CHAPTER 8: SUMMARY AND CONCLUSIONS

This dissertation was motivated by the advances in computer vision and the possibility of realizing the potential benefits of single model audio and multimodal audio-visual learning. Whether by coupling a fluid and structure to form a rigid double body for sound synthesis or fusing audio and visual inputs for object classification, tracking, and reconstruction, audio is readily available for use along with its corresponding images when datasets are generated from video.

Certain conditions also lend themselves more preferably to one modality or a combination of multiple modes. For example, vision-based methods are sufficient for most static objects and scenes. However, reflective and textureless surfaces may be better suited for audio methods since visual data may be ambiguous or changing over time and viewpoint. Finally, audio-visual techniques can use scheduling to use the appropriate inputs given the current state, account for drift error of dynamic objects and scenes, and handle occlusions from cluttered scenes.

8.1 Summary of Results

I have presented a fast and practical method for simulating the sound of non-empty objects containing fluids. This work is enhanced the sound synthesis equation in the rigid body audio pipeline method and was demonstrated for use in interactive 3D systems, where live sound synthesis is important. The key contribution was to account for the fluid force on an object at the fluid-structure boundary. This was achieved by adding pre-processing steps to identify the mesh nodes of a tetrahedralized object that are in contact with the liquid and to apply an added mass operator to those structural boundary nodes and adjacent solid domain nodes. The added mass is applied to the bounding elements in the mass matrix proportional to the liquid's volume and density, which may vary with temperature and/or type of fluid. The technique generalizes to any impermeable tetrahedral mesh representing the rigid objects and inviscid liquids.

To estimate the weight of a liquid poured into a target container, perform overflow detection, and classify liquid and target container, I introduced a novel audio-based and audio-augmented techniques,

Fluid-structure coupling used added mass operator for sound synthesis
Pouring Sequence Neural Network (PSNN) for weight estimation of liquid
Audio-Visual Object Tracker (AVOT)
Echo-Reconstruction: audio-augmented scene reconstruction on mobile devices
3D-MOV: audio-visual LSTM autoencoder for 3D reconstruction of multiple objects from video

Table 8.1: Summary of contributions

in the form of multimodal convolutional neural networks (CNNs). The audio-based neural network uses the sound from a pouring sequence—a liquid being poured into a target container. Audio inputs consist of converting raw audio into mel-scaled spectrograms. Our audio-augmented network fuses this audio with its corresponding visual data based on video images. Only a microphone and camera are required, which can be found in any modern smartphone or Microsoft Kinect. Our approach improves classification accuracy for different environments, containers, and contents of the robot pouring task. Our Pouring Sound Neural Networks (PSNN) are trained and tested using the Rethink Robotics Baxter Research Robot. To the best of our knowledge, this is the first use of audio-visual neural networks to analyze liquid pouring sequences by classifying their weight, liquid, and receiving container.

Existing state-of-the-art object tracking can run into challenges when objects collide, occlude, come close to one another, or appear similar but are of different materials. By using audio of the impact sounds from object collisions, rolling, etc., I presented an audio-visual object tracking (AVOT) neural network that can reduce tracking error and drift. AVOT is trained end to end and uses audio-visual inputs over all frames. Our audio-based technique may be used in conjunction with other neural networks to augment

visually based object detection and tracking methods. It is evaluated in terms of runtime frames-per-second (FPS) performance and intersection over union (IoU) performance against OpenCV object tracking implementations and a deep learning method. Experiments include using the synthetic Sound-20K audio-visual dataset and demonstrating that AVOT outperforms single-modality deep learning methods, when there is audio from object collisions. A proposed scheduler network to switch between AVOT and other methods based on audio onset maximizes accuracy and performance over all frames in multimodal object tracking.

I proposed "*Echoreconstruction*", an audio-visual method that uses the reflections of sound to aid in geometry and audio reconstruction. This system aids in reconstructing reflective and textureless surfaces such as windows, mirrors, and walls that are often poorly reconstructed and filled with depth discontinuities and holes. 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. The 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.

I proposed a multimodal single- and multi-frame neural network for 3D reconstructions using audio-visual inputs. The trained reconstruction LSTM autoencoder 3D-MOV accepts multiple inputs to account for a variety of surface types and views. The neural network produces high-quality 3D reconstructions using voxel representation. Based on Intersection-over-Union (IoU), it is evaluated against other baseline methods using synthetic audio-visual datasets ShapeNet and Sound20K with impact sounds and bounding box annotations. To the best of our knowledge, our single- and multi-frame model is the first audio-visual reconstruction neural network for 3D geometry and material representation.

8.2 Limitations and Future Work

Overall, the immediate next research steps to further enhance audio-visual performance and processing is further analysis of tasks that can capture audio in their datasets and benefit from its signal (e.g. inference, tracking, reconstruction), gating or scheduling of when single or multiple modalities are used, more neural network architectures, loss functions, and fusion models, and augmenting with even more modes.

Sound synthesis for fluid-structure coupling has limitations in the form of simplifications to maintain interactive performance in VR applications. First, while the work assumes that liquids are inviscid, remain steady, and are not mixed, it should be extensible to handle mixed fluids. This remains future work to evaluate. Next, the granularity of the solid mesh discretization also influences the results since the modifications to the mass matrix occur at the level of the mesh nodes. Finally, investigation of acoustic transfer, harmonic pressure, and user evaluation on auditory perception would offer additional insight.

Future directions for analyzing liquid pouring sequences include data augmentation to improve classification accuracy and generalization. As the task involves temporal data, sequential layers can be introduced into the neural network model, such as recurrent, LSTM, or GRU layers or HMM filtering and evaluated for performance. This may be especially helpful for audio only PSNN-A classification at the beginning and end of pouring sequences. Using a multiple output neural network rather than separately trained neural networks for poured weight, content, and target container classification may also help as well as using a ratio of volume over the target container volume or a combination. Finally, further research can explore if this approach can be applied to other granular materials.

Future work for audio-visual object tracking may consist of expanding the size of the training set by annotating more objects in the Sound-20K dataset, increasing the number of object classes that we are predicting, evaluating alternative fusion methods, and performing sensitivity analysis on scaling factors and aspect ratios. This object tracking has been used for audio-visual input for 3D reconstruction of tracked objects. Further investigations can examine how the error introduced by object tracking propagates to reconstruction error. Also, while audio helps classify the material of the reconstructed geometry, we assume a single material classification based on audio alone and apply that to all voxels. Research on classifying material per voxel using both audio and visual data could expand part segmentation research into reconstructing objects with different materials.

In addition to object reconstruction, I also presented enhanced scene reconstruction. To further extend this particular area of audio-visual research, a primary focus could be on the reception, and 3D reconstruction simultaneously and in real time instead of having a staged approach. An integrated approach may prove not only to be more efficient but also more effective by using audio feedback as part of the reconstruction code. Another possible avenue of exploration is to investigate the impact of live audio for training and/or testing our neural network variations.

8.3 Conclusion

In research and practice, sound is a key contributor to the level of immersion and sense of presence in virtual and imaginative environments. A distraction from any of the senses can cause a 'break in presence'. A goal in computer graphics is to continue to enhance rendering pipelines with new technology, methods, and data. While vision-based methods cover many use cases, alternate modalities such as audio can augment the level of detail and coverage of tasks in computer vision, graphics, augmented, and virtual reality. Since many sound models are physics-based and training data generated from video, established visual pipelines and datasets can be extended to generate and use sound based on the same physics and video capture used for visual data. Much of the research conducted on visual data is also relevant to sound sources. This presents opportunities for audio-based research to advance quickly based learnings from decades of vision research as well as novel directions for fusing audio, visual, and other data for multimodal learning.