From Where and How to What We See

S. Karthikeyan*, Vignesh Jagadeesh*, Renuka Shenoy*, Miguel Eckstein‡, B.S. Manjunath*

*Center for Bio-Image Informatics
*Department of Electrical and Computer Engineering
‡Department of Psychological and Brain Sciences
University of California Santa Barbara
*{karthikeyan, vignesh, renuka, manj}@ece.ucsb.edu , ‡eckstein@psych.ucsb.edu

Abstract

Eye movement studies have confirmed that overt attention is highly biased towards faces and text regions in images. In this paper we explore a novel problem of predicting face and text regions in images using eye tracking data from multiple subjects. The problem is challenging as we aim to predict the semantics (face/text/background) only from eye tracking data without utilizing any image information. The proposed algorithm spatially clusters eye tracking data obtained in an image into different coherent groups and subsequently models the likelihood of the clusters containing faces and text using a fully connected Markov Random Field (MRF). Given the eye tracking data from a test image, it predicts potential face/head (humans, dogs and cats) and text locations reliably. Furthermore, the approach can be used to select regions of interest for further analysis by object detectors for faces and text. The hybrid eye position/object detector approach achieves better detection performance and reduced computation time compared to using only the object detection algorithm. We also present a new eye tracking dataset on 300 images selected from ICDAR, Street-view, Flickr and Oxford-IIIT Pet Dataset from 15 subjects.

1. Introduction

Wearable eye tracking devices are becoming popular [4, 5] and will soon be mainstream. They will provide a platform to collect eye tracking data in a non-intrusive way when people observe multimedia content, such as web browsing. This additional information from multiple subjects can potentially be useful for challenging large scale multimedia annotation problems. Towards this, we propose a technique to obtain image-level scene semantic priors from eye tracking data, which will reduce the search space for multimedia annotation tasks.

It is known that human visual attention, irrespective of top-down task is biased towards faces and text [7]. The first step towards obtaining scene semantic prior from eye tracking information alone is to build models that predict face and text regions in images, which is the primary focus of the paper. This information is useful to improve the speed and precision of state-of-the-art detectors for challenging categories such as text, cats and dogs. We note that the performance of state-of-the-art cat and dog detectors [24] in turn depends on head (face) detection algorithm which can be enhanced using eye movement information.

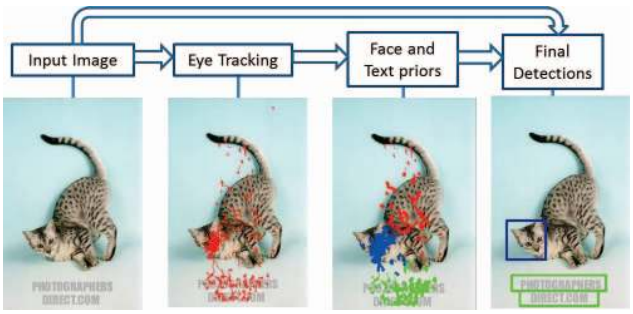


Figure 1: Left to right: 1. Input image. 2. Eye Tracking fixation samples from multiple subjects overlaid on the image 3. The eye tracking regions identified by the proposed algorithm as faces (blue) and text (green) 4. The final detection outputs of face and text detector focusing on the priors provided by eye tracking. Best viewed in color.

Related Work

Humans are able to swiftly process a rich stream of visual data and extract informative regions suitable for high level cognitive tasks. Therefore, there has been significant amount of research on human inspired visual attention models [17, 15, 18, 20]. These approaches typically predict the attention in different regions of an image given low-level saliency maps and high-level image semantics. In contrast, the proposed problem in spirit models the converse situation of predicting image semantics from eye movement data.

There have been some recent efforts which model top-down semantics by simultaneously utilizing both image and eye movement information. In this regard, Subramanian et al. [27] extract high-level information from images and verbal cues, (faces, face parts and person) and model their interrelationships using eye movement fixations and saccades across these detections. Mishra et al. [22] propose an active segmentation algorithm motivated by finding an enclosing contour around different fixations. The proposed approach distinguishes itself as it aims to speed up algorithms for high-level semantics from eye movement data alone. Bulling et al. [3] propose an activity classification method in office environments (copying text, reading, browsing web, taking notes, watching video) using eye movement data collected using electrooculography. As most of these activities follow a standard repetitive pattern, the method in [3] predicts the activities reliably for each person individually. However, due to variability in the manner in which different people view images, our approach differs from [3] and we require data from multiple observers to predict image semantics reliably. Cerf et al. [6] provide an algorithm to decode the observed image using eye movement scanpath data. However, their approach models the problem by proposing a metric between multiple saliency maps obtained from the image and the scanpath data. The saliency map generation problem again requires processing the entire image and is inherently different from the proposed approach. We make three contributions in this paper

a. We propose an algorithm to localize face and text regions in images using eye tracking data alone. The algorithm basically clusters the eye tracking data into meaningful regions using mean-shift clustering. Following which various intra- and inter-cluster fixation and saccade statistics are computed on these clusters. The final cluster labels are inferred using a fully connected MRF, by learning the unary and interaction potentials for faces and text from these statistics.

b. We demonstrate the ability of these face and text priors to improve the speed and precision of state-of-the-art text [13] and cat and dog detection [24] algorithms.

c. We also present a new eye tracking dataset, collected on images from various text, dogs and cats datasets. The dataset was collected on 300 images from 15 subjects.

Fig. 1 outlines the pipeline of the proposed approach.

2. Faces and Text Eye Tracking Database

We collected an eye tracking dataset, with primary focus on faces (humans, dogs and cats) and text using Eyelink 1000 eye tracking device. The image dataset consists of 300 images collected from ICDAR datasets (text) [21], Street view dataset (text) [30] and Oxford-IIIT Pet dataset (dogs and cats) [23] and flickr images [18]. The text images were gathered from two different datasets to ensure considerable variability in scene context. The flickr images provide sufficient representation for images without text or faces (including dogs and cats) in both indoor and outdoor scenes. The overall image dataset consists of 61 dogs, 61 cats, 35 human faces, 246 text lines and 63 images without any text or faces. Fig. 2 highlights examples for images from different categories from the dataset. The images were of dimension 1024×768 and were viewed by 15 subjects (between ages 21 and 35). The viewers sat 3 feet away from a 27 inch screen and each image was shown for 4 seconds followed by 1 second viewing a gray screen. The subjects were informed that it was a free viewing experiment and instructed to observe regions in images that gather their interest without a priori bias. Also, eye tracking calibration was performed every 50 images and the entire data was collected in two sessions (150 images each). This dataset can be downloaded from `http://vision.ece.ucsb.edu/~karthikeyan/facesTextEyetrackingDataset/`

Figure 2: Examples of images from our dataset consisting of text, human faces, dogs, cats and other background objects

Humans eye movement scanpaths typically consists of alternating fixations and saccades. Fixations represent information gathering sequences around an interest region and saccades indicate transitions between fixations. The eye tracking host computer samples the gaze information at 1000 Hz and automatically detects fixations and saccades in the data. Therefore, we have around 4000 samples per subject for every image. The fixation samples typically account for 80% of the entire data. In our analysis we only use the fixation samples and henceforth refer to these fixation samples as the eye tracking samples. The eye tracking device also clusters the fixation samples and identifies fixation and saccade points. We refer to these points as fixations and saccades hereafter. The average number of fixations and saccades per image across subjects can vary from 8 to 19. In our experiments, the first fixation and saccade was removed to avoid the initial eye position bias due to the transition gray slide in the experimental setup.

Face Regions: The dataset consists of faces of multiple

sizes, varying from about 40×40 to 360×360 pixels. In small face images, subjects look at the face as a whole. On the other hand, in larger faces there are several saccades across eyes, nose and mouth regions. As expected, face regions consistently attract attention from viewers. In addition we notice that the initial saccades are invariably directed towards face regions across subjects. In images consisting of multiple faces, rapid scanpaths moving across different faces is a common phenomenon. Fig. 3 illustrates some examples featuring some of these effects.

Text Regions: We refer to entire lines/sentences as text regions. These are present in various styles, fonts, sizes, shapes, lighting conditions and with occlusions in our image dataset. In text regions consisting of a single word, the subjects typically fixate around the center of the word and the different fixations take a nearly elliptical shape. In multiple words, we observe saccadic scanpaths from one word to another as subjects typically read the different words sequentially. Fig. 3 illustrates some example text regions in our image dataset.

3. Faces and Text Localization from Eye Tracking Data

The aim is to identify face and text regions in images by analyzing eye tracking information from multiple subjects, without utilizing any image features. Eye movements are organized into fixations and saccades. The information gathering phase is represented by the fixations, which typically group around different semantic/interesting regions as shown in Fig. 3. Therefore, we first cluster all the fixation regions using the mean-shift clustering technique [10] with a fixed bandwidth. We chose mean-shift clustering as it does not require the number of clusters and is fairly robust to multiple initializations for the selected bandwidth (50 pixels). The text and face region detection problem is mapped to a cluster labeling problem. Therefore, we compute inter-cluster and intra-cluster statistics and model the labeling problem using a fully connected Markov Random Field (MRF).

Let the i^{th}cluster in an image be denoted by $\mathcal{C}_i$. The 2D eye tracking samples (fixation samples) within the cluster are represented by E_i. The fixations (fixation points) and saccades in the entire image are denoted by $\mathcal{F}$ and $\mathcal{S}$ respectively. The fixations belonging to the i^{th} cluster are denoted by F_i and the saccades originating from i^{th} and terminating in the j^{th} by $S_{i,j}$. Finally, the fixations provided by every individual person k in cluster i is augmented giving F_i^k and the corresponding times (0-4 seconds) representing the beginning of the fixations in cluster i is given by T_i^k. The following features are used to represent inter-cluster and intra-cluster properties.

3.1. Intra-cluster features

a. Number of fixations and eye tracking samples: $|F_i|$, $|E_i|$

b. Standard deviation of each dimension of the eye tracking samples E_i

c. Shape and orientation of the cluster by ellipse approximation. Let λ_1,λ_2 and v_1,v_2 denote the two eigenvalues and eigenvectors respectively of the cluster such that $\lambda_1 > \lambda_2$. Shape of the cluster is encoded by $\frac{\lambda_2}{\lambda_1}$. The orientation is expressed as $|\angle v_1|$

d. The ratio of the eye tracking sample density in the cluster compared to its background. Let cluster $\mathcal{C}_i$ be approximated by the minimum rectangle R_i containing all the cluster points. The rectangular region centered around R_i which is twice its width and length is defined as D_i. Hence, the background region, B_i, around R_i is expressed as $D_i \setminus R_i$. The final feature is computed as $\frac{|\{E_i \in B_i\}|}{|\{E_i \in R_i\}|}$

e. Number of incoming, outgoing and within-cluster saccades, represented by $\sum_{\forall j \neq i} |S_{j,i}|$, $\sum_{\forall j \neq i} |S_{i,j}|$ and $|S_{i,i}|$ respectively

f. The number of incoming, outgoing and within-cluster saccades, (from *e*) where the saccade angle to the X-axis is less than 30 degrees (almost horizontal)

g. The percentage of incoming, outgoing and within-cluster saccades (from *e*) which are almost horizontal

h. Median of the time of first visit to the cluster across multiple subjects: $\text{median}_k \left(\min_i \left(T_i^k\right)\right)$

i. Median of the number of times each person visits a cluster: $\text{median}_k(\left|F_i^k\right|)$

In total we have 18 intra-cluster features representing each cluster's intrinsic properties. These features essentially aim to capture the eye movement attributes typical of face and text regions described in Section 2. The features *a,b,c,d* and *e* are important basic features where text and face regions exhibit characteristic responses. Features *f* and *g* are more characteristic of text regions with multiple words as nearly horizontal inter-word saccades are prevalent. Finally, features *h* and *i* are more relevant to face regions which typically immediately attract viewer attention. In addition subjects also tend to visit the face multiple times after fixating at other regions in the image, which is captured by feature *i*.

3.2. Inter-cluster features

In addition to intra-cluster features, pairwise inter-cluster features also provide useful information to identify face and text regions. In the presence of multiple faces, subjects indicate saccadic activity across the different faces. Moreover, in text images with multiple words, inter-word saccadic activity is quite common. Therefore, the following saccade centric features are computed across clusters.

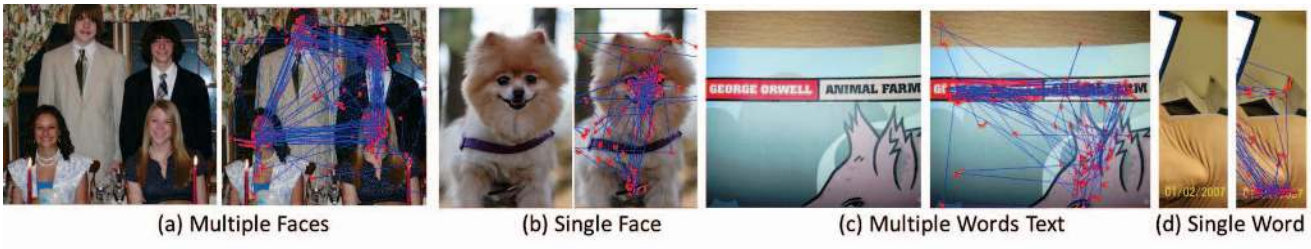


Figure 3: Shows example of faces and text in two scenarios each. The fixations are marked as red points and saccades as blue lines. (a) Multiple faces in the image where we consistently observe inter-face saccades. (b) A large single face where several saccades are observed in the eyes, nose vicinity. (c) Text with four words where a dense saccadic presence is observed between words. (d) A clip from one of the images showing a single word, whose cluster takes a nearly elliptical shape. Best viewed in color.

1. Number of saccades from the i^{th} to j^{th} cluster, $|S_{i,j}|$ and vice versa.

2. Number of almost horizontal saccades (where the saccade angle to the X-axis is less than 30 degrees) from cluster i to j and vice versa.

3. Percentage of almost horizontal saccades from cluster i to j and vice versa.

4. The number of saccades, horizontal saccades and percentage of horizontal saccades from the left cluster to the right cluster.

5. Distance between the clusters.

In total, we have 13 inter-cluster features to represent saccadic properties across multiple clusters. Specifically, features *1,2* and *3* are useful indicators of face-face and text-text regions. Also, feature *4* is targeted to capture text regions as subjects typically read text from left to right.

Figure 4: Visualizing the text MRF potentials. Left to right: 1. (Left) Input image. 2. (Center) Clustered eye tracking fixation samples from multiple subjects overlaid on the image 3. (Right) Visualizing the unary and interaction potentials of the clusters for the text MRF. The unary is color coded as red, the bright values indicating high unary potentials of a cluster belonging to text class. The interaction is marked by the blue lines between clusters, whose thickness is indicative of text-text interaction magnitude. Best viewed in color.

3.3. Learning Face and Text regions

Utilizing the features in Section 3.1 and Section 3.2, we propose a probabilistic model based technique to label the clusters provided by mean-shift algorithm [10] on the eye tracking samples. The intra- and inter-cluster features are naturally modeled as a MAP inference problem

Data: Input Images $\{\mathcal{I}^i\}$, Eye Tracking Samples $\{\mathcal{E}^i\}$, Fixations $\{\mathcal{F}^i\}$, Saccades $\{\mathcal{S}^i\}$ ground truth labels for faces and text $\{\mathcal{L}^i\}$, $i \in [1...N]$, N is the total number of images
Result: Face Cluster IDs^i, Text Cluster IDs^i,$i \in [1...N]$
Notation : Superscript indicates image number and subscripts refer to cluster IDs in an image
Precomputing Cluster Features:
for $i = 1 \rightarrow N$ **do**
  $\mathcal{C}^i$ = Mean Shift Clustering($\mathcal{E}^i$);
  for $j = 1 \rightarrow |\mathcal{C}^i|$ **do**
    $\mathcal{F}intra^i_j$ = Intra-cluster-features($\mathcal{C}^i_j, \mathcal{F}^i_j, \mathcal{S}^i_j$) ;
    $\mathcal{C}lab^i_j$ = Cluster-labels($\mathcal{L}^i_j, \mathcal{C}^i_j$);
    for k=$j+1 \rightarrow |\mathcal{C}^i|$ **do**
      $\mathcal{F}inter^i_{jk}$ = Inter-cluster-features($\mathcal{C}^i_j, \mathcal{F}^i_j, \mathcal{S}^i_j, \mathcal{C}^i_k, \mathcal{F}^i_k, \mathcal{S}^i_k$)
    end
  end
end
Learning to classify Clusters into Face and Text regions:
for $i = 1 \rightarrow N$ **do**
  TestIndex = i; TrainIndex = $\{1, 2, ..., N\} \setminus \{i\}$;
  [Unary Potentials Face, Unary Potentials Text] = QDA($\mathcal{F}intra^{\text{TestIndex}}, \mathcal{F}intra^{\text{TrainIndex}}, \mathcal{C}lab^{\text{TrainIndex}}$);
  [Pairwise Potentials Face, Pairwise Potentials Text]= QDA($\mathcal{F}inter^{\text{TestIndex}}, \mathcal{F}inter^{\text{TrainIndex}}, \mathcal{C}lab^{\text{TrainIndex}}$);
  Face Cluster IDs^i = MRF$_{\text{face}}$(Unary Potentials Face, Pairwise Potentials Face);
  Text Cluster IDs^i = MRF$_{\text{text}}$(Unary Potentials Text, Pairwise Potentials Text);
end

Algorithm 1: Proposed method to detect face and text regions in images by analyzing eye tracking samples.

using a MRF. The different clusters represent the nodes of the graph. The intra-cluster and inter-cluster features facilitate the learning of unary and pairwise potentials respectively. In addition, we utilize a fully connected graph to ensure long range interactions. Let the posterior probabilities of a quadratic discriminant analysis (QDA) classifier on intra-cluster features be denoted by p, the unaries are calculated as $-log(p)$. Similarly the pairwise potential is obtained as $-log(q)$, where q is the posterior learnt from the inter-cluster features using QDA. The problem of infer-

ring the labels y_i of $\mathcal{C}_i$ is modeled by an MRF with energy

$$E = \sum_{i \in \mathcal{C}} V_i(y_i) + \sum_{i,j \in \mathcal{C}, i \neq j} V_{ij}(y_i, y_j) \quad (1)$$

where V_i denotes the unary potential of cluster i and V_{ij} denotes the scaled pairwise potential between clusters i and j with a scaling factor λ. In order to allow overlapping text and face regions (in watermarked images), cope with limited availability of data with face-text co-occurrence, and speed up inference, we resort to separately tackle the face,non-face and text,non-text problems using two distinct MRFs. Finally, as we are dealing with a binary inference problem on limited number of clusters (< 20), we utilize fast exact inference by pre-computing all the possibilities for different number of nodes.

4. Performance of Face and Text Localization from the Eye Tracking Samples

In this section we analyze the performance of the cluster-level classification of faces and text regions in images. To enable this, we require cluster labels from ground truth bounding box annotations. The cluster labels are defined as the the label of the class (face, text and background) which has the most representation among the cluster samples. Fig. 5 shows an example of cluster labels obtained from ground truth boxes. For this experiment we fix the bandwidth of both the face and text MRFs to 50. The parameter λ which weighs the interaction relative to the unary potentials is fixed as $\frac{1}{|\mathcal{C}^i|}$ (to roughly give equal weights to unary and pairwise potentials), where $\mathcal{C}^i$ is the set of all clusters in the i^{th} image. In addition, clusters which have less than 1% of the total number of eye tracking samples are automatically marked as background to avoid trivial cases. The total number of clusters range from 3 in low entropy images to 17 in high entropy images.

Figure 5: Left: Input image with the ground truth for face (blue) and text (green). Center: Clustered eye tracking data overlay on input image. Right: Face (blue) and text (green) cluster labels propagated from ground truth. Best viewed in color.

The performance of the cluster detection problem is evaluated using a precision-recall approach for face and text detection. Precision and recall are defined as follows

$$\text{Precision} = \frac{TP}{TP + FP}, \qquad \text{Recall} = \frac{TP}{TP + FN} \quad (2)$$

where TP, FP and TN denote true positive, false positive and true negative clusters respectively in the detection problem. Finally, to get a single measure of the performance, F-measure is defined as the harmonic mean between precision and recall. In order to utilize these cluster labels to enhance text and cat and dog detection algorithms, we require high recall under reasonable precision. This ensures most of the regions containing faces and text are presented to the detector, which will enhance the overall performance.

Figure 6: Examples of good face detections from the proposed algorithm. Red fixation points correspond to face and blue corresponds to background. (a) In the presence of salient distracting object (shoe) the face (cat) is reliably detected. (b) We notice that even in challenging scenarios where multiple faces are present, the proposed approach detects reliably. Best viewed in color.

Figure 7: Examples of good text detections from the proposed algorithm. Red fixation points correspond to text and blue corresponds to background. (a) Text line is reliably detected even in the presence of several other fixations near the region of interest. (b) Text is detected correctly in the presence of more salient object (person face). Best viewed in color.

Figure 8: Example scenario where the proposed approach fails to detect face (left) and a text word (right). The eye tracking samples detected as face in (a) and text in (b) are shown in red and the samples detected as background (both (a) and (b)) are indicated in blue. Best viewed in color.

The performance of the face and text detector MRFs are shown in Table 1. The results are evaluated at two levels, cluster and image. The image level metric evaluates the presence of at least one face/text region in an image. The cluster level metric evaluates the presence of face/text in every cluster. We notice that the recall is high for both face and text detection sections. However, the precision of the face detector is also quite high (both cluster and image level), indicating that the proposed algorithm is confident about the regions which it detects as a face. In the text region as well, we observe that the precision is fairly high, indicating the

excellent localization ability of our algorithm. Fig. 6 shows some example images where the proposed approach localizes faces well. Similarly Fig. 7 highlights some text cluster detection examples. Fig. 8 also highlights a few failure cases where both the face and text localization fails. The face detector fails as many subjects do not concentrate on the face in the corner of the image. In addition the text cluster detection fails as the allocated time (4 seconds) was insufficient to scan the entire text content.

	Precision	Recall	F-Measure
Face Detection Cluster	0.671	**0.954**	0.788
Text Detection Cluster	0.748	**0.942**	0.834
Face Detection Image	0.755	**0.989**	0.856
Text Detection Image	0.610	**0.923**	0.735

Table 1: Indicates performance of cluster and image level face and text detection from the eye tracking samples.We notice that the recall (marked in bold) is high suggesting that the proposed approach seldom misses face and text detections in images. This is achieved at a sufficiently good precision ensuring that this method can be valuable to localize ROI to reduce the search space for computationally expensive face and text detectors.

5. Applications

There have been several efforts to model context [28, 1, 12, 16, 11] in single and multi-class object detection problems. The proposed faces and text eye tracking priors can be an extremely useful alternate source of context to improve detection. Therefore, we investigate the utility of these priors for text detection in natural scenes as well as cat and dog detection in images. Due to the presence of fast and robust human face detectors [29], we do not explore human face detection problem in this paper.

5.1. Detecting Cats and Dogs

Detecting cats and dogs in images is an extremely difficult task as they have high variability in appearance and pose coupled with occlusions. However, in these problems, the animal face/head is the most distinctive part and the state-of-the-art cat and dog detection algorithm proposed by Parkhi et al. in [24] makes use of this information. The final detection algorithm consists of two steps, the head/face detection and segmenting the cat/dog body by learning features from the face. The head detection used deformable parts model [14] and the body segmentation utilized iterative graph cuts [25, 2] by learning foreground and background properties . For a detailed review of the approach we refer the reader to [24].

The proposed eye tracking based face detection prior can significantly reduce the search space for cat and dog faces/heads in images. As human fixations are typically focused towards the eyes and nose of the animals, we construct a bounding box around the face clusters to localize the cat head. When the cluster is approximated by a rectangular bounding box R with width w and length l containing all the eye tracking samples, an outer bounding box B centered around R of size $2.7l \times 2.2w$ always contained the entire face within the box. Even under this conservative approximation, the search space for cat/dog faces is reduced to 15.3% of the entire dataset (image area) using the proposed eye tracking based face detection model.

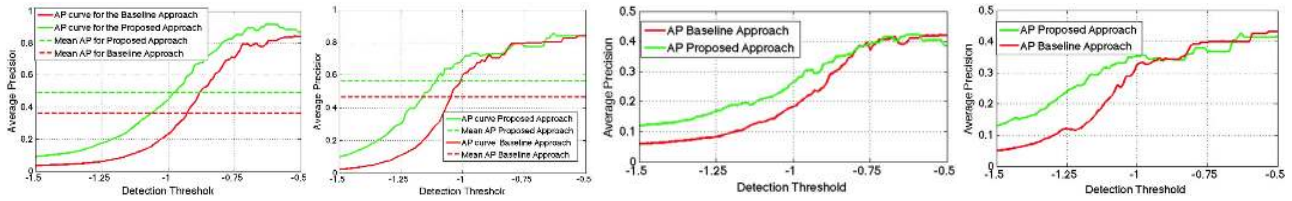
Figure 9: Example cat and dog face (blue box) and body (green box) detections from the proposed algorithm. Best viewed in color.

Figure 11: An example scenario where the head detector of the proposed approach (b) operating only in the attention region (c) marked in blue outperforms the baseline cat head detector (a). The baseline detector has a false detection as noticed in (a). Finally, red points in (c) denotes the cluster identified as face/head from which the blue attention region is constructed. Best viewed in color.

Fig. 10 shows the Average Precision curves using multiple detection thresholds for the head detection for both cats and dogs. We notice that the head detection performed only in the rectangular regions B is consistently higher than baseline (in the entire image). Especially in high recall scenarios (low detection threshold), the average precision of the proposed approach is significantly greater than the baseline approach [24]. In the whole body detection problem as well, the proposed approach outperforms the baseline approach over a larger detection threshold range. In addition, *the cat and dog head detection algorithms are 4.8 and 5.7 times faster respectively* as they operate in the reduced search space. Therefore, we achieve dual benefits of better detection performance with considerable speed-up for dog and cat detection problems. We note that the time of the proposed algorithm which we use for comparison includes the face cluster labeling overhead as well. Finally, Fig. 9 illustrates some dog and cat, head and body detection examples and Fig. 11 presents an example scenario where the proposed cat face detection approach outperforms baseline as it limits the search ROI.

5.2. Detecting Text

Detecting text in natural scenes is an important problem for automatic navigation, robotics, mobile search and several other applications. Text detection in natural scenes is challenging as text is present in a wide variety of styles, fonts and shapes coupled with geometric distortions, var-



(a) Cat Head (b) Dog Head (c) Cat Body (d) Dog Body

Figure 10: Plotting Average Precision (AP) of Cat head (a) Dog head (b), Cat Body (c) and Dog Body (d). The proposed (green) and baseline (red) curves are plotted against the detector threshold of deformable parts model. The maximum AP of baseline and proposed algorithm is comparable in all cases, however, the AP of the proposed approach is higher than baseline in high recall scenarios (low detector threshold) for both the head and body detectors of cats and dogs. Therefore, on an average the proposed approach is more stable over the detector threshold parameter than the baseline. Best viewed in color.

ied lighting conditions and occlusions. Text detection approaches are divided into texture based and connected component (CC) based approaches. Texture based approaches typically learn the properties of text and background texture,[9, 31] and classify image regions into text and non-text using sliding windows. Connected component (CC) based approaches [8, 26] group pixels which exhibit similar text properties. The grouping happens at multiple levels : character, word and sentence. This is followed by a geometric filtering technique which removes false positives. Stroke width transform (SWT) [13] is an elegant connected component based approach which groups pixels based on the properties of the potential text stroke it belongs to. We utilize SWT as the baseline text detection algorithm as it obtained state-of-the-art results in the text detection datasets [21, 30] from which we obtained the images.

The first step of SWT is edge detection and the quality of edges primarily determine the final text detection performance [19]. The presence of several false edges especially in highly textured objects leads to false detections and therefore we propose an edge subset selection procedure from text priors obtained by labeling the eye tracking samples. A connected component edge map is obtained from the canny edges and we retain connected components that are sufficiently close to regions labeled as text. This is implemented by convolving the eye tracking samples using a gaussian filter of variance 150 pixels (conservative selection) and obtaining a binary text attention map in the image plane by selecting regions which are above a threshold (0.02 in our case). In the following step, connected components of the edges which have an average text attention > 0.4 are retained for text detection.

	Precision	Recall	F-Measure	Mean Edges
SWT	0.436	0.669	0.530	6723
Our Method	0.599	0.655	0.625	19745

Table 2: Comparison of the performance of the proposed text detector with eye tracking prior and baseline SWT. There is significant gain in the precision ($\sim$37% compared to baseline) for a small loss in recall ($\sim$2%). This results in improved overall F-Measure.

The performance of the text detection is validated using standard precision-recall metrics popular in text detection literature[13]. Table 2 quantifies the improvements due to the proposed approach in precision and F-Measure of the text detector. *We notice significant gain in precision and F-Measure, about 37% and 15% respectively, compared to baseline SWT*. Table 2 also indicates that *we need to process only 34% of the edges in the dataset which makes the proposed approach 2.82 times faster than baseline SWT*. We note that the time of the proposed algorithm which we use for comparison includes the text cluster labeling overhead as well. Fig. 12 highlights some example detections from the proposed algorithm. Fig. 13 compares some results of the proposed approach to baseline SWT and indicates the utility of the text attention map to limit the ROI for text detection. In summary, we obtain significantly better detector precision than baseline SWT in considerably lower detection time.

Figure 12: Examples of images where the proposed text detection approach performs reliably. Best viewed in color.

6. Discussion, Conclusions and Future Work

This paper is the first attempt at interpreting image semantics from the manner in which multiple subjects look at these images in a free viewing task. Consequently, we generate semantic priors by analyzing eye tracking samples without image information. We focused on two semantic categories, faces and text, and collected a new eye tracking dataset. The dataset consists of 300 images with 15 subjects with specific focus on humans, dogs, cats and text in natural scenes. The eye tracking fixation samples are clustered using mean-shift. Intra- and inter-cluster features are computed which eventually maps to a labeling problem using an MRF. The proposed approach obtains promising results in classifying face and text regions from background by only analyzing eye tracking samples. This information provides a very useful prior for challenging problems which require robust face and text detection. Finally the proposed seman-

Figure 13: Two example scenarios ((a)-(c) and (d)-(f)) where SWT results ((a) and (d)) are outperformed by the proposed approach ((b) and (e)). The attention regions ((c) and (f)) shows the eye tracking samples classified as text in red and the ROI used by the text detector in blue. Therefore, as the false positive portion in SWT (red boxes in (a) and (d)) is removed by the generated text attention region, we obtain better detector precision in these images.

tic prior in conjunction with state-of-the-art detectors obtains faster detections and higher precision results for dog, cat and text detection problems compared to baseline.

The proposed approach also has a few limitations. If the face image almost occupies the entire screen, multiple clusters at different face parts will be formed and our dataset does not provide sufficient samples to model this behavior. Furthermore, if the image has a large number of text lines, the subjects do not have sufficient viewing time to gather all the information presented. This can be handled by allowing the subject to control viewing time. Both these issues will be addressed in future extensions of this work.

In addition, we will explore better localization of face and text regions for the detectors from the eye tracking information. Perhaps one could learn the relationship between the ground truth bounding boxes and the cluster properties. Additionally, an edge learning technique from the cluster labels for the text class could improve the proposed text detection algorithm. Finally, we will also investigate learning eye tracking priors for other semantic categories and over video sequences from multiple subjects.

Acknowledgements This work was supported in part by the following awards/grants: US Office of Naval Research N00014-12-1-0503, Institute for Collaborative Biotechnologies through grant W911NF-09-0001 from the U.S. Army Research Office, and the National Science Foundation awards III-0808772 and OIA-0941717. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

[1] B. Alexe et al. Searching for objects driven by context. *Advances in Neural Information Processing Systems 2012*.

[2] Y. Boykov et al. Fast approximate energy minimization via graph cuts. *Pattern Analysis and Machine Intelligence 2001*.

[3] A. Bulling et al. Eye movement analysis for activity recognition using electrooculography. *IEEE Transactions on Pattern Analysis and Machine Intelligence 2011*.

[4] A. Bulling et al. *Wearable EOG goggles: eye-based interaction in everyday environments*. ACM, 2009.

[5] A. Bulling and H. Gellersen. Toward mobile eye-based human-computer interaction. *Pervasive Computing, IEEE*, 2010.

[6] M. Cerf et al. Decoding what people see from where they look: Predicting visual stimuli from scanpaths. *Attention in Cognitive Systems 2009*.

[7] M. Cerf et al. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision 2009*.

[8] H. Chen et al. Robust text detection in natural images with edge-enhanced maximally stable extremal regions. *IEEE International Conference on Image Processing (ICIP) 2011*.

[9] X. Chen and A. L. Yuille. Detecting and reading text in natural scenes. *CVPR 2004*.

[10] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 2002*.

[11] C. Desai et al. Discriminative models for multi-class object layout. *Computer Vision–ICCV 2009*.

[12] S. K. Divvala et al. An empirical study of context in object detection. *CVPR*, 2009.

[13] B. Epshtein et al. Detecting text in natural scenes with stroke width transform. *CVPR 2010*.

[14] P. Felzenszwalb et al. A discriminatively trained, multiscale, deformable part model. *IEEE CVPR, 2008*.

[15] S. Goferman et al. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence 2012*.

[16] V. Hedau et al. Thinking inside the box: Using appearance models and context based on room geometry. *Computer Vision–ECCV 2010*.

[17] L. Itti et al. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence 1998*.

[18] T. Judd et al. Learning to predict where humans look. *IEEE International Conference on Computer Vision 2009*.

[19] S. Karthikeyan et al. Learning bottom-up text attention maps for text detection using stroke width transform. *ICIP, IEEE*, 2013.

[20] S. Karthikeyan et al. Learning top-down scene context for visual attention modeling in natural images. *ICIP, IEEE*, 2013.

[21] S. M. Lucas et al. Icdar 2003 robust reading competitions. *ICDAR*.

[22] A. Mishra et al. Active segmentation with fixation. *IEEE International Conference on Computer Vision, 2009*.

[23] O. M. Parkhi et al. Cats and dogs. *Computer Vision and Pattern Recognition (CVPR) 2012*.

[24] O. M. Parkhi et al. The truth about cats and dogs. *IEEE International Conference on Computer Vision (ICCV), 2011*.

[25] C. Rother et al. Grabcut: Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG) 2004*.

[26] P. Shivakumara et al. A laplacian approach to multi-oriented text detection in video. *IEEE Transactions on Pattern Analysis and Machine Intelligence 2011*.

[27] R. Subramanian et al. Can computers learn from humans to see better?: inferring scene semantics from viewers' eye movements. *ACM International Conference on Multimedia 2011*.

[28] A. Torralba. Contextual priming for object detection. *International Journal of Computer Vision 2003*.

[29] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. *CVPR 2001*.

[30] K. Wang and S. Belongie. Word spotting in the wild. *Computer Vision–ECCV 2010*.

[31] Y. Zhong et al. Automatic caption localization in compressed video. *IEEE Transactions on Pattern Analysis and Machine Intelligence 2000*.