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. Vision only approaches can experience challenges in cases of highly reflective, transparent, or occluded objects and scenes containing the like.

To address these challenges, this thesis explores coupling multimodal information to enhance task performance through learning-based methods for audio and visual processing. 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. When visual data is obtained from video, the corresponding audio information may be readily available to augment learning and realize the benefits from both audio and visual data. We show that fluid-structure coupling using the added-mass operator enables impact sound synthesis of objects with varying amounts and types of liquid. 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.

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.