Feature-based methods for image matching

- Bag of Visual Words approach
- Feature descriptors
  - SIFT descriptor
  - SURF descriptor
- Geometric consistency check
- Aggregation of local descriptors into global descriptors
  - Vocabulary trees
  - Fisher vectors
- Image-based retrieval
  - MPEG CDVS standard
  - Mobile visual search
  - Augmented reality
A Bag of Words

self-evident
Liberty
happiness
endowed
inalienable
Creator
pursuit
Life
Representing a Text as a “Bag of Words”

We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed, That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new Government, laying its foundation on such principles and organizing its powers in such form, as to them shall seem most likely to effect their Safety and Happiness. Prudence, indeed, will dictate that Governments long established should not be changed for light and transient causes; and accordingly all experience hath shewn, that mankind are more disposed to suffer, while evils are sufferable, than to right themselves by abolishing the forms to which they are accustomed. But when a long train of abuses and usurpations, pursuing invariably the same Object evinces a design to reduce them under absolute Despotism, it is their right, it is their duty, to throw off such Government, and to provide new Guards for their future security.
Representing an Image as a “Bag of Visual Words”
Feature descriptors

- Represent local pattern around a keypoint by a vector ("feature descriptor")
- Establish feature correspondences by finding the nearest neighbor in descriptor space
Scale/rotation invariant feature descriptors

- **Scale invariance:** extract features at scale provided by keypoint detection
- **Rotation invariance:**
  - Detect dominant orientation by finding peak in orientation histogram
  - Rotate coordinate system to dominant orientation
  - Multiple strong orientation peaks: generate second feature point
SIFT descriptors

- SIFT - Scale-Invariant Feature Transform \cite{Lowe:1999, Lowe:2004}
- Sample thresholded image gradients at 16x16 locations in scale space (in local coordinate system for rotation and scale invariance)
- For each of 4x4 subregion, generate orientation histogram with 8 directions each; each observation weighted with magnitude of image gradient and a window function
- 128-dimensional feature vector
SURF descriptors

- SURF – Speeded Up Robust Features [Bay et al. 2006]
- Compute horizontal and vertical pixel differences, $dx$, $dy$ (in local coordinate system for rotation and scale invariance, window size $20\sigma \times 20\sigma$, where $\sigma^2$ is feature scale)
- Sum $dx$, $dy$, and $|dx|, |dy|$ over 4x4 subregions (SURF-64) or 3x3 subregions (SURF-36)
- Normalize vector for gain invariance, but distinguish bright blobs and dark blobs based on sign of Laplacian (trace of Hessian matrix)
Computing feature descriptors

Orient along dominant gradient

SURF Descriptor

Orientation

Blob Response

Color

Gray

SIFT Descriptor

Computing feature descriptors
“Bag of Visual Words” Matching
Geometric mapping

- Notation:
  - Homogeneous coordinates; reference image \( \mathbf{x} = \begin{pmatrix} x & y & 1 \end{pmatrix}^T \)
  - Inhomogeneous coordinates; target image \( \mathbf{x}' = \begin{pmatrix} x' & y' \\ 1 \end{pmatrix}^T \)

- Translation
  \[
  \mathbf{x}' = \mathbf{x} + \mathbf{t} \quad \text{or} \quad \mathbf{x}' = \begin{bmatrix} 1 & \mathbf{t} \end{bmatrix} \mathbf{x}
  \]

- Euclidean transformation (rotation and translation)
  \[
  \mathbf{x}' = \begin{bmatrix} \cos \theta & -\sin \theta & t_x \\ \sin \theta & \cos \theta & t_y \end{bmatrix} \mathbf{x}
  \]

- Scaled rotation (similarity transform)
  \[
  \mathbf{x}' = \begin{bmatrix} s \cdot \cos \theta & -s \cdot \sin \theta & t_x \\ s \cdot \sin \theta & s \cdot \cos \theta & t_y \end{bmatrix} \mathbf{x}
  \]
Geometric mapping

- Affine transformation

\[ x' = \begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \end{bmatrix} x \]

- Motion of planar surface in 3d under orthographic projection
- Parallel lines are preserved

Argyropelecus olfersi.  
Sternopyx diaphana.
Geometric mapping

- Motion of planar surface in 3d under perspective projection
- Homography

\[
\begin{pmatrix}
  h_{00} & h_{01} & h_{02} \\
  h_{10} & h_{11} & h_{12} \\
  h_{20} & h_{21} & h_{22}
\end{pmatrix}
\begin{pmatrix}
  x \\
  y \\
  1
\end{pmatrix}
\sim
\begin{pmatrix}
  x' \\
  y'
\end{pmatrix}
\]

- Inhomogeneous coordinates (after normalization)

\[
x' = \frac{h_{00}x + h_{01}y + h_{02}}{h_{20}x + h_{21}y + h_{22}} \quad y' = \frac{h_{10}x + h_{11}y + h_{12}}{h_{20}x + h_{21}y + h_{22}}
\]

- Straight lines are preserved
RANSAC

- **RANdom Sample Consensus** [Fischer, Bolles, 1981]
- Randomly select subset of \( k \) correspondences
- Compute geometric mapping parameters by linear regression
- Apply geometric mapping to all keypoints
- Count no. of inliers (closer than \( \varepsilon \) from the corresponding keypoint, typical \( \varepsilon = 1\ldots3 \) pixels)
- Repeat process \( S \) times, keep geometric mapping with largest no. of inliers
- Required number of trials

\[
S = \frac{\log(1 - P)}{\log(1 - q^k)}
\]

- Total probability of success
- Probability of valid correspondence

- Use small number of correspondences

\[
P=0.99, q=0.3, k=3 \rightarrow S=168
\]

\[
P=0.99, q=0.3, k=4 \rightarrow S=566
\]
RANSAC with Affine Model
RANSAC with Homography
SURF features & affine RANSAC

Pairwise Comparison
Local Feature Descriptor Aggregation

- Nearest-neighbor matching of variable-size sets of local features is costly
- Compare images based on a global binary signature of constant size ("hash") instead
- **Simple:** VQ of feature vectors to generate histogram, compare non-empty histogram bins ("bag of features," "bag of visual words")
- **Better:** binarize gradient of log likelihood of w.r.t. to parameter vector ("Fisher vector")
Comparing Feature Histograms

- Speed up by comparing histograms of features: pairwise image comparison only for similar histograms
- Histogram intersection

Equivalent to mean absolute difference, if both histograms contain same number of samples

\[ \rho = \frac{\sum_{i=1}^{n} \min(Q_i, D_i)}{\sum_{i=1}^{n} D_i} \]

[Swain, Ballard 1991]
Growing Vocabulary Tree

[Nistér and Stewenius, 2006]
Growing Vocabulary Tree

[Nistér and Stewenius, 2006]
Growing Vocabulary Tree

[Nistér and Stewenius, 2006]
Growing Vocabulary Tree

[Nistér and Stewenius, 2006]
Growing Vocabulary Tree

\[ k = 3 \]

[Nistér and Stewenius, 2006]
Querying Vocabulary Tree

Query
Hard Binning vs. Soft Binning

**Hard Binning**

- Node 1: $w_{1db} = 0$, $w_1^q = 1$
- Node 2: $w_{2db} = 1$, $w_2^q = 0$
- Node 3: $w_{3db} = 0$, $w_3^q = 0$

**Soft Binning**

- Node 1: $w_{1db} \sim \exp\left(-\frac{(d_{1db})^2}{\sigma^2}\right)$, $w_1^q \sim \exp\left(-\frac{(d_{1q})^2}{\sigma^2}\right)$
- Node 2: $w_{2db} \sim \exp\left(-\frac{(d_{2db})^2}{\sigma^2}\right)$, $w_2^q \sim \exp\left(-\frac{(d_{2q})^2}{\sigma^2}\right)$
- Node 3: $w_{3db} \sim \exp\left(-\frac{(d_{3db})^2}{\sigma^2}\right)$, $w_3^q \sim \exp\left(-\frac{(d_{3q})^2}{\sigma^2}\right)$

**Equations**

- Hard Binning:
  \[ w_{1db} + w_{2db} + w_{3db} = 1 \]
  \[ w_1^q + w_2^q + w_3^q = 1 \]

- Soft Binning:
  \[ w_{1db} + w_{2db} + w_{3db} = 1 \]
  \[ w_1^q + w_2^q + w_3^q = 1 \]
Stanford Mobile Visual Search Dataset

**CDs**

**DVDs**

**Books**

**Landmarks**
Stanford Mobile Visual Search Dataset

Video Clips

Cards

Print

Paintings
Querying: Hard Binning vs. Soft Binning

![Graph showing recall and precision for different binning methods.]

- **SURF features**
- 6-level vocab tree
- 1M leaf nodes
- Affine RANSAC for 100 top tree results
- 25 inliers min.

Precision ~ 97%

**Graph Details:**
- **Axes:**
  - X-axis: Number of Images (million)
  - Y-axis: Recall (percent)
- **Lines:**
  - Red line: Soft Binning (m = 3)
  - Blue line: Hard Binning (m = 1)
Fisher Vector

- Discriminative score function

\[ U(X) = \frac{\partial}{\partial \Theta} \log p_{X|\Theta}(X|\Theta) \]

- Typical, we use Gaussian mixture model (GMM) for \( p_{X|\Theta}(X|\Theta) \)
- Parameters \( \Theta \): mean (and variance) of Gaussian clusters
- For GMM, feature scores \( U(X) \) are soft-assigned distance vectors (and squared distance vectors) relative to cluster centers
- Sums of feature scores of an image are “Fisher vector” that can be used to compare images
- Binarization & Hamming distance comparison results in only minor performance loss (“Binarized Fisher vector”)
MPEG standard “Compact Descriptors for Visual Search” (CDVS)

- **Interest point Detection**
  - LoG peaks
  - Non-orthogonal transform + quantization
  - xy-location needed for object location (and geometric verification)

- **Local Feature Selection**
  - Local Feature Description
  - Local Feature Descriptor Compression
  - Local Feature Location Compression

- **SIFT descriptor**
  - Statistically optimized based on peak response, scale, location, ...

- **Fisher vector**
  - based on GMM

- **Query**
  - 512, 1K, 2K, 4K, 8K, 16K bytes

- **Global Descriptor**
  - 304, 384, 404, 1117, 1117, 1117 bytes
CDVS Evaluation Framework

- Graphics
- Paintings
- Video Frames
- Landmarks
- Common Objects
MPEG CDVS Performance

The diagrams illustrate the performance of MPEG CDVS with respect to image descriptor length. The x-axis represents the image descriptor length in bits (512, 1k, 2k, 4k, 8k, 16k), while the y-axis shows the MAP (Mean Average Precision) and TopMatch values. The performance improves as the descriptor length increases, with different datasets showing varying degrees of improvement. The datasets include Graphics_1a, Paintings_2, VideoFrames_3, Landmarks_4, and CommonObjects_5.
On-Device Image Matching Demo
Database of 100K Images

Samsung Galaxy S3 Smartphone
On-Device Timing Measurements

Samsung Galaxy S3 Smartphone
1.4 GHz Processor
1 GB RAM
Database of 100K Images

400 queries

Global signature database search 54%
Feature extraction 32%
Geometric verification 14%

Augmented Reality Glasses

- Right-eye LCD
- Left-eye LCD
- Camera
- Android controller