

# NOTES FOR COMPUTER VISION SEMINAR TALK ON SIFT 2-2-2007

ROSS MESSING

## 1. INTRODUCTION

The scale-invariant feature transform (SIFT) is a feature matching framework by Lowe. Like other feature-based matching systems (I assume), it sequentially finds keypoints in an image, determines a feature vector for each, and matches those features to a database.

In this talk, I'll outline each of these steps.

## 2. IDENTIFY CANDIDATE FEATURE KEYPOINTS

**2.1. Finding extrema across scales.** Gaussian blur the scales with a bunch of stepped-variance gaussians, and take the difference to approximate a Laplacian. Find maxima and minima that are maximal and minimal across different levels of blur, and different scales, effectively finding scale-invariant illumination extrema.

**2.2. Removing undesirable extrema.** First, each extremum was adjusted to be the maximum around the original extremum of the derivative of a Taylor expansion of a difference of Gaussians. Bad extrema (as determined from a criterion based on this adjustment) were thrown out.

Next, extrema corresponding to otherwise-un-noteworthy locations along important edges were removed. Without this, very small changes to an image could lead to very different points along an important edge (that isn't around other features) being chosen as keypoints. To get rid of these, they got rid of keypoints where the ratio of the eigenvalues of a Hessian of the difference of Gaussians (by space) exceeded a threshold.

## 3. CALCULATE FEATURE VALUE AT EACH KEYPOINT

Now that we know where our features are, we calculate them. First normalizing for feature-neighborhood orientation, we create a grid of orientation histograms for local areas surrounding the keypoint. This seems to work pretty well. Intuitively, I wonder about orientation stability at finer scales (not when everything's blurred out, but pixel-scale features might not survive rotation well). By allowing a little bit of sliding of position within very localized areas, the system attempts to use the same trick the visual system does for local location independence in primary visual cortex complex cells. Sort of.

## 4. USING THE FEATURES TO DO SOMETHING USEFUL (LIKE OBJECT RECOGNITION)

Now that we've got our cool, big set of gigantic features, we want to do something with them.

**4.1. Finding candidate matches.** We find features that match features from a template set.

First, we find match candidates by doing an approximate nearest neighbor match across the whole database.

Then we throw our feature matches into a Hough transform, and find clusters that might correspond to our objects.

**4.2. Testing matches by finding affine transforms that fit.** We try to find an affine transform between the features in the image, and the template. Because we only do this for identified candidate features in the image, ignoring features of the template that aren't there in the image, this is robust to partial occlusion.

This approach acknowledges its limitations - it's not going to work for things that are rotated radically enough that full 3D transforms are much more reasonable than affine transforms, so for any potential candidate match, it tries to solve for the affine parameters of the transformation mapping a model to an image. It turns this into a simple linear equation, trying to find  $x$  where  $Ax = b$ , so  $x = A^{-1}b = A$ . I'm a little confused about why this becomes  $x = [A^T A]^{-1} A^T b$ . Are there some properties of  $A$  that make this reformulation more desirable than just taking the inverse of  $A$ ?

## 5. ISSUES / POINTS OF DISCUSSION

I had some general issues with this paper. First, it seems like a model with a bunch of parameters, most of which are chosen because they fit the test data, some of which are just picked arbitrarily, and very few of which are really explained. If I'm going to pick parameters to match my data, I'll at least speculate about, and maybe document, what happens at different values. In Lowe's defense, he does that with the choice of  $\sigma$  for extrema-detection, but nowhere else.

Second, as biological-vision-guy, I'll caution that while complex cells in primary visual cortex do, in fact, show some position invariance, and are orientation sensitive, the features used in this paper should not be taken as models of complex cells in primary visual cortex. Among other issues, complex cells in primary visual cortex are almost always motion-sensitive (and sometimes motion-direction-specific).

Lastly, I know nothing about this area. How does Lowe's SIFT technique compare to other approaches in the field?