Scene Understanding Symposium
SUnS 2011

Friday, January 28, 2011

MIT Campus, Building 46-3002, 43 Vassar Street, Cambridge, MA 02139 http://suns.mit.edu

8:30 BREAKFAST

BUILDING SCENE REPRESENTATIONS: COMPUTER VISION

9:00 Out of Context Objects
Antonio Torralba (Massachusetts Institute of Technology)

9:20 SUN Database: Large-scale Scene Categorization and Detection
James Hays (Brown University)

9:40 Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification
Fei-Fei Li (Stanford University)

10:00 Finding and Describing Objects within Broad Domains
Derek Hoiem (University of Illinois at Urbana-Champaign)

10:20 COFFEE BREAK - POSTERS

ATTENTION AND OBJECT RECOGNITION: MODELS AND NEUROPHYSIOLOGY

10:45 Untangling Object Identity Manifolds along the Ventral Visual Stream: A Direct Comparison of V4 and IT Neuronal Population Representations
James DiCarlo (Massachusetts Institute of Technology)

11:05 Object Decoding with Attention in Inferior Temporal Cortex
Robert Desimone (Massachusetts Institute of Technology)

11:25 What and Where: A Bayesian Inference Theory of Attention
Thomas Serre (Brown University)

11:45 Effects of Exploratory Saccades on Brain Activity and Visual Perception
Michael Paradiso (Brown University)

12:05 POSTER SPOTLIGHTS

12:30 LUNCH - POSTERS

BUILDING SCENE REPRESENTATIONS: COGNITIVE NEUROSCIENCE

1:45 Neural Construction of Scenes from Objects in Human Occipitotemporal Cortex
Russell Epstein (University of Pennsylvania)

2:05 Disentangling Neural Representation of Scene Content from Spatial Boundary
Soojin Park (Massachusetts Institute of Technology & Johns Hopkins University)

2:25 Property-Based Neural Representation of Space
Aude Oliva (Massachusetts Institute of Technology)

2:45 Objects, Places and Spaces: Pathways to Scene Representations
Chris I. Baker (National Institutes of Health)

3:05 COFFEE BREAK - POSTERS

3:30 Invariance to Mirror Image Reversals in Object- and Scene-Selective Cortical Regions
Daniel D. Dilks (Massachusetts Institute of Technology)

3:50 Object Ensemble Representation in the Human Brain
Jonathan Cant (Harvard University)

4:10 The Contextual Associations Network: Spatial and Temporal Characterization
Moshe Bar (Massachusetts General Hospital, Harvard Medical School)

SPECIAL TOPIC IN VISUAL COGNITION

4:30 Guided Search in Scenes
Jeremy Wolfe (Brigham and Women's Hospital, Harvard Medical School)

5:00 RECEPTION

Organized by: Aude Oliva, Thomas Serre, Antonio Torralba

The Scene Understanding Symposium series is an educational initiative from the Department of Brain and Cognitive Sciences, the McGovern Institute for Brain Research and the Computer Science and Artificial Intelligence Laboratory at MIT. We wish to thank for their sponsoring: the McGovern Institute for Brain Research, the National Science Foundation (IIS-CAREER Award to Aude Oliva, No. 0546262, and IIS-CAREER Award to Antonio Torralba, No. 0747120), the National Institutes of Health Predoctorol Training Grants (T32 EY013935 and T32 GM007484), the Department of Brain and Cognitive Sciences and the Computer Science and Artificial Intelligence Laboratory at MIT.

TALKS

9.00
Out of Context Objects
Antonio Torralba
Massachusetts Institute of Technology

Despite that objects tend to appear in typical configurations and scenes, sometimes, objects are out of place or appear in unexpected contexts. Detecting "out-of-context" objects is challenging because context violations can be detected only if the relationships between objects are carefully and precisely modeled. In this talk I will describe the failures and successes of several models of context in detecting which objects produce contextual violations in a database of real scenes. (Work in collaboration with Myung Jin Choi, and Alan S. Willsky).

9:20
SUN Database: Large-scale Scene Categorization and Detection.
James Hays
Brown University

Scene recognition is a fundamental topic in computer vision and human vision. However, scene understanding research has been constrained by the limited scope of currently-used databases. Whereas databases for object categorization contain hundreds of classes of objects, the largest available dataset of scene categories contains only 15 classes. We introduce the extensive Scene UNderstanding (SUN) database that contains 899 categories and 130,519 images and use it to evaluate numerous state-of-the-art algorithms for scene recognition. We measure human scene classification performance on the SUN database and compare this with computational methods. In addition, we introduce the concept of scene detection — detecting scenes embedded within larger scenes — in order to evaluate computational performance under a finer-grained local scene representation. Finding new global scene representations that significantly improve performance is important as it validates the usefulness of a parallel and complementary path for scene understanding that can be used to provide context for object recognition. (This is work with Jianxiong Xiao, Krista Ehinger, Aude Oliva, and Antonio Torralba. Published at CVPR 2010).

9:40
Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification
Fei-Fei Li
Stanford University

Robust low-level image features have been proven to be effective representations for a variety of visual recognition tasks such as object recognition and scene classification; but pixels, or even local image patches, carry little semantic meanings. For high level visual tasks, such low-level image representations are potentially not enough. In this work, we propose a high-level image representation, called the Object Bank, where an image is represented as a scale-invariant response map of a large number of pre-trained generic object detectors, blind to the testing dataset or visual task. Leveraging on the Object Bank representation, superior performances on high level visual recognition tasks can be achieved with simple off-the-shelf classifiers such as logistic regression and linear SVM. Sparsity algorithms make our representation more efficient and scalable for large scene datasets, and reveal semantically meaningful feature patterns. (This is joint work with Li-Jia Li, Hao Su and Eric Xing, published at ECCV'2010 (workshop) and NIPS'2010.)

10:00
Finding and Describing Objects within Broad Domains
Derek Hoiem
University of Illinois at Urbana-Champaign

In computer vision, we tend to organize objects into a taxonomy or set of dichotomies, where recognition is a process of separating the wheat from the chaff. But what do we do about objects that don't fit neatly into our pre-learned categories? What if we see something that cannot be named? We propose a domain-based approach to recognition, in which parts, pose, function, and other attributes are shared across basic categories within broad domains, such as animals or vehicles. This is a step towards creating a visual system that can at least partially recognize any object that it encounters, where precision increases with familiarity. (This work is with Ali Farhadi and Ian Endres).

10:45
Untangling Object Identity Manifolds along the Ventral Visual Stream: A Direct Comparison of V4 and IT Neuronal Population Representations
James DiCarlo
Massachusetts Institute of Technology

11:05
Object decoding with attention in inferior temporal cortex
Robert Desimone
Massachusetts Institute of Technology

Work in collaboration with Ying Zhang, Ethan Meyers, Narcisse P. Bichot, Thomas Serre and Tomaso Poggio

11:25
What and Where: A Bayesian Inference Theory of Attention
Thomas Serre
Brown University

We describe a theoretical framework, which assumes that attention is part of the inference process that solves the visual recognition problem of 'what is where' (Marr, 1982) and leads to a computational model that predicts some of the main properties of attention at both the level of psychophysics and physiology. Within this framework, spatial attention emerges as a strategy to reduce the uncertainty in shape information while feature-based attention reduces the uncertainty in spatial information. (Work in collaboration with Sharat Chikkerur, Cheston Tan and Tomaso Poggio).

11:45
Effects of Exploratory Saccades on Brain Activity and Visual Perception
Michael Paradiso
Brown University

Saccadic eye movements are fundamental to human exploration of visual scenes but they are not incorporated into most physiology and psychophysics experiments. Recording in area V1 we find that when stimuli enter receptive fields via saccades (rather than being flashed), the magnitude, timing, and selectivity of the response are altered. In corresponding psychophysics experiments we have examined the perceptual consequences of exploratory saccades. We find that even a low level measure of visual acuity – contrast sensitivity – is changed by saccades. Further, saccades reduce the influence of a stimulus on one fixation on perception during a

subsequent fixation. This suggests that the saccades may be resetting visual analysis, effectively parsing visual information on distinct fixations.

1:45
Neural Construction of Scenes from Objects in Human Occipitotemporal Cortex
Russell Epstein
University of Pennsylvania

2:05
Disentangling Neural Representation of Scene Content from Spatial Boundary
Soojin Park
Massachusetts Institute of Technology & Johns Hopkins University

Behavioral and computational studies suggest that visual scene analysis rapidly produces a rich description of both the objects and the spatial layout of surfaces in a scene. However, there is still a large gap in our understanding of how the human brain accomplishes these diverse functions of scene understanding. Here we probe the nature of real-world scene representations using multi-voxel fMRI pattern analysis. We show that natural scenes are analyzed in a distributed and complementary manner by the parahippocampal place area (PPA) and the lateral occipital complex (LOC) in particular, as well as other regions in the ventral stream. Specifically, we study the classification performance of different scene-selective regions using images that vary in spatial boundary and naturalness content. We discover that whereas both the PPA and LOC can accurately classify scenes, they make different errors: the PPA more often confuses scenes that have the same spatial boundaries, whereas the LOC more often confuses scenes that have the same content. By demonstrating that visual scene analysis recruits distinct and complementary high-level representations, our results testify to distinct neural pathways for representing the spatial boundaries and content of a visual scene. (This is work with Timothy F. Brady, Michelle R. Greene & Aude Oliva. Published in Journal of Neuroscience (2011)

2:25
Property-Based Representation of Space
Aude Oliva
Massachusetts Institute of Technology

Estimating the level of clutter and the size of the space in a scene is critical to our interaction with the world. In neuroimaging experiments, we find evidence for a distributed property-based representation of the size of space and level of clutter. Scene-selective regions of interest (e.g. retrosplenial complex, parahippocampal cortex, lateral occipital complex) show different patterns of response to pictures of indoor scenes that vary parametrically in size and level of clutter. Importantly, the differential responses of these regions are independent of the semantic category of the scene, consistent with previous results showing complementary but distinct neural representations of spatial boundary and scene content information. (Work in collaboration with Soojin Park and Talia Konkle).

2:45
Objects, Places and Spaces: Pathways to Scene Representations
Chris I Baker
National Institutes of Health

Visual scenes contain information at a variety of different levels including both semantic and spatial. In a data driven approach, we investigated the relative contribution of these factors to scene representations in the

Parahippocampal Place Area (PPA). Representations in PPA were dominated by spatial factors (expanse and distance) with no evidence for semantic information. Prior fMRI and neuropsychological studies have also implicated PPA in the representation of spatial aspects of scenes and navigation. Given the ventral visual pathway is thought to be dominated by representations of non-spatial information (object identity or stimulus quality), the source of these spatial representations is unclear. We present detailed anatomical and functional evidence from both humans and monkeys for a major pathway between the parietal and medial temporal lobes. This pathway originates in the inferior parietal lobule (7a in the monkey) passing through posterior cingulate and retrosplenial cortices into the posterior parahippocampal cortex and hippocampus. This pathway is critical for scene representations and navigation, and is one of at least three major projections from the dorsal stream conveying spatial information to the frontal, temporal and limbic lobes (Work in collaboration with Dwight Kravitz, Assaf Harel, Saleem Khadarbatcha, and Mort Mishkin).

3:30
Invariance to Mirror-Image Reversals in Object- and Scene-Selective Cortical Regions
Daniel D. Dilks
Massachusetts Institute of Technology

Electrophysiological and behavioral studies in many species have demonstrated mirror-image confusion for objects, perhaps because left/right information is rarely important in object recognition (e.g., a cup is the same cup when seen in left or right profile). However, unlike object recognition, scene recognition crucially requires left/right information; the identity and navigability of a scene are completely different when it is mirror reversed. Thus, we predicted that object representations in object-selective cortex would be invariant to left-right reversals, but scene representations in the scene-selective cortex would not be. To test for such left/right information encoding, we ran an event-related fMRI adaptation experiment. In each trial, we successively presented images of either two objects or two scenes; each pair of images was: 1) the same image (presented twice); 2) two completely different images; or 3) a scene or an object, followed by the mirror-reversed version of the same stimulus. Consistent with our prediction, we found invariance to left-right reversals in one object-selective region (the posterior fusiform sulcus – pFs), but not in the other (the lateral occipital sulcus – LO), suggesting that some left/right information is represented in at least some parts of object-selective cortex in the ventral stream. However, contrary to our predictions, we found that one scene-selective region (the parahippocampal place area – PPA) is invariant to left-right reversals; insofar as scene chirality is crucial for navigation, such information is apparently not encoded in the PPA. Such function could be fulfilled by another one of the scene-selective regions (the transverse occipital sulcus – TOS, or the retrosplenial complex – RSC), which both showed significant sensitivity to left-right reversals. These findings pose a challenge to hypotheses of the PPA's role in scene recognition, navigation, and reorientation.
(Collaborators: Joshua B. Julian, Jonas Kubilius, Elizabeth S. Spelke, & Nancy Kanwisher).

3:50
Object ensemble representation in the human brain
Jonathan Cant & Yaoda Xu
Harvard University

Many everyday activities require the encoding of distinctive visual objects. Although such object specific visual processing is important, there are also ample occasions when our visual system quickly extracts summary statistics from a large collection of objects without forming a detailed representation for the individual objects in the ensemble. Object ensemble representation complements object specific visual processing and can guide attention to specific objects for further processing. Our recent fMRI data suggest that object ensemble representation involves the collateral sulcus and the parahippocampal gyrus and is related to both texture and

scene processing. We would like to argue that object specific and object ensemble representations may constitute two independent and complimentary pathways that, together, allow an observer to perceive both the "individual trees" and the "entire forest" from a visual scene.

4:10
The Contextual Associations Network: Spatial and Temporal Characterization
Moshe Bar
Massachusetts General Hospital, Harvard Medical School

4:30
Guided Search in Scenes
Jeremy Wolfe
Brigham and Women's Hospital, Harvard Medical School

POSTERS

Does Repeated Search in Scenes Need Memory? Looking AT versus Looking FOR Objects in Scenes
Melissa L.-H. Võ and Jeremy M. Wolfe
Brigham and Women's hospital, Harvard Medical School

One might assume that familiarity with a scene or previous encounters with objects embedded in a scene would benefit subsequent search for those items. However, in a series of experiments we show that this is not the case: When participants were asked to subsequently search for multiple objects in the same scene, search performance remained essentially unchanged over the course of searches despite increasing scene familiarity. Similarly, looking at target objects during previews, which included letter search, 30 seconds of free viewing, or even 30 seconds of memorizing a scene, also did not benefit search for the same objects later on. However, when the same object was searched for again memory for the previous search was capable of producing very substantial speeding of search despite many different intervening searches. This was especially the case when the previous search engagement had been active rather than supported by a cue. While these search benefits speak to the strength of memory-guided search when the same search target is repeated, the lack of memory guidance during initial object searches – despite previous encounters with the target objects - demonstrates the dominance of guidance by generic scene knowledge in real-world search.

Global Non-selective Processing of Scenes
Karla K. Evans & Jeremy M. Wolfe
Brigham and Women's hospital, Harvard Medical School

Observers can report on non-selective properties (e.g. gist, global structure and statistical regularities) of scenes within a very short time. However we find that, under demanding, brief (20 msec) exposure conditions, categories interfere destructively when two task-relevant properties are present in the same stimulus (Evans, Horowitz & Wolfe, Psychological Science, in press). Observers can detect beaches in masked 20 msec flashes, but not if animals (relevant on other trials) are present. We demonstrate task-dependent, constructive or destructive interactions between accumulating information about each category. In the present research we show this collision of categories in saccades made to stimuli that remain visible until response. On each trial observers saw a pair (Exp. 1) or four (Exp.2 & 3) images and were asked to rapidly saccade to the image that contained a pre-cued target category (e.g. beach). Targets were always present. Critically on two-thirds of the trials an un-cued task-relevant category was also present (e.g. animal). It was in the same trial relevant image

on half of those trials and in the other field otherwise. The un-cued category slowed response and decreased accuracy regardless of its position in the display. However, the negative predictive value of the un-cued task relevant category was less effective when there were 4 images. The collision of categories is not limited to very briefly presented stimuli. Moreover, categories can collide across visual hemifields.

Depth and Size Information Reduce Effective Set Size for Visual Search in Real-World Scenes
Michelle R. Greene, Ashley Sherman, Jeremy M. Wolfe
Brigham and Women's Hospital, Harvard Medical School

Identification of classic attributes guiding search is based on experiments using unstructured displays. These attributes do not explain the efficiency of search in real scenes. We propose that "depth guidance" can massively reduce the effective set size in real scenes. Fast, non-selective processes are known to provide information about spatial layout and about proto-objects. Given approximate distances and object sizes in the image, only a very few proto-objects can possibly be the current target. To test, we had participants draw boxes on 200 real-world images (indoor and outdoor) indicating multiple possible locations and sizes for target objects (cats, cups, and people) that were not in the images. In the main experiment, one box was picked as the target. Other box locations were chosen as distractors and the boxes resized to match the target's image size. Thus, if observers were looking for cats, only one box would be the right size to just hide the cat. On each trial, observers used as few mouseclicks as possible to identify that box. At chance performance, the slope of the function relating guesses to set size is 0.5. On average, the guesses x set size slope was 0.30. If dots replaced boxes, eliminating the possibility of depth guidance, the slope increased to 0.43 (near chance performance). Our depth guidance slope estimate of 0.30 is too steep because some distractors were chosen from boxes placed at the roughly the same distance as the target object, making those boxes of an appropriate size. These are scored as distractors but be targets. An additional experiment estimated that this occurred on 42% of trials. Correcting the slope for this factor yields an estimated slope of 0.19. This shallow slope suggests that depth guidance effectively reduces the number of candidate targets in real scenes by directing attention to size-appropriate objects.

A Hierarchically Organized Cortical Network for Scene Perception
Joshua B. Julian (1), Daniel D. Dilks (1), Jonas Kubilius (1,2), Siyuan Hu (3), Jia Liu (3) & Nancy Kanwisher (1)
1 McGovern Institute for Brain Research, MIT, Cambridge, MA 02139
2 Laboratories of Biological Psychology and Experimental Psychology, Katholieke Universiteit Leuven , 3000 Leuven, Belgium
3 State Key Laboratory of Cognitive Neuroscience and Learning, Beijing Normal University, Beijing, China, 100875

It is widely believed that visual processing is hierarchical, such that the early stages of the hierarchy process simple stimulus features, while the later stages construct more complex representations. While such hierarchical organization is well established for early stages of visual processing, it remains unclear whether similar processing exists at later stages. Here we hypothesize a hierarchically organized cortical network for scene perception. Consistent with this hypothesis, using functional magnetic resonance imaging (fMRI), we found: i) a gradient of selectivity within the scene-selective cortical regions, with the more posterior, transverse occipital sulcus (TOS), exhibiting less selectivity than the more anterior regions, the parahippocampal place area (PPA), and retrosplenial complex (RSC), and ii) a functional dissociation over these regions, with TOS extracting the local components of a scene (e.g., surfaces), and PPA and RSC representing the global spatial arrangement of these components. In Experiment 2, we further tested our hypothesis of a hierarchically organized network for scene perception using resting-state fMRI correlations. Resting-state fMRI provides a unique means of mapping functional networks across cortical regions in the resting state (i.e., without an explicit task). Resting activity in TOS was strongly correlated with PPA, which, in turn was strongly correlated with RSC. Taken together, the fMRI and resting-state fMRI data provide

converging evidence for a cortical network for scene perception, with TOS serving as the initial stage in the hierarchy, followed by PPA, then RSC.

What's in a Scene? Interactions between Objects and Space in Human Visual Cortex
Assaf Harel, Dwight J. Kravitz, Chris I. Baker
Unit on Learning and Plasticity, Laboratory of Brain and Cognition, National Institute of Mental Health

Neuroimaging studies have identified a network of scene-selective cortical regions: Parahippocampal Place Area (PPA), Retrosplenial Complex (RSC) and Transverse Occipital Sulcus (TOS). However, the different contributions of these areas to scene recognition are unclear. Further, while PPA shows a stronger response to scenes than objects it also contains significant object information. To reconcile these findings and elucidate the nature of scene processing in each region, we assessed the relative impact of objects and spatial backgrounds on responses. We manipulated object and spatial information by generating minimal scenes comprising one of seven objects (or no object), presented on one of three different backgrounds differing in the spatial information they contain (room interior, horizon, and luminance gradient). The horizon and room backgrounds conveyed depth information, but only the room defined an enclosed space. Unlike other studies using real-world, but visually uncontrolled scenes, these minimal scenes provide a simple controlled test of the relative contribution of objects and backgrounds. Individual scenes were presented in an ungrouped event-related fMRI experiment and the distributed response patterns to each were analyzed using split-half correlations. Response patterns in scene-selective cortex discriminated both objects and backgrounds. However, different scene-selective regions emphasized different aspects of the scenes. RSC primarily reflected the spatial aspects of the scenes, discriminating backgrounds regardless of objects and showing no modulation of response strength by object presence. In contrast, PPA and TOS responded more strongly in the presence of an object and showed strong discrimination of objects as well as backgrounds. However, object discrimination in PPA, but not TOS, was modulated by the type of background, with the strongest discrimination against room backgrounds. We conclude that while RSC mainly represents spatial information, PPA and TOS represent both object and spatial information to provide a rich representation of the visual environment.

Canonical Views of Scenes
Krista A. Ehinger & Aude Oliva
Dept. of Brain & Cognitive Sciences, MIT

When recognizing, depicting, or imagining objects, people show a preference for particular "canonical" views. Do people have similar preferences for particular views of scenes? We investigated this question using panoramic images, which show a 360-degree view from a particular location. These images were shown in an online task: observers were asked to manipulate the view in an interactive window to produce the best possible "snapshot" of the scene. We found that agreement between observers on the "best" view of each scene was generally high, and seemed to depend on the size and shape of the space. We attempted to predict the selected views using a model based on the shape of the space and its navigational constraints (where an observer could walk in the image). We find that the shape of the space, not navigational constraints, predict the best view of the scene.

Similar Scenes Seen: What are the limits of the visual long-term memory fidelity?
Olivier Joubert & Aude Oliva
Department of Brain and Cognitive Sciences, MIT

The capacity of long-term memory (LTM) for pictures is outstanding: observers distinguish thousands of distinct pictures from foil exemplars after seeing each item only once (Standing, 1973; Brady et al., 2008). In contrast, change blindness shows that, even in short term memory, two versions of the same picture are

difficult to distinguish when they differ by only a few objects. Clearly, there are limits to the resolution of visual LTM. Here, we investigated the fidelity of LTM by using foil images representing similar versions of a scene. During a learning phase, 312 color photographs of different categories were displayed for 2 seconds each. Observers performed an N-back task to encourage sustained attention. Importantly, observers were explicitly informed prior to learning about the testing conditions. At test, they performed a 2-AFC task with one old image and a foil whose resemblance with the target was manipulated: the foil could be a mirror image of the same scene, the same scene zoomed in or out by 25 %, or a nearby scene cropped from a larger panoramic image. The control condition, a foil from a novel category, led to 93% recognition accuracy, as in related previous studies. The fidelity of memory was poorest (54%, chance level) when the foil depicted a “zoom-out” version of the old image. Participants performed well (84%) with foils depicting a translated non-overlapping version, and were moderately accurate (79%) with foils image overlapping by 50%, a zoom-in (69%) or a left-right mirror of the old image (72%). In a broader context, these results contribute to understanding the nature of stored visual representations. LTM representations have been shown to be sensitive to changes in scene viewpoint. Nevertheless, our results suggest that visual long-term memory is "open-minded" about certain kinds of viewpoint transformations: it does not mind a step backward.

Improvements in Scene Text Recognition Using Similarity, Integer Programming and Search Engine Correction

David L. Smith, Jacqueline Feild, and Erik Learned-Miller
Department of Computer Science, University of Massachusetts, Amherst

The recognition of text in everyday scenes is made difficult by viewing conditions, unusual fonts, and lack of linguistic context. Most methods integrate a priori appearance information and some sort of hard or soft constraint on the allowable strings. Weinman and Learned-Miller showed that the similarity among characters, as a supplement to the appearance of the characters with respect to a model, could be used to improve scene text recognition. In this work, we make further improvements to scene text recognition by taking a novel approach to the incorporation of similarity. In particular, we train a "similarity expert" that learns to classify each pair of characters as equivalent or not. After removing logical inconsistencies in an equivalence graph, we formulate the search for the maximum likelihood interpretation of a sign as an integer program. We incorporate the similarity information as constraints in the integer program and building an optimization criterion out of appearance features and character bigrams. Finally, we take the optimal solution from the integer program, and compare all ``nearby'' solutions using a probability model for strings derived from search engine queries. We demonstrate word error reductions of more than 30% relative to previous methods on the same data set.

Adapting to the Scene: Better Face Detection through Shared Context

Vidit Jain and Erik Learned-Miller
Department of Computer Science, University of Massachusetts, Amherst

Many classifiers are trained with massive training sets only to be applied at test time on data from a different distribution. How can we rapidly and simply adapt a classifier to a new test distribution, even when we do not have access to the original training data? We present an on-line approach for rapidly adapting a “black box” classifier to a new test data set without retraining the classifier or examining the original optimization criterion. Assuming the original classifier outputs a continuous number for which a threshold gives the class, we reclassify points near the original boundary using a Gaussian process regression scheme. We show how this general procedure can be used in the context of a classifier cascade, demonstrating performance that far exceeds state-of-the-art results in face detection on a standard data set. We also draw connections to work in semi-supervised learning, domain adaptation, and information regularization.

Distribution Fields: A Flexible Representation for Low-Level Vision Problems
Laura Sevilla Lara and Erik Learned-Miller
Department of Computer Science, University of Massachusetts, Amherst

We present a flexible new representation and set of algorithms for low level vision problems called Distribution Fields. They are quite similar to other features such as shape contexts, HOG and SIFT features, and pixel-wise mixture-of-Gaussian models, but also have important differences. We show that they are powerful representations for images that allow images to be reasonably compared even when they suffer from slight misalignments. We show applications to tracking and discuss ongoing work in background subtraction.

Blocks World Revisited: Qualitative Volumes for Scene Understanding
Abhinav Gupta, Alexei Efros and Martial Hebert
School of Computer Science, Carnegie Mellon University

Since most current scene understanding approaches operate either on the 2D image or using a surface-based representation, they do not allow reasoning about the physical constraints within the 3D scene. Inspired by the "Blocks World" work in the 1960's, we present a qualitative volumetric representation of images. Our representation allows us to apply powerful global geometric constraints between 3D volumes as well as the laws of statics in a qualitative manner. We present a novel iterative "interpretation-by-synthesis" approach where, starting from an empty ground plane, we progressively "build up" a physically-plausible 3D interpretation of the image. We demonstrate the utility of volumetric representation and reasoning for both indoor and outdoor scene understanding.

Category Independent Object Proposals
Ian Endres and Derek Hoiem
Department of Computer Science, University of Illinois at Urbana-Champaign

We propose a category-independent method to produce a bag of regions and rank them, such that top-ranked regions are likely to be good segmentations of different objects. Our key objectives are completeness and diversity: every object should have at least one good proposed region, and a diverse set should be top-ranked. Our approach is to generate a set of segmentations by performing graph cuts based on a seed region and a learned affinity function. Then, the regions are ranked using structured learning based on various cues. Our experiments on BSDS and PASCAL VOC 2008 demonstrate our ability to find most objects within a small bag of proposed regions.

How Does Text in Real-World Scenes Attract Attention?
Hsueh-Cheng Wang & Marc Pomplun
Department of Computer Science, University of Massachusetts at Boston

Intuitively, it seems plausible that in our everyday vision, attention is disproportionately attracted by texts. The present study was aimed at testing this hypothesis under free viewing conditions and examining some of the underlying factors. Texts in real-world scenes were compared with paired control regions of similar size, eccentricity, and low-level visual saliency. The greater fixation probability and shorter minimum fixation distance of texts showed their higher attractiveness. These results might be caused by high-level features such as prominent locations in a scene or special visual features of text that may differ from typical low-level saliency. In another experiment, texts were removed from their expected positions, and the results indicated that the expected locations of texts did draw more attention than controls.

Finally, texts were placed in unexpected positions in front of homogeneous and inhomogeneous backgrounds. These unconstrained texts were found more attractive than controls, with background noise reducing this difference, which indicates that the attraction by specific visual features of text was superior to typical saliency in real-world scenes.

Matching and Predicting Street Level Images

Biliana Kaneva, Josef Sivic, Antonio Torralba, Shai Avidan, and William T. Freeman
CSAIL, MIT

The paradigm of matching images to a very large dataset has been used for numerous vision tasks and is a powerful one. If the image dataset is large enough, one can expect to find good matches of almost any image to the database, allowing label transfer, and image editing or enhancement. Users of this approach will want to know how many images are required, and what features to use for finding semantic relevant matches. Furthermore, for navigation tasks or to exploit context, users will want to know the predictive quality of the dataset: can we predict the image that would be seen under changes in camera position?

We address these questions in detail for one category of images: street level views. We have a dataset of images taken from an enumeration of positions and viewpoints within Pittsburgh. We evaluate how well we can match those images, using images from non-Pittsburgh cities, and how well we can predict the images that would be seen under changes in camera position. We compare performance for these tasks for eight different feature sets, finding a feature set that outperforms the others (HOG). A combination of all the features performs better in the prediction task than any individual feature. We used Amazon Mechanical Turk workers to rank the matches and predictions of different algorithm conditions by comparing each one to the selection of a random image. This approach can evaluate the efficacy of different feature sets and parameter settings for the matching paradigm with other image categories.