcomes for users is complex and rarely predictable. Even a small change in a design parameter that improves one aspect of an interaction technique can have unwanted side effects on the others. Therefore, selecting the "best," or even a "good enough," operating point poses a central and non-trivial challenge to the design of interaction techniques.

Complicating matters further is the fact that seeking to design for more than a single objective, that is, a measurable performance metric, exposes the challenge of *Pareto-optimality*. Pareto-optimality refers to the idea that given more than one design objective, there is no longer a single best operating point. For example, a particular interaction configuration may yield good task performance but receive mediocre subjective user ratings, or vice versa. For a given Pareto-optimal design, no individual objective can be improved by adjusting the design parameters without making at least one other individual objective worse off. For instance, a well-known trade-off in interaction technique design is balancing speed and accuracy. A design instance that favors speed may lead to higher error rates, and vice versa. Such trade-offs are difficult to navigate without some degree of subjectivity and/or introducing secondary constraints.

The conventional way of tackling this challenge has been via empirical evaluation of manually selected operating points: a promising set of combinations is chosen and compared in an experiment (Hornbæk, 2013). However, because of the significant time and effort involved, a truly exhaustive study is rarely conducted, and the designer must often select a design by other means, for instance, by basing a decision on previous results (Chen et al., 2011; Gergle and Tan, 2014). This approach is not only prone to missing good designs, but potentially biased based on prior experience and personal preferences. Recently, human-in-the-loop optimization has emerged as a more systematic and unbiased design exploration process. Bayesian optimization is a particularly strong candidate for human-in-the-loop optimization given that it achieves high sample-efficiency by leveraging an iteratively refined model of the design space (Shahriari et al., 2016). Underlying this is a probabilistic surrogate model, such as a Gaussian process (GP), which offers a robust probabilistic estimate of the latent function relating the design parameters to measurable objectives. As more observations are collected, the quality of the GP's estimate is improved, which in turn enables the optimizer to make informed choices about where to sample next. Recently, researchers have investigated the advantages and disadvantages of using *multi-objective Bayesian optimization* (MOBO) in assisting design exploration (Chan et al., 2022; Liao et al., 2023). These first investigations into how designers can be assisted by MOBO reveal potential benefits in terms of the quality of designs identified, while delivering these with reduced designer workload.

Prior work has applied Bayesian optimization in the context of human-in-the-loop interface and interaction design to determine: game mechanics that maximize user engagement (Khajah et al., 2016); font features that maximize user reading speed (Kadner et al., 2021); interface features that minimize interface search time (Dudley et al., 2019); interaction settings that minimize selection time and maximize accuracy (Chan et al., 2022); and animation and image adjustments to efficiently match some desired appearance (Brochu et al., 2010; Koyama et al., 2017, 2020).

These prior applications of Bayesian optimization are either limited to a single objective and/or focus on customization of the design to the individual. In practice, the design intent for interaction techniques is often to address the needs of a group or population of users, rather than an individual. Accommodating multiple objectives further complicates such an optimization process given the concept of Pareto-optimality and the absence of a single best operating point. In this current work, we seek to bridge the gap between the demonstrable efficiency of Bayesian optimization for interaction technique design, and the practical need to focus on group requirements in contrast to individual requirements.

In this paper, we introduce two practical approaches that streamline the process of working with group data for MOBO in interaction technique design. Both approaches leverage MOBO to independently identify optimal designs for a group of individuals and subsequently offer methods for aggregating the data from these multiple individuals to assemble an *aggregated performance model* representative of the group. When there exist parameter values, or ranges of such values, that generally lead to high objective values for many users, the *aggregate model* captures these shared traits. Since they are multi-objective models they can be inspected to identify the Pareto front, which can inform trade-off decisions. We refer to these two approaches as: (i) the *Global GP*, which computes group-level optimal designs from data obtained from individual users, and (ii) the *Warm-Start GP*, which provides an initialization for the MOBO process using group-level data to support more rapid individual design optimization. The Global GP helps designers find the optimal designs for groups and populations, while the Warm-Start GP helps designers adapt these designs faster to an individual.

User-centered design distinguishes between user research, requirements, design, and evaluation. Bayesian optimization, as studied in this paper, focuses on the later stages of design, where a reasonable understanding of the design problem has been achieved, but exact decisions regarding the design have not been made. The proposed practical approaches can assist designers when they have identified which aspects of the design that are likely determinants of quality or user performance, but they do not yet know how specific design choices influence the

objectives of interest. Rather than manually searching in order to find optimal design settings or assess the objective trade-offs, the designers can instead leverage our proposed approaches.

We evaluate each capability in challenging and realistic interaction design tasks across two multi-phase user studies: (i) designing a 3D touch interaction in virtual reality, exposing a complex trade-off between selection speed and accuracy, and (ii) tuning the vibration feedback for a touchscreen button to achieve an effective compromise between temporal accuracy, consistency, and user comfort. The results reveal that the Global GP leads to significantly improved user performance. Completion time decreased by 5.5% and spatial error decreased by 48%. The Warm-Start GP also led to a significantly larger hypervolume, which is a proxy measure of the quality of a multi-objective optimization outcome. The hypervolume increased 38% and 18%, respectively, compared to a standard MOBO for two different groups of users, indicating its efficacy for faster design adaptation.

This work contributes to the study of collective intelligence by shedding light on the fundamental question of how to design interactions for the abilities and needs of groups. Inherent to our approach is that we model the diversity of responses in a user population, which is important in design as it allows consideration of difficult trade-offs. This is a problem akin to previous research that navigates trade-offs between the preferences of different stakeholders (Bose et al., 2017; Lee et al., 2019). At the same time, we demonstrate how a model that captures this group-level variation can also be exploited to adapt designs to individuals. In summary, this paper makes three key contributions:

1. We introduce the *Global GP* as a practical method for generating Pareto-optimal design instances based on user-specific optimizations performed by a group of users.
2. We introduce the *Warm-Start GP* as a practical method for deriving initializations from group-level data in order to facilitate more rapid optimization for a new user.
3. We demonstrate both the *Global GP* and the *Warm-Start GP* in two representative and challenging interaction design tasks. This provides a valuable reference on how to apply group-level multi-objective optimization more broadly to HCI design problems.

Related work

Computational methods for assisting designers have attracted substantial recent attention. In this section, we review the literature introducing and demonstrating computational approaches to interaction technique design.

Computational one-shot design

Computational one-shot design relies on prior knowledge of the function relating design choices to performance objectives. For example, one can construct a model describing the impact reducing the size or separation between buttons has on selection time. Discrete and continuous optimization methods have been extensively explored for one-shot design of user interfaces (Oulasvirta et al., 2020). These approaches have been successfully applied in a variety of HCI applications such as widget layouts (Gajos and Weld, 2004), keyboards (Karrenbauer and Oulasvirta, 2014), and context-aware interfaces in AR (Lindlbauer et al., 2019). These approaches generally assume a *predictive model* as given or learned from a pre-existing dataset. This established model can then be queried to guide the search over the design space. Often, however, it is not possible to derive or acquire a predictive model relevant to a novel design problem and so such approaches do not generalize well.

Bayesian approaches for single-objective problems

Bayesian optimization is a machine learning method that performs efficient exploration of complex or black-box objective functions to identify an optimum point. Bayesian optimization eliminates the need for an established predictive model or prior exploration of the design space. The approach is well suited to applications where the objective functions are expensive or difficult to evaluate due to the required time or effort. Bayesian optimization, therefore, has good utility in supporting interface and interaction design since many objectives can only be evaluated by conducting a test with a user. Prior work outside of HCI has also shown that Bayesian optimization can outperform other black-box optimization methods for one-dimensional design problems (Borji and Itti, 2013). Bayesian optimization has been applied to a wide variety of optimization problems, but we subsequently constrain the scope of our review to design tasks involving a user. For a more general overview, please see Shahriari et al. (2016).

Khajah et al. (2016) applied Bayesian optimization to assign parameter values dictating the mechanics of a game in order to maximize user engagement. Also using Bayesian optimization, Kadner et al. (2021) sought to customize font designs for individuals to maximize reading speed. Dudley et al. (2019) also leveraged crowdsourcing to quickly access a large number of users but evaluated a more traditional metric based on task completion time to refine design parameters for a range of simple user interfaces. Users were also given the opportunity to subjectively rate interface designs but this was not integrated into the optimization

process. Brochu et al. (2010) demonstrated a technique for allowing designers to quickly determine appropriate values for smoke animation while Koyama et al. (2017, 2020) sought to streamline user editing of photographs to achieve a desired visual appearance. The ability to tightly integrate the user into the procedure makes Bayesian optimization well suited to customizing settings to an individual. This capability has been exploited to tune hearing devices (Nielsen et al., 2015) and other assistive technologies (Snoek, 2013). Piovarci et al. (2020) used an approach influenced by Bayesian optimization to explicitly search for design parameters in surfaces and styli for drawing haptics that exhibit target friction and vibration objective values. However, the goal of Piovarci et al. (2020) was to obtain parameter values that yield predetermined friction and vibration qualities and so they do not strictly perform multi-objective optimization over the design space. Chan et al. (2022) reported the results of a between-subjects experiment that investigated the advantages and disadvantages of using Bayesian optimization in interaction technique design versus manual design space exploration. The core focus of Chan et al. (2022) was investigating when designers generate design candidates and evaluate these designs on themselves, as opposed to with a group of end users.

Multi-objective Bayesian optimization

Many methods have been proposed for multi-objective optimization of black-box objectives from evolutionary approaches (Knowles, 2006) to Bayesian approaches (Hernandez-Lobato et al., 2016; Picheny, 2015; Zuluaga et al., 2016). Multi-objective optimization has been employed in many engineering problems, such as in user interface design. It has been used to optimize linkages for haptic interfaces (Hayward et al., 1994), mid-air text entry (Sridhar et al., 2015), and keyboard layout optimization (Dunlop and Levine, 2012). These methods either relied on reducing the multiple objectives into a single objective by a linearized weighted sum (Sridhar et al., 2015) or by variations on grid search or trial-and-error (Dunlop and Levine, 2012; Hayward et al., 1994). Feit et al. (2015) provided an extensive overview of the challenges and methods available in applying multi-objective optimization to keyboard design. However, none of those methods use Bayesian optimization to converge to the Pareto-optimal parameters, without applying heuristics, by optimizing over black-box objective functions.

Most of these methods assumed that the multiple objectives are independent; however, Shah and Ghahramani (2016) described an approach that incorporates the correlation between expensive objectives. In this paper, we build on the formulation introduced by Shah and Ghahramani (2016) and apply it to interaction design for groups.

Practical approaches to MOBO: The Global GP and the Warm-Start GP

When optimizing a single objective, it is possible to determine a single "best" design.[1] By contrast, when performing multi-objective optimization, there is no single optimum but rather the set of operating points that represent optimal trade-offs between the design objectives. For example, consider three hypothetical designs A, B, and C: A delivers high speed but poor accuracy; B delivers high accuracy but poor speed; and C delivers moderate accuracy and moderate speed. All of these designs may be considered optimal if their performance in one of the two objectives cannot be improved without making the other objective worse. The set of operating points for which one objective cannot be increased without the other objectives decreasing is referred to as the set of Pareto-optimal designs or the Pareto front.

In Bayesian optimization, the goal is to optimize over black-box objective functions by sequentially choosing new test points at which to evaluate those objectives. As new samples are collected, a surrogate model relating the design parameters to their approximate objective function values is updated. It is typical to use a Gaussian process regression (GP) as this surrogate model. A GP is a non-parametric method that models functions, giving uncertainty estimates about function values and often allowing analytically tractable inference. An acquisition function is consulted to determine which point should be sampled next. Acquisition functions propose sampling points by trading off exploration (sampling where the inference uncertainty is high) and exploitation (sampling where the surrogate models predict high objective values). For multi-objective Bayesian optimization, the acquisition assessment is typically performed with reference to the Pareto hypervolume (Zitzler and Thiele, 1999). The Pareto hypervolume is the volume bounded by a fixed reference point on one side and the multidimensional Pareto front on the other side. The acquisition function effectively seeks to sample a new design point that will increase the Pareto hypervolume as this corresponds to a new point that advances the Pareto front. We leverage the acquisition function called CEIPV (Correlated Expected Improvement in Pareto hyperVolume), proposed by Shah and Ghahramani (2016).

Previous works have applied MOBO to interface design problems, but they have focused solely on optimizing for individual users (Chan et al., 2022; Liao et al., 2023). In this paper, we demonstrate two methods that facilitate the application and broaden the utility of MOBO by addressing two unique requirements commonly encountered in

interaction design problems. The first requirement is the common goal in interaction design to arrive at a configuration that performs *well* for *most* users. This reflects the need to release products that are acceptable to a broad user base. The second requirement is the goal of supporting efficient customization methods from some broadly acceptable initial configuration. This reflects a desirable quality for an interaction technique to perform generally well from the outset but to also quickly adapt to individual user needs. We refer to the two proposed methods as the *Global GP* and the *Warm-Start GP* and introduce each below.

Global GP: Concise user group modeling

Running the standard MOBO procedure produces a distinct set of Pareto-optimal designs for each individual user. Each Pareto front obtained potentially reflects the unique preferences and abilities of these individual users. In interaction design, we typically want to find a configuration that is suitable for a broad user base rather than a design that works well for one person but poorly for the majority. Therefore, we wish to combine the Pareto-optimal designs obtained for all users sampled individually into a single "global" model that reflects the broader preferences and abilities of the user group. Here we use the term "global" to refer to the user group as distinct from individual users.

To fulfill this task, we construct a *Global GP* that incorporates all the data from all users. A GP is a probability distribution over possible functions and estimates the model relating input values, x, to function values, y. In the context of interaction design, x refers to design parameter values while y is the measured performance. Since a GP captures the probability distribution over all possible functions, one can derive the means of the functions and the variances to indicate the confidence of the predictions. The Global GP is constructed by providing all observed pairings of design parameters and objectives from all participants to form a single GP model. We can then inspect the Global GP to obtain the estimated mean and variance of y (i.e., performance objectives) at any x (i.e., design parameter settings) and use this to predict the expected performance of the group given a particular design instance. To obtain the Pareto-optimal designs from this Global GP, we conduct a fine-grid sampling of the design parameter space. Given appropriate bounds for each parameter and normalization, the design parameter space is a hypercube in $\mathbb{R}^n$ and so we do an exhaustive fine-grid sampling on that hypercube with the resolution specifying the coarseness, c, of the sampling. Given c, we divide each dimension of the parameter space $\mathcal{X}$ into c equally spaced grid points, 0 and 1 inclusive. After sampling, we have c^n samples, which we use to output the set of global Pareto-optimal designs. We set $c = 16$ for our applications which was determined by the empirical trade-off between design parameter specificity and computation time.

This method is computationally expensive but since it runs as a post-processing step, it is practical and feasible to perform and provides a comprehensive summary of the optimal design parameter sets. Figure 1 illustrates the high-level process of constructing the Global GP and extracting the Pareto-optimal designs.

Besides using GPs as base models, other approaches, such as regression models and deep neural nets, may be applied to constructing a global-level model. However, these methods have various requirements and limitations. A deep neural net requires a large amount of data and high-dimensional input. That typically demands a large number of users, which is not usually feasible for human-in-the-loop optimization. Parametric regressions, such as linear regressions, require assumptions of the model (e.g., the number of degrees and the landscape of input and output), which are also not always viable for HCI problems. Furthermore, deep neural nets and parametric regression models are not capable of capturing the uncertainty of the predictions. In comparison, a non-parametric GP is a more general and natural choice, and it effectively models the uncertainty of the prediction.

Warm-Start GP: Efficient initialization for user-specific customizations

Another common use case encountered in interaction design is optimizing the parameters of a technique to the particular abilities and preferences of an individual. This customization or personalization process can also be efficiently performed with the aid of multi-objective Bayesian optimization. Ideally, this process should be as fast as possible and one way to enhance efficiency is to initialize the MOBO procedure with a generalized appreciation of the design space. In practice, this can be achieved by selecting some subset of the data collected from all users to initialize the Bayesian optimization procedure when personalizing a technique to a new user. The problem then is to select a sparse subset of size K from the whole dataset that is representative of the whole dataset so as to quickly adapt to the needs of the new user. These selected points are used to construct a *Warm-Start GP* which then provides the initialization for running Bayesian optimization with the new user. Figure 2 illustrates this general procedure.

We simplify and adapt the greedy selection approach taken by Titsias (2009) by using the approximation to the marginal likelihood of the entire dataset with the size K subset given by Seeger et al. (2003). We refer the reader to those papers for more details on the greedy algorithm in Titsias (2009) and the approximation and hyperparameter

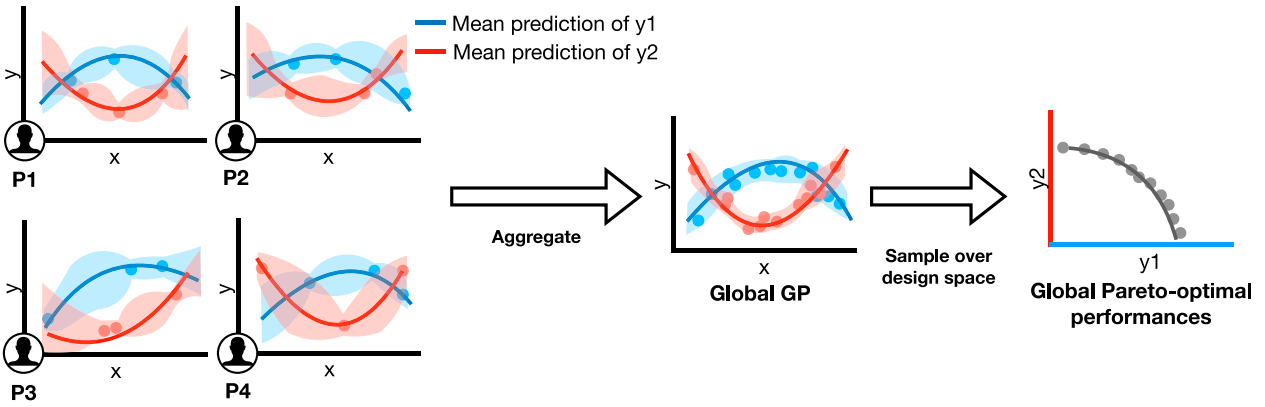

Figure 1. The Global GP aggregates all observations from the user-specific optimization processes. The consolidated model can thereby estimate the group's average performance at any given design parameter value. After constructing the Global GP, we perform a fine-grid sampling of the design space to identify the global Pareto-optimal design instances.

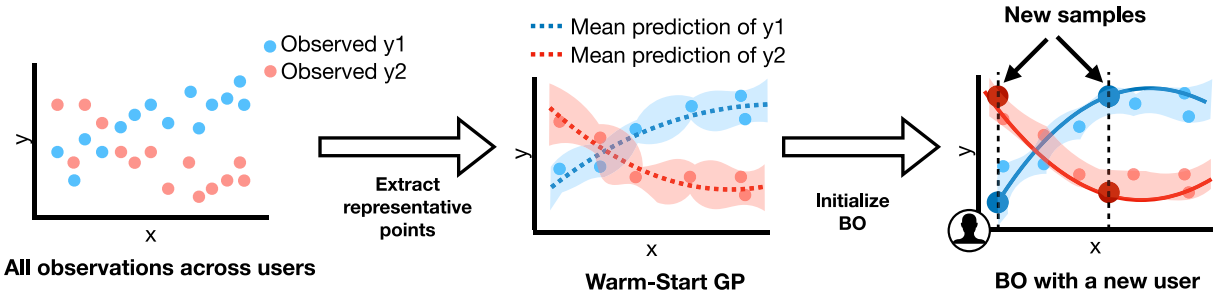

Figure 2. We extract the most representative observations from the current pool of users to form a Warm-Start GP. This Warm-Start GP serves as an informative prior, allowing rapid adaptation to individuals with fewer BO iterations.

tuning in Seeger et al. (2003). The central idea is to select the K points by iteratively adding a training point greedily from the complete dataset that maximizes the approximate marginal likelihood of the complete dataset. This approximate marginal likelihood of the complete dataset is determined from the GP constructed from the sparse subset and using the model to compute the approximate marginal likelihood over the complete dataset. We also apply the following heuristic to reduce computation time. First, instead of including the entire dataset as initial candidates for the sparse subset, we first reduce the set of all possible candidates to a randomly selected subset. For the touch-button temporal pointing task described in Study 2, we set the size of that randomly selected subset to be 100, corresponding to half of the size of the total number of data points collected (200). Next, from that reduced dataset, we apply the likelihood maximization process as stated above to greedily select the candidate point to be in the sparse subset of size K. We set K to 5 in the application described in Study 2. This method produces an appropriate prior which can adapt to a new user from newly given samples to obtain a personalized optimal design parameter set. Although there have been alternative methods proposed for sparsifying GPs (Bauer et al., 2016; Burt et al., 2020; Cao et al., 2013), we pursued this computationally efficient approach that retains representative data points from the original dataset.

Feasibility check for the proposed approaches

We now describe the key properties that render a design task suitable for the proposed techniques. As is the case with conventional Bayesian optimization, our Global GP and Warm-Start GP approaches perform better when applied to design problems characterized by a smooth objective surface. In practical terms, this implies a requirement that the user performance or experience at a given design instance is not radically different from similar nearby design instances in the parameter space. The approaches can accommodate continuous, discrete, and ordinal design objectives; however, it is worth noting that discrete and ordinal data may lead to suboptimal outcomes due to the discontinuities these objective types may introduce in the objective surface. Prior work (Frazier, 2018; Moriconi et al., 2019) suggests that MOBO may accommodate up to 20 parameters, however, this may not be practical in most human-in-the-loop applications due to the number of evaluations required to even

sparsely explore the design space. Within this paper, we examine up to five parameters and three objectives, which we demonstrate good applicability in typical interaction technique design problems. While further research is required to validate the suitability of our approaches in other domains, we anticipate good applicability across different classes of interaction design problems, such as pointing, selection, and general temporal input tasks.

Implementation details

In this section, we provide implementation details for the GP and the proposed approaches. Following Shah and Ghahramani (2016), we use correlated multi-objective GP models as surrogate models—the Multi-task GP model and the Semiparametric Latent Factor GP. GPs applied to multi-objective scenarios have to take into account the covariances between the different objectives (*inter-task*) as well as the covariances between the individual datapoints within each task (*intra-task*). The main difference between these two types of GP models is that they have different covariance matrices. The analytical form and further details are available in the original paper (Shah and Ghahramani, 2016). For Multi-task GPs, the intra-task covariance is separated from the inter-task covariance, and for Semiparametric Latent Factor GPs, linear combinations of the intra-task covariances are taken with inter-task covariances as factors. As the inter-task and intra-task covariances are decoupled in Multi-task GPs, they are more computationally efficient than Semiparametric Latent Factor GPs. We found that Multi-task GPs are computationally more efficient than Semiparametric Latent Factor GPs. However, the latter is better able to model the interdependence between each of the objectives. Therefore, to balance between computational efficiency and slightly better modeling of the distinct objectives, for $L = 2$ objectives, we use the Semiparametric Latent Factor GP, and for $L \geq 3$, we use the Multi-task GP. We use the ARD Matérn 5/2 kernel and assume that each objective is observed with Gaussian noise.

For both GP forms, there are several hyperparameters that need to be tuned. Specifically, they include the intertask covariance parameters, the kernel parameters, and the noise estimates for each objective. They are tuned at each step of the optimization, when an additional observation is added, by maximizing the log-likelihood by performing gradient ascent with L-BFGS-B.

A further simplification for implementation purposes is the conversion of the continuous design space into a discrete one. This helps avoid the requirement to exhaustively search the space when optimizing the acquisition function. The approach involves evaluating a candidate list of sample points that provide representative coverage of the design space. Appropriate bounds for each parameter are chosen, and after normalization, this sets the limits of the hypercube. The candidate list is then constructed by sampling from the parameter hypercube using a Sobol sequence as described by Snoek (2013). The choice of the number of candidates involves a trade-off between search resolution and computation time. The acquisition function is evaluated at each of these candidates and the candidate with the highest acquisition value is selected as the point to sample in the next iteration.

Study 1: Individual-to-group design with the Global GP

In this study, we seek to validate our Global GP method for deriving a set of optimal designs that are representative of a group of users. We do this within the context of a design problem in 3D touch interaction loosely based on the Go-Go technique (Poupyrev et al., 1996). In the first phase of the study, participants performed 3D touch selections while the MOBO procedure sought to identify their individual set of Pareto optimal designs. The data from the individual participants was then used to generate a set of global Pareto-optimal designs using our Global GP method. In the second phase of the study, we evaluated the performance of two designs taken from this global Pareto-optimal set against a baseline configuration roughly based on the design of the original Go-Go technique (Poupyrev et al., 1996).

3D touch interaction is a subclass of 3D object selection based on the virtual hand metaphor. A wide array of selection techniques have been proposed and tailored to different applications by trading off between accuracy and speed (Argelaguet and Andujar, 2013; Bowman et al., 1999; Poupyrev and Ichikawa, 1999). Depending on the metaphor and the implementation of the interaction, the number of design parameters can range from 3 to 10 (Argelaguet and Andujar, 2013; Frees et al., 2007; König et al., 2009; Meyer et al., 1988). A more detailed summary of 3D selection and pointing techniques can be found in Argelaguet and Andujar (2013).

The *Go-Go technique* (Poupyrev et al., 1996) is a popular technique that enables users to touch virtual targets appearing beyond their physical reach. It employs a control-display gain approach that selectively applies a linear or non-linear scaling on the virtual hand according to the physical position of the real hand. Two parameters determine the selection of the mapping schema and the degree of the non-linearity. The general task of finding ideal control-display gain function parameters can be both challenging and time-consuming. Previously, the gain function of pointing devices has been decided by either extensive trial-and-error (Casiez and Roussel, 2011) or by heuristic iteration (Nancel et al., 2015; Yun et al., 2015). These approaches are costly in terms of time and effort and difficult to conduct without prior expertise. Perhaps as a

consequence of this difficulty, the Go-Go technique as described in Poupyrev et al. (1996) recommends parameter settings without providing clear rationale. Many similar interaction techniques presented in the literature also contain parameter values that were arbitrarily chosen or derived from informal pilot testing. As an illustration of an alternative approach, this study demonstrates how MOBO can efficiently and systematically guide the identification of design parameters most suitable for a sampled user population.

Design parameterization and objectives

The original Go-Go technique scaled hand motions by computing offsets with respect to the user's chest. To better capture the direction and dynamic range of this motion, we relocated the reference frame to the hand's position when fully retracted to the shoulder, as shown in Figure 3. We defined the distal bound of the *operation range* as the distance between the origin and the hand when the arm is fully extended, as shown in Figure 3(a). The Go-Go technique's scaling mechanism was applied over this operation range.

The 3D touch design was parameterized according to four variables as described in Table 1. There are two parameters, x_1 and x_2, that determine the resulting virtual hand position in 3D space. The first parameter, x_1, is referred to as D in the original Go-Go technique publication and describes the normalized distance in the operation range at which the mapping transitions from linear to non-linear scaling. The second parameter, x_2, is the non-linear scaling factor and is referred to as k in the original publication. Appropriate parameter ranges were determined by pilot testing. We set the parameter $x_1 \in [0, 1]$ and the parameter $x_2 \in [0, 0.5]$.

In an effort to further improve selection performance, we augment the original Go-Go technique by introducing a haptic cue when the target is reached. Previous works have shown that vibration can effectively assist selection (Pfeiffer and Stuerzlinger, 2015; Sallnäs and Zhai, 2003), and it is widely employed in commercial VR controllers. We selected two parameters to describe this vibrotactile feedback: the activation-vibration gap, x_3, and the vibration amplitude, x_4. The activation-vibration gap is the distance from the target at which the cue is activated and was bound to values between 15 cm before and 5 cm after the target. The vibration amplitude is the intensity of the cue and was bound between 0 and the maximum voltage level (3.1 V, 2.6 g). The duration of the vibration feedback was fixed at 300 ms.

Input selection techniques are characterized by a trade-off between speed and accuracy. We therefore chose two proxy measures of speed and accuracy to guide the optimization process: completion time and spatial errors in target acquisition. Completion time refers to the average duration between the moment of the first movement and the moment the target is successfully selected. Spatial error is the maximum overshoot distance, that is, the 3D Euclidean distance between the cursor represented by a virtual hand and the target's position. Both completion time and spatial error are minimization metrics, i.e., a smaller value indicates better performance. We convert these metrics into objectives which we subsequently refer to as *speed* and *accuracy*, before passing them into the optimizer. Based on pilot testing, we linearly map the completion time ranged [1,600 ms, 900 ms] to speed ranged $[-1, 1]$ and linearly map the spatial error ranged [2 cm, 0 cm] to accuracy ranged $[-1, 1]$.

Phase 1: User-specific optimizations

In this first phase of the study, participants were exposed to the standard MOBO procedure and the design parameters were optimized at the individual level. A set of global designs were then derived using the Global GP method, providing the basis for Phase 2 described later.

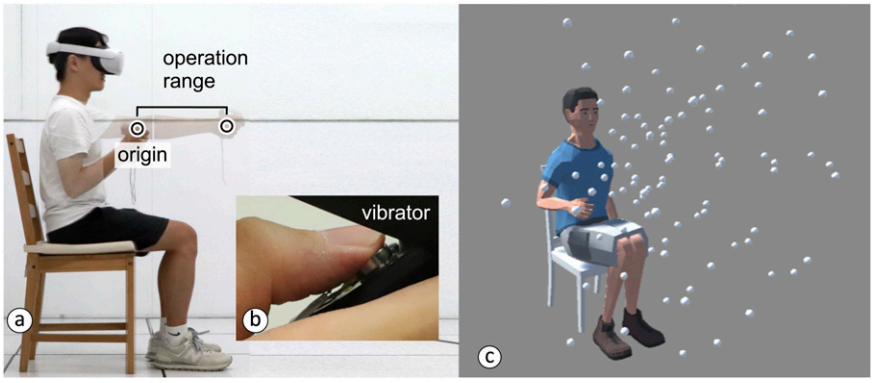

Figure 3. The experiment setup for the 3D touch task. (a) The reference origin and operation range as adapted from the original Go-Go technique design. (b) The interaction is enhanced with vibrotactile feedback via the vibrator added to the controller. (c) All possible locations of targets.

Table 1. Design parameterization of the 3D touch interaction.

Design parameter		Range
x_1:	Distance threshold, D	[0, 1]
x_2:	Scale factor, k	[0, 0.5]
x_3:	Activation-vibration gap	[15 cm, −5 cm]
x_4:	Vibration amplitude	[0 g, 2.6 g]

Participants. In total, we recruited 20 paid participants (nine female) from our university for the whole of Study 1. Their average age was 23.3 years (sd = 0.8). We randomly divided them into two groups. The group who participated in the first study will subsequently be labeled as the "experienced" group in Phase 2 of the study. The "novice" group only participated in Phase 2. Participants in the experienced group received 20, and participants in the novice group received 10 as a token of appreciation for their involvement.

Task. Participants performed a 3D touch task in a VR setup where the completion time and overshoot error were measured. Selection was performed based on a dwell threshold (0.5 s) with a cursor representing the virtual hand.

Apparatus and prototype. We built the 3D touch interaction application in Unity 3D. Participants wore a Meta Quest and performed the task with the companion hand controllers. These controllers were modified to include the custom vibration motors as shown in Figure 3(b).

Setup and procedure. We followed task arrangements used in Cha and Myung (2013) for 3D target acquisition. Each iteration of the task contained 36 trials (selecting a single target), drawn randomly from the radial distances, target widths, the azimuth, and inclination angles, to ensure the index of difficulty across trials was well distributed. The possible target locations are shown in Figure 3(c). A 5-min break was given every ten iterations. The whole procedure lasted approximately 90 min.

MOBO hyperparameters. The design configurations used in the first 10 task iterations of the experiment were randomly selected. The subsequent 40 task iterations utilized design configurations proposed by the MOBO procedure. The design space was discretized into 40 possible cue configurations.

Results of the user-specific optimizations

Participants exhibited natural differences in their performance and this is reflected in the identified optimal designs. Figure 4 shows the Pareto front and Pareto-optimal designs obtained for two illustrative participants. The optimal designs obtained for the first participant (left two plots) generally show superior performance in both objectives compared with the second participant. The two participants yielded rather distinct parameter designs suggesting that the MOBO procedure has successfully captured user-specific optimal designs.

Phase 2: Evaluation of the global designs from the Global GP

In this second phase of the study, we evaluated the performances of the global designs derived from the data collected in Phase 1 against a baseline design configuration. The purpose of this evaluation was to assess the quality of the designs produced by the Global GP method. If designs extracted by the Global GP perform as well or better than the baseline configuration, this indicates that the approach can effectively produce designs suitable for a group of users.

This evaluation was performed by both an *experienced group* who participated in Phase 1 and a *novice group* who were completely new to the study. There was a 2-day gap between Phase 1 and Phase 2. The structure of the evaluation was a 5 (designs) × 2 (groups of participants) mixed-design experiment with two independent variables. Each participant tested all of the design instances; thus, this factor is within-subjects. The two groups of participants is a between-subjects factor.

Generating global Pareto-optimal designs

We used the Global GP method based on all observations from all the participants in Phase 1 to generate a set of global Pareto-optimal designs. Each design parameter was equally divided into 16 levels for grid search in the Global GP method. The set of derived Pareto-optimal global designs is presented in Figure 5.

We grouped the global designs into a speed-oriented subset and an accuracy-oriented subset. The speed-oriented subset prioritized completion time over spatial errors, while the reverse is true for the accuracy-oriented subset. We then generated two final designs for evaluation by averaging over the individual parameter values in each subset.

These two final designs were then compared against the baseline Go-Go technique. As noted earlier, the original Go-Go technique does not include vibration feedback. To ensure a fair comparison, we augmented the standard Go-Go technique so that the vibration cue will be generated as the user contacts the target with the virtual hand (x_3 = 0 cm). This reflects a common setting used in VR interactions (Wang et al., 2020). The vibration amplitude (x_4) was set to 1 g, which was the most preferred and effective amplitude among four alternatives (0.5 g, 1 g, 1.5 g, and 2 g) presented in a pilot test. The three conditions evaluated in Phase 2 are summarized in Table 2.

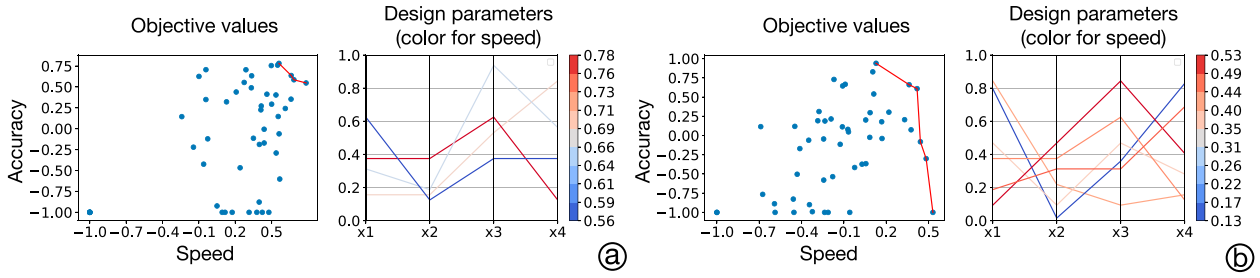

Figure 4. The Pareto front and Pareto-optimal designs obtained for two illustrative participants (a and b).

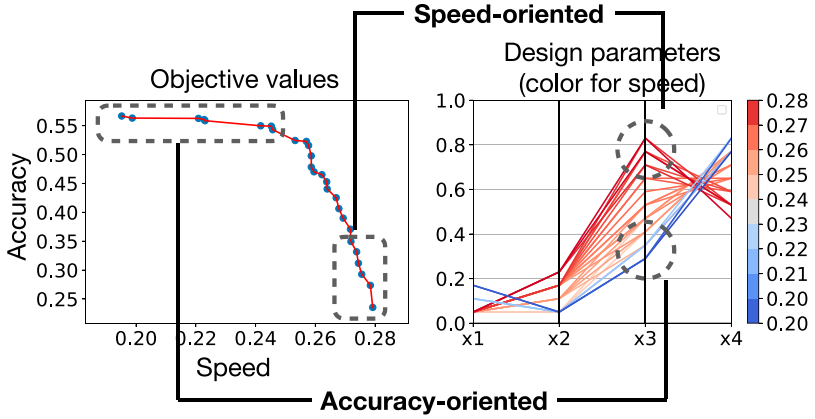

Figure 5. The predicted Pareto-optimal objective values from the Global GP and the global Pareto-optimal designs.

Table 2. The Three Design Conditions Evaluated in Phase 2.

Condition	x_1	x_2	x_3	$x_{4\ (g)}$
Speed oriented	0.05	0.10	5.77 cm	2.00
Accuracy oriented	0.09	0.04	10.76 cm	0.91
Go-Go technique	0.67	0.17	0.00 cm	1.00

Evaluation setup

The three design conditions were presented in four rounds, where each round consisted of 36 trials (target selections). The condition order was counterbalanced with a Latin square. The evaluation phase lasted approximately 15 min for each participant.

Performance of the global designs

Figure 6 shows the mean completion time and spatial error for each condition and participant group. For the mean *completion time* of the *experienced users*, the speed-oriented design, the accuracy-oriented design, and Go-Go technique were 1,038 ms (sd = 77.64), 1,081 ms (sd = 87.72), and 1,111 ms (sd = 80.32), respectively. For the mean completion time of the *novice users*, the speed-oriented design, the accuracy-oriented design, and Go-Go technique were 1,12 ms (sd = 126.28), 1,18 ms (sd = 127.10), and 1,167 ms (sd = 120.88), respectively. For the mean *spatial error* of the *experienced users*, the speed-oriented design, the accuracy-oriented design, and Go-Go technique were 2.34 cm (sd = 1.76), 0.97 cm (sd = 0.49), and 1.55 cm (sd = 0.85), respectively. For the mean spatial error of the *novice users*, the speed-oriented design, the accuracy-oriented design, and Go-Go technique were 2.03 cm (sd = 1.17), 0.96 cm (sd = 0.68), and 1.83 cm (sd = 0.78), respectively.

We conducted a mixed-design analysis of variance (mixed ANOVA) to examine the effect of interfaces and user experience levels. Sphericity was assessed with Mauchley's test, and if violated, Greenhouse-Geisser corrections were employed. The results revealed significant within-subject effects for both completion time (F (2, 36) = 7.48, $p < .005$) and spatial errors (F (1.432, 25.781) = 19.28, $p < .001$). Tests of between-subjects effects indicated there were no differences found between user experience levels (all $p > .05$). However, the generally higher completion times for novice users suggests a learning effect.

Pairwise comparisons were run for all conditions on both completion time and spatial errors and the significant

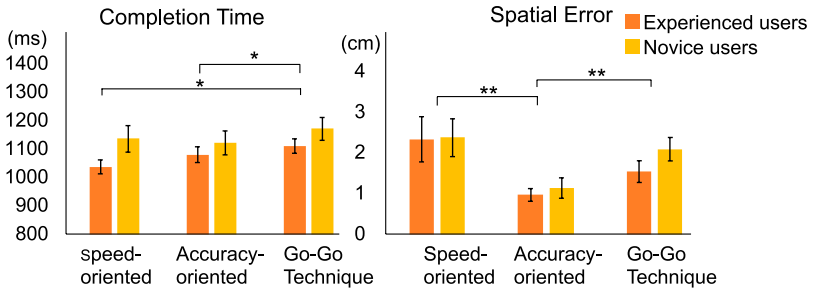

Figure 6. Results of the comparative study on three designs (two global designs and Go-Go technique) and over experienced and novice user groups. The error bars denote 1 standard deviation. The one-star (*) and two-star (**) symbols indicate $p < .05$ and $p < .001$ significant differences, respectively.

differences are noted in Figure 6. For completion time, both the speed-oriented and the accuracy-oriented designs outperformed the baseline design (all $p < .05$). No significant difference was found between the two global designs. With respect to spatial errors, the accuracy-oriented design was shown to be significantly better than the speed-oriented design and the baseline design (all $p < .001$). However, the speed-oriented design was not superior to the baseline in reducing spatial errors. Overall, both global designs brought significantly better or comparable performances to the users for both metrics. As expected, the speed-oriented design successfully delivered shorter completion times than the baseline, and the accuracy-oriented design significantly reduced spatial errors.

Summary

In this study, we showed the efficacy of deriving global Pareto-optimal design instances from user-specific observations using the Global GP method. The results of the comparative evaluation show that our global designs bring better or comparable performances for the user group compared with the baseline. This approach highlights how MOBO and the Global GP method can eliminate much of the design labor that is typically required to aggregate the preferences and behaviors of multiple individuals into a single sound design.

Study 2: Group-to-individual design with the Warm-Start GP

The second study validates our method for initializing the multi-objective Bayesian optimization procedure in order to enhance the efficiency of interaction optimization at the individual level. We refer to this initialization process as constructing a Warm-Start GP. This demonstration of the Warm-Start GP is contextualized by the design challenge of producing an *adaptive* touch-button for a temporal pointing task. To further highlight the capabilities of the MOBO procedure, we tackle this problem using a design parameterization of five variables and with respect to three objectives. One of these design objectives is based on a subjective user rating which has high relevance to many HCI design problems.

The approach we employed in this study to validate the Warm-Start GP method involved two phases. In Phase 1, participants were exposed to the standard MOBO procedure. The dataset generated in Phase 1 was then used to produce the MOBO initialization for Phase 2. In Phase 2, the same group of participants (which we refer to as the *experienced user group*) and a new group of users (which we refer to as the *novice user group*) completed the MOBO procedure in two different variants: once with the warm-start initialization and once with the default initialization. Using this protocol, we investigate whether our Warm-Start GP model can effectively leverage previously collected information on user performances and deliver more rapid adaptation to individual users.

The design problem examined has high relevance given that pressing a touch-button is a fundamental interaction on touchscreen devices. When the finger contacts a button, a key-click vibration signal is generated to notify the user of the activation of the matching function. Such a key-click signal affects the user's typing speed and errors on a soft keyboard (Ma et al., 2015) and subjective preferences (Pitts et al., 2009). The design of a touch-button is not a trivial task, though. Previous research has shown that an optimal point to trigger a button is not as the finger makes contact with the button (Kim et al., 2013; Liao et al., 2020). Rather, it is somewhere within the travel range. As our first study shows, determining proper haptic feedback for target selection is also not straightforward. Various haptic feedback leads to different sensations and performances. Further, the optimal point to render the haptic cue is not at the same point as where the selection happens (Figure 5). Additionally, vibration feedback is a continuous cue which lasts for a certain duration (Kim and Lee, 2013), and yet, the button triggering is momentary. Determining when and how to render a continuous cue to match a momentary event has not

been previously explored. Despite the prevalence of touch-buttons in our daily experiences and the challenges of designing them, there have been few attempts to investigate and iterate their design. Previous research attempted to optimize single objectives with iterative experimentation, including maximizing the button's information communication (Chen et al., 2011; Liao et al., 2017; Richter et al., 2010), minimizing typing error (Ma et al., 2015), and creating realistic physical-button sensations (Kim and Lee, 2013). However, iterative experimentation is not conducive to the efficient exploration of a multi-dimensional design space (Chang et al., 2020; Chen et al., 2011; Park et al., 2020; Richter et al., 2010) and risks omitting promising designs. Liao et al. (2020) applied Bayesian optimization to the task of designing the haptic characteristics of a push-button. However, Liao et al.'s (2020) study is on physical buttons and the method is limited to optimizing a single objective.

In this study, we sought to derive an adaptive model of touch-button pressing for temporal pointing tasks. Temporal pointing refers to tasks consisting of entering certain discrete inputs within a short time window (Lee and Oulasvirta, 2016). It is not only a common interaction for games in which a function must be elicited at a particular moment (for instance, to attack an enemy); it is also a synchronization task (Wing and Kristofferson, 1973) that occurs in daily input experiences. To our knowledge, no prior work has applied multi-objective optimization methods to search for the optimal button design for such a task.

Design parameterization and objectives

Prior work has demonstrated that the activation point of the push-button affects typing speed (Kim et al., 2018), and that the vibration emission timing impacts temporal errors (Liao et al., 2020). We translate the depth sensing of a push-button to the pressing force level on a touch-button. This approach is illustrated in Figure 7, where the activation of the button functionality and the vibration may occur at different times and after the user's first initial contact with the button. Each event is triggered when a certain level of pressing force is detected. In this illustrated example, the force threshold for activating the button is lower than the force threshold for activating vibration, so the button would be activated prior to the vibration cue being generated.

We examine five design parameters as summarized in Table 3. x_1 (Button-Activation Point) and x_2 (Vibration Point) were explained in the previous paragraph and are illustrated in Figure 7. x_3 (Initial Vibration Amplitude) and x_4 (Final Vibration Amplitude) define the amplitude level of the vibration cue after the detected force exceeds the vibration point—that is, the moment that the vibration should be generated. When x_3 and x_4 have different values, the amplitude linearly increases or decreases over the Vibration Duration, which is defined as x_5.

We aim to maximize the temporal performance (Lee and Oulasvirta, 2016) of button pressing and the user's subjective rating. The temporal performance is assessed by two separate objective measures: the mean value and the standard deviation of the temporal errors of all the presses. The three objectives that govern the optimization process are summarized in Table 4.

Phase 1: User-specific optimizations

In Phase 1 of the study, participants completed the standard MOBO procedure. The data collected from all participants in Phase 1 was then used to generate the Warm-Start initialization. This initialization was subsequently evaluated in Phase 2 as described later.

Participants. In total, 22 participants were recruited from our local institution for the whole study. Among them, 10 participants completed both phases, which were performed on different days (three female, average age = 23.5 years, sd = 5). This group of users is referred to as the

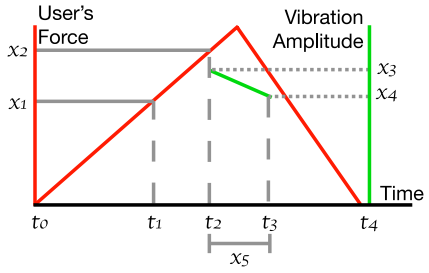

Figure 7. Illustrative design example where x_1 (Button-Activation Point) is a lower force threshold than x_2 (Vibration Point). The user starts pushing the button at t_0. The detected force exceeds the activation point at t_1 and the button is activated. At t_2, the force reaches the vibration point and the tactile cue is triggered. The initial vibration amplitude is set at x_3 and the vibration linearly decays until the amplitude becomes x_4. The vibration duration is x_5, and thus the vibration stops at t_3. The user's finger is completely lifted from the sensor at t_4, and the button is reset.

Table 3. Design parameterization of the touch-button.

Design parameter		Range
x_1:	Button activation force level	[15 g, 1515 g]
x_2:	Vibration activation force level	[15 g, 1515 g]
x_3:	Initial vibration amplitude	[0 g, 3.2 g]
x_4:	Final vibration amplitude	[0 g, 3.2 g]
x_5:	Vibration duration	[0 s, 1.5 s]

"experienced user group." The total duration for these 10 participants was 90 min, and participants received two movie tickets, worth 24, in appreciation for their involvement. Twelve additional participants were invited to participate *only in the second phase* (four female, average age = 22.9 years, sd = 2.40). The duration for these participants was about 30 min, and they received one movie ticket, worth 12. This group of users is referred to as the "novice user group."

Task. Participants (the ones in the experienced user group) were asked to perform a temporal pointing task. The graphical interface for the task is shown in Figure 8(a). A red "bullet" moves from right to left along the white bar as illustrated in Figure 8(a). Participants were instructed to activate the button when the bullet reached the center of the yellow target zone. When the button is activated, the red bullet turns blue.

Apparatus and prototype. We implemented a smartphone prototype (6 cm × 12.5 cm × 1 cm) with a force sensing resistor (FSR 402[2]) and an embedded vibration motor (Precision Microdrives 308–102,[3] rise time 21 ms) as shown in Figures 9(a) and (b). The vibration motor was driven by a motor driver (Sparkfun DRV2605L[4]), and the motor and the sensor were controlled by an Arduino Uno. The study interface shown in Figures 8(a) and (b) was implemented in processing.

Setup and procedure. In each iteration of the task, participants were presented with two levels of difficulty: easy (bullet moving at 625 pixels/second rate) and hard (1000 pixels/second). The two difficulty levels were presented in random order. Both difficulty levels required the participant to complete 12 trials (or presses). The first five presses at a given level were allocated as familiarization trials and their data were not used for performance calculation. The remaining seven presses were used to calculate the mean and standard deviation of the temporal error. After all presses were completed at both difficulty levels, participants were asked to rate the vibration design iteration they had just experienced. The statement, "*The vibration cue synchronizes (matches) with the button pressing interaction*," was presented to participants as illustrated in Figure 8(b). Participants were asked to submit their subjective rating on a scale from 0 to 100; 100 for strongly agree and 0 for strongly disagree. Five levels [1, 2, 3, 4, 5] were shown above the continuous slider to provide a coarse reference frame. A 2-min break was given after every 15 iterations of the task to avoid fatigue. Phase 1 lasted for approximately 60 min per participant.

MOBO hyperparameters. The design configurations used in the first five task iterations were randomly selected. The subsequent 45 task iterations used design configurations proposed by the MOBO procedure. The design space was discretized into 45 possible cue configurations.

Phase 2: Evaluation of the Warm-Start GP

A total of 500 (50 design configurations × 10 participants) observations were collected in Phase 1 of the study. We

Table 4. The design objectives of the touch-button.

Objective	Description
Temporal error mean	The temporal pointing is more accurate if this value is smaller.
Temporal error standard deviation	The temporal pointing is more precise if this value is smaller.
Subjective user rating	The vibration cue matches the click interaction more if this value is higher. Values from 0 to 100.

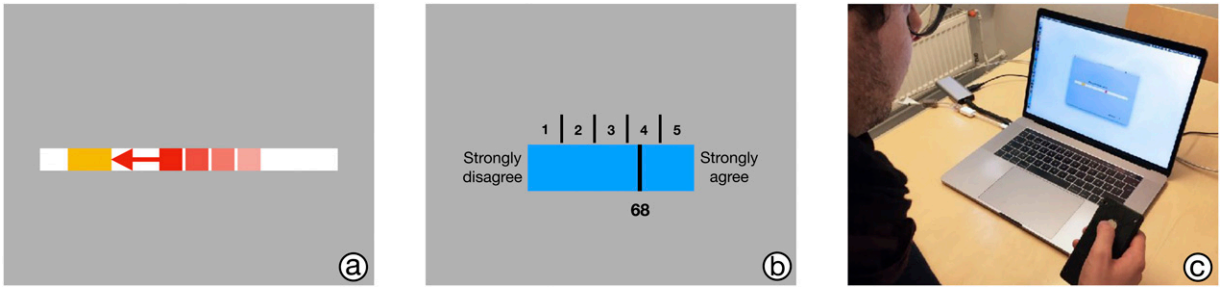

Figure 8. (a) A simplified sketch of the study interface during button pressing. The participant is asked to activate the button (the red bullet turns blue) when the red bullet reaches the yellow target area. (b) After 24 presses, the user is then asked to rate the vibration cue. (c) The study interaction.

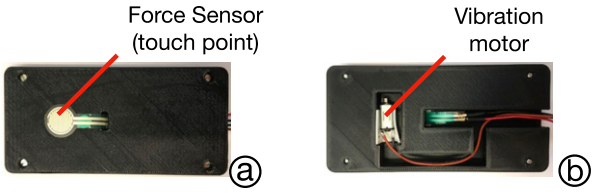

Figure 9. (a) The smartphone prototype: Users were instructed to touch the center of the force sensor. (b) The vibration motor is mounted inside the smartphone prototype.

applied the Warm-Start GP method on this dataset to select a subset of representative points as an initialization for a new GP model. We hypothesized that the number of "warm-start" points included in the initialization would influence the effectiveness of the adaptation. The reasoning behind this was as follows. If too few points are included, they will not provide a meaningful prior. Conversely, including too many "warm-start" points may provide a too strong a prior, leading to incoming observations of the new user having potentially limited influence on the model.

To select a reasonable number of points, we created three warm-start models with 5, 10, and 15 initial representative observations. We used the data from the 10 participants in Phase 1 to construct 10 surrogate GP models to simulate each individual's performance when given a design. We then isolated one GP at a time and treated it as a synthetic user. We applied the derived Warm-Start GP models (with 5, 10, or 15 observations) on these synthetic users for 15 iterations, which gave us 15 new observations for each synthetic user. We calculated the hypervolume based on these new observations. The results of the simulation indicate that five initial warm-start points is the best setting since this initialization results in the highest hypervolume increase within 15 iterations. We observed that with 10 or 15 prior observations, there is a tendency for the MOBO procedure to initially suggest similar designs for all new users, indicating that the prior is dominating and more observations from the new user will be needed to achieve more tailored optimization results.

Participants. Two groups of users were recruited for Phase 2. The *experienced users* are the same 10 participants who attended Phase 1. We further recruited 12 additional *novice users* to validate the effectiveness of the Warm-Start GP on new users (four female, average age = 22.9 years, sd = 2.4). Both groups went through an identical procedure as described below.

Evaluation setup

The evaluation was conducted 2 days after Phase 1 and employed a repeated-measures design with one factor and two conditions: MOBO performed with the Warm-Start GP initialization and MOBO performed without including any prior observations. The condition order was counterbalanced. The tasks given to the participants were the same as in Phase 1. For the MOBO condition without initialization, the design configurations in the first five task iterations were randomly selected. A further 10 task iterations were performed with designs proposed by the MOBO procedure. In the MOBO condition with the warm-start initialization, no initial random sampling was performed such that only 15 task iterations were performed with all designs proposed by the MOBO procedure.

Results of evaluating the Warm-Start GP

The hypervolume increase for each condition over the experiment iterations is plotted in Figure 10. We performed two separate two-way repeated measures ANOVAs to analyze the effect of *initialization* (with or without the Warm-Start GP initialization) and *iterations* on the *hypervolume increase* for each of the experienced user group and the novice user group. For the experienced user group, we found no statistically significant interaction between the effect of initialization and iterations (F (14, 126) = 0.38, $p >$ .05). Simple main effects analysis showed that the hypervolume was significantly higher when the experienced users start with the Warm-Start GP initialization than without (F (Dudley et al., 2019; Hornbæk, 2013) = 23.43, $p <$ .001). Simple main effects analysis also showed that there were significant differences between the iterations for the experienced user group (F (14, 126) = 31.88, $p <$ .001). We further performed paired samples t-tests to compare the hypervolume between the with and without Warm-Start GP initialization conditions at each iteration. There were significant differences throughout all the iterations (all $p <$ .05, see significance level notation in Figure 10).

For the novice user group, there was a statistically significant interaction between the effect of initialization and iterations (F (14, 154) = 8.25, $p <$ .001). Simple main effects analysis showed that the hypervolume was significantly higher when the novice users started with the Warm-Start GP initialization than without (F (Hornbæk, 2013; Koyama et al., 2020) = 19.24, $p <$ .001) and there was also a significant effect for iterations (F (14, 154) = 32.17, $p <$ .001). We then ran paired samples t-tests to compare the hypervolume between the with and without Warm-Start GP initialization conditions at each iteration. The results showed that the Warm-Start GP produced significantly higher hypervolume in iterations 1 to 11 (all $p <$ .05, see significance level notation in Figure 10). There were no significant differences in the hypervolumes at iterations 12 to 15. This result shows that Warm-Start GP effectively supported faster exploration for novice users.

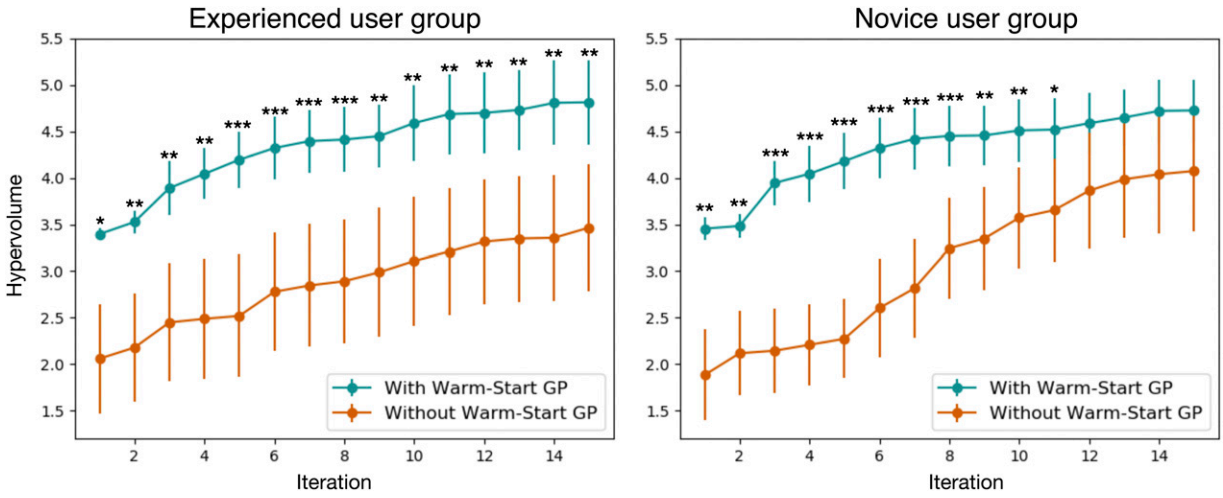

Figure 10. Impact of a Warm-Start GP initialization on experienced and novice users. The hypervolume increases throughout the 15 iterations for both experienced users and novice users during the evaluation. Error bars denote one standard error. We also indicate the significance level for the difference in hypervolume for the conditions with and without initialization at each iteration. The one-star (*) indicates $p < .05$, two-star (**) indicates $p < .01$, and three-star (***) indicates $p < .001$.

Overall, this result suggests that our selected warm-start points provided an appropriate prior and thus a useful initialization for delivering rapid adaptation to the individual users. Incorporating the Warm-Start GP initialization enabled the MOBO procedure to present more designs offering improvements in the design objectives for both novice and experienced users, manifesting as a larger final hypervolume. Another way to frame this result is that for novice users, just five iterations using the Warm-Start GP initialization yielded a set of higher-performing designs than 15 iterations without any initialization.

Summary

This study demonstrates that the Warm-Start GP method can provide an initialization delivering a faster hypervolume increase than that obtained by MOBO starting with a standard initialization, which allows faster adaptation to the preferences and abilities of an individual user. The Warm-Start GP is effective not just for the same group of users but also for new users, indicating that the Warm-Start GP is an effective method for transferring a prior understanding of the design space to different user groups. Despite the generally positive findings, different groups of users may have various preferences and optimal designs. Therefore, collecting data points from a larger pool of users may produce a more general Warm-Start GP when targeting new groups. Additionally, clustering the users and generating various Warm-Start GPs may also result in faster adaptation.

Discussion and future work

The novelty of this work chiefly lies in the demonstration of two practical approaches facilitating the application of MOBO in interaction technique design. To this end, we have (i) introduced the Global GP concept for extracting Pareto-optimal designs representative of a user group; (ii) introduced the Warm-Start GP method for initializing the MOBO procedure to enable more rapid adaptation at the individual level; and (iii) demonstrated the efficacy of both methods in two representative HCI design problems. Our methods effectively identified the group-optimized designs and reduced the time required for running individual optimizations.

The approaches introduced in this paper further provide efficient means of comprehending and predicting the outcomes associated with various design choices. A notable example is found in Study 1, where designers can refer to the global Pareto-optimal design figure (see Figure 5) and proactively make decisions favoring either speed or accuracy. In contrast, when following a conventional design process, designers may face challenges in efficiently constructing group-level aggregated results, and may need to expend additional effort in analyzing the implications of different design choices.

While the studies presented exhibit promising results, our work also highlights several open research questions. The presented approaches employ user-specific observations to construct a global model in a post-hoc step. In the future, it is worth investigating an online global model that iteratively updates as more users' data is aggregated. Further, the current method unifies the observations from all

users into a GP model. This approach may lead to unwanted outcomes if there are distinct user groups that favor drastically different designs, or if all users behave completely differently. To accommodate such cases, future research should explore the use of more advanced algorithms, such as hierarchical Gaussian process regression (Park and Choi, 2010), which may help with clustering user groups according to their distinct design preferences, and identify the group to which each new user belongs. With this potential extension, the Global GP and Warm-Start GP could potentially perform similarly to the Jury Learning concept proposed by Gordon et al. (2022) by, for example, identifying which prior cluster of users can serve as a good reference for the new user. Future work is also encouraged to integrate Bayesian optimization with meta-learning or transfer-learning techniques (Bai et al., 2023; Volpp et al., 2019), which are alternative techniques for rapid user-specific optimization based on previous optimization experience.

Our Warm-Start GP method selects a subset of prior observations to construct an adaptive model; however, determining the number of selected observations involves a trade-off between offering an informative prior versus providing a model capable of rapid adaptation. Incorporating too few prior observations may provide insufficient information for guiding productive optimization in the initial iterations. On the other hand, incorporating too many prior observations may lead to the new user's data being consistently dominated by the initial points and so no user-specific adaptation may occur. Currently, we determine the appropriate number of "warm-start" points through simulations, but designers could potentially proactively steer the optimization behavior by varying the number of observations in the Warm-Start GP based on their needs. Future work can also consider developing a dynamic approach: the Warm-Start GP initially has more prior observations to ensure meaningful acquisition, but the number of observations decreases as the adaptation continues, allowing the MOBO procedure to gradually rely more on new observations. Also, for computational efficiency, the selection of the Warm-Start observations involves randomly selecting a subset of the prior observations. Future work may also wish to explore more computationally efficient methods to consider all the prior observations without making subsets.

An assumption made throughout this study is that user performance or experience does not drastically vary due to time or order effects. In reality, there are several well-known user-related factors other than the design itself that are likely to affect the user's performance including fatigue, learning effects, and attention. Incorporating the Global GP and Warm-Start GP with non-stationary Bayesian optimization (Snoek et al., 2014) is a direction worth investigating to address this challenge. With subjective ratings, a user's preference may drift or be influenced by new exposures. For instance, the rating of a system tends to be biased by recent trials. These effects were not incorporated into our performance models. How to resolve the issues raised by the uncertainty of human users remains a compelling open research question.

Conclusions

Interaction design often involves a large number of parameters that need to be decided while also considering multiple design objectives. Although multi-objective Bayesian optimization (MOBO) offers a principled method for guiding design exploration, prior work has largely ignored the practical need for interaction technique design to meet the requirements of a population or group of users. To bridge this gap, we present (i) the *Global GP* to identify group-level optimal designs and (ii) the *Warm-Start GP* for rapid adaptation based upon a suitable prior extracted from the group-level data. We demonstrate the effectiveness of the *Global GP* and *Warm-Start GP* methods in two challenging and representative design problems. We show that the *Global GP* facilitates the identification of group-level Pareto-optimal designs, and that these designs are indeed competitive with a design arrived at by conventional means. We also show that the *Warm-Start GP* improves the efficiency of individual optimization by incorporating group-level data in the initialization for MOBO. Both methods are readily applied to other design problems involving multiple objectives and we hope that the guidance provided in this paper will promote wider uptake of MOBO in interaction technique design.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by the Research Council of Finland (flagship program: Finnish Center for Artificial Intelligence, FCAI, grants 328400, 345604, 341763; Human Automata, grant 328813; Subjective Functions, grant 357578) and the National Science and Technology Council of Taiwan (NSTC 112-2221-E-A49-125). John J. Dudley and Per Ola Kristensson were supported by the EPSRC (grants EP/S027432/1 and EP/W02456X/1).

Open science statement

The program and materials in this paper are released on our project page: https://cbl.aalto.fi/practical_mobo/. The Python library is open-sourced for designers and developers to use.

ORCID iD

Yi-Chi Liao https://orcid.org/0000-0002-2670-8328

Notes

1. Strictly, several designs may be equally good if they all achieve the same performance in terms of the chosen objective.
2. https://www.interlinkelectronics.com/fsr-402
3. https://www.precisionmicrodrives.com/product/308-102-8mm-vibration-motor-15mm-type
4. https://www.sparkfun.com/products/14538

References

Argelaguet F and Andujar C (2013) A survey of 3D object selection techniques for virtual environments. *Computers & Graphics* 37(3): 121–136. DOI: 10.1016/j.cag.2012.12.003.

Bai T, Li Y, Shen Y, et al. (2023) Transfer learning for Bayesian optimization: a survey. arXiv preprint arXiv:230205927.

Bauer M, van der Wilk M and Rasmussen CE (2016) Understanding probabilistic sparse Gaussian process approximations. NIPS'16. https://arxiv.org/abs/1606.04820.1606.04820.

Borji A and Itti L (2013) Bayesian optimization explains human active search. In Burges C, Bottou L, Welling M, et al. (eds) *Advances in Neural Information Processing Systems*. Red Hook, NY: Curran Associates, Inc, Vol. 26. https://proceedings.neurips.cc/paper/2013/file/a3f390d88e4c41f2747bfa2f1b5f87db-Paper.pdf

Bose T, Reina A and Marshall JA (2017) Collective decision-making. *Current opinion in behavioral sciences* 16: 30–34.

Bowman DA, Johnson DB and Hodges LF (1999) Testbed evaluation of virtual environment interaction techniques. In: Proceedings of the ACM symposium on virtual reality software and technology. VRST '99, New York, NY, USA: Association for Computing Machinery, 26–33. DOI: 10.1145/323663.323667.

Brochu E, Brochu T and de Freitas N. (2010) A Bayesian interactive optimization approach to procedural animation design. In: Proceedings of the 2010 ACM SIGGRAPH/eurographics symposium on computer animation. SCA '10, Goslar, DEU: Eurographics Association, 103–112.

Burt DR, Rasmussen CE and Wilk M (2020) Convergence of sparse variational inference in Gaussian processes regression. In *Journal of Machine Learning Research* 21: 1–63. https://jmlr.org/papers/v21/19-1015.html

Cao Y, Brubaker MA, Fleet DJ, et al. (2013) Efficient optimization for sparse Gaussian process regression. In: Proceedings of the 26th international conference on neural information processing systems. NIPS'13. Red Hook, NY: Curran Associates Inc., Vol. 1, pp. 1097–1105.

Casiez G and Roussel N (2011) No more bricolage! methods and tools to characterize, replicate and compare pointing transfer functions. In: Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology. UIST '11, New York, NY, USA: Association for Computing Machinery, p. 603–614. DOI: 10.1145/2047196.2047276.

Cha Y and Myung R (2013) Extended fitts' law for 3d pointing tasks using 3d target arrangements. *International Journal of Industrial Ergonomics* 43(4): 350–355. DOI: 10.1016/j.ergon.2013.05.005.

Chan L, Liao YC, Mo GB, et al (2022) *Investigating positive and negative qualities of human-in-the-loop optimization for designing interaction techniques*. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3491102.3501850

Chang Z, Ta TD, Narumi K, et al. (2020) Kirigami haptic swatches: design methods for cut-and-fold haptic feedback mechanisms. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20, New York, NY, USA: Association for Computing Machinery, p. 1–12. DOI: 10.1145/3313831.3376655.

Chen H, Park J, Dai S, et al. (2011) Design and evaluation of identifiable key-click signals for mobile devices. *IEEE Transactions on Haptics* 4(4): 229–241.

Dudley JJ, Jacques JT and Kristensson PO. (2019) Crowdsourcing interface feature design with Bayesian optimization. In: Proceedings of the 2019 CHI conference on human factors in computing systems. CHI '19, New York, NY, USA: Association for Computing Machinery, pp. 1–12. DOI: 10.1145/3290605.3300482.

Dunlop M and Levine J (2012) Multidimensional pareto optimization of touchscreen keyboards for speed, familiarity and improved spell checking. In: Proceedings of the SIGCHI conference on human factors in computing systems. CHI '12, New York, NY, USA: Association for Computing Machinery, pp. 2669–2678. DOI: 10.1145/2207676.2208659.

Feit AM, Sridhar S and Bachynskyi M (2015) Towards multi-objective optimization for UI design. CHI '15 Workshop on Principles, Techniques and Perspectives on Optimization and HCI, 4.

Frazier PI (2018) A tutorial on Bayesian optimization. 1807.02811.

Frees S, Kessler GD and Kay E (2007) Prism interaction for enhancing control in immersive virtual environments. *ACM Transactions on Computer-Human Interaction* 14(1): 2.

Gajos K and Weld DS (2004) SUPPLE: automatically generating user interfaces. IUI '04, 8.

Gergle D and Tan DS (2014) *Experimental Research in HCI*. New York, NY: Springer, 191–227. DOI: 10.1007/978-1-4939-0378-8_9.

Gordon ML, Lam MS, Park JS, et al. (2022) Jury learning: integrating dissenting voices into machine learning models. In: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems. CHI '22, New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3491102.3502004.

Hayward V, Choksi J, Lanvin G, et al. (1994) Design and multi-objective optimization of a linkage for a haptic interface. In

Lenarčič J and Ravani B (eds) *Advances in Robot Kinematics and Computational Geometry*. Dordrecht: Springer Netherlands, pp. 359–368. DOI: 10.1007/978-94-015-8348-0_36.

Hernandez-Lobato D, Hernandez-Lobato J, Shah A, et al. (2016) Predictive entropy search for multi-objective Bayesian optimization. In: International conference on machine learning. New York, NY: PMLR, pp. 1492–1501. https://proceedings.mlr.press/v48/hernandez-lobatoa16.html

Hornbæk K. Some whys and hows of experiments in human–computer interaction. *Foundations and Trends® in Human–Computer Interaction* 2013; 5(4): 299–373.

Kadner F, Keller Y and Rothkopf C. (2021) AdaptiFont: increasing individuals' reading speed with a generative font model and Bayesian optimization. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21, New York, NY, USA: Association for Computing Machinery. DOI: 10.1145/3411764.3445140.

Karrenbauer A and Oulasvirta A (2014) Improvements to keyboard optimization with integer programming. In: Proceedings of the 27th Annual ACM Symposium on User Interface Software and Technology. UIST '14, New York, NY, USA: Association for Computing Machinery, pp. 621–626. DOI: 10.1145/2642918.2647382.

Khajah MM, Roads BD, Lindsey RV, et al. (2016) Designing engaging games using Bayesian optimization. In Proceedings of the 2016 CHI conference on human factors in computing systems. CHI '16, New York, NY, USA: Association for Computing Machinery, pp. 5571–5582. DOI: 10.1145/2858036.2858253.

Kim S and Lee G (2013) Haptic feedback design for a virtual button along force-displacement curves. In Proceedings of the 26th annual ACM symposium on user interface software and technology. UIST '13, New York, NY, USA: ACM, pp. 91–96. DOI: 10.1145/2501988.2502041.

Kim S, Son J, Lee G, et al. (2013) Tapboard: making a touch screen keyboard more touchable. In: Proceedings of the SIGCHI conference on human factors in computing systems. CHI '13, New York, NY, USA: Association for Computing Machinery, p. 553–562. DOI: 10.1145/2470654.2470733.

Kim S, Lee B and Oulasvirta A (2018) Impact activation improves rapid button pressing. In: Proceedings of the 2018 CHI conference on human factors in computing systems. CHI '18, New York, NY, USA: ACM, pp. 571–578. DOI: 10.1145/3173574.3174145.

Knowles J. ParEGO: a hybrid algorithm with on-line landscape approximation for expensive multiobjective optimization problems. *IEEE Transactions on Evolutionary Computation* 2006; 10(1): 50–66. DOI: 10.1109/TEVC.2005.851274.

König WA, Gerken J, Dierdorf S, et al. (2009) Adaptive pointing–design and evaluation of a precision enhancing technique for absolute pointing devices. In IFIP Conference on Human-Computer Interaction. Berlin: Springer, pp. 658–671.

Koyama Y, Sato I, Sakamoto D, et al. (2017) Sequential line search for efficient visual design optimization by crowds. *ACM Transactions on Graphics* 36(4): 1–48. DOI: 10.1145/3072959.3073598.

Koyama Y, Sato I and Goto M (2020) Sequential gallery for interactive visual design optimization. *ACM Transactions on Graphics* 39(4): 1–88. DOI: 10.1145/3386569.3392444.

Lee B and Oulasvirta A. (2016) Modelling error rates in temporal pointing. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. CHI '16, New York, NY, USA: Association for Computing Machinery, p. 1857–1868. DOI: 10.1145/2858036.2858143.

Lee MK, Kusbit D, Kahng A, et al. (2019) Webuildai: participatory framework for algorithmic governance. *Proceedings of the ACM on Human Computer Interaction* 3: 1–35. doi: 10.1145/3359283

Liao YC, Chen YC, Chan L, et al. (2017) Dwell+: multi-level mode selection using vibrotactile cues. In: Proceedings of the 30th annual ACM symposium on user interface software and technology. UIST '17. New York, NY, USA: Association for Computing Machinery, p. 5–16. DOI: 10.1145/3126594.3126627.

Liao YC, Kim S, Lee B, et al. (2020) Button simulation and design via fdvv models. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. CHI '20. New York, NY, USA: Association for Computing Machinery, p. 1–14. DOI: 10.1145/3313831.3376262.

Liao YC, Dudley JJ, Mo GB, et al. (2023) Interaction design with multi-objective bayesian optimization. *IEEE Pervasive Computing* 22: 29–38. DOI: 10.1109/MPRV.2022.3230597.

Lindlbauer D, Feit AM and Hilliges O (2019) Context-aware online adaptation of mixed reality interfaces. In: Proceedings of the 32nd annual ACM symposium on user interface software and technology. New Orleans LA USA: ACM, pp. 147–160. DOI: 10.1145/3332165.3347945.

Ma Z, Edge D, Findlater L, et al. (2015) Haptic keyclick feedback improves typing speed and reduces typing errors on a flat keyboard. In 2015 IEEE World Haptics Conference (WHC), Evanston, IL, USA, 2015, pp. 220–227.

Meyer DE, Abrams RA, Kornblum S, et al. (1988) Optimality in human motor performance: ideal control of rapid aimed movements. *Psychological Review* 95(3): 340–370.

Moriconi R, Deisenroth MP and Kumar K (2019) High-dimensional Bayesian optimization using low-dimensional feature spaces. arXiv preprint arXiv:190210675.

Nancel M, Pietriga E, Chapuis O, et al. (2015) Mid-air pointing on ultra-walls. *ACM Transactions on Computer-Human Interaction* 22(5): 1–62. DOI: 10.1145/2766448.

Nielsen JBB, Nielsen J and Larsen J (2015) Perception-based personalization of hearing aids using Gaussian processes and active learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 23(1): 162–173. DOI: 10.1109/TASLP.2014.2377581.

Oulasvirta A, Dayama NR, Shiripour M, et al. (2020) Combinatorial optimization of graphical user interface designs. *Proceedings of the IEEE* 108(3): 434–464.

Park S and Choi S (2010) Hierarchical Gaussian process regression. In: Proceedings of 2nd Asian Conference on Machine Learning. JMLR Workshop and Conference Proceedings, Tokyo, Japan, 2010, pp. 95–110. https://proceedings.mlr.press/v13/park10a.html

Park C, Yoon J, Oh S, et al. (2020) *Augmenting Physical Buttons with Vibrotactile Feedback for Programmable Feels*. New York, NY, USA: Association for Computing Machinery, p. 924–937. DOI: 10.1145/3379337.3415837.

Pfeiffer M and Stuerzlinger W (2015) 3d virtual hand pointing with ems and vibration feedback. In 2015 IEEE symposium on 3D user interfaces (3DUI), Arles, France, 2015, pp. 117–120. DOI: 10.1109/3DUI.2015.7131735

Picheny V (2015) Multiobjective optimization using Gaussian process emulators via stepwise uncertainty reduction. *Statistics and Computing* 25(6): 1265–1280. DOI: 10.1007/s11222-014-9477-x.

Piovarči M, Kaufman DM, Levin DIW, et al. (2020) Fabrication-in-the-loop co-optimization of surfaces and styli for drawing haptics. *ACM Transactions on Graphics* 39(4): 116:1-116:16. doi: 10.1145/3386569.3392467

Pitts MJ, Williams MA, Wellings T, et al. (2009) *Assessing Subjective Response to Haptic Feedback in Automotive Touchscreens*. AutomotiveUI '09. New York, NY, USA: Association for Computing Machinery, 11–18. DOI: 10.1145/1620509.1620512.

Poupyrev I and Ichikawa T (1999) Manipulating objects in virtual worlds: categorization and empirical evaluation of interaction techniques. *Journal of Visual Languages & Computing* 10(1): 19–35. DOI: 10.1006/jvlc.1998.0112.

Poupyrev I, Billinghurst M, Weghorst S, et al. (1996) The go-go interaction technique: non-linear mapping for direct manipulation in vr. In: Proceedings of the 9th Annual ACM Symposium on User Interface Software and Technology. UIST '96, New York, NY, USA: Association for Computing Machinery, p. 79–80. DOI: 10.1145/237091.237102.

Richter H, Ecker R, Deisler C, et al. (2010) Haptouch and the 2+1 state model: potentials of haptic feedback on touch based in-vehicle information systems. In: Proceedings of the 2nd International Conference on Automotive User Interfaces and Interactive Vehicular Applications. AutomotiveUI '10. New York, NY, USA: Association for Computing Machinery, 72–79. DOI: 10.1145/1969773.1969787.

Sallnäs E and Zhai S (2003) Collaboration meets fitts' law: passing virtual objects with and without haptic force feedback. In: *INTERACT*. Amsterdam: IOS Press.

Seeger MW, Williams CK and Lawrence N (2003) Fast forward selection to speed up sparse Gaussian process regression. *AISTATS*. Key West, FL, USA: Proceedings of Machine Learning Research.

Shah A and Ghahramani Z (2016) Pareto frontier learning with expensive correlated objectives. In International Conference on Machine Learning. New York, NY: PMLR, pp. 1919–1927. https://proceedings.mlr.press/v48/shahc16.html

Shahriari B, Swersky K, Wang Z, et al. (2016) Taking the human out of the loop: a review of Bayesian optimization. *Proceedings of the IEEE* ; 104(1): 148–175. DOI: 10.1109/JPROC.2015.2494218.

Snoek J (2013) *Bayesian Optimization and Semiparametric Models with Applications to Assistive Technology*. PhD Thesis. Toronto, ON, Canada: University of Toronto.

Snoek J, Swersky K, Zemel R, et al. (2014) Input warping for bayesian optimization of non-stationary functions. In: International conference on machine learning. New York, NY: PMLR, pp. 1674–1682.

Sridhar S, Feit AM, Theobalt C, et al. (2015) Investigating the dexterity of multi-finger input for mid-air text entry. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. CHI '15. Seoul, Republic of Korea: ACM Press, pp. 3643–3652. DOI: 10.1145/2702123.2702136.

Titsias M (2009) Variational learning of inducing variables in sparse Gaussian processes. In *Artificial Intelligence and Statistics*. New York, NY: PMLR, 567–574. https://proceedings.mlr.press/v5/titsias09a.html

Volpp M, Fröhlich LP, Fischer K, et al. (2019) Meta-learning acquisition functions for transfer learning in Bayesian optimization. arXiv preprint arXiv:190402642.

Wang D, Ohnishi K and Xu W. Multimodal haptic display for virtual reality: a survey. *IEEE Transactions on Industrial Electronics* 2020; 67(1): 610–623.

Wing A and Kristofferson A (1973) The timing of interresponse intervals. *Perception & Psychophysics* 13: 455–460. DOI: 10.3758/BF03205802.

Yun J, Lim Y, Kim KE, et al. (2015) Interactivity crafter: an interactive input-output transfer function design tool for interaction designers. *Archives of Design Research* 28: 21–37. DOI: 10.15187/adr.2015.08.28.3.21.

Zitzler E and Thiele L (1999) Multiobjective evolutionary algorithms: a comparative case study and the strength Pareto approach. *IEEE Transactions on Evolutionary Computation* 3(4): 257–271. DOI: 10.1109/4235.797969.

Zuluaga M, Krause A and Püschel Me-PAL (2016) An active learning approach to the multi-objective optimization problem. *Journal of Machine Learning Research* 17(104): 1–32. https://jmlr.org/papers/v17/15-047.html