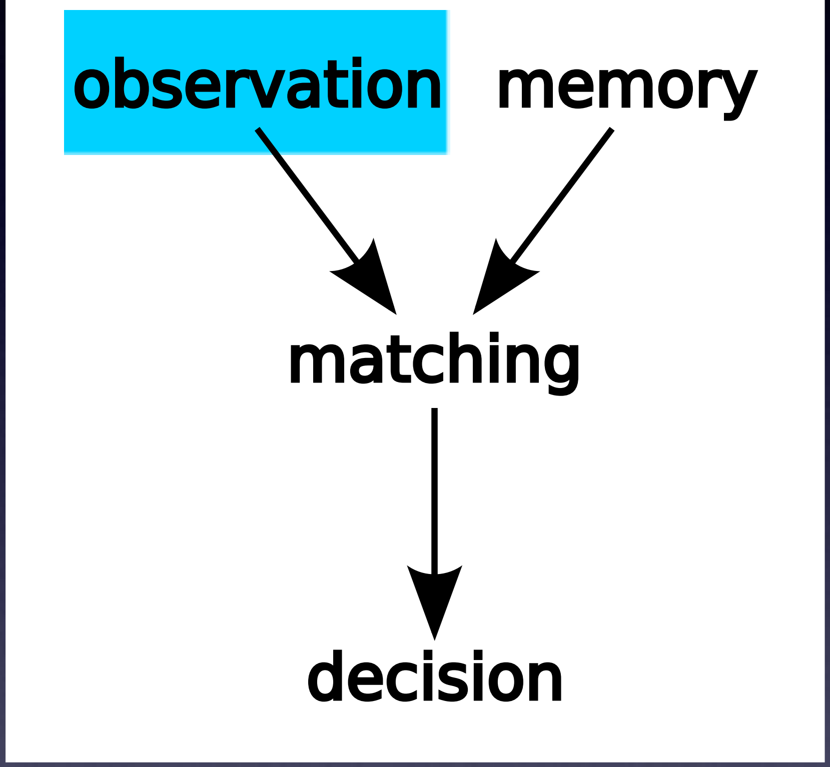
# Visual Object Recognition

## Lecture 4: Region Detectors

**Per-Erik Forssén, docent**
**Computer Vision Laboratory**
**Department of Electrical Engineering**
**Linköping University**

© 2015 Per-Erik Forssén

# Lecture 4: Region Detectors



- A region detector selects candidate regions for descriptor computation.
- Similar function to focus-of-attention system in HVS.
- An alternative is a sliding window that tests ***all*** regions.

© 2015 Per-Erik Forssén

# Lecture 4: Region Detectors

- Interest Point Detection
- Scale Selection
- Affine Adaptation
- Maximally Stable Regions
- Detection Noise

© 2015 Per-Erik Forssén

# Interest Point Detection

- In LE2 we saw how canonical frames can be found from groups of interest points (IP).
- This lecture is about estimation of a canonical frame from a region detector.
- Advantages:
  - smaller c-frames in image (better scale inv.)
  - higher frame repeatability

©2015 Per-Erik Forssén

# Interest Point Detection

- ***Repeatability*** of an IP detector (LE8)

$$\epsilon = P(\text{feature detected}|\text{feature present})$$

- feature group repeatability:

$$P(F_1 \cap F_2 \ldots F_N|\text{present}) = \epsilon^N$$

N - Number of feature points in canonical frame.

© 2015 Per-Erik Forssén

# Interest Point Detection

- Interest point groups are still useful in special applications:
  - **Bin-picking**, as they provide more accurate pose, F. Viksten, ”Local Features for Range and Vision-Based Robotic Automation”, LiU Thesis 2010
  - **Face recognition**, where a collection of detected facial landmarks are used. E.g. Tal Hassner, et al. ”Effective Face Frontalization in Unconstrained Images”, ArXiV’14

© 2015 Per-Erik Forssén

# Interest Point Detection

- A classical ”interest point detector” is the Harris/ Stephens corner detector.

- C. Harris, M. Stephens, ”A Combined Corner and Edge Detector”, Alvey Vision Conference’88

- Very similar ideas in:
  W. Förstner, E. Gülch, ”A fast operator for detection and precise location of distinct points, corners and centres of circular features”, ISPRS’87

© 2015 Per-Erik Forssén

# Interest Point Detection

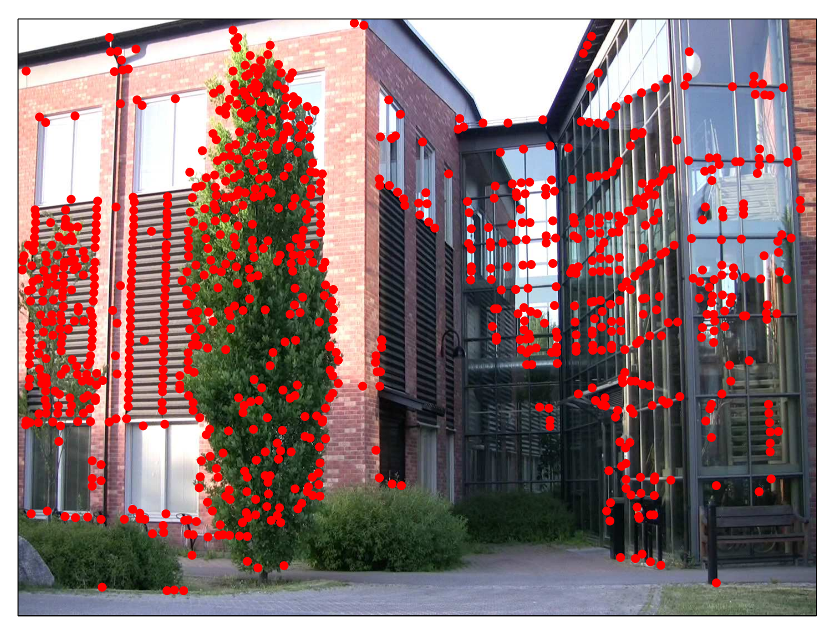
- C. Harris, M. Stephens, ”A Combined Corner and Edge Detector”, Alvey Vision Conference’88

1. Compute gradients: $\nabla f(\mathbf{x}) = (f * \begin{bmatrix} d_x \\ d_y \end{bmatrix})(\mathbf{x})$
2. Compute structure tensor: $\mathbf{T}(\mathbf{x}) = \left(\nabla f \nabla f^T * g\right)(\mathbf{x})$
3. Compute cornerness: $h(\mathbf{x}) = \det \mathbf{T}(\mathbf{x}) - \kappa \mathrm{tr}^2 \mathbf{T}(\mathbf{x})$
4. Detect local peaks by 3x3 non-max suppression on h(**x**).

© 2015 Per-Erik Forssén

# Interest Point Detection

- C. Harris, M. Stephens, ”A Combined Corner and Edge Detector”, Alvey Vision Conference’88
- Note that:
  - The Harris/Stephens detector does not just detect corners. Fires at all regions with non-simple structure.
  - As it uses an integration region, it is actually a region detector.
  - The intrinsic size of the integration region can be determined using scale selection.



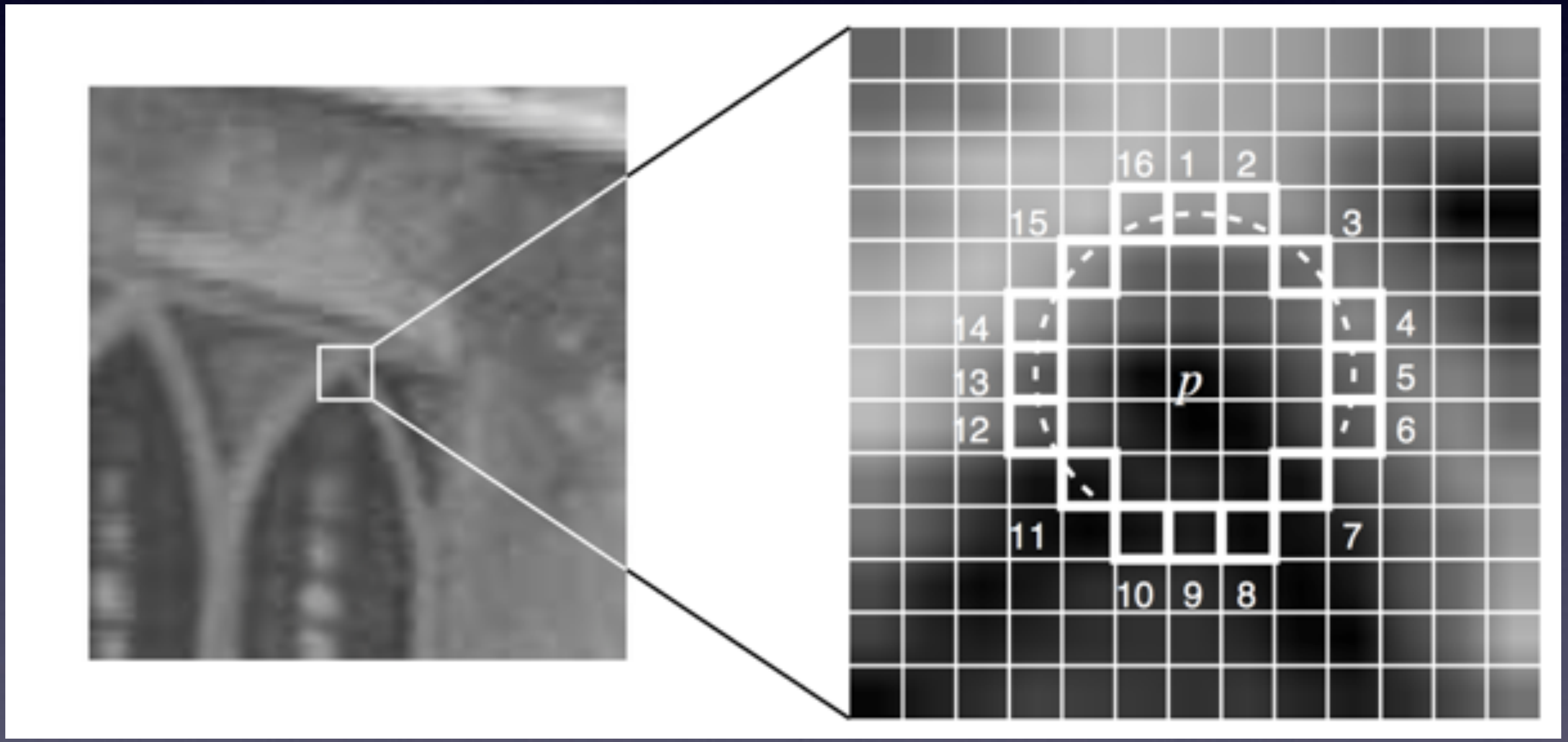
Harris output

Input image

© 2015 Per-Erik Forssén

# Interest Point Detection

- FAST: E. Rosten and T. Drummond. ”Machine learning for high-speed corner detection”. ECCV’06.



- Count contiguous sections of 9 pixels brighter/darker than centre pixel.

- Much faster than Harris corners, (1.3ms vs 24ms) ≈20x

- Not quite as repeatable.

© 2015 Per-Erik Forssén

# Interest Point Detection

- Difference of Gaussian points. DoG. Used in SIFT feature.
- Also fires along ridges.
  A second check for simple structure can help to eliminate these.
- Finds many more points than Harris and FAST.
- Similar in speed to Harris detector.
- Inherently uses scale space.

© 2015 Per-Erik Forssén

# Scale Space

$f(\mathbf{x}) \quad \Rightarrow \quad f_s(\mathbf{x},\sigma)$

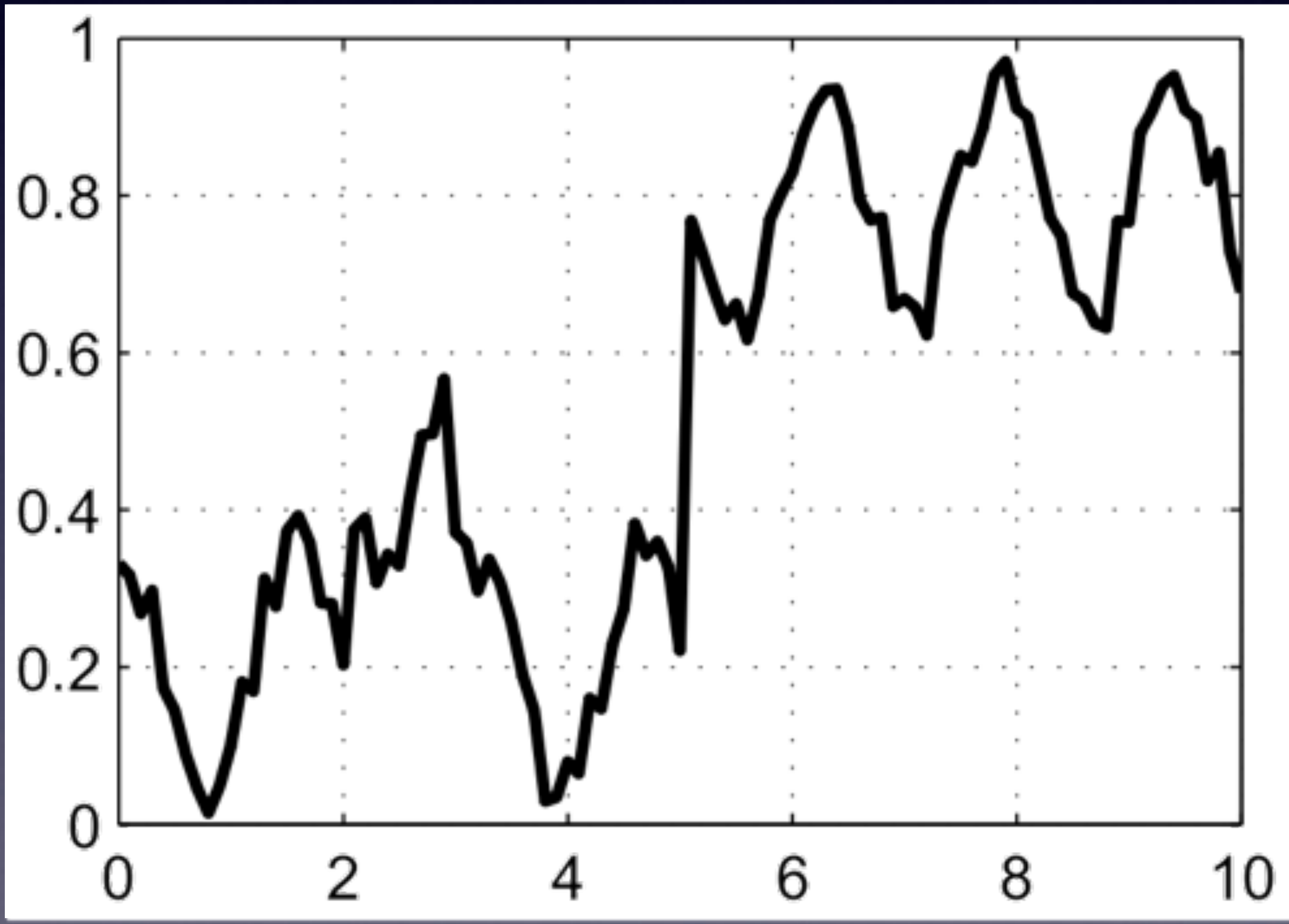
- Scale space $f_s(\mathbf{x},\sigma) = (f * g(\sigma))(\mathbf{x})$
  - The image is extended with an extra dimension, for scale/image blur. $g(\sigma)$
  - The blurring kernel $g(\sigma)$ is typically a Gaussian.

$$g(\mathbf{x},\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\mathbf{x}^T\mathbf{x}/2\sigma^2}$$

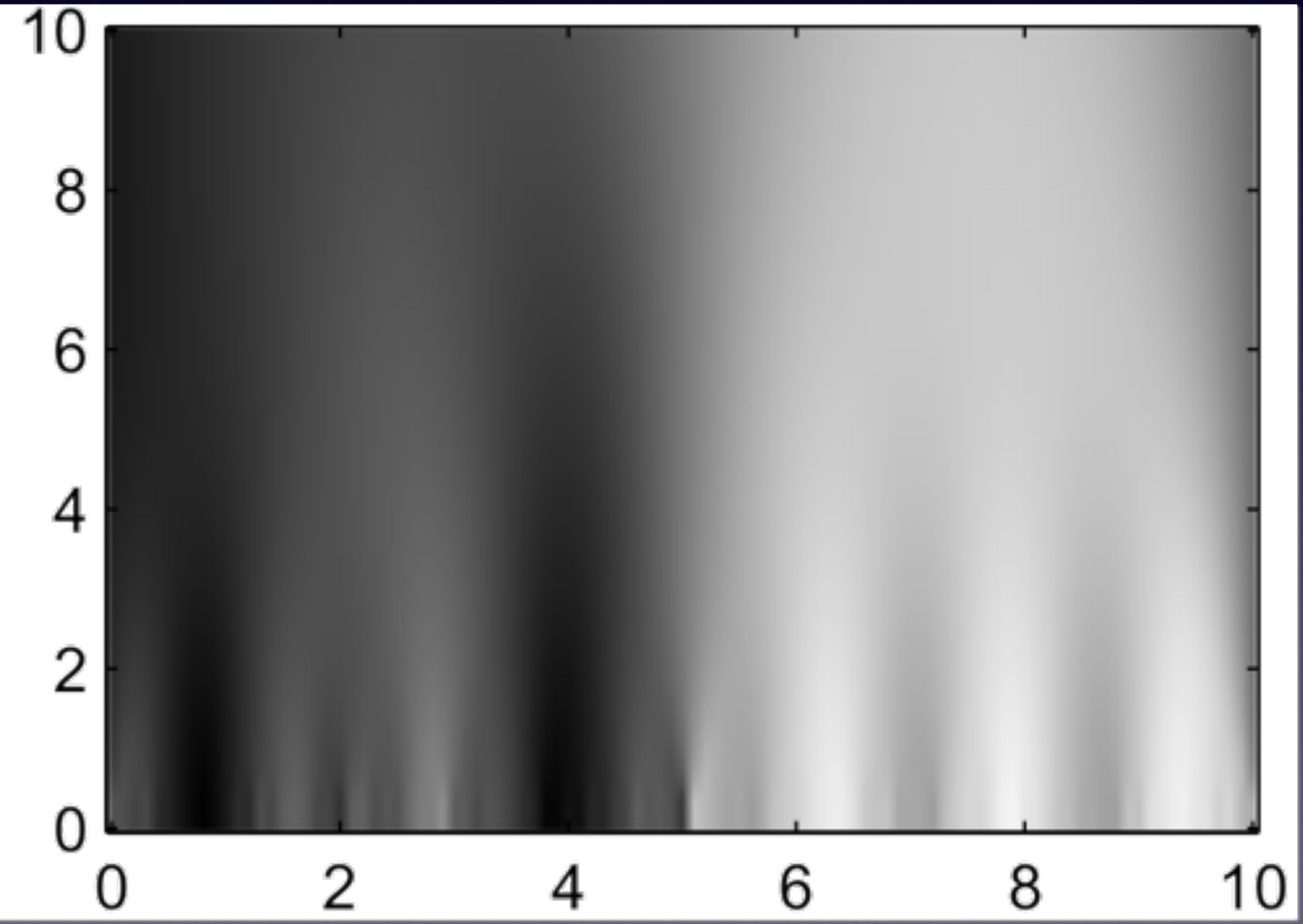
© 2015 Per-Erik Forssén

# Scale Space

$f(x)$



$f(x, \sigma)$



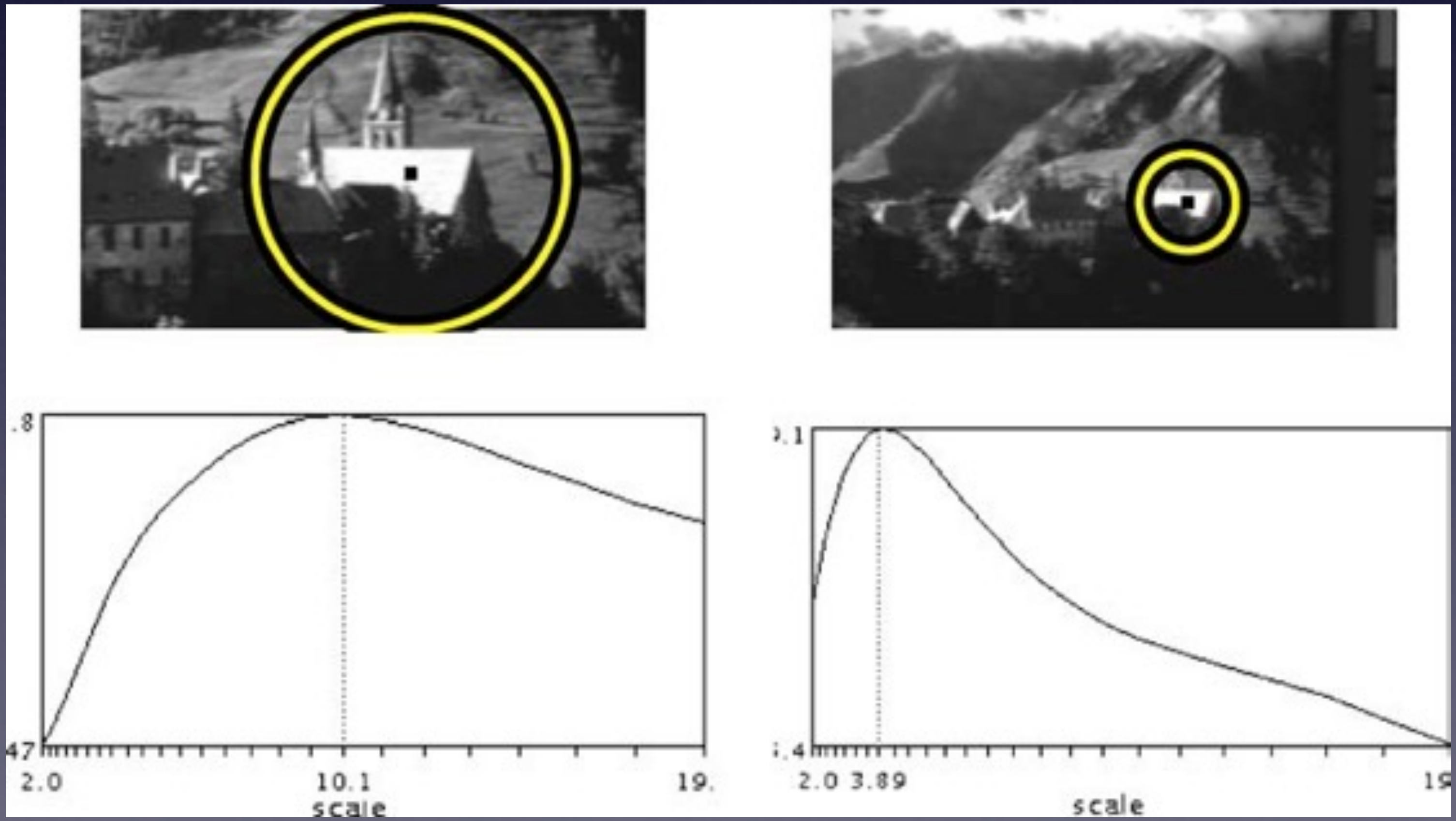
© 2015 Per-Erik Forssén

# Scale Selection

- Lindeberg, Tony (1994). Scale-Space Theory in Computer Vision. Kluwer/Springer, Boston.

$$(\hat{\mathbf{x}}, \hat{\sigma}) = \arg\max h(f(\mathbf{x}, \sigma))$$

- Find a characteristic point (e.g. max) on a function of position and scale



Idea from (Lindeberg 1993), illustration by (Mikolajczyk et al. 2005)

© 2015 Per-Erik Forssén

# Scale Selection

- Example: maximum of normalised Laplacian:

$$h(f(\mathbf{x},\sigma)) = \sigma^2(f * \nabla^2 g(\sigma))(\mathbf{x})$$

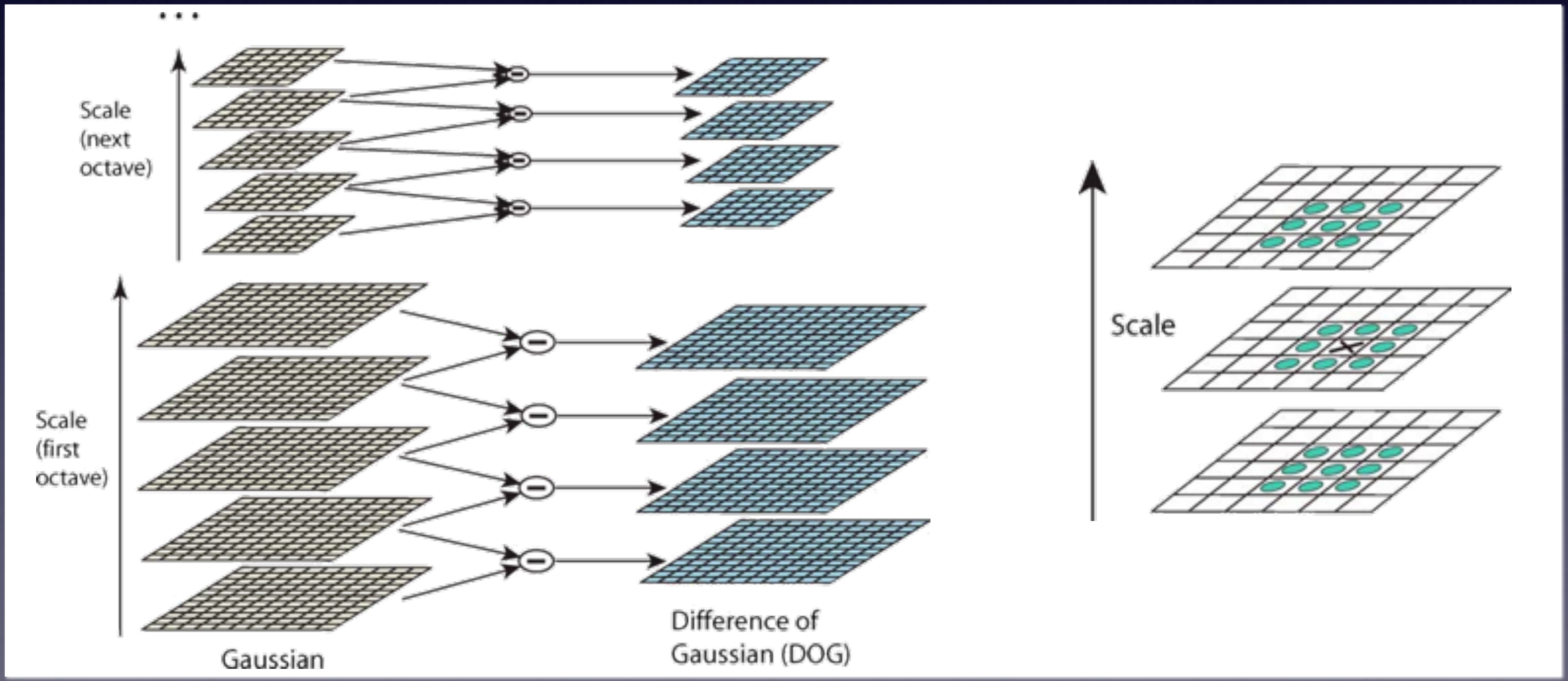
- Note the normalisation by $\sigma^2$, which is needed to compensate for decaying amplitude with scale.

- Another option (used by SIFT) is difference-of-Gaussians: $h(f(\mathbf{x},\sigma)) = (f * (g(\sigma) - g(k\sigma)))(\mathbf{x})$

© 2015 Per-Erik Forssén

# Scale Selection

- Efficient implementation using pyramids.
  David G. Lowe, ”Object recognition from local scale-invariant features”, ICCV’99



$$g(\sigma_1) * g(\sigma_2) = g(\sqrt{\sigma_1^2 + \sigma_2^2})$$

Non-max suppression in (x,y,$\sigma$)

© 2015 Per-Erik Forssén

# Scale Selection

- More accurate scale selection by polynomial fitting
  Brown& Lowe, ”Invariant Features from Interest Point Groups”, BMVC 2002 [Paper #1]

- Model for 3x3x3 neighbourhood:

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

$$\text{where } \mathbf{x} = (x\ y\ s)^T$$

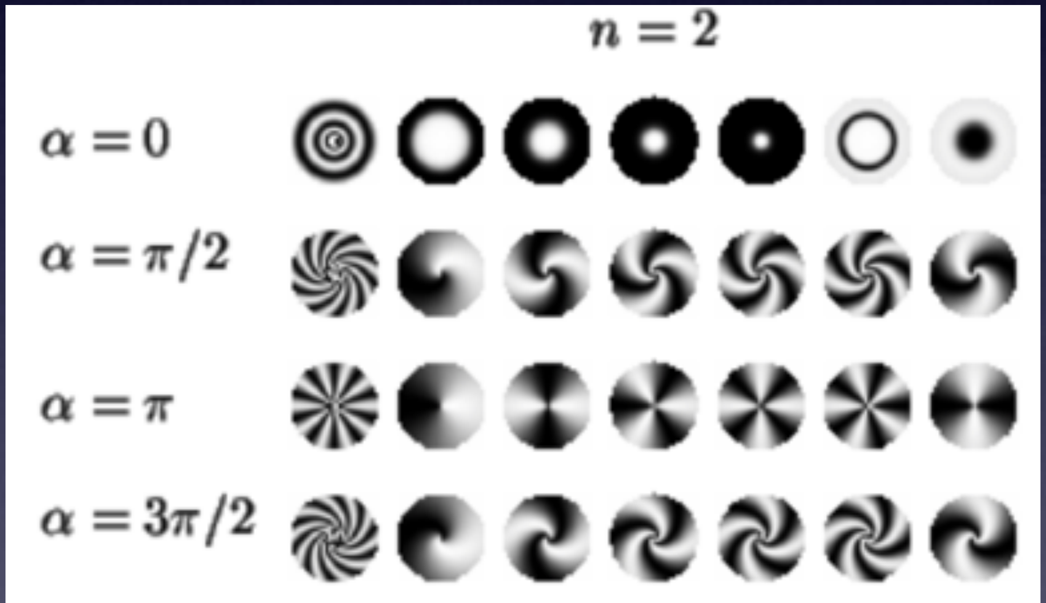
- Solution:

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

- Refines all of x,y and s.

# Scale Selection

- Scale selection for spiral features (SFOP).
  W. Förstner et al. ”Detecting Interpretable and Accurate Scale-Invariant Keypoints”, ICCV’09



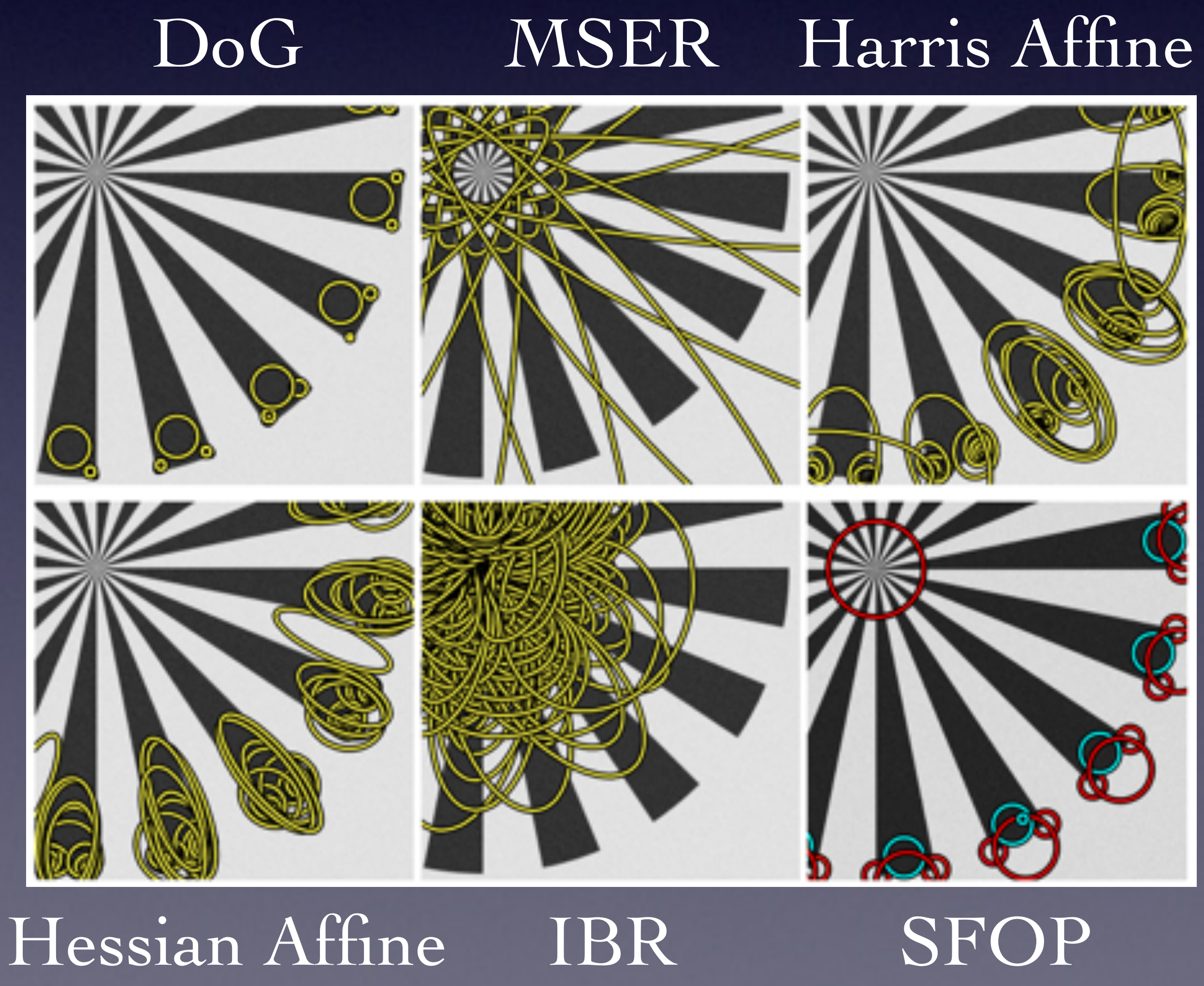
Björn Johansson, ”A quick tutorial on rotational symmetries”, CVL 2004

- 4-parameter search for (x,y,$\boldsymbol{\sigma}$,$\alpha$)

© 2015 Per-Erik Forssén

# Scale Selection

- Scale selection for spiral features (SFOP).
  W. Förstner et al. ”Detecting Interpretable and Accurate Scale-Invariant Keypoints”, ICCV’09



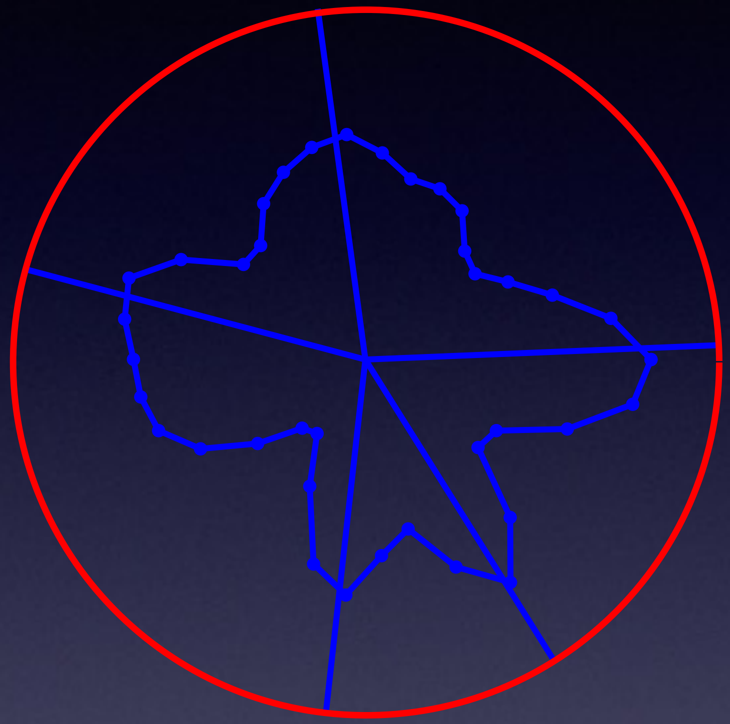
© 2015 Per-Erik Forssén

# Similarity Frames

- Scale selection gives us position and scale.

- For a similarity frame we can now determine one or more reference directions from a gradient orientation histogram at the found location in scale space.

- Idea from SIFT-paper: Generate several frames in close calls



$$h_k = \sum_{\text{patch}} |\nabla \mathbf{f}(\mathbf{x})| B_k(\tan^{-1} \nabla \mathbf{f}(\mathbf{x}))$$

© 2015 Per-Erik Forssén

# Affine Adaptation

- A. Baumberg, ”Reliable Feature Matching Across Widely Separated Views”, CVPR’00

- Affine frame by iteratively adjusting the circle defined by position and scale to an ellipse.

- In practice done by finding a resampling $\mathbf{x} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{m}$ that gives a *structure tensor* with equal eigenvalues in the c-frame.

© 2015 Per-Erik Forssén

# Affine Adaptation

- Elliptical structure in image $\mathbf{x}$
- Goal: Circular structure in canonical frame $\hat{\mathbf{x}}$
- Resampling by looping over values of $\hat{\mathbf{x}}$ and computing $\mathbf{x}$ as

$$\mathbf{x} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{m}$$

- Structure tensor $\mathbf{T}(\hat{\mathbf{x}})$ should be isotropic

© 2015 Per-Erik Forssén

# Affine Adaptation

- If we transform the input image, we will also transform the gradient field accordingly, so the following identity holds:

$$\nabla f(\mathbf{x}) \approx \mathbf{A}^{-1}\nabla f(\mathbf{A}\mathbf{x})$$

- For e.g. a rotation this is exactly true.

- Combine this with the structure tensor definition:

$$\mathbf{T}(\mathbf{x}) = \left(\nabla f \nabla f^T * g\right)(\mathbf{x})$$

- Gives us:

$$\mathbf{T}(\hat{\mathbf{x}}) \approx \mathbf{A}\mathbf{T}(\mathbf{A}^{-1}\mathbf{x})\mathbf{A}^T$$

# Affine Adaptation

- Structure tensor relation: $\mathbf{T}(\hat{\mathbf{x}}) \approx \mathbf{A}\mathbf{T}(\mathbf{A}^{-1}\mathbf{x})\mathbf{A}^T$
- Now choose: $\mathbf{A} = \lambda \mathbf{T}^{-1/2} \quad \det \mathbf{A} = 1$
- Inverse *whitening transform*. Needs to be iterated a few times, as g(**x**) should be anisotropic.
- Inherent rotation ambiguity:

$$\tilde{\mathbf{A}} = \mathbf{A}\mathbf{R},\ \mathbf{R} \in \mathrm{O}(2) \ \Rightarrow \ \tilde{\mathbf{A}}^T\tilde{\mathbf{A}} = \mathbf{T}$$

- Use reference direction(s) from gradient histogram.

# MSER

- Maximally Stable Extremal Regions

- Consider the set of all possible thresholdings of an image...

© 2015 Per-Erik Forssén

# MSER

- Maximally Stable Extremal Regions
- Consider the set of all possible thresholdings of an image...
- Connected regions form segments.
- Cf. Watershed algorithm
- Look at stability of a function of segment across image evolution. e.g. $\mathtt{area}(\mathtt{component}(t))$

© 2015 Per-Erik Forssén

# MSER

- MSERs are components that are *maximally stable*, i.e., have a local minimum of the rate of change:

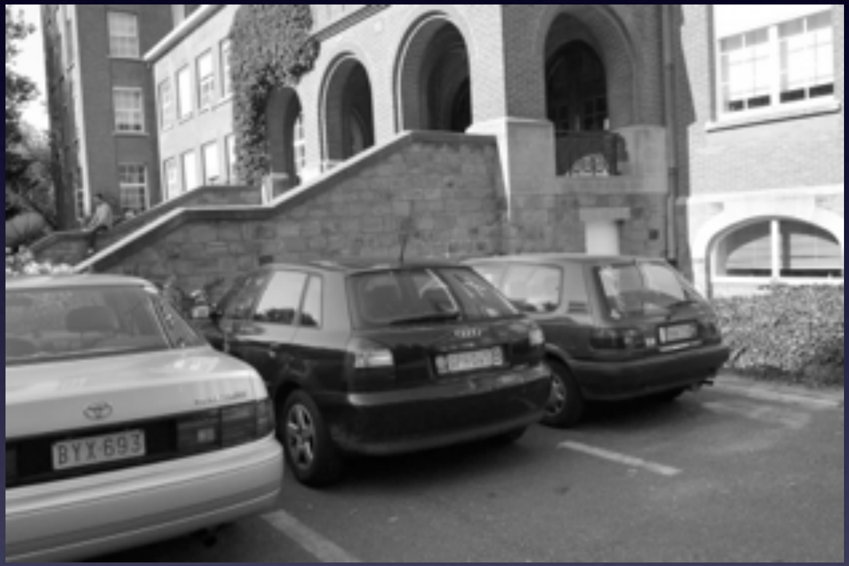
$$\frac{\partial \texttt{area}(\texttt{component}(t))}{\partial t}$$

- c.f. Scale Selection
- Stability measure: Range of stable thresholds $t_2$-$t_1$ around min is called the *margin* of the region.

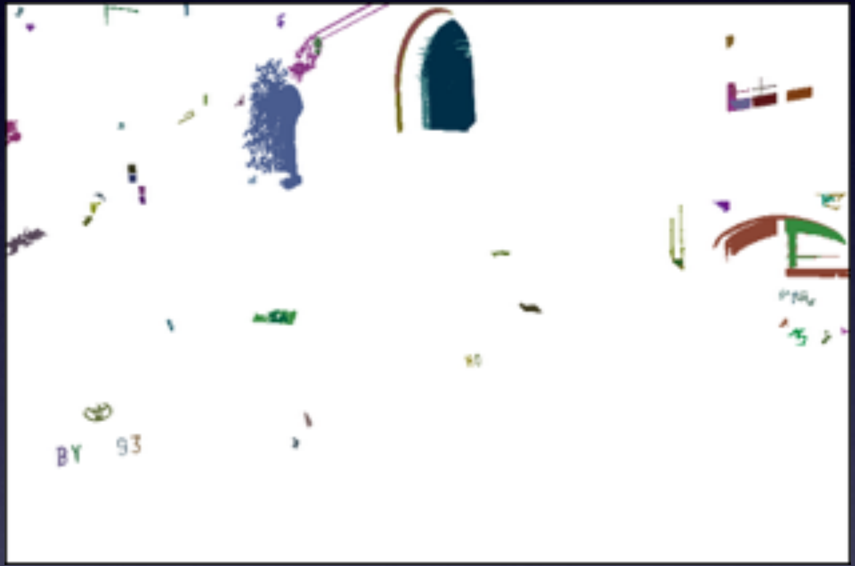
© 2015 Per-Erik Forssén

# MSER

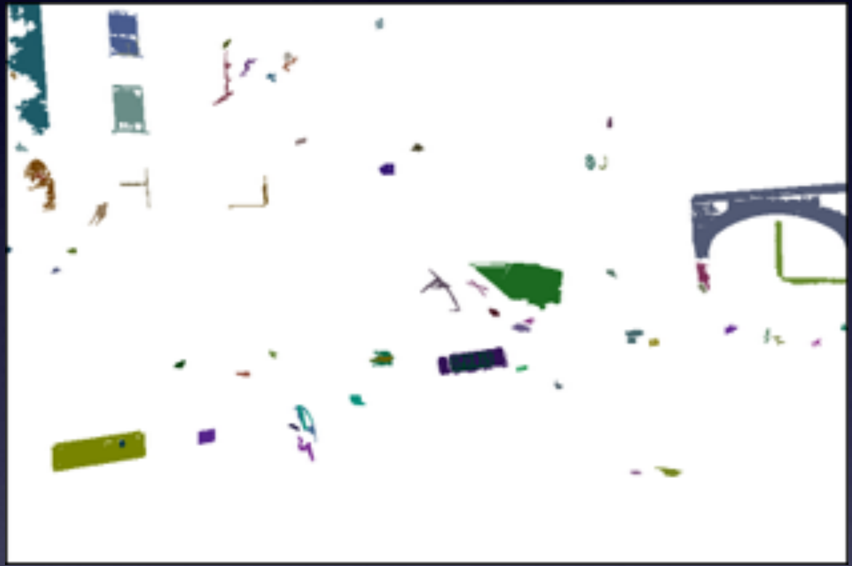
- Two possible thresholdings: $I(\mathbf{x}) < t\,, \quad I(\mathbf{x}) > t$



Input image



64 MSER- (total 272)



64 MSER+ (total 294)

- Very efficient (union/find+path compression)

- MSER type (+/-) is useful for matching (LE6,LE7)

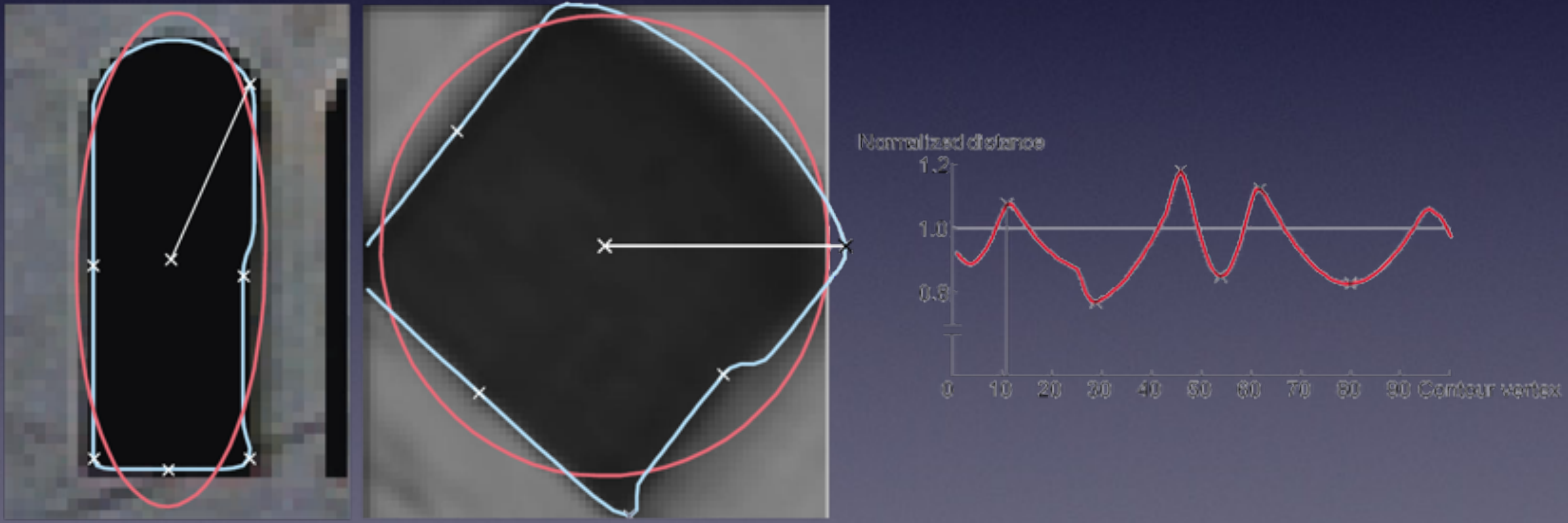
© 2015 Per-Erik Forssén

# MSER

- MSER is invariant to monotonic changes of intensity. i.e.  I(x) and f(I(x)) have the same output if
$$f(x+k) > f(x) \; \forall \; k > 0$$
- Wide range of sizes obtained without a scale pyramid. Better still with a pyramid (Forssén&Lowe ICCV'07)
- Can be used to track colour objects by computing MSERs on the Mahalanobis distance to a colour distribution. (Donoser&Bischof CVPR'06)

# Local Affine Frames

- Find approximating ellipse of region.
- Contour extrema in normalised frame give reference directions.



Matas et al. ICPR'02

© 2015 Per-Erik Forssén

# Local Affine Frames

- **Approximating ellipse**
  from moments of binary mask

$v : \Omega \mapsto \{0, 1\}$

$$\mu_{k,l}(v) = \sum_x \sum_y x^k y^l v(x, y)$$

$$\mathbf{m} = \frac{1}{\mu_{0,0}} \begin{bmatrix} \mu_{1,0} \\ \mu_{0,1} \end{bmatrix} \qquad \mathbf{C} = \frac{1}{\mu_{0,0}} \begin{bmatrix} \mu_{2,0} & \mu_{1,1} \\ \mu_{1,1} & \mu_{0,2} \end{bmatrix} - \mathbf{m}\mathbf{m}^T$$

$$\mathcal{R}(\mathbf{m}, \mathbf{C}) = \{\mathbf{x} : (\mathbf{x} - \mathbf{m})^T \mathbf{C}^{-1} (\mathbf{x} - \mathbf{m}) \leq 4\}$$

- See appendix C in thesis by Forssén 2004

© 2015 Per-Erik Forssén

# Local Affine Frames

- **Normalisation to a circle** (axis aligned)
  Compute the eigenfactorisation:

  $$\mathbf{C} = \mathbf{R}\mathbf{D}\mathbf{R}^T \quad , \quad \det\mathbf{R} > 0$$

  The circle normalisation can now be performed as:

  $$\mathbf{x} = \mathbf{A}\hat{\mathbf{x}} + \mathbf{m}, \quad \text{for } \mathbf{A} = 2\mathbf{R}\mathbf{D}^{1/2}$$

  $\hat{\mathbf{x}}$ - canonical coordinates
  $\mathbf{x}$ - image coordinates

© 2015 Per-Erik Forssén

# Local Affine Frames

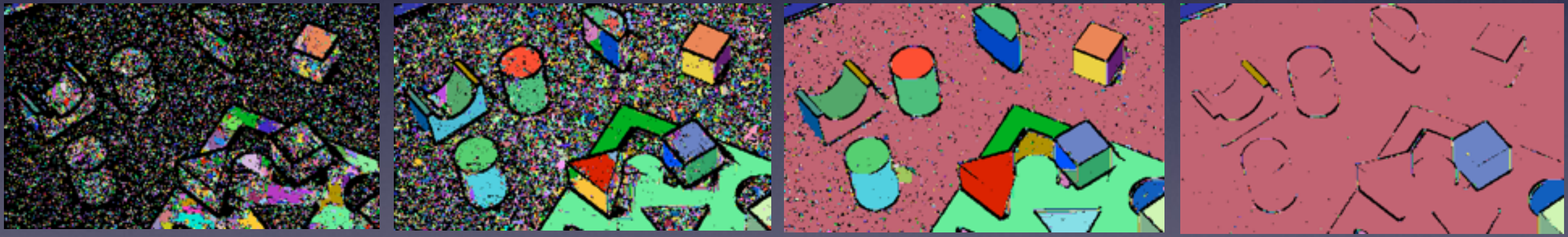
- Ellipse+extrema of distance to centre is just one frame construction option.
- Other (affine covariant) choices:
  - Points of maximum curvature.
  - Bi-tangens.
- See Obdrzalek&Matas BMVC'02

© 2015 Per-Erik Forssén

# MSCR

- Maximally Stable Colour Regions
  P.-E. Forssén, ”Maximally Stable Colour Regions for Recognition and Matching”, CVPR’07

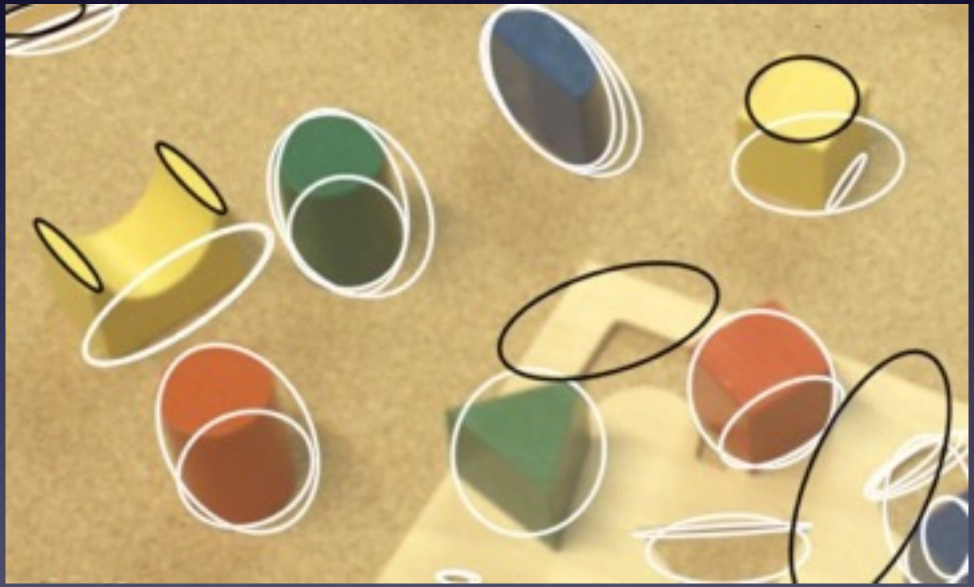
- Define evolution function on an agglomerative clustering of the image.



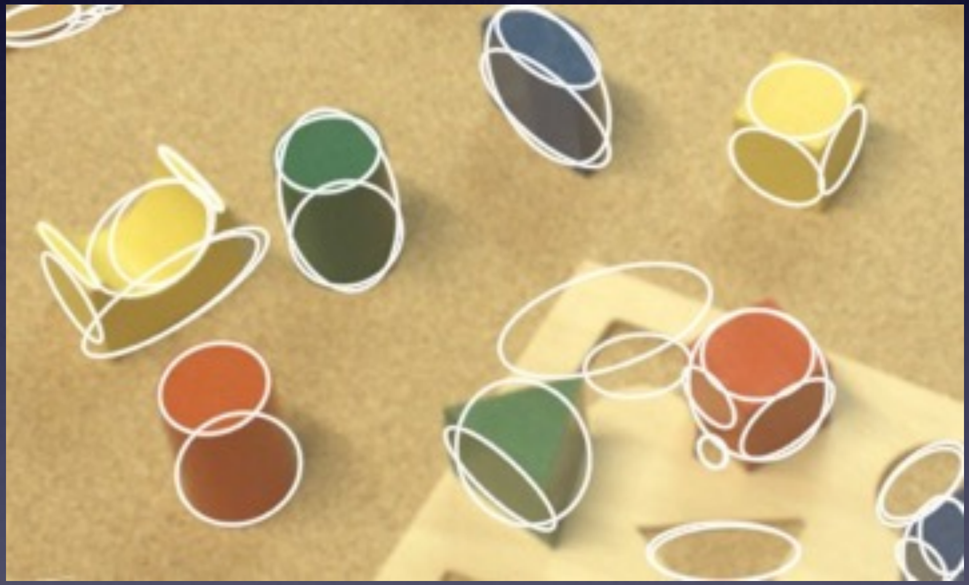
Forssén CVPR’07

# MSCR

- Improved robustness to illumination changes, and changes of background



MSER+ and MSER-



MSCR

- ~3x more computationally expensive.

© 2015 Per-Erik Forssén

# Combined detectors and descriptors

- SIFT: D. G. Lowe, ”Distinctive Features from Scale-Invariant Keypoints”, Springer IJCV 2004
- SURF: H. Bay et al., ”SURF: Speeded Up Robust Features”, ECCV’06
- ORB: E. Rublee, et al., ”ORB: an efficient alternative to SIFT or SURF”, ICCV’11
- BRISK: S. Leutenegger et al., "BRISK: Binary Robust Invariant Scalable Keypoints", ICCV’11
- FREAK: A. Alahi et al. ”FREAK: Fast Retina Keypoint”, CVPR’12
- It is also possible to switch detectors and descriptors between these…

# Detection Noise

- If frame parameters are unstable, the appearance invariance will suffer.

- Ways to compensate for an unstable frame:

  - During learning: Generate multiple training examples from a single training image by resampling (i.e. learning the invariance)

  - During recognition: A. Generate multiple frames (as in SIFT orientation). B. Generate multiple test images by resampling (slows down recognition)

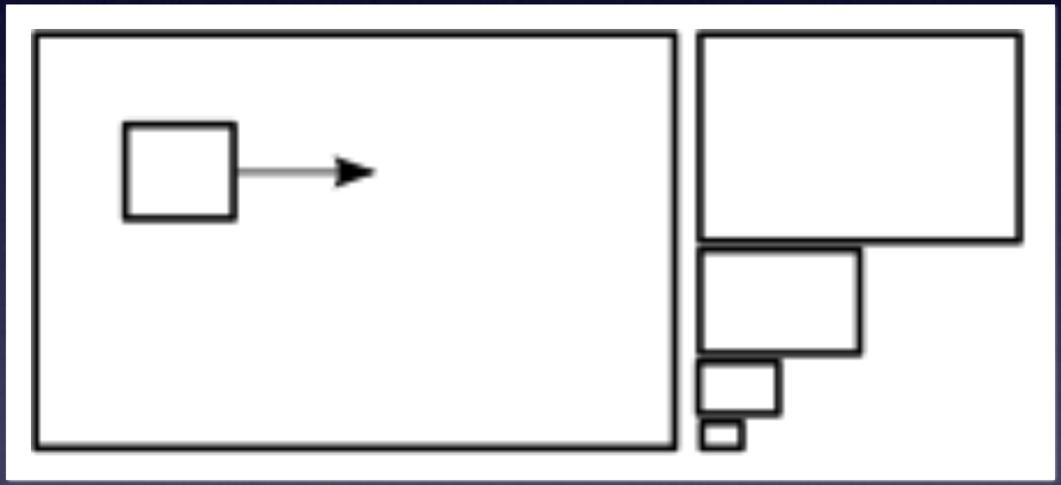
# Detection Noise

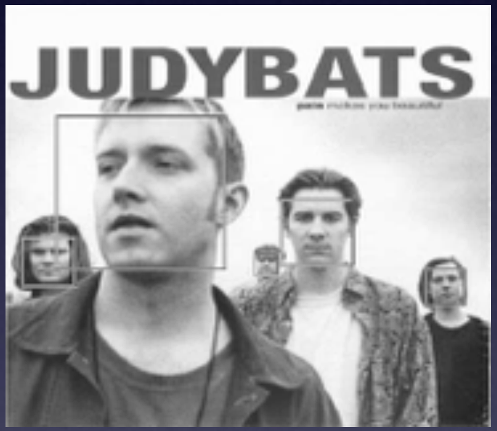
- If frame parameters are unstable, the appearance invariance will suffer.
- Ways to compensate for an unstable frame:
  - During learning: Generate multiple training examples from a single training image by resampling (i.e. learning the invariance)
  - During recognition: A. Generate multiple frames (as in SIFT orientation). B. Generate multiple test images by resampling (slows down recognition)
- Complementary, so all may be used together.

© 2015 Per-Erik Forssén

# Sliding Window

- Another way to eliminate detection noise is to remove the detector.





Example: face detection

- Sliding window approaches test ***all*** regions. Possible for translation+scale

- Extra time for octave scale search $= \frac{1}{4} + \frac{1}{4^2} \ldots = \frac{1}{3}$

© 2015 Per-Erik Forssén

# Sliding Window

- Sliding window approaches are useful for **global appearance**. For constellations (LE5) parts are defined relative to main object.
- Appearance changes from other dimensions than (x,y,s) need to be learned. E.g. separate detectors for different head poses in face detection. (But this is the case for many other aspects, such as expression anyway.)
- Efficient matching is necessary, e.g. cascaded detection, decision trees, ANN. (LE6)
- A region detector can be seen as the first step in a detection cascade.

# Summary

- The detector part of a feature finds a canonical frame in which to sample.
- **Similarity** (rotation+translation+scale) or **affine** are common choices.
- Region detection generates invariance, but also detection noise.
- Sliding window approaches combine detection and matching for global appearance.

© 2015 Per-Erik Forssén

# Discussion

- Questions/comments on paper:

  S. Leutenegger et al., "BRISK: Binary Robust Invariant Scalable Keypoints", ICCV’11

© 2015 Per-Erik Forssén

# Paper for next week

- Paper to read for next week:

  J. Sivic and A. Zisserman, ”Video Google: A Text Retrieval Approach to Object Matching in Videos”, ICCV’03

©2015 Per-Erik Forssén