

Visual Attention in Auditory Display

Thorsten Mahler¹, Pierre Bayerl², Heiko Neumann², and Michael Weber¹

¹ Department of Media Informatics

² Department of Neuro Informatics

University of Ulm, Ulm, Germany

{thorsten.mahler | pierre.bayerl | heiko.neumann |
michael.weber}@uni-ulm.de

Abstract. The interdisciplinary field of image sonification aims at the transformation of images to auditory signals. It brings together researchers from different fields of computer science like sound synthesizing, data mining and human computer interaction. Its goal is the use of sound and all its attributes to display the data sets itself and thus making the highly developed human aural system usable for data analysis. Unlike previous approaches we aim to sonify images of any kind. We propose that models of visual attention and visual grouping can be utilized to dynamically select relevant visual information to be sonified. For the auditory synthesis we employ an approach, which takes advantage of the sparseness of the selected input data. The presented approach proposes a combination of data sonification approaches, such as auditory scene generation, and models of human visual perception. It extends previous pixel-based transformation algorithms by incorporating mid-level vision coding and high-level control. The mapping utilizes elaborated sound parameters that allow non-trivial orientation and positioning in 3D space.

1 Introduction

Human actions, natural occurrences, movements of any kind produce unique acoustic events. Every acoustic event results from an action which is tightly bound to this exact effect. In the natural world this effect is taken (or should be taken) into account whenever a new product is developed: New motors are designed which reduce noise, new street delimiters cause sounds when run over etc. The interesting aspect here is that the product itself emits characteristic sound as feedback and signals for certain events. This occurrence is addressed in the relatively new field of sonification. In the computer science domain sonification becomes used more often, as well. Three classes of sonification approaches have been proposed previously [Hermann et al., 2000b]:

First, parameter mapping, where image data (e.g. position, luminance) is directly mapped to the parameters of the sound signal (e.g. amplitude, frequency, duration); second, model-based sonification, where virtual sound objects (e.g. instruments) are controlled by the visual input; third, auditory scene generation, where the input data is utilized to control the behavior of defined auditory objects which distinguishes auditory scene generation from simple parameter mapping.

2 Related Work

The starting point for our work is the sonification system vOICe introduced by Meijer [Meijer, 1992]. In his work he aims on the substitution of a missing visual sense by sound and introduces a sensor substitution device for the blind. This system is a variation of the parameter mapping approach, where image luminance steers the sound amplitude, vertical image location the sound frequency and horizontal location time and stereo. A drawback of this approach is that the entire data contained in the image is sonified, regardless of the relevance of the information.

Other researchers in this domain sonify not only two dimensional images but high-dimensional data sets. Hermann et al. present in [Hermann et al., 2000a] a model based approach in which significant structures in large datasets are detected. These significant points are integrated in a so called principal curve. This one dimensional description of the dataset is then used for sonification and thus presents a general acoustic overview over the data set.

Rath and Rocchesso present another aspect of sonification in [Rath and Rocchesso, 2005] by introducing a tangible interface. A bar is used as an natural interface for the task of balancing a virtual ball. They use an underlying physical model to immediately and accurately predict and produce rolling ball sounds whenever the bar is agitated. A moving ball is shown on the screen and the bar is used to control its motion and every motion triggers an immediate natural sound event. Thus they combine visual feedback on the screen with aural feedback and thereby could reach a significant increase in usability (average task time).

In [Martins et al., 2001] Martins et al. focus on texture analysis and point out the similarity between speech and texture generation models. Their results especially of the conducted user studies show promising results in that computer vision algorithms are of great use for sonification approaches.

The presented work has to be distinguished from other sonification approaches, in that we focus on sonification of data generated by attention driven perceptual models from images. This allows us to generate intuitively comprehensible auditory displays sonifying highly compressed visual data. Our overall investigations lead us to enhanced visual displays which can be used to support user interaction in a future user interface.

3 Image Preparation

We aim to sonify images of any kind and propose two strategies how this can be achieved. The first strategy is to model visual attention particularly concerning orientation and on this basis a second strategy uses visual grouping to dynamically select relevant visual information to be sonified. We restrict the objects of interest to be represented by elongated regions of similar contrast orientation, such as the borders of a road or the bar of a road sign. Alternatively, other approaches such as [Marr, 1982] [Krüger and Wörgötter, 2005]

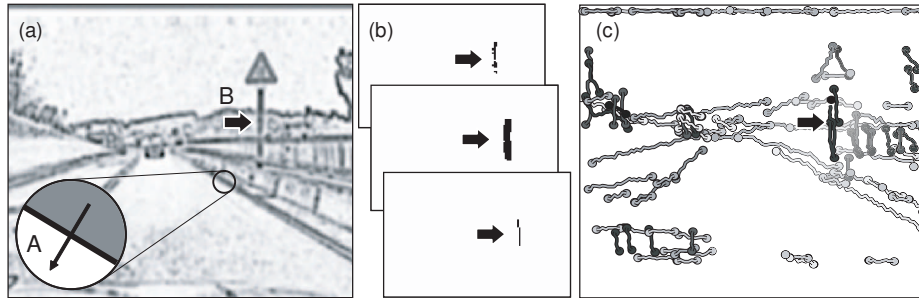


Fig. 1. Example demonstrating initial feature extraction and grouping. (a) Initial features are extracted by a center surround mechanism highlighting isolated contrasts. A feature is located at each position in the image and consists of a saliency value (encoded by darker pixels) and an orientation (depicted in inlay A). The arrow (B) points to a vertical contrast which is used to illustrate the grouping scheme in (b-c). (b) Our grouping scheme aims to bind nearby features of similar orientation to individual objects of elongated shape. Selected features are merged and finally thinned by morphological operations in order to generate a thin line depicting the object. (c) shows a set of automatically detected objects represented by connected lines of different luminance. The luminance is dependent on the underlying orientation. The vertical bar extracted in (b) is highlighted and indicated by an arrow.

[Fischer et al., 2004] [Weidenbacher et al., 2005] could also be used to obtain data for sonification.

Our approach is divided in two stages: local feature extraction (one feature per pixel) and the generation of more abstract object descriptions by binding different local features together. The first stage utilizes a local principal component analysis of the luminance gradient in the input image to extract local contrast orientation and amplitude (structure tensor approach [Trucco and Verri, 1998]). Then, attentional mechanisms are borrowed from models of visual perception, such as contrast detectors [Itti et al., 1998] and local normalization [Simoncelli and Heeger, 1998] to enhance perceptual relevant image features. For sonification, we apply a threshold to generate a sparse map of salient oriented contrasts. Individual features are described by their spatial location and orientation (Fig. 1a).

In the second stage, we apply a simple grouping scheme which aims to bind nearby features with similar properties to individual objects. Selected local features are grouped accordingly to local contrast orientation and spatial proximity. The implementation of the grouping employs morphological operations [Gonzalez and Woods, 2001] to achieve the combination of spatial connected locations (Fig. 1b). As a result we extract an image sketch describing connected lines or repeated patterns of similar orientation (Fig. 1c).

Thus, our approach generates (1) sparse features describing the underlying image scene by localizing contrasts described by the position and orientation of these contrasts. Then (2) more abstract objects are generated as combinations

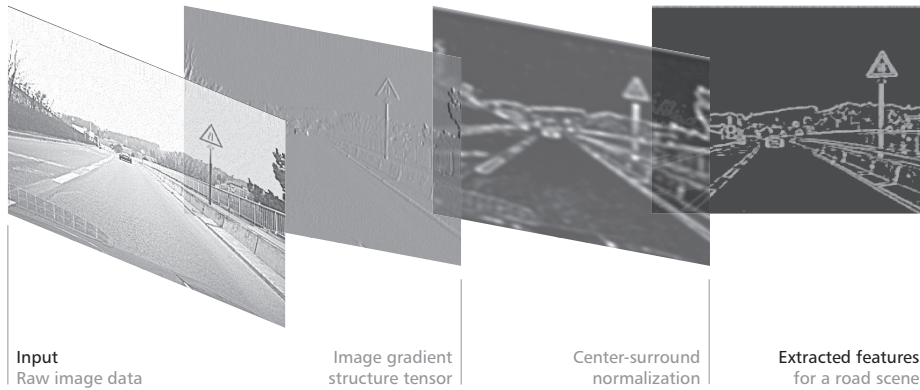


Fig. 2. Visual attention: Salient feature extraction

of local features belonging together described by the size and the shape of the objects in addition to orientation and location of underlying features. Based on the image preparation we apply two models of sonification.

4 Sound Space Modeled Sonification

Of the many different approaches to image sonification we want to present two promising ones here. The first one is an approach where the auditory space is the starting point. We use a grid to divide the auditory space into cells each representing a position in 2D sound space (a plane orthogonal to the vertical axis of the auditory observer). Now an image can be sonified by moving a cursor through the image left to right, line after line, from bottom to top. This left to right direction is used because of the most obvious reason, our natural reading order. The direction of reading an image bottom to top since it seems more natural to display foreground before background. Therefore the objects and streaks in the lower image half are naturally nearer to the spectator and thus more important and seen and therefore sonified first.

As a first step in this sonification we transform the original image. We use the thresholded and thus sparse output of the first stage of the presented fea-

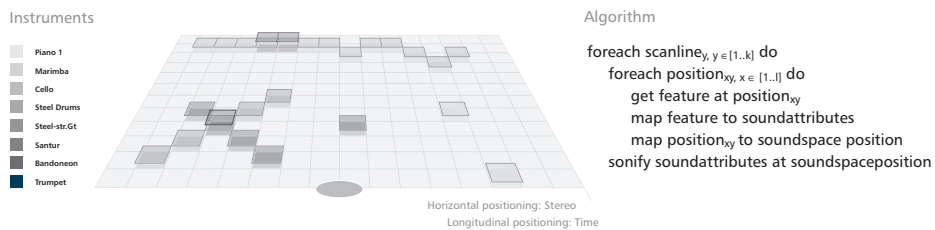


Fig. 3. Sound space modeled sonification: n channels per cell

ture extraction algorithm. The features' orientations (Fig. 2) are mapped to instruments. We use 8 different MIDI instruments corresponding to 8 different orientations. Our pilot investigations showed that raw parameter variations such as changes of tone and pitch are much more difficult to distinguish than complex sound characteristics of known instruments.

After feature extraction our sonification algorithm constantly runs through the image line by line bottom to top. For each line every cell is sonified simultaneously. Corresponding to each orientation found in the image the predefined instrument for this orientation is played at the certain position. The sounds of one scan line are played in parallel using a stereoscopic sound device for auditory localization (Fig. 3).

Using this method the user gets the impression that he is located at the edge of the soundscape. A longitudinal complex stereo sound travels into depth with constantly decreasing volume.

5 Object Modeled Sonification

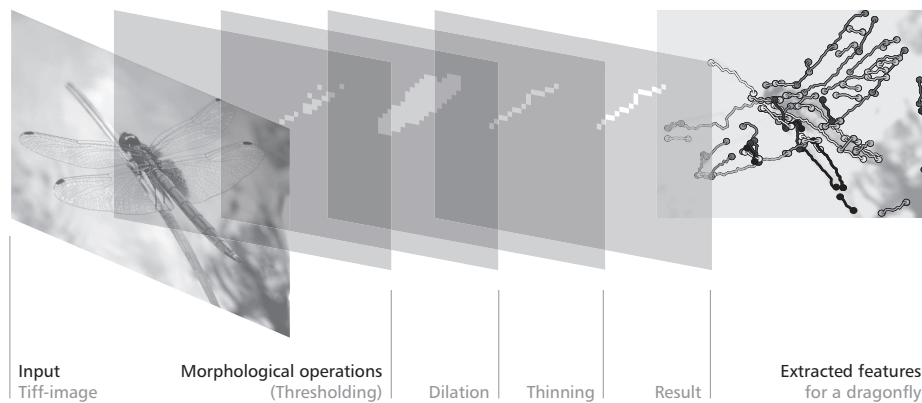


Fig. 4. Visual grouping: Perceptual binding of relatable features

Sound space modeled sonification presented in the previous section does not take into account individual objects. In contrast to this object modeled sonification uses visual features bound to individual objects in the image to generate auditory features bound to individual sound objects. Visual objects are extracted from the original picture as described above (individual steps of the algorithm are depicted in Fig. 4). Object information is stored as a separate feature set containing a consecutive number and a set of points for each object.

These extracted features are the basis for our object modeled sonification. The idea is to consecutively draw the extracted object strokes into soundspace as continuous sound streams. This is done by first mapping the object stroke to an

instrument. For example we simply use the given number of the extracted object as an index to the instrument table. Second, the discrete spatial positions within the stroke are mapped to the sound space taking into account that the listener is centered on the horizontal axis and on the edge of the longitudinal axis. This again introduces the possibility to use stereo sound which is very intuitive for horizontal positioning. The longitudinal positioning is indicated by increasing or decreasing volume.

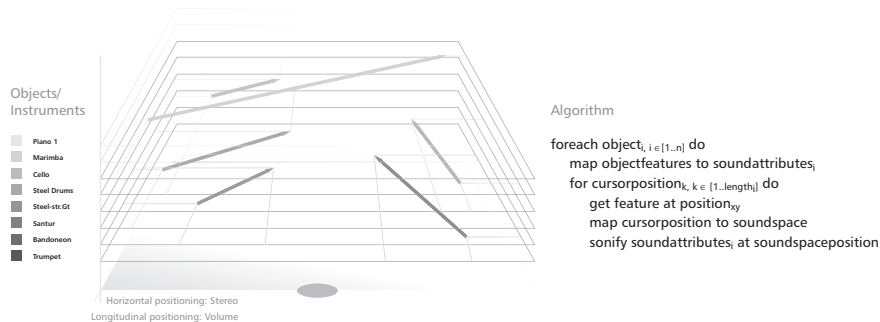


Fig. 5. Object modeled sonification: Free positioning in sound space

Thus, the listener gets the impression that a virtual instrument is moved through sound space. We again used Midi instruments for sonification after experimenting with drawing multiple strokes simultaneously into sound space (Fig. 5). This is only possible because of the sparseness of the salient feature sets we extract from the images. The bottleneck in sonification is the amount of auditory signals distinguishable by an observer. Compared to the sound space modeled approach we are able to present more compact object representations in the object modeled approach. As a consequence more objects can be sonified simultaneously with this approach.

6 Future Work and Conclusion

In contrast to classical sonification approaches which seek to replace visual displays we plan to enrich visual displays with auditory cues in line with the visual contents. We do not believe that the huge potential of sonification lies not in the substitution of the visual sense but in the extension and enhancement. Sonification techniques can be of great use in for example in user guidance or multidimensional data analysis. Spots of interest can be indicated otherwise lost in the vast amount of data. In surveillance or tracking systems small but important changes can be recognized by the application of the aural system.

We presented how salient image features can automatically be detected by computational models of visual attention to be used for sonification. We depicted different possibilities how such features can be transformed into auditory signals

and discussed other sonification approaches previously presented. Our major contribution is the generation of a comprehensible auditory signal generated from an image. We believe that the strong compression of the visual data necessary to allow an user to interpret a sonified image is possible only by employing perceptual models to extract salient and thus relevant visual information. The two presented approaches are applicable for images of any kind. However we believe that the comprehensibility of the sonification and thus which approach to favor depends on the nature of the original image.

Therefore in the near future we plan to evaluate our approaches in application scenarios in surveillance or in peripheral sensing.

In addition we plan the extension of visual display systems in the future. To overcome major limitations of two and three dimensional displays with limited space and vast degree of detail to display sonification attempts can be used for attentional guidance. They can draw attention to and sonify special information of information units. Furthermore we believe that sonification is a proper way to communicate global background information which can for instance be used with ambient displays.

The nature of sound, its linearity, its strictly increasing character, providing a natural order makes it even more interesting for interactive scenarios. Thus in the long run we plan to incorporate sonification approaches into user interfaces to provide a new way of human computer interaction.

References

- [Fischer et al., 2004] Fischer, Bayerl, Neumann, Christobal, and Redondo (2004). Are iterations and curvature useful for tensor voting? In *European Conference on Computer Vision 2004*, volume 3023, pages 158–169. LNCS.
- [Gonzalez and Woods, 2001] Gonzalez, R. C. and Woods, R. E. (2001). *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.
- [Hermann et al., 2000a] Hermann, T., Meinicke, P., and Ritter, H. (2000a). Principal curve sonification. In Cook, P. R., editor, *Proc. of the Int. Conf. on Auditory Display*, pages 81–86. Int. Community for Auditory Display.
- [Hermann et al., 2000b] Hermann, T., Nattkemper, T., Schubert, W., and Ritter, H. (2000b). Sonification of multi-channel image data. In Falavar, V., editor, *Proc. of the Mathematical and Engineering Techniques in Medical and Biological Sciences (METMBS 2000)*, pages 745–750. CSREA Press.
- [Itti et al., 1998] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Machine Intell.*, 20(11):1254–1259.
- [Krüger and Wörgötter, 2005] Krüger, N. and Wörgötter, F. (2005). Symbolic pointillism: Computer art motivated by human brain structures. *Leonardo, MIT Press*, 38(4):337–340.
- [Marr, 1982] Marr, D. (1982). *Vision*. W. H. Freeman and Company, New York.
- [Martins et al., 2001] Martins, A. C. G., Rangayyan, R. M., and Ruschioni, R. A. (2001). Audification and sonification of texture in images. *Journal of Electronic Imaging*, 10(3):690–705.
- [Meijer, 1992] Meijer, P. B. (1992). An experimental system for auditory image representation. *IEEE Transactions on Biomedical Engineering*, 39(2):112–121.

- [Rath and Rocchesso, 2005] Rath, M. and Rocchesso, D. (2005). Continuous sonic feedback from a rolling ball. *IEEE Multimedia*, 12(2):60–69.
- [Simoncelli and Heeger, 1998] Simoncelli, E. P. and Heeger, D. J. (1998). A model of neuronal responses in visual area MT. *Vision Research*, 38(5):743–761.
- [Trucco and Verri, 1998] Trucco, E. and Verri, A. (1998). *Introductory Techniques for 3-D Computer Vision*. Prentice Hall PTR, Upper Saddle River, NJ, USA.
- [Weidenbacher et al., 2005] Weidenbacher, U., Bayerl, P., Fleming, R., and Neumann, H. (2005). Extracting and depicting the 3d shape of specular surfaces. In *Siggraph Symposium on Applied Perception and Graphics in Visualization*, pages 83–86. ACM.