Inverting Audio-Visual Simulation for Shape and Material Perception

Zhoutong Zhang∗1, Jiajun Wu∗1, Qiujia Li2, Zhengjia Huang3, Joshua B. Tenenbaum2, William T. Freeman1,4

1Massachusetts Institute of Technology, 2University of Cambridge, 3ShanghaiTech University, 4Google Research

1. Introduction

Humans perceive objects through both their visual appearance and the sounds they make. Given a short audio clip of objects interacting, humans can recover rich information about the materials, surface smoothness, and the quantity of objects involved [3]. Although visual information provides cues for some of these questions, others can only be assessed with sound. Figure 1 shows an example: objects with different masses and Young’s moduli may have almost identical appearance, but they make different sounds when impacted, and vice versa.

Since collecting large-scale audio recordings with rich object-level annotations is time-consuming and technically challenging, we introduce an alternative approach to overcome such difficulties: synthesizing audio-visual data for object perception. Our data synthesis framework is composed of three core generative models: a physics engine, a graphics engine, and an audio engine. The physics engine takes objects’ shapes, material properties, and initial conditions as input, and then simulates their subsequent motions and collisions. The graphics engine renders videos based on the simulated object motion. The audio engine, built upon previous works [2], synthesizes the audio using the output of physics engine.

2. Synthesis Engine

With our generative model, we built a new synthetic dataset, Sound-20K, with audio-visual information. We show, on both Sound-20K and real-world datasets, that visual and auditory information contribute complementarily to object perception tasks and further demonstrate that knowledge learned on our synthetic dataset can be transferred for object perception on two real-world video datasets, Physics 101 [5] and The Greatest Hits [4].

In addition, the audio and physics engine can perform in real time, which enables us to infer the latent variables that define object shape, material properties and initial pose in an analysis-by-synthesis style. In short, given an audio clip, we aim to find a set of latent variables that could best reproduce it. We use Gibbs sampling over the latent variables and pass them to our synthesize engine. The likelihood function is given by the similarity between the input and output audio. We show that with simple similarity measure, such as $l_2$ distance over spectrogram, such inference scheme performs reasonably well.
The four tiles we used

<table>
<thead>
<tr>
<th>Tiles</th>
<th>Labeled real (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oak</td>
<td>45.3</td>
</tr>
<tr>
<td>Slate</td>
<td>50.5</td>
</tr>
<tr>
<td>Marble</td>
<td>52.6</td>
</tr>
<tr>
<td>Granite</td>
<td>46.3</td>
</tr>
</tbody>
</table>

Figure 3: We validate our audio synthesis pipeline through carefully-designed physics experiments. We record the sound of four tiles of different materials (a-b), and compare its spectrum with our synthesized audio (c) with corresponding physical properties. We also conducted behavioral studies, asking humans which of the two sounds match the image better. We show results in (d).

Figure 4: Sample data and material classification results

Figure 5: Human performance and Gibbs sampling result comparison. The horizontal line represents human performance for each task. Our algorithm closely matches human performance.

We validated the accuracy of our audio synthesis by comparing it with real world recordings. We recorded the sounds made by striking four plates of different materials (granite, slate, oak and marble) as shown in Figure 3b. The audio was measured by exciting the center of the plates with a contact speaker and measuring the resulting vibrations with a piezo-electric contact microphone placed adjacent to the speaker (shown in Figure 3a). All measurements were made in a sound-proof booth to minimize background noise in the recording.

We validated the accuracy of our synthetic sounds by comparing the spectrum of synthetic audio with real recordings. Figure 3c shows the spectrum comparison between the synthetic sound and the real recording of the granite tile. We also designed a human perceptual study in which 95 people were asked to judge whether the recording or the synthetic was more realistic. Table 3d shows the percentage of people who labeled synthetic sounds as real.

4. Experiment Results

We validated the accuracy of our synthetic sounds by comparing it with real world recordings. We recorded the sounds made by striking four plates of different materials (granite, slate, oak and marble) as shown in Figure 3b. The audio was measured by exciting the center of the plates with a contact speaker and measuring the resulting vibrations with a piezo-electric contact microphone placed adjacent to the speaker (shown in Figure 3a). All measurements were made in a sound-proof booth to minimize background noise in the recording.

We validated the accuracy of our synthetic sounds by comparing the spectrum of synthetic audio with real recordings. Figure 3c shows the spectrum comparison between the synthetic sound and the real recording of the granite tile. We also designed a human perceptual study in which 95 people were asked to judge whether the recording or the synthetic was more realistic. Table 3d shows the percentage of people who labeled synthetic sounds as real.

4. Experiment Results

We validated the accuracy of our audio synthesis by comparing it with real world recordings. We recorded the sounds made by striking four plates of different materials (granite, slate, oak and marble) as shown in Figure 3b. The audio was measured by exciting the center of the plates with a contact speaker and measuring the resulting vibrations with a piezo-electric contact microphone placed adjacent to the speaker (shown in Figure 3a). All measurements were made in a sound-proof booth to minimize background noise in the recording.

We validated the accuracy of our synthetic sounds by comparing the spectrum of synthetic audio with real recordings. Figure 3c shows the spectrum comparison between the synthetic sound and the real recording of the granite tile. We also designed a human perceptual study in which 95 people were asked to judge whether the recording or the synthetic was more realistic. Table 3d shows the percentage of people who labeled synthetic sounds as real.

4. Experiment Results
algorithm performs similarly as human subjects in tasks on inferring objects’ shape and material properties, as shown in Figure 5.

References


