Kinectofon: Performing with Shapes in Planes

Alexander Refsum Jensenius
University of Oslo, Department of Musicology
PB 1017 Blindern, 0315 Oslo, Norway
a.r.jensenius@imv.uio.no

ABSTRACT

The paper presents the Kinectofon, an instrument for creating sounds through free-hand interaction in a 3D space. The instrument is based on the RGB and depth image streams retrieved from a Microsoft Kinect sensor device. These two image streams are used to create different types of motiongrams, which, again, are used as the source material for a sonification process based on inverse FFT. The instrument is intuitive to play, allowing the performer to create sound by "touching" a virtual sound wall.

Keywords

Kinect, motiongram, sonification, video analysis

1. INTRODUCTION

I have for a long time been fascinated by the relationships between visuals and sound. This fascination has led to the exploration of a sonification technique called *sonomotiongram*, in which a *motiongram* is used as the starting point for an inverse FFT process [2]. A motiongram is a spatiotemporal display of motion in a video recording [3], and may visually resemble audio spectrograms. Motiongrams created from a regular video recording has the disadvantage that everything within the recorded frame is visualised and hence sonified. This paper presents a setup in which the depth sensing capabilities of the Kinect sensor is used to create an instrument using the sonomotiongram technique.

2. BACKGROUND

The sonification technique presented here is based on a fairly long tradition of turning images into sound. Notable examples include the Pattern Playback machine from the 1940s [1], the UPIC system from the 1970s [5], and a number of more recent computer-based systems, including Metasynth [7]. There have also been examples of installation pieces using a similar approach, such as Scrappler [4].

3. IMPLEMENTATION

Figure 1 shows an overview of the signal flow in the Kinectofon. The implementation has been done in Max/MSP/Jitter, using components from the open framework Jamoma [6].

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
NIME'13, May 27 – 30, 2013, KAIST, Daejeon, Korea.
Copyright remains with the author(s).

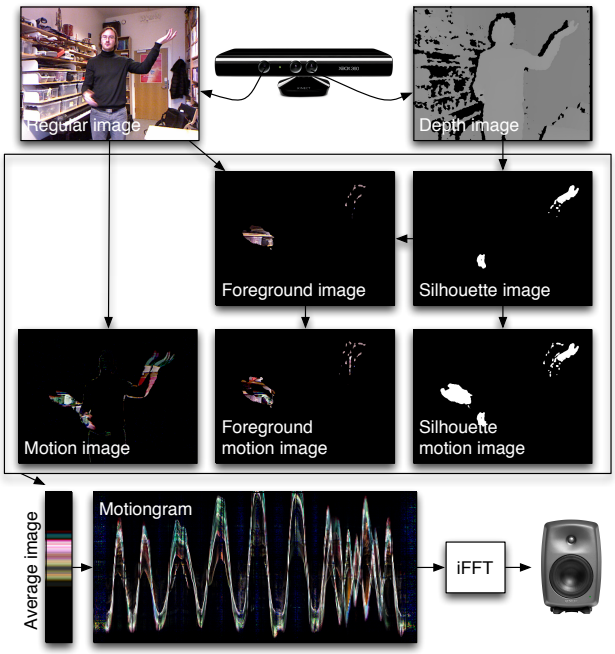

Figure 1: An overview of the Kinectofon signal flow.

Previous implementations of the sonomotiongram technique have been based on creating *motion images* by calculating the frame difference from a regular video stream (see Figure 1). The Kinect sensor device has the benefit of also supplying a depth image, which can be used to create a *silhouette image* of the parts of the body that are within the desired range. The silhouette image can again be used to filter the regular image stream to obtain a *foreground image*. Both the silhouette image and the foreground image can be used with a frame differencing filter, to present only the moving parts in the image. All in all, this gives seven possible image types that can be sent to the motiongram module for further processing. Some examples of the combinatorial outputs, and the basis for the sonification process, can be seen in Figure 2.

The sonomotiongram technique is based on the idea of treating the motiongram as if it were an audio spectrogram. Technically, however, a motiongram is very different from a spectrogram. An audio spectrogram displays the changing energy levels of the frequency bands, while a motiongram is simply a reduced display of a series of motion images. Though motiongrams and spectrograms represent different features, they share one property: the temporal unfolding of shapes of either motion or sound. Furthermore, the Y-axis in a motiongram represents vertical motion, which is often affiliated with pitch. This makes performing with the technique intuitive and easy to learn.

The regular motiongram technique is based on averaging over each row in the image matrix, which means that any motion on the horizontal axes would lead to a visual, and hence sonic, result. In addition to this "average mode," The Kinectofon also has a "slit-scan mode" for creating motiongrams by choosing only one single column in the video matrix. Examples of the visual result of these two different motiongram techniques are shown in Figure 2.

4. DISCUSSION

The sonomotiongram technique was originally developed for analytic applications, but I have later found that its greatest potential may, in fact, be for creative applications. I have performed several concerts with a setup based on a regular video camera. While this has worked well, it has not felt like a proper instrument due to the lack of being able to control the onset of sounds. This is because a motion image created from a regular video stream will represent all motion happening in the frame, and does not allow for selecting, for example, only a part of the body to use in the interaction. The depth image from the Kinect sensor makes it possible to create a virtual "box" within which it is possible to perform. The result is that the performer can move the hands in and out of this box, giving the impression of "touching" the sound. As such, the performer is not only in control of the pitch and timbre of the sound, but can also control the onset of sound more directly and intuitively.

The Kinectofon is intuitive and fun to play, but it can also offer the user a great deal of sonic complexity, dependent on the richness of the incoming video material. Having the possibility to easily switch between the different modes (7 input images and 2 motiongram techniques), makes it possible to easily explore different types of sonic qualities.

Future work includes exploring how video effects can be used to modify the images before the sonification process. For example, adding a motion blur effect to the video image will result in a delay in the sound. It will also be interesting to use gesture recognition as the basis for switching between the different modes.

5. REFERENCES

[1] F. Cooper, A. Liberman, and J. Borst. The interconversion of audible and visible patterns as a basis for research in the perception of speech. *Proceedings of the National Academy of Sciences of the United States of America*, 37(5):318, 1951.

[2] A. R. Jensenius. Motion-sound interaction using sonification based on motiongrams. In *Proceedings of the Fifth International Conference on Advances in Computer-Human Interactions*, pages 170–175, Valencia, 2012.

[3] A. R. Jensenius. Some video abstraction techniques for displaying body movement in analysis and performance. *Leonardo*, 46(1):53–60, 2013.

[4] G. Levin. The table is the score: An augmented-reality interface for real-time, tangible, spectrographic performance. In *Proceedings of the International Computer Music Conference*, New Orleans, LA, 2006.

[5] G. Marino, M.-H. Serra, and J.-M. Raczinski. The upic system: Origins and innovations. *Perspectives of New Music*, 31(1):258–269, 1993.

[6] T. Place and T. Lossius. Jamoma: A modular standard for structuring patches in Max. In *Proceedings of the International Computer Music Conference*, pages 143–146, New Orleans, LA, 2006.

[7] E. Wenger. Metasynth [computer program]. http://www.uisoftware.com/metasynth/, 1998.

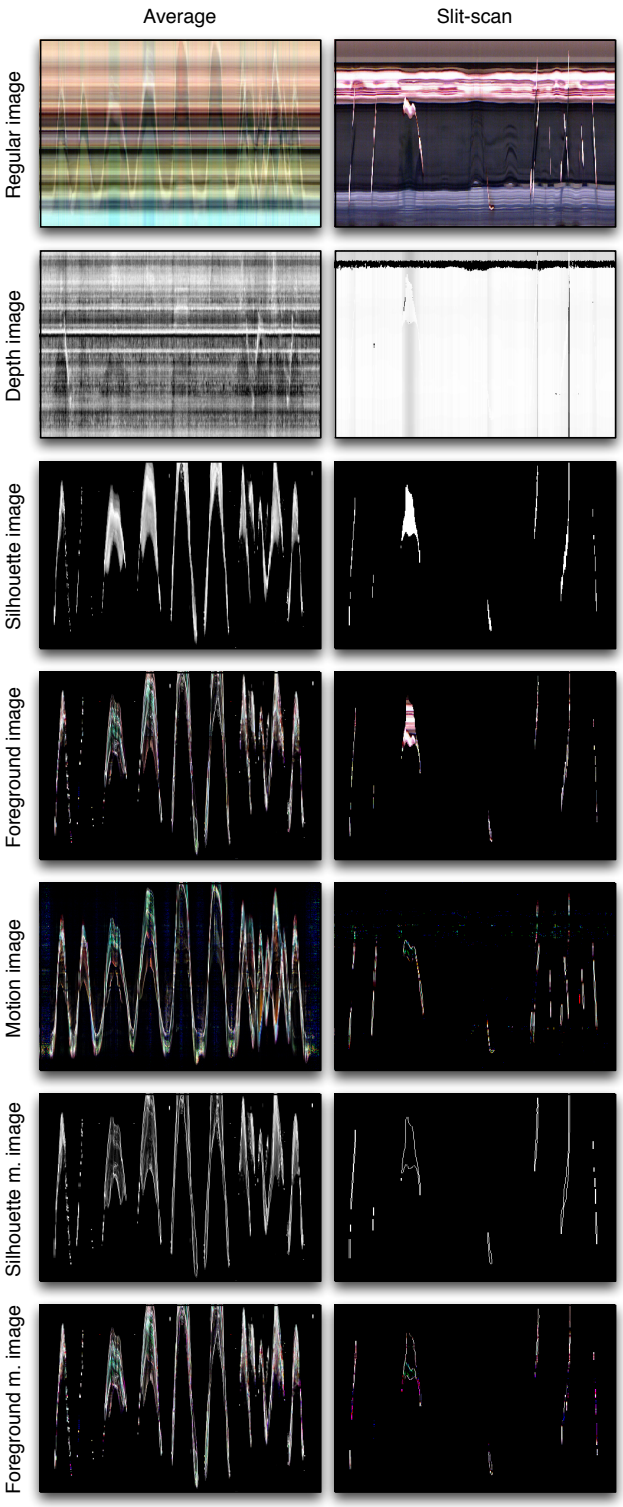

Figure 2: Examples of different types of motiongrams created from the seven different types of video images. All images are created from the same recording of a short series of hand movements. The motiongrams to the left are based on averaging over each row in the video matrix, while the motiongrams to the right are created by picking values from one single vertical line in the middle of the video image.