第 3 章

多类型数据表征

随着大数据技术的发展与应用，人们经常面临诸多不同类型的数据，如时序数据、文本数据和图像数据等。多类型数据表征是指在数据分析过程中对不同类型数据进行有效表示。本章主要介绍三种常用的数据类型及其表征方法，并结合具体的案例阐述了图像数据的处理过程及其表示方法。通过本章的学习，读者可以了解常见的数据类型及其表征方法，更好地理解数据，为执行更为复杂的数据处理任务奠定基础。

3.1 问题导入

在实际工程应用中，通过不同传感器设备采集的数据，往往具有多样性、高维度和复杂性的特点。对原始数据进行适当的表征是提升机器学习和数据分析任务性能的重要前提。为实现对不同数据类型的特征表示，需要解决以下问题：

（1）如何对时序数据进行处理以转换成适合机器学习模型输入的特征；

（2）如何对文本数据进行处理以提取出适用于自然语言处理任务的文本特征；

（3）如何从图像数据中提取特征以形成紧凑且有用的数据表示，用于图像处理任务。

针对以上三个问题，本章将从时序数据表征、文本数据表征和图像数据表征三个方面进行介绍。

3.2 时序数据表征

时序数据是指按时间顺序采集的数据集合，每个时间序列表示在不同时间点上某个观测变量的取值。从时序数据中挖掘有用的模式和规律，在金融、生物医学等领域中具有重要的意义。时序数据特征可分为如下三类：时域特征、频域特征和时频域特征。时域特征是基于时间序列原始数据获取的特征，描述了时间序列在时间维度上的统计特性，用以表征数据的整体趋势、周期性、幅度以及其他与时间相关的信息。常用的时域特征表示方法如表 2-5 所示，在此不再赘述。

3.2.1 频域特征

频域特征是指利用傅里叶变换技术将时域数据变换到频域提取的特征，它描述了信号在频率维度上的特性，用来分析信号的频率分布、能量分布等与频率相关的信息。常用的频域特征包括功率谱密度、均方频率和频率方差等。

1. 功率谱密度

功率谱密度是描述信号频域特性的一个重要概念，可以用来反映信号在不同频率上的功率强度。信号 $x(n)$ 在第 k 条谱线上的功率谱定义为

$$P(k)=\frac{\Delta t}{N}\left|\sum_{n=0}^{N-1}x(n)\mathrm{e}^{-\mathrm{j}2\pi kn/N}\right|^2 \tag{3-1}$$

其中，Δt 为采样间隔，N 为信号的长度，$x(n)$ 代表时序信号。

2. 均方频率

均方频率是均方根频率的平方，其描述了功率谱重心位置的变化。具体计算公式为

$$\mathrm{MSF}=\frac{1}{4\pi^2\Delta f^2}\frac{\sum_{k=1}^{K}f_k^2S(k)}{\sum_{k=1}^{K}S(k)} \tag{3-2}$$

其中，Δf 表示采样频率，$S(k)$ 是信号幅值谱第 k 条谱线对应的功率谱幅值，f_k 是第 k 条谱线对应的频率，K 是谱线的个数。

3. 频率方差

频率方差是描述功率谱能量分布的分散程度的特征量。具体计算公式为

$$VF=\frac{1}{4\pi^2\Delta f^2}\frac{\sum_{k=1}^{K}S(k)(f_k-S_f)^2}{\sum_{k=1}^{K}S(k)} \tag{3-3}$$

其中，S_f 是所有谱线对应幅值的均值。

3.2.2　时频域特征

时频域特征是描述信号在时域和频域上特性的一类特征，它同时提供了信号在时间和频率上的变化信息，可以用于分析信号的频率成分变化和瞬态特性等。常用的时频特征变换方法有短时傅里叶变换（Short-Time Fourier Transform，STFT）、小波变换（Wavelet Transform，WT）和 Wigner-Ville 分布等。

1. 短时傅里叶变换

短时傅里叶变换定义了一个非常有用的时间和频率分布类，指定了任意信号随时间和频率变化的复数幅度。离散短时傅里叶变换定义如下：

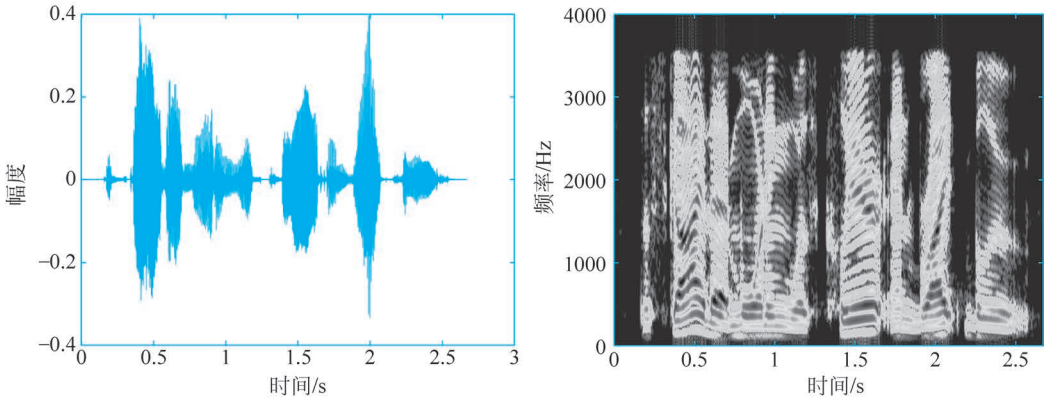
$$X(n,k)=\sum_{m=n-(N_w-1)}^{\infty}x(m)w(n-m)\mathrm{e}^{-\mathrm{j}2\pi mk/N} \tag{3-4}$$

其中，$X(n,k)$ 是与时间和频率相关的函数且是离散的，变量 n 表示时间索引，变量 k 是频率索引，有时也称频率点。$x(m)$ 表示输入信号，$w(n-m)$ 表示窗函数 $w(m)$ 在时间上翻转且有 n 个样本的偏移量，N_w 表示窗的长度。

例 3.1　考虑一个如图 3-1(a)所示的语音信号，采用短时傅里叶变换对其进行时频分析，其对应的时频图如图 3-1(b)所示。从图中可以看出，信号的频率范围主要分布在 90～3500Hz，其中颜色较深的地方展示了语音频率随时间变化的信息。

2. 小波变换

同短时傅里叶变换一样，小波变换是一种对信号进行时间－频率分析的方法，在时序信号中有着广泛的应用。离散小波变换的计算公式如下：



(a) 时域信号　　(b) 时频图

图 3-1　语音信号的短时傅里叶变换

$$W(m,n)=\frac{1}{\sqrt{a_0^m}}\sum_k x(k)\psi^*\left(\frac{n-ka_0^m}{a_0^m}\right) \tag{3-5}$$

式中，a_0 是伸缩步长，m 为尺度参数，k 为沿时间轴的平移参数，$\psi^*(x)$为小波基函数的共轭函数。

3. Wigner-Ville 分布

Wigner-Ville 分布通过计算信号的自相关函数在时间延迟和频率偏移上的傅里叶变换来获得。具体计算公式为

$$W(m,n)=\frac{1}{N}\sum_{k=0}^{N-1}x(kT)x^*((n-k)T)\mathrm{e}^{-\mathrm{j}\frac{\pi m(2k-n)}{N}} \tag{3-6}$$

式中，T 为采样周期。

需要指出的是，虽然以上方法都能用于进行时频域特征提取，但各自方法有其适用特点和缺陷。例如，短时傅里叶变换计算公式简单，相对容易实现，但其时频分析窗口不具有自适应性，无法同时获得高的时间分辨率和频率分辨率，此外由于其需要反复进行傅里叶变换，因此计算量较大；小波变换能提供多尺度的时频分析能力，但其要求时频分析窗是平行分割与等面积的，并且小波基函数的选取需要一定的先验知识，与短时傅里叶变换、Wigner-Ville 分布相比具有更好的时频分辨率，但在时频分析中会产生严重的交叉干扰项，影响时频分布解释。

3.3　文本数据表征

传统的文本数据采集往往依赖于纸质媒介，存在体量小、获取成本高以及时间相对滞后等问题。通过互联网媒介获取文本数据，可以有效从海量文本数据中提取有价值的信息，其中自然语言处理技术是进行文本处理和分析的关键技术。本节介绍与自然语言处理相关的概念及重要技术。

3.3.1　词袋模型

词袋(Bag-of-Words，BOW)模型是文本特征表示的基本方法，其目的是将文本转换为数值型向量，以便于计算机进行处理和分析。在词袋模型中，通常不考虑词语在句子中的顺序和语境关系，而是将文本看成一个由相互独立的词语组成的集合，然后通过计数的方式统计各单词在文本中出现的次数，并将其以向量的形式进行表示。如果词汇表中的某个单词没有出现在文档中，那么计数就为 0。具体来说，词袋模型包含以下几步：

第 1 步，分词。将文本按照一定的规则或算法划分成一系列由词语组成的词序列。

第 2 步，构建词表。将划分的词序列构建成一个词表，其中每个词语对应唯一的索引。

第 3 步，计算词频。统计每个词语在文本中出现的频次。

第 4 步，向量化。根据词表和词频，将文本表示成一个向量，称为词向量，其中向量的每个维度对应词表中的词语，维度对应的值表示文本中词语出现的次数。

例如，有如下两个文本：

(1) I love to eat bananas.

(2) Bananas are tasty.

图 3-2 给出了以上两个文本的词向量表示。因此，文本 1 对应的向量是[1,1,1,1,1,0,0]，文本 2 对应的向量是[0,0,0,0,1,1,1]。

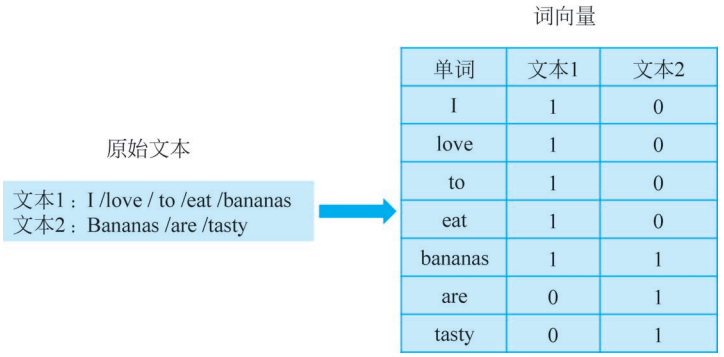


图 3-2　原始文本的词向量表示

3.3.2　TF-IDF 特征

BOW 模型操作简单，但其没有考虑单词之间的顺序，此外也无法反映一个句子中的关键词信息。例如，文本“Jack likes apples, Lily likes too”。若采用 BOW 模型，它的词表为['Jack','likes','apples','Lily','too']，对应的词向量为[1,2,1,1,1]，所以提取的关键词为“likes”。很显然，与文本所要表达的关键信息“apples”相悖。

针对以上问题，词频-逆向文件频率(Term Frequency-Inverse Document Frequency，TF-IDF)被提出，其由词频(Term Frequency，TF)和逆向文件频率(Inverse Document Frequency，IDF)两部分组成。通过 TF 来表示单词在文本中的重要性，同时引入 IDF 对文本中出现频次较高但又不含有实际意义的单词进行处理。

TF 描述了某个词在文档中出现的频率，其计算公式如下：

$$\mathrm{TF}(w)=\frac{n_w}{N_p} \tag{3-7}$$

式中，n_w 表示单词 w 在某个文档中出现的次数，N_p 表示该文档中单词的总数。

IDF 描述了某个词语出现在其他文档中的频率，用于衡量一个词语的普遍重要性。如果包含某个词条的文档越少，那么它的 IDF 值就越大，表示该词对于区分文档的重要性较高；反之，则不然。IDF 的具体计算公式如下：

$$\mathrm{IDF}(w)=\log\frac{N_z}{1+N_w} \tag{3-8}$$

其中，N_z 表示语料库中文档的总数，N_w 表示包含词 w 的文档数。

基于以上概念，TF-IDF 的计算公式如下：

$$\text{TF-IDF} = \text{TF}(w) \times \text{IDF}(w) \tag{3-9}$$

其中，TF-IDF 值越大表示该词越重要，即该词可以被认为是关键词。

3.3.3 词向量嵌入

词向量嵌入是自然语言处理中的一项关键技术，其作用是将文本数据转换成数字型的向量，使计算机能更好地理解和处理自然语言数据。其中，Word2Vec 是一种用于生成词向量的模型，已成为自然语言处理领域中的标志性工具。该模型通过神经网络构建词向量，将单词通过实数向量进行表示，并且可以学习到单词之间的语义和语法信息。Word2Vec 提供了两种算法模型：连续词袋(Continuous Bag of Words，CBOW)模型和 Skip-Gram 模型。

1. CBOW 模型

CBOW 模型的基本思想是通过上下文词语预测当前目标词语，其模型结构如图 3-3 所示。从图中可以看出，该模型由输入层、投影层和输出层组成。输入层是预测目标词 $w(t)$ 上下文对应的 One-Hot 编码表示，输出层节点对应每个词语的预测概率。

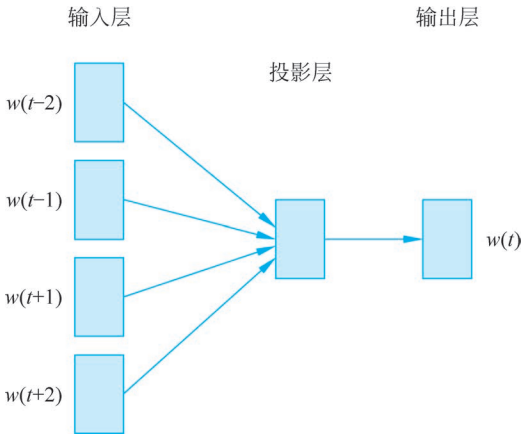


图 3-3 CBOW 模型结构

2. Skip-Gram 模型

与 CBOW 模型相反，Skip-Gram 模型的基本思想是通过当前目标词语预测上下文词语信息，其模型结构如图 3-4 所示。该模型同样由输入层、投影层和输出层组成。输入是当前目标词 $w(t)$ 对应的 One-Hot 编码表示，输出层节点对应上下文词语的预测概率。

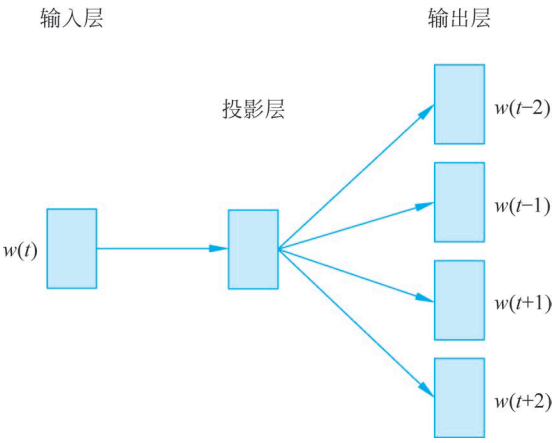


图 3-4 Skip-Gram 模型结构

3.4　图像数据表征

随着物联网技术的发展，越来越多的图像数据被生成，形成了具有大规模性、高维度性以及多样性等特点的图像数据资源。处理图像数据，需要借助图像处理技术来提取、分析和理解图像中的信息。当前，图像处理技术已在诸如计算机视觉、智能交通和智能制造等领域发挥着重要作用。本节将对图像处理的基本概念及图像数据表征的相关技术进行介绍。

3.4.1　图像处理基础

1. 采样与量化

为了生成数字图像，需要将通过传感设备采集的连续信号转换成数字信号的形式，其中采样和量化是两个重要的处理过程。

采样是指将空间上连续的图像信号转换成离散采样点集合的一种操作，以便于数字图像的存储、处理和传输。其中，图像采样分别沿着水平和垂直方向进行，得到的二维离散信号最小单位称为像素。例如，对于一幅图像进行采样，若每行像素个数为 M，每列像素个数为 N，则图像大小为 $M\times N$ 像素，从而可以构成一个 $M\times N$ 的实数矩阵。一般情况下，两个方向的采样间隔是相同的，而采样间隔由采样频率决定。在实际进行图像采样的过程中，采样频率越大，其对应的采样间隔越小，丢失的信息越小，采集的图像质量也就越高；反之，则不然。图 3-5 展示了采样的示意图。

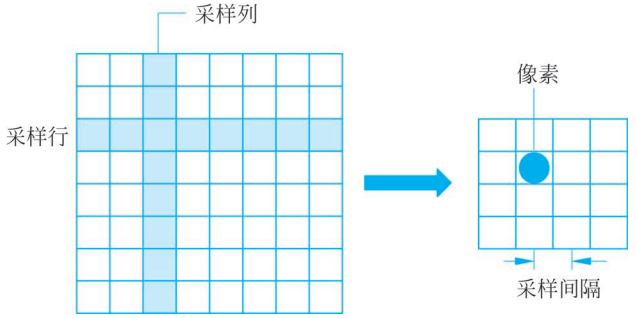


图 3-5　采样示意图

量化是指将各个像素所包含的明暗信息进行离散并以数值形式表示的过程。量化后，数字图像就可以用整数矩阵的形式来描述，其中每个像素包含位置和灰度两个属性。位置由行和列来决定，灰度表示该像素位置上的明暗程度，通常量化为一整数。灰度级别是灰度的取值范围，一般设置为 0～255，分别描述从黑到白。图 3-6 展示了量化的示意图。

2. 数值描述

数值描述是指通过数值的形式来描述一幅图像。如前面所述，图像经过采样和量化后，可以通过二维矩阵来描述图像，其中矩阵元素位置 (i,j) 对应数字图像上像素点的位置，矩阵元素的值 $I(i,j)$ 对应像素点上的像素值。图 3-7 展示了图像的像素坐标系。如果每个像素值只在 0 或 1 之间取值，其对应的是二值图；如果采用 8bit 来存储每个像素值，且像素取值范围在 0～255，其对应灰度图；如果每个像素值采用红、绿、蓝三个分量表示且每个分量用一个 0～255 的整数表示图像的颜色深度，可以表达不同的颜色，此时对应的是彩色图像。

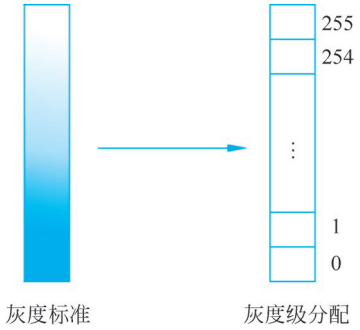


图 3-6　量化示意图

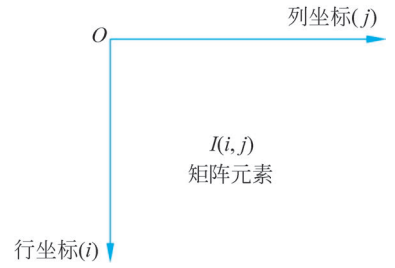


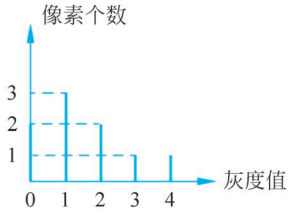
图 3-7　像素坐标系

3. 灰度直方图

灰度直方图是用于描述一幅图像灰度分布情况的统计图表，通过统计具有相同灰度值的像素个数，展示了图像中各灰度级别的像素比例。通过绘制灰度直方图，有助于了解图像的亮度特征，可用于图像增强、分割和质量评估等多方面。例如，图 3-8(a)为图像的灰度图，其对应的灰度直方图如图 3-8(b)所示。

1	4	3
2	1	0
2	0	1

(a) 灰度图



(b) 灰度直方图

图 3-8　灰度直方图

4. 图像增强

图像增强是指为了适应特定的应用需求，通过一定的技术手段对图像进行处理，突出图像中重要的目标信息，抑制不必要的细节信息，从而改善图像的质量，使处理后的图像更符合人眼的视觉特性和易于机器识别。常用的几类典型图像增强方法包括直方图均衡化方法、小波变换方法、偏微分方程算法、基于 Retinex 理论的方法以及基于深度学习的方法。接下来介绍最常用的直方图均衡化方法。

直方图均衡化是最基础的一类图像增强方法，其基本思想是将原始图像的灰度做某种映射变换，使变换后的图像灰度的概率密度呈均匀分布，从而增强图像整体对比度，增大像素灰度值的动态范围。

图像的灰度直方图描述了图像中的灰度概率分布情况，因此一幅图像中灰度级为 r_k 出现的概率 $p_r(r_k)$ 计算公式为

$$p_r(r_k)=\frac{n_k}{N},k=0,1,\cdots,L-1 \tag{3-10}$$

其中，n_k 是灰度级为 r_k 的像素个数，N 为图像中所有的像素个数，L 为图像总的灰度级数。

在直方图均衡化中，常采用原始图的累计概率分布作为映射函数。因此，对于离散的灰度级，映射函数为

$$s_k=T(r_k)=\sum_{i=0}^{k}p_r(r_i)=\sum_{i=0}^{k}\frac{n_i}{N},k=0,1,\cdots,L-1 \tag{3-11}$$

式中，s_k 表示变换之后的值。

综上所述，直方图均衡化方法仅需式(3-11)就可完成对原始图像的直方图均衡化处理，使原始图像的灰度范围扩大和对比度增强。

5. 图像变换

图像变换是指通过数学方法来改变图像的某些特性，是图像处理中的一项基本技术。在实际应用中，通过图像变换可以有效提取图像的特征，提高图像的质量，并服务于特定的图像处理任务，如图像特征提取以及图像降噪等。最常用的图像变换是图像几何变换，即对图像进行平移、旋转、缩放等，通过这些变换可以改变图像的大小和位置。接下来介绍几种常用的图像变换。

1) 平移变换

图像平移是指将图像中的所有点按指定的平移量进行水平或垂直移动。假设平移前的像素点坐标为(x,y)，经过平移量$(\Delta x,\Delta y)$，平移后的坐标为(x',y')，则有

$$\begin{cases} x' = x + \Delta x \\ y' = y + \Delta y \end{cases} \tag{3-12}$$

上式可进一步写成如下矩阵形式：

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \Delta x \\ 0 & 1 & \Delta y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3-13}$$

2) 旋转变换

旋转变换是指将图像按照某个点为中心点进行旋转，使得图像围绕这个中心点旋转一定的角度。假设旋转前的像素点坐标为(x,y)，图像绕任意中心点(x_r,y_r)旋转角度θ，旋转后的像素点坐标为(x',y')，则有

$$\begin{cases} x' = x_r + (x - x_r)\cos\theta - (y - y_r)\sin\theta \\ y' = y_r + (y - y_r)\cos\theta + (x - x_r)\sin\theta \end{cases} \tag{3-14}$$

上式可进一步写成如下矩阵形式：

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_r \\ 0 & 1 & y_r \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & -\sin\theta & 0 \\ \sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -x_r \\ 0 & 1 & -y_r \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3-15}$$

3) 缩放变换

缩放变换是指对图像大小进行调整的过程。假设缩放前的像素点坐标为(x,y)，图像按(s_x,s_y)进行缩放且缩放的中心点为(x_f,y_f)，缩放后的像素点坐标为(x',y')，则有

$$\begin{cases} x' = x_f + (x - x_f) \times s_x \\ y' = y_f + (y - y_f) \times s_y \end{cases} \tag{3-16}$$

上式可进一步写成如下矩阵形式：

$$\begin{bmatrix} x' \\ y' \\ 1 \end{bmatrix} = \begin{bmatrix} 1 & 0 & x_f \\ 0 & 1 & y_f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} s_x & 0 & 0 \\ 0 & s_y & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -x_f \\ 0 & 1 & -y_f \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \tag{3-17}$$

3.4.2　SIFT

SIFT(Scale Invariant Feature Transform，尺度不变特征变换)于1999年提出，并在2004年被完善，是用于检测和描述图像局部特征的一种算法，具有尺度不变性、旋转不变性和亮度

不变性等特点，在计算机视觉领域有着广泛应用。其特征提取包括以下步骤：尺度空间极值检测、关键点定位、关键点方向分配和生成特征描述子。利用 SIFT 提取图像特征的流程图如图 3-9 所示。

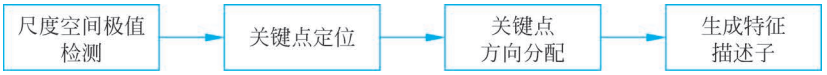


图 3-9　利用 SIFT 提取图像特征的流程图

利用 SIFT 提取图像特征的具体实施步骤如下：

1. 尺度空间极值检测

图像的尺度空间是指同一图像在不同尺度上的集合，其中尺度是指图像的模糊度。在 SIFT 算法中，通过构建图像的高斯差分金字塔，搜索所有尺度空间上的图像特征点，可以检测图像在不同尺度空间中的稳定特征点。高斯差分金字塔通过计算两个相邻尺度空间的图像表示来实现，高斯差分函数 $D(x,y,\sigma)$ 定义为

$$D(x,y,\sigma)=[G(x,y,k\sigma)-G(x,y,\sigma)]*I(x,y)=L(x,y,k\sigma)-L(x,y,\sigma) \tag{3-18}$$

式中，k 表示相邻尺度空间的因子差，$L(x,y,\sigma)=G(x,y,\sigma)*I(x,y)$ 为高斯核与原始图像 $I(x,y)$ 的卷积，即图像的尺度空间。高斯核 $G(x,y,\sigma)$ 定义为

$$G(x,y,\sigma)=\frac{1}{2\pi\sigma^2}e^{-(x^2+y^2)/2\sigma^2} \tag{3-19}$$

式中，σ 为尺度因子，决定了图像模糊的程度。

基于以上概念，进行极值点检测。在利用高斯差分函数构造的高斯差分金字塔中，将每个像素点与其所在尺度同层的 8 个邻域点和上下两层的 18 个点进行比较，从而得到极大值点和极小值点，即候选关键点。

2. 关键点定位

上一步生成的众多候选关键点，其中有一些处于边缘部位，还有一些对比度较低。为此，需要对这些候选关键点进行筛选才能获得准确稳定的关键点。具体操作包括如下两个过程：

(1) 消除对比度低的不稳定极值点。其思想是通过高斯差分函数 $D(x,y,\sigma)$ 在尺度空间的泰勒函数展开式进行拟合来对关键点的坐标准确定位，公式如下：

$$D(\boldsymbol{x})\approx D(\boldsymbol{x}_0)+\nabla D(\boldsymbol{x}_0)^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{x}_0)+\frac{1}{2}(\boldsymbol{x}-\boldsymbol{x}_0)^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x}_0)(\boldsymbol{x}-\boldsymbol{x}_0) \tag{3-20}$$

式中，$\boldsymbol{x}=(x,y,\sigma)^{\mathrm{T}}$ 为拟合后关键点的坐标，$\nabla D(\boldsymbol{x}_0)^{\mathrm{T}}=\left(\frac{\partial D}{\partial x},\frac{\partial D}{\partial y},\frac{\partial D}{\partial \sigma}\right)_{\boldsymbol{x}=\boldsymbol{x}_0}$，$\boldsymbol{x}_0$ 为初始位置点坐标，$\boldsymbol{H}(\boldsymbol{x}_0)$ 为 $D(\boldsymbol{x})$ 在 $\boldsymbol{x}_0=(x,y,\sigma)^{\mathrm{T}}$ 处的 Hessian 矩阵，表示为

$$\boldsymbol{H}(\boldsymbol{x}_0)=\begin{bmatrix}\dfrac{\partial^2 D}{\partial x^2} & \dfrac{\partial^2 D}{\partial x\partial y} & \dfrac{\partial^2 D}{\partial x\partial \sigma}\\ \dfrac{\partial^2 D}{\partial y\partial x} & \dfrac{\partial^2 D}{\partial y^2} & \dfrac{\partial^2 D}{\partial y\partial \sigma}\\ \dfrac{\partial^2 D}{\partial \sigma\partial x} & \dfrac{\partial^2 D}{\partial \sigma\partial y} & \dfrac{\partial^2 D}{\partial \sigma^2}\end{bmatrix}_{\boldsymbol{x}=\boldsymbol{x}_0}$$

对式(3.20)求导并令其为 0，可得局部极值点

$$\hat{\boldsymbol{x}}=\boldsymbol{x}_0-\boldsymbol{H}(\boldsymbol{x}_0)^{-1}\nabla D(\boldsymbol{x}_0) \tag{3-21}$$

将上式代入公式(3.20),可得

$$D(\hat{\boldsymbol{x}})=D(\boldsymbol{x}_0)-\frac{1}{2}\nabla D(\boldsymbol{x}_0)^{\mathrm{T}}\boldsymbol{H}(\boldsymbol{x}_0)^{-1}\nabla D(\boldsymbol{x}_0) \tag{3-22}$$

如果$|D(\hat{\boldsymbol{x}})|$小于设定的阈值 ,则认为该极值点为低对比度的不稳定点,将其进行剔除,反之则将极值点保留用于下一步判断。

(2) 消除边界上的不稳定极值点。当一个极值点位于边缘位置时,其对应的主曲率一般比较高,依据该特性可以消除边界不稳定极值点。主曲率可通过引入如下矩阵:

$$\bar{\boldsymbol{H}}=\begin{bmatrix}\dfrac{\partial^2 D}{\partial x^2} & \dfrac{\partial^2 D}{\partial x\partial y}\\ \dfrac{\partial^2 D}{\partial y\partial x} & \dfrac{\partial^2 D}{\partial y^2}\end{bmatrix}_{(x,y)=(x_0,y_0)} \tag{3-23}$$

得到。设α和β分别是矩阵$\bar{\boldsymbol{H}}$的最大和最小特征值,则有

$$\mathrm{Tr}(\bar{\boldsymbol{H}})=\alpha+\beta \tag{3-24}$$

$$\mathrm{Det}(\bar{\boldsymbol{H}})=\alpha\beta \tag{3-25}$$

其中,Tr(·)表示矩阵的迹,Det(·)表示矩阵行列式,它们的比值可以代表主曲率的变化。

令$\alpha=r\beta$,可以得到

$$\frac{\mathrm{Tr}(\bar{\boldsymbol{H}})^2}{\mathrm{Det}(\bar{\boldsymbol{H}})}=\frac{(\alpha+\beta)^2}{\alpha\beta}=\frac{(r+1)^2}{r} \tag{3-26}$$

从上式可以看出,式(3-26)的值是随r单调递增的,其值越大说明两个特征值的比值越大,正好符合边缘的情况。因此,为了消除边界不稳定点,只需让上式小于一定的阈值,即

$$\frac{\mathrm{Tr}(\bar{\boldsymbol{H}})^2}{\mathrm{Det}(\bar{\boldsymbol{H}})}<\frac{(r+1)^2}{r} \tag{3-27}$$

式中,r为自定义的比例系数,一般取值为10。

3. 关键点方向分配

经过以上步骤后,可以完全确定图像所有的关键点,并使这些特征点具有尺度不变性。接下来,为保证关键点对图像的旋转不变性,需要为每个关键点附加方向。关键点的方向可以通过梯度方向直方图来进行求解,其基本思想是以每个关键点为中心,计算其像素梯度的幅值和方向角度:

$$m(x,y)=\sqrt{[L(x+1,y)-L(x-1,y)]^2+[L(x,y+1)-L(x,y-1)]^2} \tag{3-28}$$

$$\theta(x,y)=\arctan\left(\frac{L(x,y+1)-L(x,y-1)}{L(x+1,y)-L(x-1,y)}\right) \tag{3-29}$$

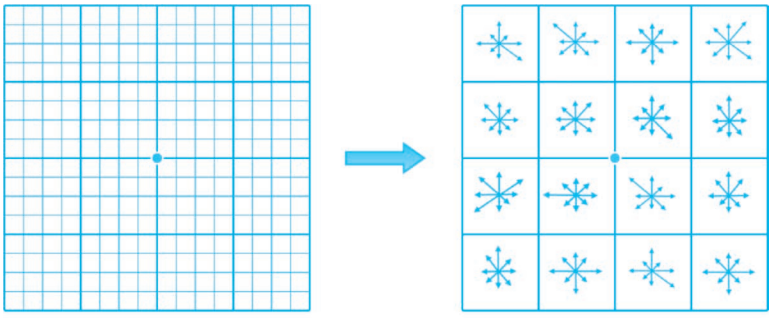
式中,$m(x,y)$为幅值,$\theta(x,y)$为方向角,L所用尺度为每个关键点各自所在的尺度。

获得关键点梯度后,以关键点为中心,采用直方图统计邻域的梯度和方向,将直方图峰值处所在方向作为关键点的主方向。至此,每个关键点都具有位置、尺度和方向信息。下一步就是根据这些信息通过一个向量来唯一表示关键点。

4. 生成特征描述子

为了进一步满足图像匹配的任务,需要为每个特征点进行描述,即创建特征描述子。其基本思想是以检测得到的关键点(如图3-10中的点所示)为中心选取16×16像素的邻域窗口,将其划分成16个子区域(每个子区域大小为4×4像素),然后对每个子区域做8个方向的梯

度幅值和方向统计，即可得到 4×4×8(128)维的特征向量，以此作为关键点的数学描述。其中，16 个子区域生成特征描述子的过程如图 3-10 所示。



(a) 关键点16×16像素邻域窗口　　(b) 特征描述子

图 3-10　特征描述子生成过程

3.4.3　HOG

HOG(Histogram of Oriented Gradients，方向梯度直方图)于 2005 年提出，是一种基于图像形状边缘的特征提取算法，具有速度快、准确率高等特点，被广泛应用于人脸识别、行人检测和目标识别等领域。其基本思想是通过计算图像的梯度并统计图像局部区域内的梯度方向分布信息来描述图像特征。图 3-11 给出了利用 HOG 提取特征的流程图。

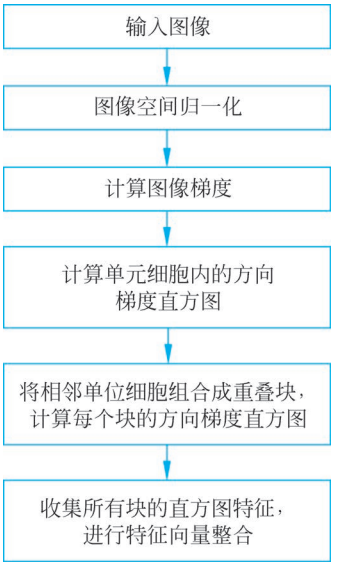


图 3-11　利用 HOG 提取特征的流程图

HOG 算法特征提取的步骤如下：

(1) 图像空间归一化。对输入图像进行灰度化处理，采用 Gamma 校正方法对图像进行处理以降低光照不均匀的干扰。Gamma 校正公式如下：

$$I_0(x,y)=I(x,y)^{\gamma} \tag{3-30}$$

其中，$I(x,y)$表示原始图像某个像素点的灰度值，$I_0(x,y)$表示校正后的灰度值，γ 为校正系数。当 $\gamma<1$ 时，图像整体灰度变亮；当 $\gamma>1$ 时，图像整体灰度变暗。通常 γ 取值为$\frac{1}{2}$。

(2) 计算图像梯度。对上述归一化后的图像，计算其每个像素点的梯度幅度和角度，计算公式如下：

$$G_f(x,y)=\sqrt{G_x(x,y)^2+G_y(x,y)^2} \tag{3-31}$$

$$\theta(x,y)=\arctan\left(\frac{G_y(x,y)}{G_x(x,y)}\right) \tag{3-32}$$

其中，$G_x(x,y)=I_0(x+1,y)-I_0(x-1,y)$和 $G_y(x,y)=I_0(x,y+1)-I_0(x,y-1)$分别表示像素点$(x,y)$在水平方向和垂直方向上的梯度。

(3) 计算单元细胞内的方向梯度直方图。将整幅图像划分成若干个同等大小的单元细胞，计算每个单元细胞的梯度信息，具体过程如下：将 360°的角平均划分为 9 份，然后根据每个像素点的方向梯度找到对应的组距(bin)，将每一个单元细胞的幅值按梯度方向对应区域进行累加，统计每一个单元细胞的 bin，最后形成每一个单元细胞的 HOG 特征。

(4) 计算每个块(block)的方向梯度直方图。将单元细胞有重叠地组成 block,把每一个 block 内的所有单元细胞的 HOG 特征级联,得到该 block 的方向梯度直方图。

(5) 特征向量整合。将所有 block 的方向梯度直方图串联起来获得整幅图像的 HOG 特征向量。

3.4.4　深度特征表示

以上传统的图像特征提取算法适用范围有限,对不同类型的图像自适应性较差。相较于以上方法,深度特征表示使用深度学习模型自动提取图像的特征,能够更好地捕捉到图像的语义信息。近年来,传统的图像特征提取算法逐渐被深度学习的方法所取代,并在图像匹配、目标检测等图像处理领域取得了显著的成果。

图 3-12 展示了基于深度学习的特征表示框架。首先,为深度学习网络模型选择合适的超参数,如网络层数、每层神经元的个数、激活函数和代价函数等,常用的深度网络学习模型包括多层感知机、卷积神经网络和深度置信网络;其次,应用选择出的超参数构建深度特征网络;最后,将原始图像数据输入深度学习网络模型进行训练,并将获取的结果与停止准则进行比较。如果训练的结果满足停止准则,则获得训练好的深度特征网络;如果训练的结果不满足停止准则,则进行超参数调优。

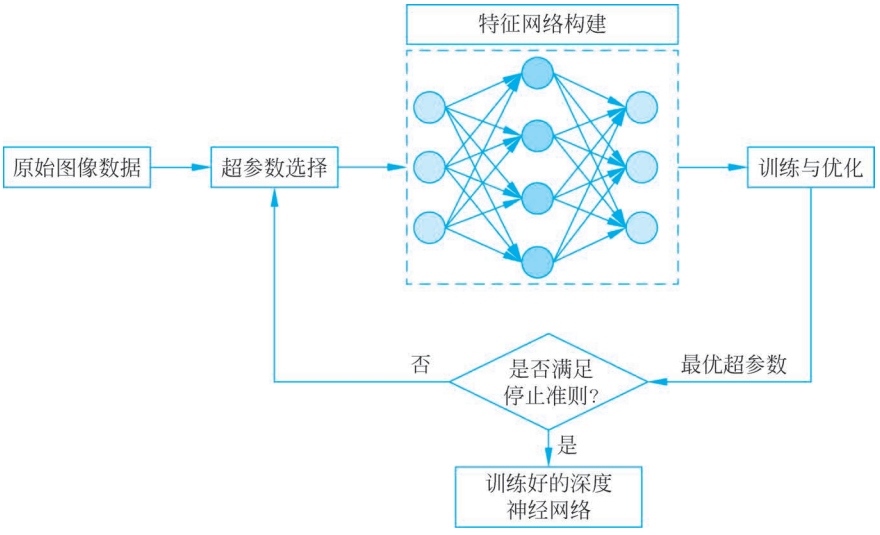


图 3-12　基于深度学习的特征表示框架

3.4.5　多模态特征融合

多模态图像特征是指从不同模态图像数据,如可见光图像、红外光图像等图像数据中提取的信息。多模态特征融合是指将来自不同模态的图像特征信息进行整合或融合的一项技术。相比于单一模态特征表示,多模态特征融合可以提供更为丰富的特征表示,让模型能够理解和处理更为复杂的问题。与多模态数据融合不同,多模态特征融合关注的是将不同模态特征进行融合,而不是对不同模态数据(如文本、音频、视频等)的整合。

多模态特征融合方法可分为模型无关的方法和基于模型的方法,其中模型无关的方法又可分为早期融合、晚期融合和混合融合。图 3-13 展示了三种模型无关的特征融合方法。早期融合,也称特征融合,它是在特征提取后立即将不同模态特征表示进行融合;晚期融合,也称决策级融合,它是先对不同模态进行训练,再融合多个模型输出;混合融合综合了早期融合和

晚期融合两者的优点，可以提高模型的性能，但增加了模型的复杂度训练难度。

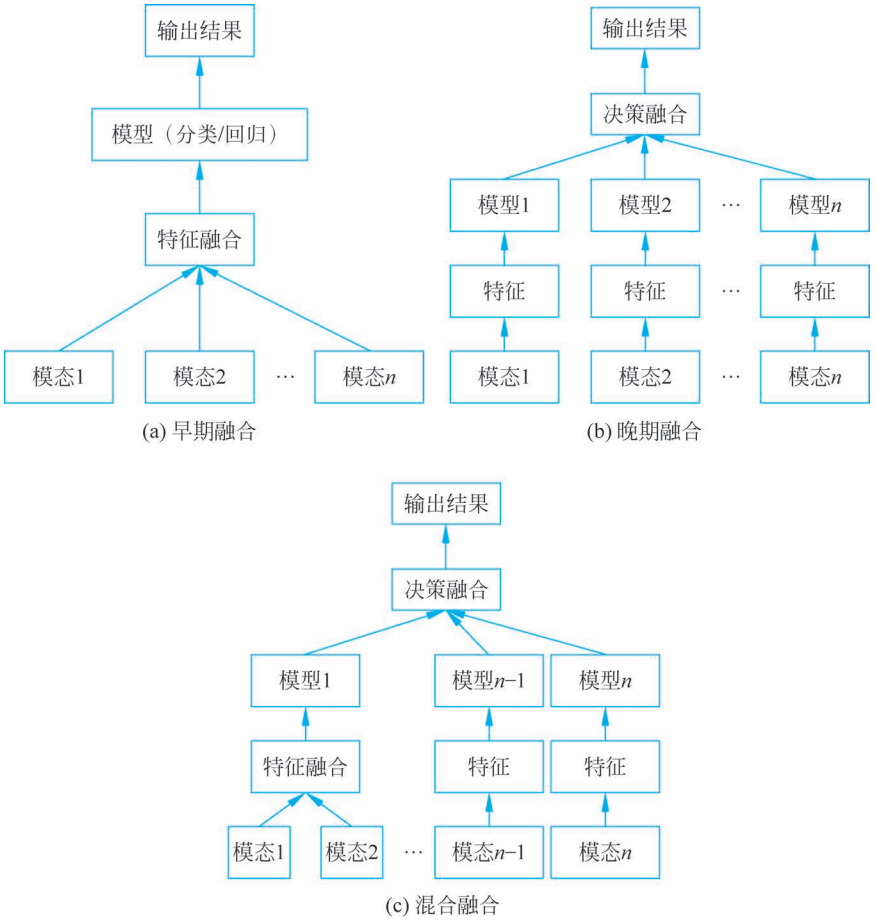


图 3-13　模型无关的多模态特征融合

基于模型的方法是从实现技术角度来解决多模态融合问题的，常用方法有神经网络方法、图像模型方法和多核学习方法。随着深度学习的发展，目前基于神经网络的方法是应用最广泛的一类方法。图 3-14 展示了基于卷积神经网络的多模态特征融合。其基本思想是首先采

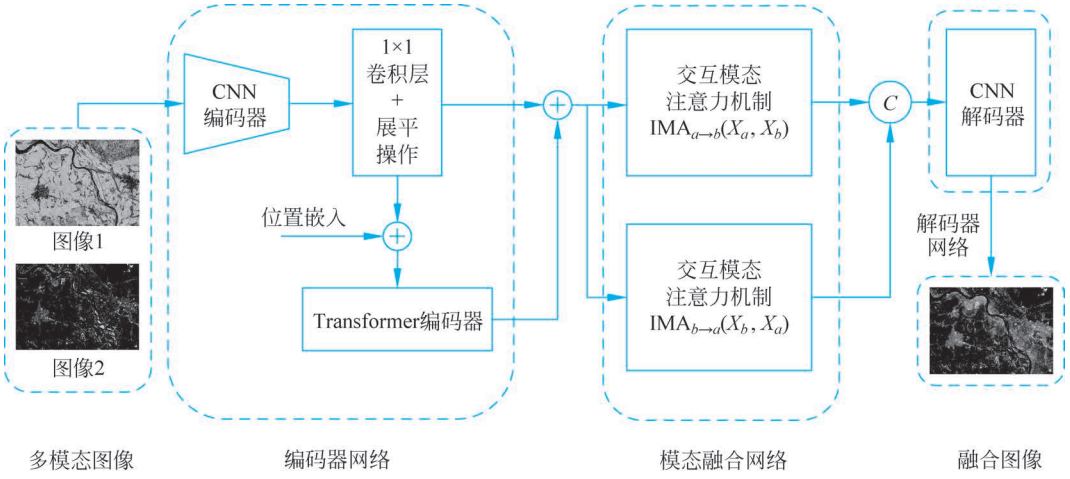


图 3-14　基于卷积神经网络的多模态特征融合

用 CNN 与 Transformer 作为编码器提取图像的特征，并将其作为编码器的输出；其次将提取的多模态特征输入到模态特征融合网络进行特征融合并对输出的特征进行拼接；最后，将融合的特征经 CNN 解码器重构出融合的图像。相比于无模型方法，基于模型的神经网络方法具有更好的特征学习能力且扩展性好，但存在模型解释性差的问题。

3.5　案例：农作物病虫害图像表征

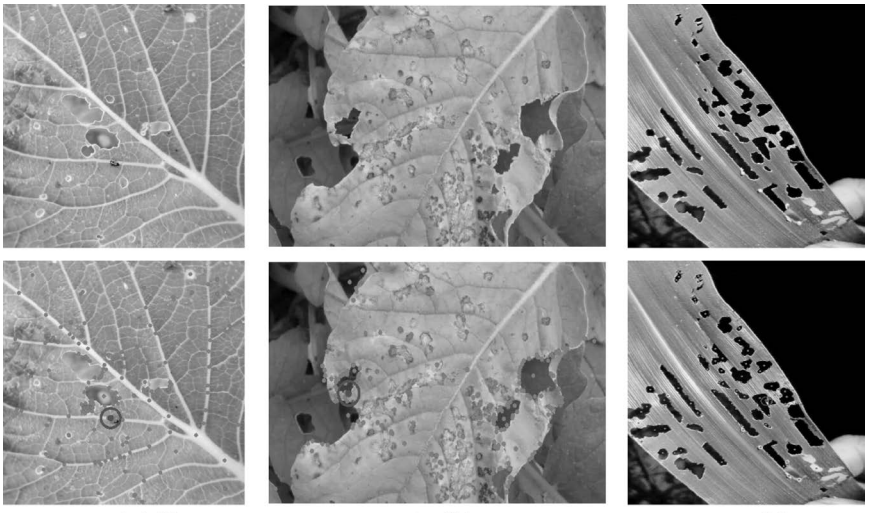
农作物病虫害检测是指利用一定的技术手段和设备检测农作物是否存在病害或虫害的过程，旨在早期发现和诊断农作物病虫害，以便及时采取有效的措施进行控制，从而减少农作物损失。因此，在农业生产管理过程中，非常有必要对农作物的病虫害进行有效检测，其中病虫害图像特征提取是进行有效检测的重要环节之一。本节将简要介绍利用 SIFT 算法提取图像特征的关键过程。

如前所述，SIFT 算法涉及步骤较多。从原始图像开始，首先通过构建图像高斯差分金字塔在高斯差分尺度空间寻找极值点作为候选关键点；其次利用式(3-20)定位关键点的位置及尺度；然后利用关键点邻域像素的梯度方向分布为每个关键点分配方向参数；最后以关键点为中心取 16×16 像素的窗口，将其划分为 4×4 个子区域，其中每个子区域包括 4×4 像素，在此基础上计算每个子区域 8 个方向的梯度直方图并绘制每个梯度方向的直方图，产生一个 4×4×8 维的特征向量。综合以上求解过程，SIFT 算法是将图像中检测到的特征点用一个特征向量进行描述。因此，一幅图像经过 SIFT 算法处理后，可得到若干个由 1×128 维的特征向量构成的集合。

利用 SIFT 算法提取图像特征的核心代码描述如下：

```
import cv2 as cv
import numpy as np
import matplotlib.pyplot as plt
# 读取灰度图像
imag1 = cv.imread("cai.jpg")
gray = cv.cvtColor(imag1, cv.COLOR_BGR2GRAY)
# SIFT 实例化
sift = cv.SIFT_create()
# 检测关键点
keypoints, descriptors = sift.detectAndCompute(gray, None)
# 绘制关键点
imag2 = imag1.copy()
cv.drawKeypoints(imag2, kp, imag2, (0, 0, 255))
```

本案例测试图像选取了小白菜、萝卜和玉米三种常见的带病虫害农作物图像进行分析。图 3-15 展示了利用 SIFT 算法提取图像特征的结果，其中上部分为原始图像，下部分为提取特征后的图像。图中红色点为利用 SIFT 算法提取的特征点，并对其中响应值最高的特征点用蓝色圆圈进行标注与特征向量提取。最终，通过 SIFT 算法提取的特征为 128 维的特征向量：小白菜对应的特征向量为[15,16,5,1,…,26,22,1,1]，萝卜对应的特征向量为[1,0,2,10,…,9,0,0,0]，玉米对应的特征向量为[1,13,98,38,…,2,3,50,14]。关于病虫害检测的完整过程，由于还涉及阈值分割、边缘检测以及分类器等相关研究内容，超出了本节的讲解范畴，在此不做进一步介绍。



(a) 小白菜　　(b) 萝卜　　(c) 玉米

图 3-15　基于 SIFT 算法的图像特征提取

3.6　本章小结

随着大数据时代的来临，大量的结构化和非结构化数据存在于各个行业和各个领域，数据种类繁多，如时序数据、文本数据和图像数据等，如何从这些不同类型的数据中提取出有价值的信息，是数据挖掘和机器学习中的重要研究内容。本章分别对时序大数据、文本大数据和图像大数据的常用特征表示方法进行了概括和总结，并通过农作物病虫害图像特征提取的案例阐述了图像特征表示方法在实际生活中的具体应用。

习题

1. 选择题

(1) 以下不属于时序数据表征方法的是(　　)。

A. 主成分分析　　B. 小波变换　　C. 均方根　　D. 傅里叶变换

(2) 在词袋模型中，如何量化词语的重要程度？(　　)

A. 使用 TF-IDF　　B. 使用词性标注　　C. 使用词向量　　D. 以上都是

(3) 关于 Word2Vec 的优缺点，说法正确的是(　　)。

A. 无法处理一词多义问题　　B. 是一种有监督的训练方式

C. 编码的词向量中不包含语义信息　　D. 不确定

(4) 以下哪种方法编码的词向量包含语义信息？(　　)

A. Word2Vec　　B. One-Hot　　C. TF-IDF　　D. Bag-of-Words

(5) 以下不属于文本数据表征的方法是(　　)。

A. MEMM　　B. BOW　　C. N-Gram　　D. TF-IDF

(6) 以下不属于数字图像处理的研究内容是(　　)。

A. 图像数字化　　B. 图像分割　　C. 图像增强　　D. 数字图像存储

（7）图像与灰度直方图间的对应关系是（　　）。

A. 一对多　　B. 多对一　　C. 一一对应　　D. 都不是

（8）图像数字化为什么会丢失信息？（　　）

A. 采样丢失数据　　B. 量化丢失数据

C. 采样和编码丢失数据　　D. 压缩编码丢失数据

（9）将像素灰度转换成离散的整数值的过程叫（　　）。

A. 采样　　B. 量化　　C. 增强　　D. 复原

（10）以下哪项不属于图像特征描述算法？（　　）

A. SIFT　　B. HOG　　C. PCA　　D. CNN

2. 简答及计算题

（1）常用的时序数据表征方法有几类？每种方法有什么特点？

（2）给定一个信号 $y=\sin(4\pi t)\cos(100\pi t)$，其中采样频率为 1024Hz，采样时间为 2s，试根据表 2-5 的计算公式，计算均值、均方根值、峭度值等信号的时域特征。

（3）假设有如下由不同频率叠加组合而成的混合信号：

$$y=\sin(10\pi t)+2.5\sin(40\pi t)+N(t)$$

其中，$N(t)$表示均值为 0、方差为 1 的随机噪声，采样频率 $f_s=100\text{Hz}$，采样时间为 5s。请计算其功率谱密度、均方频率、短时傅里叶变换系数等频域和时频域特征，并绘制出相应的图。

（4）假设有如下 3 个文档：

① Apples are a great source of fiber, which can help improve digestion and overall gut health.

② Apples have a sweet-tart taste, and are known for their juicy texture.

③ Whether eaten raw, cooked, or juiced, apples offer a refreshing and nutritious snack option.

请计算每个单词的 TF-IDF 值。

（5）输入一幅彩色图像，请将其转换为灰度图，绘制其灰度直方图，并找出出现最频繁的灰度值。

（6）假设有一个 4×4 的输入图像块（如图 3-16 所示）和 3×3 的卷积核（如图 3-17 所示），试计算卷积结果。

$$\begin{bmatrix}7 & 3 & 4 & 1\\ 3 & 5 & 3 & 0\\ 2 & 1 & 7 & 1\\ 2 & 0 & 7 & 0\end{bmatrix}$$

图 3-16　图像块

$$\begin{bmatrix}-1 & 0 & 1\\ -2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix}$$

图 3-17　卷积核

3. 思考题

（1）在电商平台上，商品可以通过文本、图片、视频等来反映。请思考如何将这些不同类型的数据进行表征与整合，用于商品的推荐。

（2）通过文献调研，编程实现一种基于深度学习的图像特征提取算法，并与经典的 SIFT 算法进行比较。