基于深度学习的人体行为识别方法研究

作者姓名 汪厚峄

指导教师姓名、职称 同　鸣　教授

申请学位类别 工学硕士

学校代码 10701　　学　号 1402121056
分 类 号 TN911.73　　密　级 公 开

西安电子科技大学

硕士学位论文

基于深度学习的人体行为识别方法研究

作者姓名：汪厚峄

一级学科：信息与通信工程

二级学科：信号与信息处理

学位类别：工学硕士

指导教师姓名、职称：同　鸣 教授

学　　院： 电子工程学院

提交日期：2017 年 06 月

Research on Human Action Recognition Method Based on Deep Learning

A thesis submitted to

XIDIAN UNIVERSITY

in partial fulfillment of the requirements

for the degree of Master

in Signal and Information Processing

By

Wang Houyi

Supervisor: Tong Ming Title: Professor

June 2017

西安电子科技大学
学位论文独创性（或创新性）声明

秉承学校严谨的学风和优良的科学道德，本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢中所罗列的内容以外，论文中不包含其他人已经发表或撰写过的研究成果；也不包含为获得西安电子科技大学或其它教育机构的学位或证书而使用过的材料。与我一同工作的同事对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文若有不实之处，本人承担一切法律责任。

本人签名：_____ 日　　期：_____

西安电子科技大学
关于论文使用授权的说明

本人完全了解西安电子科技大学有关保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权属于西安电子科技大学。学校有权保留送交论文的复印件，允许查阅、借阅论文；学校可以公布论文的全部或部分内容，允许采用影印、缩印或其它复制手段保存论文。同时本人保证，结合学位论文研究成果完成的论文、发明专利等成果，署名单位为西安电子科技大学。

保密的学位论文在___年解密后适用本授权书。

本人签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘要

行为识别在人机交互、虚拟现实、视频监控以及视频检索和分析等领域的广泛应用，引起了越来越多研究者的兴趣。行为识别具有重要的学术研究价值和很强的实用价值，是计算机视觉、模式识别和人工智能等领域的研究热点和难点。目前行为识别存在的问题主要包括：由于视角变化、复杂背景以及不同的运动速度和类型，同类行为存在较大的类内差异。此外，一些行为包含相似的运动模式，使得不同类行为具有较小的类间变化，进而引起混淆。同时，高维视频数据引起的特征冗余，摄像头运动和视频的低分辨率进一步增加了提取有效特征和设计鲁棒识别方法的难度。如何从视频中提取有效的特征和设计更为有效的行为识别构架是亟待解决的关键问题，本文对现有的行为识别方法进行分析和总结，并做出以下工作：

首先，对常见的行为识别方法进行分析和总结。针对传统描述子没考虑特征之间联合统计特性的问题，本文在稠密轨迹的基础上，将图像梯度、光流和运动边界的时间导数作为底层运动特征，然后通过计算底层特征之间的协方差矩阵，构造了TBCM(Trajectory Based Covariance Matrix)描述子，充分考虑了特征之间的联合统计特性，进一步提高对复杂环境中行为主体的描述能力。

其次，提出了一种判别性的非线性特征融合方法。本文将类别结构信息引入到KCCA(Kernel Canonical Correlation Analysis)方法的目标函数中，构造了一种新的特征融合方法。该融合方法最大化了全局和局部特征之间的非线性相关性，同时减小了类内差异性，并加大了类间差异性，进一步增强了特征的判别能力。

再次，构造了深度 3D 卷积描述子。本文分别从 C3D(Convolutional 3D)网络中提取各层的特征，并将 fc6 和 fc7 层特征向量串接作为全局特征，pool4 和 pool5 层特征向量串接作为局部特征，通过判别性的非线性特征融合方法，将全局特征和局部特征进行融合，得到一个更加完备和更具鉴别性的深度 3D 卷积描述子。在 UCF-Sports 库和 YouTube 库上，对本文行为识别方法进行验证并与现有方法对比分析。实验结果表明本文方法的有效性。

最后，总结了论文的主要研究内容，并进一步给出了未来的研究方向。

关 键 词：行为识别， 人工特征， 深度学习， 特征融合

ABSTRACT

Action recognition has been widely used in the fields of human-computer interaction, virtual reality, video surveillance, and video retrieval and analysis, which has attracted more and more researchers' attention. Action recognition has important academic research value and strong practical value, and is the research hotspot and difficulty of computer vision, pattern recognition, artificial intelligence and other fields. The exist problems of action recognition are as follow: There are large intra-class variations in the same action class, which may be caused by background clutter, viewpoint change, and various motion speeds and styles; Moreover, some actions include similar motion patterns, which make different classes of actions have less interclass variation and further cause confusion. Meanwhile, feature redundancy caused by high-dimensional video data, camera motion and low resolution of video further increase the difficulty to extract effective features and design robust recognition method. How to extract efficient features from videos and construct a more effective action recognition framework is a key issue to be solved urgently. In this thesis, the existing methods of action recognition are analyzed and summarized, and the following works have been down:

Firstly, the common action recognition methods are analyzed and summarized. As for the problem that traditional descriptors do not consider the joint statistical characteristics between features, the time derivatives of image gradient, optical flow and motion boundary are taken as low-level motion features on the basis of dense trajectory. And the covariance matrix between low-level features is calculated to construct Trajectory Based Covariance Matrix (TBCM) descriptor, which takes full account of the joint statistical characteristics between features and further improves the descriptive power for behavior subject in complex environments.

Secondly, a discriminant nonlinear feature fusion method is proposed. The category structure information is introduced into the objective function of the Kernel Canonical Correlation Analysis (KCCA) to construct a new feature fusion method. This method maximizes the nonlinear correlation between global and local features, reduces the intra-class variation, increases the inter-class variation, and thus further enhances the discriminant ability of features.

Thirdly, deep 3D convolution descriptor is constructed. In this thesis, the feature vectors of any a layer in Convolutional 3D network (C3D) are extracted. The fully connected layer-6 and layer-7 feature vectors are respectively extracted and concatenated to be taken as global feature, and the pooling layer-4 and pooling layer-5 feature vectors are respectively extracted and concatenated to be served as local feature. Through the proposed discriminant nonlinear feature fusion method, global and local features are fused to obtain a more complete deep 3D convolution descriptor. The proposed methods are compared with existing methods and verified on UCF-Sports and YouTube datasets. The experimental results demonstrate the effectiveness of our methods.

Finally, the main works of thesis are summarized, and the future research direction is also given.

Keywords: Action Recognition, Hand-crafted Feature, Deep Learning, Feature Fusion

插图索引

表格索引

符号对照表

符号	符号名称
$\cup$	求并集
max	求最大值
trace(·)	矩阵的迹
det(·)	矩阵的行列式
$\in$	属于
$\|\bullet\|_2$	求 2 范数
arctan(·)	求反正切值
log(·)	求矩阵对数
T(·)	取矩阵的上三角
R	实数集
cov(·)	计算协方差
var(·)	计算方差
min(·)	求最小值
$\sum$	求和运算
$\langle\cdot,\cdot\rangle$	计算内积
rank(·)	矩阵的秩

缩略语对照表

缩略语	英文全称	中文对照
TV-FI	Tsinghua Video Find It	清华大学视频检索引擎
SIFT	Scale-invariant Feature Transform	尺度不变特征变换
SURF	Speed Up Robust Features	加速鲁棒特征
ISA	Independent Subspace Analysis	独立子空间分析
CNN	Convolutional Neural Network	卷积神经网络
RNN	Recurrent Neural Networks	递归神经网络
LSTM	Long Short Term Memory Networks	长短时间记忆网络
TBCM	Trajectory Based Covariance Matrix	基于轨迹的协方差矩阵描述子
C3D	Convolutional 3D	3D 卷积网络
MEI	Motion Energy Images	运动能量图像
MHI	Motion History Images	运动历史图像
HOG	Histogram of Oriented Gradient	梯度方向直方图
HOF	Histograms of Oriented Optical Flow	光流方向直方图
MBH	Motion Boundary Histograms	运动边界直方图
TDD	Trajectory-pooled Deep-convolutional Descriptors	轨迹池化的深度卷积描述子
HOG3D	Histograms of Oriented 3D Gradient	三维梯度方向直方图
VCML	Video Covariance Matrix Logarithm	视频协方差矩阵对数
DBSCAN	Density-based Spatial Clustering of Application with Noise	基于密度的聚类算法
BOW	Bag of Words	词袋模型
CCA	Canonical Correlation Analysis	典型相关分析
KCCA	Kernel Canonical Correlation Analysis	核典型相关分析
FDA	Fisher Discriminant Analysis	Fisher 判别分析
KFDA	Kernel Fisher Discriminant Analysis	核 Fisher 判别分析
SVM	Support Vector Machines	支持向量机

目录

第一章　绪论

1.1 研究背景及意义

在社会的快速发展的同时，由于人类自身或外部环境的限制，需要研究一些智能机器来协助人们做一些难以达成的任务。人们迫切地想要让智能机器自主地学习、理解并分析视频或图像，从而代替人本身的视觉。视频分析和理解是人类行动识别的一个日益增长的主题，并成为计算机视觉最受欢迎的领域之一，这得益于其在监控、娱乐、医疗保健和视频检索等领域的应用。在视频监控领域中，人类行为可以从视频中被识别和分析，以确保社会安全。在娱乐方面，人机互动可以通过人类行动识别来增加娱乐体验。在医疗保健方面，人类行动识别可以通过分析患者的行为来检测异常或辅助患者康复。以下是对行为识别的几个应用的简要介绍。

相比于传统监控，智能视频监控能够运用智能机器取代人眼观测，从而帮助人们完成监控目的。当前行为识别主要对人的脸部、步态或特殊行为进行识别。智能监控通过运用相关视觉算法，可对大量数据进行及时地处理并在发现异常情况时发出警报，能对异常人员或事物实时监控、跟踪和识别。在监控短距离的人物时，通过人脸便能轻松地将身份识别。然而当监控比较远时，人脸会由于分辨率过低或者遮挡的原因而不易识别，但是进入监控区域的人物步态特征是可分辨的，因此可以将这一特征用于人物身份识别的任务中。该特征比较容易提取、并且难以掩盖等特性，引起了众多研究人员的强烈研究兴趣。

智能机器的未来必将拥有智能人机交互平台，计算机可以通过人类对其输入的相关信息提出请求或回答问题等。与计算机的交互可以通过手势、语音、头部跟踪和视觉跟踪来实现。人机交互技术在众多热门领域发挥着巨大作用, 比如用于机器人操控、长距离救护及现实虚拟化等触觉感官的交互，现代手机上定位以及跟踪的技术，还运用于研究隐身、浸入式控制的行为识别，语言障碍人群使用的无声语音识别，亦或是“意念轮椅”中的基于脑电波的交互等等技术。

视频检索作为获取视频信息的手段越来越受到重视。检索技术来源于互联网的发展需求。以此为基础的各类视频检索引擎有 TV-FI(Tsinghua Video Find It)和 Yahoo 等。通过提高网络的带宽，使得人们能更快捷地进行信息交互或共享多媒体信息。在互联网信息时代越来越多的信息通过视频等方式表现出来，因此人们对于信息的检索技术的要求也越来越高。与检索文字信息不同，视频检索通常是基于内容的，利用从图像或视频中提取的基本特征，再结合各种视觉特征进行识别行为，以便进行联合搜索。

1.2 国内外研究现状

人体行为识别在现实生活中有着广泛应用，引起了广大研究团队的兴趣。随着深度学习技术在图像领域的快速发展，研究者开始相信深度学习方法也可以用于视频分析和理解等任务。相比于传统的基于人工特征方法，采用深度学习方法的模型能够自动地获取有意义的分层特征表示。然而，从互联网或者电影中获取的视频片段比之前的标准数据中库中的视频样本更加复杂，这些视频片段包含了大量的运动成分。这些因素使得学习一个有意义的视觉表示更加困难，如何从视频中提取有效的特征仍然是众多研究人员的核心工作。

目前大部分行为识别方法主要基于两类特征表示：一类是人工特征，另一类是深度学习特征。人工特征的行为识别方法早期是在人体的几何形状或轮廓基础上来实现的。文[1]构建了运动能量图以表示图像序列中运动的位置，生成了运动历史图来代表运动强度，并将两者结合构造了时空模板用于行为识别。然而这些早期的方法只能应用于一些简单的行为识别。当面对更加复杂的真实场景时，由于缺乏运动主体的外观、尺寸和时间的信息，往往难以提取出目标对象的形状和轮廓等可靠信息，因此存在一定的局限性。

随后，局部特征开始逐渐用于行为识别。由于局部特征并不需要算法来检测人体部位，并且对光照变化和复杂背景较为鲁棒，因而成为一种有效的特征表示。典型的局部特征包括：时空兴趣点、稠密轨迹和改进的稠密轨迹。这些局部特征用于行为识别主要分为两步：首先，采用检测算法来发掘视频中的显著运动区域；然后，在获得的运动区域中提取有效的描述子。为了在视频序列中检测局部兴趣点，Laptev 等人[2]将 2D Harris 角点检测器扩展到三维空间，进行时空兴趣点的检测。Dollár 等人[3]在 3D Harris 角点检测方法的基础上，通过运用 Gabor 过滤器来获得丰富数量的兴趣点。Dalal 等人[4]提出了一种基于直方图信息的兴趣点探测器，这些直方图信息捕获了特殊的结构模型。为了描述 3D 兴趣点的局部外观和运动信息，几种局部描述子已经在过去的几年中陆续提出来。文[5]通过计算视频时空兴趣点周围的梯度方向直方图和光流方向直方图描述子，以描述人体行为的局部运动和外观信息。文[6]通过在方向直方图中添加了时间维度，将 SIFT 描述子扩展到三维。类似地，还有将 2D SURF 描述子扩展到三维的情况。文[7]将梯度方向直方图描述子扩展到三维空间，其中局部区域的三维梯度方向在子直方图集合中进行投票统计。虽然这些基于时空兴趣点的获取了行为主体的显著运动，但它们仍不足以描述更加复杂的运动。

此时，基于特征轨迹的方法已逐渐用于行为识别。为了对特征点进行跟踪，KLT 跟踪器[8]和匹配 SIFT 描述子的方法被广泛应用。Messing 等人[9]首先使用 3D Harris 角点检测方法获取视频中的时空兴趣点，然后采用 KLT 跟踪器对于兴趣点进行追踪并

生成轨迹。Matikainen 等人[10]首先采用 KLT 跟踪器提取特征轨迹，并对这些轨迹进行聚类；然后为每个聚类中心计算仿射变化矩阵，将矩阵中的每个元素来表示轨迹。Sun 等人[11]通过匹配连续两个视频帧之间的 SIFT 描述子来提取特征轨迹，并构造了时空上下文信息来表示人体行为的运动信息和结构特性。为了获取长时间运动信息，Bregonzio 等人[12]同时使用了 SIFT 描述子匹配和 KLT 跟踪器方法来提取特征轨迹，并检测时空兴趣点来捕获细微的动作。虽然这些方法取得了令人满意的行为识别结果，但是提取的特征轨迹数量仍不足以准确地描述人类行为。此外，由于真实场景视频中通常存在局部遮挡的情况，上述方法提取的轨迹往往不连续。最近，Wang 等人[13, 26]提出使用稠密轨迹和改进的稠密轨迹用于行为识别。该方法在一些具有挑战的数据库上取得了较好的识别性能。基于轨迹的行为识别方法主要包括以下几个特性：首先，提取的稠密轨迹主要位于运动显著的区域；其次，在随着轨迹弯曲的长方体中提取特征描述子，并沿着轨迹进行池化；最后，轨迹约束的采样策略充分考虑到了人体行为的时间连续性，并且能够有效地处理不同的运动速度。

然而，上述的局部特征用于行为识别也存在一定的局限性。对于行为识别而言，局部特征缺乏一定的语义信息和足够的判别能力。为了解决这些问题，一些基于中层语义和高层语义的行为识别方法逐渐被提出。这些方法通常会采用一些启发式学习方法，从视频中挖掘一些判别性的视觉元素作为特征单元。

近年来，深度学习方法在图像领域取得了重大的突破，一些学者尝试将深度学习的方法运用于视频行为识别。在深度卷积神经网络运用到行为识别之前，Taylor 等人[14]使用受限玻尔兹曼机以无监督的方式来学习视频中的运动信息，并采用卷积学习来微调模型的参数。Le 等人[15]运用堆叠的独立子空间分析(Independent Subspace Analysis, ISA)来自动地学习视频中的空时特征，虽然该方法在行为识别上有很好的性能，但是整体计算复杂度较高，并且难以将其应用到更大的行为识别数据集上。

通过采用分层可训练的滤波器以及特征池化操作，卷积神经网络(Convolutional Neural Network, CNN)能够自发地学习视觉物体识别任务中所需要的复杂特征，并且超过了现有的人工特征方法。随着卷积神经网络在图像识别领域获得了巨大成功，许多研究人员将卷积神经网络用于视频行为分类。文[16]提出了一种 3D CNN 模型提取视频的时间和空间维度特征，获取了多个相邻视频帧之间的运动信息。该模型从相邻视频帧生成多个多通道信息，并在每个通道上进行卷积和采样操作，通过结合多个通道的信息来获得特征表示。此外，该方法用辅助特征来规范化模型的输出，并结合不同模型的预测，进一步提高 3D CNN 的表现。Karpathy 等人[17]利用不同的 CNN 框架来学习视频中的局部空时信息，并提出多分辨率 CNN 来加块模型的训练速度。在 Sport-1M 数据库上进行测试，该方法的识别结果相对于人工特征有明显的提升。

Simonyan 等人[18]提出一种包含空间网络和时间网络的双流卷积网络结构，证明了利用有限的训练数据，由多个视频帧的稠密光流训练得到的卷积网络也具有很好的性能。此外，该方法还提出应用在两个不同的视频分类数据集上的多任务学习方法，既能够增加训练数据的数量，又同时提高算法性能。文[19]采用递归神经网络(Recurrent Neural Networks, RNN)对视频帧序列建模，将CNN的输出作为长短时记忆网络(Long Short Term Memory Networks, LSTM)的输入。该方法采用不同CNN框架来获取全局视频层次的描述子，并且证明了视频帧数的增加改善了分类表现。文[20]提出了一个轨迹池化的深度卷积描述子用于行为识别。该方法利用深度网络模型学习判别性的卷积特征图，并采用轨迹约束池化操作将卷积特征图聚合为有效的描述子，进一步运用时空标准化和通道标准化操作，从而提高描述子的鲁棒性。

虽然目前深度学习方法在视频行为识别方面取得了较大的进展，但是还存在以下一些问题：1）深度学习方法需要大量带标签的视频用于模型的训练，然而大部分的视频数据库很小。2）大部分深度学习下的视频行为识别方法，很大程度忽视了视频时间维和空间维的差异性，不能充分地学习视频中的时空运动信息。3）深度神经网络顶层输出作为全局特征已得到了很好应用，然而网络底层输出作为重要的局部特征，未得到充分重视。

1.3 论文主要研究内容及章节安排

在前期研究的基础上，本文分别对人工特征和深度学习的行为识别方法进行研究，构造了一个具有线性统计特性的人工特征描述子，以获取不同特征之间的联合统计特性；此外，通过一种判别性的非线性特征融合方法，将深度网络中获取的全局和局部特征进行融合，构造了一个完备的深度3D卷积描述子。本文的章节安排如下：

第一章：绪论。首先阐述了视频人体行为识别的研究背景及价值；然后分别总结了人工特征和深度学习特征用于行为识别存在的问题以及国内外研究现状，并给出了论文的章节安排。

第二章：常见的人体行为识别方法简介。本章首先介绍了本文实验仿真中所涉及的两个视频数据集UCF-Sports库和YouTube库；然后分别简述了人工特征用于行为识别的代表性方法，以及将深度网络模型用于行为识别的典型方法。

第三章：一种基于人工特征的行为识别方法。首先介绍了本文稠密光流的计算方法和稠密轨迹的提取步骤；然后给出了本文静态特征和运动学特征的提取过程；最后在底层特征的基础上，阐述了本文基于轨迹的协方差矩阵(Trajectory Based Covariance Matrix, TBCM)描述子的构造方法。

第四章：一种深度学习下的行为识别方法。首先介绍了C3D网络的结构，包括

3D 卷积和池化的操作原理，以及网络模型的训练过程；其次，详细地给出本文全局特征和局部特征的获取过程；然后，详细地介绍了典型相关分析和核典型相关分析两种融合方法以及本文判别性的非线性特征融合方法；最后，给出了本文行为识别构架以及实验仿真和分析。

第五章：全文总结和展望。对本文的主要工作进行了全面总结，针对行为识别中存在的问题和挑战，进一步给出未来研究方向。

第二章 常见的人体行为识别方法简介

2.1 引言

由于背景杂乱、遮挡及视角变换等问题，行为识别仍然是一个充满挑战性的任务。此外，快速相机运动、视角变化、视频的低分辨率、较大的类内差异和较小的类内变化也给行为识别带来了困难。随着大规模视频数据集的出现，行为识别还面临严重的计算负担。如何从视频中提取有效的特征，是解决上述疑难问题和设计更为有效的行为识别构架的关键。目前，大部分人体行为识别方法主要基于两种类型的特征：人工特征和深度学习特征。其中，人工局部特征避免了繁琐的预处理步骤，如目标跟踪和分割等，并且对光照变化和视频噪声等鲁棒。近年来，随着深度学习方法在图像分类和目标检测等领域获得了巨大的成功，人们开始将深度学习方法用于视频行为识别。

本章首先简单介绍了实验仿真所涉及的公共数据集，然后分别简述了基于人工特征和深度学习特征的行为识别中代表性方法。

2.2 人体行为识别公共数据库

近年来，视频数据库的数量和规模都在快速地增长。据YouTube报道，互联网每分钟有超过300个小时的视频上传到服务器，在这些视频中，人是最主要和引人关注的目标对象。在计算机视觉领域中，人体行为识别在智能视频监控和视频语义分析及检索方面扮演重要的角色。虽然视频数据呈现爆炸式增长，但是自动识别和分析人类行为的能力仍然有限。不同行为主体的差异性、视角变化和摄像头变化等因素都带来了巨大的挑战。大部分用于行为识别的计算机视觉算法都是在标注的视频数据集上进行测试。然而，现有的大部分数据库行为类别数目有限，其中每类行为的样本数目也有限，并且主要侧重于某类具体的行为，如运动、烹饪或其它简单动作。以下分别对本文实验仿真与分析所涉及的两个视频数据库进行简单介绍。

（1）UCF-Sports 数据库

UCF-Sports 数据库收集于2008年，包括各种体育赛事的10类行为。数据库的视频序列是从广泛的素材资源网站，包括BBC Motion gallery 和 Getty Images 上获得。数据集中的动作包括：跳水(diving)，高尔夫挥杆(golf-swinging)，踢球(kicking)，举重 (lifting)，骑马(riding horse)，跑步(running)，滑冰(skateboarding)，鞍马(swinging-bench)、单杠摆动(swinging-side)和步行(walking)。每类行为的视频数量从14到35个变化，共有超过200个视频。视频在有着复杂背景的现实场景中拍摄，并且行为表现出显著的类内变化。

图 2.1 UCF-Sports 数据库

（2）YouTube 数据库

YouTube 数据库是从 YouTube 视频网站中收集的，具体包含 11 类行为：骑自行车(cycling)、潜水(diving)、打高尔夫球(golf swinging)、足球杂耍(soccer juggling)、蹦床跳(trampoline jumping)、骑马(horseback riding)、投篮(basketball shooting)、打排球(volleyball spiking)、荡秋千(swinging)、打网球(tennis swinging)、散步(walking with a dog)。数据库中的视频通常包含摄像机运动、复杂背景和光照变化等，该数据集包含了一共 1597 个视频序列。

图 2.2 YouTube 数据库

2.3 基于人工特征的行为识别方法

一般来说，提取的特征需要对光照变化、复杂背景、视角变化和摄像头运动等鲁

棒，并且具有较好的泛化能力和判别能力。对于视频行为识别而言，时间维是一个需要重要考虑的因素，一些行为识别方法并未考虑到视频的空时特性，仅对视频序列的每帧图像单独进行处理，因而识别性能不佳。

目前人工特征可以分为两类：全局特征和局部特征。其中全局特征能够获取行为主体的大部分运动信息，但是该方法需要对人体进行准确定位，并且对光照变化和复杂背景等敏感。然而，局部特征不需要对人体部位进行检测，并且对复杂背景和视频噪声等鲁棒，因此被广泛地应用于行为识别。

2.3.1 基于时空模板的行为识别

基于时空模板的行为识别方法首先采用背景减除或跟踪方法对行为主体进行定位，然后对感兴趣的区域进行编码。一般来说，由于并不完美的提取过程，因而获取的剪影会包含一定噪声。早期，Blank 等人[21]通过堆叠视频中提取的剪影图像来构造三维时空块，并运用泊松方程获取时空显著性和方向特征。Wang 等人[22]通过计算所有视频帧的平均强度获取平均剪影，并使用所有视频帧的中心轮廓构造平均形状，随后将剪影和形状轮廓描述子用于行为识别。Bobick 等人[1]将人体剪影运用到行为识别，该方法首先提取单一视角的人体剪影，并累加动作序列的后续帧获取运动能量图(Motion Energy Images, MEI)，表明了图像序列中运动发生的具体位置；其次，构造了运动历史图(Motion History Images, MHI)，以表示历史运动的强度；最后，将MEI和MHI联合起来作为时空模板用于行为识别。该模板是一个矢量值图像，其中每个像素的分量代表该像素位置处的运动情况。以下先简单介绍MEI和MHI构造过程。

1）运动能量图的获取。图 2.3 为某人“坐”(Sitting)的一个样例。图中第一行包含了视频序列中的关键帧。图中第二行显示了累加的二进制运动图像，即由起始帧图像到上面对应帧图像计算获得。序列显示了图像的特定区域，该区域的形状可以表示动作的发生位置。

令 $\mathrm{I}(x,y,t)$ 为图像序列， $\mathrm{D}(x,y,t)$ 表示运动区域的二进制图像序列， D 由视频序列进行差分得到。将这些二值累加运动图像作为MEI，定义如下

$$E_\tau(x,y,t)=\bigcup_{i=0}^{\tau-1}D(x,y,t-i) \tag{2-1}$$

其中，τ 表示视频帧的长度。

2）运动历史图的获取。为了表示图像是如何运动的，构造了运动历史图。在 MHI 中的任意一点 H_τ 表示了该点时间历史运动情况， H_τ 的计算如下式：

$$H_\tau(x,y,t)=\begin{cases}\tau & \text{if } D(x,y,t)=1\\ \max(0,H_\tau(x,y,t)-1) & \text{otherwise}\end{cases} \tag{2-2}$$

MHI 的示例如图 2.4 所示，其中发生运动的像素更加明亮。

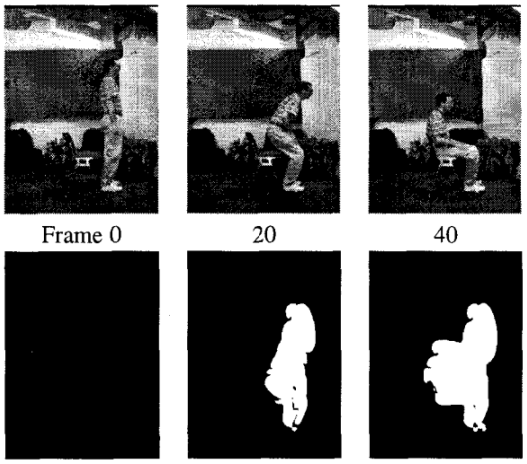


图 2.3 Sitting 的一个样例

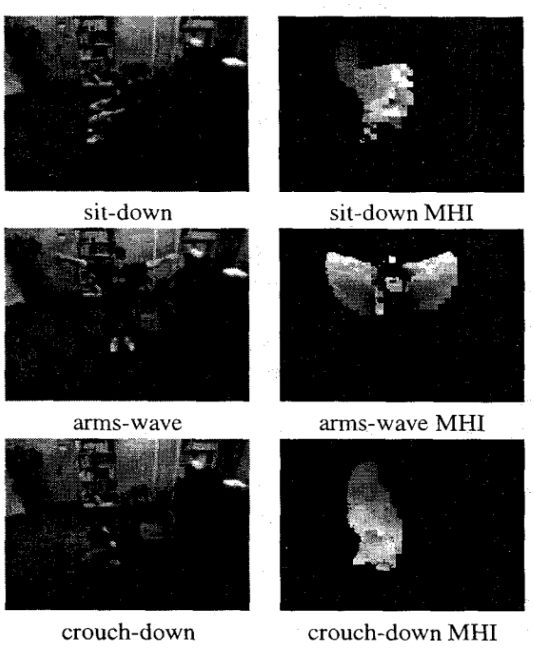


图 2.4 MHI 的示意图

给定一组运动能量图和运动历史图，分别计算 Hu 距以获取具有翻转和尺度不变性的运动描述子，并通过计算运动描述子之间的马氏距离实现行为识别

2.3.2 基于时空兴趣点的行为识别

基于时空模板的识别方法通常需要对人体进行准确定位，并对光照变化和复杂背景等敏感，因而该类方法存在一定局限性。时空兴趣点方法的提出很好地解决了这些问题。首先，在不同时空位置和尺度上检测视频中的时空兴趣点。然后根据兴趣点周围的时空邻域计算特征描述子，通过捕捉形状信息或运动信息来描述人类行为。与基于跟踪和时空形状的方法相比，这类方法对于复杂背景、镜头运动和低分辨视频具有一定的鲁棒性。为了在视频序列中检测局部兴趣点，许多方法陆续被提出。Laptev 等人[2]提出3D Harris 角点检测方法，并已得到广泛地运用。该方法通过将2D Harris 角点检测方法扩展到三维空间，从而获取视频中的时空兴趣点。此外，Laptev 等人还提出一种连接描述子来描述视频行为的局部运动和外观信息。

3D Harris 兴趣点检测方法主要是对视频序列中在空间和时间维度都有较大变化的时空块进行定位。该方法采用一个时空二阶矩阵对视频序列建模，矩阵的定义如下式：

$$\mu = g\left(.;\sigma_i^2,\tau_i^2\right) * \begin{bmatrix} L_x^2 & L_xL_y & L_xL_t \\ L_xL_y & L_y^2 & L_yL_t \\ L_xL_t & L_yL_t & L_t^2 \end{bmatrix} \tag{2-3}$$

其中，$g(\bullet)$为高斯加权函数，σ_i^2和τ_i^2为尺度参数，L_x、L_y和L_t分别代表水平、垂直和时间方向的高斯平滑。计算矩阵μ的特征值μ_1、μ_2和μ_3，当某个像素点对应的特征值μ_1、μ_2和μ_3较大时，即为待检测的时空兴趣点。通过计算响应函数的局部最大值，确定时空兴趣点位置。响应函数的定义如下式：

$$H = \det(\mu) - k \cdot trace^3(\mu) = \lambda_1\lambda_2\lambda_3 - k(\lambda_1+\lambda_2+\lambda_3)^3 \tag{2-4}$$

其中，$\det(\cdot)$和$trace(\cdot)$分别表示矩阵的迹和行列式。

然而，该算法也存在一定的局限性：响应函数H对时间维的运动变化不够敏感，并且对于一些无规律和缓慢运动也不敏感，只能检测出较少的兴趣点，从而影响到识别性能。

2.3.3 基于时空上下文的行为识别

虽然时空兴趣点的方法能够获取视频的空时信息，但是该方法并不能利用特征之间的空间几何关系，如运动轨迹和动作的时间顺序等。上下文在目标识别领域有广泛的应用，对于行为识别而言，上下文信息也是非常重要的，行为动作通常包含人与物

体之间的交互或者人体部位之间的交互。传统的上下文特征使用目标检测或分割进行预处理，并需要更为复杂的学习算法如 AdaBoost 等。尽管这些方法有较好的识别性能，但是有两个不足：首先，预处理步骤需要耗费大量的时间；其次，目标识别和机器学习的过程中容易出错，特别是遇到包含噪声的数据，会对识别结果有影响。

文[23]将上下文信息运用到行为识别。在提出的上下文模型中，存在着许多特征类别，并且每个类别都有一个独立的上下文。基于多尺度时空上下文中观察到的所有特征密度，每个兴趣点的独立上下文由该点特征类别的后验概率决定。由于特征密度一定程度反映了人与物体交互行为的变化，因此相比于传统局部外观特征，上下文特征更具判别能力。以下对该方法进行简单地介绍：

1）时空兴趣点的提取。时空特征对于行为识别非常重要，能够为视频提供一个紧凑表示，并且对类内差异性具有一定鲁棒性。采用 2.3.2 节方法获取时空兴趣点，并对每个兴趣点分别提取 HOG 和 HOF 特征，以描述兴趣点的外观和运动特征。随后，通过 K-means 聚类算法将这些局部特征聚成 N 类，其中聚类中心视为视觉单词，并且每个局部特征都可以映射到一个视觉单词上。

2）上下文特征的获取。给定一个时空兴趣点 $x=[u,v,t]$，其中 u 、v 和 t 分别表示水平、垂直和时间方向坐标。选择 P 个多尺度通道特征计算上下文特征，不同通道的上下文特征分别在不同形状和大小的长方体中进行计算，运用不同的通道可以获取不同类型的上下文信息。对于上下文特征的每个通道，采用一个规则网格对兴趣点的局部邻域时空信息进行编码。该规则网格中的每个长方体定义了上下文域，第 j 个上下文域中的特征类别 w_i 的上下文定义如下式：

$$C_{ij}=\{y \mid f(y)\in w_i, y\in \Omega_j(x)\} \tag{2-5}$$

其中，$\Omega_j(x)$ 为兴趣点 x 的第 j 个上下文域，$f(y)$ 表示局部特征向量。在 $\Omega_j(x)$ 上的总共上下文 C_j 定义如下：

$$C_j=\bigcup_{i=1}^{N} C_{ij} \tag{2-6}$$

在时空兴趣点 x 的每一个上下文通道，使用规则网格对兴趣点局部邻域的时空信息进行编码。具体来讲，每个兴趣点有 M 长方体，每个长方体对应时空上下文域 Ω_j。上下文特征包含 M 个 N 维直方图向量 $\{H_1(x),...,H_M(x)\}$。

2.3.4 基于稠密轨迹的行为识别

虽然时空兴趣点获取了视频序列中的局部感兴趣区域(Region of Interest, ROI)，但是该方法只能获取较短时间的目标动作，因此并不适合于描述较长时间范围的运动和复杂人体行为。然而，通过对特征点进行追踪，可以获取长时间范围的运动信息。例如，Sun 等人[11]首先通过匹配同一个搜索窗口内连续两帧之间的 SIFT 描述子，获取 SIFT 特征轨迹；然后，沿着运动轨迹构造了时空上下文信息来表达人体行为的运动和结构信息。Messing 等人[9]首先获取时空兴趣点；然后使用 KLT 跟踪方法得到了特征点轨迹。虽然这些基于稀疏关键点轨迹的方法产生令人满意的结果，但是相关的轨迹数量较少，并不足以表示人体行为，尤其是低分辨率视频的情况。此外，由于存在局部遮挡情形，因此存在轨迹不连续的情况。

文[13]提出一种稠密轨迹的方法对视频行为进行描述。该方法对每一个帧图像进行稠密采样，并且基于光流场的位移信息对轨迹点进行追踪。由此获得的轨迹对镜头边界变化和无规律运动鲁棒，并且覆盖了视频中的运动信息。此外，该方法还引入了一种对摄像头运动鲁棒的运动边界直方图描述子，在真实场景的数据库中获得了较好的性能。稠密轨迹特征的提取过程如图 2.5 所示，主要包含以下三个部分。

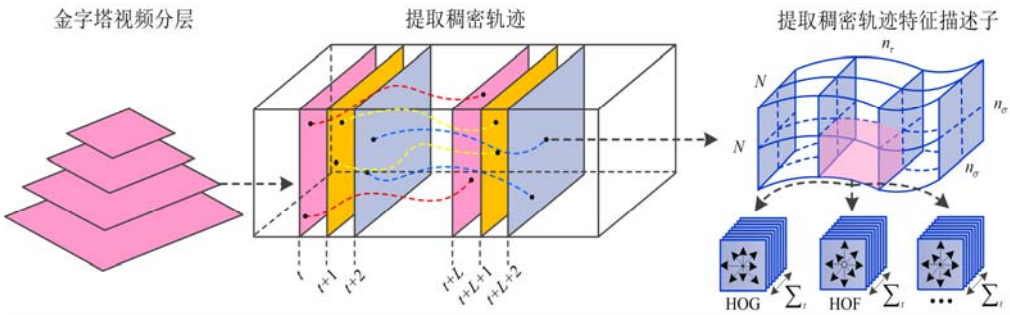


图 2.5 稠密轨迹提取示意图

1）稠密采样。在多个尺度空间对视频帧图像进行网格采样，从而保证特征点均匀地覆盖了整个视频帧的空间位置以及尺度。对处在均匀图像区域的像素点而言，由于并未包含结构信息，故需要将其移除。

2）特征点追踪。在计算光流场后，运用中值滤波器对特征点进行追踪。由于中值滤波器对异常值鲁棒，从而改善了运动边界上的轨迹点。由于特征点跟踪过程中存在位置漂移的问题，因而需要经验地选取轨迹的长度。此外，静止的轨迹并不包含任何运动信息，故被直接移除。

3）轨迹描述子构建。为了嵌入结构信息，将随着轨迹弯曲的长方体进一步分为更小的轨迹子块。在轨迹子块中提取轨迹形状描述子、梯度方向直方图、光流方向直

方图以及运动边界直方图。其中，运动边界直方图描述子通过计算光流水平和垂直方向的偏导数得到，可以表示像素之间的相对运动，因此对摄像头运动鲁棒。

2.4 基于深度学习的行为识别方法

人工特征通常是基于受控环境的领域知识而设计，然而真实的视频数据并不能总是被正确地建模，因此人造特征的泛化能力不足够。此外，人工特征直接用于行为识别，缺乏语义信息和足够的判别能力。目前，深度学习方法在图像分类和视觉目标检测等领域取得了巨大的成功，并开始运用于视频分析。深度学习通过多层非线性变化，自发地从视频数据集中学习判别性的特征表示，并获得了较好的识别性能。深度学习特征具有较强的泛化能力，并包含一定的语义信息和判别能力。以下介绍四种常见深度学习下的行为识别方法。

2.4.1 基于 3D 卷积网络的行为识别

CNN 是一种深层网络模型，能够从原始数据中学习分层的特征表示，并在视觉目标识别任务上实现优越的性能。CNN 主要应用在 2D 图像分类与检测，前期的方法将视频帧视为静止图像，并应用 CNN 来识别单个视频帧的动作。然而，这种方法没有考虑多个连续帧图像中的运动信息。为了有效地学习视频序列中的时空特性，文[16]提出了一种 3D CNN 网络模型，通过在卷积层中执行 3D 卷积操作，进而获取沿着空间和时间维度的判别性特征。该网络模型从相邻的视频帧生成多个信息通道，并在每个通道上分别进行卷积和下采样，将所有通道的信息合并得到时空描述子。此外，对 3D CNN 网络引入辅助特征，提出规范化的 3D CNN 模型，通过结合一些不同框架的输出，进一步提高 3D CNN 模型的表现。

3D 卷积操作使用 3D 卷积核与多个相邻视频帧形成的长方体进行卷积，其中卷积层中的特征映射被连接到网络上一层中的多个相邻帧，从而获取时空运动信息。图 2.6 展示了 3D CNN 网络结构。

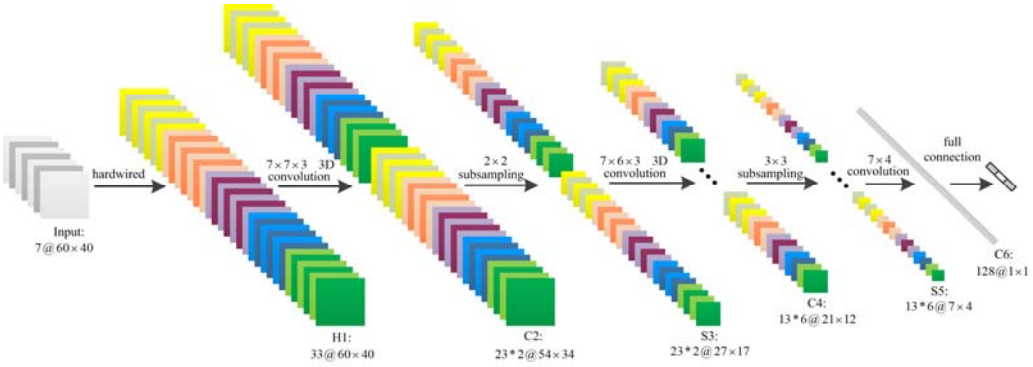


图 2.6 3D CNN 网络结构示意图

3D CNN 网络包括 1 个固定连接层、3 个卷积层、2 个下采样层和 1 个全连接层。网络中的 5 个通道分别使用大小为 7×7×3 的 3D 卷积核，其中 7×7 表示空间维大小，3 为时间维长度。为了增加特征映射的数量，两个不同的卷积集合被应用于每个位置，在 C2 层生成了两个包含 23 个特征映射的集合。在随后的下采样层 S3 上，对 C2 层的每个特征映射使用 2×2 的下采样，生成具有相同数量特征图。接下来的 C4 卷积层是通过在两个特征映射集的 5 个通道上分别使用核大小为 7×6×3 的 3D 卷积来获得。为了增加特征映射的数量，在相同位置使用 3 个不同核的卷积，在 C4 层生成 6 个不同的特征映射集，并且每个包含 13 个特征映射。S5 是在 C4 层的每个特征映射上使用 3×3 的下采样获得。C6 层包含大小为 1×1 的 128 个特征映射，并且每个映射与 S5 层的所有 78 个特征映射相连接。

网络的输出层由与动作类别相同数量单元组成，并且每个单元与 C6 层中的每一个单元相连接。该模型中所有可训练参数被随机初始化，并通过在线误差反向传播算法进行训练。此外，在训练过程中，从大量连续视频序列中提取包含长期运动信息的高层运动特征，作为辅助输出单元连接到网络最后一个隐层，使网络学习一个与该高层特征相近的特征向量，从而实现对 3D CNN 模型的规则化。

2.4.2 基于双流卷积网络的行为识别

视频能够自然地分为空间和时间成分：空间部分以单个图像的形式呈现，包含了视频中关于场景和描述对象的信息；时间部分以视频帧之间运动的形式呈现，包含观察者和物体的运动信息。文[18]提出了一个包含时间流和空间流的双流卷积网模型，并将这两个流在后期进行融合。空间流通过静态视频帧图像来实现行为识别，而时间流通过稠密光流表示的运动信息实现行为识别。此外，作者还提出了应用在两个不同的视频分类数据集上多任务学习方法，既增加了训练数据的数量，又同时提高了算法性能。以下对双流卷积网络模型的空间流和时间流进行简单介绍，其中模型结构框架如图 2.7 所示。

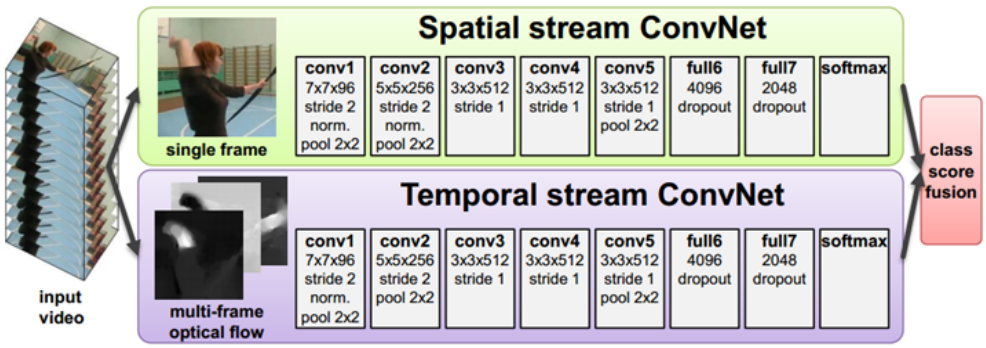


图 2.7 双流卷积网络模型结构框架图

空间流卷积网络在单一视频帧图像上进行操作，能够有效地在静态图像上实现行为识别。由于某些运动与特定目标对象有较强的联系，因此静态图像本身就是一个有用的线索。事实上，前期深度学习下的行为识别方法大部分是基于静止的视频帧图像，并取得了一定的效果。空间流卷积网络本质上就是用于图像分类的网络模型结构，因此可以直接采用最近用于大规模图像分类的深度网络模型。空间流网络模型在大型图像数据集 ImageNet 上进行预训练，随后用于视频人体行为识别。

时间流卷积网络采用堆叠的连续视频帧间光流位移场作为网络模型输入。这种输入明确地描述了视频帧图像之间的运动，使得网络不需要隐式的估计运动，从而简化行为识别。下面简单介绍时间流卷积网络的输入。

稠密光流可以被视为连续帧 t 与 $t+1$ 之间的一组位移向量场 d_t。其中 $d_t(u,v)$ 表示第 t 帧图像在 (u,v) 处的位移向量，向量场的水平分量 d_t^x 和垂直分量 d_t^y 可以被当成是图像通道，并且能够很好地用卷积网络进行识别。为了更好地描述视频帧之间的运动信息，将连续 L 个视频帧的光流场进行叠加，总共形成 $2L$ 个输入通道。w 和 h 分别表示视频帧图像的宽度和高度。对于任意一个视频帧图像 τ，时间卷积网络的输入视频块 $I_\tau \in \mathbb{R}^{w\times h\times 2L}$ 构造如下式：

$$\begin{aligned} I_\tau(u,v,2k-1) &= d^x_{\tau+k-1}(u,v) \\ I_\tau(u,v,2k) &= d^y_{\tau+k-1}(u,v), \quad u=[1;w], v=[1;h], k=[1;L] \end{aligned} \tag{2-7}$$

实际中，位移向量场成分可以取正值也可以取负值，并且包含较大的运动范围，一个方向的运动可能是其相反方向的运动。然而，对于给定的一组视频帧，其对应光流可能由一个特定位移所决定，例如摄像机的运动。作者采用了一个简单方法，即对于每一个位移场，减去其均值向量，以此来补偿摄像机运动。

空间流卷积网络能够通过在 ImageNet 数据集进行模型的预训练得到，然而，时间卷积网络则需要在视频数据集上进行训练。但是目前可用于视频分类的数据集很小，UCF-101 和 HMDB-51 数据集分别只含有 9.5K 和 3.7K 个视频。为了避免出现过拟合的现象，可以将两个数据集进行融合。但是不同类别之间存在着交集使得并不容易实现。作者采用了多任务学习方法对多个数据集进行融合，并对网络结构进行调整，使得最后一个全连接层上有两个分类层：一个计算 HMDB-51 的判分，另一个计算 UCF-101 的判分。每层都有其自身的损失函数，网络训练的总体损失是各个任务的损失之和，网络权重导数可以通过反向传播得到。

2.4.3 基于轨迹池化深度卷积描述子的行为识别

目前大多数深度学习下的人体行为识别方法很大程度上忽略了视频的时域和空

间域之间固有差异，并且在运用网络模型对视频建模时，仅将时间维度视为特征通道，因此并不能充分地学习视频中的空时特性。考虑到以上问题，文[20]结合了人工特征和深度学习特征的各自优点，构造了轨迹池化的深度卷积描述子(Trajectory-pooled Deep-convolutional Descriptors, TDD)用于行为识别。该方法结合了改进的稠密轨迹特征和双流卷积网络模型，首先在视频中提取稠密轨迹，并利用深度网络模型学习了多尺度卷积特征映射，然后引入了轨迹约束的采样和池化策略获取TDD描述子，最后采用Fisher向量编码方法对描述子进行编码，进一步输入到分类器进行行为识别。TDD描述子的提取如图2.8所示，包含三个部分：稠密轨迹的获取，多尺度卷积特征映射图的提取，TDD描述子的计算。

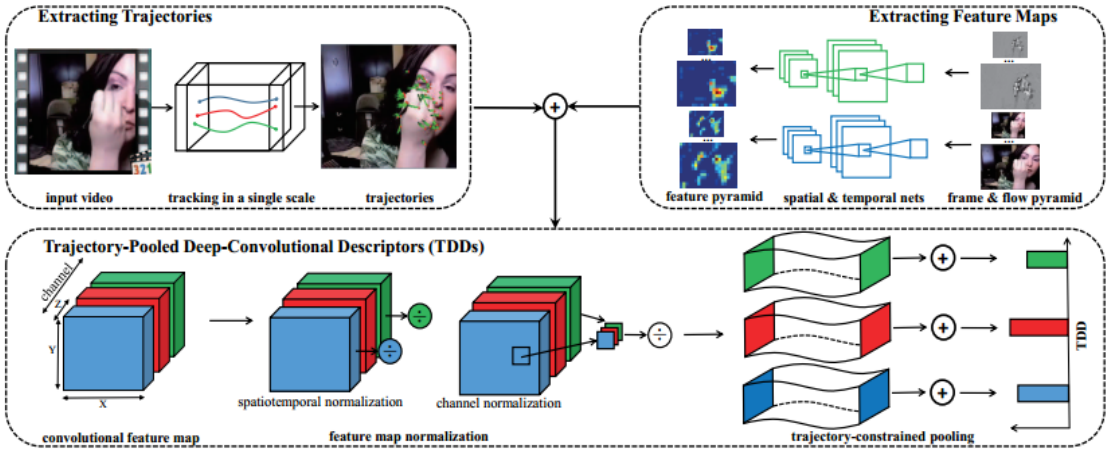


图2.8 TDD描述子的提取示意图

1）稠密轨迹的获取。为了提取稠密轨迹，首先需要对视频帧图像进行网格采样。作者采用改进的稠密轨迹方法，通过考虑摄像头运动提高了稠密轨迹的识别性能。两个连续视频帧之间的背景运动可以采用单应性矩阵来表示。为了估计单应性矩阵，首先，通过SURF特征匹配和光流匹配，找到两个连续帧之间的一致性；然后，使用RANSAC算法来估计单应性矩阵。基于单应性矩阵，改善了帧的图像以移除摄像机运动并重新计算光流。根据光流追踪算法，对特征点进行跟踪以获取多尺度运动轨迹。

2）特征映射图的提取。该方法使用双流卷积网络模型，其中空间流网络用于获取静态外观信息，时间流网络用于描述动态运动信息。将双流卷积网络模型视为特征提取器，以获取视频的卷积特征映射图。该方法对获取特征映射图的网络做出了一点改进，除了移除目标层之后的网络层外，还对每个卷积层和池化层进行补零操作，该操作使得视频中轨迹点的位置能够映射到卷积特征图中。卷积网络的底层具有较小的感受野，通常获取边缘和纹理方向信息。网络的更高层对应较大的感受野，获得更具

判别力的信息。网络的不同层描述不同层次的视觉内容，对行为识别而言，这些不同的内容是互补的。

3）TDD描述子的计算。TDD实际上就是在轨迹周围的时空视频块中计算一种局部轨迹对齐的描述子。该描述子通过空间和时间网络分别获取视频的外观和运动信息。TDD描述子的提取包含特征映射图标准化和轨迹池化两个步骤。具体来讲，标准化是一种有效的操作，可以减少光照的影响，已经在局部特征描述子中得到了广泛地应用。此外，将标准化策略应用到双流卷积网络的特征映射图中，可以有效地抑制一些神经元的激活突发，提高特征的鲁棒性。

与人工特征相比，TDD描述子由深度神经网络自动学习得到，具有较强的判别能力；此外，TDD考虑到视频时间维度的内在特点，并引入基于轨迹的池化策略以结合深度学习特征。

2.4.4 基于深度网络的行为识别

前期深度学习方法只能在较短时间段内学习视频空时特征，而不能学习有关视频时间演化的全局表示。为了实现对可变长度的视频序列建模，Ng 等人[19]提出两种方法：特征池化网络和递归神经网络。特征池化网络采用 CNN 对每一帧图像进行单独处理，再通过各种池化层进一步结合视频帧之间的信息。递归神经网络通过长短时记忆单元，对视频的长期时序关系进行探索。下面首先简单介绍几种不同的特征池化模型。

文[19]研究了 5 中不同的特征池化模型，如图 2.9 所示。卷积池化模型是在最后一个卷积层上进一步添加池化层，以保留在卷积层输出的空间信息；后期池化模型是在进行最大池化之前，将卷积特征先通过两个全连接层，从而直接结合视频帧之间的高层信息；缓慢池化模型是在较小的时间窗中分层地结合视频帧之间信息，以此方式先聚合局部时间特征，再在多帧之间结合高级信息；局部池化模型在缓慢池化模型的基础上去掉第二个最大池化层，从而避免时间信息的丢失；时域卷积模型是在对整个视频序列进行池化之前，添加了一个时域卷积层，以获取视频帧之间的局部关系。

由于视频包含动态变化的信息，因此视频帧之间的变化能够为更准确地预测提供额外信息。作者采用深度长短时记忆网络，对视频序列建模。深度 LSTM 将每一帧的最后一层 CNN 输出作为网络输入。CNN 输出向前随着时间传递，向上通过 5 层叠加LSTM，其中卷积网络和 softmax 分类器在整个时间步长内共享参数。

池化模型采用随机梯度下降方法进行优化，初始学习率为10^{-5}，权重衰减系数为0.0005。对于LSTM网络，采用相同的优化方法，初始学习率为$N*10^{-5}$，其中N为视频帧的数目。

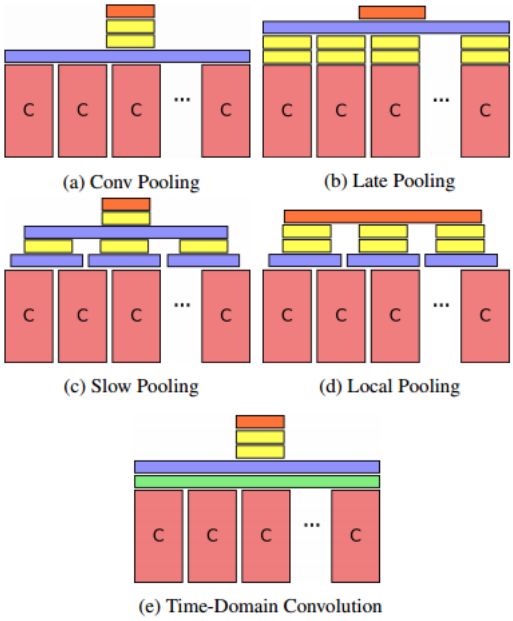


图 2.9 不同的特征池化模型

事实上，多帧模型能够获得较高的准确率，但其训练时间要比单帧模型长。由于池化操作是在共享参数的 CNN 层之后，对于单帧和多帧的池化网络的参数是很相近的。这使得由单帧模型被扩展到多帧模型成为可能。尽管扩展到多帧时，池化层中的特征分布会显著改变，但实验结果表明在这种情况下的参数传递仍是有益的。通过将一个小的网络扩展到一个大的网络，再经过调整，训练速度要明显快于从零开始训练一个大型网络。此外， LSTM 模型的训练与池化网络的训练类似。

2.5 本章小结

本章分析了行为识别中存在的问题，主要介绍了常见的人体行为识别方法。首先，详细地介绍了本文实验仿真与分析中所涉及的两个视频数据库，即 UCF-Sports 库和 YouTube 库；然后，分别介绍了人体行为识别的两类方法，即基于人工特征的行为识别和深度学习下的行为识别，并展开介绍了人工特征方法中的全局特征和局部特征方法，以及深度学习方法中具有代表性的几个网络模型结构。

第三章 一种基于人工特征的行为识别方法

3.1 引言

在行为识别领域，人工局部特征已经成为一种有效的特征表示方式。局部特征不需要特定的算法来检测人体部位，并且对复杂背景，光照变化和视频噪声等影响都鲁棒。典型的局部特征包括：时空兴趣点[2]，梯度方向直方图(Histograms of Oriented Gradient, HOG)，光流方向直方图(Histograms of Oriented Optical Flow, HOF)[5]，运动边界直方图(Motion Boundary Histograms, MBH)[13]，3D 梯度方向直方图(Histograms of Oriented 3D Gradient, HOG3D)[7]和稠密轨迹[13, 26]等。局部特征的提取主要包含两个步骤：首先发掘行为视频中显著和信息丰富的运动区域，然后在运动区域周围提取描述子。在上述人工局部特征中，基于稠密轨迹的特征在各种具有挑战的数据库中获得了较好的性能。具体来说，首先在运动区域提取轨迹立方体，然后在轨迹立方体中构建轨迹形状、HOG、HOF 和 MBH 描述子用于行为识别。事实上，上述的 HOG、HOF 和 MBH 等描述子均为基于单个特征的 1 维统计直方图表示，并没有考虑到不同特征之间的联合统计特性。然而，这些统计特性对于行为识别也许非常重要。

协方差为特征之间的线性相关性提供了一种度量方式。Guo 等人[24]使用局部特征的协方差矩阵对视频序列建模，并应用一个稀疏线性表示的框架来实现行为识别。随后，在稠密轨迹的基础上，Bilinski 等人[25]提出视频协方差矩阵对数(Video Covariance Matrix Logarithm, VCML)描述子，来模拟不同底层静态特征之间线性关系。本文在稠密轨迹的基础上，将图像梯度、光流和运动边界的时间导数作为底层运动特征，获取了运动主体的速度和加速度信息。在此基础上，通过计算底层特征之间的协方差矩阵，构造了基于轨迹的协方差描述子，获取了不同特征之间的联合统计特性，具备了更加丰富的二阶统计信息。

本章节后续安排如下：3.2 节给出了底层特征提取步骤；3.3 节详细介绍了本文基于轨迹的协方差矩阵(Trajectory Based Covariance Matrix, TBCM)描述子的构建过程；3.4 节介绍了将 TBCM 描述子用于行为识别方法；3.5 节给出本章的总结。

3.2 底层特征提取

为了分析视频中的人体行为，一个主要步骤就是提取简单有效的底层特征。由于视频空间域和时间域具有不同的特性，因此需要以不同的方式来处理它们。通过视频序列追踪兴趣点是一个简单的选择，目前一些方法[13, 26, 27]利用轨迹的运动信息进行行为识别，并获得了不错的效果。本文首先对视频序列进行稠密采样，并计算采样点处

的稠密光流和筛选特征点；然后采用特征点跟踪方法对轨迹点进行追踪，以获取行为主体的运动轨迹；最后在随轨迹弯曲的长方体体中，提取运动目标的速度信息并作为底层特征。

3.2.1 稠密光流提取

光流的定义源自视觉感知的生理描述，早在 20 世纪中期，Gibson 首次提出光流，并将其表示为运动目标对于观察者的相对运动。光流是一个矢量，并且包含个两个分量。然而，由于目标的运动，在图像平面中像素点处的亮度变化仅只能产生一个约束，即只能计算其中一个运动分量。因此若不引入附加约束，就无法计算另一个运动分量，也就不能获取该点处的光流。随后，Horn 和 Schunck [28] 通过引入了空间平滑性约束，提出了经典的 H-S 光流计算方法。在 1981 年，Lucas 和 Kanade [29]提出了 L-K 光流法，该算法假设在一定范围的空间领域内运动矢量保持不变，并采用最小二乘法估计光流。事实上，光流的产生是因为在观测场景中的运动目标和观察者之间存在相对运动，然而它仅只能表示在图像平面上的运动强度，并不能解释在真实物理场景中三维运动。随后，光流场的概念被进一步给出，即由二维平面和三维目标的运动而产生的运动场，光流场类似于由运动估计技术得到的稠密运动场。

本文采用 Gunnar Farnebäck 算法[30]计算稠密光流。Gunnar Farnebäck 算法通过二次多项式来估计像素点的邻域值，然后通过多项式展开系数求像素点处的光流。根据局部信号模型，将图像中每个像素点的邻域值近似表示为二次多项式形式：

$$f\left(x\right)=x^{T}\boldsymbol{A}x+b^{T}x+c \tag{3-1}$$

其中，$\boldsymbol{A}$ 为对称矩阵，b 为矢量，c 为标量，通过加权的最小二乘法来估计系数。权重由两部分组成，分别称为确定性和适用性。确定性被耦合到图像邻域像素值。比如，通常设置图像外邻域像素点的确定性为 0，因此这些像素点对系数的估计没有影响。基于邻域中像素点的位置，适用性决定了邻域中像素点的相对权重。对于邻域中心位置的像素点，通常会给以最大的权重值，沿着径向权重依次减小。

当前视频帧的图像可表示为：

$$f_{1}\left(x\right)=x^{T}\boldsymbol{A}_{1}x+b_{1}^{T}x+c_{1} \tag{3-2}$$

经过全局位移 d 后，后面一个视频帧图像表示为：

$$\begin{aligned} f_2(x) = f_1(x-d) &= (x-d)^T \boldsymbol{A}_1 (x-d) + b_1^T (x-d) + c_1 \\ &= x^T \boldsymbol{A}_1 x + (b_1 - 2\boldsymbol{A}_1 d)^T x + d^T \boldsymbol{A}_1 d - b_1^T d + c_1 \\ &= x^T \boldsymbol{A}_2 x + b_2^T x + c_2 \end{aligned} \tag{3-3}$$

在式（3-3）中，二次多项式中由对应的系数相等可得下式：

$$\boldsymbol{A}_2 = \boldsymbol{A}_1 \tag{3-4}$$

$$b_2 = b_1 - 2\boldsymbol{A}_1 d \tag{3-5}$$

$$c_2 = d^T \boldsymbol{A}_1 d - b_1^T d + c_1 \tag{3-6}$$

由式（3-5）可知，当$\boldsymbol{A}_1$为非奇异矩阵时，便可求出全局位移d，即为像素点的光流值：

$$d = -\frac{1}{2}\boldsymbol{A}_1^{-1}(b_2 - b_1) \tag{3-7}$$

3.2.2 稠密轨迹提取算法

本文提取稠密轨迹的步骤如下：首先，对视频帧图像进行稠密采样[13]。然后，采用 3.2.1 节中的 Gunnar Farnebäck 算法计算采样点处的光流，若光流大于阈值T_{flow}，表明该点为运动特征点并保留，否则将其舍去；进一步地，若剩下的特征点数目超过阈值T_{feat}，则对运动特征点处的光流进行基于密度的聚类[31]，以移除摄像头运动造成干扰；最后，采用特征点跟踪方法，对轨迹点进行追踪，获取在不同尺度下长度为L的轨迹。具体过程如下：

（1）稠密采样

对视频帧每隔w个像素进行网格采样，参数w的值设为 5，从而获取足够多的稠密轨迹并捕获视频中的显著运动。由于难以决定最佳的空间尺度来追踪特征点，因此对视频帧图像在不同尺度下进行稠密采样。根据视频的分辨率，选择多个空间尺度进行计算，相邻空间尺度的比率为$1/\sqrt{2}$。该方法确保了特征点覆盖了视频序列的空间位置和尺度。

通常地，特征点追踪算法尝试获取特征点周围的结构信息。然而，在均匀的图像区域，并没有结构信息。本文采用Shi－Tomasi 算法[32]移除在这些平滑区域的采样点，即当自相关矩阵的特征值小于阈值T时，移除在网格中的采样点。

$$T = 0.001 \times \max_{i \in I} \min(\lambda_i^1, \lambda_i^2) \tag{3-8}$$

其中，λ_i^1 和 λ_i^2 为在图像中点 i 的特征值。参数 0.001 为采样点稠密性和显著性的一个折衷，通过实验进行选取。

（2）特征点筛选

令采样点 p 处的光流为 $\boldsymbol{flow}_{\mathrm{p}} = \left(flow_{\mathrm{px}}, flow_{\mathrm{py}} \right)$，其中 $flow_{\mathrm{px}}$ 和 $flow_{\mathrm{py}}$ 分别表示采样点处光流的水平和垂直分量。设定光流的阈值为 T_{flow}，若 $\left\| \boldsymbol{flow}_{\mathrm{p}} \right\|_2 \geq T_{\mathrm{flow}}$，则认为该采样点为运动特征点，将其保留下来并计入特征点数目 Num_{feat}；否者将该点直接舍弃。设定运动特征点的阈值为 T_{feat}，视频帧图像的运动特征点阈值由下式确定：

$$T_{\mathrm{feat}} = \alpha \times \frac{W_{\mathrm{Vid}} \times H_{\mathrm{Vid}}}{w \times w} \tag{3-9}$$

其中，w 表示稠密采样过程中的网格间隔；W_{Vid} 和 H_{Vid} 分别表示视频帧图像的宽和高；α 表示运动特征点数目占所有采样点的比例。若 $Num_{\mathrm{feat}} > T_{\mathrm{feat}}$，则认为该视频帧存在摄像头运动。此时，对运动特征点的光流采用 DBSCAN (Density-based Spatial Clustering of Application with Noise)算法[31]进行聚类，从而进一步移除摄像头运动的干扰。图 3.1 给出了当特征点的数目大于阈值时，即 $Num_{\mathrm{feat}} > T_{\mathrm{feat}}$，使用 DBSCAN 算法聚类后的效果图。

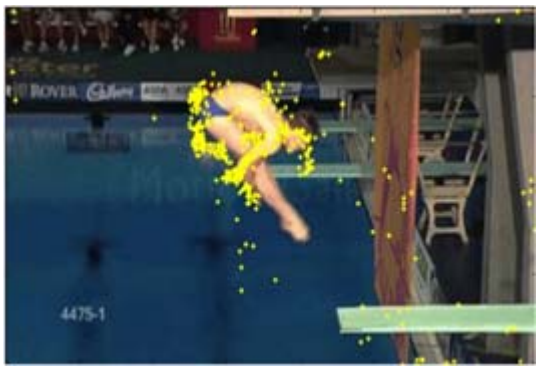


图 3.1 DBSCAN 聚类进行特征点筛选

DBSCAN 算法是一种基于密度的聚类算法。事实上，由于摄像头运动而产生的运动特征点，具有较大的相似性，分布在高密度区域；另外，由运动目标主体产生的特征点，分布较为分散在低密度区域。因此，采用 DBSCAN 算法可以很好地移除由于摄像头运动引起的特征点。

（3）特征点追踪

根据以上方法筛选特征点后，在随后的连续帧图像中对特征点进行跟踪。首先，

采用 Gunnar Farnebäck 算法由连续的两个视频帧 I_t 和 I_{t+1} 计算光流 w_t。给定一个视频帧 I_t 的特征点 $P_t=(x_t,y_t)$，则在下一帧 I_{t+1} 中该点的新位置 P_{t+1} 计算如下：

$$P_{t+1}=(x_{t+1},y_{t+1})=(x_t,y_t)+(M*w_t)|_{(x_t,y_t)} \tag{3-10}$$

其中，M 为 3×3 大小的中值滤波器，$w_t=(u_t,v_t)$ 为稠密光流场。

对于识别一些持续时间较短的行为如微笑和简单的手势等动作，提取较短的轨迹是很有必要的。此外，在包含一些快速无规则运动的情况下，较短的轨迹会更加鲁棒。因此，本文轨迹的长度限制为 $L=15$ 帧。如果水平和垂直位置的标准偏差小于 $\sigma_{\min}$，那么这条轨迹被认为是静止轨迹并且不包含任何运动信息，从而被移除。如果轨迹的水平和垂直位置标准偏差大于 $\sigma_{\max}$，或者连续两帧之间的位移矢量大于这条轨迹的整体位移 $T_{\max}$，那么这条轨迹被认为是异常轨迹，因此被移除。当在 $W\times W$ 的邻域中没有追踪点时，在随后的视频帧图像中新的特征点被采样并进行跟踪。

3.2.3 静态特征和运动学特征提取

本节主要关注行为视频的外观和运动信息表示，这些特征信息对于行为识别非常重要。本文的静态特征和运动学特征提取过程如下：

（1）计算视频帧图像梯度，以获取外观信息。对于视频的每一帧，在 x 和 y 方向上，采用简单的一维 Sobel 算子 $[-1,0,1]$ 计算图像梯度。对于视频帧的任一像素点 P，其 x 和 y 方向的梯度 P_x 和 P_y 计算下式：

$$P_x=\frac{\partial P}{\partial x},\quad P_y=\frac{\partial P}{\partial y} \tag{3-11}$$

（2）计算图像梯度的时间偏导，获取运动的边界信息。在梯度图像的基础上，对两个连续的梯度图像采用 $[-1,1]$ 时间滤波器以计算时间偏导。事实上，人体梯度边界的变化，反映了运动部位的速度，强调了运动边界。对于图像梯度 P_x 和 P_x 对时间方向 t 的偏导数 $P_{t,x}$ 和 $P_{t,y}$ 计算如下式：

$$P_{t,x}=\frac{\partial}{\partial t}\left(\frac{\partial P}{\partial x}\right),\quad P_{t,y}=\frac{\partial}{\partial t}\left(\frac{\partial P}{\partial y}\right) \tag{3-12}$$

（3）计算光流的时间偏导，获取行为的时间相对运动信息。本文首先计算视频帧的光流，并进一步分别对 x 和 y 方向的连续光流采用 $[-1,1]$ 时间滤波器，得到光流在时间 t 方向上的偏导。光流的时间偏导数变化，反映了运动部位的加速度信息。对

于视频帧的光流 $f=(u,v)$，其中 u 和 v 表示光流沿 x 和 y 方向的分量。光流对时间方向 t 的偏导数 $f_{t,x}$ 和 $f_{t,y}$ 计算如下式：

$$f_{t,x}=\frac{\partial u}{\partial t},\quad f_{t,y}=\frac{\partial v}{\partial t} \tag{3-13}$$

（4）计算运动边界的时间偏导，获取目标运动边界的加速度信息。首先通过计算光流的空间导数以获取运动边界。然后对连续的运动边界采用 $[-1,1]$ 时间滤波器，计算得到运动边界的时间偏导。运动边界的时间偏导数变化，反映了目标运动边界的速度变化。运动边界在时间 t 方向的偏导 $(u_{t,x},u_{t,y})$ 和 $(v_{t,x},v_{t,y})$ 计算如下式：

$$u_{t,x}=\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial x}\right),\ u_{t,y}=\frac{\partial}{\partial t}\left(\frac{\partial u}{\partial y}\right),\ v_{t,x}=\frac{\partial}{\partial t}\left(\frac{\partial v}{\partial x}\right),\ v_{t,y}=\frac{\partial}{\partial t}\left(\frac{\partial v}{\partial y}\right) \tag{3-14}$$

对于视频帧图像中的每一个像素点，本文提取 20 维的底层特征。将视频帧图像沿 x 和 y 方向的导数作为静态特征，将图像梯度、光流和运动边界沿时间方向的导数以及相应的模值和方向作为运动学特征。此外，对行为识别而言，底层特征与像素点空间位置之间的关系也是有用信息。因此，视频帧图像中像素点的空间位置也作为底层特征，每个像素点的底层特征表示为以下形式：

$$\begin{aligned}F=\Bigg[&X,Y,P_x,P_y,P_{t,x},P_{t,y},\sqrt{P_{t,x}^2+P_{t,y}^2},\arctan\left(\frac{P_{t,y}}{P_{t,x}}\right),f_{t,x},f_{t,y},\sqrt{f_{t,x}^2+f_{t,y}^2},\arctan\left(\frac{f_{t,y}}{f_{t,x}}\right),\\&u_{t,x},u_{t,y},v_{t,x},v_{t,y},\sqrt{u_{t,x}^2+u_{t,y}^2},\sqrt{v_{t,x}^2+v_{t,y}^2},\arctan\left(\frac{u_{t,y}}{u_{t,x}}\right),\arctan\left(\frac{v_{t,y}}{v_{t,x}}\right)\Bigg]\end{aligned} \tag{3-15}$$

其中，X 和 Y 表示视频中像素点的空间位置；$\sqrt{P_{t,x}^2+P_{t,y}^2}$ 和 $\arctan\left(\frac{P_{t,y}}{P_{t,x}}\right)$ 分别表示 $P_{t,x}$ 和 $P_{t,y}$ 的幅值和方向角；$\sqrt{f_{t,x}^2+f_{t,y}^2}$ 和 $\arctan\left(\frac{f_{t,y}}{f_{t,x}}\right)$ 分别表示 $f_{t,x}$ 和 $f_{t,y}$ 的幅值和方向角；$\sqrt{u_{t,x}^2+u_{t,y}^2}$ 和 $\arctan\left(\frac{u_{t,y}}{u_{t,x}}\right)$ 分别表示 $u_{t,x}$ 和 $u_{t,y}$ 的幅值和方向角；$\sqrt{v_{t,x}^2+v_{t,y}^2}$ 和 $\arctan\left(\frac{v_{t,y}}{v_{t,x}}\right)$ 分别表示 $v_{t,x}$ 和 $v_{t,y}$ 的幅值和方向角。

3.3 构造 TBCM 描述子

传统的 HOG、HOF 和 MBH 特征在单像素方向上进行量化统计，没有考虑不同

特征之间的联合统计特性，对于一些相似的行为，这些特征描述子并不具备足够的判别性。本文在稠密轨迹的基础上，首先将图像梯度、光流和运动边界的时间导数作为底层运动特征，然后通过计算各个特征之间的协方差矩阵，构建本文 TBCM 描述子。

在 3.2 节中提取的底层特征基础上，首先计算底层特征之间的协方差矩阵，然后利用 Log-Euclidean 度量方式[33]将协方差矩阵投影到欧式空间，最后计算所有轨迹子块的描述子，并串接得到本文 TBCM 动态描述子。TBCM 描述子为基于轨迹的协方差矩阵表示，旨在获取底层特征之间的联合统计特性。

3.3.1 协方差矩阵的计算

给定一个大小为$n_x \times n_y$的视频帧t，对于帧图像中的每一个像素点，根据 3.2 节提取d维的底层特征，因此视频帧图像可表示为一个d维特征向量集合$\{F_{(x,y.t)} \in \mathrm{R}^d\}_{1\le x\le n_x, 1\le y\le n_y}$。由于图像以较高的维度表示（$n_x \times n_y \times d$），有必要将其转换为一个更加紧凑的表示。为了方便起见，本文将特征集合$\{F_{(x,y.t)} \in \mathrm{R}^d\}_{1\le x\le n_x, 1\le y\le n_y}$表示为$\{F_{(k.t)} \in \mathrm{R}^d\}_{k=1,\ldots,n}$形式，其中$n$表示每个帧图像中像素点个数，即$n = n_x \times n_y$。

通过计算协方差矩阵，对每个特征的方差以及不同特征之间的协方差进行编码，以一个更加紧凑的形式来表示帧图像。本文协方差矩阵的定义如下：

$$C_t = \frac{1}{n-1}\sum_{k=1}^{n}\left(F_{(k,t)} - \mu_t\right)\left(F_{(k,t)} - \mu_t\right)^{\mathrm{T}} \tag{3-16}$$

其中，$\left\{F_{(k,t)} \in \mathrm{R}^d\right\}_{k=1\ldots n}$为视频帧$t$的底层特征向量集合，$d = 20$为底层特征的维数，$n$为视频帧的像素点个数；$\mu_t$为特征向量均值，即$\mu_t = \frac{1}{n}\sum_{k=1}^{n} F_{(k,t)}$。

3.3.2 投影协方差矩阵到欧式空间

协方差矩阵可以表示为一个连通的黎曼流形。欧式度量准则不能很好地衡量两个协方差矩阵之间的距离，为了方便对基于协方差矩阵描述子进行特征编码，此时需要采用黎曼度量方式。通常对于协方差矩阵有两种经典的距离度量方式：affine-invariant 黎曼度量方式[34]和 Log-Euclidean 黎曼度量方式[33]。根据文[33]可知，这两种度量方式有着相似的性能表现，但是 Log-Euclidean 度量方式比 affine-invariant 度量方式更加简单并且有效。因此，本文选取 Log-Euclidean 度量方式将协方差矩阵投影到欧式空间，以方便进一步地对基于协方差矩阵的描述子进行聚类并构造码书。首先对协方差矩阵进行奇异值分解，得到：

$$C_t = U\Sigma V^{\mathrm{T}} \tag{3-17}$$

其中，U 和V 为正交矩阵，$\Sigma=\mathrm{diag}\left(\lambda_1,\lambda_2,...,\lambda_d\right)$为协方差矩阵奇异值构成的对角矩阵。矩阵对数$\log\left(C_t\right)$计算如下：

$$\log\left(C_t\right)=U\cdot\log\left(\Sigma\right)\cdot V^{\mathrm{T}}=U\cdot\mathrm{diag}\left(\log\left(\lambda_1\right),\log\left(\lambda_2\right),...,\log\left(\lambda_d\right)\right)\cdot V^{\mathrm{T}} \quad (3\text{-}18)$$

由于协方差矩阵是一个$d\times d$ 维的对称矩阵，因此它由$d(d+1)/2$ 个值决定。为了以更加紧凑形式表示视频帧图像，仅取协方差矩阵的上三角部分来表示视频帧。

$$V_t=\mathrm{T}\left(\log\left(C_t\right)\right) \quad (3\text{-}19)$$

其中，$\mathrm{T}(\bullet)$表示将矩阵的上三角部分转换为一个矢量。

3.3.3 获取轨迹立方体描述子

给定一个 3.2 节中获取一个$W\times H\times L$ 的轨迹长方体，将其均分为$W\times H\times l$ 的m 个子块。为了使每个子块均有紧凑的表达方式，将子块中所有视频帧特征表示的平均矢量作为子块的描述子，即按照公式（3-20）计算每个轨迹子块的描述子。

$$D_{Sub}=\frac{1}{l}\sum_{t=1}^{l}\mathrm{T}\left(\log\left(C_t\right)\right) \quad (3\text{-}20)$$

其中，l为轨迹子块的帧长度。将m 个子块的描述子串接，从而得到本文的 TBCM 描述子。

$$D_{TBCM}=\left[D_{Sub\_1},D_{Sub\_2},...,D_{Sub\_\mathrm{m}}\right]^T \quad (3\text{-}21)$$

其中，$m=L/l$，以下图 3.2 给出本文 TBCM 描述子的构造示意图。

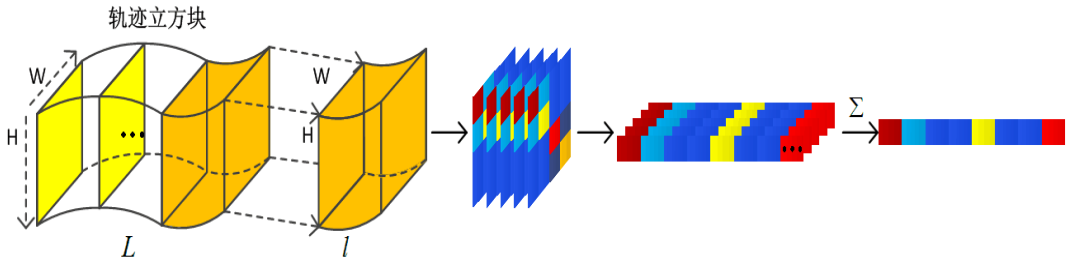


图 3.2 TBCM 描述子的构造示意图

3.4 TBCM 描述子用于行为识别

为了更好地将 TBCM 描述子用于行为识别，本文采用词袋模型(Bag of Words, BOW)的特征编码方法，对提取的 TBCM 描述子进行编码。

BOW 编码方法最初应用在文本处理领域，随着计算机视觉技术的发展，BOW 模型开始运用到图像分类、视频检索与分析和目标识别等领域。BOW 编码方法首先需要构建码书，目前常见的构建码书方法有：K-means 聚类[35, 36]、层次聚类[37]和谱聚类[38]等。在上述这些聚类方法中， K-means 聚类是构造码书的常用方法。

给定特征集合 $\{x_1, x_2, ..., x_M\}$，$x_m \in R^D$，采用 K-means 聚类算法将特征集合聚为 K 个簇 $\{d_1, d_2, ..., d_K\}$，并对每个特征 x_m 引入相对应的变量 $r_{mk} \in \{0,1\}$。若特征向量 x_m 隶属于聚类簇 k，则有 $r_{mk} = 1$ 和 $r_{mj} = 0, j \neq k$。定义以下目标函数：

$$\min J(\{r_{mk}, d_k\}) = \sum_{m=1}^{M}\sum_{k=1}^{K} r_{mk} \left\| x_m - d_k \right\|_2^2 \tag{3-22}$$

通过迭代寻优方法得到 $\{r_{mk}\}$ 和 $\{d_k\}$，以最小化目标函数 J。由此得到 K 个码字的码书 $\mathrm{D} = \{d_1, d_2, ..., d_K\} \in \mathrm{R}^{D \times K}$，对特征集合进行矢量量化编码，最终得到码字直方图，即视频的特征向量表示。

根据上述的 BOW 编码方法首先对本文的两个描述子进行编码；然后将编码后的特征向量串接作为视频的表示；最后将串接后的特征向量输入到线性支持向量机(Support Vector Machines, SVM)中进行行为识别的训练和测试。

3.5 本章小结

本章详细介绍了一种基于人工特征的行为识别方法，首先介绍了本文使用的底层特征，并详细地介绍了光流和稠密轨迹的提取算法，以及说明了在获取的轨迹长方体基础上，静态特征和运动学特征的提取方法，本文将图像梯度、光流和运动边界的时间导数作为运动学特征；接着考虑特征之间的线性关系，给出了本文 TBCM 描述子的具体构造过程；最后详细地介绍了将 TBCM 描述子用于行为识别的方法。

第四章 一种深度学习下的行为识别方法

4.1 引言

目前人工特征通常是基于受控环境的领域知识而设计，事实上真实场景中的视频数据并不能总是被正确地建模，因此人造特征缺乏一定的泛化能力。视频包含了非常丰富的语义信息，传统人工特征直接用于行为识别，缺乏一定的语义信息和足够的判别能力，容易引起行为识别混淆。最近，在图像处理，音频和文本数据分析等领域，深度学习方法取得了巨大的成功和进步。在行为识别领域，一些方法提出将网络的最后全连接层输出作为特征向量用于行为识别，并取得了较好的性能。事实上，网络底层特征可以类比于“浅层”人造特征，也包含一些重要的局部信息，但这些局部信息并未引起人们的重视。

网络的顶层特征和底层特征分别从不同角度学习视频的特征表示，存在必然的互补信息，两者融合可以得到更加完备的特征表示。本文在深度 3D 卷积网络的基础上，提出了一种判别性的非线性特征融合方法，构造了一个更具判别性的深度 3D 卷积描述子，本文深度 3D 卷积描述子构建方法如下：首先，C3D(Convolutional 3D)网络 fc6 和 fc7 层的特征向量串接作为全局特征，pool4 和 pool5 层的特征向量串接作为局部特征；然后，通过判别性的非线性特征融合方法，将全局和局部特征融合，获取本文的深度 3D 卷积描述子。

本章的后续安排如下：4.2 节介绍了 C3D 网络的结构及特点；4.3 节给出了 C3D 网络的全局和局部特征提取方法；4.4 节介绍了经典的线性和非线性特征融合方法，以及本文判别性的非线性特征融合方法；4.5 节给出本文行为识别新构架和实验仿真与分析；4.6 节对本章进行总结。

4.2 C3D 网络的结构

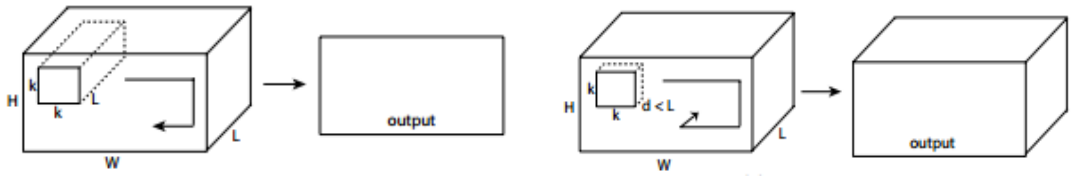
随着深度学习方法在图像领域获得重要突破，近年来特征学习取得了快速地发展，各种已经训练好的卷积网络模型[39]可用于提取图像特征。深度学习通过多层非线性变化将数据转换为一种更高层的抽象表示。深度学习的核心在于，网络各层特征均不是人工设计，而是从原始数据中自发地学习到的。一些方法[40,41]将网络最后全连接层的输出作为特征向量用于迁移学习任务，获取了较好的性能。然而，基于图像的深度学习特征并不直接适用于视频分类，相对于图像，视频中包含大量的空时信息，这些特征不能学习到多个连续视频帧之间的运动信息。

为了有效地考虑视频中的运动信息，ji 等人[16]提出一种 3D 卷积神经网络模型，

以获取沿视频空间维和时间维的判别性特征。实验表明 3D 卷积网络模型要优于传统的网络模型。然而，该方法需要使用人体检测器和头部跟踪算法对视频进行分割，并将分割后的视频片段作为 3D 卷积神经网络的输入，存在一定的局限性。随后，Tran 等人[42]提出了深度 3D 卷积网络模型即 C3D 网络，学习视频的空时信息。与之前的 3D 卷积网络相比，C3D 网络具有更深的网络结构，并且将完整的视频帧作为输入而不依赖于任何预处理，因此更容易扩展到大规模数据集。本文采用 C3D 网络模型来构建深度 3D 卷积描述子，以下先阐述 3D 卷积和池化操作，随后进一步介绍 C3D 网络结构。

4.2.1 3D 卷积和池化

C3D 网络能够很好地对视频的时间信息建模，非常适合于时空特征的学习。与传统的卷积神经网络相比，C3D 网络的卷积和池化操作是在时空维度上进行的，而传统卷积网络的卷积和池化仅仅是在空间维度上执行的。图 4.1 分别展示了传统 2D 卷积与 3D 卷积的差异。如图 4.1 所示，对多个视频帧进行 2D 卷积操作，这里仅将不同的视频帧视为不同通道，其结果仍为一副图像。因此，传统的卷积神经网络在进行卷积操作后，失去了输入视频片段的时间信息。3D 卷积网络可以很好地保持输入视频的时空信息，对视频片段进行 3D 卷积操作，其结果仍然为视频块。同样情况也适用于传统的池化和 3D 池化。



(a) 2D 卷积操作　　(b)3D 卷积操作

图 4.1 2D 卷积和 3D 卷积操作

4.2.2 C3D 网络模型和训练

C3D 网络模型结构如图 4.2 所示。该网络包含 8 个 3D 卷积层，5 个池化层，2 个全连接层以及 1 个 softmax 分类层。8 个卷积层的滤波器数目依次为 64、128、256、256、512、512、512 和 512。所有的 3D 卷积核的大小为$3\times3\times3$，在空间和时间维度的步长均为 1。所有的池化层为最大池化，除了第一层外，所有池化核的大小为$2\times2\times2$，步长为 1。为了避免过早地合并时间信号并满足视频片段的长度要求，第一层的池化核大小设置为$1\times2\times2$。此外，两个全连接层均有 4096 个输出。

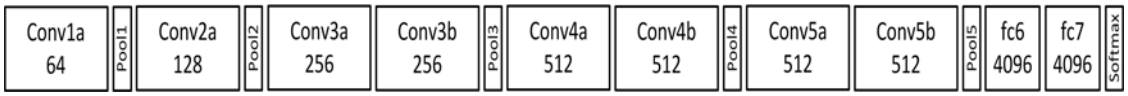


图 4.2 C3D 网络模型结构

为了学习视频中的时空信息，Tran 等人在 Sports-1M 数据库[17]上训练 C3D 网络。Sports-1M 数据库是目前最大的视频分类标准库，该数据集包含近 110 万的体育视频，487 个行为类别。由于 Sports-1M 数据集有许多长视频，Tran 等人随机地从每个训练视频中提取 5 个时间长度为 2 秒的片段，视频帧的大小调整为128×171。在训练过程中，随机地将输入视频帧裁剪为$16\times112\times112$的图像，并以 50%的概率随机翻转裁剪后的图像。采用随机梯度下降算法训练网络，初始学习率设为 0.003，每迭代 15 万次后学习率减半，经过 190 万次迭代后停止训练。

4.3 全局特征和局部特征的提取

深度神经网络的每一层都学习不同的特征表示，网络底层的激活单元通常对感受野中的边缘和纹理信息较敏感，学习到局部的特征表示。然而，在网络更深层的激活单元对应更大的感受野，能够学习到更加全局和高层的特征表示，获取更加复杂的不变性信息。

C3D 网络经过训练后，可作为一个特征提取器。该网络任意一层的特征向量提取过程如下：首先，将视频分为多个连续 16 帧的视频片段；然后，将视频片段作为 C3D 网络的输入以提取该层的激活值，并对所有视频片段的激活值求和并取平均，以获得矢量表示；最后，对获得的矢量进行 L2 范数操作得到该层的特征向量。本文分别提取 C3D 网络全连接层 fc6 和 fc7 的特征向量，并串接作为全局特征，旨在获取行为视频的高层语义信息。此外，由于底层特征的维度太大而无法提取，因此本文分别提取池化层 pool4 和 pool5 的特征向量，并串接作为局部特征，旨在获取纹理和边缘方向信息。

4.4 特征融合

全局特征和局部特征能够从不同角度对视频行为进行描述。对于行为识别而言，全局特征包含全局身体信息，对行为有一个整体抽象的描述，但是缺乏局部结构信息；局部特征主要包括边缘和纹理等细节信息，因而对光照和视角变化不敏感。两类特征之间存在一定的互补性，将两者以合适地方式融合能够得到更加完备的特征表示，以进一步准确地描述行为主体的运动。

由于 C3D 网络的各层之间存在非线性映射关系，因此，池化层 pool4 和 pool5 组成的局部特征与全连接层 fc6 和 fc7 组成的全局特征之间也存在一定的非线性关系。若将 4.3 节中提取的全局特征和局部特征进行线性融合，如线性加权方法[43]或典型相关分析(Canonical Correlation Analysis, CCA)方法[44]等，则无法体现全局和局部特征之间的非线性关系。核典型相关分析方法(Kernel Canonical Correlation Analysis, KCCA)[45]可以解决这个问题，该方法通过隐性的非线性映射，将两组特征集合分别映射到高维特征空间进行 CCA 融合，并将这种非线性相关性作为有效的判别信息，实现了不同特征之间的融合。然而，KCCA 方法没有考虑特征之间的类别信息，因此，融合后的特征判别能力不足够。

本文将类别结构信息引入到 KCCA 中，提出了一种判别性的非线性特征融合方法，将 C3D 网络中提取的全局和局部特征进行非线性特征融合，得到本文判别性的深度 3D 卷积描述子。

4.4.1 典型相关分析算法

近年来，CCA 方法已经运用于图像分析和处理、模式识别和语音识别等领域。基于 CCA 的特征融合方法通过两组特征集合之间的相关性找到两组变换，使得变换后的特征集合之间具有最大相关性，并且每个特征集合内部是不相关的。

设 $X=[x_1,x_2,...,x_N]\in \mathrm{R}^{p\times N}$ 和 $Y=[y_1,y_2,...,y_N]\in \mathrm{R}^{q\times N}$ 分别表示两个特征集合，每个特征集合包含 N 个特征向量，p 和 q 分别表示两个不同模态的特征维数。令 $S_{xx}\in\mathbb{R}^{p\times p}$ 和 $S_{yy}\in\mathbb{R}^{q\times q}$ 分别表示集合 X 和 Y 的协方差矩阵，即 $S_{xx}=\mathrm{cov}(x)$ 和 $S_{yy}=\mathrm{cov}(y)$。$S_{xy}\in\mathbb{R}^{p\times q}$ 和 $S_{yx}\in\mathbb{R}^{q\times p}$ 表示 X 和 Y 之间的互协方差矩阵，即 $S_{xy}=\mathrm{cov}(x,y)$ 和 $S_{yx}=\mathrm{cov}(y,x)$，并且满足 $S_{xy}=S_{yx}^{\mathrm{T}}$。

CCA 算法旨在求取转换矩阵 W_x 和 W_y，使得在子空间中新的两组特征集合 $X^*=W_x^{\mathrm{T}}X$ 和 $Y^*=W_y^{\mathrm{T}}Y$ 具有最大的相关性。CCA 特征融合方法的目标函数如下：

$$corr\left(X^*,Y^*\right)=\frac{cov\left(X^*,Y^*\right)}{var\left(X^*\right)\bullet var\left(Y^*\right)} \tag{4-1}$$

其中，$corr\left(X^*,Y^*\right)$ 表示特征集合 X^* 和 Y^* 之间的相关性，$cov\left(X^*,Y^*\right)=W_x^{\mathrm{T}}S_{xy}W_y$，$var\left(X^*\right)=W_x^{\mathrm{T}}S_{xx}W_x$ 和 $var\left(Y^*\right)=W_y^{\mathrm{T}}S_{yy}W_y$，$cov\left(X^*,Y^*\right)$ 表示特征集合 X^* 和 Y^* 之间的协方差，$var\left(X^*\right)$ 和 $var\left(Y^*\right)$ 分别表示集合 X^* 和 Y^* 的方差。通过拉格朗日乘子法求解目标函数，即在约束条件为 $var\left(X^*\right)=var\left(Y^*\right)=1$ 下，最大化 X^* 和 Y^* 之间的互协方差。通过求解以下式（4-2）特征值方程，可求得转换矩阵 W_x 和 W_y。

$$\begin{cases} S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx}\hat{W}_x = R^2\hat{W}_x \\ S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy}\hat{W}_y = R^2\hat{W}_y \end{cases} \tag{4-2}$$

其中，$\hat{W}_x$ 和 $\hat{W}_y$ 为特征值向量，R^2 为特征值的对角矩阵。在每个方程中非零特征值数目为 $d = rank\left(S_{xy}\right) \le \min\left(n, p, q\right)$，其中特征值以降序顺序排列。转换矩阵 W_x 和 W_y 由非零特征值对应的特征向量组成，$X^*, Y^* \in \mathbb{R}^{d\times n}$ 称为典型变量。

根据文[44]的定义，通常对转换后的特征向量进行串接或者相加，以实现特征级融合。

$$Z_1 = \begin{pmatrix} X^* \\ Y^* \end{pmatrix} = \begin{pmatrix} W_x^{\mathrm{T}}X \\ W_y^{\mathrm{T}}Y \end{pmatrix} = \begin{pmatrix} W_x & 0 \\ 0 & W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{4-3}$$

$$Z_2 = X^* + Y^* = W_x^{\mathrm{T}}X + W_y^{\mathrm{T}}Y = \begin{pmatrix} W_x \\ W_y \end{pmatrix}^T \begin{pmatrix} X \\ Y \end{pmatrix} \tag{4-4}$$

其中，Z_1 和 Z_2 被称为典型相关判别特征。

4.4.2 核典型相关分析算法

CCA 方法只能获取两组特征变量之间的线性关系，并不能提取复杂的非线性关系。KCCA 通过使用核方法，将两组特征集合映射到高维特征空间，并在高维空间中进行 CCA 融合，从而获取特征集合间的非线性关系。

KCCA 方法通过非线性映射函数 ϕ 和 φ，将特征集合 $X = [x_1, x_2, ..., x_N] \in \mathrm{R}^{p\times N}$ 和 $Y = [y_1, y_2, ..., y_N] \in \mathrm{R}^{q\times N}$ 映射到高维空间，如式（4-5）和（4-6）所示。

$$\begin{cases} \phi : x \to \phi(x) \\ X \to \phi(X) = [\phi(x_1), \phi(x_2), ..., \phi(x_N)] \end{cases} \tag{4-5}$$

$$\begin{cases} \psi : y \to \psi(y) \\ Y \to \psi(Y) = [\psi(y_1), \psi(y_2), ..., \psi(y_N)] \end{cases} \tag{4-6}$$

设核函数分别为 k_x 和 k_y，核矩阵的表示如式（4-7）。

$$\begin{cases} K_X = \phi^{\mathrm{T}}(X)\phi(X) \\ K_Y = \varphi^{\mathrm{T}}(Y)\varphi(Y) \end{cases} \tag{4-7}$$

其中，$(K_X)_{ij} = k_x(x_i, x_j)$ 和 $(K_Y)_{ij} = k_y(y_i, y_j)$，$i, j = 1, 2, .., N$。

KCCA 方法旨在寻找投影方向 w_x 和 w_y，最大化以下目标函数式（4-8）：

$$J(w_x, w_y) = \frac{w_x^T \phi(X)\varphi(Y)^T w_y}{\sqrt{w_x^T \phi(X)\phi(X)^T w_x \cdot w_y^T \varphi(X)\varphi(Y)^T w_y}} \tag{4-8}$$

其中，投影方向 w_x 和 w_y 分别位于集合 $\phi(x_1),\phi(x_2),...,\phi(x_N)$ 和 $\psi(y_1),\psi(y_2),...,\psi(y_N)$ 的特征空间中。根据核再生理论可知，必定存在 N 维向量 ξ 和 η，满足 $w_x = \phi(X)\xi$ 和 $w_y = \psi(X)\eta$ 成立，将其带入上式（4-8）可得：

$$J(\xi,\eta) = \frac{\xi^T K_X K_Y \eta}{\sqrt{\xi^T K_X^2 \xi \cdot \eta^T K_Y^2 \eta}} \tag{4-9}$$

由于奇异矩阵的影响，在求解目标函数时可能会产生没有意义的典型相关向量，因此需要在式（4-9）中引入正则约束项。

$$J(\xi,\eta) = \frac{\xi^T K_X K_Y \eta}{\sqrt{\xi^T \left((1-\tau)K_X^2 + \tau K_X\right)\xi \cdot \eta^T \left((1-\tau)K_Y^2 + \tau K_Y\right)\eta}},\ 0 \le \tau \le 1 \tag{4-10}$$

通过拉格朗日乘子法求解目标函数，有以下推导：

$$L(\xi,\eta) = \xi^T K_X K_Y \eta - \frac{\lambda_1}{2}\left(\xi^T \left((1-\tau)K_X^2 + \tau K_X\right)\xi - 1\right) - \frac{\lambda_2}{2}\left(\eta^T \left((1-\tau)K_Y^2 + \tau K_Y\right)\eta - 1\right) \tag{4-11}$$

其中，λ_1 和 λ_2 均为拉格朗日乘数。令 $\partial L / \partial \xi = 0$ 和 $\partial L / \partial \eta = 0$，式（4-10）的求解等价于以下广义特征方程对应的特征向量求解问题。

$$\begin{cases} K_X K_Y \eta = \left((1-\tau)K_X^2 + \tau K_X\right)\xi \\ K_Y K_X \xi = \left((1-\tau)K_Y^2 + \tau K_Y\right)\eta \end{cases} \tag{4-12}$$

求解式（4-12）可得到 ξ 和 η，因而变化后特征向量分别为 $X' = \xi K_X$ 和 $Y' = \eta K_Y$。对变化后的特征向量进行串接，实现特征级融合：

$$Z_1^* = \begin{pmatrix} X' \\ Y' \end{pmatrix} = \begin{pmatrix} \xi K_X \\ \eta K_Y \end{pmatrix} = \begin{pmatrix} \xi & 0 \\ 0 & \eta \end{pmatrix}\begin{pmatrix} K_X \\ K_Y \end{pmatrix} \tag{4-13}$$

KCCA 方法是 CCA 方法的进一步扩展，提取了两组特征集合之间的非线性关系，

有效地消除了特征之间的冗余。此外，融合后的特征具有较好表达能力。

4.4.3 核 Fisher 判别分析算法

核 Fisher 判别分析(Kernel Fisher Discriminant Analysis, KFDA)是 Fisher 判别分析(Fisher Discriminant Analysis, FDA)的非线性扩展，KFDA 将低维样本空间映射到高维特征空间，并在高维特征空间进行 FDA。

KFDA 方法在理论和应用方面引起了众多学者的关注。为了降低方法的计算复杂度，Billings 等人[46]将核矩阵替换为其子矩阵。Wang 等人[47]将判别性向量视为部分训练样本的线性结合，从而提出一种快速的 KFDA 技术。此外，最优核的选择已成为理论研究领域的热点。Fung 等人[48]提出一种基于 FDA 二次规划公式的迭代方法。Khemchandani 等人[49]通过二阶锥规划，找到了数据依赖的最优核函数。近年来，KFDA 方法已经广泛地应用在人脸识别、图像分类和故障诊断等领域。

设特征集合 $X=[x_1,x_2,...,x_N]\in \mathrm{R}^{p\times N}$ 包含 C 个类别，其中 P 为特征向量的维数，N 表示样本总数。FDA 方法旨在寻找最优投影方向 $w\in\mathbb{R}^p$，从而有效地区分不同类的样本，使得类内样本的离散程度尽可能小。FDA 方法的目标函数如下：

$$J(w)=\frac{w^T S_b w}{w^T S_w w} \tag{4-14}$$

其中，S_b 和 S_w 分别为类间和类内散度矩阵。最大化目标函数 $J(w)$，获得投影方向 w。通过拉格朗日乘子法求解，可得：

$$L(w)=w^T S_b w-\lambda\left(w^T S_w w-1\right) \tag{4-15}$$

其中，λ 为拉格朗日乘数，令 $\partial L(w)/\partial\xi=0$，上式可转换 $S_w^{-1}S_b w=\lambda w$，即为求矩阵 $S_w^{-1}S_b$ 的特征值问题。FDA 本质上是一个线性方法，因此很难分离非线性可分的样本。通过使用核方法，KFDA 方法能够充分考虑样本间的非线性关系。

KFDA 通过非线性映射 $\phi: x\to\phi(x)$，将原始空间的样本映射到高维特征空间 F。在特征空间中，类间散度矩阵 S_b^ϕ 和类内散度矩阵 S_w^ϕ 的定义如下式。

$$S_b^\phi=\frac{1}{N}\sum_{i=1}^{C}N_i\left(m_i^\phi-m^\phi\right)\left(m_i^\phi-m^\phi\right)^{\mathrm{T}} \tag{4-16}$$

$$S_w^\phi=\frac{1}{N}\sum_{i=1}^{C}\sum_{j=1}^{N_i}\left(\phi\left(x_j^i\right)-m_i^\phi\right)\left(\phi\left(x_j^i\right)-m_i^\phi\right)^{\mathrm{T}} \tag{4-17}$$

其中，$\phi\left(x_j^i\right)$为在第i类中第j个样本的投影值，m_i^ϕ为第i类样本的均值向量，m^ϕ为所有样本的均值向量；N_i表示第i类包含的样本数，N为总样本数。KFDA 方法通过最大化以下目标函数，来寻找最佳的投影方向v。

$$J(v)=\frac{v^T S_b^\phi v}{v^T S_w^\phi v} \tag{4-18}$$

然而，在高维特征空间中，并不能直接通过求解式（4-18）来获得投影方向v，核函数的引入可以很好地解决这个问题。根据核再生理论，存在N维向量γ，满足$v=\phi(X)\gamma$，将其带入式（4-18）中，可得：

$$J(\gamma)=\frac{\gamma^T K_b \gamma}{\gamma^T K_w \gamma} \tag{4-19}$$

其中，K_b和K_w分别为核类间和核类内散度矩阵，K_b和K_w的定义如下式：

$$K_b=\frac{1}{C}\sum_{i=1}^{C}(m_i-m)(m_i-m)^{\mathrm{T}} \tag{4-20}$$

$$K_w=\frac{1}{C}\sum_{i=1}^{C}\frac{1}{N_i}\sum_{j=1}^{N_i}\left(\xi_j^i-m_i\right)\left(\xi_j^i-m_i\right)^{\mathrm{T}} \tag{4-21}$$

其中，$\xi_j^i=\left[k(x_1,x_j),k(x_2,x_j),...,k(x_N,x_j)\right]^{\mathrm{T}}$，$m_i=\frac{1}{N_i}\sum_{j=1}^{N_i}\xi_j^i$和$m=\frac{1}{C}\sum_{i=1}^{C}m_i$，$k(x_1,x_j)=\left\langle\phi(x_1),\phi(x_j)\right\rangle$表示为核函数计算得到的两个向量$\phi(x_1)$和$\phi(x_j)$在高维空间中的内积。根据广义瑞丽商的性质，最大化式（4-19）以求解投影向量γ的问题，可转换为广义特征方程的求解问题，即$K_w^{-1}K_b\gamma=\lambda\gamma$。

4.4.4 判别性的非线性特征融合算法

本文将类别信息引入到 KCCA 的目标函数中，提出了一种判别性的非线性融合方法。通过将视频的全局和局部特征进行非线性融合，构造了本文的深度 3D 卷积描述子。本文融合方法结合了 KFDA 和 KCCA 的共同优点，将核类内和核类间散度矩阵引入 KCCA 的目标函数中，最大化了全局和局部特征之间的非线性相关性，同时缩小了类内差异性，并加大了类间差异性。

令全局特征和局部特征集合分别为$X=[x_1,x_2,...,x_N]\in\mathrm{R}^{p\times N}$和$Y=[y_1,y_2,...,y_N]\in\mathrm{R}^{q\times N}$，$p$和$q$分别为全局和局部特征的维数，$N$为总样本数目。通

过非线性映射 ϕ 和 φ，将特征集合 X 和 Y 分别映射到高维空间，即 $\phi(X)=[\phi(x_1),\phi(x_2),...,\phi(x_N)]$ 和 $\varphi(Y)=[\varphi(y_1),\varphi(y_2),...,\varphi(y_N)]$。设核函数分别为 k_x 和 k_y，核矩阵为 $K_X=\phi^{\mathrm{T}}(X)\phi(X)$ 和 $K_Y=\varphi^{\mathrm{T}}(Y)\varphi(Y)$，其中 $(K_X)_{ij}=k_x(x_i,x_j)$ 和 $(K_Y)_{ij}=k_y(y_i,y_j)$，$i,j=1,2,..,N$。本文采用高斯核函数，即 $k_x\left(x_i,x_j\right)=\exp(-\|x_i-x_j\|^2/2\sigma_x^2)$ 和 $k_y\left(y_i,y_j\right)=\exp(-\|y_i-y_j\|^2/2\sigma_y^2)$，其中 σ_x 和 σ_y 为高斯核参数。本文判别性的非线性特征融合方法的目标函数如式（4-22）。通过最大化以下目标函数，得到最优投影向量 α 和 β。

$$\begin{aligned}&\max_{\alpha,\beta}\ J(\alpha,\beta)=\alpha^T K_{xy}\beta+\alpha^T K_b^x\alpha+\beta^T K_b^y\beta+\beta^T K_{yx}\alpha\\&s.t.\ \alpha^T K_w^x\alpha+\beta^T K_w^y\beta=1\end{aligned}\tag{4-22}$$

其中，$K_{xy}=\mathrm{cov}\left(K_b^x,K_b^y\right)$，$K_{yx}=\mathrm{cov}\left(K_b^y,K_b^x\right)$，且 $K_{xy}=K_{yx}{}^{\mathrm{T}}$，这里 K_{xy} 和 K_{yx} 均表示 K_b^x 和 K_b^y 的互协方差矩阵，K_w^x 和 K_w^y 分别表示 X 和 Y 的核类内散度矩阵，K_b^x 和 K_b^y 分别表示 X 和 Y 的核类间散度矩阵，K_b^x、K_b^y、K_w^x 和 K_w^y 均为对称矩阵。核类内和核类间散度矩阵的计算如式（4-20）和（4-21）。

通过拉格朗日乘子法求解目标函数，有以下推导：

$$L(\alpha,\beta)=\left(\alpha^T K_{xy}\beta+\alpha^T K_b^x\alpha+\beta^T K_b^y\beta+\beta^T K_{yx}\alpha\right)-\lambda\left(\alpha^T K_w^x\alpha+\beta^T K_w^y\beta-1\right)\tag{4-23}$$

令 $\partial L(\alpha,\beta)/\partial\alpha=0$ 和 $\partial L(\alpha,\beta)/\partial\beta=0$，以上求解可转换为以下广义特征值问题：

$$\begin{bmatrix}K_b^x & K_{xy}\\ K_{yx} & K_b^y\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}=\lambda\begin{bmatrix}K_w^x & 0\\ 0 & K_w^y\end{bmatrix}\begin{bmatrix}\alpha\\ \beta\end{bmatrix}\tag{4-24}$$

当 K_w^x 或 K_w^y 为奇异矩阵时，此时并不可逆，故无法进行下一步求解，本文采用添加正则项的方法来解决该问题，即 $K_w^{x\prime}=K_w^x+\mu_1 I$ 和 $K_w^{y\prime}=K_w^y+\mu_2 I$，$\mu_1$ 和 μ_2 均为较小的常数，I 为单位矩阵。通过计算式（4-24），取前 q 个最大特征值对应的特征向量分别构成投影矩阵 $W_X=[\alpha_1,\alpha_2,...,\alpha_q]$ 和 $W_Y=[\beta_1,\beta_2,...,\beta_q]$，其中 $q=\min(d_x,d_y)$。最后，由求得的投影矩阵和核矩阵，得到本文的深度 3D 卷积描述子 $D_{C3D}=[W_X^{\mathrm{T}}K_X,W_Y^{\mathrm{T}}K_Y]$。

本文在 C3D 网络的基础上，提取了视频行为的全局和局部特征，并通过判别性的非线性特征融合方法，得到更具判别性和完备性的深度 3D 卷积描述子，该描述子充分考虑了全局和局部特征之间的非线性关系，消除了特征之间的冗余信息。

4.5 本文行为识别架构与仿真

4.5.1 本文识别算法新构架

本文在稠密轨迹的基础上，通过计算底层特征之间的联合统计特性，构造了TBCM 描述子。本文基于人工特征的行为识别构架如图 4.3 所示，分为三个部分：首先，进行稠密采样、特征点筛选和追踪以获取运动轨迹，并在随轨迹弯曲的长方体中提取静态特征和运动学特征；然后，通过计算底层特征之间的协方差矩阵，构造了TBCM 描述子。最后，采用 BOW 方法分别对该描述子进行特征编码，并将编码后的特征向量作为 SVM 分类器的输入，进行人体行为识别。本文的 TBCM 描述子富含与行为密切相关的速度、加速度和丰富的底层空时运动信息，充分考虑了底层特征之间的联合统计特性。

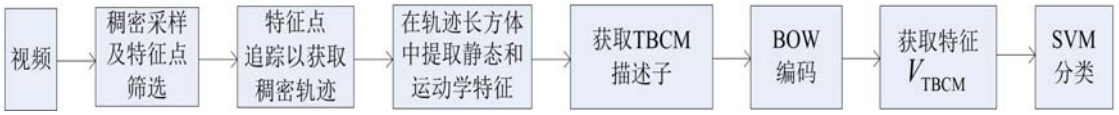


图 4.3 基于人工特征的行为识别构架图

此外，本文还提出了一种深度学习下的人体行为识别构架，通过构造一种判别性的非线性特征融合方法，将视频的全局和局部信息进行融合，构造了一个更加完备的深度 3D 卷积描述子用于行为识别。深度学习下的人体行为识别构架如图 4.4 所示，分为三个部分：首先，分别提取 C3D 网络的池化层 pool4 和 pool5 特征向量，以及全连接层 fc6 和 fc7 的特征向量；然后，将 fc6 和 fc7 层特征向量串接作为全局特征，pool4 和 pool5 层特征向量串接作为局部特征；最后，通过判别性的非线性特征融合方法，将全局和局部特征融合构造了深度 3D 卷积描述子。本文提出的判别性的非线性特征融合方法，减少了特征之间的冗余信息，并引入了类别信息以增强了描述子的判别能力。此外，网络的顶层特征和底层特征分别从不同方面学习视频的特征表示，将两者融合可以得到更加完备的特征表示。

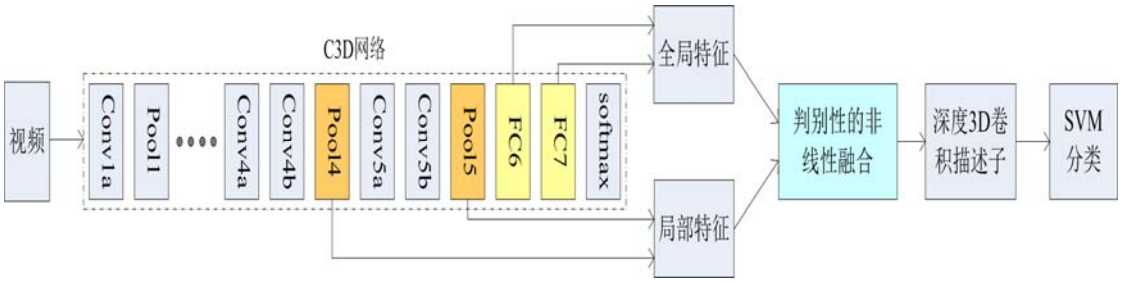


图 4.4 深度学习下的行为识别构架图

4.5.2 实验仿真分析

（1）参数设置

本文构造 TBCM 描述子用于行为识别的参数设置如下。对于提取稠密轨迹，采样间隔 $w=5$ 以保证采样点的稠密性，光流阈值 $T_{\text{flow}}=0.4$ 以筛选出运动特征点，并设置 $\alpha=0.1$。特征点跟踪的轨迹长度为 $L=15$，用于轨迹跟踪的矩形框大小设置为 32×32，即 $W=32$ 和 $H=32$。此外，将 $32\times32\times15$ 的轨迹长方体均分为 3 个 $32\times32\times5$ 的轨迹子块，也就是 $m=3$ 和 $l=5$。为了获取视频特征表示，采用 BOW 模型对特征进行编码，并采用 K-means 方法来构造码书。对于 UCF-Sports 库和 YouTube 库，码书大小分别为 1000 和 1100，并采用线性 SVM 进行分类。

有关深度学习下的行为识别相关参数设置如下。与用于图像分类的 ImageNet 数据库相比，视频行为识别的数据库都非常小，并且视频中人体行为比图像中的物体更加复杂和多变。因此，对于行为识别而言，训练一个深度卷积网络是一个复杂并且充满挑战的任务。本文使用 Tran 等人提供的 C3D 网络模型，该模型在 Sports-1M 库上进行训练，并且基于 Caffe 工具箱[39]来实现。本文 C3D 网络模型作为特征提取器，以获取网络各层的特征。由于 C3D 网络中 fc6、fc7、pool4 和 pool5 层的特征向量维数较高，因此采用主成分分析方法对提取的特征向量进行降维。在对 C3D 网络的全局特征和局部特征进行判别性的非线性融合时，正则项参数设置为 $\mu_1=\mu_2=10^{-4}$，本文采用了高斯核函数，核参数 σ 的值为 1。

（2）基于人工特征的行为识别实验

为了表明本文提取的 TBCM 描述子用于行为识别的有效性，表 4.1 和表 4.2 分别给出了在 UCF-Sports 库和 YouTube 库上该描述子各自的识别结果。由表 4.1 和表 4.2 可以看出，本文 TBCM 描述子有较好的识别性能，并且优于其它对比方法。分析原因在于：本文提出的描述子不仅获取了更多与运动相关的信息，而且还充分考虑了不同特征之间的联合统计特性，因而具有较强的判别能力。

表 4.1 TBCM 描述子在 UCF-Sports 库上的识别结果

识别方法	准确率 (%)
文[50]	88.00
文[51]	90.30
文[52]	91.30
TBCM	**94.00**

表 4.2 TBCM 描述子在 YouTube 库上的识别结果

识别方法	准确率 (%)
文[50]	84.10
文[53]	86.20
文[54]	85.40
TBCM	**87.91**

（3）深度学习下的行为识别实验

为了说明本文判别性的非线性融合方法的优势，表 4.3 给出了本文融合方法与其它经典融合方法在两个视频库上的行为识别结果。表 4.4 和表 4.5 给出了深度 3D 卷积描述子用于行为识别的结果，表明本文深度学习特征的有效性。

表 4.3 不同特征融合方法的实验结果

数据库 / 融合方法	UCF-Sports	YouTube
局部特征	90.67	87.29
全局特征	91.33	88.23
线性加权融合 [43]	92.00	88.60
CCA [44]	92.00	88.92
KCCA [45]	92.67	89.04
判别性的非线性融合	**94.67**	**89.54**

由表 4.3 可知，1）线性加权融合的结果要优于仅用全局或局部特征的识别结果，表明两个特征之间存在一定的互补信息；2）KCCA 特征融合方法的结果要优于线性加权融合和 CCA 方法的结果，表明全局和局部特征之间存在一定的非线性相关性；3）本文判别性非线性特征融合方法的行为识别结果要优于其他对比方法的融合结果。分析原因在于：本文实现了互补信息之间的非线性融合，充分利用了特征之间的非线性关系，消除了特征之间的冗余；同时，本文在目标函数中引入类别约束，减少了类内差异，增加了类间差异，加强了特征的判别能力，进一步提高了行为识别率。

表 4.4 深度 3D 卷积描述子在 UCF-Sports 库上的识别结果

识别方法	准确率 (%)
文[50]	88.00
文[51]	90.30
文[52]	91.30
深度 3D 卷积描述子	**94.67**

表 4.5 深度 3D 卷积描述子在 YouTube 库上的识别结果

识别方法	准确率 (%)
文[50]	84.10
文[53]	86.20
文[54]	85.40
深度 3D 卷积描述子	**89.54**

由表 4.4 和表 4.5 实验结果可知，本文深度 3D 卷积描述子的识别性能要优于传统描述子，表明基于深度学习的特征相对于人工特征包含了更多判别性。

4.6 本章小结

本章详细地介绍了一种深度学习下的行为识别方法。首先，介绍了 C3D 网络的结构，包括 3D 卷积和池化操作的原理以及 C3D 网络训练过程。其次，详细地给出网络的全局特征和局部特征提取步骤。然后，详细阐述了经典的特征融合方法，以及本文判别性的非线性特征融合方法。在本章最后，介绍了本文行为识别算法构架，并分给出了本文 TBCM 描述子的行为识别实验，以及判别非线性特征融合实验和深度 3D 卷积描述子用于行为识别的实验。实验结果表明：1) 本文 TBCM 描述子充分考虑了不同特征之间的联合统计特性，具有较好的识别性能；2) 本文特征融合方法能够利用特征之间的非线性关系，消除特征之间的冗余信息，通过引入类别信息进一步提高特征的判别能力；3) 深度 3D 卷积描述子的识别性能要优于大多数传统的人工特征。

第五章 全文总结与展望

5.1 全文总结

在社会快速发展的同时，人们迫切地想要让智能机器自主地学习、理解并分析视频或图像，从而代替人类本身的视觉。视频分析和理解是人类行为识别的一个日益增长的主题，并成为计算机视觉最受欢迎的领域之一，这得益于其在监控、娱乐、医疗保健和视频检索等领域的应用。本文对基于人工特征和深度学习特征的行为识别方法进行总结与分析，并主要做了以下几个方面的工作：

1）针对传统的描述子并没考虑特征之间的联合统计问题，本文在稠密轨迹的基础上，首先将图像梯度、光流和运动边界的时间导数作为底层运动特征，然后通过计算底层特征之间的协方差矩阵，构造了 TBCM 描述子，从而提高对复杂环境中行为主体的描述能力。

2）提出了一种判别性的非线性特征融合方法。本文将类别结构信息，引入到 KCCA 方法的目标函数中，构造了一种新的特征融合方法。该融合方法最大化了全局和局部特征之间的非线性相关性，同时缩小了类内的差异，并加大了类间的变化，进一步增强了特征的判别能力。

3）构造了深度 3D 卷积描述子。本文分别从 C3D 网络中提取各层的特征，并将 fc6 和 fc7 层特征向量串接作为全局特征，pool4 和 pool5 层特征向量串接作为局部特征，通过判别性的非线性特征融合方法，将全局特征和局部特征进行融合，得到一个更加完备和更具鉴别性的深度 3D 卷积描述子。

5.2 未来展望

尽管目前在视频行为识别方面已经有了大量的研究并取得了一定的进展，但是由于存在摄像头运动、局部遮挡、复杂背景和较大的类内差异等情况，真实场景中的视频行为识别仍然有较多问题亟需解决和完善。未来对行为识别方法的研究可从以下几个方面展开：

1）传统人工特征通常是基于人类先验知识而设计，缺乏一定的泛化能力。此外，将底层人工特征直接用于行为识别，存在语义鸿沟和判别能力不足的问题。近年来，虽然基于深度学习的行为识别方法取得了一定的进展，但是大部分方法仍不能充分地学习视频中的空时特性。可以充分考虑了以上两类方法的优势和不足，构造一种新的行为识别架构。

2）目前大部分基于深度学习方法的行为识别方法忽视了视频时域和空域的内在

差异性。因此，设计更加有效的深度学习网络结构，以更好地学习视频行为中的空时信息，也是未来的一个研究方向。

3）大部分行为识别方法主要研究单个对象的运动模式，事实上在现实中存在更为复杂的行为，如异常行为动作以及群体行为活动等，有待更加深入地研究。

参考文献

[1] Bobick A F, Davis J W. The recognition of human movement using temporal templates[J]. IEEE Transactions on pattern analysis and machine intelligence, 2001, 23(3): 257-267.

[2] Laptev I. On space-time interest points[J]. International journal of computer vision, 2005, 64(2-3): 107-123.

[3] Dollár P, Rabaud V, Cottrell G, et al. Behavior recognition via sparse spatio-temporal features[C]//Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on. IEEE, 2005: 65-72.

[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on. IEEE, 2005, 1: 886-893.

[5] Laptev I, Marszalek M, Schmid C, et al. Learning realistic human actions from movies[C]//Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on. IEEE, 2008: 1-8.

[6] Scovanner P, Ali S, Shah M. A 3-dimensional sift descriptor and its application to action recognition[C]//Proceedings of the 15th ACM international conference on Multimedia. ACM, 2007: 357-360.

[7] Klaser A, Marszałek M, Schmid C. A spatio-temporal descriptor based on 3d-gradients[C]//BMVC 2008-19th British Machine Vision Conference. British Machine Vision Association, 2008: 275: 1-10.

[8] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision[J]. 1981.

[9] Messing R, Pal C, Kautz H. Activity recognition using the velocity histories of tracked keypoints[C]//Computer Vision, 2009 IEEE 12th International Conference on. IEEE, 2009: 104-111.

[10] Matikainen P, Hebert M, Sukthankar R. Trajectons: Action recognition through the motion analysis of tracked features[C]//Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on. IEEE, 2009: 514-521.

[11] Sun J, Wu X, Yan S, et al. Hierarchical spatio-temporal context modeling for action recognition[C]//Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, 2009: 2004-2011.

[12] Bregonzio M, Li J, Gong S, et al. Discriminative Topics Modelling for Action Feature Selection

and Recognition[C]//BMVC. 2010: 1-11.

[13] Wang H, Kläser A, Schmid C, et al. Action recognition by dense trajectories[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 3169-3176.

[14] Taylor G, Fergus R, LeCun Y, et al. Convolutional learning of spatio-temporal features[J]. Computer Vision–ECCV 2010, 2010: 140-153.

[15] Le Q V, Zou W Y, Yeung S Y, et al. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 3361-3368.

[16] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 221-231.

[17] Karpathy A, Toderici G, Shetty S, et al. Large-scale video classification with convolutional neural networks[C]//Proceedings of the IEEE conference on Computer Vision and Pattern Recognition. 2014: 1725-1732.

[18] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]//Advances in neural information processing systems. 2014: 568-576.

[19] Yue-Hei Ng J, Hausknecht M, Vijayanarasimhan S, et al. Beyond short snippets: Deep networks for video classification[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 4694-4702.

[20] Wang L, Qiao Y, Tang X. Action recognition with trajectory-pooled deep-convolutional descriptors[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015: 4305-4314.

[21] Blank M, Gorelick L, Shechtman E, et al. Actions as space-time shapes[C]//Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. IEEE, 2005, 2: 1395-1402.

[22] Wang L, Suter D. Informative shape representations for human action recognition[C]//Pattern Recognition, 2006. ICPR 2006. 18th International Conference on. IEEE, 2006, 2: 1266-1269.

[23] Wang J, Chen Z, Wu Y. Action recognition with multiscale spatio-temporal contexts[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 3185-3192.

[24] Guo K, Ishwar P, Konrad J. Action recognition from video using feature covariance matrices[J]. IEEE Transactions on Image Processing, 2013, 22(6): 2479-2494.

[25] Bilinski P, Bremond F. Video covariance matrix logarithm for human action recognition in videos[C]//IJCAI 2015-24th International Joint Conference on Artificial Intelligence (IJCAI). 2015.

[26] Wang H, Schmid C. Action recognition with improved trajectories[C]//Proceedings of the IEEE

International Conference on Computer Vision. 2013: 3551-3558.

[27] Yu J, Jeon M, Pedrycz W. Weighted feature trajectories and concatenated bag-of-features for action recognition[J]. Neurocomputing, 2014, 131: 200-207.

[28] Horn B K P, Schunck B G. Determining optical flow[J]. Artificial intelligence, 1981, 17(1-3): 185-203.

[29] Lucas B D, Kanade T. An iterative image registration technique with an application to stereo vision[J]. 1981.

[30] Farnebäck G. Two-frame motion estimation based on polynomial expansion[J]. Image analysis, 2003: 363-370.

[31] Ester M, Kriegel H P, Sander J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]//Kdd. 1996, 96(34): 226-231.

[32] Shi J. Good features to track[C]//Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on. IEEE, 1994: 593-600.

[33] Arsigny V, Fillard P, Pennec X, et al. Log‐Euclidean metrics for fast and simple calculus on diffusion tensors[J]. Magnetic resonance in medicine, 2006, 56(2): 411-421.

[34] Förstner W, Moonen B. A metric for covariance matrices[M]//Geodesy-the Challenge of the 3rd Millennium. Springer Berlin Heidelberg, 2003: 299-309.

[35] Kanungo T, Mount D M, Netanyahu N S, et al. An efficient k-means clustering algorithm: Analysis and implementation[J]. IEEE transactions on pattern analysis and machine intelligence, 2002, 24(7): 881-892.

[36] Bishop C. Pattern Recognition and Machine Learning (Information Science and Statistics), 1st edn. 2006. corr. 2nd printing edn[J]. Springer, New York, 2007.

[37] Johnson S C. Hierarchical clustering schemes[J]. Psychometrika, 1967, 32(3): 241-254.

[38] Ng A Y, Jordan M I, Weiss Y. On spectral clustering: Analysis and an algorithm[C]//NIPS. 2001, 14(2): 849-856.

[39] Jia Y, Shelhamer E, Donahue J, et al. Caffe: Convolutional architecture for fast feature embedding[C]//Proceedings of the 22nd ACM international conference on Multimedia. ACM, 2014: 675-678.

[40] Zhang N, Paluri M, Ranzato M A, et al. Panda: Pose aligned networks for deep attribute modeling[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014: 1637-1644.

[41] Zhou B, Lapedriza A, Xiao J, et al. Learning deep features for scene recognition using places database[C]//Advances in neural information processing systems. 2014: 487-495.

[42] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3d convolutional

networks[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 4489-4497.

[43] Atrey P K, Hossain M A, El Saddik A, et al. Multimodal fusion for multimedia analysis: a survey[J]. Multimedia systems, 2010, 16(6): 345-379.

[44] Sun Q S, Zeng S G, Liu Y, et al. A new method of feature fusion and its application in image recognition[J]. Pattern Recognition, 2005, 38(12): 2437-2448.

[45] Hardoon D R, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: An overview with application to learning methods[J]. Neural computation, 2004, 16(12): 2639-2664.

[46] Billings S A, Lee K L. Nonlinear Fisher discriminant analysis using a minimum squared error cost function and the orthogonal least squares algorithm[J]. Neural networks, 2002, 15(2): 263-270.

[47] Liu J, Zhao F, Liu Y. Learning kernel parameters for kernel Fisher discriminant analysis[J]. Pattern Recognition Letters, 2013, 34(9): 1026-1031.

[48] Fung G, Dundar M, Bi J, et al. A fast iterative algorithm for fisher discriminant using heterogeneous kernels[C]//Proceedings of the twenty-first international conference on Machine learning. ACM, 2004: 40.

[49] Khemchandani R, Chandra S. Learning the optimal kernel for Fisher discriminant analysis via second order cone programming[J]. European Journal of Operational Research, 2010, 203(3): 692-697.

[50] Wang H, Kläser A, Schmid C, et al. Dense trajectories and motion boundary descriptors for action recognition[J]. International journal of computer vision, 2013, 103(1): 60-79.

[51] Cho J, Lee M, Chang H J, et al. Robust action recognition using local motion and group sparsity[J]. Pattern Recognition, 2014, 47(5): 1813-1825.

[52] Wu X, Xu D, Duan L, et al. Action recognition using context and appearance distribution features[C]//Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on. IEEE, 2011: 489-496.

[53] Xing D, Wang X, Lu H. Action recognition using hybrid feature descriptor and VLAD video encoding[C]//Asian Conference on Computer Vision. Springer International Publishing, 2014: 99-112.

[54] Liu L, Shao L, Zheng F, et al. Realistic action recognition via sparsely-constructed Gaussian processes[J]. Pattern Recognition, 2014, 47(12): 3819-3827.

致谢

时光荏苒，不知不觉三年的研究生生活已接近尾声，回首求学历程，我收获颇多、感慨万千。值此论文即将完成之际，向所有在研究生期间关心、理解、支持和帮助过我的老师、同学以及家人致以我最真挚的感谢。

首先，衷心感谢同鸣教授三年来在学术以及生活上的悉心指导和无微不至的关怀。在学术上，同老师耐心地引导我入题，在我遇到难题和瓶颈之时，同老师不辞劳累地指导，点拨迷津，使我找到正确的科研方向，我所取得的一切成绩都倾注了同老师大量的心血。同老师严肃的科学态度、敏锐的学术思维、严谨的治学精神以及精益求精的工作作风，深深地感染和激励着我。在生活上，同老师从容淡定的处事风格和乐观开朗的生活态度令我受益匪浅，受用终身。师从同老师，我感到万分庆幸。再次向同老师致以我最诚挚的敬意和感谢！

其次，感谢一起朝夕相处和共同学习的研究生同学，尤其感谢实验室每一位同学，共同营造了一个积极进取、和谐融洽的学习生活环境，感谢你们给予我的所有关心和帮助，你们将是我一生的财富。

特别感谢含辛茹苦养育我的父母，他们无私的关怀是我二十多年求学路上的坚强后盾，我将会更加努力的学习和工作，不辜负父母对我的期望！

最后，衷心感谢在百忙之中评阅学位论文的各位专家、教授！

作者简介

1. 基本情况

汪厚峄，男，湖北鄂州人，1992 年 2 月出生，西安电子科技大学电子工程学院信号与信息处理专业 2014 级硕士研究生。

2. 教育背景

2010.08～2014.07 西安科技大学，本科，专业：电子信息工程
2014.08～ 西安电子科技大学，硕士研究生，专业：信号与信息处理

3. 攻读硕士学位期间的研究成果

3.1 发表学术论文

[1] Tong M, Wang H, Tian W, et al. Action recognition new framework with robust 3D-TCCHOGAC and 3D-HOOFGAC[J]. Multimedia Tools and Applications, 2017, 76(2): 3011-3030. (SCI 检索)

[2] Tong M, Tian W, Wang H, et al. A compact discriminant hierarchical clustering approach for action recognition[J]. Multimedia Tools and Applications, 2017: 1-26. (SCI 检索)