Capturing Experts' Mental Models to Organize a Collection of Haptic Devices: Affordances Outweigh Attributes

Hasti Seifi
MPI for Intelligent Systems
Stuttgart, Germany
seifi@is.mpg.de

Michael Oppermann
University of British Columbia
Vancouver, Canada
opperman@cs.ubc.ca

Julia Bullard
University of British Columbia
Vancouver, Canada
julia.bullard@ubc.ca

Karon E. MacLean
University of British Columbia
Vancouver, Canada
maclean@cs.ubc.ca

Katherine J. Kuchenbecker
MPI for Intelligent Systems
Stuttgart, Germany
kjk@is.mpg.de

ABSTRACT
Humans rely on categories to mentally organize and understand sets of complex objects. One such set, haptic devices, has myriad technical attributes that affect user experience in complex ways. Seeking an effective navigation structure for a large online collection, we elicited expert mental categories for grounded force-feedback haptic devices: 18 experts (9 device creators, 9 interaction designers) reviewed, grouped, and described 75 devices according to their similarity in a custom card-sorting study. From the resulting quantitative and qualitative data, we identify prominent patterns of tagging versus binning, and we report 6 uber-attributes that the experts used to group the devices, favoring affordances over device specifications. Finally, we derive 7 device categories and 9 subcategories that reflect the imperfect yet semantic nature of the expert mental models. We visualize these device categories and similarities in the online haptic collection, and we offer insights for studying expert understanding of other human-centered technology.

Author Keywords
Haptics; haptic hardware collection; expert-sourced categorization; mental model; information visualization.

CCS Concepts
•**Human-centered computing** → **Human computer interaction (HCI);** *Haptic devices;* User studies;

INTRODUCTION
The recent surging interest in virtual reality, physical computing, and robotics has enticed a large number of researchers and practitioners from diverse fields to create new experiences or technical solutions with haptics. Selecting the right hardware is a particularly important design decision for these people.

CHI '20, April 25–30, 2020, Honolulu, HI, USA.
Copyright is held by the author/owner(s).
ACM ISBN 978-1-4503-6708-0/20/04.
http://dx.doi.org/10.1145/3313831.3376395

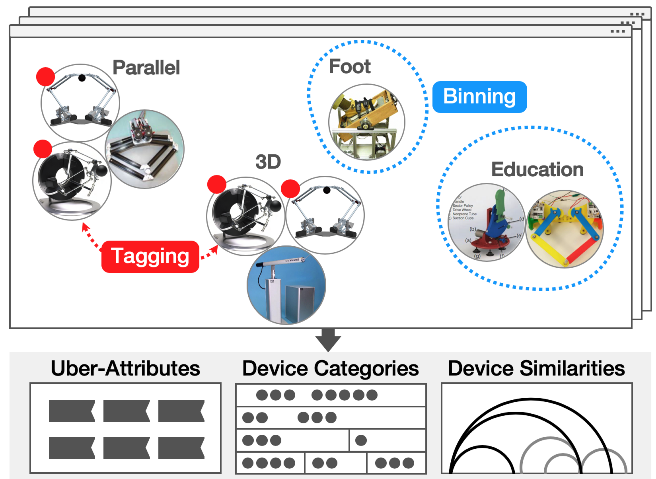


Figure 1: Expert mental categories for haptic devices. To depict the device similarities, experts sometimes created non-overlapping "bins" and sometimes "tagged" devices by copying them into multiple bins. We derived device uber-attributes, categories, and similarities from the responses of 18 experts.

One effective way to understand and find the appropriate haptic hardware is by reviewing the literature and/or device collections for relevant attributes and examples. Haptics surveys and books define about 60 commonly reported device attributes that come primarily from engineering specifications. A recently developed library, *Haptipedia*, provides an online catalogue of over 100 haptic devices painstakingly annotated with their attributes [33]. Users can browse the devices in different visualizations, see their detailed attribute specifications (e.g., degrees of freedom (DoF), peak force, actuator type(s), spatial resolution) and filter and/or compare them to select one for a particular project [33]. Both the haptics literature and *Haptipedia* intentionally provide a detailed view of each device to ensure consistent reporting in the research community and enable precise filtering and search.

Given enough time, a practitioner with sufficiently broad and deep experience can distill important information from these detailed specifications. Small insights about haptic hardware gradually develop into a complex mental map of how different attributes interact and also how they relate to a device's capabilities and affordances. Experts know the influential devices in the field and can detect their common variations as well as uncommon devices with new features. After reviewing a device's specifications, they can identify its key characteristics, evaluate its novelty, and place it in the right mental cluster(s) for later retrieval.

While useful for characterizing an individual device, these specifications may not be an effective way to browse or search a large collection. Each attribute describes one detail that may or may not contribute greatly to the overall experience of a particular device. Together, these attributes form a high-dimensional space that is not easy to visualize and takes time to search. Naturally, these challenges are amplified for those with less experience in the field. Such fine-grained detail may even obscure the overall capabilities of a device. For example, two similar devices might appear substantively different because they use electrical amplifiers with different maximum current output, an attribute that can easily be modified.

The goal of this work was to capture expert mental models for a major type of haptic hardware technology, grounded force-feedback (GFF) devices, in order to use the captured schemas to organize a large haptic collection such as *Haptipedia*. GFF haptic devices have a wide range of applications (e.g., education, surgery, entertainment), profit from a thirty-year history of development with many variations, and are well-catalogued in *Haptipedia*, which includes more than 54 attribute values per device. To support a wide range of users, we sought a data-driven project-independent organization for GFF devices.

To this end, we collected data around the following three questions: Q1. How do haptics experts perceive the similarity of GFF devices? Q2. What device attributes define GFF device categories? Q3. Do interaction designers (IxD) and device creators (DevD) categorize GFF devices differently?

Specifically, we derived similarity and semantic categories for 75 distinct GFF devices taken from *Haptipedia* by conducting a custom online card-sorting study. During the session, each of 18 experts (9 device creators and 9 interaction designers) first rated their familiarity with the devices. They then reviewed, grouped, and labeled the devices according to perceived similarity, and finally they described their groupings in an individual interview. To identify the aggregate device categories and similarities, we applied a clustering algorithm to the expert grouping data and complemented it with thematic analysis of the interview data.

Our results highlight two distinct expert approaches for describing the device similarities, suggest that device affordances are not dictated by the common definitions of attributes, and present a set of aggregate GFF device categories. To depict the intricate device relationships, some experts focused on a small number of key (to them) attributes for each device and created primarily non-overlapping groups (i.e., binning), while others perceived many attributes for each device and copied them into multiple groups (i.e., tagging). Most used a mix of binning and tagging. Focusing on affordances, the experts primarily used 6 high-level attributes (which we henceforth call uber-attributes), namely body-device interconnection, kinematic structure, motion range, versatility, unique engineering features, and complexity of building and using the device. Rather than following the literature's strict attribute definitions, the experts employed fuzzy interpretive definitions that consider the gestalt of the device's form and function, as well as the interaction of multiple attributes. Our clustering results support these qualitative findings through 7 categories and 9 subcategories that reflect the uber-attributes. We propose new visualizations for structuring GFF device collections with these semantic categories and similarities.

Organizing other evolving human-centered hardware (e.g., robotic hands, 3D printers, virtual reality gear) poses similar challenges: a high-dimensional attribute space with a mix of engineering and interaction design perspectives that contribute to the overall device affordances. To inform future studies, we reflect on our methods and present guidelines for capturing the expert mental models of such technologies. Our contributions include:

- Qualitative and quantitative synthesis of the expert mental organization for GFF haptic devices (Q1, Q2), linked to the expert's device or interaction design background (Q3)
- An interactive visualization of the GFF device categories and similarities (visualization of answers to Q1, Q2)
- Insights on the study design and interface for capturing the expert mental models of other high-dimensional interactive technologies (generalizable methods)

RELATED WORK

Below, we present existing literature on haptic device categories and attributes, algorithmic and user-centered approaches to making sense of complex high-dimensional item sets, and related theories from the categorization literature.

Conventions for Haptic Device Categories and Attributes

Some haptics review papers and books suggest primary categorizations for haptic hardware. One common scheme separates devices by the haptic sub-sense they target – kinesthetic versus tactile [14, 10, 11]. The literature further categorizes kinesthetic or force-feedback technology into grounded devices that are attached to a stable surface and ungrounded devices that are held by or mounted to the human body (e.g., wearables, exoskeletons) [10, 22]. While reviews acknowledge the diversity of GFF device designs, authors typically do not categorize them beyond this point [14, 10, 11, 22]. One reason for the lack of more finely grained device categories could be the multiplicity of expert opinions about the correct categorization. In this paper, we derive GFF categories based on an experiment with 18 experts from different backgrounds.

In contrast to these minimal categorizations, the haptics literature and device collections mention a plethora of GFF *attributes*, most of which focus on engineering specifications [12, 32]. *Haptipedia* expands that convention by visualizing a taxonomy of both engineering and interaction attributes for accessibility to experts of different communities [33]. However, *Haptipedia* does not prioritize specific attributes, so the resulting high-dimensional specification is time consuming to navigate. Our results identify attribute constellations (i.e., uber-attributes) that can define GFF similarity and categories.

Methods for Finding Categories in High-Dimensional Data

Previous studies propose a variety of algorithmic and user-centered methods for finding perceptual and semantic categories and similarities of items. Existing algorithmic approaches typically demand a numerical representation of items. Popular algorithms such as principal component analysis (PCA), t-distributed stochastic neighbor embedding (t-SNE), and multidimensional scaling (MDS) help define a low-dimensional representation of a high-dimensional data set of items [39, 24, 21]. Alternatively, clustering algorithms help identify distinct groups in a data set [19, 20]. The *Haptipedia* database is too small (about 100 devices), too varied (a mix of categorical and numerical attributes), and too sparse (many missing values) for these methods to be effective. In addition, *Haptipedia*'s attributes may not fully capture overall device affordances.

Similarly, analysis of the citation patterns or publication text with natural language processing (NLP) techniques might detect overall development trends but is not a reliable measure of device similarity [7, 38, 15]. A quick check of the device citations that are visualized in *Haptipedia* shows that the authors commonly cite unrelated influential devices and contrasting designs, and they sometimes fail to cite similar devices that were not known to them. An alternative approach for capturing item similarities is to rely on human judgment.

The psychophysics and human-computer interaction (HCI) literature offer a variety of methods for estimating perceptual and semantic distances of items through user studies. The most common approach, pairwise comparison, is prone to noise from local judgments and does not scale to large item sets [36, 17]. In the spatial arrangement method (SpAM), participants place items on a 2D canvas according to their similarity [17, 8]. This method is faster and allows for global judgments but is cognitively demanding and error prone when the underlying dimensions of the data are higher than two. Finally, sorting methods are typically employed with both perceptual and simple cognitive items [31, 25, 29]. In open card sorting, the participants can create any number of groups and label them [31]. This method can accommodate a large set of items (≥ 30) at the cost of a less granular (i.e., binary) distance matrix. To improve matrix resolution for large sets, Ternes proposed and validated a method for creating random subsets of the stimuli and aggregating the results from different participants into a distance matrix [35]. Our stimuli are complex devices that need to be reviewed to be understood. In contrast, these studies used perceptual (e.g., visual icons, vibration waveforms) or simple cognitive stimuli (e.g., website menu items) that can be judged in a matter of seconds. Therefore, we combined and adapted the above methods.

Categorization Literature and Methods

The field of haptics has not yet settled on holistic mid-level models of device categories. While haptic technology has a three-plus decade history, the participation and influence of interaction designers is relatively new and poses a particularly stark contrast to the engineering view of devices. Most attributes in *Haptipedia* are from engineering papers about the devices themselves [33]. These sources conform to academic genre conventions and are sufficient for populating an engineering database. Although *Haptipedia* was designed to meet the needs of diverse viewpoints, we believe the centrality of engineering papers to its shaping biases the platform to an engineer's granular mental model of haptic devices.

In categorization theory for resource description, specialists contrast structures of classification and categorization with the means by which such structures can be generated. Generally, a structure may either organize a universe of resources according to numerous qualities that have been prioritized to give a full and orderly description of each item (appropriate to a context-independent view of a field), or it can assign items to categories based on particularly salient qualities (appropriate to flexible and creative understandings) [18]. We refer to these two approaches as "tagging" and "binning", respectively. Generation of such structures may rely primarily on either the literature or on user vocabulary [3, 26]. Ideally, generating categories and labels from the literature requires minimal interpretation and produces a system that represents the resources consistent with the terms (and even the mental models) of their creators [1]. On the other hand, generating categories and labels from user vocabulary engages users who are not yet familiar with the resources and so are likely to describe them with less precision or jargon [2, 23].

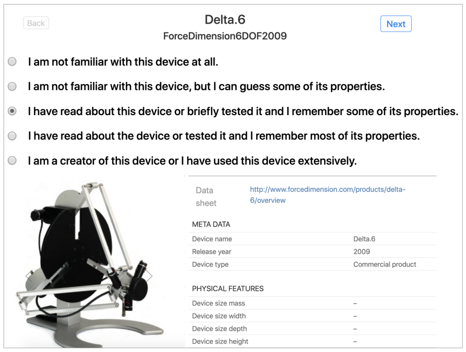
One prominent critique of the user-centered approach is the impossibility of representing a generic user [9, 16]. As with any system, haptic device collections have pluralities of users with diverse vocabularies and mental models. To investigate the descriptive architecture of *Haptipedia* and improve its utility across user groups, we adopted methods to identify categories and labels from the mental models of relevant experts [4, 34], and we analyzed the divergence or convergence of these models across expert user groups.

METHODS

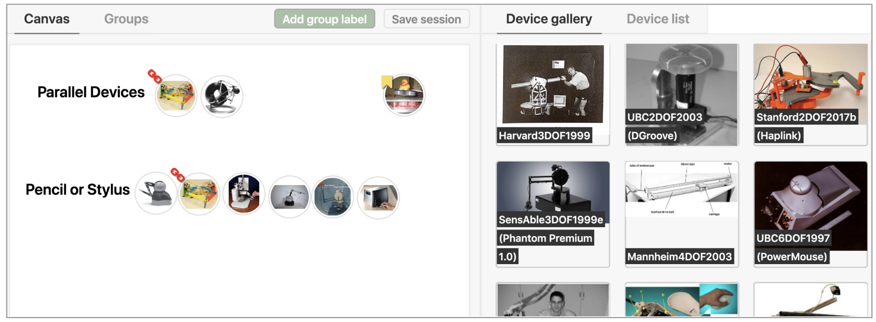
To empirically derive mental models of GFF devices, we selected 75 *Haptipedia* devices, divided them into random subsets, designed a custom card-sorting interface, and ran a two-hour-long online study session with 18 haptics experts that we recruited through email and snowball sampling.

Curating the Device Set

To reduce the duration and cognitive load of the experiment, we pruned the set of 105 GFF devices that were present in *Haptipedia* in April 2019 in the following ways: 1) When a device had several versions with minimal differences (e.g., the Stanford Haptic Paddle has five similar successors [30, 27]), we kept only the most representative and recognizable one (Hapkit 3.0 in this case). 2) For commercial devices that had multiple versions with different performance characteristics (e.g., Omega and Delta devices from Force Dimension Inc.), we kept only the simplest and the most complex devices in the set (Omega.3 and Delta.6 in this case). 3) We removed 7 more



(a) The familiarity rating interface shows one device per page, with a 5-point rating scale at the top, plus the device images and specifications from *Haptipedia*.



(b) Our custom card-sorting interface allows the user to drag and drop device icons on a 2D canvas, duplicate (red mark) and flag (yellow mark) the devices, and label them. The user can see the device images and specifications in the gallery or list view or open a detailed view with the device specifications. Hovering over a device on the canvas opens a pop-up view with five key attributes that the user selected at the start of the grouping task.

Figure 2: User interface for the study.

devices with poor-quality or unclear images based on a pilot study, leaving a final set of 75.

Estimating Device Subset Size

Through internal pilot studies and discussions, we determined 40 to be a reasonable trade-off between the number of devices that an expert participant can review in one session and the number of participants needed to determine similarity between all possible pairs of the 75 devices in our set. We initially aimed to recruit six participants from each of our two target populations. Based on the responses from our snowball sampling, we increased this number by 50% to get additional coverage for more device pairs. With 18 participants in our study (9 device creators, 9 interaction designers), all the pairwise similarities are covered from 2 to 8 times.

User Interface for the Study

To gather detailed recordings from the remote experts, we developed two custom web interfaces for rating device familiarity and sorting devices (Figure 2).

Familiarity rating interface – The interface displays each device in a separate page that replicates the *Haptipedia* specification page. Specifically, the interface shows the device ID, name, release year, multiple images, a link to the corresponding publication or datasheet, publication title and list of authors, and a table of device attribute values (Figure 2a). The user can rate each device on an integer scale from 1 (I am not familiar with this device at all) to 5 (I am a creator of this device or I have used this device extensively). The user can go back to change his or her previous ratings and is directed to the card-sorting interface after rating all 40 devices.

Custom card-sorting interface – An initial pop-up window invites the user to choose up to five attributes from a list of 54 taken from *Haptipedia*; this selection can be revised throughout the study. A two-column interface then displays an empty white canvas on the left side where the user can drag and drop device icons (Figure 2b). Three tabs to the right show a *gallery view* with the device thumbnails and names, a *list view* showing the user-selected five attributes for all devices, and a *detailed device view* showing the same specifications as in the familiarity rating phase. All devices placed on the canvas can be rearranged, and their positions are automatically saved. Devices on the canvas can be bookmarked or duplicated via the right-click context menu. Hovering over a device icon reveals its name, its five attributes, and a larger image; the user can also add text labels to their groups.

Procedure

During a two-hour study session, the expert participants responded to background questions, rated their familiarity with 40 GFF devices, grouped these devices based on similarity, and described their groupings in a follow-up interview. All sessions were conducted over Skype and were audio and video-recorded with participant consent. These steps are detailed below and depicted in Figure 3.

1) Background interview (15 minutes) – After explaining the study goal, we asked the expert about their years of experience, previous projects with GFF devices, and other haptic technologies they may have used. If needed, we asked them to self-identify as a device creator or interaction designer.

2) Familiarity rating (20-30 minutes) – Next, we sent them a link to the familiarity rating interface and asked them to share their screen and review the rating scale and the attribute specifications for each device. After we answered questions, they rated their familiarity with 40 devices in a random order.

3) Card sorting with duplication (40-50 minutes) – To introduce this task, the experimenter shared their screen and used a dummy set of five devices to demonstrate the interface features and grouping task. Specifically, we showed how one could choose five attributes in the initial pop-up view and explained the gallery, list, and detailed device views. We also demonstrated the duplication, flagging, and labeling functionalities. The experts were allowed to use any number of groups and to have single-item and/or "do not know" groups.

4) Interview about the groups (25-30 minutes) – At the end, we invited the expert to describe each group and their grouping criteria. In addition, we asked if all the devices fully belonged to their groups, if the spatial layout of the devices between and within the groups was meaningful, which of the *Haptipedia* specifications were useful for the grouping, how satisfied they were with their groups, and how their groups might change

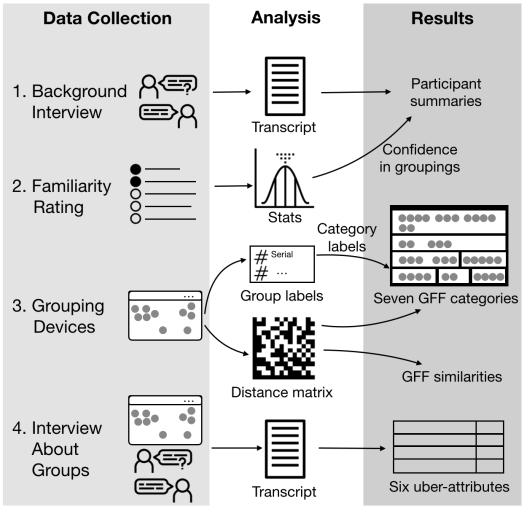


Figure 3: Overview of the data collected during our study as well as our analytical procedures and results.

in a future session. Finally, we asked them if and how the experiment format and interface limited them in depicting the device relationships they wanted to portray.

ANALYSIS

We analyzed different portions of the collected data using quantitative and qualitative methods.

Quantitative Analysis of Familiarity Ratings and Groups

Averaging familiarity ratings – We calculated the median, mean, and standard deviation of the device familiarity ratings for each expert and for the entire participant pool as an indication of their confidence in the groups they created.

Creating a device distance matrix from the grouping data – To obtain one best-fit categorization for all 75 GFF devices, we aggregated data from all of the experts into one 75-by-75 device distance matrix and analyzed it with a hierarchical clustering algorithm. We built the matrix by creating CSV files that included all the groups created by each expert and the device IDs assigned to each group. Specifically, when an expert had a hierarchical structure in their groups, we included separate groups for each level of the hierarchy in the CSV file and duplicated the device ID for each leaf node in all its parent groups. Also, we created a separate group for each of the devices in the "no group", "do not know", "miscellaneous", or similarly labeled groups to reflect a lack of similarity among these devices. Next, we calculated a device distance matrix for each expert using the formula presented in [29], summed the matrices for all the experts, and normalized the resulting matrix according to the number of observations per matrix cell. Finally, we applied hierarchical clustering to this distance matrix to obtain one categorization for all 75 devices.

Qualitative Analysis of Interviews and Group Labels

Thematic analysis of interviews – We used thematic analysis to identify meaningful patterns and themes in the participant descriptions of the device categories [6, 5]. One of the authors watched and transcribed all of the interviews, iteratively applied descriptive codes to the data (open coding, focused coding), wrote memos for connecting the codes and describing the data, and discussed the results with the team, which in turn led to merging some of the codes and creating new ones in the next round. We converged on themes around the expert perception of GFF similarities, the uber-attributes and their descriptions, and differences among device creators and interaction designers in grouping and describing the devices.

Aggregating group labels – To supplement the results from our cluster and interview analyses, we aggregated the expert labels in two ways. First, we labeled the categories from our clustering analysis by compiling the most frequent labels for the devices in that group (Figure 4). Second, we compiled all the expert labels for each uber-attribute to identify all the instances of an attribute (e.g., foot is an instance of the body-device interconnection) and their frequencies (see label frequencies and average percentages of groups in Table 2).

RESULTS

The 18 participants (4 female: 3 IxD, 1 DevD) were haptics specialists who had done at least one significant project (e.g., a Ph.D. thesis) with GFF devices. We did not require them to be previously familiar with all the devices but expected a deep understanding of GFF device attributes and performance so they could use the *Haptipedia* specifications to evaluate devices they had not previously seen or tested. Recruits included haptics industry practitioners (N = 2, average 9.5 years of experience), senior Ph.D. students (N = 3, avg. 4.5 years), postdocs (N = 5 , avg. 7.5 years), and faculty members (N = 8, avg. 18.5 years). They participated from the 9 following countries: Canada, France, Germany, Italy, South Korea, Spain, Turkey, the UK, and the USA. While some experts had worked on various aspects of a GFF device, half described their focus to be on building, modifying, or otherwise engineering GFF devices, while the other half primarily used existing GFF devices for perception studies and/or interaction design.

The distribution of the familiarity ratings peaked around the three inner statements of the rating scale with an overall median of 2 ("I am not familiar with this device, but I can guess some of its properties."). The five statements on the scale (Figure 2) received 99, 752, 366, 392, and 130 votes respectively. These ratings suggest that the experts could guess or identify the attributes of most presented devices to perform the grouping task. The "do not know" group allowed in the card-sorting interface helped remove potential noise from a lack of prior familiarity. Below, we present the expert perception of the device similarities, the uber-attributes and GFF categories, and similarities and differences between the device creators and interaction designers in defining the device relationships.

Q1. How do haptics experts perceive the similarity of grounded force-feedback devices?

The next three paragraphs describe how the experts *depicted* device similarities using the card-sorting interface, and the following two paragraphs provide evidence that these depictions adequately reflected their *perception* of the GFF devices.

Table 1: The experts varied in their approach for grouping devices and linking the groups. Overall, the taggers used more device copies and had fewer leftover devices, which we estimated as a sum of the number of items that were in single-item, miscellaneous, or "do not know" groups.

Expert	Grouping Approach	Linking Groups	Copies	Leftovers
$DevD_1$	Tagging	Hierarchy	28	0
$DevD_2$	Tagging	Proximity	17	1
$DevD_3$	Tagging	None	18	7
$DevD_4$	Binning	Hierarchy	5	7
$DevD_5$	Tagging	Proximity	11	2
$DevD_6$	Binning	Hierarchy	2	2
$DevD_7$	Binning	Proximity	0	0
$DevD_8$	Mix	Proximity	5	3
$DevD_9$	Binning	Proximity	1	7
IxD_1	Mix	Proximity	12	5
IxD_2	Mix	Proximity	8	7
IxD_3	Tagging	Prox., Hier.	27	3
IxD_4	Binning	Proximity	4	8
IxD_5	Binning	Proximity	3	6
IxD_6	Mix	None	6	4
IxD_7	Binning	Hierarchy	2	13
IxD_8	Mix	Prox., Hier.	5	4
IxD_9	Mix	Proximity	8	3

Grouping similar devices – The experts varied along a spectrum of using a tagging or a binning approach for grouping the devices. The taggers perceived many attributes for a device and included the device in many overlapping groups. *"This one has aspects of a lot of different kinds of my groupings here... it has high DoF, some kind of cable actuation, some kind of a serial linkage... so maybe I put this one in a bunch of different categories ($DevD_2$)."* The tagging approach is similar to a specification table (e.g., *Haptipedia*) where a device is described by all the attributes of interest. At the other end of the spectrum, the binners perceived a primary home category for each device. $DevD_7$ placed each device in only one group; in response to experimenter inquiry, this expert acknowledged this choice, saying *"Maybe I could have paid attention more to the secondary or third properties that are in common with the other groups ($DevD_7$)."* $DevD_9$ described little overlap as a sign for a good categorization. The majority of experts fell somewhere along the spectrum but were typically closer to the binning approach. The taggers used duplication more frequently than the binners (Table 1). We identified the approach for each expert based on their verbal descriptions and later cross-checked it with their number of duplicated devices.

Leftover devices – The experts described some devices as "leftovers" or "loners" and had difficulty grouping them. In some cases, the device was one of a kind. *"As you know we have many, many haptic devices. Most of them have some unique features, and some of them are very unique to that device only... only one or two... such devices are not easy to classify or group (IxD_4)."* In other cases, the expert had difficulty telling whether the device was truly unique or just hard to understand. To handle leftovers, some experts created a "do not know" group or a "miscellaneous" or "loose group" that could vanish in a future grouping: *"This is misc... it's got levitation, it's got handheld, I don't know... I don't know how to group this so I just put them all together (IxD_5)."* Finally, some created a single-item group for each unknown device. We compiled all the devices in the "do not know", "loose", and single-item groups as leftovers. Overall, the average number of leftover devices seems to be higher for the binners compared to the taggers (2.6 for taggers and 7.2 for binners, Table 1).

Linking similar groups – While two experts randomly placed the groups, the others used proximity or a hierarchy to relate similar groups (Table 1). All the experts perceived a relationship between at least two of their groups; the groups either had similar characteristics (e.g., joysticks and mice) or used the same grouping criteria (e.g., serial and parallel devices). However, the experts varied in whether and how they tried to depict this relationship among the groups. At one extreme, the groups were randomly placed on the canvas ($DevD_3$, IxD_6). Other experts created a hierarchy of groups. *"I divided the devices into 3d space, 2d space, 1d space ... and inside the groups I again divided the groupings into two different groups, one that has a big space... and the one that has very small environment ($DevD_1$)."*. The majority were in-between; they placed similar groups close to each other on the canvas and verbally described the boundaries of the related groups. We did not find a direct relationship between the tagging versus binning approaches and the use of space to show group relationships, i.e., both taggers and binners were likely to use a random layout, proximity, or a hierarchy.

Robustness of the groups – At the end of the task, the majority (16 of 18) were satisfied with their groups and anticipated creating similar groups if repeating the task with the same or a different device subset. The other two were taggers. They described their groups as a starting point and imagined redoing the task to identify other, more interesting, device attributes.

Impact of the card-sorting interface – The experts found the grouping interface effective for depicting device similarities and suggested additional features for facilitating the task. At the end of the session, we asked them if there was any device relationship they could not depict with the interface. 14 of 18 experts found the 2D canvas and duplication adequate: *"I don't feel constrained for grouping the devices with this way because you can do more groups and duplicate... then it's like 3, 4 dimensions you can use ($DevD_5$)."* When we asked them about the possible use of a third dimension or grouping physical devices in the real world, they responded that they might use the additional dimension for spacing out the unrelated groups but did not find it necessary. The other four experts requested a mechanism for linking the groups by drawing lines between or boundaries around the related groups ($DevD_4$, $DevD_9$, IxD_3, IxD_8). Furthermore, one tagger, IxD_3, wanted support for tracking the device copies to make sure that the group assignments were complete.

Given these results, we consider the groupings as a proxy for the experts' mental models in the following analysis.

Q2. What device attributes define GFF device categories?

We present *6 uber-attributes* from a thematic analysis of the interviews, and *7 device categories* from a quantitative clustering analysis of the expert groupings (Figure 3). While the categories are derived independently from the uber-attributes, they reflect combinations of the uber-attributes that are meaningful and present in GFF devices to date.

Six uber-attributes and their fuzzy definitions – Our analysis of the verbal descriptions and group labels indicate that

Table 2: Frequently used device attributes for the groupings (i.e., uber-attributes). The left column shows each uber-attribute, its instances (in the square brackets) and its "fuzzy" description and links to the device affordances. The right column shows the relevant group labels that were used by more than one expert and the label frequencies (in parentheses).

Uber-Attributes and Descriptions	Associated Group Labels
1. Body-device interconnection [Foot, fingers, pinch-grasp, power-grasp, joystick, pen or stylus, mouse]: describes the activated muscles and the range of motion afforded by the device. The experts commonly separated interfaces for the foot and those that sense and actuate individual fingers. The hand-held devices were categorized according to the grip posture and end-effector. Specifically, handles that require power-grasp and arm movement were separated from those with a pen or stylus, which offers more movement flexibility for the wrist. Joysticks and mice were described to have a small workspace for the wrist and friction feedback, respectively. **Average percentage of groups:** 29%, **Number of experts:** 16/18, ***Haptipedia* attributes:** Body part (5)	finger feedback (13), foot feedback (10), stylus or pen (9), joystick (6), tactile feedback(6), power-grasp (5), pinch-grasp (3), mouse (3), tool-mediated (2), paddle (2)
2. Kinematic structure [serial, parallel, wire] [Pantograph, Phantom, Delta, SPIDAR]: uses labels from the robotics literature to describe the configuration of the mechanical links, but the grouping was based on overall device affordances. The devices were sometimes grouped around the prototypical devices in that category such as the Pantograph [13]. The kinematic structure directly impacts the motion range, performance, and the complexity of building and programming the device. **Average percentage of groups:** 25%, **Number of experts:** 13/18, ***Haptipedia* attributes:** Type of links (3), Device structure (3)	parallel devices (11), serial devices (10), cable-driven devices (6), pantographs (4), x-y(-z) table (4), delta (2)
3. Motion range [1D or 1DoF, 2D or planar, 3D, 6DoF+] [small workspace, large workspace]: is a mix of degrees of freedom and workspace. One-DoF devices were often separated due to their simplicity of building/using and different design goals (e.g., demonstrating a new actuator). Planar and 3D devices were grouped around their translational motion and could have more than 2 or 3 DoF. Some experts used overall complexity to separate out the devices with 6+ DoF and/or the ones with very small or large workspace. **Average percentage of groups:** 22%, **Number of experts:** 13/18, ***Haptipedia* attributes:** Translational Workspace (12), Actuated DoF (10), User-reachable DoF (9), Sensed DoF (5), Rotational Workspace (4)	2d or planar (13), 1 dof (7), 6+ dof (5), 3d (5), 3 dof (4), large workspace (4), small workspace (3), translational/rotational (2), 1d (3), rotation (2), low dof (2)
4. Versatility [Specialized tool or application, generic, commercial]: defines the extent to which a device can be used for multiple purposes and is widely available to experts. Specialized devices are designed to simulate a tool (e.g., an endoscope) or satisfy the requirements of an application (e.g., education), and they can have properties that are unusual or undesirable for a typical haptic device (e.g., being flexible or low fidelity). Commercial devices are the most available and are commonly used for a variety of projects. **Average percentage of groups:** 14%, **Number of experts:** 13/18, ***Haptipedia* attributes:** Anticipated applications (5), Device type (3)	specialized tools or applications (7), education (5), surgical (4), general (3), medical (3), entertainment (3), rehabilitation (2), commercial (2)
5. Unique features [Magnetic levitation, brakes, steering-wheel, admittance-type]: captures rare engineering features of a device. Magnetic levitation devices were commonly separated due to their different operating principles. Few GFF devices use brakes instead of motors, a steering wheel for guiding user movements (i.e., co-bots), or are admittance type (i.e., measure force and output position). **Average percentage of groups:** 7%, **Number of experts:** 10/18, ***Haptipedia* attributes:** Force or torque (6), Actuator types (5)	magnetic levitation devices (8), brakes (2), admittance type (2),
6. Complexity of building and using: was a meta-attribute that guided the expert decisions. The expert grouping according to motion range, kinematic structure, and unique features reflected the complexity of building and using the device and its potential for being commercialized. Only one expert explicitly grouped the devices according to their complexity. **Average percentage of groups:** 2%, **Number of experts:** 3/18, ***Haptipedia* attributes:** -	ease of use/complexity (1)

experts agreed on what device attributes were important (Table 2), but did not follow strict attribute definitions to group devices. Instead, they used interpretation and a "fuzzy" definition to capture device gestalt.

For example, the experts used the robotic terms for device kinematic structure (e.g., serial, parallel), but found the strict mechanical definition inadequate. *"I classify them from the point of view of the end-effector... not from the kinematic point of view... because from the kinematic point of view there are many devices [such] as this Phantom [device name] with this parallelogram here but it is mainly a serial device... from the user point of view ($DevD_5$)."* As another example, the expert definition for planar and 3D devices did not have a one-to-one link to the device degrees of freedom (DoF). *"They [planar devices] allow the user to move just on the plane on the translational space... in some case [the designers] can add some other DoF [which] can be generally a torque [rotation] on the axis normal to the plane or even the small translational DoF normally to the plane ($DevD_6$)."*

Finally, experts wondered about grouping devices that modified another well-known device. *"The Reading [device ID] device is really... they took two Phantoms [a commercial device] that are not finger devices and put them together to make a grasping device... In my mind, I'm looking at this and say OK, the interface which is being actuated was used in this finger capacity but at the end of the day the actual underlying device behind it is more general ($DevD_4$)."* Some included this item in two groups according to affordances of base and modification; others treated the modified device as a whole. Table 2 summarizes the most frequent uber-attributes along with their "fuzzy" definitions, associated labels, and placement frequency.

Aggregate categories for the GFF devices – Our clustering analysis resulted in 7 categories and 9 subcategories for the 75

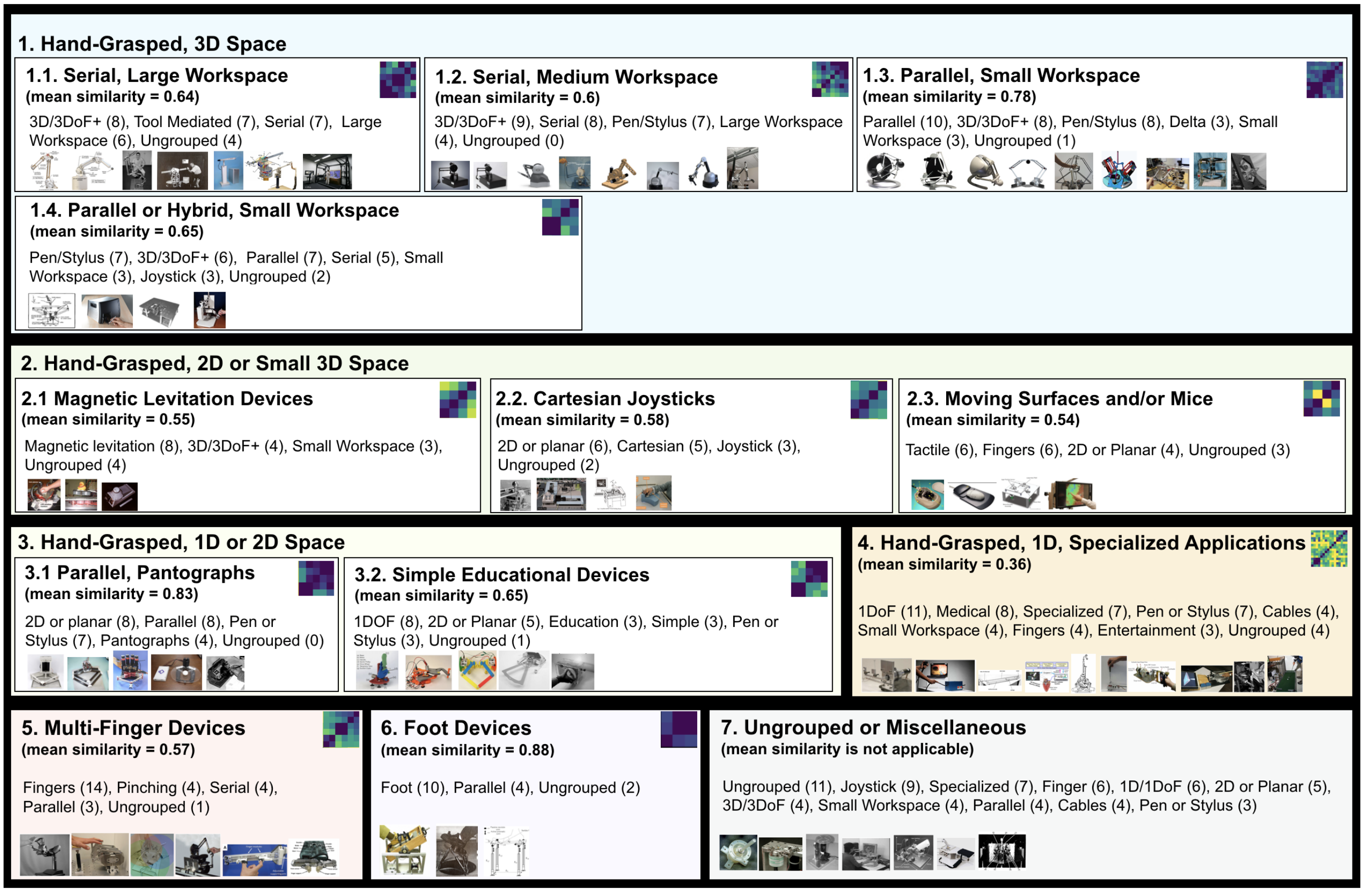


Figure 4: The seven (including 1 'miscellaneous') GFF categories that we derived for the 75 devices using a clustering analysis. We defined the category labels using the most frequent labels for the devices in each category. We calculated the mean similarity for each category by averaging the similarity between all the device pairs in the category. The devices with an average similarity of $\leq$0.3 with the other devices in their category are separated as Ungrouped or Miscellaneous (category 7). The colored squares show the distance matrix for the category. Dark colors indicate high device similarity.

GFF devices (Figure 4). As described above, we derived these categories by applying hierarchical clustering to the aggregate distance matrix from all the experts. To determine the appropriate number of clusters, we examined the clusters generated by cutting the hierarchy at each level and compared them with the uber-attributes reported above. Our proposed categories (Figure 4) reflect the uber-attributes and are consistent with the number of groups created by the experts (median = 14). We labeled each category based on the most frequent labels that the experts assigned to the devices in that category.

Our clustering analysis captures the binning model used by the majority of the experts, yet it does not fully capture the taggers' views. To assess the fitness of our approach, we also examined the parts of the device distance matrix that are not explained by our proposed categories. Figure 5 shows the residual distance matrix that was constructed by subtracting the distance matrix for the proposed categories from the distance matrix built from the expert groupings. The annotations provide examples of the unexplained similarities and show that these residual values are also useful. In the next section, we propose interactive visualizations around these GFF categories and similarities.

Q3. Do interaction designers and device creators categorize GFF devices differently?

To address this question, we separately analyzed the groupings created by the device creators and the interaction designers.

Similarities – The dichotomy of device creators and interaction designers could not fully predict the groups created by the experts. For example, some interaction designers mainly categorized devices by their kinematic structure (IxD_1, IxD_8), and some device creators mainly used interaction features ($DevD_1$, $DevD_6$). Furthermore, the average number of groups created according to each of the 6 uber-attributes was similar between the device creators and the interaction designers.

Differences – Overall, the device creators noted more individual features than the interaction designers. Due to the long engineering history of GFF devices, we anticipated that the device creators would have higher agreement with each other compared to the interaction designers. Surprisingly, the distance matrix for the device creators was notably less coherent than the one for the interaction designers. In Figure 6, the blue cells that denote similar devices are spread all over the device

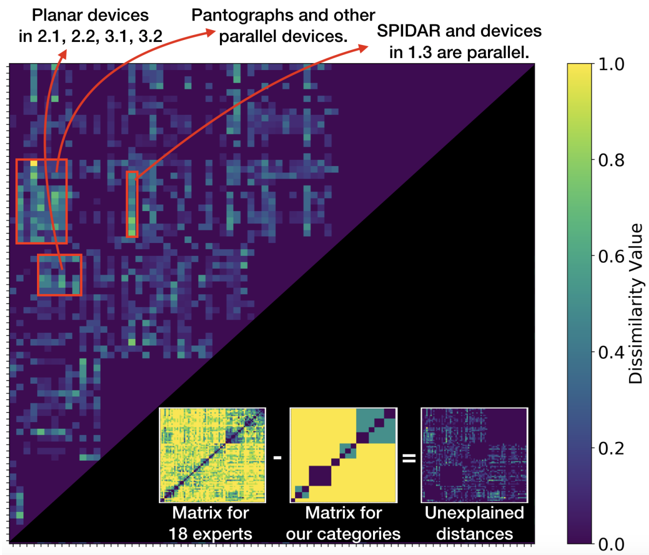


Figure 5: The residual distance matrix depicts device similarities that are not explained by our proposed categories. We mark sample areas where the residual values are $\geq$0.4 for a group of devices to show that these residuals are not noise. Only half of the symmetric matrix is displayed. In the other half, we show the miniaturized distance matrix for the 18 experts and the distance matrix from our proposed categories. The residual distance matrix is the result of subtracting these two matrices. All the matrices use the same sorting to place similar devices close to each other for clustering.

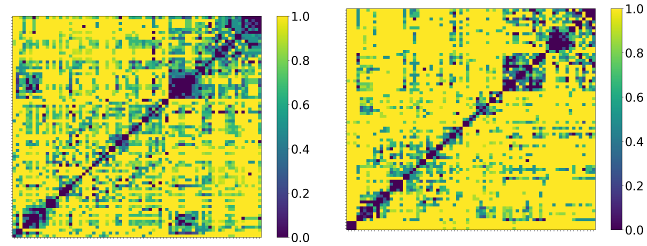
creators' matrix. One can note that the majority of the taggers were among the device creators (Table 1), i.e., they perceived many attributes for the devices, whereas the majority of the interaction designers considered the overall characteristics of the device. *"It [the categorization] was less about the properties and more about the gestalt of what the device was (IxD$_8$)."*

These results suggest that while the 6 uber-attributes were similarly important for both groups, the device creators tended to have a more nuanced view of the GFF similarities. Our analysis did not suggest any differences among the experts according to their years of experience.

INTERACTIVE VISUALIZATION OF THE RESULTS

We designed new visualizations to demonstrate how the expert-sourced categories and similarities can support browsing of a large collection like *Haptipedia* (Figure 7). To develop these visualizations, we created alternative prototypes and followed the literature guidelines [28].

A structure for the gallery view – At the start of this study, the *Haptipedia* homepage displayed all device thumbnails in a uniform grid ordered alphabetically or by release year. We suggest a new gallery structure based on the expert-sourced categories from Figure 4 (Figure 7a). In the new gallery view, users can change the thumbnail size dynamically or switch back to the original grid layout.



(a) Distance matrix for device creators

(b) Distance matrix for interaction designers

Figure 6: The distance matrix for interaction designers has a more coherent similarity pattern.

Arc diagram – We additionally visualize the device similarities in an interactive arc diagram [37]. An arc connects two devices if their similarity is above a user-adjustable threshold (0.7 by default) on a scale from 0 to 1. The user can hover over a node to highlight its connections and load additional device information (Figure 7b).

Device recommendations – On *Haptipedia*, the device detail page displays all the specifications, figures, videos, and CAD files for a device. We extended this page to recommend the three most similar devices ($\geq$ 0.7 similarity) according to the device distance matrix obtained in our study.

Device tags – For the detail page of each device, we compiled a list of tags from the experts' group labels. We show the most common tags to help the user detect important affordances of a device before checking its detailed specifications.

DISCUSSION

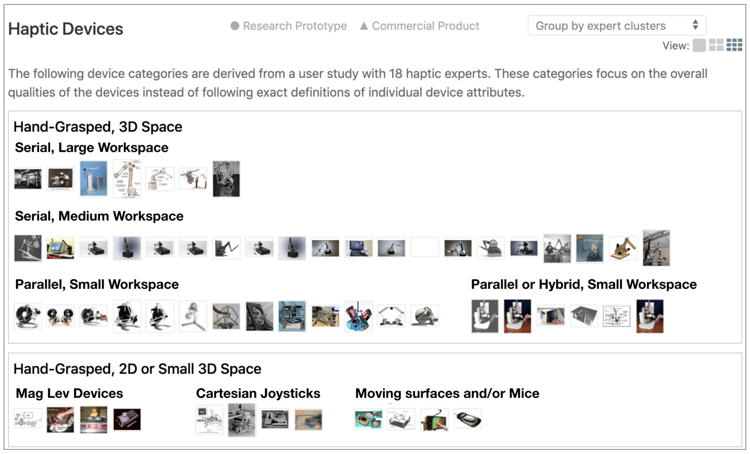
We discuss how our results contribute to the haptics literature and present guidelines for capturing the expert mental models of other complex interactive technologies.

Reflections on Haptic Categories

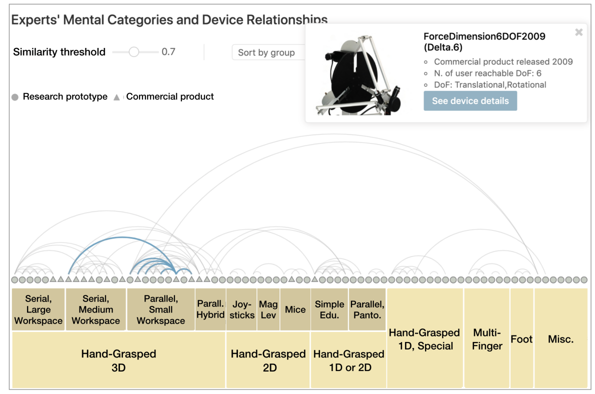
Nature of the categories – The expert descriptions are holistic rather than precise. In our study, the experts often used abstraction and interpretation to categorize the devices based on their gestalt characteristics and affordances. Three of the uber-attributes in our results (body-device interconnection, motion range, and kinematic structure) are presented in existing haptics books and surveys among many other attributes [14, 11]. However, their definitions differ from our results. For example, the literature defines motion range of a device according to its degrees of freedom (DoF) and workspace size. In contrast, in our results the motion range refers to the overall movement constraints of a device (e.g., 1D, 2D, 3D) which does not have a 1-to-1 mapping to the DoF values, e.g., a device with three translational DoFs is considered planar by an expert if the range of motion in the third dimension is small.

Use cases – We anticipate three uses for expert categories:

1) Efficient navigation: Experience designers could skim the GFF categories to decide whether their project requirements (e.g., 3D space, large motion) are supported by existing devices. They can quickly identify the most relevant category and narrow down their search to a fraction of the devices in a large collection such as *Haptipedia*.



(a) The restructured gallery view organizes devices into the identified categories. Note that this screenshot shows only a subset of the categories.



(b) The arc diagram connects the device pairs that have a similarity value above a user-defined threshold. The devices are sorted according to the GFF categories. Hovering over a mark highlights the connections for the corresponding device.

Figure 7: The gallery view and arc diagram that we created visualize the GFF device categories and similarities from the study.

2) Identifying a gap: Device creators can see the distribution of the devices in the categories and identify areas with little work in the literature.

3) Teaching haptics: Educators can use the categories and the uber-attributes to highlight prominent features of previously developed GFF devices.

Growing the collection – To add a new device, an expert can determine the primary and secondary categories for a new device and/or select the most similar devices from the collection. As more devices are added, we can re-evaluate the fit of the proposed categories. Rather than a rigid prescription, we anticipate the GFF categories to be a living system that is revised and extended based on future inventions and input from a larger group of experts.

Insights for Capturing an Expert Mental Model

Thinking more broadly, we conjecture that other interactive technologies (e.g., wearables, robots, 3D printers, virtual reality) with high-dimensional specifications, rapid innovation, and contributions from multiple communities of practice can benefit from capturing expert mental models. To inform future studies, we reflect on our methods.

Process – The familiarity ratings and the device subsets greatly increased the feasibility of the study. The familiarity rating task served as an effective warm-up for the grouping task, reduced its cognitive load, and contributed to the quality of the results. The size of the device subsets allowed everyone to complete the study within the two-hour time limit.

Card-sorting interface – The flexibility of our custom card-sorting interface allowed us to observe differences in how the experts used duplication (tagging vs. binning), the 2D space (proximity vs. hierarchy), and the labels. They found the interface adequate for grouping the devices but wanted more support for linking their groups (e.g., by drawing lines and/or circles) and tracking the device copies. During the sorting task, they mainly relied on the device images and rarely checked the specifications in detail. Thus, we suggest that future studies focus on compiling a comprehensive set of examples with effective media and publications, using pilot studies to test the necessity of any further information.

Analysis – Dividing the items into random subsets was an effective strategy for aggregate analysis but complicated analysis of the individual devices. While the expert groups and criteria were preserved at an aggregate level (i.e., the clustering results are consistent with the uber-attributes from the thematic data), the device subsets led to higher variation at lower levels of the clustering hierarchy (e.g., “specialized tools or applications” group, Figure 4). Furthermore, the differences in the device subsets across the experts made it challenging to analyze the leftover devices, as we were unsure whether a device would still be difficult to group if it was part of a different subset. An open question for future work is how to devise the subsets to mitigate these problems.

CONCLUSION

We present the expert mental organization for GFF haptic devices based on a custom card-sorting study with 18 experts. The resulting device categories and similarities contribute a descriptive layer that goes beyond attributes and specifications to be more holistic and interpretive. Our visualizations propose a new structure for large haptic device collections, and our guidelines can inform future studies in haptics and other HCI subfields. Finding good descriptions for existing technologies is the first step toward innovating new solutions and evaluating them. While our work focuses on describing the GFF design space as is, the descriptions can highlight areas with little work and therefore indirectly help researchers invent new solutions. Breaking out from existing ideas ultimately requires more than good descriptions and is a fruitful area for future work.

ACKNOWLEDGMENTS

We would like to thank NSERC Canada, UBC ICICS, and the Max Planck Society for providing facilities and funding for this work. We also thank our participants and reviewers for their scientific input.

REFERENCES

[1] Hanne Albrechtsen and Elin K Jacob. 1998. The dynamics of classification systems as boundary objects for cooperation in the electronic library. *Library Trends* 47, 2 (1998). http://cat.inist.fr/?aModele=afficheN&cpsidt=1768725

[2] Marcia J Bates. 1986. Subject access in online catalogs: A design model. *Journal of the American Society for Information Science* 37, 6 (1986), 357–376. DOI: http://dx.doi.org/10.1002/asi.4630370602 ISBN: 0002823118610.

[3] Clare Beghtol. 1986. Semantic validity: Concepts of warrant in bibliographic classification systems. *Library Resources & Technical Services* 30, April/June (1986), 109–125.

[4] Frank Branch, Theresa Arias, Jolene Kennah, Rebekah Phillips, Travis Windleharth, and Jin Ha Lee. 2017. Representing transmedia fictional worlds through ontology. *Journal of the Association for Information Science and Technology* 68, 12 (2017), 2771–2782. DOI: http://dx.doi.org/10.1002/asi.23886

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101.

[6] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. Thematic analysis. https://www.psych.auckland.ac.nz/en/about/our-research/research-groups/thematic-analysis.html. (2019). [Online; accessed 15-September-2019].

[7] Chaomei Chen. 2006. CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for information Science and Technology* 57, 3 (2006), 359–377.

[8] Çağatay Demiralp, Michael S Bernstein, and Jeffrey Heer. 2014. Learning perceptual kernels for visualization design. *IEEE Transactions on Visualization and Computer Graphics* 20, 12 (2014), 1933–1942.

[9] Melanie Feinberg. 2007. Hidden bias to responsible bias: an approach to information systems based on Haraway's situated knowledges. *Information Research* 12, 4 (2007). http://informationr.net/ir/12-4/colis07.html

[10] Blake Hannaford and Allison M. Okamura. 2016. Haptics. In *Springer Handbook of Robotics*. Springer, 1063–1084.

[11] Christian Hatzfeld and Thorsten A Kern. 2016. *Engineering haptic devices*. Springer.

[12] Vincent Hayward and Oliver R. Astley. 1996. Performance measures for haptic interfaces. In *Proceedings of the International Symposium on Robotics Research*, Vol. 7. MIT Press, 195–206.

[13] Vincent Hayward, Jehangir Choksi, Gonzalo Lanvin, and Christophe Ramstein. 1994. Design and Multi-Objective Optimization of a Linkage for a Haptic Interface. In *Advances in Robot Kinematics and Computational Geometry*, Jadran Lenarčič and Bahram Ravani (Eds.). Springer, Dordrecht, 359–368. DOI: http://dx.doi.org/10.1007/978-94-015-8348-0_36

[14] Vincent Hayward and Karon E. MacLean. 2007. Do it yourself haptics: Part I. *IEEE Robotics and Automation Magazine* 14, 4 (2007).

[15] Florian Heimerl, Qi Han, Steffen Koch, and Thomas Ertl. 2015. CiteRivers: Visual analytics of citation patterns. *IEEE transactions on visualization and computer graphics* 22, 1 (2015), 190–199.

[16] Birger Hjørland. 2013. User-based and cognitive approaches to knowledge organization. *Knowledge Organization* 40, 1 (2013), 11–27.

[17] Michael C Hout, Stephen D Goldinger, and Ryan W Ferguson. 2013. The versatility of SpAM: A fast, efficient, spatial method of data collection for multidimensional scaling. *Journal of Experimental Psychology: General* 142, 1 (2013), 256.

[18] Elin K Jacob. 2004. Classification and categorization: A difference that makes a difference. *Library Trends* 52, 3 (2004), 515–540. Publisher: Graduate School of Library and Information Science. University of Illinois at Urbana-Champaign.

[19] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. 1999. Data clustering: a review. *ACM Computing Surveys (CSUR)* 31, 3 (1999), 264–323.

[20] Stephen C Johnson. 1967. Hierarchical clustering schemes. *Psychometrika* 32, 3 (1967), 241–254.

[21] Joseph B Kruskal and Myron Wish. 1978. *Multidimensional Scaling*. Vol. 11. Sage.

[22] Katherine J. Kuchenbecker. 2018. Haptics and haptic interfaces. In *Encyclopedia of Robotics*. Springer. https://link.springer.com/referenceworkentry/10.1007%2F978-3-642-41610-1_19-1

[23] María J. López-Huertas. 1997. Thesaurus structure design: a conceptual approach for improved interaction. *Journal of Documentation* 53, 2 (1997), 139–177.

[24] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research* 9, Nov (2008), 2579–2605.

[25] Karon MacLean and Mario Enriquez. 2003. Perceptual design of haptic icons. In *Proceedings of EuroHaptics*. 351–363.

[26] Jens-Erik Mai. 2005. Analysis in indexing: document and domain centered approaches. *Information Processing & Management* 41, 3 (May 2005), 599–611. DOI:http://dx.doi.org/10.1016/j.ipm.2003.12.004

[27] Melisa Orta Martinez, Tania K Morimoto, Annalisa T Taylor, Aaron C Barron, JD Akzl Pultorak, Jeanny Wang, Agnes Calasanz-Kaiser, Richard Lee Davis, Paulo Blikstein, and Allison M Okamura. 2016. 3-D printed haptic devices for educational applications. In *Proceedings of the IEEE Haptics Symposium*. 126–133. DOI:http://dx.doi.org/10.1109/HAPTICS.2016.7463166

[28] Tamara Munzner. 2014. *Visualization Analysis and Design*. AK Peters.

[29] Anshul Vikram Pandey, Josua Krause, Cristian Felix, Jeremy Boy, and Enrico Bertini. 2016. Towards Understanding Human Similarity Perception in the Analysis of Large Sets of Scatter Plots. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3659–3669. DOI: http://dx.doi.org/10.1145/2858036.2858155

[30] Christopher Richard, Allison M Okamura, and Mark R Cutkosky. 1997. Getting a feel for dynamics: Using haptic interface kits for teaching dynamics and controls. In *IEEE ASME IMECE Annual Symposium on Haptic Interfaces*. 15–21.

[31] Gordon Rugg and Peter McGeorge. 1997. The sorting techniques: a tutorial paper on card sorts, picture sorts and item sorts. *Expert Systems* 14, 2 (1997), 80–93.

[32] Evren Samur. 2012. *Performance metrics for haptic interfaces*. Springer Science & Business Media.

[33] Hasti Seifi, Farimah Fazlollahi, Michael Oppermann, John Andrew Sastrillo, Jessica Ip, Ashutosh Agrawal, Gunhyuk Park, Katherine J. Kuchenbecker, and Karon E. MacLean. 2019. Haptipedia: Accelerating Haptic Device Discovery to Support Interaction & Engineering Design. In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems (CHI '19)*. ACM, New York, NY, USA, Article 558, 12 pages. DOI: http://dx.doi.org/10.1145/3290605.3300788

[34] Smiraglia, Richard P. 2015. *Domain Analysis for Knowledge Organization*. Chandos Publishing. DOI: http://dx.doi.org/10.1016/C2014-0-02303-9

[35] David Richard Ternes. 2007. *Building large sets of haptic icons: Rhythm as a design parameter, and between-subjects MDS for evaluation*. Ph.D. Dissertation. University of British Columbia.

[36] L Tsogo, MH Masson, and Anne Bardot. 2000. Multidimensional scaling methods for many-object sets: A review. *Multivariate Behavioral Research* 35, 3 (2000), 307–319.

[37] Martin Wattenberg. 2002. Arc diagrams: Visualizing structure in strings. In *IEEE Symposium on Information Visualization (InfoVis)*. 110–116.

[38] Jevin Darwin West. 2010. *Eigenfactor: ranking and mapping scientific knowledge*. University of Washington.

[39] Svante Wold, Kim Esbensen, and Paul Geladi. 1987. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems* 2, 1-3 (1987), 37–52.