Human-Computer Interaction -- INTERACT'03
M. Rauterberg et al. (Eds.)
Published by IOS Press, (c) IFIP, 2003, pp. 607-614

Design and Evaluation of a Multimodal System for the Non-Visual Exploration of Digital Pictures

Patrick Roth & Thierry Pun

Computer Science Department
CUI, University of Geneva
CH – 1211 Geneva 4, Switzerland

Patrick.Roth@staff.hu-berlin.de

Abstract: This paper reports on the design and the evaluation of an audio-haptic tool that enables blind computer users to explore digital pictures using the hearing and feeling modalities. The tool is divided in two entities: a description tool and an exploration tool. The description tool allows moderators (sighted persons) to describe a scene. Firstly, the scene is manually segmented into a set of objects (car, tree, house, etc.). For each object, the moderator can define a behavior, which corresponds either to an auditory rendering (i.e., using speech or non-speech sounds) and/or to a kinesthetic one. The blind person uses the exploration tool in order to obtain an audio-haptic rendering of the segmented image, as previously defined by the moderator. This interactive exploration is obtained by means of a planar force feedback device. The system was experimented by a group of ten blind participants. Results revealed that the audio encoding of visual objects combined with an active kinesthetic exploration enables blind people to obtain a fairly good mental image of the explored scene.

Keywords: Blind users, digital pictures accessibility, audio and haptic feedback

1 Introduction

Multimedia plays an important role in today's communication society. A growing number of documents available on the Web are based on a combination of different kind of media such as sound, picture and video. Unfortunately, the increase of information bandwidth for the majority of population becomes a source of gap for some communities of disabled people. For instance, picture based documents on the Web seriously decrease the amount of information transmitted to blind computer users.

Current methods for non-visual picture representation use two different rendering models, either tactile or by verbal description. Tactile pictures are obtained by thermal printers that use a special paper that swells when heated (Edman, 1992). Before being printed, the pictures are previously simplified by specialists (sighted people) in order to be understandable by blind people. One problem with this technique is due to the apparatus, which produce uniquely static pictures.

The verbal description method became very useful with the emergence of the Web. This technique allows Web designers to add a short text that describes the image. This text does not appear on the screen but blind users can read it. In reality however, few Web designers use a verbal description. Another problem stems from the nature of the output information. Representing images in a textual form removes the spatial layout of the picture, which bears essential information for the sighted.

To solve the problems mentioned above, we have developed a system that enables blind computer users to access both the content and the spatial organization of digital images. This work is an extension of our previous investigations on a web sonification tool named "WebSound" (Petrucci, 2000).

In section 2, we introduce other work that has been done to represent spatial information of graphics via non-visual modalities. In section 3, we describe the design of our non-visual representation system. Section 4 covers the system hardware modifications made on the force feedback device. Sections 5 and 6 report our experiments and present a

discussion of our findings. Finally, in section 7, we draw our conclusion and briefly introduce our future perspectives.

2 Past Research

A significant amount of research has been done to study the non-visual presentation of images to the blind people. In general, those researches fall into two categories depending on the output modality they use, i.e. auditory or haptic.

2.1 Auditory Displays

The use of auditory displays (AD) is the domain where most of the investigations have been made. Scadden (Scadden, 1984) was the first to report the use of sonification to access data on visual interfaces.

As for the non-visual representation of diagrams, they have been investigated by linking touch (using a graphical tablet) with auditory feedback. Kennel (Kennel, 1996) represented diagrams (e.g., flowcharts) to blind people using multi-level audio feedback and a touch panel. Touching objects (e.g., diagram frames) and applying different pressures triggered three types of feedback. The first was the information regarding the frame. The second was the interrelation between frames. The third type of feedback used speech output to express the textual content of the frame.

Using speech output, Mikovec (Mikovec, 1999) defined an object-oriented language for picture description. In his approach, an image is defined by a list of objects in the picture. Every object is specified by its definition (position, shape, color, texture, etc.), its behavior ("is driving") and by its interrelations with other objects. These interrelations are either hierarchical ("is in") or not (for groups of objects without hierarchical relation). This description is then stored into an XML document. To obtain the picture description, the blind user works with a special browser. This tool browses through the objects that compose the image and reads their information.

The direct use of the physical properties of the sound is another method to represent spatial information. Meijer (Meijer, 1992) designed a system that uses a time-multiplexed sound to represent a 64*64 gray level picture. In his system, each pixel is associated with a sinusoidal tone, where the frequency corresponds to its vertical position and the amplitude corresponds to its brightness. Each column of the picture is defined by superimposing the vertical tones. The final signal is obtained by concatenating each resulting column together. Hollander (Hollander, 1994) represented shapes using a "virtual speaker array". This environment is defined with a virtual auditory spatialization system based on specific head-related transfer function (HRTF) (Hollander, 1983). This auditory environment directly mapped the visual counterpart. A pattern is rendered by a moving sound source that traces out the segments belonging to the pattern.

2.2 Haptic Displays

Jason Fritz investigated different methods for representing various forms of picture characteristics (boundary or shape, color, texture) using haptic rendering techniques. For example, to render lines, Fritz developed a virtual fixture mechanism (Fritz, 1999). When the force feedback pointer is close enough to the line, this mechanism pulled the end-effector towards the line. Another type of visual data rendered by Jason Fritz was surfaces. In his model, a surface is represented as a texture. This texture was implemented using a bump-mapping algorithm (Minsky, 1995). Furthermore, Grabowski extended the system developed by Fritz by adding sonification to the haptic representation (Grabowski, 1998). In his approach, the haptic component was used to represent topological properties (size, position) while sonification mapped purely visual characteristics such as colors or textures.

Colwell et al. (Colwell, 1998) carried out a series of studies on virtual textures and 3-D objects. In their studies, Colwell et al. tested the accuracy of such haptic interface for displaying size and orientation of geometrical objects (cube, sphere). They also studied whether blind people could recognize simulated complex objects (i.e., sofa, armchair and kitchen chair). Results from their experiments showed that participants might perceive the size of larger virtual objects more accurately than for the smaller ones. Users also may not understand complex objects from purely haptic information. Therefore, additional information has to be supplied before the blind user can explore the object.

3 Design of the Non-Visual Representation Tool

By tacking all researches on auditory and haptic displays into account, we have implemented a non-visual picture representation system. The tool that we have developed is based on the concepts formulated by Martin Kurze (Kurze, 1995), which concern the presentation of graphical information to blind persons in a non-visual way. Applying these concepts, our tool operates in two phases. In a first phase, a sighted person (moderator) provides the

corresponding model of the displayed visual graphic. In the second phase, the blind person can explore the graphic without any external help. Our tool can be seen as two different separate systems, that is a description tool and an exploration tool. The following sections detail the design of these two systems.

3.1 Description tool

As mentioned before, the description tool allows the moderator to adapt a picture for non-visual presentation. When adapting the image, the moderator first segments it manually into a list of objects (house, car, etc.). We have opted for a manual segmentation instead of automatic algorithms, in order to keep the semantic consistency of the images.

After segmenting the scene into objects, the moderator secondly defines, for each object, a unique behavior. This behavior model is represented either by an auditory cue or/and by a kinesthetic constraint.

3.1.1 Manual segmentation technique

For the manual segmentation, the moderator defines with the mouse the set of control points that represent regions surrounding the object (see figure 1.b). Shapes can be represented either by a polygon, a spline, or by a circle.

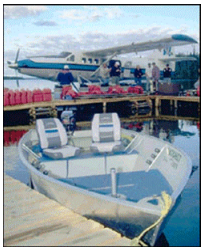

Figure 1.a: Original image.

Figure 1.b: Manual segmentation.

3.1.2 Verbal summary

This feature allows moderators to insert a short text that defines the nature of the picture (scene, diagram, etc.) as well as the list of objects included. Using this feature, the blind user quickly obtains a general idea of the image that he will explore.

3.1.3 Object sonification

Investigations on auditory display (Kramer, 1994) have shown that the use of non-speech sounds in computer interfaces can increase the amount of information bandwidth transmitted to the user. These results were also confirmed during our previous studies (Petrucci, 2000).

Taking these results into account, we have implemented a feature by which the moderator can assign an auditory cue to a corresponding object. This sound can be applied on two different aspects of the object that are its contour and its surface. Depending on the aspect, the blind user obtains the auditory feedback when crossing the object and/or during the whole time he points to the surface of the object (see section 3.2 for more details).

Concerning the auditory cue selection, we have created a sound database that contains environmental sounds. These sounds are classified into different categories of events such as alarms, weather, engine, etc. The moderator can also add new sounds into the database.

We also provide an alternative to the sound selection by means of verbal label. In this case, the moderator assigns a short text to the object. This text can be read by the system during the exploration process.

3.1.4 Haptic rendering

As complement to the auditory feedback, we use the haptic sense by means of a planar force feedback interface. Using this modality, we also render two aspects of the objects: the contour and the surface (see figure 1.b).

3.1.4.1 Contour rendering

The contour modeling is based on a virtual force of fixture that we implemented as a result from one of our previous investigations (Roth, 2002). To summarize the method, this force of fixture is based on a virtual spring that attracts the mouse cursor towards the line. The point **p** on the line with the minimum distance dc from the mouse cursor **c** is first determined (see figure 2).

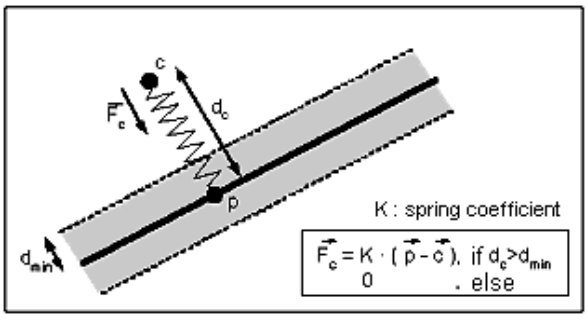

Figure 2: Force of fixture rendering.

If the distance is above a given threshold (i.e., $\mathbf{d_{min}}$ in our case), an attraction force $\mathbf{F_c}$ is activated; $\mathbf{F_c}$ is orthogonal and directed towards the line, as well as proportional to the distance $\mathbf{d_c}$.

3.1.4.2 Surface modeling

The surface rendering is provided when the cursor is located within the object. At this moment, the physi-

cal constraint is activated. The moderator has the choice between two categories of constraints that are friction or texture. As for the friction constraint, the moderator can apply different forces, $\mathbf{F_{max}= m_s \cdot F_n}$ by varying the coefficient of dynamic friction $\mathbf{m_s}$ as well as the normal constraint force $\mathbf{F_n}$. The friction algorithm that we use in our system is provided by Immersion [16].

The haptic texture constraint that we provide is based on the model developed by Minsky (Minsky, 1995) on a two-degrees of freedom haptic interface. Basically, the force $\mathbf{F_x}$ opposing the cursor motion is calculated from the translation direction **x**, the surface height **y** (making $\mathbf{\Delta y/\Delta x}$ the local gradient) and is given by $\mathbf{F_x=k(\Delta y/\Delta x)}$ where **k** is a stiffness constant (see figure 3).

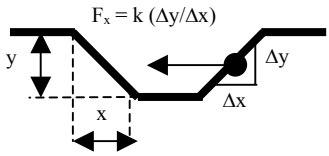

Figure 3: Force determination for haptic texture rendering.

The moderator defines different textures by varying the values of **x**, **y** and **k**.

3.2 Exploration tool

As we have mentioned before, we have also developed an exploration tool that displays the non-visual image to the blind user. To obtain the full non-visual rendering, the user goes through two consecutive stages. In the first stage, the user passively listens to the verbal summary in order to obtain a general overview of the image.

In the second stage, the blind user actively explores the content of the image by means of a planar force feedback device. In our system, this device provides the absolute positioning input as well as force feedback effects. Thus, when the user explores the scene using the planar force feedback, the system responds with auditory cues whose pitch depends on the touched object. The absolute position of the object on the scene is given by the location of the manipulandi on the planar force feedback device. As for the kinesthetic feedback, the user explores the picture using 2D planar force feedback device. During this exploration, the user obtains kinesthetic sensations that represent the object shapes and surfaces as specified by the moderator.

4 Adaptation of the Planar Force Feedback Device

The results obtained during our previous investigations on the haptic presentations of graphs revealed that the planar force feedback we used (i.e., Logitech WingMan Force Feedback mouse, see figure 4.a.) was not adapted for such type of representation (Roth, 2002). This was due to several psychophysical limitations such as the small size (i.e., 2.5[cm]*2[cm]) of the workspace or the nature of grasping.

Figure 4.a: Planar force feedback device used in our previous investigations.

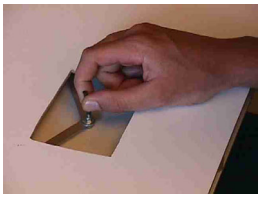

Figure 4.b: Modified device.

Taking these results into account, we have performed several modifications on this device, in particular by increasing the size of the workspace (i.e., 11.5[cm]·8[cm]). We have also replaced hand grasping by finger grasping (see figure 4.b.), since as reported by Cutkosky finger grasping provides a better precision compared with hand grasping (Cutkosky, 1990).

5 Experiments

We have conducted a series of experiments in order to study the efficiency of the system described above. The system was experimented by ten blind people aged between 20-43 years old. Five of the participants were congenitally blind. No participant had prior experience using planar force feedback devices.

5.1 Methodology

All participants were given twenty minutes to familiarize themselves with the exploration rendering system. Participants were allowed a maximum amount of time not exceeding 30 minutes to complete each task. While testing, participants were allowed to use the haptic device with the one hand they were accustomed to use the most. Upon completing the tasks, participants were asked to reproduce their mental image of the scene using a raised line drawing pad (Edman, 1992).

No hints were given at any time. Task completion times and the participants' comments were noted

down. All sessions were video recorded for further analysis.

5.2 Tasks

Each participant was given two different pictures one at a time for evaluation. The two pictures were:

- Landscape: a scene composed of a lake, a boat, a dock, a road, a tree, a house and a car (i.e., 7 objects) (see figure 5.a.);
- Plan: a map comprising 2 traffic lights, 6 trees, 2 roads, an island, a tram stand, 4 sidewalks and 4 crosswalks (i.e., 20 objects) (see figure 5.b.).

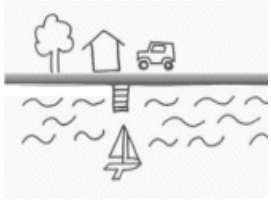

Figure 5.a: Picture of a *landscape*.

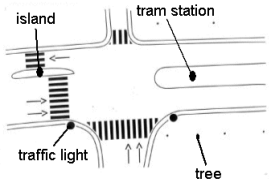

Figure 5.b: *Plan* of a crossroads.

We choose to work with these pictures because of their contextual diversities as well as their differences in complexity. In addition, the figure 5.b basically represents a tactile transcription used by blind people to obtain the orientation of a new district.
The two pictures were segmented then encoded as described above. For the overall summary, we verbally defined their contexts and enumerated all the objects that were included. For the audio encoding of each scene, we sonified each objects using environmental sound as well as a verbal label. As for the surface modeling, we used both texture and friction constraints to encode the objects. The Table 1 and Table 2 summarize the audio and haptic encoding that we used for the experiment.

6 Results and Discussion

6.1 Results

The success of the reproduction was assessed by means of three complementary criteria. The first criterion (variable "found") reports on the percentage of participants that found the corresponding object. As for the second and third criteria, we evaluated subjectively two spatial aspects of the objects drawn by the users, that is the precision of their location (variable "located") and their morphological similarity with the original picture (variable "contour").

The chart in Figure 6 reports the results obtained for the landscape. They reveal that participants had no difficulties in finding all the objects that are included in the picture (90% of success). We also noticed that participants had no problem either to establish a connection between the objects and their auditory counterpart.

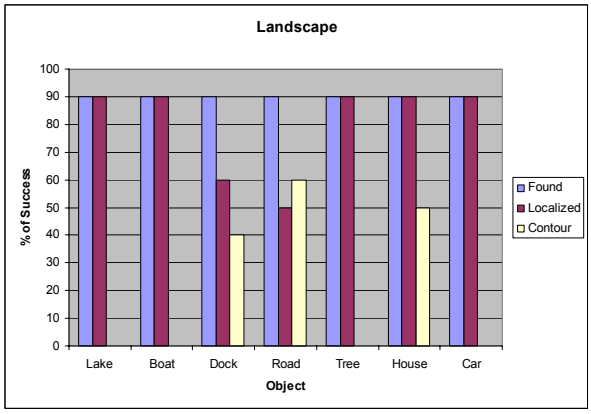

Figure 6: For the landscape, percentage of participants that correctly performed with the three criteria.

Object	Rendering			
	Sonic			Haptic
	Speech	Non speech		
		Contour	Surface	
Lake	"lake"	Object water	Waterfall	Friction
Boat	"Boat"	Ship whistle	Motor boat start	Texture
Dock	"Dock"	Walk	Walk	Texture
Road	"Road"	Traffic jam	Traffic jam	Friction
Tree	"Tree"	Small fire	Tree fall	Texture
House	"House"	Close door	Steps walk	Texture
Car	"Car"	Car start	Car pass	Friction

Table 1. Auditory and haptic encoding for the landscape.

Object	Rendering			
	Sonic			Haptic
	Speech	Non speech		
		Contour	Surface	
Traffic light	"Traffic light"	Vibration	Car brake	Friction
Tree	"Tree"	Small fire	Tree fall	Texture
Road	"Road"	Traffic jam	Traffic jam	Texture
Island	"Island"	Thump	Traffic jam	Friction
Tram station	"Tram station"	Subway halt	Subway halt	Texture
Sidewalk	"Sidewalk"	Walk	Walk	Texture
Crosswalk	"Crosswalk"	Walk	Traffic jam	Friction

Table 2. Auditory and haptic encoding for the plan.

Participants also have successfully localized all the objects on the scene (90% of success), except for the dock (60% of success) and for the road (50% of success). For the latest, as can be seen on figure 7, some participants were able to perceive only a portion of the road.

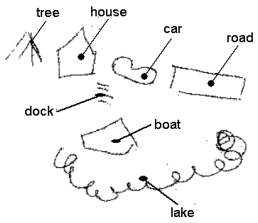

Figure 7: Landscape reproduced by a congenitally blind participant.

Oppositely, except for the road (50% of success), the house (50% of success) and the dock (40% of success), the system did not allow participants to distinguish the boundary of the objects. For some participants, as it is showed on figure 8, the object contours were reproduced by privileging the semantic aspect of the sound instead of their kinesthetic properties.

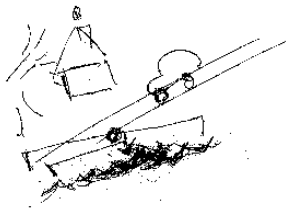

Figure 8: Landscape reproduced by a late blind participant.

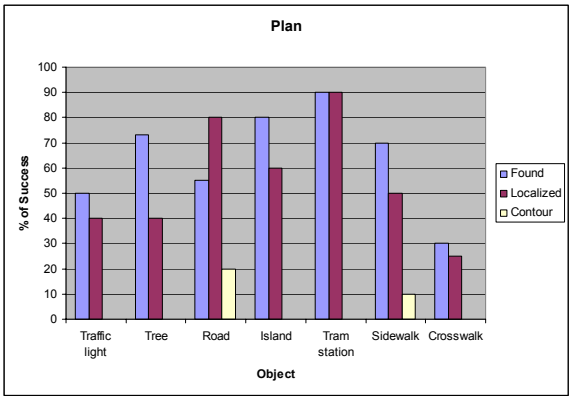

Figure 9: For plan, percentage of participants that correctly performed with the three criteria.

The charts in Figure 9 summarize the results obtained for the plan. Here, the results showed that people had more difficulties in finding and positioning the objects. We noted two main reasons for these failures. The first is that the number of objects was too high. The second is that the size of some objects was too small: participants had difficulties in localizing small sized objects such as trees and traffic lights.

Finally, as can be seen on figure 10, except in some case such as with the road, the participants were not able using the system to determine the boundary of the objects.

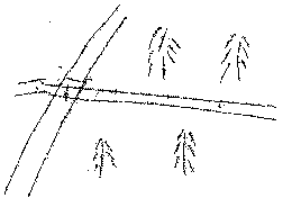

Figure 10: Plan reproduced by a congenitally blind participant.

6.2 Discussion

In general, we observed that to obtain a mental representation of the picture, the participants analyzed the pictures following four consecutive steps, as it is showed on figure 11.

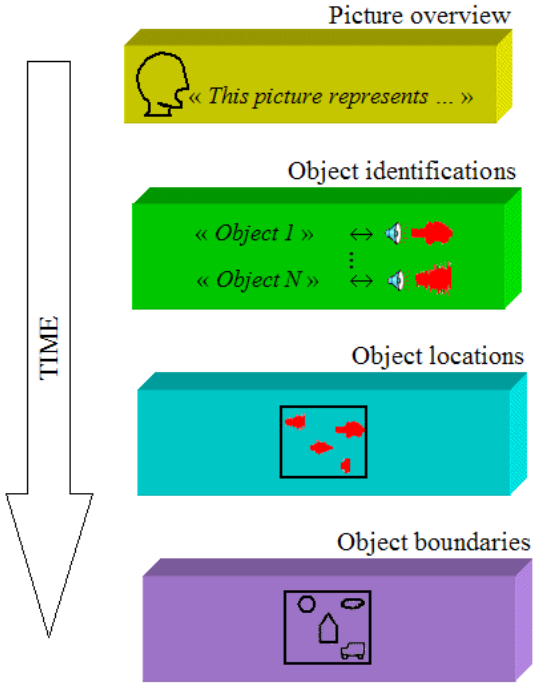

Figure 11: Mental image representation process.

The following sections elaborate on each of these four steps used for building the mental representation of the image.

6.2.1 Verbal overview

The blind participants started their mental image construction by working with the verbal summary

functionality. This functionality was very useful for them to get a global overview of the picture as well as of the objects included in it.

6.2.2 Object identifications and non-speech attributes

Following the verbal overview, they actively explored the image by means of the planar feedback interface. They have performed this exploration in order to identify all the objects in the picture. The principal reason they performed this step was to quickly learn the non-speech language. Then, at each time they identified a new object, they learned its corresponding non-speech "word". We have noted that the participants used the verbal labeling functionality only when the non-speech mapping did not semantically correspond well enough to the visual object. This was the case for three objects: "traffic light", "island" and "crosswalk".

6.2.3 Object locations

When the auditory language had been completely learned, the participants focussed their attention on the kinesthetic aspects of the exploration. Therefore, the people determined the position of all the objects, this according to a unique object of reference. For example, as can be seen in figure 12, the location of all objects is obtained relatively to the house.

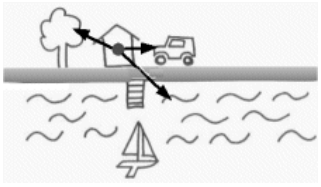

Figure 12: Localization technique used by the participants.

This technique however has showed two drawbacks. First, the participants indicated having difficulties in precisely moving the manipulandi on a horizontal, respectively a vertical axis. This lack of precision explained that in some case, such as shown in Figure 13, the participants did not place the tree, the house and the car on the same horizontal axis.

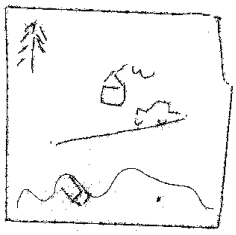

Figure 13: Object positions reproduced by a congenitally blind participant.

To solve this problem, we developed a haptic navigation help functionality. When activated, this help allowed blind users to move precisely according three orientations: horizontal, vertical and oblique. These three orientations are modeled as straight lines that provide a force of fixture similar to the one used for the contour rendering.

The second problem that occurred during the experiments concerned the difficulties encountered by the participants to distinguish between similar objects such as sidewalk and crosswalk. For the latest, the objects were rendered by the same behavior and participants could not differentiate between them. A solution to such problem would simply consist in adding a unique label (e.g., number) for redundant objects.

6.2.4 Object boundaries

Finally, after locating all the objects, the participants tried to determine their corresponding boundaries. To obtain this information, they used the contour rendering functionality. Unfortunately, the participants showed difficulties in following the boundaries of several objects. The two main reasons for this are:

- By contrast to natural tactile perception, our haptic interface provides users with only a single point of contact;
- The complexity of the boundaries of some objects. For the same reasons as the one mentioned by Colwell (Colwell, 1998), we have seen that participants were able to recognize simple shapes (i.e., square, rectangle); they were however not able to recognize more complex objects.

To solve these problems, a possible solution lies in the use of an auditory language that could help for the navigation. This language would include four auditory cues, one for each orientation (see figure 14.a.).

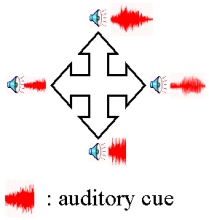

Figure 14.a: Auditory cues for representing an orientation.

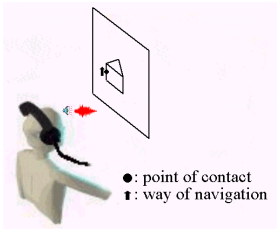

Figure 14.b: Active navigation using auditory navigation aid.

The sounds in Figure 14.a represent the orientation of the pixels that follow the current point of contact. Therefore, in relation with the direction of navigation, the blind user would hear a sound that would inform him as to the direction to take (see figure 14.b.).

7 Conclusion

We have described in this paper the design of a tool that aims at providing blind computer users access to the content of digital pictures. Our approach is based on audio and haptic interaction. Results from the experiment have shown that in order to represent digital pictures in a non-visual way, the auditory modality can be useful to define the semantic properties of the objects. As for the kinesthetic modality, it can be used to provide spatial information such as object positions. However, the contour rendering technique that we have developed did not appear to provide enough capability to represent the boundaries of complex objects. Here, a possible solution could be based on the adaptation of an auditory navigation aid that would cooperate with the kinesthetic exploration. Another solution could be the use of a higher quality force feedback device (e.g., the haptic glove provided by Immersion), which offers a better kinesthetic resolution.
As an extension to this research, we plan to compare our non-visual presentation method with other traditional approaches such as tactile and verbal descriptions.

8 Acknowledgments

This project is actually financed by the "Swiss National Science Foundation".

References

Blauert, J. (1983), *Spatial Hearing*. MIT Press, MA.

Colwell, C., Petrie, H., Kornbrot, D., Hardwick, A., Furner, S. (1998), Haptic Virtual Reality for Blind Computer Users, *Proc. ASSETS '98*, pp. 24-30.

Cutkosky, M. R., Howe, R. D. (1990), Human Grasp Choice and Robotic Grasp Analysis, in S.T. Venkataraman and T. Iberall, (eds.), *Dextrous Robot Hands*, Springer-Verlag, New York.

Edman, P. K. (1992), *Tactile Graphics*, American Foundation for the Blind.

Fritz, J. P., Barner, K. E. (1999), Design of a Haptic Visualization System for People with Visual Impairments, *IEEE Transactions on rehabilitation engineering* 7(3), 372-384.

Grabowski, N., Barner, K. E. (1998), Visualization Methods for the Blind Using Force Feedback and Sonification, *Proceedings of SPIE International Symposium on Intelligent Systems and Advanced Manufacturing*, pp. 131-139.

Hollander, A. J. (1994), *An Exploration of Virtual Auditory Shape Perception*, Diploma thesis, University of Washington.

Immersion corporation, http://www.immersion.com.

Kennel, A. (1996), AudioGraf: A diagram reader for blind people, *Proc. ASSETS'96*, pp. 51-56.

Kramer, G. (1994), *Auditory Display, Sonification, Audification, and Auditory Interfaces*, Addison-Wesley.

Kurze, M. (1995), Giving Blind People Access to Graphics (Example: Business Graphics), *Proc. Software-Ergonomie '95 Workshop*, Darmstadt, Germany.

Meijer, P. B. L. (1992), An Experimental System for Auditory Image Representations, *IEEE Transactions on Biomedical Engineering* 39(2), 112-121.

Mikovec, Z., Slavik, P. (1999), Perception of pictures without graphical interface. in *5th ERCIM Workshop on User Interfaces for All*, Dagstuhl, Germany.

Minsky, M. (1995), *Computational Haptics: Texture*, PhD Thesis, MIT Media Labs.

Petrucci, L. S., Harth, E., Roth. P., Assimacopoulos, A., Pun, T. (2000) WebSound: a generic Web sonification tool, and its application to an auditory Web browser for blind and visually impaired users. *Proc. ICAD 2000*, pp. 6-9.

Roth, P., Kamel, H., Petrucci, L. & Pun, T. (2002), A Comparison of Three Nonvisual Methods for Presenting Scientific Graphs, *Journal of Visual Impairment & Blindness* 96(6), 420 - 428.

Scadden, L. A. (1984), Blindness in the information age: equality or irony?, *Journal of Vision Impairment and Blindness* 78(9), pp. 394-400.

Sjöström, C. (1999), *Non-Visual Haptic Interaction Design*, PhD Thesis, Center for Rehabilitation Engineering Research, Sweden.