Local Image Features

Computer Vision

James Hays

Read Szeliski 7.1.2 and 7.1.3
“Flashed Face Distortion”
2nd Place in the 8th Annual
Best Illusion of the Year
Contest, VSS 2012
Keep your eyes on the cross
Project 2

The top 100 most confident local feature matches from a baseline implementation of project 2. In this case, 93 were correct (highlighted in green) and 7 were incorrect (highlighted in red).

Project 2: Local Feature Matching
This section: correspondence and alignment

- Correspondence: matching points, patches, edges, or regions across images
SIFT is 20+ years old. Is it still useful?
SIFT is 20+ years old. Is it still useful?

- Let’s look at some trendy research on Neural Radiance Fields (NERF)
SIFT is 20+ years old. Is it still useful?

• Let’s look at some trendy research on Neural Radiance Fields (NERF)
• Let’s look under the hood

Tips for training NeRF models with Instant Neural Graphics Primitives

Our NeRF implementation expects initial camera parameters to be provided in a transforms.json file in a format compatible with the original NeRF codebase. We provide a script as a convenience, scripts/colmap2nerf.py, that can be used to process a video file or sequence of images, using the open source COLMAP structure from motion software to extract the necessary camera data.
SIFT is 20+ years old. Is it still useful?

• COLMAP is the “standard” way to do structure from motion these days

“Structure-From-Motion Revisited”. Johannes L. Schonberger, Jan-Michael Frahm; CVPR 2016
3k+ citations
SIFT is 20+ years old. Is it still useful?

- COLMAP is the “standard” way to do structure from motion these days.
SIFT is 20+ years old. Is it still useful?

• COLMAP is the “standard” way to do structure from motion these days

You can either detect and extract new features from the images or import existing features from text files. COLMAP extracts SIFT [lowe04] features either on the GPU or the CPU. The GPU version requires an attached display, while the CPU version is recommended for use on a server. In general, the GPU version is favorable as it has a customized feature detection mode that often produces higher quality features in the case of high contrast images. If you import existing features, every image must have a text file next to it (e.g., /path/to/image1.jpg and /path/to/image1.jpg.txt) in the following format:

```
NUMFEATURES 128
X Y SCALE ORIENTATION D_1 D_2 D_3 ... D_128
...
X Y SCALE ORIENTATION D_1 D_2 D_3 ... D_128
```

“Structure-From-Motion Revisited”. Johannes L. Schonberger, Jan-Michael Frahm; CVPR 2016
3k+ citations
Overview of Keypoint Matching

1. Find a set of distinctive keypoints

2. Compute a local descriptor from the region around each keypoint

3. Match local descriptors

\[ d(f_A, f_B) < T \]
Review: Harris corner detector

- Define distinctiveness by local auto-correlation.
- Approximate local auto-correlation by second moment matrix.
- Quantify distinctiveness (or cornerness) as function of the eigenvalues of the second moment matrix.
- But we don’t actually need to compute the eigenvalues. Instead, we use the determinant and trace of the second moment matrix.
Review: Harris corner detector

• We want to find *distinctive* patches that don’t look self-similar to neighboring patches
• If there are *gradients* in a patch, those gradients indicate distinctiveness in a particular direction.
• We want to check that we have strong, independent gradients in all directions.
• The eigenvalues of a the collection of gradients in a patch tell us this.
What do the gradients / structure matrix look like?
If you’re not comfortable with Eigenvalues and Eigenvectors, Gilbert Strang’s linear algebra lectures are linked from the course homepage.

Lecture 21: Eigenvalues and eigenvectors

Linear Algebra Lecture 21
Eigenvalues - Eigenvectors
\[ \det(A - \lambda I) = 0 \]
\[ \text{TRACE} = \lambda_1 + \lambda_2 + \ldots + \lambda_n \]
Harris Detector [Harris88]

- **Second moment matrix**

\[
\mu(\sigma_I, \sigma_D) = g(\sigma_I)^* \begin{bmatrix}
I_x^2(\sigma_D) & I_xI_y(\sigma_D) \\
I_xI_y(\sigma_D) & I_y^2(\sigma_D)
\end{bmatrix}
\]

1. Image derivatives (optionally, blur first)

\[
det M = \lambda_1\lambda_2
\]

\[
\text{trace } M = \lambda_1 + \lambda_2
\]

2. Square of derivatives

3. Gaussian filter \(g(\sigma_I)\)

4. Cornerness function – both eigenvalues are strong

\[
har = \det[\mu(\sigma_I, \sigma_D)] - \alpha[\text{trace} (\mu(\sigma_I, \sigma_D))^2] =
\]

\[
g(I_x^2)g(I_y^2) - [g(I_xI_y)]^2 - \alpha[g(I_x^2) + g(I_y^2)]^2
\]

5. Non-maxima suppression
Harris Detector: Steps
Harris Detector: Steps

Compute corner response \( R \)
Harris Detector: Steps

Find points with large corner response: $R > \text{threshold}$
Harris Detector: Steps

Take only the points of local maxima of $R$
Harris Detector: Steps
Invariance and covariance

- We want corner locations to be *invariant* to photometric transformations and *covariant* to geometric transformations
  - **Invariance:** image is transformed and corner locations do not change
  - **Covariance:** if we have two transformed versions of the same image, features should be detected in corresponding locations
Affine intensity change

- Only derivatives are used => invariance to intensity shift $I \rightarrow I + b$
- Intensity scaling: $I \rightarrow aI$

Partially invariant to affine intensity change
Image translation

• Derivatives and window function are shift-invariant

Corner location is covariant w.r.t. translation
Scaling

Corner location is not covariant to scaling!

All points will be classified as edges.
Image rotation

Second moment ellipse rotates but its shape (i.e. eigenvalues) remains the same

Corner location is covariant w.r.t. rotation
So far: can localize in x-y, but not scale
Automatic Scale Selection

How to find corresponding patch sizes?

\[ f(I_{i_1...i_m}(x, \sigma)) = f(I_{i_1...i_m}(x', \sigma')) \]
Automatic Scale Selection

- Function responses for increasing scale (scale signature)
Automatic Scale Selection

- Function responses for increasing scale (scale signature)
Automatic Scale Selection

• Function responses for increasing scale (scale signature)
Automatic Scale Selection

- Function responses for increasing scale (scale signature)
Automatic Scale Selection

- Function responses for increasing scale (scale signature)
Automatic Scale Selection

- Function responses for increasing scale (scale signature)
Orientation Normalization

- Compute orientation histogram
- Select dominant orientation
- Normalize: rotate to fixed orientation

[Low, SIFT, 1999]
Maximally Stable Extremal Regions

• Based on Watershed segmentation algorithm
• Select regions that stay stable over a large parameter range

"Robust Wide Baseline Stereo from Maximally Stable Extremal Regions", Matas, Chum, Urban, and Pajdla, BMVC 2002
6k+ citations
Example Results: MSER
Comparison

Harris

LoG

Hessian

MSER
Local features: main components

1) Detection: Identify the interest points

2) Description: Extract vector feature descriptor surrounding $\mathbf{x}_1 = [x_1^{(1)}, \ldots, x_d^{(1)}]$ for each interest point.

3) Matching: Determine correspondence between descriptors in two views

Kristen Grauman
Image representations

• Templates
  – Intensity, color, gradients, etc.
  – Keeps spatial layout

• Histograms
  – Distribution of intensity, color, texture, SIFT descriptors, etc.
  – Discards spatial layout
Image Representations: Histograms

Histogram: Probability or count of data in each bin

- **Joint histogram**
  - Requires lots of data
  - Loss of resolution to avoid empty bins

- **Marginal histogram**
  - Requires independent features
  - More data/bin than joint histogram

Images from Dave Kauchak
Image Representations: Histograms

Clustering

Use the same cluster centers for all images
Computing histogram distance

$$\text{histint}(h_i, h_j) = 1 - \sum_{m=1}^{K} \min(h_i(m), h_j(m))$$

Histogram intersection (assuming normalized histograms)

$$\chi^2(h_i, h_j) = \frac{1}{2} \sum_{m=1}^{K} \frac{[h_i(m) - h_j(m)]^2}{h_i(m) + h_j(m)}$$

Chi-squared Histogram matching distance

Cars found by color histogram matching using chi-squared
Histograms: Implementation issues

- **Quantization**
  - Grids: fast but applicable only with few dimensions
  - Clustering: slower but can quantize data in higher dimensions

- **Matching**
  - Histogram intersection or Euclidean may be faster
  - Chi-squared often works better
  - Earth mover’s distance is good for when nearby bins represent similar values

**Few Bins**
- Need less data
- Coarser representation

**Many Bins**
- Need more data
- Finer representation
What kind of things do we compute histograms of?

- Color
  - L*a*b* color space
  - HSV color space

- Texture (filter banks or HOG over regions)
What kind of things do we compute histograms of?

- Histograms of oriented gradients

SIFT – Lowe IJCV 2004
SIFT vector formation

- 4x4 array of gradient orientation histogram weighted by magnitude
- 8 orientations x 4x4 array = 128 dimensions
- Motivation: some sensitivity to spatial layout, but not too much.

Showing only 2x2 here, but typical feature would be 4x4
Ensure smoothness

• Gaussian weight
• Interpolation
  – a given gradient contributes to 8 bins:
    4 in space times 2 in orientation
Reduce effect of illumination

- 128-dim vector normalized to 1
- Optionally, threshold gradient magnitudes to avoid excessive influence of high gradients
  - after normalization, clamp gradients >0.2
  - renormalize
Local Descriptors: Shape Context

Count the number of points inside each bin, e.g.:

- Count = 4
- Count = 10

Log-polar binning: more precision for nearby points, more flexibility for farther points.

Belongie & Malik, ICCV 2001
Shape Context Descriptor
Self-similarity Descriptor

Figure 1. These images of the same object (a heart) do NOT share common image properties (colors, textures, edges), but DO share a similar geometric layout of local internal self-similarities.

Matching Local Self-Similarities across Images and Videos, Shechtman and Irani, 2007
Self-similarity Descriptor

Matching Local Self-Similarities across Images and Videos, Shechtman and Irani, 2007
Self-similarity Descriptor

Matching Local Self-Similarities across Images and Videos, Shechtman and Irani, 2007
Learning Local Image Descriptors, Winder and Brown, CVPR 2007

Image Patch → Smooth G(x,σ) → T-Block Filter → S-Block Pooling → N-Block Normalize → Descriptor

- Image Patch: 64x64 Pixels
- Smooth G(x,σ): ~64x64 vectors of dimension k
- T-Block Filter: N histograms of dimension k

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
We obtained a mixed training set consisting of tourist photographs of the Trevi Fountain and of Yosemite Valley (920 images), and a test set consisting of images of Notre Dame (500 images). We extracted interest points and matched them between all of the images within a set using the SIFT detector and descriptor [9]. We culled candidate matches using a symmetry criterion and used RANSAC [5] to estimate initial fundamental matrices between image pairs. This stage was followed by bundle adjustment to reconstruct 3D points and to obtain accurate camera matrices for each source image. A similar technique has been described by [17].
Local Descriptors

• Most features can be thought of as templates, histograms (counts), or combinations
• The ideal descriptor should be
  – Robust
  – Distinctive
  – Compact
  – Efficient
• Most available descriptors focus on edge/gradient information
  – Capture texture information
  – Color rarely used

K. Grauman, B. Leibe
Local features: main components

1) Detection: Identify the interest points

2) Description: Extract vector feature descriptor surrounding \( \mathbf{x}_1 = [x^{(1)}_1, \ldots, x^{(1)}_d] \) each interest point.

3) Matching: Determine correspondence between descriptors in two views

Kristen Grauman
Matching

• Simplest approach: Pick the nearest neighbor. Threshold on absolute distance
• Problem: Lots of self similarity in many photos
Nearest Neighbor Distance Ratio

\[ \frac{NN_1}{NN_2} \]

where NN1 is the distance to the first nearest neighbor and NN2 is the distance to the second nearest neighbor.

• Sorting by this ratio (into ascending order) puts matches in order of confidence (in descending order of confidence).
Matching Local Features

- Nearest neighbor (Euclidean distance)
- Threshold ratio of nearest to 2\textsuperscript{nd} nearest descriptor

\[
\text{PDF for correct matches} \quad \text{PDF for incorrect matches}
\]

Lowe IJCV 2004
6.4 Matching to large databases

An important remaining issue for measuring the distinctiveness of features is how the reliability of matching varies as a function of the number of features in the database being matched. Most of the examples in this paper are generated using a database of 32 images with about 40,000 keypoints. Figure 10 shows how the matching reliability varies as a func-
SIFT Repeatability

Lowe IJCV 2004
SIFT Repeatability

Correct nearest descriptor (%) vs. Width n of descriptor (angle 50 deg, noise 4%)

- With 16 orientations
- With 8 orientations
- With 4 orientations

Lowe IJCV 2004
Choosing a detector

• What do you want it for?
  – Precise localization in x-y: Harris
  – Good localization in scale: Difference of Gaussian
  – Flexible region shape: MSER

• Best choice often application dependent
  – Harris-/Hessian-Laplace/DoG work well for many natural categories
  – MSER works well for buildings and printed things

• Why choose?
  – Get more points with more detectors

• There have been extensive evaluations/comparisons
  – [Mikolajczyk et al., IJCV’05, PAMI’05]
  – All detectors/descriptors shown here work well
Comparison of Keypoint Detectors

Table 7.1. Overview of feature detectors.

<table>
<thead>
<tr>
<th>Feature Detector</th>
<th>Corner</th>
<th>Blob</th>
<th>Region</th>
<th>Rotation invariant</th>
<th>Scale invariant</th>
<th>Affine invariant</th>
<th>Repeatability</th>
<th>Localization accuracy</th>
<th>Robustness</th>
<th>Efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harris</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
<td>++</td>
</tr>
<tr>
<td>Hessian</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td>SUSAN</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>Harris-Laplace</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td>Hessian-Laplace</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
</tr>
<tr>
<td>DoG</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>SURF</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>Harris-Affine</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>Hessian-Affine</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>+++</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
</tr>
<tr>
<td>Salient Regions</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>+</td>
<td>+</td>
<td>++</td>
<td>+</td>
</tr>
<tr>
<td>Edge-based</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>+++</td>
<td>+++</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>MSER</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>+++</td>
<td>+++</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>Intensity-based</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>++</td>
<td>++</td>
<td>++</td>
<td>++</td>
</tr>
<tr>
<td>Superpixels</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
</tbody>
</table>
Choosing a descriptor

• Again, need not stick to one

• For object instance recognition or stitching, SIFT or variant is a good choice

• Learning-based methods are taking over this space, although not as quickly as one might expect.
Things to remember

• Keypoint detection: repeatable and distinctive
  – Corners, blobs, stable regions
  – Harris, DoG

• Descriptors: robust and selective
  – spatial histograms of orientation
  – SIFT