Distinctive Image Features from Scale-Invariant Keypoints

David G. Lowe

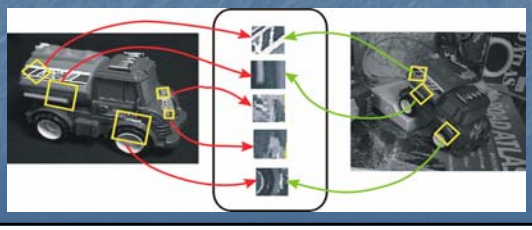
presented by,
Sudheendra

Introduction

- Invariance
  - Intensity
  - Scale
  - Rotation
  - Affine
  - View point

Introduction

- SIFT (Scale Invariant Feature Transformation) approach
  - Local features that are invariant to
    - translation, rotation, scale, and other imaging parameters
  - Features are highly distinctive, each feature finds its correct match in the database with high probability
    - Robust against occlusion and clutter

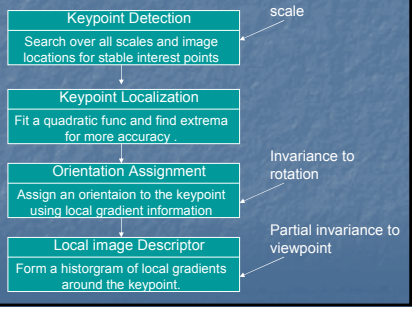


Related Research

- Local interest points
  - Rover visual obstacle avoidance, Moravec 1981 – corner detectors
  - A combined corner and edge detector, Harris and Stephens 1988 – Harris corner detector
- Rotationally invariant points
  - Local greyvalue invariants for image retrieval, Schmid and Mohr 1997
- Scale invariance
  - Scale space theory, Lindeberg 1993 – identifying appropriate scale for features
- Invariance to affine transformation

SIFT method

1. Feature Extraction



SIFT method

2. Object recognition
   - Match key descriptors of test image to the database of features
     - Euclidean distance – ratio of nearest to second nearest neighbor
     - Best-Bin-First approximate algorithm
   - Hough transform
     - Cluster matched features with a consistent interpretation
     - Vote for all object poses consistent with the feature
     - Select clusters with more than 3 features
   - Affine transformation
     - Solve for affine parameters and perform geometric verification of the object's pose in the test image

Keypoint detection

- Goal – detect keypoints that are invariant to scale
- If the scale of the data set is not available then the only way to work with it is to represent it in all possible scales – scale space theory
- Scale-space representation – one parameter family of images where fine-scale image is information is successively suppressed (smoothened)
- Using some constrainsts on scale space functions Lindeberg defines the scale space representation of an image as

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

- And that the maxima and minima of the laplacian of L produce stable image features
- Difference of gaussian approximates laplacian of gaussian which is required for true scale invariance

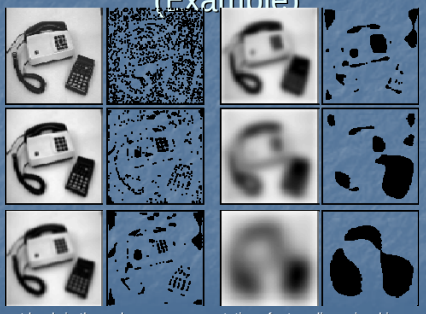
$$\sigma\nabla^2 G = \frac{\partial G}{\partial \sigma} \approx \frac{G(x, y, k\sigma) - G(x, y, \sigma)}{k\sigma - \sigma}$$

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma), \end{aligned}$$

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

- Extrema of the difference of gaussian gives interest points

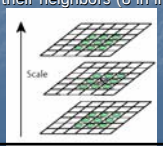
Scale space representation (Example)



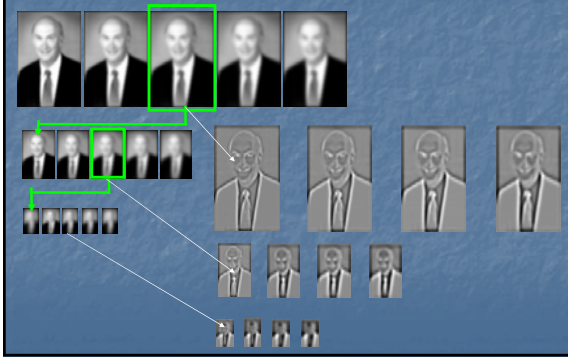
- Different levels in the scale-space representation of a two-dimensional image at scale levels t = 0, 2, 8, 32, 128 and 512 together with grey-level blobs indicating local minima at each scale

Keypoint detection

- Incrementally convolve with gaussians seperated by a constant factor k
  - Octave – doubling of sigma
  - Choose k such there are s images in a particular octave ($k=2^{1/s}$)
- Compute DoG from adjacent scales for entire octave
- Choose every second row and column from the image convolved with gaussian with twice the width and repeat above steps for the next octave
- Extrema detection - select points in each octave that are greater than or less than all their neighbors (8 in image space, 9+9 with different scales)

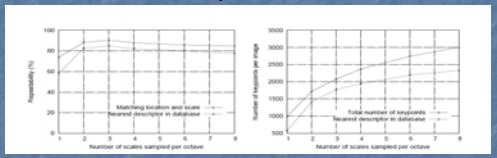


Keypoint Detection (Diagram)



Keypoint detection

- Issues
  1. Frequency of sampling in scale
     - Choosing s (defined previously)
     - Affects the accuracy of the extrema



     - Highest repeatability for 3 scales per octave
  2. Frequency of sampling in the spatial domain
     - Choosing the initial width of the gaussian
     - Similar graph for repeatability against initialwidth provides an optimal value of 1.6

Keypoint Localization

- Local extrema of DoG gives approximate location of keypoints
  - Accuracy – pixel, scaling factor level
- Quadratic function is fit using the 1st and 2nd derivatives at the obtained keypoints
- Extrema of quadratic gives more accurate and stable keypoint location in scale space (x, y, s)
  - Accuracy - sub-pixel and sub-scale level

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

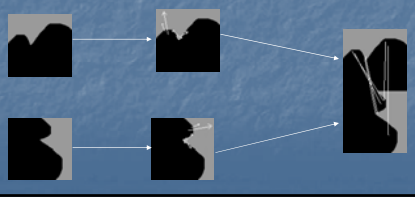
$$\hat{x} = -\frac{\partial^2 D}{\partial x^2}^{-1} \frac{\partial D}{\partial x}$$

Keypoint Elimination

- Function value at extrema used to reject unstable keypoints
  - $D(\bar{x}) = D + \frac{1}{2}\frac{\partial D^T}{\partial x}\bar{x}$
  - discard if |D(x)| < 0.03
- Eliminate edge responses
  - Edges have strong responses even if the location along the edge is poorly determined
  - Principle curvatures
    - Large along the edge, small perpendicular to it
    - Eigenvalues of the Hessian proportional to principle curvatures
    - Ratio of principle curvatures found from the sum and product of eigenvalues
    - All points with a ratio greater than 10 are rejected

Orientation assignment

- Goal – to provide invariance to rotation
- Assign orientation based on local gradients
- Match features with respect to the orientation



Orientation assignment

- To make the feature invariant to rotation
- If L is the Gaussian smoothed image in some scale
- Points in region around keypoint are selected and magnitude and orientations of gradient are calculated

$$m(x,y) = \sqrt{(L(x+1,y)-L(x-1,y))^2 + (L(x,y+1)-L(x,y-1))^2}$$

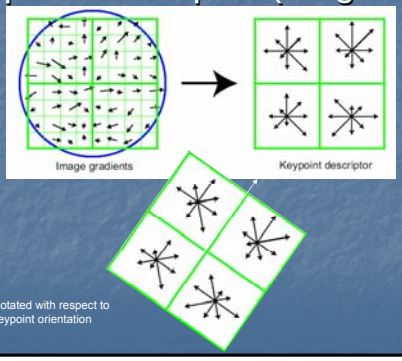
$$\theta(x,y) = \tan^{-1}((L(x,y+1)-L(x,y-1))/(L(x+1,y)-L(x-1,y)))$$

- Orientation histogram formed with 36 bins of points within a region around the keypoint
- Sample is added to appropriate bin and weighted by gradient magnitude and Gaussian-weighted circular window with a of σ 1.5 times scale of keypoint
- Highest peak in the histogram is selected along with any peak that is 80% of the highest peak to form multiple keypoints with different orientations
  - Quadratic fitting performed to get peaks with higher accuracy

Keypoint Descriptor

- Goal – distinctive feature with invariance to viewpoint and intensity
- Motivation – complex cells in the visual cortex
  - Simple cells – respond to a bar of light at the center
  - Complex cells – respond to a bar of light anywhere in the receptive field
- 16x16 region around keypoint divided into 16 4x4 regions
- Histogram of 8 orienations with gaussian weighted magitudes are formed (8 x 4 x 4 = 128 dimensional vector)
  - Gradients rotated relative to keypoint orientation
- Similar to receptive field idea, allows upto four shifts of a point while still being in the same histogram
  - Entries weighted using the distance from central point to avoid boundary effects
- Invariance to intensity – vector is normalized

Keypoint Descriptor (Diagram)



Keypoint selection (Diagram)

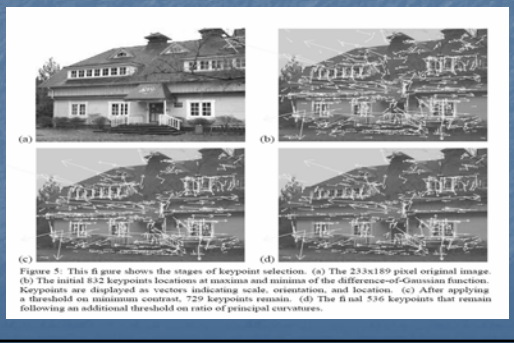
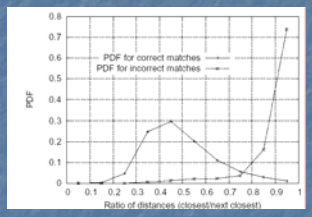


Figure 5: This figure shows the stages of keypoint selection. (a) The 233x189 pixel original image. (b) The initial 832 keypoints locations at maxima and minima of the difference-of-Gaussian function. Keypoints are displayed as vectors indicating scale, orientation, and location. (c) After applying a threshold on minimum contrast, 729 keypoints remain. (d) The final 536 keypoints that remain following an additional threshold on ratio of principal curvatures.

Matching Keypoints

- Feature – location, scale, orientation of keypoint and 128 dimensional keydescriptor
- Independently match all keypoints of test image over all octaves with all keypoints of training image over all octaves
  - Matching function - Euclidean distance
  - Ratio of distance of closest neighbor to second-closest
    - Eliminates matches with background noise and clutter
- Exhaustive search – not feasible
- Best bin first approximate algorithm
  - Similar to KD tree algorithm

Matching Keypoints (Diagram)

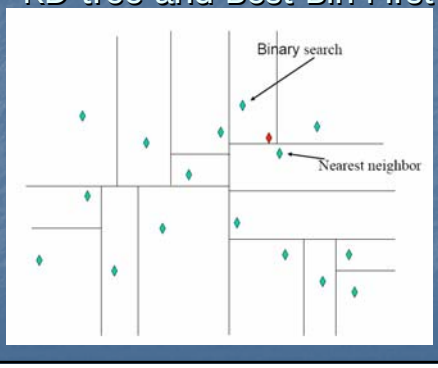


- Threshold at 0.8
  - Eliminates 90% of false matches
  - Eliminates less than 5% of correct matches

KD – tree and Best bin first

- KD tree
  - Data structure for searches involving multi-dimensional keys
  - Split data into two sets based on the median value for a particular dimension
  - Choose another dimension and repeat on both sets
  - Cycle thru the dimensions till all points have been covered
  - Search performed by starting at the root and finding the closest leaf and then backtracking along the parent and pruning impossible branches
- Best-bin-first
  - Instead of backtracking according to tree structure choose branches/bins that are close to the query point
  - Search 200 nearest candidates and stop
  - Provides a speedup of about 2 orders of magnitude

KD tree and Best Bin First



Hough transform

- Keypoint parameters – location, scale and orientation available in test image
- Using this and the keypoint parameters relative to the training image object position in the test image can be found
- Bins created for object parameters in test image
  - 30 degrees for orientation
  - A factor of 2 for scale
  - .25 times the size of the training image (according to scale) for location
- If more than 3 features fall into a bin the bin is subject to geometric verification for affine transformations

Affine transformation

- Accounts for the 3d rotation of a planar surface
- The transformation of a model point [x y] to an image point [u v] is given by

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

where m_i are the affine rotation, stretch and scale parameters and t_x, t_y are the translational parameters

- The equation is solved for the least squares solution of these parameters
- Using this solution the error between each projected image feature and model feature can be calculated

$$e = \sqrt{\frac{2\,\|\mathbf{Ax}-\mathbf{b}\|^2}{r-4}}$$

- The features and parameters are accepted if e < 0.05

Results

Results

Results

Conclusion

- Advantages
  - Invariance to
    - Scale
    - Rotation
    - Intensity
    - Viewpoint (partially)
- Disadvantages
  - Large feature size (128 dimensional floating opint vector)
  - Large number of features (2000 for typical images)