Sound provides supervision for visual learning

Abstract: From the clink of a mug placed onto a saucer to the bustle of a busy café, our days are filled with visual experiences that are accompanied by characteristic sounds. These sounds, when paired with their corresponding videos, can provide a rich training signal that allows us to learn visual representations of objects, materials, and scenes. In this talk, I'll first address the material-understanding task of predicting what sound an object makes when it is hit or scratched. I'll present an algorithm that learns to predict plausible soundtracks for silent videos of people striking objects with a drumstick. The sounds predicted by this model convey information about materials and physical interactions, and they frequently fool human subjects in "real or fake" psychophysical studies. I will then apply similar ideas to show that ambient audio -- e.g., crashing waves, people speaking in a crowd -- can be used to learn about objects and scenes. By training a convolutional network to predict held-out sound for internet videos, we can learn image representations that perform well on object recognition tasks, and which contain units that are selective for sound-producing objects.

Bio: Andrew Owens is a graduate student at the MIT Computer Science and Artificial Intelligence Laboratory, working under the supervision of Bill Freeman and Antonio Torralba. Before that, he obtained his B.A. in Computer Science at Cornell University in 2010. He is a recipient of a Microsoft Research PhD Fellowship, an NDSEG Fellowship, and a Best Paper Honorable Mention Award at CVPR 2011.