华南理工大学

South China University of Technology

本科毕业设计（论文）

基于生物启发模型的时空特征提取

作者姓名	蔡博仑
学生学号	200930251439
学科专业	集成电路设计与集成系统
指导教师	徐向民　　教授
所在学院	电子与信息学院
提交日期	2013 年 6 月 16 日

Biologically Inspired Model based Spatio-temporal Feature Extraction

A Dissertation Submitted for the Degree of Bachelor

Candidate: Cai Bolun

Supervisor: Prof. Xu Xiangmin

South China University of Technology

Guangzhou, China

摘　要

伴随着科技的进步与发展，数字摄像头、互联网技术、多媒体技术、信息处理技术的逐渐成熟，数字视频已成为一种广泛应用的媒体。人工智能的迅猛发展促使机器视觉、视频感知成为近十年来的学术界研究热点与重点。人体行为识别在智能监控、人机交互和视频检索等方面具有广阔的应用前景，因此其成为计算机智能视觉监控领域中最活跃的研究主题之一。特征提取与行为分类识别算法是人体行为识别的关键技术所在。

然而，受限于传统信号处理与模式识别架构，仅仅依靠现有机器视觉范畴的研究理论，很难实现特征提取上“区分性”和“不变性”的统一。本文引进生物学模型，模拟灵长动物视觉神经系统对视觉信息的层次化处理过程。基于生物启发模型的时空兴趣点框架包括，时空兴趣点检测和时空描述符提取。提出联合生物启发模型与时空兴趣点方法进行时空兴趣点的检测和时空描述符的提取。首次采用初级运动检测器对背侧流的 MT 区进行建模，构建背侧流模型进行人体行为识别。生物启发模型的引进改善了现有局部时空兴趣点方法在“区分性”和“不变性”的不统一，得到更为有效的特征向量。

通过常用人体行为识别数据库实验表明，本文提出的基于生物启发模型的时空兴趣点方法具有较好的鲁棒性，准确性。相比于现有时空兴趣点方法，本文方法具有以下优点：首先，本方法基于生物研究基础，不存在参数主观选择问题；其次，在相同的参数设置下对于多个数据库均有较好效果；最后，本方法模拟灵长动物大脑皮层，现处于研究初级阶段，仍有广阔的提升空间。

关键词：生物启发模型；时空兴趣点；人体行为识别

Abstract

With the advancement and development of technology, digital cameras, Internet technology, multimedia technology and information processing technology have matured, so digital video has become a medium used widely. The rapid development of artificial intelligence promotes machine vision and video perception to become hot and priorities over the past decade. Human behavior recognition has broad application prospects in intelligent surveillance, human-computer interaction and video retrieval, so it became the most active research themes in computer intelligent visual surveillance. Feature extraction and classification algorithm are the key technology for human behavior recognition.

However, limited conventional signal processing and pattern recognition framework, relying solely on existing research areas of machine vision theory, it is difficult to achieve a feature, which is unitary on the "distinction" and "invariance". This paper introduces biologically inspired model to simulate the primate visual system's hierarchical visual information processing. Spatio-temporal interesting points based on biologically inspired model framework include spatio-temporal interesting points' detection and spatio-temporal descriptor extraction. Make a joint bio-inspired model of temporal and spatial temporal interest points' method for the detection and temporal points of interest descriptor extraction. In this paper, MT area is modeled to simulate the primary dorsal stream motion detector. The introduction of the biologically inspired model improves the existing methods of local spatio-temporal feature's uniformity in the "distinction" and "invariance". Then the framework can extract the more effective feature vectors.

Through two human activity common databases recognition experiments show that the proposed model, spatio-temporal interesting points based on biologically inspired model, has better robustness and accuracy for human activity recognition. Compared to the existing spatio-temporal interesting points method, this method has the following advantages: Firstly, the method is based on biological research, there is no subjective parameter selection; Secondly, the same parameter settings have good effect for multiple databases; Finally, this method simulating primate cerebral cortex, is now in the initial stage of research, and there is still vast room for improvement.

Keyword: Biologically inspired model, Spatio-temporal interesting points, Human action recognition

目 录

第一章 绪论

1.1 课题背景

伴随着科学技术的进步与发展，数字摄像头、互联网技术、多媒体技术、信息处理技术的逐渐成熟，数字视频已成为信息领域的一种广泛应用的媒体。此外，人工智能的迅猛发展推动其在众多学科领域的广泛应用，机器视觉、视频感知逐渐成为近十年来的学术界研究热点与重点。出于对反恐、安保、侦察等方面的急切需求，各国家、地区均纷纷建立了庞大的视频监控网络，对道路、公共场所、事故多发地等进行 24 小时不间断的视频监控。由于我国新型城镇化与智慧城市建设需求，以智能视频监控为核心的城市安防监控产业迎来快速发展期。我国新型城镇化与智慧城市的建设需要，为智能化公共安全视频监控解决方案提供了重要市场机遇。

当前视频监控产业快速扩张主要局限于视频记录为主的常规监控技术，现阶段的视频监控设备仅对数据的进行压缩编码、高速传输，部分智能监控产品也仅采用初级的机器视觉技术实现简单的行为识别。伴随着视频信息量的指数性地增长，有且只靠人工对视频数据监控分析满足不了“大数据”时代的需求。“需求是科学发展的驱动力”，面对海量视频数据分析所带给我们的极大困难，视频监控领域的刚性需求为人体行为识别（Human Action Recognition，HAR）的发展提供了良好的契机。各国的相关研究机构、大学、企业均投入了巨大资源支持该领域的研究。同时，HAR 技术也可以应用于人机交互和、视频检索等方面[1-3]，具有广阔的应用前景。然而，HAR 研究领域属于交叉学科研究，既包含图像处理、机器视觉等知识，同时也涉及机器学习、模式识别、甚至数据挖掘等知识，此外还需要借鉴生物学、认知学和心理学等交叉学科的研究成果。因此，HAR 技术尚未成熟，相关研究尚处于起步阶段，有着较为广阔的研究空间。

HAR 属于对时序数据的识别分类问题，与基本模式识别框架一致，主要包括特征提取和分类识别两个部分。HAR 的根本目的是为了识别和理解人类行为，其中包括个体行为、人与人之间、以及人与外部环境之间的相互行为。然而，客观环境的多样性和人体运动的复杂性，为 HAR 提供了巨大的挑战。因此，提取对于不同的行为类型具有良好的“区分性”，对于同一种行为具有良好的“不变性”的分类特征是 HAR 的研究重点和难点。一个优秀的特征描述是提高行为识别准确度的首要条件，描述特征的性能将决定着分类器的最终行为识别分类结果。

有别于静止图像感知技术，HAR 数据主要通过视频采集设备获取，如网络摄像头、模拟摄像头、红外摄像头等。视频数据除了包含二维图像信息外，还增加了图像信息之间的

时间相关度，提取包含时间维度的时空特征是本课题研究的关键。现阶段，二维图像的特征提取算法已相对成熟，其中包括 SIFT（Scale-invariant Feature Transform）特征[4]，SURF（Speeded Up Robust Features）特征[5]，HOG（Histograms of Oriented Gradients）特征[6]等在内的二维图像特征描述符在静止图像识别领域已取得很好的效果与工业化应用。二维图像特征的成功应用，表明了提取时空描述符的方法在机器视觉识别领域具有可行性，为构建有效的时空特征描述符提供了理论基础和研究指导。

表观特征、“时空体”特征、语义分析、人体肢体运动特征、神经网络等方法逐渐应用在 HAR 领域之中。然而，由于受限于基础学科发展以及计算机科学发展的约束，单纯局限于计算机视觉领域的研究很难从根本上满足人体行为识别特征的需求，无法很好地保证特征描述符在“不变性”和“区分性”上的统一。生物学研究表明，哺乳动物，尤其是灵长类是拥有很强的视觉模式识别功能。借鉴生物视觉系统对于视觉信息的抽象认知模型，提取不易受到客观环境影响的高层行为特征是人体行为识别特征提取的发展方向。该思想是当前 HAR 领域中运动特征提取模型的发展的主要方向，也是本论文的研究方向。

1.2 研究现状

1.2.1 人体行为识别研究现状

伴随着机器视觉和人工智能研究的高速发展，许多运动信息提取和分类方法被广泛地应用到视频中的人体行为识别领域。

概率网络模型是被最广泛且成功地应用于处理时序特征分类问题。对于视频中的时序人体行为特征分类问题，首先在视频序列中提取运动特征，并使用概率模型对人体行为进行建模分类。在 Tsukuba 大学的一项研究[7]中，该作者提出使用基于光流场统计方向直方图的方法对视频序列进行运动特征提取，并利用隐马尔科夫模型（Hidden Markov Model，HMM）对行为进行分类识别。Liang 等[8]则利用局部保持投影（Locality Preserving Projections，LPP）匹配动态形状流形实现行为识别。在 Monash 大学的一篇研究论文[9]中，作者提取星形骨骼的人体姿势的特征描述符，并基于隐马尔科夫方法进行分类识别。然而，基于概率网络模型的方法存在对于状态匹配结果敏感的缺点，当运动特性发生轻微的尺度和方向偏差变化而影响到某一状态的匹配结果，将影响概率模型中所有状态的遍历过程。

在视频中提取局部时空特征并对其进行模板匹配，也是行为识别的通用方法之一。Paul 等[10]使用三维尺度不变转换（3D Scale-invariant Feature Transform，3D-SIFT）特征描述符表征人体行为，并用特征包匹配方法进行行为分类识别。在 Columbia 大学的一项研究[11]中，局部运动特征（Location Motion Pattern，LMP）描述符被用于构建超完备的稀疏响应字典用于人体行为识别。Willemset 等[12]提出基于图像 SURF 特征改进的扩展 SURF 特征

（Extend Speeded Up Robust Features，E-SURF）描述符用于人体行为识别。以上提到的这些特征描述，均独立地处理时间维度信息和空间维度信息，因此无法很好地融合时空特征，不能很好地描述相似的不同行为。

此外，文本语义技术[13, 14]，人体肢体运动特征[15, 16]，人工神经网络[17]等方法也被逐步应用于 HAR 领域。然而，这些方法都没有超越计算机视觉的范围，仍停留在传统的信号处理和模式识别的架构上，无法真正地解决 HAR 领域所面临的瓶颈。

由于物理学，数学和计算机科学的客观条件的限制，传统的信号处理架构不能满足“不变性”和“区分性”的运动特征信息提取需求。众所周知，灵长类动物是视觉模式识别功能最强大的动物，将生物视觉感知机制应用到传统的信号处理模型中，提取一种保持尺度和方向不变的新运动信息特征成为解决 HAR 领域瓶颈的思路。Jhuanget 等[18]提出基于分层前馈结构和神经逻辑模型的生物启发系统用于行为识别。在智利玛利亚大学的研究[19]中，研究人员提出基于 MT 细胞平均活跃度的生物启发特征方法应用于视频的行为识别。然而，以上提出的方法都是基于生物启发模型的全局特征提取，因此并不能很好地适应局部遮挡，局部光照变化和复杂场景的应用。

1.2.2 生物视觉皮层研究现状

随着生物学领域对于生物神经研究的进展，灵长动物视觉皮层信息处理机制的研究已初见成果。建立于神经生物学和心理物理学的生物的理论基础上，生物视觉皮层的数学模型已初步形成。1855 年，Panizza[20]通过生物学实验发现大脑内部存在专司与视觉信号处理的视觉皮层。二十世纪 50 年代，Barlow[21]通过青蛙视网膜的微电极研究，发现了视网膜神经结细胞的放电现象； Mcilwain [22]对猫视网膜神经结细胞的研究发现了感受野结构。60 年代初，Roddick 首次提出生物视觉皮层感知的数学模型，用于估算视网膜神经结细胞对空间刺激的响应。60 年代末到 70 年代初，Hubel 等人[23]发现了大脑视觉皮层的各视觉功能区之间存在响应联系，并提出了感受野综合的假设。80 年代，Mishkin 和 Ungerleider[24]对猴脑视觉皮层进行深入研究，发现灵长类动物视觉皮层存在两条视觉通路，分别为负责静态特征提取的腹侧流和负责运动信息分析的背侧流。90 年代末，Riesenhuber 和 Poggio[25]发现腹侧流和背侧流并不是简单的各司其职，两条视觉通路会在中高层的视觉皮层发生信息融合和相互作用，共同完成对物体和行为的识别。与此同时，中科院李朝义院士[26]提出整合野概念，初级视觉皮层神经元通过整合野与感受野的相互作用对图像进行高层特征提取。生物学研究的成熟和基本数学模型的提出，促使国内外各大科研机构将生物学启发模型应用于 HAR 领域。

1.3 论文研究内容与及成果

应用生物学研究成果于人体行为识别领域是本论文的研究重点。本论文研究旨在基于生物学研究基础之上，提出一种新的局部时空兴趣点描述符的行为识别方法。采用视觉感受野思想检测局部时空兴趣点，使得时空特征描述符在局部光照变化，局部遮挡，尺度变化等复杂场景中也具有较好的行为识别效果。引入层次化的生物启发模型提取人体行为的时空特征，将得到具有“区分性”与“不变性”的高层特征描述符，更好地进行行为识别。

在本论文研究成果是将生物学研究成果应用于 HAR 领域，提出了一种基于生物启发模型的时空兴趣点（Spatio-Temporal Interesting Points base on Biologically Inspired Model，BIM-STIP）框架用于图像序列中的人体动作识别。BIM-STIP 框架主要包含两大部分：第一部分，基于生物启发模型的时空兴趣点检测方法；第二部分，提取基于生物启发模型的时空特征描述符。本论文贡献点如下：

（1）提出基于生物启发模型的时空兴趣点检测（BIM-STIP Detector）方法。模拟灵长动物视觉皮层的层次化，多尺度，局部感受野模型检测时空兴趣点。首次引入初级运动检测器作为视皮层模拟，进行运动感兴趣区域检测。初级运动检测器的引入改进了原有三维 Harris 角点检测存在的时间维度、空间维度独立处理缺点，更好地融合时空信息，得到对人体行为更有表征效果的时空兴趣点。此外，时空兴趣点检测方法与后续的局部时空特征提取方法建立于相同的生物启发框架下，能更好地保证时空兴趣点的检测结果为局部时空特征提取的较优解。

（2）制定基于生物启发模型的时空特征描述（BIM-STIP Descriptor），准确地捕捉图像序列的时空特征。时空特征提取模拟生物视觉细胞感受野，层次化模型，提取具有“不变性”和“区分性”的高层特征。现有行为识别特征描述符的构造方法均建立在二维静止图像特征提取的基础上，简单地对二维静止图像特征提取方法进行时间维度的三维推广。此外，基本上目前所有的二维静止图像特征描述符均存在凭经验设计，模型参数选取，缺乏理论支撑等缺点。引入生物启发模型，为机器视觉在时空特征的构造提供生物学依据，具有广阔的改进空间和提升空间。

（3）比较分析的基于生物启发模型的时空兴趣点框架（BIM-STIP frame）与现有的时空兴趣点描述符的性能。实验数据表明，相比现有时空兴趣点描述符方法，在 BIM-STIP 框架下具有很高的识别效果。

1.4 论文结构安排

本论文的其余章节组织结构如下：

第二章主要介绍灵长动物视觉系统的内容。首先，从生物学研究的角度介绍灵长动物如何将外界视觉刺激转换为神经电刺激，通过各个视觉皮层的逐层处理，最终实现对视觉的感知。其次，介绍几十年来神经解剖学和心理物理学对灵长动物大脑视皮层的理论研究成果，从生物学角度介绍视皮层建模的依据。最后，将逐一对背侧流中不同视觉皮层神经细胞建模，提供各个皮层的数学模型。

第三章主要引入生物启发模型到时空兴趣点方法。提出基于生物启发模型的时空兴趣点检测方法，在视频序列中检测具有强表征的多尺度时空兴趣点。对应所检测所得的时空兴趣点，同样采用生物启发架构提取具有“不变性”和“区分性”的多尺度特征。最后，采取特征包模型对提取的特征进行分类，以实现人体行为的分类。

第四章中通过对人体行为识别领域公认的 Weizmann 和 KTH 数据库的测试、本文提出的特征提取算法将与目前较成功的人体行为检测模型在特征的“区分性”和“不变性”上作性能比较。

最后，第五章对全文的研究成果进行了总结，指出本文主要的贡献，并对未来需要继续开展的工作做出展望。

第二章 灵长动物视觉通路

2.1 灵长动物视觉皮层

灵长类动物的视觉通路是一个跨学科的复杂的系统，它不仅涉及到眼球运动，而且还与下丘脑，大脑皮层等其他神经活动相关。从神经解剖学的角度来看，灵长类动物的视觉系统是由大量的有序排列、层次化排列的神经元细胞组成，其中包括眼球光学系统，视网膜，外侧膝状体（Lateral Geniculate Nucleus，LGN）和视觉皮层等。图 2-1 示出从视觉光刺激到神经信号的转换过程[27]。

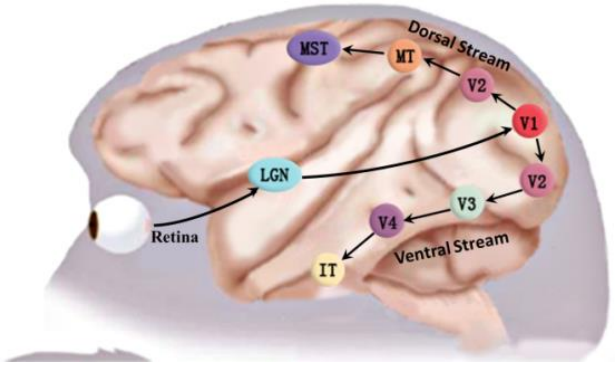

图 2-1 视觉信号处理流程

通过的瞳孔和眼睛晶状体的调整，自然界的光信号以视觉刺激的形式投射到视网膜。在视网膜上，光信号转变为对视网膜神经节细胞的神经脉冲信号，然后发送到位于丘脑处的 LGN。神经脉冲信号通过视觉 LGN 的初级处理，经神经元被传送到视觉皮层。视觉皮层[28]是指大脑皮层中主要负责处理视觉信息的一种典型的感觉型粒状神经皮（Koniocortex cortex），其位于大脑后部的枕叶的距状裂周围，包括初级视皮层（Primary Visual Cortex）、纹前皮层和纹外皮层。视觉皮层存在两种不同的视觉通路：腹侧流（Ventral Stream）和背侧流（Dorsal Stream）。腹侧流常被称为“内容通路”（What pathway），参与物体识别；背侧流常被称为“空间通路”（Where pathway），参与处理物体的空间位置信息以及相关的运动信息提取。这两条通路均起源于初级视皮层（V1 区），亦称为纹状皮层，负责提取多尺度，多方向的局部感受野特征。初级视皮层处理结果通过未予纹前皮层的复杂视觉细胞（V2 区）进行最大池（Maximum Pool）操作，起中继区作用，负责存储视觉信息并确定信息流方向。其中一部分视觉信息通过 V3 区和 V4 区，流入对物体空间信息敏感的腹侧流；另外一部分视觉信息通过 MT 区和 MST 流入对物体运动方向敏感的背侧流。

2.2 视觉通路的功能

生物学研究表明，灵长动物视觉通路的神经细胞的层次化处理模型可以通过逐层视觉特征提取，逐步获取具有“不变性”与“区分性”的高层视觉特征。本论文旨在模拟视觉通路中背侧流通路对运动信息处理过程实现人体行为识别与分类，现简要介绍背侧流通路中各层次皮层的功能。

眼球光线系统由巩膜、角膜及其内容物等组成，其主要部分大体上像球状体。视网膜（Retina）由色素上皮层和视网膜感觉层组成，居于眼球壁的内层。色素上皮层由色素上皮细胞组成，与脉络膜紧密相连，具有支持和营养光感受器细胞、遮光、散热以及再生和修复等作用。光学信息通过眼球聚焦投影到视网膜上，形成视觉神经冲动，沿视觉通路将视信息传递到视觉中枢形成视觉，从而在大脑中建立起图像。

LGN 位于视束的后端、丘脑枕的外侧，其形如马鞍状。McAlonan 和 Cavanaugh 等[29]通过对猕猴的视觉皮层进行研究表明 LGN 起到空间注意力调控作用。根据视觉注意力机制，通过 LGN 的空间注意力调控，视觉神经将着重关注运动感兴趣区域，对运动感兴趣区域做高层特征提取。

初级视皮层位于 Brodmann 17 区，也称为 V1 区，其输入主要来自位于丘脑的 LGN。生物研究表明，位于 V1 区的神经元细胞具有两个特性：一、神经元细胞仅对处于其感受野中的刺激产生响应，即单个神经元仅对与其相关的局部空间敏感；二、神经元细胞在感受野范围内对纹理方向特征敏感，即单个神经元仅对某一频域段呈现较强的响应。因此，初级视皮层的神经元细胞感受野可以被描述为具有局部性、方向性和带通特性的一系列信号编码滤波器，即提取某一频段的，某一方向的边缘、线段、条纹的卷积模板。

视皮层复杂细胞位于 V2 区，相比于简单视皮层细胞其感受野面积较大，接受来自 V1 区细胞的视觉信息，对其刺激呈非线性响应。由于复杂细胞结构复杂且各功能区域重叠分化不严格，其在感受视野内对具体位置信息的灵敏度相对简单细胞较低。同时，复杂细胞当有视觉信号经过视野区域的时间段内能对适当方向向量信号刺激产生一个持续性的响应。综上所述，简单细胞只能对某种方向向量信号产生响应，而复杂细胞却能在具有此方向向量信号的刺激发生运动时产生一个持续性的响应。从视觉细胞对外界信号响应上讲，复杂细胞提供了一种不受位置局限的抽象化刺激的方向向量信息。

视觉皮层中颞（MT）区视觉，也被成为 V5 区域。该纹外皮层区属于运动知觉，在将局部运动信号整合到全局视觉以及眼部运动的引导中起重要的作用。对 MT 区电生理特性的神经元属性的研究表明，大部分细胞会对移动物体的速度和方向刺激有较强反应。这些结果表明，MT 在活动视像处理上扮演了极其重要的角色。病变研究结果也支持 MT 在物体活动感知和眼球运动中的作用，神经心理学研究表明 MT 区受损的患者看不到运动的视像而是一系列静态的“帧”组成。

2.3 生物视觉系统模型

2.3.1 LGN 与运动显著区域

McAlonan 和 Cavanaugh 等人[29]用猕猴所做实验表明，空间注意力调控发生在外侧膝状核中。根据视觉注意力机制，对运动序列的时空体进行分析，提取出运动显著区域（Movement Significant Regional，MSR）[30]。

MSR 是人体行为的稀疏表示形式，可以精确地表征行为，因此对于 MSR 提取该区域的局部特征描述符可以避免背景环境和行为不规范的影响，具有较强的不变性。MSR 提取实验表明，人体行为的运动显著区域集中于人体头部和四肢部分，与人眼直观感受一致。通过将 MSR 的应用可以初步定为时空兴趣点位置，降低后续层次模型的感受野计算个数，从而减少了信息冗余，提高了模型的识别率，降低了运算量。

2.3.2 V1 区与 Gabor 滤波

Gabor 变换[31]属于加窗傅立叶变换，Gabor 函数可以在不同频域、不同尺度、不同方向上提取相关的特征。Gabor 滤波器在频率和方向类似于人类的视觉系统，所以常用于纹理识别。在空间域，二维 Gabor 滤波器是一个高斯核函数和正弦平面波的乘积。

初级视觉皮层，V1 区中简单细胞感受野具有多方向选择的基本特性。生物研究表明，因为 V1 区简单细胞具有的局部性，多尺度性和多方向性等特性，所以其感受野可以由一系列的尺度不一的二维 Gabor 滤波函数来逼近，并满足时频联合测不准原则。Granlund[32]首次提出二维 Gabor 小波变换函数的数学表示形式，并将其应用于图像处理领域。后续地，KnutSSon[33]和 Daugman[34]等人从生物学方向对二维 Gabor 滤波器进行更深入的研究，结果表明哺乳动物初级视觉皮层的感受野响应能够被二维 Gabor 函数很好的拟合,如图 2-2。如下图所示，从空域分布特性、方向选择特性、频域覆盖范围上，二维 Gabor 函数都能很好地模拟简单细胞感受野特性。他们的研究成果推动了二维 Gabor 小波分析方法在机器视觉领域的高速发展。现如今，二维 Gabor 函数已被广泛应用于图像的边缘检测和纹理检测中。

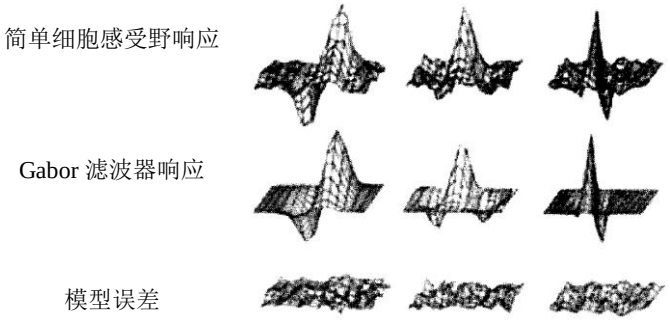

图 2-2 Gabor 滤波器与生物电模型

2.3.3 V2 区与神经元竞争

视觉皮层 V2 区的复杂视觉神经细胞的感受野范围大于 V1 区简单细胞，接受简单细胞的响应输入，综合线性特征，起到范围检测器的作用，即利用神经元间的竞争机制对于 V1 区的输出进行筛选。生物学研究表明，V2 区复杂视觉神经细胞的竞争机制可分为尺度竞争，方向竞争和位置竞争。

（1）相同方向，相同视野位置的不同尺度神经细胞之间的竞争成为尺度竞争。尺度竞争旨在得到更具有尺度不变性的更高层特征。

（2）复杂细胞具有比简单细胞更大的感受野，复杂位置竞争在其感受野内接受简单细胞刺激，刺激之间相互抑制，这种成为位置竞争。

（3）相同视野位置，不同方向的神经细胞之间的竞争，这种称为方向竞争。尺度竞争旨在保证

神经元间存在相互作用，可分为兴奋性和抑制性，当某一神经元处于兴奋的同时将对周围神经元释放抑制性递质，抑制性的后果是神经抑制。在视觉神经竞争过程中，对于视觉刺激响应较弱的神经元细胞将被抑制，其响应信息将不被传递至下后续各视皮层；对于视觉刺激响应较强的神经元细胞将保持原有响应状态，并将神经冲动经突触传入下一层次处理。基于生物神经元间抑制机理，其数学模型可表示为：对 V1 区简单视觉细胞的输出响应做数值比较，被抑制的神经元赋值为零，兴奋的神经元保持原有响应。该方法可以去除冗余信息，保留有效信息，避免信息爆炸，提取更具表征能力的高层的信息。

2.3.4 MT 区与初级运动感受器

今年来的一些研究表明，Reichardt 的初级运动检测器（Elementary Motion Detector，EMD）[35]与灵长动物视觉系统中运动检测单元在算法层次上是等价的。运动检测器理论最初是由 Reichardt 提出，用于解释昆虫的运动感知。如今 Reichardt 运动检测器模型已经得到生物视觉研究领域的认同，当前的一些时空能量模型和一些经典的生物运动检测模型在数学上都与 Reichardt 运动检测器模型等同[36]。

进一步研究和仿真表明[37]，对于单一频域的运动图像，Reichardt 运动检测器的速度估计与运动图像的空间频域有关；而对于宽频运动图像的速度估计，速度估计有且仅与图像的空间频谱有关。由于现实中的自然图像是宽频图像，且其频谱满足特定的形式， 因此 Reichardt 运动检测器它可以精确估计出自然图像的运动速度。该结论极大地丰富了初级运动检测器理论，是对运动检测器理论的完善。所以，在此采用 EMD 阵列从生物学上逼真地模拟了视皮层运动分量的神经元，对 MT 区神经元细胞进行建模。

2.4 本章小结

本章首先从生物学研究角度简单介绍了灵长动物视觉皮层的功能，介绍视觉通路的概念和两条主要的视觉通路。其次，以背侧流通路为主，介绍各个层次皮层在生物学上的功能和视觉刺激信号的在各层之间的流动方向。最后，根据现有的生物学研究依据，为各个层次皮层信号处理机制选取合适的数学模型。在下章中将着重介绍如何应用已有的生物学模型与时空兴趣点联合解决人体行为识别问题。

第三章 基于生物启发模型的时空特征框架

在本章中，将详细介绍所提出基于生物启发模型的时空兴趣点（Spatio-temporal Interesting Points based on Biologically Inspire Model，BIM-STIP）框架。根据灵长类动物的视觉通路研究表明，如图 3-1 所示，生物启发模型与灵长动物视觉皮层对应如下：其中运动显着区域模拟 LGN；多层背侧流模型模拟背侧流通路；特征包分类器用于模拟前额叶皮层（Pre-frontal Cortex，PFC）。与一般的时空特征点方法相似，在我们的框架中，对人体动作识别分为三个阶段。在第一阶段，我们提出了一种 BIM-STIP 检测方法，在视频序列中找到对该视频具有较强表征能力的兴趣点集。在第二阶段，在兴趣点的临域内，构造提取多尺度 BIM-STIP 时空描述符。在第三阶段中，利用特征包分类器在视频通过 BIM-STIP 描述符区分不同的人体行为动作。

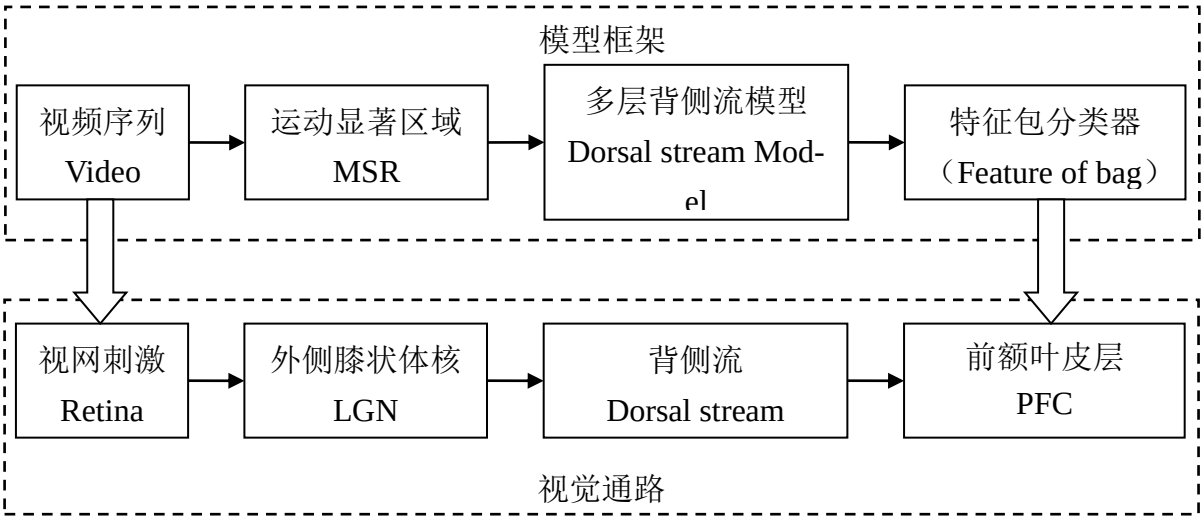

图 3-1 生物启发模型框架

3.1 BIM-STIP 检测

基于生物启发模型的时空兴趣点检测模拟 LGN 和背侧流通路的生物视皮层处理过程，如下图所示。运动显著区域是用于模拟 LGN，粗略找到感兴趣区域以减少背侧流模型的计算量。通过多层的背侧流模型，包括层 V1 区，V2 区和 MT 区，精确地提取出感兴趣的时空兴趣点。

3.1.1 LGN：空间注意调节

根据视觉注意力，MSR 模型用于模拟 LGN 的空间注意调节机制，粗糙提取视频序列

中的时空兴趣点。由于背侧流模型采用多尺度卷积模板，具有时间复杂度高的缺点，引入MSR 模拟 LGN 旨在降低时空兴趣点检测的时间复杂度，采用逐步求精的方法，初步定位时空兴趣点的位置。因为时空兴趣点检测将在紧接的背侧流模型中做进一步求精，因此在这个框架下采用一种简单的像素变化概率图（PCPM）[38]的方法提取 MSR，公式如下：

$$P(x, y, t) = \eta \times P(x, y, t-1) + (1-\eta) \times \left| I(x, y, t) - I(x, y, t-1) \right| \quad (3\text{-}1)$$

其中η为遗忘因子，$I(x, y, t)$为原始灰度图像。

为了降低算法时间复杂度，引进积分图（Integral image）[5]的思想对局部运动能量（Location Motion Energy，LME）的统计做算法上的加速。积分图又称总和面积表，是一种对一个网格中的矩形子区域中和计算的快速且有效的数据结构和算法。积分图的每一点的值是原图中对应位置的左上角区域的所有值得和：

$$PI(x, y, t) = \sum_{(x,y)=(0,0)}^{(x,y)} P(x, y, t) \quad (3\text{-}2)$$

因此，如图 3-2 所示，局部运动能量（LME）可以表示为：

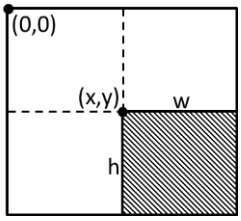

图 3-2 积分图原理

$$LME(x, y, t, w, h) = \frac{PI(x+w, y+h, t) + PI(x, y, t) - PI(x+w, y, t) - PI(x, y+h, t)}{w \times h} \quad (3\text{-}3)$$

通过 LME 统计粗略确定感兴趣点位置，再采用背侧流模型对确定的粗糙区域做进一步求精，可以很大程度上减少背侧流模型的计算量，降低整体框架计算时间复杂度。

3.1.2 V1 区：初级视觉特征提取

如上章 2.3.2 所述，生物学研究表明 V1 区各种性能的初级视觉细胞与 Gabor 能量滤波模板相似。V1 区的细胞响应可用一系列的 Gabor 滤波器对输入的原始灰度图像进行处理所得，Gabor 滤波器的其表达式如下[39]：

$$G(x, y, \theta, s(\delta, \lambda, \gamma)) = \exp\left(-\frac{X^2 + \gamma^2 Y^2}{2\delta^2}\right) \times \cos\left(\frac{2\pi}{\lambda} X\right) \quad (3\text{-}4)$$

，其中$X = x\cos\theta + y\sin\theta$，$Y = -x\cos\theta + y\sin\theta$，$(x, y)$为 Gabor 滤波器卷积核的坐标，$\theta$为方向角度，$s$为尺度空间大小（其中$\delta$为带宽，$\lambda$为波长，$\gamma$为纵长比）。

根据灵长动物初级视觉细胞感受野的特性，在此选取一系列尺寸大小空间从 7 到 37 的 Gabor 滤波器模板，对 V1 区细胞感受野（Receptive Field，RF）ξ进行建模，参数选择如

表 3-1[39, 40]。其中 Gabor 滤波器共选取 4 个方向角度θ，16 个尺度空间s，共 16×4=64 个子图，如图 3-3 所示。相邻的两级尺度空间分布在同一个子带ε中，共组成 8 个子带。V1 区的细胞响应可以描述为

$$V1(x, y, t, \theta, s) = I(x, y, t) * G(x, y, \theta, s) \quad (3\text{-}5)$$

其中，$V1(x, y, t, \theta, s)$ 是 V1 区的输出结果，$I(x, y, t)$是视频中的原始灰度图像。

表 3-1 基于生物启发模型的时空兴趣点框架参数选择

ε	1		2		3		4		5		6		7		8	
s	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
ξ	7	9	11	13	15	17	19	21	23	25	27	29	31	33	35	37
δ	2.8	3.6	4.5	5.4	6.3	7.3	8.2	9.2	10.2	11.3	12.3	13.4	14.6	15.8	17	18.2
λ	3.5	4.6	5.6	6.8	7.9	9.1	10.3	11.5	12.7	14.1	15.4	16.8	18.2	19.7	21.2	22.8
γ	0.23	0.28	0.32	0.37	0.41	0.46	0.51	0.55	0.60	0.64	0.69	0.74	0.78	0.83	0.87	0.92
Σ	8		12		16		20		24		28		32		36	
θ	0 $\quad \frac{\pi}{4} \quad \frac{\pi}{2} \quad \frac{3\pi}{4}$															

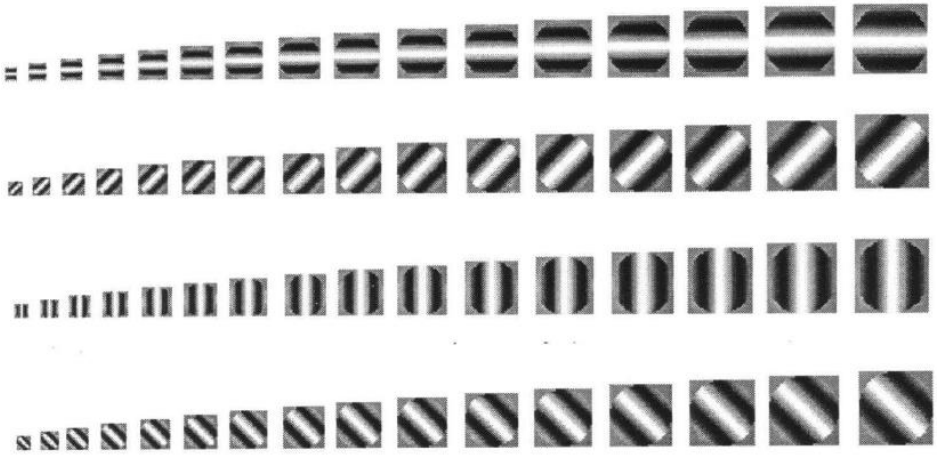

图 3-3 各尺度空间 Gabor 滤波模板样式

3.1.3 V2 区：尺度不变性

Ilan 等[41]通过复杂的细胞（V2）空间整合特性的研究，结果表明复杂细胞存在对于简单视皮层细胞输出信号的最大池操作（Maximum Pool Operation）。复杂细胞有更大的感受野，对应于图像特征表现为位置不变性；此外复杂细胞融合各个方向的纹理、边缘信息，表现为方向不变性；此外，复杂细胞适应比简单细胞更广泛的空间频率，对应于图像处理的尺度不变性。

由于位移和方向信息将在 MT 做进一步处理，因此在 BIM-STIP 框架下在 V2 区仅做尺度竞争操作：

$$V2(x, y, t, \theta, \varepsilon) = Max(V1(x, y, t, \theta, s = 2\varepsilon - 1), V1(x, y, t, \theta, s = 2\varepsilon)) \quad (3\text{-}6)$$

位置和方向竞争将在 MT 区中高级运动分析之后进行操作。

3.1.4 MT 区：高级运动分析

Reichardt 等提出的初级运动检测器（Elementary Motion Detector，EMD）等价于灵长类动物的视觉通路的运动检测单位[42]。Ron 等[37]的对初级运动检测器的研究表明，现实中的 Reichardt 检测器可以在自然的视觉环境中提供准确速度估计值。在此框架下，采用 EMDs 用来模拟 MT 层，很好地结合了空间和时间信息。

EMD 是一种二维的运动检测器，其结构可由下式和图 3-4 表示。

$$R(x,t) = F_A(t-\tau) \times F_B(t) - F_A(t) \times F_B(t-\tau) \tag{3-7}$$

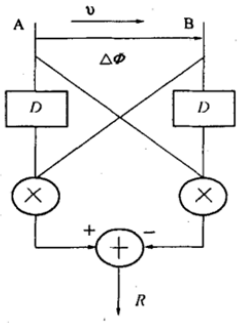

图 3-4 初级运动感受器

其中，$F_A(t) = F(x,t)$，$F_B(t) = F(x+\Delta\varphi, t)$，$F(x,t)$是输入信号，$\tau$ 是时间偏移量，$\Delta\varphi$ 是信号平移量。MT 层输入信号来自 V2 层，包括 8 个子带和 4 个方向的响应结果。在此，选取 4 个方向的初级运动检测器对应处理 V2 层不同方向的输入响应。当 EMD 的平移方向与 Gabor 滤波器的方向是正交的，此时的 EMD 对该方向的运动是最敏感的。因此，MT 层响应信号可以被表示为如下所示：

$$F'(x,y,t,\theta,\Delta\Phi) = V2(x+\Delta\Phi\sin\theta, y+\Delta\Phi\cos\theta, t, \theta, \varepsilon) \tag{3-8}$$

$$MT(x,y,t,\theta,\varepsilon) = F'_A(t-\tau) \times F'_B(t) - F'_A(t) \times F'_B(t-\tau) \tag{3-9}$$

其中，$F'_A(t) = F'(x,y,t,\theta,\varepsilon,0)$ ，$F'_B(t) = F'(x,y,t,\theta,\varepsilon,\Delta\varphi)$。根据公式可知，在此存在两个参数选择（$\tau$ 和$\Delta\varphi$）。根据生物学研究表明，人眼的视觉残留时间是 0.17s，对应于 25 fps 的数字视频图像可算得$\tau = 4.25 \approx 4$帧；$\Delta\varphi$参数的选取依据 Gabor 滤波器的主瓣宽度，则 $\Delta\varphi = 4$。

3.1.5 BIM-STIP 检测策略

通过生物启发模型各个步骤，将得到生物启发响应图。在此，提出了一系列的检测策

略以精确定位时空兴趣点，其中包括方向竞争、位置竞争、局部求精和全局最优。BIM 的时空兴趣点检测策略将显示在算法 3-1。

首先，对于 MT 层的输出响应作方向竞争，以保证方向不变性，其方向竞争（Orientation Competition，OC）等式表示如下：

$$OC(x, y, t, \varepsilon) = Max\left\{MT\left(x, y, t, \theta = (\frac{\pi}{4}, \frac{\pi}{2}, \frac{3\pi}{4}, \pi), \varepsilon\right)\right\} \quad (3\text{-}10)$$

其次，初步兴趣点通过不同尺度空间区域的极值检测，模拟神经元间的位置竞争。其中，极值检测的空间区域选择参见表 3-1 中∑值。此外，对初步兴趣点进行局部求精，全局最优等操作得到视频中具有强表征的时空兴趣点。

（1）局部求精过程是通过比较相近的两个时空兴趣点，当两个粗糙兴趣点距离小于某一阈值，剔除掉响应强度较小的兴趣点，保留强响应点。此操作与神经元间的位置竞争相似，为更好地减少弱兴趣点的影响。

（2）全局最优过程是当同一视频帧中存在过多的时空兴趣点，提取视频帧中响应强度较强的兴趣点，剔除响应强度较小的点。此操作可以确保不发生时空兴趣点爆炸性增长，并削减弱特征对于分类的影响。

算法 3-1 BIM-STIP 检测策略

输入： $OC(x, y, t, \varepsilon)$， $LME(x, y, w, h)$
输出：vector <KeyPoint> points
算法伪代码：

```
//局部最大响应点检测
vector <KeyPoint>   coarse
for ε=1:8
    w = h = ∑(ε)
    for x = 1 : image_width, y = 1 : image_hight
        if LME(x, y, w, h) < LME_threshold
            continue
        end if
        compete OC(x, y, t, ε)
        find the max respond OC(x, y, t, ε) position Keypoint(x, y, ε) in
(x, y, w, h)
        coarse. push( Keypoint(x, y, ε))
    end for
end for
//局部兴趣点求精
for i = 1 : coarse.size
    for i = 1 : coarse.size − i
```

```
        if Distance(coarse[i],coarse[j]) < D _ threshold
            if coarse[i].respond > coarse[j].respond
                coarse. pop(j)
            else
                coarse. pop(i)
            end if
        end if
    end for
end for
vector <KeyPoint> refine=coarse
//全局最优点
grade-down sort(refine)
for i = 1 : Maxnum
    points. push(refine[i])
end for
```

3.2 BIM-STIP 描述符提取

为了描述时空兴趣点，本文提出基于生物启发模型的时空描述符。根据各个时空兴趣点的空间尺度 s，我们定义一个 $\Sigma \times \Sigma \times T$ 的时空立方体，其中 Σ 的选取根据表 3-1，而 T 在本文中采用经典时空网格划分参数，选取为 8 帧。首先，时空立方体将被划分为 $M \times M \times N$ 个子块，其中 M 和 N 分别对应空间和时间方向的子块划分个数，在这里选取 $M = 2$，$N = 2$。如图 3-5 所示，各个子块为对应原始灰度图像时空兴趣点的临域。与生物启发模型时空兴趣点检测方法类似，模拟各层视觉皮层对时空立方体进行层次化响应提取。

首先，对时空立方体提取 V1 区响应。为了使描述符更好地保留细节特征，有别于生物启发模型的兴趣点检测采用多尺度空间方法，在此选取生物模型最小尺度空间 $s=1$（最小感受野 $\xi = 7$）以保证描述符对细节的表征能力，公式如下：

$$VR(x, y, t, \theta) = I(x, y, t) * G(x, y, \theta, s = 1) \qquad (3\text{-}11)$$

其次，采用初级运动感受器模拟 MT 区对 V1 区输出响应做处理。初级视觉特征被传输到 EMD，为了更好地得到时空细节，在这里选取最小时间偏移量 $\tau=1$。空间偏移量 $\Delta\varphi=4$，此为最小尺度空间所对应 Gabor 滤波卷积模板的主瓣宽度。

$$MTR(x,y,t,\theta)=F'_A(t-1)\times F'_B(t)-F'_A(t)\times F'_B(t-1) \qquad (3\text{-}12)$$

其中，$F'_A(t)=VR(x,y,t,\theta)$，$F'_B(t)=VR(x+4\sin\theta,y+4\cos\theta,t,\theta)$

对于初级运动感受器，响应强度的符号代表运动的方向。即因初级运动检测器对于统

计 8 个子块分 4 个方向和正负号统计的响应强度，得到 8×4×2=64 维的统计直方图，构成 64 维 BIM-STIP 特征描述符。最后，采用最大最小值法对特征向量进行归一化。

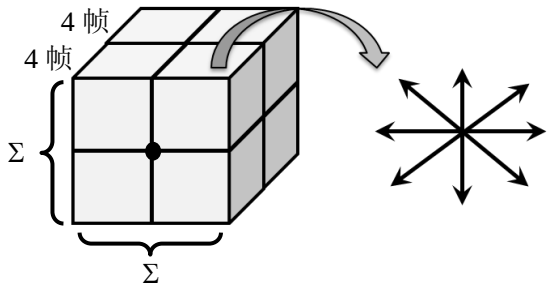

图 3-5 时空网格划分与特征提取

3.3 特征包分类器

特征包（Bag of Features，BoF）模型[43]，也被称为词袋（Bag of Words，BoW）模型。在自然语言处理与信息检索领域，特征包模型得到了广泛应用，随着机器视觉的发展，其在目标识别、人体行为识别等研究中取得很好的效果。

本文采用特征包模型来对视频序列进行建模，从而实现人体行为的识别。基于特征包模型的人体行为识别和基于特征包模型的图像建模相似，将一个视频看作一个文本，在视频中提取特征向量构建成特征包。对视频图像中提取的局部时空特征，以此建立特征包词典。基于特征包模型的人体行为识别通常包括 3 个步骤：

（1）提取所有视频，包括训练集和测试集中的时空兴趣点，根据时空兴趣点所属感受野，在兴趣点的领域内提取局部时空特征来描述该时空兴趣点，建立描述子矢量。

（2）对训练集中的所有特征描述符进行 K-means 聚类[44]，聚类中心作为词典单词，把所得词汇合并组成特征包。

（3）采用 FLANN（Fast Library for Approximate Nearest Neighbors）[45]方法对测试集的时空描述符与特征包做比较，计算最小欧式距离，找到测试集中各个描述符所对应的特征包。最后得到测试视频关于特征包的统计直方图，直方图反映每个视频的词频分布。

（4）对统计直方图进行最大归属分类，选取统计数最多的特征包所属行为类别作为测试视频的分类结果。

3.4 本章小结

本章从生物启发模型的角度建立起全新的时空兴趣点人体行为识别框架，其中包括时空兴趣点检测和时空特征的提取。详细介绍本文时空兴趣点框架与生物启发模型的对应关

系，逐层介绍生物启发模型的时空兴趣点框架的各层次数学模型与生物学依据。此外，首次引入初级运动检测器作为 MT 区数学模型，构建整条灵长动物视皮层背侧通路的结构，实现对运动信息与图形信息的有机融合。

第四章 实验结果与分析

4.1 常用测试数据库

4.1.1 Weizmann 人体行为数据集

Weizmann 人体行为数据集[46]包括 10 种行为：弯腰（bend）、交叉跳（jack）、跳走（jump）、原地跳（pjump）、跑（run）、侧跳（side）、单脚跳（skip）、走（walk）、单手挥（one-hand wave）、双手挥（two-hands wave），每种行为由 9 个人完成，每个视频中只有一个人。背景静止，背景简单，摄像头视角不变，存在光线变化。其样例图像如下图：

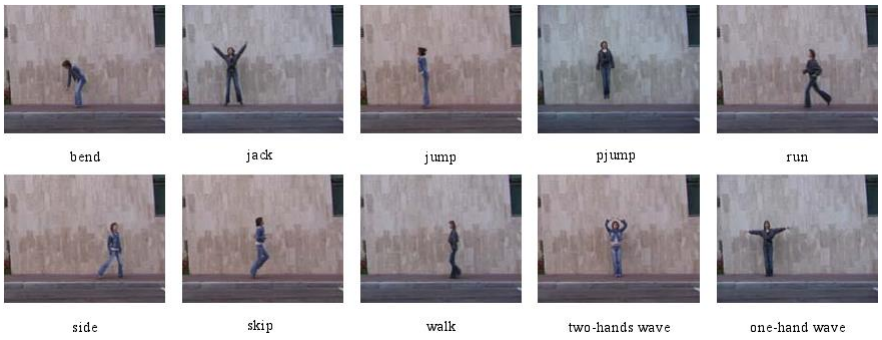

图 4-1 Weizmann 人体行为数据集

4.1.2 KTH 人体行为数据集

KTH 人体行为数据集[47]包含 6 种行为：走路（walking）、慢跑（jogging）、快跑（running）、拳击（boxing）、挥手（hand-waving）、鼓掌（hand-clapping），每种行为由 25 个不同的人去完成；包含 4 种场景：室内、户外、户外摄像机变焦、户外不同衣服穿着，共计 6×25×4=600 个视频。在该数据集中，不同人对于同种行为存在较大的差异，户外场景拍摄中摄像头存在微小抖动，室内场景中存在阴影干扰。其样例如图 4-2 所示。

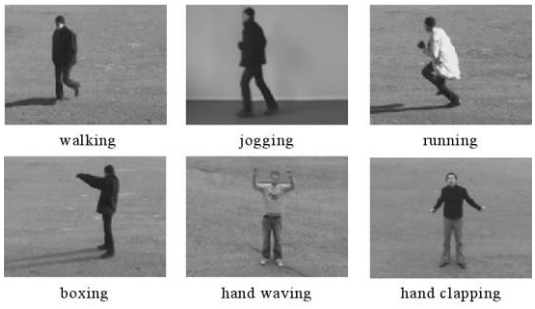

图 4-2 KTH 人体行为数据集

4.2 BIM-STIP 检测实验

如上一章所述，基于生物启发模型的时空兴趣点检测方法需要确定一些阈值，其中包括 LME 粗糙检测阈值、求精临域范围。LME 粗糙检测阈值的选取对于兴趣点检测结果影响不大，但选取好的 LME 阈值有利于降低时间复杂度，减少计算开销。求精临域范围选取与时空兴趣点所处的尺度空间有关。在本文中，选取感受野尺寸的 1/2 作为求精临域的范围。时空兴趣点的数目与人体行为分类结果和算法运行效率有着密切联系。当时空兴趣点过少时，则所提取兴趣点并不能充分表征视频序列中的人体行为特征，造成分类准确率的下降；如果时空兴趣点太多，则增加计算时长和存储开销，降低算法运行效率。在实验中，本文的时空兴趣点检测方法在 Weizmann 数据集上检测到的兴趣点。

如下图 4-3 所示，第一行为本文基于生物启发模型的时空兴趣点检测方法在 Weizmann 数据集检测到的结果，第二行为经典 Harris3D 检测器[48]在 Weizmann 数据集检测到的时空兴趣点。由图可知，基于生物启发模型的时空兴趣点方法在 Weizmann 数据集上检测到的时空兴趣点明显比 Harris3D 检测器更为丰富，一般情况下，检测到的时空兴趣点越丰富，最终的行为分类效果越好。

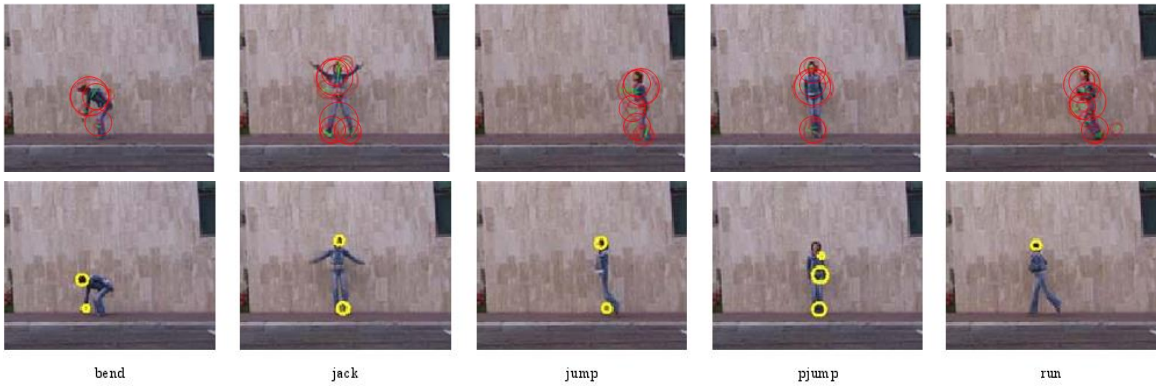

图 4-3 时空兴趣点检测方法比较

4.3 BIM-STIP 框架行为识别实验结果

依据本文提出的时空兴趣点检测方法检测局部时空特征，并计算基于生物启发模型的时空描述子，采用特征包模型进行行为分类。我们的算法在以上的两个数据集上取得较好的识别精度，Weizmann 数据集平均分类精度分别为 87.1%，KTH 数据集为 88.12%，其混淆矩阵分别如图 4-3，图 4-4 所示。

如图 4-4 所示，采用 BIM-STIP 框架和特征包模型对 Weizmann 数据集的分类。其中选取 6 个人的行为作为特征包训练集，3 个人的动作行为作为测试集，特征包尺寸为 1000 个单词，交叉实验 3 次得到下图混淆矩阵。由混淆矩阵可得，其中五种行为：bend、jack 、

pjump、one-hand wave、two-hands wave，分类精度最高，均达到了 100%的识别率；识别准确率较低的人体行为是奇偶 skip，准确率为 50%，与其最容易混淆的行为是 run。

	bend	jack	jump	pjump	run	side	skip	walk	wave1	wave2
bend	1.00									
jack		1.00								
jump			0.88		0.11	0.11				
pjump				1.00						
run			0.10		0.80		0.10			
side			0.11			0.78	0.11			
skip			0.10		0.40		0.50			
walk							0.10	0.90		
wave1									1.00	
wave2										1.00

图 4-4 Weizmann 数据库识别结果混淆矩阵

如图 4-5 所示，对 KTH 数据集进行行为分类实验，其中选取 16 个人的行为作为特征包提取训练集，9 个人的动作行为作为测试集，特征包尺寸选取为 12000 个单词，交叉实验 3 次得到该混淆矩阵。在 KTH 数据集中，三种行为的识别分类准确率较高，分别为 boxing、hand-clapping、hand-waving、walking，均达到了 90%以上； jogging 和 running 这两种行为的识别准确率较低。从测试视频中可知，jogging 和 running 在运动细节上有很大的相似，对比其他时空兴趣点方法可知，在其他的时空兴趣点方法中，joging 和 running 同样是易混淆的两种行为。

	box	clap	wave	jog	run	walk
box	0.94	0.06				
clap	0.05	0.95				
wave	0.04	0.05	0.92			
jog				0.64	0.27	0.09
run				0.10	0.89	0.01
walk				0.06		0.94

图 4-5 KTH 数据库识别结果混淆矩阵

4.4 实验结果对比

4.4.1 Weizmann 数据库对比分析

对于 Weizmann 数据库，将本文的算法与其他局部时空兴趣点方法作比较，如下表所示：

表 4-1 Weizmann 数据库实验对比

方法	准确率
BIM-STIP	87.1%
SCG(CVPR'07)[49]	72.8%
3D-SIFT(ICPR'04)[10]	82.6%
Spin Images(CVPR'08)[50]	74.2%
ST(CVPR'08)[50]	68.4%
3d-gradients (BMVC'08)[51]	84.3%
STW(IJCV'08)[52]	**90.0%**

如上图所示，BIM-STIP 方法在该数据库有较好的表现，其中在 one-hand wave、two-hands wave 这两个动作的识别效果上相比于其他时空兴趣点方法有明显的提高。可见本文方法采用生物启发模型的优势。其中，行为分类效果略低于 STW，但在 KTH 数据集下表现明显好于 STW 特征。

4.4.2 KTH 数据库对比分析

对于 KTH 数据库，将本文方法与其他局部时空兴趣点方法做比较，实验结果如下表所示：

表 4-2 KTH 数据库实验对比

方法	准确率
BIM-STIP	88.6%
LF(ICCV'04)[47]	71.7%
VF(ICCV'05)[53]	62.9%
Dollar(VS-PETS'05)[54]	81.2%
DSM(ICCV'07)[55]	84.7%
E-SURF(ECCV'08)[12]	81.4%
STW(IJCV'08)[52]	83.3%
SPREF(CVPR'08)[56]	86.6%
HOG/HOF(CVPR'08)[3]	**91.8%**
MHCF(PAMI'11)[57]	89.8%

如上表所示，BIM-STIP 方法得到了 88.6%的识别精度。对比与 2008 年以前的研究成果

有了很大的提高。然而，对比与 HOG/HOF 方法仍有不及。对比与 HOG/HOF 方法，本文方法有以下优点。首先，本文所提方法不存在参数选取问题，HOG/HOF 方法对于不同的测试数据库需要对参数做相应的调整以达到效果最优，且该论文中的实验表明不同参数选取会对实验结果产生重大影响。其次，本文方法在时空兴趣点检测上采用逐步求精策略，特征维度为 64 维，计算时间复杂度较低。HOG/HOF 方法中 HOG 特征提取和 HOF 提取得到 192 维特征，算法相对复杂。最后，本文方法引入生物启发模型，在后续的研究中有较为广阔的提升空间。

4.5 本章小结

本章通过 Weizmann 和 KTH 数据库对本文提出的基于生物启发模型的时空兴趣点框架进行验证。实验表明，在该框架下有较好的表现结果，对于，相比于现有算法存在以下优点：1、算法采用逐步求精思想，时间复杂度低；2、本框架下所有参数均建立于生物启发模型框架下，相同参数选择对于不同数据库均有较好效果；3、基于生物启发模型的时空描述符仅 64 维，维度远低于现行方法，避免维度灾难，可使用简单的分类方法进行行为识别。

第五章 论文总结及展望

5.1 论文总结

人体行为识别是机器视觉领域近十年来最具潜力的研究方向。在本论文中提出了一种新的时空兴趣点检测和时空特征描述符提取方法。引入生物启发模型为时空兴趣点框架提供一种全新的解决思路和方法，现对论文总结如下：

首先，本论文从生物学研究角度简单介绍了灵长动物视觉皮层的功能，介绍视觉通路的概念和两条主要的视觉通路。其次，主要以背侧流为主，介绍各个层次皮层在生物学上的功能和视觉刺激信号的在各层之间的流动方向。最后，根据现有的生物学研究依据，为各个层次皮层信号处理机制选取合适的数学模型。在本文引入初级运动检测器模拟 MT 区，构建背侧流中的完整通路模型。

其次，应用生物启发模型中背侧流通路的生物研究成果，结合传统时空兴趣点人体行为识别框架。构建基于生物启发模型的时空特征点方法，其中包括时空兴趣点检测和时空特征的提取。详细介绍本文时空兴趣点框架与生物启发模型的对应关系，逐层介绍生物启发模型的时空兴趣点框架的各层次数学模型与生物学依据。

最后，采用 KTH 和 Weizmann 人体行为数据库对本文方法进行评价，实验表明基于生物启发模型的时空兴趣点方法有着较好效果。由于引入生物启发模型，层次化提取不变的时空特征，因此能很好地保证特征的“不变性”与“区分性”。

5.2 展望

在日新月异到的科技发展中，人们是越来越需要更加智能化的信息处理方式，在这样的大环境下，人体行为识别作为机器视觉领域的重要一块，也是担当着比较重要的角色，需要更有力的发展。但是目前行为识别的发展仍处于起步阶段，仍处于理论研究阶段，没有很好地与工程性应用做有机的结合。其主要原因是因为复杂环境和人体行为多样性等不确定的因素的影响，造成实际应用的时候，识别度精度不高。这是需要不断地去努力克服现有的难题，实现技术上的突破。

结合生物视觉系统进行视频特征提取是机器视觉研究领域的一个新方向和突破口。引入生物启发模型进入人体行为识别是一个跨学科、跨领域的研究工作，是一个具有长远研究意义和极富挑战性的课题。现如今，国内外各个研究团队从不同的切入点进行探索，应用现有的生物学研究成果于机器视觉中，旨在提取更具“区分性”和“不变性”的特征。

生物学研究成果很大程度上推动着机器视觉向更高的程度发展，逐步摆脱传统信号处理领域的局限，为视频中的人体行为识别提供了深度挖掘的方向，推动着这个研究领域的进步。然而，现有的研究仍处于起步阶段，同时也面临着如何将生物学研究与视觉感知研究有机结合的困境。

本论文为生物启发模型与时空兴趣点特征方法的结合提供了一种有效方法，在下一步的研究中，将着重于联合更高层次的生物模型对特征进行更有效的融合。如引入慢特征分析法（Slow Feature Analysis），稀疏编码（Sparse Coding），深度学习（Deep Learning）等大脑生物学模型以提高生物启发框架的准确率。此外，对于现有 Gabor 滤波存在时间复杂度高的缺点，此后将着重于改进 Gabor 滤波器使用卷积模板求解的方法，采用近似 Gabor 滤波的方法提高算法效率。

参考文献

[1] Ayers D, Shah M. Monitoring human behavior from video taken in an office environment[J]. Image and Vision Computing. 2001, 19(12): 833-846.

[2] Duchenne O, Laptev I, Sivic J, et al. Automatic annotation of human actions in video[C]. 2009.

[3] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies[C]. 2008.

[4] Lindeberg T. Scale invariant feature transform[J]. Scholarpedia. 2012, 7(5): 10491.

[5] Bay H, Tuytelaars T, Van Gool L. Surf: Speeded up robust features[M]. Computer Vision--ECCV 2006, Springer, 2006, 404-417.

[6] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]. 2005.

[7] Li X. HMM based action recognition using oriented histograms of optical flow field[J]. Electronics Letters. 2007, 43(10): 560-561.

[8] Liang W, Suter D. Learning and Matching of Dynamic Shape Manifolds for Human Action Recognition[J]. Image Processing, IEEE Transactions on. 2007, 16(6): 1646-1661.

[9] Chen H, Chen H, Chen Y, et al. Human action recognition using star skeleton[C]. New York, NY, USA: ACM, 2006.

[10] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition[C]. New York, NY, USA: ACM, 2007.

[11] Guha T, Ward R K. Learning Sparse Representations for Human Action Recognition[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2012, 34(8): 1576-1588.

[12] Willems G, Tuytelaars T, Gool L. An Efficient Dense and Scale-Invariant Spatio-Temporal Interest Point Detector[M]. Computer Vision – ECCV 2008, Forsyth D, Torr P, Zisserman A, Springer Berlin Heidelberg, 2008: 5303, 650-663.

[13] Cho K, Cho H, Um K. Human Action Recognition by Inference of Stochastic Regular Grammars[M]. Structural, Syntactic, and Statistical Pattern Recognition, Fred A, Caelli T, Duin R, et al, Springer Berlin / Heidelberg, 2004: 3138, 388.

[14] Kitani K M, Sato Y, Sugimoto A. Deleted interpolation using a hierarchical Bayesian grammar network for recognizing human activity[C]. 2005.

[15] Haritaoglu I, Harwood D, Davis L S. W^4: real-time surveillance of people and their activities[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2000, 22(8): 809-830.

[16] Davis J W, Taylor S R. Analysis and recognition of walking movements[C]. 2002.

[17] Weiming H, Xie D, Tieniu T, et al. Learning activity patterns using fuzzy self-organizing neural network[J]. Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on. 2004, 34(3): 1618-1626.

[18] Jhuang H, Serre T, Wolf L, et al. A Biologically Inspired System for Action Recognition[C]. 2007.

[19] Escobar M, Kornprobst P. Action recognition via bio-inspired features: The richness of center–surround interaction[J]. Computer Vision and Image Understanding. 2012, 116(5): 593-605.

[20] Panizza B. Osservazioni sul nervo ottico[J]. GI r. lst Lombardo Sci. Lett. Arti. Bibl. Ital. 1855, 7: 237-252.

[21] Barlow H B. Summation and inhibition in the frog's retina[J]. The Journal of physiology. 1953, 119(1): 69-88.

[22] Mcilwain J T, Buser P. Receptive fields of single cells in the cat's superior colliculus[J]. Experimental

Brain Research. 1968, 5(4): 314-325.
[23] Hubel D H, Wiesel T N. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex[J]. The Journal of physiology. 1962, 160(1): 106.
[24] Mishkin M, Ungerleider L G, Macko K A. Object vision and spatial vision: two cortical pathways[J]. Trends in neurosciences. 1983, 6: 414-417.
[25] Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex[J]. Nature neuroscience. 1999, 2(11): 1019-1025.
[26] Li C. Integration fields beyond the classical receptive field: organization and functional properties[J]. Physiology. 1996, 11(4): 181-186.
[27] 李宁. 基于视觉认知的人体行为特征提取模型研究[D]. 北京交通大学, 2010.
[28] Grill-Spector K, Malach R. The human visual cortex[J]. Annu. Rev. Neurosci. 2004, 27: 649-677.
[29] Mcalonan K, Cavanaugh J, Wurtz R H. Guarding the gateway to cortex with attention in visual thalamus[J]. Nature. 2008, 456(7220): 391-394.
[30] Bregler C. Learning and recognizing human dynamics in video sequences[C]. 1997.
[31] Gabor D. Theory of communication. Part 1: The analysis of information[J]. Electrical Engineers-Part III: Radio and Communication Engineering, Journal of the Institution of. 1946, 93(26): 429-441.
[32] Granlund G H. In search of a general picture processing operator[J]. Computer Graphics and Image Processing. 1978, 8(2): 155-173.
[33] Knutsson H, Wilson R, Granlund G. Anisotropic Nonstationary Image Estimation and Its Applications: Part I--Restoration of Noisy Images[J]. Communications, IEEE Transactions on. 1983, 31(3): 388-397.
[34] Daugman J G, Others. Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters[J]. Optical Society of America, Journal, A: Optics and Image Science. 1985, 2(7): 1160-1169.
[35] Reichardt W. Autocorrelation, a principle for the evaluation of sensory information by the central nervous system[J]. Sensory communication. 1961: 303-317.
[36] Adelson E H, Bergen J R. Spatiotemporal energy models for the perception of motion[J]. J. Opt. Soc. Am. A. 1985, 2(2): 284-299.
[37] Dror R O, O'Carroll D C, Laughlin S B. Accuracy of velocity estimation by Reichardt correlators[J]. JOSA A. 2001, 18(2): 241-252.
[38] Qin Y, Li H, Liu G, et al. Human action recognition using PEM histogram[C]. 2010.
[39] Serre T, Wolf L, Poggio T. Object recognition with features inspired by visual cortex[C]. 2005.
[40] Serre T, Wolf L, Bileschi S, et al. Robust object recognition with cortex-like mechanisms[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2007, 29(3): 411-426.
[41] Lampl I, Ferster D, Poggio T, et al. Intracellular measurements of spatial integration and the MAX operation in complex cells of the cat primary visual cortex[J]. Journal of neurophysiology. 2004, 92(5): 2704-2713.
[42] Reichardt W, Guo A. Elementary pattern discrimination (behavioural experiments with the fly Musca domestica)[J]. Biological cybernetics. 1986, 53(5): 285-306.
[43] Lewis D D. Naive (Bayes) at forty: The independence assumption in information retrieval[M]. Machine learning: ECML-98, Springer, 1998, 4-15.
[44] Leung T, Malik J. Representing and recognizing the visual appearance of materials using three-dimensional textons[J]. International Journal of Computer Vision. 2001, 43(1): 29-44.
[45] Muja M, Lowe D G. Fast approximate nearest neighbors with automatic algorithm configuration[C]. 2009.

[46] Blank M, Gorelick L, Shechtman E, et al. Actions as space-time shapes[C]. 2005.
[47] Schuldt C, Laptev I, Caputo B. Recognizing human actions: a local SVM approach[C]. 2004.
[48] Laptev I, Lindeberg T. Space-time interest points[C]. 2003.
[49] Niebles J C, Fei-Fei L. A hierarchical model of shape and appearance for human action classification[C]. 2007.
[50] Liu J, Ali S, Shah M. Recognizing human actions using multiple features[C]. 2008.
[51] Klaser A, Marszalek M. A spatio-temporal descriptor based on 3D-gradients[J]. 2008.
[52] Niebles J C, Wang H, Fei-Fei L. Unsupervised learning of human action categories using spatial-temporal words[J]. International Journal of Computer Vision. 2008, 79(3): 299-318.
[53] Ke Y, Sukthankar R, Hebert M. Efficient visual event detection using volumetric features[C]. 2005.
[54] Doll A R P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features[C]. 2005.
[55] Nowozin S, Bakir G O K, Tsuda K. Discriminative subsequence mining for action classification[C]. 2007.
[56] Wong S, Cipolla R. Extracting spatiotemporal interest points using global information[C]. 2007.
[57] Gilbert A, Illingworth J, Bowden R. Action recognition using mined hierarchical compound features[J]. Pattern Analysis and Machine Intelligence, IEEE Transactions on. 2011, 33(5): 883-897.

致　谢

时光似箭，岁月如梭，弹指一挥间四年即逝。值此拙文完成之际，回想起在华南理工大学度过的难忘的四年，心中颇为感慨，衷心地向各位亲友、老师和同学致以诚挚的敬意和谢意。

首先，感谢家人对我无微不至的关怀。感谢父母的养育之恩，父母的期望指引着我二十三年的人生路。感谢我的爷爷奶奶从小对我的悉心照顾，我的成长离不开他们的关爱。此外，特别要感谢我的兄长蔡博昆，23 年来与我相伴，一起学习，一起生活，一起玩耍，一同长大；此外，当我在本论文完成过程中遇到一些困惑时，昆哥总为我提供一些开阔性的思路。

其次，感谢华南理工大学电子与信息学院诸位老师的悉心培养和指导。老师们渊博的知识、严谨的治学态度以及人格的魅力深深激励着我，让我受益匪浅。特别感谢我的指导老师徐向民教授，在我过去一年中的关心和帮助。在毕设的过程中，徐老师给了我不少的思想上的指导和精神上的支持，使我可以更顺利开展研究，如期完成我的毕设。

另外，感谢实验室的师兄、师姐们在学术上对我的指导与支持，感谢实验室的朋友们在研究上的交流与互助。特别感谢郭锴凌师兄对我研究方向的指导和学术上的帮助。正是在郭师兄的指导下，我走进了生物启发模型与机器视觉的领域。

最后，感谢 09 集成班全体同学，你们的陪伴让我度过了终身难忘的大学四年，我们一起学习，一起生活，一起笑过，傻过，疯癫过的大学。感谢美丽的华园，在这里有一起上过课的教学楼，一起蜗居过的自习室，一起通宵过的实验室，一起流汗过的球场……此外，特别感谢 233 的舍友们，感谢成哥、秋哥和远在异国他乡的胖子。

木棉依旧花开花落，棉絮飞扬；中山像依旧风吹雨打，日晒雨淋；校巴依旧海纳百川，有容乃大；北区小路依旧坎坷颠簸，曲折艰辛；校园网依旧时断时续，时续时断；北二饭堂依旧食材丰富，菜量十足……四年间，似乎一切都没有变，回想起四年前踏入华园的那一刻，似乎还是那么的清晰可见。然而，一切都变了，我已经不是四年前的我，四年的同学也将各奔东西。感谢所有关心我的人，路漫漫其修远兮，我会在大家的支持下越走越远！