How do you like this course so far?
Please give your feedback!

https://docs.google.com/forms/d/e/1FAipQLScsgR-2civzlDB-SC0zhUnof7FkgRsUOB-eA2zOPXXOSeG4cg/viewform
Internship opportunities at HUAWEI and SONY AI Zurich on Computer Vision and Event Cameras

Internship: Computer Vision and Machine Learning
Zurich Research Center - Zürich, Zurich, Switzerland

Description

Huawei's vision is to enrich life through communication. We are a fast growing and leading global information and communications technology solutions provider. With our three business units Carrier, Enterprise and Consumer, we offer network infrastructure, cloud computing solutions and devices such as smartphones and tablet PCs.

Among our customers are 45 of the world's top 50 telecom operators, and one third of the world's population uses Huawei technologies. Huawei is active in more than 170 countries and has over 180,000 employees of which more than 80,000 are engaged in research and development (R&D). With us you have the opportunity to work in a dynamic, multinational environment with more than 150 nationalities worldwide. We seek and reward talent. At Huawei, if you are dedicated to creativity, engagement of technical risks and delivery of target-driven results, your efforts will be rewarded with outstanding career prospects.

Qualifications & Experience

- Background in Computer Science or Engineering (Robotics, Me
- Experience with computer vision and/or robotics
- Experience with programming in Python or C++
- Experience with application development on Linux
- Good communication skills in English (German is a plus)
- Experience with event-based cameras is a plus

Contact

Dario Brescianini
Please email your CV and transcript dario.brescianini@sony.com.

https://apply.workable.com/huawei-16/j/D4A20DBF3F/
Internship opportunities on Event Cameras at NASA/JPL

• **Goal:**
  - explore the feasibility of event cameras for space exploration for next Mars helicopter mission

• **Tasks:**
  - implement VIO pipeline using frames and events and compare with different sensors (thermal)
  - Collect dataset relevant to Mars scenario to test algorithms

• **Start:** December 2020 to March 2021

• **Requirements:**
  - Proved experience with computer vision and field robotics. Strong coding skills in C++, Linux, ROS, and version control (git).
  - Experience with event cameras are a plus but not required.

• **Contact:**
  - Email Prof. Davide Scaramuzza with subject: “JPL internship”

- **Topics**: machine learning, computer vision, control, planning, robot integration

- **Highlights**: many of our Master students have published their thesis in international conferences and won prestigious awards (e.g., ETH Medal, Fritz-Kutter Award, conference paper awards)
Announcement

• Given the recent increases of Covid-19 cases in Switzerland, there is a chance that at some point over the next days or weeks ETH will enforce online teaching only, in which case, we will switch to ZOOM

• If anything happens, you will be informed by email
Vision Algorithms for Mobile Robotics

Lecture 06
Point Feature Detection and Matching – Part 2

Davide Scaramuzza

http://rpg.ifi.uzh.ch
Lab Exercise 4 – This afternoon

Implement SIFT blob detection and matching
Outline

• Automatic Scale Selection
• The SIFT blob detector and descriptor
• Other corner and blob detectors and descriptors
Scale changes

How can we match image patches corresponding to the same feature but belonging to images taken at different scales?
Scale changes

How can we match image patches corresponding to the same feature but belonging to images taken at different scales? Possible solution: rescale the patch.
Scale changes

How can we match image patches corresponding to the same feature but belonging to images taken at different scales? Possible solution: rescale the patch.
Scale changes

How can we match image patches corresponding to the same feature but belonging to images taken at different scales? Possible solution: rescale the patch.
Scale changes

• Scale search is time consuming (needs to be done individually for all patches in one image)

• **Complexity** is $(NS)^2$ assuming $N$ features per image and $S$ rescalings per feature

• **Solution**: **automatic scale selection**: automatically assign each feature its own “scale” (i.e., size)
Automatic Scale Selection

- Idea: Design a function on the image patch, which is **scale invariant** (i.e., it has the same value for corresponding patches, even if they are at different scales)
Automatic Scale Selection

• Idea: Design a function on the image patch, which is **scale invariant** (i.e., it has the same value for corresponding patches, even if they are at different scales)
  • Find **local extrema** of this function
  • The **patch size** at which the local extremum is reached should be **invariant** to image rescaling
  • **Important**: this scale invariant patch size is found in each image **independently**
Automatic Scale Selection: Example

Image 1

Image 2

\[ f(I(x, y, \sigma)) \]

\[ f(I'(x', y', \sigma')) \]

\[ \sigma \]

\[ \sigma' \]
Automatic Scale Selection: Example

\[ f(I(x, y, \sigma)) \]

\[ f(I'(x', y', \sigma')) \]
Automatic Scale Selection: Example

Image 1

Image 2
Automatic Scale Selection: Example

Image 1

Image 2

\[ f(I(x, y, \sigma)) \]

\[ f(I(x', y', \sigma')) \]
Automatic Scale Selection: Example

Image 1

Image 2

\[ f(I(x, y, \sigma)) \]

\[ f(I(x', y', \sigma')) \]
Automatic Scale Selection: Example

Image 1

\[ f(I(x, y, \sigma)) \]

\[ S_1 \]

Image 2

\[ f(I(x', y', \sigma')) \]

\[ S_2 \]
When the right scale is found, the patches must be normalized to a canonical size so that they can be compared by SSD.
Automatic Scale Selection: Example
• A “good” function for scale detection should have a single & sharp peak

• What if there are multiple peaks? Is it really a problem?

• What is a good function?
  • **Sharp intensity changes** are good regions to monitor in order to identify the scale
Automatic Scale Selection

- The **ideal function** for determining the scale is one that highlights sharp discontinuities.
- **Solution**: convolve image with a **kernel that highlights edges**

\[ f = \text{Kernel} \ast \text{Image} \]

- It has been shown that the **Laplacian of Gaussian kernel** is optimal under certain assumptions [Lindeberg’94]:

\[
\text{LoG}(x, y, \sigma) = \nabla^2 G_\sigma(x, y) = \frac{\partial^2 G_\sigma(x, y)}{\partial x^2} + \frac{\partial^2 G_\sigma(x, y)}{\partial y^2}
\]

Automatic Scale Selection

The **correct scale(s)** is (are) found as **local extrema across consecutive smoothed patches**
Main questions

• What features are repeatable and distinctive?
• How to describe a feature?
• How to establish correspondences, i.e., compute matches?
Feature descriptors

We know how to detect points, but **how to describe** them for matching?
Patch and Census Descriptors

• **Patch descriptor** (i.e., patch of intensity values, integer values)

• **Census descriptor** (binary values)
HOG Descriptor (Histogram of Oriented Gradients)

- The image is divided into a **grid of cells** and for each cell a **histogram of gradient directions** is compiled.
- The HOG descriptor is the **concatenation of these histograms** (used in SIFT).
- Differently from the patch and Census descriptors, HOG has **float values**.

**Example of gradient histogram with 8 orientation bins.**
Each vote is weighted by the gradient magnitude.
Feature Descriptor Invariance

• The ideal feature descriptor should be **invariant** to:
  • **geometric changes:** rotation, scale, view point
  • **photometric changes:** illumination

• **Most feature methods** are designed to be invariant to
  • 2D translation,
  • 2D rotation,
  • Scale
• Some of them can also handle
  • **View-point changes** (e.g., SIFT & LIFT work with up to 50 degrees of viewpoint changes)
  • **Affine illumination changes**
How to achieve invariance to Rotation and Scale?

De-rotation:
- **Determine patch orientation**
  - e.g., eigenvectors of M matrix of Harris or dominant gradient direction (see next slide)
- **De-rotate patch through “patch warping”**
  - This puts the patches into a **canonical orientation**

Re-scaling:
- **Detect scale** using LoG operator
- **Rescale the patch** to a canonical size (e.g., 8×8 pixels)

Example of patch after de-rotation and re-scaling
How to determine the patch orientation?

1. First, **multiply the patch by a Gaussian kernel** to make the shape circular rather than square.
2. Then, **compute gradients vectors** at each pixel.
3. Build a **histogram of gradient orientations**, weighted by the gradient magnitudes. The histogram represents the HOG descriptor.
4. Extract **all local maxima in the histogram**: each local maximum above a threshold is a candidate dominant orientation.
5. Construct a **different keypoint descriptor for each dominant orientation**.

Dominant gradient direction
How to achieve invariance to small view point changes?

**Affine warping** provides invariance to small view-point changes

- The second moment matrix $M$ of Harris detector can be used to identify the two directions of fastest and slowest change of SSD around the feature
- Out of these two directions, an elliptic patch is extracted at the scale computed by with the LoG operator
- The region inside the ellipse is normalized to a canonical circular patch
Recap: De-rotation, Re-scaling, and Affine Un-warping
How to warp a patch?

• Start with an “**empty**” canonical patch (all pixels set to 0)
• For each pixel \((x, y)\) in the empty patch, apply the **warping function** \(W(x, y)\) to compute the corresponding position in the source image. It will be in floating point and will fall between the image pixels.

• **Interpolate** the intensity values of the 4 closest pixels in the detected image using either of:
  • Nearest neighbor interpolation
  • Bilinear interpolation
  • Bicubic interpolation
Example: Similarity Transformation (rotation, translation & rescaling)

- Warping function $W$ (counterclockwise rotation plus rescaling and translation):
  
  $$x' = s(x \cos \theta - y \sin \theta) + a$$
  $$y' = s(x \sin \theta + y \cos \theta) + b$$

Patch detected in the image

Empty canonical patch
Bilinear Interpolation

• It is an extension of linear interpolation for interpolating functions of two variables (e.g., x and y) on a rectilinear 2D grid.

• The key idea is to perform linear interpolation first in one direction, and then again in the other direction.

• Although each step is linear in the sampled values and in the position, the interpolation as a whole is not linear but rather quadratic in the sample location.

\[
I(x, y) = I(0, 0)(1 - x)(1 - y) + I(0, 1)(1 - x)y + I(1, 0)x(1 - y) + I(1, 1)x y
\]

In this geometric visualization, the value at the black spot is the sum of the value at each colored spot multiplied by the area of the rectangle of the same color.

This formula won’t be asked at the exam 😊
Nearest Neighbor vs Bilinear vs Bicubic Interpolation

1D nearest-neighbour

Linear

Cubic

2D nearest-neighbour

Bilinear

Bicubic
Example: Rescaling

Original image: x 10

Nearest-neighbor interpolation  Bilinear interpolation  Bicubic interpolation
Disadvantage of Patch Descriptors

• **Disadvantage of patch descriptors:**
  • If the warp is not estimated accurately, very **small errors** in rotation, scale, and viewpoint will **affect matching score significantly**
  • **Computationally expensive** (need to unwarp every patch)
Outline

• Automatic Scale Selection

• The SIFT blob detector and descriptor

• Other corner and blob detectors and descriptors
SIFT Descriptor

- **Scale Invariant Feature Transform**
- Invented by David Lowe in 2004 (now at Google)

SIFT Descriptor

Descriptor computation:

- Multiply the patch by a Gaussian filter, compute dominant orientation, and de-rotate patch
- Consider a $16 \times 16$ pixel patch
- Compute HOG descriptor
  - Divide patch into $4 \times 4$ cells
  - Use 8 bin histograms (i.e., 8 directions)
  - Concatenate all histograms into a single 1D vector
  - Resulting SIFT descriptor: $4 \times 4 \times 8 = 128$ values
- Descriptor Matching: SSD (i.e., Euclidean-distance)
- Why $4 \times 4$ cells and why 8 bins? See later
Descriptor Normalization

• The descriptor vector $\nu$ is then normalized such that its $l_2$ norm is 1:

$$\bar{\nu} = \frac{\nu}{\sqrt{\sum_i^n \nu_i^2}}$$

• This guarantees that the descriptor is invariant to linear illumination changes

• The descriptor is already invariant to additive illumination because it is based on gradients

• We can conclude that the SIFT descriptor is invariant to affine illumination changes
SIFT Matching Robustness

• Can handle severe **viewpoint changes** (up to 50 degree out-of-plane rotation)
• Can handle even **severe non affine changes in illumination** (low to bright scenes)
• Computationally **expensive**: 10 frames per second (fps) on an i7 processor
• Original SIFT binary files: [http://people.cs.ubc.ca/~lowe/keypoints](http://people.cs.ubc.ca/~lowe/keypoints)
• OpenCV C/C++ implementation: [https://docs.opencv.org/master/da/df5/tutorial_py_sift_intro.html](https://docs.opencv.org/master/da/df5/tutorial_py_sift_intro.html)
SIFT Detector

• SIFT uses the **Difference of Gaussian (DoG) kernel** instead of Laplacian of Gaussian (LoG) because computationally cheaper

\[
\text{LOG}(x, y) \approx \text{DoG}(x, y) = G_{k\sigma}(x, y) - G_{\sigma}(x, y)
\]

• The proof that LoG can be approximated by a difference of Gaussian comes from the Heat Equation:

\[
\frac{\partial G_\sigma}{\partial \sigma} = \sigma \nabla G_\sigma
\]
SIFT Detector (location + scale)

SIFT keypoints: **local extrema in both space and scale** of the DoG images

- **Each pixel is compared to 26 neighbors** (in green below): its 8 neighbors in the current image + 9 neighbors in the adjacent upper scale + 9 neighbors in the adjacent lower scale

- If the pixel is a global maximum or minimum (i.e., extrema) with respect to its 26 neighbors then it is selected as SIFT feature

For each extrema, the output of the SIFT detector is the location \((x, y)\) and the scale \(s\)
Example
Magnitude of $(G(k\sigma) - G(\sigma)) \mid s = 4; \sigma = 1.6$
DoG Images example

Magnitude of \((G(k^2\sigma) - G(k\sigma))\) | \(s = 4; \sigma = 1.6\) |
DoG Images example

Magnitude of \( (G(k^3 \sigma) - G(k^2 \sigma)) \mid s = 4; \sigma = 1.6 \mid
DoG Images example

Magnitude of \( (G(k^4 \sigma) - G(k^3 \sigma)) \mid s = 4; \sigma = 1.6 \mid
DoG Images example

Magnitude of \((G(k^5\sigma) - G(k^4\sigma))\) \(s = 4; \sigma = 1.6\)
(second octave shown at the input resolution for convenience)
DoG Images example

Magnitude of \((G(k^6\sigma) - G(k^5\sigma))\) | \(s = 4; \sigma = 1.6\) |
(second octave shown at the input resolution for convenience)
DoG Images example

Magnitude of $(G(k^7\sigma) - G(k^6\sigma))$ | $s = 4; \sigma = 1.6$ |
(second octave shown at the input resolution for convenience)
DoG Images example

Magnitude of \((G(k^8 \sigma) - G(k^7 \sigma))\) \(s = 4; \sigma = 1.6\)
(second octave shown at the input resolution for convenience)
DoG Images example

Magnitude of \((G(k^8\sigma) - G(k^8\sigma))\) \(s = 4; \sigma = 1.6\)
(third octave shown at the input resolution for convenience)
Local extrema of DoG images across Scale and Space

What are SIFT features like?
Hint: Remember the definition of filtering as template matching
How it is implemented in practice

1. Build a Space-Scale Pyramid:
   - The initial image is incrementally convolved with Gaussians $G(k^l \sigma)$ to produce blurred images separated by a constant factor $k$ in scale space (shown stacked in the left column).
     - The initial Gaussian $G(\sigma)$ has $\sigma=1.6$
     - $k$ is chosen: $k = 2^{1/s}$, where $s$ is the number of intervals into which each octave of scale space is divided
     - For efficiency reasons, when $k^l$ equals 2, the image is downsampled by a factor of 2 and then the procedure is repeated again up to 5 octaves (pyramid levels)
   - Adjacent blurred images are then subtracted to produce the Difference-of-Gaussian (DoG) images

2. Scale-Space extrema detection
   - Detect local maxima and minima in space-scales (see previous slide)
Scale (Gaussian blurring: $G(k\sigma)$)

Octaves

DoG images
SIFT: Recap

• SIFT: Scale Invariant Feature Transform

• An approach to detect and describe regions of interest in an image.
  • SIFT detector = DoG detector

• SIFT features are invariant to 2D rotation, and reasonably invariant to rescaling, viewpoint changes (up to 50 degrees), and illumination

• It runs in real-time but expensive (10 Hz on an i7 laptop)
  • The expensive steps are the scale detection and descriptor extraction
Original SIFT Demo by David Lowe

Download original SIFT binaries and Matlab function from:
http://people.cs.ubc.ca/~lowe/keypoints

>>[image1, descriptor1s, locs1] = sift('scene.pgm');
>>showkeys(image1, locs1);
>>[image2, descriptors2, locs2] = sift('book.pgm');
>>showkeys(image2, locs2);
>>match('scene.pgm','book.pgm');
SIFT Repeatability with Viewpoint Changes

Repeatability =

\[
\frac{\text{# correspondences detected}}{\text{# correspondences present}}
\]

![Diagram showing repeatability with viewpoint changes](image)
SIFT Repeatability with Number of Scales per Octave

Repeatability = \frac{\text{# correspondences detected}}{\text{# correspondences present}}
Influence of Number of Orientations and Number of Sub-patches

The graph shows that a single orientation histogram ($n = 1$) is very poor at discriminating. The results improve with a 4x4 array of histograms with 8 orientations.

Figure 8: This graph shows the percent of keypoints giving the correct match to a database of 40,000 keypoints as a function of width of the $n \times n$ keypoint descriptor and the number of orientations in each histogram. The graph is computed for images with affine viewpoint change of 50 degrees and addition of 4% noise.
What’s the output of SIFT?

- **Descriptor**: 4x4x8 = 128-element 1D vector
- **Location** (pixel coordinates of the center of the patch): 2D vector
- **Scale** (i.e., size) of the patch: 1 scalar value (high scale corresponds to high blur in the space-scale pyramid)
- **Orientation** (i.e., angle of the patch): 1 scalar value
Application of SIFT to Object recognition

• Can be implemented easily by returning object with the largest number of correspondences with the template

• For planar objects, 4 point RANSAC can be used to remove outliers (see Lecture 8).

• For rigid 3D objects, 5 point RANSAC (see Lecture 08).
Application of SIFT to Panorama Stitching

AutoStitch: http://matthewalunbrown.com/autostitch/autostitch.html
Main questions

• What features are repeatable and distinctive?
• How to describe a feature?
• How to establish correspondences, i.e., compute matches?
Feature Matching
Feature Matching

- Given a feature in $I_1$, how to find the best match in $I_2$?
  1. **Define distance function** that compares two descriptors ((Z)SSD, (Z)SAD, (Z)NCC or Hamming distance for binary descriptors (e.g., Census, ORB, BRIEF, BRISK, FREAK))
  2. **Brute-force matching:**
     1. Compare each feature in $I_1$ against all the features in $I_2$ ($N^2$ comparisons, where $N$ is the number of features in each image)
     2. Take the one at minimum distance, i.e. the **closest descriptor**
Feature Matching

• **Issues with closest descriptor**: can occasionally return good scores for false matches

• **Better approach**: compute ratio of distances to 1\textsuperscript{st} to 2\textsuperscript{nd} closest descriptor

\[
\frac{d_1}{d_2} < \text{Threshold (usually 0.8)}
\]

where:

- \(d_1\) is the distance from the closest descriptor
- \(d_2\) is the distance of the 2\textsuperscript{nd} closest descriptor
Distance Ratio: Explanation

• In SIFT, the nearest neighbor is defined as the keypoint with minimum Euclidean distance. However, many features from an Image 1 may not have any correct match in Image 2 because they arise from background clutter or were not detected in the Image 1.

• An effective measure is obtained by comparing the distance of the closest neighbor to that of the second-closest neighbor. This measure performs well because correct matches need to have the closest neighbor significantly closer than the closest incorrect match to achieve reliable matching.

• For false matches, there will likely be a number of other false matches within similar distances due to the high dimensionality of the feature space (this problem is known as curse of dimensionality). We can think of the second-closest match as providing an estimate of the density of false matches within this portion of the feature space and at the same time identifying specific instances of feature ambiguity.
The SIFT paper recommends to use a threshold on 0.8. Where does this come from?

“A threshold of 0.8 eliminates 90% of the false matches while discarding less than 5% of the correct matches.”

“This figure was generated by matching images following random scale and orientation change, a depth rotation of 30 degrees, and addition of 2% image noise, against a database of 40,000 keypoints.”

Figure 11: The probability that a match is correct can be determined by taking the ratio of distance from the closest neighbor to the distance of the second closest. Using a database of 40,000 keypoints, the solid line shows the PDF of this ratio for correct matches, while the dotted line is for matches that were incorrect.
Outline

• Automatic Scale Selection
• The SIFT blob detector and descriptor
• Other corner and blob detectors and descriptors
“FAST” Corner Detector

- **FAST**: Features from Accelerated Segment Test
- Analyses intensities along a ring of 16 pixels centered on the pixel of interest \( p \)
- \( p \) is a FAST corner if a set of \( N \) contiguous pixels on the ring are:
  - all brighter than the pixel intensity \( I(p) + \text{threshold} \),
  - or all darker than \( I(p) - \text{threshold} \)
- Common value of \( N: 12 \)
- A simple classifier is used to check the quality of corners and reject the weak ones

- **FAST is the fastest corner detector ever made**: can process 100 million pixels per second (<3ms per image)
- **Issue**: it is very sensitive to image noise (high in low light). This is why Harris is still more common despite a bit slower

Rosten, Drummond, Fusing points and lines for high performance tracking, International Conference on Computer Vision (ICCV), 2005. [PDF]

“SURF” Blob Detector & Descriptor

- **SURF**: Speeded *Up* Robust *Features*
- Similar to **SIFT** but much faster
- **Basic idea**: approximate Gaussian and DoG filters using box filters
- Results comparable with SIFT, plus:
  - Faster computation
  - Generally shorter descriptors

Bay, Tuytelaars, Van Gool, "Speeded Up Robust Features", European Conference on Computer Vision (ECCV) 2006. [PDF]
“BRIEF” Descriptor (can be applied to corners or blobs)

• **BRIEF**: Binary Robust Independent Elementary Features

• **Goal**: high speed (in description and matching)

• **Binary descriptor formation**:
  • Smooth image
  • for each detected keypoint (e.g. FAST),
  • sample 128 intensity pairs \((p_{1i}, p_{2i})\) \((i = 1, ..., 128)\)
    within a squared patch around the keypoint
  • Create an empty 128-element descriptor
  • for each \(i^{th}\) pair
    • if \(I_{p_{1i}} < I_{p_{2i}}\) then set \(i^{th}\) bit of descriptor to 1
    • else to 0

• The pattern is generated randomly (or learned) only once; then, the same pattern is used for all patches

• Pros: **Binary descriptor**: allows very fast Hamming distance matching (count of the number of bits that are different in the descriptors matched)

• Cons: **Not scale/rotation invariant**

Pattern for intensity pair samples – generated randomly
“ORB” Descriptor (can be applied to corners or blobs)

- **ORB**: Oriented FAST and Rotated BRIEF
- Keypoint detector based on FAST
- Binary descriptor based on BRIEF but adds an orientation component to make it rotation invariant

Rublee, Rabaud, Konolige, Bradski, “ORB: an efficient alternative to SIFT or SURF” (PDF). IEEE International Conference on Computer Vision (ICCV), 2011. [PDF](#).
“BRISK” Descriptor (can be applied to corners or blobs)

- **BRISK**: Binary Robust Invariant Scalable Keypoints
- Keypoint detector based on **FAST**
- Both rotation and scale invariant
- Binary descriptor, formed by pairwise intensity comparisons (like BRIEF) but on a radially symmetric sampling pattern
  - Red circles: size of the smoothing kernel applied
  - Blue circles: smoothed pixel value used
- Detection and descriptor speed: **10 times faster than SURF**
- Slower than BRIEF, but scale- and rotation- invariant
“FREAK” Descriptor (can be applied to corners or blobs)

- **FREAK**: Fast Retina Keypoint
- **Rotation and scale invariant**
- Binary descriptor
- **Sampling pattern** similar to BRISK but uses a more pronounced “retinal” (i.e., log-polar) sampling pattern inspired by the human retina: higher density of points near the center
- **Pairwise** intensity comparisons form **binary** strings similar to BRIEF
- Pairs are **learned** (as in ORB)
- Circles indicate size of smoothing kernel
- **Coarse-to-fine matching** (cascaded approach): first compare the first half of bits; if distance smaller than threshold, proceed to compare the next bits, etc.
- Faster to compute, less memory and **more robust than SIFT, SURF or BRISK**

Alexandre Alahi, Raphael Ortiz, Pierre Vandergheynst, FREAK: Fast Retina Keypoint, CVPR 2012. [PDF](#).
“SuperPoint”: Self-Supervised Interest Point Detection and Description

• Joint regression of keypoint **location and descriptor**. **Self-supervised**.
• Trained on synthetic images and refined on homographies of real images
• **Detector less accurate than SIFT**, but **descriptor outperforms SIFT**
• But slower than SIFT

Detone, Malisiewicz, Rabinovich. SuperPoint: Self-Supervised Interest Point Detection and Description. CVPRW 2018. [PDF](#).
“LIFT” Descriptor (can be applied to corners or blobs)

- **LIFT**: Learned Invariant Feature Transform
- **Learning-based descriptor**
- **Rotation, scale, viewpoint and illumination invariant**
- First a network predicts the **patch orientation** which is used to derotate the patch.
- Then another neural network is used to generate a **patch descriptor** (128 dimensional) from the derotated patch.
- **Illumination invariance** is achieved by randomizing illuminations during training.
- **LIFT descriptor beats SIFT in repeatability**

Kwang Moo Yi, Eduard Trulls, Vincent Lepetit, Pascal Fua, LIFT: Learned Invariant Feature Transform, European Conference on Computer Vision (ECCV) 2016. [PDF]
LIFT vs SIFT

Matching features on ‘DTU’, sequence #19. Correct matches shown with green lines.

SIFT. Average: 34.1 matches

LIFT (Ours). Average: 98.5 matches

https://youtu.be/hhxAttChmCo
## Recap Table

<table>
<thead>
<tr>
<th>Detector</th>
<th>Localization Accuracy of the detector</th>
<th>Descriptor that can be used</th>
<th>Efficiency</th>
<th>Relocalization &amp; Loop closing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Harris</td>
<td>++++</td>
<td>Patch, SIFT/LIFT, BRIEF, ORB, BRISK, FREAK</td>
<td>++++</td>
<td>++++++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SIFT</td>
<td>++</td>
<td>++++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BRIEF</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORB</td>
<td>++++</td>
<td>+++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BRISK</td>
<td>++++</td>
<td>+++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FREAK</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td>Shi-Tomasi</td>
<td>++++</td>
<td>Patch, SIFT, BRIEF, ORB, BRISK, FREAK</td>
<td>++</td>
<td>+++++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SIFT</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BRIEF</td>
<td>++++</td>
<td>+++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORB</td>
<td>++++</td>
<td>+++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BRISK</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FREAK</td>
<td>++++</td>
<td>+++</td>
</tr>
<tr>
<td>FAST</td>
<td>++++</td>
<td>Patch, SIFT/LIFT, BRIEF, ORB, BRISK, FREAK</td>
<td>++++</td>
<td>++++++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>SIFT</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BRIEF</td>
<td>++++</td>
<td>+++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>ORB</td>
<td>++++</td>
<td>+++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BRISK</td>
<td>+++</td>
<td>+++</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FREAK</td>
<td>++++</td>
<td>+++</td>
</tr>
<tr>
<td>SIFT</td>
<td>+++</td>
<td>SIFT</td>
<td>+</td>
<td>+++</td>
</tr>
<tr>
<td>SURF</td>
<td>+++</td>
<td>SURF</td>
<td>++</td>
<td>+++</td>
</tr>
<tr>
<td>SuperPoint</td>
<td>++</td>
<td>SuperPoint</td>
<td>+</td>
<td>+++</td>
</tr>
</tbody>
</table>

86
Summary (things to remember)

- Similarity metrics: NCC (ZNCC), SSD (ZSSD), SAD (ZSAD), Census Transform
- Point feature detection
  - Properties and invariance to transformations
    - Challenges: rotation, scale, view-point, and illumination changes
- Extraction
  - Moravec
  - Harris and Shi-Tomasi
    - Rotation invariance
- Automatic Scale selection
- Descriptor
  - Intensity patches
    - Canonical representation: how to make them invariant to transformations: rotation, scale, illumination, and view-point (affine)
    - Better solution: Histogram of oriented gradients: SIFT descriptor
- Matching
  - (Z)SSD, SAD, NCC, Hamming distance (last one only for binary descriptors)
    - ratio 1st /2nd closest descriptor
- Depending on the task, you may want to trade off repeatability and robustness for speed: approximated solutions, combinations of efficient detectors and descriptors.
  - Fast corner detector: FAST;
  - Keypoint descriptors faster than SIFT: SURF, BRIEF, ORB, BRISK
Reading

- Ch. 4.1 and Ch. 8.1 of Szeliski book, 1st Edition
- Chapter 4 of Autonomous Mobile Robots book: link
- Ch. 13.3 of Peter Corke book
Understanding Check

Are you able to answer:

• How does automatic scale selection work?
• What are the good and the bad properties that a function for automatic scale selection should have or not have?
• How can we implement scale invariant detection efficiently? (show that we can do this by resampling the image vs rescaling the kernel).
• What is a feature descriptor? (patch of intensity value vs histogram of oriented gradients). How do we match descriptors?
• How is the keypoint detection done in SIFT and how does this differ from Harris?
• How does SIFT achieve orientation invariance?
• How is the SIFT descriptor built?
• What is the repeatability of the SIFT detector after a rescaling of 2? And for a 50 degrees viewpoint change?
• Illustrate the 1st to 2nd closest ratio of SIFT detection: what’s the intuitive reasoning behind it? Where does the 0.8 factor come from?
• How does the FAST detector work? What are its pros and cons compared with Harris?