A computational model for top down attention

Abhijit Sharang

Abstract

Modelling top down visual attention is a challenging task owing to the huge dimensionality of the input involved.Moreover,since it is task specific,the variables to be taken into consideration might differ with every task.This project aims at developing a general model for top down attention and testing the effectiveness of the model on two different kinds of tasks.Results show that the model works well when the person is dynamic with respect to the environment,but does not perform upto the expectations when the person is static.

1 Introduction

Attention involves selectively processing the visual stimuli resulting in the dimensionality reduction of the input,aiding in fast and real time processing of the same in subsequent stages of vision.It is widely accepted that there are two aspects of attention:*bottom up*,which is data driven,and *top down*,which is task driven.It is a combination of these that is responsible for the eye gaze in everyday vision.However,when the vision is goal oriented,there is a domination of the top down aspect.

It is easier to model the bottom up aspect of visual attention,since it only involves the manipulation of the scene information while ignoring the context.However,the success of these models in predicting the eye gaze fixations is limited to free viewing.Moreover,since the context is ignored,the bottom up saliency maps might emphasise the areas of the scene which are not relevant to the task at all.

Top down saliency can be modelled in a Bayesian formulation using features in the scene(both bottom up and task driven) which can influence the gaze pattern.[Torralba et al., 2006] employed a discriminative model to maxmimise the probability P(O=1,X|L,G) ,where O=1 indicates the object presence,X is the location of the gaze in the image,and L and G denote the local and global features respectively.[Navalpakkam and Itti, 2005] proposed to maximise the signal to noise ratio in the image for object detection.In these approaches,visual search is an important component.

Many of the existing models for top down attention are task specific,where the global features are used for understanding the context.Peters and Itti[Peters and Itti, 2007] developed a spatial model by mapping the global features to eye fixation in navigation and exploration tasks.[Coen-Cagli et al., 2009] used a discriminative model for sensory-motor coordination in drawing tasks.[Ballard et al., 1995] developed a dynamic Bayesian network for sandwich making task.

To construct a general model for attention,[Borji et al., 2012] propose that apart from the global scene information,we also need to take into consideration the objects relevant to goal completion.This information,combined with the eye gaze information can be used to generate a Bayesian model that exploits the sequentiality of the task driven aspect of attention.While their Bayesian framework is restricted to the objects,we hypothesise that the events occuring in the scene are equally important in driving the gaze.Thus,our generative model also takes the event information into consideration.

The rest of the report is divided into four sections.In section 2,the data for developing the models is elaborated.In section 3,the models developed for comparison are discussed.In section 4,the experimental details and results obtained.Finally,in section 5,the relevant discussion and conclusion is put.

2 Dataset

The data for the generation of the models was made available by [Borji et al., 2012].It consisted of the eye gaze density map for five subjects playing video games,and the tagged event,object and feature information of the videos.Each video was downscaled to 15 X 20 for reducing the complexity involved in model generation.The data for two video games,*Hot Dog Ambush*(HDB) and *3D Driving School*(3DDS) were taken into consideration.In the former game,the players had to serve food and drinks to the customers on demand and in the latter,they had to drive a car around the city,following the appropriate instructions.Hence,in the former case the players were static with respect to the environment while in the latter case they were dynamic with respect to the environment.

3 Methodology used

The data obtained was iteratively divided into test data,consisting of the data from one of the participants,and training data,consisting of the data from the others.The partition which gave the best model in each of the models described next was considered for result analysis in the subsequent stages.For the purpose of comparison,five different models were generated,out of which four were control models and the fifth was the generative model.Their details are described below.

3.1 Control models

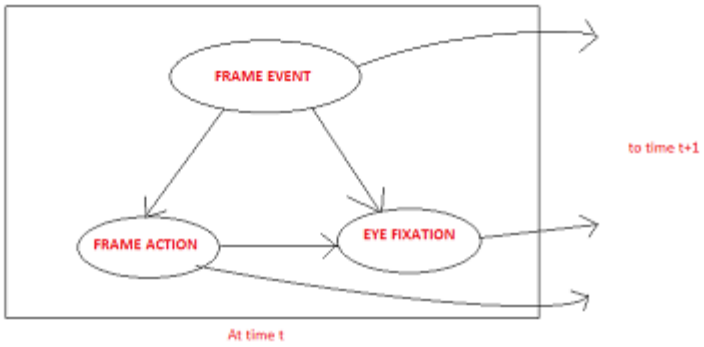
- **Random eye position** This is the simplest model.Here,the training data is ignored and a random position is chosen to be the fixation point on the eye gaze map
- **Mean Eye position** This model utilises the eye gaze map of all the frames from the test data and averages them to obtain a mean eye gaze map.Again,the feature,event and object information is ignored.
- **Regression model** We assume that there exists a linear relationship between the features and eye gaze map.We model this with the equation M X $W = E$ where,
  F = matrix consisting of the feature vectors of the present frame(F_t) and the eye gaze of the previous frame(E_{t-1})
  E = eye gaze map of the current frame.
  The solution for the same is given by M' X E,where M' is the pseudo inverse matrix of M obtained by singular value decomposition.Having learned this matrix,the eye gaze map for the test data can be generated.
- **k Nearest neighbours** In this model,the gaze density map of each frame is estimated from the gaze density map of its k most similar frames obtained through Euclidean distancebetween the features in the training data.The value of k was fixed at 40.The weighted average of the gaze density map of these frames is taken to be the gaze density map of these frames.Hence,

  $$E_j = \frac{\sum_{i=1}^{k} (D(F_i,F_j))^{-1} E_i}{k},$$

  where $D(F_i, F_j)$ is the feature distance between the test frame and i_{th} training frame.

3.2 Generative Model

The idea behind the generative model is that top down attention tends to be sequential in nature.Hence,most of the eye fixations occur in continuity.This fact can be exploited to formulate a Bayesian Network,where the current gaze map depends on the features present in the current frame and the eye gaze map of the previous frames.We extend the Bayesian Network to the temporal domain through the Dynamic Bayesian Network(DBN) where there is an intra frame dependency between eye gaze map and the features,and an inter frame dependency between the eye gaze map of the consecutive frames.The network for the game 3DDS is shown in the following page.The network for HDB is similar,except that instead of using the frame events and actions,we use the objects present in the frame and the attended object in the frame to conditionally determine the eye gaze map of the current frame.



DBN for 3DDS

Let $F_{1:t}$ denote the events occuring in the frames from 1 to t,$A_{1:t}$ denote the actions occuring in the frames for 3DDS and objects attended to in HDB from 1 to t and $E_{1:t}$ denote the eye gaze map of frames from 1 to t.Through the network,the following conditional dependencies are derived:

1. $F_t \rightarrow A_t$
2. $F_t \rightarrow E_t$
3. $A_t \rightarrow E_t$
4. $F_{t-1} \rightarrow F_{t-1}$
5. $A_{t-1} \rightarrow A_t$
6. $E_{t-1} \rightarrow E_t$

Since the structure of the network is already estimated,we now need to adjust the state transition matrix $(P(S_t^i|parent(S_t^i))$ and the observation matrix $(P(E_t^i|parent(E_t^i))$ for maximising $P(E|V)$,where $V = (m; \theta)$ represents the model parameters.This is done using the *junction tree algorithm*[Cowell et al., 2007].Having learned the parameters,the evidence is entered from the test data to obtain the predicted eye gaze map.

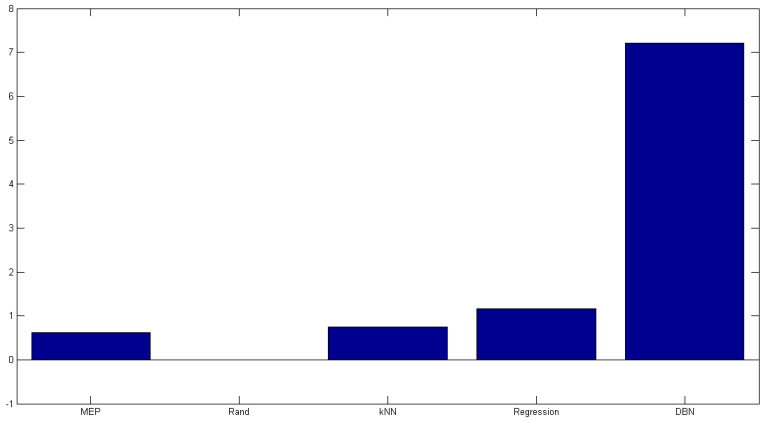
4 Experimentation and results

In each of the models developed,the training was done for two cases.In the first case,the training set included all the frames present in the video for each subject.The model thus generated predicted general fixations.In the second case,the training was restricted to the frames where *saccades* occured and two frames in the neighbourhood of these frames.Hence,here the fixations were predicted when saccades occured.The accuracy of the saccades thus depended on the accuracy of the fixations.The code for the models was largely based on the code made available by [Borji et al., 2012] for HDB game,with suitable modifications made for 3DDS game.The results are described below.

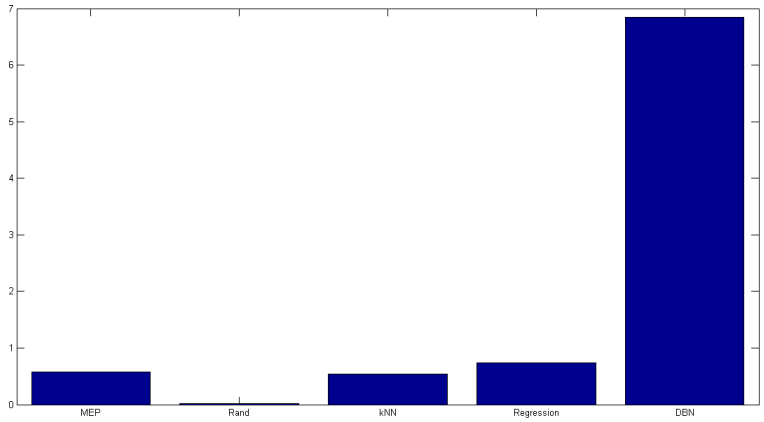
4.1 Normalised scan path saliency(NSS)

NSS[Peters and Itti, 2007]is the response value at the actual human eye fixation in the gaze density map that the model has predicted,normalised to zero mean and unit standard deviation.A higher value implies greater ability of the model to predict correct eye fixations.A value around zero implies that the model only picks some arbitrary position on the map.

4.1.1 Scores for 3DDS

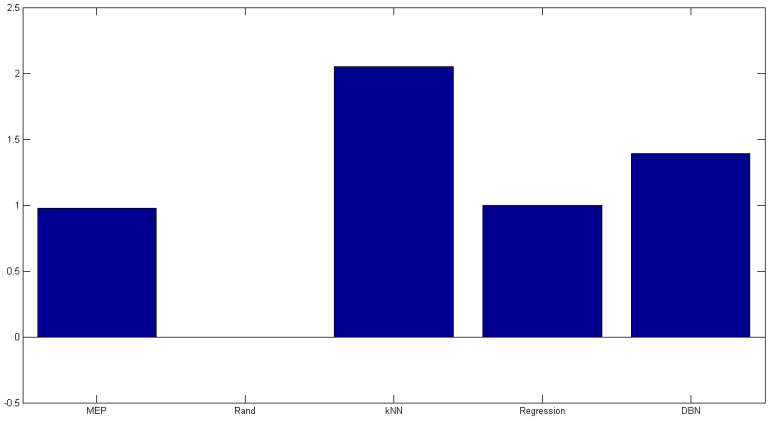


Fixation

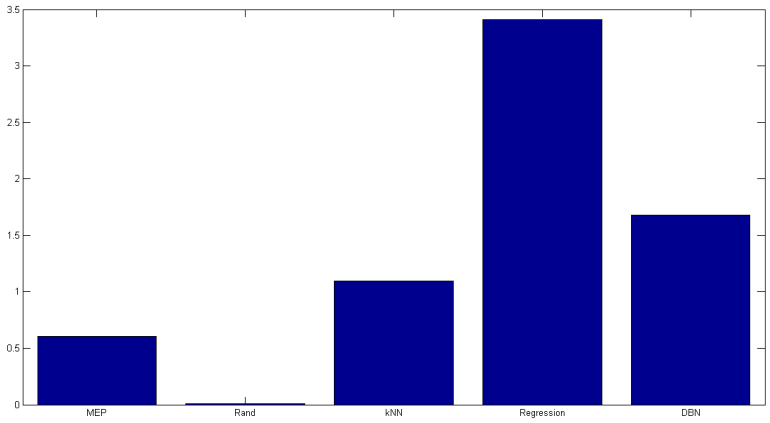


Saccades

4.1.2 Scores for HDB



Fixation

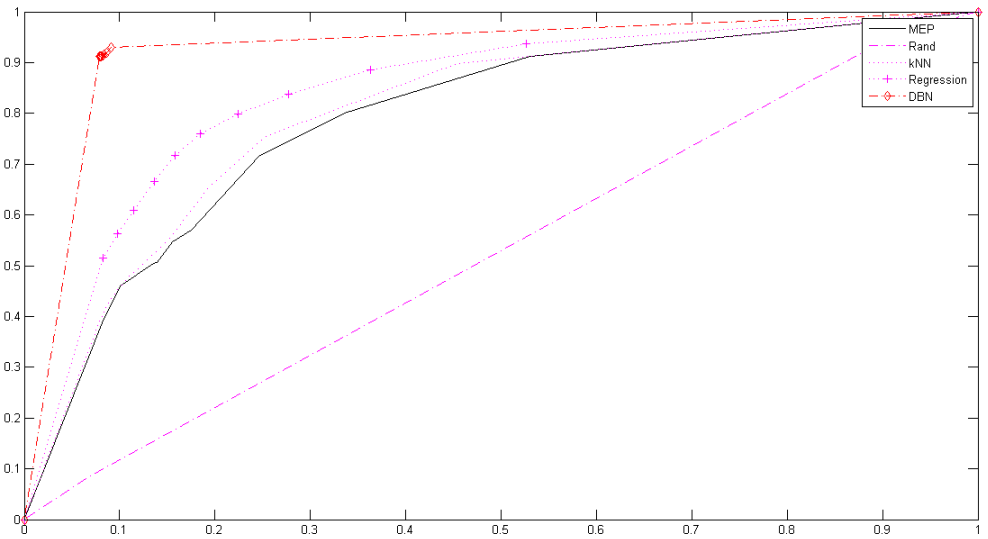


Saccades

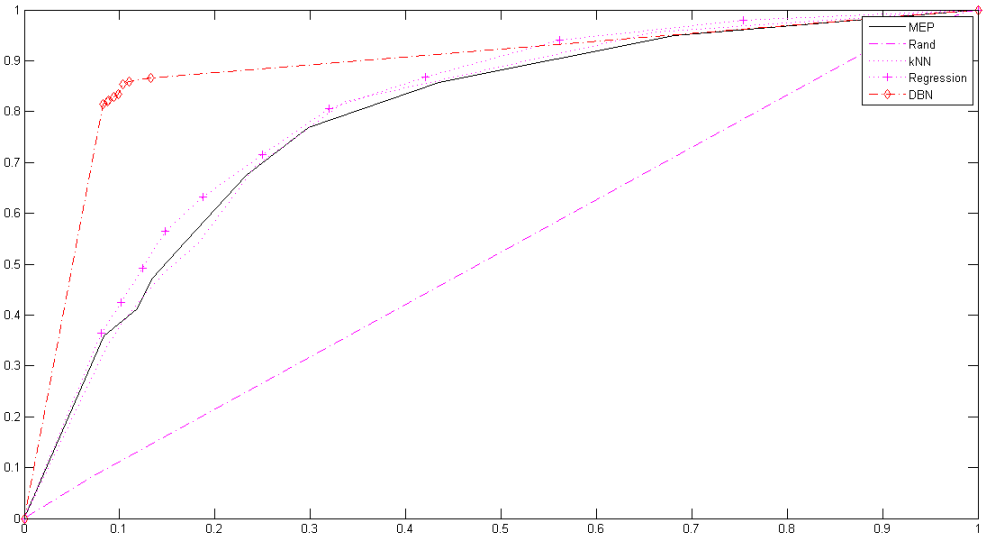
4.2 Receiver Operating characteristic curve(ROC)

For plotting ROC,the gaze density map computed by the model serves as a binary classifier for fixation occuring at a pixel.The actual gaze density map serves as the ground truth.The threshold is varied for obtaining the true positive rate from the model and the false positive rate generated from a random gaze density map.

4.2.1 Scores for 3DDS

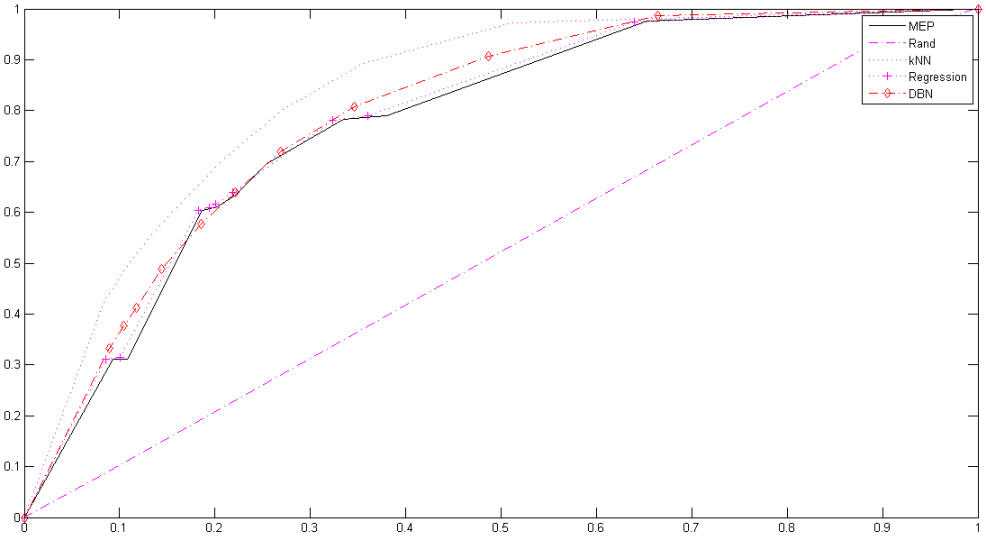


Fixation

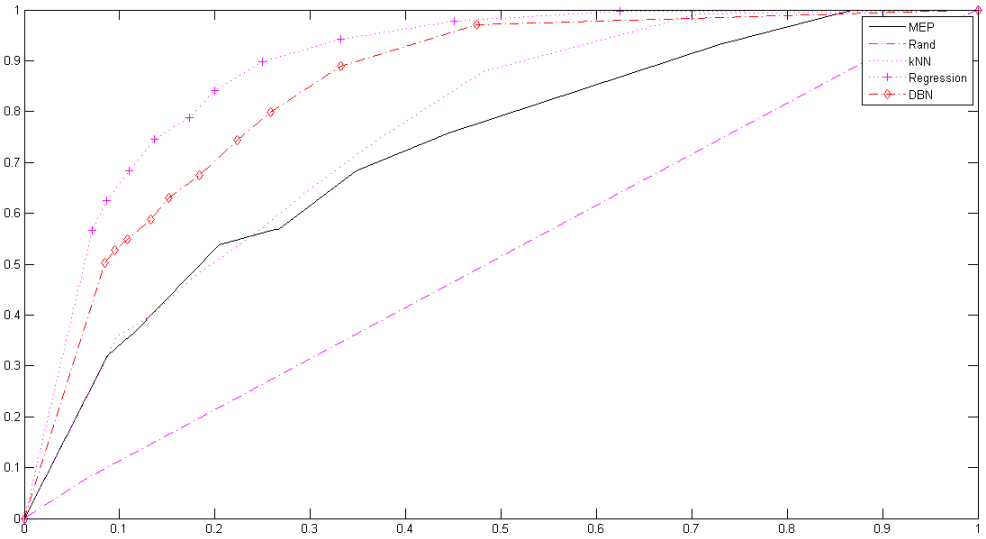


Saccades

4.2.2 Scores for HDB



Fixation



Saccades

5 Discussion and conclusion

In the case of the 3DDS game,the generative model outperforms the control models by a large margin in the case of general predictions as well as in the case of saccades.Whereas,in the case of HDB,the results for the generative model are good,but it is outperfomed by some other control model.

This result tends to bring out the dissimilarity in the fixation patterns between the case where the person is dyanmic with respect to the environment and the case when the person is static with respect to the enviroment.In the former case,much of the fixations are sequential in nature,leading to less saccades and more smooth movements across the frames.In the latter case,when some task has to be performed in a stipulated amount of time,the saccades number more.This affects the sequentiality paradigm of the hypothesis and though the generative model is able to predict the fixations and saccades with reasonable accuracy,it might not be generalised to such cases.

Hence,the Dynamic Bayesian Network can be easily incorporated in cases where the eye movement is generally smooth.The generalisation to the whole domain of top down attention seems to require the incorporation of the dependence of the gaze density map on more features.An important issue to handle in such a model would be the discretisation of such features for fast learning of the model parameters,which would again tend to make the model domain specific.

References

[Ballard et al., 1995] Ballard, D. H., Hayhoe, M. M., and Pelz, J. B. (1995). Memory representations in natural tasks. *Journal of Cognitive Neuroscience*, 7(1):66–80.

[Borji et al., 2012] Borji, A., Sihite, D. N., and Itti, L. (2012). An object-based bayesian framework for top-down visual attention. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*.

[Coen-Cagli et al., 2009] Coen-Cagli, R., Coraggio, P., Napoletano, P., Schwartz, O., Ferraro, M., and Boccignone, G. (2009). Visuomotor characterization of eye movements in a drawing task. *Vision research*, 49(8):810–818.

[Cowell et al., 2007] Cowell, R. G., Dawid, P., Lauritzen, S. L., and Spiegelhalter, D. J. (2007). *Probabilistic networks and expert systems: Exact computational methods for Bayesian networks*. Springer.

[Navalpakkam and Itti, 2005] Navalpakkam, V. and Itti, L. (2005). Modeling the influence of task on attention. *Vision research*, 45(2):205–231.

[Peters and Itti, 2007] Peters, R. J. and Itti, L. (2007). Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE.

[Torralba et al., 2006] Torralba, A., Oliva, A., Castelhano, M. S., and Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological review*, 113(4):766–786.