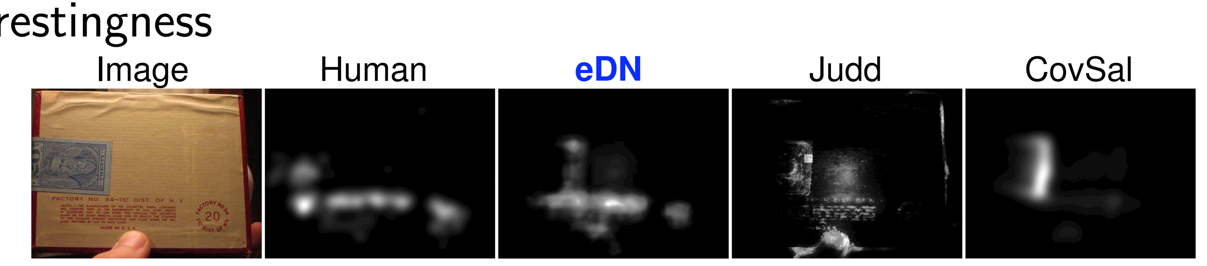
Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images

Eleonora Vig[1,2], Michael Dorr[1], and David Cox[1]

[1]Harvard University [2]Xerox Research Centre Europe

Saliency Prediction – Current Trends

- saliency map: topographic map that assigns to each scene location a measure of interestingness



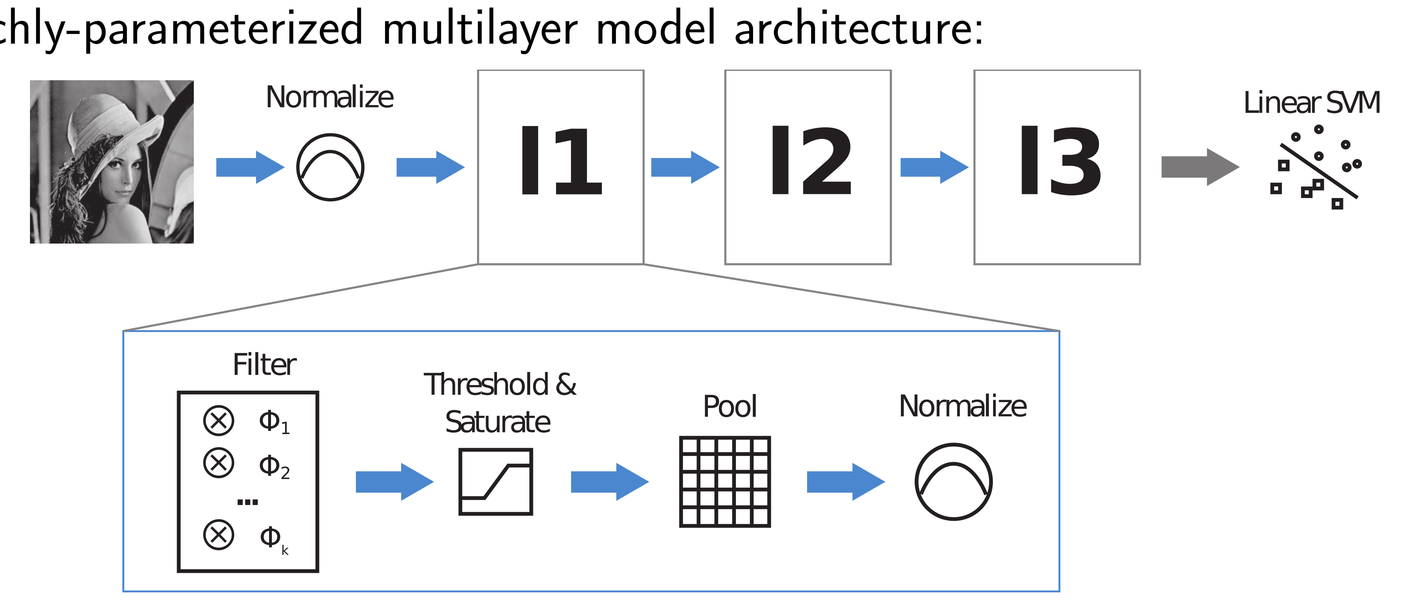
- current trends in saliency prediction:
  - incrementally add more and more hand-tuned features to existing models, e.g. face-, horizon-, text-, object detectors
  - combine many good models

Motivation – Learn Features Automatically

- our approach:
  - is entirely automatic, data-driven
  - performs a large-scale search for optimal features
  - identifies those instances of a richly-parameterized bio-inspired model family that successfully predict saliency
  - automatically derives their optimal combinations

Model Architecture

Richly-parameterized multilayer model architecture:



Standard set of operations in each layer l, $l \in \{1, 2, 3\}$:

Operations	Details	Parameters
1. **Filtering** $F^l = Filter(N^{l-1}, \Phi^l)$	$F_i^l = N^{l-1} * \Phi_i^l$ N^{l-1} normalized input of layer l $\Phi_i^l, i \in \{1, \ldots, k^l\}$ random filter	filter size # of filters k^l
2. **Activation** $A^l = Activate(F^l)$	$Activate(x) = \begin{cases} \gamma_{max}^l & \text{if } x > \gamma_{max}^l \\ \gamma_{min}^l & \text{if } x < \gamma_{min}^l \\ x & \text{otherwise} \end{cases}$	thresholds γ_{min}^l, γ_{max}^l
3. **Pooling** $P^l = Pool(A^l)$	$P^l = Downsample_\alpha \left(\sqrt[p^l]{(A^l)^{p^l} * \mathbf{1}_{a^l \times a^l}} \right)$	neighb. size $a^l \times a^l$ exponent p^l, α
4. **Normalization** $N^l = Normalize(P^l)$	$N^l = \begin{cases} \frac{C^l}{\Vert\hat{C}^l\Vert_2} & \text{if } \rho^l \Vert\hat{C}^l\Vert_2 > \tau^l \\ \rho^l C^l & \text{otherwise} \end{cases}$ $C^l = P^l - \delta^l \hat{P}^l$, $\hat{C}^l = C^l * \mathbf{1}_{b^l \times b^l \times k^l}$	stretching param. ρ^l threshold τ^l $\delta^l \in \{0, 1\}$ neighb. size $b^l \times b^l$

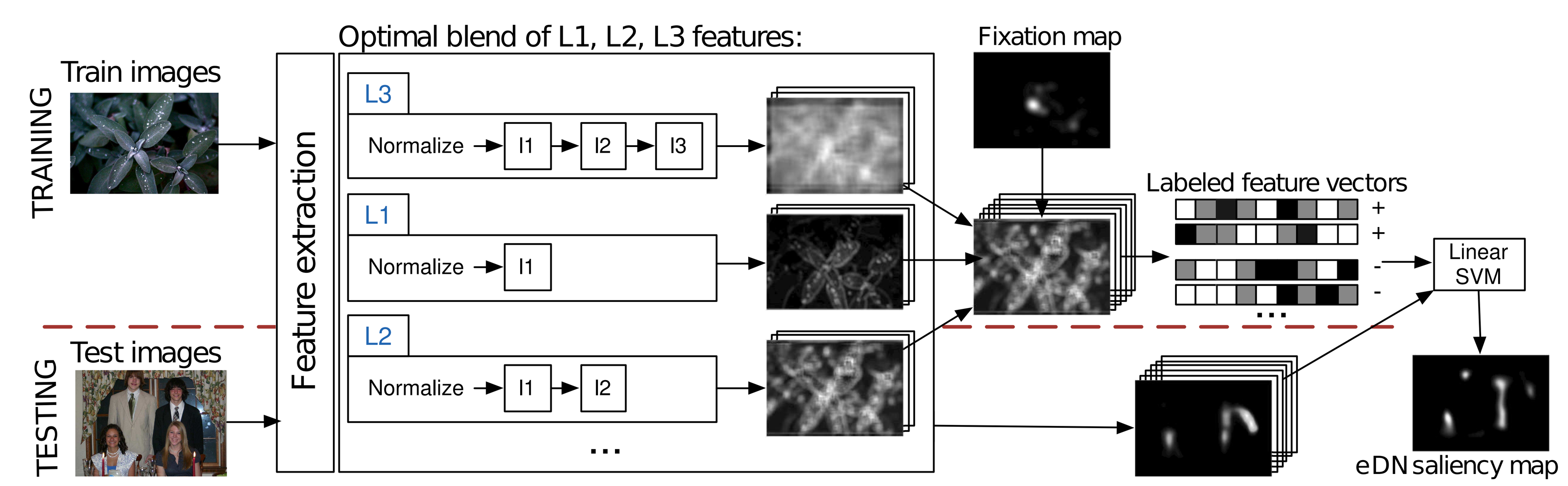
Automatic Hyper-Parameter Optimization

- highly configurable architecture, but many hyperparameters to tune
- perform an efficient search for best architecture(s)
- **hyperopt**: "library for optimizing over awkward search spaces with real-valued, discrete, and conditional dimensions" (Bergstra *et al.*, ICML'13)
- optimization algorithm used: Tree of Parzen Estimators

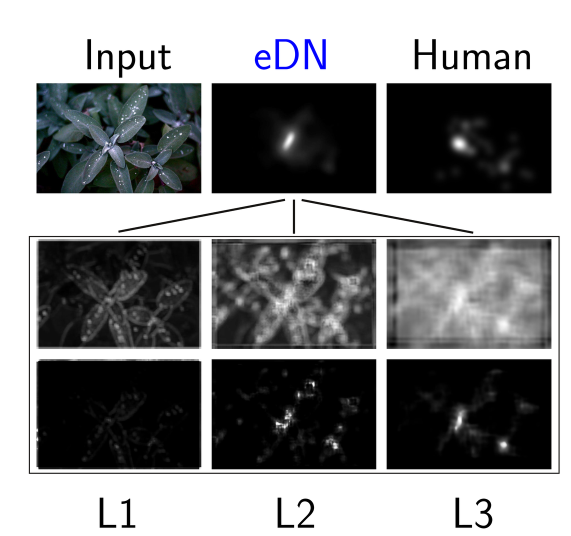
Feature Search Pipeline

- performed on a subset of the MIT1003 data set (600 images)
- two-stage search:
  - search for individual L1, L2, L3 models (RGB and YUV input)
  - search for ensembles of best individual models

Saliency Prediction Pipeline



Individual Models and their Blend



AUC scores:

Model	RGB	YUV
L1	0.6744	0.6705
L2	0.6737	0.7401
L3	0.7207	0.7977
eDN	0.8227	

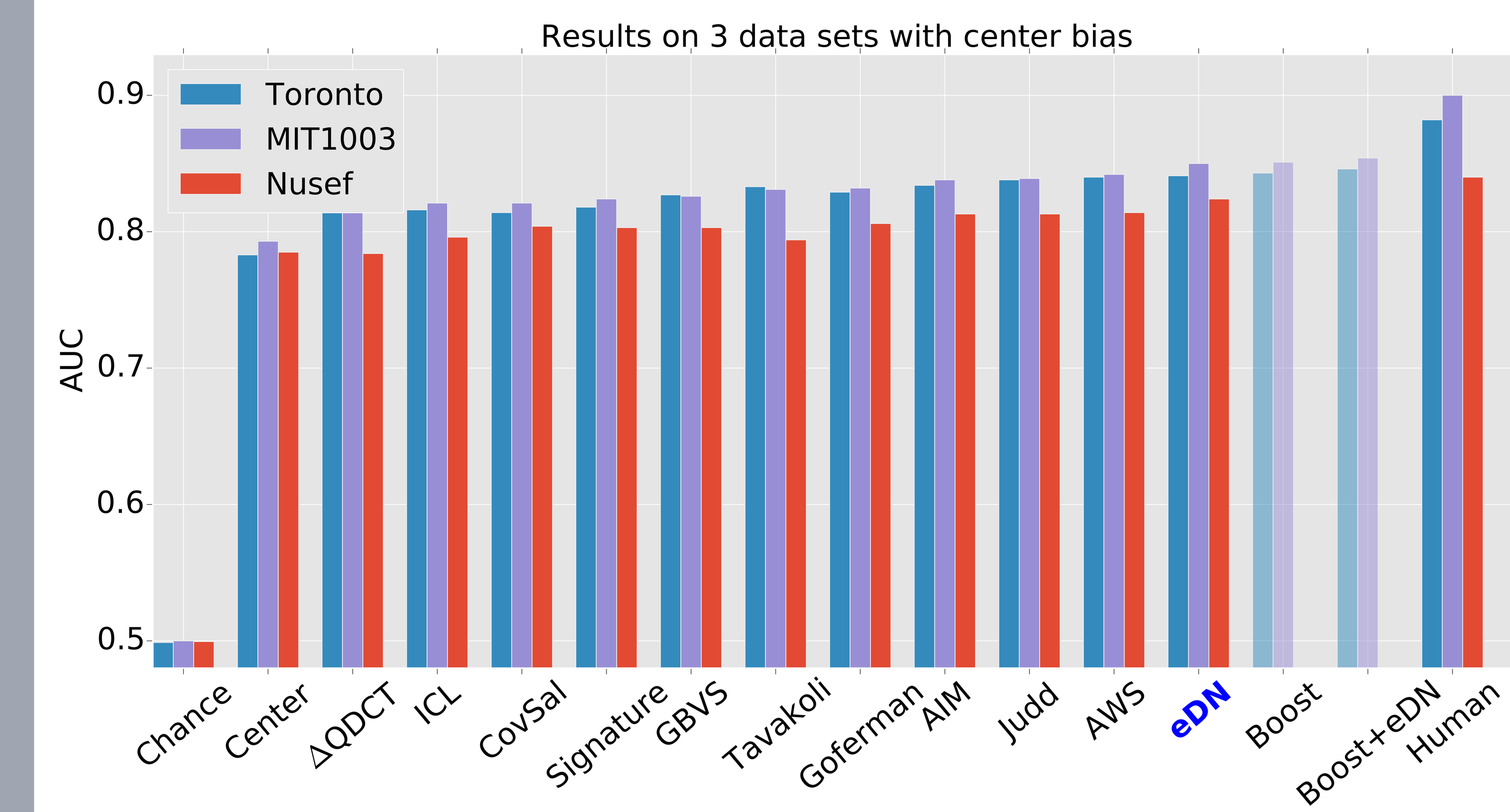
eDN: ensemble of Deep Networks

Evaluation – Eye Movement Benchmarks

1. MIT1003: 1003 images, 15 viewers, many faces
2. Toronto: 120 images, 20 subjects, no faces
3. Nusef: 758 images (affective content), 75 viewers
4. MIT300: 300 images, 39 viewers (gaze data not public)

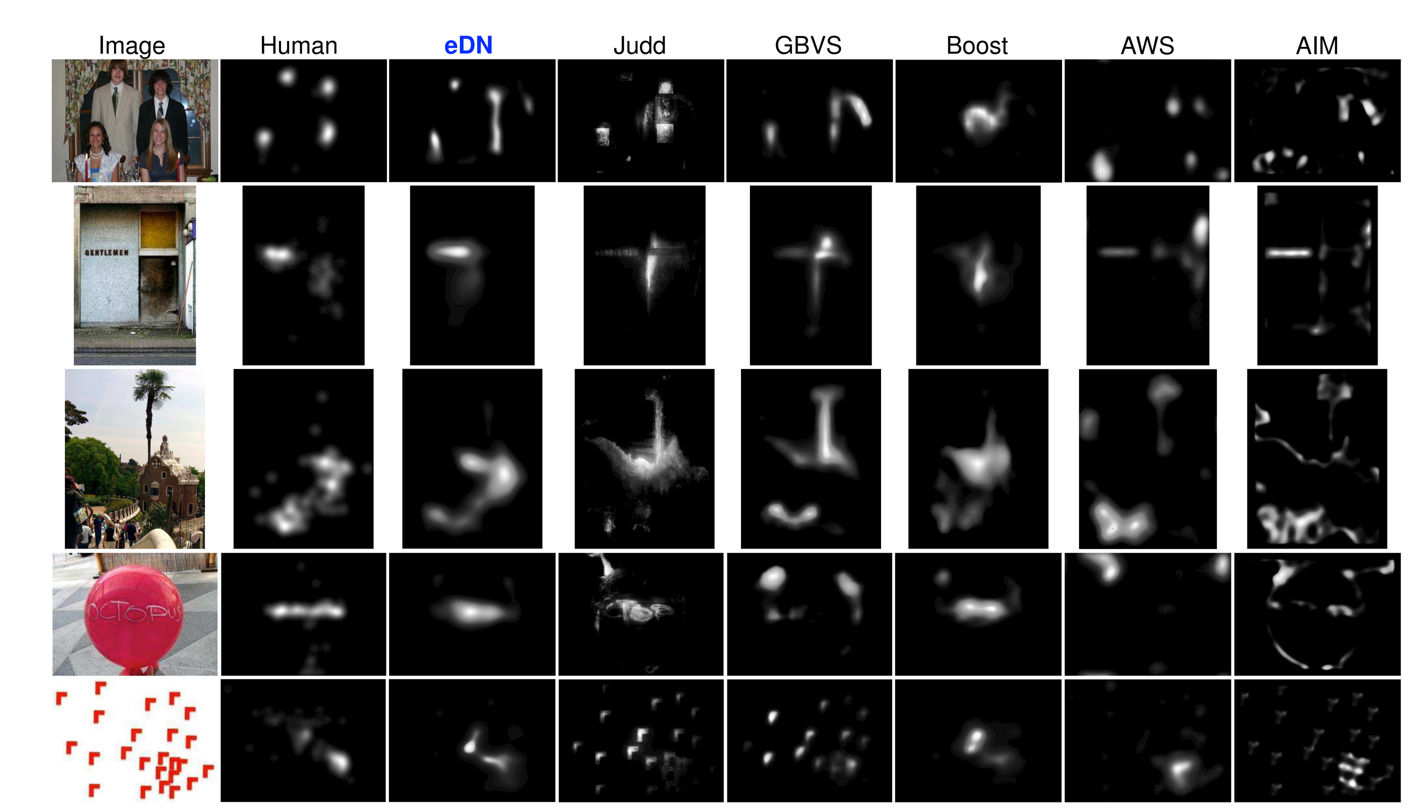
+ 4 metrics (AUC, EMD, similarity, NSS)

Results



Best of 23 models on MIT300 (2^{nd} as of 29.05.2014):

	Model Name	Link to code	Area under ROC* curve (higher is better)	Similarity* (higher is better)	Earth mover's distance* (lower is better)
	Humans**	code	0.922	1	0
eDN	Bio-inspired hierarchical features	(coming soon)	0.8192	0.5123	3.0129
	Judd et al.	code	0.811	0.506	3.13
	CovSal	paper, website	0.8056	0.5018	3.1092
	Tavakoli et al. 2011	paper and website	0.8033	0.4952	3.3488



Conclusions

- efficient search in a large pool of richly-parameterized neuromorphic models
- automated blending of individual models → diversity, multiple scales
- no assumptions on what features/objects attract attention → learn them
- best performance on several benchmarks

http://coxlab.org/saliency/ eleonora.vig@xrce.xerox.com