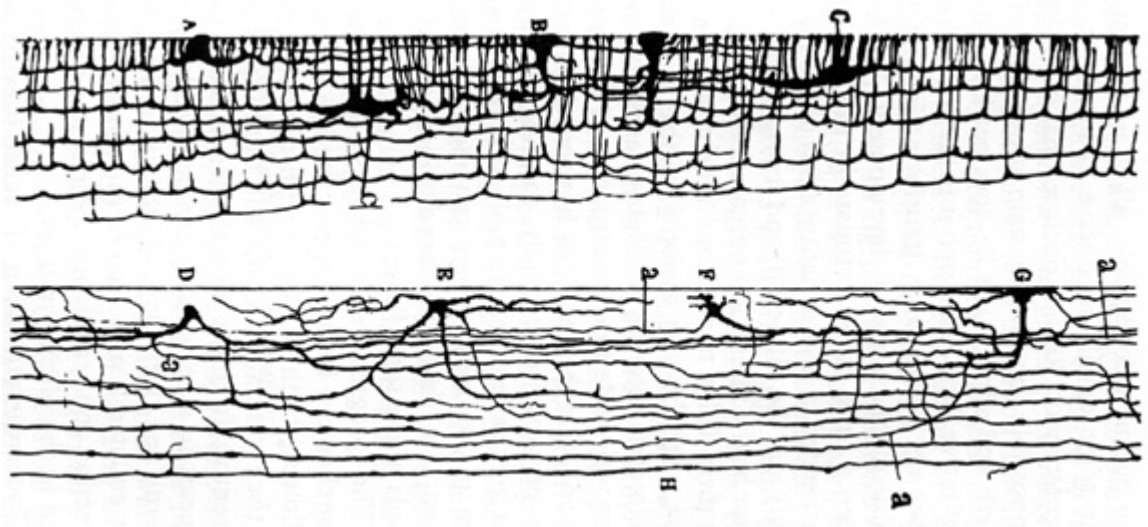


Robust Object Recognition with Cortex-Like Mechanisms

Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber and Tomaso Poggio

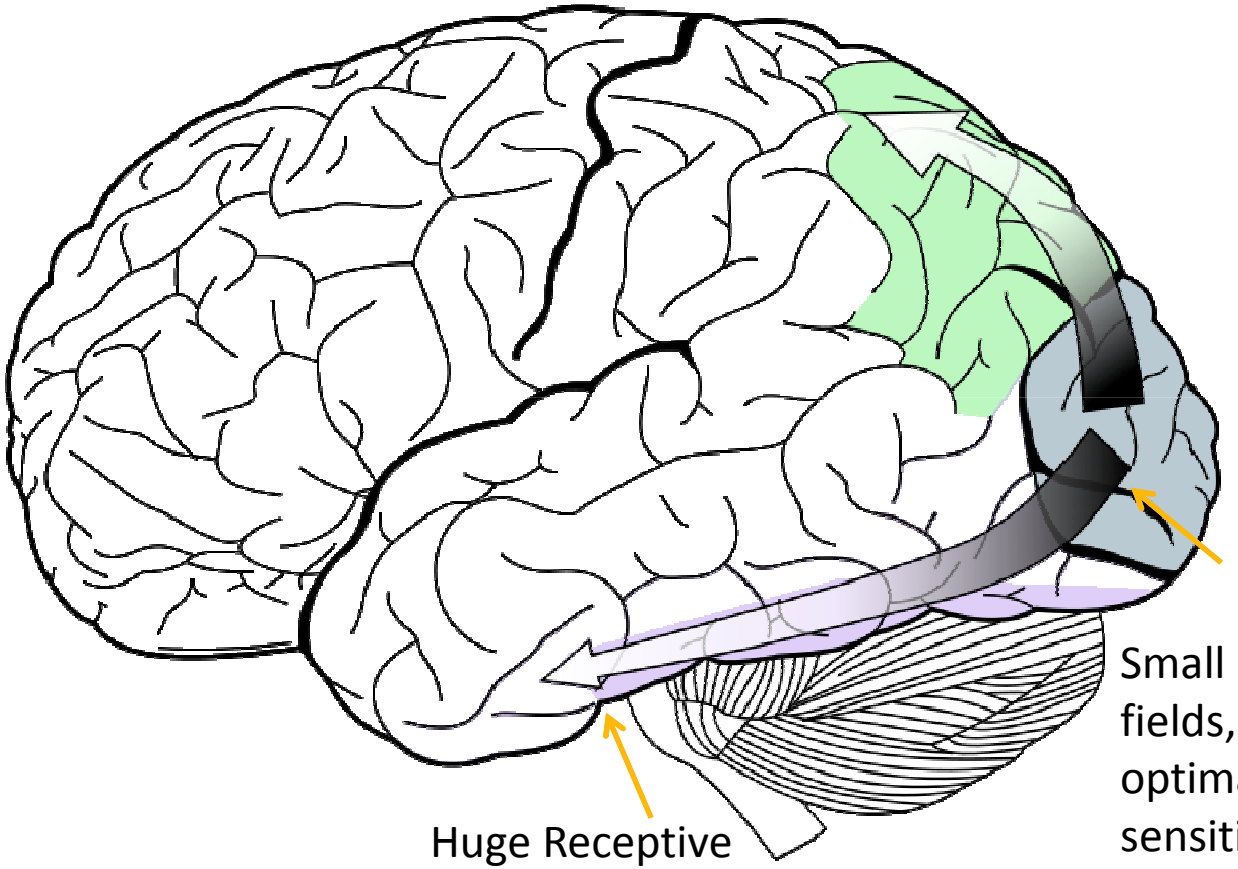
PAMI 2007

Presented by Tal Golan

Motivation:

- Humans and primates outperform the best machine vision systems.
- The goal – building a system that emulates object recognition in the cortex.

Computational principles of the ventral stream of visual cortex

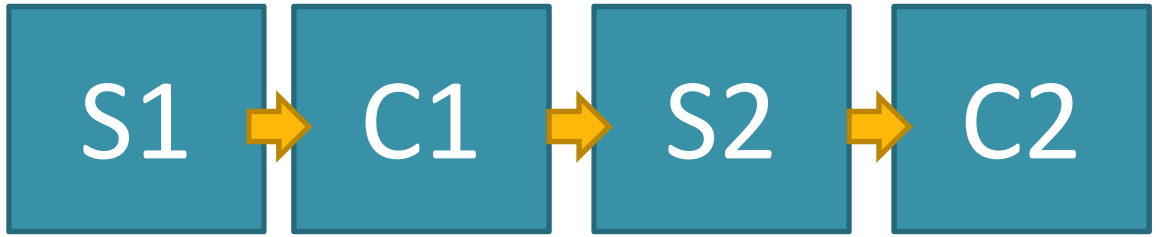


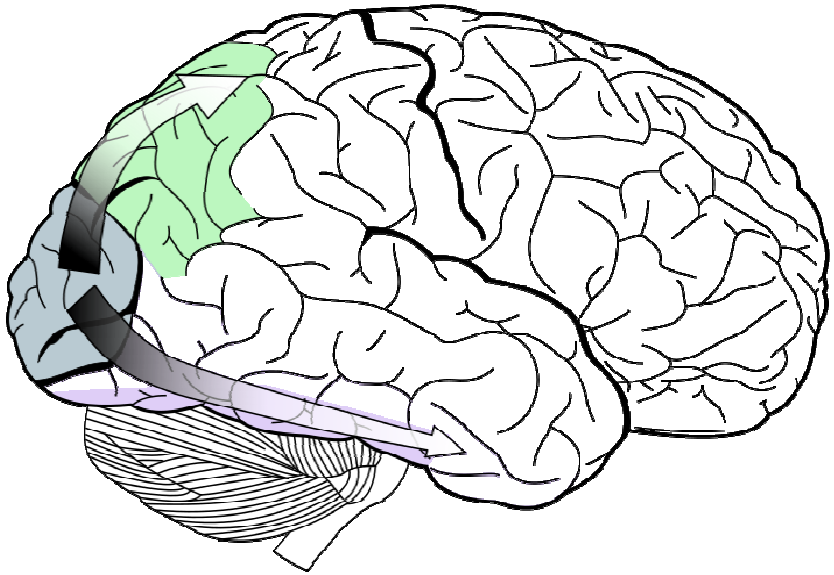
Small Receptive fields, simple optimal stimuli, sensitivity to transformations

Huge Receptive fields, complex optimal stimuli, invariance to transformations

Image: Wikimedia

The Model

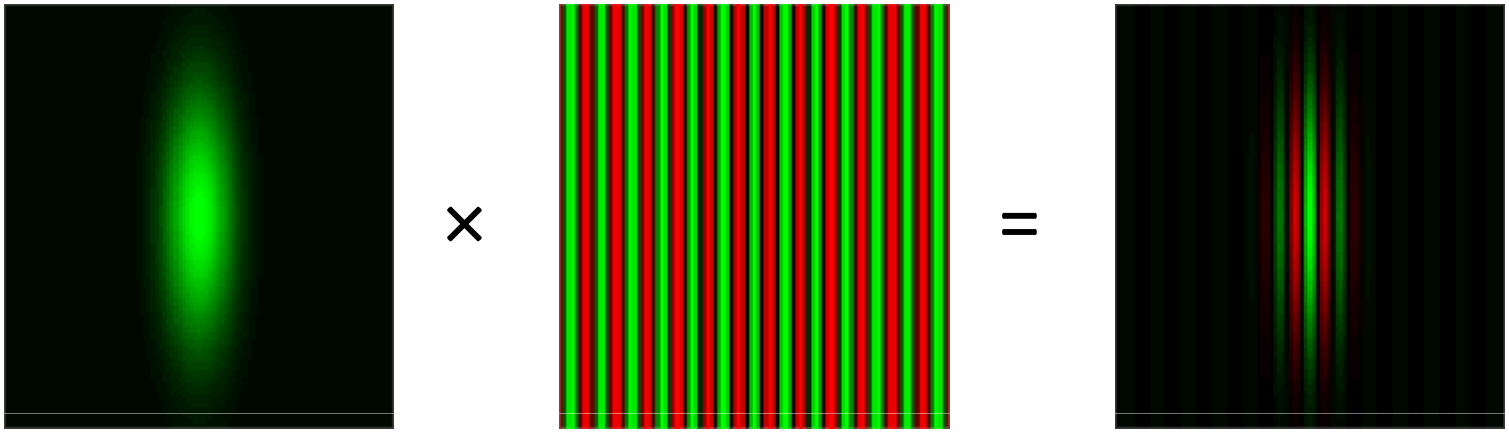




S1 layer

- Corresponds to the simple cells of the primary visual cortex (V1).
-  Hubel & Wiesel
- Gabor filters are used to model their receptive fields.

2D Gabor filter



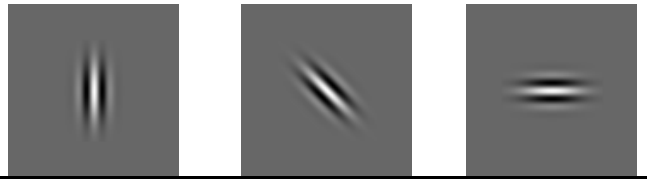
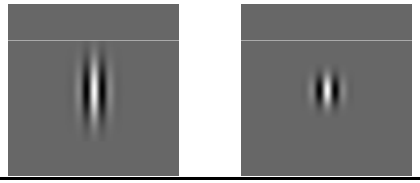
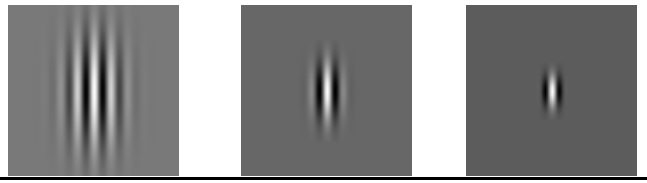
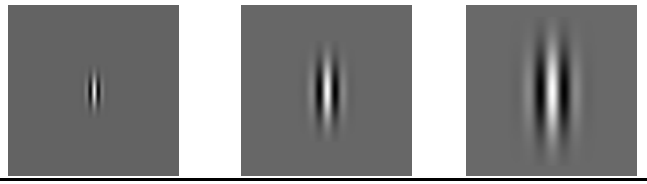
2D Gaussian × Cosine grating = Gabor filter

$$F(x, y) = \exp\left(\frac{-\left(x_0^2 + \gamma^2 y_0^2\right)}{2\sigma^2}\right) \times \cos\left(\frac{2\pi}{\lambda} x_0\right)$$

$$\begin{pmatrix} x_0 \\ y_0 \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix}$$

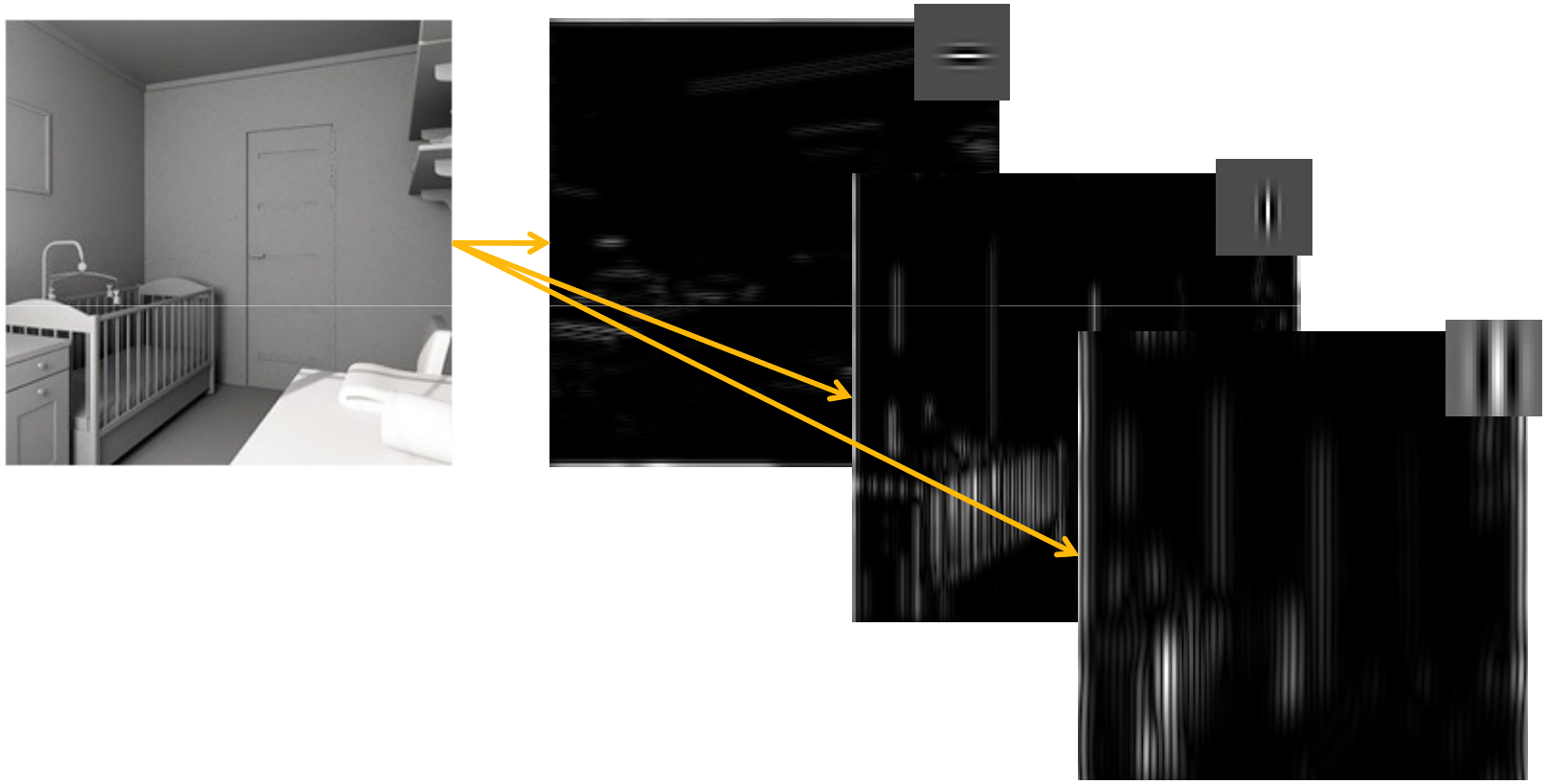
Gabor filter parameters

$$F(x,y)=\exp\left(\frac{-\left(x_0^2+\gamma^2 y_0^2\right)}{2\sigma^2}\right)\times\cos\left(\frac{2\pi}{\lambda}x_0\right) \qquad \begin{pmatrix}x_0\\y_0\end{pmatrix}=\begin{pmatrix}\cos\theta & \sin\theta\\-\sin\theta & \cos\theta\end{pmatrix}\begin{pmatrix}x\\y\end{pmatrix}$$

Θ	Orientation	
γ	Spatial Ratio	
σ	Gaussian SD	
λ	Wavelength*	

Examples taken from http://matlabserver.cs.rug.nl/

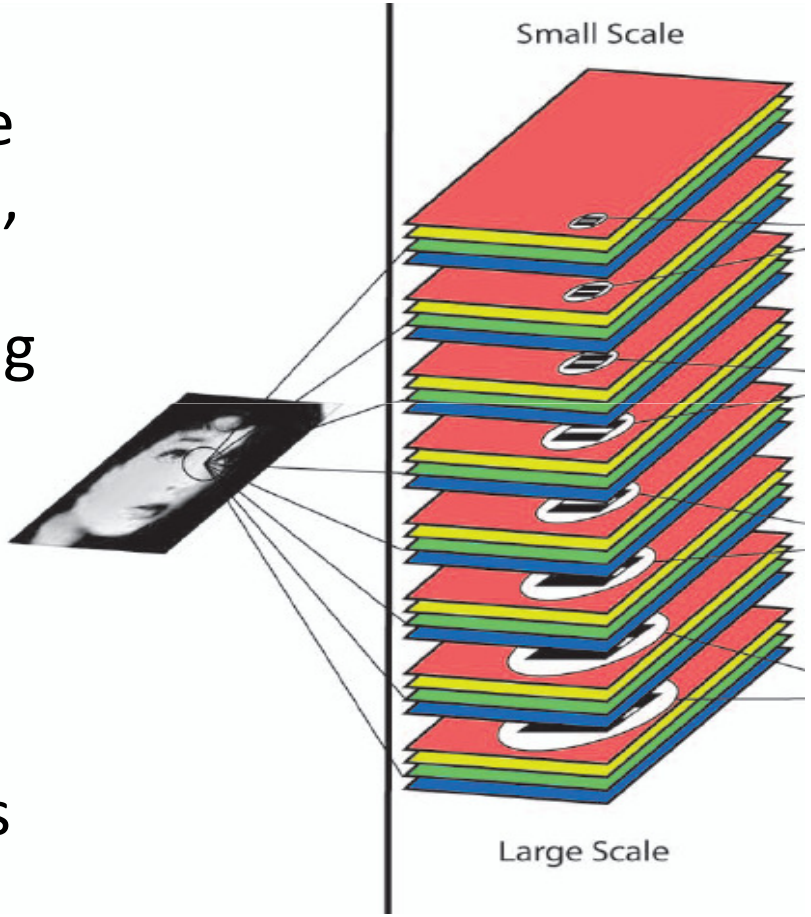
Effect of Gabor filter on Natural Images



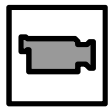
Examples taken from
http://matlabserver.cs.rug.nl/

S1 layer

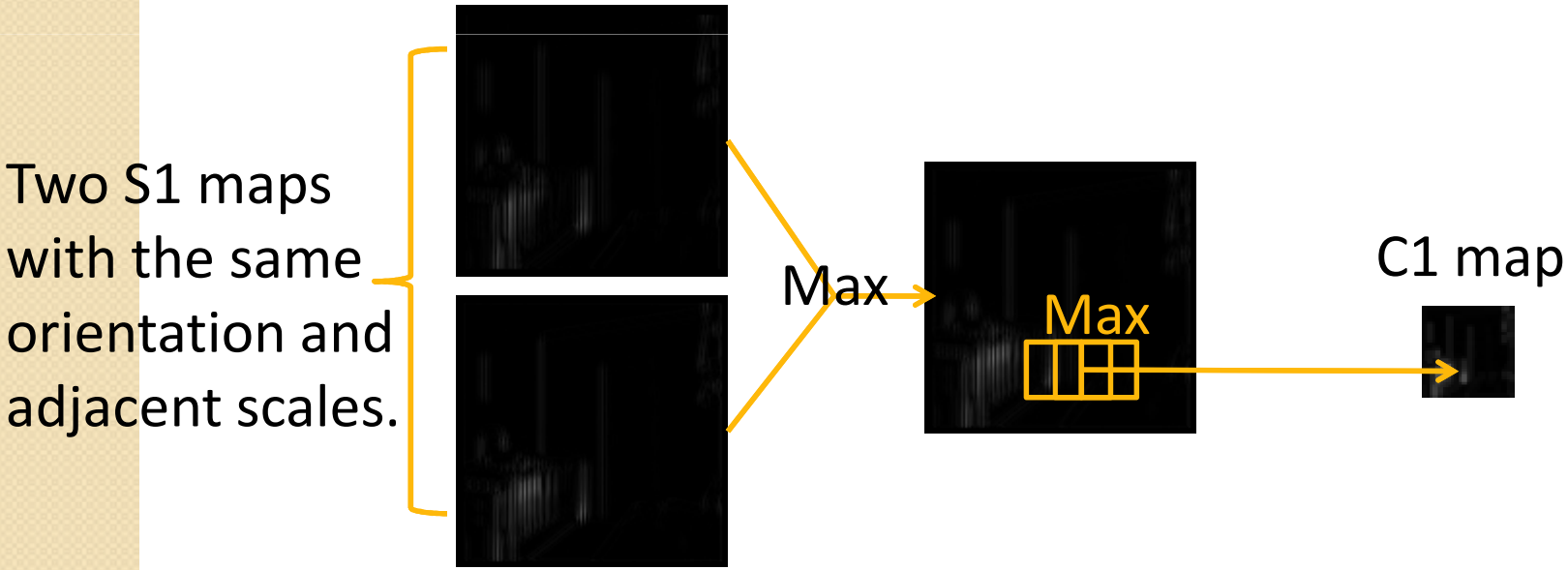
- A battery of filters is applied on the grayscale image. 4 orientations (0, 45, 90 & 135) and 16 scales are used, resulting in 64 different maps.
- The distribution of the filters' parameters is adjusted to match the distribution of parameters of monkey's parafoveal V1 simple cells.



C1 Layer

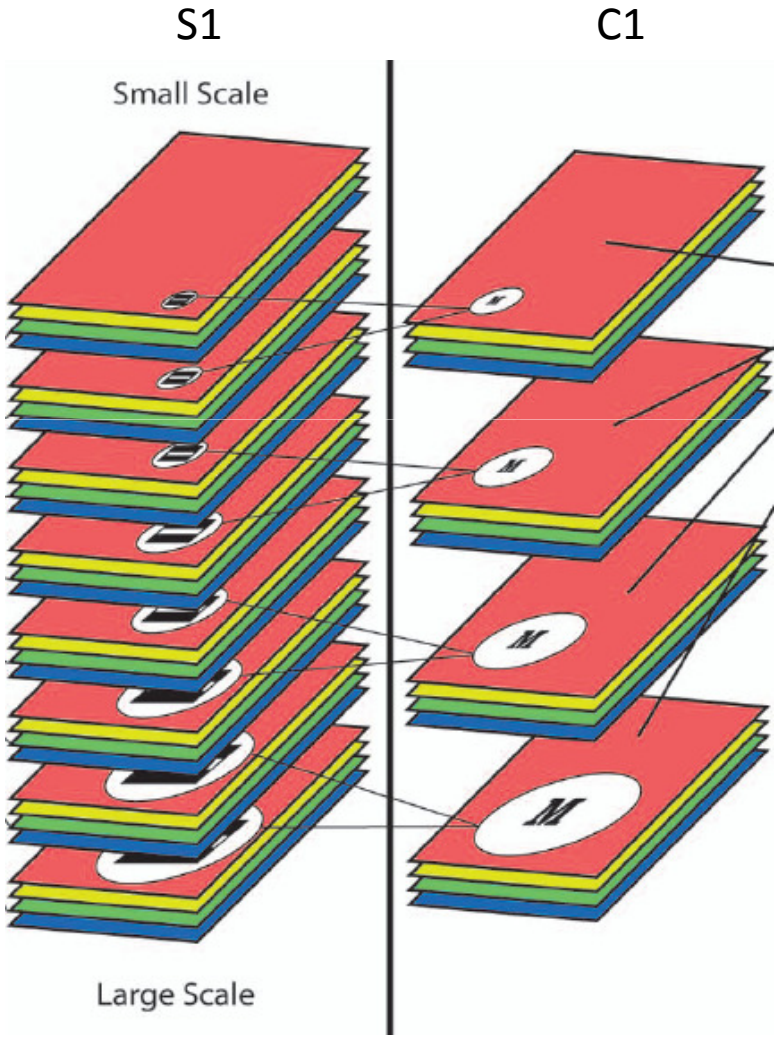


- Corresponds for complex cortical cells. These cells exhibit some tolerance to size and position shifts.



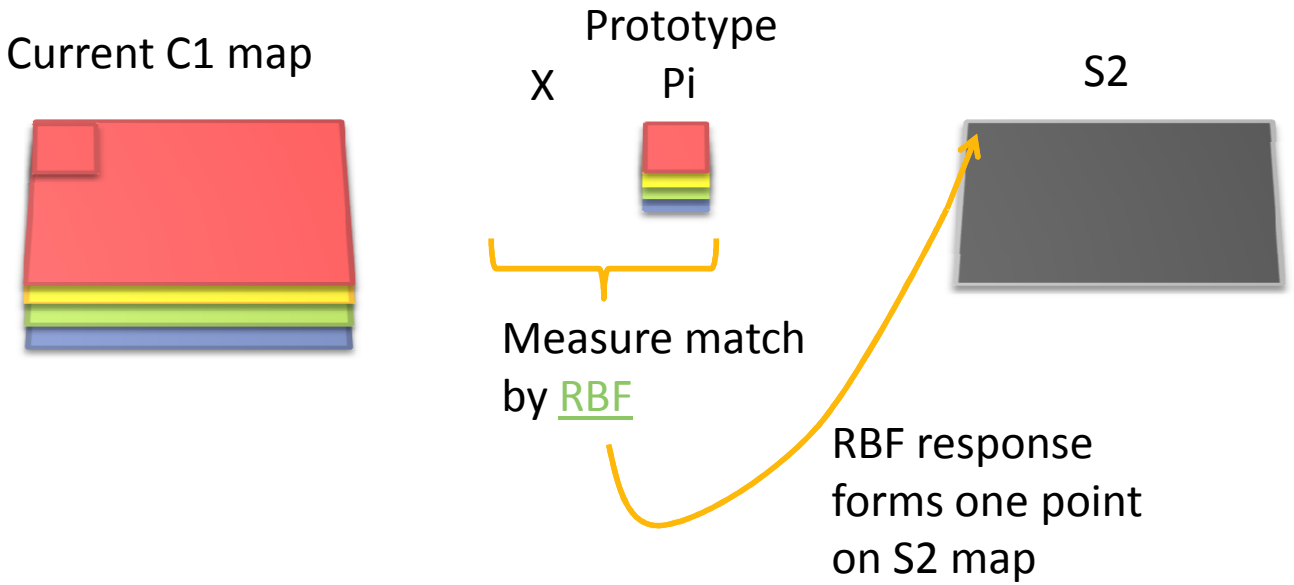
C1 Layer

- 8 Scales bands (pairs of S1 scales) are pooled. With 4 orientation per each bend, we get 32 maps.
- Parameters are again fitted to receptive fields of monkey's complex cells.



S2 Layer

- Uses N prototypes - previously learnt image patches.
- For each scale band, each prototype Pi is compared to all crops of the current image.



Radial Basis Function (RBF)

$$r = \exp\left(-\beta \left\| X - P_i \right\|^2\right)$$

X is current image in C1 format, in a specific scale band and position.
Pi is previously learnt patch in C1 format.
β is tuning parameter.

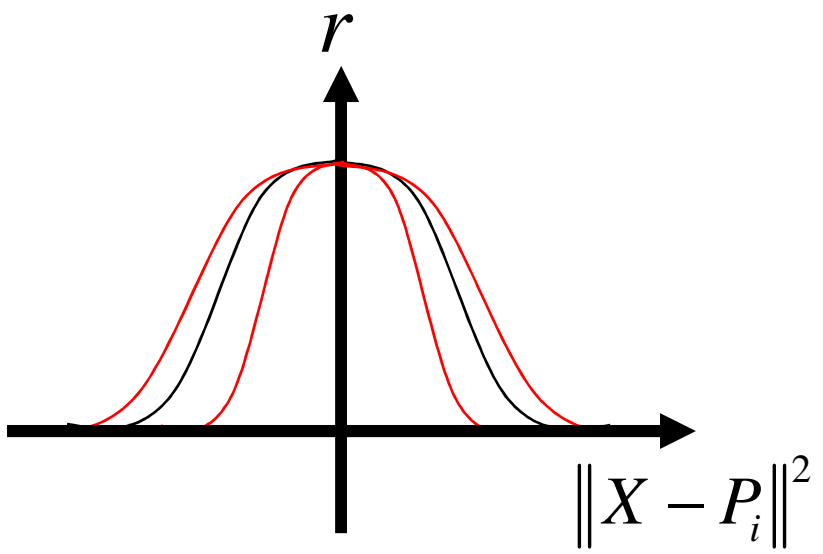
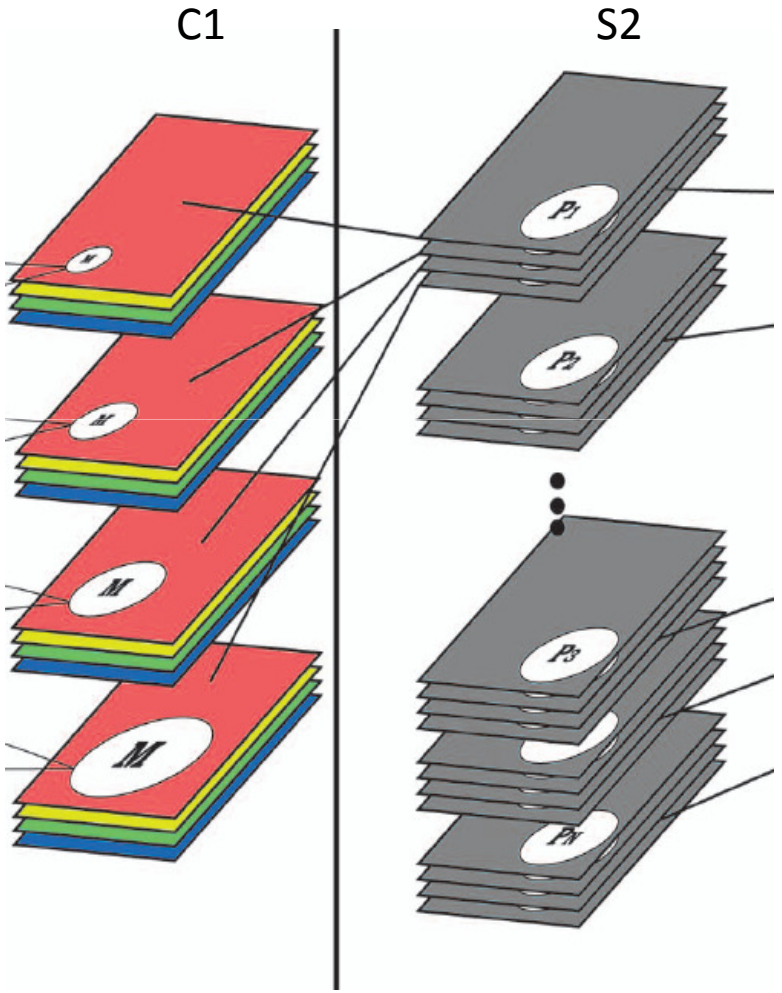


Figure is adapted from Michael Fink's neural computation course

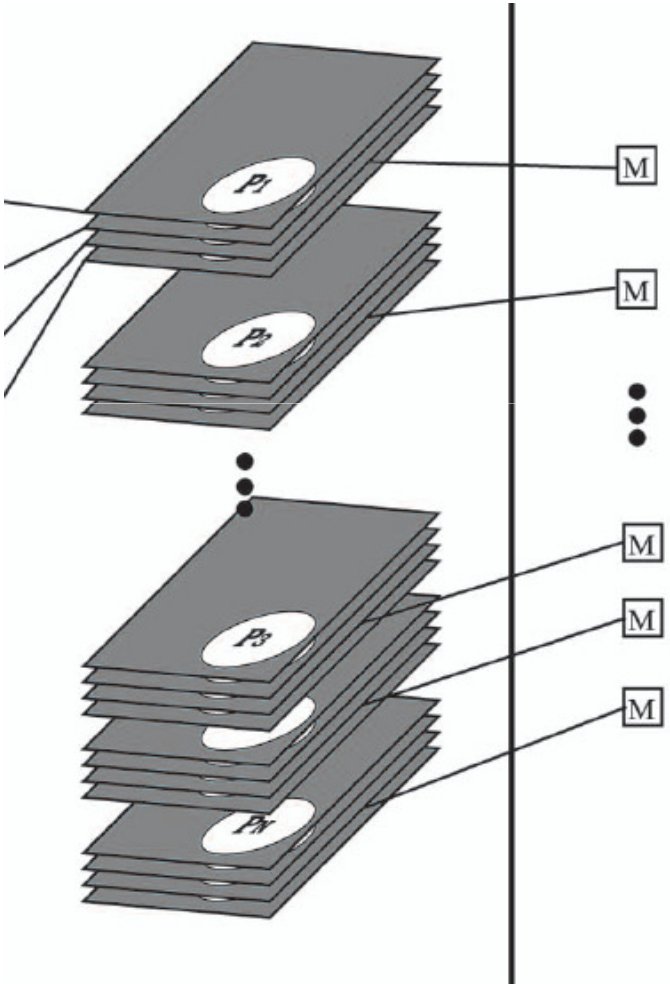
S2 Layer

- For N prototypes, 8N S2 maps are produced.

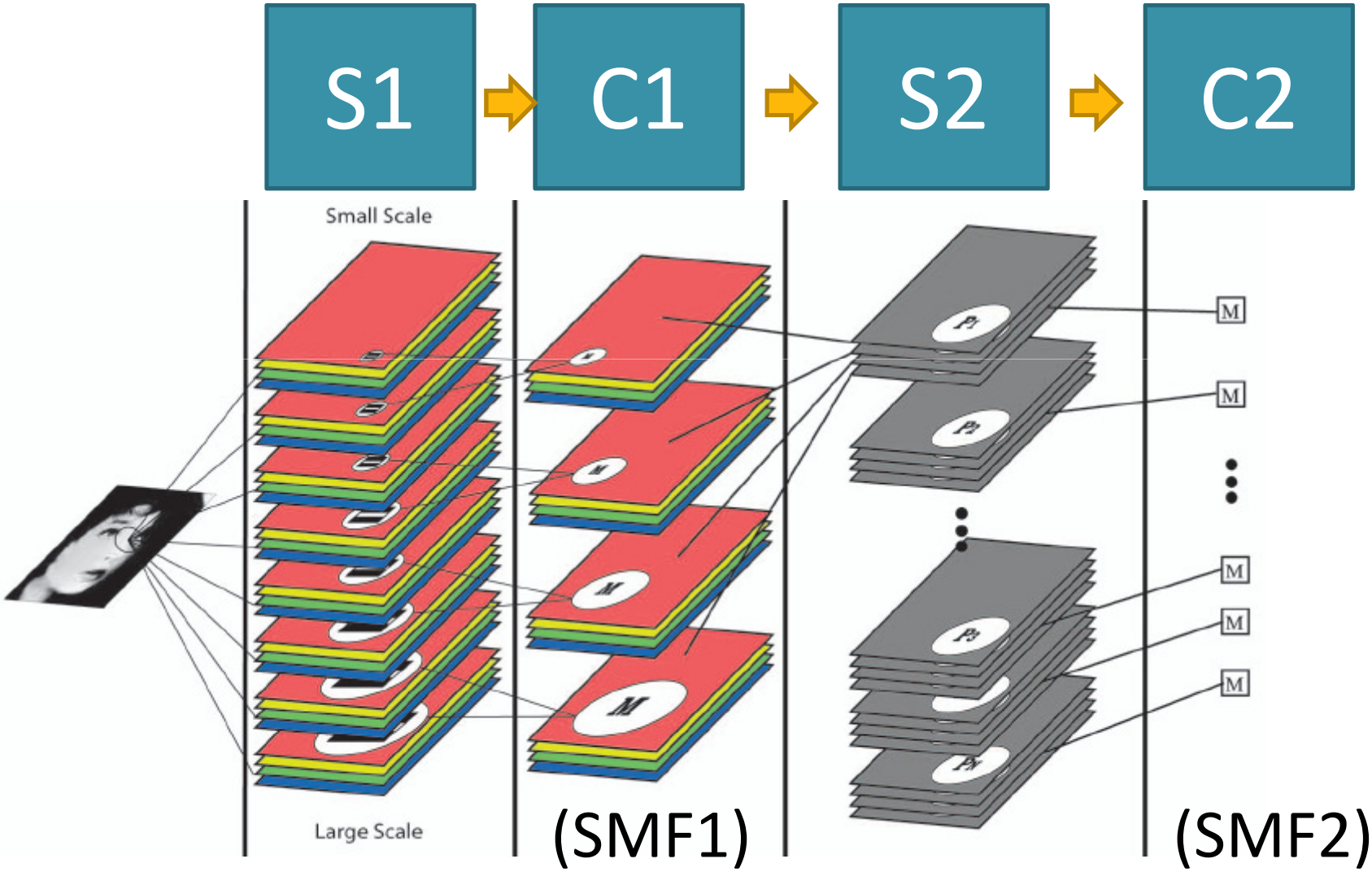


C2 Layer

- For each prototype Pi, maximum value is taken from the entire S2 lattice.
- For N previously learnt patches, C2 is a N-tuple.



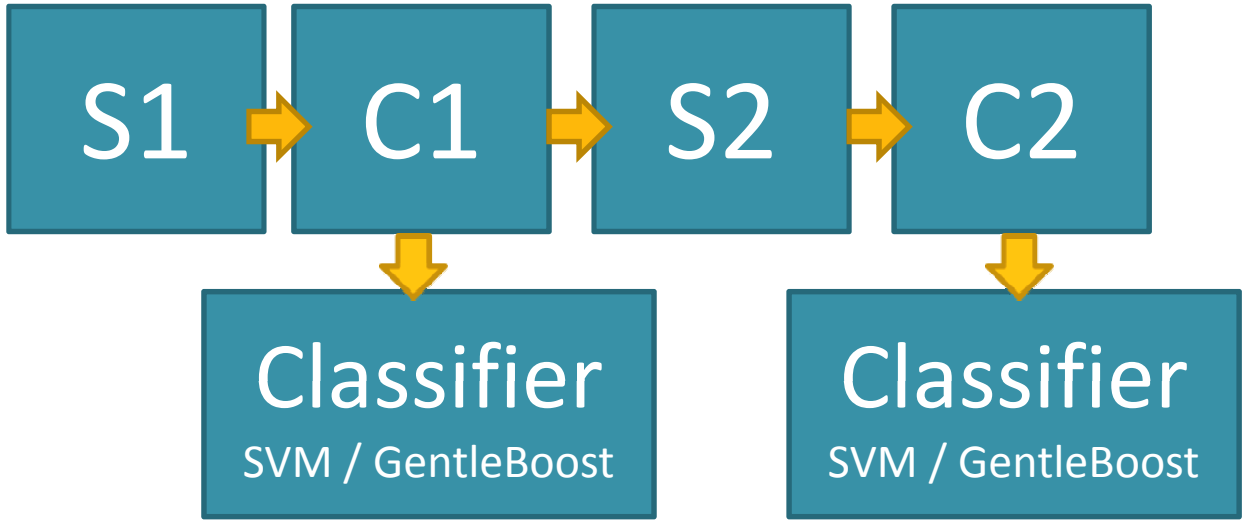
Overview



Prototype selection

- Prototypes can be sampled from the positive training set (weakly supervised learning) or from a random set of natural images (unsupervised learning).
- Image patches are extracted at random positions and sizes and stored in C1 format.

Classification



Learning → Training → Classification

Empirical Evaluation - Object Recognition In Clutter

Datasets	Benchmark	C_2 features	
		boost	SVM
Leaves [19]	84.0	**97.0**	95.9
Cars [20]	84.8	99.7	**99.8**
Faces [20]	96.4	**98.2**	98.1
Airplanes [20]	94.0	**96.7**	94.9
Motorcycles [20]	95.0	**98.0**	97.4
Faces [17]	90.4	**95.9**	95.3
Cars [18]	75.4	**95.1**	93.3

Constellation models by Perona et al.

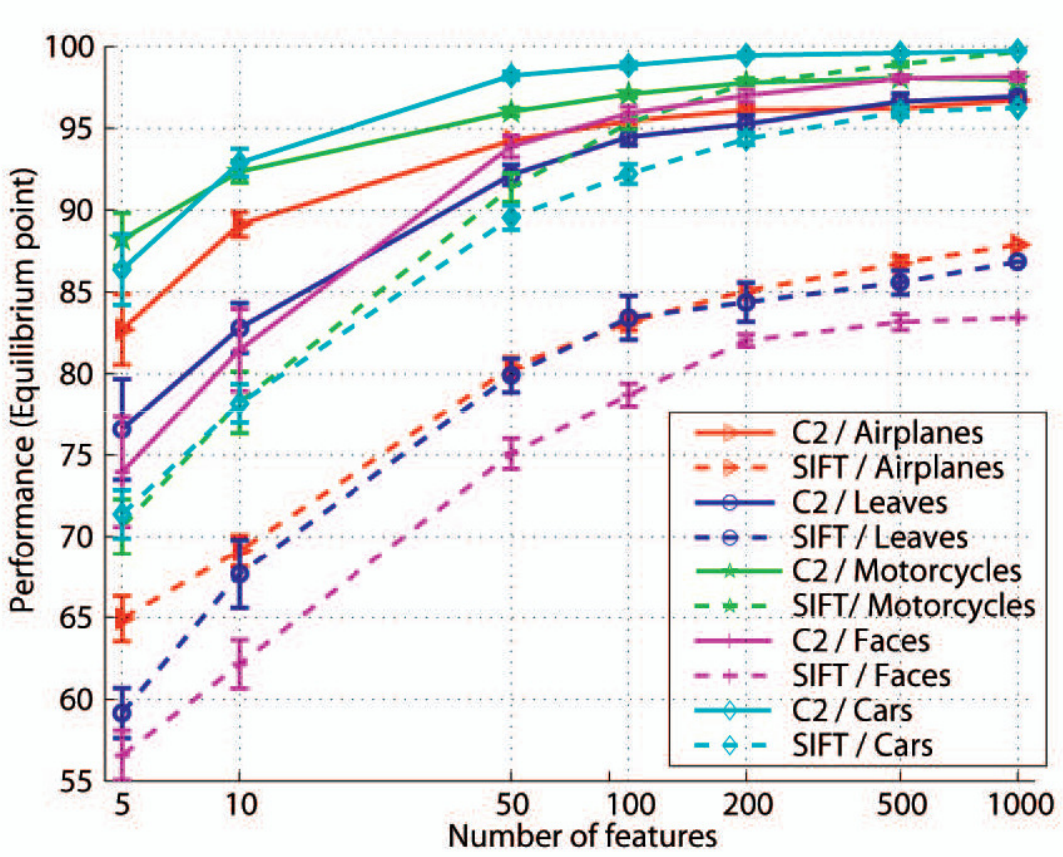
Hierarchical SVM-based face detection by Heisele et al.

Ullman et al.'s fragments

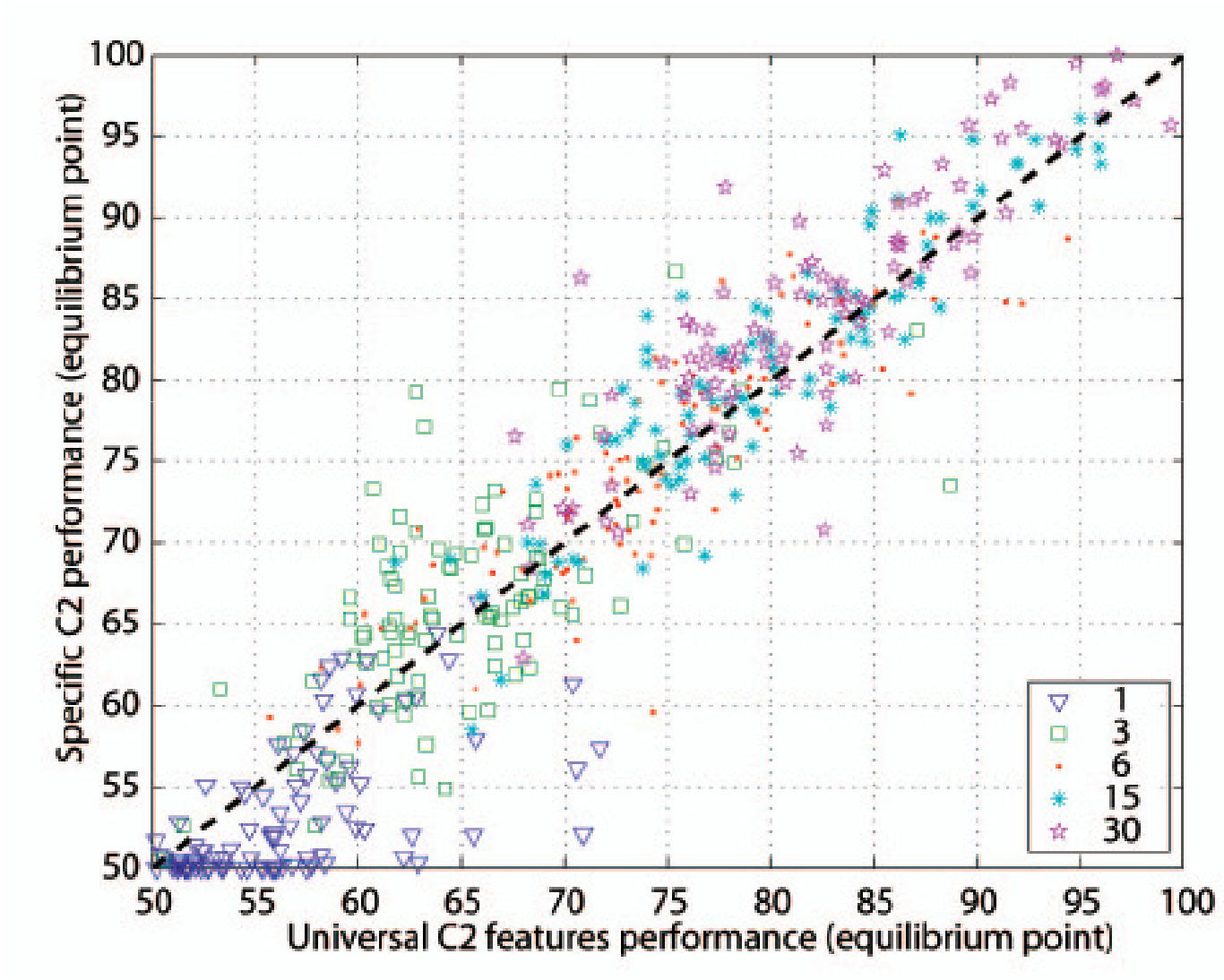
Comparison with SIFT

- N reference key-points were sample from the training dataset.
- Given a new image, the minimum distance between all its key-points and the N reference key-points thus obtaining an N-tuple feature vector.
- Only SIFT descriptors used, no position information.

Comparison with SIFT

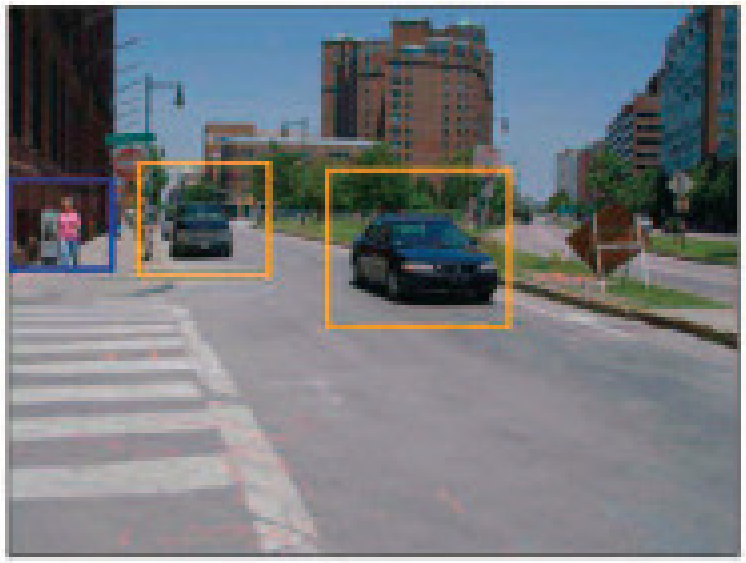


Using universal features



Empirical Evaluation – Objects Recognition without Clutter

- Car, pedestrian and bicycles detection using sliding window. C1 and C2 SMF's were tested.
- C1 SMF's are better than all the benchmarks at car and bicycle recognition. Histogram of Gradients is better on pedestrians.



Benchmarks:

- Gray scale template matching
- Local Patch Correlation
- Leibe et al.'s part-based system
- Histogram of Gradients.

Discussion

- Shortcomings
- Strengths
- What cognitive function does the model model?