Lecture 6. Feature Descriptors

SIFT

Juan Carlos Niebles and Jiajun Wu

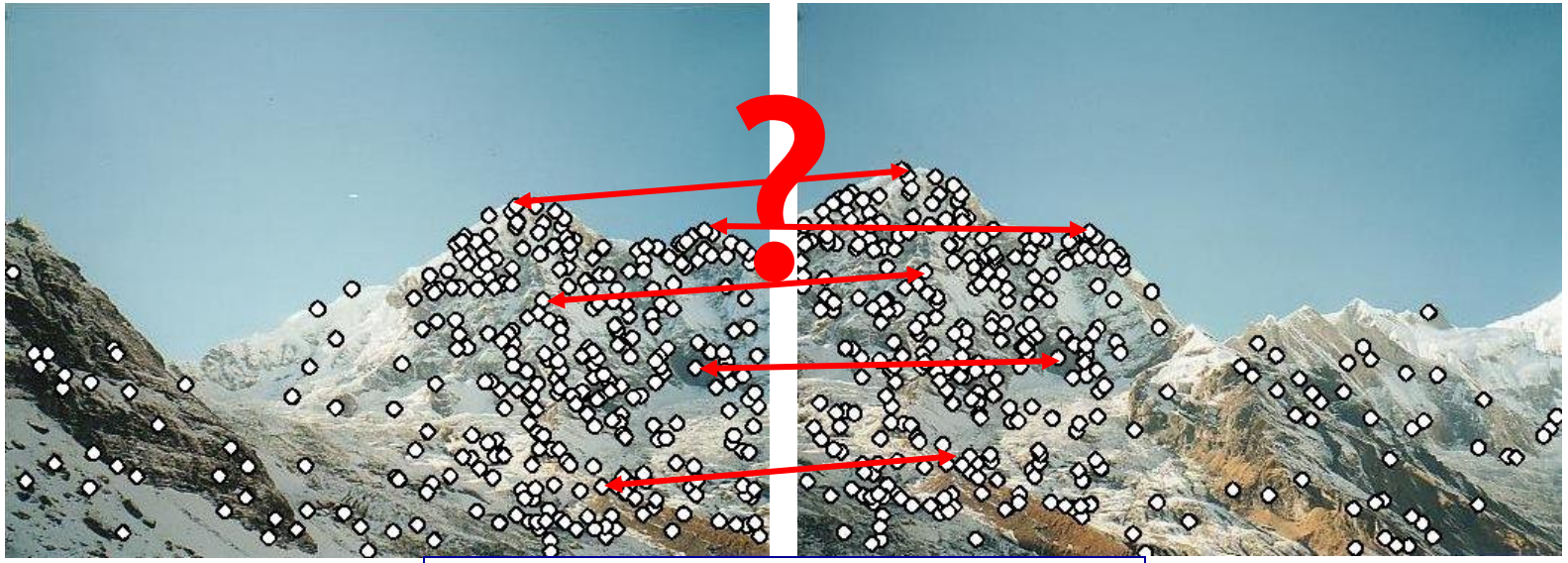
CS131 Computer Vision: Foundations and Applications

What will we learn today?

- SIFT: an image region descriptor

Local Features and Descriptors

- We know how to detect points
- Next question: How to describe them for matching?
- Descriptor: Vector that summarizes the content of the keypoint neighborhood.



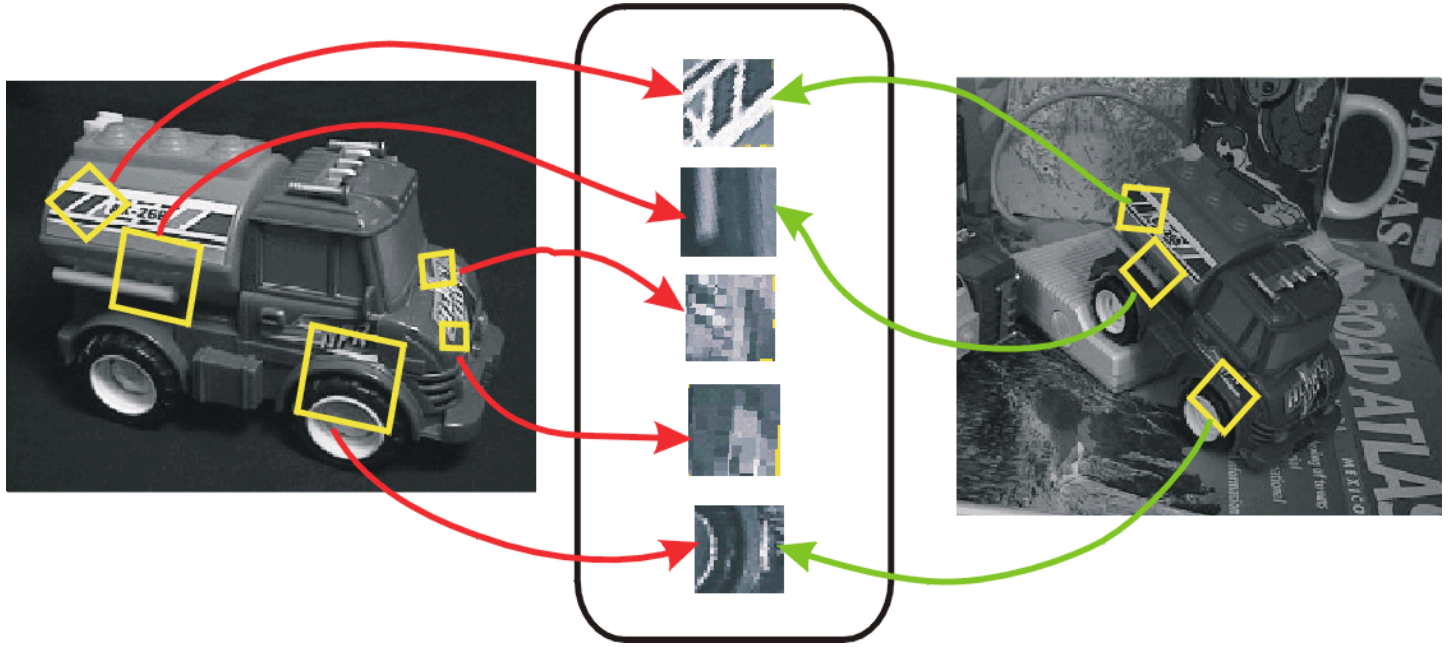
Point descriptor should be:

1. **Invariant**
2. **Distinctive**

Slide credit: Kristen Grauman

Invariant Local Features

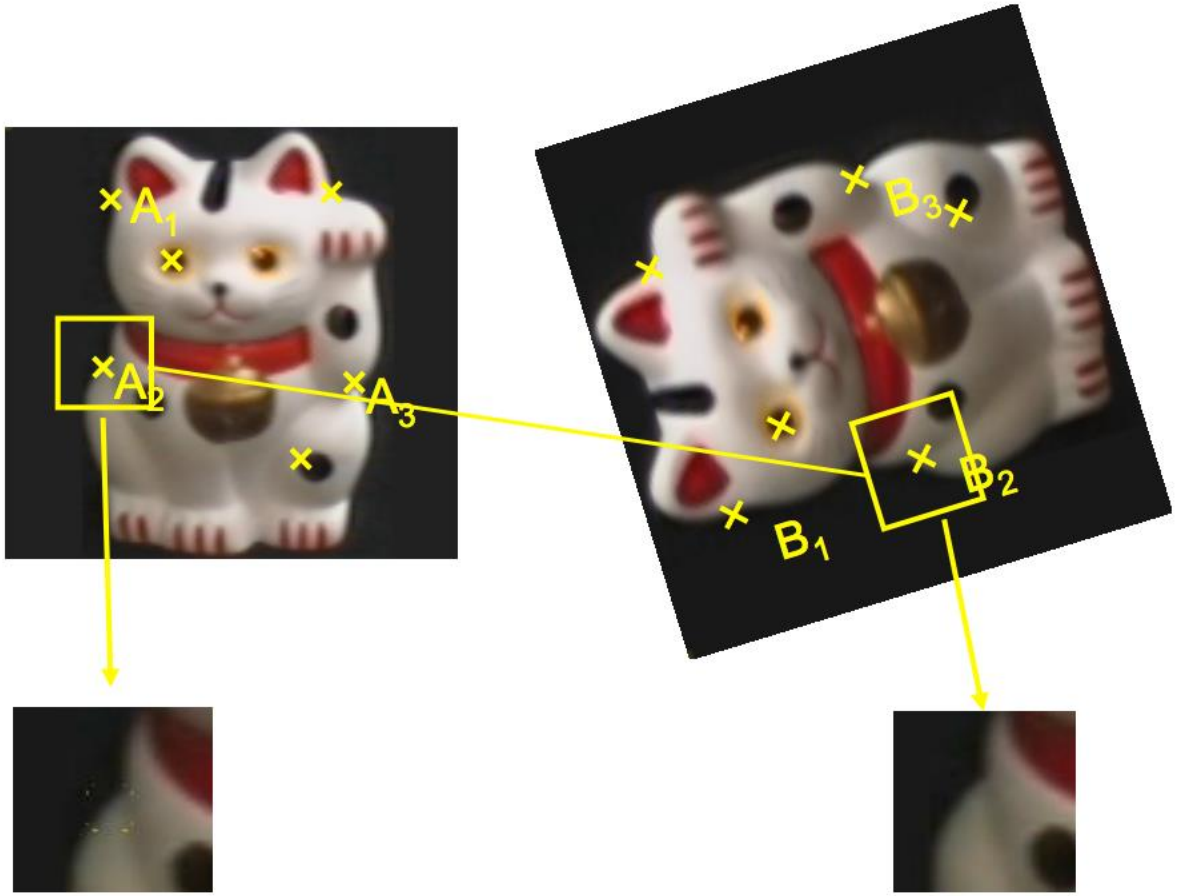
- Image content is transformed into local feature coordinates that are invariant to translation, rotation, scale, and other imaging parameters



Following slides credit: CVPR 2003 Tutorial on **Recognition and Matching Based on Local Invariant Features** David Lowe

Rotation invariant descriptors

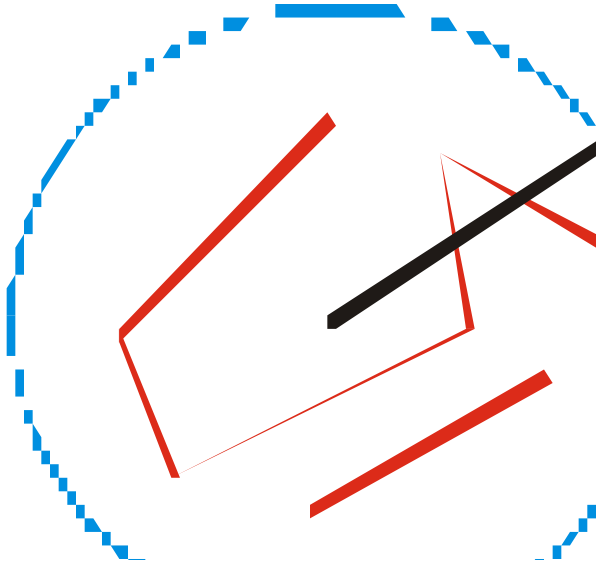
Option 1: Normalize patches by rotating them



Rotation invariant descriptors

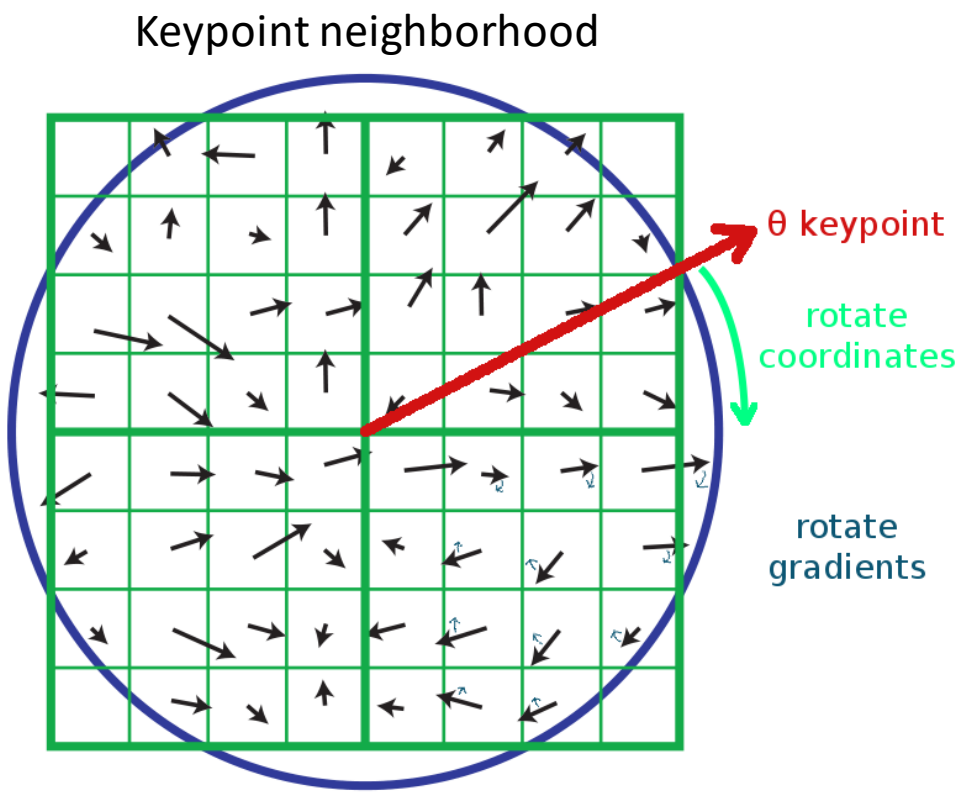
Option 2: Construct a rotation invariant descriptor

- We are given a keypoint and its scale from DoG
- We will select a characteristic orientation for the keypoint (based on the most prominent gradient there; discussed next slide)
- We will describe all features ***relative*** to this orientation
- Causes features to be rotation invariant!
  - If the keypoint appears rotated in another image, the features will be the same, because they're **relative** to the characteristic orientation

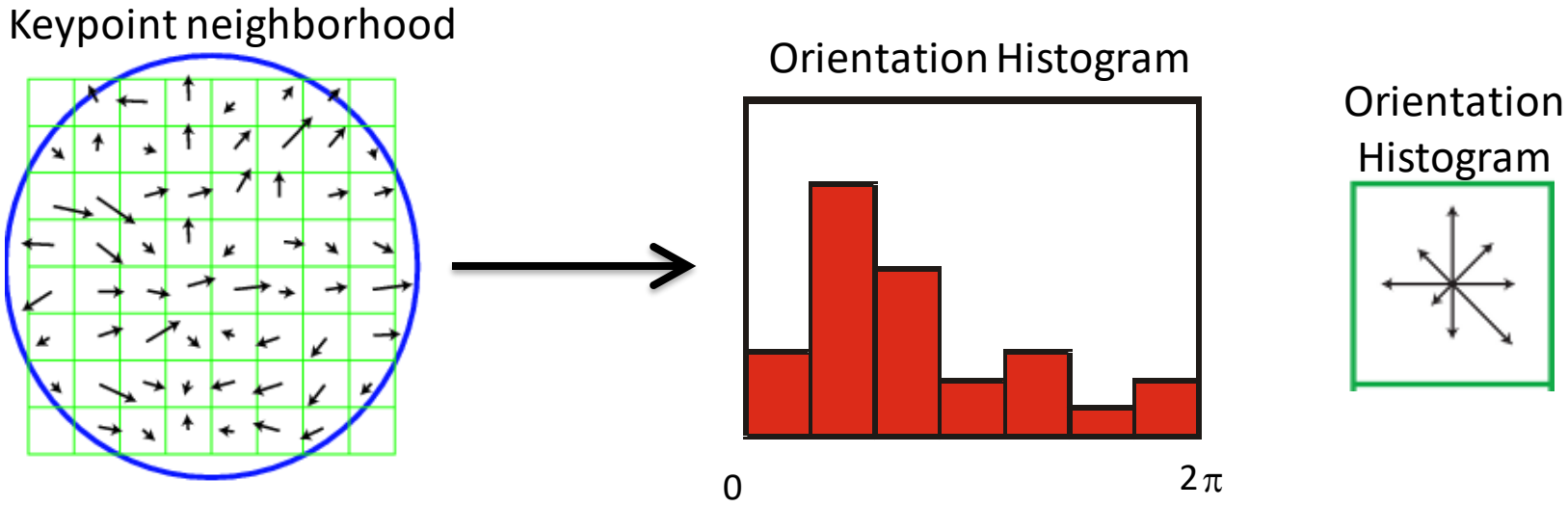


SIFT descriptor (SIFT=Scale-Invariant Feature Transform)

- Gradient-based descriptor to capture texture in the keypoint neighborhood
- Use the blurred image associated with the keypoint's scale (8x8 pixels in this example)
- Take image gradients over the keypoint neighborhood.
- To become rotation invariant, rotate the gradient directions AND locations by (-keypoint orientation)
  - Now we've cancelled out rotation and have gradients expressed at locations relative to keypoint orientation θ
  - We could also have just rotated the whole image by -θ, but that would be slower.

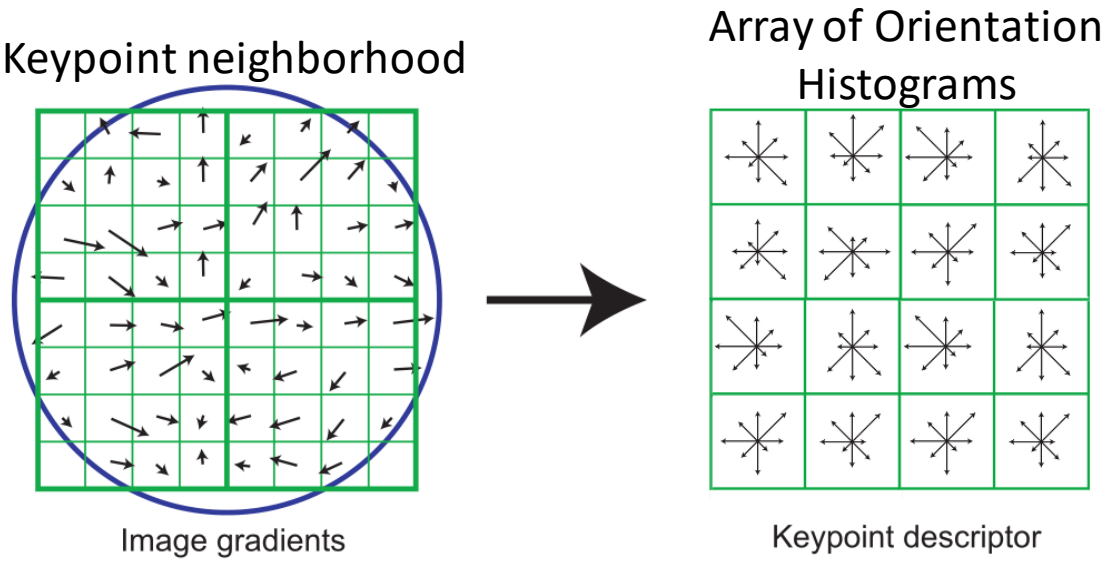


SIFT descriptor formation



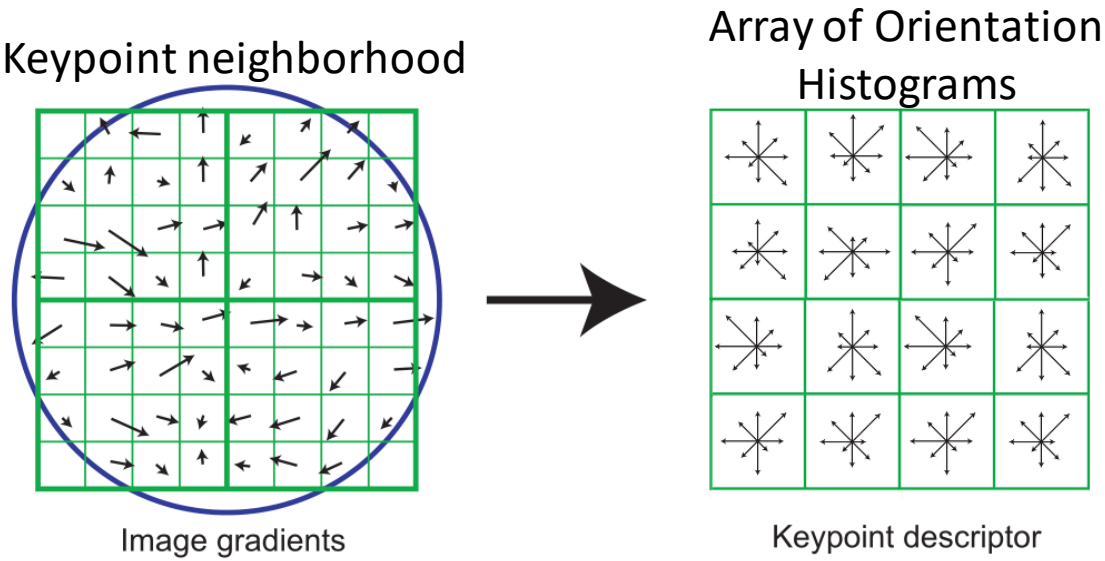
- Stack all gradients into a single vector?
- Using precise gradient locations is fragile. We'd like to allow some "slop" in the pixel configurations within the image, but still produce a very similar descriptor
- Create array of orientation histograms (a 1x1 array is shown), with 8 orientation bins

SIFT descriptor formation



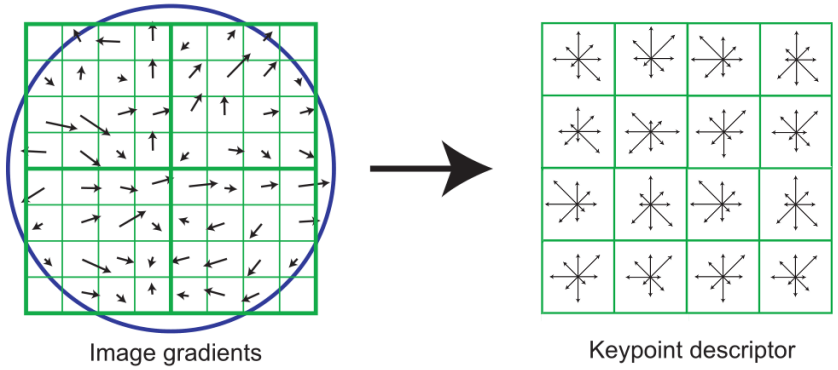
- Create array of orientation histograms (a 4x4 array is shown)
- Put the rotated gradients into their local orientation histograms
  - A gradients's contribution is divided among the nearby histograms based on distance. If it's halfway between two histogram locations, it gives a half contribution to both.
  - Also, scale down gradient contributions for gradients far from the center
- The SIFT authors found that best results were with 8 orientation bins per histogram, and a 4x4 histogram array.

SIFT descriptor formation



- 8 orientation bins per histogram, and a 4x4 histogram array, yields 8 x 4x4 = 128 numbers.
- So a SIFT descriptor is a length 128 vector, which is invariant to rotation (because we rotated the patch) and scale (because we worked with the scaled image from DoG)
- We can compare each vector from image A to each vector from image B to find matching keypoints!
  - Euclidean “distance” between descriptor vectors gives a good measure of keypoint similarity

SIFT descriptor formation



- Adding robustness to illumination changes:
- Remember that the descriptor is made of gradients (differences between pixels), so it's already invariant to changes in brightness (e.g. adding 10 to all image pixels yields the exact same descriptor)
- A higher-contrast photo will increase the magnitude of gradients linearly. To correct for contrast changes, normalize the histogram (scale to magnitude=1.0)
- Very large image gradients are usually from unreliable 3D illumination effects (glare, etc). To reduce their effect, clamp all values in the vector to be ≤ 0.2 (an experimentally tuned value). Then normalize the vector again.
- Result is a vector which is fairly invariant to illumination changes.

Sensitivity to number of histogram orientations

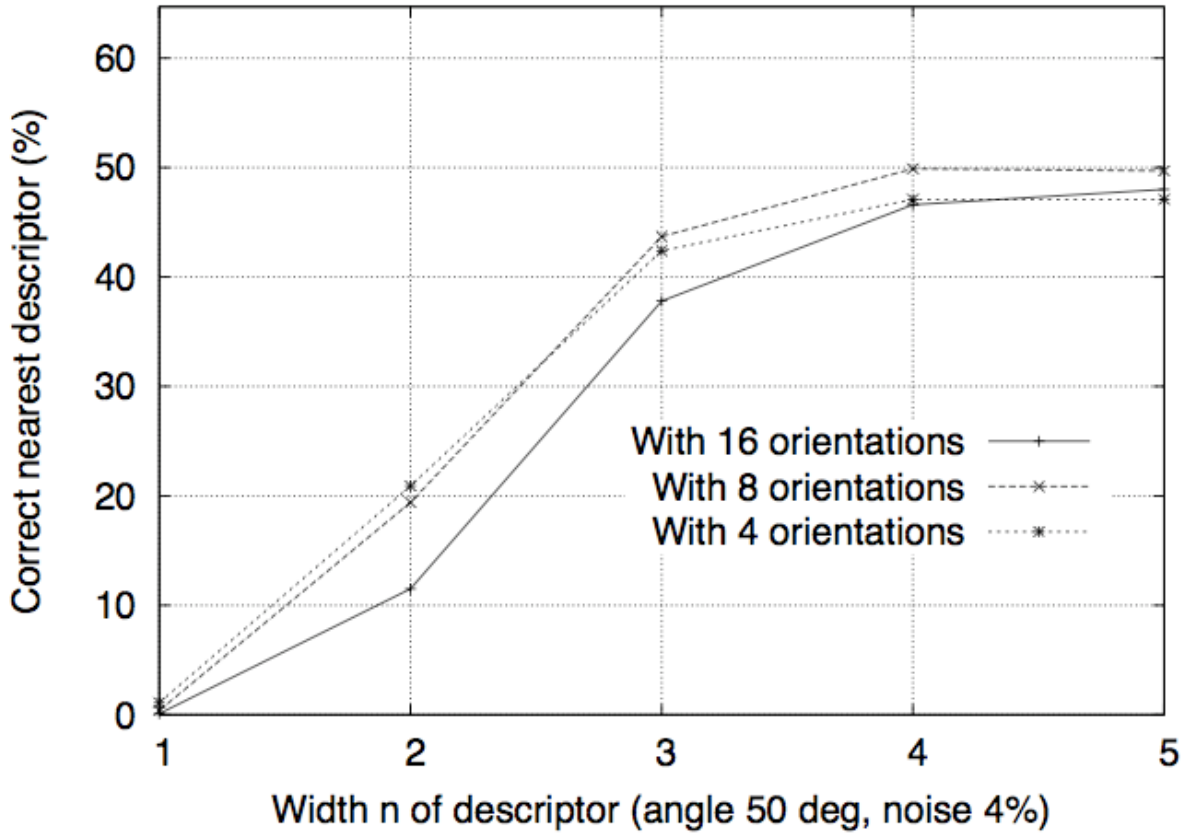
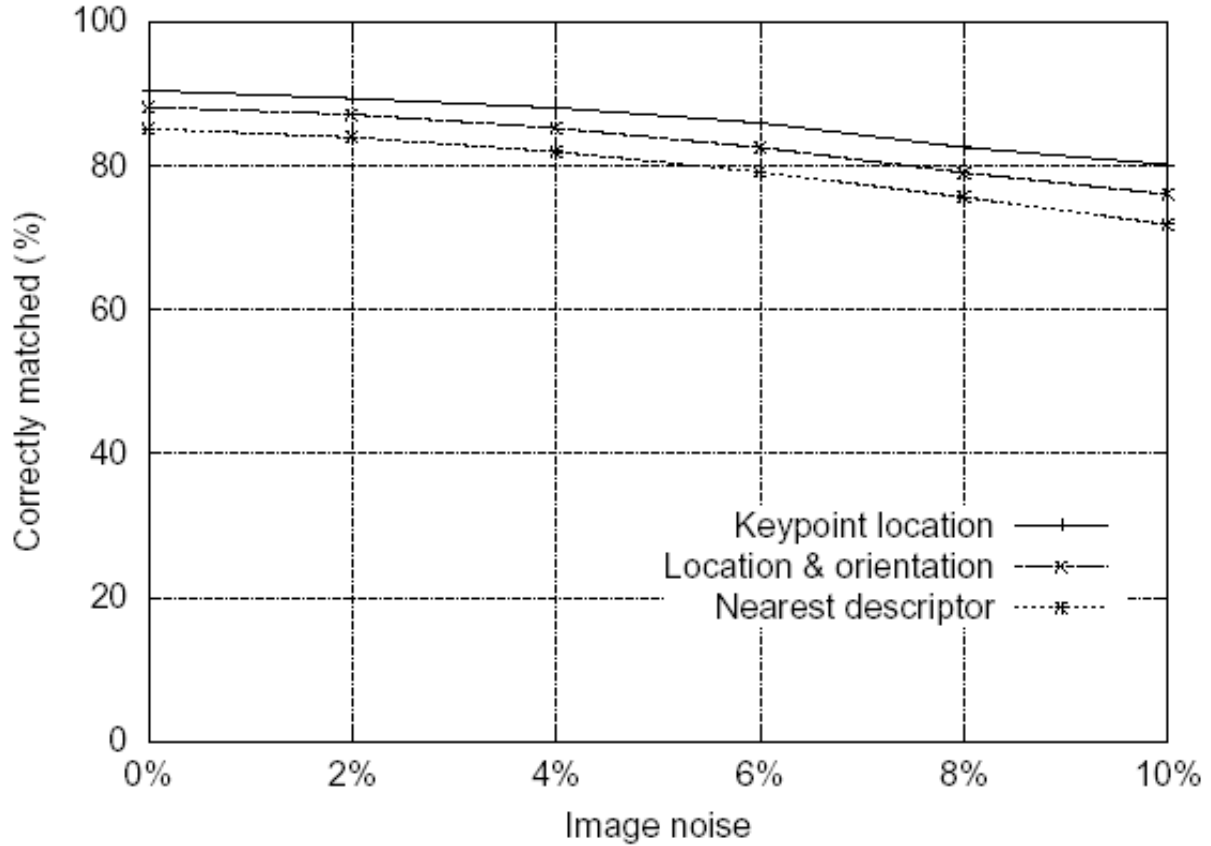


Figure 8: This graph shows the percent of keypoints giving the correct match to a database of 40,000 keypoints as a function of width of the $n \times n$ keypoint descriptor and the number of orientations in each histogram. The graph is computed for images with affine viewpoint change of 50 degrees and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

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

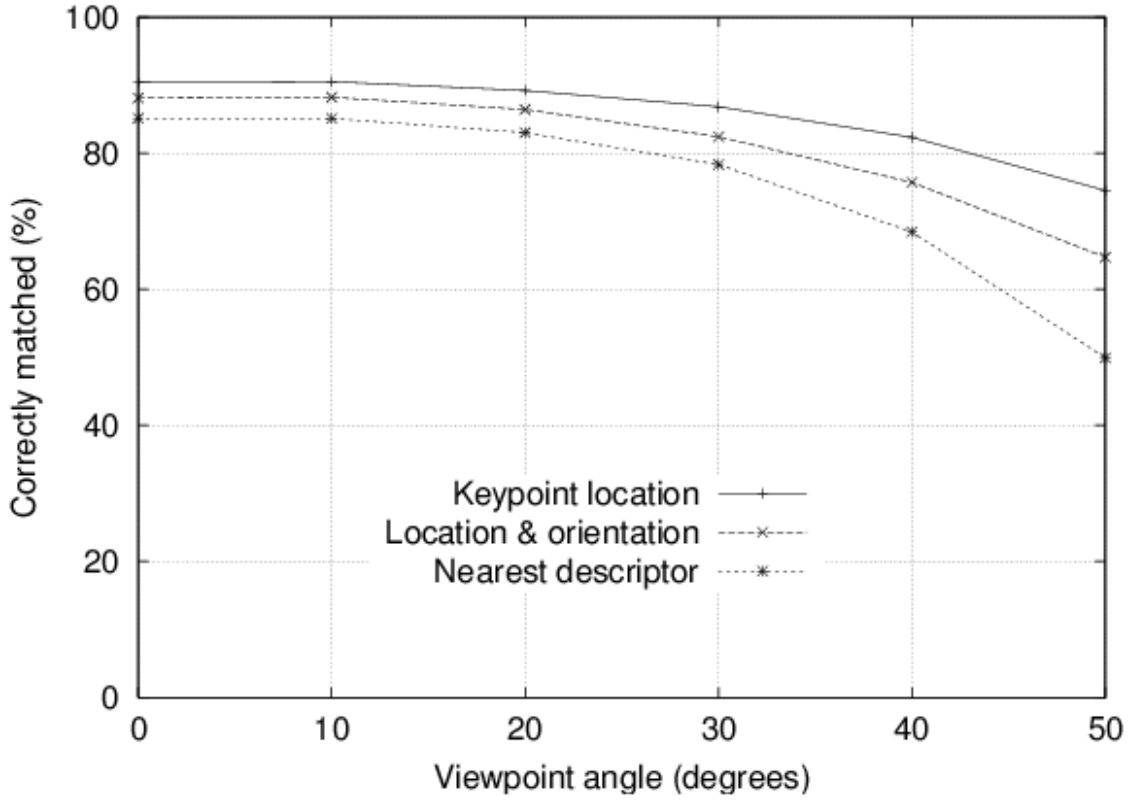
Feature stability to noise

- Match features after random change in image scale & orientation, with differing levels of image noise
- Find nearest neighbor in database of 30,000 features



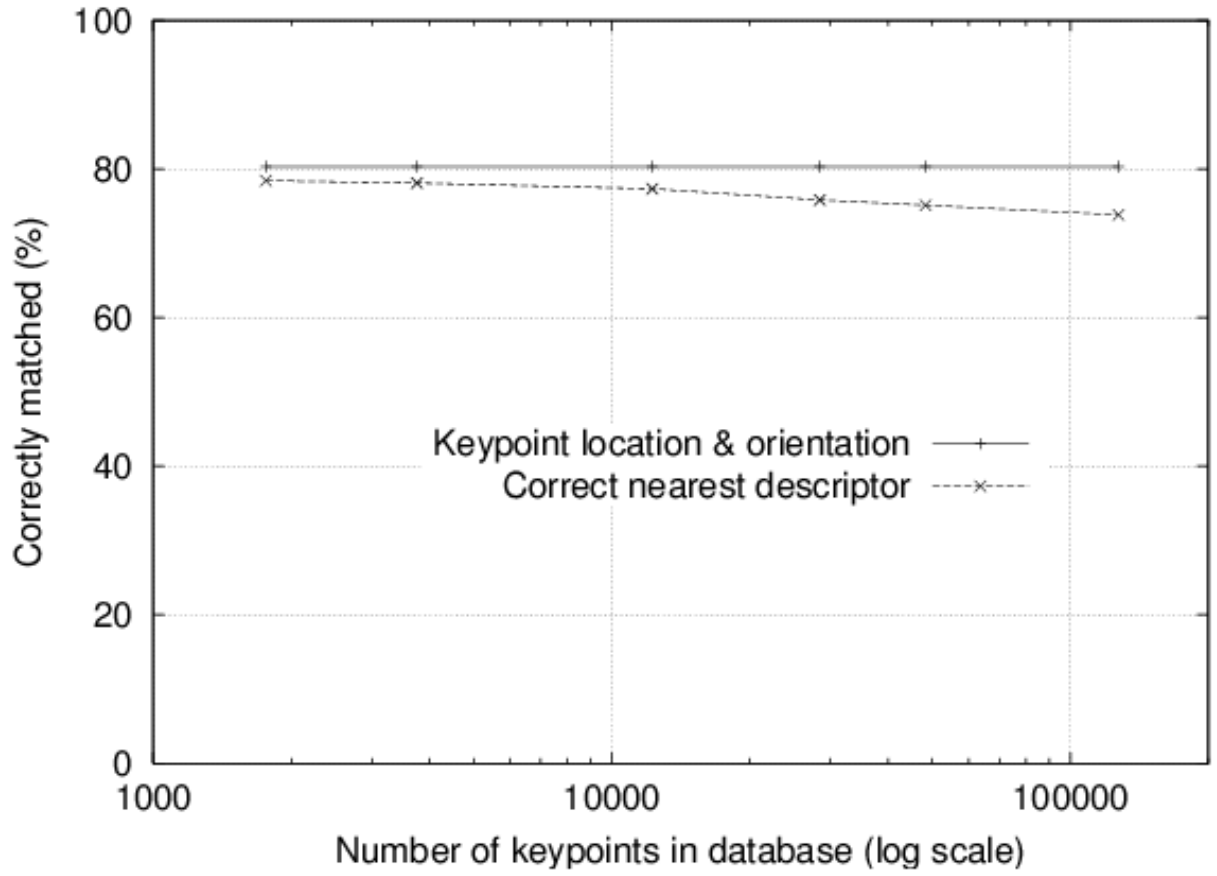
Feature stability to affine changes

- Match features after random change in image scale & orientation, with 2% image noise, and affine distortion
- Find nearest neighbor in database of 30,000 features



Distinctiveness of features

- Vary size of database of features, with 30 degree affine change, 2% image noise
- Measure % correct for single nearest neighbor match



Useful SIFT resources

- An online tutorial: http://www.aishack.in/2010/05/sift-scale-invariant-feature-transform/
- Wikipedia: http://en.wikipedia.org/wiki/Scale-invariant_feature_transform

Figure 12: The training images for two objects are shown on the left. These can be recognized in a cluttered image with extensive occlusion, shown in the middle. The results of recognition are shown on the right. A parallelogram is drawn around each recognized object showing the boundaries of the original training image under the affi ne transformation solved for during recognition. Smaller squares indicate the keypoints that were used for recognition.

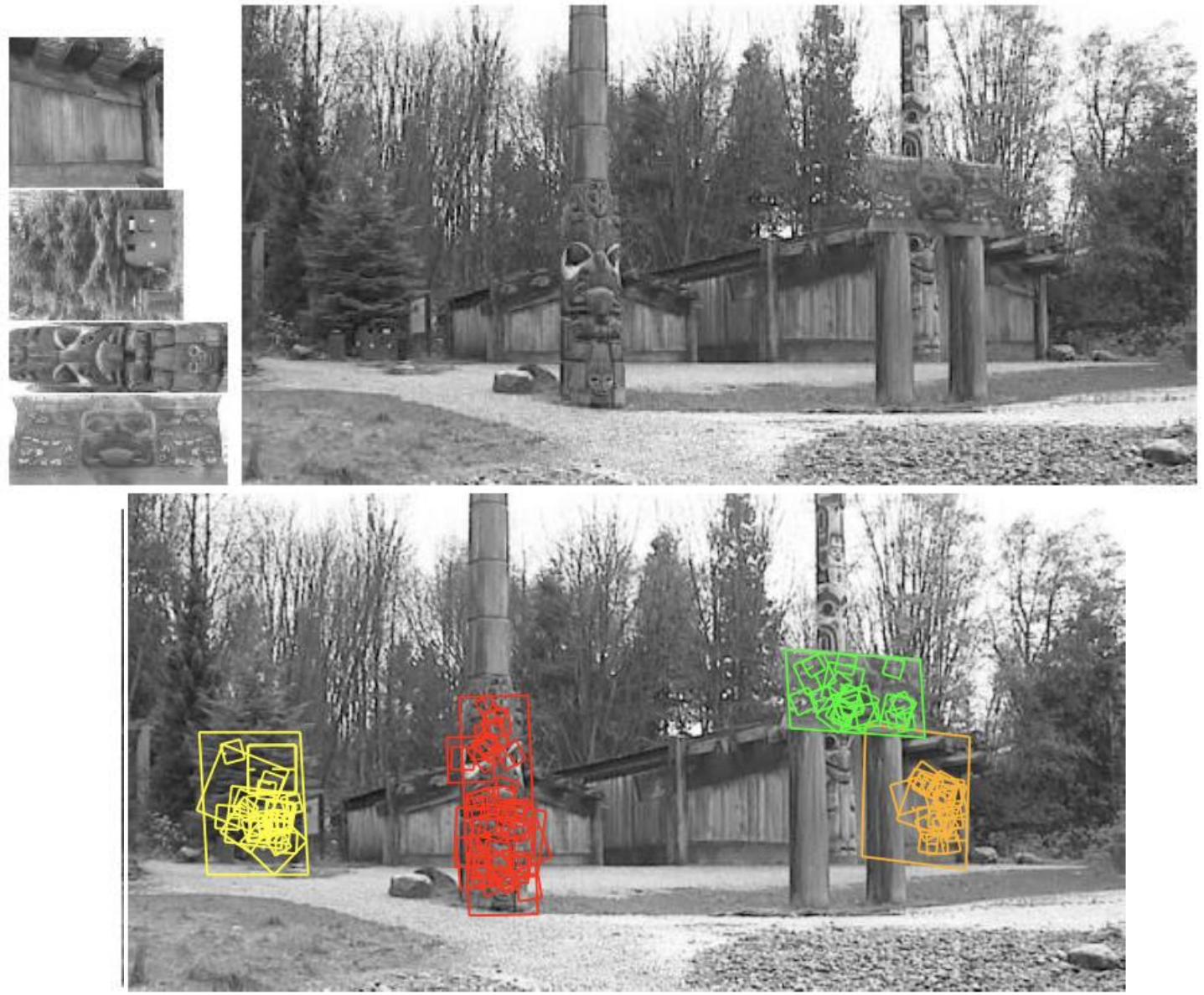
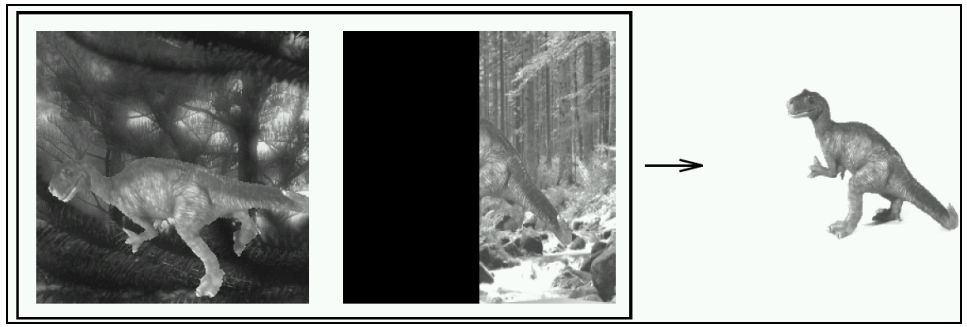


Figure 13: This example shows location recognition within a complex scene. The training images for locations are shown at the upper left and the 640x315 pixel test image taken from a different viewpoint is on the upper right. The recognized regions are shown on the lower image, with keypoints shown as squares and an outer parallelogram showing the boundaries of the training images under the affi ne transform used for recognition.

Recognition of specific objects, scenes



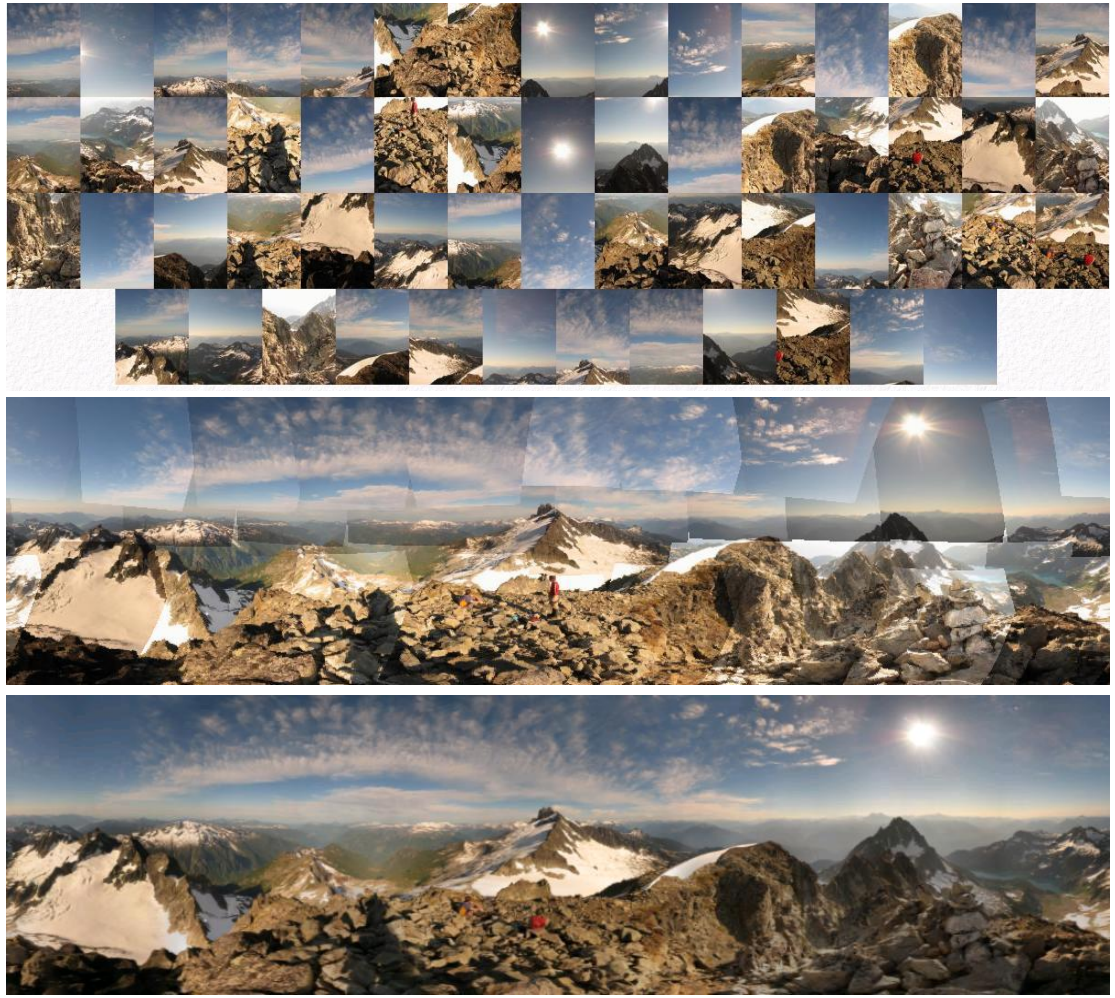
Schmid and Mohr 1997

Sivic and Zisserman, 2003

Rothganger et al. 2003

Lowe 2002

Panorama stitching/Automatic image mosaic



http://matthewalunbrown.com/autostitch/autostitch.html

Wide baseline stereo

[Image from T. Tuytelaars ECCV 2006 tutorial]

Applications of local invariant features

- Recognition
- Wide baseline stereo
- Panorama stitching
- Mobile robot navigation
- Motion tracking
- 3D reconstruction
- …

Summary

- SIFT Descriptor
  - Descriptor calculation
  - Experimental Analysis
  - Applications