# Object Recognition from Local Scale-Invariant Features

## Abstract

An object recognition system has been developed that uses a new class of local image features. The features are invariant to image scaling, translation, and rotation, and partially invariant to illumination changes and affine or 3D projection. These features share similar properties with neurons in inferior temporal cortex that are used for object recognition in primate vision. Features are efficiently detected through a staged filtering approach that identifies stable points in scale space. Image keys are created that allow for local geometric deformations by representing blurred image gradients in multiple orientation planes and at multiple scales. The keys are used as input to a nearest-neighbor indexing method that identifies candidate object matches. Final verification of each match is achieved by finding a low-residual least-squares solution for the unknown model parameters. Experimental results show that robust object recognition can be achieved in cluttered partially-occluded images with a computation time of under 2 seconds.

## 1. Introduction

Object recognition in cluttered real-world scenes requires local image features that are unaffected by nearby clutter or partial occlusion. The features must be at least partially invariant to illumination, 3D projective transforms, and common object variations. On the other hand, the features must also be sufficiently distinctive to identify specific objects among many alternatives. The difficulty of the object recognition problem is due in large part to the lack of success in finding such image features. However, recent research on the use of dense local features (e.g., Schmid & Mohr [19]) has shown that efficient recognition can often be achieved by using local image descriptors sampled at a large number of repeatable locations.

This paper presents a new method for image feature generation called the Scale Invariant Feature Transform(SIFT). This approach transforms an image into a large collection of local feature vectors, each of which is invariant to image translation, scaling, and rotation, and partially invariant to illumination changes and affine or 3D projection. Previous approaches to local feature generation lacked invariance to scale and were more sensitive to projective distortion and illumination change. The SIFT features share a number of properties in common with the responses of neurons in inferior temporal (IT) cortex in primate vision. This paper also describes improved approaches to indexing and model verification.

The scale-invariant features are efficiently identified by using a staged filtering approach. The first stage identifies key locations in scale space by looking for locations that are maxima or minima of a difference-of-Gaussian function. Each point is used to generate a feature vector that describes the local image region sampled relative to its scale-space coordinate frame. The features achieve partial invariance to local variations, such as affine or 3D projections, by blurring image gradient locations. This approach is based on a model of the behavior of complex cells in the cerebral cortex of mammalian vision. The resulting feature vectors are called SIFT keys. In the current implementation, each image generates on the order of 1000 SIFT keys, a process that requires less than 1 second of computation time.

The SIFT keys derived from an image are used in a nearest-neighbour approach to indexing to identify candidate object models. Collections of keys that agree on a potential model pose are first identified through a Hough transform hash table, and then through a least-squares fit to a final estimate of model parameters. When at least 3 keys agree on the model parameters with low residual, there is strong evidence for the presence of the object. Since there may be dozens of SIFT keys in the image of a typical object, it is possible to have substantial levels of occlusion in the image and yet retain high levels of reliability.

The current object models are represented as 2D locations of SIFT keys that can undergo affine projection. Sufficient variation in feature location is allowed to recognize perspective projection of planar shapes at up to a 60 degree rotation away from the camera or to allow up to a 20 degree rotation of a 3D object.

## 2. Related research

Object recognition is widely used in the machine vision industry for the purposes of inspection, registration, and manipulation. However, current commercial systems for object recognition depend almost exclusively on correlation-based template matching. While very effective for certain engineered environments, where object pose and illumination are tightly controlled, template matching becomes computationally infeasible when object rotation, scale, illumination, and 3D pose are allowed to vary, and even more so when dealing with partial visibility and large model databases.

An alternative to searching all image locations for matches is to extract features from the image that are at least partially invariant to the image formation process and matching only to those features. Many candidate feature types have been proposed and explored, including line segments [6], groupings of edges [11, 14], and regions [2], among many other proposals. While these features have worked well for certain object classes, they are often not detected frequently enough or with sufficient stability to form a basis for reliable recognition.

There has been recent work on developing much denser collections of image features. One approach has been to use a corner detector (more accurately, a detector of peaks in local image variation) to identify repeatable image locations, around which local image properties can be measured. Zhang *et al.* [23] used the Harris corner detector to identify feature locations for epipolar alignment of images taken from differing viewpoints. Rather than attempting to correlate regions from one image against all possible regions in a second image, large savings in computation time were achieved by only matching regions centered at corner points in each image.