FISHEARS – The Design of a Multimodal Focus and Context System

David K McGookin *Stephen A Brewster*

Department of Computing Science
University of Glasgow
Glasgow,G12 8QQ
UK
mcgookdk, stephen@dcs.gla.ac.uk
http://www.dcs.gla.ac.uk/~stephen

SUMMARY
In this paper we describe a new focus and context visualization technique called multimodal focus and context. This technique uses a hybrid visual and spatialized audio display space to overcome the limited visual displays of mobile devices. We demonstrate the technique by applying it to maps of theme parks.

KEYWORDS: Focus and context, earcons, PDA, small screen devices, multimodal interaction, sonification

INTRODUCTION
Each year manufacturers are producing smaller and more powerful mobile computing devices. Palm Pilots, Pocket PC's and mobile phones have become ubiquitous. For example 5.5 million mobile phones were sold in the 3 months before Christmas 2000 [9]. Manufacturers are now looking to produce multi-purpose mobile devices that will act as digital music players, mobile phones and web browsers. There is a potential limit to this growth – screen space. The main advantage of these devices, that they are small and hence mobile, severely limits the size of the visual display. For example, the Palm m105 has a screen size of only 5cm x 5cm.

Solutions to the problem of limited screen space have long been around for desktop systems in the form of focus and context visualization techniques [8]. Mobile computing is, however, very different from desktop computing. Of great importance is the ability of users to employ their visual sense for safe navigation of their environment. For example, if you are checking your email on the move you must split your visual resources between the reading of your mail and not falling down flights of stairs, getting run over by a car or any of the other dangers we can fall victim to by not looking where we are going. Even if we attempt to reduce these dangers by staying stationary whilst checking our mail, people could still walk into us or a car could mount the pavement and hit us. In short we need our eyes for much more important tasks than using a mobile computing device. Because of these issues it is necessary for a user to continually avert their eyes from reading email to monitor the environment and confirm that all is well. In doing this it is likely that when they return to viewing the email it will take some time to relocate where they were.

In an attempt to reduce the visual load on users we have designed a hybrid visual and spatialized audio focus and context technique called multimodal focus and context. Multimodal focus and context should not only increase the mobile device's display space, allowing more information to be displayed but also reduce the demands on the user's visual sense by providing a constant audio context, allowing users to more quickly relocate where they were if and when their eyes are averted from the PDA (personal digital assistant) display. This should allow users to better and more safely navigate the physical environment.

In the remainder of this paper we will explain the relevant history of focus and context visualisation before describing the multimodal focus and context system. We shall then describe how data is represented in the spatialized audio space.

FOCUS AND CONTEXT
Focus and Context visualisation was originally, independently proposed by both Furnas [5,6] and Spence & Apperley [10]. Each of their proposed systems share the same common features but differ in key aspects.

All focus and context representations of information spaces share the same basic premise that more information is required to be presented than can be adequately presented simultaneously. In order to maximise the visual display space the information to be presented is split into two parts:

- **Focus:** That part of the information space that is of most interest to the user. This part is presented in maxmum detail**.**
- **Context:** The rest of the information to be displayed. In order to allow all of the required information to be displayed this information is displayed in much less detail than the focus.

The way in which the visual display is split between the focus and context largely determines whether the representation would be considered as Furnas's Fisheye [5] or Spence and Apperley's Bifocal Lens representation [10]. In the Fisheye representation the focus and context are merged such that the detail of the information being displayed is gradually reduced the further away from the focus we move. This makes Fisheyes useful where the focus and context share the same visual representation. The Bifocal Lens has a much stricter visual disparity between the focus and context. In this system the focus and context can have different visual representations. Hence it is easy to tell if data is in the focus or the context (there is no merging of the two as with the Fisheye). For example, Spence and Apperley [10] demonstrated a visual bookshelf representation. Books were dragged from the bookshelf to another part of the screen where they were "opened" so that they could be read. As was noted in [2], the bifocal lens style of focus and context means the representation of the data in the focus and context do not need to be the same.

There has been little research on applying focus and context to mobile computing devices. Notably the work of Bjork *et al*. [2] has attempted to apply Flip Zooming [7] focus and context visualisation to PDA's. However, this work still suffers from the issues previously outlined involving the demands on the visual sense.

MULTIMODAL FOCUS AND CONTEXT

Our new Focus and Context system augments the visual display with new modalities, specifically spatialized (3D) audio, to increase the available display area for information presentation. However, we only use a 2D transverse audio plane at the level of the ears.

Overview

We decided to apply the Bifocal Lens concept to the multimodal display platform. There are several advantages to this approach. Firstly, as with the disparity between the focus and context on the bifocal display, there is a disparity between the visual and audio modalities. In other words it is not possible to display visual representations in audio and *vice versa*. Another advantage is that the focus is high detail whereas the context is of lower detail. This fits well with the display platform in that it is not possible to display aural information in as much detail as visual information. These advantages mean that it is convenient to make the visual display the focus and the audio display the context. The splitting of the focus and context in this way should mean that the visual demand of the user is lowered and that they will be able to retain their position in the map even when their visual attention is distracted by the environmental stimuli.

Fitting together the focus and context

The focus essentially "floats" over the context. In essence users see the focus on the screen. The data which is to the right and up from the focus is 'played' in audio to the right and forward of the user. The data that is to the left and down from the focus is played to the left and rear of the user. Users navigate through the space via scrollbars on the visual display, or a movement sensor mounted on the mobile device. The act of moving a part of the display from the focus to the context actually means moving map items from the visual to the audio modality. When this occurs the visual representation of the map item is replaced with a spatialized audio representation. For example, scrolling to the right will cause the left part of the focus to move from the visual display to the audio display (and hence move from the focus to the context).

Audio representations of map items remain the same relative distance from each other that they did when they were displayed in the visual modality. In essence we are moving a lens (the visual display) over a large information space. The data that the visual display is over is represented visually; the rest of the information space is represented in audio.

APPLICATION TO THEME PARKS

In order to properly explain the rest of multimodal focus and context we shall use it as a means of presenting theme park visitor maps on PDA's.

By their very nature theme parks are large and thus difficult to navigate, requiring visitors to use maps. The use of electronic maps will allow dynamic information such as the length of queues for rides to be displayed.

Cluttered contexts

Because of the display limits of mobile devices, far more information will be represented in the audio modality (context) than in the visual modality (focus). This could lead to a very cluttered audio space which might be unusable to the user. To overcome this problem we apply parts of the Level of Detail concept as proposed by Furnas [6]. Here each map item (theme park ride in our example) is given a measure of importance. How important something is directly relates to how much visual (or audio) display resource it gets. The greater the level of detail an object is given, the more important it is. Two types of importance were defined [6]:

- ***a priori* importance**: How important something is globally. E.g. in a city the town hall, hospital, railway station etc would be given high *a priori* importance. Places such as a particular restaurant or particular house would have low *a priori* importance as for the majority of visitors, these places would be unimportant.
- ***a posteriori* importance**: How important something is to an individual based on what they are currently doing. For example, if you were going to visit your

relations in a city, your relation's house would have a high *a posterori* importance whereas (hopefully) the hospital would have a low *a posteriori* importance as it would not be important to what you were currently doing.

In our system we deal only with *a priori* importance and use the term priority to describe it. Each theme park ride in our display space is given a priority level based on its global importance. The higher the priority level the more important the ride is. Note that we use the term ride in a very general way to include such things as toilets and food stalls. Hence a food stall or toilet would have a low priority level whereas large roller coasters would have high priority levels.

In order to enforce priority levels we define priority zones that extend outwards from the focus (see Figure 1).

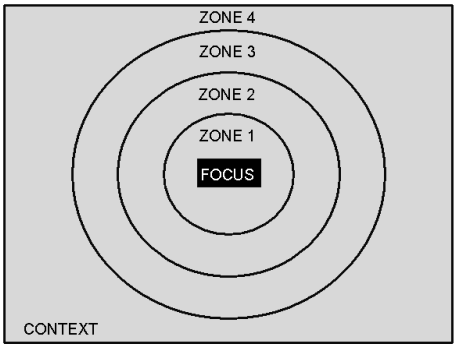

Figure 1: Diagram showing the priority zones. Priority zones are fixed relative to the focus

For a sound (representing a ride) to be played it must lie in a priority zone that has a priority less than or equal to it. In other words if a ride (represented by a sound) lies in a zone with a higher priority than it, the sound is not played. Priority zones are located relative to the focus so as the user changes the focus, sounds will pass between priority zones resulting in them either being switched on or off. Note that no ride that lies in the focus region of priority zone 1 is played in sound, such rides are presented visually.

The advantage of this approach is that we can remove from the context those things that will be of little interest, in most cases, for the user, and hence reduce the clutter in the context. For example, assuming we connected a GPS (global positioning system) device to our multimodal focus and context system such that, unless the user explicitly changed the focus by scrolling, the focus on the mobile device would reflect the user's physical position in the theme park. Now, suppose that the user was hungry and wanted a hamburger. It is likely that they would want to go to the hamburger stall closest to their current location, as opposed to the hamburger stall at the other end of the park. There would be little reason for them to be able to hear the hamburger stalls that are far away from them in the context. On the other hand, if a user were looking for a large roller coaster it is likely that they would want to find it no matter where it was on the map display, i.e. the distance the roller coaster was from the focus would be largely irrelevant. It would therefore be required to display the roller coaster either visually or aurally (depending if it was in the focus or context) no matter where the roller coaster was on the map.

DISPLAYING RIDES

Displaying rides visually is straightforward; we simply provide a pictorial representation of the ride on the visual display. However how do we represent rides in audio? In order to answer this question we must first consider what attributes of rides to communicate. In other words what things about a ride might a user of our system want to know? In Table 1 we present four possible attributes.

Attribute	Description
Type	As stated previously, we define "ride" in very general terms, including toilets and food stalls. It is necessary therefore to give the user some understanding of what type the ride is. Possible examples of type might be roller coaster, water ride, toilet, food stall, etc.
Intensity	How intense the ride is. Large, fast, roller coasters equal high intensity, whereas a miniature railway designed to transport customers around the park would be given a low intensity.
Cost	How much the ride costs.
Queue Size	How long the queue for the ride is. Or rather, how long it is necessary to queue before being able to get on the ride.

Table 1: Useful ride attributes a user might want to know.

This list is by no means exhaustive, and not all of the attributes would be suitable in all cases. For example, intensity is probably irrelevant for food stalls, and ride cost is unimportant for theme parks that charge only for park entry and not for rides. However, as will be shown later, the more information we try to communicate about rides, the more complex understanding the attributes will become.

The obvious way to present our ride attributes in audio is through speech. However speech is slow. For example, in speech a simple audio message representing the four attributes represented above might be "Ride Type: Roller coaster, Intensity: High, Cost: High, Queue Size: High.". Using the AT&T TTS (text to speech converter) [1] this phrase took 5 seconds to be played. Since there will be multiple rides in the context, each being represented by a spoken phrase like that described, using speech will re-

quire significant attention of the user of the system. This will obviously cause an increase in workload for the user. We propose instead to use Blattner's Earcons [3] to represent the attributes of rides in the multimodal focus and context system. Earcons are structured abstract audio messages [4] that can be constructed to be shorter than equivalent spoken messages. We propose to use the compound form of earcons. Compound earcons are composed of motives. Motives represent primitive object e.g. ride type is represented by a motive. In our example each earcon would be composed of a maximum of four motives, one for each of the attributes. The motives are played serially, one after another, to form the earcon. There are several ways in which the earcons could be designed. We shall describe one possible method in order to demonstrate the concept.

We give each of the attributes a different timbre, e.g. the type could be a piano, the intensity could be a trumpet, the cost a guitar and the queue size a drum. In order to communicate different states for each of the attributes we could use the pitch of the motive. Since it is difficult to make absolute judgements on audio pitch we shall reduce the allowable states of each of the attributes to no more than 3-4 states. For example, intensity could be reduced to low, medium, or high, each of which would be represented by a different pitch. Other attributes could be classified in a similar way.

EVALUATION

We intend to compare the multimodal focus and context system to the standard approach for displaying large amounts of data on small visual displays – scrolling. We hope to show that multimodal focus and context reduces visual workload and allows more effective navigation around large theme park maps.

CONCLUSIONS

We have described a technique for expanding the limited visual displays of mobile computing devices. This technique is based on focus and context visualization, using the visual display as the focus and 3D spatialized audio space as the context. We term this multimodal focus and context. We have also described how multimodal focus and context could be applied to a theme park's visitor map.

ACKNOWLEDGEMENTS

This work was supported by EPSRC award 800055301.

BIBLIOGRAPHY

1. AT&T Text to speech generator. Accessible at *http://www.research.att.com/~mjm/cgi-bin/ttsdemo*. 2001.

2. Bjork, S. & Redstrom, J. Redefining the Focus and Context of Focus+Context Visualizations. In *Proceedings of IEEE Symposium on Information Visualization 2000* IEEE, 2000.

3. Blattner, M.M., Sumikawa, D. A. & Greenberg, R. M. Earcons and Icons: Their Structure and Common Design Principles. *Human Computer Interaction* 4, 1 (1989), pp. 11-44.

4. Brewster, S.A. *Providing a structured method for integrating non-speech audio into human-computer interfaces*. University of York, 1994.

5. Furnas, G.W. Generalized Fisheye Views. In *Proceedings of Conference on Human Factors in Computing Systems* ACM, 1986, pp. 16-23.

6. Furnas, G.W. The FISHEYE view: a new look at structured files. In *Readings in information visualization: using vision to think*, Morgan Kaufmann, San Francisco, California, 1999, pp. 312-330.

7. Holmquist, L.E. Focus+Context Visualization with Flip Zooming and the Zoom Browser. In *Proceedings of Conference of Human Factors in Computing Systems* (Atlanta, Georgia USA) ACM, 1997, pp. 263-264.

8. Leung, Y.K. & Apperley, M. D. A Review and Taxonomy of Distortion-Oriented Presentation Techniques. *ACM Transactions on Human-Computer Interactions* 1, 2 (1994), pp. 126-160.

9. BBC News, Accessible at *http://news.bbc.co.uk/hi/english/business/newsid1100000/1100250.stm*. 2000.

10. Spence, R. & Apperley, M. D. Database Navigation: An Office Environment for the Professional. *Behaviour and Information Technology* 1, 1 (1982), pp. 43-54.