Biologically Plausible Detection of Amorphous Objects in the Wild

Sunhyoung Han, Nuno Vasconcelos
Statistical Visual Computing Lab.
University of California, San Diego

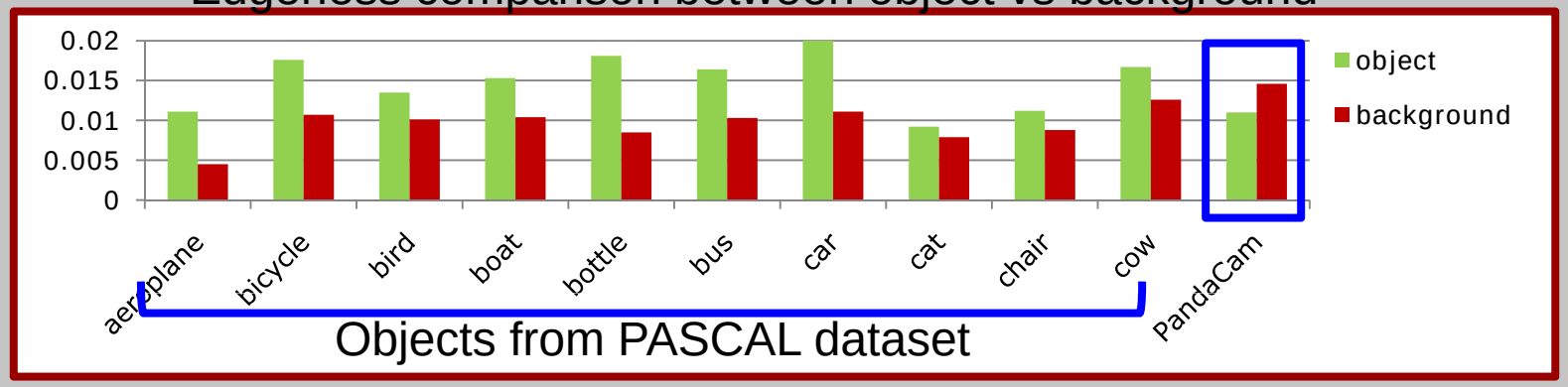
SVCL UCSD

PandaCam dataset

PandaCam dataset

- Video feed provided by San Diego Zoo showing real time movement of a Panda family in a natural habitat
- 24/7 surveillance type real application
  - Dramatic change of illumination
  - Low quality images
  - Video is a collection from multiple cameras and viewpoint
- Textureless amorphous object in highly structured background
  - Highly deformable shape of panda bear
  - Background have richer combination of structure, shape, texture

Edgeness comparison between object vs background



Outline

- Biologically plausible Discriminant saliency network
- Connection with existing object recognition models
  - HMAX, convolutional neural network
- Unique property of DiscSalNet
  - Detection of feature absence
- Experimental results & conclusion

Discriminant Saliency network

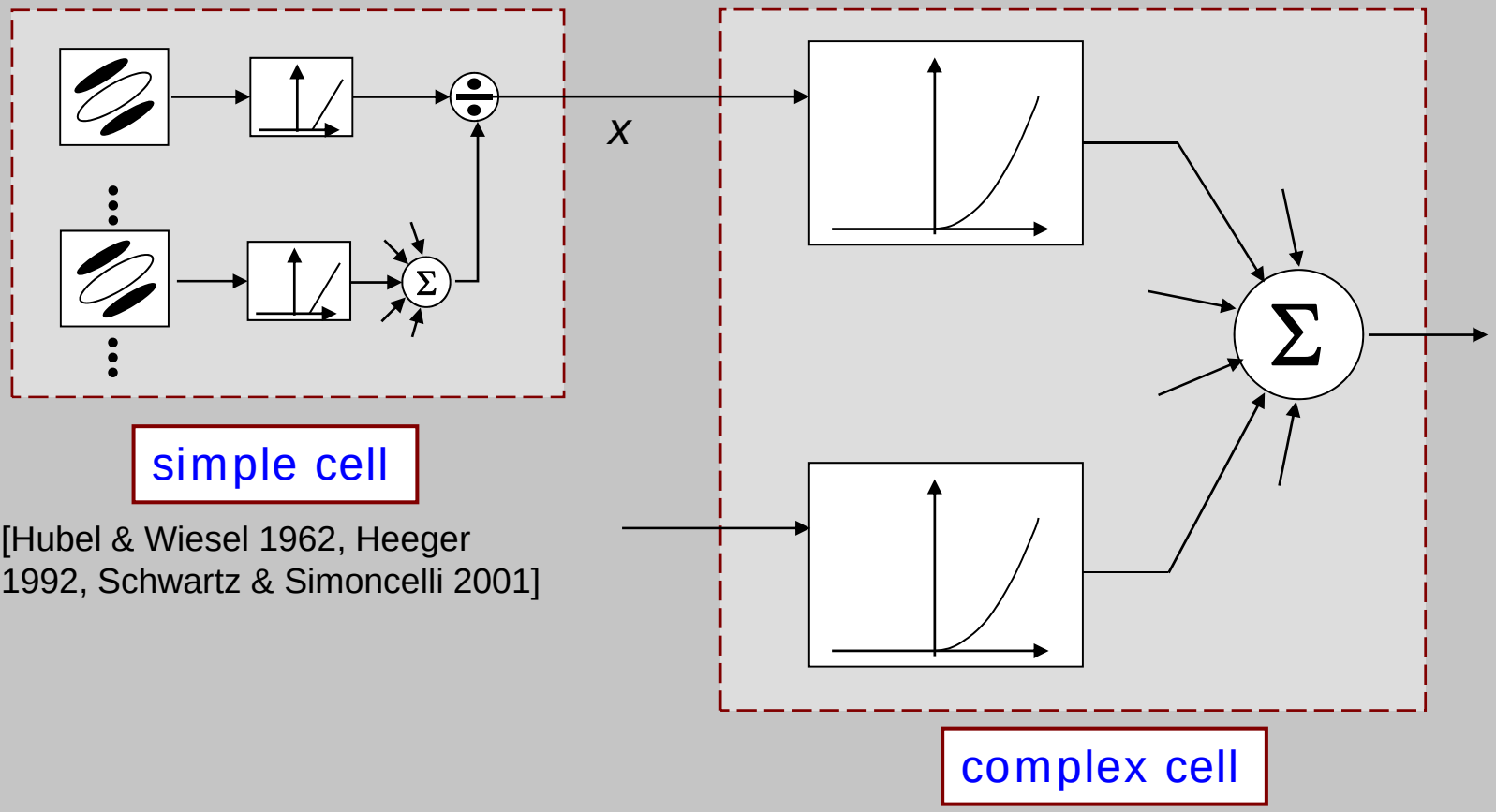
- Discriminant saliency implemented in standard neurophysiological model

- we are investigating a discriminant view of perception
  - brains are optimized for decision-making
  - this holds at all levels of neural circuitry from single cells to populations and high-level circuits

- Hierarchical network architecture

Neurophysiology of vision

- standard architecture of the primary visual cortex (V1)
  - a cascades of linear filter, divisive normalization, a quadratic nonlinearity, and spatial pooling [Carandini et al., 2005]



[Hubel & Wiesel 1962, Heeger 1992, Schwartz & Simoncelli 2001]

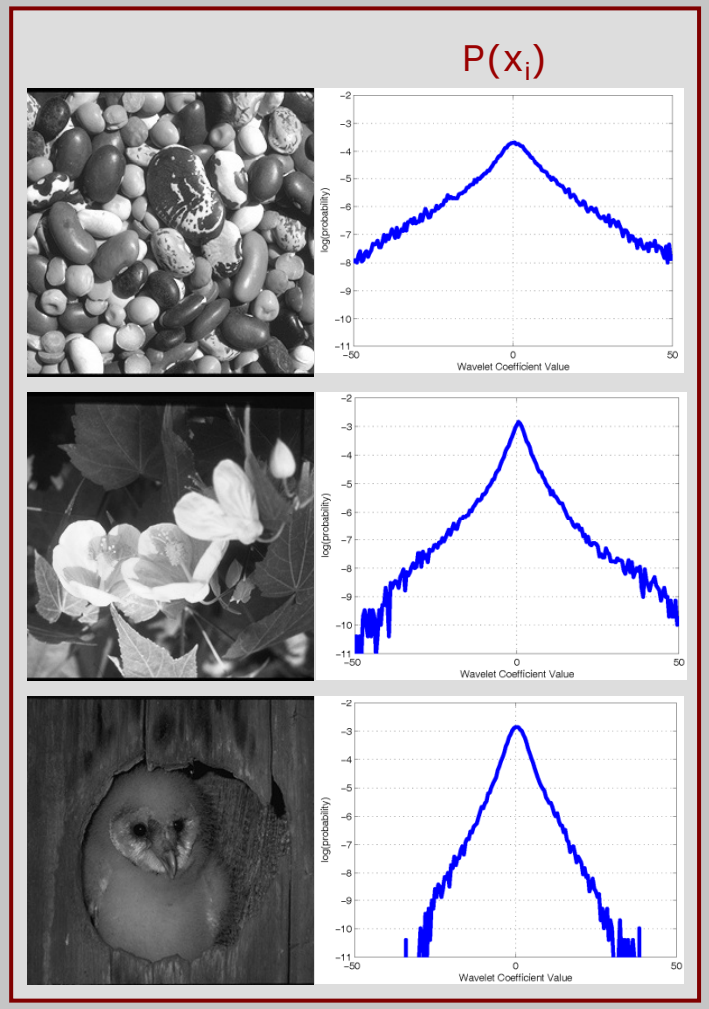
[Energy model: Adelson-Bergen 1985]

Natural stimuli statistics

- whenever we pass a natural image through a band-pass filter (Gabor, wavelet, DCT, PCA, etc.)
- response follows generalized Gaussian density (GGD)

$$p(x;\alpha,\beta) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left(-(|x|/\alpha)^{\beta}\right)$$

  - parameters: α scale, β shape
  - heavy tails for $\beta < 1$
- extremely consistent observation (any natural images, any filter)



Generalized Gaussian Density (GGD)

- GGDs have two interesting properties
  - negative log-likelihood is proportional to response

$$\boxed{-\log p(x;\alpha,\beta) = \left(\frac{|x|}{\alpha}\right)^{\beta} + K}$$

  - parameter estimation:
    - using a conjugate (Gamma) prior for the inverse scale ($\theta = 1/\alpha^{\beta}$)

$$P_{\Theta}(\theta) = Gamma\left(\theta, 1+\frac{\eta}{\beta}, \nu\right) = \frac{\nu^{1+\eta/\beta}}{\Gamma(1+\eta/\beta)}\theta^{\eta/\beta}e^{-\nu\theta}$$

    - MAP estimate of α from sample of iid responses $\{x(1),\ldots,x(n)\}$

$$\boxed{\hat{\alpha} \propto \left(\sum_j |x(j)|^{\beta} + \nu\right)^{1/\beta}}$$

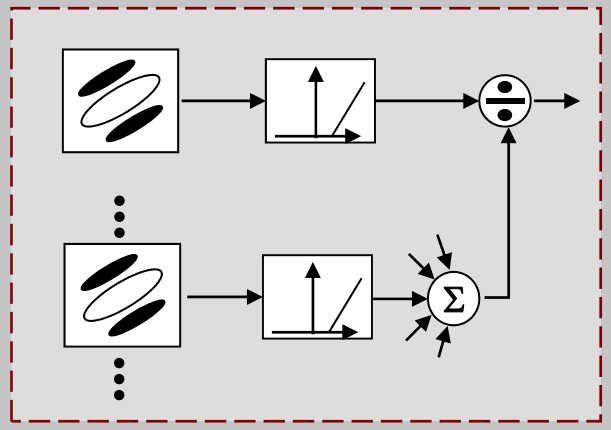
is the norm of the vector of those responses

Simple cells

- even more interesting:
  - when we put the two together we obtain divisive normalization

$$-\log p(x; \{x(j)\}) \propto \frac{|x|^\beta}{\sum_j |x(j)|^\beta + \nu}$$

  - simple cell computations
    - probability of stimulus in cell's receptive field
    - under the GGD of the "normalizing responses" $\{x(j)\}$



simple cell

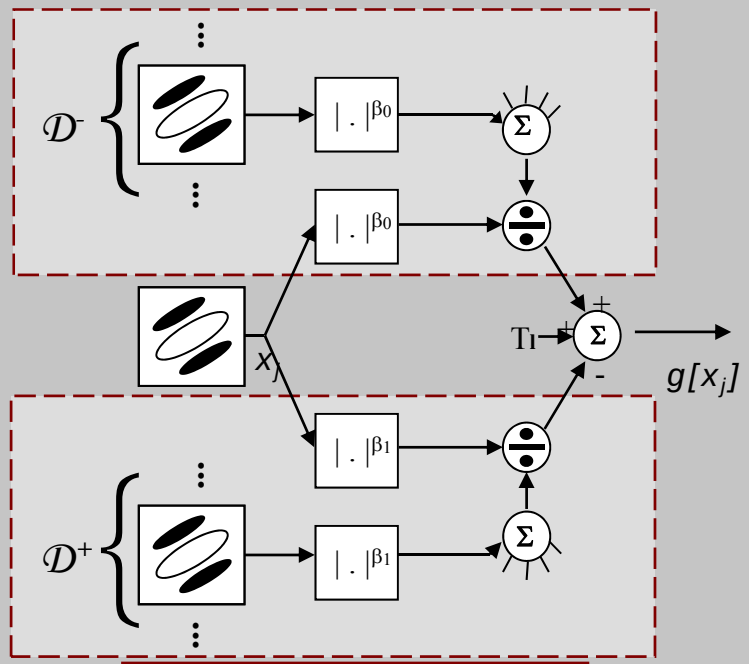
[Hubel & Wiesel 1962, Heeger 1992, Schwartz & Simoncelli 2001]

Optimal decision-making

- from this
  - we can build all components of statistical inference
- classification/decision rules:
  - Log-likelihood ratio

$$g(x) = \log \frac{P_{X\,Y}(x|1)}{P_{X|Y}(x|-1)}$$

  - inhibitory connections: learning of null hypothesis *(Y=-1)*
  - excitatory connections: learning alternative hypothesis *(Y=1)*

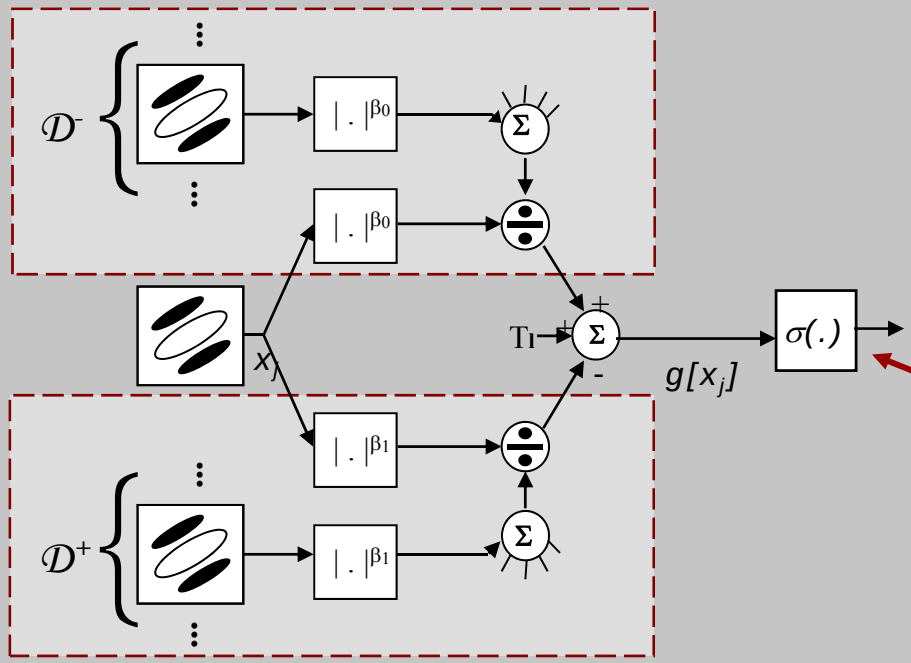


differential simple cell

Optimal decision-making

- classification:
  - Log-likelihood ratio

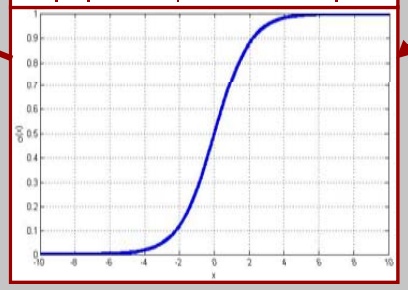
$$g(x) = \log \frac{P_{X|Y}(x|1)}{P_{X|Y}(x|-1)}$$



- classification:
  - posterior probability

$$P_{Y|X}(1|x) = \frac{P_{X|Y}(x|1)}{P_{X|Y}(x|1) + P_{X|Y}(x|-1)} = \frac{1}{1 + \frac{P_{X|Y}(x|-1)}{P_{X|Y}(x|-1)}} = (1 + \exp\{-g(x)\})^{-1} = \sigma(x)$$

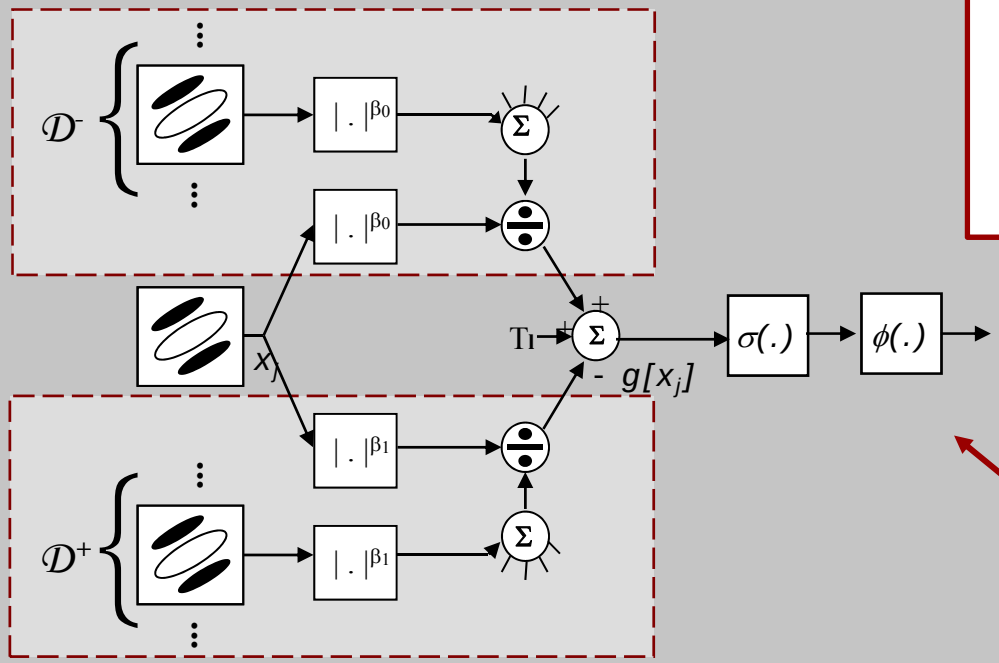
$$\sigma(x) = (1 + e^{-x})^{-1}$$



Optimal decision-making

- classification:
  - posterior probability

$$g(x) = \sigma[g(x)]$$

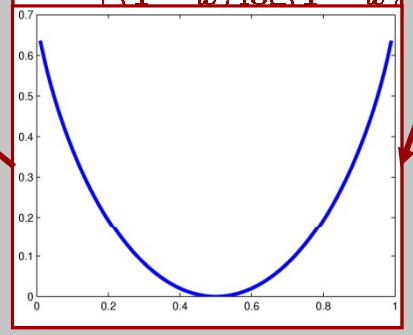


- confidence measures

		$\phi(.)$
information	$I(Y;X=x)$	$\xi\{\sigma(.)\}$

$$\begin{aligned} I(Y, X = x) &= \\ &= H(Y) - H(Y|X = x) \\ &= \log 2 - P_{Y|X}(1|x) \log P_{Y|X}(1|x) \\ &\quad - P_{Y|X}(-1|x) \log P_{Y|X}(-1|x) \\ &= \xi\{P_{Y|X}(1|x)\} \end{aligned}$$

$$\xi(x) = \log 2 + x \log x + (1-x)\log(1-x)$$



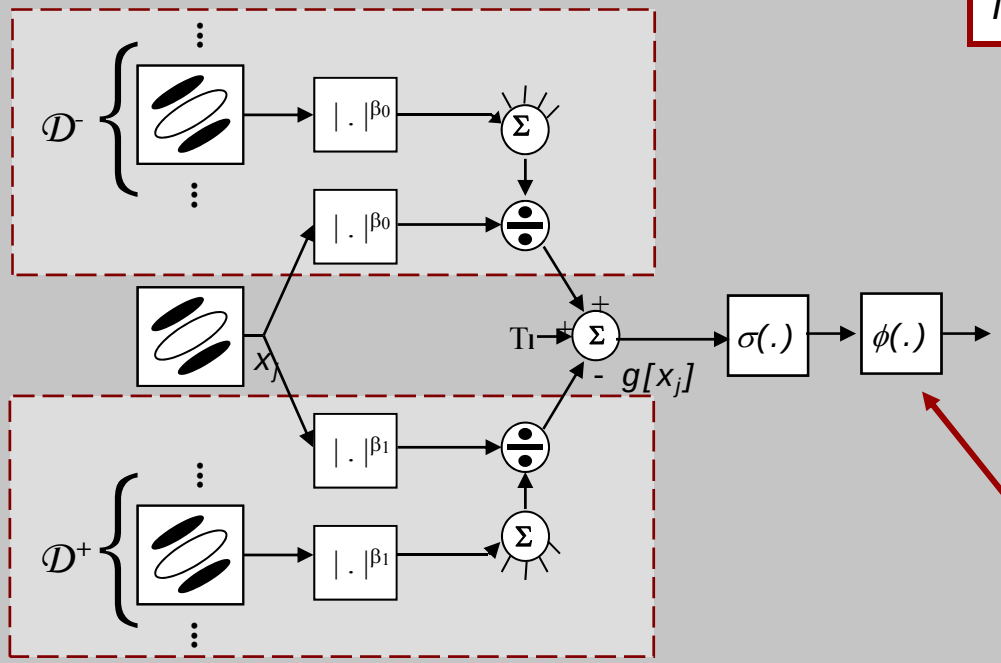
Optimal decision-making

- classification:
  - posterior probability

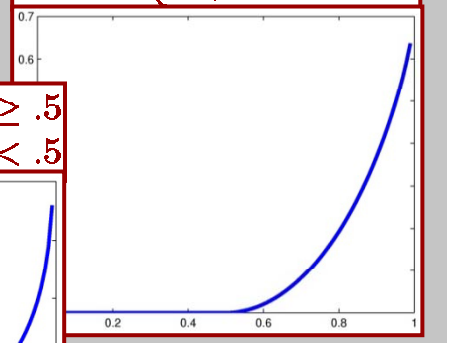
$$g(x) = \sigma[g(x)]$$

- confidence measures

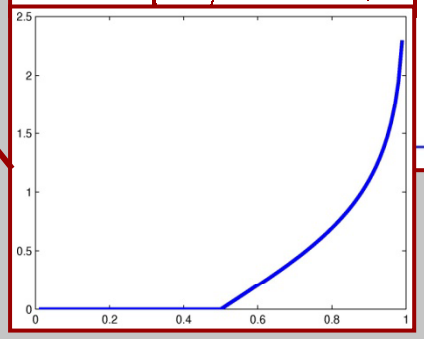
		$\phi(.)$
Info(x)	$I(Y;X=x)$	$\xi(.)$
LLRC (x)	$[g(x)]_+$	$\tilde{\psi}(\cdot)$
IC(x)	$[I(Y;X=x)]_+$	$\tilde{\xi}(\cdot)$



$$\tilde{\xi}(x) = \begin{cases} \xi(x), & x \geq .5 \\ 0, & x < .5 \end{cases}$$



$$\tilde{\psi}(x) = \begin{cases} \psi(x), & x \geq .5 \\ 0, & x < .5 \end{cases}$$



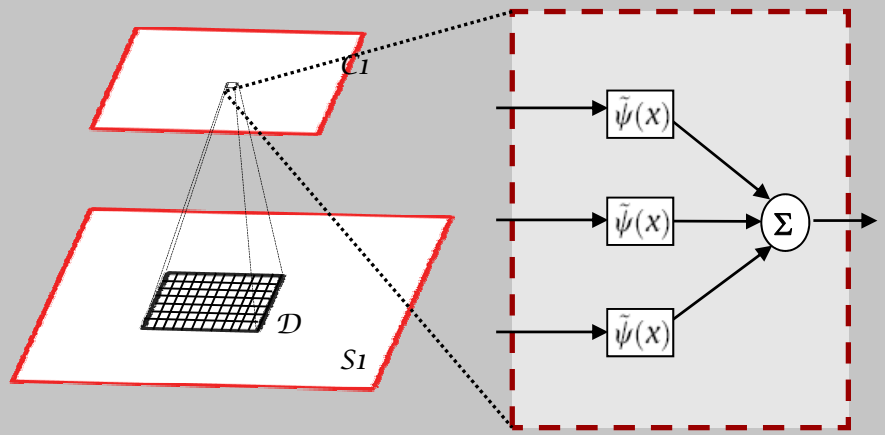
Risks

- given a sample $\mathscr{D} = \{x_1, \ldots, x_n\}$
  - these can be used to compute various risks

$$\tilde{\psi}(x) = \begin{cases} \frac{1}{2}\log\frac{x}{1-x}, & x \geqslant .5 \\ 0, & x < .5 \end{cases}$$

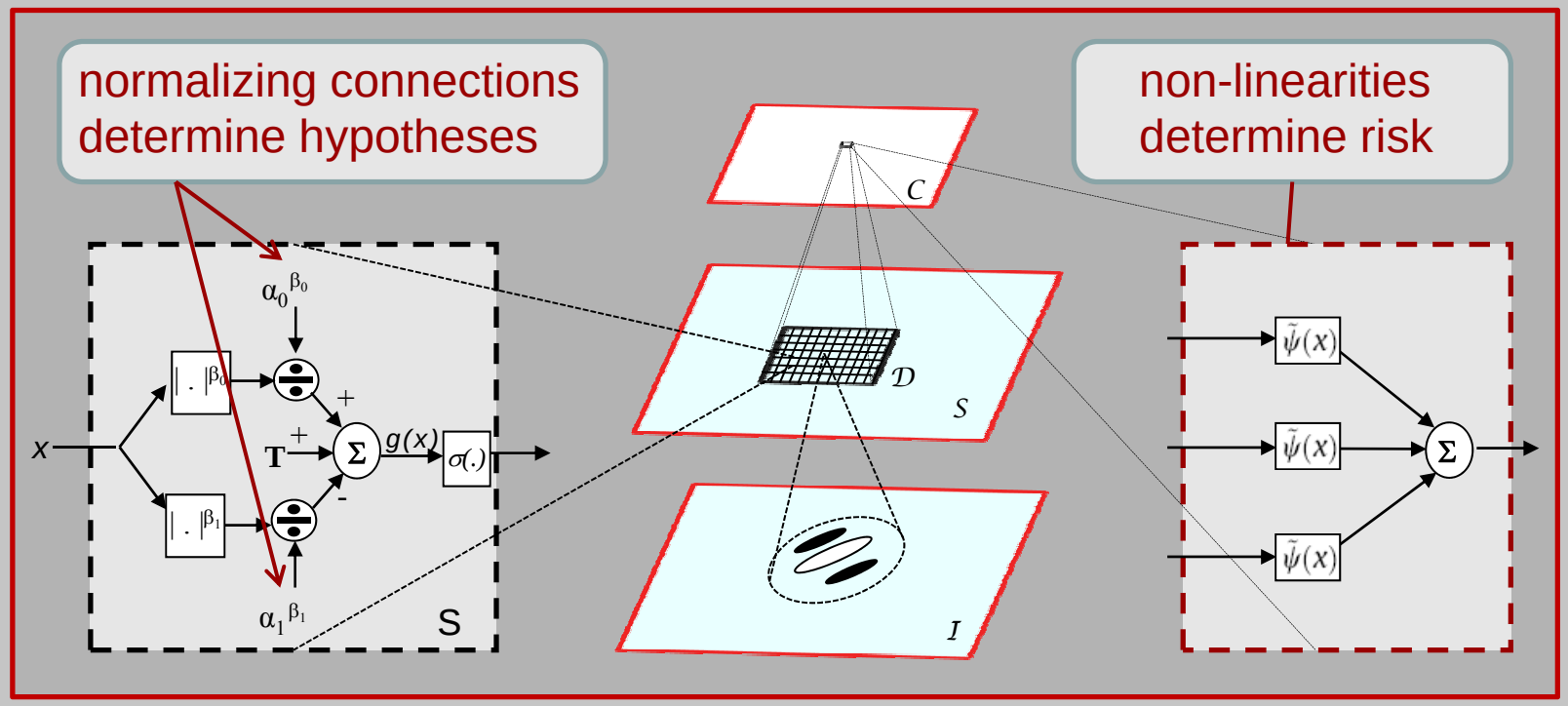
Empirical risks based on sample $\mathcal{R} = \{x_1, \ldots, x_n\}$			
expected NLL	$E_X\left[-\log P_X(x)\right]$	$\frac{1}{n}\sum_{i=1}^{n} l_{\alpha,\beta}(x_i)$	entropy $H[X]$
expected LLR	$E_X\left[\log\frac{P_{X\|Y}(x\|1)}{P_{X\|Y}(x\|0)}\right]$	$\frac{1}{n}\sum_{i=1}^{n} g(x_i)$	$KL[P_X(x)\|\|P_{X\|Y}(x\|0)]$ $-KL[P_X(x)\|\|P_{X\|Y}(x\|1)]$
MI	$E_X\left[I(Y;X=x)\right]$	$\frac{1}{n}\sum_{i=1}^{n} \xi\{\sigma[g(x_i)]\}$	$I(Y;X)$
expected confidence (LLR)	$E_X[LLR(x)]$	$\frac{1}{n}\sum_{i=1}^{n} \tilde{\psi}\{\sigma[g(x_i)]\}$	$KL[P_{X\|Y}(x\|1)\|\|P_{X\|Y}(x\|0)]$
expected confidence (MI)	$E_X[IC(x)]$	$\frac{1}{n}\sum_{i=1}^{n} \tilde{\xi}\{\sigma[g(x_i)]\}$	

  - non-linearity plus pooling
  - computations of the complex cell



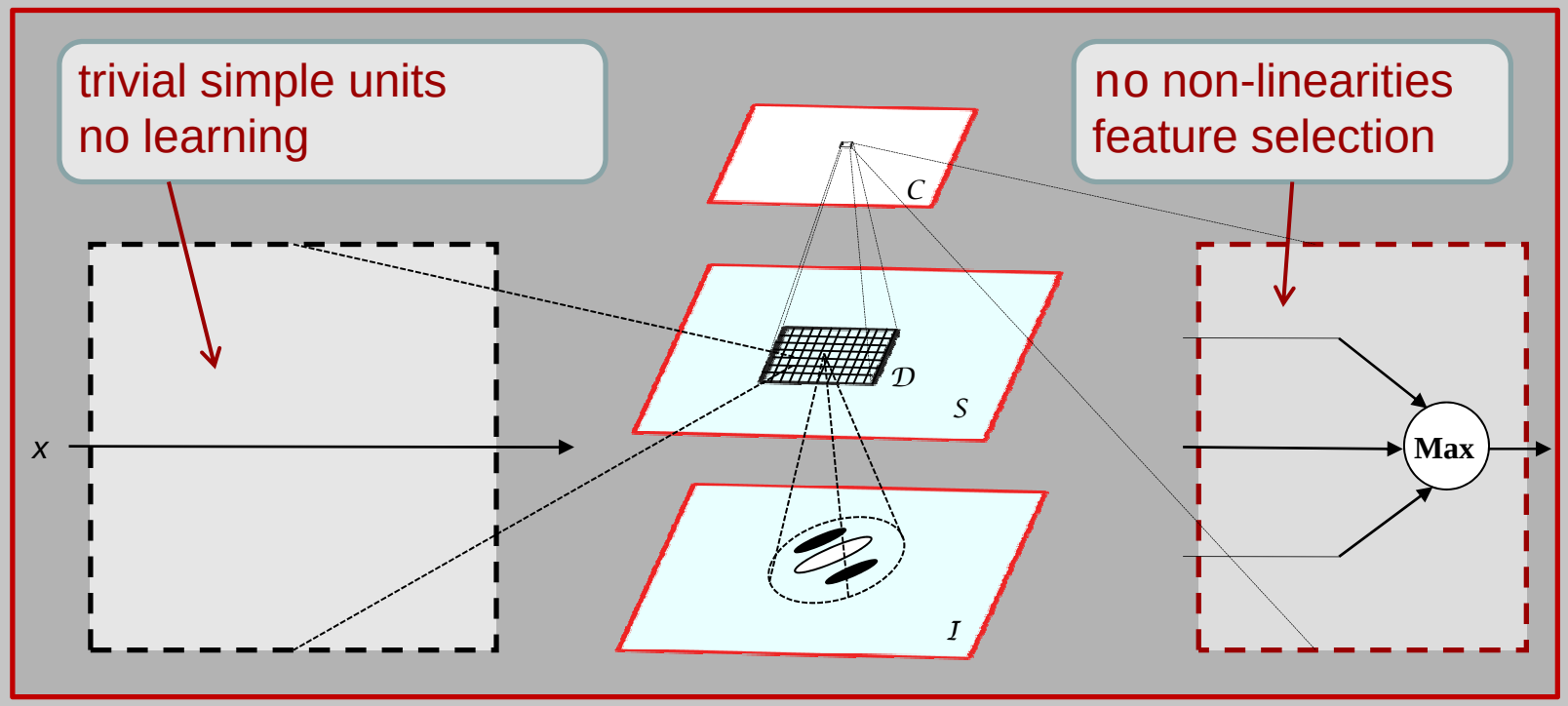
Discriminant saliency Layer

- SC network:
  - Gabor decomposition into orientation channels
  - simple units compute target probabilities $P_{Y|X}(1|X)$
  - complex units compute risk



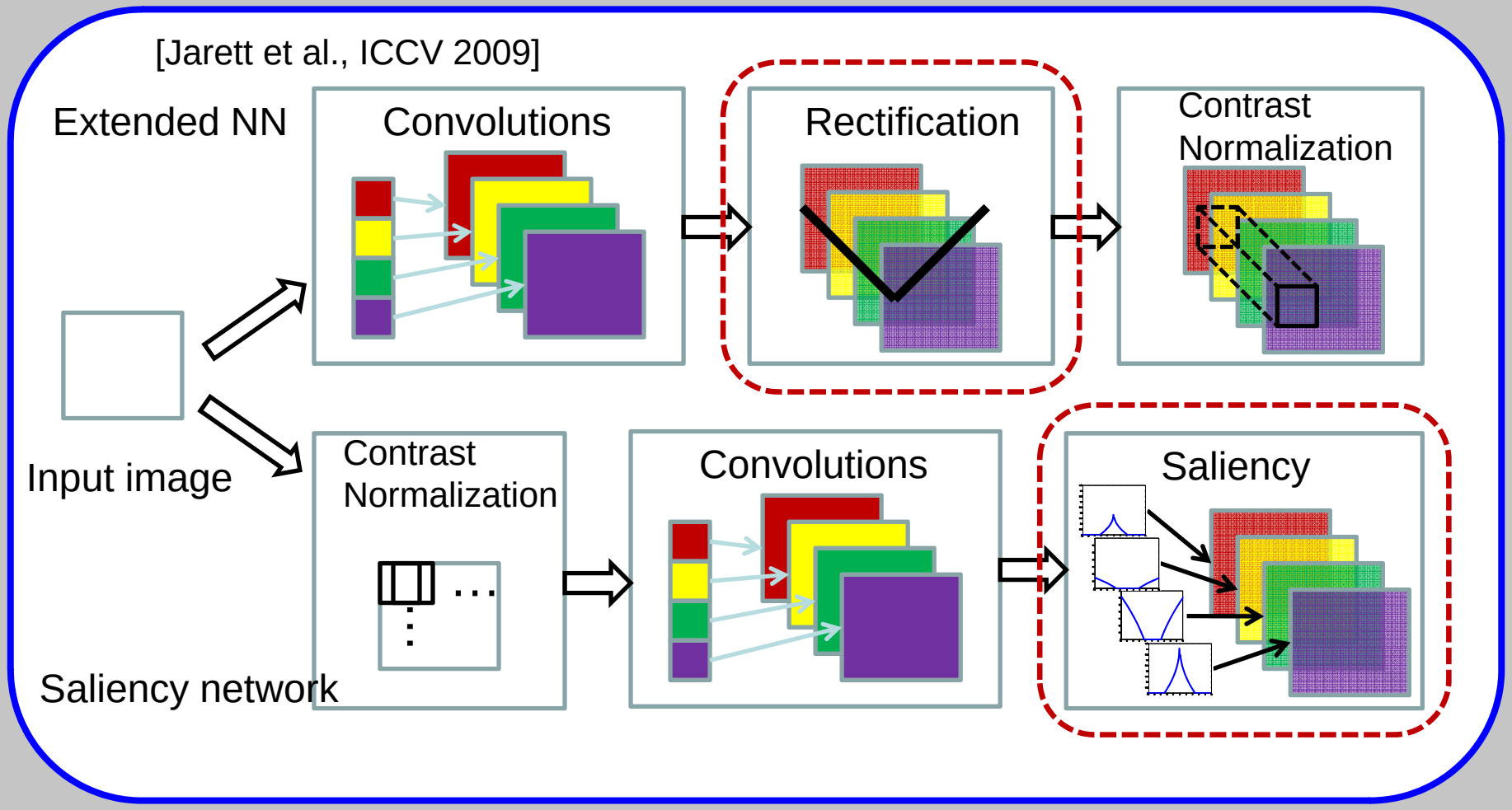
HMAX units

- similar to DS units but:
  - simple units do trivial Gabor filtering
  - complex units more like feature detectors (max pooling) than risk assessors

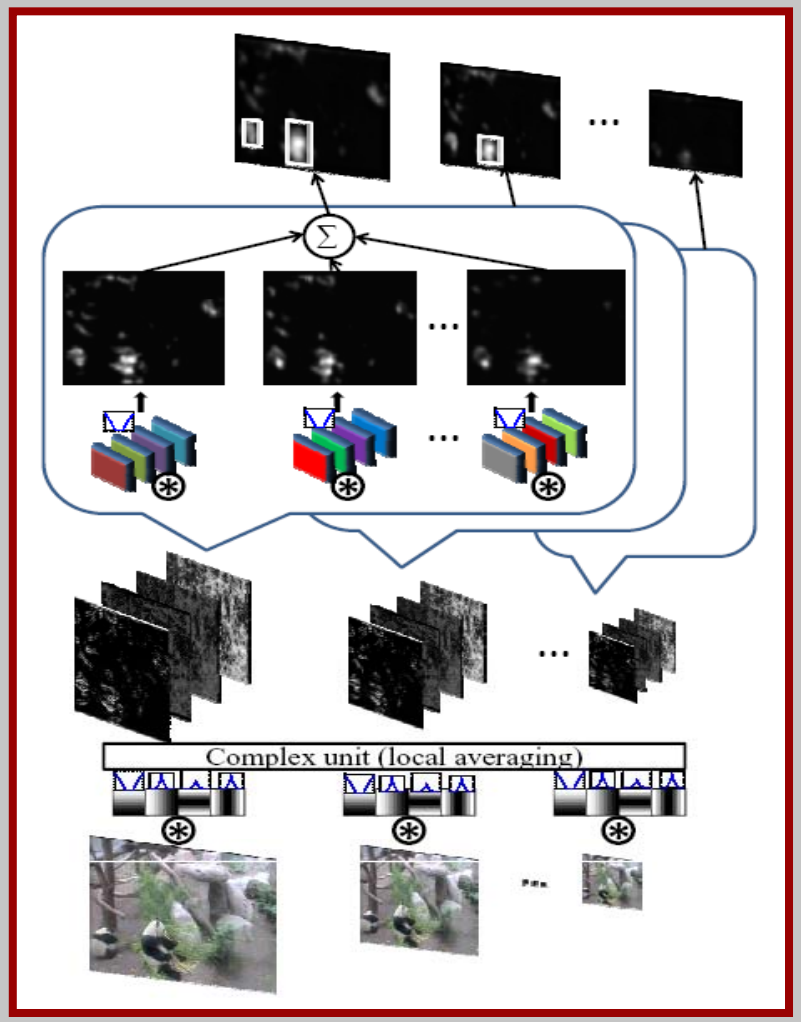


Comparison with other algorithms

Discriminant Saliency layer vs. extended neural network

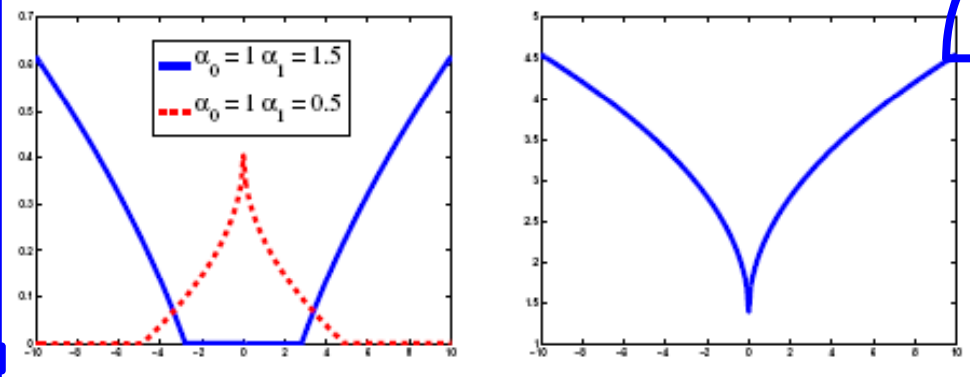


Discriminant Saliency network



- two layers of simple + complex units
  - layer 1: simple feature-based DS units (gabor, DCT basis)
  - layer 2:
    - templates sampled randomly from L1 training responses
    - template-based DS units
- Hierarchical architecture shared with HMAX, convNN, SPMK

Feature absence detection



(a) discriminant saliency (b) entropy

$$H(l) = -\int_{\mathcal{A}(l)} P_X(x) \log P_X(x) dx$$
$$= K + \frac{1}{2|\mathcal{A}(l)|} \sum_{i \in \mathcal{A}(l)} \left(\frac{|x_i|}{\alpha} \right)^{\beta}$$

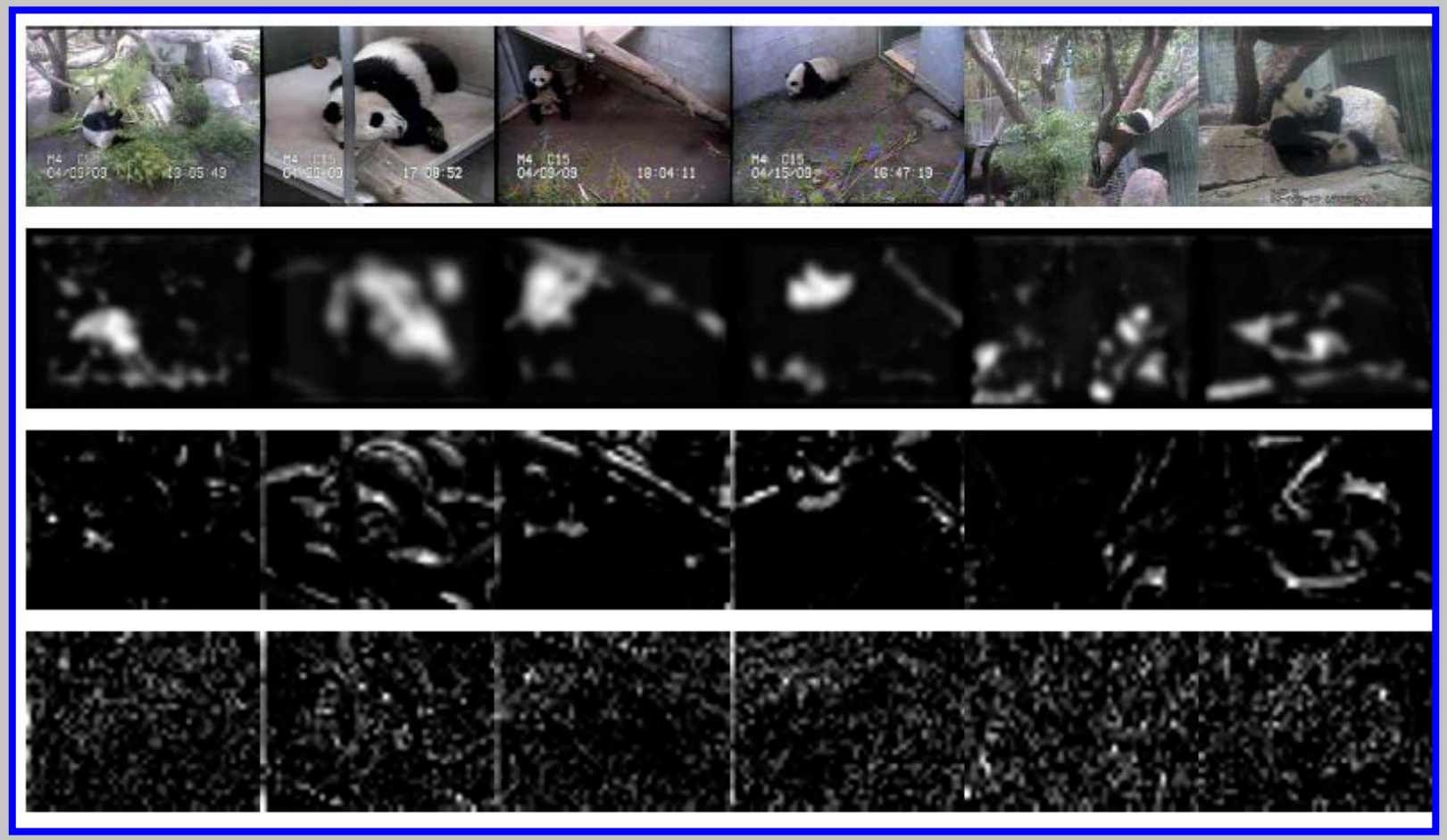
$$S(l) = \int_{x \in \mathcal{A}(l)} P_{X|Y}(x|1) \log \frac{P_{X|Y}(x|1)}{P_{X|Y}(x|0)} dx. \approx \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}(l)} \delta_i \log \frac{P_{X|Y}(x_i|1)}{P_{X|Y}(x_i|0)}$$

where $\hat{\delta}_i = \begin{cases} 1 & \text{if } P_{Y|X}(1|x_i) \geq 0.5 \\ 0 & \text{otherwise.} \end{cases}$

Range of non-zero

$|x|^{\beta} \geq \frac{\alpha_0^{\beta} \alpha_1^{\beta}}{\alpha_1^{\beta} - \alpha_0^{\beta}} \log \frac{\alpha_1}{\alpha_0}$ when $\alpha_1 > \alpha_0$ $|x|^{\beta} \leq \frac{\alpha_0^{\beta} \alpha_1^{\beta}}{\alpha_1^{\beta} - \alpha_0^{\beta}} \log \frac{\alpha_1}{\alpha_0}$ when $\alpha_1 < \alpha_0$

Experiment: localization



Image

Sal/Sal

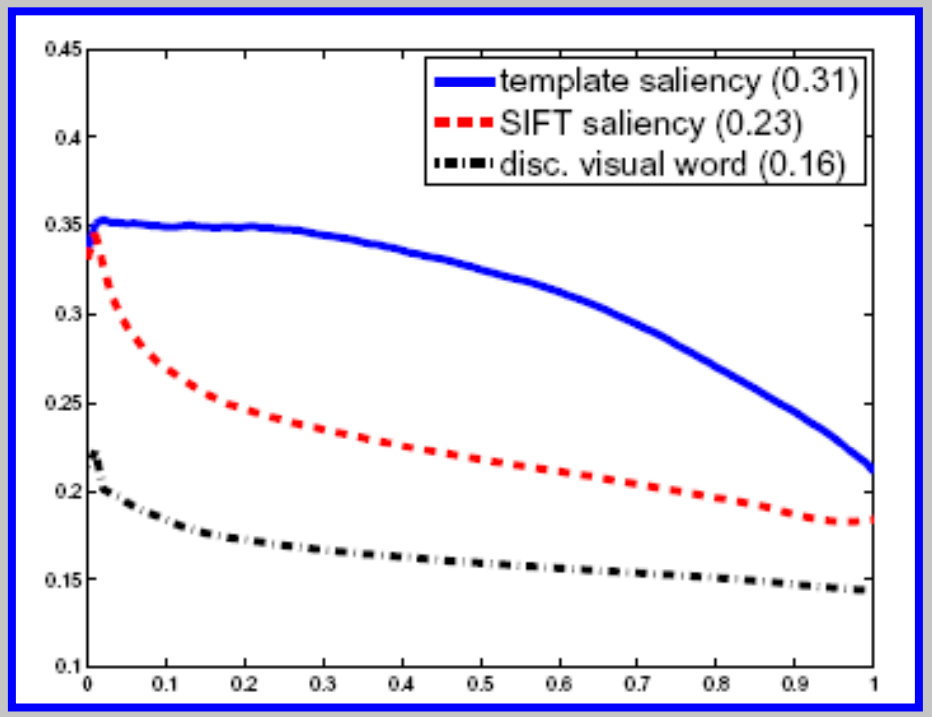
SIFT/Sal

SIFT/DW

Discriminant Word $D(w) = \frac{\#\text{ target images containing } w}{\#\text{ images containing } w}.$

[Dorko & Schmid PAMI, 2004]

Results: localization performance

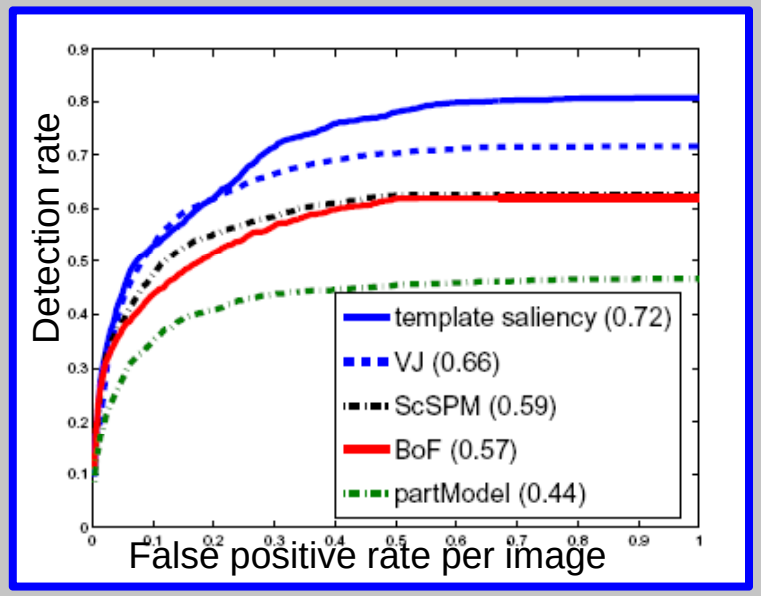


Precision Recall curve

Results: Panda detection

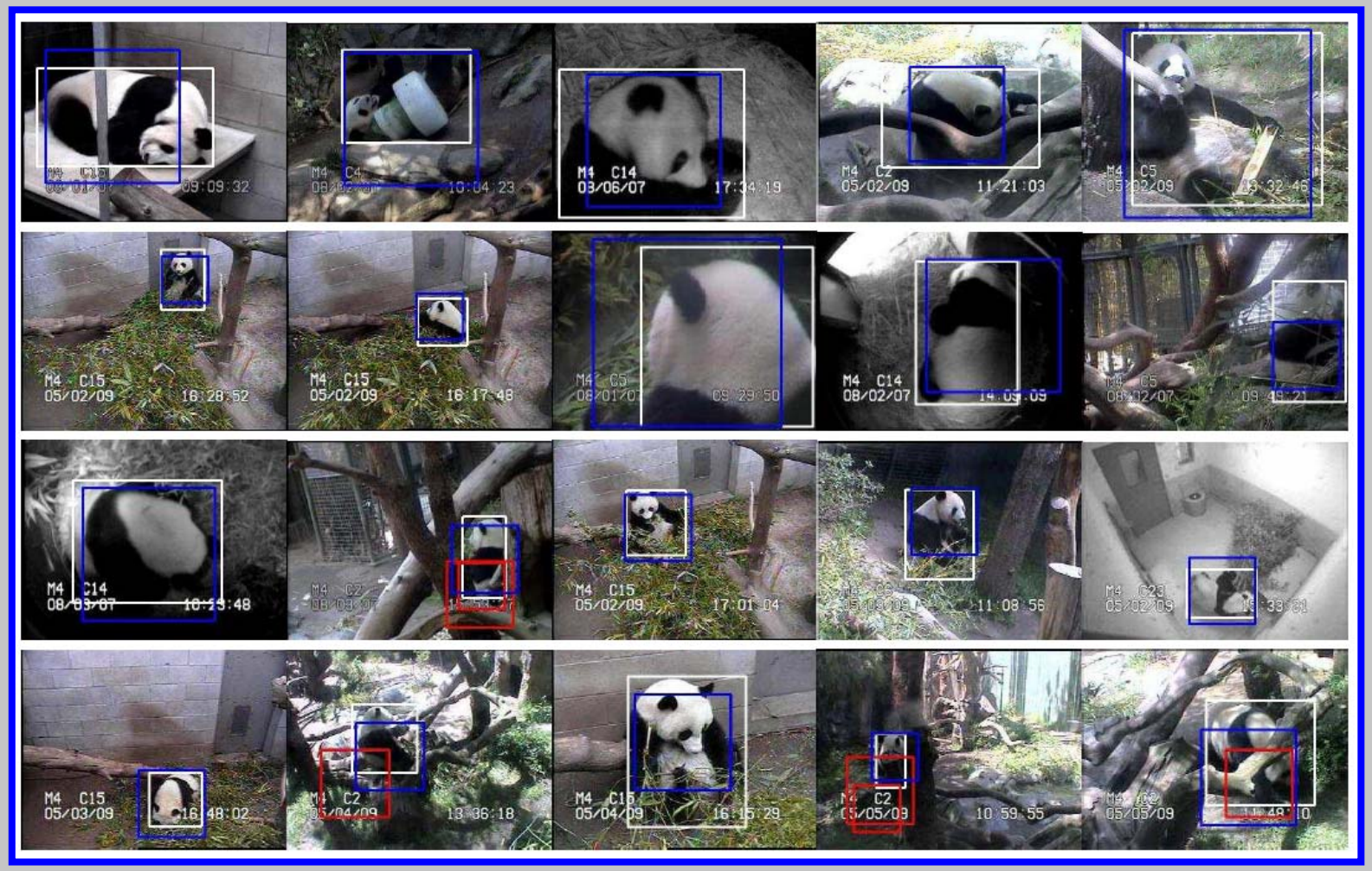
Post processing

- box filter is applied to 7 scale saliency maps
- center location and scale of largest saliency is detected
- non-maximum suppression is applied



VJ: Viola & Jones CVPR 2001
ScSPM : Yang et al. CVPR 2009
SPMK: Lazebnik et al. CVPR 2006
partModel: Felzenszwalb et al CVPR 2008

Results: Panda detection



Results: comparison on Caltech101 and Natural scene 15 classification dataset

Method	# Layer 2 units		Recognition rate		# Layer 2 units		Recognition rate	
	N15	C101	N15	C101	N15	C101	N15	C101
V1 model[1]	-	4000	-	42±0.5	-	86K	-	65
HMAX[2]	-	4075	-	56	-	-	-	-
convNN[3]	-	4096	-	65.5	-	-	-	-
BoF[4]	-	4200	-	64.6	8400	-	81.4±0.5	-
ScSPM[5]	5120	5120	75.3±0.5	64.8±0.7	21504	21504	80.28±0.9	**73.2±0.5**
NBNN[6]	-	-	-	-	-	10M	-	70.4
HGMM[7]	-	-	-	-	46080	310272	**85.2**	**73.1**
DiscSalNet	600	4040	**82±0.5**	**70±0.5**	22500	20200	**85.4±0.3**	**73.1±0.6**

[1] Pinto et al. PLoS 2008 [2] Mutch & Lowe CVPR 2006 [3] Jarrett et al. ICCV 2009
[4] Lazebnik et al. CVPR 2006 [5] Yang et al. CVPR 2009 [6] Boiman et al. CVPR 2008
[7] Zhou et al. ICCV 2009

- Performances are **highly dependent** to the # of L2 units
- With similar L2 unit number, **Saliency Net achieves the best performance**

Conclusion

- top-down discriminant saliency model for amorphous object detection
  - Cell level optimal tuning for given problem
  - Complex feature in higher level (increase selectivity)
  - Absence of feature can also be exploited
  - Achieves best performance among state-of-the-arts computer vision algorithms on standard test benchmarks
- Edge based (SIFT, HOG, etc) feature may not be enough
  - State-of-the-arts method could be tuned too much to the specific dataset

Thank you!