Audio-Visual Panoramas and Spherical Audio Analysis using the Audio Camera

Adam E. O'Donovan
Perceptual Interfaces and Reality Laboratory
UMIACS, University of Maryland
adam.o@visisonics.com

Ramani Duraiswami
Perceptual Interfaces and Reality Laboratory
UMIACS, University of Maryland
ramani@umiacs.umd.edu

Abstract

Capturing a scene for later or contemporaneous display needs to capture the complex interactions between the source(s) in the scene and the environment. High order spherical Ambisonics and plane-wave analysis are powerful mathematical tools for such scene analysis. The spherical microphone array (and its embodiment in the Audio Camera) is a useful tool for capture and analysis of scenes. Further information about the environment is available from the visual scene.

We present the audio-visual panoramic camera as a tool that greatly simplifies the task of processing audio visual information by providing one common framework for both modalities. Via the Audio Camera [1], we show that microphone arrays can be viewed as a central projection camera that can effectively image the audible acoustic frequency spectrum. We demonstrate a new device, the audio visual Panoramic camera that is composed of a 64 channel spherical microphone array combined with a 5 element video camera array. The combined sensor is capable of real-time audio visual panoramic image generation using state of the art NVidia Graphics cards. It also provides an order-7 ambisonic description of the scene.

1. Introduction

Nearly every biological creature senses the world with both eyes and ears. This is due to the tremendous amount of complementary information in each of these modalities. The visual system conveys pinpoint geometric information about objects in our environment. The acoustic environment conveys information such as speech and does not suffer as badly as vision from issues such as occlusions. For these reasons and many others it is attractive to investigate utilizing both modalities in problems of scene understanding. Microphone arrays have been an attractive tool for audio processing as they provide geometric information about acoustic sources in an environment as well as the ability to spatially suppress noise. However, it is often difficult to calibrate and utilize both microphone arrays and video cameras to perform multi-modal scene understanding. We take the approach that both microphone arrays and video cameras are central projection devices [1] and therefore can be treated in a common imaging framework. This allows the creation of a pre-calibrated multimodal panoramic sensor, The Panoramic Audio Visual Camera, which significantly simplifies the fusion of both audio and visual information.

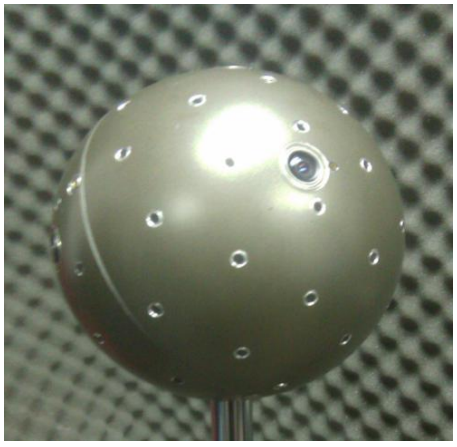

Figure 1: The Audio Visual Panoramic Camera.

2. The Spherical Microphone Array

To generate the acoustic images in the audio visual panoramic camera we utilize the spherical microphone array. There are several benefits to the spherical geometry [2]. The first is that it provides equal spatial resolution in all directions. Additionally, several mathematical simplifications presented in [3] provide efficient algorithms for processing the acoustic information in parallel on commercial graphics cards thus providing real-time capability. The array consists of 64 microphones distributed over the surface of an 8 inch sphere. The microphone signals are then amplified via individual pre-amp circuits and sent to an array of analog to digital converters. The digitized acoustic data is sampled at 44.1 kHz per channel and collected by an FPGA where it is interleaved into a single USB 2.0 Stream. This provides the interface to the PC where the data can be immediately shipped to the graphics processor for real-time processing. Figure 1 shows the Audio Visual Panoramic Camera.

3. Auditory Scene Capture and Playback

The auditory scene can be decomposed into its spherical harmonic components up to order 7. Further, the scene can

Figure 2: Example of the panoramic video image acquired by our device.

also be decomposed into its filtered plane-wave components [4]. These can be used as inputs to creating ambisonic displays using spherical arrays or mixed with HRTFs as discussed in [5] and recreate the auditory scene over headphones.

Figure 3. External ports of the Audio Visual Camera.

4. Panorama Stitching

Due to the fact that the spherical microphone array provides an omni-directional acoustic image of the environment it is highly beneficial to have an omni-directional image of the visual environment as well. In order to generate a panoramic image of the scene we utilize a 5 camera array. The placement of the cameras was selected to avoid all microphones in the spherical microphone array via a spatial optimization. Additionally, the placement was selected such that all directions except those present around the mounting handle are seen by at least one camera. Each of the 5 video cameras are 752x480 color Firewire cameras. The frame acquisition is triggered by the internal audio FPGA to provide synchronization of both the audio and video components of the device. The 5 camera image streams are collected via an internal Firewire that allows an interface to the PC consisting of a single Firewire cable. Figure 2 shows an example of the stitch achieved using our 5 camera panoramic video camera.

5. The Panoramic Audio Visual Camera

Given both the omni-directional acoustic and video images we perform a one time joint audio visual calibration to bring both modalities into a single global coordinate system [1]. Because both the audio and visual cameras are collocated and share a common center of projection we can perform acoustic image transfer onto the video stream as described in [3] to provide a final image that represents both acoustic and visual information present in the environment. The extent of the external cable connections of the device consist of a single USB 2.0 port and a single Firewire port as well as external power. Figure 3 shows the interface ports present at the base of the handle.

6. Conclusion

We present a multimodal panoramic audio-visual camera. We demonstrate that we can present both acoustic and visual panoramic video streams in real-time. By combining both the spherical microphone array and an omni-directional camera array we provide a simple means of sensing the world of light and sound using a single common framework. Many applications of the device are possible, including in auditory display.

Acknowledgement: Partial ONR support is gratefully acknowledged.

7. References

[1] A. O'Donovan, R. Duraiswami, J. Neumann, Microphone arrays as Generalized Cameras for Integrated Audio Visual Processing. IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, 2007

[2] J.Meyer and G. Elko, "A highly scalable spherical microphone array based on an orthonormal decomposition of the soundfield," IEEE ICASSP 2002, 2:1781-1784

[3] A. O'Donovan, R.Duraiswami, N. Gumerov, "Real Time Capture of Audio Images and Their Use with Video" Proceedings 2007 IEEE WASPAA.

[4] .D.N. Zotkin, R. Duraiswami, N.A. Gumerov. Plane-wave decomposition of acoustical scenes via spherical and cylindrical microphone arrays. IEEE Transactions on Audio, Speech & Language Processing, 18:2-18, 2010.

[5] R. Duraiswami, D.N. Zotkin, Z. Li, E. Grassi, N.A. Gumerov, L. Davis, High Order Spatial Audio Capture and its Binaural Head-Tracked Playback over Headphones with HRTF Cues. 119th AES Convention, 2005