第 3 章　基于语义的视觉媒体处理

3.1　特征提取及表示

在多媒体分析领域，特征用于从原始的多媒体数据中获得后续应用的元数据。从多媒体对象（如视频、图像等）获得特征的过程称为特征提取[1]。这个过程通常是自动完成的，因此，可灵活应用于可穿戴式感知这种对效率要求较高的应用中。特征提取常常分为两层特征（即低层特征和高层特征）进行探讨，以反映所提取出的特征与媒体语义相关性的程度。

3.1.1　低层特征

总的来说，低层特征指的是数据出现的模式和统计结果，这些特征相比与对媒体内容的文字描述来说具有更少的含义。由于低层特征提取完全是一种数学计算，因此很容易实现提取过程的计算机自动处理。以文本文档为例，低层特征可以从文档中每个文字出现的频率进一步导出，并去除文档中对表达文档语义没有贡献的停用词，如英文单词中的“the”“a”等。对于其他媒体如视频和图像来说，常用的低层特征包括语音中的平均能量、过零率、静音率等，以及视觉中的颜色、纹理、形状等[1]。

尽管低层特征通常不直接用于检索，更具含义的特征可以建立在低层特征之上，并用于进一步的分析。使用低层特征的优势可以归纳如下。

（1）强表达能力。对比原始的输入数据，低层特征可以更加准确地表示数据各方面的属性。

（2）低存储空间。从原始数据（如图像）中提取出来的低层特征，相比原始图像的像素来说，需要少得多的存储空间。

（3）维度降低。同样以图像为例，由于原始图像的像素阵列通常维度非常高，所提取的特征可以极大降低计算维度。同时，一些处理技术如隐语义分析、主成分分析等也可用于所提取的特征以进一步实现降维。

（4）较少的计算开销。由于维度的降低，采用提取后的低层特征可以使两个特征向量之间的比较更加容易和迅速。

视觉媒体（如图像）是可穿戴式感知的重要语义信息来源，本书将详细介绍图

像表示所需的典型特征。

1. 颜色特征

每幅图像都由特定数量的像素组成。由于每个像素都有一个在颜色范围内的值(对于黑白图像是灰度值),颜色特征可用于描述图像的内容[2],包括:

(1) 颜色直方图。颜色直方图可以反映在离散颜色取值上的像素分布。该直方图通过简单对给定一组颜色范围中具有某个颜色值的像素进行计数来获得。颜色直方图广泛用于通过视觉相似度对图像进行区分。

(2) 可伸缩颜色。可伸缩颜色是另一个在整个图像上对颜色分布进行度量的描述符。在进行计算时,将颜色空间固定为 HSV 空间,并均匀量化为 256 份,其中包括 H 中的 16 个层级,S 中的 4 个层级,以及 V 中的 4 个层级。可伸缩颜色的直方图基于 Haar 变换进行编码,以减少这种表示的大小,同时允许可伸缩的编码[3]。

(3) 颜色布局。类似于颜色直方图和可伸缩颜色,颜色布局也作为 MPEG-7 视觉描述符进行设计,以捕捉一幅图像或者任意形状区域中的颜色分布。它是在 YCbCr 颜色空间中定义的一种紧凑和不变分辨率的颜色描述符。颜色布局使用 8×8 网格,并采用 DCT 变换,将得到的系数进行编码。少量低频系数被采用 Z 字形扫描方式选择,其中亮度的 6 个系数和每个色度的 3 个系数被保留,构成一个十二维的颜色布局矩阵[3-4]。

2. 纹理特征

类似于颜色特征,纹理特征是另一种用于图像检索的低层描述符,并可以自动完成抽取。纹理描述符将一幅图像当成不同纹理区域的马赛克图案[5],与这些区域对应的图像特征被用于图像查找。三种纹理描述符在 MPEG-7 中被采用,包括纹理浏览、均匀纹理和边缘直方图。如本章文献[4]所描述,当一幅图像中存在均匀区域、优势方向等模式时,所有这些描述符都被计算。

3. 形状特征

图像形状通常由一组点的样本所表示,这些点样本从形状轮廓中抽取出来,例如从一个边缘检测结果采样 100 个像素位置。这些表示点的选择并没有特别的需求,也就是说,它们不必是地标点或者曲线的极值等[6]。基于形状的特征,使用形状边界或者整个形状区域用于捕捉图像中的局部几何属性。傅里叶描述符是一个基于边界的形状特征表示,而不变矩使用基于区域的矩,这些矩对于变换来说是恒定的[7]。形状上下文是另一种形状描述符,用于描述对于一个给定点的其余形状点的粗略分布。这种描述符已经在人类动作识别[8]、商标检索[9]等应用中被采纳。采用形状描述,可以将两个形状的比较转换成从两个形状中寻找具有相似形状上下文的样本点。

3.1.2 高层特征

高层特征指的是对终端用户具有语义内含的特征。尽管低层特征对于终端用户来说是不可读的，然而高层特征能够以一种更方便接受的方式将媒体的语义表示为“概念”，例如室内、室外、植被、计算机屏幕等。这些特征可以为低层特征和用户期望提供一个有含义的链接。对高层特征的提取需要连接低层特征与高层特征之间的鸿沟，该鸿沟在多媒体信息检索领域被称为语义鸿沟。

语义概念通常采用数学的方法进行自动的探测，即将低层特征映射到高层特征上。研究人员往往常采用机器学习的方法，如支持向量机，确定给定的抽取出来的特征所对应的最可能的概念[10]。支持向量机是一种判别式的模型，该模型具有更多面向任务的特点，而马尔可夫模型等生成式统计模型用于分析变量的联合概率，这类方法也在概念标注的研究中得到应用[11]。生成式和判别式方法都有各自的优缺点。生成式模型是一个多变量的完全概率模型，而判别式模型则具有有限的建模能力。这是因为判别式模型提供仅仅以观察变量为条件的目标变量的模型，因此很难表达观察变量和目标变量的复杂关系。然而，判别式模型通常更容易学习，并且执行起来也比生成式模型更快。并且研究表明，在大训练数据集（通常包含正样本和负样本）的情况下，判别式分类器常常比生成式分类器获得更好的分类性能。在很多机器学习算法中，支持向量机是一种高效的判别式方法，并且具有很强的理论基础，在诸如手写识别、图像检索和文本分类等任务中具有出色的表现[12]。很多研究表明，支持向量机在概念探测任务中是一个高效的框架[13-14]，并且在本书概念索引任务中也被用作一种基础分类算法。

通过学习进行分类的模型一般包含一个大量标注数据集构成的语料库。由于不可能为所有概念构建探测器，因此需要为能够满足不同应用域的概念构建探测器。一般情况下，多媒体内容检索的解决方案集中于特定的域。例如，LSCOM 概念本体和 MediaMill 的 101 个概念探测器主要解决电视新闻检索领域的问题。在本书中，我们将探讨日常行为分析领域所需要的高层特征，例如 SenseCam 图像中的语义概念。

3.2 基于内容和基于概念的检索

3.2.1 基于内容的检索

由于低层特征可以从媒体对象中自动抽取出来，因此基于这些特征对媒体对象的比较则形成了基于内容的检索。使用基于内容的检索，一个多媒体信息系统可以以隐含的方式处理概念。尽管在基于内容检索中假定低层特征对应于查询语义，然而这种映射并没有被建模。颜色、纹理、形状等特征被广泛应用于基于内容

的检索[2,15-16]。在视频检索中,从语音中提取的文本特征、字幕等[17-23]也被用于和图像特征进行结合并用于内容检索。

3.2.2　基于概念的检索

更多的研究表明了基于内容的检索的局限性,即使用低层特征无法解决语义鸿沟的问题。高层特征的引入,表明了其在填补或者至少减少这种语义鸿沟过程中的优势。采用高层特征的检索称为基于概念的检索。基于概念的检索以一种显式的方式来处理概念,并将用户查询表达为高层概念,而不是低层特征[10]。

进行高层特征识别的方法可以归纳为两种类型,即专用的方法和通用的方法。专用的方法尝试捕捉不同领域中从低层特征到高层特征的直接映射[24-26]。这些方法通常是基于规则的方法,因此对于一个新的概念,则需要开发新的映射规则。如此多的特定方法或者专门方法导致的多样性问题,可以通过采用通用的方法得到解决[27-31]。一系列概念探测器可以在基于概念的检索中学习得到并作为词汇使用,这些词汇还可以通过多用途的词典(如 WordNet)或者特定领域的本体得到丰富。通过采用通用的机器学习范例,尤其是对于那些具有足够标注训练数据的部分概念和相关的任务[32],在研究中得到了比较满意的结果。通过将用户查询准确表示为概念探测器,基于高层概念的检索已经获得比较成功的进展,这种对用户查询进行准确表示的过程称为概念选择[10,33]。

3.2.3　概念选择/查询扩展

查询扩展是在检索过程中常用的一种方法,该方法尤其在文档检索的应用中得到了发展。查询扩展的出发点是通过为查询增加更多的词,从而更明确地表达该查询,以达到提高检索效果的目的。在基于概念的多媒体检索中也存在类似的情况,即通过自动地扩展视频探测中的概念集合,使更多关于相同检索对象的概念被包含在内。这种扩展通常会得到更加精确的查询表示。理论上,查询—概念的映射用于将用户的期望翻译为一组概念集合,这个过程称为概念选择。尽管很多研究表明,通用的方法可以从大量人工标注的语料库中学习概念识别方法,但是构建与人类词汇相当数量级的概念探测器仍然是不现实的。文本方法[10]和基于集合的统计方法[34-35]均可以被用于概念选择,然而更多的研究则倾向于使用本体来选择相关的概念[10,36]。

3.3　以事件为中心的媒体处理

研究人员已经广泛地认识到,事件是人类组织自身记忆的基本单元[37]。对个人照片组织的研究也表明,人们通常根据事件考虑他们自己的照片,例如这些事件

对应于特定的主题(如婚礼、假期、生日等),尽管这些主题可能在人类的意识里只是松散地被定义[38-40]。在现代多媒体处理中,事件通过采用不同的方式被表示出来,例如文本、图像、视频,以及其他一些传感器数据等。尽管如此,目前并没有一种通用的事件模型真正意义上构成一个事件并被不同的领域所接受。这一问题在前期的研究中已经逐渐受到业界的重视。例如,ACM 多媒体事件国际研讨会(EiMM09)从 2009 年开始每年举办一次。

由于人类的日常生活在记忆中作为事件进行组织,因此,事件在可穿戴式感知的研究中起到了非常重要的作用。并且人类还经常以事件的形式对未来的生活进行计划和预见。因此,在可穿戴式感知的研究中需要一个一致的事件模型以及事件为重心的观点,以指导对可穿戴式生活记录进行语义解释和处理。

在传统日记中,我们将有意义或者重要的行为或观点写下来,以用于一段时间之后进行回顾。为了生成一个数字化的日志并反映用户生活的各个方面,主要的事件尤其是最让人感兴趣和不同寻常的事件应该被识别出来,并表示为日记的一部分。采用可穿戴式传感器,由此采集的数据不但可以用来记录每天发生的主要活动,并且可以包含位置、周边的人等事件细节,由事件所得到的图像也使得有效重构人们生活中最重要的事件成为可能。对日常事件和事件边界进行可靠的和准确的识别与理解,有助于在大量行为的数据记录中进行更好的事件管理与检索。

本书采用图 3-1 事件模型和分层结构来表示可穿戴式感知过程中的事件,以及事件所包含的媒体内容。该事件结构包括下面三层。

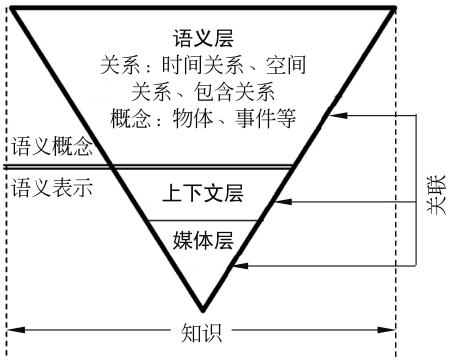


图 3-1 事件模型和分层结构

(1) 语义层。语义层表示了数据本身所蕴含的语义内容。在该层中,如对象、活动等概念语义以及事件主题和关系(如时间空间关系、对等和包含关系等)将被进一步解释,以在更高的层次进行理解。

(2) 上下文层。上下文层包含了能够表示事件不同侧面的上下文。时间空间方面是描述事件的基本物理上下文,即事件是沿着时间和空间轴延伸而开展的。在可穿戴式感知研究中,这两个上下文与感知媒体的时间属性和空间位置紧密相

关。除此之外，事件所涉及的人员和实体以及进一步的信息也需要在上下文层中包含，从而和时空上下文一起回答关于事件的“Who，What，Where，When”问题。

(3) 媒体层。物理的和形式化的内容在媒体层中被媒体所表示，例如像素、传感器数据、编码机制等。尽管语义事件是在任何媒体之外独立存在和发生的，丰富的媒体文档对于用户进行可穿戴式生活记录中事件的浏览也是必不可少的。

上述的媒体层和上下文层强调了可穿戴式生活记录的感知和语法两个方面。然而，语义层则表示了更适合人类对感知内容进行理解的含义方面。在图 3-1 中，由水平范围所表示的知识在整个事件模型中覆盖了所有的三个层次，以表示在每一层中知识内容和推理的重要性。从图上可以注意到，知识的数量在从上到下的过程中是逐渐减少的。模型顶部的语义层包含了更丰富的知识，这可以从图 3-1 中更宽的知识跨度反映出来。这也意味着语义层包含更多可以从隐含内容中得到的概念和关系，这些知识在对用户情况理解的过程中比上下文信息更具有决定作用。这三个层次相互关联，从而可以为生成一种数字化的事件生活记录，满足结构和经验上的需要。从上面可以看出，为获得事件理解所需的丰富语义，对从媒体层识别到的上下文进行融合是非常关键的步骤，本书也是在此认识的基础上开展后续可穿戴式日常行为语义感知和增强处理的。

3.4　日常行为感知及挑战

3.4.1　日常行为感知——以 SenseCam 为例

作为一种信息量丰富的可穿戴式设备，可穿戴式视觉传感器可以在自由环境中进行日常行为的连续记录并通过处理完成行为识别。在可穿戴式生活记录中，通常采用移动传感设备，这些设备可以直接被用户以穿戴的方式进行固定，例如通过固定在头部[41-42]或者固定在胸前[43-44]的方式对外界变化进行采集。在本章文献[45]中，作者对生成自动的日记以及构建生活日志系统中的关键问题和挑战进行了探讨。在本章文献[46]中，其他多种传感器(如加速度计、GPS、图像和音频等)通过智能手机进行了记录，并基于标注的日常行为进行了实际应用。尽管这些工作在某种程度上体现了其方法的有效性，但是这种直接从低层特征(如颜色和纹理)映射到语义标识的方法在描述行为语义，例如在更好地理解场景、物体等内容出现的过程中仍缺乏灵活性。近期在本章文献[47]中的工作也同样强调了这个问题的重要性。

作为日常活动的一种真实反映，这些记录下来的多媒体数据内容需要通过浏览、索引的方式进行管理，以获得大量日常事件的深层次含义。当前，在多媒体信息检索研究领域，对图像和视频进行有效索引的途径是，采用统计学习的方法将低层图像特征(如形状和颜色等)映射到更高层的语义概念(如“室内”“建筑”“走路”

等)。根据 TRECVid 国际标准评测的结果[32],一些概念的探测结果已经达到可以接受的程度,尤其是对于某些具有足够标注数据用于训练的概念。将语义概念的自动探测技术引入视觉生活记录,使得对大量传感器记录进行内容查询变得更加可行,在此基础上,可以将索引结果应用于对日常生活规律进行有效描述等上层应用。然而,由于人类从事行为的多样性特点,并且人与人在行为表现方面的不同,大量不同的语义概念可能出现在视觉传感记录中,这大大增加了自动概念探测的难度,从而提高了由此分析行为的难度。另外,由于图像在捕获过程中穿戴者自身的运动,即使在同一个行为事件片段中的图像也会有很显著的视觉差异。这就给基于探测出的语义概念进行日常行为特征刻画,带来了极大的挑战。

图 3-2 所示的 SenseCam 具有体积小、重量轻、可穿戴式,以及集成多种传感器的特点。SenseCam 集成有自动照相设备,采用鱼眼光学镜头,以穿戴者即第一人称的视角,对穿戴者所看到的外界环境进行图像记录。SenseCam 能在穿戴者不做任何干预的情况下,以大约 40 秒/次的速率连续进行拍照。另外,SenseCam 内嵌的红外传感器、温度传感器、加速度计等能够在感知穿戴者外界环境突然变化时,自动触发照相设备进行即时图像采集。由于其自身的优越性,SenseCam 在支持对过去事情的辅助回忆[43,48]、膳食监测[49]、行为识别[50]、体育训练[49]等方面都表现出有效性。由于 SenseCam 的多传感器感知能力、重量轻、电量续航时间长等特点,本书的研究工作采用 SenseCam 作为可穿戴式设备对用户的日常行为进行详细的实验记录。

图 3-2 SenseCam 可穿戴式视觉传感设备及穿戴示例

应用可穿戴式视觉传感构建的“数字记忆”面对的最大问题就是对大量数据的有效检索问题,例如用户使用 SenseCam 在一天中平均可以采集和记录多达 2500 幅图像。单纯通过人工的方式浏览这些图像并查找感兴趣的内容是非常耗时的,从几个月或几年的数字记忆中查询目标事件,或者以人工方式进行统计分析,其难度更是可想而知。不同于传统的视频或图像处理,对这些视觉传感数据的处理往往还要结合各种不同传感器产生的大量异质数据。以 SenseCam 为例,其内嵌的

温度、加速度计、红外传感器等同时记录了各种上下文数据，并且和采集的视觉图像进行同步。我们在实验中发现，如果结合 GPS 及蓝牙等外部传感数据，这些传感器在一天中就要产生大约 6000 条 GPS 记录、3000 条蓝牙探测结果，以及 SenseCam 产生的 16000 条加速度计信息。另外，由于人们个人行为事件的多样性，使用可穿戴式视觉采集设备记录的媒体中蕴含大量的语义概念，这就给自动概念探测带来了很大的困难。

对多媒体数据自动进行概念标注，即概念探测，是将非结构化的多媒体数据转换成计算机可理解的语义内容进行索引的关键环节。作为对可穿戴式传感器采集的多媒体数据的一项处理，语义概念探测是后期在事件层进行语义融合建模进而完成个人行为识别的基础。也就是说，语义概念出现的时间规律可以对图像序列从更高的层次进行刻画。例如，在“做饭”行为中，视觉概念（如“冰箱”“微波炉”等）通常以序列方式交替出现，并且与其他概念如“手”等频繁进行交互。例如，一个常见的规律是在打开冰箱之后，往往紧接着会出现启动微波炉的事件发生。这些规律可以被认为是概念的时间语义特征，并用于对更复杂时间序列（如行为事件）的进一步识别。

3.4.2　可穿戴式行为感知处理框架

精确度量人类的日常行为在很多方面可以带来重大的价值，在分析人类行为规律、膳食监测、职业治疗、辅助独立生活等领域都有广泛的应用前景。对日常行为度量的传统方法是采用个人报告的方法，即利用人工调查的方法对参与人员进行的某些值得关注的行为（如饮食行为等）进行回忆和统计。由于这种个人报告的方式在很大程度上受到个人经历、知识水平、回答者的理解程度等多方面的限制，客观的观测方法，如采用传感器记录并分析，在这些领域中获得了很强烈的需求，以提供对日常行为各方面的客观评估。因为频繁的现场观察非常耗时，并且受到评估人员各方面条件（如场地和时间）的限制，对日常行为的度量需要一种有效的解决方案，以帮助从大量传感器记录的行为数据中定位出具有实际意义的行为片段。当前移动智能设备和可穿戴式传感器的广泛普及，以及内嵌的多传感能力和强大计算能力，使得这种精确度量日常行为的方法逐渐成为可能。

由于精确度量日常行为的重要性和业界需求，研究人员开始采用数字传感设备记录多种上下文数据，以从更细粒度反映物理行为片段的各方面要素。目前，可穿戴式传感设备具有体积小、重量轻、电池续航时间长的特点，将这些传感设备应用于自动行为度量开辟了新的研究领域。在实际应用中，可穿戴式视觉传感大多数采用图像或视频等视觉媒体数据，这些数据往往包含更丰富的语义信息，因此，本书将重点针对此类非结构化的海量多媒体数据蕴含的丰富事件语义的有效检索技术展开探讨。

当前可穿戴式视觉传感研究中的一种日常行为识别和度量的方法可以由

图 3-3所示的框架进行处理。该解决方案包括低层特征提取、概念探测、时序动态过程建模等主要组成部分。概念动态融合过程可以桥接低层视觉特征和高层语义内容，并提供了对日常行为更直观的理解，这个过程即前面提到的基于语义概念的建模方法。在这种方法中，先通过机器学习分类模型构建一个低层特征和概念识别的映射，识别的概念通过在时间维度上的进一步融合对行为样本中的视觉语义动态规律进行建模，从而将索引结果通过查询和交互界面为更复杂的日常行为分析应用提供支持。

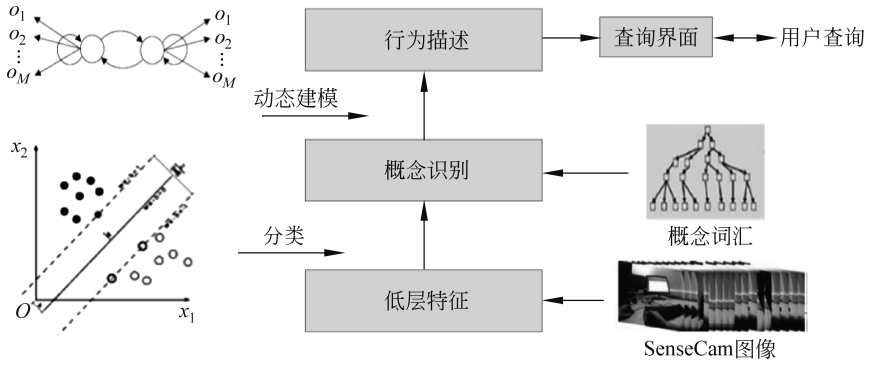


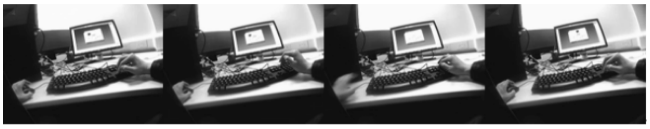
图 3-3 可穿戴式视觉研究中行为感知处理框架

3.4.3 面临的新挑战

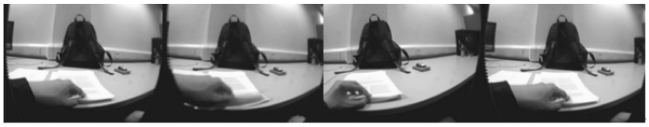
尽管经过多年的研究，从视觉图像中进行概念探测已经取得了较大的进步，但是最先进的概念探测方法仍然存在很多不尽如人意的地方。如何有效地从存在噪声和错误的概念探测结果中有效构建高层的行为识别算法，仍然是一个亟待解决的问题，尤其是对于可穿戴式视觉传感这样充满挑战的应用。在这种应用场景下，视觉传感器所捕捉的图像或视频内容是穿戴者从事日常行为的反映，由于穿戴者的不断移动以及从事行为的多样性，记录的视觉媒体中往往存在大量不同的视觉概念，甚至在同一个行为进行过程中所出现的概念也会有很大的视觉差异。

在研究过程中我们发现，可穿戴式视觉传感设备（如 SenseCam）记录的单个事件，如“走路”“驾驶”“做饭”“用电脑”“阅读”等，都有可能包含非常多通常达上百幅图像。在“用电脑”“阅读”等事件中，由于用户通常处于比较静止的状态，大多数图像记录在视觉上非常相似，如图 3-4(a)和图 3-4(b)所示。然而，在“走路”“做饭”等事件中，用户的不断移动产生了大量不同内容的图像，如图 3-4(c)和图 3-4(d)所示。从图 3-4 可以看出，在“走路”“做饭”等很多事件中，即使连续的图像无论在内容还是视觉特征上都有很大的不同。而传统多媒体信息检索研究的电视新闻视频、电影视频等都是经过预先编辑的，这些视频的单个镜头中连续帧之间在视觉特征上都非常相似。可以预见，可穿戴式传感在视觉数据上的多样性将导致在事件

探测过程中面临更多传统视频分析中未曾出现的困难。



(a) “用电脑”事件



(b) “阅读”事件

(c) “走路”事件

(d) “做饭”事件

图 3-4　SenseCam 传感图像示例

彩图 3-4

另一方面，由可穿戴式设备记录的视觉媒体本身的质量问题，也是影响整个算法效果的重要因素。如图 3-3 所示，在解决方案中，首先需要从原始视觉传感数据中提取低层特征。然而，由视觉传感设备采集的原始媒体数据所具有的低质量问题，是影响分析效果的一个与生俱来的挑战。以图 3-2 所示的 SenseCam 为例，尽管其中内嵌的运动传感器可以在一定程度上减轻由用户移动所带来的图像模糊问题，但是所捕捉的图像仍然存在很多质量问题。如图 3-5(a)、(b)和(c)列举了运动模糊造成图像降质的三个样例。并且，由于穿戴方式的问题（如图 3-2 中 SenseCam 悬挂在穿戴者胸前），视觉传感器的镜头经常被穿戴者的衣服或胳膊遮挡，这就使捕捉的有些图像具有较窄的视野范围，如图 3-5(d)和(e)两个样例所示。此外，图 3-5(f)所示的过度曝光问题也使得从图像中精确获得语义内容变得非常困难。

虽然通过预处理的方法（如对低质量的图像进行过滤）可以在一定程度上避免低质量图像带来的干扰，如采用对图像对比度（Contrast）和显著性（Saliency）融合的方法[51-52]，但是不能进一步解决概念探测精度受限的问题。本书后续的章节将


彩图 3-5

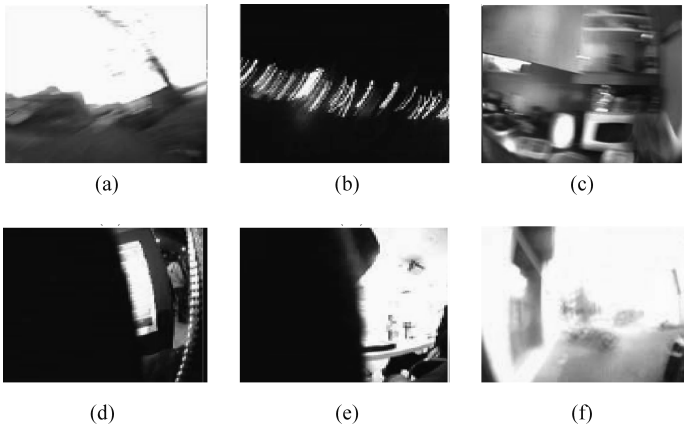


图 3-5 可穿戴式行为研究中面临的典型图像质量问题

从概念出现的上下文相关性出发，探索有效的概念探测增强方法，研究概念出现的时间动态规律，并且提出鲁棒的行为识别方法。需要说明的是，由于可穿戴式记录本身的特点，常常需要在数据容量和图像时间粒度间进行折中，即在降低数据存储空间的同时牺牲了图像采集速率。例如，SenseCam 大约每 40 秒采集一幅图像，这就导致了一些动态描述符，如 HOG（Histograms of Oriented Gradients）和 HOF（Histograms of Optical Flow）等特征不可用，虽然这些特征在传统的视频分类任务中被广泛采用并表现出很好的效果。

为了有效缓解上述挑战，本书将在第 5 章研究语义索引即概念探测的增强方法。本书将在第 7 章介绍不同的时序建模方法，并通过建模概念属性随时间变化的动态规律，构建日常活动的识别方法。本书的第 8 章将通过系统的分析实验，研究这种基于概念的行为识别方法的影响因素，以及这些影响因素与最终行为识别效果的关联关系，以提供给其他研究人员在实际应用中的指导性建议。

3.5 本章小结

本章介绍了与可穿戴式行为语义感知相关的多媒体信息检索背景知识，以及在可穿戴式感知方面的应用。这些背景知识包括低层特征提取、高层特征提取，以及基本的检索方式和方法等。由于对日常行为的处理通常以事件作为基本的语义单元，本章专门介绍了以事件为中心的媒体处理。作为一种新的多媒体形式，可穿戴式采集设备所记录的日常行为数据具有其自身的特点。这些特点与传统多媒体数据（如广播电视新闻、电影等）在模态、图像质量、视觉多样性等方面都有很大的差别。本章以 SenseCam 可穿戴式采集设备为例，分析了在可穿戴式行为语义感知方面的对应挑战。

参 考 文 献

[1] Blanken H, de Vries A P, Blok H E, et al. Multimedia Retrieval[M]. Berlin: Springer-Verlag, 2007.

[2] Gevers T, Smeulders A W M. PicToSeek: Combining color and shape invariant features for image retrieval[J]. IEEE Transactions on Image Processing, 2000, 9(1): 102-119.

[3] Spyrou E, Borgne H, Mailis T, et al. Fusing MPEG-7 visual descriptors for image classification: Proceedings of the 15th International Conference on Artificial Neural Networks: Formal Models and Their Applications[C]. Berlin: Springer, 2005.

[4] Manjunath B S, Ohm J, Vasudevan V V, et al. Color and texture descriptors[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2001, 11(6): 703-715.

[5] Manjunath B S, Ma W Y. Texture features for browsing and retrieval of image data[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1996, 18: 837-842.

[6] Belongie S J, Malik J M, Puzicha J. Shape matching and object recognition using shape contexts[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2002, 24: 509-522.

[7] Rui Y, Huang T S, Chang S F. Image retrieval: current techniques, promising directions and open issues[J]. Journal of Visual Communication and Image Representation, 1999, 10(4): 39-62.

[8] Conaire C Ó, Connaghan D, Kelly P, et al. Combining inertial and visual sensing for human action recognition in tennis: Proceedings of the 1st ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams[C]. New York: ACM, 2010.

[9] Rusinol M, Aldavert D, Karatzas D, et al. Interactive trademark image retrieval by fusing semantic and visual content: Proceedings of the 33rd European Conference on Advances in Information Retrieval[C]. Berlin: Springer, 2011.

[10] Snoek C G M, Huurnink B, Hollink L, et al. Adding semantics to detectors for video retrieval[J]. IEEE Transactions on Multimedia, 2007, 9(5): 975-986.

[11] Li J, Wang J Z. Automatic linguistic indexing of pictures by a statistical modeling approach[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25: 1075-1088.

[12] Li B, Goh K, Chang E Y. Confidence-based dynamic ensemble for image annotation and semantics discovery: Proceedings of the 11th Annual ACM International Conference on Multimedia[C]. New York: ACM, 2003.

[13] Li X, Wang D, Li J, et al. Video search in concept subspace: A text-like paradigm: Proceedings of the 6th ACM International Conference on Image and Video Retrieval[C]. New York: ACM, 2007.

[14] Snoek C G M, Gemert J C, Gevers T, et al. The MediaMill TRECVid 2006 semantic video

search engine: Proceedings of the 4th TRECVid Workshop[C].[S.l.]:[s.n.],2006.

[15] Ma W Y, Manjunath B S. NeTra: a toolbox for navigating large image databases[J]. Multimedia Systems,1999,7(3): 184-198.

[16] Bimbo A D,Pala P. Visual image retrieval by elastic matching of user sketches[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,1997,19: 121-132.

[17] Brown M G,Foote J T,Jones G J F,et al. Automatic content-based retrieval of broadcast news: Proceedings of the 3rd ACM International Conference on Multimedia[C]. New York: ACM,1995.

[18] Adams B, Amir A, Dorai C, et al. IBM Research TREC-2002 video retrieval system: Proceedings of the TREC-2002[C].[S.l.: s.n.],2002.

[19] Westerveld T,Vries A P,Ballegooij A,et al. A probabilistic multimedia retrieval model and its evaluation[J]. EURASIP Journal on Applied Signal Processing,2003,(2): 186-198.

[20] Chua T S,Neo S Y,Li K Y,et al. TRECVid 2004 search and feature extraction task by NUS PRIS: Proceedings of NIST TRECVid Workshop[C].[S.l.: s.n.],2004.

[21] Yan R, Yang J, Hauptmann A G. Learning query-class dependent weights in automatic video retrieval: Proceedings of the 12th Annual ACM International Conference on Multimedia[C]. New York: ACM,2004.

[22] Natsev A, Naphade M R, Tesic J. Learning the semantics of multimedia queries and concepts from a small number of examples: Proceedings of the 13th Annual ACM International Conference on Multimedia[C]. New York: ACM,2005.

[23] Kennedy L S,Natsev A,Chang S F. Automatic discovery of query-class-dependent models for multimodal search: Proceedings of the 13th Annual ACM International Conference on Multimedia[C]. New York: ACM,2005.

[24] Lienhart R,Kuhmunch C,Effelsberg W. On the detection and recognition of television commercials: Proceedings of the IEEE International Conference on Multimedia Computing and Systems[C].[S.l.]: IEEE,1997.

[25] Smith J R,Chang S F. Visually searching the web for content[J]. IEEE Multimedia, 1997,4: 12-20.

[26] Rui Y,Gupta A,Acero A. Automatically extracting highlights for TV baseball programs: Proceedings of the 8th ACM International Conference on Multimedia[C]. New York: ACM,2000.

[27] Naphade M R, Huang T S. A probabilistic framework for semantic video indexing, filtering,and retrieval[J]. IEEE Transactions on Multimedia,2001,3: 141-151.

[28] Amir A,Berg M,Chang S F,et al. IBM research TRECVid-2003 video retrieval system: Proceedings of NIST TRECVid Workshop[C].[S.l.: s.n.],2003.

[29] Fan J,Elmagarmid A K,Zhu X,et al. ClassView: Hierarchical video shot classification, indexing,and accessing[J]. IEEE Transactions on Multimedia,2004,6: 70-86.

[30] Snoek C G M,Worring M,Geusebroek J M,et al. The semantic pathfinder: Using an

authoring metaphor for generic multimedia indexing[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2006,28(10): 1678-1689.

[31] Gemert J C,Geusebroek J,Veenman C J,et al. Robust scene categorization by learning image statistics in context: Proceedings of the 2006 Conference on Computer Vision and Pattern Recognition Workshop[C].[S.l.]: IEEE,2006.

[32] Smeaton A F,Over P,Kraaij W. High level feature detection from video in TRECVid: a 5-year retrospective of achievements[M]//Divakaran A. Multimedia Content Analysis: Theory and Applications.[S.l.]: Springer US,2009: 151-174.

[33] Neo S Y,Zhao J,Kan M Y,et al. Video retrieval using high level features: Exploiting query matching and confidence-based weighting: Proceedings of the 5th International Conference on Image and Video Retrieval[C]. Berlin: Springer,2006.

[34] Lin W H,Hauptmann A G. Which thousand words are worth a picture? Experiments on video retrieval using a thousand concepts: Proceedings of the IEEE International Conference on Multimedia and Expo[C].[S.l.]: IEEE,2006.

[35] Hauptmann A,Yan R,Lin W H. How many high-level concepts will fill the semantic gap in news video retrieval?: Proceedings of the 6th ACM International Conference on Image and Video Retrieval[C]. New York: ACM,2007.

[36] Wei X Y,Ngo C W. Ontology-enriched semantic space for video search: Proceedings of the 15th International Conference on Multimedia[C]. New York: ACM,2007.

[37] Westermann U,Jain R. Toward a common event model for multimedia applications[J]. IEEE Multimedia,2007,14:19-29.

[38] Frohlich D,Kuchinsky A,Pering C,et al. Requirements for photoware: Proceedings of the 2002 ACM Conference on Computer Supported Cooperative Work[C]. New York: ACM, 2002.

[39] Rodden K,Wood K R. How do people manage their digital photographs?: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems[C]. New York: ACM,2003.

[40] Naaman M,Song Y J,Paepcke A,et al. Automatic organization for digital photographs with geographic coordinates: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries[C]. New York: ACM,2004.

[41] Hori T,Aizawa K. Context-based video retrieval system for the life-log applications: Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval[C]. New York: ACM,2003.

[42] Mann S,Fung J,Aimone C,et al. Designing EyeTap digital eye-glasses for continuous lifelong capture and sharing of personal experiences: Proceedings of the CHI 2005 Conference on Computer Human Interaction[C]. New York: ACM,2005.

[43] Sellen A,Fogg A,Aitken M,et al. Do life-logging technologies support memory for the past? An experimental study using SenseCam: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems[C]. New York: ACM,2007.

[44] Blum M, Pentland A S, Tröster G. InSense: Interest-based life logging[J]. IEEE Multimedia, 2006, 13(4): 40-48.

[45] Machajdik J, Hanbury A, Garz A, et al. Affective computing for wearable diary and lifelogging systems: an overview: Workshop of the Austrian Association for Pattern Recognition[C].[S.l.: s.n.], 2011.

[46] Hamm J, Stone B, Belkin M, et al. Automatic annotation of daily activity from smartphone-based multisensory streams: Proceedings of the 4th International Conference on Mobile Computing, Applications, and Services[C]. New York: ACM, 2012.

[47] Song S, Chandrasekhar V, Cheung N M, et al. Activity recognition in egocentric life-logging videos: Proceedings of Computer Vision - ACCV 2014 Workshops[C]. Cham: Springer, 2014.

[48] Silva A R, Pinho S, Macedo L M, et al. Benefits of SenseCam review on neuropsychological test performance[J]. American Journal of Preventive Medicine, 2013, 44(3): 302-307.

[49] O'Loughlin G, Cullen S J, McGoldrick A, et al. Using a wearable camera to increase the accuracy of dietary analysis[J]. American Journal of Preventive Medicine, 2013, 44(3): 297-301.

[50] Wang P, Smeaton A F. Using visual lifelogs to automatically characterize everyday activities[J]. Information Sciences, 2013, 230: 147-161.

[51] Doherty A R, Byrne D, Smeaton A F, et al. Investigating keyframe selection methods in the novel domain of passively captured visual lifelogs: Proceedings of the 2008 International Conference on Content-based Image and Video Retrieval[C]. New York: ACM, 2008.

[52] Wang P, Smeaton A F. Aggregating semantic concepts for event representation in lifelogging: Proceedings of the International Workshop on Semantic Web Information Management[C]. New York: ACM, 2011.