A Computational Framework for Driver’s Visual Attention Using a Fully Convolutional Architecture

Ashish Tawari and Byeongkeun Kang, 2017

Honda Research Institute

2017 IEEE Intelligent Vehicles Symposium
Main Keywords:

- Driver Visual Attention - Regions is a scene / image which attracts driver’s attention.

- Visual Saliency Detection - Regions is a scene / image which are salient features and attract driver’s attention.
Objectives:

- Predict the driver’s visual attentions / eye fixations in the real world driving scene.

- Present a Bayesian framework to model visual attention of the driver.

- Develop a fully convolutional deep neural network to predict a part of the visual attention model.
Background:

**Bottom Up Saliency**
- Parts or events in the visual field that stand out from background

**Top Down Attention**
- Task oriented / goal oriented attention
- Lane keeping - Front of the vehicle
- Lane changing - Sides of the vehicle

This paper
Methodology - 1
Bayesian Framework for Visual Attention

- Saliency of a pixel at location ‘z’ is modeled as:

\[ s_z = p(O = 1|F = f_z, L = l_z) \]

where,
O = presence of relevant object in scene
\( f_z \) = visual features at ‘z’
\( l_z \) = (x,y) location of point ‘z’
Applying Bayes rule and considering a few approximations, the final equation of the saliency of a pixel is reduced to

\[ s_z = \frac{1}{Z} \sum_{T_i} p(O|f_z, T_i) p(O|l_z, T_i) p(T_i) \]

- **Bottom Up Saliency**: Modelled using a fully convolutional neural network
- **Top Down Attention**: Location prior and Task probability obtained from driving data

\( T_i = \) specific task
Methodology - 2
Fully Convolutional Neural Network

- Skip connections and Deconvolution required.
- Model should not learn trivial central bias solution.

❖ Hence, FCN - Fully Convolutional Network is adopted.
FCN-8s architecture

Convolution layer

Pooling layer

Deconvolution layer
Training Details:

- FCN loss function is changed to L2 loss function to model FCN as a pixel-wise regression problem -
  \[ L = \frac{1}{2N} \sum_{n=1}^{N} ||\hat{y}_n - y_n||_2^2 \]

- Deconvolution = Transposed convolution.

- Weights of FCN-8s are initialized to a FCN-8s trained on PASCAL VOC segmentation dataset.
Methodology - 3
Location Priors

- Location priors are calculated from the yaw-rate obtained from the GPS sensor.
Analysis and Results:

- Dataset used: DR(eye)VE dataset
  - 74 sequences of 5 mins each.
  - 1920x1080 images from head mounted camera.
  - Gaze locations captured using wearable eye tracking device - to get ground truth (GT).

- Correlation Coefficient (CC) computed between estimated and GT saliency map.

<table>
<thead>
<tr>
<th>Method</th>
<th>Correlation Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline-Center</td>
<td>0.47 ± 0.24</td>
</tr>
<tr>
<td>Itti [35]</td>
<td>0.16 ± 0.10</td>
</tr>
<tr>
<td>Image Signature [36]</td>
<td>0.14 ± 0.12</td>
</tr>
<tr>
<td>GBVS [37]</td>
<td>0.20 ± 0.10</td>
</tr>
<tr>
<td>DR(EYE)VE [38]</td>
<td>0.55 ± 0.28</td>
</tr>
<tr>
<td>Proposed Approach</td>
<td>0.55 ± 0.28 (0.55 ± 0.28)</td>
</tr>
</tbody>
</table>
Analysis and Results:
Advantages of the proposed approach:

- From gaze data it is clear that the current driving task is an important factor. Hence, calculating location priors from available sensor data is important.
- As the velocity of the car increases the performance of the model increases.
- Modeling the problem as a pixel-wise regression instead of a pixel-wise classification task helps in obtaining a better saliency map.
Disadvantages of the proposed approach:

- More comprehensive on-road data is required to highlight non-trivial reactive events.
- Better quantitative metric should be used for a comprehensive comparison between predictions and GT.
- End to end training of the model is not done. Using multi-modal / multi-task neural networks would be a better approach.
Key takeaways:

- In general, any information independent of visual features can be incorporated as prior information and learned from the data. This helps the model to learn more meaningful features.

- Data imbalance needs to be taken care of before training or generating a model.
Question to answer -

Why is it important to use top-down attention/cues in addition to bottom-up saliency/cues for a visual attention model?

Thank You