Andrew Owens

Research Summary
June, 2019

My ultimate goal is to create a multisensory perception system that learns from sight, sound, and touch to physically interact with its environment. So far, my focus has been on creating computer vision methods that learn from other sensory modalities.

Today, computer vision methods need human supervision, such as object labels, to learn about the world. A major goal of the research community has been to remove the need for this supervision—creating systems that, instead, teach themselves by analyzing unlabeled images. This focus on learning from vision alone, however, is likely making the perception problem harder, not easier! Humans, by comparison, have access to many sensory streams, and they learn from *associations between senses*. When a child eats an apple, for instance, she'll not only taste it—she'll also hear it crunch, see its shiny skin, and feel its smooth surface [15]. Psychologists have suggested that these co-occurring sensations provide her with "self-supervision" [5]: that after her snack, she'll associate shininess with smoothness, and crunching with pulp. Vision trains hearing, touch trains vision, *etc.*

Inspired by this idea, I've developed computational models that learn about the world by finding structure in multimodal sensation, and that use what they learn for "downstream" applications. In my research, these have primarily been computer vision applications, such as object and action recognition, but I've also been motivated by machine hearing and robotic manipulation—domains where human-labeled data is often scarce, but unlabeled multimodal data is plentiful.

1 Multimodal perception

Learning sight from sound In [13], I used sound produced by physical interactions to train a computer vision model about material properties. Material recognition is usually posed as a supervised learning problem: someone annotates a photo by hand—say, by assigning labels like *hard* or *soft* to each object—and then trains a model to predict these labels. Instead of having the computer predict labels, I ask it to answer a question that nonetheless requires an understanding of material properties: *what would this object sound like if you hit it with a drumstick?* I trained a model to predict soundtracks for silent videos in which a human physically interacted with a scene by hitting and scratching objects with a drumstick. After training, the sound predictions convey material properties of the objects that were struck. In Figure 1, for instance, the model predicts a low-pitched thud sound when the drumstick strikes dirt, and a high-pitched rattle when it strikes ivy.

Figure 1: What sound do these objects make when you hit them with a drumstick? Our model learns about material properties by predicting "visually indicated" sounds from silent videos.

Physical interactions like these, however, represent only a fraction of the sounds we experience. Often what we hear instead are ambient sounds, such as the babble of a busy café, or the rustle of trees in the wind. In follow-up work [14], I showed that ambient sounds provide a "free" source of supervision for teaching visual models about objects and scenes. I trained a computer vision model to predict audio statistics from video frames, using a large dataset of unlabeled internet videos. This is a task that can only be solved by recognizing objects that make sound. A model that solves it will also have to generalize over a wide range of visual transformations. For instance, we could change an

(a) Localizing sound sources (b) Separating on- and off-screen sound

Figure 3: In [10], we learn an audio-visual representation that (a) localizes sound sources, (b) can be used to separate on- and off-screen sound: *e.g.*, the on-screen speech of a diplomat and off-screen speech of a translator.

image drastically—dim the lighting, rotate the camera, move the objects around—without affecting the ground-truth sound that the model must predict.

Through this process, our model learned features that were useful for solving downstream object- and scene-recognition tasks. It also learned an internal representation that coded for objects that are associated with characteristic sounds, such as faces, crowds, and waterfalls (Figure 2). I've been excited to see that researchers have taken inspiration from our approach, and have used similar ideas to obtain the state-of-the-art unsupervised features for several tasks [1, 8].

Figure 2: Our audio prediction model learns a representation that codes for sound-making objects. These image regions each produce a strong activation for an internal unit of the network [14].

Learning multisensory representations More recently, I've been developing models that not only learn from cross-modal correlations, but that also fuse information from multiple modalities when making decisions. While we normally think of vision and hearing as separate systems, psychologists suggest they are closely intertwined: the motion of a speaker's lips, for instance, can profoundly change what we hear them say [9]. Despite these findings, today's perceptual models tend to be unimodal—there are vision models, hearing models, and not much in between. In [10], I showed that we could learn a video representation that *fuses* these two sensory streams by finding correspondences between visual motion and audio events. To obtain this correspondence, the model predicts whether the visual and audio streams of a video are temporally aligned, or have been synthetically misaligned by an adversary. The model learns to localize *sound sources* such as moving lips or axes striking wood (Figure 3(a)). It also learns features that allow us to solve audio-visual learning tasks, such as action recognition, with high accuracy.

Visual speech understanding As an application of these techniques, I have been developing methods for understanding the *visual* aspects of speech. I created a method for separating on- and off-screen sound sources—for example, removing a translator's voice from a foreign official's speech (Figure 3(b)). The model, which was based on my multisensory video representation [10], was the first method that worked successfully on real-world video footage, *e.g.* television broadcasts. I've also recently studied how humans gesture when they speak [**?**]. My collaborators and I created a "person-specific" gesturing model that, after analyzing hours of footage of a person talking, predicts how they will move their arms during speech. Given only an audio clip, our model predicts the speaker's body pose, and synthesizes a plausible video that depicts them speaking.

Grasping with vision and touch When humans grasp objects, they use many modalities: vision, for example, is useful for planning our grasp, and touch for positioning our hands and selecting which forces to exert. I've explored this idea by developing methods for multisensory grasping.

In [18], my collaborators and I showed that we could infer a fairly subtle material property—how hard or soft an object is—by applying video convolutional networks to optical touch sensing. Then, in a series of papers [2, 3], we applied similar learning methods to robotic grasping (Figure 4). We trained a robot to lift objects by "self-supervised" trial and error: we placed objects on a table, and had the robot repeatedly attempt to lift them, recording which of these attempts were successful and which failed. From these outcomes, we learned a policy that tells the robot which actions it should execute to lift new objects. For example, if the robot determines from touch that it is only gripping an object by the very tips of its fingers, the policy guides it down (toward the object center) to improve its grip.

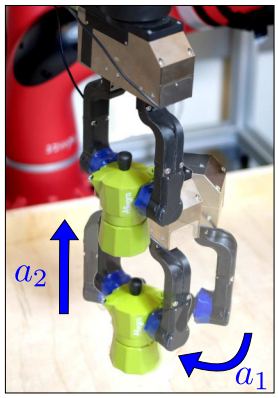

Figure 4: We trained a robot to grasp objects using vision and touch.

2 Other work

My other work deals with creating 3D representations, and detecting fake images.

3D reconstruction To interact with the world, we need to know not just *what* is in a scene, but *what is where*. In my work on grasping (Figure 4), this what-is-where knowledge was acquired implicitly—lacking a 3D model of the world, the robot had to learn how to move its gripper in space via trial and error. In another line of work, I've developed methods that, instead, create and use explicit 3D representations. My work combines *multi-view* reconstruction methods, which use projective geometry to infer highly accurate depth (but only sparsely), and *single-view* methods, which recognize the dense shape of objects by their appearance (but only coarsely).

In [4], my collaborators and I proposed a method for estimating camera pose from photos, known as the *structure from motion* problem. We showed that the problem could be formulated as a graphical model, in a way that let us perform efficient inference with discrete optimization methods. This formulation allowed us to incorporate single-view cues, such as vanishing points, as well as information from other sensors, like GPS. To support work in reconstruction, my collaborators and I created a dataset containing 3D reconstructions of large indoor spaces [16], such as apartments, hotel rooms, and offices. I then used this dataset to train a 3D reconstruction method that combines cues from single-view and multi-view reconstruction. My method recognizes image patches that have a distinctive 3D shape, such as corners and folds [11], while using cues from multi-view geometry to resolve ambiguities.

Visual perception of 3D structure Once we have a 3D model of a scene, we can use it to solve graphics and visualization problems. In [12], I used these models to create a computational model of *camouflage*. While computer vision provides methods for detecting objects, camouflage does the opposite. We posed what we called the *object non-detection* problem: creating object whose appearance is not detectable. I created a texture synthesis method for texturing a 3D object that, when placed into a scene, will be hard for humans to see from every viewpoint they could observe it from. One object that my method created, a box hidden on a bookshelf, is shown in Figure 5.


Figure 5: Two views of a camouflaged box. Its surface texture makes it hard to see from many of the viewpoints that someone might observe it from.

Later, I used 3D models to create visualizations. In [17], my collaborators and I designed a stereo matching algorithm for re-rendering scenes from novel viewpoints without blurring artifacts, and in [19]

Figure 6: The image on the left was created by a computer graphics algorithm [6] that spliced together the two images on the right. Our forensics method [7] detects this manipulation by finding *inconsistency* in the properties of the camera that purportedly took the photo.

we created 3D *motion sculpture* visualizations that reveal the subtle motions of a human body as it performs a complicated action, such as running or dancing.

Detecting fake images Computer vision researchers face an ethical dilemma: as our methods get better, so too do the tools for malicious image manipulation. While malicious editing was once only the domain of the highly skilled—dictators, spy agencies, and unscrupulous photojournalists—recent advances have made it possible to create fake images with only basic computer skills, and social networks have made it easier than ever to disseminate them. One might have hoped that these same advances could also be used to *detect* fake images, but this hasn't been the case: the space of fake images is so vast and diverse that it's not clear how to obtain a representative dataset of "ground truth" fake images to train supervised learning methods. And whatever methods we *do* deploy will quickly become obsolete as our adversaries adapt!

To address these problems, I've begun studying *unsupervised* image forensics methods. As an initial step, my collaborators and I created a method that detects fake images *without any training examples of fake images* [7] (Figure 6). Our model learns to predict photographic metadata (EXIF tags) from images—*e.g.* camera model, compression scheme—and flags photos as being fake if these predictions are inconsistent, such as when a photo's statistics look more like a mixture of two camera models than one. Our method was the first unsupervised forensics method to work on "in-the-wild" manipulations, and it outperformed prior methods trained on *labeled* data.

References

[1] R. Arandjelović and A. Zisserman. Look, listen and learn. *International Conference on Computer Vision (ICCV)*, 2017.

[2] R. Calandra, A. Owens, M. Upadhyaya, W. Yuan, J. Lin, E. H. Adelson, and S. Levine. The feeling of success: Does touch sensing help predict grasp outcomes? *Conference on Robot Learning (CoRL)*, 2017.

[3] R. Calandra, A. Owens, D. Jayaraman, J. Lin, W. Yuan, J. Malik, E. H. Adelson, and S. Levine. More than a feeling: Learning to grasp and regrasp using vision and touch. *Robotics and Automation Letters (RA-L)*, 2018.

[4] D. Crandall, A. Owens, N. Snavely, and D. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *Computer Vision and Pattern Recognition (CVPR)*, 2011.

[5] V. R. de Sa. Learning classification with unlabeled data. *Neural Information Processing Systems (NIPS)*, 1994.

[6] J. Hays and A. A. Efros. Scene completion using millions of photographs. In *Transactions on Graphics (TOG)*, 2007.

[7] M. Huh, A. Liu, A. Owens, and A. A. Efros. Fighting fake news: Image splice detection via learned self-consistency. *European Conference on Computer Vision (ECCV)*, 2018.

[8] B. Korbar, D. Tran, and L. Torresani. Cooperative learning of audio and video models from self-supervised synchronization. *Neural Information Processing Systems (NIPS)*, 2018.

[9] H. McGurk and J. MacDonald. Hearing lips and seeing voices. *Nature*, 1976.

[10] A. Owens and A. A. Efros. Audio-visual scene analysis with self-supervised multisensory features. *European Conference on Computer Vision (ECCV)*, 2018.

[11] A. Owens, J. Xiao, A. Torralba, and W. T. Freeman. Shape anchors for data-driven multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2013.

[12] A. Owens, C. Barnes, A. Flint, H. Singh, and W. T. Freeman. Camouflaging an object from many viewpoints. In *Computer Vision and Pattern Recognition (CVPR)*, 2014.

[13] A. Owens, P. Isola, J. H. McDermott, A. Torralba, E. H. Adelson, and W. T. Freeman. Visually indicated sounds. In *Computer Vision and Pattern Recognition (CVPR)*, 2016.

[14] A. Owens, J. Wu, J. H. McDermott, W. T. Freeman, and A. Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision (ECCV)*, 2016.

[15] L. Smith and M. Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 2005.

[16] J. Xiao, A. Owens, and A. Torralba. SUN3D: A database of big spaces reconstructed using SfM and object labels. In *International Conference on Computer Vision (ICCV)*, 2013.

[17] T. Xue, A. Owens, D. Scharstein, M. Goesele, and R. Szeliski. Multi-frame stereo matching with edges, planes, and superpixels. *In Submission*.

[18] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson. Shape-independent hardness estimation using deep learning and a GelSight tactile sensor. In *International Conference on Robotics and Automation (ICRA)*, 2017.

[19] X. Zhang, T. Dekel, T. Xue, A. Owens, J. Wu, S. Mueller, and W. T. Freeman. MoSculp: Interactive visualization of shape and time. *User Interface Software and Technology (UIST)*, 2018.