# CHAPTER 1: INTRODUCTION

## 1.1 Motivation

> *' The screen is a window through which one sees a virtual world. The challenge is to make that world look real, act real, sound real, feel real. '* (Sutherland, 1965)

**Virtual environments:** in addition to sight, sound is also integral to the level of immersion and sense of presence in virtual and augmented reality (Cummings and Bailenson, 2015). Experiencing audio and visual information are synergestic because a distraction from any of the senses can cause a 'break in presence' (Sanchez-Vives and Slater, 2005).

**Sound synthesis:** since both sound and graphics can be physics-based, the graphics pipeline can be naturally extended to generate sound. By performing modal analysis to precompute frequency and damping, real-time sound synthesis can be achieved (O'Brien et al., 2002; Ren et al., 2013a; van den Doel et al., 2001). This is important such that the audio and visual information rendered from interactions of virtual objects with other objects, liquids, and the user reflect the current state of the virtual environment.

**Multimodal learning:** learning-based methods often primarily rely on visual feedback and human interaction. State-of-the-art vision-based techniques for image classification (Deng et al., 2009) and object detection (Liu et al., 2016; Redmon et al., 2016; Ren et al., 2015a) in images and video. The liquid pouring task in robotics is another example of using visual sensing for volume estimation. With many of these methods using video as an input, multimodal learning with both audio and visual data can improve processing and performance. Fused modalities also cover edge cases that can be a challenge for a single model; for example, noise from blur or illumination for visual and noise from other sound sources for audio.

**3D reconstruction:** reconstruction of 3D geometries are both outputs and inputs. A number of algorithms exist to generate 3D shape from 2D and other sensory information. On the other hand, these 3D points are also used as inputs to train neural networks for other downstream tasks such as object classification, segmentation, and tracking.

## 1.2 Scope of this dissertation

There are a number of training datasets, neural network architectures, technologies, and active research areas for multimodal learning, especially in the area of audio and visual data from video. Applications in these areas range from Virtual and Augmented Reality, e.g., sound synthesis, reconstruction, inference, etc. to expanding methods to handle a wider variety of surfaces and scenes, such as illumination, reflectivity, texture, and occlusion. This dissertation focuses on coupling fluid-structure (Section 3) and audio-visual classification (Section 4), tracking (Section 5), and reconstruction (Sections 6 and 7) with demonstrations in multimodal learning and virtual reality.

## 1.3 Thesis Statement

Coupling multimodal information enhances task performance and processing of audio-visual learning-based methods for fluid-structure sound synthesis, analyzing liquid pouring sequences, object tracking, and 3D reconstructions while also allowing for single mode application for special cases.

## 1.4 Contributions of this dissertation

For sound synthesis: (1) Transforming the problem into a single fluid-structure system using the *added mass operator*. (2) Enhancing the rigid-body sound synthesis pipeline with pre-processing steps for objects containing a liquid. (3) Demonstrating the proposed method in interactive 3D VR applications.

For analyzing pouring sequences: (1) Training, validation, and test data generated from audio recordings and video images with ground truth measurements from a digital scale. (2) Audio-based convolutional neural network for multi-class weight estimation and binary classification for overflow detection by robotic systems. (3) Audio-augmented neural network enhancing the audio only based method with fused visual inputs for robots pouring contents into various target containers. (4) Pouring content and target container classification for robots, based on pouring sequence audio data.

The broad contributions of this dissertation are new real-time fluid-structure coupling methods, new audio-visual classification and tracking, and prototype audio-augmented scene reconstruction on mobile devices.