Visual Object Recognition

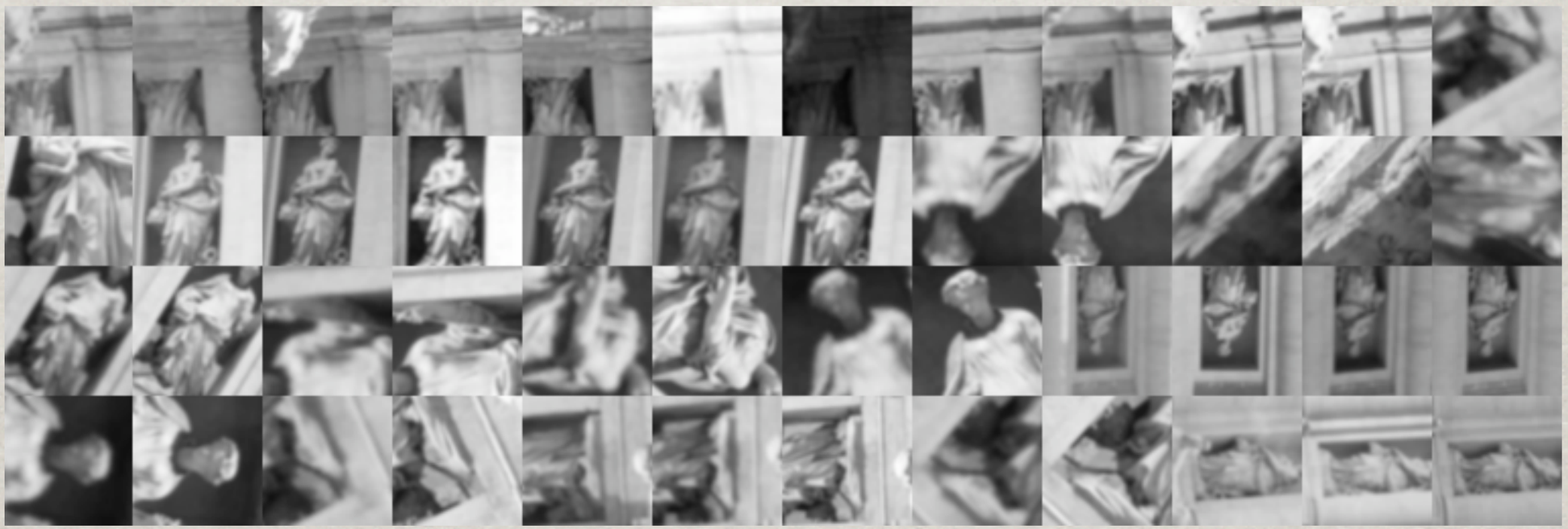
State-of-the-art techniques and performance evaluation

Lecture 4:Descriptor Construction

- An opportunity for machine learning
- Scale Invariant Feature Transform (SIFT)
- Speeded Up Robust Features (SURF)
- Geometric Blur and Log-polar grids
- ContourSIFT
- Results from learning descriptors

Opportunity for Machine Learning

- When we have a system that automatically detects canonical frames, we could potentially use machine learning to find a good way to describe the invariant image patches.



Opportunity for Machine Learning

- Any method used to learn global appearance (LE1) could potentially be used.
- But, high-dimensional learning requires a large amount of training data.
- Solution: Parametrise space of solutions in intelligent ways.
- See: Winder&Brown, *Learning Local Image Descriptors*, CVPR'07 (more at end of lecture)

The SIFT descriptor

- Scale Invariant Feature Transform

- Converts an image patch sampled in canonical frame to a 128-byte *descriptor vector*.

- Inherits geometric invariances from c-frame.

The SIFT descriptor

- Compute x- and y-gradients through convolution:

$$\nabla \mathbf{f}(\mathbf{x}) = \begin{bmatrix} (d_x * f)(\mathbf{x}) \\ (d_y * f)(\mathbf{x}) \end{bmatrix}$$

- Rotate gradient map to direction from orhist:

$$\nabla \hat{\mathbf{f}}(\mathbf{x}) = \mathbf{R} \nabla \mathbf{f}(\mathbf{R}^{\mathbf{T}} \mathbf{x})$$

- Compute gradient orientation histograms in 4x4 spatial regions:

$$h_{kl} = \sum_{\mathbf{x} \in \text{patch}_l} |\nabla \hat{\mathbf{f}}(\mathbf{x})| w(\mathbf{x} + \mathbf{d}_l) B_k(\tan^{-1} \nabla \hat{\mathbf{f}}(\mathbf{x}))$$

The SIFT descriptor

- Compute gradient orientation histograms in 4x4 spatial regions :

$$h_{kl} = \sum_{\mathbf{x} \in \text{patch}_l} |\nabla \hat{\mathbf{f}}(\mathbf{x})| w(\mathbf{x} + \mathbf{d}_l) B_k(\tan^{-1} \nabla \hat{\mathbf{f}}(\mathbf{x}))$$

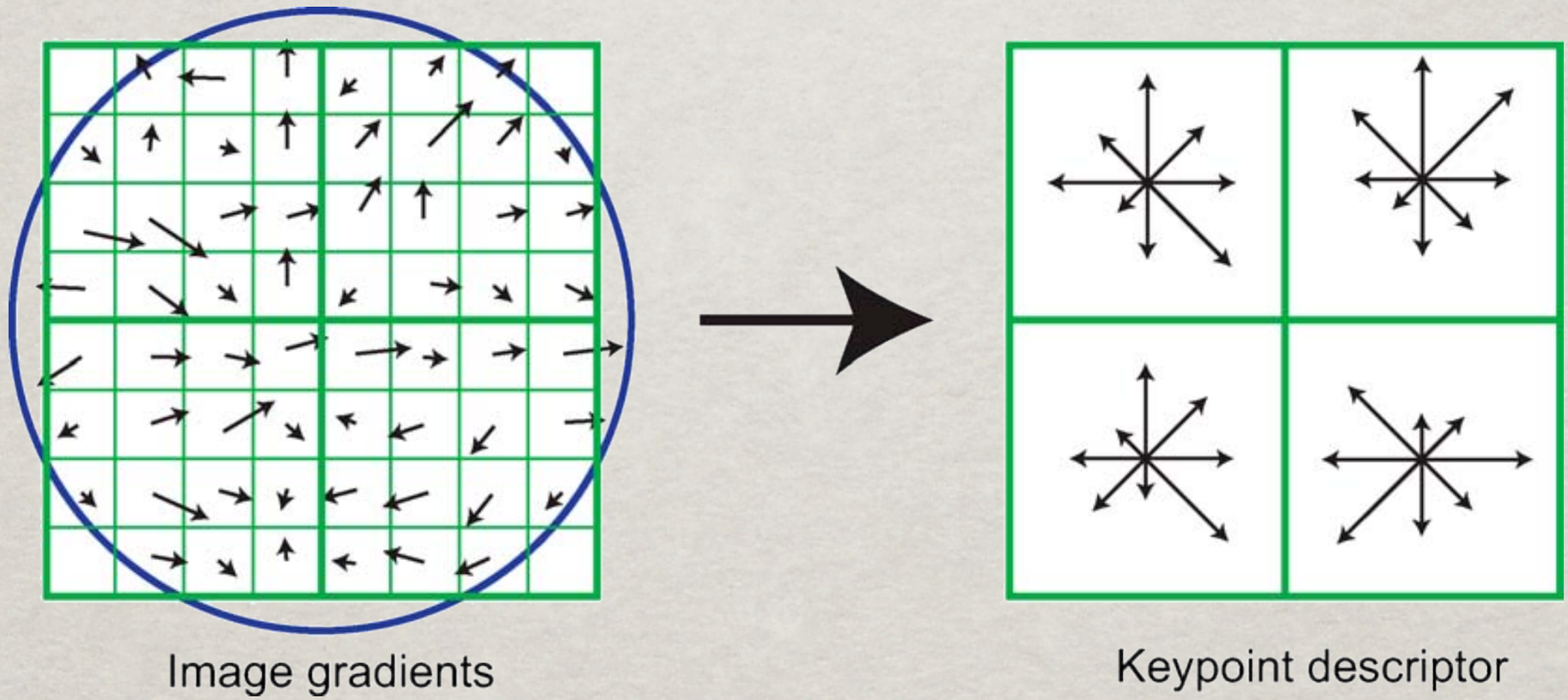
- $B_k(x)$ linear interpolation kernel
  Quadratic is better (Jonsson&Felsberg)

- Subwindows $l \in [1 \ldots 16]$ directions $k \in [1 \ldots 8]$

- Spatial weight $w(\mathbf{x} + \mathbf{d}_l)$ (Gaussian decay)

The SIFT descriptor

- Implementation with source code (A. Vedaldi): http://vision.ucla.edu/~vedaldi/code/sift/sift.html



Note that 4x4 regions are actually used, with 8 orientations, 128 elements

The SIFT descriptor

- Affine illumination invariance by using gradients and normalising descriptor $\hat{\mathbf{h}} = \mathbf{h}/||\mathbf{h}||$

- Some robustness by truncating and normalising again $\hat{\mathbf{h}} = \min(\mathrm{t}, \hat{\mathbf{h}})/||\hat{\mathbf{h}}||$

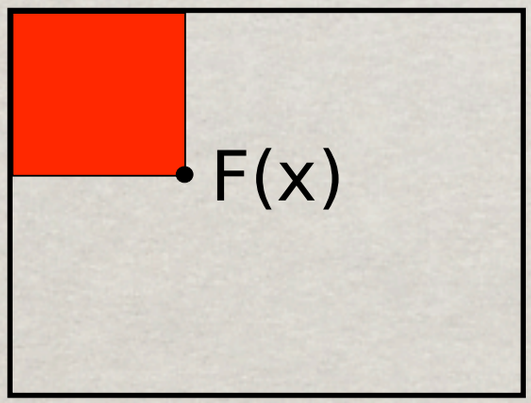
- The spatial histogramming gives robustness to scale/rotation/translation errors.

- Bears some similarity to the "Standard model" (see LE1)

SURF

- Bay&Tuytelaars&vanGool, *SURF: Speeded Up Robust Features*, ECCV06

- Both detector and descriptor optimised for speed using *Integral images*. (Viola&Jones)



$$F(\mathbf{x}) = \sum_{k=0}^{x_1} \sum_{l=0}^{x_2} f(k,l)$$

- Very fast to compute:

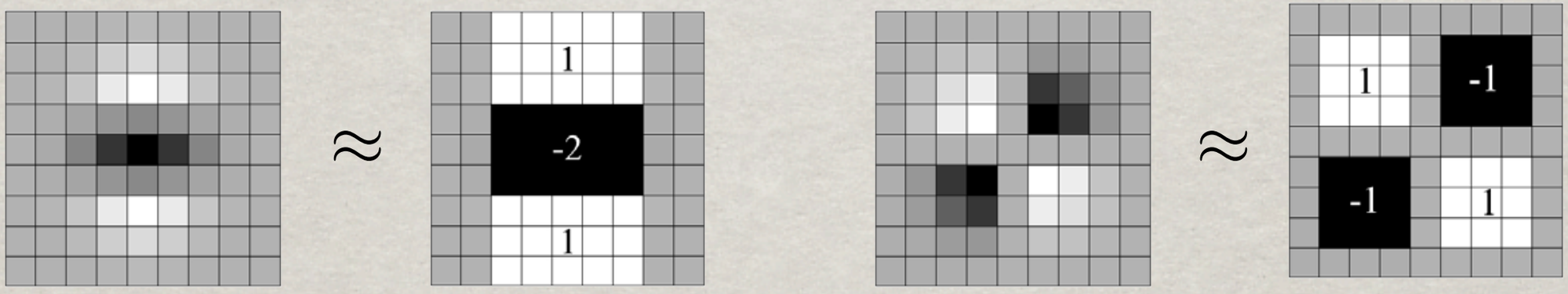
$$F(\mathbf{x}) = F(\mathbf{x} - \binom{1}{0}) + F(\mathbf{x} - \binom{0}{1}) - F(\mathbf{x} - \binom{1}{1}) + f(\mathbf{x})$$

The SURF Detector

- Find interest points using Hessian matrix in scale space:

$$\mathbf{H}(\mathbf{x}, \sigma) = \begin{bmatrix} f_{xx}(\mathbf{x}, \sigma) & f_{xy}(\mathbf{x}, \sigma) \\ f_{xy}(\mathbf{x}, \sigma) & f_{yy}(\mathbf{x}, \sigma) \end{bmatrix}$$

- Approximate filters using I-images:



- Scale space by increasing filter size.

The SURF Detector

- Scale space maxima of det(H)

$$\det(\mathbf{H}) \approx H_{11}H_{22} - 0.9H_{12}^2$$

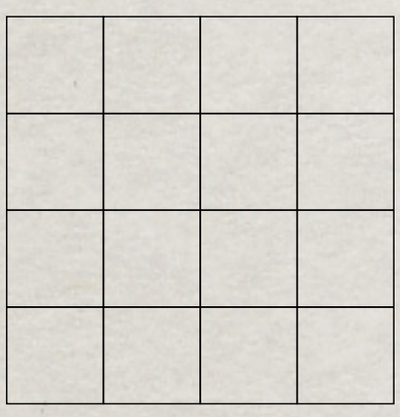
- Detection like in SIFT:

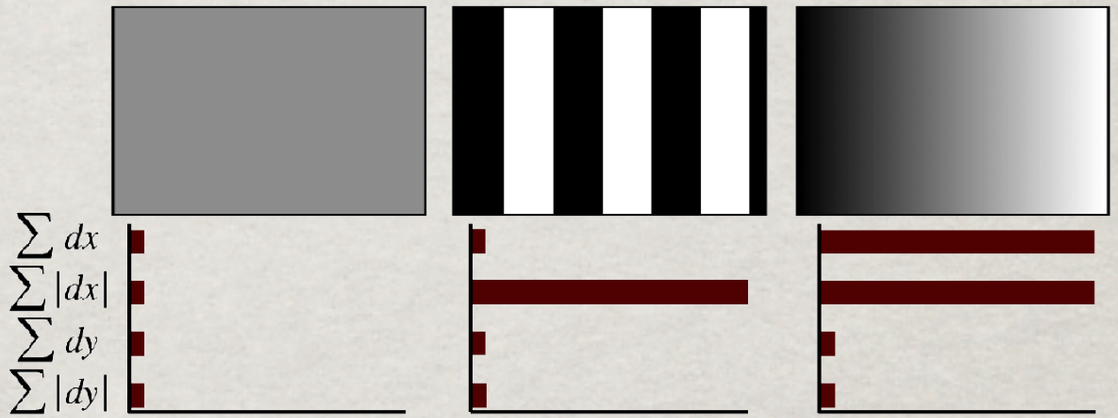
A.3x3x3 non-max-suppression

B.Quadratic polynomial interpolation in scale space (Brown&Lowe'02)

The SURF Descriptor

- Again using I-images





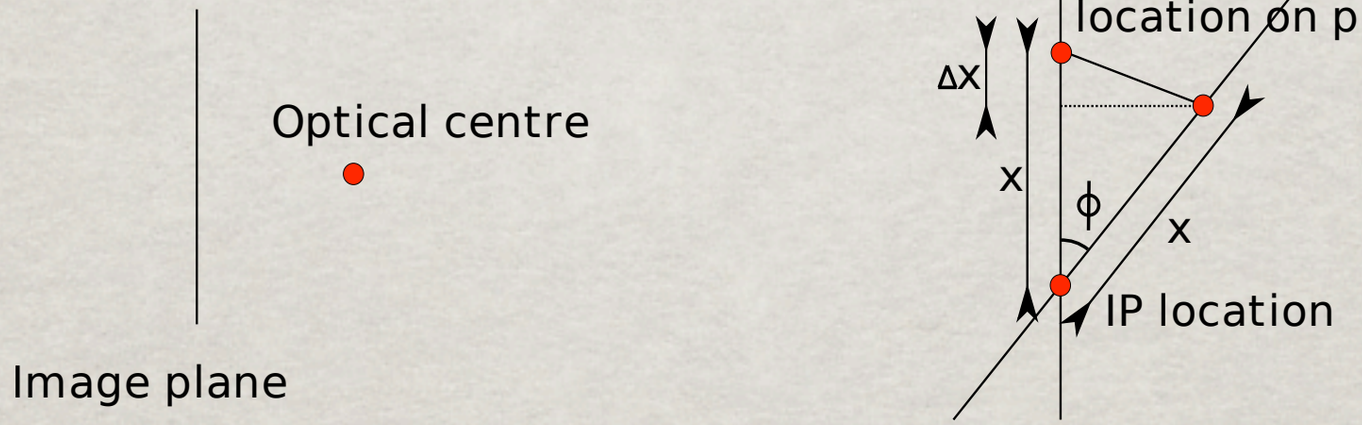
- 4 responses in each 4x4 block ⇒ 64 elements in descriptor.

- Descriptor is normalised.

- Sign of Laplacian tells dark blobs from bright cf. MSER+ and MSER-

Geometric Blur

- Introduced by Berg&Malik CVPR'01. As a descriptor at CVPR'05.

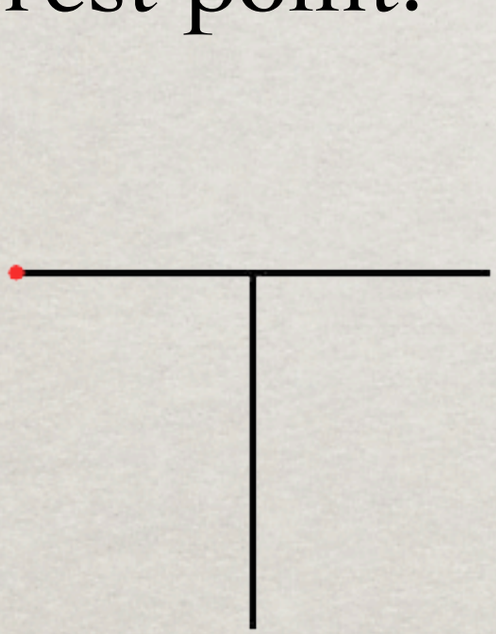


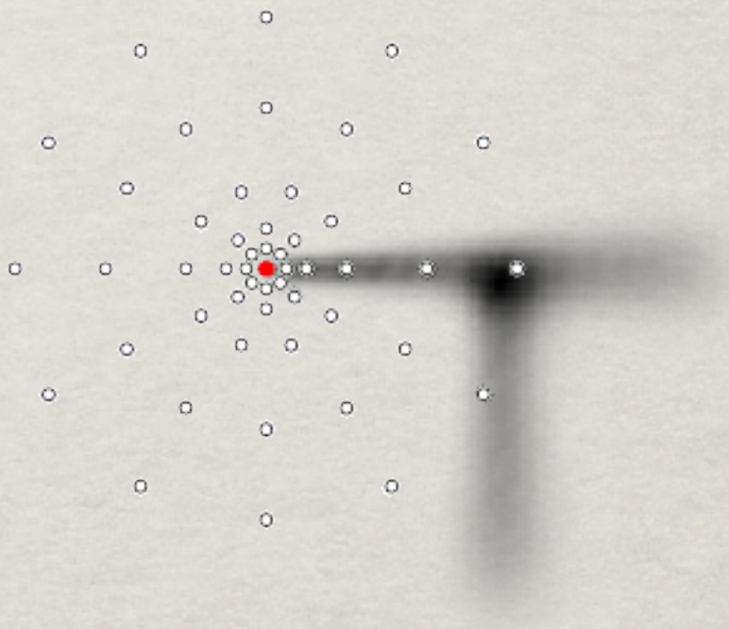
- If orthogonal projection: Point on a planar scene is displaced proportionally to the distance from the reference point:

$$\Delta x = (1 - \cos\phi)x$$

Geometric Blur

- Increase blur linearly with distance from interest point.





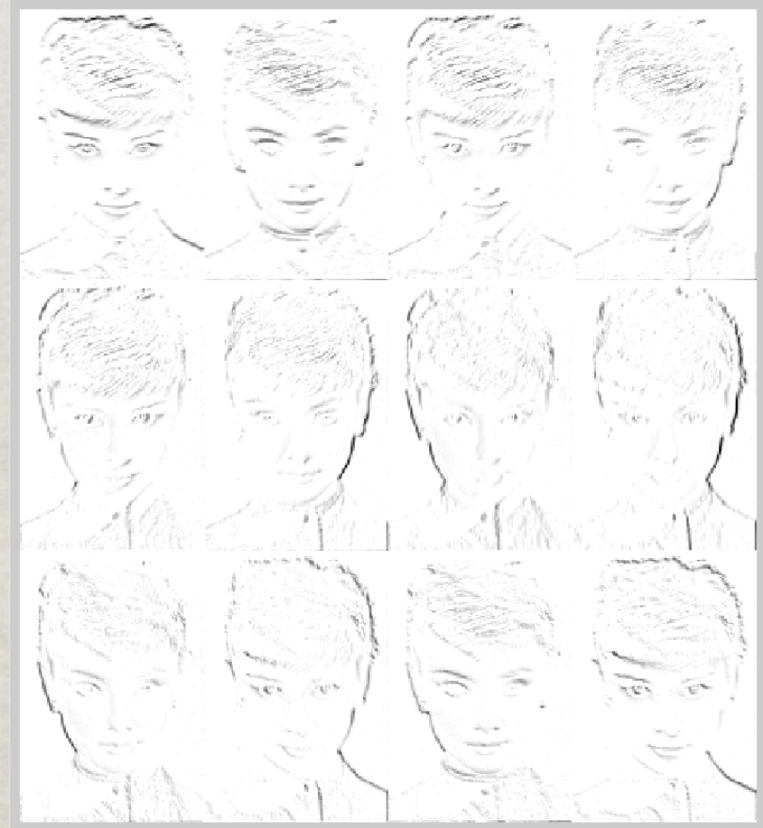
- Sample in 38 sparse points

Geometric Blur descriptor

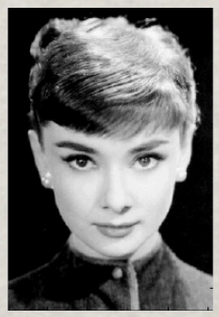
- Use *half-wave rectified* gradient responses in six directions:

$$h_k^+(\mathbf{x}) = \max(0, (f * g_k)(\mathbf{x}))$$

$$h_k^-(\mathbf{x}) = \max(0, -(f * g_k)(\mathbf{x}))$$



- Each is blurred with geometric blur before sampling.

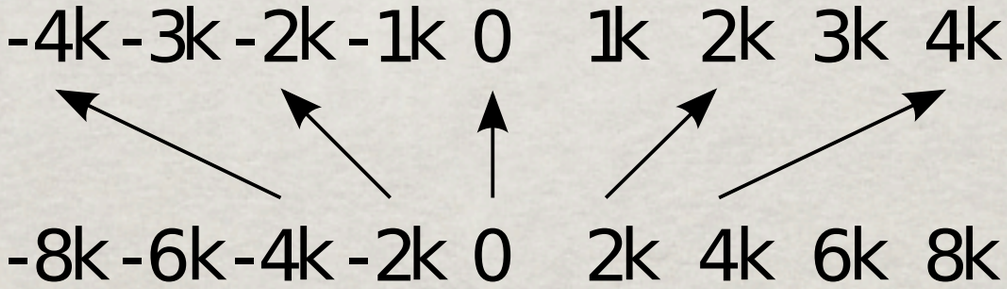


- 228 elements

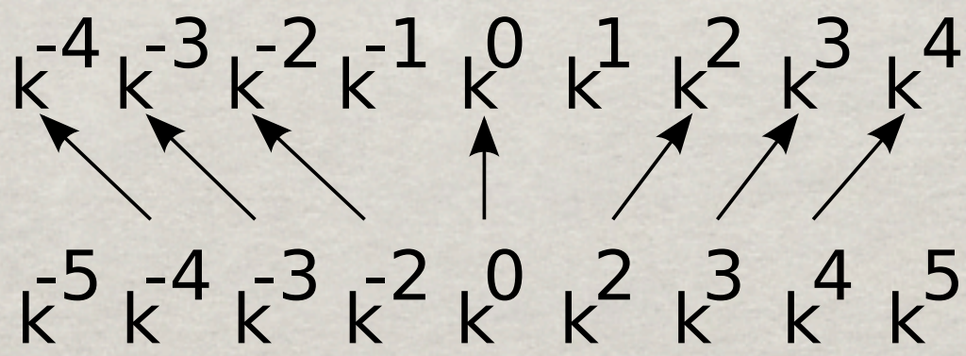
Log-polar patches

- A similar sampling pattern can also be motivated from scale invariance.

- With linear distances, matching across scale is difficult:



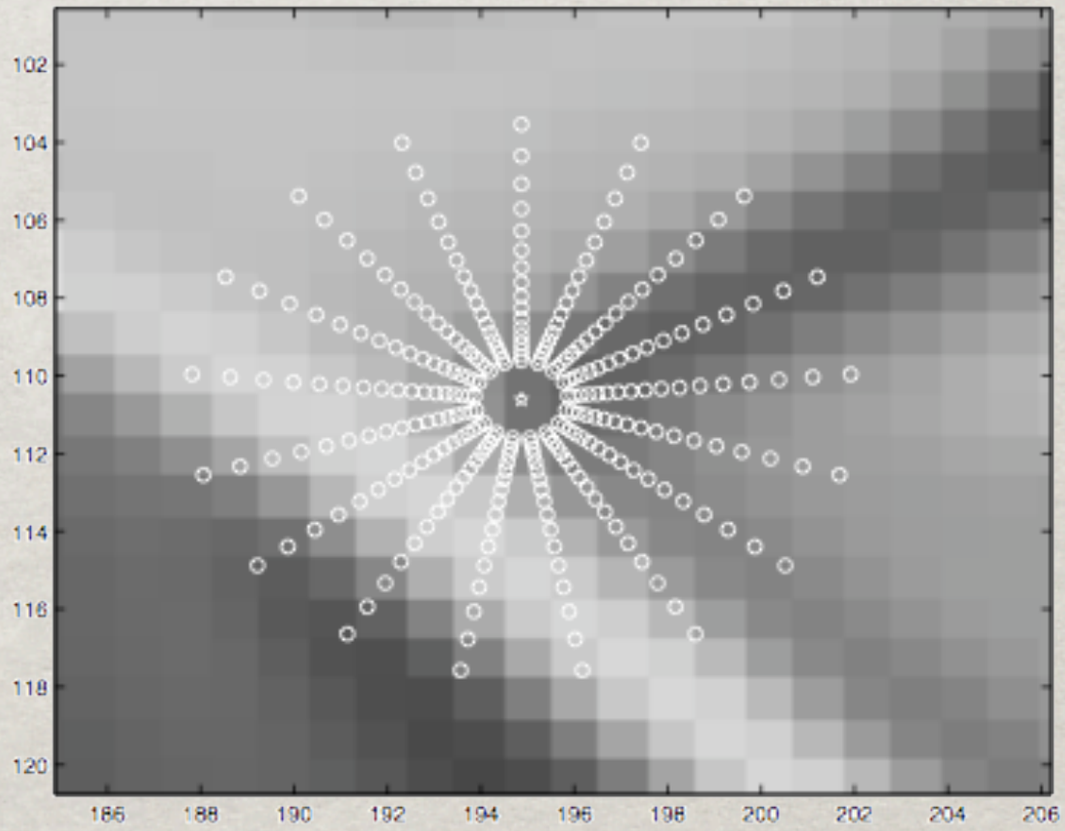
- With exponential distances, scaling becomes a shift:



Log-polar patches

- Viksten&Moe, *Local single-patch features for pose estimation, using the log-polar tranform.* IbPRIA'05

(a) Harris points.

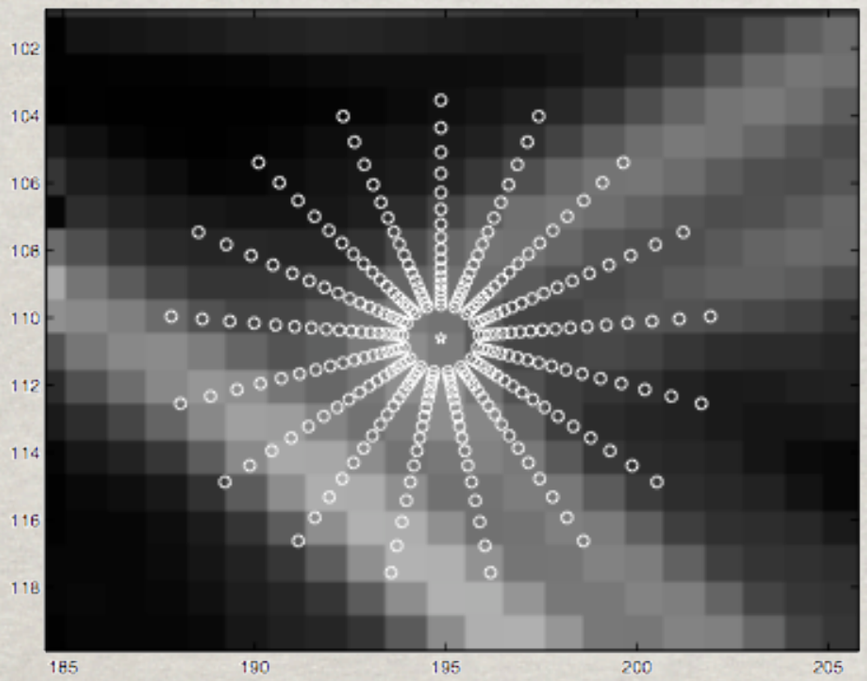


(b) Sample grid around Harris point.

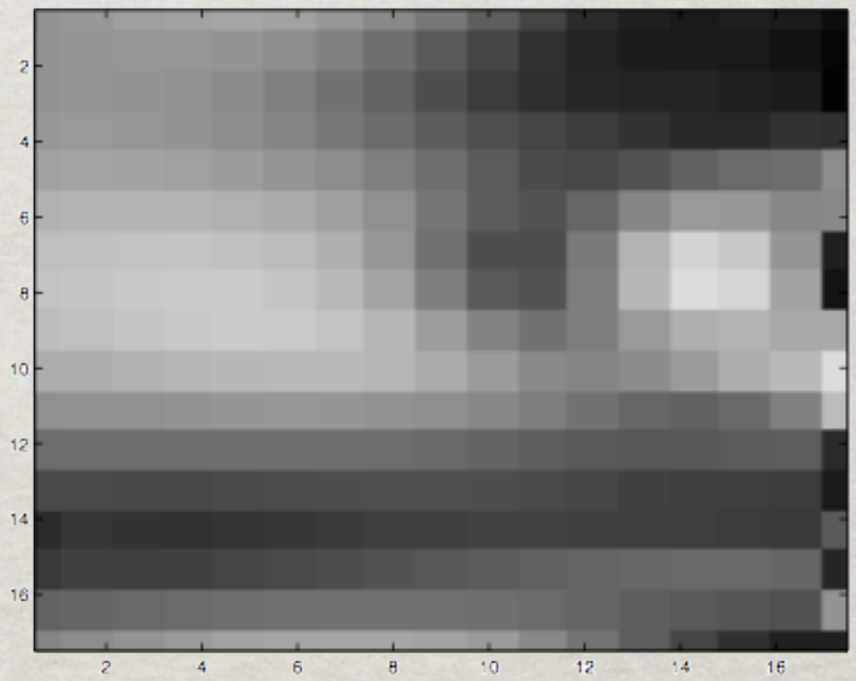
Log-polar patches

- Sample gradient image around Harris points in a log-polar pattern.

Figure 6.1: Harris points and log-polar sampling grid.



(a) Edge image with sample grid.



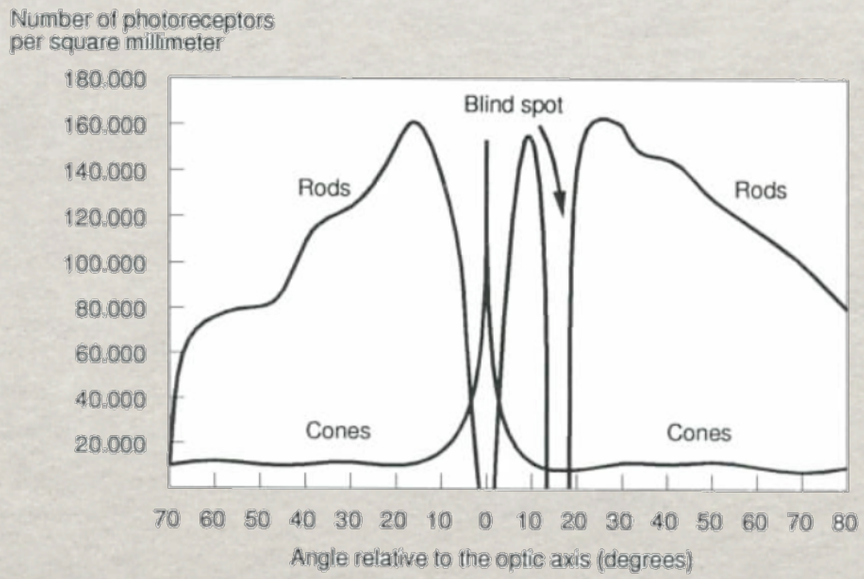
(b) Log-polar sampled edge image.

- Scale change can now be found using correlation.

Log-polar patches

- Same sampling pattern is used in GLOH descriptor. Mikolajczyk&Schmid "A performance evaluation of local descriptors", TPAMI 2005

- Fun fact: The eye also has a (partially) log polar sampling pattern.



Difficult cases for Descriptors

- Background clutter in 3D scenes

- Patches cut out around features will have varying background.

Difficult cases for Descriptors

- Large illumination changes

- Gradient strength changes non-uniformly.
- Contrast may be inverted.

Contour SIFT

- Idea: Use a detector that produces contours, e.g. MSER or MSCR

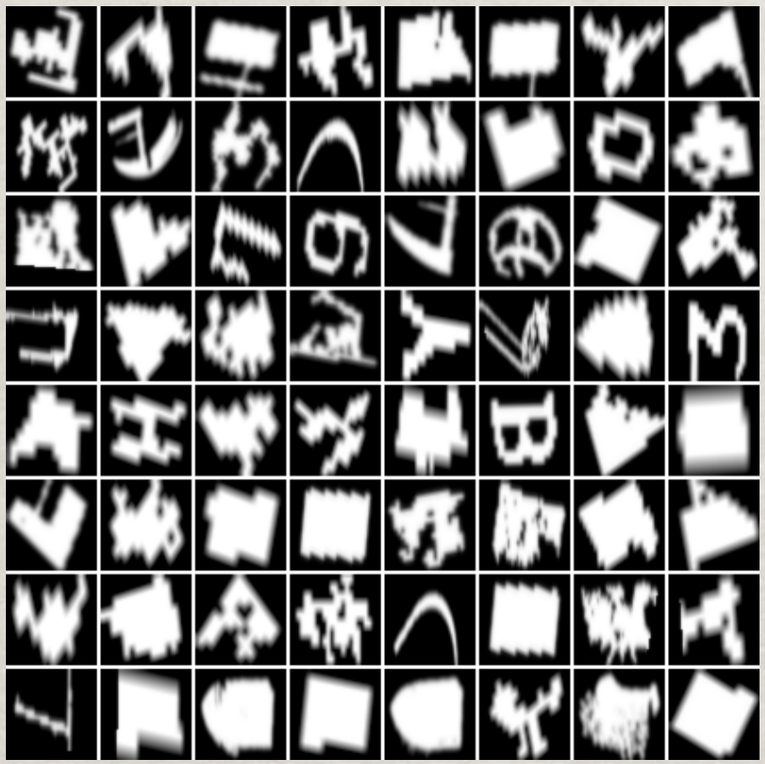
Input image

64 random MSER- regions

- Region shape is robust to changes outside the region

Contour SIFT

- Compute a descriptor from the binary mask of the region instead of the grey-scale patch.



- Less descriptive patches, but more robust to illumination and background clutter

Contour SIFT

- Shape Descriptors for Maximally Stable Extremal Regions, Forssén&Lowe, ICCV'07

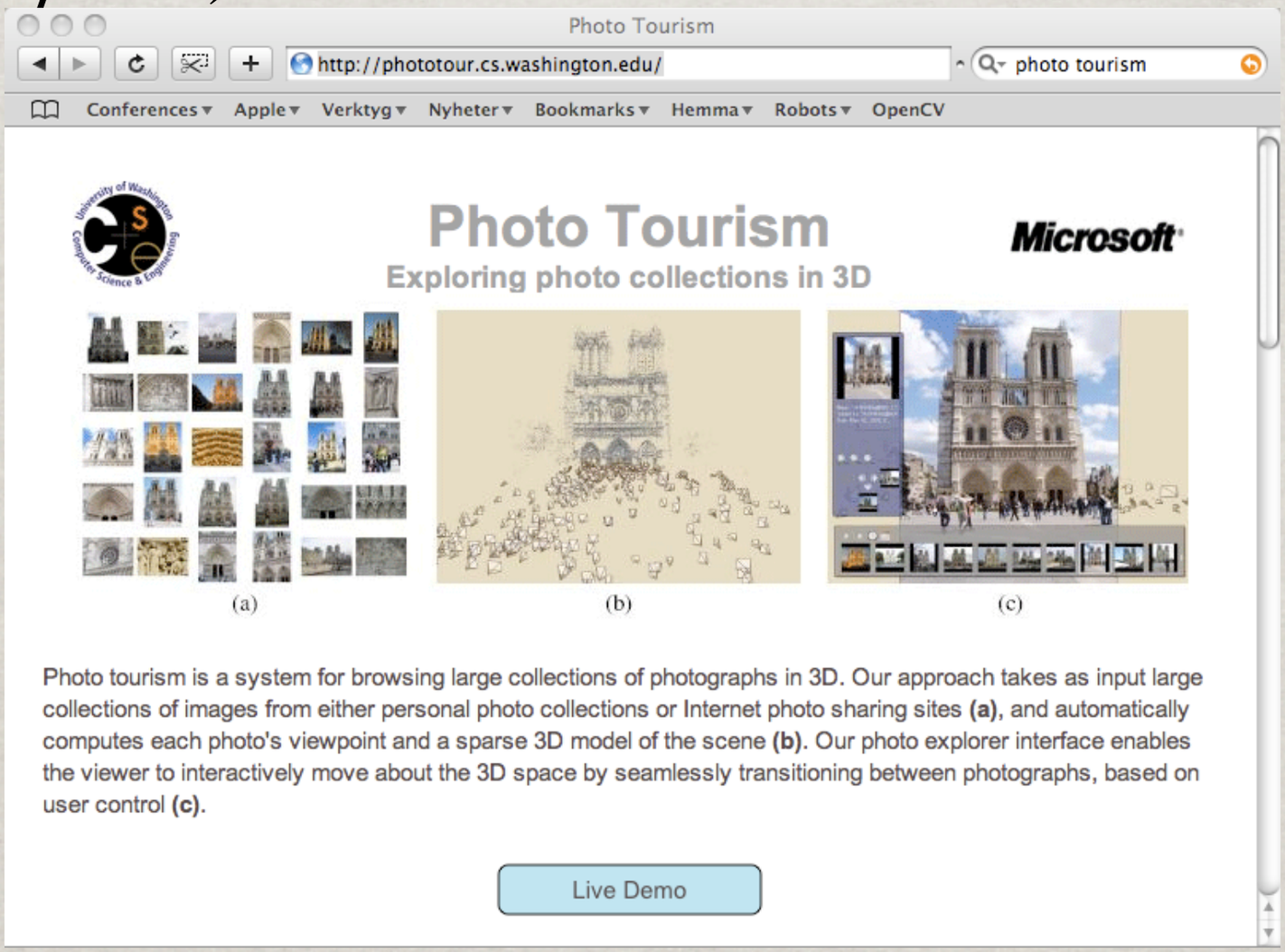
- Use the "standard SIFT pipeline"

- Re-tune all parameters to maximise performance on binary patches.

- Use Mikolajczyk's data set for training.

Learning local Descriptors

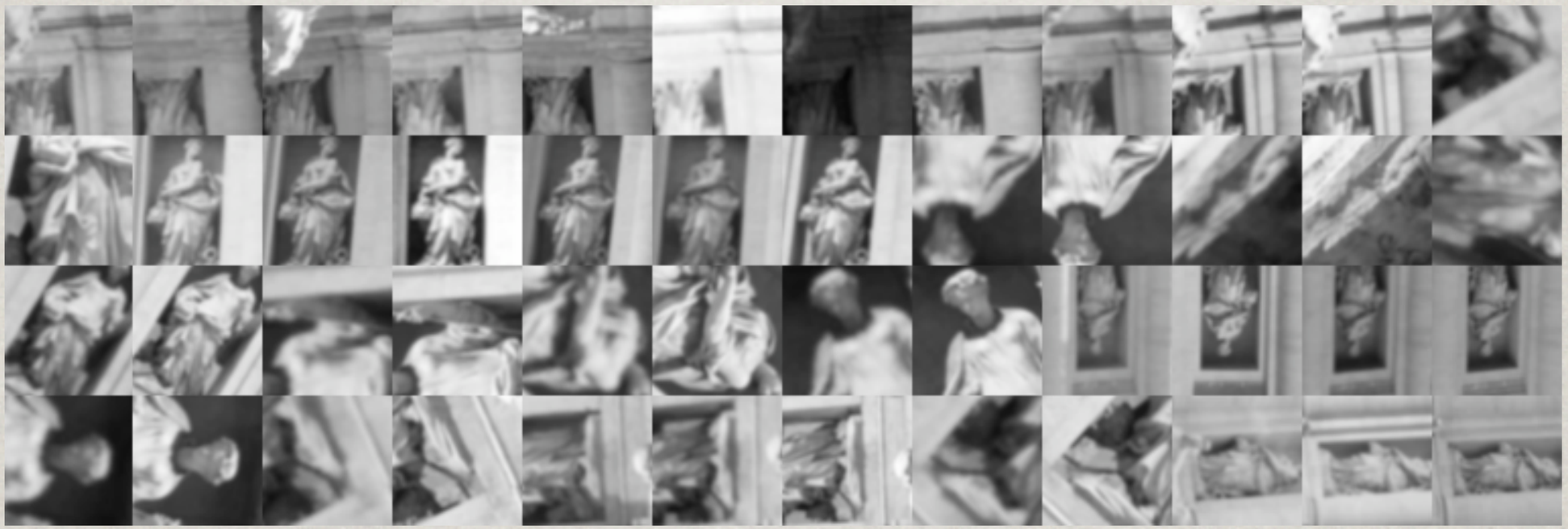
- Winder Brown, *Learning Local Image Descriptors*, CVPR07



- Training data from Photo-tourism dataset.

Learning local Descriptors

- Training data from Photo-tourism dataset.

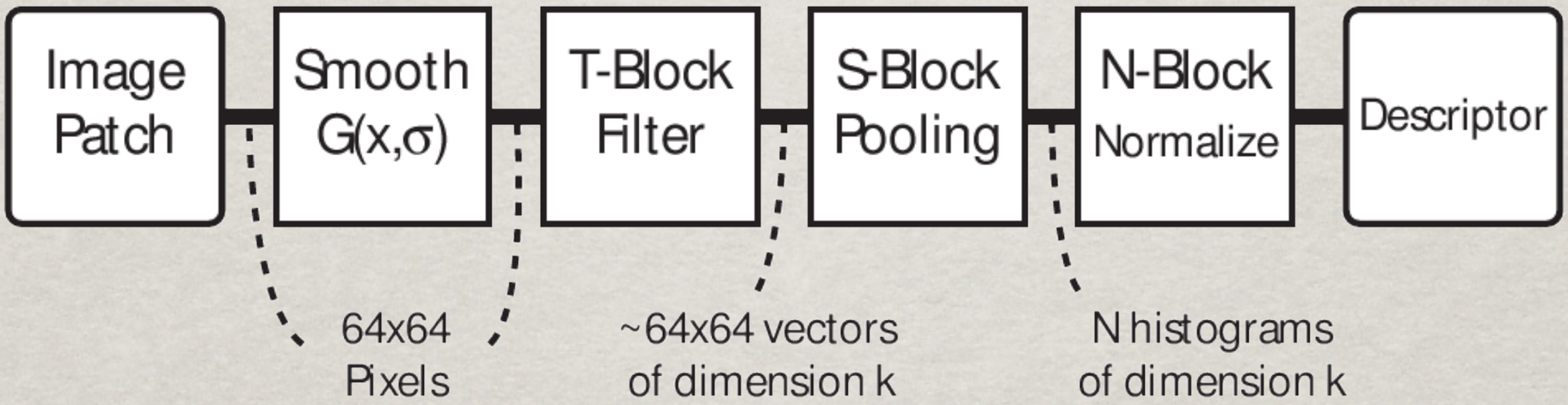


- A research predecessor of MS Photosynth
  http://livelabs.com/photosynth/

- A neat way to arrange photo collections in 3D

Learning local Descriptors

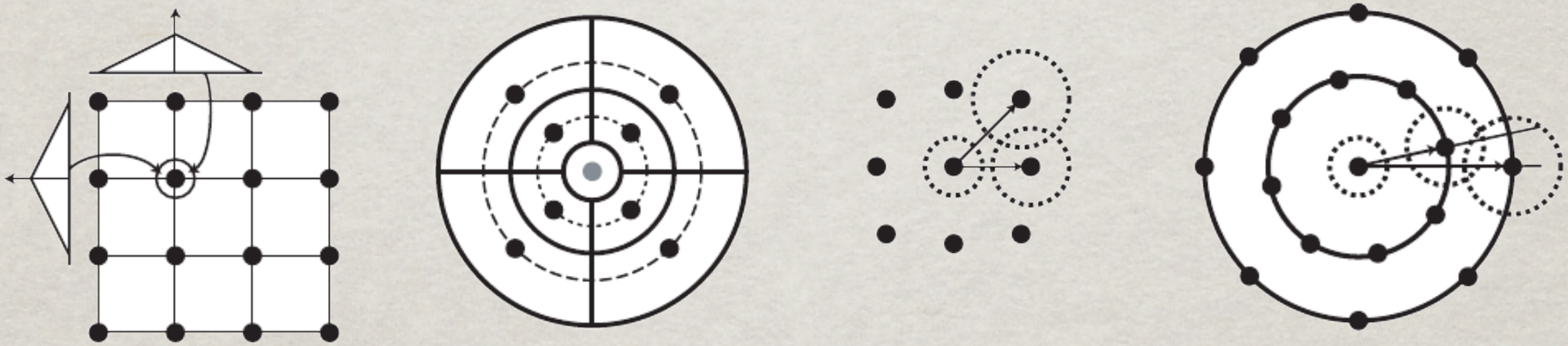
- Most descriptor algorithms can be described by the pipeline:



- Tune all steps of the algorithm with a robust Newton method (Powell's direction set method).

Learning local Descriptors

- Each block has many discrete variants, as well as continuous parameters.



S1: SIFT grid with bilinear weights

S2: GLOH polar grid with bilinear radial and angular weights

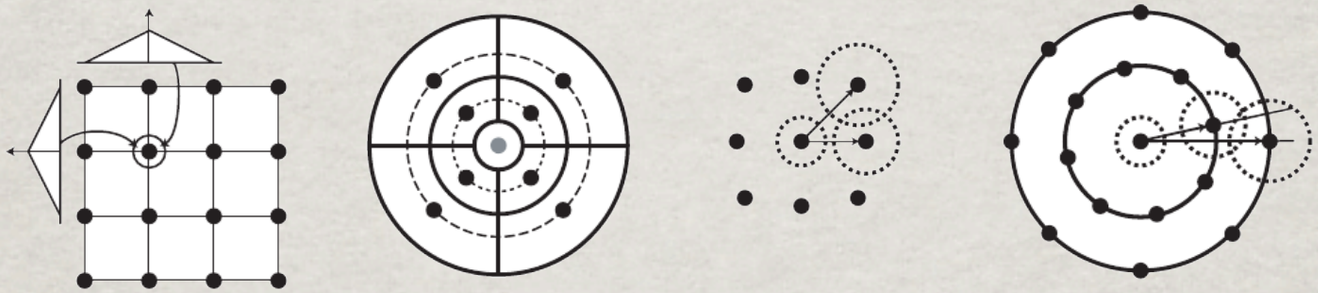
S3: 3x3 grid with Gaussian weights

S4: 17 polar samples with Gaussian weights

- E.g. S-block (Spatial pooling).

Learning local Descriptors

- Results for S-block:

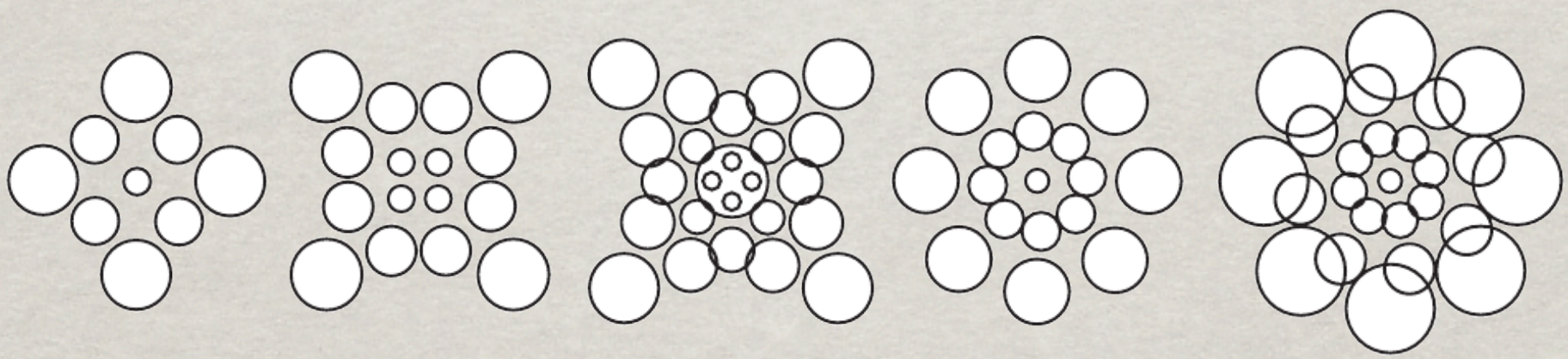


S1: SIFT grid with bilinear weights

S2: GLOH polar grid with bilinear radial and angular weights

S3: 3x3 grid with Gaussian weights

S4: 17 polar samples with Gaussian weights



S3-9 (3x3) S3-16 (4x4) S3-25 (5x5) S4-17 S4-25

- Radially increasing blur seems useful.

Learning local Descriptors

- Experiments used DoG detector.
- Winner was a 4th order quadrature descriptor with 400 elements
- Winner had an error rate of 2%, where the standard SIFT had 6%

Discussion

- Questions/comments on paper and lecture.