ABSTRACT

Justin Alden Wilson: Multimodal Learning for Audio and Visual Processing
(Under the direction of Henry Fuchs and Ming C. Lin)

The world contains vast amounts of information which can be sensed and captured in a variety of ways and formats. Virtual environments also lend themselves to endless possibilities and diversity of data. Often our experiences draw from these separate but complementary parts which can be combined in a way to provide a comprehensive representation of the events. Multimodal learning focuses on these types of combinations. By fusing multiple modalities, multimodal learning can improve results beyond individual mode performance. However, many of today's state-of-the-art techniques in computer vision, robotics, and machine learning rely solely or primarily on visual inputs even when the visual data is obtained from video where corresponding audio may also be readily available to augment learning. Vision only approaches can experience challenges in cases of highly reflective, transparent, or occluded objects and scenes where, if used alone or in conjunction with, audio may improve task performance. To address these challenges, this thesis explores coupling multimodal information to enhance task performance through learning-based methods for audio and visual processing using real and synthetic data.

Physically-based graphics pipelines can naturally be extended for audio and visual synthetic data generation. To enhance the rigid body sound synthesis pipeline for objects containing a liquid, I used an added mass operator for fluid-structure coupling as a pre-processing step. My method is fast and practical for use in interactive 3D systems where live sound synthesis is desired. By fusing audio and visual data from real and synthetic videos, we also demonstrate enhanced processing and performance for object classification, tracking, and reconstruction tasks.

As has been shown in visual question and answering and other related work, multiple modalities have the ability to complement one another and outperform single modality systems. To the best of my knowledge, I introduced the first use of audio-visual neural networks to analyze liquid pouring sequences by classifying their weight, liquid, and receiving container. Prior work often required predefined source weights or visual data. My contribution was to use the sound from a pouring sequence—a liquid being poured into a target container- to train a multimodal convolutional neural networks (CNNs) that fuses mel-scaled spectrograms as audio inputs with corresponding visual data based on video images.

I described the first use of an audio-visual neural network for tracking tabletop sized objects and enhancing visual object trackers. Like object detection of reflective surfaces, object trackers can also run into challenges when objects collide, occlude, appear similar, or come close to one another. By using the impact sounds of the objects during collision, my audio-visual object tracking (AVOT) neural network can correct trackers that drift from their original objects that were assigned before collision.

Reflective and textureless surfaces not only are difficult to detect and classify, they are also often poorly reconstructed and filled with depth discontinuities and holes. I proposed the first use of an audio-visual method that uses the reflections of sound to aid in geometry and audio reconstruction, referred to as ”*Echoreconstruction*”. The mobile phone prototype emits pulsed audio, while recording video for RGB-based 3D reconstruction and audio-visual classification. Reflected sound and images from the video are input into our audio (EchoCNN-A) and audio-visual (EchoCNN-AV) convolutional neural networks for surface and sound source detection, depth estimation, and material classification. EchoCNN inferences from these classifications enhance scene 3D reconstructions containing open spaces and reflective surfaces by depth filtering, inpainting, and placement of unmixed sound sources in the scene.

In addition to enhancing scene reconstructions, I proposed a multimodal single- and multi-frame reconstruction LSTM autoencoder for 3D reconstructions using audio-visual inputs. Our neural network produces high-quality 3D reconstructions using voxel representation. It is the first audio-visual reconstruction neural network for 3D geometry and material representation.

Contributions of this thesis include new neural network designs, new enhancements to real and synthetic audio-visual datasets, and prototypes that demonstrate audio and audio-augmented performance for sound synthesis, inference, and reconstruction.