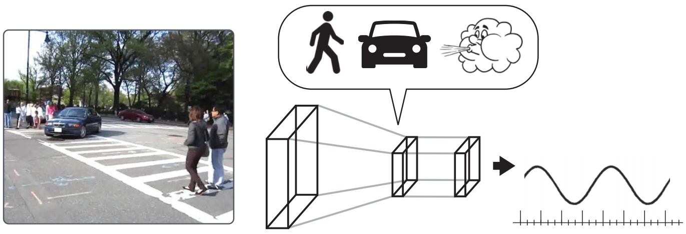
Ambient Sound Provides Supervision for Visual Learning

Andrew Owens Jiajun Wu Josh McDermott William Freeman Antonio Torralba

Motivation

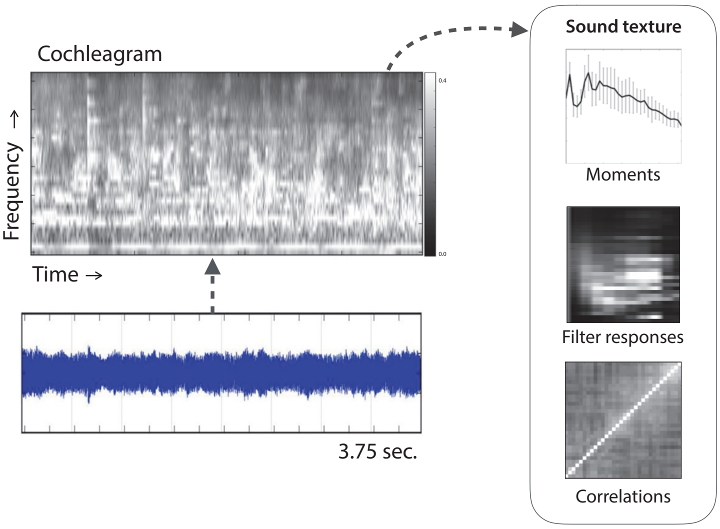
Can we learn image representations using ambient sound — instead of manual annotations — as a supervisory signal?

Task: Predict sound from a video frame. To perform this task well, the model should learn to recognize objects and scenes.



Audio representation

We represent audio using *sound textures* — collections of time-averaged summary statistics [1].



We create a discrete label space by clustering the audio features with k-means, or with an LSH-like binary code.

Sound prediction model

Images grouped by audio cluster

Clustered audio stats.

Results

The image features that our model learns perform comparably to state-of-the-art unsupervised methods on recognition tasks.

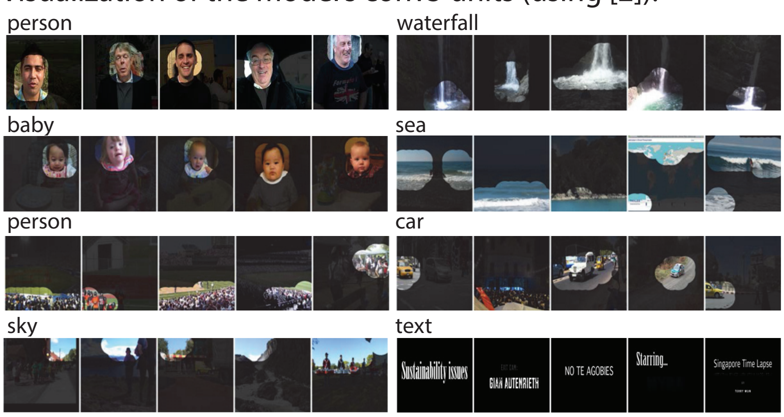
Classification with linear SVM

Method	VOC Cls. (%mAP)				SUN397 (%acc.)			
	max5	pool5	fc6	fc7	max5	pool5	fc6	fc7
Sound (cluster)	36.7	45.8	44.8	44.3	**17.3**	**22.9**	20.7	14.9
Sound (binary)	**39.4**	**46.7**	**47.1**	**47.4**	17.1	22.5	**21.3**	**21.4**
Sound (spectrum only)	35.8	44.0	44.4	44.4	14.6	19.5	18.6	17.7
Texton-CNN	28.9	37.5	35.3	32.5	10.7	15.2	11.4	7.6
K-means (Krähenbühl et al.)	27.5	34.8	33.9	32.1	11.6	14.9	12.8	12.4
Tracking (Wang and Gupta)	33.5	42.2	42.4	40.2	14.1	18.7	16.2	15.1
Patch pos. (Doersch et al.)	26.8	46.1	-	-	9.8	22.2	-	-
Egomotion (Agrawal et al.)	22.7	31.1	-	-	9.1	11.3	-	-
ImageNet (Krizhevsky et al.)	**63.6**	**65.6**	**69.6**	**73.6**	29.8	34.0	37.8	37.8
Places (Zhou et al.)	59.0	63.2	65.3	66.2	**39.4**	**42.1**	**46.1**	**48.8**

Fine-tuning Fast R-CNN

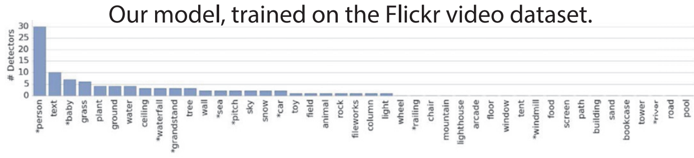
Method	(%mAP)
Random init. (Krähenbühl et al.)	41.3
Sound (cluster)	44.1
Sound (binary)	43.3
Tracking (Wang and Gupta)	44.0
Egomotion (Agrawal et al.)	41.8
Patch pos. (Doersch et al.)	46.6
Calib. + Patch (Krähenbühl et al.)	**51.1**
ImageNet (Krizhevsky et al.)	**57.1**
Places (Zhou et al.)	52.8

Visualization of the model's conv5 units (using [2]):

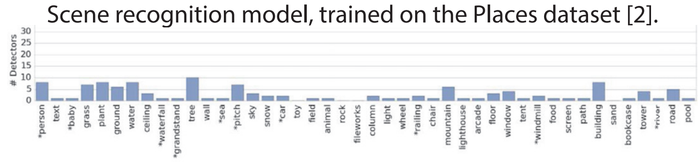


The units tend to be selective for objects that are associated with a characteristic sound (e.g. people and waterfalls).

Our model, trained on the Flickr video dataset.



Scene recognition model, trained on the Places dataset [2].



[1] J. H. McDermott and E. P. Simoncelli. Sound texture perception via statistics of the auditory periphery: evidence from sound synthesis. Neuron, 2011.

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba. Object detectors emerge in deep scene CNNs. ICLR 2015.

[3] A. Owens, P. Isola, J. McDermott, A. Torralba, E.H. Adelson, W.T. Freeman. Visually indicated sounds. CVPR, 2016.

[4] W. Gaver. What in the world do we hear?: An ecological approach to auditory event perception. *Ecological psychology*, 1993.

[5] J. Ngiam, A. Khosla, M. Kim, J., Nam, H. Lee, A.Y. Ng. Multimodal deep learning. ICML 2011.