计算机视觉报告：SIFT 特征提取与匹配

Technical Report on Computer Vision Programming Assignment: SIFT detection and matching

叶翰樵

中国科学院大学-人工智能学院-*202218020629017*

中国科学院自动化研究所

yehanqiao22@mails.ucas.ac.cn

Contents

PART

I

Preliminaries

对图像底层的局部特征进行检测和描述是解决很多计算机视觉问题的基础，例如物体识别、图像匹配和图像复原。在对这些局部特征进行匹配后，算法就能够对不同视角图像当中的特殊区域进行识别和比较。

在图像底层局部特征提取的诸多算法当中，尺度不变特征变换（Scale-Invariant Feature Transform, SIFT）是迄今使用最为广泛的一种特征，它具有以下优点：

- 旋转不变性（*rotation invariance*）和尺度不变性（*scale invariance*）：对图像的旋转和尺度变化具有不变性；
- 对三维视角变化、光照变化以及噪声具有很强的适应性；
- 在存在遮挡或场景杂乱的情况下，底层的局部特征具有不变性；
- 辨别力强：特征之间相互区分的能力强，有利于匹配；
- 扩展性强：能够与其他形式的特征向量联合
- 易获取：一般 500×500 的图像能提取出约 2000 个特征点

下面对 SIFT 背后的原理进行简单论述。

Section 1

LoG 的空间选择特性

回忆当我们在进行边缘检测时，会使用高斯偏导（Derivative of Gaussian）对图像做卷积，得到的图像与在原图像基础上先后进行高斯平滑与求导两个操作等价，最终响应值最高处即被确定为边缘所在位置。

其实，高斯二阶导也可用于边缘检测。高斯二阶导也被叫做拉普拉斯核（Laplacian of Gaussian, LoG），与之卷积所获得的图像与在原图像基

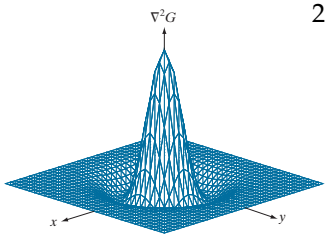

图 1. 3-D plot of the negative of the LoG.

础上先后进行高斯平滑与连续求二阶导两个操作等价，最终响应值过零点处即被确定为边缘所在位置。

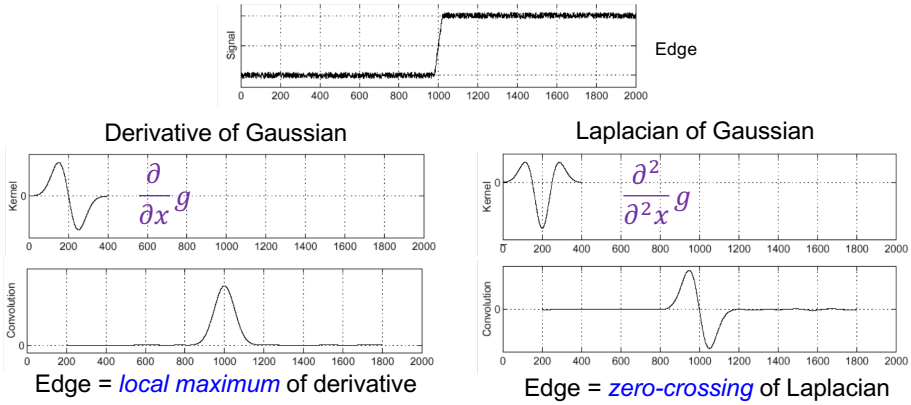

图 2. 高斯偏导和 LoG 分别进行边缘检测 [1]。

不仅如此，LoG 相比于高斯偏导还拥有一个更好的特性，即具有空间选择特性（*Spatial Selection*）：当 LoG 的尺度 σ 与信号的宽度匹配时，两者卷积会在原信号的中心位置有最大响应。

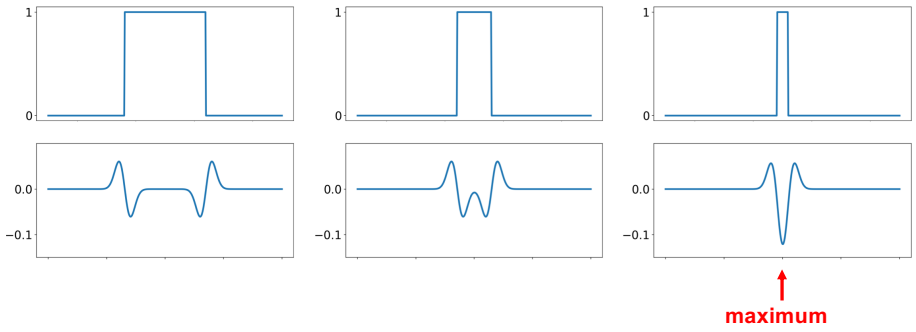

图 3. LoG 具有空间选择特性。

因此容易想到由 LoG 的空间选择特性，用一组尺度不一的 LoG 分别与某一尺度未知信号做卷积，通过判断响应值的大小最终确定该信号的尺度。而响应值最大时 LoG 与信号在尺度上到底存在什么数量关系呢?

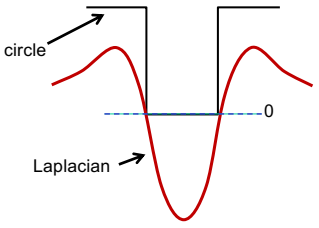

图 4. LoG 的过零点刚好与信号卡住的情形。

Intuition 1

当 LoG 的过零点刚好与信号卡住的时候，会在信号的中心点产生最大响应。于是，令

$$\nabla^2_{\text{norm}} g = 0$$

得到 $x^2 + y^2 = 2\sigma^2$，即要想产生最大响应，信号的半径 r 应等于 LoG

过零点围成的半径 $\sqrt{2}\sigma$。换句话说，与半径为 r 的信号产生最大响应的 LoG 参数 σ 应当满足：$\sigma = r/\sqrt{2}$。

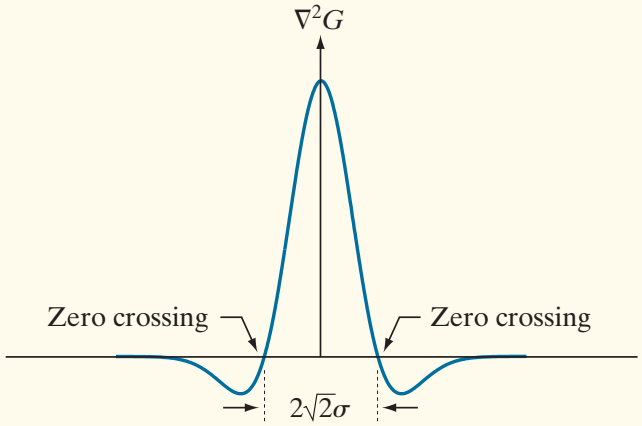

图 5. LoG 过零点

直觉上看，信号的尺度和与之产生最大响应值的 LoG 的 σ 存在上述关系。但事实上，卷积的响应值会随着尺度 σ 的增大不断衰减，因此在此之前还必须对响应值进行尺度补偿（*scale normalize*）：

$$\nabla^2_{\text{normalized}} g = \sigma^2 \left(\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 g}{\partial y^2} \right) \tag{1.1}$$

Section 2

图像尺度空间

对于图像信号，如果希望找到图像上以某一点为中心所在局部特征的尺度 r，就可以使用一组尺度（σ）不同的归一化 LoG 分别对图像做卷积。**图** 6表示了一组用多尺度 LoG 相卷积得到的图像尺度空间，在纵向找到该位置处响应值最大时的 LoG，然后根据其参数 σ 即可计算局部特征的尺度 $r = \sqrt{2}\sigma$。

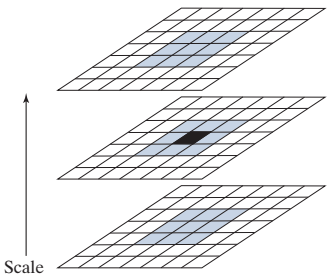

图 6. 图像尺度空间特征检测

要对图像提取局部特征时，就对每一个像素坐标进行上面的操作。但是实际中并不会将 LoG 组的尺度划分得过细以逼近响应极值，这样将造成过大的计算消耗，同时意味着每一个像素坐标只能对应一个局部特征。因此，通常每三个 LoG 尺度进行比较，如果中间的响应值大于上下，则

认为中间这个位置为响应极值，从而可确定出其所对应局部特征的尺寸。

然后横向在响应极值所在的尺度平面上看，周围像素其对应的响应也可能是极值，这也就导致了图像中局部特征可能会大量重叠。因此还需要进行*非极大值抑制*，其思路就是在确定响应极值时，不仅纵向跟自己位置进行比较，还需要跟图像尺度空间当中周围的 26 个像素相比，如果该响应值仍然最大，则保留这一处这一尺度的局部特征，否则遗弃该位置。

至此介绍了在图像尺度空间进行*尺度搜索*和*非极大值抑制*两个关键环节。但是上述过程仍然需要进行大量计算，为此 [2] 提出了 Harris-Laplacian 方法，先找出图像当中所有的 Harris 角点，然后只在这些点周围建立尺度空间，进行 LoG 的尺度分析。Harris 对光照、平移、旋转具有不变性，再加上 LoG 的空间选择特性带来具有尺度不变性的局部特征。再到后来 [3] 提出了更加高效的 SIFT 特征提取方法。

PART

II

Keypoint Detection

正如上一章所讲，因为我们希望提取到的局部特征具有尺度不变性，所以既要找局部特征所在位置，还要确定它们的尺度。LoG 已经被说明能够很好地完成上述任务，但它还不是最优的。本章将介绍 SIFT 是如何优化的，以及程序在实现关键点检测时的一些细节。

Section 3

构造高斯金字塔

SIFT 注意到一个明显的问题是当 LoG 尺度越来越大 σ 时，卷积窗口 w 也会越来越大（一般认为 $w \approx 6\sigma$），计算消耗也会越来越明显。为了解决这一问题，SIFT 指出可以采用高斯差分（Difference of Gaussian, DoG）完成如下近似：

$$\underbrace{G(x,y,k\sigma) - G(x,y,\sigma)}_{\text{DoG}} \approx (k-1)\underbrace{\sigma^2\nabla^2 G}_{\text{scale-normalized LoG}} \tag{3.1}$$

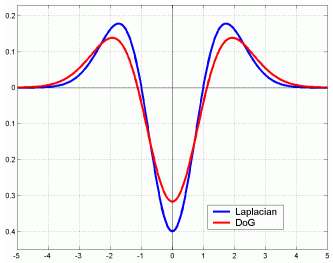

图 7. Approximate LoG with a difference of Gaussians

上式表明，在不考虑尺度因子 $k-1$ 的情况下，DoG 约等于尺度归一化后的 LoG。利用 DoG 近似的好处是，由高斯卷积的性质，大的高斯核（$G(x,y,k\sigma)$）可以拆成两个小高斯核的卷积（**定理** 1），即要想获得方差更大高斯滤波，只需要在小方差高斯滤波的结果上继续进行一个小方差高斯滤波即可。

Theorem 1 用两个方差分别为 σ_1^2 和 σ_2^2 的高斯核与图像进行连续的卷积滤波操作，该过程等价于直接用一个方差为 $\sigma_1^2+\sigma_2^2$ 的高斯核与图像进行一次卷积滤波操作。即：

$$G(x,y,\sigma_1^2+\sigma_2^2)=G(x,y,\sigma_1^2)*G(x,y,\sigma_2^2)$$

于是 $G(x,y,k\sigma)$ 可以表示为：

$$G(x,y,k\sigma)=G(x,y,\sqrt{k^2-1}\sigma)*G(x,y,\sigma) \tag{3.2}$$

这是提升 DoG 图像金字塔创建效率的有效方法。

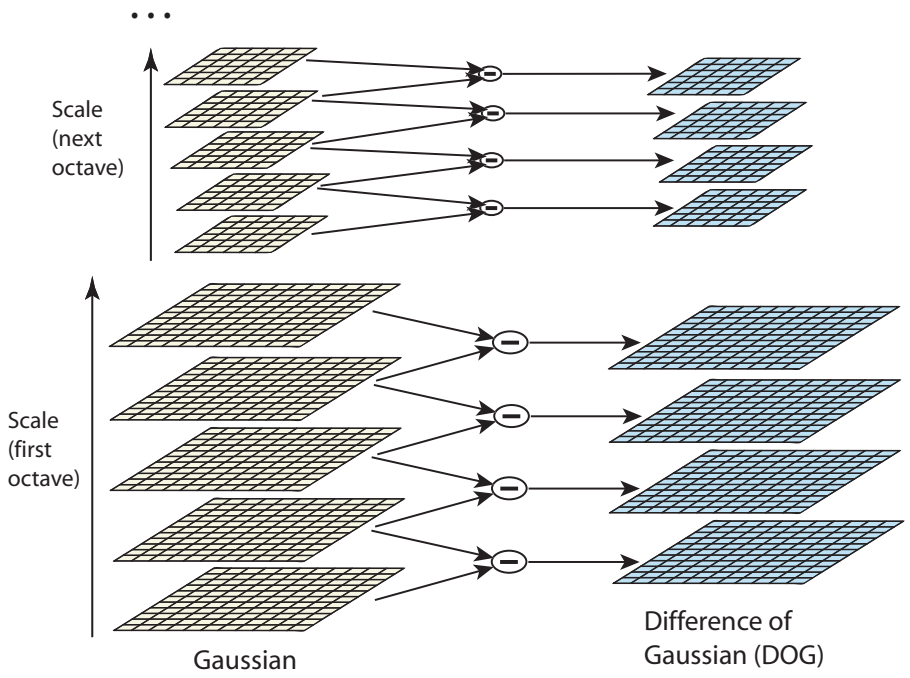

图 8. DoG 图像金字塔

DoG 图像金字塔的结构如**图** 8，它基本由如下三个参数确定：

- 每一个 *octave* 中包含的尺度个数 S；

- *octave* 的个数 O；
- 标准差基本值 σ_0。

如**图** 8所示，在每一个 *octave* 的高斯金字塔（左侧）内一共需要进行 $S+3$ 次高斯操作，高斯核的标准差从下往上依次扩大 $k=2^{1/S}$ 倍（这样设置能够确保待检测的尺度在每个 *octave* 之间刚好连续），相邻两层相减就得到了右侧的 DoG。通过直接对一个 *octave* 的高斯金字塔的第 S 层进行两倍下采样，获得下一个 *octave* 的最底层。下面是构建 DoG 图像金字塔的步骤：

1. 加载图像；
2. 对于索引 $s=[0,1,\ldots,S+2]$，用标准差 $k^s\sigma_0$ 对图像进行高斯模糊。这里可以采用**式** (3.2) 介绍的方法提高效率，每一层直接在上一层的基础上进行模糊；
3. 相邻的两张高斯模糊图像做差得到 DoG，每三层 DoG 就可以叠放成**图** 6所示的 3D 张量；
4. 重复 2-3 步以最终得到 O 个 *octave*。每一个 *octave* 的底层高斯模糊图像可以直接从上一 *octave* 的第 S 层两倍下采样获得。

SIFT Descriptor

PART III

Matching Descriptors

PART IV

Experiment

PART V

参考文献

[1] D. Marr and E. Hildreth, “Theory of edge detection,” *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 207, no. 1167, pp. 187–217, 1980.

[2] K. Mikolajczyk and C. Schmid, “Indexing based on scale invariant interest points,” in *Proceedings Eighth IEEE International Conference on Computer Vision. ICCV 2001*, vol. 1. IEEE, 2001, pp. 525–531.

[3] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, pp. 91–110, 2004.