Crafting Diversity in Radiology Image Stack Scrolling: Control and Annotations

Louise Oram[1], Karon Maclean[1], Philippe Kruchten[2], Bruce Forster[3]
University of British Columbia
Departments of computer science[1], electrical and computer engineering[2], and radiology[3]
{olouise, maclean}@cs.ubc.ca pbk@ece.ubc.ca bruce.forster@vch.ca

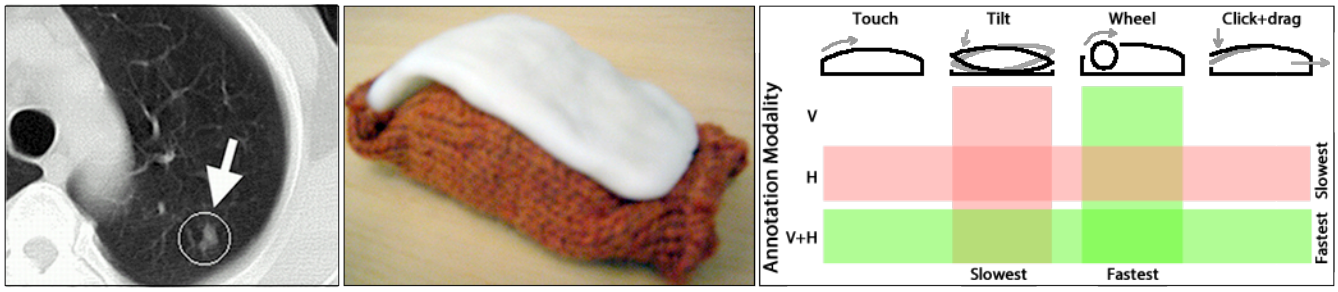

Figure 1: Left: early sketch (Polymorph™, knitted wool, polymer stuffing with many potential interaction movements). Middle: image of lungs with annotated potential nodule, from Amato S.G. et al. Radiology 225: 685-692, 2002. Right: Shows the relative scrolling speed for a task for the four scrolling techniques and three annotation modalities (visual (V), Haptic (H), and both (V+H))

ABSTRACT
To make a single diagnosis, today's radiologists must examine thousands of images; yet little effort has been put into refining this time-consuming, repetitive task. Meanwhile, automatic or radiologist-generated annotations may impact how radiologists navigate image stacks as they review lesions of interest. Observation and/or interviews of 19 radiologists revealed that stack scrolling dominated the resulting task examples. We iteratively crafted and obtained radiologist feedback for a variety of prototypes, then evaluated their scrolling and annotation-review support for lay users. With a simplified stack seeded with correct / incorrect annotations, we compared the effect of four *scrolling techniques* (traditional scrollwheel and click-and-drag, plus sliding-touch, and tilt rate control) and *visual vs. haptic annotation cues* on scrolling dynamics, detection accuracy and subjective factors. Scrollwheel was fastest overall, and combined visual / haptic annotation cues sped target-finding relative to either modality alone. We share insights on integrating our findings into radiologist practice.

Author Keywords
Input device; prototyping; scrolling; haptic; tactile; mouse; computer-aided diagnosis (CAD); radiology.

DIS 2014, June 21–25, 2014, Vancouver, BC, Canada.

ACM 978-1-4503-2902-6/14/06...$15.00.
http://dx.doi.org/10.1145/2598510.2598585

ACM Classification Keywords
H.5.m. Information interfaces and presentation (e.g., HCI).

INTRODUCTION
To utilize the detailed information provided by today's high-resolution image capture technologies, such as Magnetic Resonance Imaging (MRI) and Computed Tomography (CT), radiologists must examine ever-larger image sets. It is not uncommon for multi-trauma CT scans or coronary CT angiograms to have data sets of 4000 images [5]. Diagnosis entails a complex, time-pressured visual search task. Target conspicuity, background clutter and other attentional factors can influence the radiologist's ability to detect anomalies [5], and radiologists are put at substantial risk of repetitive strain injury [13].

Radiology images (e.g. Figure 1) are currently viewed as single 2D slices [5, 6], arranged in a *stack* through which the user scrolls depthwise. The main interaction tool, a scrollwheel mouse, has not changed since ~1995; but in contrast to x-ray images, image stacks are continuous media streams. Efficient perusal demands fluid, controllable interaction akin to video scrubbing [19], as has been demonstrated with a haptic scrollwheel [24]. A conventional mouse's restricted input mobility (x-y hand movement and finger-level scrollwheel movement) has limited ability to support this.

Meanwhile, the daunting scope of the image-viewing task makes it a candidate for semi-automation, e.g. computer-aided detection (CAD) of anomalies in images [10]. Such algorithms are tuned to find all real anomalies (true positives), at the cost of substantial rates of false positives, which radiologists must then discriminate. While accuracy and number of events varies, at 5-7 seconds to re-evaluate a

CAD-identified nodule [22] there is clearly a cost to potential time and accuracy gains. Similar issues exist for annotations from other sources, e.g. other radiologists, in redundant procedures and peer reviews or training reviews.

Stack annotation can affect detection accuracy [2, 10]. Of concern is *context bias* (radiologists' diagnostic sensitivity depends on expected prevalence of a given anomaly [12]); and *automation bias* (CAD misses particular cancer types). Learned dependency can also lead the user to miss.

How might alternative annotation *presentation* affect bias? CAD data is now presented as visual highlights, which may be more likely than another modality to influence what the radiologist sees at perceptual, attentional and strategic levels. Integrated with care, haptic highlights might also avoid an identified risk of degrading the decision process through simple sensory overload [18]: visual systems are highly tasked, and the hospital environment is noisy.

Neither *adding* a specialized device nor compromising familiar mouse functions are likely to be accepted. Radiologists heavily use other manual tools (keyboard, dictaphone), and oscillate swiftly between GUI pointing and stack strolling. Proprietary data systems enforce device standards. The x-y mouse is best for pointing [13], and its ease of use and familiarity make it favored relative to alternative input devices in this setting (e.g. [23]).

APPROACH

With a focus on ergonomic stressors and opportunities for aligning tasks with interface advances, we observed and interviewed radiologists and analyzed their manual tasks. We identified a design space, then brainstormed and prototyped alternative input models with different mobility affordances (ways in which the user's hand can move when interacting with the device). We reviewed these prototypes with domain experts for improved support of key tasks.

We next analyzed 19 radiologists' work via observation and/or interviews, leading to mouse augmentations which we hypothesized could support (a) more efficient image scrolling (with more fluid interaction) and (b) attentionally improve annotation display (using the haptic modality). After a round of qualitative feedback and iteration on our prototypes and the interactive techniques they support, we examined the impact of interaction and display on detection rates in a controlled, study with lay users as proxies for hard-to-access radiologists. Here, we used an abstracted detection task whose representative nature we confirmed with experts. Finally, we integrate these findings with radiologist feedback into recommendations for next steps.

We contribute:

- A set of verified task examples (Table 1) that capture the most important manual radiology image interactions.
- Prototypes representing a set of novel scrolling inputs.
- An abstracted task suitable for screening scrolling- and annotation-type candidates on lay users.
- Quantitative data on detection accuracy and subjective reactions to scrolling type and annotation modality (summarized in right of Figure 1).
- A proposal for how haptics can increase effectiveness by minimizing bias in review of annotated data.

BACKGROUND

The Radiologist' Work Environment and Constraints

To view images, radiologists use two or three high-resolution LCD monitors, a mouse for stack navigation and GUI navigation, and keyboard and dictaphone to transcribe diagnoses. Data is provided via a *Picture Archiving and Communication System (PACS):* workstation, software, and network for image storage and retrieval according to industry standards. PACS are sourced by health authorities as major capital investments from a small number of medical imaging vendors, and have proprietary elements.

Viewing Images by Scrolling

Scrolling is integral to image review. CT image consumption is faster in a (serial) stack than in parallel as multiply-visible tiles, probably due to eased perception of 3D structures [20]. Radiologists scroll at different speeds, stop, and reverse to compare or examine locations. They are trained to review specific anatomical structures, and make successive passes focusing on each in turn.

PACS workstations typically support two scrolling techniques: scrollwheel and click-&-drag. Both employ *position control* (scrolling distance is proportional to the position of mouse or angle traversed by scrollwheel). Atkins et al. [6] compared scrollwheel and click-&-drag techniques to a jogwheel (a *rate control* device: scrolling *rate* is proportional to input position), and found that most radiologists preferred the more familiar position control even though some were faster with rate control. Relative movement rates were generally fastest for the wheel/click-&-drag combination, slowest with wheel alone, and in between for jogwheel [6]. Sherbondy et al. used a tablet and stylus for scrolling, and found that position was faster than rate control for finding a target in a CT stack [23].

When the user knows where the target is, scrolling is modeled by Fitts' Law for techniques including rate control, scrollwheel, and wheel with acceleration [14]. But with visual target search (e.g., reviewing an annotated stack), scrolling time depends on distance to target [4].

Beyond the Mouse, and Direct-Touch Sensing

Multi-touch sensing has become a ubiquitous manual control. In an early mouse example, Hinkley et al. explored touch sensing near the scrollwheel, and found it a useful discrete alternative, e.g. tapping to page up/down [15]. Villar et al. found that multi-touch in five form factors could extend control degrees of freedom and support different input modes [26], mitigating the need to switch between devices. They advised locating touch-sensed areas in easy reach of one hand posture, and cuing their location.

Flying mice and other tracked devices can be lifted off the table surface. Direct mapping to a 3D space makes them easy to learn [29]; but fatigue in maintaining cursor persistence make them unsuitable for radiology interaction.

A pen and tablet solution showed decreased times relative to a mouse for the radiology task of outlining a *region of interest* [9]. However, switching between different devices may hinder radiologists' workflow. Direct-touch reduces the need for device switching, but creates occlusion [27] and fatigue from unsupported hands [28].

Other desk-supported concepts have diversified interaction. The "Rockin' Mouse" adds a degree of freedom; faster than a normal mouse in 3D, scrolling was not studied [7]. Many other control movements could be used with a mouse-like device, but have not been explored in the radiology setting.

Haptic Feedback in Support of Scrolling

Akamatsu et al. found that for a pointing task with a mouse, tactile feedback (pin pushing into fingerpad when on target) was quickest, and no feedback slowest for final positioning times [1]. Levesque et al. saw variable friction feedback speed up target selection on a touch screen [17]. For a different mobile device, tilting to scroll was augmented with a vibrotactile (VT) buzz at transition to the next item on the list. VT feedback lowered task completion time, and position was faster than rate control [21].

These results suggest that haptic feedback on possible targets will give modest performance gains (and not losses), even if the system does not know where the user is heading. The prevalence of detents on a mouse in a radiology setting indicates radiologists may be receptive to this.

Computer Aided Detection (CAD)

Most CAD research focuses on validating that CAD information, provided as visual image annotations, improves radiologist detection sensitivity and/or speed [10]. However, annotations overlaid on the stack affect what radiologists see. Even when biased towards finding everything, CAD misses 20% [10], and leads to automation bias. Radiologists attending to annotated areas are more likely to miss artifacts not found by the CAD. Alberdi et al. found a lower detection rate for users given CAD information in comparison to those who were not; here, the largest difference was seen in cancers not found by CAD. They hypothesized a bias effect, where users calibrate to the expected prevalence of cancers and expected proportion of cancers missed by CAD in the current data set [2]. Additionally, a criticism of many CAD studies is that the data sets used contain an unrealistic proportion of cancers, and radiologists know this [2]. We have not seen studies that modified how CAD annotations are displayed; yet this may help mitigate the detection bias that CAD produces.

Table 2: Aggregate results of task example questionnaire (mean / SD of all responses, after averaging by task example)

Topic	M	SD	Response Range of Mean
Importance	4.55	0.29	very – extremely
Frequency	4.08	0.61	very – extremely
Difficulty	2.52	0.21	not very – somewhat
Support	3.22	0.26	somewhat -- very

Table 1: Task Examples, and the mean responses for each from the Likert scale questionnaire (1=not at all, 5=extremely).

1. Identifying or finding a specific piece of anatomy The radiologist looks for an object or area of interest in one anatomical plane, looking through several slices to find and properly identify it. If unsure, or things are unusual, then s/he may look at the area in another plane (or several other planes if they are available). Can cross-reference a point between different planes, to see the location in other planes. Additionally, they may adjust the window/level to get better contrast between the object and its surrounds.				*4. Comparing two images (old and new)* The goal is to look for interval change: differences between the sets of image. Do new objects appear, have old objects enlarged? The radiologist brings up both sets of diagnostic images and looks at the same plane and area in each image side by side. They scroll back and forth in each set of images, comparing the areas of interest (can link the two images so they scroll together, but the slices may not land at exactly the same spots). They may re-measure objects that were found in the first diagnostic to see if they have changed in size.			
Importance: 4.7	Frequency: 4.4	Difficulty: 2.4	Support: 3.4	Importance: 4.9	Frequency: 4.6	Difficulty: 2.5	Support: 3.0
2. Defining the edge / size of something The radiologist may want to know the size of an object, or if it is encroaching on the area of other anatomy. Window/level may be used to get better contrast of the object to its surrounds. After looking at the object in several planes, they choose a specific image, or multiple images, to outline, circle, or measure the diameter of the object.				*5. Identifying the makeup of something* The radiologist may want to know what something abnormal is composed of. They look at the item in several planes, and see the attenuation of the item. They may adjust the window/level to get the best contrast with the surrounds, or to see colour differences within the object. To know the density of the item from the imaging they can select part or all of it and see the density number.			
4.2	3.8	2.5	3.6	4.3	3.7	2.2	3.3
3. Tracking / connecting objects The radiologist follows a part of the anatomy through several slices to check for abnormalities. The radiologist moves back and forth through the image slices while watching the area of interest. If they feel they have missed something, or loose track of the object they may slow down and watch more carefully for a subset of the image slices. This is repeated as many times as needed for different anatomical parts, usually by organ system but sometimes by area (such as in the brain).				*6. Getting a second opinion* If the radiologist is unsure of something, less familiar with it, or finds something unusual, they may ask the opinion of another radiologist. Another option is to look up papers on the topic to help confirm the diagnosis or learn about more nuanced aspects they cannot remember off the top of their head.			
4.8	4.8	2.8	3.1	4.4	3.2	2.7	2.9

Rubin et al. [22] saw CAD had a significantly higher sensitivity to finding lesions missed by a first human reader, in comparison to a second human reader. However, this comparison posits unrealistically that the user of the CAD annotations would accept all true positives and reject all false positive CAD detections.

In low-dose CT images, a CAD scheme detected 83 percent of lung nodule cancers (in stacks with on average 1-2 nodules), with 5.8 false positives per scan [10]. Another scheme (run on different scans, containing some potentially more subtle cancers) detected 80 percent, with 2.7 false positives per scan. In our experiment we therefore manipulate annotation display assuming a detection ratio of 80% to align with current CAD performance.

TASK ANALYSIS

We analyzed physical interactive elements of the radiologists' workflow in a two-stage process.

Task Example Creation

We informally observed and interviewed 12 radiologists within a variety of work settings, over 1-3 sessions in blocks of around 30 minutes. They had many suggestions for PACS software improvements (out of our scope) as well as for physical image interaction. We noticed some disparities between observation and self-report in activities (e.g. percentage of scrollwheel vs. click-&-drag use), which may point to subjective importance. We captured this domain-expert input in a set of task examples (Tables 1-2).

Task Example Validation

To verify that our task examples faithfully represented the most important elements of radiology image interaction, we took them more formally to ten radiologists (8 male; including 3 from the original 12), recruited by email from hospital administration and word of mouth. Our participants had experienced a variety of work settings in a mid-sized North American city (e.g. academic hospital radiology department, private lab, city hospital emergency room) and professional roles (interventional radiology, diagnostics, neuroradiology). Career experience ranged from 0-31 years (avg. 12.7). All were familiar with touch devices and owned and/or often used one. Six reported ergonomic issues from extended PACS use, including shoulder pain, eyestrain, and repetitive use of the scrollwheel.

In ~15-minute workplace sessions, volunteers read the task examples, answered a questionnaire, and were interviewed.

Questionnaire: Four 5-point Likert scale questions, repeated for each task example, asked how *important*, *frequent*, *difficult*, and *well supported* that example was. Tables 1-2 detail and summarize this quantitative data.

Interview: We voice-recorded discussion of a set of open-ended questions, asking them to identify:

- Missing tasks they find important, frequent, or difficult.
- What is well and poorly supported by PACS they have used (many had experience with different PACS brands).
- Mouse interactions they found tiring or repetitive.
- Any issues with repetitive strain injuries.

Importance

Each of the six tasks was rated as very or extremely important by at least 8/10 participants. P1 summarized that *"they all seem extremely important to me"*. Participants either said no important tasks had been overlooked (2), or gave examples of very specialized or specific tasks (8).

Frequency and Repetitiveness

Tasks 1-5 were labeled very or extremely frequent by 6-10 participants, with Tasks 3 and 4 rated highest. Task 6 was less frequent (but of high importance). We note that area of specialization is likely to play a role in these assessments.

Participants verbally identified the most repetitive task as scrolling: *"When you are looking at [a] CT that has 350 images in it, and you are looking at every image, that takes a lot of scrolling up and down"* [P7]. P6 noted that scrolling is very mouse-intensive and therefore a way to end up with an injury, then suggested having a way *"to scroll through a large amount of data set with minimal hand motion"*.

On scrolling and speed, P2 commented: *"I use scroll-wheel way more often than the drag stuff."* When asked if it was hard to go fast enough, P2 replied, *"Yeah… But it's too hard to go slow enough with the click-&-drag… something in between, so if you had a dual function?"*

Difficulty

Generally, task difficulty arose from diagnostic complexity, e.g. *"when there is complex anatomy, complex disease processes"* [P7]; or ambiguity: *"to know what is normal, or what is in the range of normal, or where it starts to be abnormal or pathologic"* [P8].

Discussion resolved the potential ambiguity of responses indicating both low-difficulty and low-support (s): radiologists have figured out ways to perform necessary tasks, accommodating non-optimal support, and no longer find them difficult; but still wish for better support.

Device Interaction

Participants suggested device improvements, with many relating to functional specificity: *"I would prefer to have more buttons, with less functionality per button"* [P5]. P1 mentioned speed interfering with functional mapping; a rapid clicker, double clicking was mapped to a function he did not usually mean to invoke. P2 had even considered adding his own accessory to the PACS workstation: *"At one point I was considering getting a gaming accessory pad… so you could mouse or move over to the pad"*.

Summary:

Scrolling is an essential and frequent part of radiology interaction: validated tasks 1-5 require scrolling, and radiologists confirmed their frequency and importance. Discussion confirmed both centrality of scrolling in routine activities, and the need and possibility for improvement of image interaction via device and/or software.

Figure 3: Prototypes. From left: *Touch*, *Tilt,* W*heel / Click+Drag*

The most crucial challenges in current scrolling technology identified were: reducing repetitive movements, more easily varying the speed of scrolling, and more functionality mapped to the device.

Many PACS aspects are somewhat personalizable: e.g. *"I can't imagine using PACS without having my custom way of looking at it"* [P4]. Radiologists were generally receptive to the idea that the input device could be more personalized, for instance with pre-set scrolling speeds.

DESIGN SPACE AND PROTOTYPES

We identified a scrolling-input-mobility design space to explore for possible improvements to identified challenges, which includes current baseline methods and adds diversity in input control (Table 3). We populated this input-modality design space with three exploratory prototypes (one representing two design dimensions), constructed by modifying existing mice (Figure 3).

(All): VT Annotation Display*:* A pager motor generated a vibrotactile buzz in all prototypes, perceptible in all hand positions observed. In pilots, we arrived at a 200ms (pager motor supplied 3V, ~200 Hz) cue at the annotated image, with a 1-image advance for fast scrolling (<10 images/sec), so that the majority of the buzz was felt on the image.

Touch: An Apple Magic Mouse™ (curved multi-touch surface) was modified by adhering a pager motor to the underside of the touch surface, adding ~1cm to its height. The multi-touch surface was of interest because custom gestures (the requested extra 'buttons') could be mapped onto it, but this ability was not tested here.

Tilt: A curved top surface with profile matching the Magic Mouse was 3D-printed and a pager motor placed on its underside. Springs at either end achieved stable centering of a curved bottom surface. An accelerometer, read by an Arduino Uno, detected its tilt angle which, configured for rate control, controlled rate of movement through the stack.

Wheel / Click-&-Drag: To provide baseline comparisons at a comparable level of prototype polish, we replaced the top of a traditional mouse with a 3D-printed surface identical to *Tilt's* but with a slot for the scrollwheel, and attached a pager motor to the underside of this surface.

(All): Connectivity: Prototypes communicated with a custom image-viewing program (written in C++) on a control laptop via an Arduino microprocessor (Uno or Micro). This program commanded a vibration via USB-2, and received input from existing x-y, scrollwheel and multi-touch mouse channels and *Tilt's* accelerometer.

Table 3: Scrolling input mobility design space and study factor

Type	***Prototype Name:*** **Motion description**
Wheel scrolling	*Wheel (Baseline):* Traditional scrollwheel mouse functionality.
Dragging of whole mouse	*Click & Drag (Baseline):* Traditional dragging and pointing functionality (combinable with *Wheel*).
Sliding on mouse surface	*Touch:* Sensing of a finger sliding on a smooth surface, as in current mobile touch screens; multi-touch can map gestures to specialized functions.
Rocking	*Tilt:* Maps forward/back rocking to scrolling up/down; also uses rate rather than position control

EXPERIMENT: DETECTION PERFORMANCE

We conducted a study to compare usability of our four prototypes (representing points in the scrolling input design space), as well as the impact of both scrolling method and annotation modality on the human viewer's detection performance in the face of imperfect annotation (false positives and true positives). In constructing annotation modality conditions, we aimed to hold perceptual salience constant.

We hypothesized that:

1. *Haptic+Visual* will afford faster detection together.
2. Annotation modality will not affect error rates.
3. *Wheel* and *Touch* will afford similar accuracy, because they both clutch through the images.
4. *Click-&-Drag* and *Tilt* will be fastest in approaching an area, but perform poorly in finer adjustments.

We also sought subjective input that would elucidate ergonomic factors, but did not test them directly.

Table 4: Annotation Modality study factor

Visual	Dashed green circle around the target (Figure 4), visible when the image itself appeared.
Haptic	A 200ms buzz (pager motor) as user approached annotated image. The buzz started one image before the anomaly image if the user was scrolling rapidly (majority of the buzz felt on image), and on the image if the user was scrolling slowly.
Combined	Combination of *Visual* and *Haptic*

Radiology Proxy: Abstracted Task for Lay Subjects

It was infeasible to access professional radiologists for repetitive performance tests, so we created a version of the stack-scrolling task in which non-expert performance would indicate qualitative, first-pass trends of trained radiologist performance at a level which could guide a next round of development. From the observation of radiologists it is impossible to determine their error rates, but qualitatively the scrolling passes performed by lay subjects appear similar to radiologists: users scroll at their preferred speed, with slow downs on areas of interest. We removed the need for radiology knowledge, with visual complexity that could be quickly learned; and reduced the task to one of *scrolling* and *signal detection* to focus our findings on the *relative* roles played by annotation modality and interaction type in performance and usability. Validation took two forms: pre-test expert task assessment; and adjustment of difficulty to match lay with published expert performance (see also Results).

With the help of our expert radiologist co-author (~20 years in practice, leads a university team) we dissected the radiology stack-scrolling task to its most basic element: *searching for something visually specific among similar objects / distractors.* The expert confirmed the final task abstraction's suitability for a first lay assessment. We have since discussed the validity of the task as a facsimile for their work with 2 more radiologists; both felt it was valid and sensible to use this type of visual search task with lay users to initially predict their own expert performance.

The task emulated scrolling through a lung CT image stack while looking for potentially cancerous nodules (as in Task Example 1). In real stacks, lung images exhibit a bronchial tree (bronchi tubes feed into smaller "bronchioles"). The alveoli sacks at the ends of this tree can look similar to, but have slightly different characteristics, than cancerous nodules (Figure 1 shows a potential nodule).

Our more learnable version entailed small greyscale rectangles placed throughout a 60-image stack with a uniform black field (256x256px, rendered at ~8cm/side on a laptop screen; Figure 4). The task was to find the *true target* (a perfect square of 5-10px, medium grey), of which there would be exactly one per stack, among 50 *distractor noise* rectangles (sizes randomly chosen between 4-12 px/side with aspect ratio 40% larger and smaller than the *true target*, and either lighter or darker grey); then click one of four buttons on a numeric keypad, indicating the quadrant where it was seen. In pilots, we adjusted task difficulty (varying distractor shape, frequency and contrast) to the settings described here. These supported a ~10% error rate performance. This was slightly better than the 20-30% documented for radiologists [16], deemed appropriate given a cognitively easier task, and conscientious pilots subjects.

For tractable analysis, we constrained target stack index to four values (20, 30, 40, 50). In pilots (confirmed in study results), participants did not appear to learn target locations, i.e. they continued to make errors at a uniform rate. The order the stacks where presented in was shuffled randomly.

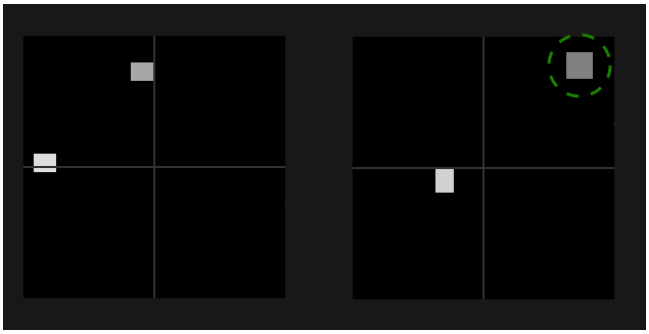

Figure 4. Study 2: Images from abstracted task. Right image shows visual target annotation.

For each combination of scrolling input and feedback, participants saw a learning example plus 20 test stacks. These comprised 5 with a highlight at each of the four stack indices: four where the *true target* (perfect square) was present and highlighted (16 total), and one where a *distractor target* (aspect ratio 16% smaller / larger than *true target)* was present / highlighted, and the *true target* was located later in stack (4 total). This ratio (80%/20%) matches current published CAD performance [10].

The *distractor target* (closer to a perfect square than the distractor rectangles) always appeared before the *true target*, with an advance randomly selected between image 5 in the stack, and 5 images before the *true target*.

Experiment Design

A Latin square produced 4 orderings for scrolling input (*Touch, Tilt, Wheel, Click-&-Drag*) and 3 for annotation modality (*Visual, Haptic, Combined*). The latter were blocked on scrolling input to minimize device switching, for a total of 12 orderings of the 12 condition combinations.

Metrics were: *task completion time,* measured from start of scrolling to keypad click, and *accuracy* (did they indicate both correct image and correct quadrant of the *true target*, or not). 12 lay participants (1/ordering) were recruited via campus posters and emails, and compensated with $15.

Protocol

An experiment session took up to 1.5 hours. Participants were seated in a quiet room, asked to complete a demographic questionnaire, and instructed to complete the task quickly and accurately. They then carried out target-search on 12 sets of 20 image stacks, while listening to white noise through noise cancelling headphones to mitigate any auditory vibration feedback as the background noise at a hospital likely would.

Between each set of 20 tasks, participants were surveyed on their scrolling accuracy, frustration, confidence, and attentional needs; how easy it was to notice the annotations, and how helpful they were. Upon completion, they were asked to rank their preference of device and annotation.

RESULTS

We replaced one subject who did not understand the task. 9/12 participants had error rates <20%, and 3 in the range of 30-50%. 5 did not complete the final set (scroll input) due to a time restriction of 1.5 hours.

Task Completion Time

Completion time exhibited a broad and heavily skewed distribution: targets were placed at different distances from the start point, and participants varied in the care they took, with trials tending to go long if they did not find the square in the first pass. Conventional models like ANOVA and GLM (general linear modeling) require normality. ANOVA can also only treat whether or not they got the trial correct as a variable, whereas a Cox model can use this factor to censor the data. Further, completion time and accuracy were not fully independent since with enough time a correct target could always be found in our abstracted task.

The legitimacy of some trials comes into question if we do not censor the times by whether or not the subject found the target correctly (censoring is a statistical situation wherein only partial information is known about a data item, e.g. that up to time x, the user had not completed the task [11]). Fast responses where an apathetic participant chose a non-target image would skew results, but censoring essentially removes this data by only taking it as partial information.

We therefore used a proportional hazards model (Cox regression [3, 8]) for completion time, which assumes that if given more time users could answer correctly. Non-error trials have all information needed, and error trials partial (they did not find it up to a certain time). Thus the model censors time by whether or not they got the trial correct:

$$T_{comp} = P + S + A + Ti + Th + N^2 + Th*A + S*A \quad [Eq.\ 1]$$

where model parameters are **P**articipant, **S**crolling input condition, **A**nnotation modality condition, **T**arget **i**ndex, **T**arget **h**ighlighted, and trial **N**umber.

The hazard rate from the Cox regression can be plotted as a survival curve, which shows the likelihood a task would be completed at a certain time (Figure 5). Most tasks are completed within 40 seconds, and it is apparent that combined annotations (haptic and visual) make it more likely that the task is completed earlier.

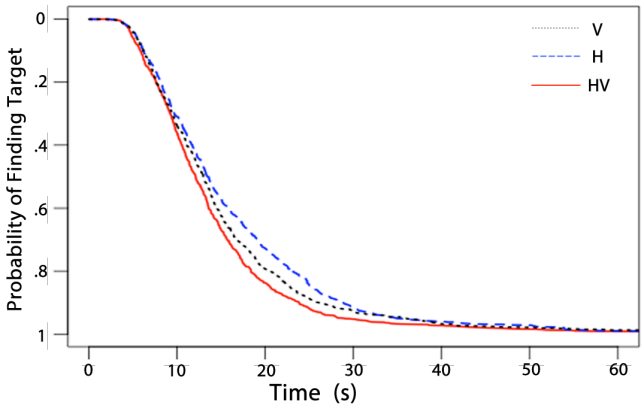

Figure 5. Survival likelihood (Cox regression) vs. projected completion time. Visual, Haptic, and combined.

The Cox regression delivered the following statistically significant results for completion time ($p<.05$):

- *S: Wheel* scrolling was faster than the baseline *Click-&-Drag* (Z=2.48, p=0.013), and *Tilt* was slower than *Click-&-Drag* (Z=-2.47, p=0.014).
- *A: Combined* annotations were faster than *Visual* (Z=2.59, p=0.0096), and *Haptic* was slower than *Visual* (Z=-2.82, p=0.0048).
- *Th*: For false alarms (false-target highlighted) trials were slower (Z=-2.27, p=0.023).
- *Th* x *A*: For false alarms, *Combined* annotation was slower than for true positives (Z=-3.30, p=0.00096).

Regarding individual variance and task validation,

- *P:* Participants varied widely in completion time (T_{comp} SD: 12913ms). E.g. P10 was faster than P1 (Z=5.29, p<0.0001), and P8 slower than P1 (Z=-2.82, p=0.0048).
- $N^{2:}$ Trial number reaching significance (Z=-5.48, p<0.0001) indicates T_{comp} fit a t^2 distribution: earlier trials were slower, middle trials fastest, and later trials slower again. This suggests learning followed by boredom.
- *Ti:* The shortest target index distances (20) had faster trials than the two longest (40; Z=-5.46, p<0.0001) and (50: Z=-8.11, p<0.0001).

Approach analysis: To get a sense of the motion dynamics as a function of scrolling method and annotation modality, we defined T_{app} as the period of time a user proceeded forward measured from the trial's start to a first direction reversal. To reduce noise, trajectories shorter than 10 images were removed from this analysis.

Analyzed with a GLM, *Haptic* had slower approaches than *Visual* (t=2.46, p<0.0014). *Wheel, Touch* and *Tilt* had slower approaches than *Click-&-Drag* (t=3.88, 5.02, 8.30, all p<0.0001), but there was less data for *Click-&-Drag* following short-trajectory removal; we conjecture that its motion was jerkier.

Accuracy

To analyze trial accuracy (a binomial distribution of right/wrong) we used a GLM with the same parameters as for T_{comp} (Eq. 1). Significant results ($p<.05$) are as follows.

- *P:* Participants varied widely in accuracy (average 17% error rate, min 2%, max 55%). E.g. P11 had significantly fewer errors than P1 (Z=8.90, p<0.0001).
- N^2 (Z=-2.28, p=0.023): there is likely a learning then boredom effect (consistently with T_{com}).

Questionnaire Results

Ten participants preferred *Combined* annotation modalities; one preferred *Haptic*, and one *Visual*.

Likert scale responses were analyzed using a proportional odds logistic regression, accounting for scale ordering along with experiment factors (scrolling input, annotation modality). This indicated ($p<0.05$):

- *Wheel* was deemed the most accurate device ($Z=-4.79$, $p<0.0001$) with *Touch* the runner-up ($Z=-1.97$, $p=0.0493$). Users had more confidence in *Wheel* ($Z=-4.45$, $p<0.0001$) and felt they required less attention ($Z=3.03$, $p=0.0025$).
- *Wheel* was rated the least frustrating ($Z=4.79$, $p<0.0001$), with *Touch* 2nd least frustrating ($Z= 3.07$, $p=0.0021$).
- *Combined* (haptic and visual) annotation was most noticeable ($Z=3.27$, $p=0.0011$), as well as most helpful ($Z=2.34$ $p=0.0191$).

There were generally more positive responses for *Wheel* in comparison to the other scrolling inputs. *Combined* annotation received higher ratings than either alone.

Figure 6. Modified prototype. Images show some of the movements /interactions that can be performed with it.

FOLLOW UP WITH RADIOLOGISTS

Modified prototype: We combined the best performing features found in evaluation, to create a prototype that worked as a conventional mouse with the added abilities to (a) touch-scroll, and (b) tilt backwards to access rate control scrolling (with a switch to control direction). We began with a Microsoft Wedge mouse, added a rocking base (Polymorph™), and sensed tilt with a potentiometer (an accelerometer would confound translation with tilt). An Arduino relayed mouse signals, and a tactor was installed underneath the touch surface (Figure 6).

Method: We took the modified prototype to the workplaces of 3 radiologists (2 previously interviewed, 1 new), demonstrated its movement and haptic feedback (in the context of our abstracted test task) and informally discussed its potential usefulness with them.

Highlights: Given existing customizability of PACS setups, radiologists reiterated their receptivity to the idea of a personalizable mouse. Their preferred speed of scrolling is highly personal and varies depending on the type of stack, so the rate control could have several preset speeds (e.g., controlled via a slider on the side of the mouse). *"The goal should be to customize the mouse… in a perfect world once, and then to not have to fool with it after that"* [P1].

P2, an emergency radiologist, stated *"The way that I look at a large data set study is I fly through it once and get a birds eye view… I want to exclude any immediately life-threatening conditions"*. Further, in a diagnosis he needed to access multiple stacks, and felt the haptic feedback would help re-orient him upon switching. He also indicated aesthetic appreciation: *"Ooh the haptic feedback I love*".

Sometimes radiologists need to re-read other radiologist's image sets, e.g. with trainees, to ensure quality of care. The haptic annotations could help speed this review: *"You mark up the image in a peer review, and then I go through it to check whoevers work, and I can find immediately what they were looking at – that is valuable"* [P1].

P3 noted there might be *"a temptation to go really fast"*, and worried that the haptic cues would encourage this, resulting in missing anomalies. However, he further mused that it would be useful for very large data sets, such as the lungs. He generally felt that *"You have a problem and you are trying to find a solution to the problem, and here we have a potential solution to many problems"*.

Unsurprising was some mention of potential integration issues: *"Many of our workflows are so refined over the years… because we are just used to going through data sets in a certain way"* [P2].

DISCUSSION

Value of Haptic Feedback

Hyp. 1: Combined (Haptic+Visual) will afford faster detection than either alone - Accepted

Results from our non-expert, abstracted study suggests that for a task similar to image-stack scrolling, multimodal annotations (*combined*) are most noticeable, most helpful, and improved detection times. *Haptic* annotation was slower than *Visual*. We can infer performance relative to no annotation from the cases where the true target was not annotated (distractor target highlighted); having just haptic or visual annotations showed no differences in speed, but multimodal annotation slowed the user relative to when the target was correctly highlighted. Overall, using both types of annotation together was still fastest.

A possible explanation, in addition to simple cuing redundancy, is that each modality provided slightly different speed-related benefits. *Visual* annotations told the user exactly where in the image the target was; *Haptic* may allow faster motor responses. *Combined* annotations benefited from both.

The timing of the *Haptic* annotations here was devised to match *Visual* as closely as possible. However, the haptic annotation could be given earlier, allowing the user to slow down pre-emptively and search more carefully through the next few images. In our abstracted task, the context of the perfect square does not matter, but a radiologist might tweak the timing of the feedback to help view and understand the context of the potential anomaly.

An important emerging source of annotations is other radiologists. Trainees must have their diagnoses checked by

a board-certified radiologist, and can be required to provide annotations in key images for the 2nd radiologist to review. Also, there is widespread pressure within diagnostic imaging [25] and medicine as a whole to increase peer review activities as a quality assurance measure.

Effect on Decisions

Hyp. 2: There will be no effect of annotation modality on error rates - Accepted

Lay participants had the same accuracy for trials annotated correctly and incorrectly, as there was no significance found for the target highlighted (Th) term. However, they made slower detections in trials containing false positives (for completion time the Th term did reach significance). Annotation modality did not affect the lay users' ability to make a decision, as it did not impact accuracy. Overall, having *Combined* annotation speeds them up and they show a preference towards it, in comparison to *Visual* alone.

Scrolling Type

Hyp 3: Wheel and Touch will afford similar accuracy, because they both clutch through the images – Partially supported

No device emerged as the most accurate, but subjectively *Wheel* was felt to be the most accurate, with *Touch* next.

Hyp. 4: Click-&-Drag and Tilt will be fastest in approach, but perform poorly in finer adjustments – Partially supported

Click-&-Drag was fastest for approaching an area. *Tilt* was slowest in task completion time, so appears to be weaker for finer adjustments for the implementation we tested; however it was also the least familiar to users, and had the least refined implementation (the others being minor revisions of commercial products).

The traditional and most familiar (*Scroll*) supported the fastest task completion times by lay users, and was preferred. In most metrics, sliding-touch scrolling (*Touch)* was ranked second. However, *Click-&-Drag* supported faster initial approach (even if it was to the wrong area). This, along with familiarity, is likely why *Scroll* and *Click-&-Drag* work well together in the radiology environment.

Combining scrolling input methods

Radiologists were interested in reducing the repetitive movements associated with the mouse that occur often with scrolling (e.g. clutching with the mouse wheel). This encourages us to continue to refine our *Tilt* implementation and test it following longer learning, as its rate control approach while continuing to support other functionality. Multi-touch would also allow many more potential improvements in radiology image interaction, via the mapping of gestures to different tools that could reduce the need for modal interaction with PACS workstations.

Validity of Abstracted Task + Lay Users

How are our lay subjects like/unlike radiologists? Our lay users' error rates varied wildly, and we would expect radiologists to show more homogeneity because of their training, and studies indicate consistent error rates of 20-30% [16]. Our lay users varied more, ranging from less than 2% error to 55% error (average 17% error rate). We would expect professionals to have fewer slower outliers, and less inter-person variability.

We must therefore take care in generalizing to radiologists. We saw little effect in error rate, but there may be effects for radiologists. Future validation includes a compacted study to look at the effect of annotation modality on errors for trained radiologists.

CONCLUSION

We analyzed radiologists' work and found a high prevalence of scrolling, poorly supported by traditional scrolling input devices with negative ergonomic and productivity implications that can be expected to grow in the future. The radiologists we interviewed were highly interested in seeing improvements to their working tools, and some had experimented with this on their own. Creation and iteration of the various interaction ideas (that eventually became the 3 prototypes used in the experiment) was central to our discussion with the radiologists.

In our study comparing the four scrolling input motions, the scrollwheel emerged as the fastest (with lay users), and confirms our early observation that augmentation of established tools should be explored rather than replacement. However, novel input methods (e.g. a tilt or rocking motion associated with rate control scrolling) were disadvantaged by their newness and less optimized implementation. Because the scrollwheel has known ergonomic issues from excessive repetitive movement, alternate methods still need to be explored.

In the emerging practice of incorporating annotations (from CAD or other radiologists) into radiologists' workflow, we have shown that multimodal cues are a promising approach, showing task speedup without error degradation, for a task abstracted to non-experts. Radiologists are heavily visually loaded, and may benefit from information provided through a less loaded modality, even when redundant.

Because of the many factors (economic, training) making the radiologist work environment highly change-resistant, introduction of new input devices must be undertaken with care. Our participatory approach, which revealed enthusiasm for change, seems a promising avenue for this.

Future Work

In addition to addressing the improvements and caveats mentioned throughout this paper, some development directions have emerged.

In our study we required our lay users to use each scrolling input type separately. A more realistic scenario is for the user to access them all in a seamless manner: some methods are better for scanning the stack, others for fine adjustments, and yet others for other GUI uses. A next step

will entail further integrated prototype refinement in collaboration with expert users, followed by its use in a similar task, in order to compare its performance to the individual scrolling types.

The effectiveness of the haptic feedback could be increased by personalization, to accommodate individual differences in reaction times. One could create a program that logs the reaction to the haptic cue, and adjusts the timing of the feedback based on this. Other types of haptic cues might improve attentionally on the simple buzz we used, such as a vibration fading in upon approaching a region of interest.

ACKNOWLEDGMENTS

This research was partially funded by NSERC, and done with the cooperation of McKesson Medical Imaging. We thank the radiologists who volunteered for this research.

REFERENCES

[1] Akamatsu, M. et al. A comparison of tactile, auditory, and visual feedback in a pointing task using a mouse-type device. *Ergonomics*. 38 (1995), 816–827.

[2] Alberdi, E. et al. Effects of incorrect computer-aided detection (CAD) output on human decision-making in mammography. *Academic Radiology*. 11 (2004), 909–918.

[3] Andersen, P.K. and Gill, R.D. Cox's Regression Model for Counting Processes: A Large Sample Study. *The Annals of Statistics*. 10, 4 (1982), 1100–1120.

[4] Andersen, T.H. A simple movement time model for scrolling. *Extended abstracts CHI*. (2005), 1180-1183.

[5] Andriole, K.P. et al. Optimizing analysis, visualization, and navigation of large image data sets: one 5000-section CT scan can ruin your whole day. *Radiology*. 259, 2 (2011), 346–362.

[6] Atkins, M.S. et al. Evaluating Interaction Techniques for Stack Mode Viewing. *J. of Digital Imaging*. 22 (2009), 369–382.

[7] Balakrishnan, R. et al. The Rockin' Mouse : Integral 3D Manipulation on a Plane. In *Proc. of CHI* (1997), 311–318.

[8] Cox, D.R. Regression Models and Life-Tables. *Journal of the Royal Statistical Society*. 34 (1972), 187–220.

[9] Dix, A. et al. Human-Computer Interaction in Radiotherapy Target Volume Delineation: A Prospective, Multi-institutional Comparison of User Input Devices. *J. Digital Imaging.* 24 (2010), 794–803.

[10] Doi, K. Current status and future potential of computer-aided diagnosis in medical imaging. *The British J. radiology*. 78 (2005), S3–S19.

[11] Efron, B. The Efficiency of Cox's Likelihood Function for Censored Data. *J. American Statistical Association*. 72, 359 (1977), 557–565.

[12] Egglin, T.K.P. Context Bias, A Problem in Diagnostic Radiology. *J. of the American Medical Association*. 276, 21 (1996), 1752-1755.

[13] Goyal, N. et al. Ergonomics in radiology. *Clinical Radiology*. 64, 2 (2009), 119–126.

[14] Hinckley, K. et al. Quantitative analysis of scrolling techniques. *Proc. of CHI*. 4 (2002), 65-72.

[15] Hinckley, K. and Sinclair, M. Touch-sensing input devices. *Proc. of CHI.* (1999), 223–230.

[16] Kundel, H.L. et al. Visual scanning, pattern recognition and decision-making in pulmonary nodule detection. *Investigative Radiology*. 13 (1978), 175–181.

[17] Lévesque, V. et al. Enhancing physicality in touch interaction with programmable friction. *Proc. of CHI*. 31 (2011), 2481–2490.

[18] Manning, D.J. et al. Perception research in medical imaging. *The British J. of radiology*. 78, 932 (2005), 683–5.

[19] Matejka, J. et al. Swifter: improved online video scrubbing. *Proc. of CHI.* (2013), 1159-1168.

[20] Mathie, A.G. and Strickland, N.H. Interpretation of CT scans with PACS image display in stack mode. *Radiology*. 203 (1997), 207–209.

[21] Oakley, I. et al. Tilt and Feel : Scrolling with Vibrotactile Display. *EuroHaptics.* (2004), 316-323.

[22] Rubin, G.D. et al. Pulmonary nodules on multi-detector row CT scans: performance comparison of radiologists and computer-aided detection. *Radiology*. 234 (2005), 274–283.

[23] Sherbondy, A.J. et al. Alternative Input Devices for Efficient Navigation of Large CT Angiography Data Sets. *Radiology*. 234 (2005), 391–398.

[24] Snibbe, S.S. et al. Haptic techniques for media control. *Proc. of UIST*. (2001), 199-208.

[25] Swanson, J.O. et al. Optimizing peer review: A year of experience after instituting a real-time comment-enhanced program at a children's hospital. *American J. Roentgenology*. 198, 5 (2012), 1121–1125.

[26] Villar, N. et al. 2009. Mouse 2.0: multi-touch meets the mouse. *Proc. of UIST*. (2009), 33–42.

[27] Vogel, D. and Baudisch, P. Shift: A Technique for Operating Pen-Based Interfaces Using Touch. *Proc. of CHI* (2007), 657-666.

[28] Wang, F. and Ren, X. Empirical evaluation for finger input properties in multi-touch interaction. *Proc. of CHI.* (2009), 1063-1072.

[29] Zhai, S. User performance in relation to 3D input device design. *ACM SIGGRAPH Computer Graphics*. 32 (1998), 50–54.