The Scene Understanding Symposium series is an educational initiative from the Department of Brain and Cognitive Sciences at MIT. The symposium is supported by the McGovern Institute for Brain Research and the Department of Brain and Cognitive Sciences at MIT, the Center for Biological & Computational Learning at MIT, the National Institutes of Health Graduate Training Grant (T32 EY013935 and T32 MH020007), and the National Science Foundation (No. 0546262). Artwork by Bettlann McKay.

80 million tiny images
Antonio Torralba\textsuperscript{1}, Rob Fergus\textsuperscript{2}, Yair Weiss\textsuperscript{3} and William Freeman\textsuperscript{1}
\textsuperscript{1}Massachusetts Institute of Technology, \textsuperscript{2}New York University, \textsuperscript{3}Hebrew University

Object detection and recognition is generally posed as a matching problem between the object representation and the image features (e.g., aligning pictorial cues, shape correspondence, constellations of parts, etc.) while rejecting the background features using an outlier process. In this work, we take a different approach: we formulate the object detection problem as a problem of aligning elements of the entire scene. The background, instead of being treated as a set of outliers, is used to guide the detection process. Our approach relies on the observation that when we have a big enough database then we can find with high probability some images in the database very close to a query image, as in similar scenes with similar objects arranged in similar spatial configurations. If the images in the retrieval set are partially labeled, then we can transfer the knowledge of the labeling to the query image, and the problem of object recognition becomes a problem of aligning scene regions. But, can we find a dataset large enough to cover a large number of scene configurations? Given an input image, how do we find a good retrieval set, and, finally, how do we transfer the labels to the input image? We will use two datasets; 1) the LabelMe dataset, which contains more than 10,000 labeled images with over 180,000 annotated objects. 2) The tiny images dataset: A dataset of weakly labeled images with more than 79,000,000 images. We use this database to perform object and scene classification, examining performance over a range of semantic levels.

Schedule of Talks

Natural scenes: The power of 10s

8:55 Opening remarks

9:00 80 million tiny images
Remembering thousands of natural images with high fidelity
Aude Oliva, George Alvarez, Talia Konkle and Timothy Brady
Massachusetts Institute of Technology

The human visual system has been extensively trained to deal with objects and natural images, giving it the opportunity to develop robust strategies to quickly identify categories and exemplars. Although it is known that the memory capacity for images is massive (Standing, 1973), the fidelity with which human memory can represent such a large number of images is an outstanding question.

We conducted three large-scale memory experiments to determine the details remembered per item, by systematically varying the amount of detail required to succeed in subsequent memory tests. Our results show that contrary to the commonly accepted view that long-term memory representations contain only the gist of what was seen, long-term memory can store thousands of items with a large amount of detail per item. Further, item analyses reveal that memory for an object or a natural image depends on the extent to which it is conceptually distinct from other items in the memory set, and not necessarily on the featural distinctiveness along shape or color dimensions. These findings suggest a “conceptual hook” is necessary for maintaining a large number of high-fidelity representations. Altogether, the results present a great challenge to models of object and natural scene recognition, which must be able to account for such a large and detailed storage capacity.

Creating and exploring a large photorealistic virtual space
Josef Sivic1, Biliana Kaneva1, Antonio Torralba1, Shai Avidan2 and William Freeman1
1Massachusetts Institute of Technology, 2Adobe

We present a system for exploring large collections of photos in a virtual 3D space. Our system does not assume the photographs are of a single real 3D location, nor that they were taken at the same time. Instead, we organize the photos according to scene types, such as city streets or skylines, which we call themes, and let users navigate within each theme using intuitive 3D controls that include pan, zoom and rotate. Themes allow us to maintain a coherent semantic meaning to the tour, while matching on the coarse geometry of the scene allows us to create a "being there" impression, as if the images were of a particular location. We present results on a collection of several millions of images downloaded from Flickr and broken into themes that consist of a few hundred thousands images each.

Using explicit encoding models to identify natural images from human brain activity
Jack L. Gallant, Kendrick N. Kay, Thomas Naselaris and Ryan J. Prenger
University of California, Berkeley

An important line of current research in visual neuroscience focuses on developing an algorithm that can accurately decode perceptual content from evoked brain activity (as measured by fMRI). Current classification-based decoding approaches have been limited to decoding brain activity evoked by simple stimuli (e.g. gratings, pre-segmented objects) that were similar to those used to train the decoder. We are developing a more general Bayesian approach that can decode brain activity evoked by completely novel natural images. In this approach we estimate explicit receptive field models that describe how the pooled activity of a latent population of visually-selective neurons is reflected in the activity of individual voxels. In theory these encoding models can be used to solve many decoding problems. Here we show that they can be used to identify which specific natural image was seen by an observer, even if that image was drawn from a vast pool of completely novel images. Our results suggest that it may soon be possible to reconstruct a picture of a person’s visual experience from brain activity measurements alone.
Perceptual organization in images of natural scenes
Jitendra Malik
University of California, Berkeley

The Gestaltists argued for the primacy of generic cues such as proximity, similarity and good continuation for determining perceptual organization. Max Wertheimer was well aware of the arguments in favor of top down knowledge of "familiar configurations", and while he allowed a role for it, he denied its primacy on both logical and empirical grounds. While their view is still taken seriously in perception and psychophysics, in computer vision the experience of dealing with real images has led to an opposing world view: Low or Mid-level cues provide but little power in the presence of local ambiguity in images. So the best practical strategy is to run a Canny edge detector or some other mechanism for estimating local oriented edges, and rely on the contact with memory of previously seen objects and scenes to do the real work.

At Berkeley we have taken a contrarian view for more than a decade, and finally we have clear evidence that one can indeed do significantly better. The key ingredients are (1) carefully modeling and using color and texture for finding boundaries and not just brightness (2) incorporating global information such as curvilinear continuity and closure from a spectral partitioning framework (3) capturing mid-level cues in a conditional random field framework. The F-measure (harmonic mean of precision and recall) of performance on the Berkeley Segmentation dataset is 0.58 for Canny edges, and 0.70 for our latest boundary finder, whereas humans (making use of high level knowledge as well) have a F-measure of 0.79. There is direct evidence from research in the Cipolla and Schmid groups that shows that better contours translate to better performance at object recognition.

Over the years many colleagues have contributed to this research agenda: Jianbo Shi, Thomas Leung, Serge Belongie, David Martin, Charless Fowlkes, Xiaofen Ren, Michale Maire and Pablo Arbelaez.
Attention and the neural representation of natural images

1:30 How are visual signals transformed along the primate object recognition pathway?
Nicole C Rust and James J DiCarlo
Massachusetts Institute of Technology

Current hypotheses suggest that the pathway involved in object recognition, the ventral visual stream, transforms a “local” representation of image elements into a “global” representation based on their arrangement. Because we lack a systematic understanding of what comprises the “global” configurations of natural images, these hypotheses have been difficult to test. In contrast to traditional single-neuron approaches, we probed the ability of population neurons at different levels of the ventral pathway to represent classes of images with natural and manipulated statistics. Specifically, we constructed scrambled images that have the same local, oriented frequency structure as natural images but configured randomly [Portilla & Simoncelli, 2000]. We recorded the responses of neurons in a mid-level visual area (V4) and a high-level visual area (anterior inferotemporal cortex, IT) to sets of natural and scrambled images while monkeys performed an object detection task. We found that V4 neurons responded similarly to both image sets whereas IT neurons responded much more robustly to the natural over the scrambled images. Likewise, discrimination between different members of each image class by the V4 population was only slightly reduced for the scrambled versus the natural images, whereas discriminability by the IT population was considerably degraded for the scrambled as compared to the natural images. These results suggest that the representation of images in the ventral stream is indeed transformed from “local” early on to “global” at later stages of processing, and that we can experimentally observe that transformation across cortical stages. Furthermore, these results suggest that information about non-natural images is encoded at early stages of processing but is then disregarded as signals are transformed along the ventral visual stream.

1:55 Salience maps and the neural representation of attention
James Mazer
Yale School of Medicine

During natural vision a vast amount of data impinges on the visual system. Attention is one mechanism the brain uses to reduce computational demands imposed by this large data volume by selectivity filtering visual inputs based on behavioral demands. We are exploring the idea that the visual system instantiates one or more salience maps containing neurons able to dynamically change stimulus preferences to facilitate target detection during visual search and natural visually guided behavior. We are also investigating the coordinate system used to represent attentional signals and find that the neural representation of salience and spatial attention is intrinsically retinotopic, even when task demands require subjects to use a spatiotopic or head-centered reference frame. This means that salience maps must be recomputed or updated after each eye movement.

2:20 The ‘Parahippocampal Place Area’ responds selectively to high spatial frequencies in humans and monkeys
Reza Rajimehr, Kathryn Devaney, Jeremy Young, Gheorghe Postelniciu and Roger Tootell
NMR Martinos Center, Massachusetts General Hospital

Defining the exact mechanisms by which the brain processes visual objects and scenes remains an unresolved challenge. Valuable clues to this process have emerged from the demonstration that clusters of neurons (‘modules’) in primate inferior temporal (IT) cortex apparently respond selectively to specific categories of visual stimuli such as places. Here we show that at least part of the higher-order ‘place-selective’ response in the Parahippocampal Place Area (PPA) results from lower-level selectivity for high spatial frequencies in 3D shapes and natural images. This spatial selectivity was demonstrated in multiple fMRI tests, in both humans and macaques. High spatial frequencies may be emphasized in PPA to enhance object borders and scene details, during place-related cortical processing (e.g. for spatial navigation).
Scene recognition: Behavioral perspective

3:05 Highly Efficient Search for Arbitrary Objects in Natural Scenes
Jeremy M. Wolfe¹, George Alvarez² and Aude Oliva²
¹Harvard Medical School, ²Massachusetts Institute of Technology

Most visual search experiments involve search for a fixed target (e.g. red vertical lines) or target set (e.g. vowels) in a highly artificial display of items isolated on a blank background. Even when the search stimuli are scenes, the target set tends to remain fixed (e.g. search for people) while a new scene is presented on each trial. In contrast, many, if not most, real search tasks involve a relatively stable scene (e.g. the kitchen) and a succession of searches for arbitrarily different targets (e.g. where is my spoon? the cat? the can of cat food? the plate? Etc.). We report on the surprisingly high efficiency of such searches, as calculated by the slope of RT x set size functions. Additionally, we find that RT does not decrease over repeated searches through the same scene for different objects. However, there is a massive RT benefit on the second search for an object in a fixed scene.

3:30 Memory for viewpoint changes in naturalistic scenes
Monica S. Castelhano¹, Alexander Pollatsek² and Keith Rayner²
¹Queen's University, ²University of Massachusetts at Amherst

How do we represent viewpoint changes of a scene? To investigate this question, two study images were shown from either the same or different viewpoints of a scene. Immediately following, participants indicated whether a test image was identical to a study image. When shown a different viewpoint of the scene, the viewpoint could be taken between the two study images (Interpolation condition) or outside the two study images (Extrapolation condition). In Experiment 1, there were many more errors for interpolated than extrapolated views. In Experiment 2, the two study images were from two different scenes, and were tested with the equivalent different viewpoint conditions. Results confirmed that this effect was not due to a stimulus anomaly or to a lag effect on memory accuracy. We developed memory models of two hypotheses: 1) the views in memory are stored and compared to the test image independently (View-centered hypothesis) or 2) the views in memory are compared to the test image as a function of the sum of the strengths of the degree of match of the memory views to the test image (Combined hypothesis). We found that the memory performance was consistent with the Combined hypothesis. Therefore, we proposed that the visual system does not construct a common memory representation; however both images are simultaneously involved in the retrieval process.

BCS special seminar

4:00 Minimal scenes, maximal challenges
Irving Biederman
University of Southern California

What distinguishes a scene from a collection of objects? A dramatic cognitive benefit is gained when objects are depicted in a relation that tends to be described with a contact preposition or a gerund (“cup on a clock,” “hand holding shoe”) compared to a listing or a conjunction (“cup, clock,” “hand and shoe”). However, objects in any array, even when not in close proximity, can be described with spatial prepositions (“above,” “side of”). We show that such spatial relations are explicitly represented as observed in fMRI-adaptation experiments. The activation of these relations may be a component of the structural descriptions mediating the understanding of complex, novel scenes in a fraction of a second.
Poster session

**Sex differences in visual cue use during spatial navigation**
Xiaoqian J. Chai and Lucia F. Jacobs
*University of California, Berkeley*

Sex differences in visual cue use during spatial navigation were investigated in several 3D virtual environments. We examined two distinct classes of environmental cues based on the parallel map theory (PMT) by Jacobs and Schenk (2003): directional (global, compass) cues and positional (local landmark) cues. Directional cues include gradient (graded information) and distant objects. Positional cues are proximal objects which provide accurate position of the target. Our data suggest that directional cues are used more heavily by males than females. Males outperformed females when an additional directional cue (geographic slant) was added into the test environment, and were more affected when directional cues were removed from the environment. When participants were forced to learn the target location using either positional cues or directional cues, males were more accurate at estimating the target location overall and the difference in male-female performance was greater when only the directional cues were present. Self-report spatial representation revealed the same pattern. These findings support the PMT model which predicts male preference on directional cues and provide insights for future research direction on the neural basis of spatial navigation.

**Visual recognition memory in the absence of meaning**
Garga Chatterjee, Richard Russell, Anna Tong and Ken Nakayama
*Harvard University*

Recognition memory for pictures of verbalizable and categorizable objects is excellent. (Standing, 1973). Here we study visual memory for images that are heterogenous and easy to distinguish perceptually, yet are non-verbalizable and non-categorizable. Subjects viewed 100 abstract art images, for 1s each followed by 1s interval. Immediately thereafter subjects were tested on the set of 100 pictures in a 3AFC old/new task. The two new foil images at each test trial were also abstract art. Neither the old images nor the foils were repeated in other test trials. Matching performance by converting to d’ scores, the performance for our task was 65.6% (d=1.1), which was far worse than the 90% (d’=1.95) for regular images and 96% (d’= 2.5) for vivid images reported by Standing. The results indicate that lack of verbalizability and categorizability impairs visual recognition memory of otherwise highly distinctive images. In the absence of meaning, recognition is hampered.

**Dude, where’s my car? Developmental object agnosia without prosopagnosia**
Laura Germine (1), Nathan Cashdollar (2), Emrah Duzel (2) and Brad Duchaine (2)
(1) Harvard University, (2) University College London

Pure cases of developmental prosopagnosia demonstrate that developmental deficits can selectively impair face recognition while sparing non-face recognition. It is not clear, however, if the opposite developmental dissociation exists. Here, we present the case of A.W., a 19 year old woman with significantly impaired recognition of visual object exemplars and scenes, but with normal face recognition and spatial memory. Together with developmental prosopagnosia, A.W.’s results reveals a developmental double dissociation between face and object recognition at the within-category level. Her dissociation demonstrates that the development of face recognition does not require intact non-face recognition. Further, her data hint at dissociations between other types of visual memory, such as memory for complex versus purely topographical scene information.
The galois lattice model of visual search
Severine Merand and Adam Reeves
Northeastern University

We present a new formalism for visual search (oddity) in which targets and distractors can be classified according to a Galois lattice. Features are presumed to be all about equally salient, so only feature combinatorics matters. Features placed at lattice nodes at locations between the union and the intersection of all the features, arrows between the modes indicating connections. Search difficulty is predicted by the complexity of the lattice, the position of the target within the lattice, and the consistency of lattice structures across trials.

The role of color in real-world scene contextual cueing
Krista A. Ehinger (1) and James R. Brockmole (2)
(1) Massachusetts Institute of Technology, (2) University of Edinburgh.

Although color is a common manipulation in visual experiments with abstract stimuli, the importance of color in natural scene viewing tasks has been debated. In a series of experiments, we sought to determine whether colour is used to guide visual search in contextual cueing with real-world scenes. In Experiment 1, participants searched for targets in natural scenes shown repeatedly in one of three conditions: natural colors, unnatural colors, and unnatural colors which changed with every repetition of the scene. We found a reliable contextual cueing effect, but no difference in the rate of learning across the three groups. In Experiment 2, we demonstrated that learned target locations transfer when colors of the scenes are altered. In Experiments 3 and 4, we repeated the previous experiments using scenes from a color-diagnostic natural image category (parks) and obtained similar results. These findings indicate that color is not used to guide search in real-world scene contextual cueing and suggest that color plays a limited role in scene memory and recognition.

Exploiting domain specific saliency for event classification
Vidit Jain (1), Amit Singhal and Jiebo Luo (2)
(1) Univ. of Massachusetts Amherst, (2) Eastman Kodak Company

Classifying an event captured in an image is interesting for many reasons. It facilitates generation of meta-information such as annotation in the form of textual tags, or cluster the images according to a variety of criteria. Consider a typical set of consumer images with sports related content. These images are mostly taken by amateur photographers, and often at a distance. Typical recognition tasks are formidable on these images in the absence of manual annotation or other sources of information such as GPS. Identifying the sporting event in these images can be seen as building context for further identification and annotation tasks. We propose to discriminate among sports by exploiting the domain specific saliency in the appearance of the playing surfaces and markings on them. To this end, we present a variation of the hidden-state conditional random field that selects a subset of the observed features suitable for classification. In particular, the inferred hidden variables represent the surface orientation, and select the features on the playing surface (assumed to be horizontal) for the sports images. We demonstrate the utility of this model on consumer images collected from the internet.
Objects in context
Andrew Rabinovich, Carolina Galleguillos and Serge Belongie
University of California, San Diego

Abstract: In the task of visual object categorization, semantic context can play the very important role of reducing ambiguity in objects’ visual appearance. However, the majority of newly proposed object recognition models -- whether generative or discriminative -- only consider the intrinsic object features. In this work we introduce a novel approach to object categorization that incorporates two types of context -- co-occurrence and relative location -- with local appearance-based features. Our approach, named CoLA (for Co-occurrence, Location and Appearance), uses a conditional random field (CRF) to maximize object label agreement according to both semantic and spatial relevance. We compare two sources of semantic context: one learned from training data and another queried from Google Sets. To capture spatial constrains, we model relative location between objects using simple pairwise features. By vector quantizing this feature space, we learn a small set of prototypical spatial relationships directly from the data. The overall performance of the proposed framework is evaluated on the PASCAL and MSRC datasets. Our findings conclude that incorporating context into object categorization greatly improves categorization accuracy. We evaluate our results on two challenging datasets: PASCAL 2007 and MSRC. The results also show that combining co-occurrence and spatial context improves accuracy in as many as half of the categories compared to using co-occurrence alone.

A multinomial processing tree model of change blindness and change detection
Emmanuelle Boloix
Massachusetts Institute of Technology

In most change blindness studies, the failure to report changes is attributed to either a representational failure (the pre-change scene is not represented in memory) or to a comparison failure (pre- and post-change scenes are not compared). Here, I propose a Multinomial Processing-Tree (MPT) model, which determines the relative contributions of representational and comparison failures to change blindness. MPT models are statistical models used to measure latent cognitive processes from observable raw data. Cognitive processes are represented as model parameters, their respective weights can be assessed, and the fit of the model to the empirical data can be evaluated via goodness-of-fit tests (Batchelder & Riefer, 1999). The MPT model I propose assumes that visual information can be either represented in memory or non-represented, and, if represented, that visual representation can be compared to the currently-displayed view or not. Model parameters have been set to be consistent with various theories of scene representation (Hollinworth, 2004; Irwin & Zelinski, 2002; Rensink, 2000; Triesch et al, 2003). Here, I ran one change blindness experiment to test the model’s predictions. Observers (N=132) were shown a map of a virtual village and had to perform a 7-stage route in the map. Changes occurred on task-relevant or task-irrelevant objects. Observers were instructed to complete the route as quickly as possible, and to report perceived changes. Our results showed that observer’s change detection performance was best modelled when MPT parameters were set according to a Task-Relevance theory of scene representation. We then used the model to quantify how much of change blindness is due to a comparison failure or to a representational failure. Results show that differences in change detection performances are due to the quality of representation in memory, which depends on object’s relevance for the task, and not to a comparison process (the model’s weighting of the memory representation parameter varies from 19% to 82%, whereas the weighting of the memory comparison parameters remains stable in the range from 28% to 38%). Implications for several theories of scene representation are discussed.
Visual scene and object analysis from web images and text
Nicholas Morsillo, Chris Pal and Randal Nelson
University of Rochester

Abstract: Web image search based on keywords is a powerful way to gather data for scene understanding research. However, the data gathered presents difficulties due to substantial label noise. We propose a technique for semi-supervised visual concept learning from noisy web search results. Our approach is based on a novel probabilistic graphical model composed of both directed and undirected elements. We use this approach to learn visual object and scene categories from images and text, robustly accounting for noisy retrieval results returned by search engines. Web images belonging to the learned categories are suitable for building large image datasets with accurate labels.

Discovering visual senses using hidden topic models of images and associated text context
Kate Saenko and Trevor Darrell
Massachusetts Institute of Technology

We propose a method for discovering visual category senses using hidden topic models of images and their associated text context. Online image search engines such as Google often return results containing multiple sense of the search term (e.g. 'mouse' can refer to a device or an animal). Hidden topic models have the potential to separate the senses, but are typically applied to only images or only text documents. We take a multi-modal approach, using similarity in both text and image space, with the images and their textual contexts represented as bags of "words". Our method discovers latent factors in both domains, and uses this structure to improve retrieval of the correct object class. We present a qualitative analysis of the different senses of each keyword discovered by the text model, as well as a quantitative analysis of re-ranking performance on a dataset consisting of the results of five keyword searches.

Transfer learning for image classification with sparse prototype representations
Ariadna Quattoni, Trevor Darrell and Michael Collins
Massachusetts Institute of Technology

To learn a new visual category from few examples, prior knowledge from unlabeled data as well as previous related categories may be useful. We develop a new method for transfer learning which exploits available unlabeled data and an arbitrary kernel function; we form a representation based on kernel distances to a large set of unlabeled data points. To transfer knowledge from previous related problems we observe that a category might be learnable using only a small subset of reference prototypes. Related problems may share a significant number of relevant prototypes; we find such a reduced representation by performing a joint loss minimization over the training sets of related problems with a shared regularization penalty that minimizes the total number of prototypes involved in the approximation. This optimization problem can be formulated as a linear program that can be solved efficiently. We conduct experiments on a news-topic prediction task where the goal is to predict whether an image belongs to a particular news topic. Our results show that when only few examples are available for training a target topic, leveraging knowledge learnt from other topics can significantly improve performance.
Unsupervised distributed feature selection for multi-view scene recognition
C. Mario Christoudias, Raquel Urtasun and Trevor Darrell
Massachusetts Institute of Technology

Scene recognition accuracy can be improved when information from multiple views is integrated, but information in each view can often be highly redundant. We consider the problem of distributed scene recognition or indexing from multiple cameras, where the computational power available at each camera sensor is limited and communication between sensors is prohibitively expensive. In this scenario, it is desirable to avoid sending redundant visual features from multiple views, but traditional supervised feature selection approaches are inapplicable as the class label is unknown at the camera. In this paper we propose an unsupervised multi-view feature selection algorithm based on a distributed compression approach. With our method, a Gaussian Process model of the joint view statistics is used at the receiver to obtain a joint encoding of the views without directly sharing information across encoders. We demonstrate our approach on recognition and indexing tasks with multi-view image databases and show that our method compares favorably to an independent encoding of the features from each camera.

A topological approach to hierarchical segmentation using mean-shift
Sylvain Paris (1) and Fredo Durand (2)
(1) Adobe, (2) Massachusetts Institute of Technology

Mean shift is a popular method to segment images and videos. Pixels are represented by feature points, and the segmentation is driven by the point density in feature space. In this paper, we introduce the use of Morse theory to interpret mean shift as a topological decomposition of the feature space into density modes. This allows us to build on the watershed technique and design a new algorithm to compute mean-shift segmentations of images and videos. In addition, we introduce the use of topological persistence to create a segmentation hierarchy. We validated our method by clustering images using color cues. In this context, our technique runs faster than previous work, especially on videos and large images. We evaluated accuracy with a classical benchmark which shows results on par with existing low-level techniques, i.e. we do not sacrifice accuracy for speed.

Look before you leap: Effects of delaying visual search on scene context learning
Barbara Hidalgo-Sotelo and Aude Oliva
Massachusetts Institute of Technology

One characteristic of many real-world search tasks (e.g. finding a pen on cluttered desk, or a remote control in the living room) is that they are performed in visual environments that are familiar to the observer. Although the content of a familiar environment will differ across individuals, one shared attribute is that a prior history of interacting within that environment may provide a source of information. In this study, we investigate how memory retrieval contributes to search performance while experience is accumulating in familiar environments.

We tracked observer's eye movements while they searched for a target-person in scenes; a subset of the scenes were repeated throughout the experiment. A delay at central fixation on the scene was enforced before observers could initiate their first eye movement. Critically, we compared search performance between two groups of observers who were delayed for either a short (300 ms) or long (1300 ms) duration. At the beginning of the experiment, no difference in search performance is observed between the groups. Following several block of repetition, however, faster search times, fewer fixations, and shorter cumulative saccade distance were exhibited by the group with a long delay preceding (overt) search initiation. These results suggest that achieving memory retrieval prior to initiating search may enhance the efficacy of eye guidance in familiar environments.
Thousands of scenes
Talia Konkle, Tim Brady, George Alvarez, and Aude Oliva
Massachusetts Institute of Technology

A poster illustrating the 4672 natural scenes used in the massive memory experiment of the talk "Remembering thousands of natural images with high fidelity".

Calculating scene context: What 47,928 objects in 3500 scenes can tell us about natural scene categories
Michelle R. Greene, Aude Oliva and Antonio Torralba
Massachusetts Institute of Technology

Context is critical to our ability to recognize environments and to search for objects within them. Indeed, contextual associations have been shown to modulate reaction time and object recognition accuracy (Biederman, 1981), as well as influencing the distribution of eye movements (De Graaf et al, 1990; Friedman, 1979) and patterns of brain activations (Bar & Aminoff, 2003). In spite of all of these results, there has yet to be a systematic quantification of context given object-object and object-scene relationships in the world. Here we seek to fill this gap by providing both an operational definition of types of context, as well as data for the object regularities existing in natural scene categories. A large scene database of 3500 real-world scenes from 16 basic-level categories was hand segmented and annotated using the LabelMe online tool. A total of 47,928 objects were annotated. From these data, we computed a variety of descriptive statistics including frequency and entropy; relations between scene categories and objects including object consistency (p(object | scene category)), diagnosticity (p(scene category | object)), and the mutual information between scene categories and objects; ensemble statistics including mean and variance of object sizes across categories, and variance of spatial distances between objects and the density of objects in scenes; and dependencies between pairs of objects. These data provide valuable information on natural scene redundancy that can be exploited for machine vision, and will help researchers in visual cognition design experiments based on contextual data rather than intuition.