An in-the-wild study of learning mid-air gestures to browse hierarchical information at a large interactive public display

Christopher Ackad¹, Andrew Clayphan¹, Martin Tomitsch², Judy Kay¹
1. Faculty of Engineering and Information Technologies
2. Faculty of Architecture, Design, and Planning
The University of Sydney, NSW, 2006, Australia.
“firstname.secondname”@sydney.edu.au

ABSTRACT
This paper describes the design and evaluation of our Media Ribbon, a large public interactive display for browsing hierarchical information, with mid-air gestures. Browsing a hierarchical information space is a fundamental form of interaction. Designing learnable mid-air gestures is a current challenge for large display interaction. Our in-the-wild evaluation draws on 41 days of quantitative log data, with 4484 gestures detected, and qualitative data from 15 interviews, and associated video. We explored: whether our design enabled people to learn the gestures; how our tutorial and feedback mechanisms supported learning; and the effectiveness of support for browsing hierarchical information. Our contributions are: (1) design of large public display for browsing of hierarchical information; (2) with its gesture set; (3) insights into the ways people learn and use this interface in our context; and (4) guidelines for designing learnable mid-air gestures.

Author Keywords
Interactive public information displays, gestural interaction, user centred design and pervasive computing.

ACM Classification Keywords
H.5.2 Information Interfaces and Presentation: User Interfaces—Graphical user interfaces, input devices and strategies, interaction styles, screen design, user-centered design.

INTRODUCTION
This paper describes our work towards creating a large interactive public display that enables people to browse a hierarchical information space. This is a generic interaction paradigm that could play a promising role for such displays. It is applicable to many contexts, each with content collections tailored to the specific physical location in which the display is installed. To achieve this, we needed to tackle several challenges.

Although it is increasingly common that large displays are embedded in public spaces [6, 21, 23], these displays are still mostly limited to displaying non-interactive content. Previous research in the field has identified a sequence of four key challenges that stand in the way of effectively deploying interactive displays in public space. First, the theory of display blindness [15, 16, 17, 25] suggests that people do not notice public displays, since there is an increasing number of displays being deployed that show information with little to no relevance for passers-by. The theory is based on a similar effect known in online advertising as ‘banner blindness’ and suggests that people deliberately avoid looking at displays because they already expect that their content will be of little interest. Second, the theory of interaction blindness [17, 30] suggests that even when people look at a public display with interactive capabilities, they may not realise that they can indeed interact with the display. Third, related to interaction blindness, is the challenge of enticing interaction [6], which describes the phenomenon that even if people do interact with a public display, their interactions tend to be opportunistic [39] and of short duration [9]. Fourth, it is necessary to design intuitive interaction [13] that addresses the location, context and demographic of the public space, and allows passers-by to quickly learn how to interact with the system.

The study presented in this paper focuses on the last of these challenges – intuitive interaction. More specifically, our study involved the design of learnable mid-air gestures for interacting with a large public display and an evaluation of a gesture-based public display in-the-wild to better understand how people learn gestures in a public space. We focus on mid-air gestures, since the size and location of our public display required people to interact from a distance. Mid-air gestures have further been proposed as a promising interaction technique for large or wall-sized displays, where people need to stand well away in order to take in their entire content [4, 18, 27]. In such situations, touch interaction – although inherently more intuitive – is unsuitable as people may either not be able to touch the display surface or be required to walk back and forth, to move between viewing and interacting states [41]. The use of mid-air gestures presents challenges compared to touch interfaces, since these gestures represent a less common form of human–computer interaction and there is no set of well-known mid-air gestures. This means that people are not readily able to draw on their previous experience to learn or discover mid-air gestures.

An exception is the use of gestures as the input mechanism in gaming environments, as in the popular Microsoft Xbox Kinect platform. However, the default metaphor in those interfaces is point-and-dwell, which mimics mouse-like interaction by representing the user’s hand through an on-screen cursor. Activating an element in the interface, such as buttons or menu
items, is done by hovering the cursor over the element for a set amount of time. Although this interaction might offer some advantages for learning, it conflicts with the third challenge of enticing interaction. Interacting with a dwell-and-point interface is slow, tiring and cumbersome. This is because it demands focused hand-eye coordination for careful placement of the cursor, with the arm in a static mid-air hold, waiting for the interface element to activate [14, 20, 27].

We therefore arrived at the design decision to implement a combination of mid-air gestures that would allow passers-by to quickly learn and apply these gestures. The gestures were designed for a public information display, installed on a glass wall of the University of Sydney’s School of Information Technologies (IT) building facing the courtyard of a nearby theatre/performing arts centre (Figure 1). It consists of two high-intensity 1080p projectors covering a total area of 1.2 by 4.2 metres with a small gap in the centre. The display is back-projected onto a film laminated on the glass, which is optimised for high visibility in dark light conditions. The display operates from 6pm until midnight. This ensures optimal visibility of the display, and avoids problems of sun interference with the depth camera used to track people in front of the display. A Kinect for Windows device, was used as the depth camera, and was encased in an enclosure above the display. Being located near a theatre with several performance spaces, featuring diverse events from theatre to concerts, there is a steady stream of passers-by after 6pm.

The application running on the public display was designed to allow passers-by to explore a hierarchical information space. For this study, the display ran with with 132–155 items, with the content updated over time. The content included details of degree programs from Engineering and IT, the work of the Faculty’s research groups, entries from a University Photo Competition, and information regarding events at the adjacent theatre. Individual items included pictures and videos, with a paragraph of information. The material was prepared by marketing staff to ensure a professional and engaging appearance.

The remainder of the paper is structured as follows: First, we present a review of related work on public displays, designing gestures, and studies of interactive public displays conducted in the wild. We then describe our design of the gestures and our study setup. This is followed by an overview of our findings, their discussions, and a set of recommendations that we derived from the study.

BACKGROUND
There has been a burgeoning of research on large screen displays, including a recent review [8]. We have already noted the well studied challenges of display and interaction blindness. This section will focus on the core challenges we faced: designing a suitable mid-air gesture set; ways to teach gestures at large displays; and support for browsing a hierarchical information space using mid-air gestures.

For our context, it is critical that each gesture is quick to learn, easy to perform and is reliably recognised. Users must be comfortable executing the gestures in a public space. It is acknowledged that the design of mid-air gesture sets is challenging [8, 10, 28, 44, 45]. But there has been progress on many fronts. For example, Chattopadhyay and Bolchini recently reported work towards understanding gestures that are easy to do reliably [8]. Walter et al. [45] point to the challenge of designing gestures that are “immediately usable”. It has been recognised that the design of gestures needs to consider reliability combined with learnability [5]. Much of that work has assumed that the user controls a cursor [28]. For fast learning and easy recall, the gesture set should be small [7]. Rico et al. [34] identified the types of gestures people consider socially acceptable in public spaces, with people favouring small, unobtrusive gestures that are everyday actions. Nielsen et al. [28] concluded that people consider gestures laborious when they require two hands and large movements.

There has been a substantial body of work that has aimed to identify intuitive gestures by conducting elicitation studies [22, 26, 29]. Much of this was for interactive tabletops. For example, Morris et al. [22] found users preferred user-designed gestures, especially those elicited from many people and Nacenta [26] examined longer term gesture recall. These works have a recurring theme, relevant to mid-air gestures, that is – the value of simple gestures that people consider intuitive.

There have been many attempts to define categorisations for gesture sets, as an aid to designers. For example, Quek [33] distinguished manipulative and semaphoric gestures. Manipulative gestures involve a relationship between movements and the entity being manipulated, where semaphoric gestures are stylised static or dynamic hand/arm gestures. Aigner
We now consider work that supports browsing information. This metaphor was core to the work of Hespanhol et al. [13] comparing four mid-air gestures for intuitiveness, concluding that dwelling was most intuitive for selecting an item, followed by grabbing. Some participants preferred grabbing gestures for item rearrangement and dwelling gestures for dropping the item, which supports Aigner’s findings for “select” and “release” actions.

Much of the above work concerned learnability and intuitiveness of gestures. There has been far less about how to teach people mid-air gestures for a public display [10]. Early work used video to teach a set of flat-hand gestures where the orientation determined the command [42]. In a 4-user informal lab study, participants learnt the palm down gesture (used to provide more information). When a similar approach, with video icons, was used to teach four gestures [11], an in-the-wild study with 2312 users indicated this was effective for the two simpler gestures but it proved unsuccessful for the two more complex gestures. In-the-wild studies explored teaching a single “teapot” gesture for registration [44]. The work explored three approaches to “teaching” the gesture, by varying the way they presented the message, Touch your hip, along with a silhouette. They describe these as: spatial division (permanently on-screen in a reserved screen area); temporal division (where the information takes over the screen for a short period) and integration (where the information is integrated into the display content). The first proved more effective. But the authors noted the importance of group interaction and playfulness.

Indeed, there have been many papers reporting the ways that people learn through play [38] or by observing others [32, 37]. CityWall [32] explored aspects of social behavior change with the introduction of a large touch-enabled display in a public space to facilitate interaction. Chained Displays [37] explored the effects of the arrangement of multiple displays with mid-air gesture interfaces. Looking Glass [24] studied how the depiction of the user, affected interactivity at public mid-air gesture displays. Notably, these systems studied interactivity for games rather than our focus on an information display.

We now consider work that supports browsing information spaces at public displays. Early work by Vogel and Balakrishnan [42] focused on proxemic interaction for several feeds of information, such as calendars. Additional personal information was overlaid when the user was close to the screen. Information wall [19] is most similar to our work. It provided information such as restaurant menus and events. Using point-and-dwell gestures, users rotated a cube and expanded elements on the facing side. This cube interface gave a 2-level hierarchy. Other work has demonstrated effective support for browsing a single-level information space [11, 12], with left or right arm sweeping gestures to move the display. Grace et al. [11] designed a pair of 2-arm gestures to expand or contract a central item. These proved too hard to learn. To improve on these 2-arm gestures, Ackad et al. [1] designed static gestures and also extended the interface so that these new gestures supported browsing up and down the hierarchical information space. This work also explored new teaching elements: static icons showing the gestures plus dynamic tutorial text. The in-the-wild study over 120 days tested 4 conditions: tutorial on and off; and a single level vs a multilevel content set. No benefit was seen from the tutorial, and it indicated people had trouble learning the static gestures. The mix of fluid and static gestures seemed problematic for learning. The work pointed to the need for a greater understanding into the learning of gestures at public displays.

In summary, there is no previous work that has described the design of effective, fluid mid-air gestures, as distinct from point-and-dwell, for controlling a hierarchical information space. Our work aims to fill this gap, with a particular focus on gaining insights into the learning processes and the overall learnability of the gestures, in-the-wild.

**DESIGNING THE MEDIA RIBBON**

This section explains our design of the Media Ribbon interaction, starting with details of the context that influenced its design. The Media Ribbon is situated in a pedestrian corridor adjacent to a theater complex, as shown in Figure 2. Yellow lines show the thoroughfare-routes. The leftmost vertical thoroughfare takes people right past the Media Ribbon. The other path directs to a passageway at the right of the figure. These passersby are able to stop, but most do not. Some will use this route regularly, so are potential repeat users. There has been important in-the-wild work in such walk-through contexts [11, 24, 31, 37], including the impact of the location [3, 30]. The most likely users of the Media Ribbon are patrons attending the theatre. They typically have free time before or after a performance, at intermission or during an event in the courtyard. They are time-constrained by performance start times. Most come to this place irregularly and infrequently. In summary, this context demands designing for the new or intermittent user and many are likely to be time-constrained.

![Figure 2: Location of the Media Ribbon (blue bar and projectors at left), its depth camera’s field of view (pink) with pedestrian walking paths (yellow) and theatre patrons walking paths (purple).](image-url)
Supporting hierarchical information browsing

One of the core functions of desktops, tablets and phones is to support browsing of the hierarchical store of materials. Key concepts are that the user can browse through the collection of objects at the current level in the hierarchy, where some are containers for more objects while others are leafs in the hierarchy. For browsing the full information space, there must be a mechanism to move between hierarchy levels. Our Media Ribbon builds on this conceptual model.

We now explain this in terms of our interface, as shown in Figure 3. Each Media Ribbon session starts as in Figure 3a, with the top level of the hierarchy displayed as a “ribbon” across the screen. The central (focus) item is larger than the others. For browsing within the current level of the hierarchy, we define two abstract commands. Left and Right move the ribbon back and forth, within the current hierarchy level, as in Figures 3b and 3c. As an item comes into focus (centre), it unfolds, presenting the ribbon-image at the left and additional information to the right. The top level of our display is wrap-around; so a user can browse the top level using just one of these commands, for example, just Left.

To delve into the hierarchy, we defined the More command. This moves the current ribbon to the top of the display, becoming smaller, replacing it with the next level of the hierarchy as in Figure 3d. The user can then browse this level with Left and Right commands. The Back command reverses More, moving the small ribbon down to its former position (Figure 3e).

These four abstract commands are sufficient to move around the hierarchical information space. We defined one additional command – Vote. Voting is available on the focused leaf-item. Figure 3f shows an example with 22 votes (See the text near the green-background icon – midway down the image, slightly right from centre). Voting augments pure browsing, giving users a more active role, for example contributing opinions about a community issue [35, 36, 40]. This enabled us to explore how to design one active command beyond the bare minimum for browsing the hierarchical store.

Overall, we aimed to design interaction for a large screen public display that can support the fundamental role of a computer, browsing an information space. We also added one additional command – Vote, providing a very basic action.

Gesture design

We now describe how we designed the 5 gestures. We first identified the following, potentially competing goals:

1. Gesture recognition should be fast and accurate with only a Kinect device [22, 29].
2. Gestures should be socially acceptable to do in the display context [34].
3. Gestures should be possible with just one arm since people often carry things [37].
4. Gestures should be easy to execute, avoiding tiring static gestures [1, 11, 14, 19, 20, 27].
5. The gesture set should be small, simple and memorable [28].
6. Aim for similarity and symmetry for associated gesture pairs (Left/Right, More/Back).
7. Avoid overloading gestures.
8. Provide feedback as gestures are detected and recognised.

While all of these are important, the last four focus on learnability. As already mentioned, we rejected gestures based on
point-and-dwell, where the user’s hand becomes an on-screen cursor. This has the merit of drawing on a familiar WIMP mental model [43]. We rejected this in favour of more fluid gestures (Goal 4). We designed the Right and Left as whole arm swipe actions, the left or right arm moving as shown in Figures 3b and 3c. These are similar to single-finger swiping gestures on many touch devices, and so may draw on these existing mental models. The gestures meet our symmetry goal. Our wrap-around display of the top level partially meets Goal 3, for a single arm being sufficient. To ensure that users can quickly revisit an item, content designers would need to avoid large numbers of items in any one level. These gestures are similar to early work [12] where sweeping motions proved effective to move a single level carousel of photographs in the direction of the wave and for the case of browsing within a single level of a hierarchy [1, 11]. We rejected small gestures, such as pure hand movement [42] because that was incompatible with reliable recognition of the gestures. We considered a one arm swipe, with the direction defining the movement of the ribbon, but this poses problems for reliable gesture recognition. For example, the Left gesture requires the user to move their arm out to the start position and this should not be interpreted as a gesture (Goal 1).

In designing the More and Back gestures, we again chose a pair of symmetric sweeping gestures. For More, the user sweeps upwards with either arm (Goal 3) and then holds it briefly. The hold is needed to ensure reliable gesture recognition (Goal 1) but somewhat compromises the goal for fluid action (Goal 4). A similar downward sweep and brief hold is used for Back. The design works with either arm, and so meets Goal 5.

The Vote gesture is the same as More and to undo a vote the user repeats it. This violates Goal 7 as the upwards sweep means More, Vote and undo-Vote, but it makes for a small and simple set of gestures (Goal 5) that can be reliably recognised (Goal 1). This gesture is similar to a voting gesture designed by Parra et al. [31].

Support for learning

Our design provides four forms of learning support: feedback indicating how the system interprets user actions, the dynamic icon-based tutorial at the lower centre of the screen, a sequence of short duration tutorial messages, and just-in-time information on the voting gesture.

The first form of learning support is a real time representation of the user as a skeleton, using joint data detected by the Kinect sensor over the content as can be seen in each screen of Figure 3. This is a important form of feedback [11, 25, 38].

The second form of learning support is a dynamic icon-based tutorial. Figures 3a-e shows the icon-based tutorial as a user works through the tutorial. Figure 3a shows the start state, when the user first faces the screen. All icons are grey. After 2 seconds, the Left-icon changes to white to catch the user’s attention and encourage them to try the Left gesture. Once they have done this, the icon becomes green, as in Figure 3b, where the user has just completed their first Left gesture, using their left arm to swipe, as in the icon image. This causes the ribbon to move, following their left arm. We have added the red arrow to the image to show the movement of the ribbon. The next step of the tutorial is the Right gesture, which is signalled to the user by the Right-icon turning white. When the user has completed the gesture, the Right icon becomes green. Figure 3c shows this – where the user has just completed the second step of the tutorial. The third step of the tutorial is the More gesture and it follows the same pattern: the icon is highlighted in white to draw attention to it; then it becomes green when the gesture has been completed (as shown in 3d). Figure 3e shows the screen completion of the fourth tutorial step, the Back gesture, which moves the ribbon back down. Completed gestures become green. The icon block (bottom-center) is always present, a strategy that also proved effective in StrikeAPose [44]. Previous work [1] explored gesture uptake with and without the guided tutorial, using log data over 30 days for each condition. In the prior design the icon-tutorial was attached to the text messages for the tutorial condition and separated for the non-tutorial condition. For this design, we positioned the icon-tutorial carefully, to always be in a fixed location, avoiding obstruction of content and, importantly, near the skeleton, as previous work indicates this area draws user attention [1, 11, 25, 38].

The third form of learning support is a sequence of messages, one for each step of the tutorial. They are at the top of the screen and can be seen throughout Figures 3a-e. Each is a short encouraging message with advice to try the next tutorial step. We placed these messages above the content, to make them visible for both the active user and potential users watching from a distance, helping both sets of users learn.

Once the user has completed all 4 gestures in the order of the icon-tutorial, the icon-display changes to the mode shown in Figure 3f. This is where the last-recognised gesture turns blue. In the figure, the user just completed the More gesture. This gives additional feedback about the way the system interprets each action.

The fourth form of learning support is a context-dependent tutorial shown on the lower left of the additional information for an item in-focus. Figure 3a shows one for each focus item that have child-items in the hierarchy. This icon is accompanied by the message ‘Hold your Arm Up to find out more’. These are close to the attention-attracting skeleton. They are also right below important information that may engage the user to want to find out more. The other such message is for items in the leaf of the hierarchy, explaining voting as in Figure 3f.

Overcoming display and interaction blindness

The final design element deals with enticing users and informing them that the display is interactive. The very size of our display should facilitate this, based on reports that a large display (or series of displays) gives passing users time to notice the display and turn to it [24, 37]. We also aim to attract attention with a scrolling effect. To support this, the top level of the ribbon wraps around. Whenever a user is detected walking past the ribbon – it will appear to follow them [11]. Once they are detected facing the screen, a welcome message appears at the top of the display, as in Figure 3a. In addition, as discussed earlier, the skeleton, mirroring the user’s action should attract the user’s attention [11, 25, 38].
STUDY DESIGN

We designed a study to answer questions around the following themes in the context of our Media Ribbon.

1. **Learning**: Do people attempt to learn the gestures? When they do, how successful are they and how long does that take? What is the effect of groups?

2. **Tutorial**: In what ways does each tutorial element support learning?

3. **Feedback**: How effective is the skeleton mirror representation for providing direct feedback on user interactions?

4. **Browsing**: To what extent does the Media Ribbon achieve its core goal of supporting browsing hierarchical information, with mid-air gestures?

5. **Difficulties**: What problems do people face in using the Media Ribbon? What were the problems in regards to learning the gestures? Was the content adequate?

The first and second themes tackle the recognised challenge of creating a gesture set that is easy and quick to learn and apply [4, 10, 45]. This is particularly important in public environments, where people may be time-limited. We investigate the impact of groups since they play an important role for public display interaction [32, 44]. The third theme builds on previous work that showed the effectiveness of using a mirror representation as a feedback mechanism. We investigate whether such a representation can assist with browsing informational content, rather than just play. The fourth theme assesses the effectiveness of the Media Ribbon as a whole. This involves tackling challenges of display and interaction blindness right through to successfully using the Media Ribbon [17, 24, 30, 31] in a utilitarian display application. The final theme investigates aspects that may have been barriers for people learning the gestures or interacting with the Media Ribbon.

To investigate these themes, we drew upon multiple data sources. We collected *quantitative long-term data* of people interacting with the Media Ribbon. We automatically captured time-stamped log data of start time for each session (i.e. when an individual or group came up to the display); the number of people in the session; the time each gesture was recognised; and session end time. Before further analysis, we filtered out false positives triggered by the depth camera, such as a skeleton staying for less than one second. For *qualitative analysis*, we captured the depth stream from the Kinect camera at five frames per second whenever a person was interacting with the system, as well as the user view of the system, i.e. a screen capture of the application.

To develop a stronger understanding, we conducted 15 interviews with people immediately after they had interacted with the Media Ribbon. After seeking their consent to participate in the interview, we asked questions about the gestures, how they learnt them, in particular, whether they attempted the tutorial, what attracted them to the display, to interact, and whether they encountered any problems. Responses were recorded and later transcribed.

To analyse the data, we conducted a thematic analysis of the interviews, which led us to identify key themes for coding the data. For each interview session, we created a side-by-side view of the two synchronised video streams and a summary of the session’s quantitative data. This meant the coder could see the Kinect depth image that showed what the participants were doing, the display as the participants would have seen it, and automatically extracted statistics.

We coded the video streams matching the 15 interviews to assess whether participants appeared to: (a) follow the tutorial, either partially or completely; (b) was able to browse through the content; and (c) engage in playful interactions, such as observed previously [38].

FINDINGS

We now report the findings in terms of our five themes: learning, tutorial, feedback, browsing and difficulties. The quantitative data is from the 41 days, October 30 to December 16, 2014 (with two weeks missing for maintenance). The depth camera logs indicate that it detected 786 people in a total of 410 sessions. A session is a period of time where at least one person stood continuously in front of the display for at least 1 second. In 376 sessions, at least one person faced the display and in 284 sessions, at least 1 gesture was triggered. Of the 410 sessions, 217 (53%) involved just an individual, 193 were groups of two or more. Table 1 overviews the context and basic learning results for our interviews. The interviews fall into three groups, split by horizontal lines. For interviews 1 to 7, the theatre hosted a local technology conference. This had events in the courtyard, with crowds of 70-100 (Column 5 in the table). For this period, the display had 23 items related to the conference, bringing the content to 155 items. At the time of the 8th interview, the content had just the core 132 items. The content was then upgraded with the theatre’s summer program, bringing it again to 155 items. Timing of interviews was dictated by factors like weather and scheduled theatre events.

The table shows that interviews were held in the early evening. The sessions of interaction with the Media Ribbon ranged from 40 seconds (S7) to 6 minutes (S15). The next column shows the size of the crowd. This has the potential to impact learning by observation. The next two columns show the group size and gender. The 15 interviews involved 28 participants (20 male, 8 female). Just four involved individuals, with pairs dominating and just S1 and S15 being groups of three. This indicates our interviews involved more groups than the full log data set. The next column shows that all but three sessions included participants carrying things, a potential problem for two-arm interaction. The last column before the vertical bar, comes from the interviews and indicates that seven participants had previously used a Kinect. S3 is the only repeat user having observed, but did not use, the system the night before and came back to use it. The next two columns of Table 1 show gestures triggered, based on log data. The last column summarises key behaviours based on coded video analysis. This was based on two coders, assessing whether participants did the tutorial (T), browsed (B) or played (P). Inter-coder agreement using Scott’s Pi of 0.856 for Tutorial, 0.842 for Browsing and 0.864 for Play. The table shows differences between the coders: in S12, only Coder 1 coded T and B; in S4, only Coder 2 had P.
Learning
This section investigates learning of the gestures, combining our multiple sources of evidence. We first consider the degree to which people progress through the gestures. Then we consider effects of time and groups.

We first looked at how many gestures were triggered, based on log-data. (Table 2). The first row shows the number and proportion of sessions where each gesture was triggered at least once. The second row shows gestures that were repeatedly triggered. This is more likely to indicate intentional gestures. Both rows show a similar pattern, with much higher rates for Left/Right than More and a further drop when Back was used. Of 193 sessions where the More gesture was triggered, 105 also triggered Back, leaving 88 sessions, which ended without doing a Back gesture. A number of factors may have contributed to this, including time constraints, lack of interest in content, and problems in learning the gestures. We will explore each of these in turn.

We were initially surprised that the count for Right is higher than Left, since Left is the first step in the tutorial sequence, Left-Right-More-Back. We studied the video to see the potential effect of people carrying items and summarise the results in Table 1. In 8 sessions, people carried items in one hand; all but one (P2 in S8) held the item in their left hand to start.

<table>
<thead>
<tr>
<th>ID</th>
<th>Day</th>
<th>Session Length</th>
<th>Crowd Size</th>
<th>Group Size</th>
<th>Gender</th>
<th>Items Carried</th>
<th>Used</th>
<th>Kinect</th>
<th>Gestures</th>
<th>Repeated</th>
<th>Video Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>1</td>
<td>5:29</td>
<td>2:00</td>
<td>70-100</td>
<td>M M M</td>
<td>None</td>
<td>No</td>
<td>LRMB</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TBP</td>
</tr>
<tr>
<td>S2</td>
<td>1</td>
<td>5:50</td>
<td>0:46</td>
<td>70-100</td>
<td>M M</td>
<td>P2:Bag, Drink[L-N/A]</td>
<td>No</td>
<td>LRMB</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TB</td>
</tr>
<tr>
<td>S3</td>
<td>1</td>
<td>6:08</td>
<td>1:02</td>
<td>70-100</td>
<td>M</td>
<td>P1:Drink[L-S]</td>
<td>Kinect</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TB</td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>1</td>
<td>6:39</td>
<td>1:58</td>
<td>70-100</td>
<td>M M</td>
<td>P1:Bag, Drink[L-S]; P2:Bag, Drink[L-N/A]</td>
<td>Kinect</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TB</td>
<td>B[P(c)]</td>
</tr>
<tr>
<td>S5</td>
<td>1</td>
<td>6:41</td>
<td>1:11</td>
<td>70-100</td>
<td>M</td>
<td>P1:Backpack</td>
<td>Kinect</td>
<td>LRMB</td>
<td>LRMB</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>1</td>
<td>7:10</td>
<td>2:57</td>
<td>70-100</td>
<td>F F</td>
<td>P1:Drink[L-S]</td>
<td>No</td>
<td>LRM</td>
<td>LRM -</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>2</td>
<td>5:35</td>
<td>0:40</td>
<td>70-100</td>
<td>F M</td>
<td>P1:Backpack, P2: Bag</td>
<td>Kinect</td>
<td>LRM</td>
<td>LRM -</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>9</td>
<td>7:30</td>
<td>3:01</td>
<td>10-15</td>
<td>F F</td>
<td>P1:Phone[L-SI], P2:Phone[R-SI]</td>
<td>Xbox</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TP</td>
<td></td>
</tr>
<tr>
<td>S9</td>
<td>24</td>
<td>7:16</td>
<td>2:03</td>
<td>10-15</td>
<td>F M</td>
<td>P1:Bag, Phone[L-I]</td>
<td>P2:Bag</td>
<td>No</td>
<td>LRM -</td>
<td>LRMB</td>
<td>B</td>
</tr>
<tr>
<td>S10</td>
<td>24</td>
<td>7:06</td>
<td>1:31</td>
<td>7</td>
<td>F M</td>
<td>P2:Bag</td>
<td>No</td>
<td>LRM</td>
<td>LRM -</td>
<td>L - - -</td>
<td>T</td>
</tr>
<tr>
<td>S11</td>
<td>24</td>
<td>7:48</td>
<td>1:29</td>
<td>30</td>
<td>M M</td>
<td>None</td>
<td>Kinect</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TB</td>
<td></td>
</tr>
<tr>
<td>S12</td>
<td>24</td>
<td>7:55</td>
<td>2:22</td>
<td>30</td>
<td>F M(a)</td>
<td>P1:Bag[L-ID]</td>
<td>No</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TB</td>
<td></td>
</tr>
<tr>
<td>S13</td>
<td>29</td>
<td>6:59</td>
<td>3:11</td>
<td>7</td>
<td>M</td>
<td>P1:Bag</td>
<td>No</td>
<td>L - M</td>
<td>L - M -</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>S14</td>
<td>37</td>
<td>7:11</td>
<td>3:46</td>
<td>10</td>
<td>M</td>
<td>None</td>
<td>No</td>
<td>LRM</td>
<td>LRM</td>
<td>(TB(d))P</td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Overview of interview participants. Notes: Items Carried column has: L/R indicates hand; I, they used that hand to interact; S, they swapped item to other hand; D, they put item on the ground. Video Coding column has: T, they appeared to follow the tutorial (partially or completely); B, they appeared to browse the content; P, they have appeared to play at the display. (Notes: (a) S12 male was a child, but interview was only his mother; (b) – S7 and S10 had difficulties throughout; and (c) and (d) are discrepancies between coders: (c) – coded only by Coder 2 and (d) coded only by Coder 1.)

The table also shows that in 4 sessions, people swapped their item to the other hand to make it easier to interact and the participant in S12 also put the item down on the ground. So, it seems that about half the people carried items, most in their left hand, and about half of them either swapped hands or put the item down. This may explain the somewhat higher rates of the Right gesture recorded in our data.

We now consider the time it took people to progress through the gestures. This is important as some people may have only had short period available and this may have been the reason they did not progress. The median time for first triggering each gesture was: 9 seconds for Left, 6 for Right, 11 for More, and 26 for Back. This suggests that for those sessions triggering all 4 gestures, this commonly happens within half a minute. To explore this, we grouped the sessions by time. Figure 4 shows the repeat-use of each gesture, split by sessions < 30 seconds, 30 – 60 seconds and > 60 seconds. It shows the interview sessions in small overlaid bars. Unsurprisingly, those willing to be interviewed tended to have more time, most staying over 60 seconds. The distribution of the repeated gestures was similar to the overall population, skewed towards 4 gestures.

Figure 4 shows that in sessions of < 30 seconds, the largest group only completed 1 gesture. For sessions of 30-60 seconds, two and three gestures dominate. But for those staying over one minute, the largest group did all 4 gestures. It is unsurprising that short sessions are associated with less time to learn, and less learning. The figure suggests high success in learning the gestures in sessions over 1 minute.

We now discuss the findings from combined analysis of the interviews and associated video. The high level summary of triggered gestures is in two columns at the right of Table 1. This shows the letters LRMB indicating single and repeats of each gesture. This indicates that 10 of the 15 may have

<table>
<thead>
<tr>
<th>ID</th>
<th>Day</th>
<th>Session Length</th>
<th>Crowd Size</th>
<th>Group Size</th>
<th>Gender</th>
<th>Items Carried</th>
<th>Used</th>
<th>Kinect</th>
<th>Gestures</th>
<th>Repeated</th>
<th>Video Coding</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>1</td>
<td>5:29</td>
<td>2:00</td>
<td>70-100</td>
<td>M M M</td>
<td>None</td>
<td>No</td>
<td>LRMB</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TBP</td>
</tr>
<tr>
<td>S2</td>
<td>1</td>
<td>5:50</td>
<td>0:46</td>
<td>70-100</td>
<td>M M</td>
<td>P2:Bag, Drink[L-N/A]</td>
<td>No</td>
<td>LRMB</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TB</td>
</tr>
<tr>
<td>S3</td>
<td>1</td>
<td>6:08</td>
<td>1:02</td>
<td>70-100</td>
<td>M</td>
<td>P1:Drink[L-S]</td>
<td>Kinect</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TB</td>
<td></td>
</tr>
<tr>
<td>S4</td>
<td>1</td>
<td>6:39</td>
<td>1:58</td>
<td>70-100</td>
<td>M M</td>
<td>P1:Bag, Drink[L-S]; P2:Bag, Drink[L-N/A]</td>
<td>Kinect</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TB</td>
<td>B[P(c)]</td>
</tr>
<tr>
<td>S5</td>
<td>1</td>
<td>6:41</td>
<td>1:11</td>
<td>70-100</td>
<td>M</td>
<td>P1:Backpack</td>
<td>Kinect</td>
<td>LRMB</td>
<td>LRMB</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>S6</td>
<td>1</td>
<td>7:10</td>
<td>2:57</td>
<td>70-100</td>
<td>F F</td>
<td>P1:Drink[L-S]</td>
<td>No</td>
<td>LRM</td>
<td>LRM -</td>
<td>B</td>
<td></td>
</tr>
<tr>
<td>S7</td>
<td>2</td>
<td>5:35</td>
<td>0:40</td>
<td>70-100</td>
<td>F M</td>
<td>P1:Backpack, P2: Bag</td>
<td>Kinect</td>
<td>LRM</td>
<td>LRM -</td>
<td>T</td>
<td></td>
</tr>
<tr>
<td>S8</td>
<td>9</td>
<td>7:30</td>
<td>3:01</td>
<td>10-15</td>
<td>F F</td>
<td>P1:Phone[L-SI], P2:Phone[R-SI]</td>
<td>Xbox</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TP</td>
<td></td>
</tr>
<tr>
<td>S9</td>
<td>24</td>
<td>7:16</td>
<td>2:03</td>
<td>10-15</td>
<td>F M</td>
<td>P1:Bag, Phone[L-I]</td>
<td>P2:Bag</td>
<td>No</td>
<td>LRM -</td>
<td>LRMB</td>
<td>B</td>
</tr>
<tr>
<td>S10</td>
<td>24</td>
<td>7:06</td>
<td>1:31</td>
<td>7</td>
<td>F M</td>
<td>P2:Bag</td>
<td>No</td>
<td>LRM</td>
<td>LRM -</td>
<td>L - - -</td>
<td>T</td>
</tr>
<tr>
<td>S11</td>
<td>24</td>
<td>7:48</td>
<td>1:29</td>
<td>30</td>
<td>M M</td>
<td>None</td>
<td>Kinect</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TB</td>
<td></td>
</tr>
<tr>
<td>S12</td>
<td>24</td>
<td>7:55</td>
<td>2:22</td>
<td>30</td>
<td>F M(a)</td>
<td>P1:Bag[L-ID]</td>
<td>No</td>
<td>LRMB</td>
<td>LRMB</td>
<td>TB</td>
<td></td>
</tr>
<tr>
<td>S13</td>
<td>29</td>
<td>6:59</td>
<td>3:11</td>
<td>7</td>
<td>M</td>
<td>P1:Bag</td>
<td>No</td>
<td>L - M</td>
<td>L - M -</td>
<td>P</td>
<td></td>
</tr>
<tr>
<td>S14</td>
<td>37</td>
<td>7:11</td>
<td>3:46</td>
<td>10</td>
<td>M</td>
<td>None</td>
<td>No</td>
<td>LRM</td>
<td>LRM</td>
<td>(TB(d))P</td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Counts (and percentages) of sessions with each gesture triggered, at least once (Row 1) and repeatedly (Row 2).
learnt all four browsing gestures (S1-5, S8, S11-12, S14-15). This observation is supported by comments in the interviews. Several participants described the gestures as easy to use (S2-3, S5, S11-12) and intuitive (S8, S12-S15). S3 and S8 associated this with their familiarity with touch device gestures. S15 said the More gesture reminded them of putting their hand up in school, and that this made it intuitive and natural.

We now consider the voting gesture. Logs show 79 (21%) sessions with votes, with a total of 183 items liked (255 raw likes with 72 undos). Of the interviewees, 10 (S1, S3-4, S6, P8-9, S11-12, S15) voted. Of the 5 that did not vote, 2 (S5, S14) did not view any content that had voting enabled. For the other 3, S2 said they were feeling rushed (session was 46 seconds) and was not interested in the content or voting, S10 had difficulties, while S13 played ignoring the content. Overall, this indicates people actively used and understood the Vote gesture, even though it was overloaded, i.e. sharing the same gesture with More and Undo.

We now consider the 5 other sessions, where all gestures were not learnt. There were three classes of reasons for this. S7 and S10 had serious problems overall. In S7, the first participant held their arm up through the entire session while the second observed. In S10 both participants stood too close to the display (failing to see warning messages about this). The second class of problems were with the Back gesture. This affected S6 and S9. The video showed the primary participant clearly and repeatedly try to do it. They made a swipe down and forward, rather than to the side. Interview comments by S7 and S8 indicated they found Back was hard to do (although S8 did it successfully). The third reason for failing to progress through the gestures was seen in S13, where the individual was more interested in playing.

We now consider the effect of groups. In the full log data, about half the sessions were individuals (208) and the groups were mainly pairs (128) with 40 groups of 3 and 5 groups of 4. Both the number of gestures triggered and session length seem linked to group size. The median session length was 20 seconds for individuals, 43 seconds for pairs, 74 seconds for groups of 3, and 103 seconds for groups of 4. Figure 5 shows that larger group sizes were associated with more progress through the gestures; 29% of groups of 3 and 4 triggered all 4 gestures, but just 9% of individuals did so. Sessions where not even a single gesture was triggered occurred for 51% of single-person session but only in 9% of groups of 3 and 4. So groups tended to stay longer than individuals and progress through more gestures. To gain insights into how this interacts with learning the gestures, we turn to the interviews of each group. Of 11 interview sessions, all were pairs, bar the two triads (S1,S15). In five of these (S2,4,7,9,11), the second user was purely an observer and did not interact. For the remaining 6 sessions (S1,6,8,10,12,15) all members interacted with the display. We saw cases of people helping each other. For example, in S1, P1 completed the tutorial and browsed, while P2 observed. Then P2 took over, but had problems; P1 and P3 came back and helped them. There were several cases with observing, then mimicking successfully (S6,8,12,15) and of people trying to help each other, even in the problematic S10. Overall, the behaviour of these 6 groups points to learning benefits from groups, mostly in learning by observing but also in some helping.

Summary. This section started with the log data, complementing it with data from the interviews and videos. The log data suggests that the Right gesture was somewhat favoured over the Left but that both were learnt readily. Logs suggest More and Back are mastered in longer sessions, but that there may be problems with the Back gesture. The video and interviews confirm this and point to problems with Back for a minority of users performing it.

Tutoral

We now examine whether and how people used the two main forms of tutorial, the icon-tutorial, with dynamic feedback, and the short term messages that prompted the user to try each gesture. Both were designed to encourage the user to progress in a strict order: Left, Right, More then Back. However, the icon-tutorial was always visible, and so invited users to try gestures in any order. We placed the icon-tutorial near the skeleton as we expected that would help ensure users noticed it.

First, we consider the 10 sessions with success in learning all 4 gestures. We studied the videos to see if they used the tutorial
to do this, purposefully working through each gesture in order. The results, in Table 1, showed this was a dominant behaviour, seen in S1-3,7-12. So, it seems that the tutorial served these people well.

We now consider the sessions where the users appeared to learn all the gestures but not use the tutorial sequence (S4-5,15). S5 took 26 seconds to use all 4 gestures, S4 under a minute and S15 took 2 minutes. In the interview, S4 stated the gestures were simple and easy to learn, and referred to the icons providing the information needed. S5 and S15 both stated they learned the icon-tutorial. So they seemed to learn from this but without following the tutorial in order.

In the 5 problems sessions, the three reasons for failure had different interactions with the tutorial content. S7 and S10 stated that they did not see the on screen instructions; the users had the following problems: S7 held their arm up and S10 stood too close. S7 was very focused on watching the skeleton. It seems that our tutorial information failed for visibility to these users. For S6 and S9, our information failed to help them to learn the Back gesture correctly, even though S9 tried to work through the tutorial. S13, who played at the display, stated that they ignored the tutorial.

The interview comments also indicate that the icon-tutorial was judged helpful. Twelve participants (S1-6, S8-9, S11-12, S14-15) said they used the gesture icons to guide them while interacting with the display and eight (S1-2, S5-6, P8-9, S11, S14) stated that it was easy to follow. There were no corresponding comments about the text on the top of the screen. Four participants (S6-7, S9, S15) stated the text instructions were not visible.

Summary. We designed the tutorial elements for a single order of learning: Left-Right-More-Back. We now conclude that this was needlessly restrictive. Fortunately, the icon-tutorial did enable users to see all the browsing gestures at once. Overall, the icon-tutorial seems to have been largely successful. The value of its adaptive colouring was less clear. We should consider altering its adaptive colouring of “learnt” gestures to support learning in any order. Our results are in line with those of Walter et al. [44] – in that spatial division of tutorial information is more effective for learning than temporal division. But unlike that work, we wanted people to learn a set of gestures. In such contexts, the always visible icon-tutorial seems even more important. In the context of our public display, it seems important to facilitate free exploration and learning of gestures.

Feedback

The design made heavy use of the skeleton, mirroring the user actions. Participants from 10 of the sessions (S2-4, S6-9, S12-13, S15) stated that they found the skeleton helpful and four (S7-8, S12-13) said it was the skeleton that made them realise the display was interactive. Participants from 8 of the sessions (S2, S4-6, P8-9, S13, S15) stated that the skeleton helped them to understand how the computer interpreted their actions, with three (S2-4) saying this helped them learn the gestures. S2 also explained how the skeleton showed him the problem with his initial attempts, when his gestures were too small and not detected.

At the same time, 4 participants (S5, S10-11, S14) considered the skeleton was not useful and detracted from the displayed content. Even S3, who used and liked the skeleton for learning the gestures, stated after, it was too distracting to the overall experience and took people away from the content.

Summary. The skeleton appears to be effective in attracting attention and providing feedback. However, some people found it distracting after the learning phase.

Browsing

Since a core goal of this work was to enable people to browse a hierarchical information space, we now consider how well it did this. We report on video analysis and interview comments about the information explored.

Of all the users who triggered at least one gesture, logs indicate that an average of 7 became the focus. This rose to 15 for sessions with all 4 gestures. In 79 sessions, people descended 1 level in the hierarchy and in 112 sessions, 2 levels. Combining this with the level of voting reported above, the logs suggest many sessions browsing the information space.

The video analysis is summarised in the last column of Table 1. This shows 10 sessions with people browsing the content. In 4 of these sessions (S10, S12, S14, S15) people were observed to appear to be carefully reading the content. In interviews, 7 (S2,4-6,10,12,15) stated that they had read the content and 3 (S7,10,12) stated they found it interesting. This indicates that a majority of people succeeded in browsing.

Although our core goal was to support browsing, we need to acknowledge that some users simply did not want to do that. This is reflected in the case of S13, who preferred to play. But there may also be synergies between play and learning. This should be explored in the future.

Summary. We designed the Media Ribbon to enable people to quickly learn the gestures needed to browse a hierarchical information space. Overall, it seemed to achieve this with a majority of people doing so.

Difficulties

Of the 6 sessions that did not repeatedly use all 4 gestures, all 6 (S6-7, S9-10, S12-13) did not use Back, 2 (S12-13) did not use the Right gesture and 1 (S10) did not use More. From the video logs we see S10 appeared to be too close to the display throughout the entire session (with implications as explained above). S13 and S12 used only two gestures, S13 initially swiped left as he entered, however for the rest of the session he played purely with his skeleton, ignoring the content and gestures. S12 attempted to use the display, however she was caring a shoulder bag that impeded the mobility of her right hand. She navigated using Left swipes and used More to transverse down a level and liked a piece of content before giving control back to her son, who then proceeded to both use and play with the display. In this case she was unable to use her right hand, however in the interview she mentioned that she did not see the visual aid and was not aware of the Back gesture.
SUMMARY OF FINDINGS AND RECOMMENDATIONS

We started from existing recommendations for the design of our gestures and interaction. Then, the design required trade-offs between these. Our study supported several of these design decisions. This paper reports lessons learnt from 41 days of log data, interviews and video. It builds upon previous work on teaching mid-air gestures. We note that our results without a ‘no-tutorial’ baseline may not be transferable to other contexts. Our design and study were for a place where most users are first time visitors with a few minutes around their evening at the theatre. So they are time-constrained and need to be able to learn gestures quickly. Many are in small groups and most are carrying items. So it is valuable to support single-handed operation. Unlike much large display literature which involves game interfaces, we aimed to support the quite fundamental task of browsing a hierarchical information space and doing a basic action at leaf nodes. When combined with creative content, these can provide a foundation for many valuable and interesting ways to interact with a large public display.

**Keep the gesture set small.** Our work demonstrates that our five abstract gestures were sufficient for supporting browsing a hierarchical information space and voting at leafs.

**Left and right sweep gestures are easy to learn.** Both the log and qualitative data point to the learnability of these gestures.

**Provide always-present dynamic on-screen tutoring for a small set of core gestures.** While a majority of interviewees worked through the tutorial, the icon-tutorial allowed more flexibility in order of learning.

**Provide context-sensitive help, to introduce additional gestures and consolidate core gestures.** Two-thirds of the interview participants used the Vote gestures and their comments indicate they intended to do so and found it easy to learn. We propose to extend the advice on the Back gesture, adding it near the Vote information when relevant.

**Overloading gestures is an effective way to keep the gesture set small.** We designed the same gesture for More and voting. Participant comments indicate that this form of overloading worked fine.

Harness playful and whimsical features for feedback. Participant comments and our observations confirm previous findings that the skeleton attracts attention, is pleasing to many and provides valuable feedback on the gestures.

**Strive for consistency and symmetry in the gesture sets.** Our two pairs of gestures, Left/Right and More/Back, were both symmetric. They are also complementary to each other. Each pair is strictly consistent. The Left/Right pair is flowing. In order to get reliability in recognition, we had to add a pause at the end of the flowing gestures for More/Back. This works well but it would be preferable to have reliable gesture recognition for flowing gestures.

**Design for single-handed operation.** Our design only does this in part. For the case of browsing within the top level of the hierarchy, a user could go back to an item using the wrap-around. All four gestures were single handed and More/Back could be accomplished with either hand. However, the use of Left and Right required people to swap hands to reverse direction.

CONCLUSION

We presented results from a 41 day in-the-wild study of our Media Ribbon, an interactive public information display in a public space. We aimed to design a learnable gesture set that would enable people to explore a hierarchical information space. To that end, the work fills a gap in the literature, since most of the work on gesture-based public display interaction has focused on games and playful interaction. Furthermore, previous work investigating user interaction with multi-hierarchical information on large displays has focused on touch input only. Our study is the first to our knowledge to investigate a generic set of gestures for browsing multi-hierarchical content on a large public display.

To ensure the ecological validity of our data, we evaluated the Media Ribbon in a public space. We collected data through interaction logs, video logs, and from a series of interviews, to answer questions around the themes of learning, feedback, the use of a tutorial, and people’s browsing behaviour.

Our findings reveal that the Left and Right gestures were easy to learn for navigating within one level of the hierarchy. In contrast, More was used less often, and this seems linked to the time people spent at the display, rather than the design of the gesture. The Back gesture posed some learnability problems. Although we did not compare our tutorial approach with a ‘no tutorial’ baseline, which is a limiting factor to our findings, interviews showed that the icon-tutorial was effective and that the skeleton representation also helped. Our work points to the need to take account of different sub-populations, some focused more on play, as well as those who focus on learning the gestures and exploring the content.

ACKNOWLEDGEMENTS

This research was supported by funding from the Smart Services CRC and the Faculty of Engineering and Information Technologies, The University of Sydney, under the Faculty Research Cluster Program.
REFERENCES


42. Vogel, D., and Balakrishnan, R. Interactive public ambient displays: transitioning from implicit to explicit, public to personal, interaction with multiple users. In the 17th annual ACM symposium on User interface software and technology, ACM (2004), 137–146.

43. Vogel, D., and Balakrishnan, R. Distant freehand pointing and clicking on very large, high resolution displays. In the 18th annual ACM symposium on User interface software and technology, ACM (2005), 33–42.
