Biologically Inspired Model for Object Recognition

Xiaobai Chen

COS 598B

The “Evolution”

- Fukushima, Biol. Cybernetics 80
  - Early attempts with neural network to mimic hierarchical model of Hubel & Wiesel 65’
- Et al. Poggio, Nature Neuroscience 99
  - **Max** Vs. **Sum** pooling
- Et al. Poggio, PAMI 07
  - State-of-art neural network
  - Extensive experiments

Neocognitron (Fukushima)

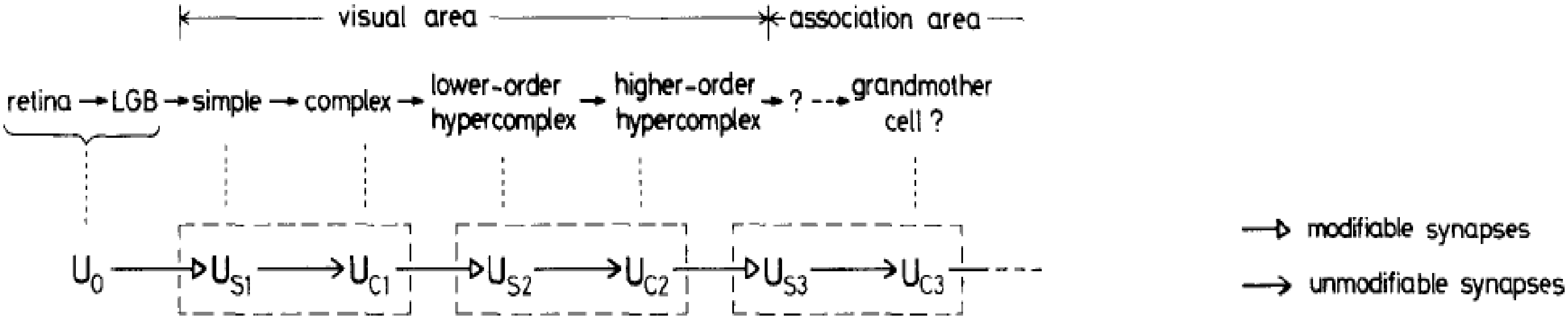


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

- Hierarchical structure
- From “simple” to “complex”

Simple & Complex

- Along the hierarchy, two functional stages are interleaved:
  - □ **Simple** (*S*) units build an increasingly complex and specific representation by combining the response of several subunits with different selectivity
  - □ **Complex** (*C*) units build an increasingly invariant representation (to position and scale) by combing the response of several subunits with the same selectivity but at slightly different position and scales

Neocognitron (Fukushima)

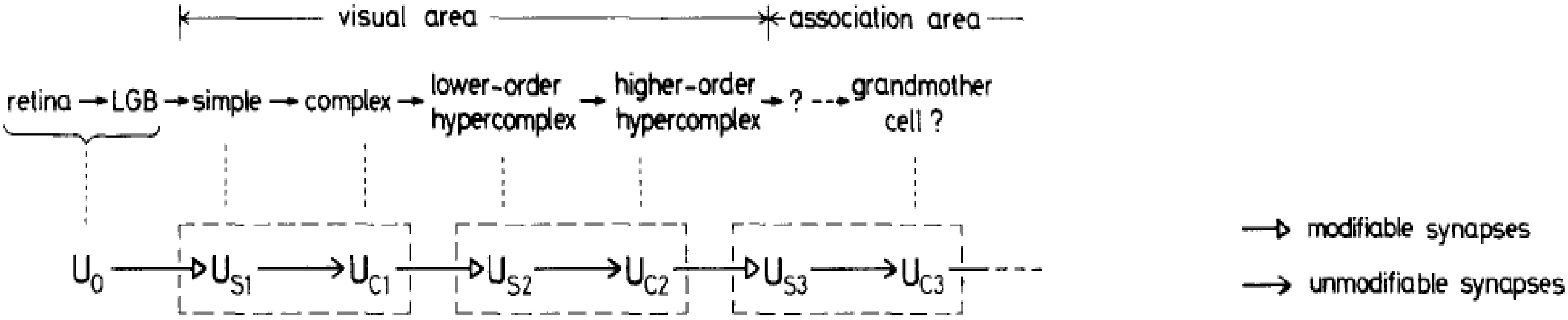


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

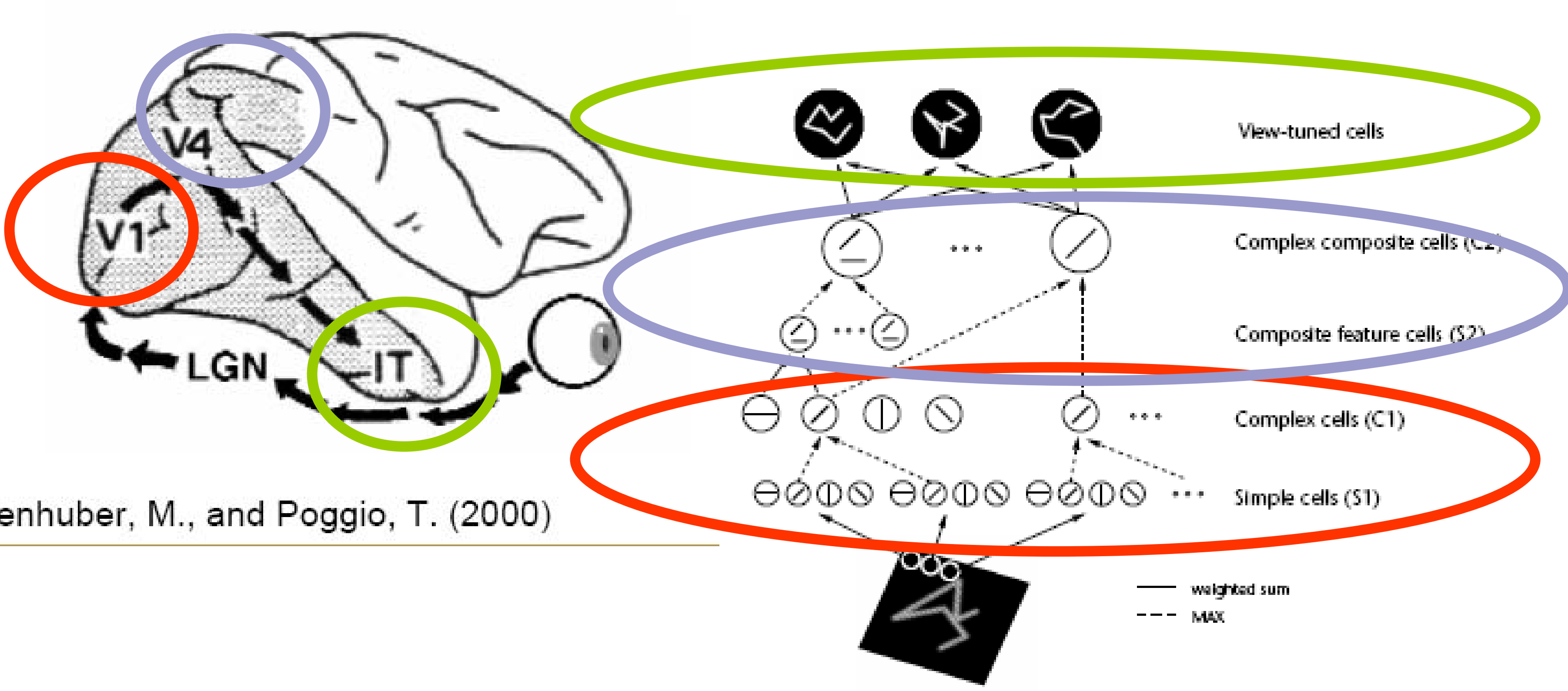
- Hierarchical structure
- From “simple” to “complex”
- Increase invariance

The basic idea and goals persists till now!

Then…What has evolved?

- How much more we know about human visual system?
- For neural network model
  - How to connect each layer?
  - What is the computing model of each layer?
  - How many layers?

The hierarchy based on the brain model

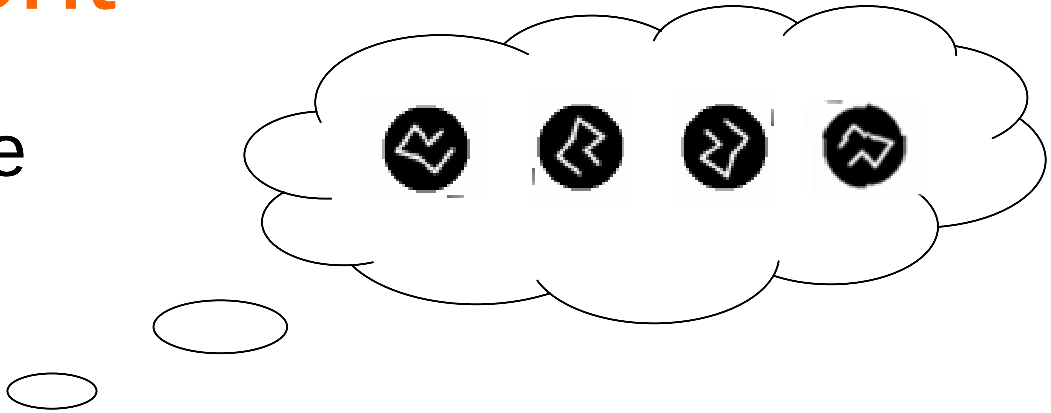


Riesenhuber, M., and Poggio, T. (2000)

Et al. Poggio 1999.

Experiment

Training stage

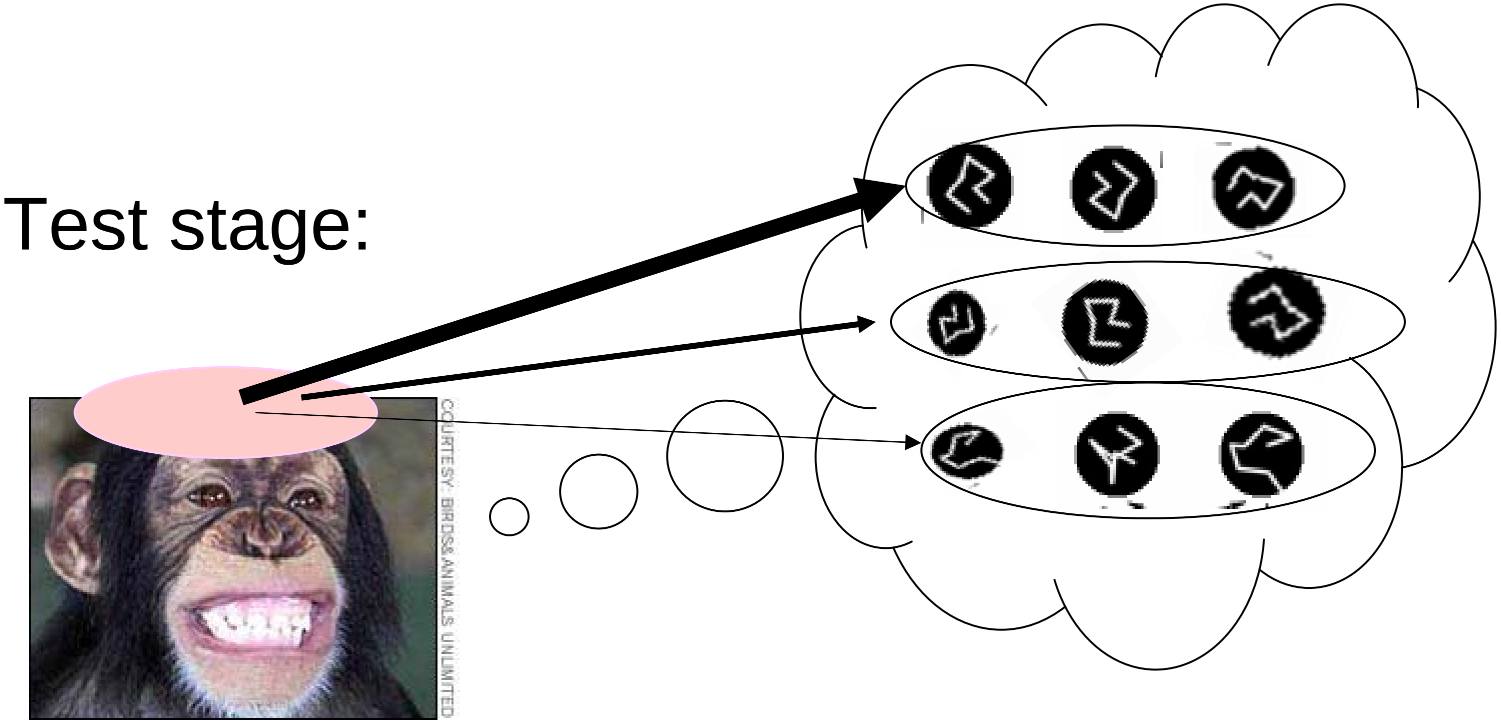


COURTESY: BIRDS&ANIMALS UNLIMITED

The monkey was trained to recognize restricted set of views of unfamiliar target stimuli resembling paperclips. They check which IT cell responds best to all views. The cell that responded the most was picked for the study.

Credit to Tomer Livne and Maria Zeldin

Test stage:



The best reaction of the cell was to the trained data.
The second best was to new transformations of the trained object.
And very little response to new objects (distractors)

Credit to Tomer Livne and Maria Zeldin

Quantitative results

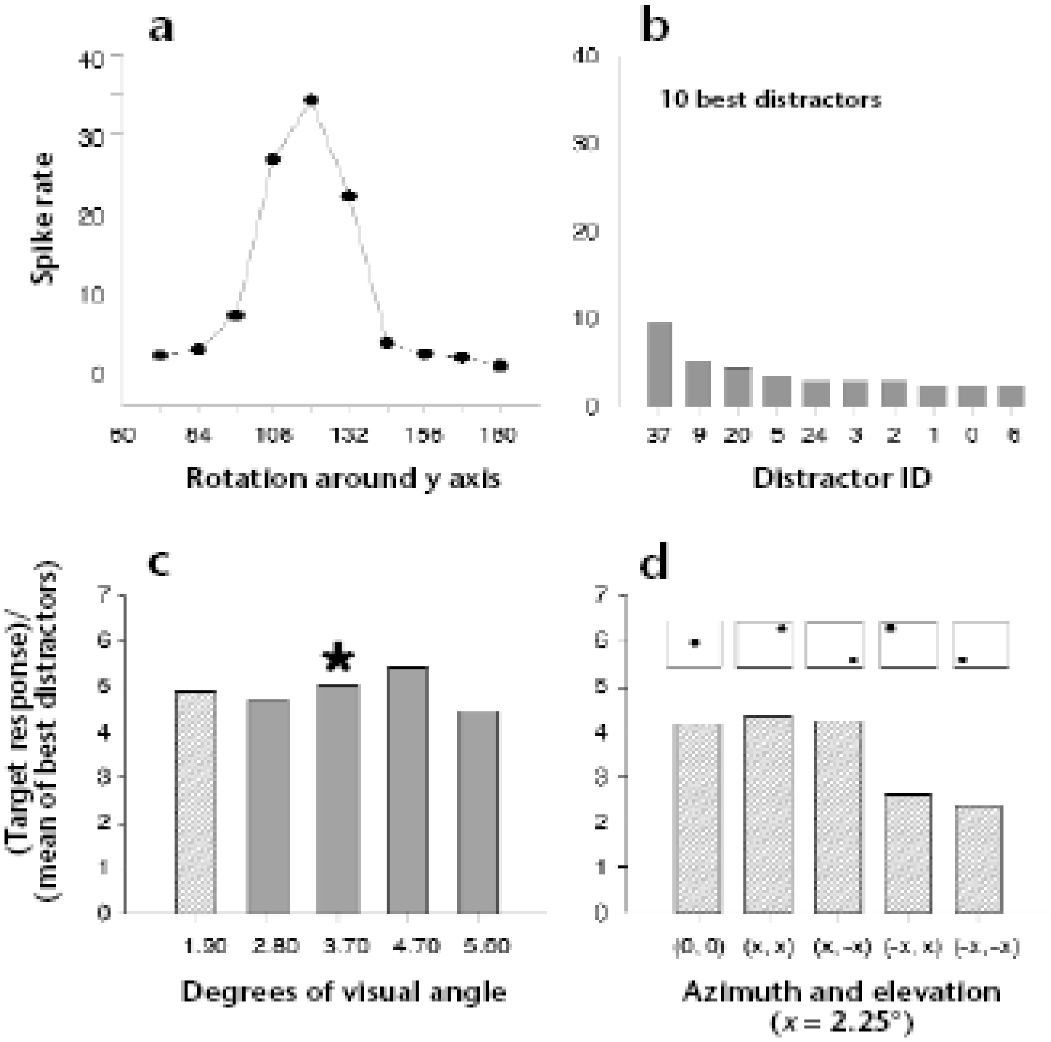


Fig. 1. Invariance properties of one neuron (modified from Logothetis *et al.*[21]). The figure shows the response of a single cell found in anterior IT after training the monkey to recognize paperclip-like objects. The cell responded selectively to one view of a paperclip and showed limited invariance around the training view to rotation in depth, along with significant invariance to translation and size changes, even though the monkey had only seen the stimulus at one position and scale during training. (a) Response of the cell to rotation in depth around the preferred view. (b) Cell's response to the ten distractor objects (other paperclips) that evoked the strongest responses. The lower plots (c, d) show the cell's response to changes in stimulus size (asterisk shows the size of the training view) and position (using the 1.9° size), respectively, relative to the mean of the ten best distractors. Defining 'invariance' as yielding a higher response to transformed views of the preferred stimulus than to distractor objects, neurons showed an average rotation invariance of 42° (during training, stimuli were actually rotated by ±15° in depth to provide full 3D information to the monkey; therefore, the invariance obtained from a single view is probably smaller), translation and scale invariance on the order of ±2° and ±1 octave around the training view, respectively (J. Pauls, personal communication).

Hierarchical models of object recognition in cortex. Reisenhuber and Poggio. Nature America Inc, november 1999.

Invariance

- All kinds of invariance
  - Translation
  - Scale
  - Rotation
- How to add invariance in the NN model?
  - Pooling (Perrett & Oram 93')

Pooling

- Pooling over afferents tuned to various transformed versions of the same stimuli
- Two idealized mechanism
  - Linear – Sum
    - Suitable to increase complexity
  - Nonlinear – Max
    - Selectivity
- Which is a better fit for the complex cell to achieve invariance?

Argu(1): position invariance

- Both lead to position invariance
- Sum
  - Specificity is lost
  - Case-by-case parameter adjustments in clutter
- Max
  - signal the best match of any part of the stimulus to the afferents’ preferred feature
  - More robust in clutter

Argu(2): size invariance

- Sum
  - More afferents will be excited if the same object increases size;
  - hence excitation of the cell will increase
- Max
  - Cell response is determined by the best-matching afferent
  - Not influenced much by more afferents

Argu(3):neurophysiological data

- An IT neuron’s response seems to be dominated by the stimulus producing a higher firing rate
- Theoretical investigation on V1 also supports a MAX-like pooling mechanism (Sakai & Tanaka 97)

Argu(4):Experiment

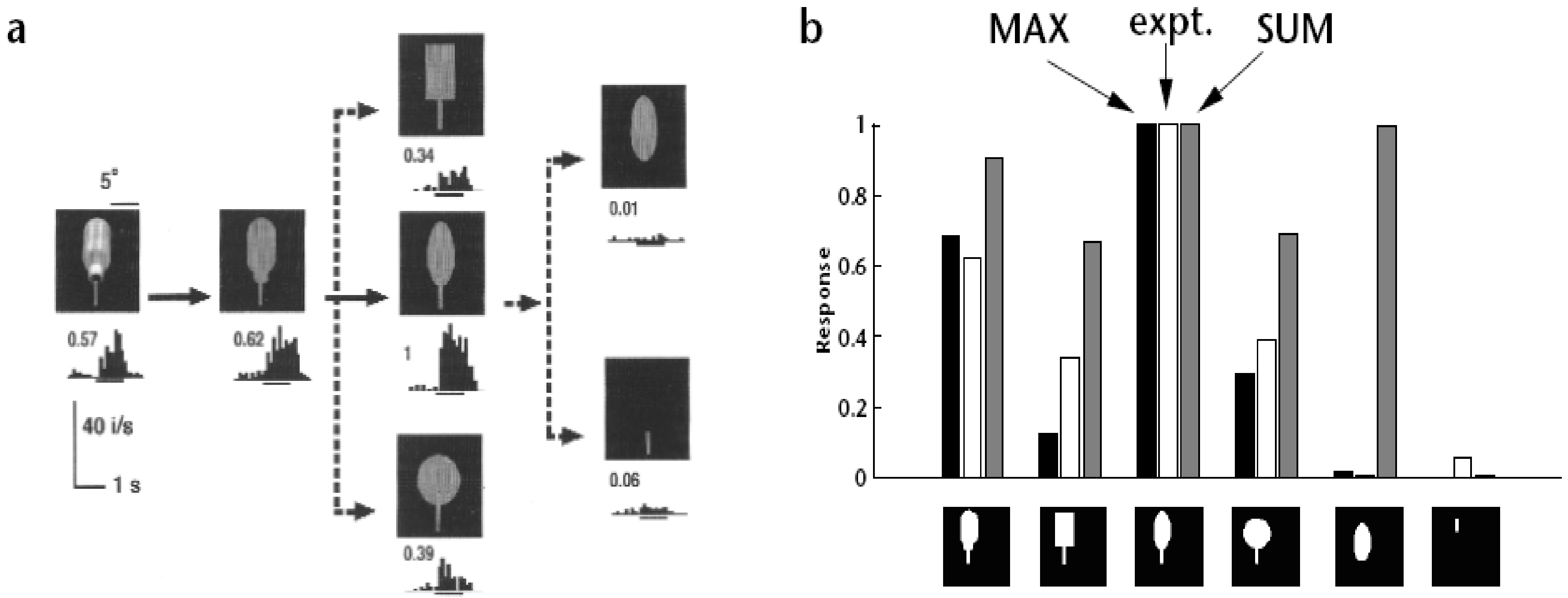
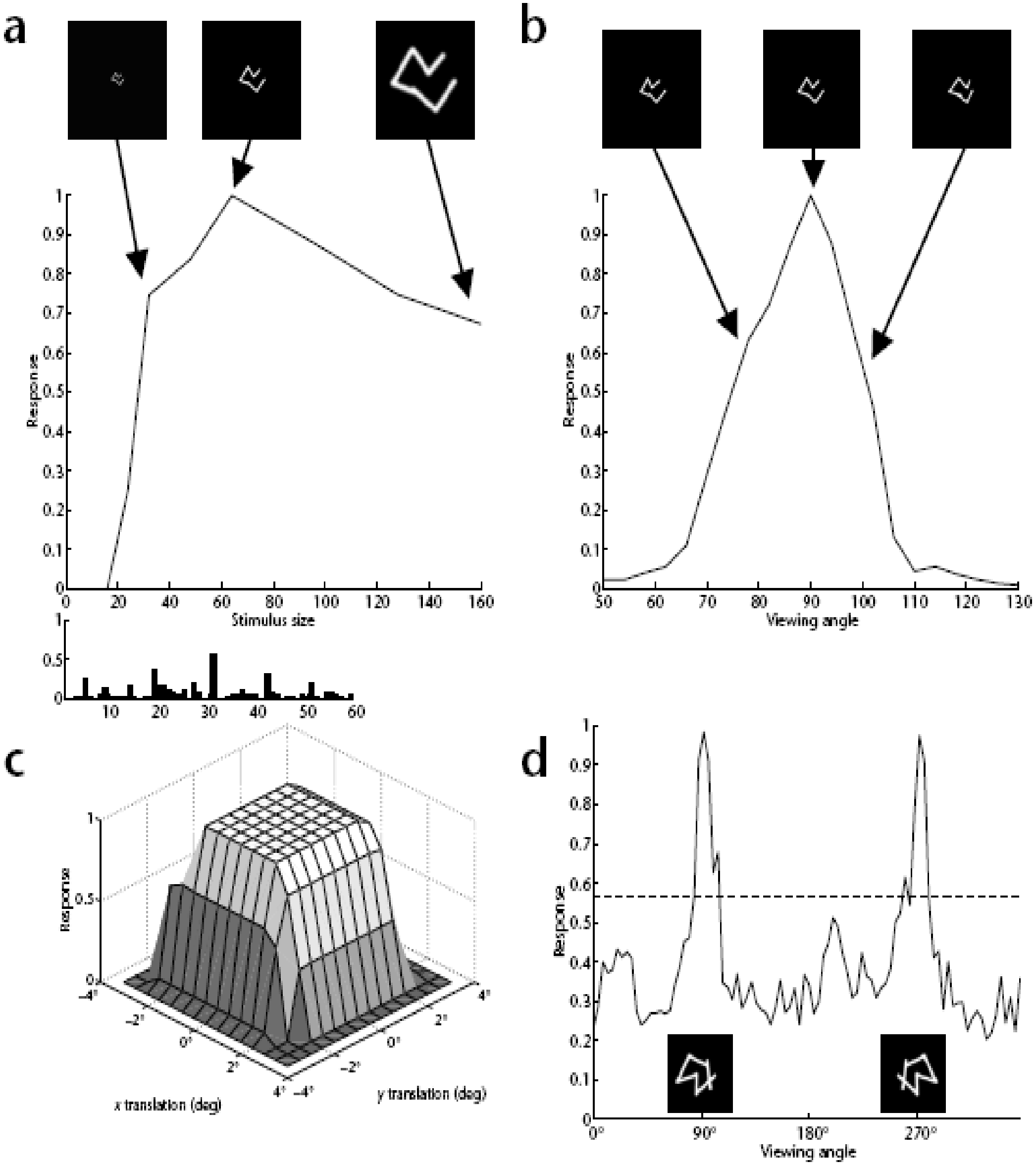


Fig. 3. Highly nonlinear shape-tuning properties of the MAX mechanism. (**a**) Experimentally observed responses of IT cells obtained using a 'simplification procedure'[26] designed to determine 'optimal' features (responses normalized so that the response to the preferred stimulus is equal to 1). In that experiment, the cell originally responded quite strongly to the image of a 'water bottle' (leftmost object). The stimulus was then 'simplified' to its monochromatic outline, which increased the cell's firing, and further, to a paddle-like object consisting of a bar supporting an ellipse. Whereas this object evoked a strong response, the bar or the ellipse alone produced almost no response at all (figure used by permission). (**b**) Comparison of experiment and model. White bars show the responses of the experimental neuron from (**a**). Black and gray bars show the response of a model neuron tuned to the stem-ellipsoidal base transition of the preferred stimulus. The model neuron is at the top of a simplified version of the model shown in Fig. 2, where there were only two types of S1 features at each position in the receptive field, each tuned to the left or right side of the transition region, which fed into C1 units that pooled them using either a MAX function (black bars) or a SUM function (gray bars). The model neuron was connected to these C1 units so that its response was maximal when the experimental neuron's preferred stimulus was in its receptive field.

Argu(5):Simulation



Softmax approximation

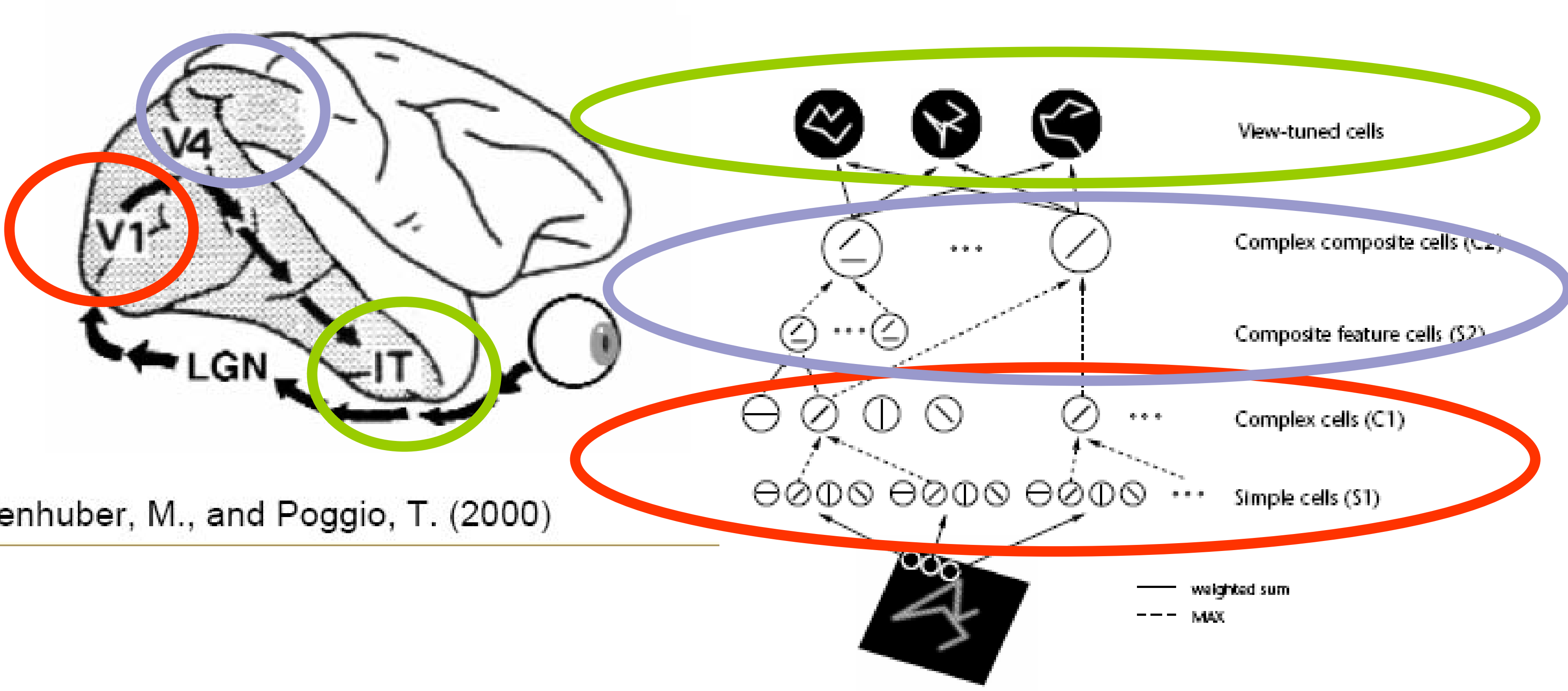
$$c_i^1 = \sum_j \frac{\exp(p \cdot |s_j|)}{\Sigma_k \exp(p \cdot |s_k|)} s_j,$$

- P=0, linear sum
- P->∞,MAX

Stop for a while…

- NN can really look like the visual pathway
- Alternative Max + Sum seems to work for a NN
- Recognition of different transformations of an object is similar to the problem of classification
- Use NN to learn features and do classification with linear classifiers
  (et al. Poggio 07)

The hierarchy based on the brain model



Riesenhuber, M., and Poggio, T. (2000)

Et al. Poggio 1999.

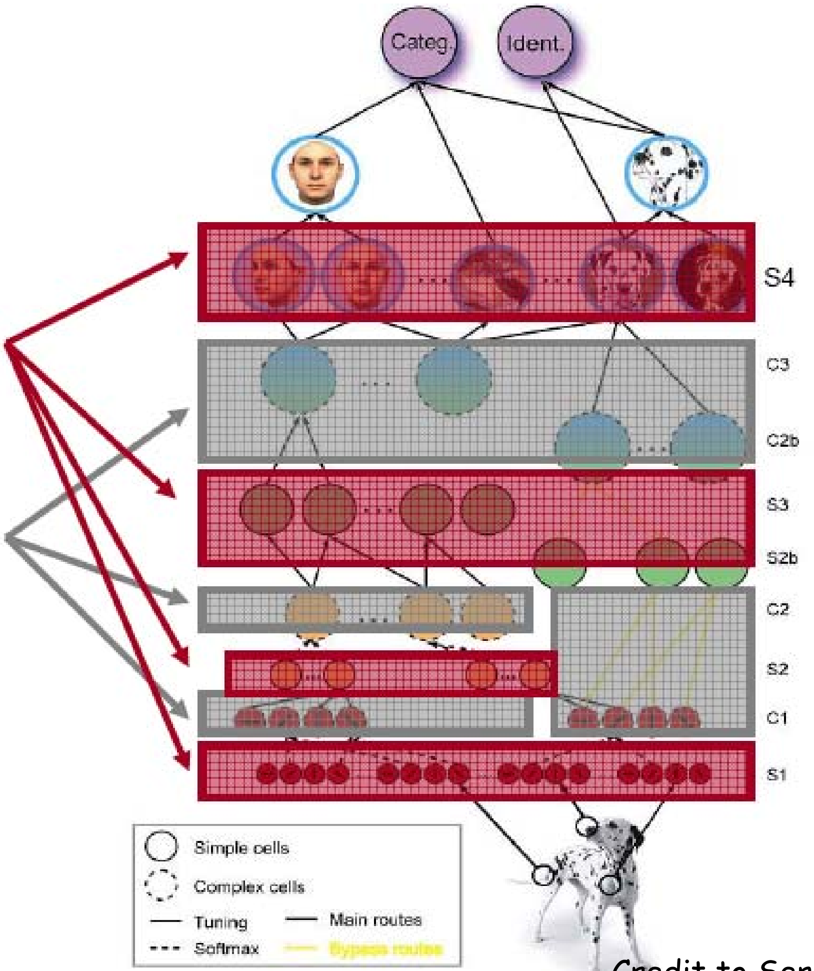
Implementation Details

- Along the hierarchy, from V1 to IT, two functional stages are interleaved:
  - **Simple** (*S*) units build an increasingly complex and specific representation by combining the response of several subunits with different selectivity with TUNING operation.
  - **Complex** (*C*) units build an increasingly invariant representation (to position and scale) by combing the response of several subunits with the same selectivity but at slightly different position and scales with a MAX-like operation.

Implementation Details

(1) **Selectivity (AND-like):** Gaussian-like function for tuning and specificity

(2) **Tolerance (OR-like):** Maximum-like operation for invariance over positions and scales



Credit to Serre and Poggio

Implementation Details

- By interleaving these two operation, an increasingly complex and invariant representation is built.
- Two routes:
  - □ Main route
    - follows the hierarchy of cortical stages strictly.
  - □ Bypass route
    - skip some of the stages
    - Bypass routes may help provide richer vocabulary of shape-tuned units with different levels of complexity and invariance.

Credit to Serre and Poggio

Implementation Details

- *S₁* units:
  - □ Correspond to the classical simple cells of Hubel and Wiesel found in the primary visual cortex (V1)
  - □ *S₁* units take the form of Gabor functions

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

$$x_0 = x\cos\theta + y\sin\theta \quad \text{and} \quad y_0 = -x\sin\theta + y\cos\theta$$

The aspect ratio: γ

The orientation: θ

The effective width: σ

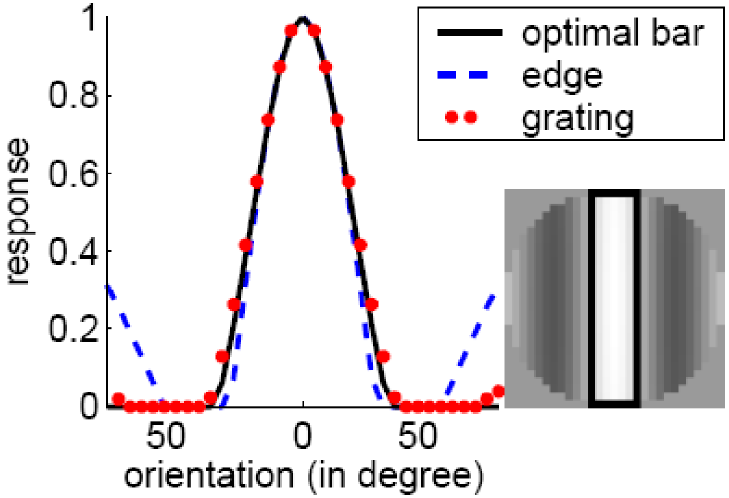
The wavelength: λ

3/31/2008

Credit to Serre and Poggio

Implementation Details

- □ Perform TUNING operation between the incoming pattern of input *x* and there weight vector *w*.
- □ The response of a S_1 unit is maximal when x matches w exactly.

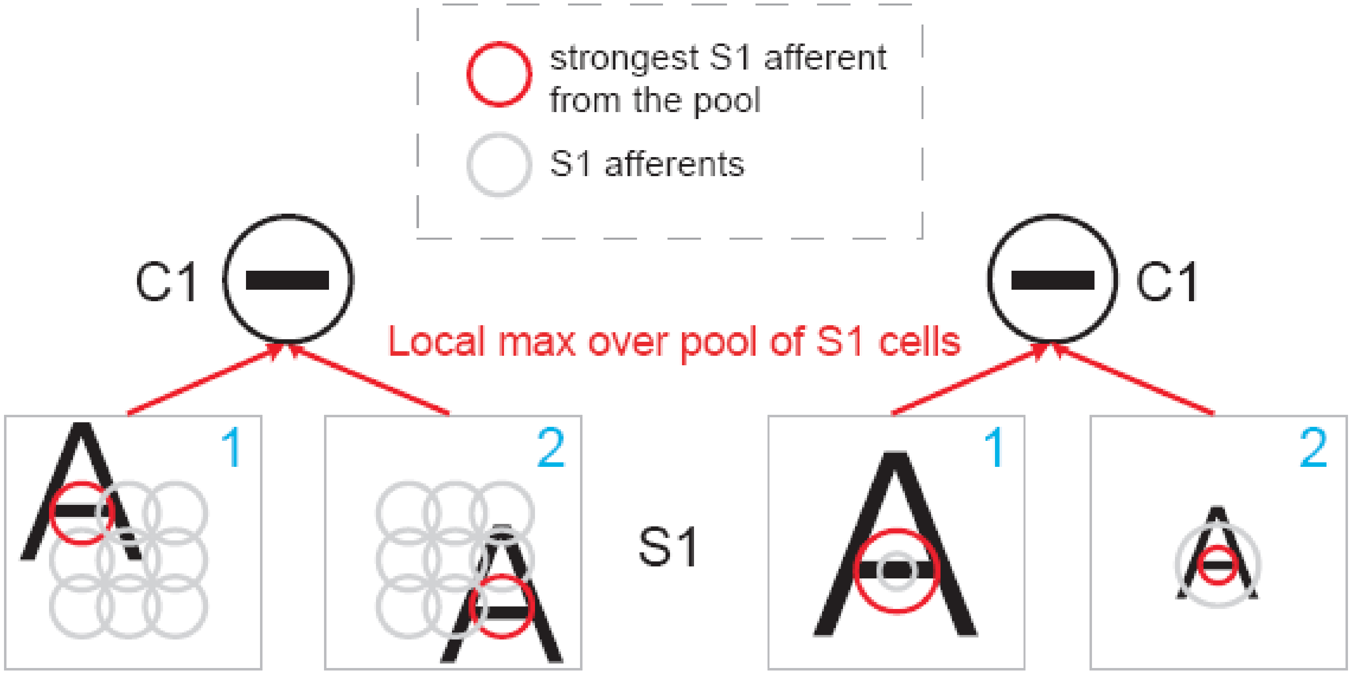


Credit to Serre and Poggio

Implementation Details

- C_1 units:
  - Corresponds to cortical complex cell which show some tolerance to shift and size.
  - Each of the complex C_1 unit receives the outputs of a group of simple S_1 units from the first layer with the same preferred orientation but at slightly different positions and sizes.
  - The operation by which the S_1 unit responses are combined at the C_1 level is a nonlinear MAX-like operation.

Implementation Details

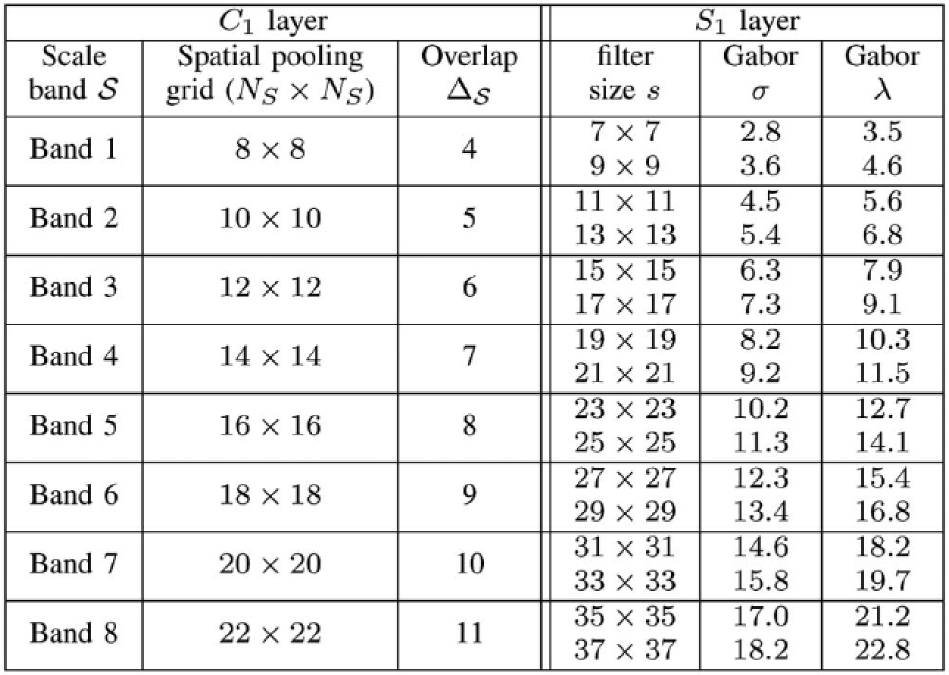

strongest S1 afferent from the pool
S1 afferents
C1
C1
Local max over pool of S1 cells
1
2
S1
1
2

(a) Tolerance to position

(b) Tolerance to scale

Implementation Details

- □ This process is done for each of the four orientations and each scale band independently.



C_1 layer			S_1 layer		
Scale band $\mathcal{S}$	Spatial pooling grid ($N_{\mathcal{S}} \times N_{\mathcal{S}}$)	Overlap $\Delta_{\mathcal{S}}$	filter size s	Gabor σ	Gabor λ
Band 1	8×8	4	7×7 9×9	2.8 3.6	3.5 4.6
Band 2	10×10	5	11×11 13×13	4.5 5.4	5.6 6.8
Band 3	12×12	6	15×15 17×17	6.3 7.3	7.9 9.1
Band 4	14×14	7	19×19 21×21	8.2 9.2	10.3 11.5
Band 5	16×16	8	23×23 25×25	10.2 11.3	12.7 14.1
Band 6	18×18	9	27×27 29×29	12.3 13.4	15.4 16.8
Band 7	20×20	10	31×31 33×33	14.6 15.8	18.2 19.7
Band 8	22×22	11	35×35 37×37	17.0 18.2	21.2 22.8

Implementation Details

- For instance
  - The first band: S=1.
    two S_1 maps: the one obtained using a filter of size 7x7 and 9x9.
  - For each orientation, the C_1 unit responses are computed by subsampling these maps using N_sxN_s=8x8.
  - One single measurement is obtained by taking the maximum of all 64 elements.
  - As a last stage, we take a max over the two scales from within the same spatial neighborhood.

Implementation Details

- S_2 unit:
  - □ A TURNING operation is taken over C_1 units at different preferred orientations to increase the complexity of the optimal stimulus.
  - □ S_2 level units becomes selective to more complex patterns – such as the combination of oriented bars to form contours or boundary-conformations.

Implementation Details

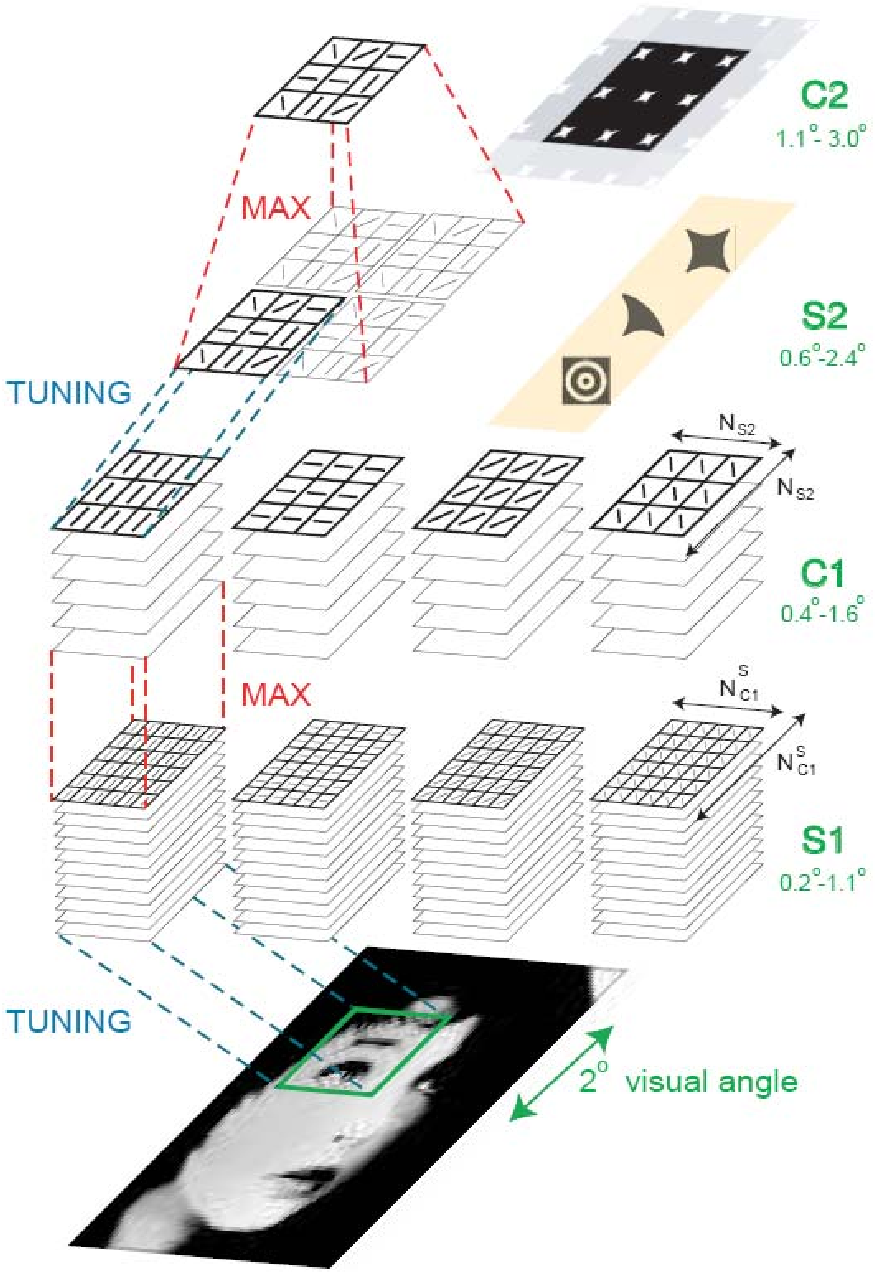
- Each S_2 units response depends in a Gaussian-way on the Euclidean distance between a new input and a stored prototype .

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

- P_i is one of the N features learned during training.
- patch X from the previous C_1 layer at a particular scale S

Implementation Details

- C_2
  - Our final set of shift- and scale-invariant C_2 responses is computed by taking a global maximum over all scales and position for each S_2 type over the entire S_2 lattice.
  - Units that are tuned to the same preferred stimulus but at slightly different positions and scales.


C2
1.1°- 3.0°
MAX
S2
0.6°-2.4°
TUNING
N_{S2}
N_{S2}
C1
0.4°-1.6°
MAX
N^S_{C1}
N^S_{C1}
S1
0.2°-1.1°
TUNING
2° visual angle

Implementation Details

- The learning stage
  - Corresponds to selecting a set of N prototypes $\mathbf{P}_i$ for the S_2 units.
- The classification stage
  - The C_1 and C_2 standard model features (SMF) are then extracted and further passed to a simple linear classifier.

Model Summary

- 4 Layers of processing
- 2 types of operations (Max, Sum)
- Output – N dimensional vector

Experiment

- Object Recognition in Clutter
- Object Recognition without Clutter
- Object Recognition of Texture-Based Objects
- Toward a Full System for Scene Understanding

Exp. with clutter

- Data Sets
  - Caltech5, Caltech101, MIT-CBCL
- Training and test images contains both targets and distractors

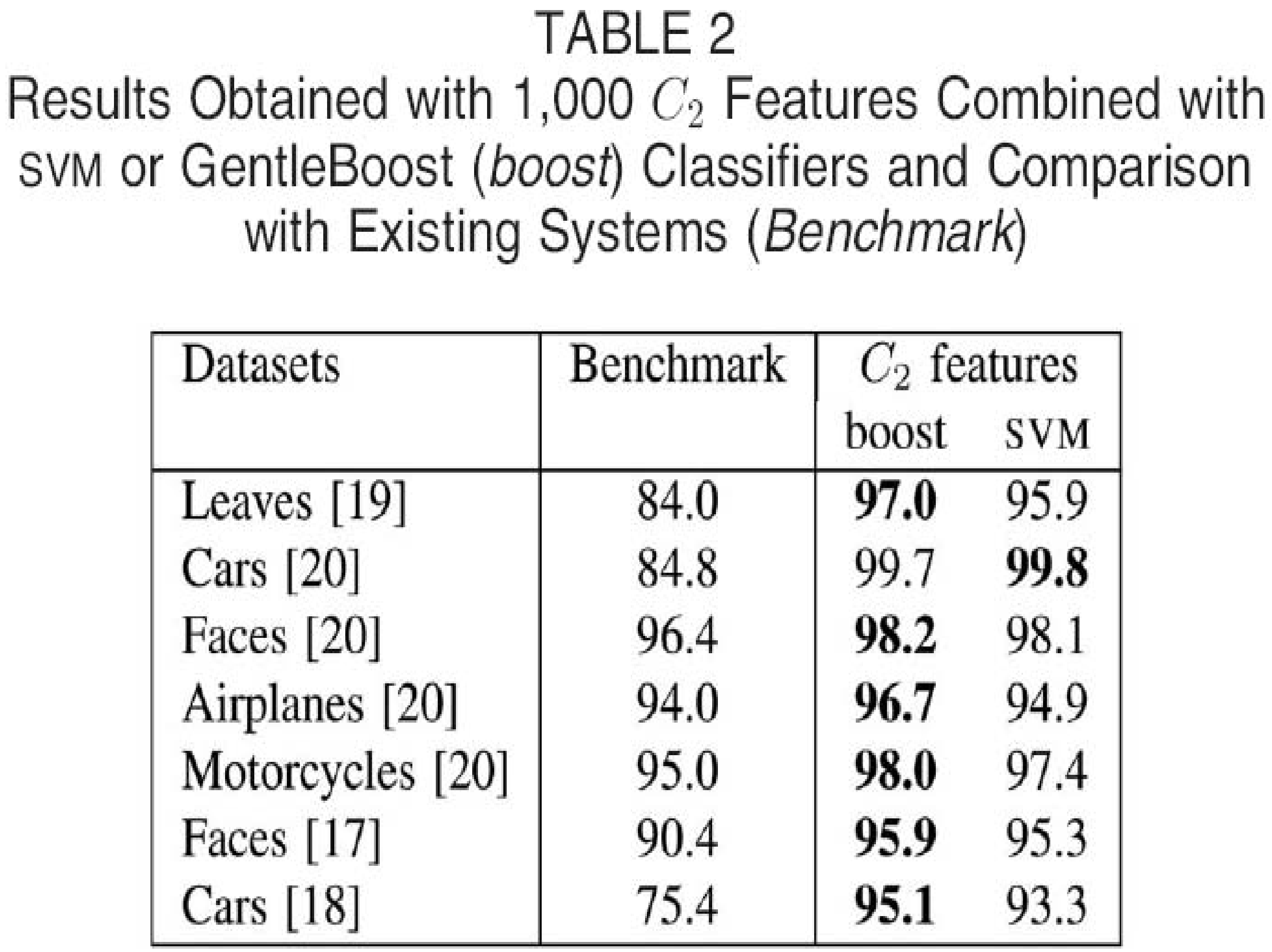


TABLE 2
Results Obtained with 1,000 C_2 Features Combined with SVM or GentleBoost (*boost*) Classifiers and Comparison with Existing Systems (*Benchmark*)

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

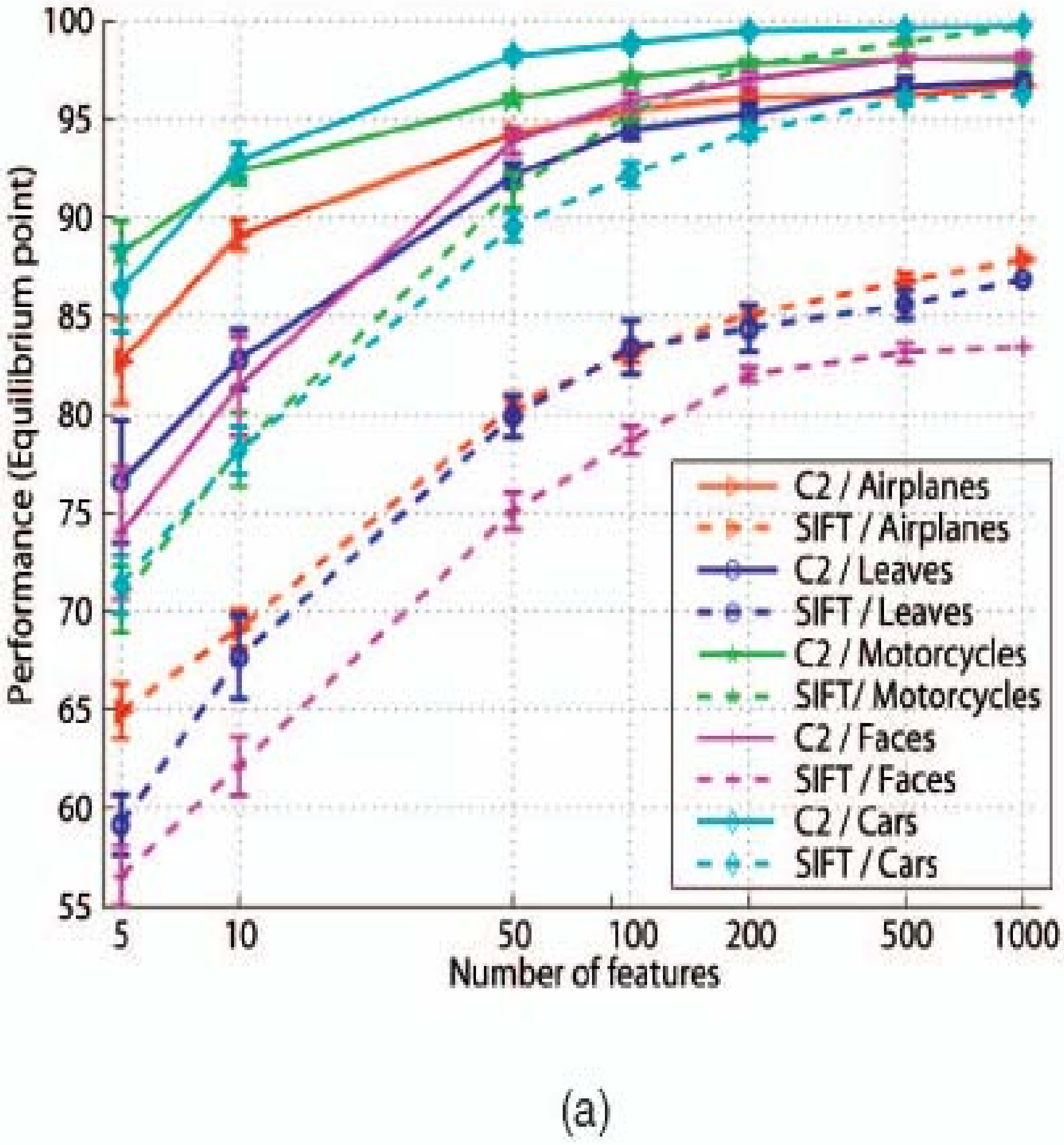


Fig. 3. Comparison between the SIFT and the C_2 features on the *CalTech5*

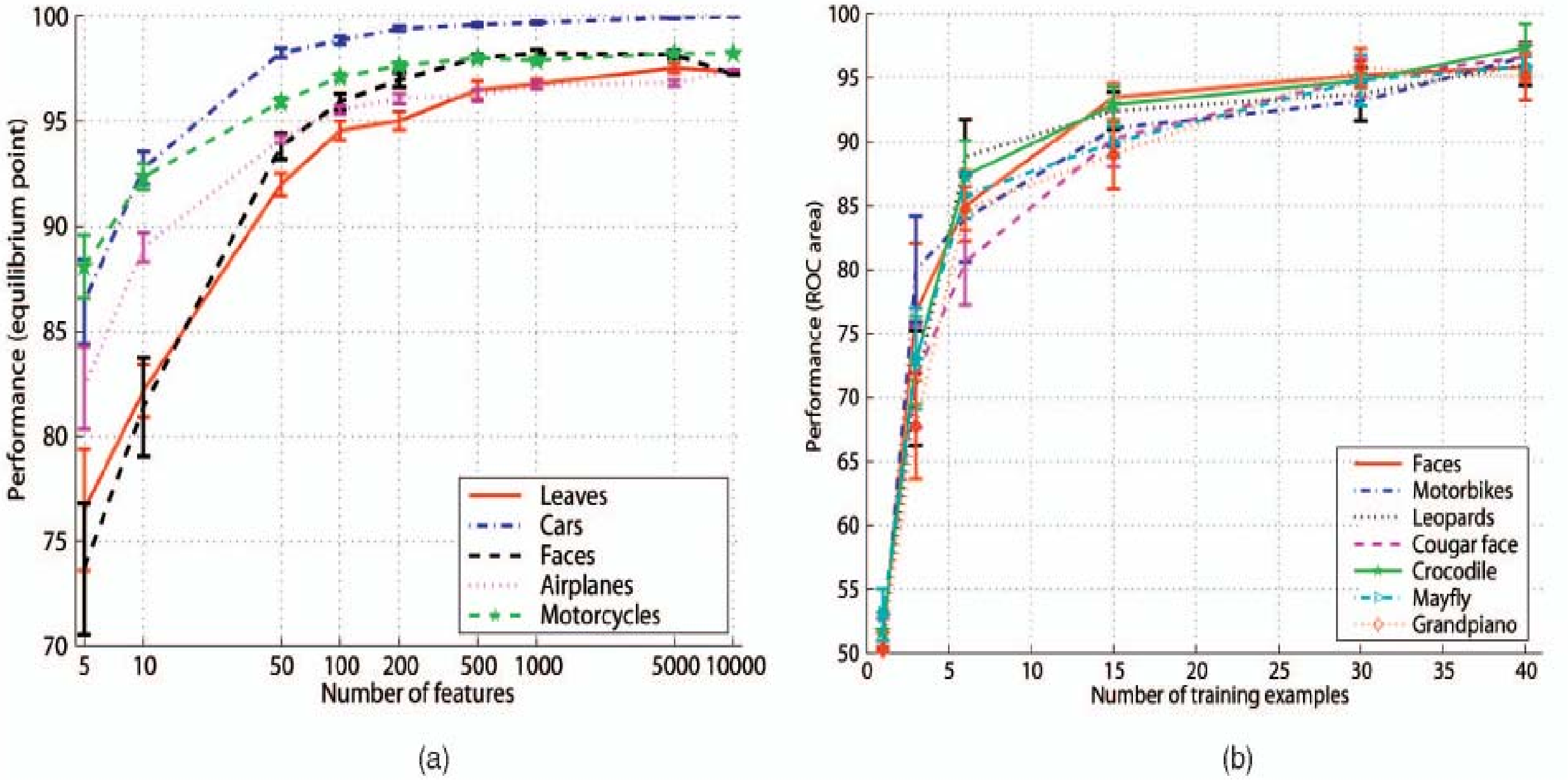
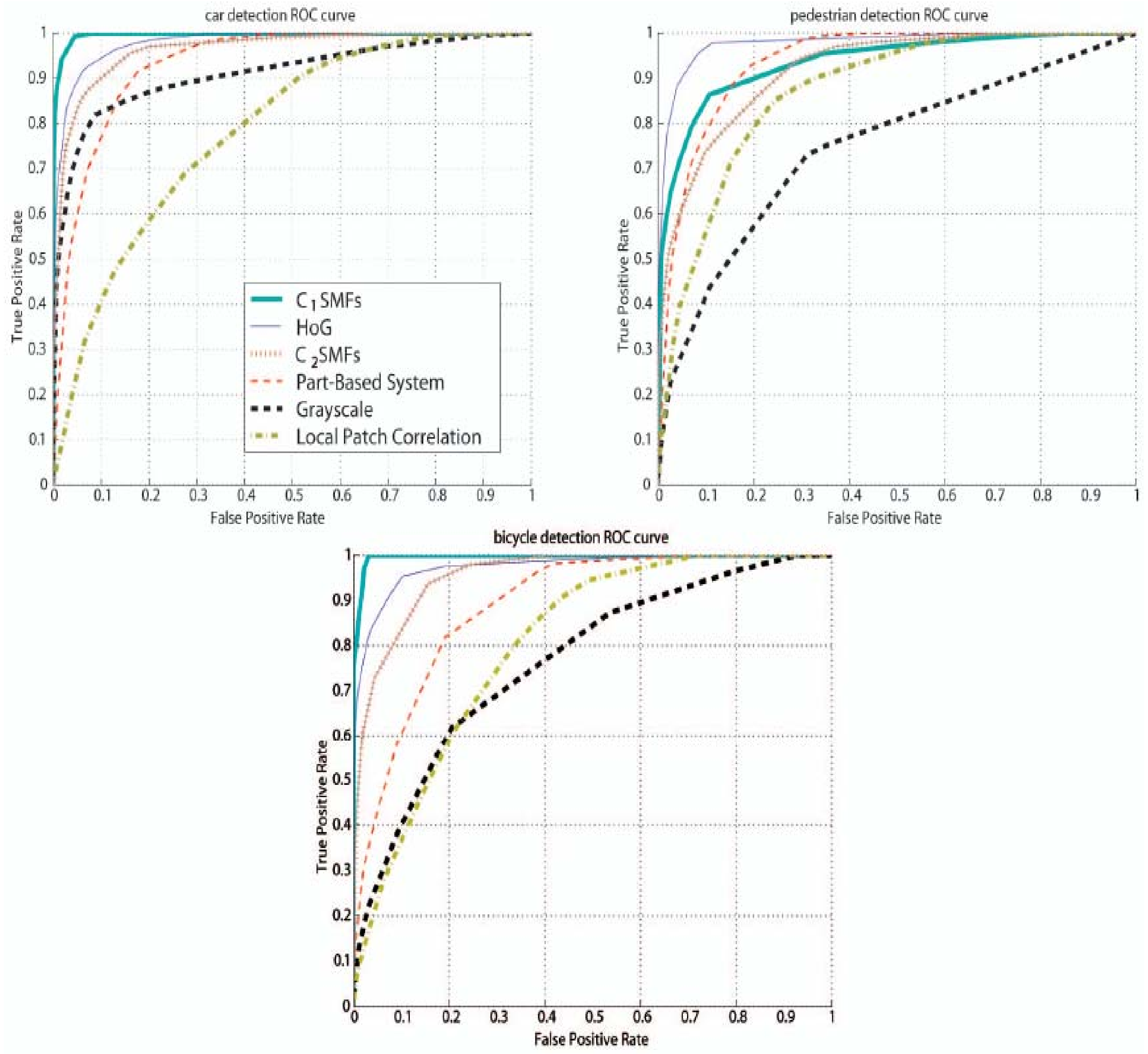


Fig. 4. Performance obtained with gentleBoost and different numbers of C_2 features on the (a) *CalTech5* and on sample categories from the (b) *CalTech101* for a different number of training examples.

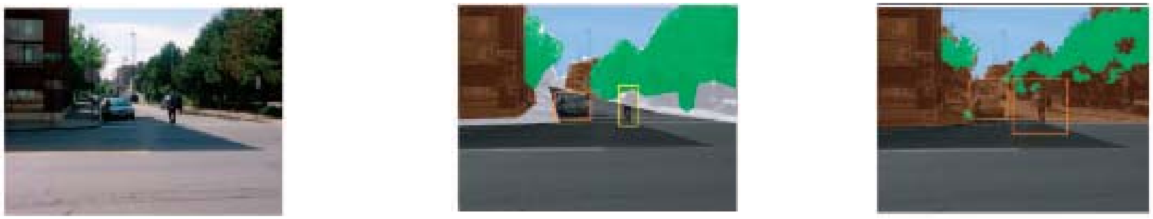
Exp. without clutter

- Data Sets
  - StreetScenes Database
  - Car, pedestrian, bicycle
- Training with gentleBoost
- Training sets of only either pos. or neg.
- Randomized training(1/3) and testing(2/3)
- Windowing approach


car detection ROC curve
pedestrian detection ROC curve
bicycle detection ROC curve
True Positive Rate
False Positive Rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
C_1 SMFs
HoG
C_2SMFs
Part-Based System
Grayscale
Local Patch Correlation

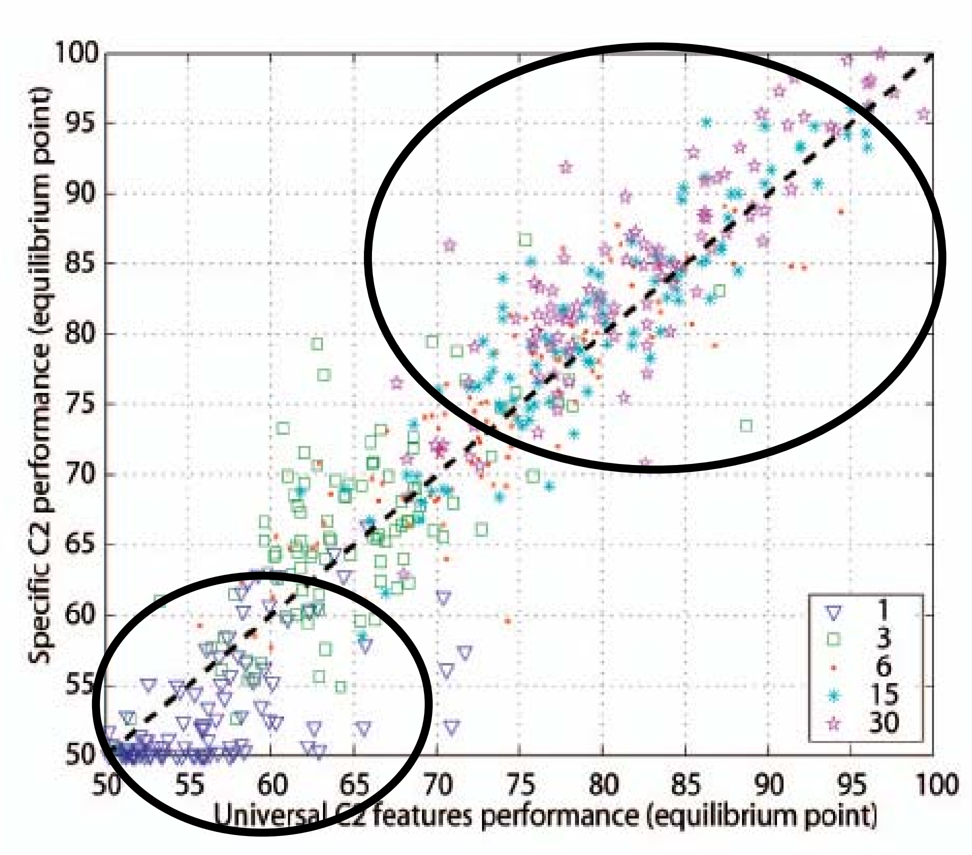
Exp. on Texture based objects

- Again C1 and C2 based classifiers
- C2 features are now evaluated only locally, not over all image locations
- C2 based classification is better (the features are more invariant and complex)
- Evaluated by correct labeling of pixels in the image



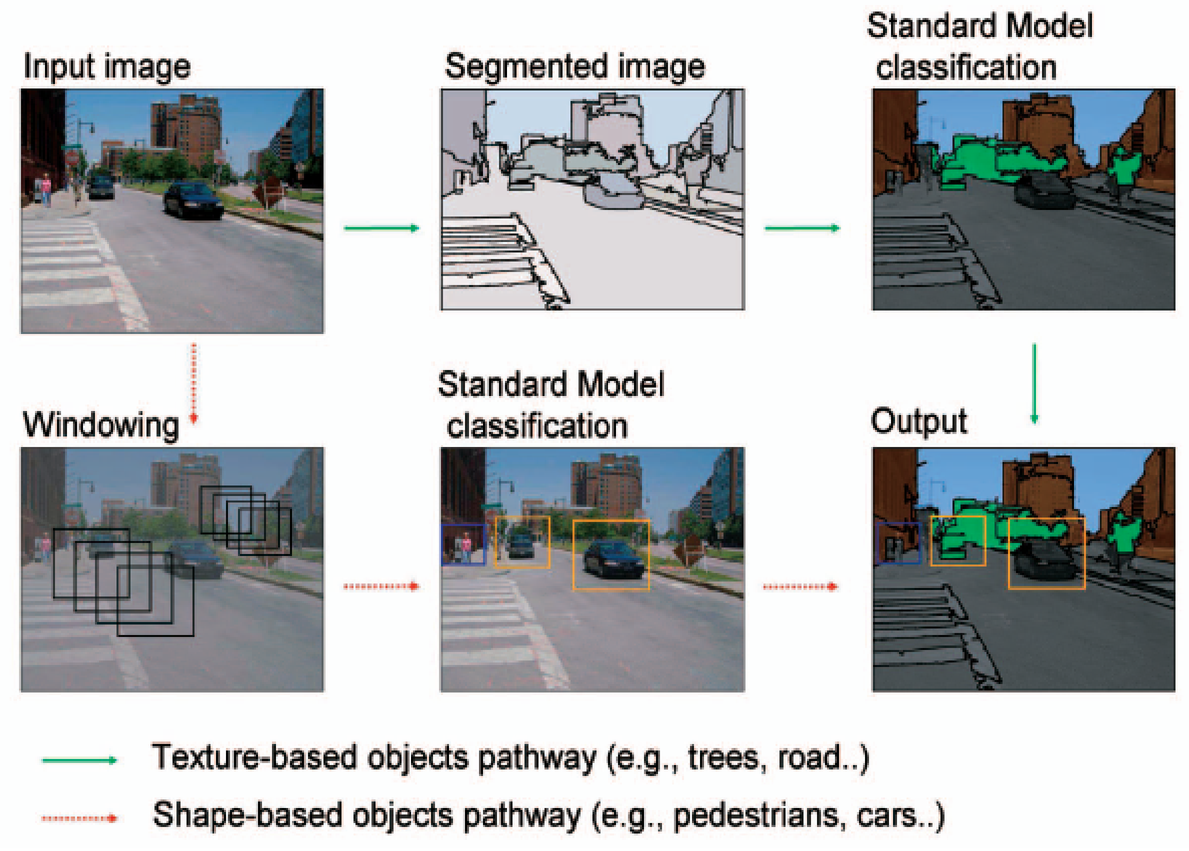
Object specific features or a universal dictionary

- A Universal dictionary based system is good for small training sets (10,000 features)
- An object specific based system is better when using large training sets (improves with practice – increased number of features [200 an image])


Specific C2 performance (equilibrium point)
Universal C2 features performance (equilibrium point)
100
95
90
85
80
75
70
65
60
55
50
50
55
60
65
70
75
80
85
90
95
100
▽ 1
□ 3
• 6
✱ 15
☆ 30

A unified system – looking at multiple processing levels

- The hierarchical nature of the described system enables the use of multiple levels of feature
- Recognizing both shape and texture based objects in the same image
- Two processing pathways



Input image

Segmented image

Standard Model classification

Windowing

Standard Model classification

Output

→ Texture-based objects pathway (e.g., trees, road..)

⇢ Shape-based objects pathway (e.g., pedestrians, cars..)

Scene understanding task

- Complex scene understanding requires more than just detection of objects, location information of the detected objects is also required
- Shape-based objects
  - C1 based classification, using a windowing approach, for both identification and localization
  - Local neighborhood suppression by the maximal detected result
- Texture-based objects
  - C2 based classification
  - texture boundaries posses a problem (solved by additionally segmenting the image and averaging the responses within each segment)

Improvement?

- Only feedforward
- Processing speed

Thank you!