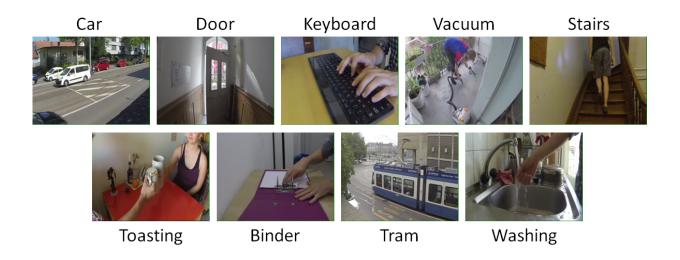# Cow goes moo: Artificial intelligence-based system associates images with sounds

November 15 2016

Credit: Disney Research

The cow goes "moo." The pig goes "oink." A child can learn from a picture book to associate images with sounds, but building a computer vision system that can train itself isn't as simple. Using artificial intelligence techniques, however, researchers at Disney Research and ETH Zurich have designed a system that can automatically learn the association between images and the sounds they could plausibly make.

Given a picture of a car, for instance, their system can automatically return the sound of a car engine.

A system that knows the sound of a car, a splintering dish, or a slamming door might be used in a number of applications, such as adding sound effects to films, or giving audio feedback to people with visual disabilities, noted Jean-Charles Bazin, associate research scientist at Disney Research.

To solve this challenging task, the research team leveraged data from collections of videos.

"Videos with audio tracks provide us with a natural way to learn correlations between sounds and images," Bazin said. "Video cameras equipped with microphones capture synchronized audio and visual information. In principle, every video frame is a possible training example."

One of the key challenges is that videos often contain many sounds that have nothing to do with the visual content. These uncorrelated sounds can include background music, voice-over narration and off-screen noises and sound effects and can confound the learning scheme.

"Sounds associated with a video image can be highly ambiguous," explained Markus Gross, vice president for Disney Research. "By figuring out a way to filter out these extraneous sounds, our research team has taken a big step toward an array of new applications for [computer vision](#)."

"If we have a video collection of cars, the videos that contain actual [car engine](#) sounds will have audio features that recur across multiple videos" Bazin said. "On the other hand, the uncorrelated sounds that some videos might contain generally won't share any redundant features with other videos, and thus can be filtered out."

Once the video frames with uncorrelated sounds are filtered out, a

computer algorithm can learn which sounds are associated with an image. Subsequent testing showed that when presented an image, the proposed system often was able to suggest a suitable sound. A user study found that the system consistently returned better results than one trained with the unfiltered original video collection.

Combining creativity and innovation, this research continues Disney's rich legacy of inventing new ways to tell great stories and leveraging technology required to build the future of entertainment.

These results were recently presented at a European Conference on Computer Vision (ECCV) workshop in Amsterdam. In addition to Jean-Charles Bazin, the research team included Matthias Solèr and Andreas Krause of ETH Zurich's Computer Science Department, and Oliver Wang and Alexander Sorkine-Hornung of Disney Research.

  More information: "Suggesting Sounds for Images from Video Collections-Paper" [PDF, 4.93 MB]

Provided by Disney Research