SIFT and Object Recognition

Dan O'Shea
Prof. Fei Fei Li, COS 598B

Distinctive image features from scale-invariant keypoints

David Lowe. *International Journal of Computer Vision*, 2004.

Towards a Computational Model for Object Recognition in IT Cortex

David Lowe. *Proceedings of the First IEEE international Workshop on Biologically Motivated Computer Vision*, 2000.

Detectors vs. Descriptors

Challenge: Computationally inefficient to characterize entire image

Detectors: Find key points of interest which most distinctly identify the target object

Descriptors: Characterize the image around each key point in an invariant fashion

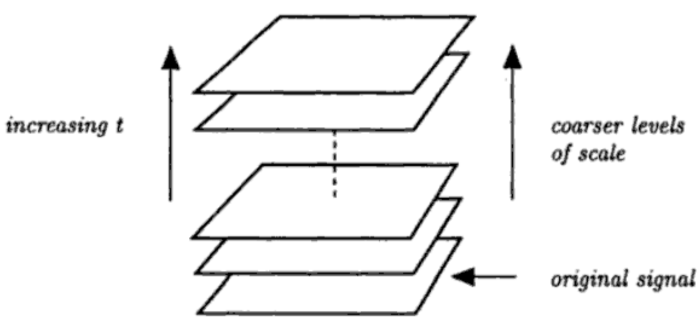
Lowe’s techniques encompass both!

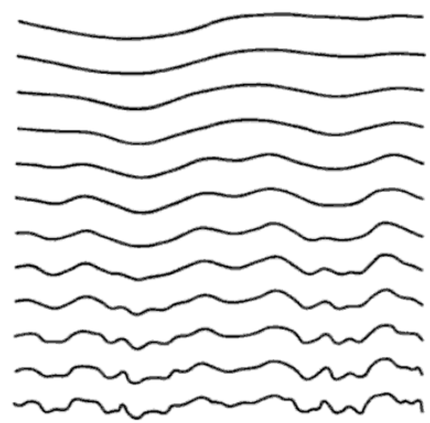
SIFT Features

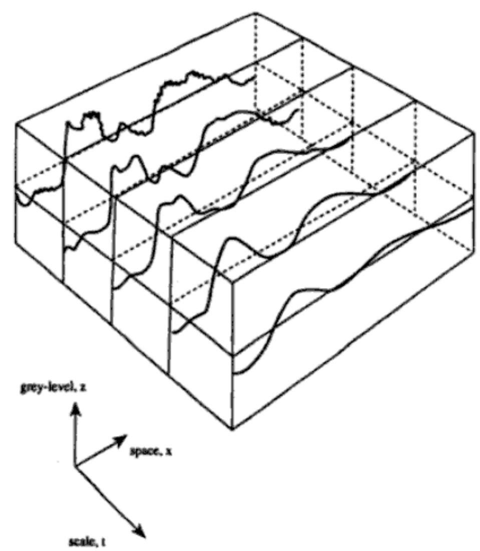
- Localize stable key points in scale space
- Perform feature detection only relative to canonical scale and orientation
- Emphasize local image gradient orientation, allow for small shift in position (like complex cells)

Scale-Space Theory

- Multi-scale signal representation
- Achieved via smoothing operation
- Gaussian kernel is unique in that increasing the width monotonically blurs fine detail







Source: Lindenberg, 1994.

Keypoint Detection

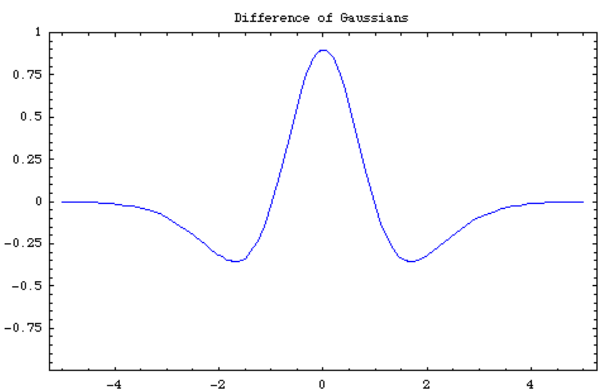
- Precompute pyramid of Gaussian filtered images at increasingly coarse scales
- Downsample by 2 each octave before convolution

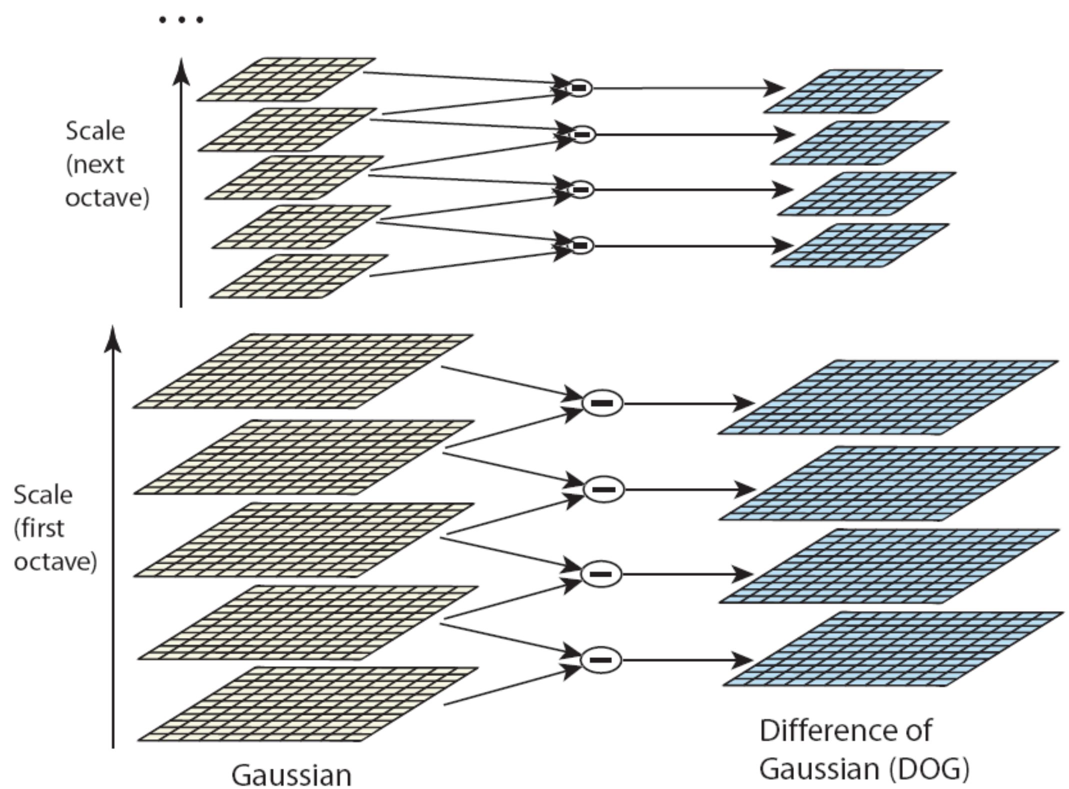
$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y),$$

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2}$$

Locating Keypoints

- Stability --> Must be reliably assigned
- Difference of Gaussians to find edges

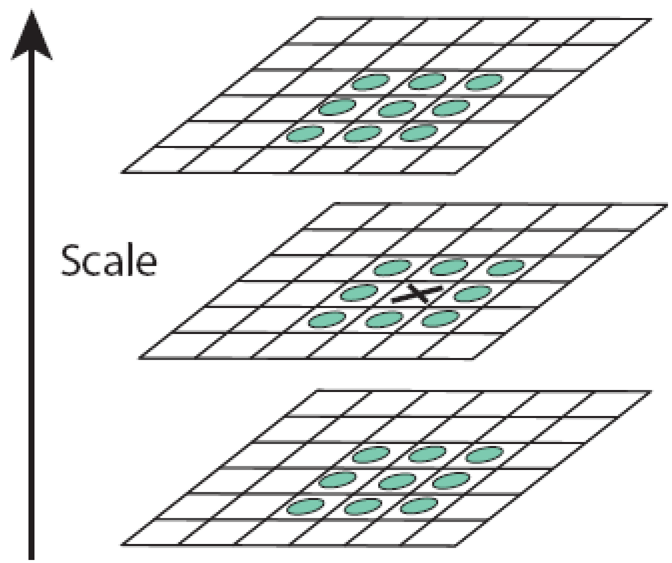




Difference of Gaussians

Scale-space Extrema

- Find points which are extrema within surrounding 3x3 cube (26 neighbors)

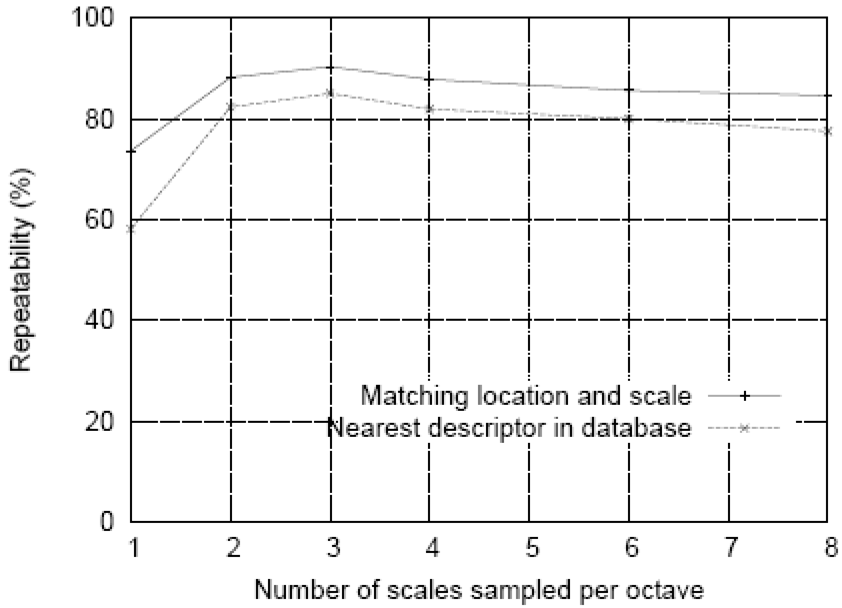


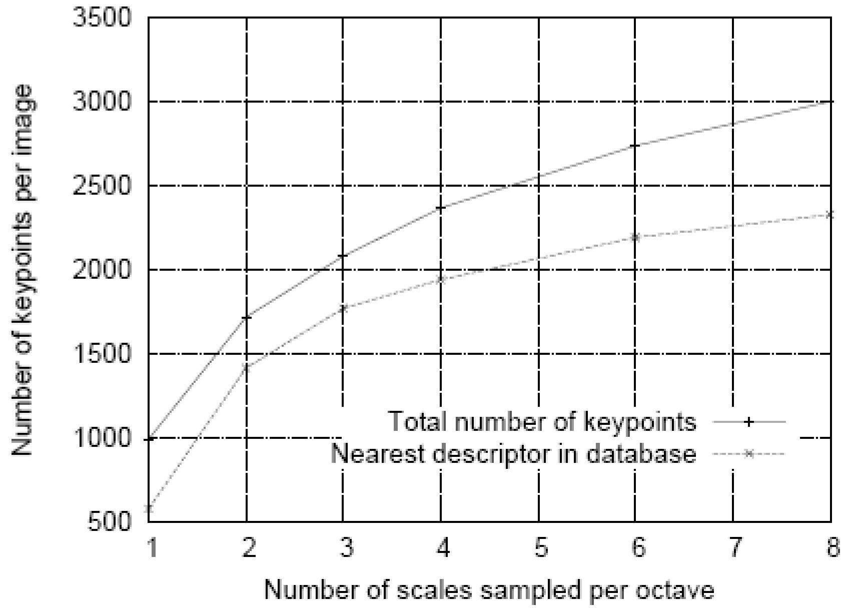
Sampling Frequency

- Extrema can be arbitrarily close together, but may be sensitive to small perturbations
- Test keypoint reliability across rotation, scaling, stretch, brightness, contrast, and in the presence of additive noise

Scale Sampling

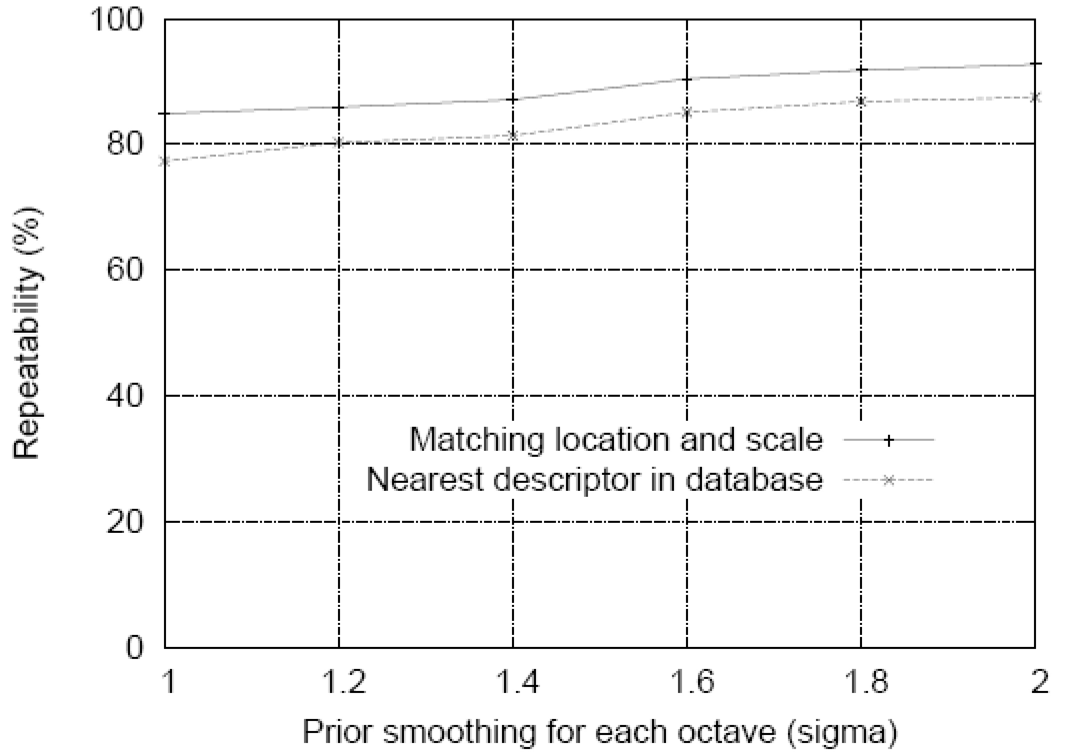
- 3 scales/octave empirically chosen





Spatial Sampling

- $\sigma = 1.6$ empirically chosen



Keypoint Localization

- Fit 3D quadratic function to DoG space magnitudes to interpolate extrema locations

$$D(\mathbf{x}) = D + \frac{\partial D}{\partial \mathbf{x}}^T \mathbf{x} + \frac{1}{2}\mathbf{x}^{\mathbf{T}} \frac{\partial^2 D}{\partial \mathbf{x}^2} \mathbf{x}$$

$$\hat{\mathbf{x}} = -\frac{\partial^2 D}{\partial \mathbf{x}^2}^{-1} \frac{\partial D}{\partial \mathbf{x}}$$

Low Contrast Rejection

- Points with low contrast are sensitive to noise
- Calculate DoG Value at extremum, disgard all below threshold as having low contrast

$$D(\hat{\mathbf{x}}) = D + \frac{1}{2}\frac{\partial D}{\partial \mathbf{x}}^T \hat{\mathbf{x}}$$

Edge Response Rejection

- Locations along edges are poorly determined and very sensitive to noise
- Use principal curvature: direction along edge large, orthogonal to edge weak

$$\mathbf{H} = \begin{bmatrix} D_{xx} & D_{xy} \\ D_{xy} & D_{yy} \end{bmatrix}$$

$$\mathrm{Tr}(\mathbf{H}) = D_{xx} + D_{yy} = \alpha + \beta,$$

$$\mathrm{Det}(\mathbf{H}) = D_{xx}D_{yy} - (D_{xy})^2 = \alpha\beta$$

$$\frac{\mathrm{Tr}(\mathbf{H})^2}{\mathrm{Det}(\mathbf{H})} = \frac{(\alpha + \beta)^2}{\alpha\beta} = \frac{(r\beta + \beta)^2}{r\beta^2} = \frac{(r+1)^2}{r}$$

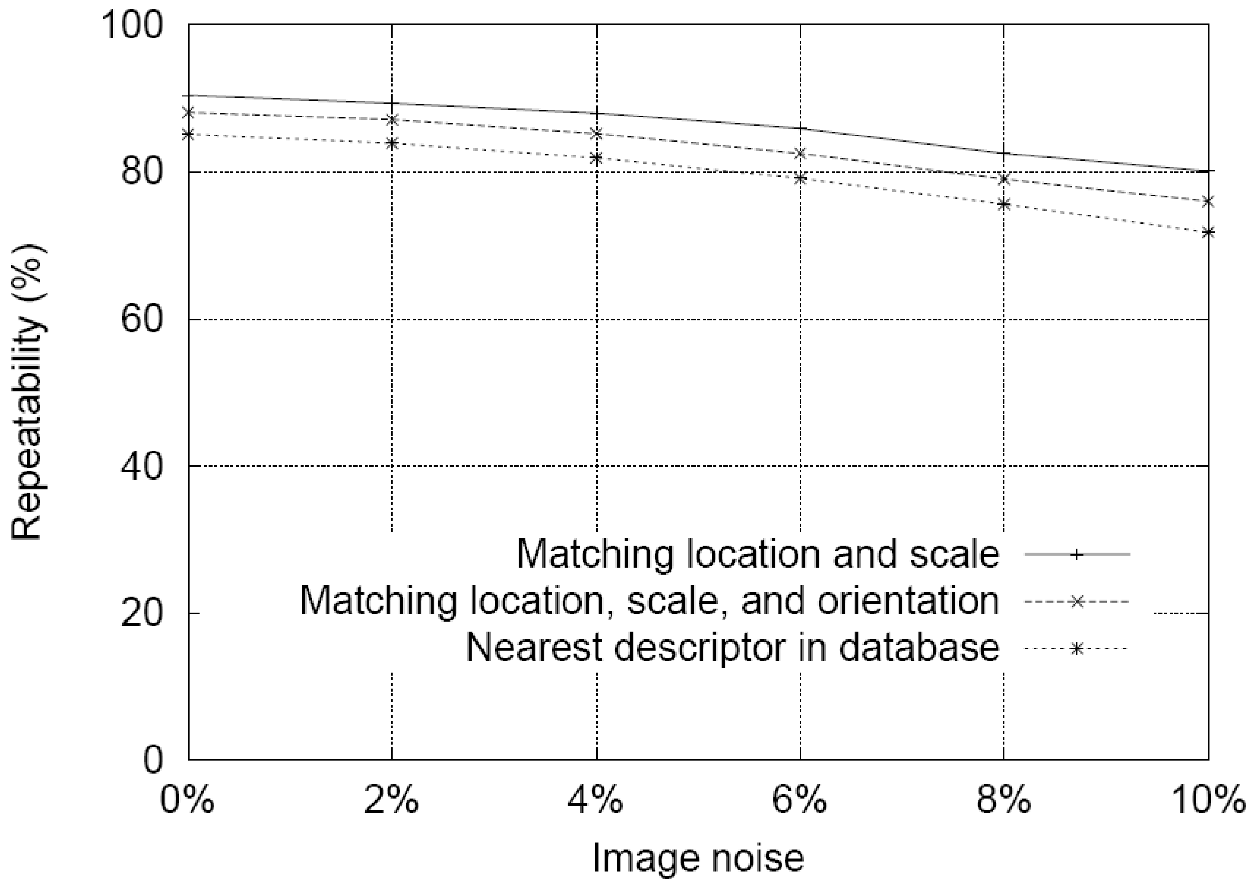
$$\frac{\mathrm{Tr}(\mathbf{H})^2}{\mathrm{Det}(\mathbf{H})} < \frac{(r+1)^2}{r}$$

Orientation Assignment

- Assign orientation to each keypoint based on local image properties
- Construct weighted gradient orientation histograms about each keypoint at closest scale
- Create keypoint with orientation at each major peak in histogram (> 80% of maximum)

Orientation Reliability

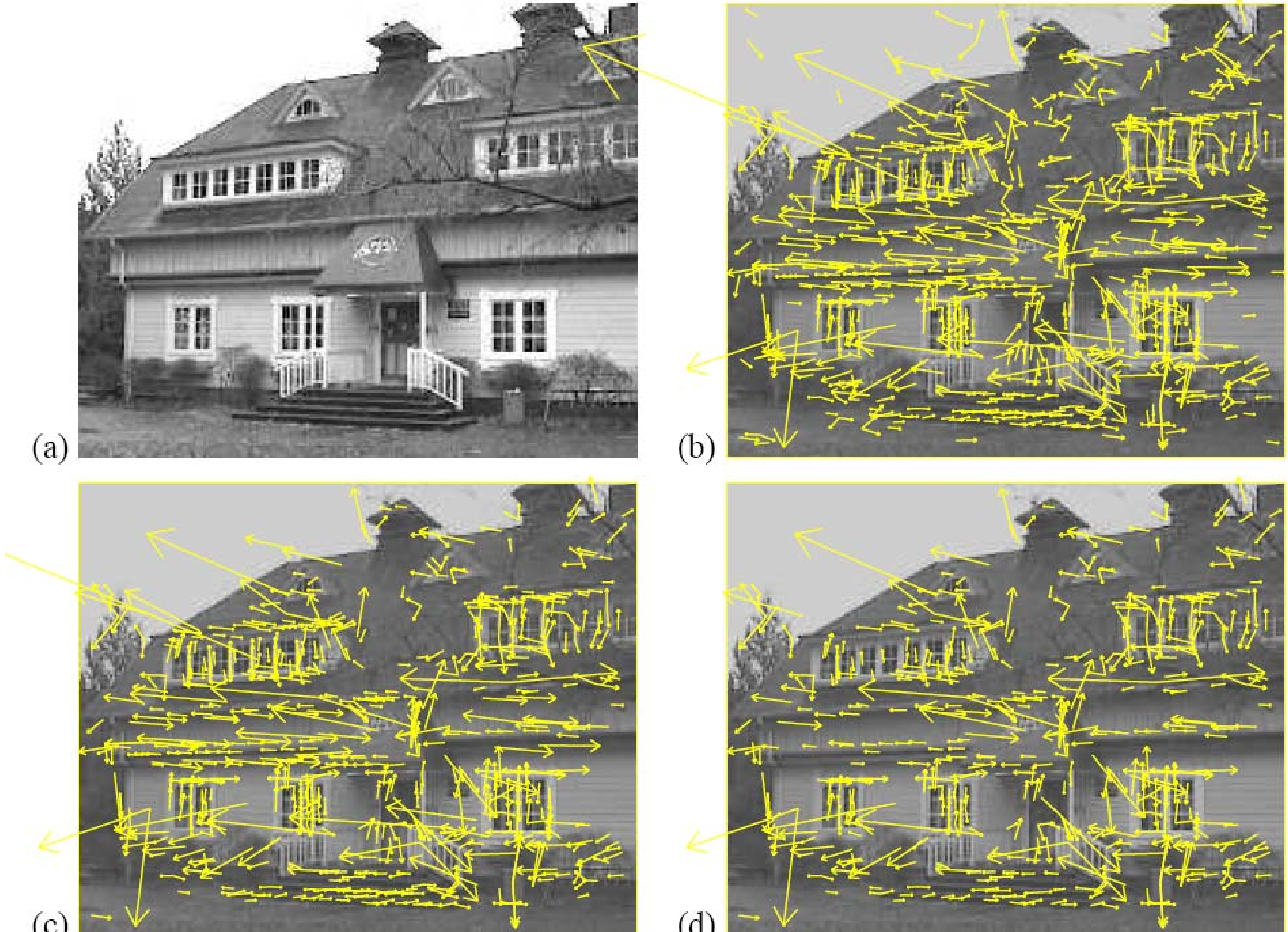
- Orientation more reliable than location/scale



Keypoint Example

Original

Initial
Keypoint
s



Low
Contrast
Rejection

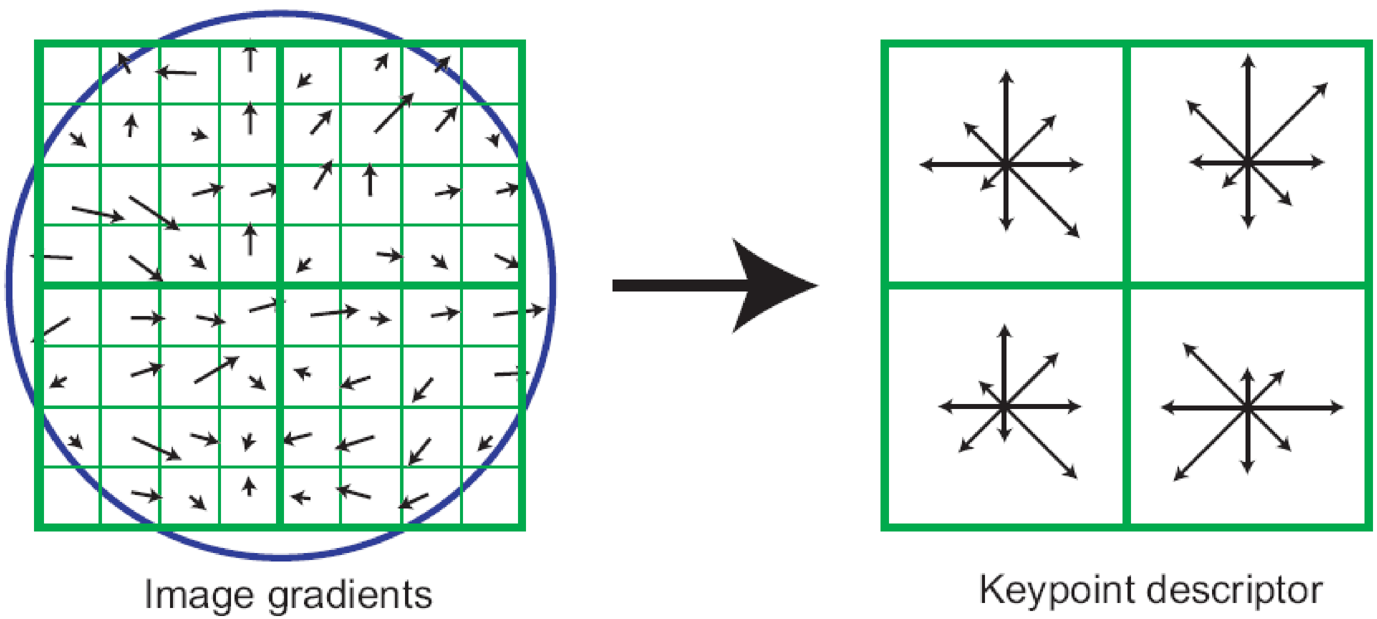
Principal
Curvatur
e
Threshol
d

Local Image Descriptor

- Image Patch Technique – store pixel intensities surrounding keypoints, use simple correlations for comparison
  - Sensitive to affine and 3d viewpoint changes
- Local Gradient Technique – record surrounding gradients, allow for some spatial translation
  - Based off complex neuron responses

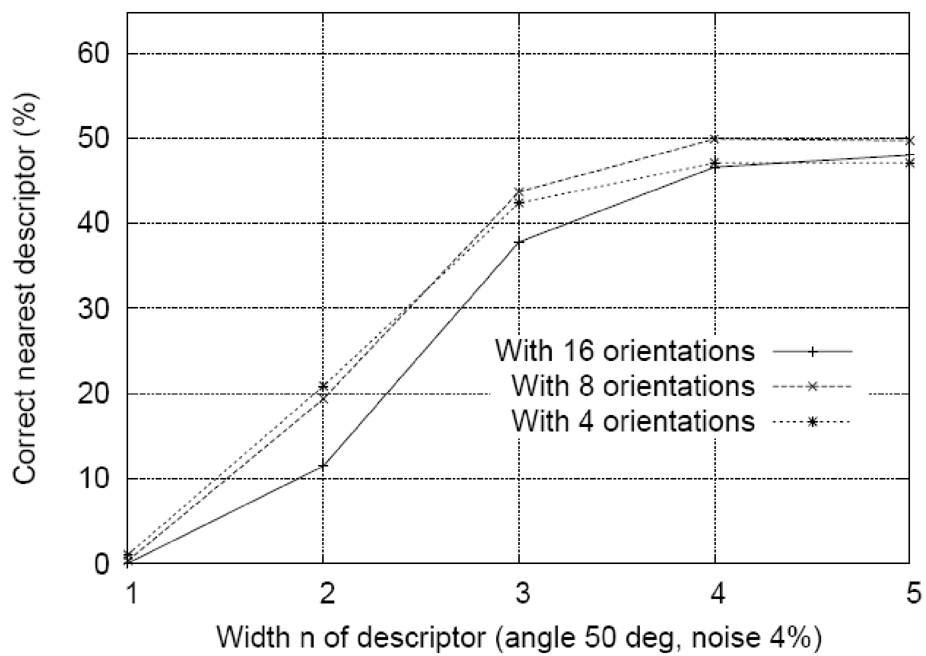
Gradient Histograms

- Sample gradient magnitude orientation (relative to keypoint orientation) in 16x16 window around key
- Intelligently arrange into 4x4 histograms with 8 bins



Descriptor Size

- R bins * N^2 sample grid: $R*N^2$ element vector
- Used 4x4 grid, 8 orientation bins: 128 element vector



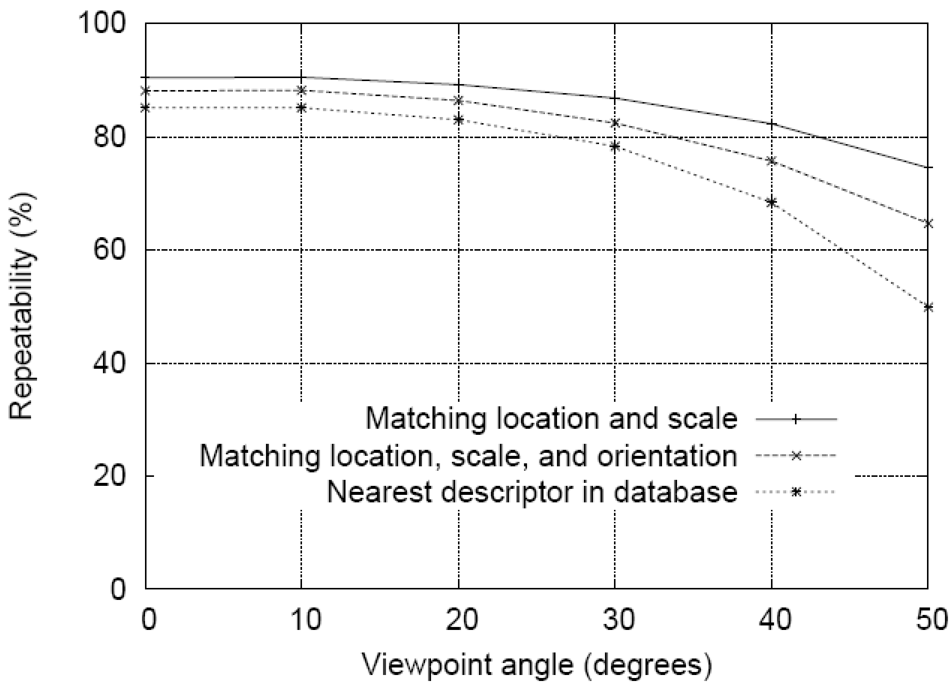
At 4x4:
8 best, 16 worst

Descriptor Subtleties

- **Gradients far from keypoint less reliable:**
  - Use Gaussian kernel to weight magnitudes
- **Boundary effects at 4x4 grid division:**
  - Use trilinear interpolation to distribute across bins/histograms
- **Contrast Changes:** normalize to unit length
- **Illumination saturations:** affect large gradient magnitudes but not orientations
  - Saturate large magnitudes, emphasize orientation

3D Viewpoint Angle Performance

- 50% Reliability out to 50 degree rotation in depth
- Could simply store SIFT features for multiple model views independently

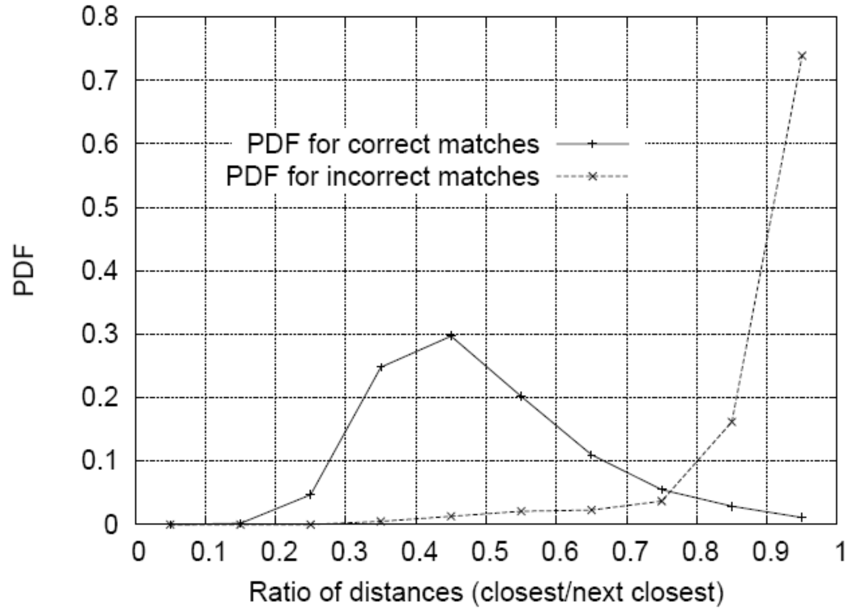


Object Recognition Overview

- Store SIFT vectors for each keypoint for each model object in database
- Generate keypoints in test image
- Use nearest neighbor to find feature matches
- Cluster features that agree on object pose
- Affine projection estimate
- Geometric verification

Keypoint Matching

- Similarity metric is Euclidean distance
- Global thresholds work poorly as discriminative ability of descriptors varies: use ratio of 1st to 2nd closest neighbors
- Best-Bin First: approximate NN search algorithm



Keypoint Clustering

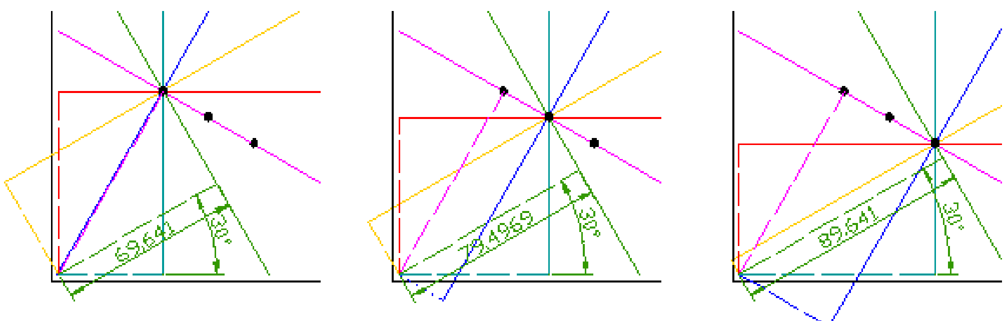
- Find groups of keypoint matches that agree on an object and its pose (location, orientation, scale)
- Each match casts a 4-element vote, tally in histogram, select clusters
- Accomplished with Hough transform and hash table

Reliable object detection with only 3 feature matches!

Hough Transform Example

- Application: detecting lines in the 2d plane
- Find point closest to origin (intersection by orthogonal), describe by radius and a

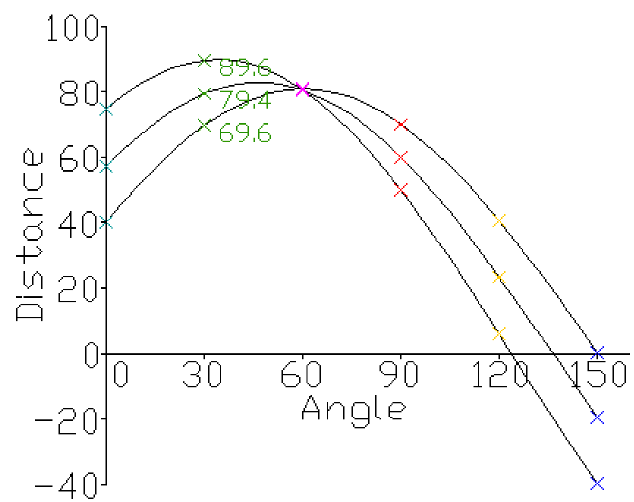
$$y = \left(-\frac{\cos\theta}{\sin\theta}\right)x + \left(\frac{r}{\sin\theta}\right)$$



Angle	Dist.
0	40
30	69.6
60	81.2
90	70
120	40.6
150	0.4

Angle	Dist.
0	57.1
30	79.5
60	80.5
90	60
120	23.4
150	−19.5

Angle	Dist.
0	74.6
30	89.6
60	80.6
90	50
120	6.0
150	−39.6



Source: Wikipedia

Affine Transformation Estimate

- Least-squares fit to affine projection from model to test image coordinates

$$\begin{bmatrix} u \\ v \end{bmatrix} = \begin{bmatrix} m_1 & m_2 \\ m_3 & m_4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}$$

$$\begin{bmatrix} x & y & 0 & 0 & 1 & 0 \\ 0 & 0 & x & y & 0 & 1 \\ & & \cdots & & & \\ & & \cdots & & & \end{bmatrix} \begin{bmatrix} m_1 \\ m_2 \\ m_3 \\ m_4 \\ t_x \\ t_y \end{bmatrix} = \begin{bmatrix} u \\ v \\ \vdots \end{bmatrix}$$

$$\mathbf{A}\mathbf{x} = \mathbf{b}$$

$$\mathbf{x} = [\mathbf{A}^{\mathbf{T}}\mathbf{A}]^{-1}\mathbf{A}^{\mathbf{T}}\mathbf{b}$$

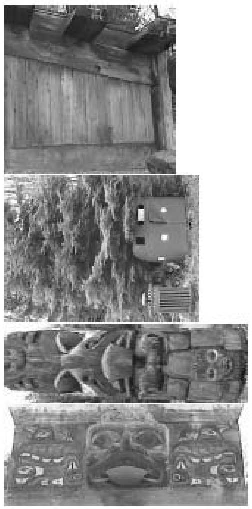
Geometric Verification

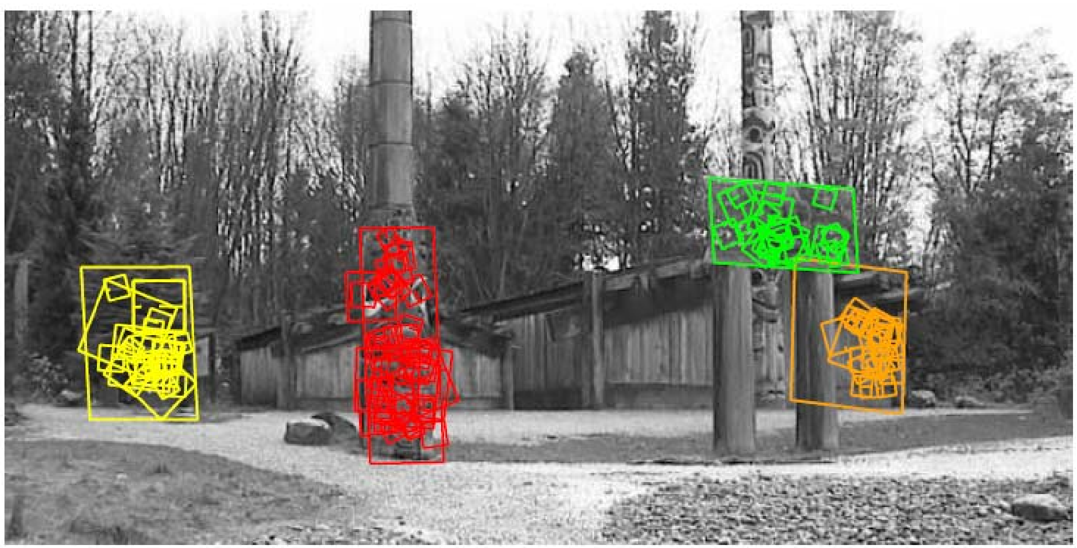
- Calculate residual error from least-squares fit, reject outliers above threshold
- Repeat fit, add features that agree with new estimate
- Recognition fails if less than 3 features remain
- Final decision based on probabilistic learning model described in Lowe, 2001 (maximum-likelihood)

Recognition in Occlusion

Recognition in Occlusion (2)

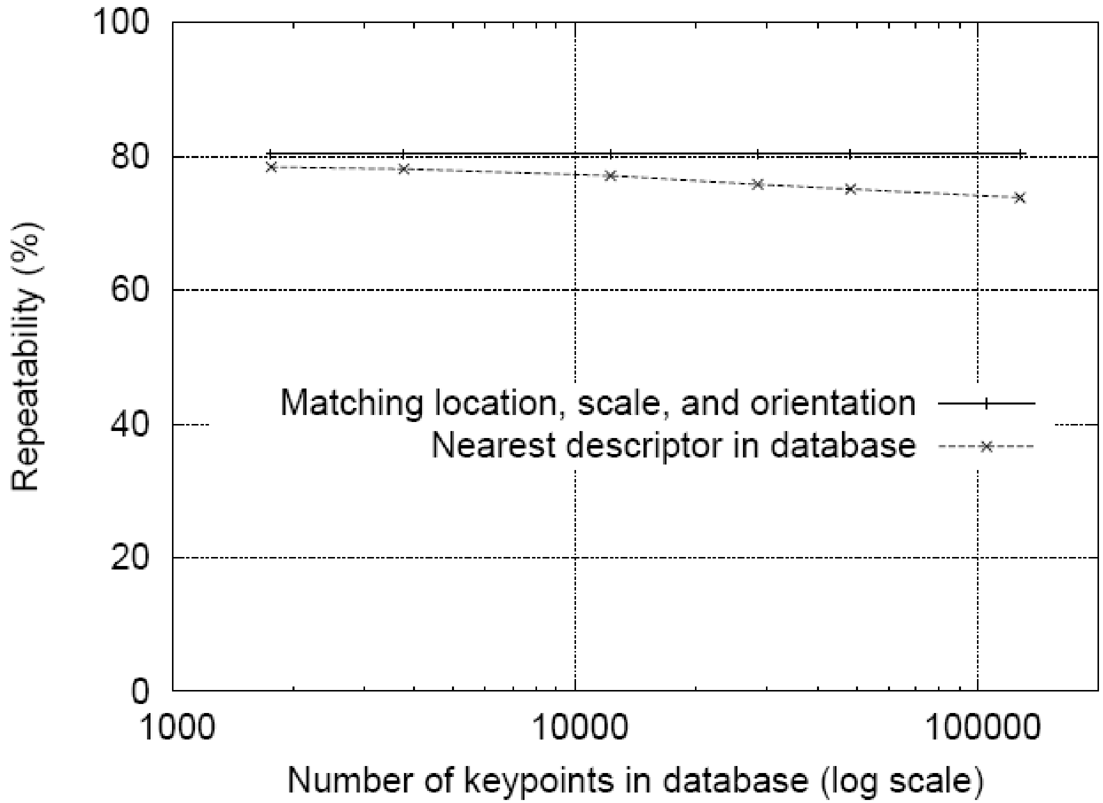
Recognition in Complex Scenes





Large Database Performance

- Nearest Neighbor matching with Euclidean distance
- Perf izes



Future Directions

- Full 3D viewpoint representation (4D to 6D pose)
- Better invariance to nonlinear illumination changes
- Extension to 3 channel color
- Inclusion of local texture measures
- Class-specific features for categorization
- Edge groupings at object boundaries

Binding and Attention

- **Humans:**
  - – Detect features in parallel
  - – Serial attention required to bind features to object, determine pose, and segregate background
- **SIFT:**
  - – Detect keypoints and compute features in parallel
  - – Hough transform binds features to object
  - – Probabilistic EM framework optimizes decision

Conclusions

- SIFT finds stable keypoints in scale-space at suitable difference of Gaussian extrema
- Local descriptor invariant to: scale, invariance, affine transformations, brightness, contrast
- Computationally efficient
- Requires labeled, clutter-free model images

Bottom-Up Attention?

Is bottum-up attention useful for object recognition?

Ueli Rutishauser, Dirk Walther, Cristof Koch, and Pietro Perona. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2004.

- Attention: selection and gating of visual information
  - Top-down: prior knowledge about the scene
  - Bottom-up: saliency in image
- Idea: use bottom-up attention to highlight regions where objects are likely to be found

Saliency Model

- Construct across-scale center-surround feature maps
- Use RGB [illegible] tion, intensity

Center-surround feature maps:

$$
\begin{aligned}
\mathcal{F}_{I,c,s} &= \mathcal{N}\left(|I(c) \ominus I(s)|\right) \\
\mathcal{F}_{RG,c,s} &= \mathcal{N}\left(|(R(c) - G(c)) \ominus (R(s) - G(s))|\right) \\
\mathcal{F}_{BY,c,s} &= \mathcal{N}\left(|(B(c) - Y(c)) \ominus (B(s) - Y(s))|\right) \\
\mathcal{F}_{\theta,c,s} &= \mathcal{N}\left(|O_\theta(c) \ominus O_\theta(s)|\right)
\end{aligned}
$$

Sum across maps: $\bar{\mathcal{F}}_l = \mathcal{N}\left(\bigoplus_{c=2}^{4} \bigoplus_{s=c+3}^{c+4} \mathcal{F}_{l,c,s}\right)$ with $l \in L_I \cup L_C \cup L_O$

$L_I = \{I\},\ L_C = \{RG, BY\},$
$L_O = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$

Conspicuity maps: $C_I = \bar{\mathcal{F}}_I,\ C_C = \mathcal{N}\left(\sum_{l \in L_C} \bar{\mathcal{F}}_l\right),\ C_O = \mathcal{N}\left(\sum_{l \in L_O} \bar{\mathcal{F}}_l\right)$

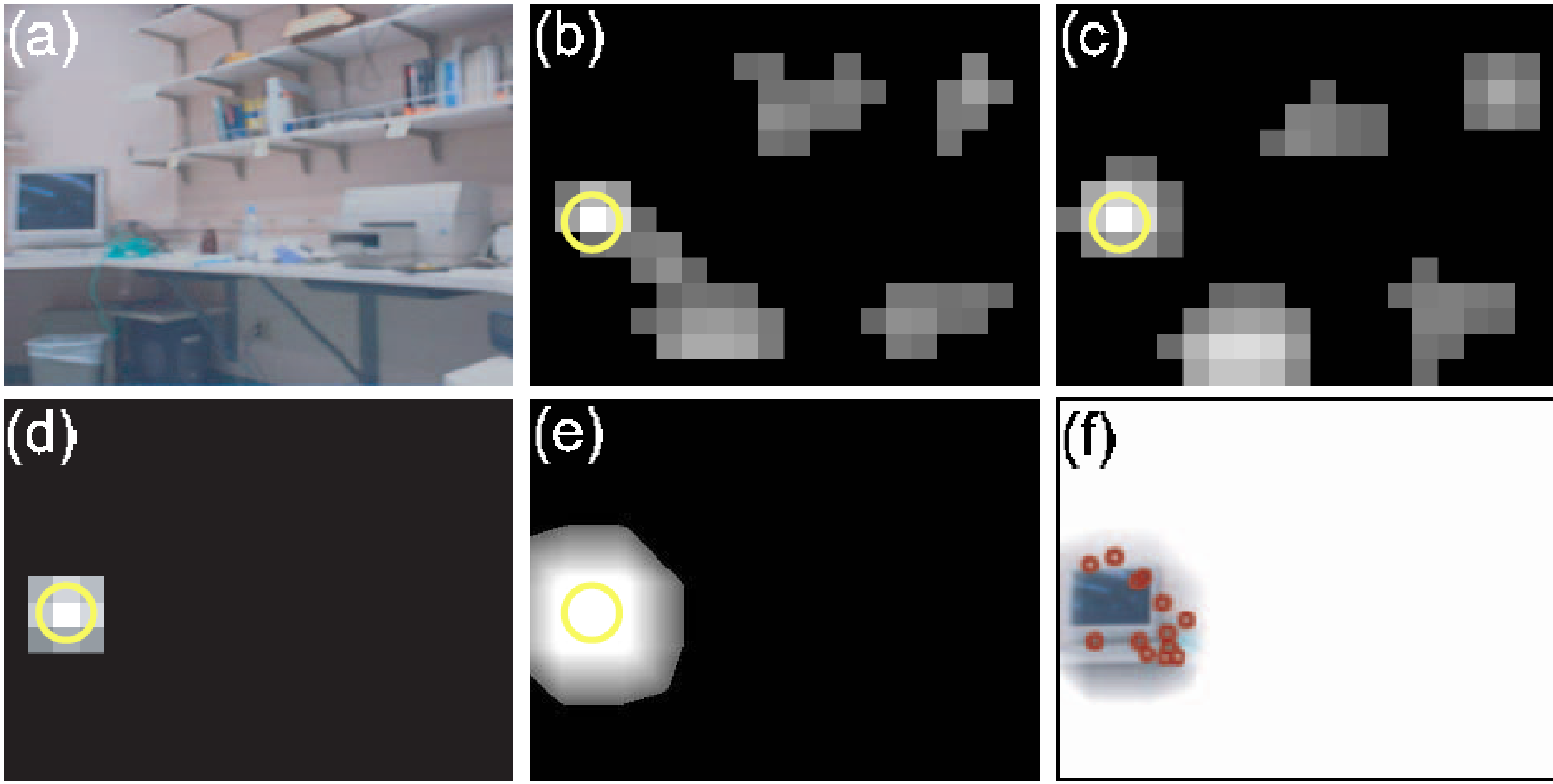
Saliency map: $\mathcal{S} = \frac{1}{3} \sum_{k \in \{I,C,O\}} C_k$ $\longrightarrow$ Winner Take All (WTA) Competition

Regions of Saliency

- WTA chooses most salient point (x_w, y_w)
- Use adaptive thresholding to grow region around point at feature map level (sparser representation)
- "Remove" influence within WTA competition → multiple salient regions

Use salient regions to train SIFT: unlabeled model images!

Saliency Example



Inventory Learning Example

Landmark Learning