Abstract. The subject of this chapter is image key points which we define as a distinctive point in an input image which is invariant to rotation, scale and distortion. In practice, the key points are not perfectly invariant but they are a good approximation. To make our discussion more concrete we shall concentrate on two key point algorithms: SIFT and SURF and their use in spatial alignment.

13.1 Scale-Invariant Feature Transform

The scale-invariant feature transform [5] (SIFT) algorithm provides a robust method for extracting distinctive features from an input image $I$ which are invariant to rotation, scale and distortion. These points (known as “key-points”) are found by detecting local extrema in a multi-scale representation of $I$:

$$\{D(m,n|\sigma_1), D(m,n|\sigma_2), \ldots, D(m,n|\sigma_K)\}$$

where $D(m,n|\sigma_k)$ is the difference-of-Gaussian (DoG) representation of $I$ at a scale $\sigma_k$:

$$D(m,n|\sigma_k) = I(m,n) \otimes G(\sigma_k) - I(m,n) \otimes G(\sigma_{k+1})$$

and $I(m,n) \otimes G(\sigma_k)$ denotes the convolution of $I(m,n)$ with the two-dimensional zero-mean Gaussian $G(\sigma_k)$ and $\sigma_{k+1} = 2^{1/3} \sigma_k$. The parameter $K$ is specified by the user and is based on the maximum width in pixels.

The local extrema in (13.1) are defined as points $(m,n|\sigma_k)$ for which $D(m,n|\sigma_k)$ is greater than its 26 neighbors. This includes eight immediate neighbors from the $D(p,q|\sigma_k)$ and nine neighbors from $D(p,q|\sigma_{k+1})$ and nine neighbors from $D(p,q|\sigma_{k-1})$. This is followed by accurate interpolation of scale space using the Taylor series expansion upto a second degree of $D(m,n|\sigma_k)$ in the neighborhood of $(m,n)$ and $\sigma_k$.

Stability of the extrema is further ensured by rejecting key-points with low contrast and key points localized along edges. For a descriptor of the key-point, an orientation histogram is computed of the area surrounding the key-point. Gradient
magnitude and the weight of a Gaussian window originating at the key-point add to the value of each sample point within the considered region.

Mathematically the SIFT operator is computed by partitioning the image region surrounding each detection key point into a $4 \times 4$ grid of subregions, and computing an orientation histogram of 8 bins in each sub-region (Fig. 13.1).

![Fig. 13.1](image-url)

*Fig. 13.1* Shows the formation of the SIFT descriptor for a key-point located at $(m, n)$. In the figure we show the $4 \times 4$ sub-regions with their orientation vectors.

The grid is square, with the $x$-axis oriented along the key-point gradient direction and the width of the grid being approximately 12 times the detected scale of the key-point. Within each subregion, the gradient orientation of each pixel is entered into the orientation histogram, with weighted vote proportional to the gradient magnitude. A normalized 128 component vector is formed by concatenating the 16 region containers.

### 13.1.1 Hyperspectral Images

The SIFT operator has been extended to color images [1, 3] and hyperspectral images [6] as follows. Given a hyperspectral image

$$I(m, n|l), l \in \{1, 2, \ldots, L\},$$

with $L$ bands, we separately perform the DoG operation on each band [6]:

$$D(m, n|l) = I(m, n|l) \otimes G(\sigma_k) - I(m, n|l) \otimes G(\sigma_{k+1}),$$

and then combine the $D_l(m, n|l)$ using a non-linear function $f$: