Modeling search for people in 900 scenes: A combined source model of eye guidance

Krista A. Ehinger[1]*, Barbara Hidalgo-Sotelo[1]*, Antonio Torralba[2], & Aude Oliva[1]

[1]Department of Brain & Cognitive Sciences, MIT, [2]Computer Science & Artificial Intelligence Laboratory, MIT

*These authors contributed equally to the work

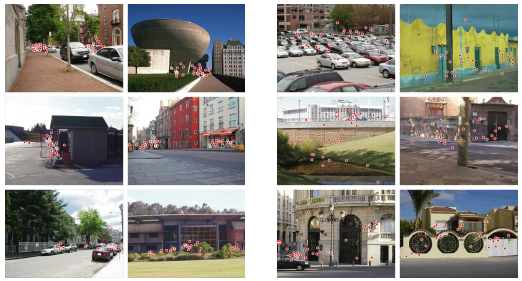
Massachusetts Institute of Technology

Modeling Search Fixations

Experiment:

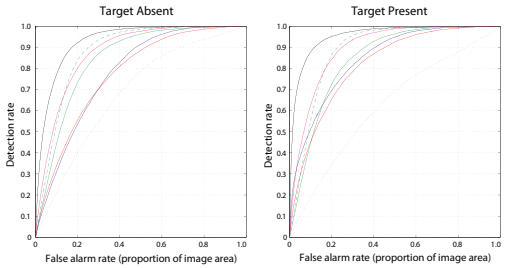
14 observers searching for pedestrians in 912 outdoor scenes (half target present)
Eyetracking using ISCAN video-based eyetracker

Examples of search fixations on target-absent scenes:



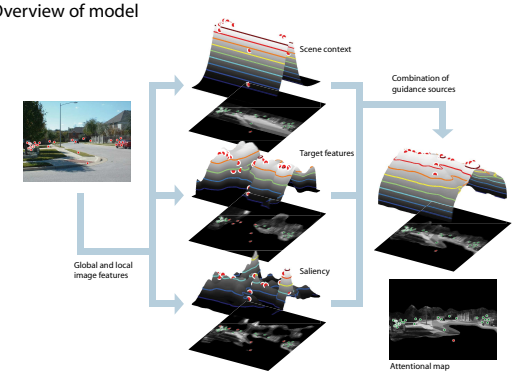
High inter-observer agreement | Low inter-observer agreement

ROC curves: Human and model performance



Human Agreement AUC	Cross-Image Control AUC
Target absent: 0.93 Target present: 0.96	Target absent: 0.68 Target present: 0.62

Overview of model



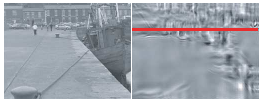
Context Model

Implementation:

Trained on 10 crops of 1880 target-present images

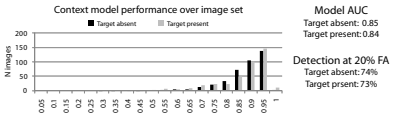
Relationship between global features and target location modeled as a mixture of Gaussians as in Torralba et al. (2006)

Context region in a novel image is computed by comparing global features to prototpyes in model



Target-present scene with its gist representation; context region is represented by red line

Best performance of context model | Worst performance of context model



Model AUC
Target absent: 0.85
Target present: 0.84

Detection at 20% FA
Target absent: 74%
Target present: 73%

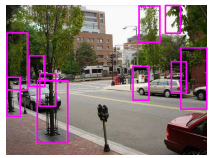
Target Features Model

Implementation:

Trained on 4000 cropped pedestrians and 60,000 random windows from pedestrian-free scenes as in Dalal & Triggs (2005)

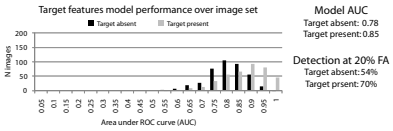
Detection based on a dense grid of histograms of gradients (HOGs)

Person detection rate was about 90% at 10% false positives per window (FPPW) on our stimuli



Output of the Dalal & Triggs (2005) pedestrian detector

Best performance of context model | Worst performance of context model



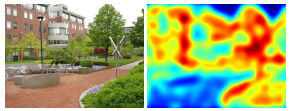
Model AUC
Target absent: 0.78
Target present: 0.85

Detection at 20% FA
Target absent: 54%
Target present: 70%

Saliency Model

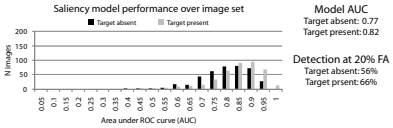
Implementation:

Compute local outliers of color, edge orientations and spatial scales as in Torralba et al. (2006)



Examples of images and their saliency map (more salient regions in red).

Best performance of context model | Worst performance of context model



Model AUC
Target absent: 0.77
Target present: 0.82

Detection at 20% FA
Target absent: 56%
Target present: 66%

Context Oracle

Manually corrects failures of scene context model

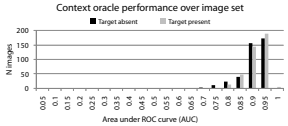
Scene context model | Context oracle

Best performance of context model | Worst performance of context model

Method: 7 observers indicated context region (marked by a horizontal line) in each scene

Context oracle created by overlaying regions selected by each observer

Illustration of context oracle task



Model AUC
Target absent: 0.88
Target present: 0.89

Detection at 20% FA
Target absent: 84%
Target present: 87%

Results:

Performance of the context oracle was not significantly different from the combined source model

Adding saliency and target features models to the context oracle gives only a tiny boost in performance (AUC = 0.89 for TA, 0.90 for TP)

Combined Source Model

Implementation:

Linear combination of component models: saliency, target features, and context

Model weights optimized on an independent validation set (100 scenes)

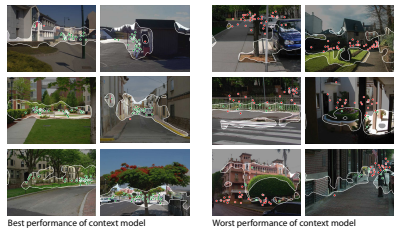
Results:

Performs at 94% of Human Agreement

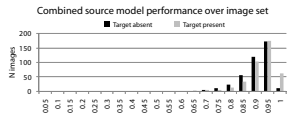
Removing Context component produces largest drop in performance

Approximates human selectivity but does not fully capture fixation clumping

AUC = 0.95 | AUC = 0.98



Best performance of context model | Worst performance of context model



Model AUC
Target absent: 0.88
Target present: 0.90

Detection at 20% FA
Target absent: 80%
Target present: 85%

Conclusions

Of the single models, the Context model is the best predictor of human fixations in this search task (but model weights may vary according to task)

Empirically-based Context oracle performs as well as the Combined source model

The combined source model, which is primarily driven by the Context model, achieves 94% of Human Agreement

Dalal, N., & Triggs, B. (2005). Histograms of Oriented Gradients for human detection. *IEEE Conference on Computer Vision and Pattern Recognition*, **2**, 886-893.

Torralba, A., Oliva, A., Castelhano, M., & 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**, 766-786.