Indian Institute of Technology Kanpur

Top Down Attentional Guidance During Visual Search

Ankit Awasthi (Y8084),Keerti Choudhary(Y8244)

SE 367 : Introduction to Cognitive Sciences

Supervisor:
Dr. Amitabha Mukerjee
Department of Computer Science Engineering
IIT Kanpur

ABSTRACT

While seraching for an object in a scene humans make use various sources of information for narrowing down their search to certain areas which are more probable to have that object. Various sources of "top down" influences have been investigated in the literature including contexual guidance and appearance based influence to attentional guidance. In our work we use both the top down cues to predict eye fixations of humans beings during visual search. We also propose a confidence measure in the context model and weigh down the contexual influence accordingly. A simple object detector with boosting is used to obtain scores for images object based information. The results show that contextual guidance is better at predicting the eye fixations than the object based information and the combination of the influences shows a marginal increase in accuracy.

Contents

1 Introduction

Attention broadly means selecting stimuli which are relevant among others which are irrelevant. Here,attention refers to visual attention i.e, selecting relevant stimuli from the available visual data. In the literature there are two contrasting views on models of attentional selection with a continuos spectrum of models which combine both of these The view that has dominated the research is Visual Saliency Hypothesis.According to it bottom up stimulus based information drives our attention. The bottom up influences include sharp contrast points in terms of various features like intensity, color, orientation. It does not take into account the pre-acquired knowledge of objects and is task and goal independent. One of the most famous models of bottom-up saliency is by Koch & Itti [7].

The other paradigm is of Cognitive Relevance Hypothesis.This model incorporates top down influences on attentional guidance. Top down influences include task based influences. For example, if asked to search for an object say a chair, we are more likely to fixate on objects which look similar to a chair. It supports object based attentional selection i.e, we pay attention to objects rather than low level features such as contrast, color and orientation. Top Down attention mechanism is guided by knowledge of the visual appearance of objects; and how and where objects appear in a given scene. Scene context guides our attention to regions having high probability of containing target objects.

In our work we have presented a model for attentional guidance during a quite constrained visual task of object search. Torralba et al,2006 [2] and Kanan et al,2009[3] have independently looked at Contexual guidance and target object based information respectively as possible sources of influence on attentional selection during visual search. The results in both cases achieve nearly the same accuracy.

As indicated by some of the recent experiments [4],the bottom-up influence on attentional selection may not be direct but only through its correlation with the target objects. Moreover,the two top down factors - contexual guidance and target feature based information, have been suggested in [5] to be independent and additve sources of information for attentional guidance during visual search.

1.1 Motivation

Objects we pay attention can give an important insight into how perceptual data might be represented in our mind. It has been found that attention is affected by language processing and vice-versa . If someone says ,T̈here's a snake in this room; all people would start attending to snake like objects (cables etc.). Thus unraveling the mystery of attentional selection would go a long way in making clear the nature of visual representation and language vision interaction in humans. During visual search Top-down information plays a guiding role in

our attention. Therefore, in this project we focus only on Top Down attentioanl mecanisms, namely, template based and context based object search mechanisms. We combine the two models and investigate to what extent top down mecanism explains our eye movement during the visual search.

2 Contexual Guidance

The context guidance used in our model was given by Torralba [2] and the main motivation behind this model is that the gist of the scene is acquired in the first few seconds of view.The details of the context model are as follows. For each image steerable pyramidal filters are applied at four different scales and six different orientations. The resulting 24 images are downsampled into 4x4 blocks and filtered response is averaged over the block. As a result, for a given image we have a global feature vector of size 384 which we further reduce to 100 dimensions using Principal Component Analysis.

The context model is trained on the Label Me dataset and we consider a pairs of global feature vector of an image and corresponding target object locations. A Gaussian Mixture Model is fitted to the joint probability of global feature vectors of images and correspondinng object locations. The mixture of gaussians allow us to look at the model as a set of scene prototypes. For each scene prototype we have a corresponding distribution of target location.

Thus given a training image the context model outputs the probabilistic distribution of target object location. This distribution corresponds to the Contextual Guidance Map.The figure 1 shows the result of the contexual guidance

3 Object Appearance Based Information

The target features come up as an obvious influence while considering top-down factors. Although,a lot psychological experiments have been done to investigate the influence of target features not many computaional models look at target features as a possible influence on attentional guidance. A simple object detector with boosting [6] is used to obtain target feature based information. First a dictionary of target features is made using 10 images/instances of the target object. Each image of is filtered using 3 filters and the cropped object region (available in Label Me dataset) is divided into 3x3 blocks. The filtered response over these blocks is averaged and we have a dictionary of 270 (10x3x9) features. Using this dictionary of features the object detector is trained for all the training instances. For each positive sample 20 randomly cropped images not overlaaping with the target object are taken as

Figure 1: Context Map for the given image

negative examples. A number of weak classifiers are learned using these features and these classifiers are cascaded at different scales as in boosting. The final score obtained across multiple scales is used as the target feature based saliency map.

4 Combining Top Down Cues

Inspired by some of the recent experiments on eye movements during visual search[5], we combined the two sources of top-down infulences - contextual guidance and target feature based information. According to [5],these sources are used independently and additively for facilitating attentional guidance during visual search. We independently compute the two influences as explained above and then use a combination of these features for modelling attentional guidance during visual search. Once we have the Contexual guidance map and target feature based saliency map, we combine these models in the following manner.

M_c : contextual guidance map
M_t : target feature based map
M : combined map
$M = M_c^k * M_t^{(1-k)}$

Here the value of k may be chosen empirically or using the confidence measure as described in the next section. The results reported have used k=0.15(empirically) and k derived using the confidence measure.

5 Confidence in Context

We propose a confidence measure for the context model as the context model relies on the training it has undergone. For modelling the confidence in the training images we fit a Gaussian Mixture Model to the global feature vectors of the training images. Depending upon whether the global feature vector of the image lies within a certain threshold from peaks. We use the following scheme for weighing the contexual guidance model.

k = 0.25 if $d < \sigma$ for any gaussian in the mixture of gaussians
k= 0.15 if $d < 2\sigma$ for any gaussian in the mixture of gaussians
k= 0.1 otherwise

Here, σ refers to the variance of a gaussian and d refers to the distance of the point from the mean of that gaussian.

6 Eye Movement Data

The eye movement data used for testing the model is openly available [2]. The eyedata wasa taken for people, painting and mug search in the scenes. Three groups of eight participants each was taken for different search tasks. The participants were asked to search for target object in the given scenes. The first 5-6 fixations were taken into cnsideration, as after that fixations tend to become quite random. Eye movement analysis had shown that fixations weere guided by context as well as search task. To find how consistent our fixations across participants, fixations of 7 participants was studied to predict the fixation of eighth participant. Participants showed around 80-90 percent consistency.

7 Results

Once we have the map from the combined model. We find the top 20% salient region in the map and the accuracy signifies the number of eye fixations that lie in that region. The results obtained in Torralba,2006 [2] are shown in the figure 2. The results show that for pedestrains search the results are quite good but the results not good for later fixations as the fixations are more and more scattered and random as we consider later fixations. Figure 3 shows our

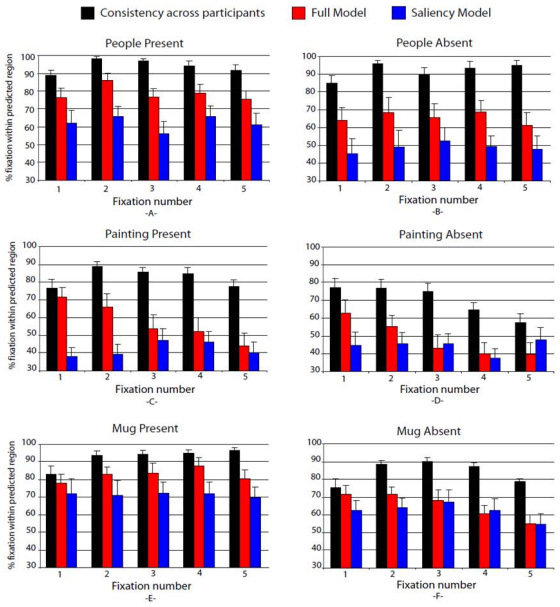

Figure 2: Results as reported in Torralba[2]

results for the person search for the first five fixations for the context model,target fetaure based model and our combined model. The figures 4 and 5 shows the results for mug search and painting search. The results here are with k=0.15. The results using the confidence measure are very similar and are not reported here but if we have images context different to the context presnt in the training images we expect to obtain better results with the confidence measure.

8 Discussion and Future Work

The debate over the possible sources of influences to our attentional guidance has been going on since the earliest works in the field. The biological plausibile bottom-up visual saliency dominated for a long time in the cognitive science community. Some of the recent experiments [4] have shown that it is possible that bottom-up salinecy may be influencing our eye movements in an indirect manner, i.e, through its correlation with objects present in the scene. Compuational studies on modelling attention have largely focused on bottom up information mostly because of the ease of incorporating bottom up information as compared to top down information which includes Contexual information and Object Appearance based infromation. Over the past few years there have some attempts to model attentional guidance using some Top-down features(Torralba[2],Kanan [3]) . While eye movements during tasks like free viewing can better be explained using bottom up information,some other tasks like search for a target object certainly involves top-down influences. Another issue in Compuational models of visual attention is to consider consistency among humans in terms of the eye movement data. Eye movements while free viewing and memory tests vary a lot among

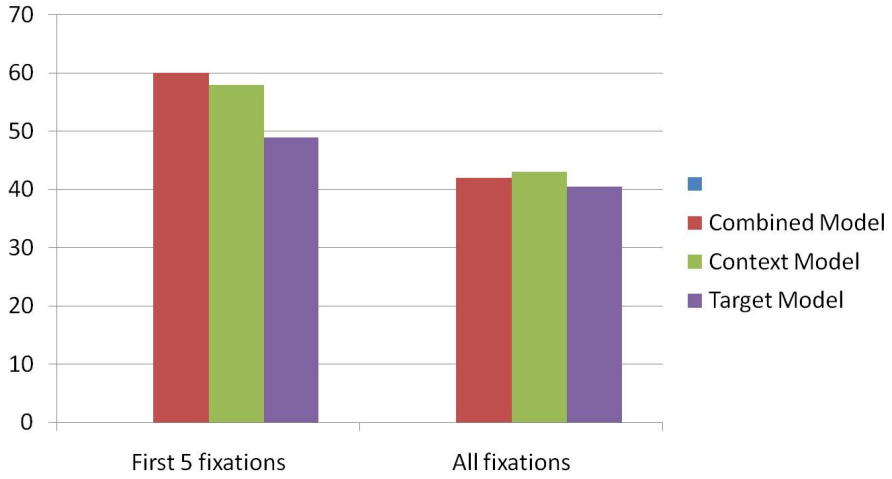

Figure 3: Accuracy of Combined model, Context model and Target feature based information for people search

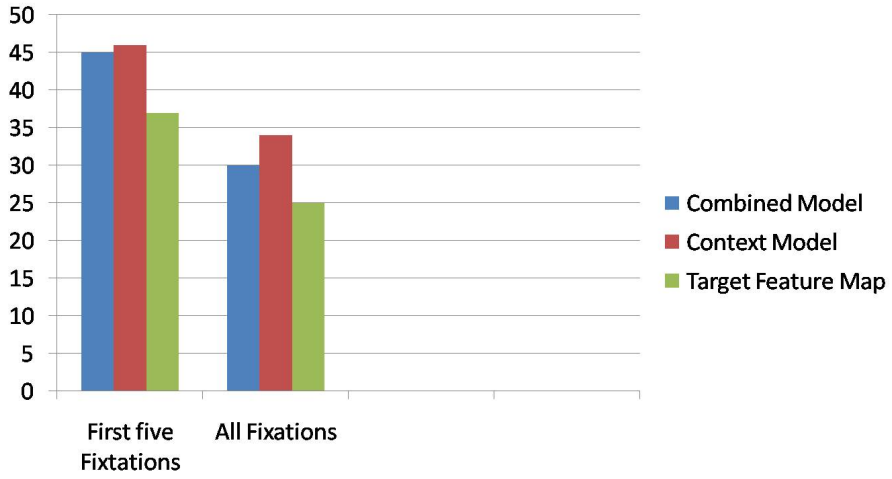

Figure 4: Accuracy of Combined model, Context model and Target feature based information for painting search

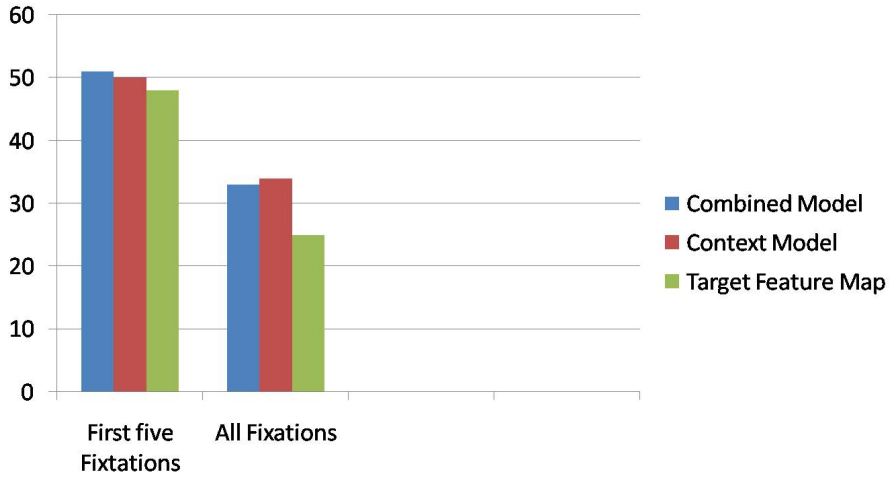

Figure 5: Accuracy of Combined model, Context model and Target feature based information for mug search

humans . This constrains the accuracy of models for such tasks. To avoid this inconsistency it is better to consider constrained tasks such as search for a target object which indeed shows high consistency among humans as reported in [2]. It is quite clear that top-down influences play a role in attentional guidance during visual search, but the question is what is the extent of this influence. It is even believed that attentional selection is largely object based or top down and whatever influence bottom information has is through its correlation with the target objects [4] As found out in our results contextual guidance model alone gives good accuracy. Moreover, the combined model does not show much improvement. Considering the moderate accuracy of state of the art models, future work should include a better model for context which may consider object associations among other things. In general, better and more general ways of incorporating top-down influences would help in coming up with a generalised model for the attentional guidance in humans. Our future work will investigate how the transition from bottom up attention in infants to top down attention in adults takes place. We hope such studies would explain better the role of individual visual experiences in a better and a more genralised model could then made.

9 Acknowledgement

We are thankful to Antonio Torralba for making openly available the code of 'Context model of attentional guidance' and 'Object Detector' , and also making available the eyedata for visual search and 'Label Me' dataset.

References

[1] B. C. Russell, A. Torralba, K. P. Murphy, W. T. Freeman, *LabelMe: a database and web-based tool for image annotation*. International Journal of Computer Vision, pages 157-173, Volume 77, Numbers 1-3, May, 2008.

[2] A. Torralba, A. Oliva , M. S. Castellhano, J. M. Henderson, *Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search*, Psychological Review, pages 766-786, Volume 113, Number 4, October 2006.

[3] Kanan C., Tong M., Zhang L., Cottrell G. (2009). *SUN: Top-down saliency using natural statistics. Visual Cognition*, 17, 9791003.

[4] John M Henderson & Antje Nuthman(2010), *Object-based attentional selection in scene viewing*, Journal of Vision(2010), 10(8):20, 1-19.

[5] Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real world scene search. Journal of Vision, 10(2):4, 111, doi:10.1167/10.2.4.

[6] A. Torralba, K. P. Murphy and W. T. Freeman. (2004). *Sharing features: efficient boosting procedures for multiclass object detection*. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR). pg 762-769.

[7] Koch & Itti (2001), *Computational modelling of visual attention*Nature Reiviews and Neuroscience,2001, Vol 2; Part 3, pages 194-204