Image Processing and Computer Graphics

Image Processing

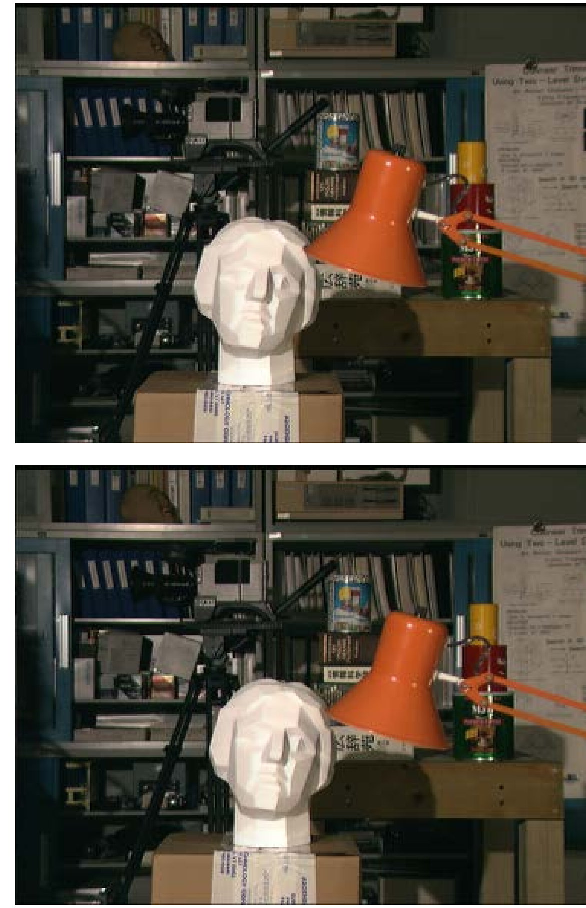
Class 6

Matching and local descriptors

Matching of local structures

- Key problem in computer vision appearing in:
  - Motion estimation
  - Camera calibration
  - Stereo
  - Image retrieval
  - Object recognition

Object recognition: training image on the left, test image on the right. Matching here is quite hard.



Stereo pair: point matching needed to compute depth

Outline

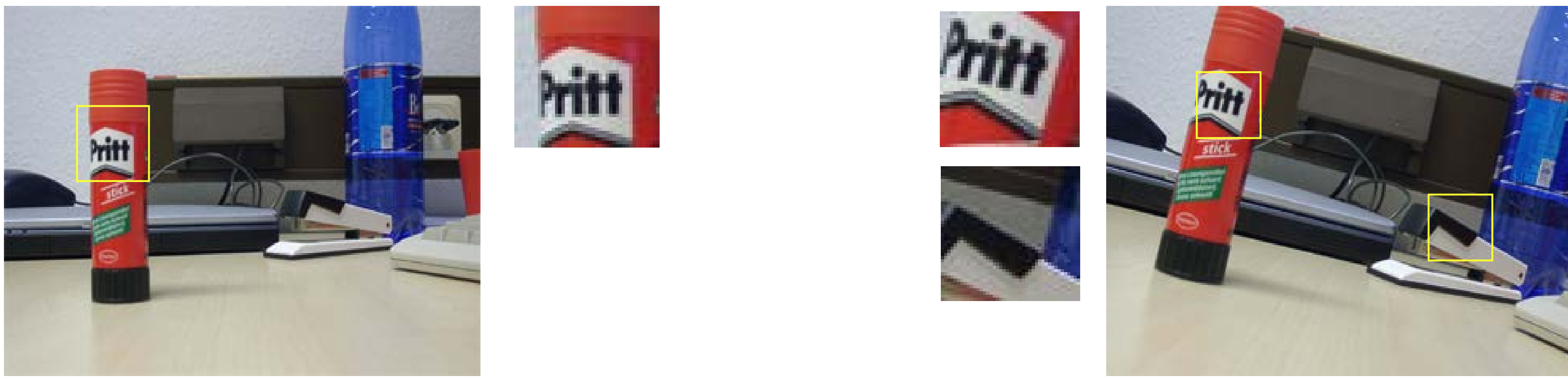
1. Sparse matching
   – What are good regions to match?

2. Manual descriptor design
   – What is required for a good descriptor?

3. Feature learning
   – How can we optimize descriptors for a particular task?

Block matching

- Straightforward way to match points in images:
  - Regard the image patch around each point in image 1
  - Compare it to the image patches around all points in image 2



- Computationally expensive
  $O(kN^2)$, k: size of patch, N: size of image (in pixels)
- Normal image patches are not **invariant** to typical appearance changes
  → normalization or invariant descriptors

Sparse matching via interest points

- Often we need only a limited number of matches

- Idea: Do not match all points in the images, but only promising subsets
  → significantly reduced complexity

- Requirements for good interest points:
  1. Points must come with enough information for unique matching


  2. Subset in image 2 must contain matches from subset in image 1

Corner detection

- Choose points with high information content and clear localization
  → typically corner points

- Corner detection with the structure tensor:
  (Förstner-Gülch 1987, Harris-Stevens 1988)

$$J_\rho = K_\rho * (\nabla I \nabla I^\top) = \begin{pmatrix} K_\rho * I_x^2 & K_\rho * I_x I_y \\ K_\rho * I_x I_y & K_\rho * I_y^2 \end{pmatrix}$$

- Measure of cornerness (unintuitive, but fast to compute):

$$c = \det J_\rho - \alpha \boxed{\mathrm{tr} J_\rho}$$

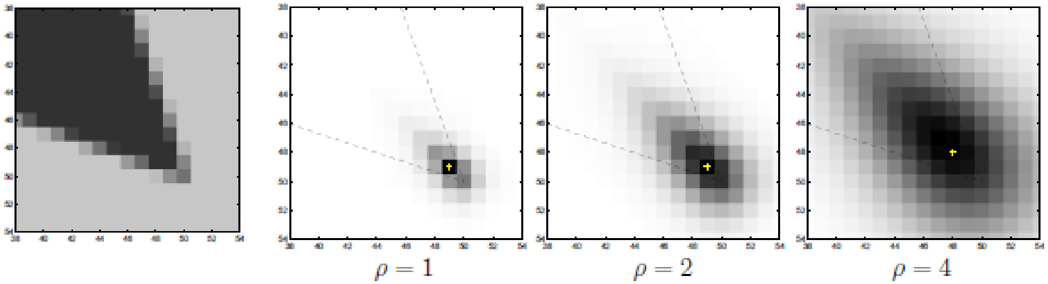
= gradient magnitude

Corner detection

- Eigenvalue decomposition of the structure tensor:

$$J_\rho = T \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} T^\top$$

$$c = \lambda_2$$



Input and second eigenvalue for different ρ

- Interpretation:
  - Smoothing of J integrates gradients from the neighborhood
  - Eigenvectors in T yield the dominant orientation in this neighborhood and the perpendicular orientation
  - Eigenvalues yield the structure magnitude in these directions
  - A large second eigenvalue indicates strong structures in multiple directions → corners

Corner detection

- Corners: local maxima of the second eigenvalue

Scale invariance

- Problem: Detected corners depend on the image scale

- Solution: Create and compare descriptors at multiple scales

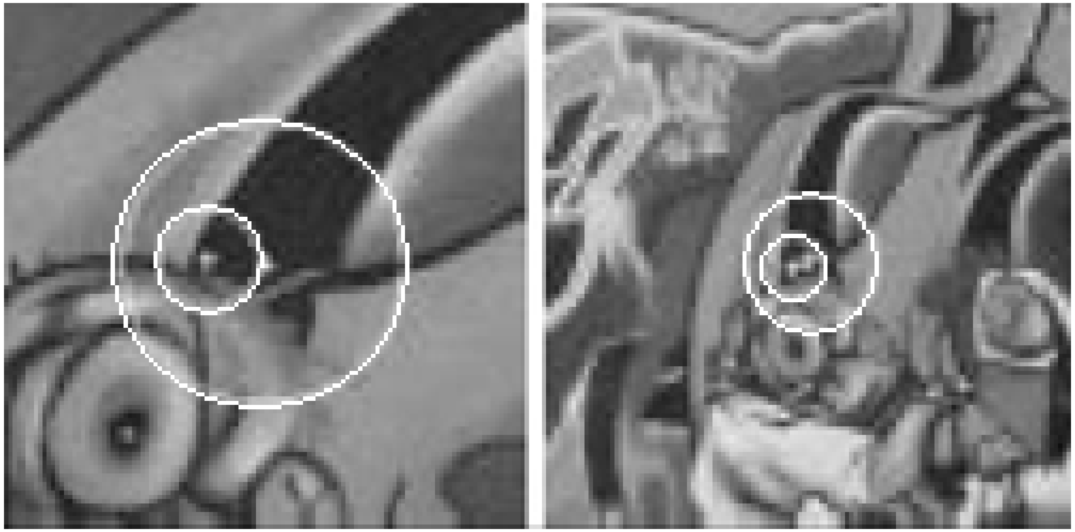
Scale invariant feature transform (SIFT)

- Considers local maxima of the Laplacian in scale space:

$$x^*, y^*, \sigma^* = \text{argmax}_{x,y,\sigma} \left(\sigma^2 \cdot |\partial_{xx}(K_\sigma * I(x,y)) + \partial_{yy}(K_\sigma * I(x,y))| \right)$$

- Laplacian is rather a blob detector than a corner detector

- Detects regions with multiple sizes
  → good to normalize out scale changes



Authors: Krystian Mikolajczyk and Cordelia Schmid

Affine region detector

- Maximally stable extremal regions
  (Matas et al. 2002)
  - Regions encircled by large gradients
  - Obtained by low-level segmentation algorithm

Maximally stable extremal regions and fitted ellipses. Author: Andrea Vedaldi

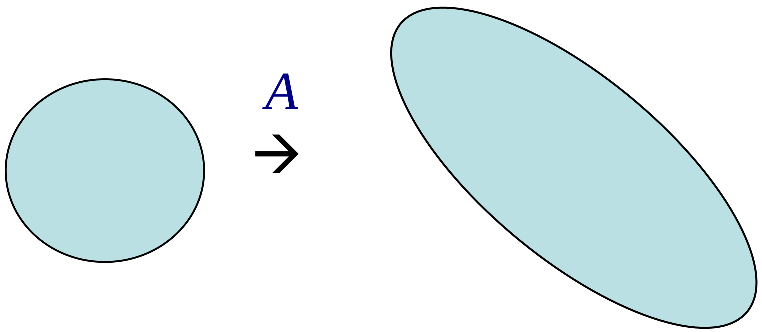
- Apart from scale also yields elongation of fitted ellipses
  → allows for affine invariance

Affine invariance

- Affine transformation (6 degrees of freedom in 2D):

$$f(x) = Ax + t$$

- Maps a circle to an ellipse (or vice-versa):



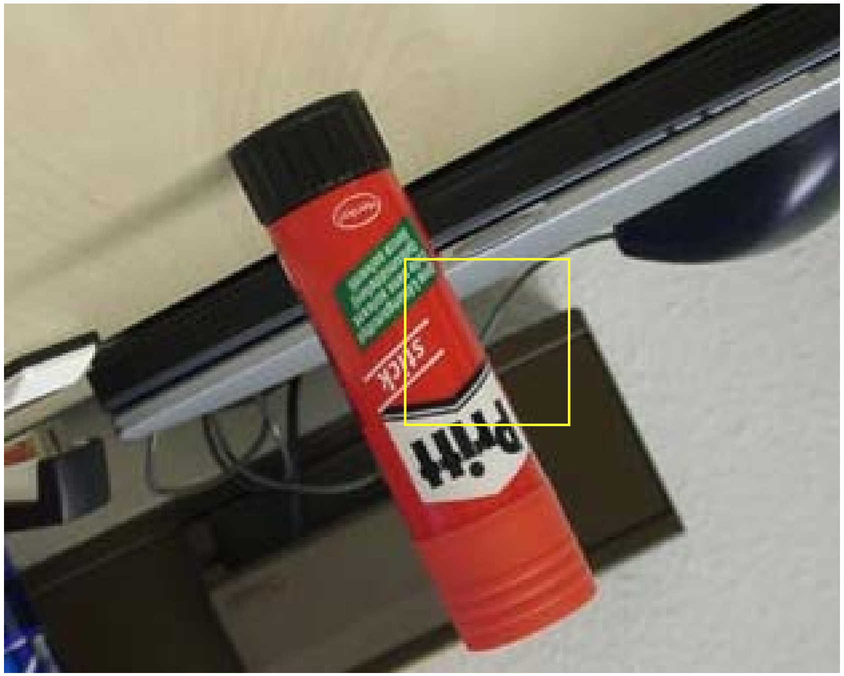
- Approximation of a projective transformation

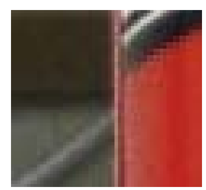
- Parameters can be estimated from a region detector

Summary sparse matching

- Significantly reduced complexity: with 100 detected points in both images, one has to compare only 10000 patches instead of 96 billion(!) in 640x480 images
- Allows us to efficiently normalize out some variations (scale, affine transformations)
- Non-dense displacement fields (important matches might be missed)
- Corresponding patches can be slightly shifted
- Other variations (than affine transformations) not covered

Example: rotation and scaling

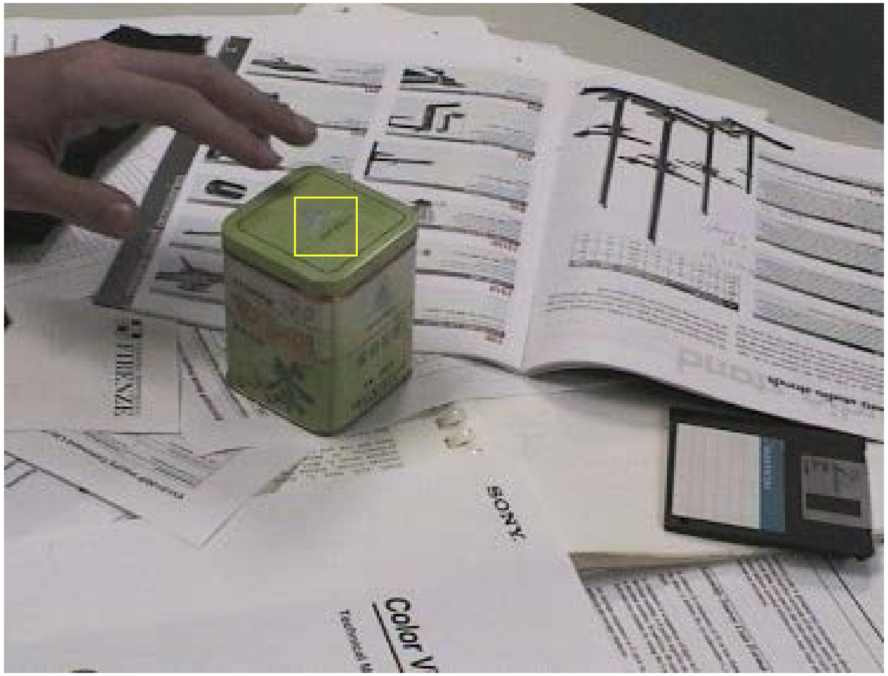


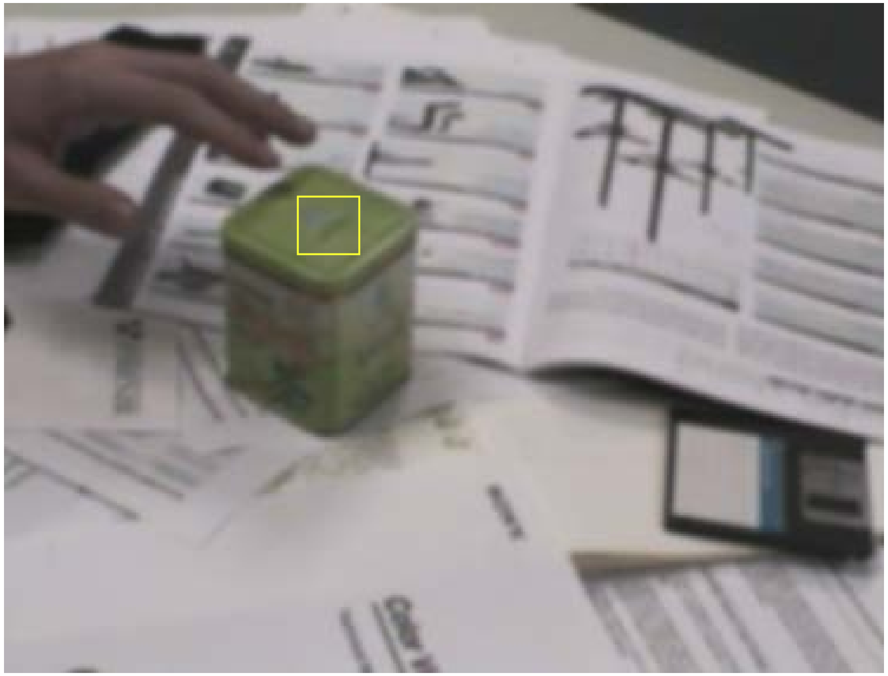


Example: projective transformation

Example: lighting changes

Example: blurring





Invariant local descriptors

- **Local descriptors**: vectors that contain information about the local neighborhood of a point or region of interest

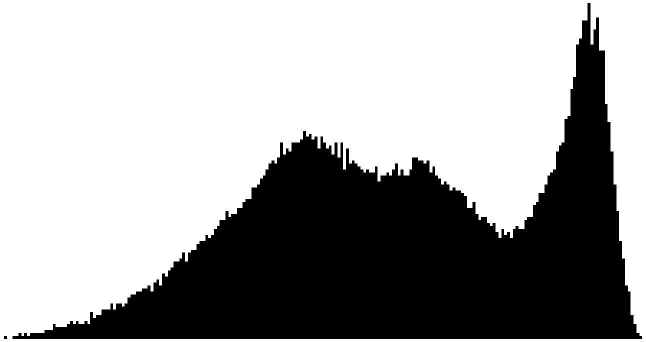
- Goal: design local descriptors that are **invariant** under the mentioned selected transformations

- Be Careful: If a descriptor is invariant to all sorts of transformations, it may not be descriptive anymore

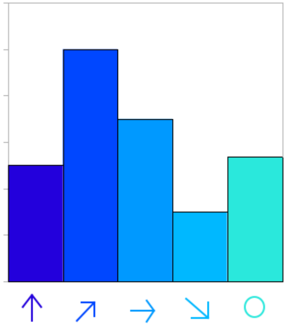
Histograms

- Alternative to a normalized neighborhood:
  derive invariant features within the fixed block

- Gray value histogram:
  - Rotational invariance (good)
  - Invariant to blurring (good)
  - Sensitive to lighting changes (bad)
  - Significant loss of information (very bad)



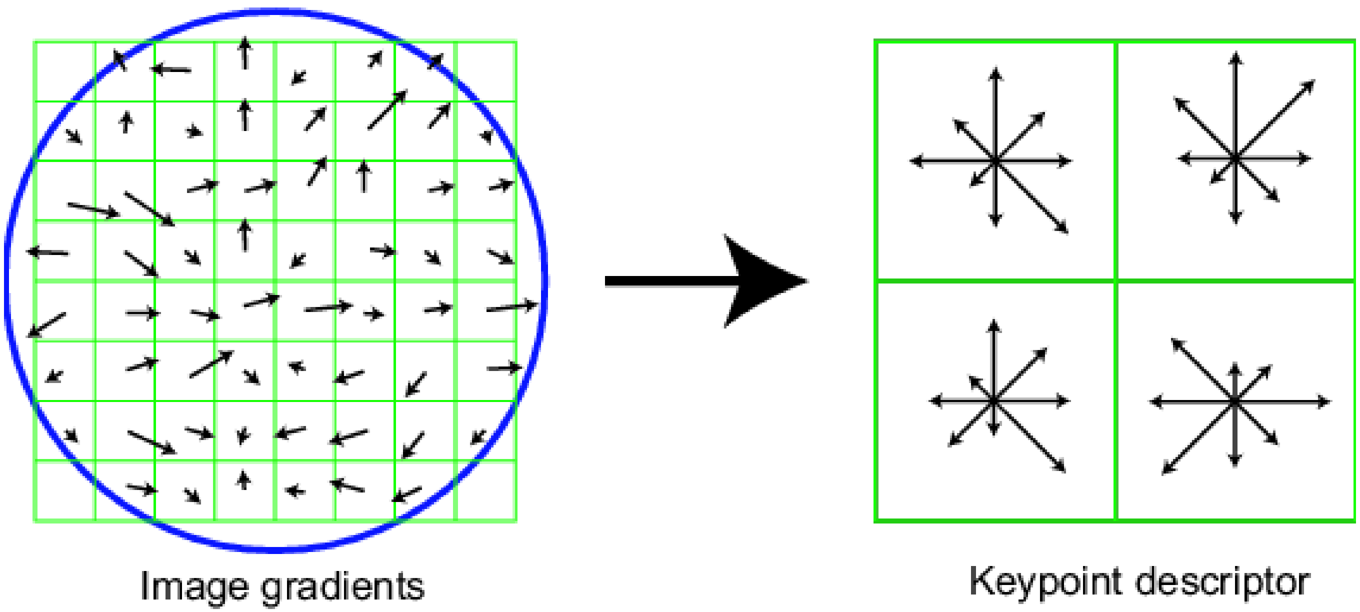
- Histogram of the gradient direction (**orientation histograms**)
  - Invariant to (additive) lighting changes
  - Building block of many successful descriptors



SIFT descriptor

- Popular local descriptor (several variants exist)

- Based on local assembly of orientation histograms and adaptive local neighborhoods



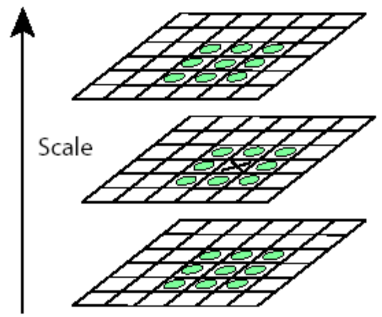
Author: David Lowe

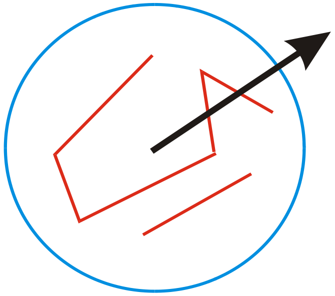
SIFT detector

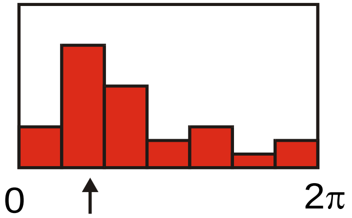
- Extract SIFT feature points
  - Strongest responses of Laplacian in scale space → position and scale
  - Fit quadratic function to obtain subpixel accuracy
- Create orientation histogram at selected scale
  - Peak of smoothed histogram estimates orientation
  - In case of two peaks, create two feature points

→ Estimation of position, scale, and orientation

- Affine invariance can be provided with MSER
- In object recognition: dense sampling of such points at all positions and all scales, no rotation invariance

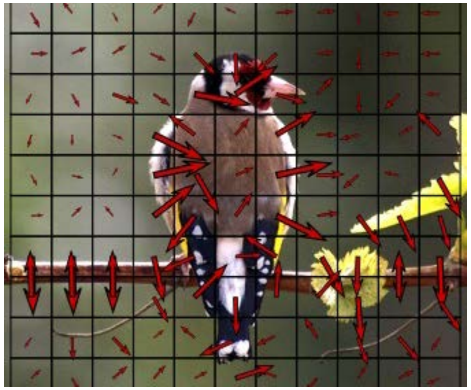




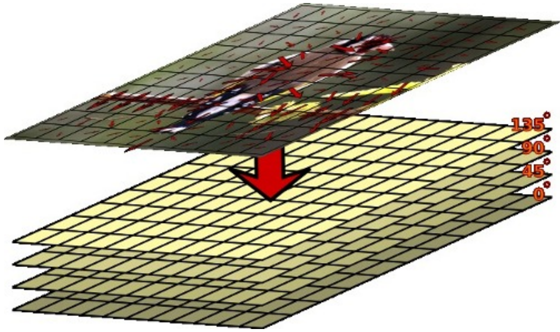


Author: David Lowe

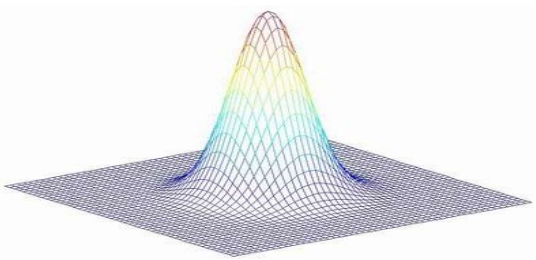
Dense computation of SIFT/HOG descriptors



1. Compute gradient orientation and magnitude at each pixel



2. Compute orientation indicator at each pixel
   - Create NxMx8 array and initialize with zero
   - Quantize the orientation at each pixel (here 8 bins) and add the respective magnitude to the respective entry in the array



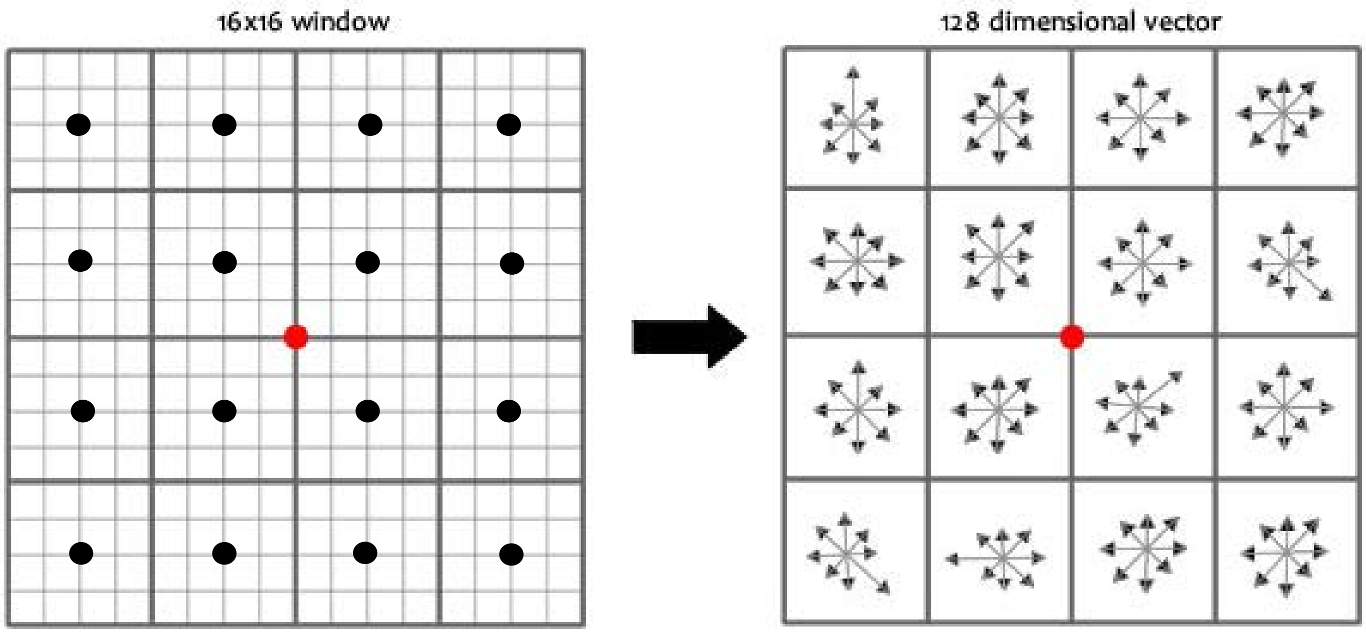
3. Local integration → orientation histogram
   Smooth array with a Gaussian kernel

4. Smooth in orientation direction
   (among neighboring channels)

Dense computation of SIFT descriptors

5. Sample feature vectors from the histogram image
   – Original SIFT:
     4 pixel spacing, 4x4 histogram array
     → 128-D vector

6. Normalize the feature vector to unit length



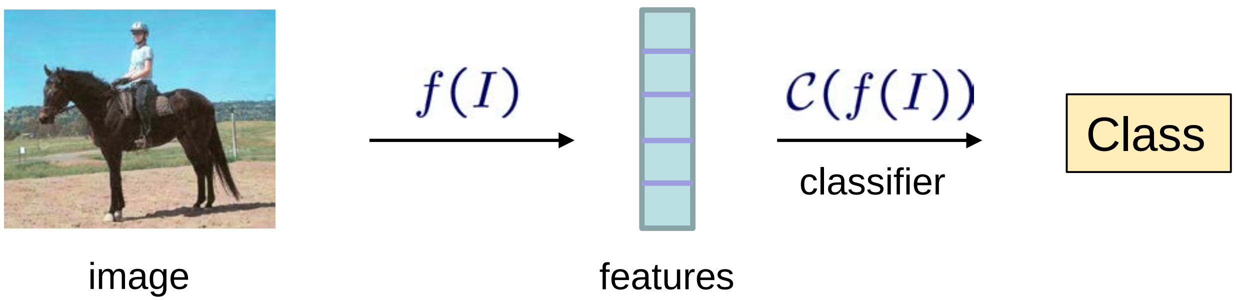
Author: Utkarsh Sinha

http://aishack.in/tutorials/sift-scale-invariant-feature-transform-features/

Feature learning

- Instead of manual descriptor design, let the computer find the optimal descriptor for a defined task and training set

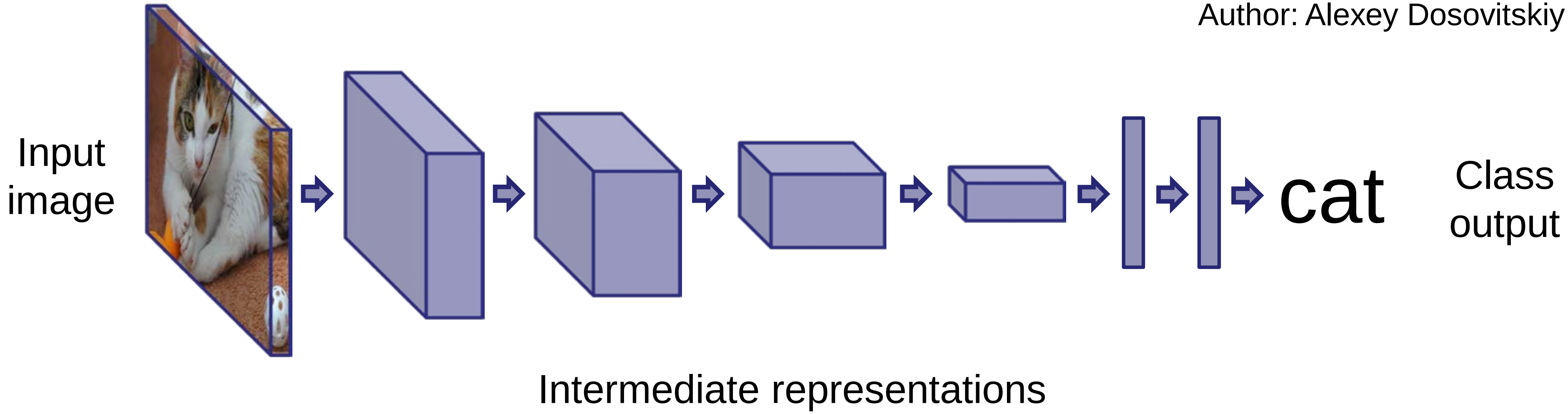
- Example task: object classification
  → training set consists of <u>images and their class labels</u>



- Shallow modeling of the function $f(I)$ is not efficient to cover all the variation that appears in an object class
  → hierarchy of functions, “deep” representation

Descriptors learned with convolutional networks

Author: Alexey Dosovitskiy

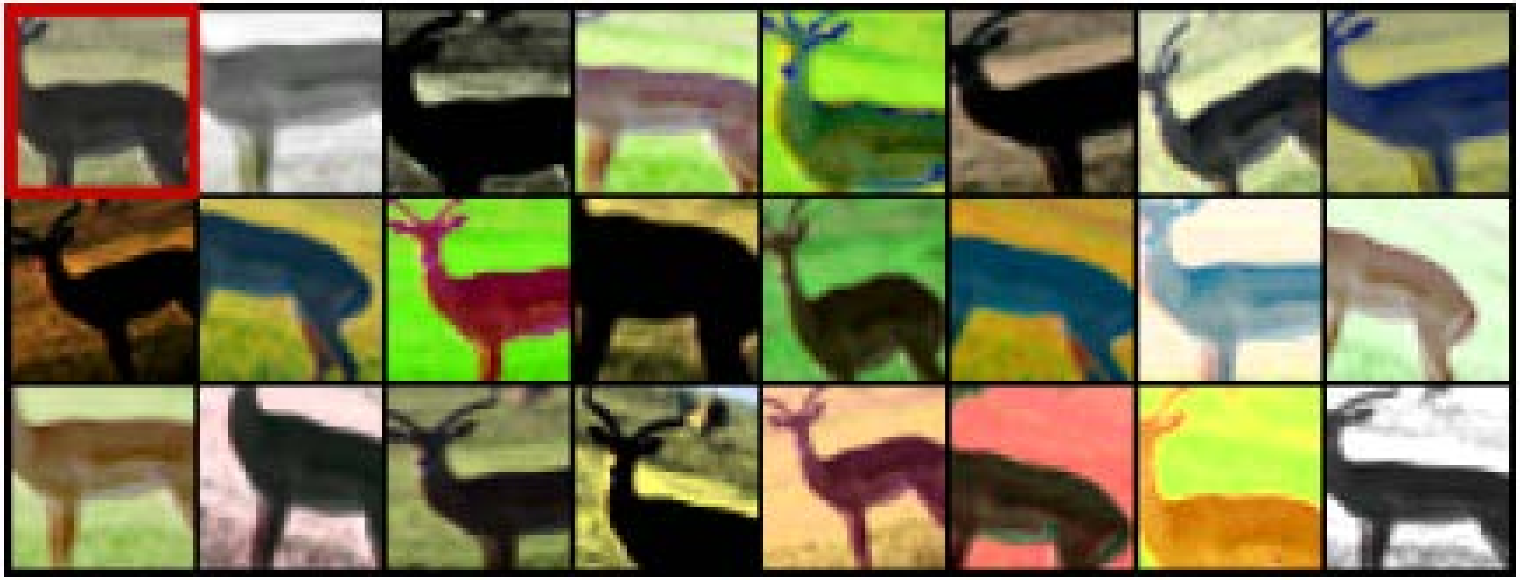


- Classification networks are trained on large datasets with class labels (e.g. ImageNet with 1M images)
  → network learns a representation that is good for object classification

- Intermediate layer outputs turn out to be also good generic descriptors (still a bit mystic what is represented)

Unsupervised training to trigger invariant features

- Train CNN to discriminate **surrogate classes** (Dosovitskiy et al. 2015)

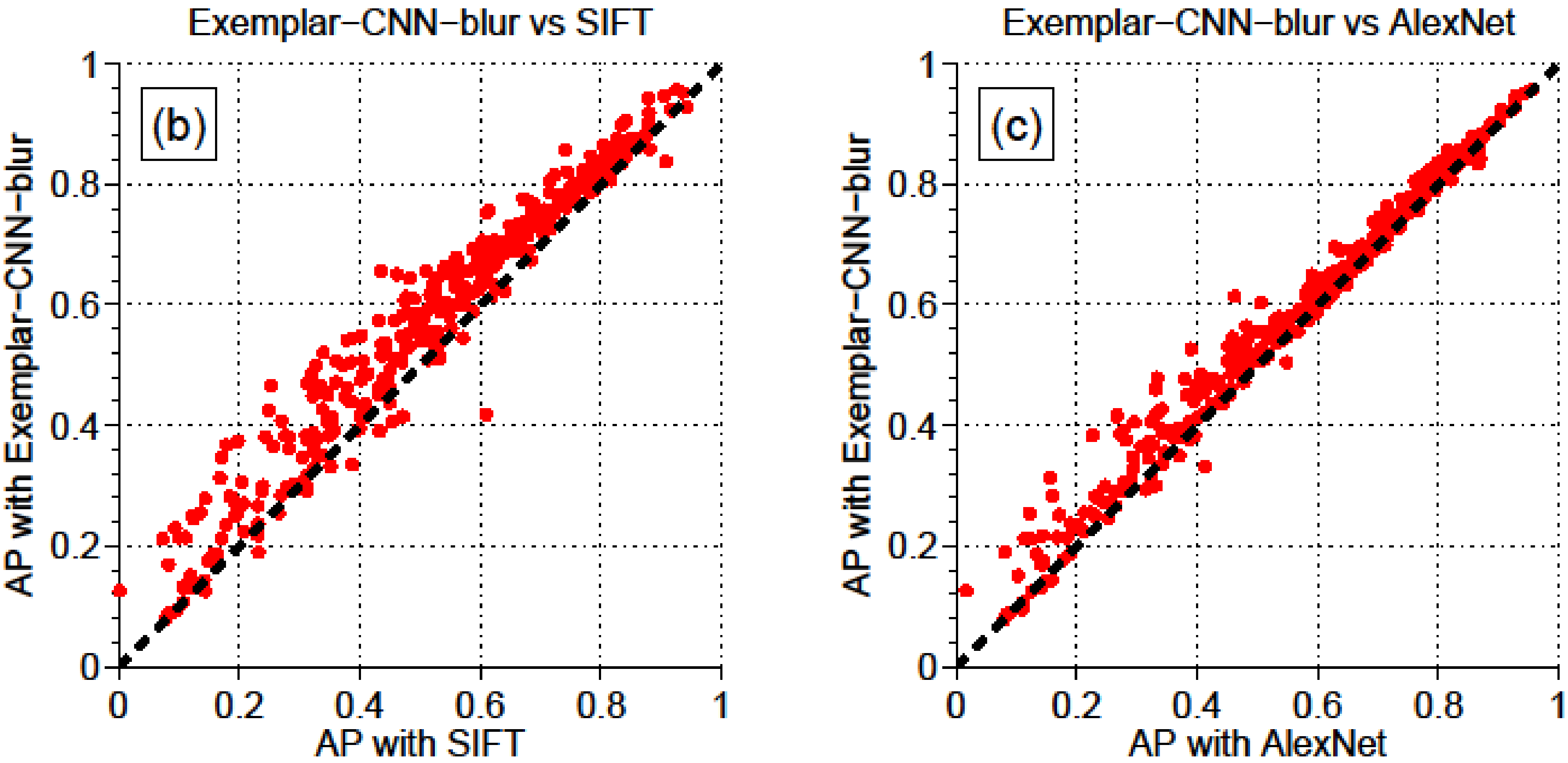


Seed patch and transformed versions of it make up a surrogate class

- Applied transformations:
  translation, rotation, scaling, color, contrast, brightness, blur

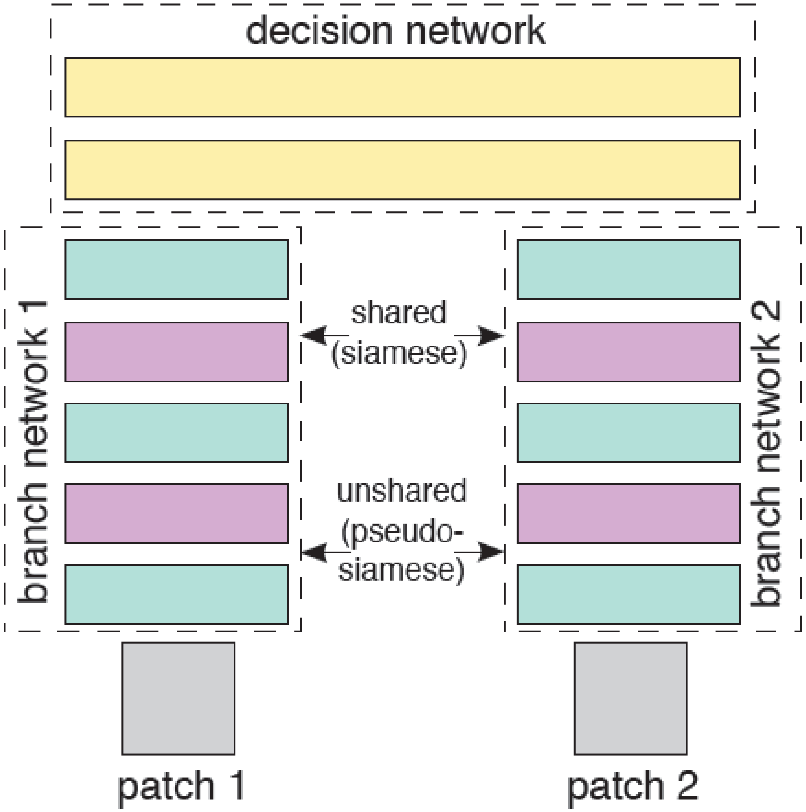
- Transformations define invariance properties of the features to be learned by the network → early version of today's contrastive learning

Descriptor matching performance



Siamese networks

- Trained directly on matching and non-matching patches



Learns the metric

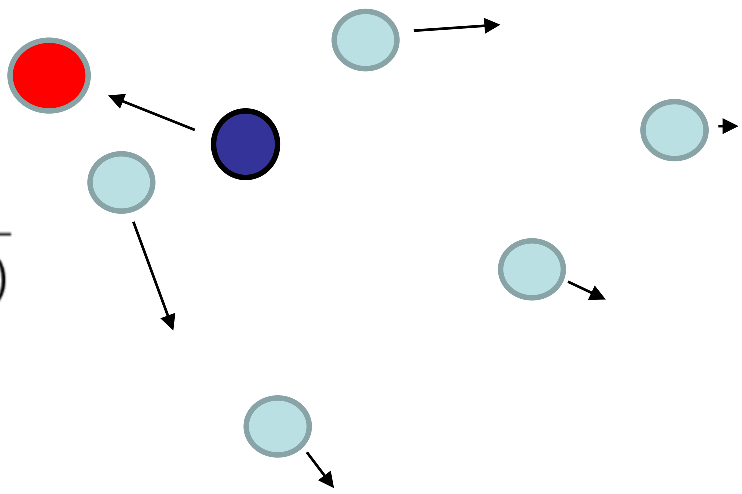
Learns the descriptor

Example from Zagoruyko&Komodakis 2015

- Issue class imbalance: far more non-matching patches than matching ones

Contrastive learning

- The principle of surrogates can be generalized to a contrastive loss to learn a feature embedding.
- Main idea: for a sample define another positive sample that should be close in embedding space, and multiple negative samples that should be far
- Contrastive loss:



$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)}$$

where τ is a scaling parameter that can be reduced over time

- Positives and negatives can often be defined without supervision → self-supervised learning
- Are negative samples needed?

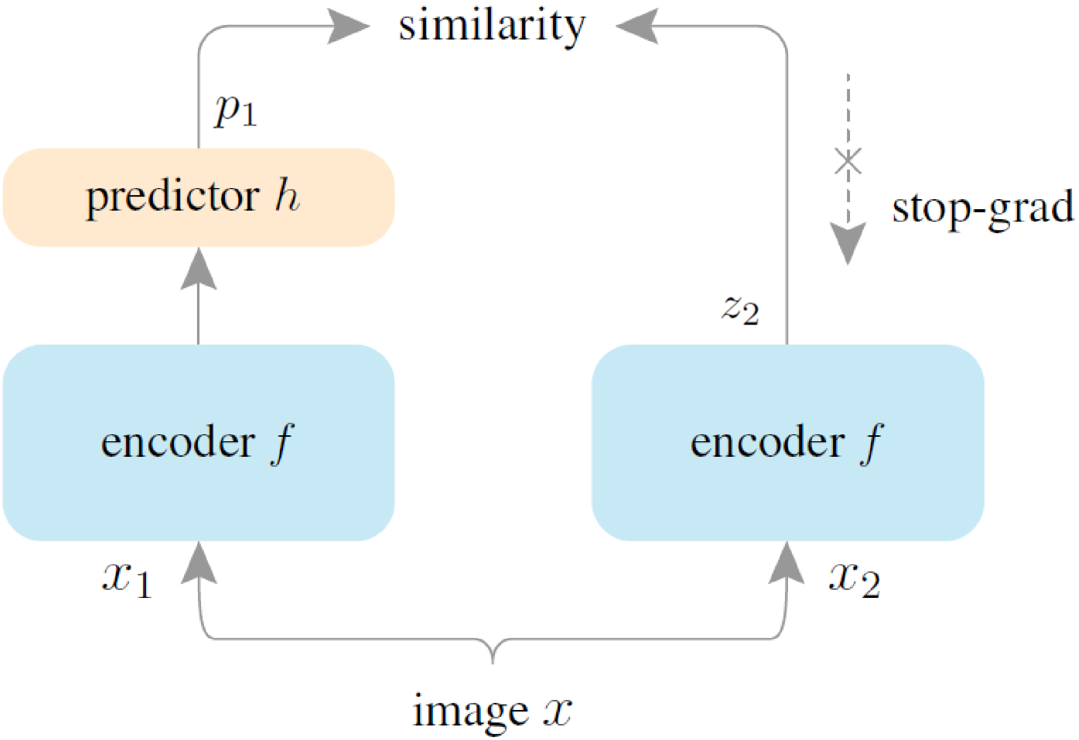
Simple Siamese approach (Chen et al. 2021)

- Put two matching samples (positives) as x_1 and x_2

- Train the encoder with a similarity loss

$$\mathcal{D}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}$$

- Predictor is an additional 2-layer network



- Predictor is flipped alternatingly between x_1 and x_2

$$\mathcal{L} = \frac{1}{2}\mathcal{D}(p_1, z_2) + \frac{1}{2}\mathcal{D}(p_2, z_1)$$

- This strategy (predictor and stop-gradient switching permanently) avoids collapse to a trivial representation

Summary

- Interest points are distinctive points in an image with a significant information content in their neighborhood
- Interest point detection can help establish invariance to certain image transformations.
- Local descriptors describe a local area in the image for the purpose of matching.
- The SIFT descriptor is based on a grid of orientation histograms
- Intermediate layers of convolutional networks yield good descriptors
- Unsupervised learning strategies can learn targeted invariance

References

- D. G. Lowe: Distinctive image features from scale-invariant keypoints, International Journal of Computer Vision 60(2):91-110, 2004.

- J. Matas, O. Chum, M. Urban, T. Pajdla: Robust wide baseline stereo from maximally stable extremal regions, Proc. British Machine Vision Conference, 2002.

- A. Dosovitskiy, P. Fischer, T. Springenberg, M. Riedmiller, T. Brox: Discriminative unsupervised feature learning with convolutional neural networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2016.

- S. Zagoruyko, N. Komodakis: Learning to Compare Image Patches via Convolutional Neural Networks, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

- X. Chen, K. He: Exploring Simple Siamese Representation Learning, *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Programming assignment

- Implement the corner detector based on the second eigenvalue of the structure tensor
  - For computing derivatives and for smoothing images you can make use of the predefined filter masks as well as the convolution operations in `CFilter.h`.
  - The structure tensor of a color image is the sum of tensors over all channels
  - See an online math lecture if you do not remember how to compute the eigenvalues of a matrix http://www.khanacademy.org/
- Apply the corner detector to the images in `ImageProcessing08Ex03.zip` and play with the parameters
- Implement the dense SIFT descriptor (without the detector). Use a 4 pixel spacing and a 3x3 grid of histograms. You can ignore scale and rotation invariance and even skip normalization for this exercise.
- Run your corner detector on `tennis500.ppm` and manually select among the interest points the 10 visually most interesting ones. Extract SIFT descriptors for these points.
- Compute SIFT descriptors for all points in `tennis505.ppm`. For each descriptor in `tennis500.ppm` find the best match in `tennis505.ppm` and visualize the correspondences in your result image.
- Play with the amount of smoothing, the spacing, and the number of histograms per descriptor.