Context-based Object Recognition - Putting context into vision

Phil Crosby and Zhe Lin

CMSC 828J

May 04, 2006

Outline

- Introduction to context
- Overview of Cognitive Science Literature
- Matching words and images
- Statistical contextual priming
- Contextual priors for object recognition
- Other contextual models
- Summary
- References

Definitions of Context

- Any data or meta-data not directly produced by object of interest. Includes:
  - Neighbor-based context: Nearby image data
  - Scene-based context: Scene information
  - Object-based context: Presence, locations of other objects

Contextual Viewing of Visual Attention

Marvin Chun

Chun, M. M. (2000). Contextual cuing of visual attention. Trends in Cognitive Science, 4(5), 170–178.

Chun on Attention

- Attention must prioritize relevant information in the face of information overload
- Bold and flamboyant bottom-up cues can produce rapid search
- Chun argues normal scenes have **too many** bottom-up cues to be useful

Chun on Attention…

- Context dictates what should receive attention and what should be ignored
  - Lights while driving
- Context guides eye movements to the important parts of a scene, which are then fixated upon with foveal vision

Yarbus, A.L. (1967) *Eye Movements and Vision*, Plenum Press

Biederman - Context

- If context guides our attention, it basically facilitates a more efficient search
- Biederman's (1981) influential theory: schema representations for a scene specify the range of plausible objects that can occur, and their positioning relative to each other
- Schemas acquired rapidly, with a glance

Biederman, I. (1981) On the semantics of a glance at a scene. In *Perceptual Organization* (Kubovy, M. and Pomerantz, J., eds), PP. 213–253, Erlbaum

Potential Benefits of Context

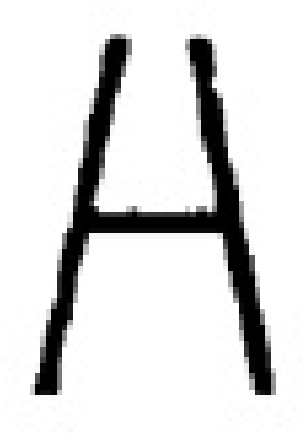
- Eliminate ambiguity!
- Place constraints on the types of objects (Chun, Biederman)
  - Reduces computational complexity when searching memory for an identity match for a new object

Potential Benefits of Context…

- Focus attention, giving more perceptual processing to more relevant image areas
- Exploit redundancy and invariance in a scene.
  - Discard irrelevant or non-changing details
  - Increase predictability to save time

Cognitive science studies on context

What letter is this?

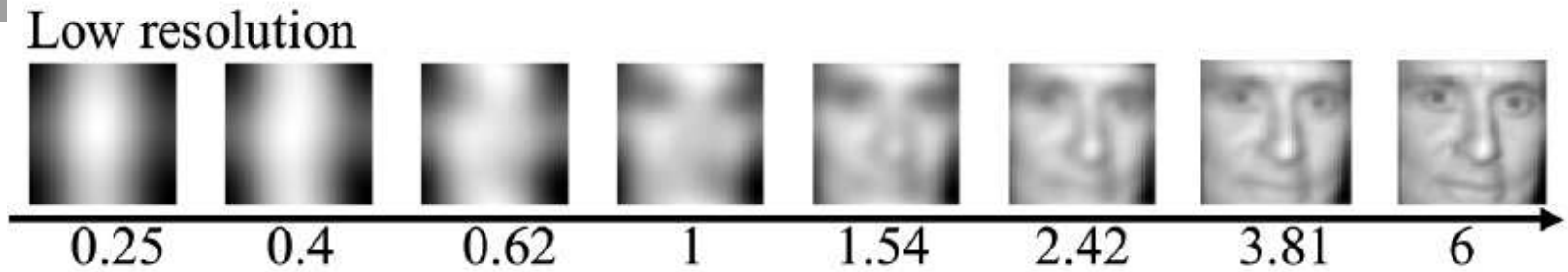


Selfridge, O.G. (1955) *Pattern Recognition and Modern Computers*,IEEE

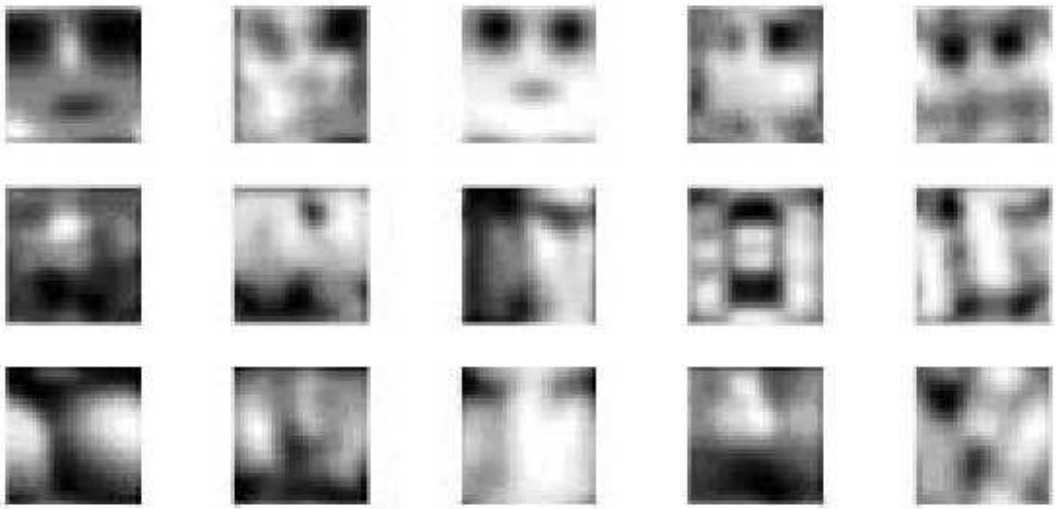
Selfridge - Disambiguation

Selfridge, O.G. (1955) *Pattern Recognition and Modern Computers*,IEEE

Torralba – Face detection in impoverished images



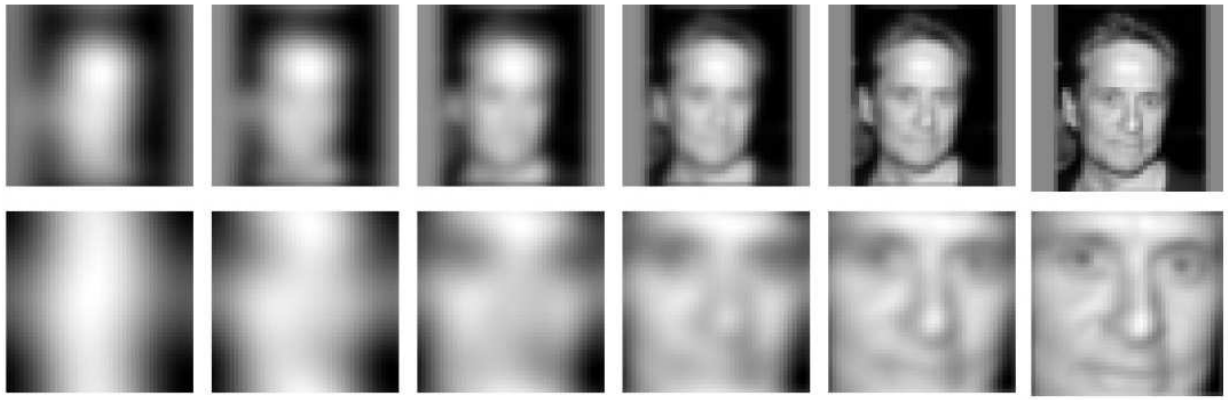
Versus distracter shapes:

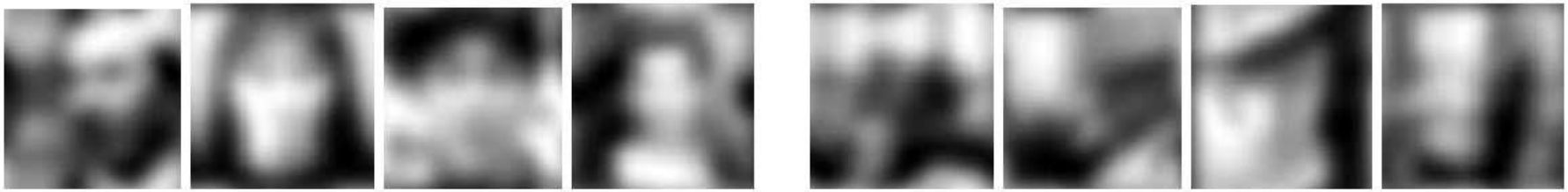


A. Torralba, P. Sinha. (2001). **Detecting faces in impoverished images**. AI Memo 2001-028, CBCL Memo 208, November 2001.

Torralba – Face detection in impoverished images

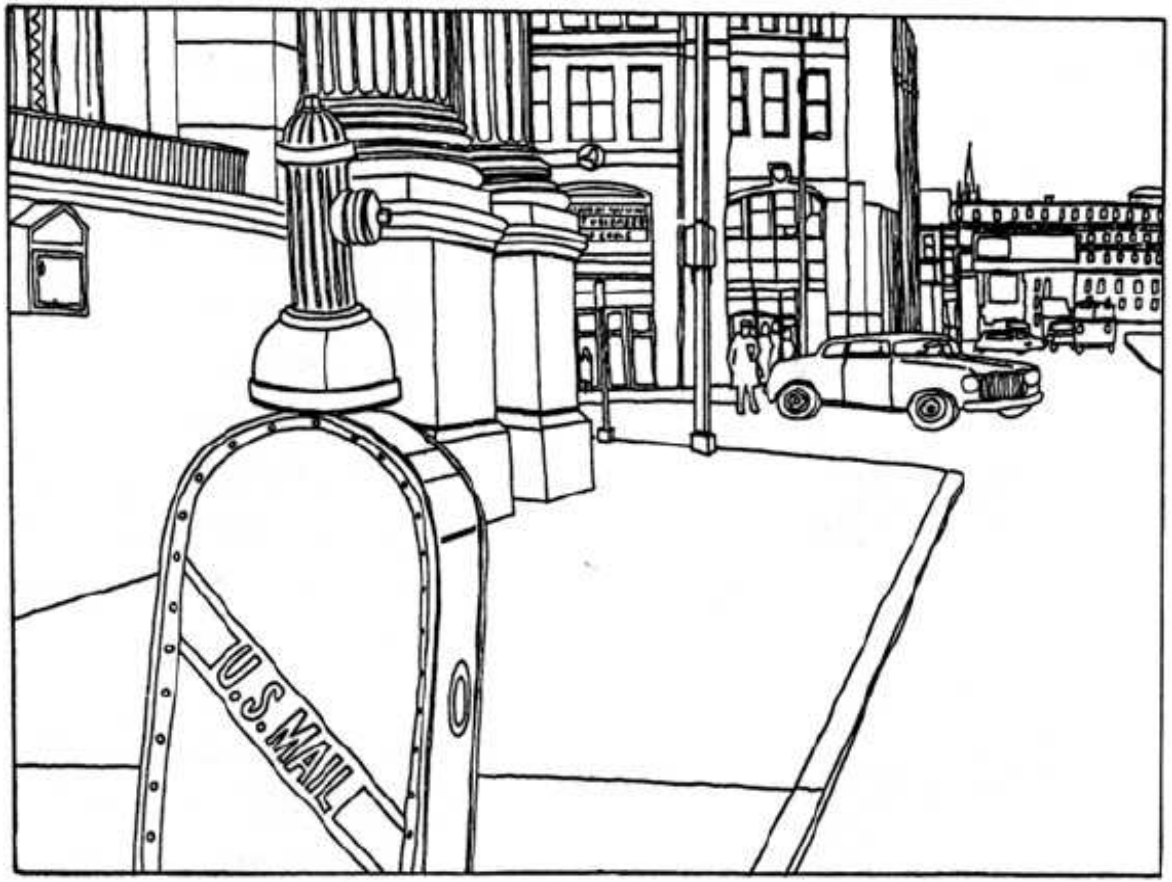
Same resolution, but trade details of face for contour of head





A. Torralba, P. Sinha. (2001). **Detecting faces in impoverished images**. AI Memo 2001-028, CBCL Memo 208, November 2001.

Where is the fire hydrant?

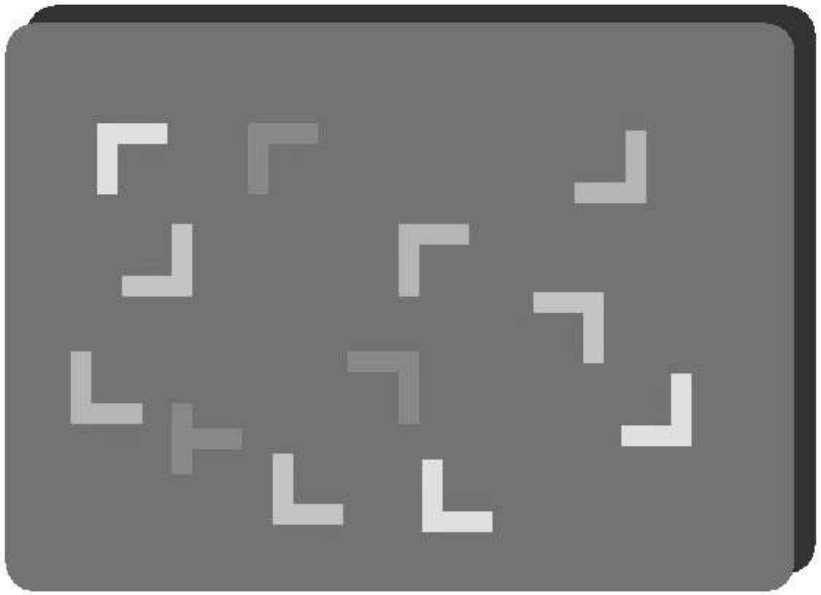


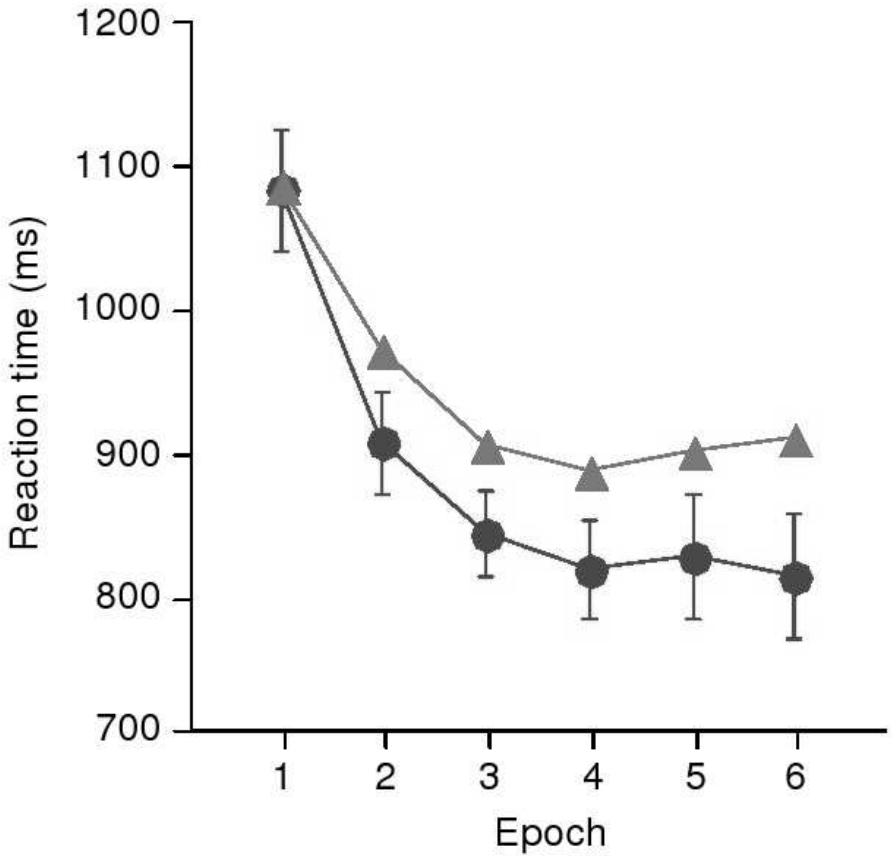
Biederman, I. et al. (1982) Scene perception: detecting and judging objects undergoing relational variations. Cognit. Psychol. 14, 143-177

I think I know that face…

Sinha, P. and Poggio, T. I think I know that face… *Nature* 384, 404

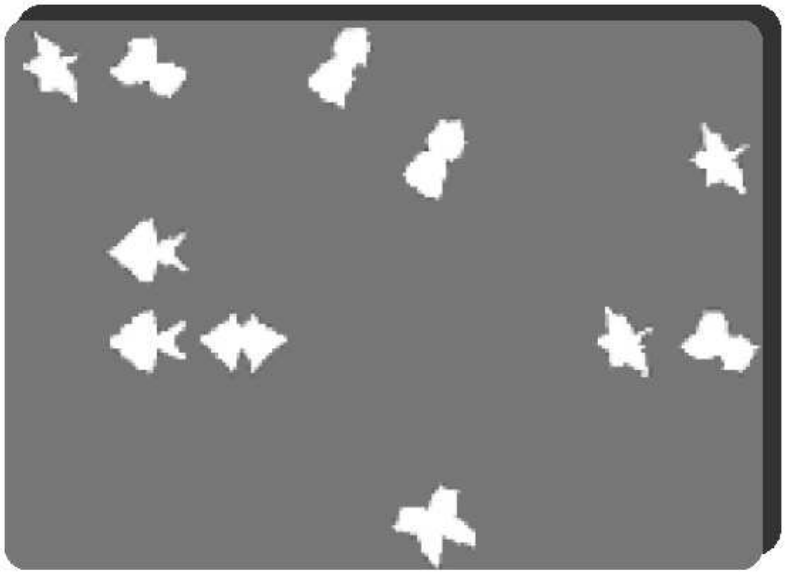
Spatial contextual cueing

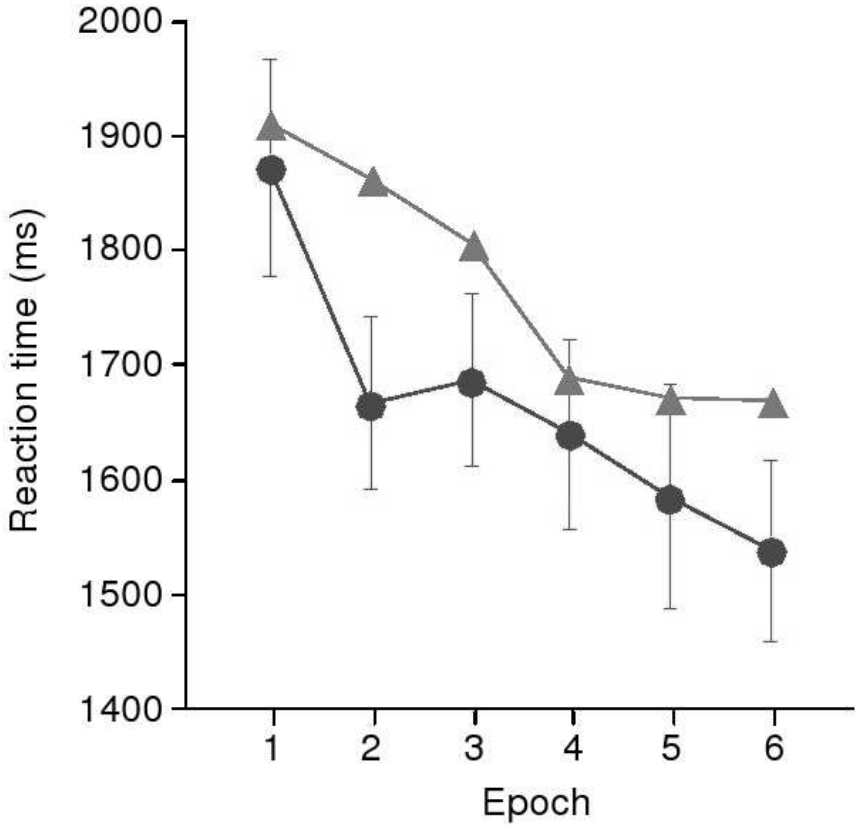




Chun, M.M. and Jiang, Y. (1998) Contextual cueing: implicit learning and memory of visual context guides spatial attention. *Cognit. Psychol.* 36, 28–71

Object cueing





Chun, M.M. and Jiang, Y. (1999) Top-down attentional guidance based on implicit learning of visual covariation. *Psychol. Sci.* 10, 360–365

Henderson, Hollingworth – High Level Scene Perception

- How do we explain the results of these studies?

- Three hypotheses for human object identification in scenes:

Henderson, J.M. and Hollingworth, A. (1999) High-level scene perception. *Annu. Rev. Psychol.* 50, 243–271

Perceptual Schema Model (Biederman's view)

- Expectations derived from knowledge about the composition of a scene type interact with the perceptual analysis of object tokens in that scene.
  - Facilitates perceptual analysis of scene-consistent objects
  - Inhibits analysis of scene-inconsistent objects

Henderson, J.M. and Hollingworth, A. (1999) High-level scene perception. *Annu. Rev. Psychol.* 50, 243–271

Priming Model

- Recognition of a scene "primes" the memory associated with objects that should occur in that scene.
- Criterion for matching a visual observation to an object in memory **is lower** for object memories that are consistent with that scene

Henderson, J.M. and Hollingworth, A. (1999) High-level scene perception. *Annu. Rev. Psychol.* 50, 243–271

Functional Isolation Model

- Object identification is isolated from expectations derived from scene knowledge

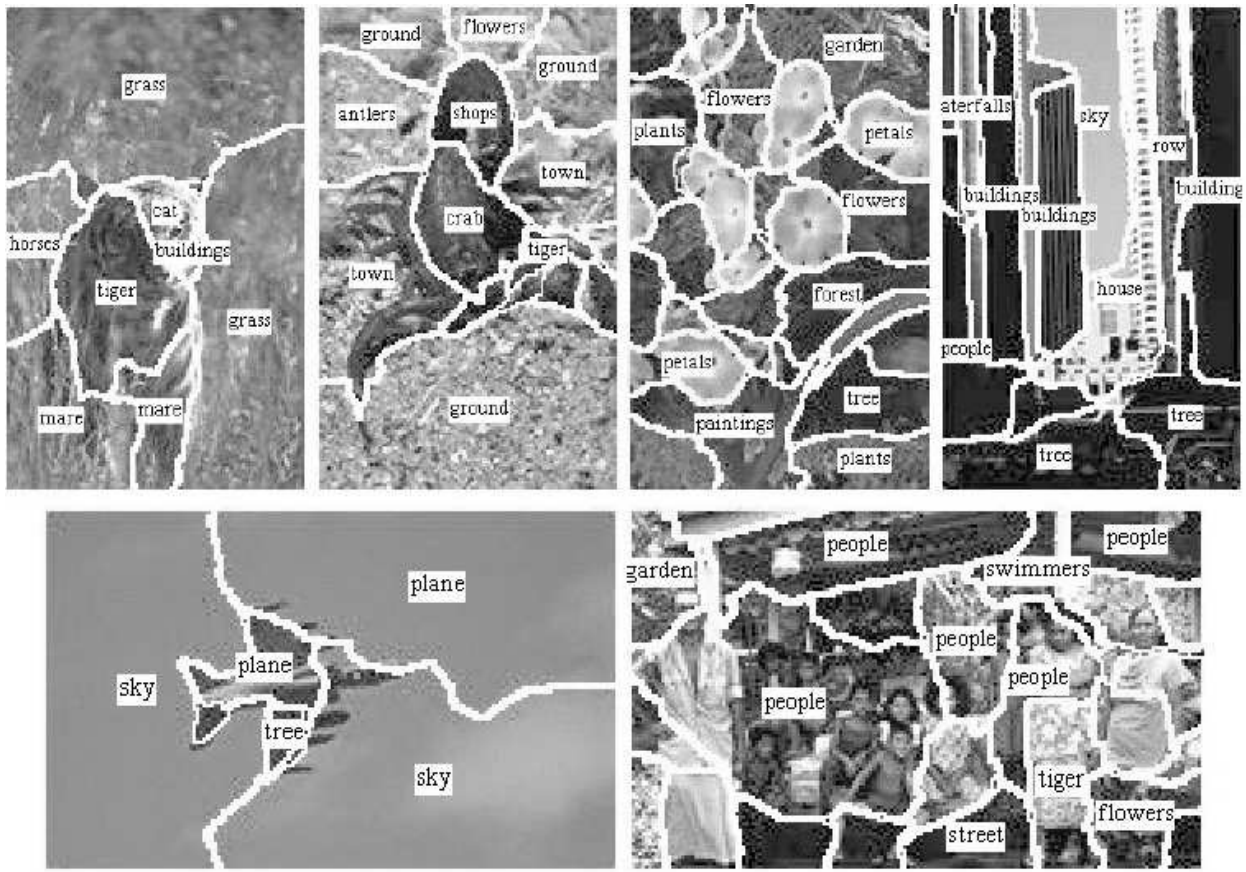
Henderson, J.M. and Hollingworth, A. (1999) High-level scene perception. *Annu. Rev. Psychol.* 50, 243–271

Which model is accurate?

- Who knows?

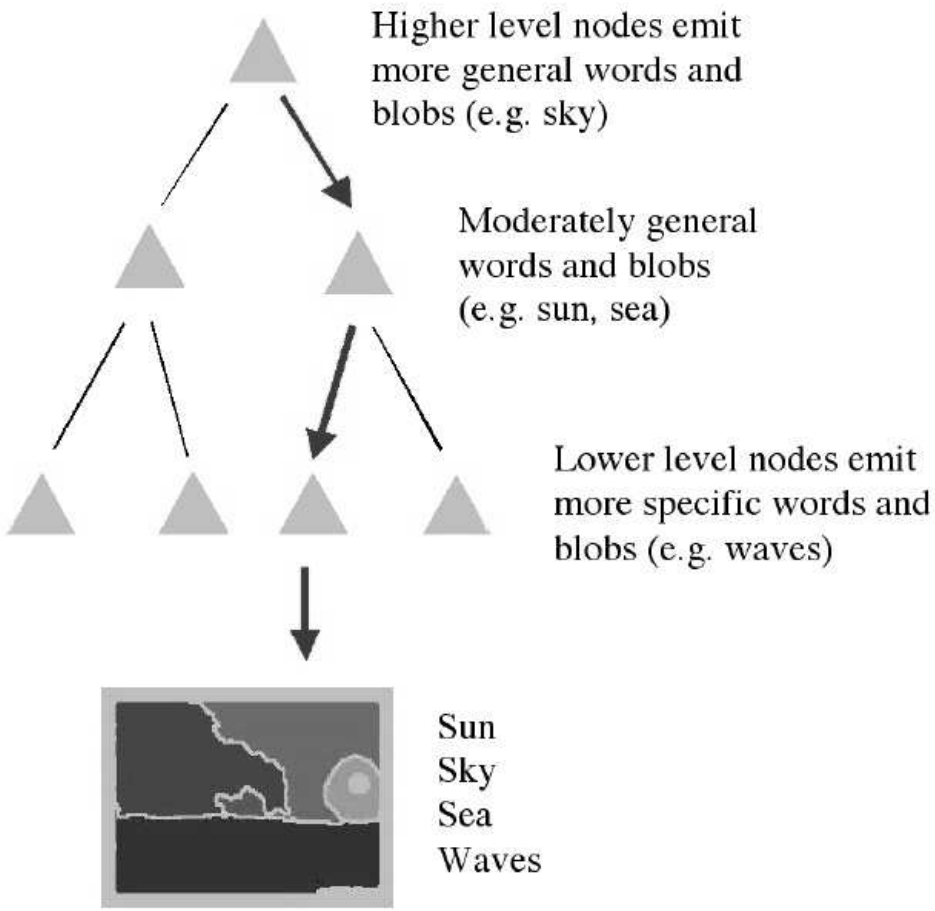
- Henderson & Hollingworth seem to find that the functional isolation model best describes how humans do object recognition

Barnard et al – Matching words and Pictures



Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. 2003. Matching words and pictures. *J. Mach. Learn. Res.* 3 (March 2003), 1107-1135.

Matching words and Pictures - Hierarchical model for annotation



Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. 2003. Matching words and pictures. *J. Mach. Learn. Res.* 3 (March 2003), 1107-1135.

Matching words and Pictures - Hierarchical model for annotation

$$p(D|d) = \sum_c p(c) \prod_{w \in W} \left[\sum_l p(w|l,c)p(l|d) \right]^{\frac{N_w}{N_{w,d}}} \prod_{b \in B} \left[\sum_l p(b|l,c)p(l|d) \right]^{\frac{N_b}{N_{b,d}}}$$

$$\begin{aligned} p(w|B) &\propto \sum_c p(c)p(w|c)p(B|c) \\ &= \sum_c p(c) \left[\sum_l p(w|l,c)p(l|c) \right] \prod_{b \in B} \left[\sum_l p(b|l,c)p(l|c) \right]^{\frac{N_b}{N_{b,d}}} \end{aligned}$$

Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. 2003. Matching words and pictures. *J. Mach. Learn. Res.* 3 (March 2003), 1107-1135.

Matching words and Pictures - leveraging context

- Leverage clusters in the hierarchical model to help make clustering more accurate. Clusters are the context.
- **Approach 1** – linking word emission and region emission probabilities with **mixture weights**:

$$p(D|d) = \sum_c p(c) \prod_{w \in W} \left[\sum_l p(w|l,c) p(l|B,c,d) \right]^{\frac{N_w}{N_{w,d}}} \prod_{b \in B} \left[\sum_l p(b|l,c) p(l|d) \right]^{\frac{N_b}{N_{b,d}}},$$

where we stipulate that

$$p(l|B,c,d) \propto \sum_{b \in B} p(l|b,c,d).$$

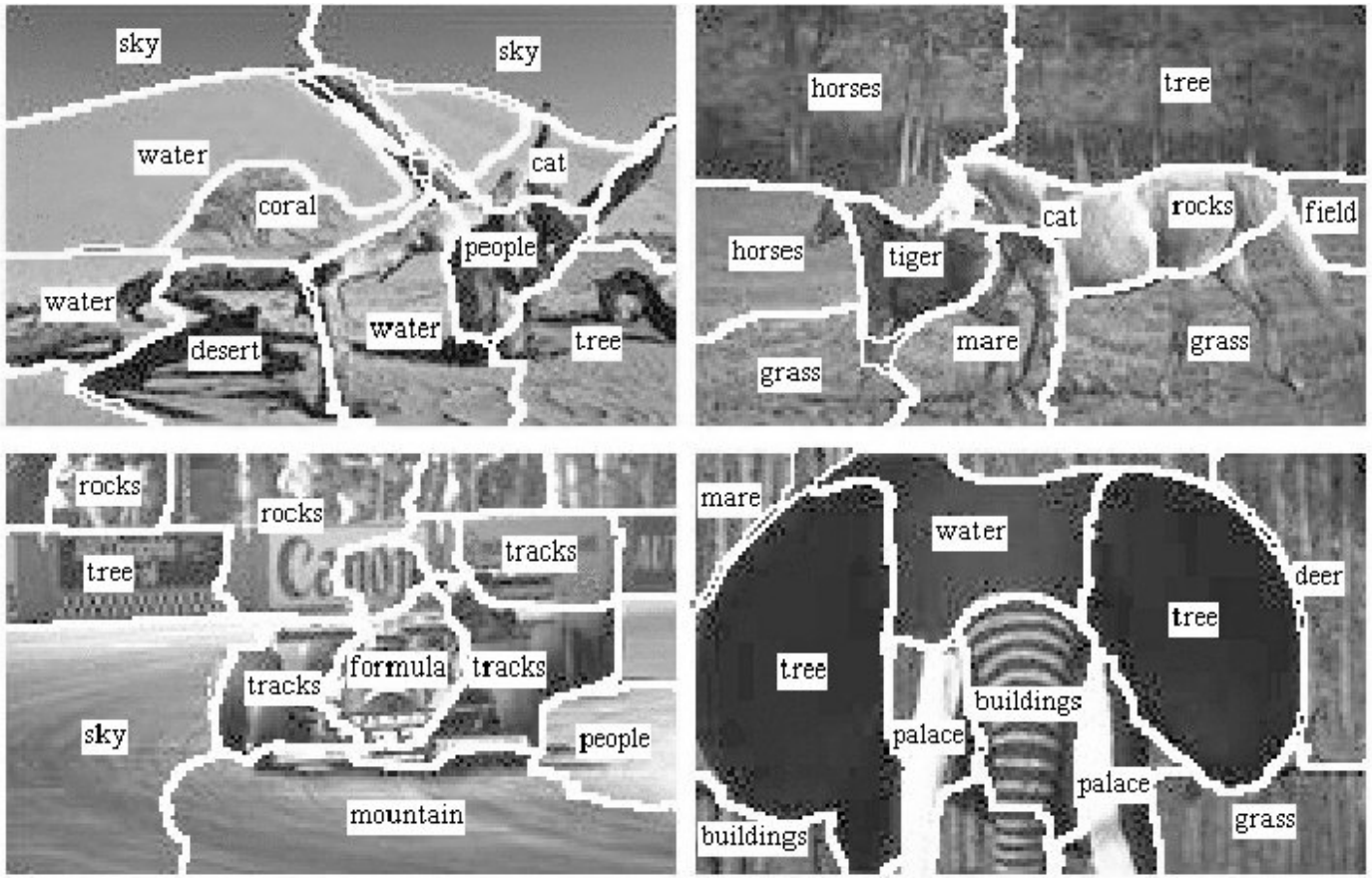
Matching words and Pictures - leveraging context

- **Approach 2** – paired word and region emission at nodes

$$p(D|d) = \sum_c p(c) \prod_{(w,b)\in D} \left[\sum_l p((w,b)|l,c)p(l|d) \right]$$

$$p(w \Leftrightarrow b) \approx \sum_c p(c) \sum_l p((w,b)|l,c)p(l|d).$$

Some more results



Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D. M., and Jordan, M. I. 2003. Matching words and pictures. *J. Mach. Learn. Res.* 3 (March 2003), 1107-1135.

Traditional Approach to Object Detection

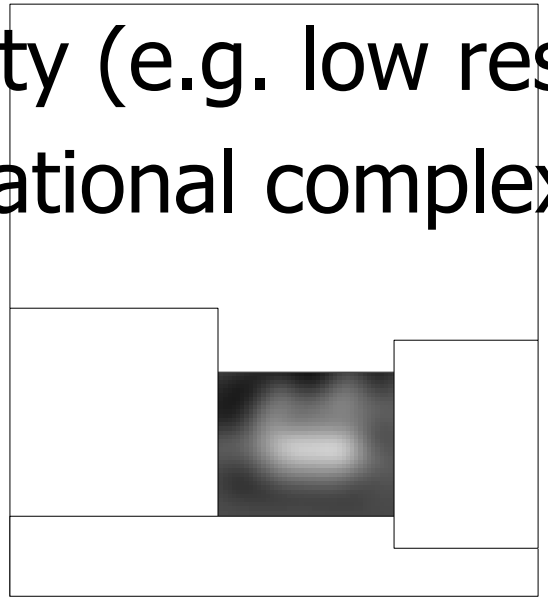
- Uses only local object properties
- Classify local image patches at each location and scale.

$\mathbf{V}_L \rightarrow$ Classifier $p(\text{car} \mid \mathbf{V}_L) \rightarrow$ no car

Images by K. Murphy

Problems of Local Methods

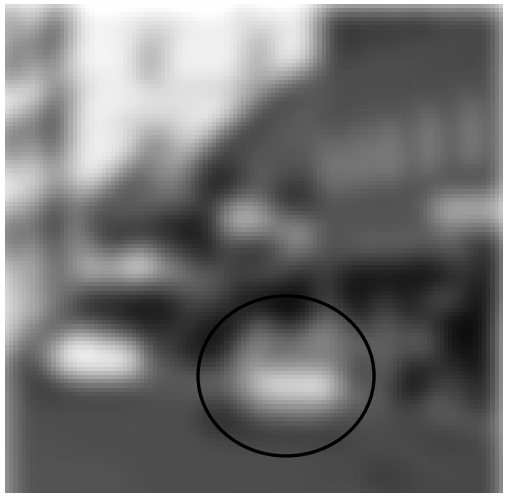
- Ambiguity (e.g. low resolution)
- Computational complexity



What's this?

Images by A. Torralba

Resolving Ambiguity



With contextual information, you can tell what it is.

Images by A. Torralba

Resolving Ambiguity

pedestrian

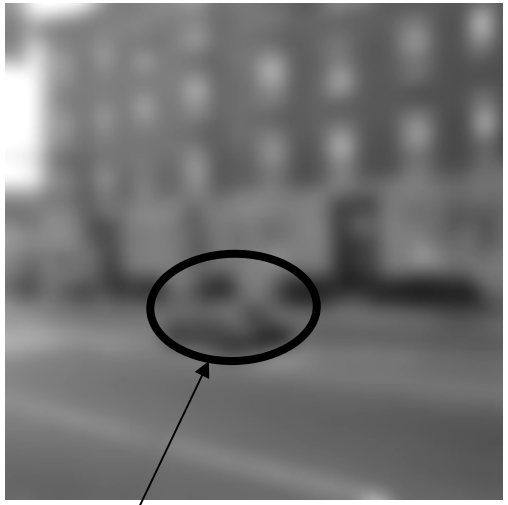
car

ash tray

Images by A. Torralba

Resolving Ambiguity

pedestrian



car

ash tray

Identical local image features!

Images by A. Torralba

Context – Reduces Search Space

- Context can provide a prior on what to look for, and where to look for it.
  - Example
    - Office à desk, chair, computers
    - Outdoor à cars are usually on the road

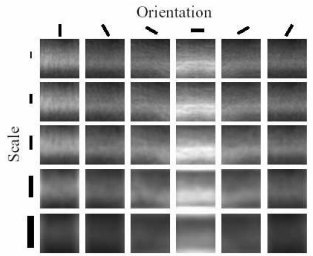
Context Features: Holistic Representation

- Spatial layout of spectral components
  - Responses of filters with different scales and orientations to image
  - Filter examples
    - Steerable Pyramid (Torralba et al., 2003)
    - Weighted Fourier Transform (Torralba & Sinha 2001)
    - Gabor filter banks (Torralba 2003)
  - PCA dimensionality reduction

Context Features

After filtering…

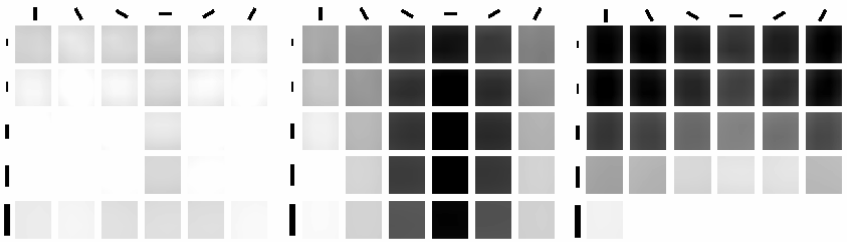
$v(x,k)$



$$v(x,k) \cong \sum_{n=1}^{D} a_n \cdot \psi_n(x,k)$$

PCA

$\psi_1(x,k)$ $\psi_2(x,k)$ $\psi_3(x,k)$



a1 a2 a3

$$Context\ \ features:$$

$$v^C = [a_{1\sim N}]$$

A. Torralba, “Contextual priming for object detection,” IJCV 2003.

Statistical Context Priming

- Context can be a rich source of information about object's identity, location and scale.
- Statistical framework for modeling the relationship between context and object properties.

Statistical Context Priming

We are interested in $P(\vec{p}, \sigma, \vec{x}, o_n \mid \vec{v})$

Local evidence $P_l(\vec{p}, \sigma, \vec{x}, o_n \mid \vec{v}_L)$

Context priming $P_c(\vec{p}, \sigma, \vec{x}, o_n \mid \vec{v}_C)$

$$P_c(\vec{p}, \sigma, \vec{x}, o_n \mid \vec{v}_C) = P_p(\vec{p} \mid \sigma, \vec{x}, o_n, \vec{v}_C) P_s(\sigma \mid \vec{x}, o_n, \vec{v}_C) P_f(\vec{x} \mid o_n, \vec{v}_C) P_o(o_n \mid \vec{v}_C)$$

Pose and Shape Priming | Scale Selection | Focus of Attention | Object Priming

A. Torralba and P. Sinha, “Statistical Context priming for object detection,” ICCV 2001.

Object Priming

- Annotated database for learning
- Learning by MOG

$$P(o / \mathbf{v}_C) = \frac{P(\mathbf{v}_C / o)P(o)}{P(\mathbf{v}_C)}$$

$$P(\mathbf{v}_C) = P(\mathbf{v}_C / o)P(o) + P(\mathbf{v}_C / \neg o)P(\neg o)$$

$$P(\mathbf{v}_C / o) = \sum_{i=1}^{M} b_i \cdot G(\mathbf{v}_C; \mu_i, \Sigma_i) \quad P(\mathbf{v}_C / \neg o) = \sum_{i=1}^{M} b_i \cdot G(\mathbf{v}_C; \mu_i, \Sigma_i)$$

Model parameters $(b_i, \mu_i, \Sigma_i)_{i=1,M}$ are learned from **EM algorithm**.

A. Torralba, “Contextual priming for object detection,” IJCV 2003.

Object Priming Result

Figure 7. Random selection of images from the test set showing the results of object priming for four superordinate object categories (o_1 = people, o_2 = furniture, o_3 = vehicles and o_4 = trees). The bars at the right-hand of each picture represent the probability $P(o \mid \mathbf{v}_C)$.

A. Torralba, “Contextual priming for object detection,” IJCV 2003.

Object Priming Result

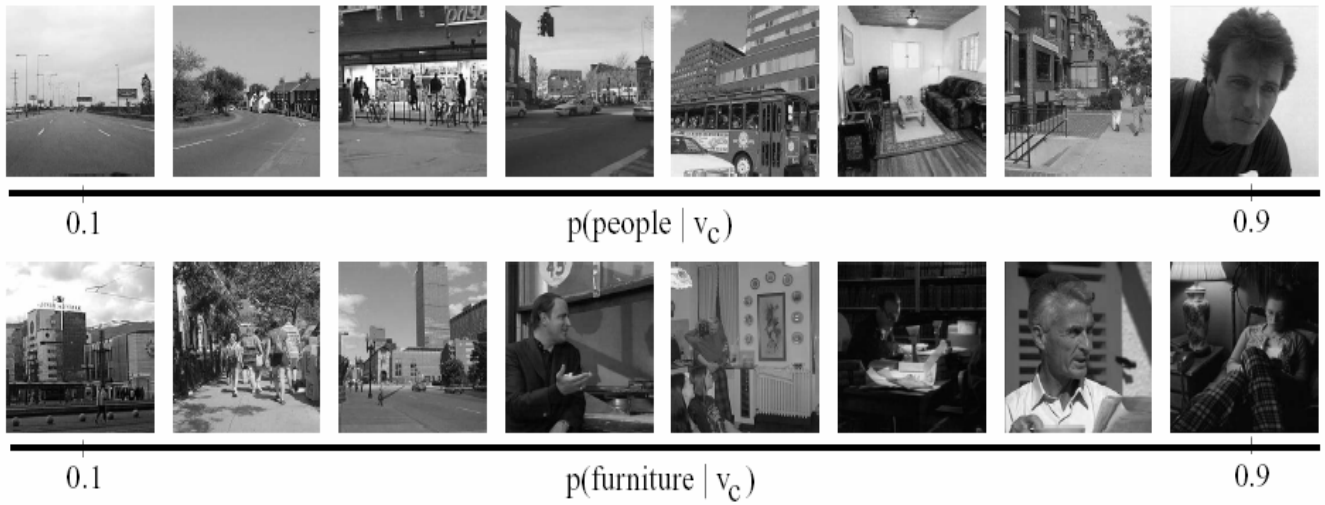


Fig. 8. Random selection of images from the test set organized with respect to the probability $P(o|\mathbf{v}_C)$ for $o =$ people and furniture.

A. Torralba, “Contextual priming for object detection,” IJCV 2003.

Context and Task-driven FOA

- Model: Gaussian Mixture Model

$$P(\mathbf{x} \mid o, \mathbf{v}_C) = \frac{\sum_{i=1}^{M} b_i \, G(\mathbf{x}; \mathbf{x}_i, \mathbf{X}_i) G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i \, G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)} \qquad \mathbf{x}_i = \mathbf{a}_i + \mathbf{A}_i(\mathbf{v}_C - \mathbf{v}_i)$$

$$(\bar{x}, \bar{y}) = \int \mathbf{x} P(\mathbf{x} \mid o, \mathbf{v}_C) \, d\mathbf{x}$$

$$= \frac{\sum_{i=1}^{M} b_i \mathbf{x}_i G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)} \qquad \sigma_r^2 = \int r^2 P(\mathbf{x} \mid o, \mathbf{v}_C) \, d\mathbf{x}$$

- Model learning by EM

$$(b_i, a_i, \mathbf{A}_i, \mathbf{X}_i, \mathbf{v}_i, \mathbf{V}_i)_{i=1,M}$$

A. Torralba, “Contextual priming for object detection,” IJCV 2003.

Context and Task-driven FOA

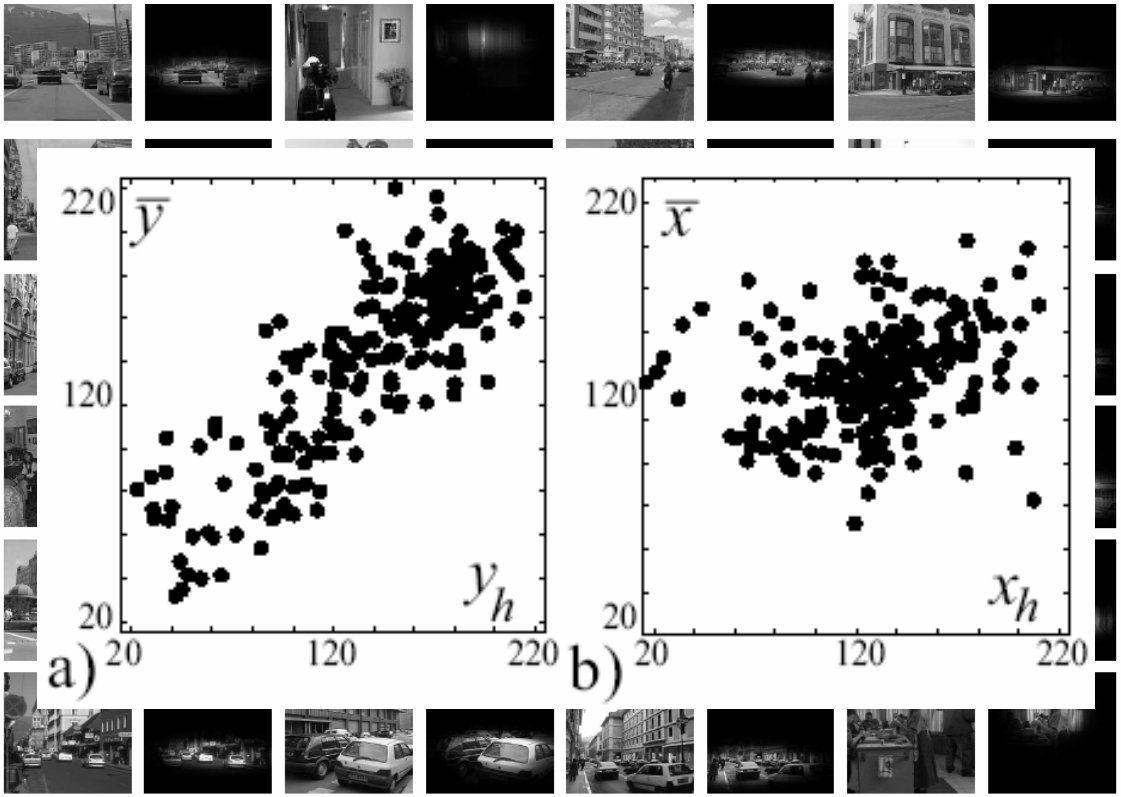


Fig. 14. Focus of attention based on global context configuration. Each pair shows the original image and the image multiplied by the function $P(\mathbf{x} \mid \mathbf{v}_C, o = heads)$ to illustrate the primed regions.

A. Torralba, “Contextual priming for object detection,” IJCV 2003.

Context-driven Scale Selection

- Model (GMM)

$$P(\sigma \mid o, \mathbf{v}_C) = \frac{\sum_{i=1}^{M} b_i \, G(\sigma; \sigma_i, \mathbf{S}_i) G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)} \qquad \sigma_i = a_i + \mathbf{A}_i(\mathbf{v}_C - \mathbf{v}_i)$$

$$\bar{\sigma} = \int \sigma \, P(\sigma \mid o, \mathbf{v}_C) \, d\sigma = \frac{\sum_{i=1}^{M} \sigma_i b_i G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}{\sum_{i=1}^{M} b_i G(\mathbf{v}_C; \mathbf{v}_i, \mathbf{V}_i)}$$

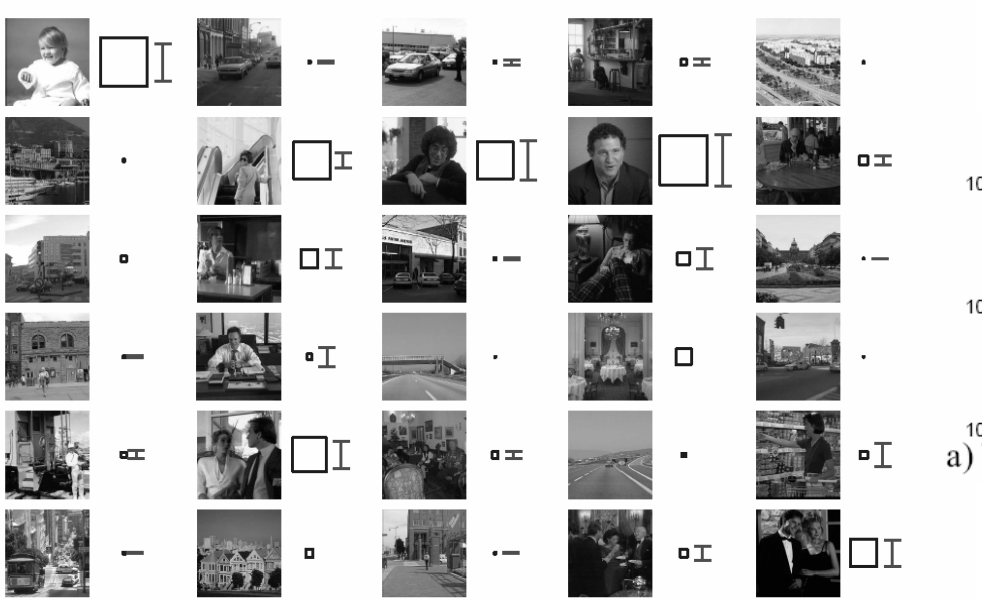
$$\sigma_h^2 = \int (\sigma - \bar{\sigma})^2 P(\sigma \mid o, \mathbf{v}_C) \, d\sigma$$

- Model Learning by EM

$$(b_i, a_i, \mathbf{A}_i, \mathbf{S}_i, \mathbf{v}_i, \mathbf{V}_i)_{i=1,M}$$

A. Torralba, “Contextual priming for object detection,” IJCV 2003.

Context-Driven Scale Selection



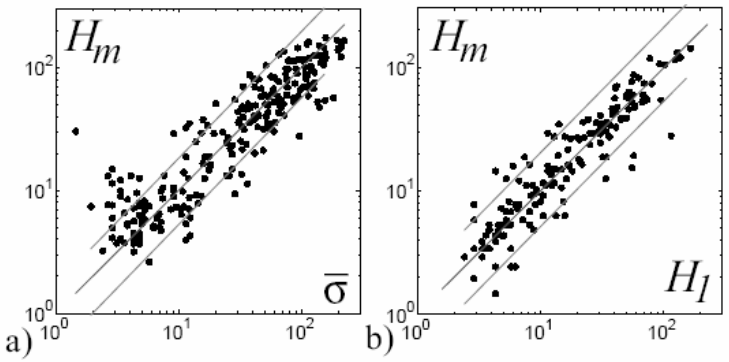


Fig. 19. Results for scale selection given global context information for rand

Fig. 22. Each row shows a set of 8 pictures sorted according to the predicted size of human heads (top) and cars (bottom).

A. Torralba, "Contextual priming for object detection," IJCV 2003.

Context-based Place and Object Recognition

- Testbed
  - Wearable device: helmet-mounted mobile system
- Place and Scene Recognition
  - Global features + PCA
  - Wearable test-bed
  - Model of place recognition – HMM
- Object Recognition
  - Individual object recognition using contextual information

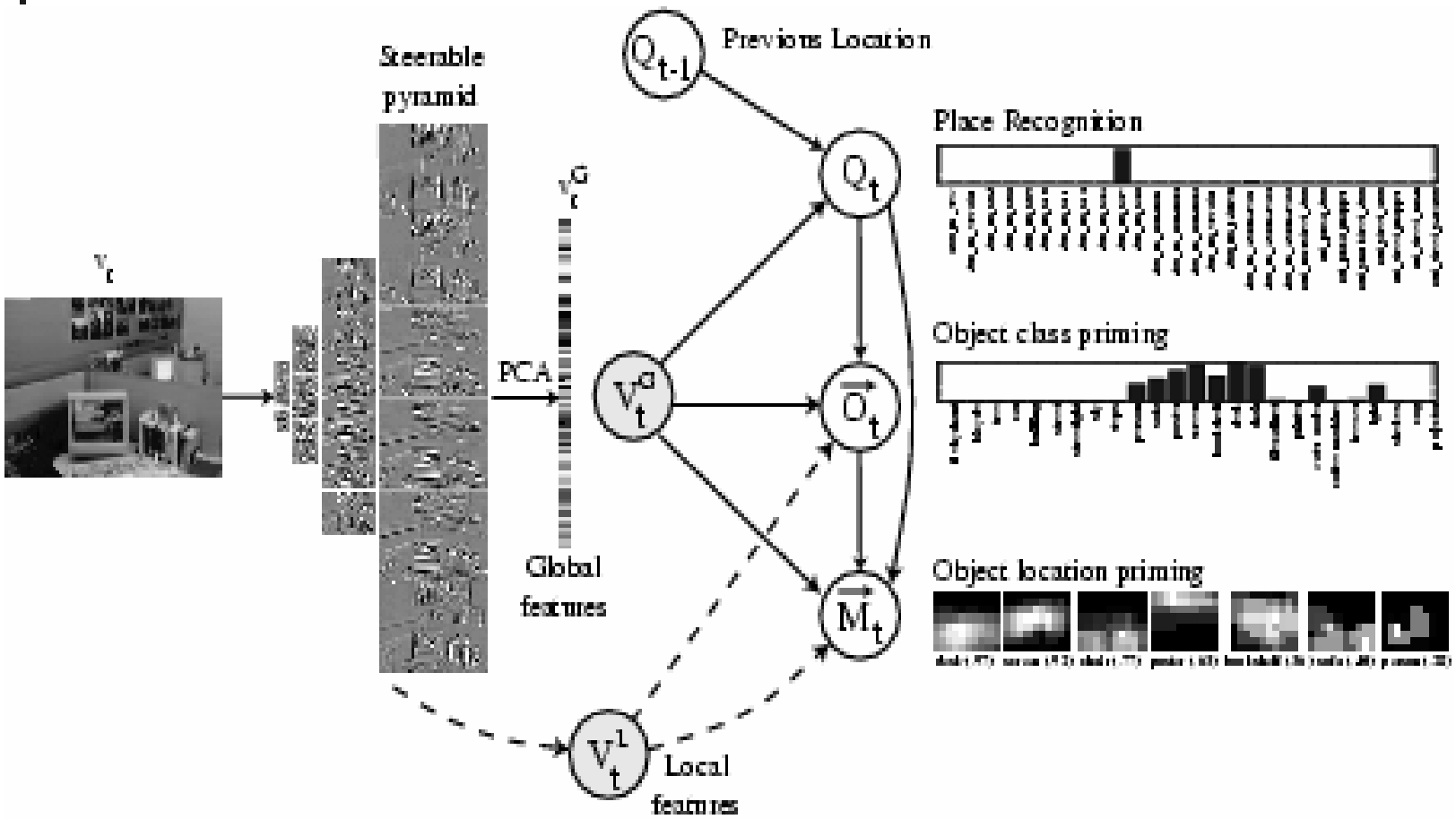
A. Torralba et al., “Context-based vision system for place and object recognition,” ICCV 2003.

Place, Category Recognition

- Place: Instance of a scene category
  - E.g. office 610, main street
- Scene category: type of place
  - E.g. office, street

A. Torralba et al., “Context-based vision system for place and object recognition,” ICCV 2003.

System Framework



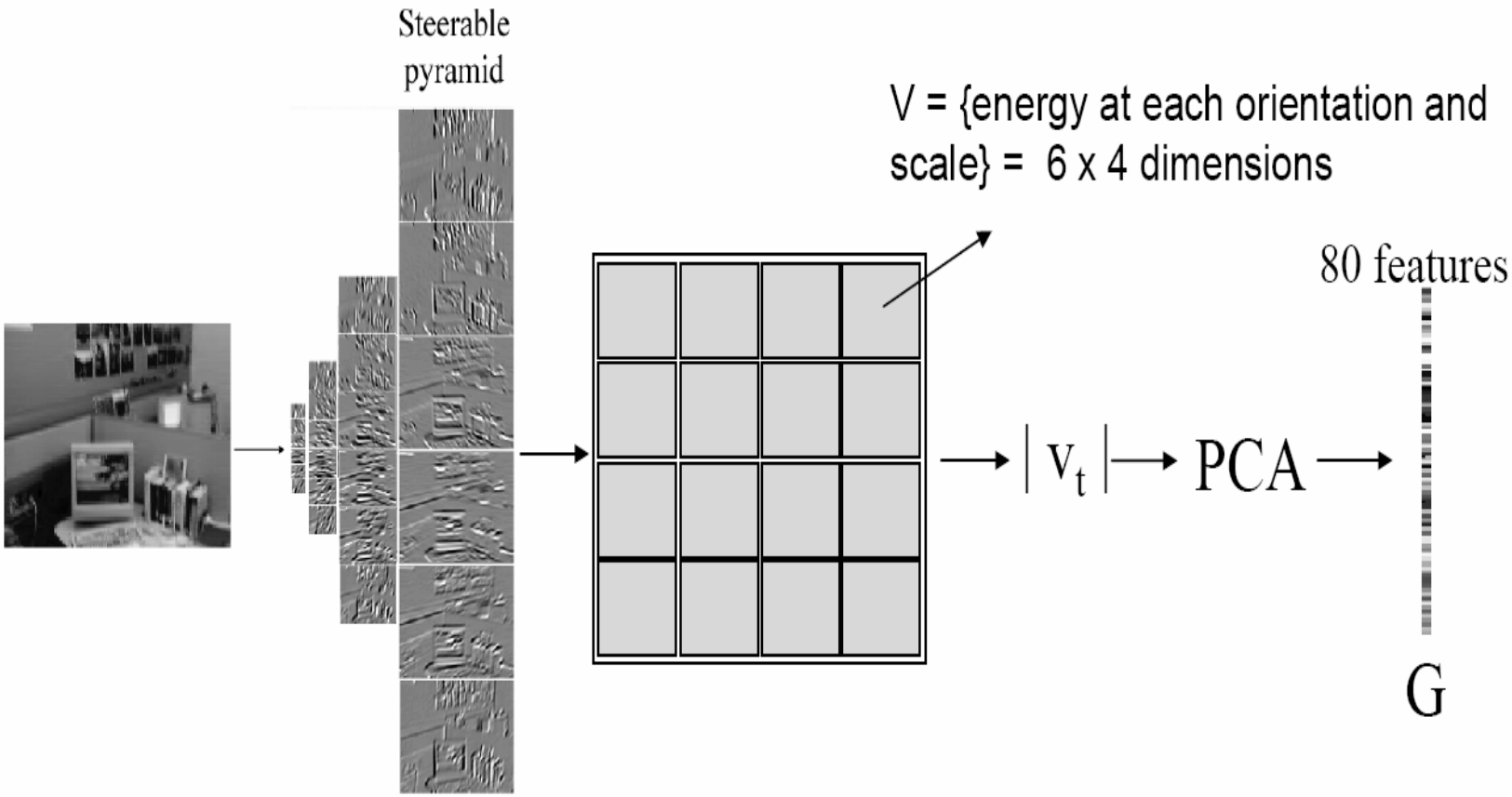
A. Torralba et al., “Context-based vision system for place and object recognition,” ICCV 2003.

Low-dimensional Scene Representation

- Compute image intensity (no color)
- Steerable pyramid (6 orientations, 4 scales)
- Compute magnitude of average filter responses
- Downsample to 4 x 4Ł 384 dimensions
- PCA to 80 dimensions

A. Torralba et al., “Context-based vision system for place and object recognition,” ICCV 2003.

Global Feature (Interpretation)



A. Torralba et al., “Context-based vision system for place and object recognition,” ICCV 2003.

Visualizing the Global Features

Images

Gist: 80-dim Representation

A. Torralba et al., “Context-based vision system for place and object recognition,” ICCV 2003.

Models for Place Recognition

- Transition model: HMM

$$P(Q_t = q|v_{1:t}^G) \quad \propto \quad p(v_t^G|Q_t = q)P(Q_t = q|v_{1:t-1}^G)$$
$$= p(v_t^G|Q_t = q)\sum_{q'} A(q', q)P(Q_{t-1} = q'|v_{1:t-1}^G)$$

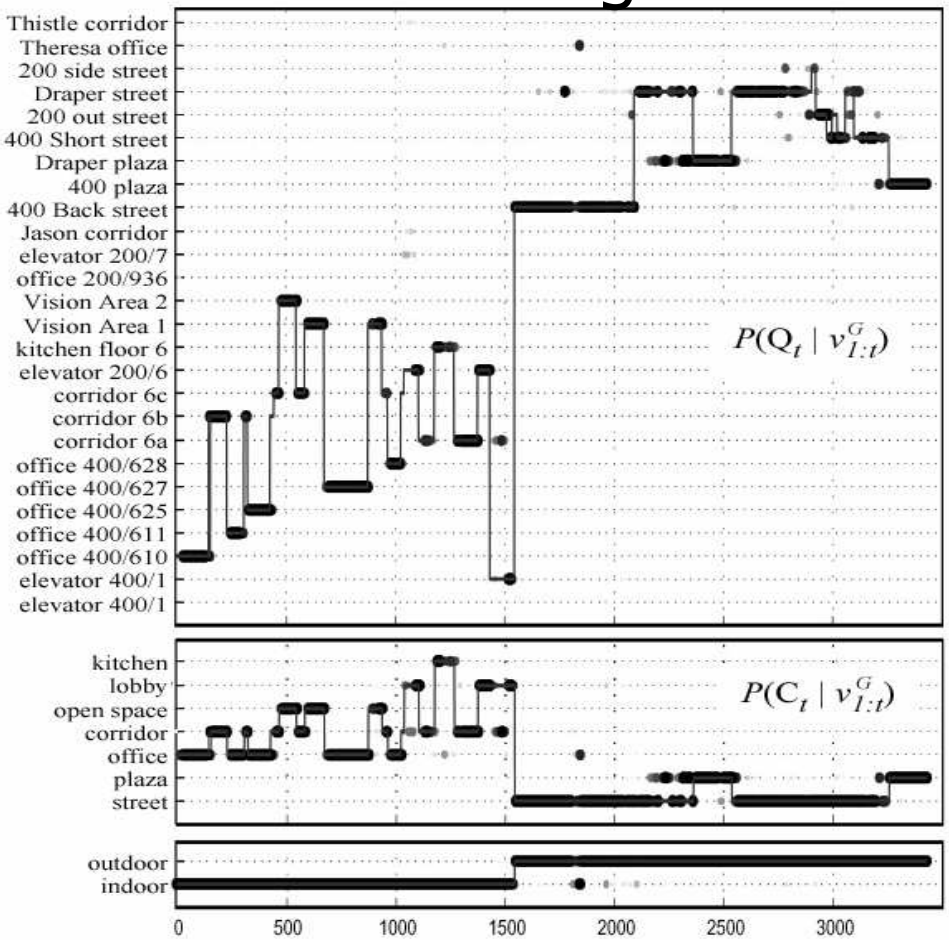
$$A(q', q) = P(Q_t = q|Q_{t-1} = q')$$

- Observation model $p(v_t^G|Q_t)$
  - Mixture of Gaussian (100views/Places)

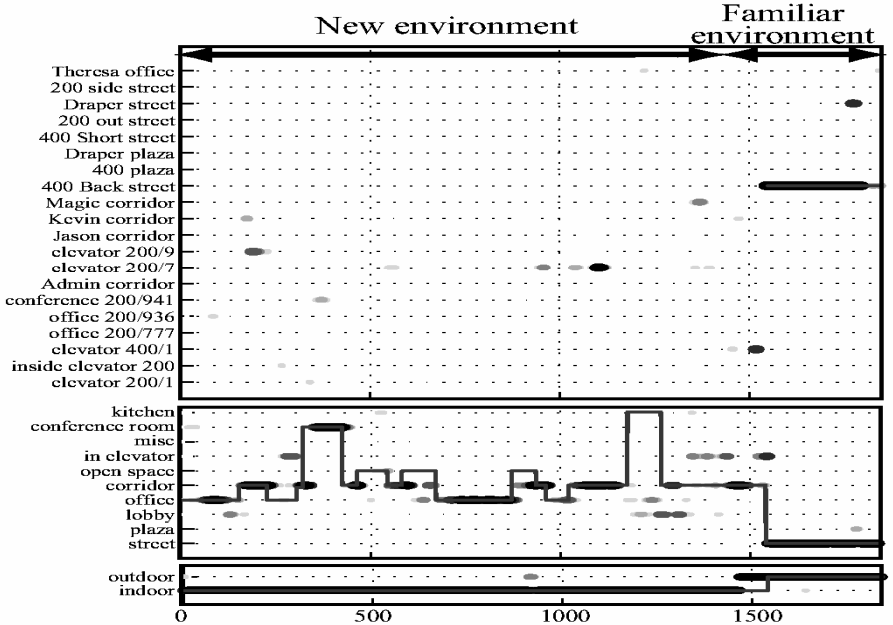
A. Torralba et al., “Context-based vision system for place and object recognition,” ICCV 2003.

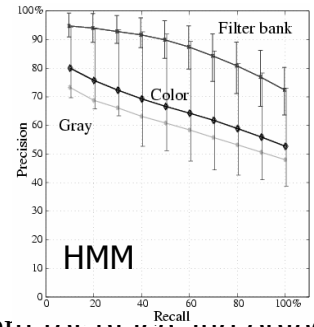
Recognition Performance

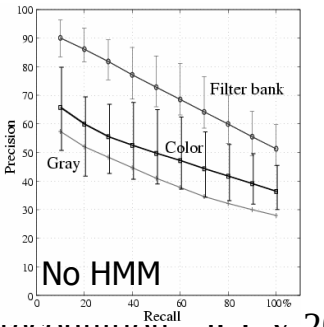
Place Recognition



Category Recognition







A. Torralba et al., "Context-based vision system for place and object recognition," ICCV 2003.

Contextual Priming for Object Detection

Object properties

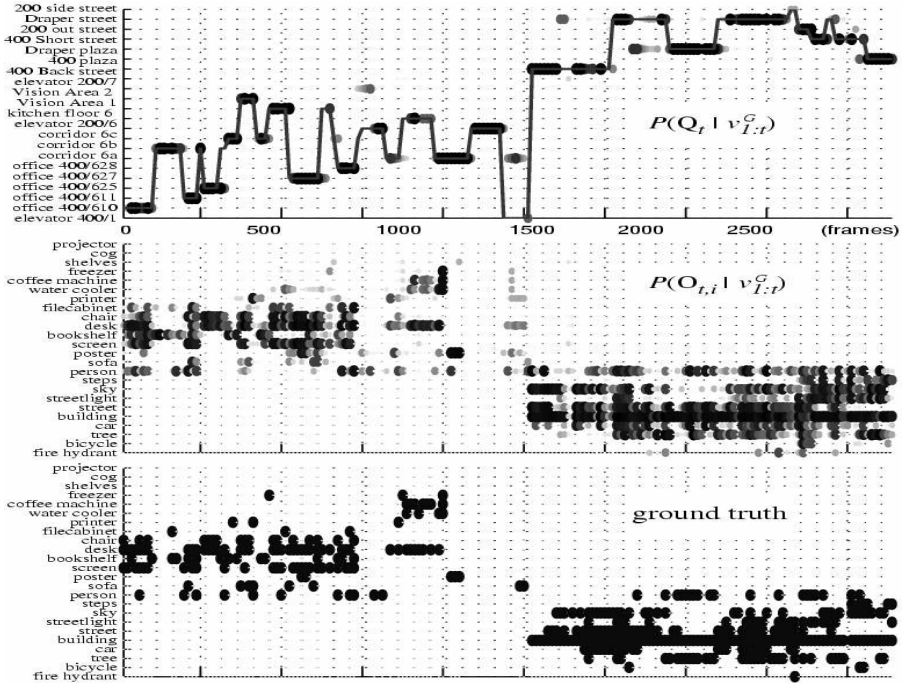
$$\vec{O}_t = (O_{t,1}, \ldots, O_{t,N_o})$$

Object priming using context ONLY

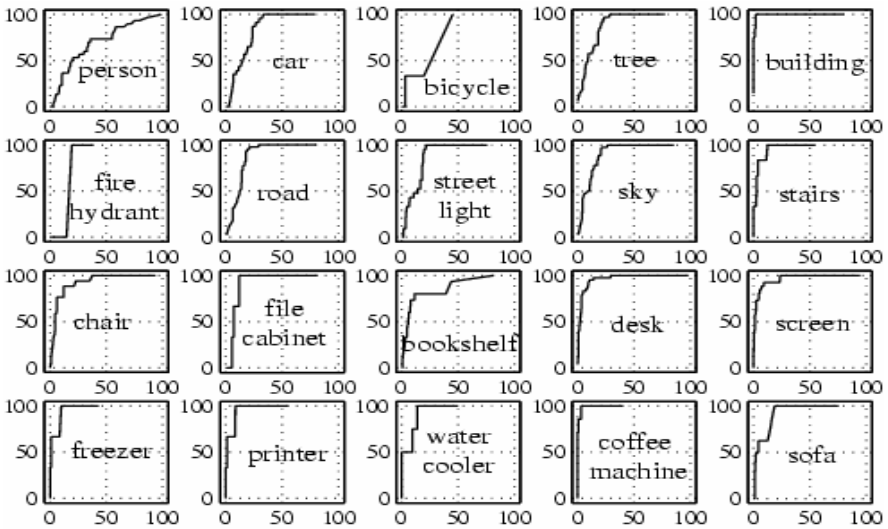
$$P(O_{t,i}|v_t^G, Q_t = q) = \frac{p(v_t^G|O_{t,i}, j)P(O_{t,i}|q)}{p(v_t^G|O_{t,i}, q)P(O_{t,i}|q) + p(v_t^G|\bar{O}_{t,i}, q)P(\bar{O}_{t,i}|q)}$$

A. Torralba et al., “Context-based vision system for place and object recognition,” ICCV 2003.

Predicting Object Presence



Object presence prob.



Perf. by ROC curve

A. Torralba et al., “Context-based vision system for place and object recognition,” ICCV 2003.

Contextual Priors for Object Localization

- Goal: $P(X_{t,i}|v_t^G, O_{t,i} = 1)$
- Approach:
  - Divide images into 8*10 binary masks

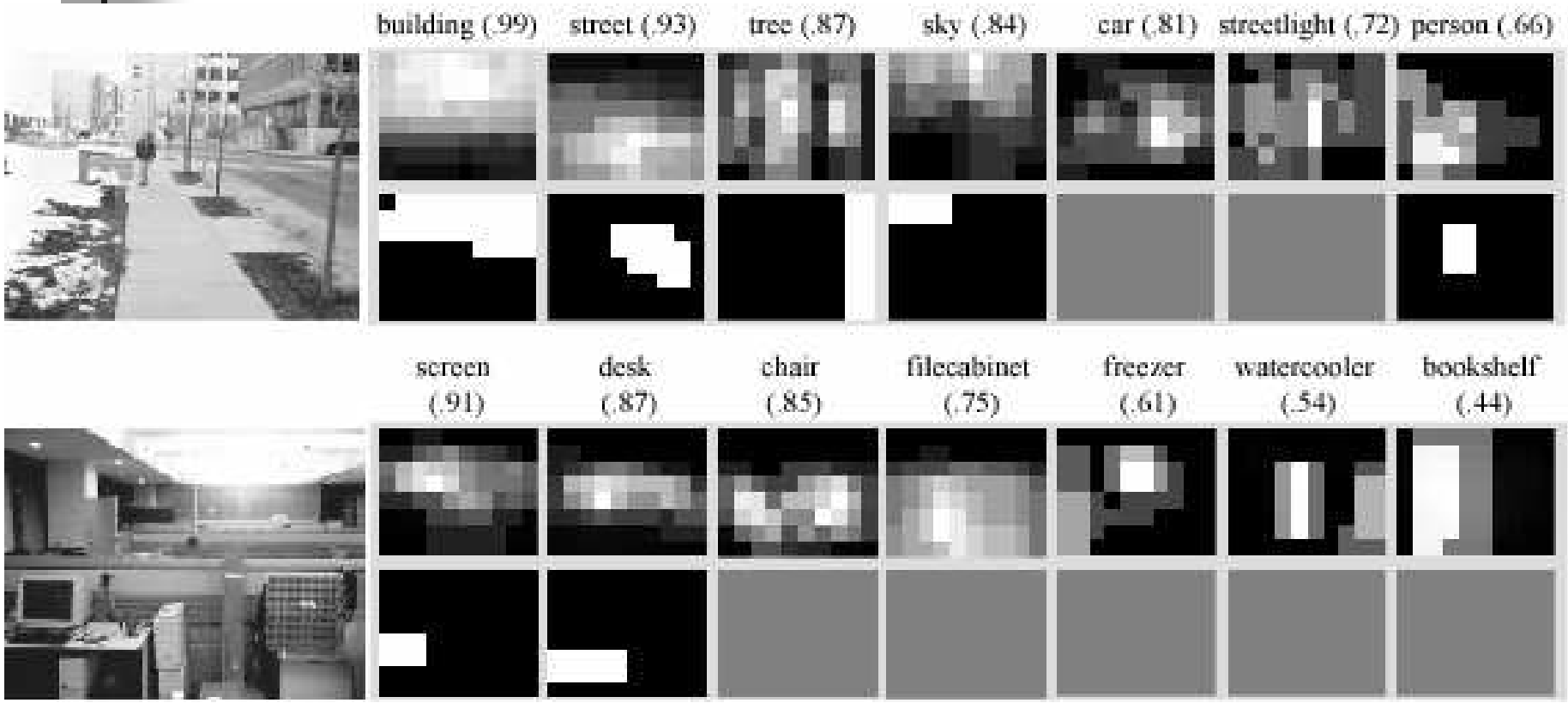
$$E[M_{t,i}|v_{1:t}^G] = \sum_{o\in\{0,1\}} \sum_q P(O_{t,i} = o, Q_t = q|v_{1:t}^G) \times E[M_{t,i}|v_t^G, Q_t = q, O_{t,i} = o]$$

  - Likelihood is weighted sum of prototypes

$$E[M_{t,i}|v_t^G] = \sum_q \sum_k w_{k,i,q} \times \mu_{k,i,q}^m$$

A. Torralba et al., “Context-based vision system for place and object recognition,” ICCV 2003.

Results for Object Localization



A. Torralba et al., “Context-based vision system for place and object recognition,” ICCV 2003.

Other Models of Context

- Using forest to see the trees: graphical model, K. Murphy 2003.
- Boosted random field, K. Murphy 2004.

Summary

- We introduced basic concepts of context and its role in human visual perception.
- We also discussed some recent probabilistic approaches to application of context in computer vision.
  - Matching words and pictures
  - Statistical context priming for object recognition
- Context is important in object recognition, especially when visibility of object appearance is low. It reduces ambiguity, search space, and also make a complementary role to the local approach.

References

- Antonio Torralba and Pawan Sinha, "Statistical context priming for object detection," ICCV 2001.
- Antonio Torralba, "Contextual priming for object detection," IJCV 2003.
- Antonio Torralba et al., "Context-based vision system for place and object recognition," ICCV 2003.
- Kevin Murphy et al., "Using the forest to see the trees: a graphical model relating features, objects, and scenes," NIPS 2003.

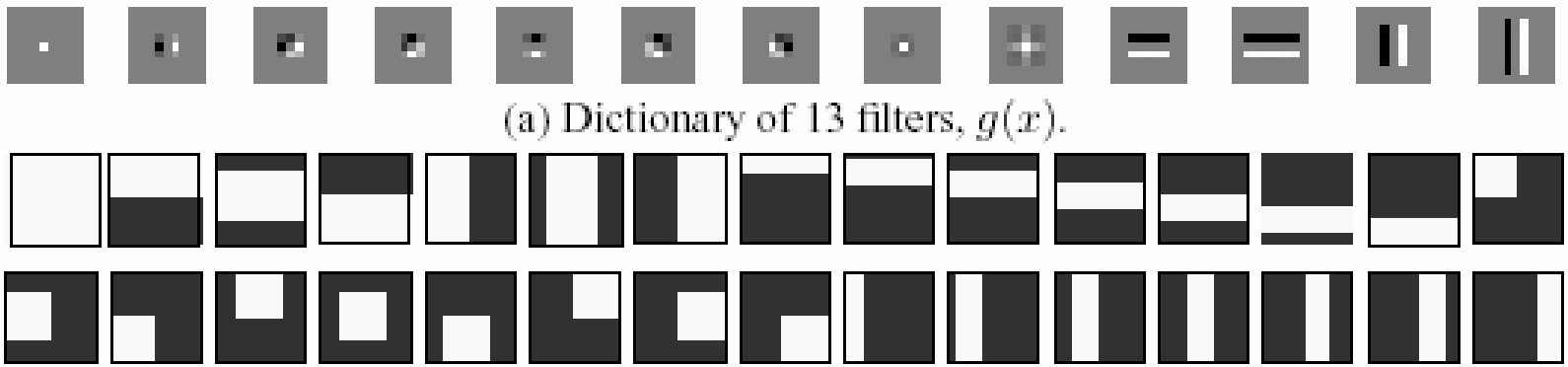
Graphical Model Relating Features, Objects, and Scenes

- Combining local, bottom-up information with global, top-down information using a graphical model
  - Boosting-based object detection
  - Scene learning and classification
  - Joint classification and detection

Kevin Murphy et al., “Using the forest to see the trees: a graphical model relating features, objects, and scenes,” NIPS 2003.

Filters and Spatial Templates

$$f_i(k) = \sum_x w_k(x) \left(|I(x) * g_k(x)|^{\gamma_k}\right)_i \quad \gamma_k \in \{2, 4\}$$



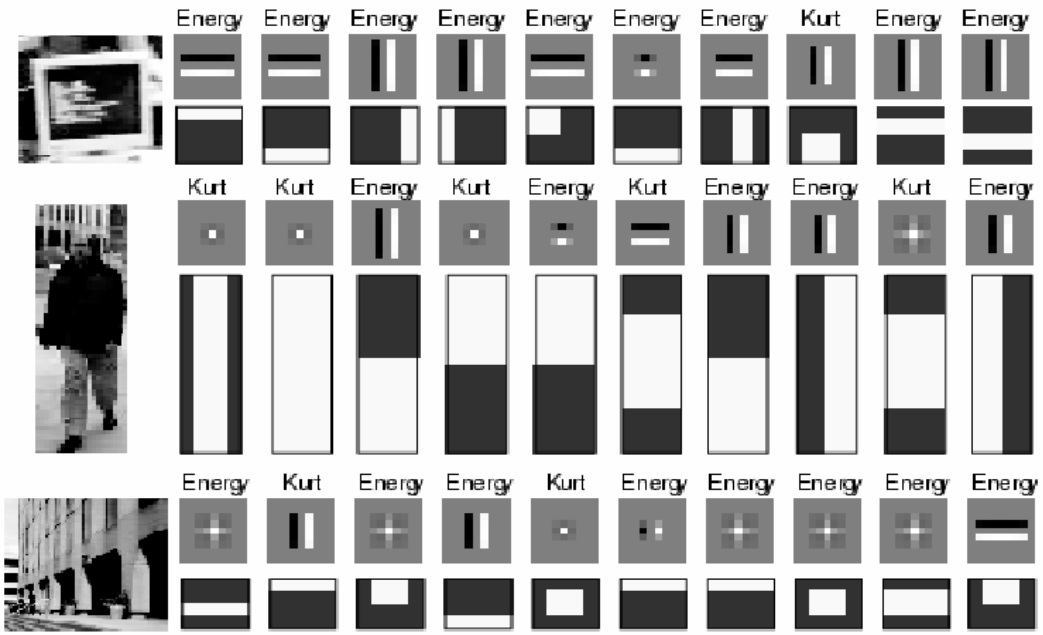
(a) Dictionary of 13 filters, $g(x)$.

(b) Dictionary of 30 spatial templates, $w(x)$.

Kevin Murphy et al., “Using the forest to see the trees: a graphical model relating features, objects, and scenes,” NIPS 2003.

Examples of Learned Features

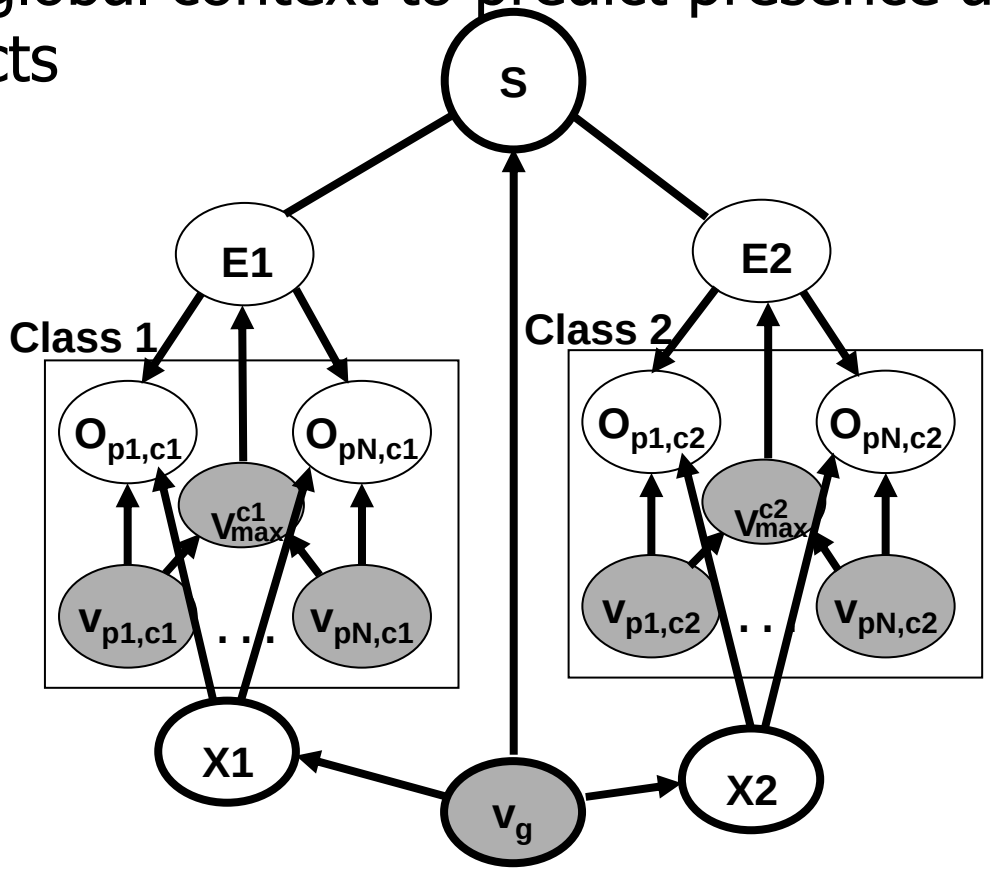
- Some features after 100 rounds of boosting



Kevin Murphy et al., “Using the forest to see the trees: a graphical model relating features, objects, and scenes,” NIPS 2003.

Tree-structured Graphical Model for Joint Classification and Detection

- Use global context to predict presence and location of objects



Kevin Murphy et al., “Using the forest to see the trees: a graphical model relating features, objects, and scenes,” NIPS 2003.

Backup Slides

Steerable Pyramid

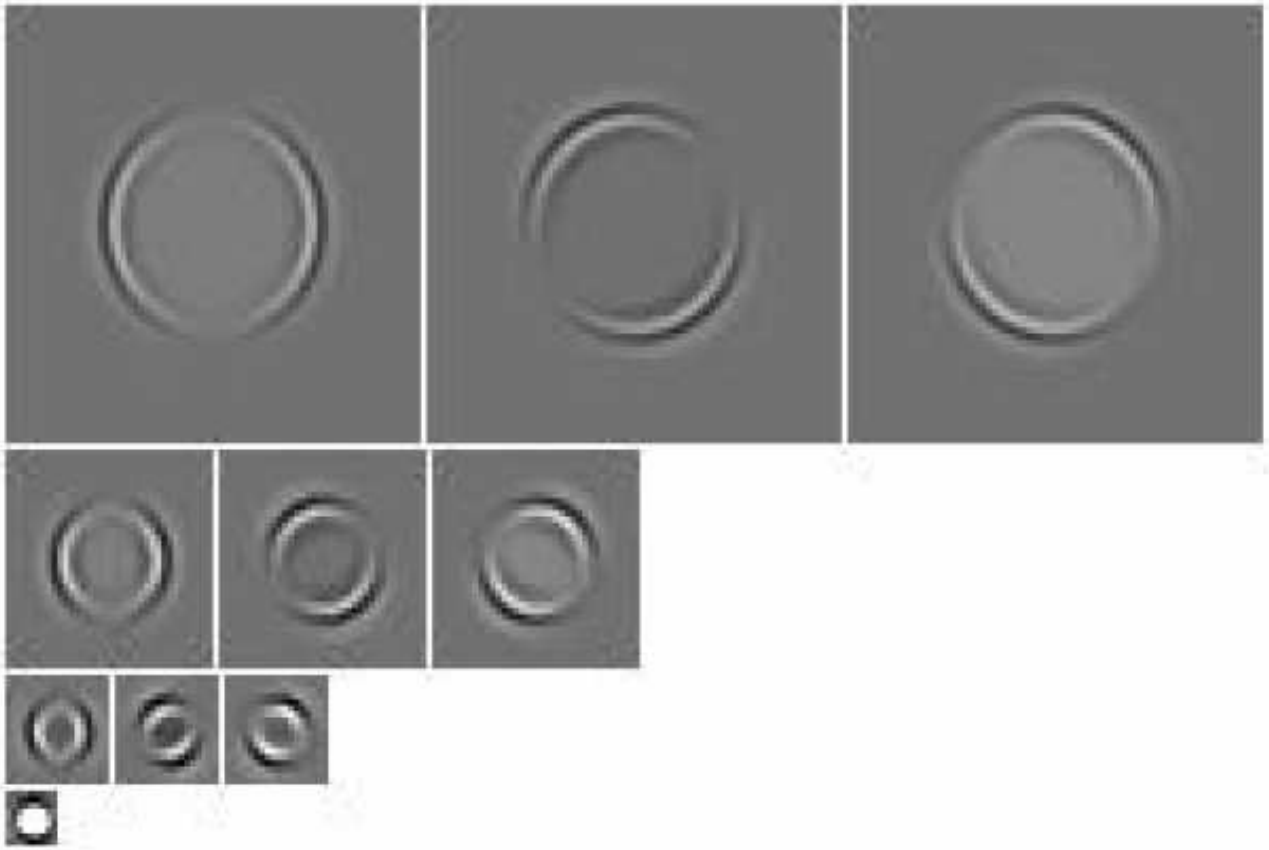


Image from Eero P Simoncelli & W.T. Freeman 1995

Weighted Fourier Transform

- WFT

$$I(x, y, f_x, f_y) = \sum_{x',y'=0}^{N-1} i(x', y')\, h_r(x' - x, y' - y)\, e^{-j\,2\pi(f_x x' + f_y y')}$$

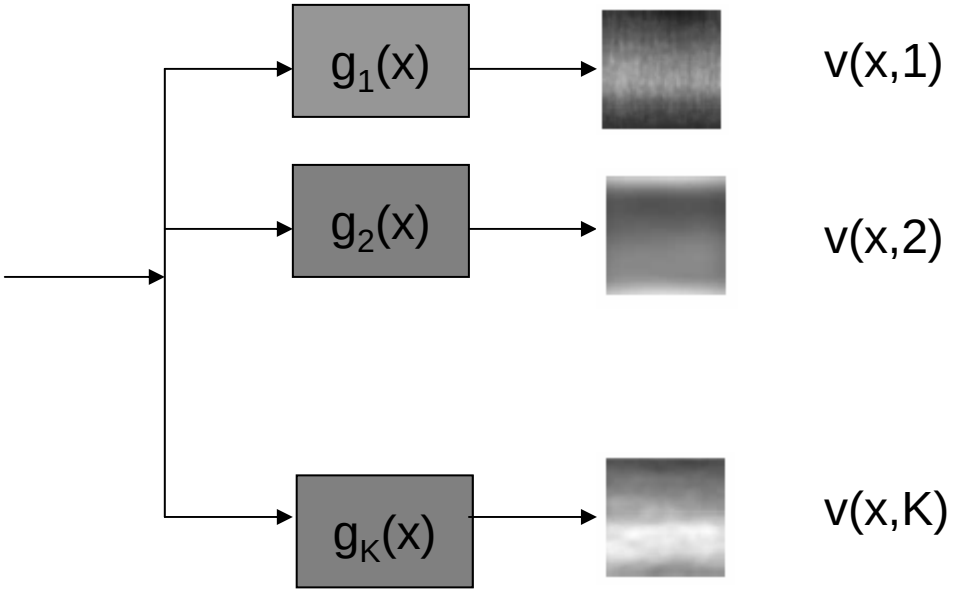
$$A(x, y, f_x, f_y) = \frac{|I(x, y, f_x, f_y)|}{I(x, y, 0, 0)\; std(x, y, f_x, f_y)}$$ Normalization

$$A(x, y, f_x, f_y) \simeq \sum_{n=1}^{N} a_n \psi_n(x, y, f_x, f_y)$$ PCA with N=60 PCs

$$\boxed{\vec{v}_C = \{a_n\}_{n=1,N}} \longrightarrow$$ Context features

Gabor Filter Bank

- Gabor filter bank



$$v(x,k) = \left| \sum_{x'} I(x') g_k(x - x') \right|$$

Magnitude of averaged filter output

$$g_k(x) = g_0 \cdot e^{-\frac{\|x\|^2}{\sigma_k^2}} \cdot e^{2\pi i < f_k, x >}$$

Gabor (oriented band-pass) filters

Images by T. Karra