Visual Object Recognition

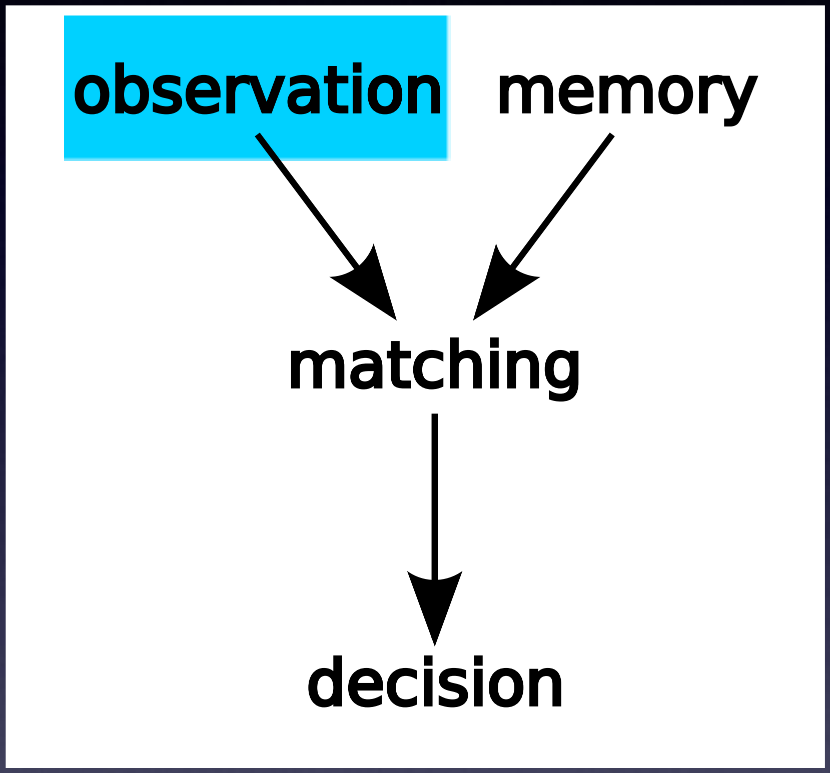
Lecture 3: Descriptors

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

Lecture 3: Descriptors

- Terminology
- An opportunity for machine learning
  DeCAF
- Some common descriptors
  HOG/SIFT, Detector+descriptor pairs, BRIEF, Random Ferns, GaborJet, GIST, Colour Histograms, Shape descriptors

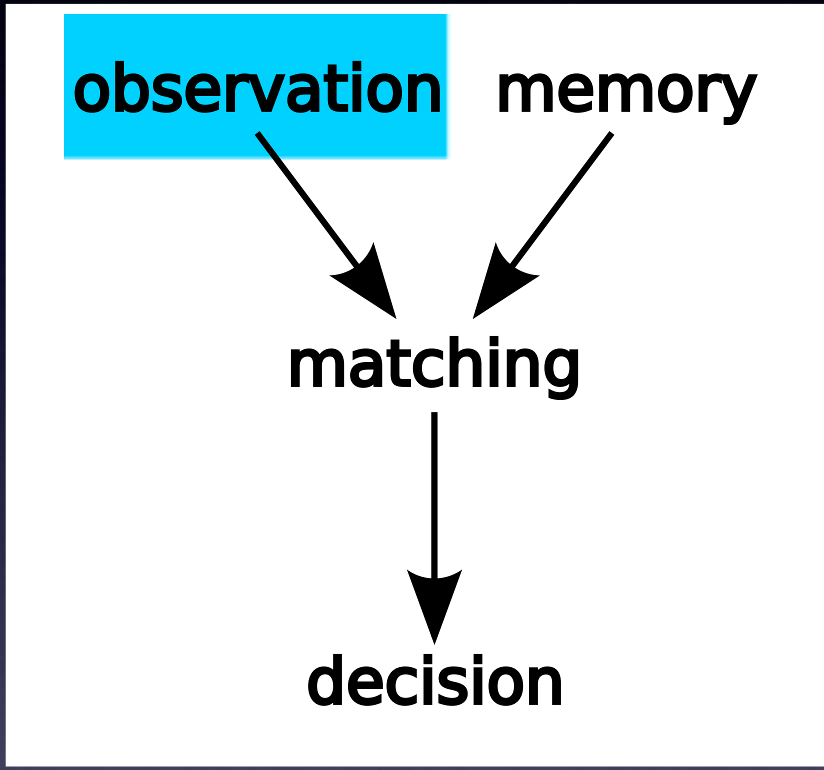
Terminology



- An observation is constructed by **detection** (deciding where to sample) followed by **description** (deciding how to sample)
- Detection is e.g. a canonical frame (LE2), or local affine region detection (LE4)
- The resulting **descriptor** is a vector **v** that can be compared to memory, e.g.

$$\text{match=True, if } ||\mathbf{v} - \mathbf{m}_k|| < \varepsilon$$

Terminology

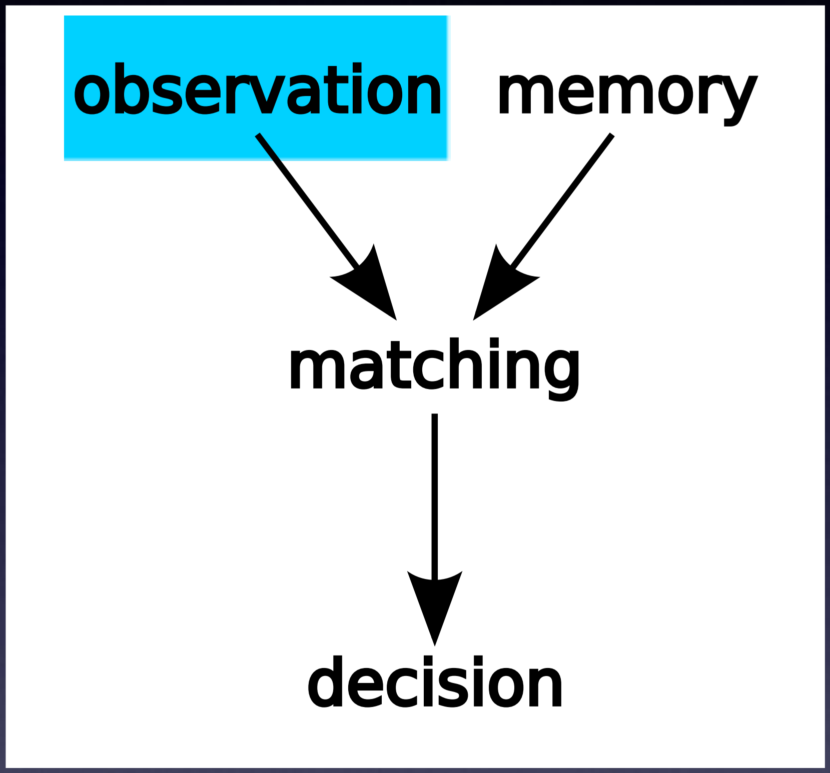


- Desirable properties of a **descriptor** vector:

    1. **invariance to nuisance parameters** such as illumination, small shifts in position and scale of the region

    2. **discriminative power** such that different objects can be told apart

Terminology



- Desirable properties of a **descriptor** vector:

  1. **invariance to nuisance parameters** such as illumination, small shifts in position and scale of the region

  2. **discriminative power** such that different objects can be told apart

$$d(\mathbf{q}, \mathbf{p}_{\text{right\_model}}) = \texttt{small}$$

$$d(\mathbf{q}, \mathbf{p}_{\text{wrong\_model}}) = \texttt{large}$$

Terminology

- Nomenclature for **descriptor** properties:

   1. **Texture**
      Fine details, e.g. wrinkles

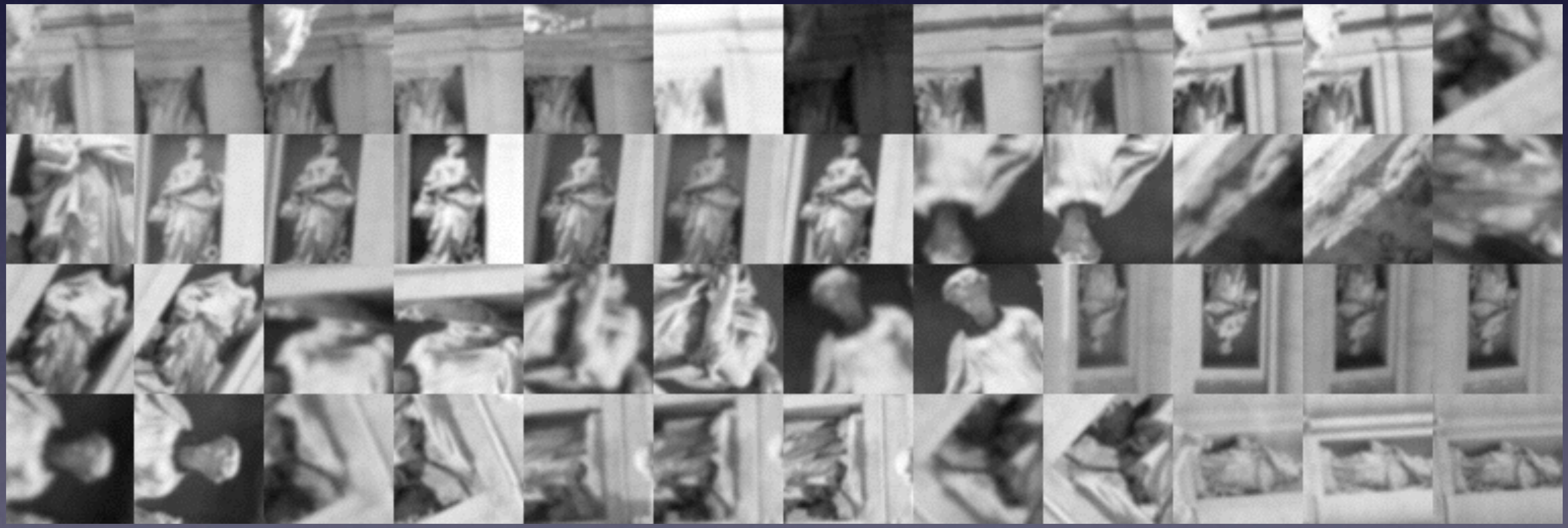
   2. **Colour**
      Surface reflectance properties.

   3. **Shape**
      Coarse details, e.g. contours and depth boundaries

- In there is overlap, caused by the estimation process.

Opportunity for Machine Learning

- With access to a large set of labeled examples, it is possible to use machine learning to find good image descriptors.



Dataset from:
Brown, Hua, Winder, "Discriminative Learning of Local Image Descriptors", PAMI 2011

Opportunity for Machine Learning

- Methods to learn patch appearance (LE4,LE7) can be used.

  + a learned descriptor can improve performance significantly, compared to a hand-coded one.

  - high-dimensional learning requires large amounts of training data.

  - learned descriptors are computationally expensive.

- Using hand-coded descriptors saves computations and is thus very common for practical applications.

Opportunity for Machine Learning

- Example: Jeff Donahue, Yangqing Jia et al., "DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition", ArXiv'13

- DeCAF_6 and DeCAF_7 are pre-trained feature sets (i.e. descriptors) obtained by training the Convolutional Neural Network Classifier CAFFE on the ImageNet database (14M images, 1000 categories)

- The CNN had 5 convolutional layers, and three fully connected layers, 6-8, DeCAF_6 and DeCAF_7 are the outputs from layers 6&7.

- Demonstrated usefulness as generic descriptors, for object recognition, subcategory recognition, and scene recognition.

Designed descriptors

- Most descriptors in use today are still designed
- In practise, all designed descriptors have parameters that have been tuned, i.e. a form of learning is also used here

Intensity normalisation

- A very simple descriptor is the intensity normalized patch we saw in LE2

$$\mathbf{v} = \frac{\tilde{\mathbf{v}} - \mu(\tilde{\mathbf{v}})}{\sigma(\tilde{\mathbf{v}})}$$

- where $\tilde{\mathbf{v}} = \begin{bmatrix} f(\mathbf{x}_1) & \dots & f(\mathbf{x}_n) \end{bmatrix}^T \quad \mathbf{x}_n \in \text{patch}$

Intensity normalisation

- A very simple descriptor is the intensity normalized patch we saw in LE2

$$\mathbf{v} = \frac{\tilde{\mathbf{v}} - \mu(\tilde{\mathbf{v}})}{\sigma(\tilde{\mathbf{v}})}$$

- where $\tilde{\mathbf{v}} = \begin{bmatrix} f(\mathbf{x}_1) & \dots & f(\mathbf{x}_n) \end{bmatrix}^T \quad \mathbf{x}_n \in \text{patch}$

- Why not use ZNCC? (see LE6)

$$d(\mathbf{v}_1, \mathbf{v}_2) = \mathtt{zncc}(\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2)$$

Intensity normalisation

- A very simple descriptor is the intensity normalized patch we saw in LE2

$$\mathbf{v} = \frac{\tilde{\mathbf{v}} - \mu(\tilde{\mathbf{v}})}{\sigma(\tilde{\mathbf{v}})}$$

- where $\tilde{\mathbf{v}} = \begin{bmatrix} f(\mathbf{x}_1) & \dots & f(\mathbf{x}_n) \end{bmatrix}^T \quad \mathbf{x}_n \in \text{patch}$

- Why not use ZNCC? (see LE6)

$$d(\mathbf{v}_1, \mathbf{v}_2) = \mathbf{zncc}(\tilde{\mathbf{v}}_1, \tilde{\mathbf{v}}_2)$$

descriptor comparison should be separable over descriptor dimensions.

Intensity normalisation

- A very simple descriptor is the intensity normalized patch we saw in LE2

$$\mathbf{v} = \frac{\tilde{\mathbf{v}} - \mu(\tilde{\mathbf{v}})}{\sigma(\tilde{\mathbf{v}})}$$

- where $\tilde{\mathbf{v}} = \begin{bmatrix} f(\mathbf{x}_1) & \dots & f(\mathbf{x}_n) \end{bmatrix}^T \quad \mathbf{x}_n \in \text{patch}$

- We will now go through some commonly used, and more advanced descriptors.

The HOG descriptor

- Nearly identical to the SIFT-descriptor (LE4), but adapted to dense grids

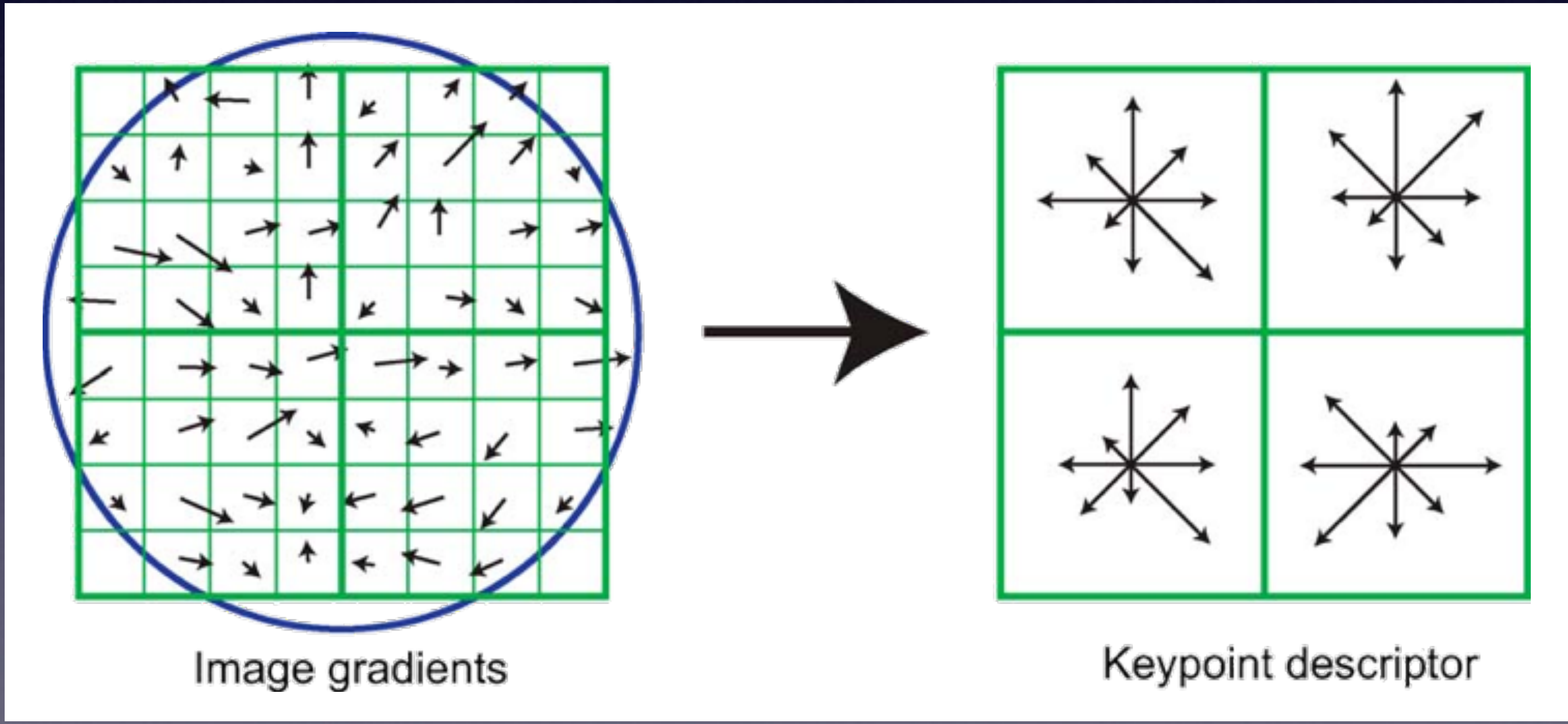


Image from SIFT paper [Lowe IJCV'04]

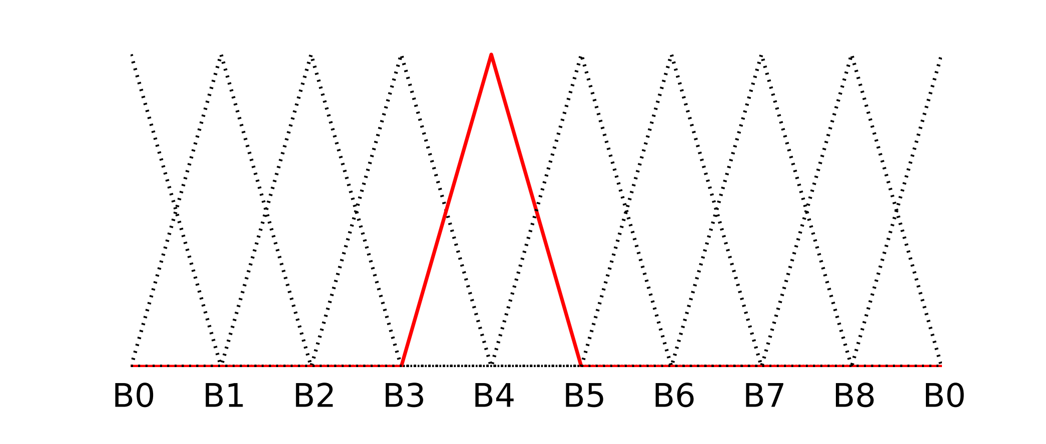
The HOG descriptor

- Compute gradient with small filters

$$\nabla f(\mathbf{x}) = (f * \begin{bmatrix} d_x \\ d_y \end{bmatrix})(\mathbf{x}) \qquad \begin{aligned} d_x &= [-1\ 0\ 1] \\ d_y &= [-1\ 0\ 1]^T \end{aligned}$$

- Perform orientation binning with

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



The HOG descriptor

- Each cell now contains K values (K=9)

$$\mathbf{h}_l = \begin{bmatrix} h_{l,0} & \dots & h_{l,9} \end{bmatrix}^T$$

- These are grouped into 2x2 blocks

$$\tilde{\mathbf{b}} = \begin{bmatrix} \mathbf{h}_1^T & \mathbf{h}_2^T & \mathbf{h}_3^T & \mathbf{h}_4^T \end{bmatrix}^T$$

- and finally, the blocks are normalized

$$\mathbf{b} = \tilde{\mathbf{b}} / \|\tilde{\mathbf{b}} + \epsilon\|$$

The HOG descriptor

- Each cell now contains K values (K=9)

$$\mathbf{h}_l = \begin{bmatrix} h_{l,0} & \dots & h_{l,9} \end{bmatrix}^T$$

- These are grouped into 2x2 blocks

$$\tilde{\mathbf{b}} = \begin{bmatrix} \mathbf{h}_1^T & \mathbf{h}_2^T & \mathbf{h}_3^T & \mathbf{h}_4^T \end{bmatrix}^T$$

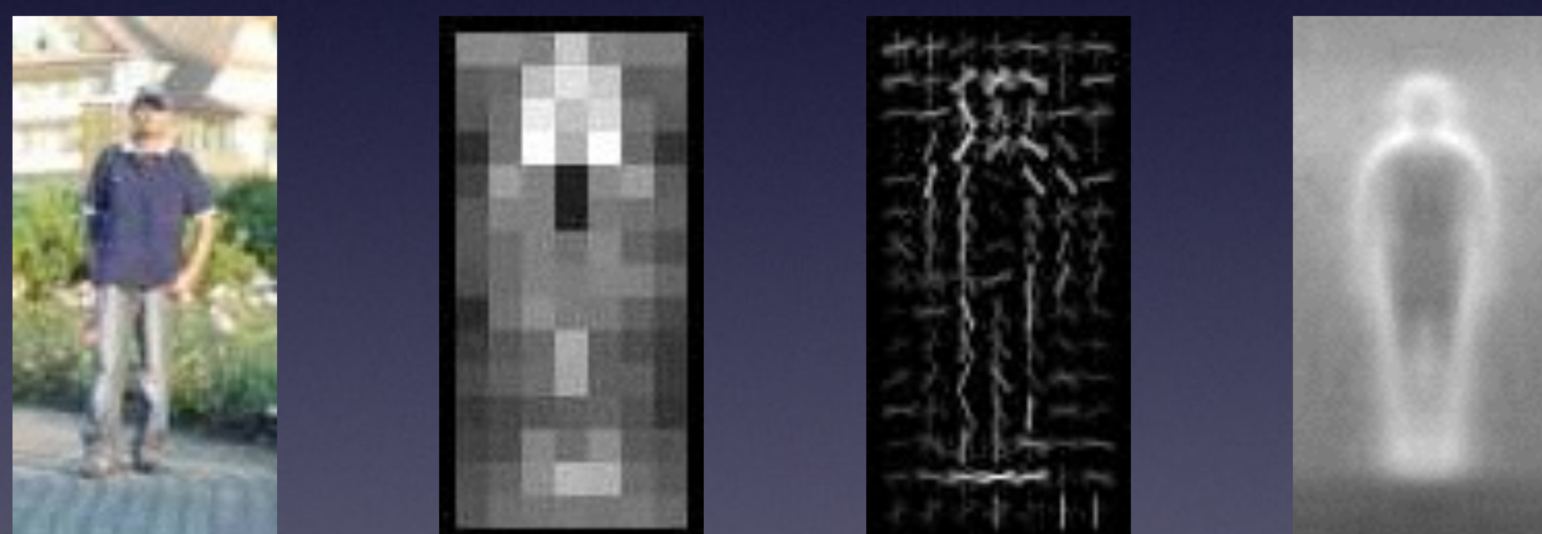
- and finally, the blocks are normalized

$$\mathbf{b} = \tilde{\mathbf{b}} / \|\tilde{\mathbf{b}} + \epsilon\|$$

- Blocks typically overlap, so each cell belongs to several blocks

The HOG descriptor

- The HOG descriptor was introduced in the paper ”Histograms of Oriented Gradients for Human Detection”, Dalal & Triggs, CVPR’05



- Still very common (>9500 citations in Google Scholar)

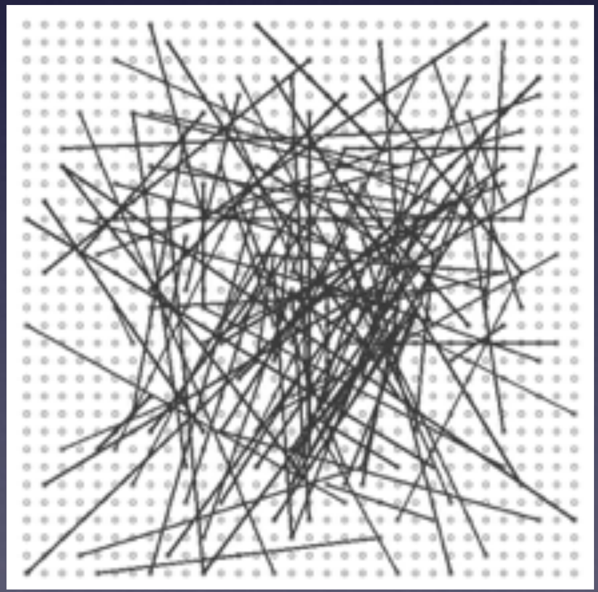
Detector+descriptor pairs

- SIFT, Scale Invariant Feature Transform [D. Lowe ICCV'99, IJCV'04]
- An interest point detector (DoG) + a descriptor
- 4x4 HOG blocks, with a single common normalization
- Other common detector+descriptor features: SURF, BRISK, ORB, SFOP, FREAK (Covered in LE4)

BRIEF

- M. Calonder et al., ”BRIEF: Binary Robust Independent Elementary Features”, ECCV’10, (also PAMI’12)
- A binary descriptor based on intensity differences of pixel pairs, **x**,**y**

$$\tau(\mathbf{x}, \mathbf{y}) = \begin{cases} 1 & \text{if } I(\mathbf{x}) < I(\mathbf{y}) \\ 0 & \text{otherwise.} \end{cases}$$



- $\mathbf{p} = \sum_{i=1}^{n_d} 2^{i-1} \tau(\mathbf{x}_i, \mathbf{y}_i)$
- $\mathbf{x}, \mathbf{y} \in \mathcal{N}(0, S^2/25)$ for an SxS patch.

BRIEF

- M. Calonder et al., ”BRIEF: Binary Robust Independent Elementary Features”, ECCV’10, (also PAMI’12)
- 256 **bit** instead of e.g. 128 **byte** for SIFT (4x size reduction)
- Descriptor comparison is done with

$$d(\mathbf{p},\mathbf{q}) = \mathrm{bitcnt}(\mathrm{XOR}(\mathbf{p},\mathbf{q}))$$

- Very efficient when supported by machine SIMD instructions (e.g. SSE4+ and ARM NEON)

BRIEF related

- Detector+descriptors: BRISK, FREAK, ORB are all based on BRIEF.

- Census transform (R. Zabih, J. Woodfill, ECCV'94)
  Compare central pixel to neighbours in patch and check signs.

- Local binary pattern (LBP) (T. Ojala et al., JMLR'96)
  Compare central pixel to neighbours in a circle and check signs.

- Maximum entropy matching by F. Lundberg at CVL ("Vision for a UAV helicopter", K. Nordberg et al. IROS'02 ws.) describes 256 bit descriptor with **x**,**y** uniformly sampled in 32x32 patch.

Random Ferns

- M. Özuysal, P. Fua, V. Lepetit, ”Fast Keypoint Recognition in Ten Lines of Code”, CVPR’07
- Treats descriptor matching as a classification problem.
  Each patch on an object is treated as a class.
- Split BRIEF style bit tests f_j into groups called **ferns** (a fern is typically S=10 bit tests)

$$F_k = \sum_{j=1}^{S} 2^{i-1} f_{j,k}$$

- Train patch appearance on re-sampled local neighbourhood with added noise.

Random Ferns

- Train patch appearance on re-sampled local neighbourhood with added noise.

$$P(\text{patch}_i | \{F_k\}) \approx P(\{F_k\} | \text{patch}_i) \approx \prod_k P(F_k | \text{patch}_i)$$

- Many samples are needed (2^S=1024 bins to populate, for S=10)

- Frequency count with rule of succession bias (aka. Dirchlet prior)

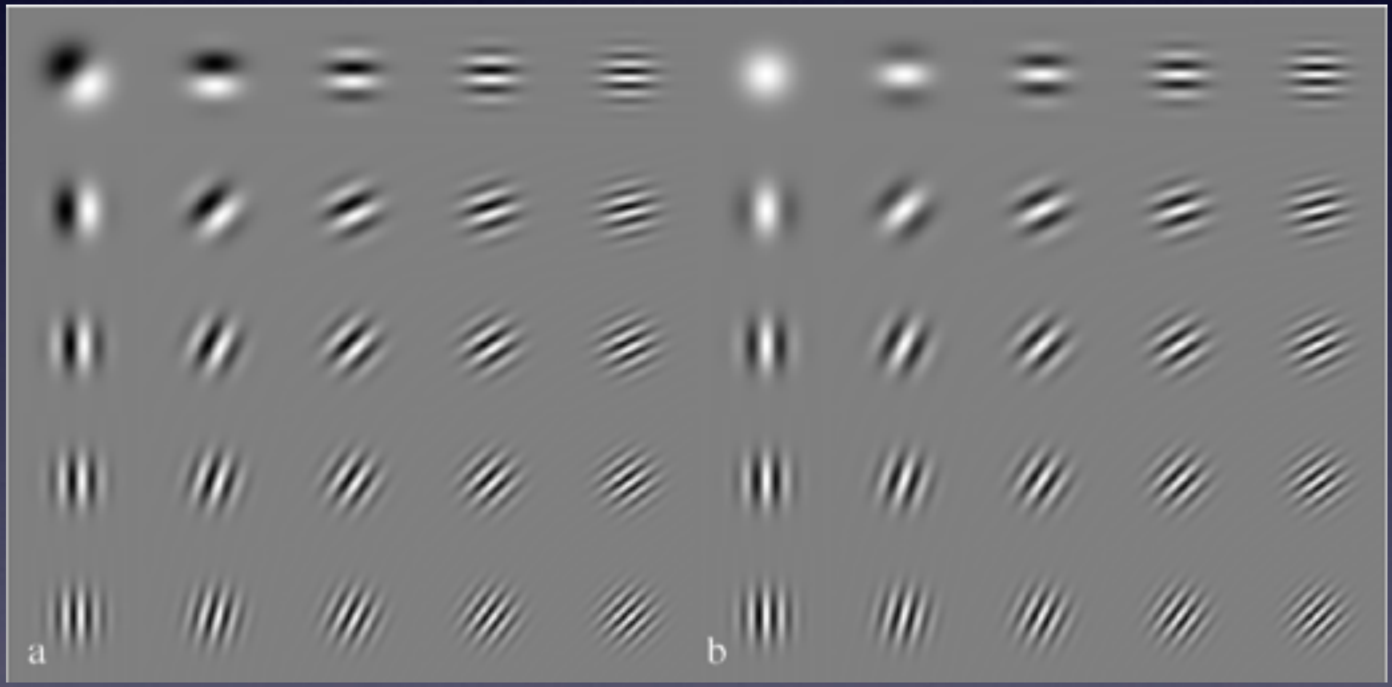
$$P(F_k | \texttt{patch}_i) = \frac{n_{k,i} + 1}{\sum_j n_{k,j} + 1}$$

- i.e. for unpopulated bins, a uniform class distribution is assumed.

Gabor Jet

- A set of responses from filters that are oriented and localized wavelets

$$g(\mathbf{x}, \omega, \hat{\mathbf{n}}, \phi, \sigma) = \exp(i\omega\hat{\mathbf{n}}^T\mathbf{x} + i\phi)\exp(-\mathbf{x}^T\mathbf{x}/2\sigma^2)$$



Tai Sing Lee, "Image Representation using Gabor Wavelets", PAMI'96

- A **filter bank**. Other filter banks include e.g. derivative filters in multiple scales, and wavelets.

Gabor Jet

- Filter banks are typically used to classify texture, e.g. E. Hayman et al. ”On the Significance of Real-World Conditions for Material Classification”, ECCV’04

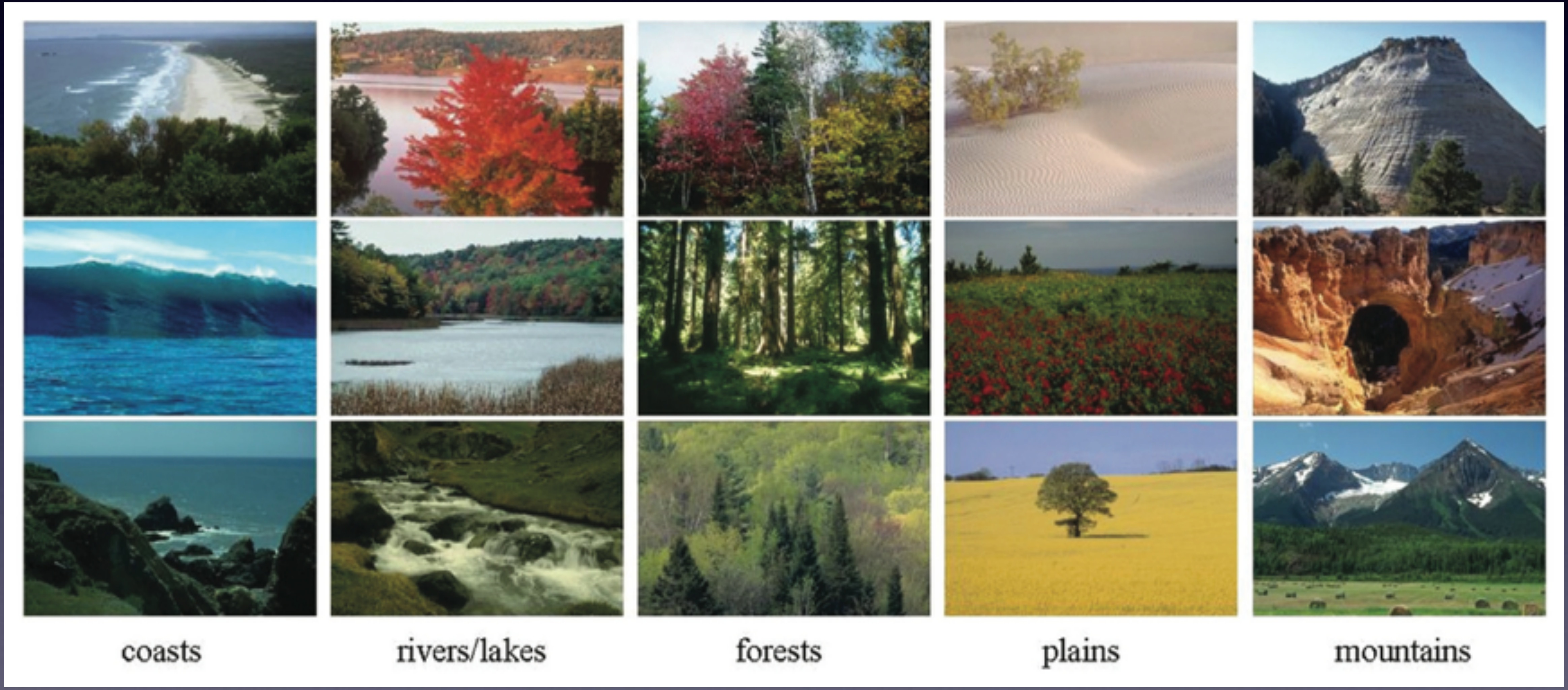
KTH TIPS2 dataset

GIST

- A. Olivia and A. Torralba, ”Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope”, IJCV’01
- A global feature for images that is useful in **scene categorization**.
- Motivation: Perceptual studies indicate that scene category is recognized before semantic information such as objects and their relations.

GIST

- Scene categorization dataset



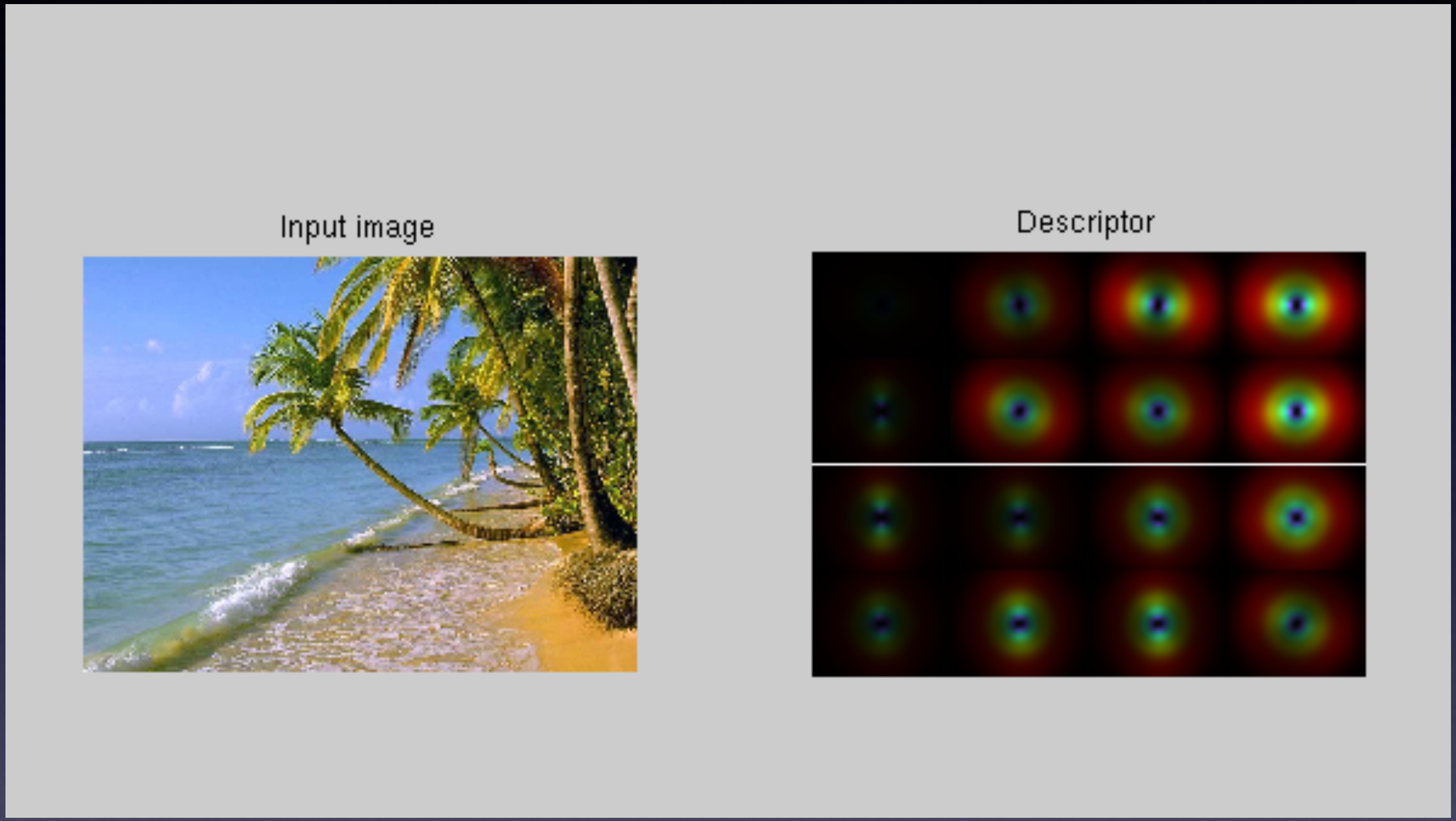
J. Vogel et al. "Categorization of Natural Scenes: Local versus Global Information and the Role of Color", Applied Perception 2007

GIST examples

http://people.csail.mit.edu/torralba/code/spatialenvelope/

- Gabor jets in 4x4 grid (4 scales, 8 directions) on downsampled images (128x128) 512 element descriptor.

GIST examples



http://people.csail.mit.edu/torralba/code/spatialenvelope/

- Gabor jets in 4x4 grid (4 scales, 8 directions) on downsampled images (128x128) 512 element descriptor.

Colour histograms

- Many different variants. E.g. from C. Carson et al. ”Blobworld: A system for region-based image indexing and retrieval”, ICVIS’99

  1. Transform region of interest into La*b* colour space.
  2. Use coarse binning of Lab space, 5x10x10 bins
  3. select the 218 bins that fall within the RGB gamut.

- Spatial position is discarded.
  ⇒ Shift insensitive, scale insensitive.

Colour histograms

- Colour Names, J. van de Weijer et al. ”Learning Color Names for Real-world Applications”, TIP’09
- Label pixels as one of 11 different colours:

- Non-uniform decision regions in Lab space.
- Descriptor by histogramming.

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 detected correspondences on Mikolajczyk's data set for parameter tuning.

Shape descriptors

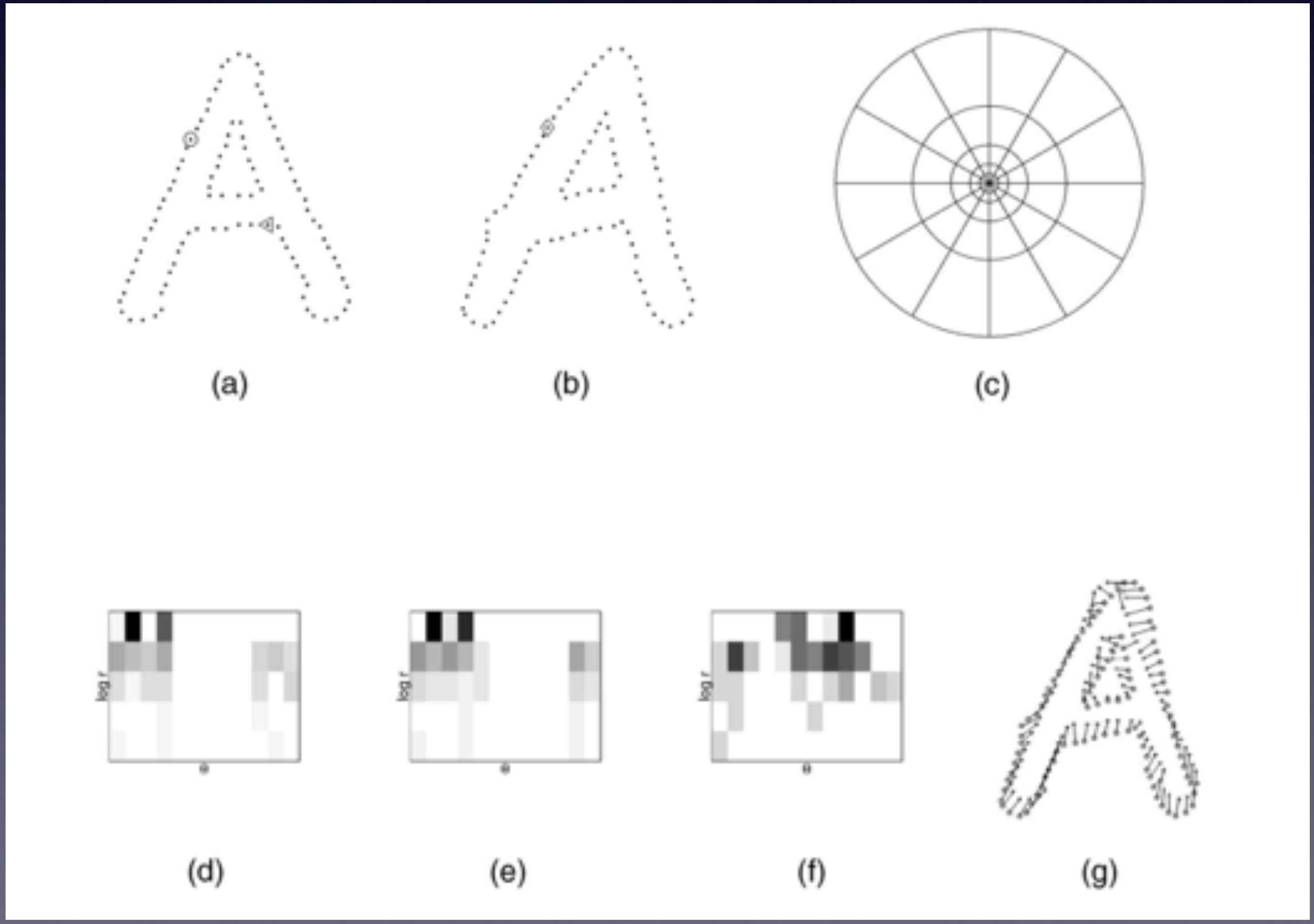
- Other common shape descriptors are: the shape context descriptor, and Fourier descriptors.

- Shape Context descriptors:
  S. Belongie, J. Malik, J. Piuzicha, ”Shape Matching and Object Recognition Using Shape Contexts”, IEEE TPAMI 2002

- Fourier descriptors:
  Granlund, G.H.: ”Fourier Preprocessing for Hand Print Character Recognition”. IEEE Trans. on Computers C–21(2), 195–201 (1972)

Shape descriptors

- S. Belongie, J. Malik, J. Piuzicha, ”Shape Matching and Object Recognition Using Shape Contexts”, IEEE TPAMI 2002
- Log-polar histogram of points along the contour of a binary mask.



Shape descriptors

- F. Larsson, M. Felsberg, P.-E. Forssén, ”Correlating Fourier descriptors of local patches for road sign recognition”, 2011, IET Computer Vision, (5), 4, 244-254.

- Represent points along contour as complex numbers z(t)=x(t)+iy(t), and apply the Fourier transform on the resultant periodic signal.

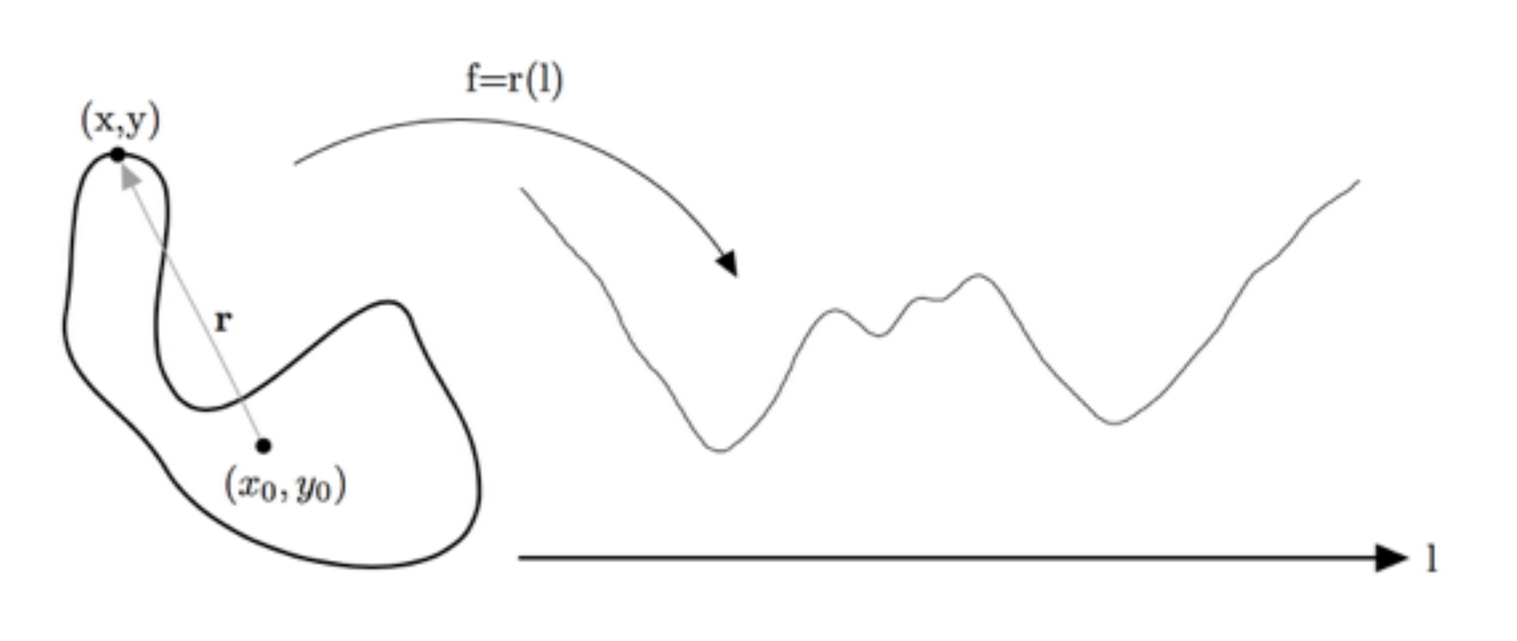


Figure 2.2: Shape signature based on distance to the center of gravity.

Summary

- Descriptors estimate **shape**, **texture** and **colour**.

- Descriptors can be learned, but for speed, and in practise, hand coded descriptors are more common.

- Descriptors where comparison is separable allow fast ANN search.

Discussion

- Questions/comments on paper:

  M. Calonder et al., ”BRIEF: Binary Robust Independent Elementary Features”, ECCV’10, (or extended version in PAMI’12)

Next week

- paper to read for next week:

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