Studying Relationships Between Human Gaze, Description, and Computer Vision

Kiwon Yun[1], Yifan Peng[1], Dimitris Samaras[1], Gregory J. Zelinsky[2], Tamara L. Berg[1]
[1]Department of Computer Science, [2]Department of Psychology
Stony Brook University, Stony Brook, NY 11794, USA
{kyun, yipeng, samaras, tlberg}@cs.stonybrook.edu, {gregory.zelinsky}@stonybrook.edu

Abstract

We posit that user behavior during natural viewing of images contains an abundance of information about the content of images as well as information related to user intent and user defined content importance. In this paper, we conduct experiments to better understand the relationship between images, the eye movements people make while viewing images, and how people construct natural language to describe images. We explore these relationships in the context of two commonly used computer vision datasets. We then further relate human cues with outputs of current visual recognition systems and demonstrate prototype applications for gaze-enabled detection and annotation.

1. Introduction

Every day we consume a deluge of visual information by looking at images and video on the web and more generally looking at the visual world around us in our daily lives. In addition, the number of cameras that could conceivably *watch us back* is increasing greatly. Whether it is webcams on laptops, or front-facing cell phone cameras, or Google Glass, the media that we use to access imagery increasingly has the potential to observe our viewing behavior. This creates the unprecedented opportunity to harness these devices and use information about eye, head, and body movements to inform intelligent systems about the content that we find interesting and the tasks that we are trying to perform. This is particularly true in the case of gaze behavior, which provides direct insight into a person's interests and intent.

We envision a day when reliable eye tracking can be performed using standard front facing cameras, making it possible for visual imagery to be tagged with individualized interpretations of content, each a unique "story" simply through the act of a person viewing their favorite images and videos. In this paper we provide a glimpse into this exciting future by analyzing how humans interact with visual imagery in the context of object detection, and how this symbiotic relationship might be exploited to better analyze and index content that people find important. Understanding how humans view and interpret images will lead to new methods to design, train, evaluate, or augment computer vision systems for improved image understanding.

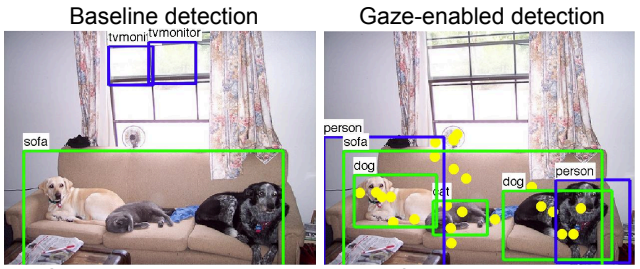


Figure 1: Left: baseline detection results including correct detections (green) and incorrect detections (blue). Right: gaze-enabled detection results with fixations (yellow). Bottom: objects described by people and detected objects from each method (green - correct, blue - incorrect).

1.1. Visual Recognition and Detection

In computer vision, visual recognition algorithms are making significant progress. Recent advances have started to look at problems of recognition at a human scale, classifying or localizing thousands of object categories with reasonable accuracy [19, 24, 5, 6, 18]. However, despite rapid advances in methods for object detection and recognition in images [11, 5], they are still far from perfect. As evidenced in Figure 1, running object detectors (20 deformable part models [12] with default thresholds) on an image, still produces unsatisfactory results. Detectors still produce noisy predictions. In addition, even if the detectors were completely accurate, they would produce an indiscriminate labeling of *all* objects in an image. For some applications, such as image retrieval, a more human-centric annotation of the most important content might be desired.

1.2. Information from Gaze

It has long been known that eye movements are not directly determined by an image, but are also influenced by task [33]. The clearest examples of this come from the extensive literature on eye movements during visual search [8, 21, 34, 35]; specifying different targets yields different patterns of eye movements even for the same image. However, clear relationships also exist between the properties of an image and the eye movements that people make during *free viewing*. For example, when presented with a complex scene, people overwhelmingly choose to direct their initial fixations toward the center of the image [27], probably in an attempt to maximize extraction of information from the scene [27]. Figure/ground relationships play a role as well; people prefer to look at objects even when the background is made more relevant to the task [22]. All things being equal, eye movements also tend to be directed to corners and regions of high feature density [20, 29], sudden onsets [30, 31], object motion [14, 15], and regions of brightness, texture, and color contrast [16, 17, 23]. These latter influences can all be considered saliency factors affecting object importance. The focus of our experiments is on less well explored *semantic factors* – how categories of objects or events might influence gaze [9] and how we can use gaze to predict semantic categories.

Eye movements can inform image understanding in two different but complementary ways. First, they can be used to indicate the relative importance of content in an image by providing a measure of how a person's attention was spatially and temporally distributed. Second, the patterns of saccades and fixations made during image viewing might be used as a direct indication of content information. To the extent that gaze is drawn to oddities and inconsistencies in a scene [28], fixations might also serve to predict unusual events [1].

1.3. Human-Computer Collaboration

In this paper, we explore the potential for combining human and computational inputs into integrated collaborative systems for image understanding. There are many recognition tasks that could benefit from gaze information. The prototype system in [4] looked at methods for human-computer collaborative image classification. In this paper, we focus on object detection and annotation (Figure 1 suggests potential benefits of such a system). Rather than applying object detectors at every location in an image arbitrarily, they could be more intelligently applied only at important locations as indicated by gaze fixations. This would not only minimize the potential for false positives, but also constrain the true positives to only the most user-relevant content. It might also have implications for efficiency in real-time detection scenarios.

Central to making these systems work is our belief that humans and computers provide complimentary sources of information for interpreting the content of images.

Humans can provide:

- Passive indications of content through *gaze* patterns. These cues provide estimates about "where" important things are, but not "what" they are.
- Active indications of content through *descriptions*. These cues can directly inform questions of "what" is in an image as well as indicating which parts of the content are important to the viewer.

Computer vision recognition algorithms can provide:

- Automatic indications of content from recognition algorithms. These algorithms can inform estimates of "what" might be "where" in visual imagery, but will always be noisy predictions and have no knowledge of relative content importance.

It is our position that image understanding is ultimately a human interpretation, making it essential that inputs from humans be integrated with computational recognition methods. Attempts to solve this problem through analysis of pixels alone are unlikely to produce the kind of image understanding that is useful to humans, the ultimate consumers of imagery. In order to build such a human-computational collaborative system we first have to comprehend the relationship between these disparate modalities.

In this paper we describe several combined behavioral-computational experiments aimed at exploring the relationships between the pixels in an image, the eye movements that people make while viewing that image, and the words that they produce when asked to describe it. To the extent that stable relationships can be discovered and quantified, they can be integrated into image interpretation algorithms, used to build better applications, and generally contribute to basic scientific knowledge of how humans view and interpret visual imagery. For these experiments we have collected gaze fixations and some descriptions for images from two commonly used computer vision datasets. Our data, the SBU Gaze-Detection-Description Dataset, is available at http://www.cs.stonybrook.edu/∼ial/gaze.html

2. Dataset & Experimental Settings

We investigate the relationships between eye movements, description, image content, and computational recognition algorithms using images from two standard computer vision datasets, the Pascal VOC dataset [10] and the SUN 2009 dataset [3].

PASCAL VOC: The PASCAL VOC is a visual recognition challenge widely known in the computer vision community for evaluating performance on object category detection (among other tasks). We use 1,000 images from the

2008 dataset [10], selected by Rashtchian et al [26] to contain at least 50 images depicting each of the 20 object categories. For each object category, Felzenszwalb et al. [12] provide a pre-trained deformable part model detector. For each image, we also have 5 natural language descriptions obtained by Rashtchian et al [26] using Amazon's Mechanical Turk (AMT) service. These descriptions generally describe the main image content (objects), relationships, and sometimes the overall scene.

SUN09 Dataset: The second dataset we use is a subset of the SUN09 dataset [3] of scene images with corresponding hand labeled object segmentations. In our experiments we use 104 images of 8 scene categories selected from the SUN09 dataset, each having hand-labeled object segmentations. We train 22 deformable part model object detectors [12] using images with associated bounding boxes from ImageNet [7]. These categories were selected to cover, as much as possible, the main object content of our selected scene images.

Experimental Settings

PASCAL VOC: On this dataset we explore short timeframe viewing behavior. Each of 1,000 images is presented for 3 seconds to 3 human observers. The observers' task is to freely view these images in anticipation of a memory test. Eye movements were recorded during this time using a remote eye tracker (EL1000) sampling at 1000 Hz. Image descriptions were not collected from observers during the experiment, as we wanted to examine the general relationships between gaze and description that hold across different people.

SUN09 Dataset: On this dataset we explore somewhat longer timeframe viewing behavior. Each image is presented to 8 human observers for 5 seconds. Subjects are instructed to freely view these images. After presentation subjects are asked to describe the image they previously saw and are given 20 seconds to provide an oral description. Descriptions are then manually transcribed to text. In addition, we also collect text descriptions via AMT in a similar manner to Rashtchian et al [26]. Figure 2 shows an example gaze pattern and description.

3. Experiments & Analysis

In this section, we address several general questions relating gaze, description, and image content. 1) What do people look at? (Sec 3.1) 2) What do people describe? (Sec 3.2), and 3) What is the relationship between what people look at and what they describe? (Sec 3.3).

3.1. What do people look at?

Gaze vs Selected Objects: To determine whether the objects we have selected for consideration (the 20 Pascal

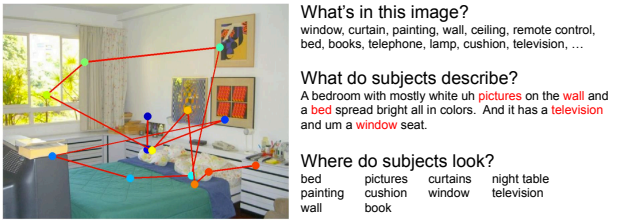


Figure 2: Left: An example of a gaze pattern and description. Each dot indicates a fixation. Colors indicate earlier (blue) to later (red) fixations. Right: A person's description of the image, together with the object ground truth and the objects that were fixated. Red words indicate objects automatically extracted from the sentence.

categories, and 22 classes from SUN09) represent the interesting content of these images, we first need to validate to what degree people actually look at these objects. For example, Pascal was collected to depict certain objects for evaluating detection algorithms, but it also contains other unrelated objects. The SUN09 dataset has labels for almost every object including background elements like floor, or tiny objects like remote control, most of which we have not selected for consideration in our experiments. Hence, we first compute how many fixations fall into the image regions corresponding to selected object categories. We find that 76.33% and 65.57% of fixations fall into selected object category bounding boxes for the PASCAL and Sun09 datasets respectively. Therefore, while these objects do reasonably cover human fixation locations they do not represent all of the fixated image content.

Gaze vs Object Type: Here we explore which objects tend to attract the most human attention by computing the rate of fixation for each category. We first study the per image fixation rate for each category, that is, given an image what is the rate at which each object category will be fixated, $NF(I,b)$:

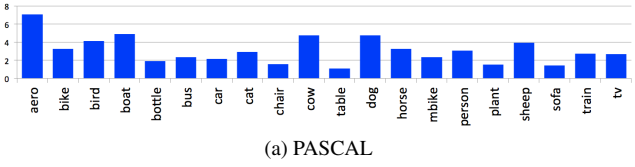
$$F(I,b) = \frac{\text{\# fixations in bounding box } b}{\text{\# fixations in image } I} \quad (1)$$

$$B(I,b) = \frac{\text{size of bounding box } b}{\text{size of image } I} \quad (2)$$

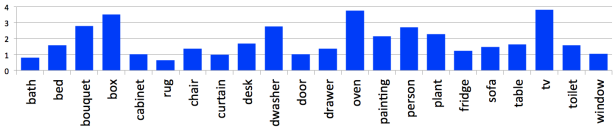
$$NF(I,b) = \frac{F(I,b)}{B(I,b)} \quad (3)$$

where $F(I,b)$ indicates the percentage of fixations falling into bounding box b in image I, and $B(I,b)$ indicates the ratio of the size of bounding box b to the whole image. $NF(I,b)$ denotes the normalized percentage of fixations of bounding box b in image I.

Figure 3 shows the results. In the Pascal dataset people preferentially look at animals like cow or dog, or (relatively) unusual transportation devices like boats or airplanes



(a) PASCAL



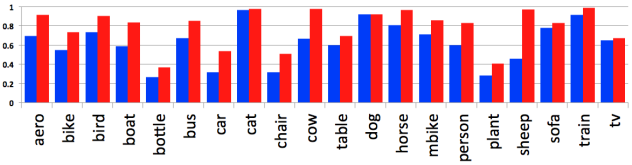
(b) SUN09

Figure 3: (a) In the PASCAL dataset, given an image people preferentially look at some object categories (dog, cat, person) over others (chair, potted plant). (b) Similar patterns can be seen in the SUN09 dataset.

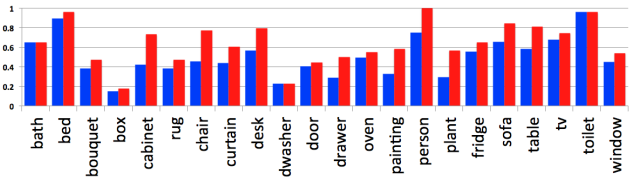
over other common scene elements in an image like dining tables, chairs, or potted plants. In the SUN dataset, people are more likely to look at content elements like televisions (if they are on), people, and ovens than objects like rugs or cabinets.

We also study the overall fixation rate for each category (results are shown in Figure 4). We evaluate this in two ways, 1) by computing the average percentage of fixated instances for each category (blue bars), and 2) by computing the percentage of images where at least one instance of a category was fixated when present (red bars). We calculate the second measure because some images contain many instances of a category, e.g. an image containing a number of sheep. While viewers will probably not take the time to look at every single sheep in the image, if sheep are important then they are likely to look at at least one sheep in the image. We find that while only 45% of all sheep in images are fixated, at least one sheep is fixated in 97% of images containing sheep. We also find that object categories like person, cat, or dog are nearly always fixated on while more common scene elements like curtains or potted plants are fixated on much less frequently.

Gaze vs Location on Objects: Here we explore the gaze patterns people produce for different object categories, examining how the patterns vary across categories, and whether bounding boxes are a reasonable representation for object localization (as indicated by gaze patterns on objects). To analyze location information from fixations, we first transform fixations into a density map. For a given image, a two-dimensional Gaussian distribution that models the human visual system with appropriately chosen parameters is centered at each fixation point. Specifically, sigma was chosen by 7.0% of the image height – to be slightly larger than fovea size. Then, a fixation density map is calculated by summing the Gaussians over the entire image. For each category, we average the fixation density maps

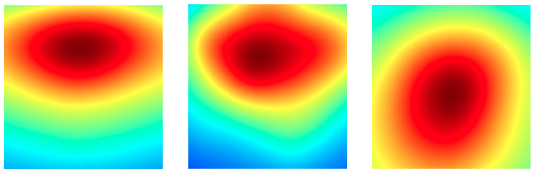


(a) PASCAL

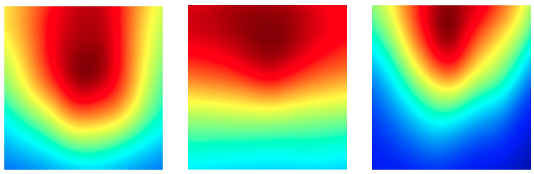


(b) SUN09

Figure 4: Blue bars show the average percentage of fixated instance per category. Red bars show the percentage of images where a category was fixated when present (at least one fixated instance in an image).



(a) person (b) horse (c) tvmonitor



(d) bicycle (e) chair (f) diningtable

Figure 5: Examples of average fixation density maps. Fixation patterns tend to be category dependent.

over the ground truth bounding boxes to create an "average" fixation density map for that category. Figure 5 shows how gaze patterns differ for example object categories. We find that when people look at an animal such as a person or horse (5a, 5b), they tend to look near the animal's head. For some categories such as bicycle or chair (5d, 5e), which tend to have people sitting on them, we find that fixations are pulled toward the top/middle of the bounding box. Similarly, there are often objects resting on top of dining tables (5f). For other categories like tv monitor (5c), people tend to look at the center of the object. This observation suggest that designing or training different gaze models for different categories could potentially be useful for recognizing what someone is looking at.

We also analyze the relationship between gaze, bounding boxes, and object segmentations in the SUN09 dataset which provides segmentations of all labeled objects. We compute the percentage of fixations that fall into the true

	All	Person	Chair	Painting
% of area	68.41%	52.74%	57.51%	91.09%
% of fixations	68.97%	58.84%	59.14%	91.47%

Table 1: Comparison between segmentations and bounding boxes. We measure what percentage of the bounding box is part of the segmented object, and what percentage of the human fixations in that bounding box fall in the segmented object.

object segmentation compared to the entire bounding box (results are shown in Table 1). We find that the percentage of fixations in the object segmentation compared to the bounding is similar to their ratios in area, indicating that while human gaze cues can help provide some rough localization information, they will not necessarily be useful for refining bounding box predictions to object segmentations.

3.2. What do people describe?

In this section we study what people describe in image descriptions. To extract object words from descriptions, we use a Part of Speech tagger [25] to tag nouns. We compare the extracted nouns to our selected object categories using WordNet distance [32] and keep nouns with small WordNet distance. Since WordNet distance is not perfect, we further manually correct the extracted word-object mappings. Experimentally, we find that 85.4% and 58.75% of the ground truth objects are described in the PASCAL and SUN09 datasets respectively. Since the depictions in the SUN09 dataset are somewhat more complex and cluttered, subjects are less likely to describe all selected objects all of the time. Previous work has shown that object categories are described preferentially [2]. For example, animate objects are much more likely to be described than inanimate categories.

3.3. What is the relationship between gaze and description?

We examine the relationship between gaze and description by studying: 1) whether subjects look at the objects they describe, and 2) whether subjects describe the objects they look at. We quantify the relationship between specific gaze patterns and word choices for description by computing the probability that someone will look at the described objects, $P(\textit{fixated} \mid \textit{described})$ and the probability that someone will describe the fixated objects, $P(\textit{described} \mid \textit{fixated})$. Note that we look at these probabilities in two different, but interesting scenarios: when the viewer and describer are the same individual (SUN09) and when they are two different individuals (PASCAL) – to determine whether relationships hold across people. Results are shown in Table 2. We find that there is a strong relationship between gaze and description in both datasets. However, since the Pascal dataset tends to contain cleaner and less cluttered images than those in our SUN images, the correlation in PASCAL is higher than in SUN09.

	P(fixated|described)	P (described|fixated)
PASCAL	86.56%	95.22%
SUN09	73.67%	72.49%

Table 2: The relationship between human description and fixation.

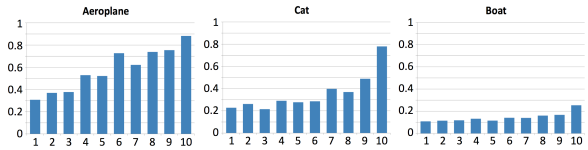


Figure 6: Fixation percentage versus detection score. Scores in the top 10% (bin 10), top 10%-20% (bin 9), etc. In the PASCAL dataset, for categories aeroplane, bus, cat, cow, horse, motorbike, person, sofa, people tends to look much more in the detection boxes with high scores. For other categories, people tend to fixate evenly at detection boxes.

4. Gaze-Enabled Computer Vision

In this section, we discuss the implications of human gaze as a potential signal for two computer vision tasks – object detection and image annotation.

4.1. Analysis of human gaze with object detectors

We first examine correlations between the confidence of visual detection systems and fixation. Positive or negative correlations give us insight into whether fixations have the potential to improve detection performance. In this experiment, we compute detection score versus fixation rate (Equation 3). Results are shown in Fig 6. in general, we find that observers look at bounding boxes with high confidence scores more often, but that detections with lower confidence scores are also sometimes fixated. As indicated by our previous studies, in general some categories are fixated more often than others, suggesting that we might focus on integrating gaze and computer vision predictions in a category specific manner.

Given these observations, we also measure for what percentage of cases fixations could provide useful or detrimental evidence for object detection. In this experiment, we select the bounding boxes output by the detectors at their selected default thresholds. Results are shown in Fig 7 evaluating the following scenarios: 1) There is no predicted detection box overlapping with the ground truth object (blue bars). For these cases, gaze cannot possibly help to improve the result, 2) There are both true positive (TP) and false pos-

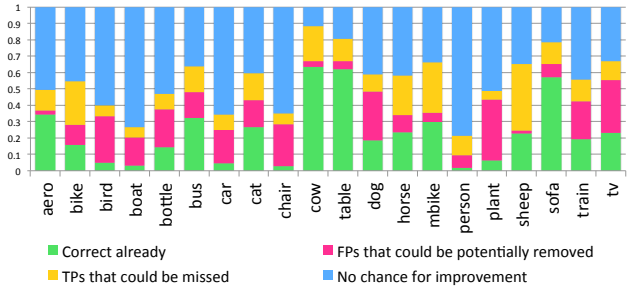


Figure 7: Analysis of where gaze could potentially decrease (yellow), increase (pink), or not affect (green & blue) performance of detection.

itive (FP) boxes overlapping with the ground truth. In some of these cases there will be more fixations falling into a FP box than into a TP. In these cases it is likely that adding gaze information could hurt object detection performance (yellow bars). 3) In other cases, where we have more fixations in a TP box than in any other FP box, gaze has the potential to improve object detection (pink bars). 4) Green bars show detections where the object detector already provides the correct answer and no FP boxes overlap with the ground truth (therefore adding gaze will neither hurt nor help these cases).

4.2. Object Detection

In this section, we employ simple methods for gaze-enabled object detection, using deformable part models [12] with detections predicted at their default thresholds. We first consider the simplest possible algorithm – filter out all detected bounding boxes that do not contain any fixations (or conversely run object detectors only on parts of the image containing fixations). This algorithm filters out many false positive boxes, especially for detectors with lower performance such as bottle, chair, plant, and person. At the same time, it also removes a lot of true positive boxes for objects that are less likely fixated such as bottle and plant, resulting in improvements for some categories, but overall decreased detection performance (Table 3 shows detection performance on the 20 PASCAL categories).

Thus, we also propose a discriminative method where we train classifiers to distinguish between true positive detections and false positive detections output by the baseline detectors. Features for classification include the detection score and features computed from gaze. For gaze features, we first create a fixation density map for each image (as described in Section 3.1). To remove outliers, fixation density maps are weighted by fixation duration [13]. Then, we compute the average fixation density map per image across viewers. To compute gaze features of each detection box, we calculate the average and the maximum of the fixation density map inside of the detection box. Then, the final gaze feature of each box is a three dimensional feature vector (eg. detection score, and the average and maximum of the fixation density map).

For the PASCAL dataset, we split the 1,000 image dataset equally into training and testing sets. Testing evaluation is performed as usual with the standard 0.5 overlap required for true positives. However, for training, we also consider bounding boxes with detection scores somewhat lower than the default threshold for training our gaze classifier and consider a more generous criterion (ie. Pascal overlap > 0.30) for positive samples so that we obtain enough samples to train our classifier. On the other hand, a more strict criterion (ie. Pascal overlap < 0.01) is applied for negative samples. Then, we use hard-negative mining to iteratively add hard negatives (we use 3 iterations of negative mining). Finally, we train 20 classifiers, one per object category, using Support Vector Machines (SVMs) with RBF Kernel, and set parameters with 5-fold cross validation.

Table 3 shows results for baseline detectors, our simple filtering technique, and gaze-enabled classification. Gaze-enabled classifiers outperform the baseline detectors for some animal categories (eg. bird, cat, dog, and horse), train and television, while performance decreases for the plane, boat, car and cow. We generally find gaze helps improve object detection on categories that are usually fixated while it can hurt those that are not fixated (e.g. chair). Additionally, we observe some performance decrease due to detector confusion. For example, the boat detector fires on planes. Since people often look at planes, gaze-enabled classifiers could increase this confusion. Although overall performance (ie. the mean of average precision across categories) is not greatly increased, we believe gaze-enabled algorithms could potentially be useful for many categories.

4.3. Annotation Prediction

We evaluate applicability of gaze to another end-user application, image annotation – outputing a set of object tags for an image. Here, we consider a successful annotation to be one that matches the set of objects a person describes when viewing the image. To transform detection to annotation we output the unique set of categories detected in an image. Using our simple filtering and gaze-enabled classification methods (described in Sec 4.2), we find gaze to be a useful cue for annotation. Overall, both simple filtering and classification improve average annotation performance (Table 4), and are especially helpful for those categories that tend to draw fixations and description, e.g. bird, cat, dog, tv. For inanimate or everyday object categories, e.g. bike, table, sofa we do see some drop in performance, but not a significant amount.

5. Conclusion and Future work

In this paper through a series of behavioral studies and experimental evaluations, we explored the information con-

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	overall (mAP)
baseline detection	63.6	61.7	38.2	44.1	27.9	55.0	50.8	42.9	30.3	66.6	55.2
simple filtering	63.6	62.5	39.7	38.8	15.2	55.3	41.9	44.1	24.6	67.4	52.3
gaze-enabled detection	60.4	61.1	**40.9**	42.2	27.8	**55.5**	49.4	**47.1**	29.6	64.8	**55.3**
	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	
baseline detection	78.7	65.7	65.7	63.3	43.9	32.7	45.3	82.2	72.7	72.5	
simple filtering	79.3	67.5	63.8	60.2	40.6	16.6	38.5	82.6	73.9	70.4	
gaze-enabled detection	78.5	**66.3**	**66.1**	63.1	43.6	**32.9**	45.0	**83.4**	**75.2**	**73.4**	

Table 3: Average precision of detection in the PASCAL dataset

	aero	bike	bird	boat	bottle	bus	car	cat	chair	cow	overall (mAP)
baseline detection	67.6	75.8	42.6	57.1	49.3	74.9	71.4	44.8	49.2	84.9	66.0
simple filtering	67.6	76.8	44.8	51.9	51.8	75.1	76.1	46.1	48.6	85.4	66.8
gaze-enabled detection	66.4	72.9	**47.2**	55.0	**49.5**	**75.2**	**72.7**	**49.1**	**50.3**	**85.2**	**66.9**
	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	
baseline detection	76.3	66.2	85.9	81.9	64.5	39.8	63.3	73.0	82.9	68.7	
simple filtering	76.9	67.9	86.2	82.3	65.1	41.1	63.5	73.3	84.5	71.0	
gaze-enabled detection	76.3	**67.9**	**87.1**	**82.6**	**65.6**	38.6	**63.8**	72.9	**85.1**	**74.1**	

Table 4: Average precision of annotation prediction in the PASCAL dataset

tained in eye movements and description and analyzed their relationship with image content. We also examined the complex relationships between human gaze and outputs of current visual detection methods. In future work, we will study the relationship between temporal order of narrative decryption and the temporal order of fixations. Moreover, we will build on this work in the development of more intelligent human-computer interactive systems for image understanding.

Acknowledgements This work was supported in part by NSF Awards IIS-1161876, IIS-1054133, IIS-1111047, IIS-0959979 and the SUBSAMPLE Project of the DIGITEO Institute, France. We thank J. Maxfield, Hossein Adeli and J. Weiss for data pre-processing and useful discussions.

References

[1] P. F. Baldi and L. Itti. Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666, 2010. 2

[2] A. C. Berg, T. L. Berg, H. Daume, J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, A. Sood, K. Stratos, et al. Understanding and predicting importance in images. In *CVPR*, pages 3562–3569. IEEE, 2012. 5

[3] M. J. Choi, J. J. Lim, A. Torralba, and A. S. Willsky. Exploiting hierarchical context on a large database of object categories. In *CVPR*, 2010. 2, 3

[4] T. De Campos, G. Csurka, and F. Perronnin. Images as sets of locally weighted features. *Computer Vision and Image Understanding*, 116(1):68–85, 2012. 2

[5] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and F.-F. Li. Large scale visual recognition challenge. In *http://www.image-net.org/challenges/LSVRC/2012/index*, 2012. 1

[6] J. Deng, A. C. Berg, K. Li, and F.-F. Li. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010. 1

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR*, 2009. 3

[8] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, 17(6):945–978, 2009. 2

[9] W. Einhäuser, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *Journal of Vision*, 8(14), 2008. 2

[10] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2008 (VOC2008) Results. 2, 3

[11] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. 1

[12] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, pages 1627–1645, 2009. 1, 3, 6

[13] J. M. Henderson. Human gaze control during real-world scene perception. *Trends in cognitive sciences*, 7(11):498–504, 2003. 6

[14] L. Itti. Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. *Vision Cognition*, 12:1093–1123, 2005. 2

[15] L. Itti and P. F. Baldi. Bayesian surprise attracts human attention. *Vision Research*, 49(10):1295–1306, 2009. 2

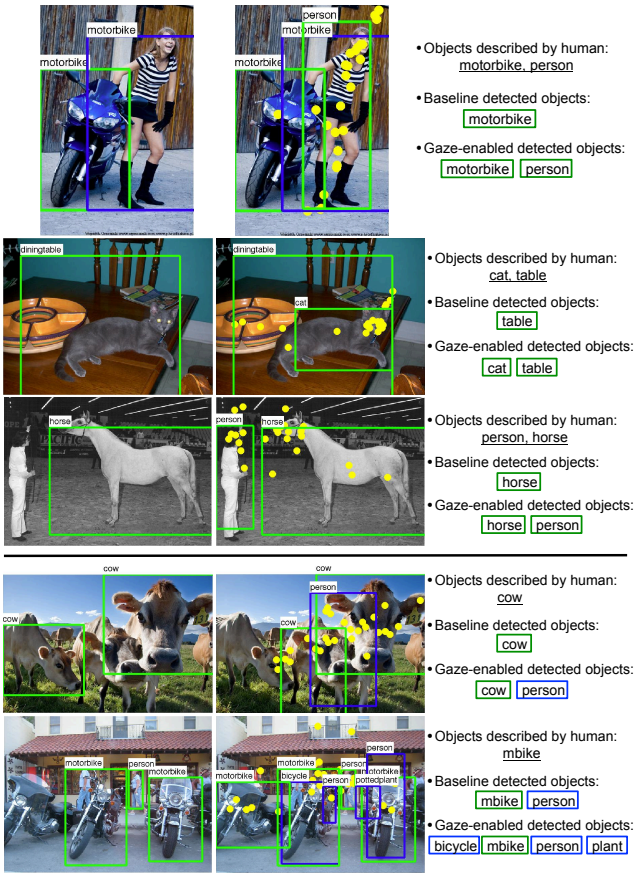


Figure 8: Results of annotation prediction on the PASCAL dataset. Left: baseline detection, Right: Gaze-enabled detection. Gaze-enabled detection improves over the baseline for objects that people often fixate on (*e.g.* cat and person in top three rows). Gaze also sometimes help remove false positives (*e.g.* tv in Figure 1), but sometimes hurts performance by enhancing detector confusion (*e.g.* cow versus person in the 4th row, and bicycle and motorbike in the 5th row). Moreover, sometimes gaze adds additional false positives (*e.g.* plant in the 5th row)

[16] L. Itti and C. Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40:1489–1506, 2000. 2

[17] L. Itti and C. Koch. Computational modeling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001. 2

[18] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 1

[19] Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang. Large-scale image classification: Fast feature extraction and svm training. In *CVPR*, 2011. 1

[20] N. Mackworth and A. Morandi. The gaze selects informative details within pictures. *Perception and Psychophysics*, 2:547–552, 1967. 2

[21] M. B. Neider and G. J. Zelinsky. Scene context guides eye movements during search. *Vision Research*, pages 614–621, 2006. 2

[22] M. B. Neider and G. J. Zelinsky. Searching for camouflaged targets: Effects of target-background similarity on visual search. *Vision Research*, 46:2217–2235, 2006. 2

[23] D. J. Parkhurst, K. Law, and E. Niebur. Modeling the role of salience in the allocation of overt visual selective attention. *Vision Research*, 42:107–123, 2002. 2

[24] F. Perronnin, Z. Akata, Z. Harchaoui, and C. Schmid. Towards good practice in large-scale learning for image classification. In *CVPR*, 2012. 1

[25] X.-H. Phan. CRFTagger: CRF English POS Tagger. http://crftagger.sourceforge.net/, 2006. 5

[26] C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. Collecting image annotations using amazon's mechanical turk. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, 2010. 3

[27] L. W. Renninger, P. Vergheese, and J. Coughlan. Where to look next? eye movements reduce local uncertainty. *Journal of Vision*, 3:1–17, 2007. 2

[28] B. W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):1–17, 2007. 2

[29] B. W. Tatler, R. J. Baddeley, and B. T. Vincent. The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, 46:1857–1862, 2006. 2

[30] J. Theeuwes. Stimulus-driven capture and attentional set: selective search for color and visual abrupt onsets. *Journal of Experimental Psychology: Human Perception and Performance*, 20:799–806, 1994. 2

[31] J. Theeuwes, A. Kramer, S. Hahn, D. Irwin, and G. Zelinsky. Influence of attentional capture on oculomotor control. *Journal of Experimental Psychology: Human Perception and Performance*, 25:1595–1608, 1999. 2

[32] Z. Wu and M. Palmer. Verbs semantics and lexical selection. In *ACL*, pages 133–138, Stroudsburg, PA, USA, 1994. 5

[33] A. Yarbus. *Eye movements and vision*. Plenum Press, 1967. 2

[34] G. J. Zelinsky. A theory of eye movements during target acquisition. *Psychological Review*, 115(4):787–835, 2008. 2

[35] G. J. Zelinsky and J. Schmidt. An effect of referential scene constraint on search implies scene segmentation. *Visual Cognition*, 17(6):1004–1028, 2009. 2