Making Computers Look the Way We Look: Exploiting Visual Attention for Image Understanding

Harish Katti[1], Ramanathan Subramanian[2], Mohan Kankanhalli[1], Nicu Sebe[2],
Tat-Seng Chua[1], K. R. Ramakrishnan[3]
[1]School Of Computing, National University of Singapore,
[2]Department of Information Engineering and Computer Science, University of Trento,
[3]Department of Electrical Engineering, Indian Institute of Science
harishk,mohan,chuats@comp.nus.edu.sg, subramanian,sebe@disi.unitn.it,
krr@ee.iisc.ernet.in

ABSTRACT

Human Visual attention (HVA) is an important strategy to focus on specific information while observing and understanding visual stimuli. HVA involves making a series of *fixations* on select locations while performing tasks such as object recognition, scene understanding, *etc.* We present one of the first works that combines fixation information with automated concept detectors to (i) infer *abstract image semantics*, and (ii) enhance performance of object detectors.

We develop visual attention-based models that sample fixation distributions and fixation transition distributions between *regions-of-interest* (ROI) to infer abstract semantics such as *expressive* faces and object-interactions (such as *look*, *read*, *etc.*). We also exploit eye-gaze information to deduce possible locations and scale of *salient* concepts to aid state-of-the-art detectors. We observe a 18% performance increase with over 80% reduction in computational time for the state-of-the-art object detector in [4].

Categories and Subject Descriptors

H.1.2 [**User/Machine Systems**]: Human information processing; I.5.4 [**Pattern Recognition Applications**]: Computer vision

General Terms

Algorithms, Measurement, Human Factors

Keywords

Visual attention, fixations, *salient* regions, eye-tracker, *abstract* semantics, concept detection

1. INTRODUCTION

As humans, *we understand what we see.* Nevertheless, since our cognitive system is designed to assimilate only some of the abundant visual information from the outside world, *we only see what we attend to.* Human Visual Attention (HVA) is the strategy employed to allocate cognitive resources for visual processing. Eye movements are an important artifact of HVA [2], and consist of stationary phases called *fixations* and rapid, ballistic eye movements called *saccades*. Visual information assimilation happens mainly for the portion of scene close to the center of gaze (foveal region), and detailed visual information is assimilated exclusively during fixations [6]. Popular computational techniques predict visual attention on the basis of bottom-up or early saliency [9]. However, the regions-of-interest predicted by such saliency algorithms often do not match with those fixated by humans [7] as HVA is dominated by top-down factors in semantically rich images, leading to characteristic gaze patterns [1].

Eye-gaze measurements have been employed in [3] to establish that visual attention is driven by the recognized and *interesting* objects in semantically rich images. Inherent association of an 'order of importance' to objects, even in everyday scenes, has been shown in [12]. Fixations on *salient* image regions have been found to be consistent across a subject population for semantically rich images [13]. We term this phenomenon as **attentional bias**.

This paper presents one of the first works to exploit attentional-bias in image understanding. We summarize by stating the key contributions of this work as follows:

- We demonstrate the extraction of *abstract image semantics* from fixation information by combining automated concept detection with fixation analysis. This is opposed to previous works that essentially employed fixation clusters as a handle to identify salient image regions [7, 13].
- We show that the *fixation sequence* can be exploited through the 'binning' algorithm. This be used to deduce object interactions for characterizing actions such as *look*, *read*, *shoot*, *etc.*.
- This is also one of the first works to investigate how fixation information can be used to enhance the performance of concept detectors.

2. RELATED WORK

An exhaustive review of research works that model the eyes and gaze is presented in [5]. *Salient* image region estimation using low-level image information has been shown

MM'10, October 25–29, 2010, Firenze, Italy.

in [9, 14]. Recently, [7] motivated salient region estimation from fixations and trained a saliency predictor using compiled eye-tracking data.

Recently, a biological perspective to actively segment images using 'fixation seeds' has also been proposed in [8], based on the assumption that eye fixations invariably fall on the interior of *salient* objects. In [11], gaze is used to achieve eye-gaze driven, interactive semi-automated cropping of images. Fixations observed consistently on *salient* image regions are exploited for semi-automated localization of image-caption labels in [13]. We now describe how fixations can be combined with concept detectors to infer abstract image semantics, beginning with a brief outline of the NUSEF [10] database used for our experiments.

3. IMAGE SEMANTICS FROM GAZE AND CONCEPT DETECTORS

3.1 Data and Experimental Protocol

We used the NUS Eye Fixation database (NUSEF) [10] compiled from a pool of 75 undergraduate and graduate volunteers aged 18-35 years, for our experiments. Fixations were acquired non-invasivelyusing the $\mathbf{ASL}^{TM}$ eye-tracker as subjects freely-viewed images. The eye-tracker is accurate within the nearest 1^o visual angle at 3 feet viewing distance leading to an on-screen error radius of 5 pixels. NUSEF comprises fixations for a pool of 758 images from diverse semantic categories, capturing objects at varying scale, illumination and orientation. The semantic image categories include *faces- normal* (neutral, smiling) and *expressive* (angry, disgust, surprise, fear), *portraits* showing both the face and body of mammals and *nudes*, *action* images containing a pair of interacting objects as in *look*, *read*, *shoot*, *world* images comprising living and non-living entities, *reptiles*, *injury*, *etc.*

3.2 Description of gaze-based measures

Semantically rich images can be represented using regions-of-interest (ROIs), with each ROI denoting a unique concept. ROIs may be overlapping, and can be generated automatically using concept detectors such as the face [15] and person [4] detectors. We observe that fixations are strongly driven by image semantics [13], Also, fixations on *salient* image regions have been found to be consistent such as 'man', 'book' in Fig.1(a). We term this *attentional-bias*.

To quantitatively model and exploit *attentional-bias* for inferring image semantics, in addition to the *fixation duration*, *bias weight* and *conditional probability* definitions introduced in [13], we also use the ROI-interaction measure defined as follows.

Let image I comprise of n ROIs, the representative interaction measure, $Int_{(l,m)}I$, which models the interaction between each key ROI pair a_l, a_m, is then defined as,

$$Int_{(l,m)I} = \overline{CP(m/l)}_I + \overline{CP(l/m)}_I \quad (1)$$

$CP(i,j)$ being the conditional probability of transition from ROI a_i to a_j. When there is a strong interaction observed between a pair of entities (concepts), extensively high number of eye-gaze transitions are observed between the entity-pair as illustrated by the *'man looks at book'* image (Fig. 1(a)), resulting in high $Int_{(l,m)I}$ values. We define *action* images as those that are characterized by a noticeable interaction between the *source* and *recipient*, as denoted by the thick green arrows in Fig.1(a,b).

(a) (b)

Figure 1: *Action* vs multiple non-interacting entities. A *read* image is shown in (a) with *man, book* being key interacting elements. (b) is an example image containing multiple non-interacting entities. Fixations are shown as yellow dots superimposed on the images. Fixation cluster centroids are marked with red circles with radius proportional to cluster size. The thickness green arrows are indicative of fixation transition probabilities between them.

3.3 'Binning' algorithm for automated ROI detection

We now describe a novel 'binning' procedure that we adopt to automatically determine spatially distinct ROIs based on *time-sequence* information. A majority of fixation transitions occur between locations corresponding to distinct, but related image ROIs, due to exploratory behavior by humans. We exploit this property of eye-gaze to discover and bound semantically related image ROIs.

The binning algorithm assigns a set of P fixation points to N bins. The algorithm begins with NULL bins. Bins are created with time, based on the spatial distribution of fixation points. Due to the exploratory behavior exhibited by users, two consecutive fixations hardly fall on the same ROI. Therefore, if fixation S_j has been assigned to bin_k, the algorithm will attempt to assign S_{j+1} to to bin_l, such that $k \neq l$, based on Euclidian distance, implying a fixation transition from ROI k to l. If the closest bin is bin_k itself, it's assumed that the subject's eye-gaze hasn't transitioned to another ROI in the image. Large distances between S_{j+1} and bin_l lead to the formation of a new bin with S_{j+1} as centroid. Bins with high membership counts, represent most *salient* image concepts. $BinAdj$ is a matrix that stores the number of transitions between bins $l, m \forall\ l, m = 1..N$. The binning procedure is summarized in Algorithm.3.1.

As illustrated in Fig.1(a), the binning procedure enables automated estimation of the ROIs as well as the extent of transitions between the ROIs. This is especially interesting because it is extremely difficult to infer object interactions such as *look*, by applying computer-vision based techniques on image or video data. Fig.1(a) is an example of a *read action* image, while Fig.1(b) shows an image containing multiple non-interacting entities. While a high number of fixations are observed around *salient* objects, the fixation density alone is insufficient to infer object interactions.

Fig.1(a),(b) show the computed bin-centroids as red circles with radius proportional to the fixation cluster size. Also, the green arrows denote the directions of fixation transitions between ROIs with the arrow thickness denoting inter-

ROI transition count. For *action* images, the fixation transitions between interacting entities are symmetrically high, while for images having multiple non-interacting objects, like in Fig.1(b), the symmetrical transitions are missing. This phenomenon permits the automated classification of *action* vs non-action images.

Algorithm 3.1: ClusterFixations($FixationData$)

$bins \leftarrow$ [**NULL**]
$BinAdj \leftarrow$ [**NULL**][**NULL**]
for each S_j *in* S
do {
  $prevBin \leftarrow null$
  for each S_j *in pattern*
  do {
    if $isempty(bins)$
    then
      $bins.create()$
      $bins(1).add(S_j)$
      $prevBin \leftarrow 1$
    else
      $foundBin \leftarrow$ *bin closest to* S_j
      $prevBin \leftarrow foundBin$
      if $foundBin == null$
      then
        $prevBin \leftarrow bins.addNewBin()$
      else
        if $foundBin \neq prevBin$
        then
          $bins(foundBin).add(S_j)$
          $BinAdj(prevBin, foundBin).addEdge()$
        else
          if $foundBin == prevBin$
          then
            $bins(foundBin).add(S_j)$
  }
}
return $(bins, BinAdj)$

3.4 Experiments and Results

In this section, we discuss how eye fixations can be utilized to infer abstract image semantics. In particular, we discuss classification of *normal* vs *expressive* face, *portrait* vs *nude* and *action* vs non-action image categories. The ratio of *attentional bias* (w_i) values between *eyes* and *nose+mouth* is employed for *normal* Vs *expressive* discrimination. A similar w_i ratio between *face* and *body* is used for *portrait* Vs *nude* classification. We use automated detectors to infer the necessary ROIs. However, for *action* images, where the interacting entities are spatially separated, concept detectors alone, are insufficient. This is owing to the fact that while concept detectors can only identify that there is a 'Man' and 'Book' for Fig.1(a), the presence or absence of inter-entity interactions has to be purely determined using gaze information. The methodologies for determining ROIs automatically in *face* and *person* images are the same as described in Section 3.2.

For classification, we perform leave-one-out cross-validation, *i.e.*, all but one instance is used training data, while the chosen one is used as test data. The training data is then used to learn representative $w_i/Int_{(l,m)I}$ for the classes involved ($Int_{(l,m)I}$ is employed for *action* images only). This process is repeated until all images are chosen for the test data. Table.1 (rows 1-4) presents the classification results for *face* and *person* images. An overall accuracy of 69.6% and 60.2% are obtained for the *face* and *person* classes respectively. Results obtained for 70 *action* images are also presented in Table. 1 (row 5,6). Overall, correct action classification is achieved for 62.5% of the images.

Category	Instances	Correctly Classified	Accuracy
Normal Face	**37**	**28**	**0.76**
Expressive Face	**25**	**15**	**0.6**
Nude	**32**	**18**	**0.57**
Person	**36**	**23**	**0.63**
Action	**34**	**21**	**0.62**
No Action	**36**	**23**	**0.63**

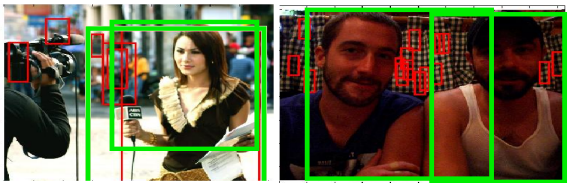
Table 1: Row 1-4 demonstrate the combination of concept detectors and fixations to classify *face* and *person* images. Rows 5,6 demonstrate classification results for *Action* and *no-Action* images.

4. USING VISUAL ATTENTION TO GUIDE OBJECT DETECTION

We now present a novel framework demonstrating the effectiveness of human eye-gaze in guiding state-of-the-art object detectors. Sliding window based object detectors such as [4], are essentially image classifiers. A trained classifier is used to exhaustively inspect rectangular regions over successively scaled down versions of the input image. Detection scores from detections at successive levels are then combined across multiple scales to identify image regions with maximum likelihood of the object-presence. Lacking prior knowledge of the location or size of key objects in the image, object detectors search exhaustively through an exponentially large search space of windows. For example, a 1024x768 image consumes 15-20 seconds to be searched in totality by the [4] detector on a standard PC (Pentium Core 2 Duo, 2 Ghz, 2 Gb RAM). In the next section, we demonstrate the effectiveness of visual attention in guiding a state-of-the-art detector [4]. We show how object detectors can achieve higher detection rates within shorter time-spans, when guided by fixation clusters.

4.1 Using eye-gaze information to guide object detection

If object search is limited to within *ROIs* obtained as described in Sec. 3.3, detectors are then limited to operate on a fraction of the scales in the image pyramid. The scales chosen are the ones where the sliding window size is close to the size of ROIs, which in most cases, is close to the size of the *salient* object. More formally, trained model is a filter F of size $w \times h$. Let H be the feature pyramid extracted from the successively resized images and $pos(x, y, l)$ be a position (x, y) in the lth level of the pyramid. The score of F at $pos(x, y, l)$ then is $F * \psi(H, pos, w, h)$, where $\psi(H, pos(x, y, l), w, h)$ is the vector obtained by concatenating feature vectors in the $w \times h$ sub-window at level l. The final likelihood is then obtained by combining scores so obtained across different levels. The exact score generation and combination strategy can vary across specific detector implementations.



(a) (b)

Figure 2: visual attention greatly enhances performance of the baseline detector [4] are shown in *red* and those after enhancement using HVA is shown in *green*.(b) Fine grained scale selection enables detection of key objects missed by [4].

4.2 ROI size estimation and scale control to reduce false positives

Eye fixations are most useful for controlling false positives obtained from object detectors (red boxes in the Fig. 2 (*a*). Scale selection is enforced on ROIs by choosing levels l from all pyramid levels L such that the area of resized ROI is close to the sliding window area at these levels, i.e; $\frac{area(l)}{w \times h} \approx 1$. This is akin to creating a partial pyramid with finer grained resizing scales.

4.3 Experimental results and Discussion

We demonstrate the applicability of our method using a generic object detector that has been the top performing system in the recent PASCAL VOC 2009 challenge [4]. A subset of 200 images from our dataset is chosen for evaluation corresponding to *person*, *dog*, *cat* and *bird*.

We combine $precision$ and $recall$ using an $fmeasure$ score computed over detection boxes ($bbox$) with respect to human annotated ground-truth ($gtruth$) boxes as, $fmeasure = \frac{2*precision*recall}{precision+recall}$. Where, $precision = \frac{bbox \cap gtruth}{bbox}$ and $recall = \frac{bbox \cap gtruth}{gtruth}$. The evaluation over 120 images from the concept *person* yields a 18% improvement in $fmeasure$.

Our method is independent of the application that the ROIs are put to, and in this case, the specific object-detector employed therein. We demonstrate this by considering $fmeasures$ for ROI boxes generated from eye-gaze data as detections by a hypothetical detector and comparing against human annotated ground-truth boxes as shown in Table. 2.

Category	Images	Fmeasure-VA	Fmeasure-VOC
Person	120	**0.3**	**0.34**
Bird	40	**0.41**	**NA**
Cat or Dog	40	**0.43**	**NA**

Table 2: Evaluation of the visual attention guided *ROIs* against human annotated ground truth by considering ROI as a detection. This is especially significant in cases like *bird and cat/dog*, where the automatic detector [4] fails completely.

5. CONCLUSION

In this paper, we present one of the first works that employs gaze information in conjunction with concept detectors to enhance image understanding. While fixation distribution amongst salient ROIs is exploited to distinguish between normal/expressive face and portrait/nude images, timing information of fixation data is critical for discovering inter-entity interactions using the novel binning procedure, in *action* images. Incorporating fixation information in the detection framework improves the accuracy of concept detectors, and significantly reduces computational time.

6. REFERENCES

[1] D. Agrafiotis, S. J. C. Davies, N. Canagarajah, and D. R. Bull. Towards efficient context-specific video coding based on gaze-tracking analysis. *ACM Trans. Multimedia Comput. Commun. Appl.*, 3(4):1–15, 2007.

[2] G. Buscher, E. Cutrell, and M. R. Morris. What do you see when you're surfing?: using eye tracking to predict salient regions of web pages. In *CHI '09: Proceedings of the 27th international conference on Human factors in computing systems*, pages 21–30, 2009.

[3] W. Einhï£¡user, M. Spain, and P. Perona. Objects predict fixations better than early saliency. *J. Vis.*, 8(14):1–26, 11 2008.

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 99(PrePrints), 2009.

[5] D. Hansen and Q. Ji. In the eye of the beholder: A survey of models for eyes and gaze. *PAMI*, 32(3):478–500, March 2010.

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

[7] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *ICCV*, 2009.

[8] A. Mishra, Y. Aloimonos, and C. L. Fah. Active segmentation with fixation. In *ICCV*, 2009.

[9] R. J. Peters, A. Iyer, L. Itti, and C. Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(8):2397–2416, Aug 2005.

[10] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua. An eye fixation database for saliency detection in images. In *ECCV 2010 (accepted)*, 2010.

[11] A. Santella, M. Agrawala, D. DeCarlo, D. Salesin, and M. Cohen. Gaze-based interaction for semi-automatic photo cropping. In *CHI '06: Proceedings of the SIGCHI conference on Human Factors in computing systems*, pages 771–780, 2006.

[12] M. Spain and P. Perona. Some objects are more equal than others: Measuring and predicting importance. In *ECCV '08*, pages 523–536, 2008.

[13] R. Subramanian, H. Katti, R. Huang, T.-S. Chua, and M. Kankanhalli. Automated localization of affective objects and actions in images via caption text-cum-eye gaze analysis. In *ACM Multimedia 2009 - Human-Centered Multimedia Short Paper Track*, 2009.

[14] R. Valenti, N. Sebe, and T. Gevers. Image saliency by isocentric curvedness and color. In *ICCV*, 2009.

[15] P. Viola and M. J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, May 2004.