PET: AN EYE-TRACKING DATASET FOR ANIMAL-CENTRIC PASCAL OBJECT CLASSES

Syed Omer Gilani[1], *Ramanathan Subramanian*[2], *Yan Yan*[3], *David Melcher*[4], *Nicu Sebe*[3], *Stefan Winkler*[2]

[1] SMME, National University of Sciences & Technology, Islamabad, Pakistan
[2] Advanced Digital Sciences Center, University of Illinois at Urbana-Champaign, Singapore
[3] Department of Computer Science & Information Engineering, University of Trento, Italy
[4] Department of Cognitive Sciences, University of Trento, Italy
[1]omer@smme.nust.edu.pk, [2]{subramanian.r, stefan.winkler}@adsc.com.sg,
[3]{yan, sebe}@disi.unitn.it, [4]david.melcher@unitn.it

ABSTRACT

We present **PET**– the Pascal animal classes Eye Tracking database. Our database comprises eye movement recordings compiled from forty users for the *bird*, *cat*, *cow*, *dog*, *horse* and *sheep* trainval sets from the VOC 2012 image set. Different from recent eye-tracking databases such as [1, 2], a salient aspect of PET is that it contains eye movements recorded for both the ***free-viewing*** and ***visual search*** task conditions. While some differences in terms of overall gaze behavior and scanning patterns are observed between the two conditions, a very similar number of fixations are observed on target objects for both conditions. As a utility application, we show how feature pooling around fixated locations enables enhanced (animal) object classification accuracy.

Index Terms— PET, Pascal VOC, Animal-centric object classes, Eye movements, Free-viewing vs. Visual search

1. INTRODUCTION

The notion of utilizing implicit user inputs such as eye movements [3, 4] or brain responses [5] for automated scene understanding has gained in popularity recently. With reference to the use of eye movements, a number of algorithms that learn from eye *fixations* and *saccades* to infer salient image regions and objects have been proposed in literature [6, 7, 8]. Fixations denote stationary phases during scene viewing during which humans encode visual information, while saccades denote ballistic eye movements to encode information pertaining to different scene regions.

Of late, some works have specifically looked at utilizing eye fixations for facilitating object detection on the Pascal VOC image set [9]. Two notable works in this respect are [1] and [2]. Both these works are based on the assumption that eye movements are concentrated on the salient object(s), and can therefore enable (i) implicit and fast annotation of objects for model training, and (ii) enhanced object detection performance.

However, there is an important difference between the manner in which [1] and [2] compile user eye movements. In [1], the authors record eye movements under a free-viewing paradigm hypothesizing that natural gaze patterns are capable of directly revealing salient image content. In contrast, [2] employs a visual search paradigm to compile eye movements based on the argument that free-viewing may not provide optimal data for training object detectors. While the impact of the task-on-hand on eye movements has been known and studied for long [10, 11, 12], no empirical study has evaluated the optimality of the visual search or free-viewing paradigms in the context of viewer-based object detection.

To this end, we present **PET** or Pascal animal classes Eye Tracking database[1]. PET comprises eye movement recordings compiled for the *bird*, *cat*, *cow*, *dog*, *horse* and *sheep* training+validation (or trainval) sets from the VOC 2012 image set. These six animal-centric classes were chosen from the 20 object classes in VOC2012 owing to the following reasons: (i) Animal classes such as *cats*, *dogs* and *birds* are particularly difficult to detect using traditional supervised learning methods (*e.g.*, deformable parts model) owing to large intrinsic shape and textural variations [13], and (ii) It would be highly beneficial to incorporate human knowledge to train object detectors for these classes as many psychophysical studies [14] have noted our tendency to instantaneously detect animals (which are both predators and prey).

A salient aspect of the PET dataset is that it contains eye movements recorded under both *free-viewing* (no task) and *visual search* (task-based) paradigms. In all, eye movements were recorded from a total of 40 users who viewed each image for 2 seconds, so that four gaze patterns are available per image and condition. Comparing high-level eye movement statistics as well as temporal scan paths observed for the two conditions, we observe systematic differences in line with Yarbus' observations [10]. However, very similar proportions of eye fixations are observed on the target objects in both conditions, implying that both free-viewing and visual search could be equally suitable when fixated locations are used for subsequent model training. Finally, we show how the spatial pooling of visual features around fixated eye locations enhances object classification performance for animal classes. Overall, we make the following contributions:

1. While the fact that task influences eye movement patterns is widely known, the explicit impact of the visual search task on target detection in images (animal-centric VOC classes in our case) has never been empirically studied and quantified. PET

This study is supported by the research grant for the Human-Centered Cyber-physical Systems Programme at the Advanced Digital Sciences Center from Singapores Agency for Science, Technology and Research (A*STAR).

[1]http://vintage.winklerbros.net/pet.html

represents the first work in this direction, expressly considering approaches that have focused on viewer-centered object detection.

2. We systematically analyze eye movement behavior in the free-viewing and visual search conditions to show that while the visual search paradigm may in general be more beneficial for target detection, there is little difference between the two paradigms if only the fixated scene locations are used for subsequent learning.
3. We show that feature pooling around fixated locations enhances (animal) object classification performance. More compact scan paths are observed for the visual search task due to the propensity of viewers to return to informative scene regions, which contributes to slightly higher classification for fixations recorded during visual search.

The paper is organized as follows. Section 2 presents related work, while Section 3 describes the experimental protocol employed for compiling the PET data. Section 4 compares and contrasts the free-viewing and visual search conditions based on statistical observations, while Section 5 discusses how pooling of features extracted around the fixated locations improves classification accuracy for the animal classes. We conclude the paper in Section 6.

2. RELATED WORK

Given that humans tend to understand complex scenes by focusing their attentional resources on a small number of *salient* objects [15], the practice of employing user inputs to interactively understand visual content has been in vogue for sometime now as exemplified by the *ESP game*. While the impact of high-level factors such as cognitive task on eye movements has been extensively studied since the pioneering work of Yarbus [10], understanding and predicting the salient content in the scene has only been attempted for over a decade now. Even as early works on saliency modeling such as [16] hypothesized that our visual attention is guided by low-level image factors such as brightness, contrast and orientation, more recent works [14, 17] have noted that we are equally likely to focus on semantic entities such as faces, animals and vehicles such as cars.

The above observations have encouraged direct incorporation of eye movement information for model training in object detection. Two such works that have expressly studied the use of eye movements for improving object detection performance on the Pascal Visual Object Classes (VOC) image set are [1] and [2]. In [1], the authors perform several experiments to explore the relationship between image content, eye movements and text descriptions. Eye movements are then used to perform gaze-enabled object detection and produce image annotations. [2] uses eye movements for training a model to draw bounding boxes on target objects, upon learning the spatial extent of these objects from fixation and content-based cues.

However, the point of contention between [1] and [2] is that [1] assumes that natural gaze patterns are already capable of revealing the locations of target objects in the image, while [2] explicitly instructs observers to perform a visual search on the images. In contrast, we record eye movements under both paradigms, and analyze if a visual search task explicitly improves user-centric target detection performance. Table 1 compares various aspects concerning [1, 2] and our work. A detailed explanation of the experimental protocol adopted for PET is as follows.

3. MATERIALS AND METHODS

Stimuli and Participants: 4135 images from the Pascal VOC 2012 dataset [9] were selected for the PET study. These images contained one or more instances of the *bird, cat, cow, dog, horse,* and *sheep* categories, and also humans. 2549 images contained *exactly* a single instance of the above target classes, while 1586 images contained either multiple instances from the animal classes, or a mix of animals and humans. Considering only images that contained multiple animals, the mean number of animals per image was 3.1 ± 2.68, which covered a 0.45 ± 0.28 fraction of the image area based on bounding box annotations available as part of the VOC dataset. A total of 40 university students (18–24 years, 22 males) took part in the experiments.

Experimental protocol: Each participant performed the eye-tracking experiment over two sessions spanning about 40 minutes with a short break in-between. They were required to view about 800 images in two blocks, with each image displayed for a duration of 2 seconds and a blank screen displayed in between each image for 500 milliseconds. All participants were instructed to 'free-view' the first block, and asked to 'find all animals in the scene' (visual search) for the second block. The visual search task was always enforced *after* free-viewing to avoid any viewing biases. Also, to minimize boredom, a few *distractor* images which did not contain a single instance of the animal classes were included in the two blocks. The order of the two blocks of images was counterbalanced across a set of subjects, while the images in each block were shown randomly. All images were displayed at 1280×1024 resolution on a 17" LCD monitor placed about 60 cm away from the subjects. Their eye movements were recorded using a Tobii desktop eye tracker, which has a 120 Hz sampling frequency and is accurate to within 0.4° visual angle upon successful calibration.

Pre-processing: Prior to our analysis, we left out the first fixation on each image, and those fixations with invalid (x, y) coordinates. This resulted in total of 28733 fixations for free viewing, and 29901 fixations for the visual search task. Upon pre-processing, for the free-viewing condition, fixation data was available for 3.8 ± 0.4 users per image, while the number of fixations per image was between 5–45 (mean 24.7 ± 6.1). For the visual search task, 4 ± 0.6 gaze patterns were available per image with the number of fixations ranging between 4–52 (mean 22.6 ± 6.4).

4. FREE-VIEWING VS VISUAL SEARCH

In this section, we systematically compare gaze behavior in the free-viewing and visual search paradigms and examine if either task benefited viewer-based object detection.

4.1. Fixation density maps and overall statistics

Fixation density maps (or heat maps) qualitatively reveal those regions that are most frequently visited in a scene, and also provide a measure of fixation dispersion in the scene. Fig. 1 shows fixation density maps for the six animal categories considered in PET. Considering pairs of rows, the top row shows raw eye fixations made by observers during the free-viewing and visual search tasks, while

Table 1. Overview of the three datasets containing eye movement recordings for the VOC image set.

Attribute	**SBU GDD** [1]	**POET** [2]	**PET** (ours)
Objective	Using eye movements and descriptions to improve object detection performance.	Using eye fixations to implicitly annotate bounding boxes for training object detectors.	Using eye movements to improve detection/classification of animal categories in the VOC image set.
Stimuli	1000 images from 20 object categories (50/class) in VOC2008.	6270 images from *cat*, *dog*, *horse*, *cow*, *bicycle*, *motorbike*, *sofa* and *diningtable* classes.	4135 images from the *cat*, *dog*, *bird*, *cow*, *horse* and *sheep* classes in VOC2012.
Task	Free-viewing	Visual search	*Both* free-viewing and visual search.
Number of gaze patterns/image	3	5	4 each for free-viewing and visual search.
Stimulus protocol	Each image presented for 3 seconds.	Pairs of images presented until user responds to indicate presence of a target object class.	Each image presented for 2 seconds.

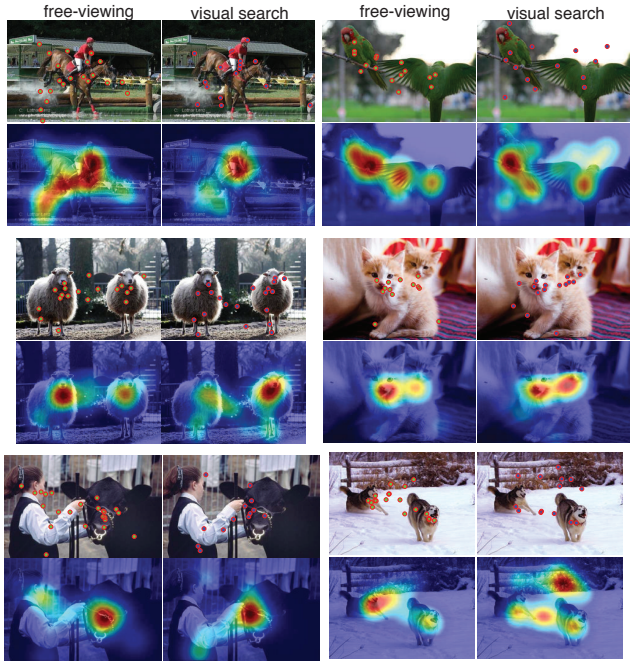


Fig. 1. Recorded eye fixations and fixation density maps for the six animal categories considered in PET.

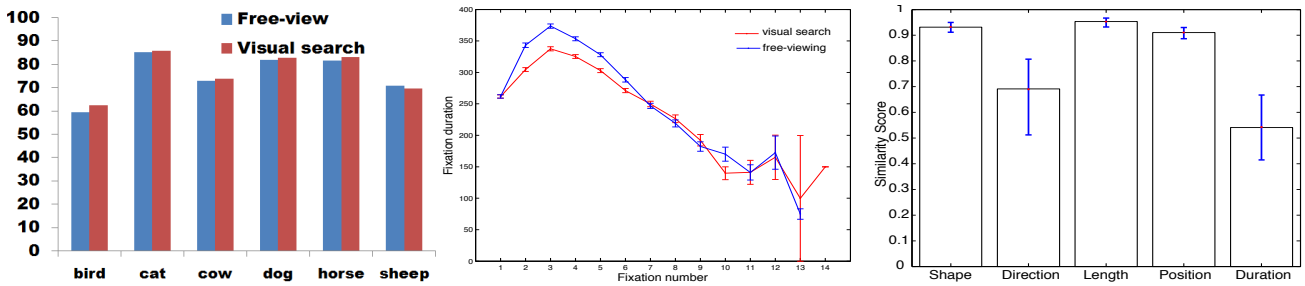
the bottom row presents fixation density maps obtained on convolving fixated scene locations with a Gaussian kernel of 2° visual angle bandwidth. Visual inspection of Fig. 1 reveals that roughly similar density maps are obtained for both conditions, and that fixations mainly lie on animal faces.

Bounding box annotations for objects from the 20 VOC classes are available for the trainval sets that are part of Pascal VOC [9]. Using bounding box coordinates in images containing *multiple instances* of the target object classes and considering only the *first five fixations* made by each user[2] we determined (i) the proportion of fixations that fell within the bounding boxes for the two conditions– the proportion was found to be 0.33 ± 0.26 for both conditions. (ii) the proportion of target objects that had at least one fixation falling on them– this was again found to be 0.73 ± 0.26 for both conditions. (iii) the time taken by each user to fixate on at least half the number of target objects in the image, termed as *saccadic latency*. Saccadic latency for visual search (0.40 ± 0.34) was found to be lower than for free-viewing (0.48 ± 0.35), and a two-sample t-test confirmed that this difference was highly significant ($p < 0.000001$). (iv) the mean fixation durations per target object in the two conditions– mean durations were 0.51 ± 0.26 and 0.47 ± 0.28 for the free-viewing and visual search conditions, and the difference was again highly significant ($p < 0.000001$). We also computed proportion of fixations falling within the target bounding box for the two conditions for each of the animal classes, to examine if any content-related artifacts affected the overall observed statistics. Fig. 2(a) presents the class-wise distribution for both conditions, and shows that the proportions are roughly equal for all of the classes.

4.2. Per-fixation durations

Moving on from overall statistics, we now focus on fine-grained analysis of gaze patterns to examine the free-viewing and visual search paradigms. Firstly, we examined the duration of each fixation made by the population of users for the two conditions. As seen from Fig. 2(b), the first few fixations were longer and are followed by progressively shorter fixations. Also, consistent with the overall numbers, per-fixation durations for free-viewing were consistently higher as compared to visual search up to the sixth fixation, while subsequent fixations were similar in duration.

[2]We considered the first five fixations since they are most likely to convey the intent of the viewer.



(a) (b) (c)

Fig. 2. **Free-view vs visual search:** Comparing (a) Class-wise distribution of fixation proportions (in %), (b) Per-fixation durations and (c) Multimatch similarity scores between scan-path attributes.

4.3. MultiMatch and Recurrence Quantification analysis

In multimatch analysis [18], the series of fixations made by a viewer is treated as a vector (denoting the scan path). Then, the set of scan paths obtained for two conditions are processed to quantify the inter-conditional differences in terms of saccade shape, length and direction, aligned fixation locations and durations. The algorithm returns a similarity score in the range [0-1]. Figure 2(c) presents the similarity between the above measures characterizing viewing behavior during the free-viewing and visual search tasks (considering all subjects and images). In general, these results show that the gaze behaviors in the free-viewing and visual search conditions are similar in a number of respects. Considerable differences are observed only with respect to saccade direction and fixation duration, for which low similarity scores are obtained.

The recurrence quantification analysis (RQA) technique examines the dynamics of a single scan path [19], and provides a measure of the *compactness* in viewing behavior. The compactness is quantified using measures such as *Recurrence*– the proportion of fixations that are repeated on previously fixated locations; *Determinism*– the proportion of recurrent fixations representing repeated gaze trajectories; *Laminarity* denoting the proportion of fixations in a region being repeatedly fixated, and *center of recurrent mass* (CROM), which measures the temporal proximity of recurrent fixations (*i.e.*, time-interval between recurrent fixations). Figure 3 presents the RQA results comparing the visual search and free-viewing conditions. Interestingly, viewing behavior in the visual search scenario was found to be significantly more compact than for free-viewing, in terms of all four measures ($p < 0.01$ with two-tailed t-tests).

4.4. Discussion

Analysis of the viewing behavior in the free-viewing and visual search tasks reveals that viewers in general, tended to fixate on the target objects quicker (lower saccadic latency) and showed a greater urgency in moving around the scene (lower overall fixation and per-fixation duration) during visual search. The multimatch and recurrence quantification analyses show up differences in terms of saccade direction, and the compactness of fixated locations for the two tasks. The fact that scan paths were found to be more compact during visual search suggests that viewers tended to recurrently traverse (what they perceived as) the informative parts of the scene instead of focusing on peripheral scene details. Based on these observations, we make the following comments

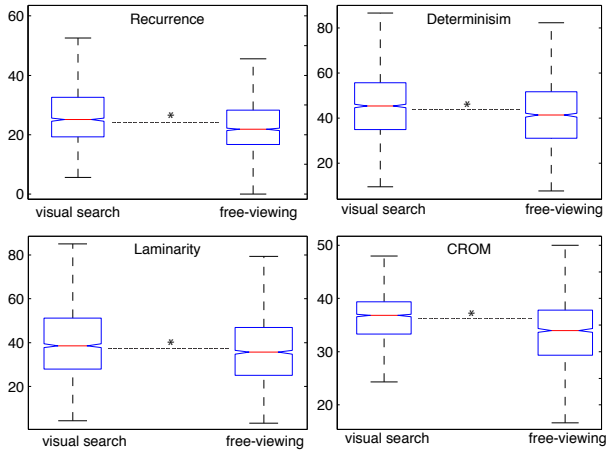


Fig. 3. RQA results for the free-viewing and visual search conditions. Error bars denote unit standard deviation.

regarding the suitability of the free-viewing and visual search tasks for viewer-based target detection.

1. The visual search task appears to motivate viewers to preferentially fixate on designated targets (animals), and traverse the scene with more urgency as compared to free-viewing.

2. Nevertheless, there is not much to choose between the two paradigms if only the fixated locations are to be considered for model training. The proportion of fixations observed on target objects as well as the proportion of target objects fixated are very similar for both conditions.

5. GAZE-ASSISTED OBJECT CLASSIFICATION

In this section, we show how the eye fixations made by viewers in the target detection task can be useful for enhancing object classification accuracy. The Pascal visual object classes (VOC) challenge consists of two main tasks, namely *object classification*, where the objective is to predict the presence or absence of an instance from the target class(es) in an image, and *object detection*, where the task is to spatially localize the target instances via bounding boxes. The

bag-of-words model is extremely popular for object classification. However, it does not encode any spatial information. Spatial pyramid histogram representation is a more sophisticated approach in this respect, as it includes spatial information for object classification and consists of two steps– *coding* and *pooling*.

The coding step involves point-wise transformation of descriptors to a representation better adapted to the task. The pooling step combines outputs of several nearby feature detectors to synthesize a local or global bag of features. Pooling is employed to (i) achieve invariance to image transformations, (ii) arrive at more compact representations, and (iii) achieve better robustness to noise and clutter. However, the spatially pooled regions are usually naively defined in literature. Spatial pyramid match [20] works by partitioning the image into increasingly finer sub-regions, and computing histograms of local features inside each sub-region. These regions are usually *squares* which are *sub-optimal* due to the inclusion of unnecessary background information. Given that viewers tend to fixate on meaningful scene regions, the fixated locations can provide a valuable cue regarding the image features to be used for learning. Therefore, in this work, we pool features from regions around the fixated locations instead of sampling from all over the image. Fig. 4 illustrates the architecture of our proposed fixation-based feature pooling approach.

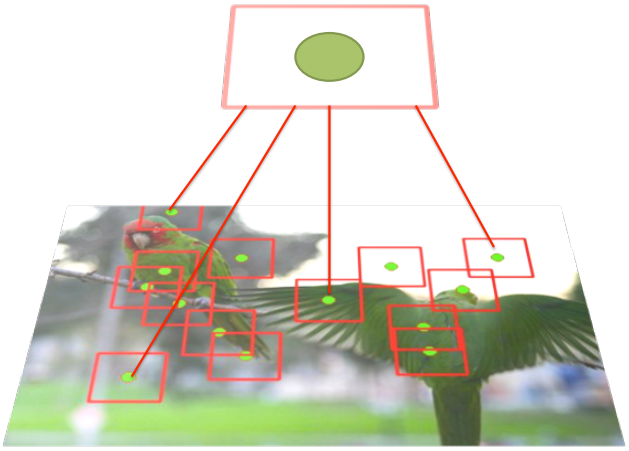


Fig. 4. Feature pooling based on fixated locations: Green dots denote eye fixations. SIFT features are pooled within a window of size 30 $\times$ 30 pixels around the fixated location.

Linear spatial pyramid matching with sparse coding [20] has been successfully used for object classification. Sparse coding has been shown to find succinct representations of stimuli, and model data vectors as a linear combination of a few dictionary codewords. In this paper, sparse coding is adopted for the coding step. We then evaluate the effect of different pooling strategies. Sparse coding is defined as follows:

$$\min_{\mathbf{D},\mathbf{C}} \|\mathbf{X} - \mathbf{CD}\|_F^2 + \lambda_1 \|\mathbf{C}\|_1$$

$$s.t. \quad \mathbf{D_j}.\mathbf{D_j^T} \leq 1, \quad \forall j = 1, ..., l$$

where $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_n}]^T \in R^{d \times n}$, $\mathbf{x_i}$ is the d-dimensional feature vector and n is the number of training samples. $\mathbf{D} \in R^{l \times d}$ is an overcomplete dictionary ($l > d$) with l prototypes. $\mathbf{C} \in R^{n \times l}$ corresponds to the sparse representation of $\mathbf{X}$. λ_1 is the regularization parameter, while $\mathbf{D_j}$ denotes the j-th row of $\mathbf{D}$. The sparsity constraint prevents the learned dictionary from being arbitrarily large.

Popular pooling strategies are average and max pooling. Average pooling is defined as $\mathbf{p} = \frac{1}{m} \sum_{i=1}^{m} \mathbf{c_i}$, while max pooling is defined as $p_k = \max\{|c_{1k}|, |c_{2k}|, ..., |c_{mk}|\}$, where p_k is the k-th element of $\mathbf{p}$, c_{ij} is the element at position (i, j) in $\mathbf{C}$. m is the local number of local descriptors in the considered image region.

5.1. Experimental Results

Instead of pooling features from a regular spatial pyramid, we pool features around viewer-fixated locations. Sparse representations of SIFT features are pooled within a window of size 30 $\times$ 30 pixels around each fixation. Table 2 compares the impact of different pooling strategies on animal classification for the PET images. All experiments were repeated five times and average accuracies and standard deviations are reported. A linear SVM classifier is used as in [20]. As in [20], we observe that max pooling based on sparse codes generally outperforms average pooling on the PET image set. Moreover, eye fixation-based max pooling achieves the best results on four of the six considered animal classes. More than 3% improvement in average classification accuracy is generally achieved with respect to max pooling using a regular spatial pyramid. Finally, given the compactness of gaze patterns in the visual search task, we achieve slightly better classification accuracy using eye fixations recorded from visual search as compared to free-viewing.

Overall, the observed classification results confirm that the use of eye fixations as implicit annotations allows for better characterization of the salient visual image content, namely the animal object instances in this work. Max pooling works better in practice as compared to average pooling of the visual features around fixated locations (or from a grid partition). Also, even though the proportion of fixations on the target objects is roughly similar in the visual search and free-viewing conditions, the propensity of viewers to return back to informative scene regions indirectly contributes to a marginally better classification performance with eye fixations recorded during visual search.

6. CONCLUSIONS

The presented PET database contains eye movement recordings compiled exclusively for trainval images from the six animal categories in the Pascal VOC 2012 dataset. A salient aspect of PET is that it contains eye movements recorded under both *free-viewing* and *visual search* conditions. Systematic comparison of gaze patterns for the two conditions suggests that while visual search appears to motivate the viewer better to perform target detection, target objects are fixated in equal measure under both conditions. Object classification accuracy is found to improve by pooling SIFT features around fixated locations, and pooling features around fixations acquired during visual search is more beneficial given the compactness of gazed locations observed for this condition as compared to free-viewing.

7. REFERENCES

[1] Kiwon Yun, Yifan Peng, Dimitris Samaras, Gregory J. Zelinsky, and Tamara L. Berg, "Studying relationships between human gaze, description, and computer vision," in *CVPR*, 2013.

Table 2. Object classification accuracy with different pooling strategies on PET images.

	bird	cat	cow	dog	horse	sheep	avg+std
max pooling [20]	0.333 ± 0.018	0.263 ± 0.058	0.220 ± 0.032	$\mathbf{0.522 \pm 0.024}$	0.589 ± 0.036	0.437 ± 0.045	0.394 ± 0.007
avg pooling	0.323 ± 0.035	0.283 ± 0.026	0.222 ± 0.041	0.463 ± 0.018	0.517 ± 0.044	0.402 ± 0.051	0.368 ± 0.012
max pooling @ eye fixation (visual search)	$\mathbf{0.357 \pm 0.022}$	0.278 ± 0.034	$\mathbf{0.253 \pm 0.023}$	0.517 ± 0.016	$\mathbf{0.659 \pm 0.021}$	$\mathbf{0.472 \pm 0.031}$	$\mathbf{0.423 \pm 0.011}$
avg pooling @ eye fixation (visual search)	0.346 ± 0.021	$\mathbf{0.291 \pm 0.014}$	0.247 ± 0.036	0.508 ± 0.009	0.547 ± 0.022	0.441 ± 0.038	0.396 ± 0.021
max pooling @ eye fixation (free-viewing)	0.348 ± 0.019	0.264 ± 0.023	0.242 ± 0.019	0.499 ± 0.025	0.635 ± 0.018	0.457 ± 0.028	0.408 ± 0.009
avg pooling @ eye fixation (free-viewing)	0.341 ± 0.017	0.251 ± 0.018	0.224 ± 0.031	0.487 ± 0.012	0.526 ± 0.015	0.428 ± 0.026	0.376 ± 0.014

[2] Dim P. Papadopoulos, Alasdair D. F. Clarke, Frank Keller, and Vittorio Ferrari, "Training object class detectors from eye tracking data," in *ECCV*, 2014, pp. 361–376.

[3] Ramanathan Subramanian, Divya Shankar, Nicu Sebe, and David Melcher, "Emotion modulates eye movement patterns and subsequent memory for the gist and details of movie scenes," *Journal of Vision*, vol. 14, no. 3, 2014.

[4] Syed Omer Gilani, Ramanathan Subramanian, Huang Hua, Stefan Winkler, and Shih-Cheng Yen, "Impact of image appeal on visual attention during photo triaging," in *ICIP*, 2013, pp. 231–235.

[5] Mojtaba K. Abadi, Ramanathan Subramanian, Syed M. Kia, Paolo Avesani, Ioannis Patras, and Nicu Sebe, "DECAF: MEG-based multimodal database for decoding affective physiological responses," *IEEE Transactions on Affective Computing*, 2015.

[6] Laurent Itti and Christoph Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, Mar 2001.

[7] Ramanathan Subramanian, Victoria Yanulevskaya, and Nicu Sebe, "Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements," in *ACM MM*, 2011, pp. 33–42.

[8] Ali Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *CVPR*, 2012.

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results," http://www.pascalnetwork.org/challenges/VOC/voc2012/workshop/index.html.

[10] A. L. Yarbus, *Eye Movements and Vision*, Plenum. New York., 1967.

[11] Marianne DeAngelus and Jeff B Pelz, "Top-down control of eye movements: Yarbus revisited," *Visual Cognition*, vol. 17, no. 6-7, pp. 790–811, 2009.

[12] Benjamin W Tatler, Nicholas J Wade, Hoi Kwan, John M Findlay, and Boris M Velichkovsky, "Yarbus, eye movements, and vision," *i-Perception*, vol. 1, no. 1, pp. 7, 2010.

[13] O. M. Parkhi, A. Vedaldi, C. V. Jawahar, and A. Zisserman, "The truth about cats and dogs," in *ICCV*, 2011.

[14] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba, "Learning to predict where humans look," in *ICCV*, 2009.

[15] Merrielle Spain and Pietro Perona, "Some objects are more equal than others: measuring and predicting importance," in *ECCV*, 2008.

[16] Laurent Itti, Christof Koch, and Ernst Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[17] Ramanathan Subramanian, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua, "An eye fixation database for saliency detection in images," in *ECCV*, 2010, pp. 30–43.

[18] Thomas Foulsham, Richard Dewhurst, Marcus Nystrm, Halszka Jarodzka, Roger Johansson, Geoffrey Underwood, and Kenneth Holmqvist, "Comparing scanpaths during scene encoding and recognition: A multi-dimensional approach," *Journal of Eye Movement Research*, vol. 5, no. 4:3, pp. 1–14, 2012.

[19] Nicola C. Anderson, Walter F. Bischof, Kaitlin E.W. Laidlaw, Evan F. Risko, and Alan Kingstone, "Recurrence quantification analysis of eye movements," *Behavior Research Methods*, vol. 45, no. 3, 2013.

[20] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang, "Linear spatial pyramid matching uisng sparse coding for image classification," in *CVPR*, 2009.