Attention in Low Resolution: Learning Proto-Object Representations with a Deep Network

Chengyao Shen, Xun Huang, Qi Zhao

Department of Electrical and Computer Engineering, National University of Singapore

Introduction

What is Proto-Object?

Proto-objects can be seen as pre-attentive structures coherent in limited space and time.

Proto-objects can bind various low-level features over a small region of space and a short period of time, becoming “highest-level output of low-level vision’’.

Contrary to precise object recognition after the deployment of attention, the notion of proto-object is more like object-level gist that can be computed rapidly in parallel over the entire visual field (as illustrated in Fig. 1).

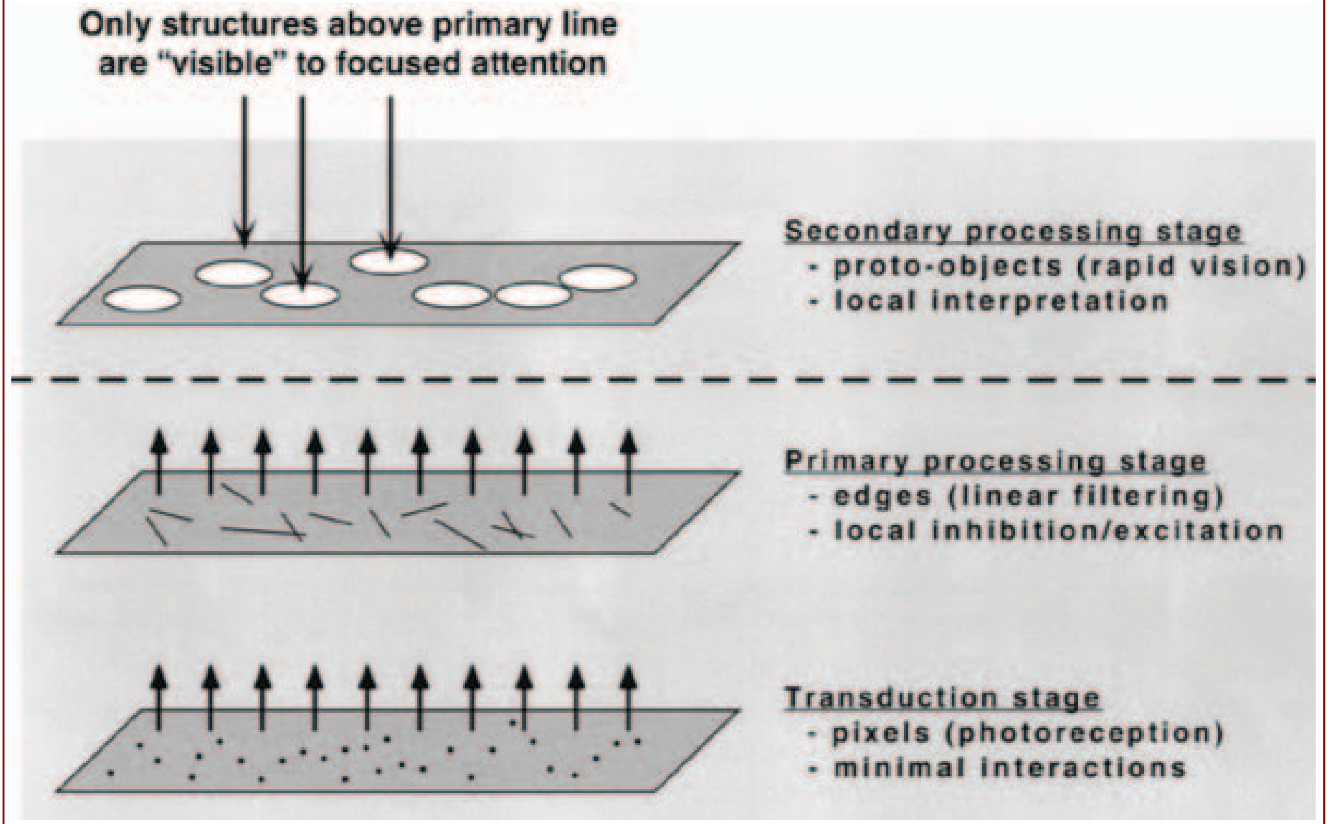
Only structures above primary line are “visible” to focused attention
Secondary processing stage
- proto-objects (rapid vision)
- local interpretation
Primary processing stage
- edges (linear filtering)
- local inhibition/excitation
Transduction stage
- pixels (photoreception)
- minimal interactions

Figure 1: Schematic of the “pre-attentive” stage described in Coherence Theory [1]. Proto-object can be seen as the “highest-level output of low-level vision’’ and can be computed in parallel.

Why Low Resolution?

Proto-objects are computed in parallel over the entire visual field where most regions are in lower resolution than the fovea area [1].

Human can perceive objects well even they are in low a resolution of 16x16 [2].

Fixations from lower resolution images can predict well fixations on corresponding higher resolution images [3].

Data Preparation

Large-scale attention data from SALICON dataset [4]:

Salient patches: multi-scale patches in low resolution sampled from top five local maxima in the blurred ground truth maps.

Non-salient patches: randomly sampled from the positions where saliency values are less than the mean of the blurred ground truth maps.

The Model

Convolutional Neural Network (CNN): Model saliency prediction as a binary classification problem on salient and non-salient patches in low resolution. Multiple scales in low resolution are concatenated and linear integrated at the final stage.

Two CNN structures are used for each single scale:

2-layer model: Input Size 16x16, C(5,64)-MP(2)-C(5,512)-MP(2)

3-layer model: Input Size 36x36, C(5,64)-MP(2)-C(5,128)-MP(2)-C(5,512)-MP(2)

where C(*f*,*n*) indicates *n* convolution kernels in size of *f* x *f*, MP(*f*) indicates non-overlap max pooling in *f* x *f*.

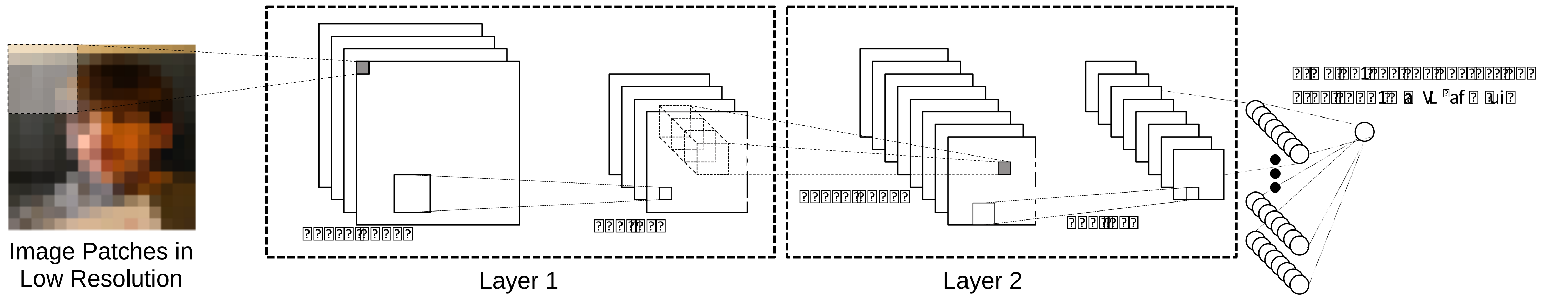

Figure 2: Network structure of the 2-layer model. For the 3-layer model, the structure is similar, with one more layer.

Results

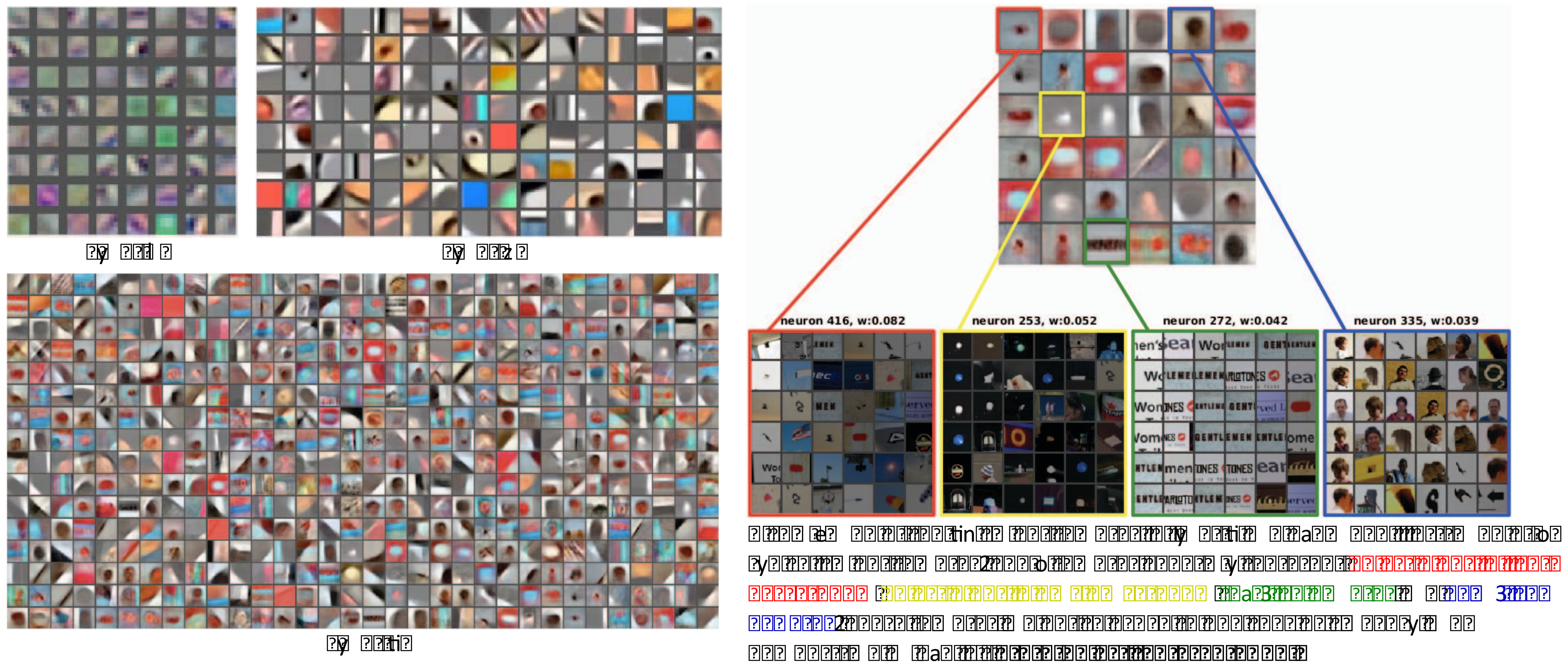

Figure 3: Visualization of features in layer 1, layer 2, layer 3 and the top salient features in layer 3.

	OSIE			MIT1003			NUSEF			FIFA		
	sAUC	CC	NSS	sAUC	CC	NSS	sAUC	CC	NSS	sAUC	CC	NSS
3-layer	**0.820**	**0.604**	**2.280**	**0.716**	0.529	**1.496**	**0.656**	0.609	**1.467**	**0.820**	**0.602**	**2.398**
2-layer	0.783	0.567	2.010	0.694	**0.533**	1.438	0.646	**0.610**	1.426	0.790	0.555	2.172
BMS	0.764	0.468	1.478	0.687	0.491	1.234	0.632	0.546	1.203	0.756	0.422	1.359
AWS	0.764	0.453	1.452	0.686	0.445	1.107	0.628	0.492	1.096	0.745	0.370	1.216
eDN	0.730	0.375	1.129	0.675	0.458	1.063	0.621	0.502	1.057	0.736	0.362	1.115
SigSal	0.732	0.423	1.319	0.666	0.465	1.085	0.614	0.495	1.094	0.747	0.402	1.268
GBVS	0.697	0.431	1.359	0.643	0.502	1.254	0.591	0.559	1.204	0.716	0.425	1.352
ITTI	0.644	0.294	0.851	0.645	0.468	1.127	0.577	0.305	0.642	0.690	0.384	1.165

Table 1: Performance of different models on MIT1003, OSIE, NUSEF and FIFA datasets. The highest scores are in bold.

Results

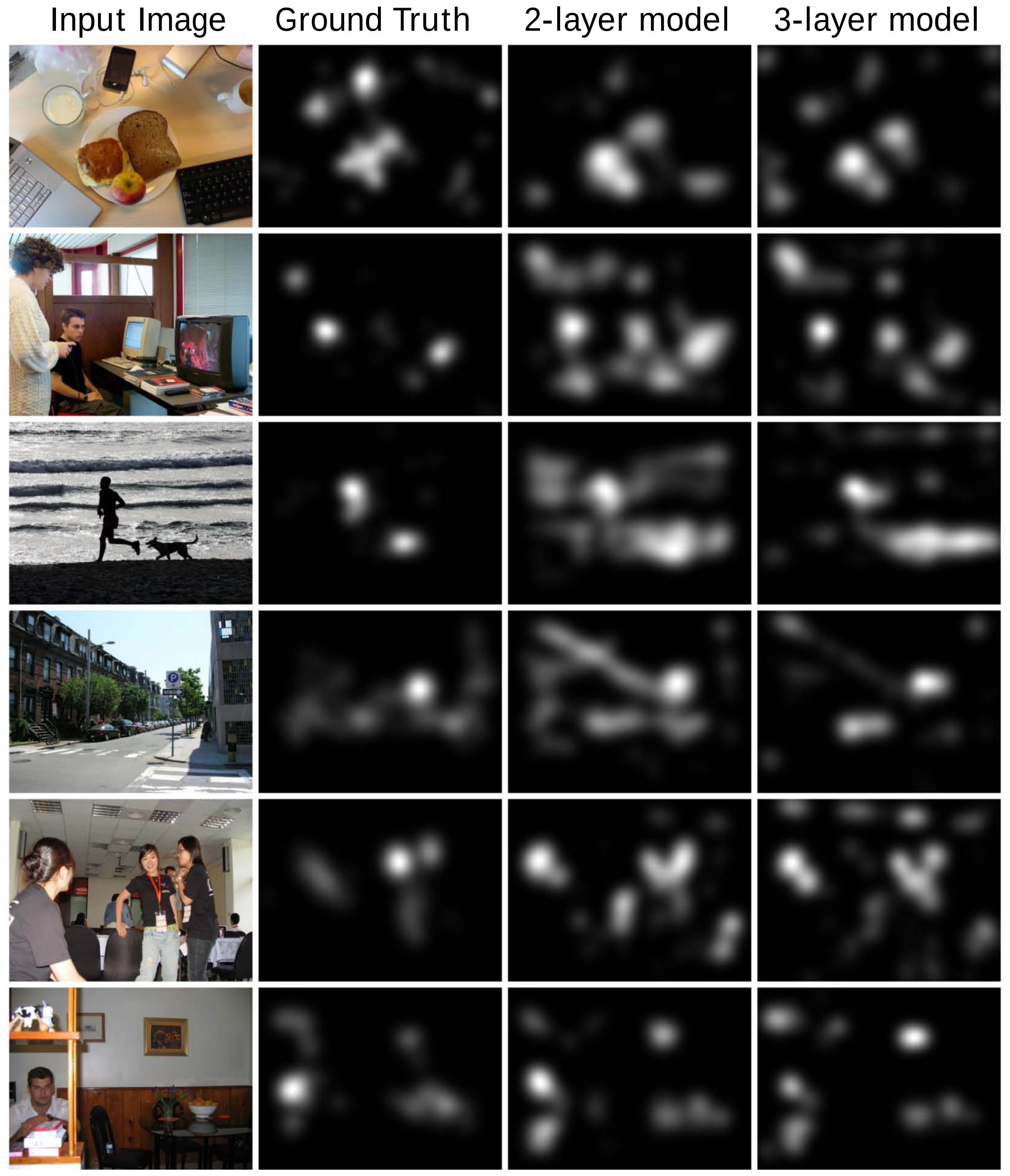

Figure 4: Qualitative comparison of our models with human ground truth. The models are in general able to detect various objects in natural scene images.

Conclusion

By training on salient and non-salient patches in low resolution, proto-object representations can be learned out in a deep architecture similar to the conceptual schematic described in [1].

The proposed models are competitive in predicting eye fixations in natural scenes compared with state-of-the-art saliency models.

This poster can be downloaded at:

http://bit.ly/1P7OJJS

Reference

[1] Rensink, R. A. (2000). The dynamic representation of scenes. *Visual cognition*,*7*(1-3), 17-42.

[2] Torralba, A. (2009). How many pixels make an image?. *Visual neuroscience*,*26*(01), 123-131.

[3] Judd, T., Durand, F., and Torralba, A. (2011). Fixations on low-resolution images.*Journal of Vision*, *11*(4), 14.

[4] Jiang, M., Huang, S., Duan, J., and Zhao, Q. (2015). SALICON: Saliency in Context. In *Proceedings of CVPR* (pp. 1072-1080).