Bottom-up Attention Guidance for Recurrent Image Recognition

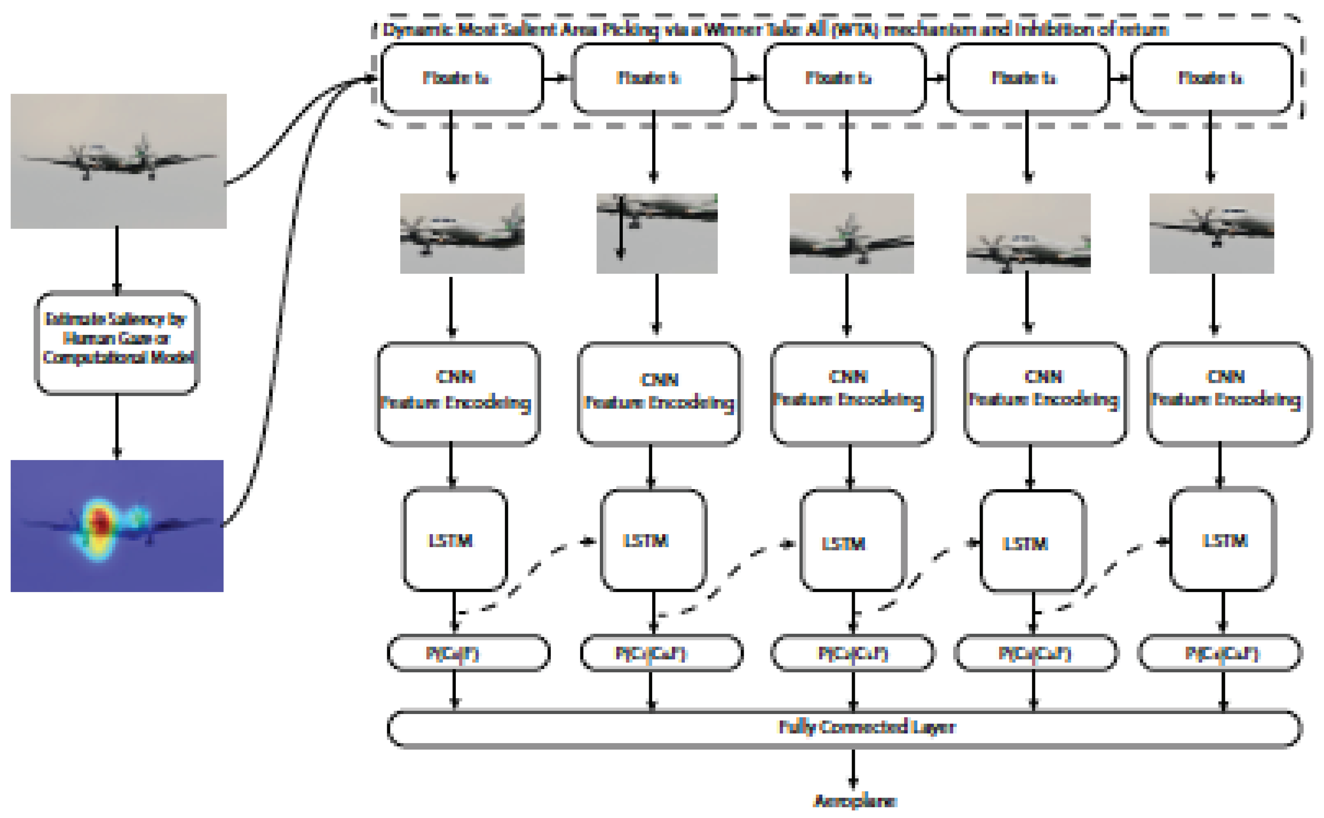
Hamed R. Tavakoli*† Ali Borji + Rao Muhammad Anwer* Esa Rahtu† Juho Kannala*

* Department of Computer Science, Aalto University

\+ Center for Research in Computer Vision, University of Central Florida

† Department of Signal Processing, Tampere University of Technology

Overview



Contributions

(1) A recurent neural architecture guided by bottom-up attention is proposed

(2) Comparing patch selection mechanism based on human gaze maps with machine predicted gaze maps

Problem:

$$\theta^\star = \operatorname*{argmax}_\theta P(C|\{F\};\theta),$$

Class Label

Set of Image Patches

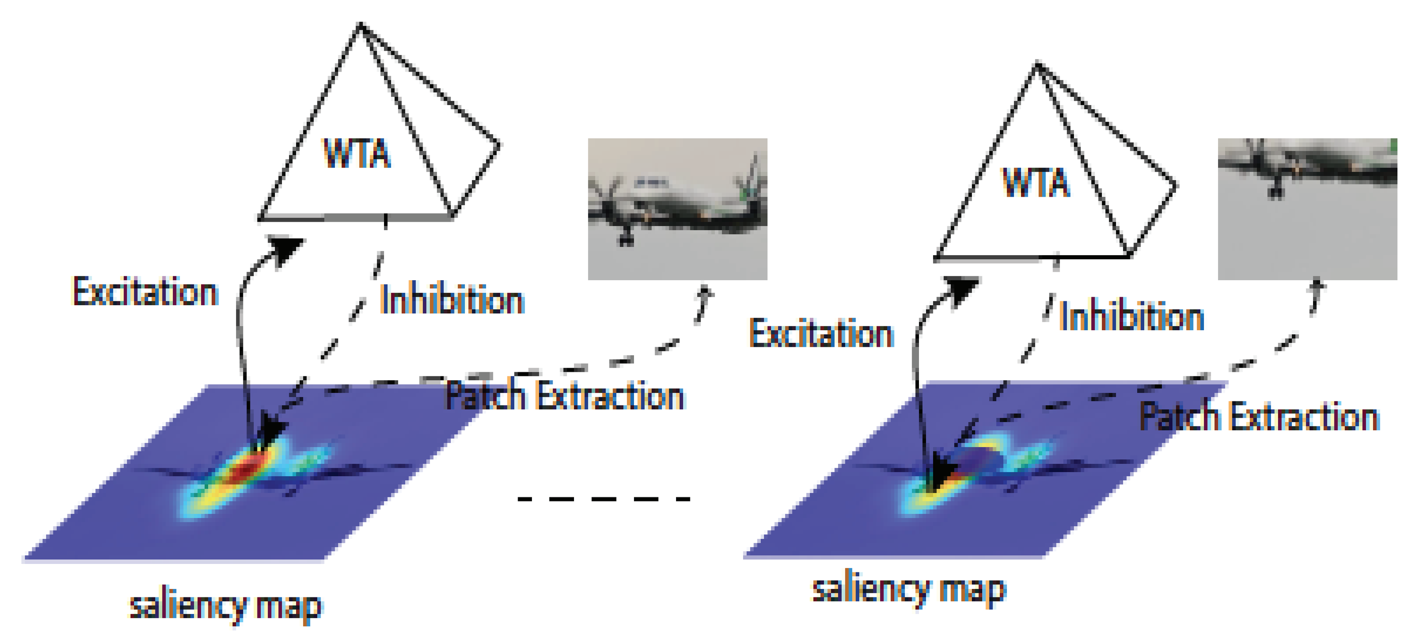
Parameters

Using a sequence of image patches, we can write,

$$P(C|\{F\};\theta) = \phi(P(C_N|C_0,\dots,C_{N-1},F_N;\theta),\dots,\\ P(C_t|C_0,\dots,C_{t-1},F_t;\theta),\dots,\\ P(C_0|F_0;\theta)),$$

Neural mapping with softmax

Patch selection



Findings

(1) The best informative patch is better than the whole image in training a feed- forward network,

(2) A recurrent model based on a sequence of informative image patches is superior to a feed-forward model and a sequence of randomly chosen image patches,

(3) Despite the gap between saliency models and human has become smaller in fixation prediction task, there is a larger gap in performance of gaze-driven maps (maps from human) and saliency models for selecting informative patch sequences in recognition task.

Results

(1) How many fixations are needed?

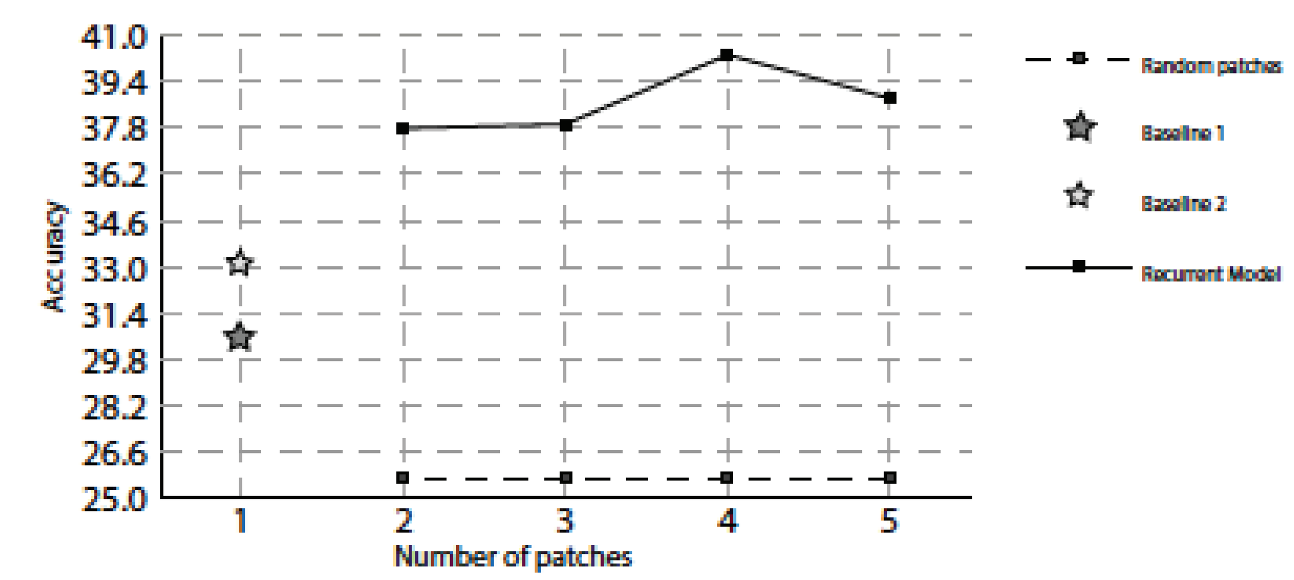


Fig. 4. The performance of recurrent recognition on human-driven image patches in comparison to two baselines on POET data. Baseline 1 is the feedforward network, trained with the whole image as input; Baseline 2 is the feedforward network trained with the first salient patch as input.

(2) Human-driven gaze maps vs machine driven maps

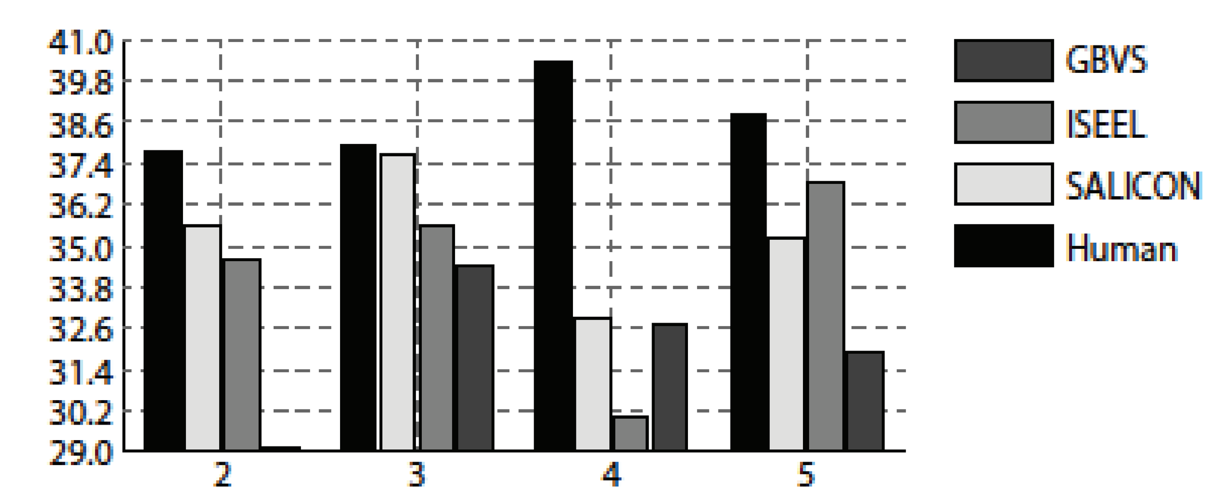


Fig. 5. The performance of recurrent recognition using computational saliency models for patch selection and human as upper-bound. The results of the recurrent approach are shown using 2, 3, 4 or 5 patches (as in Fig. 4).