南 京 师 范 大 学

毕 业 设 计（论 文）

（2019 届）

题　　目： 基于 Kinect 体感信息的动作及行为识别技术研究

学　　院： 计算机科学与技术学院

专　　业： 计算机科学与技术

姓　　名： 仇思宇

学　　号： 21150611

指导教师： **宋凤义**

南京师范大学教务处　　制

摘　要

随着计算机视觉技术的发展，让机器能够充分理解人的动作及行为，成为构建良好人机交互接口以及实现图像智能分析的关键。Kinect 作为一款人机交互产品，提供了包含图像景深的丰富图像信息以及人体结构信息，为后续的动作及行为识别提供了丰富的数据基础。

本论文聚焦于基于 Kinect 体感信息的动作及行为识别技术研究，探究了典型思路及代表性方法，包括使用骨架关节特征、三维模型特征、时-空特征和学习特征的动作及行为识别方法，并分析了不同方法的性能及其在不同环境中的适用性，包括鲁棒性和实时性，为典型应用场景提供算法设计及系统构建的决策依据，如，视频监控场景和体感游戏设计。

关键词：计算机视觉；体感信息；人体动作及行为识别；视频监控；

Abstract

With the development of computer vision technology, it is the key of constructing a good human-computer interaction interface and realizing intelligent image analysis to make the machine fully understand the actions and behaviors of people. As a human-computer interaction product, Kinect provides rich image information including image depth of field and human body structure information, providing a rich data foundation for subsequent action and behavior recognition.

This thesis focuses on the research of action and behavior recognition technology based on Kinect's somatosensory information, and explores typical ideas and representative methods, including action and behavior recognition methods with skeleton joint features, 3D model features, space-time features and learning features. The performance of different methods and their applicability in different environments, including robustness and real-time performance, are provided to provide algorithmic design and system construction decision basis for typical application scenarios, such as video surveillance scenarios and somatosensory game design.

Key words: computer vision; somatosensory information; human action and behavior recognition; video surveillance;

目　　录

第 1 章　绪论

目前，体感识别技术的课题主要是研究人体姿态和手势信息提取与识别等相关技术，如基于 Kinect 传感器深度信息的手势检测和识别技术[1]，为人机交互提供了新的方法和思考。在识别简单手势和动作识别技术逐渐成熟并广泛运用于人们日常生活中后，基于 Kinect 传感器的人体动作识别技术开始出现[2][3]。与此同时，识别和分析生物行为信息的技术也开始逐渐发展，如：针对小型动物的行为识别和分析系统[4]；利用 Kinect 深度传感器得到的深度图像，对猪群的攻击行为进行检测和辨别[5]；以及对老年人日常生活的深度图像进行分析，从而发现他们身体功能恶化的早期迹象，从而对可能产生的疾病进行预测[6]。

1.1 本课题的目的及研究意义

人类行为识别研究在过去十年取得了重大进展，并在各种学科中得到越来越多的关注。从诸如彩色相机，深度相机，距离传感器，可穿戴惯性传感器或其他类型传感器中获取相关数据[8]，进而利用这些数据进行人体动作和行为识别和分析。而由于从不同类型传感器中获取的数据处理方法不同、获取并利用的信息不同、使用的任务范围也不尽相同。从行为监视，视频分析，人机交互[7]，人类的动作和行为识别技术已经被广泛应用于日常生活和各个应用领域，同时，辅助生活，健康监控，危险行为预警等相关技术也应运而生[6]。对应于不同的传感器类型，用于识别人体动作的主要有基于视觉的动作识别和基于惯性的动作识别这两种主要的技术。

传统彩色相机捕获的图像序列信息的方法用于动作或手势识别的主要限制在于处理彩色图像时的高计算需求以及对图像质量敏感等相关挑战。本课题将从适用性、可靠性、效率等角度对比现有的基于视觉的动作识别技术和基于惯性的动作识别技术，并针对现有人体动作识别技术存在的问题和局限进行相关算法的改进。

1.2 国内外研究现状

从获取动作信息的图像种类上，可以将用于动作识别数据分为彩色图像（RGB 图像）和深度图像。从动作及行为表示方法上，可以将现有方法分为三个大类：骨架关节特征、三维模型特征、时-空特征和学习特征，如表 1-2 所示。

表 1-2 动作及行为识别方法的分类与分类思路

动作及行为的表示方法分类	动作及行为的表示方法名称
骨架关节表示方法	特征关节[15]、三维关节直方图[16]
三维模型表示方法	DMM-HOG[17]、局部占用模式[19]、随机占用模式[18]
时-空特征表示方法	时空占用模式[21]、傅里叶时间金字塔[19]
学习特征表示方法	三维卷积神经网络特征提取[25]

1.2.1 动作及行为数据的种类与特点

1) 彩色图像的特点

基于传统彩色传感器的动作识别方法主要有：时空体积、时空特征和轨迹，它们被广泛用于传统彩色图像传感器捕获的视频序列中的人体动作识别。局部特征与 SVM 分类器的结合使用[10]，证明了可以通过度量局部特征实现动作识别。在[11]中提供了一种对噪声和姿势变化具有更强鲁棒性的算法，这种算法使用时空特征点（单张图像上局部特征按时间组合）来表征行为。

时空特征包(Bag Of Features，简称 BoF，即一组时空特征的集合) [13]，则避免了恢复人的身体二维模型或三维模型，从而降低了动作识别结果对背景杂乱，遮挡和比例变化的敏感度。随后，动态能量图像(MEI)和运动历史图像(MHI)[13]作为运动模板，以模拟已知的视频中人类行为的空间和时间特征，从而进行动作匹配。这些方法都基于强度或基于颜色，因此也具有相同的缺点，即：识别结果对照明变化的敏感性，限制了识别稳健性。

虽然基于视觉的人类动作识别技术作为模式识别和计算机视觉研究的重要组成部分仍在持续发展，但识别性能正在受到各种挑战。除去上一段中所介绍的，动作识别面临的挑战还有例如遮挡，摄像机位置，执行动作中的主体变化，背景杂乱等因素影响识别结果[8]。实际上，除此之外，使用者或研究者还需要拥有大量的硬件资源才

能运行计算密集型图像处理和计算机视觉算法，并且还需要处理传统图像中缺少 3D 动作数据的问题。

2) 深度图像的特点

近年来，低成本深度传感器的出现，使它们大量被应用于人体动作识别及其相关领域。利用深度传感器提取的深度图像，可以解决传统 RGB 图像中缺失的 3D 动作数据，也因此具备可以更加精确识别人体动作的潜能。

1.2.2 动作及行为识别的代表性方法

1) 骨架关节表示方法

利用在[14]中的从单个深度图像快速准确地预测身体关节的空间位置的方法，提取出由关节点构成的人体骨架，并利用以关节位置差异计算姿态特征f_{cc}，运动特征f_{cp}和偏移特征f_{ci}。对三种特征归一化并使用主元素分析方法（PCA）降低数据维度后，该特征即是特征关节[15]。在结合两个人之间的距离和相对位置后，利用动作森林模型[2]，可以识别两个人的交互行为特征，并且具有更高的整体识别效率。由于骨架估计的不准确性，这种基于骨架的方法具有局限性。并且，骨架信息在许多应用场合中并不总是可用。在[16]中，使用 3D 关节位置直方图（Histograms Of 3D Joint, 简称 HOJ3D）表示姿态，通过对深度图像序列的每一帧计算 HOJ3D 并使用线性判别分析（LDA）重新投影，然后聚类成若干个姿势视觉词。人体的静态姿势便由这些姿势视觉词序列构成。由离散隐马尔可夫模型（HMM）建模分析这些视觉词的时间序列，将其识别为若干已知动作。

2) 三维模型表示方法

一幅深度图像可以视作由点云构成三维模型，将深度图像分别投影到三个坐标平面上，并利用投影图像计算运动能量，被称为深度运动图（DMM）。从 DMM 中提取定向梯度柱状图（HOG）并将其组合为 DMM-HOG，以用于表示动作[17]。与投影方法将三维图像转变为二维图像的思路不同，将关节附近的局部空间划分为若空间网格，落入空间网格的点云数量作为姿态特征，被称为局部占用模式（Local Occupancy Pattern，简称 LOP）[19]。在 LOP 特征的基础上，将全局空间划分为若干子空间，在所有子空间中寻找最具识别力的占用模式，其被称为随机占用模式（Random

Occupancy Pattern，简称 ROP）[18]。在使用稀疏编码对该特征进行编码后，使用 SVM 对编码系数进行分类，从而实现动作识别。

3) 时-空特征表示方法

由于动作信息往往具有连贯性，因此从连续多帧深度图像获取的动作特征具有更加紧凑的特性。利用动态时间规整算法（Dynamic Time Warping，简称 DTW）[26]消除动作的各个子序列执行时间的不同造成的误差后，利用傅里叶时间金字塔[19]可以有效的提取动作的时间模式。时-空占用模式（Space-Time Occupancy Pattern，简称 STOP）[21]则将相同时间段内的点云信息集中在一帧图像中，以统计单位空间点云数量的方式表示动作序列的子序列。

4) 学习特征表示方法

利用深度学习的方法可以避免手动设计动作的特征提取算法，且同样有卓越的性能。利用三维卷积神经网络[24]进行动作识别是一种具有代表性的方法。与二维卷积神经网络不同，三维卷积神经网络的卷积层使用三维的卷积核进行运算。同时还可以利用神经网络提取的高级学习特征与其他动作表示特征分别训练 SVM，并进行决策融合[25]。

第 2 章　动作及行为识别方法评价体系

机器学习模型的评价体系一般包括测试数据集、测试方法与性能评价。本章主要介绍基于深度体感信息的动作识别评价体系，其包括：人体深度信息数据集、测试方法设计和性能评价与可视化等。

2.1 深度图像数据集

人体动作识别技术的巨大进步得益于各种公用标准测试数据集的建立，而用于该技术的数据往往包含着相同的分类结构，即以动作和动作执行者进行分类。通常，即使是同类型的动作，由于动作执行者在身体和执行动作时的差异性，采集的样本具有较大的类内方差。将相同动作分为不同执行者便可以更加合理的评价动作识别模型的泛化能力和鲁棒性。

综述文献[23]中，对引用的部分数据集从所包含的动作类别数、样本数和特性等角度进行了总结，如表 2-1 所示。这些数据集中的绝大多数均采用微软的 Kinect 传感器作为采集工具，它们为各种动作识别算法的性能分析搭建了一个公平的环境，并将继续推动和促进相关研究工作的进一步发展。

表 2-1 深度数据集资料汇总[23]

数 据 集	类别数	样本数	特　　性
MSR Action3D	20	567	10 个演员，每类动作每个演员执行 2~3 次；提供 20 个关节点的 3D 坐标数据、深度图像与 RGB 图像；视频序列为无背景的纯人体运动目标。
UTKinct Action	10	200	10 个演员，每类动作每个演员执行 2 次；提供 20 个关节点的 3D 坐标数据。
Florence 3D Action	9	215	10 个演员，每类动作每个演员执行 3 次；动作相似性大，包含人与物体的交互，同类动作具有不同的执行方式。
CAD-60	12	60	4 个演员，在 5 个不同的场景中执行动作；提供 15 个关节点的 3D 坐标数据、深度图像与 RGB 图像。

在本文中主要使用 MSR Action3D 数据进行动作及行为结果的评估。该数据集内包含深度图像与骨架关节信息，其中深度图像为无背景的纯人体运动目标，因此无需对该数据集进行前景提取。

2.2 识别模型的评价

2.2.1 数据集的测试方法

对于动作识别精确度的判断，目前主要采用跨目标验证的方法。跨目标验证的思想是：训练样本与测试样本分别来自不同动作执行者的动作序列。此方法便是为了解决上一节提到的，同类型的动作不同动作执行者模型评价问题。

跨目标验证实验分多批进行，每批测试使用不同的动作执行者作为测试集与训练集，正如表 2-2 所示。在所有批次测试结束后，对正确率和混淆矩阵求均值，作为最终的模型评价指标。在我的实验中，总共有 10 批这样的训练。

表 2-2 跨目标测试动作执行者的划分

分组序号	训练动作执行者序号	测试动作执行者序号
1	3、1、10、5、2	4、6、7、8、9
2	10、5、2、4、3	1、6、7、8、9
3	3、5、6、2、7	1、4、8、9、10
……	……	……

2.2.2 测试准确率的表示方法

常用的用于评价机器学习模型的性能指标有正确率、召回率、混淆矩阵、ROC 曲线和 AUC 曲线[22]，本文中使用正确率和混淆矩阵进行模型的评价。

对一批二分类样本进行分类后，对于每一个样本，其分类结果必然属于以下表 2-3 四种情况之一：

表 2-3 二分类结果可能出现的情况

真正例（True Positive，简称 TP）	将一个正例正确判断成一个正例
伪正例（False Positive，简称 FP）	将一个反例错误判断为一个正例
真反例（True Negtive，简称 TN）	将一个反例正确判断为一个反例
伪反例（False Negtive，简称 FN）	将一个正例错误判断为一个反例

在此给出模型分类的准确率定义为：

$$\text{accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

将四种情况以如表 2-4 二维表格形式表示，便可以清晰地表示出模型分类性能，以及哪些类更加容易混淆。

表 2-4 分类准确率的二维分布

		预测分类	
		0	1
实际分类	0	TN	FP
	1	FN	TP

除去标签，$\begin{bmatrix} TN & FP \\ FN & TP \end{bmatrix}$便是二分类的混淆矩阵定义。对于M分类问题，混淆矩阵为一个M × M的矩阵。在本文中，以关节的空间绝对位置作为骨架关节特征，利用朴素贝叶斯最近邻（NBNN）分类器对 MSR Action3D 数据集进行动作分类后的结果如图 2-1：

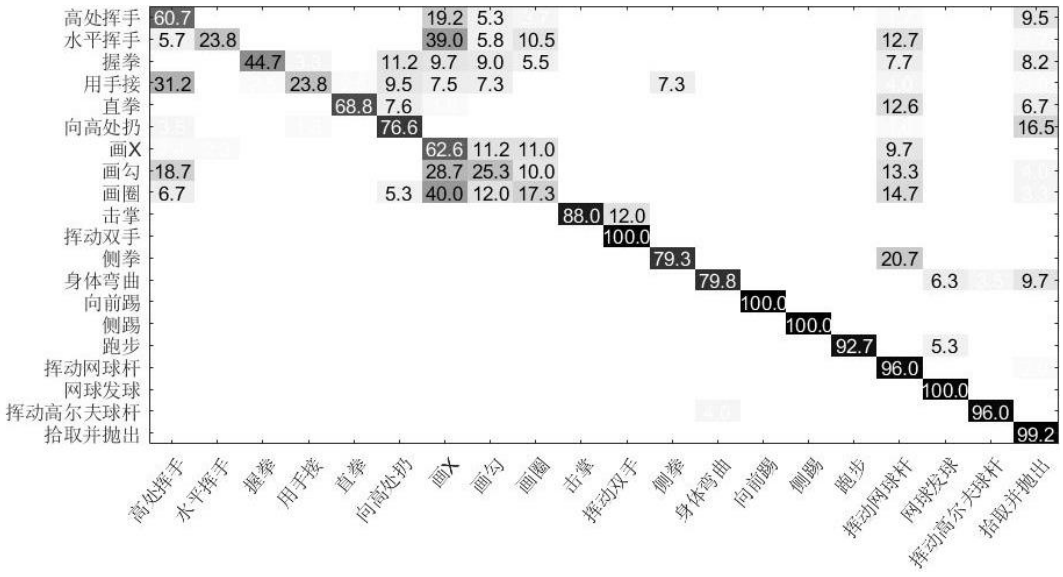


图 2-1 评估动作及行为识别模型的混淆矩阵

由此可见，使用关节绝对位置和 NBNN 分类器对动作进行分类时，向高处扔（high throw）、画 X（draw X）、画勾（draw tick）、画圈（draw circle）、挥动网球杆（tennis swing）和拾取并抛出（pickup and throw）具有较强的易混淆性需要使用更具有识别力的姿态和动作特征。

第 3 章　基于骨架关节特征的动作识别方法

早期的研究人员发现了一种有趣的现象，黑暗中若有人穿着关节处附有光源的衣服，其动作便很容易观察[20]。这种现象使人们认识到可以通过观察关节运动判断人类行为。

3.1 身体部位判断和关节点获取

本节应用随机决策森林进行身体部位识别，并使用模式发现算法生成关节位置。

3.1.1　关节特征获取

在对身体部位进行判断前，应对身体部位进行简单标注，从而能够得到训练集图像中将像素所属的身体部位标签，并利用每个身体部位的深度信息进行特征提取。

函数f_θ表示关节点的平面位置x关节点的深度特征，其计算方法如公式 1 所示。$d_I(x)$是图像I中平面位置x处的像素深度，参数$\theta = (u, v)$中包含两个包含随机偏移向量u和v。通过$\frac{1}{d_I(x)}$使偏移规范化，确保特征与深度相机的位置的相对不变。

$$f_\theta(I, x) = d_I\left(x + \frac{u}{d_{I(x)}}\right) - d_I\left(x + \frac{v}{d_{I(x)}}\right) \tag{1}$$

如果对图像I中的每个像素进行特征提取，将会产生巨大的计算量和内存空间占用。因此，应在整张图像范围内进行随机采样，产生N个样本点，并获得每个样本点位置x_i属于身体部位c的概率$P(c|I, x_i)$，以此进行概率密度估计。

3.1.2　关节位置预测

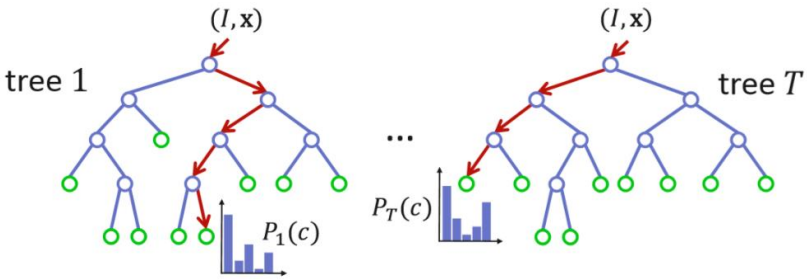


图 3-1 随机决策森林原理示意图[14]

分类器使用随机决策森林模型，该模型是若干决策树的集合。如图 3-1 所示，每棵树由中间节点（蓝色）和叶节点（绿色）组成。红色箭头表示不同树对特定输入可

能采用的不同路径。决策树在叶子节点产生输入x的预测$P_t(c|I,x)$，即x所属身体部位标签c的概率。

文献[9]采用基于均值漂移的局部模式发现算法获得关节位置，该算法使用的核函数为加权高斯核函数。身体部位c的密度估计函数定义为：

$$f_c(\hat{x}) \propto \sum_{i=1}^{N} w_{ic}\, exp\left(-\left\|\frac{\hat{x}-\hat{x}_i}{b_c}\right\|^2\right) \tag{2}$$

其中$\hat{x}$是为三维世界坐标，N是图像像素的数量，w_{ic}是像素的权重，$\hat{x}_i$是图像像素x_i在给定深度$d_I(x_i)$下重新投影到世界坐标空间中的位置，并且b_c是核函数密度估计后确定的核函数带宽。像素权重w_{ic}的计算如公式 3 所示：

$$w_{ic} = P(c|I,x_i)\cdot d_I(x_i)^2 \tag{3}$$

密度估计值高于概率阈值λ_c的像素被用作部位c的概率密度估计。利用均值漂移迭代算法不断更新x_i，最终概率密度的极大值点将作为预测的关节位置。

3.2 动作及行为的表示和识别

3.2.1 姿态特征获取

深度动作序列的每一帧c，都包含N个关节的空间坐标：$X=\{x_1,x_2,\dots,x_N\}$，这些关节坐标均以髋关节中心（Hip Center）的位置作为原点。以关节的空间位置差异表示动作信息，需包含三个特征：姿态特征f_{cc}，运动特征f_{cp}和偏移特征f_{ci}。

其中姿态特征f_{cc}也称为相对关节位置。这种表示方法由成对关节绝对位置的差值构成，用于表征当前帧的静态姿势信息。其计算方法为：

$$f_{cc} = \{x_i - x_j | i,j = 1,2,\dots,N; i \neq j\} \tag{4}$$

f_{cp}则为了捕获当前帧c的运动属性，在当前帧和前一帧p之间计算成对关节差异:

$$f_{cp} = \{x_i^c - x_j^p | x_i^c \in X_c; x_j^p \in X_p\} \tag{5}$$

f_{ci}捕获帧c和初始帧i之间的成对关节差异，表征当前帧c中整体位移：

$$f_{ci} = \{x_i^c - x_j^i | x_i^c \in X_c; x_j^i \in X_i\} \tag{6}$$

将姿态特征f_{cc}，运动特征f_{cp}和偏移特征f_{ci}三者结合后得到每一帧的初步特征表示f_c。但是对任意关节点$x=(u,v,d)$，其三个值可能是不同坐标系中的坐标。例如

(u, v)为屏幕坐标，d为深度坐标。为避免不同坐标系的噪声影响，需要进行标准化，即将关节点x中的每个值缩放为$[-1, +1]$，从而得到f_{norm}。

对于包含N=20 个关节点的深度图像，f_{norm}的维度为$(190 + 200 + 200) \times 3 = 2970$，需要进行数据降维。因此，在得到$f_{norm}$后需要进行主元素分析，从而实现关节点姿态的紧凑表示。

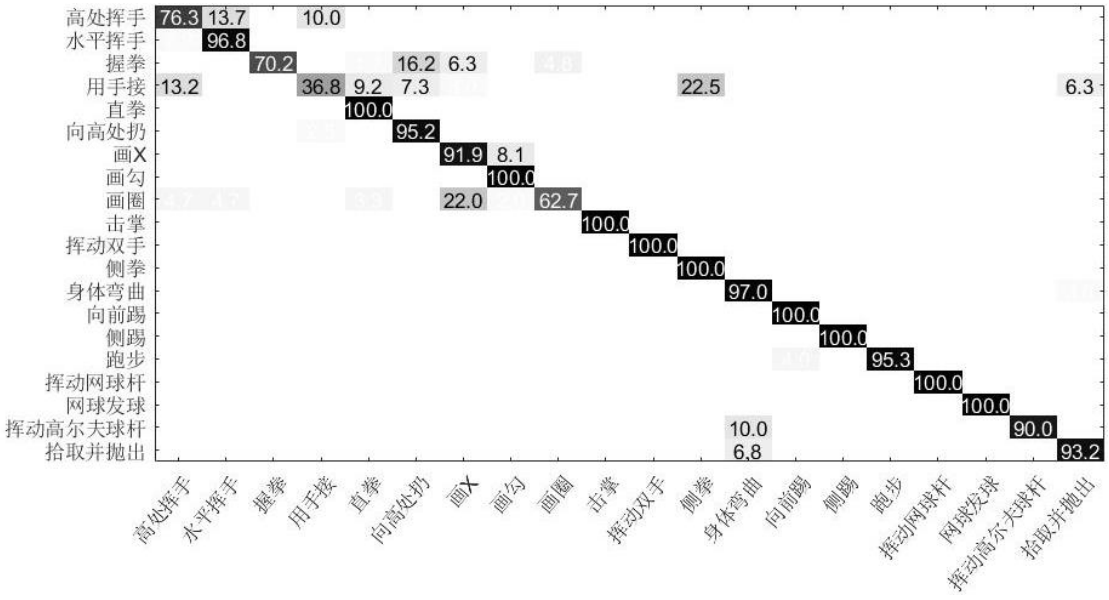


图 3-2 使用 NBNN-特征关节识别结果的混淆矩阵

3.2.2 动作序列识别

在文献[15]中提到的 NBNN 分类器使用的是类似于最邻近方法对视频序列进行分类，即找出与待分类样本v^*“距离”最近的样本v_C，C为样本v_C所属的类别。对每一个可能的分类C求出“距离”后，找到“距离”最近样本v^*。该样本所属的类别便是待分类样本的类别。该距离dist的算法如下：

$$\mathrm{dist}(v_C, v^*) = \sum_{i=1}^{M} \|d_i - NN_c(d_i)\|^2 \tag{7}$$

其中M为每个视频样本所拥有的帧数，d_i为视频序列每一帧的描述符，$NN_c(d_i)$指的是描述符d_i在C类内的最近邻。d_i即是姿态特征的抽象表示，本节中使用上一节中计算得出的f_{norm}作视频每帧的描述符。最终，可得待分类视频样本 v^*所属的分类$C^* = \underset{C}{\mathrm{argmin}}\, \mathrm{dist}(v_C, v^*)$。

表 3-1 不同骨架关节特征与分类器搭配进行动作识别的准确率

	关节绝对位置	关节相对位置	三维关节直方图	特征关节
SVM 分类器	0.65	0.67	0.68	0.58
HMM 模型	0.57	0.60	0.71	0.73
NBNN 分类器	0.72	0.78	0.81	0.84

使用隐马尔可夫模型的动作分类大致包含三个过程，即特征抽取、向量量化和离散隐马尔可夫建模。特征抽取使用线性判别分析法（Linear Discriminant Analysi-s，简称 LDA），也叫做 Fisher 线性判别(Fisher Linear Discriminant，简称 FLD)。线性鉴别分析的基本思想是将高维的特征向量投影到最具鉴别力的向量空间，以达到抽取分类信息和压缩特征维数的效果，投影后保证样本特征在新的子空间有最大的类间距离和最小的类内距离。向量量化则使用 K-均值聚类（K-means）算法，将训练集中的每一个视频样本中的每一帧的特征描述符量化为视觉词。

我们将三元组$\lambda = \{A, B, \pi\}$视作一个 HMM 模型H_i，对向量量化后的视频特征训练M个 HMM 模型。对于一个输出序列，其分类方法如公式 8 所示：

$$\text{decision} = \underset{i=1,2,\ldots,M}{\operatorname{argmax}}\{\Pr(O|H_i)\} \tag{8}$$

使用特征关节特征与 NBNN 分类器得到的混淆矩阵如图 3-2 所示，与图 2-1 对比可得结论：使用特征关节特征与 NBNN 分类器具有更好的识别性能。在不同的骨架关节特征与分类器的对比中，如表 3-1，特征关节和关节位置直方图作具有较强的识别力。使用 HOJ3D 和特征关节作为骨架特征，NBNN 分类器和 HMM 模式动作分类的性能好于 SVM 分类器。

骨架关节特征具有数据量小的特点，因此在实时性上有不俗地表现，并且适合用于实时人机交互。但骨架特征的使用存在场景限制，其不可用于存在遮挡或多人近距离接触（如：拥抱等），因为在这些场景中关节点的预测准确率较低。由于骨架关节特征无法表示所接触物体的特征，因此在人与物体的交互行为场景中其鲁棒性存在一定问题。

第 4 章　基于三维模型特征的动作识别方法

单纯地使用人体骨架关节信息作为特征不能够利用人体与环境交互信息，例如使用电脑打字和伏案写作的动作差异就难以通过关节的位置信息捕捉。而三维模型则可以融合与人体相接触的物体信息，如：键盘和笔。因此，使用三维模型作为动作特征对关节运动相似的的动作具有敏锐的识别力。

4.1 深度信息到三维模型的转化

原始数据是由T帧深度图像组合而成的视频序列，对于每一帧图像我们需要将深度图像的二维坐标转化为三维坐标。

4.1.1　深度图像到点云的转化

假设(u, v)图像中的二维坐标点，z_c为(u, v)的深度，其对应的空间坐标点为(x_w, y_w, z_w)。从二维坐标到空间坐标的转化公式如下：

$$\begin{cases} x_w = z_c \cdot (u - u_0) \cdot C_x \\ y_w = z_c \cdot (v - v_0) \cdot C_y \\ z_w = z_c \end{cases} \tag{9}$$

其中u_0、v_0、C_x和C_y均为传感器的内部参数。此时，人体和环境中的物体均以点云的形式表现出来，人体和环境物体的点云相互作用也成为了识别特征，此时便可使用处理三维模型的思路进行动作及行为的表示。

4.1.2　占用空间的统计

利用上一节在t帧提取的点云，对每个关节提取局部占用模式。局部占用模式（Local Occupancy Patterns，简称 LOP）通过将关节j附近的空间分为$N_x \times N_y \times N_z$个空间网格，并使每个网格拥有$(S_x, S_y, S_z)$个像素。例如，如果$(N_x, N_y, N_z) = (12,12,4)$并且$(S_x, S_y, S_z) = (6,6,80)$，则意味着关节$j$周围的局部区域$(72,72,320)$被划分为$12 \times 12 \times 4$个网格，每个网格大小为$6 \times 6 \times 80$个像素。

计算当前帧中落入每个空间网格b_{xyz}中的点的数量，并且应用sigmoid归一化函数以获得该网格的特征o_{xyz}。如图 4-1 所示，对于每一个网格，其占用信息如公式 10 所示：

$$o_{xyz} = \delta(\sum_{q \in b_{xyz}} I_q) \tag{10}$$

其中，如果点云坐标点位于像素q中，则$I_q = 1$，否则$I_q = 0$。$\delta(.)$是sigmoid归一化函数：$\delta(x) = \frac{1}{1+e^{-\beta x}}$。关节$i$的 LOP 特征是由关节周围的所有空间网格的特征组成的向量，由o_i表示。

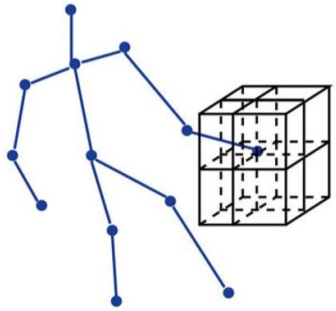


图 4-1 局部占用模式的原理示意图

4.2 三维模型数据冗余度的降低

人体的姿态是由大量关节连接而成，但对于特定动作而言，只有少部分关节点具有识别力，如：以站立的姿势打电话或者喝水仅仅涉及头部、手部和肘部关节。其他关节相对而言不具有识别力，因此可作为冗余数据。这种人体姿态表示方法被称为 Actionlet 集合模型，具有识别力的关节点组成的集合被称为 Actionlet。由于可能的 Actionlet 数量巨大，为了从动作信息中高效的提取 Actionlet，需要使用 Aprior 数据挖掘算法。

4.2.1 动作与身体部位关系的挖掘

对于训练集合中的第i个样本，其内容为$(x^{(i)}, y^{(i)})$。其中，$x^{(i)}$是第i个样本的特征，是一个由各关节特征描述符组成的向量，$y^{(i)}$是第i个样本的标签。为了确定最有

辨别力的 Actionlet，需要对每个关节j的特征G_j训练 SVM 模型，得到分类标签$y^{(i)}$等于动作标签c的概率$P_j(y^{(i)}=c|x^{(i)})$。

对于每个的 Actionlet S，当且仅当每个关节$j \in S$（包含在该 Actionlet 的所有关节）都预测$y^{(i)}=c$时，该 Actionlet S才预测$y^{(i)}=c$。且每个关节预测$y^{(i)}=c$的概率是独立的，所以在 Actionlet S中，分类标签$y^{(i)}$等于动作分类c的概率为：

$$P_S\left(y^{(i)}=c\middle|x^{(i)}\right)=\prod_{j\in S}P_j(y^{(i)}=c|x^{(i)}) \tag{11}$$

将χ_c定义为具有类标签c：$\{i:y^{(i)}=c\}$的训练集合。对于有判别力的 Actionlet，χ_c中的样本应该有较大的$P_S\left(y^{(i)}=c\middle|x^{(i)}\right)$，不属于$\chi_c$的样本则较小。

将 Actionlet S置信评分定义为：

$$\mathrm{Conf}_S=\max_{i\in\chi_c}\log P_S\left(y^{(i)}=c\middle|x^{(i)}\right) \tag{12}$$

Actionlet S歧义评分定义为：

$$\mathrm{Amb}_S=\frac{\sum_{i\notin\chi_c}\log P_S\left(y^{(i)}=c\middle|x^{(i)}\right)}{\sum_{i\notin\chi_c}1} \tag{13}$$

Actionlet S的辨别力可以通过其置信评分Conf_S和歧义评分Amb_S评价，通过这两个评分可以挑选出最好的 Actionlet。但由于一个动作包含指数级数量的 Actionlet，枚举所有 Actionlet 非常耗时，因此需要借助一种基于 Apriori 的数据挖掘算法，实现高效地发现高判别力的 Actionlet。

基于 Aprior 的算法本质上是一种分支定界算法，通过消除小于置信评分阈值的路径，有效地减少搜索空间。如果 Actionlet S'的置信评分$\mathrm{Conf}_{S'}$已经小于置信阈值，我们不需要考虑$S'\subset S$的任何 Actionlet S。

图 4-2 展现了利用 Actionlet-LOP 进行动作识别的性能，其对应的准确率可以达到0.83。相比于骨架关节特征，三维模型特征在整体识别准确率上没有太大的改进，但其对人与物体交互的行为更具识别力。然而，在三维模型特征提取过程中，需要借助占用大量内存空间的点云数据，为实时动作识别造成一定影响。

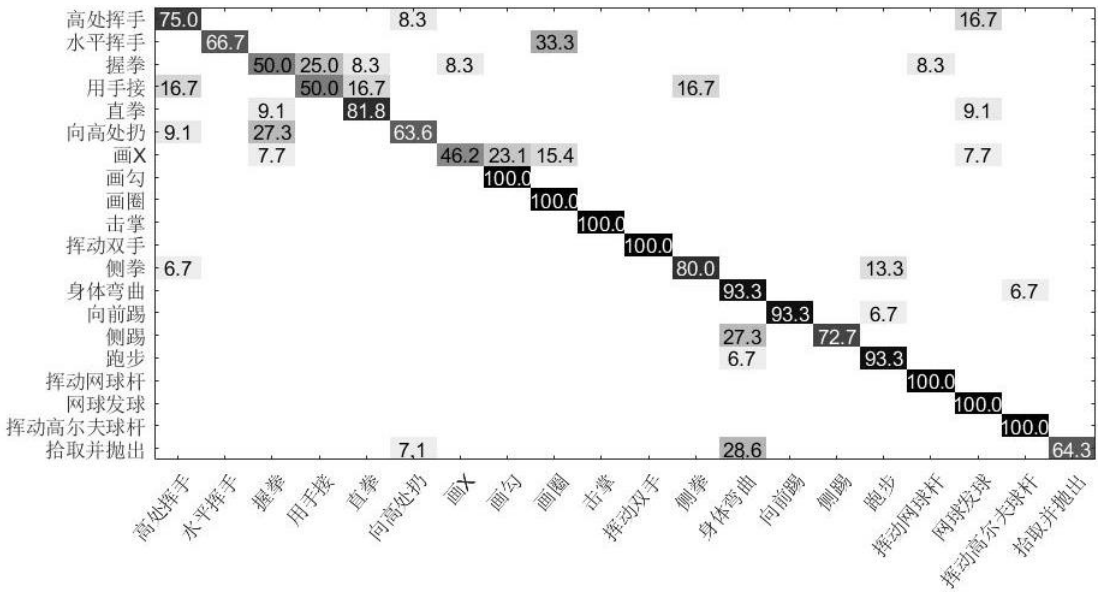


图 4-2 使用 Actionlet-随机占用识别结果的混淆矩阵

第 5 章　基于时-空特征的动作识别方法

动作是一个连续的过程，所以与单帧图像相比，从连续的动作序列提取的特征更具有识别力。与单帧骨架特征和三维模型特征相比，时-空特征产生更高维的数据，因此需要有效的提取动作时间模式的方法。为了识别精度的提高，本章还简要介绍视频序列的标准化方法。

5.1 视频序列的标准化

动作的视频序列的标准化分为两步：长度标准化和时间规整。

5.1.1　视频序列的长度标准化

当前用于动作识别的深度图像数据集种类繁多，且形式不一。例如，本文使用的 MSR Action3D 数据集，每个样本最多有 50 帧，而小样本只有 20 多帧。对于不同的样本，需要进行样本大小（即视频序列的长度）的标准化。为了使样本总数不减少，过小的样本不能直接丢弃。

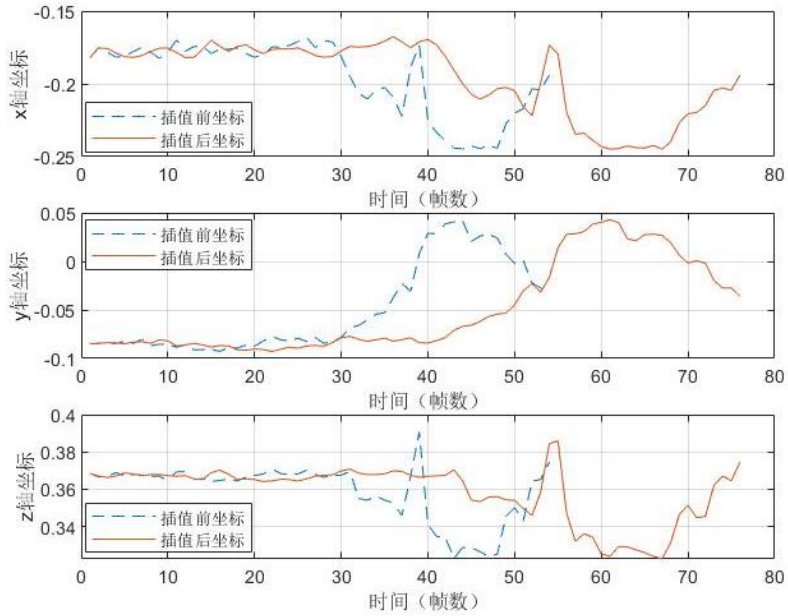


图 5-1 关节点坐标值随时间变化的插值结果示意图

对于骨架关节特征，每一个关节点随时间的变化可看作一条连续的曲线。因此可以考虑使用样本点的插值算法。如图 5-1，图中虚线展现了原始关节坐标值随时间的变化，实线则代表插值后的关节坐标值随时间的变化。插值前的原始坐标序列为 54

帧，插值后为 76 帧。插值曲线使用三次样条曲线，使坐标的运动轨迹具有更平滑的特点。

对于深度图像序列，目前没有较好的长度标准化手段，在实验中使用简单的最近邻插值算法。对于时间长度为T的视频序列$V(t)$，$t = 0,1,\dots,T$，插值后的视频序列$V^*(t) = V(\text{round}(\frac{t\times(l_0-1)}{l_1}))$。$l_0$和$l_1$分别是视频原始长度和插值后的长度，round函数用于数值的取整。

5.1.2 视频序列的时间归整

在语音识别的相关研究中，不同的讲话习惯会造成识别上的误差。这是由于一句话中每个词停留时间不同导致的。在动作识别中，这种现象也同样存在：动作各部分执行的快慢同样影响着识别精确度。为了解决这种现象导致的误差，需要使用一种经典的动态时间规整算法（Dynamic Time Warping，简称 DTW）[26]。

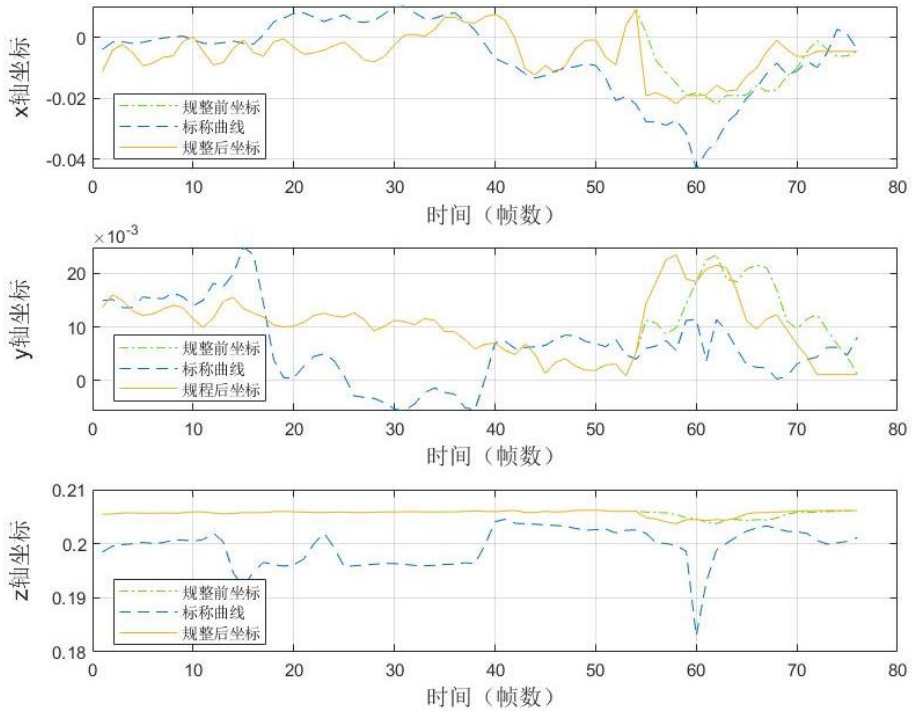


图 5-2 关节点坐标值的时间规整结果示意图

时间规整首先需要对每段动作序列求得具有“代表性”的表示，随后利用这种表示对所有样本进行时间规整。这种特征被称为标称特征（Nominal Feature），即对每类动作的每个关节点求得标称曲线（Nominal Curve）。如图 5-2 所示，虚线为某关节

三维坐标的标称曲线，点划线为规整前的坐标变化曲线，实线代表规整后的坐标变化曲线。计算标称曲线算法流程如表 5-1 所示。

表 5-1 计算标称曲线的算法流程[26]

输入：在$t = 0,1,\dots,T$范围内的曲线$\mathfrak{C}_1(t),\dots,\mathfrak{C}_J(t)$、最大迭代次数$max$和阈值$\delta$。
输出：在$t = 0,1,\dots,T$范围内的标称曲线$\mathfrak{C}(t)$
初始化：$\mathfrak{C}(t) = \mathfrak{C}_1(t)$，迭代次数$iter = 0$。 **当**$iter < max$**时，** 使用 DTW 算法，并按照平方欧式距离，将每条曲线$\mathfrak{C}_j(t)$以$\mathfrak{C}(t)$为标称曲线进行规整，并得到一条规整后的曲线$\mathfrak{C}_j^w(t)$。 通过$\mathfrak{C}'(t)$更新标称曲线，$\mathfrak{C}'(t) = \frac{1}{J}\sum_{j=1}^{J}\mathfrak{C}_j^w(t)$。 如果$\sum_{t=0}^{T}\|\mathfrak{C}'(t) - \mathfrak{C}(t)\|_2^2 \le \delta$（$\|\cdot\|_2^2$代表$\ell_2$范数），**终止迭代**。 $\mathfrak{C}(t) = \mathfrak{C}'(t)$；$iter = iter + 1$；

5.2 视频序列的时间模式提取

对于从每个帧t中提取多种类型的特征，如：3D 关节位置特征、LOP 特征和 ROP 特征等。在这个小节中，我们介绍文献[19]提出的傅立叶时间金字塔模型，来表示动作序列的时间模式。

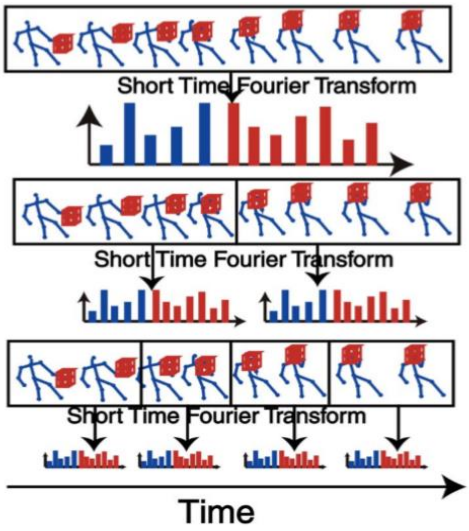


图 5-3 傅里叶时间金字塔原理示意图[19]

傅立叶时间金字塔是一种能够良好展示动作时间结构的描述性特征。为了构建动作的时间结构，除了对整段视频计算傅里叶系数之外，还递归地构建一个金字塔结构，

金字塔的每一层均为视频的若干子分段。在对所有分段使用快速傅里叶变换后，如图5-3所示，最终特征是来自所有分段的傅里叶系数的组合。

对于动作视频样本v，令f_v表示其特征向量，N_v表示特征向量f_v的维数，于是有$f_v = (f_v^1, f_v^2, \ldots, f_v^{N_v})$。其中，每个元素$f_v^n$在完整的视频中是时间的函数，我们可以将其写为$f_v^n(t)$。对于金字塔每层的每个视频分段，快速傅里叶变换被应用于元素$f_v^n(t)$，并获得其傅里叶系数。最终，视频样本v的傅立叶时间金字塔特征被定义为金字塔所有层次低频系数的组合。

使用傅立叶时空金字塔特征有几个好处。首先，通过丢弃高频傅立叶系数，所提取的特征对噪声具有鲁棒性。其次，该特征对时间错位不敏感，因为时间转换的时间序列具有相同的傅里叶系数幅度。

表 5-2 傅立叶时间金字塔对骨架特征提取时间特征的动作识别准确率

	关节绝对位置	关节相对位置	特征关节
傅立叶时间金字塔模型	0.89	0.89	0.93

如表 5-2 所示，使用傅立叶时间金字塔模型对动作序列进行时间特征提取对动作识别性能具有极大的提升。在图 5-4 中，只有用手接（hand catch）和握拳（hammer）这两种动作出现易混淆的情况。相比于骨架关节特征和三维模型特征，动作的时-空特征明显地消除了动作混淆的情况，但其识别过程也较为复杂，不满足实时性要求。

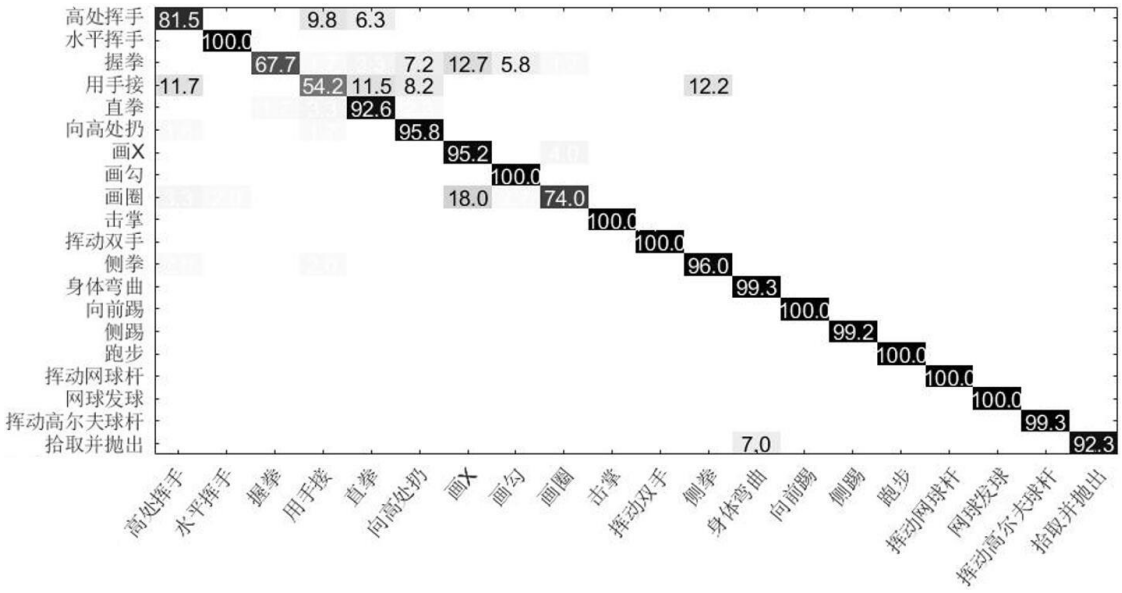


图 5-4 使用傅里叶时间金字塔-特征关节的识别结果混淆矩阵

第 6 章　基于学习特征的动作识别方法

深度学习模型可以通过神经网络的层次结构，直接提取原始数据的高级特征。这种高级特征被称为学习特征。学习特征不再需要我们设计算法来量化或表示姿态和动作的特征，其提取过程由神经网络完成。对于深度动作序列，本文使用三维卷积神经网络提取动作的学习特征。

6.1 原始数据的处理

原始数据的处理涉及数据预处理和数据增强等过程。

6.1.1　数据预处理

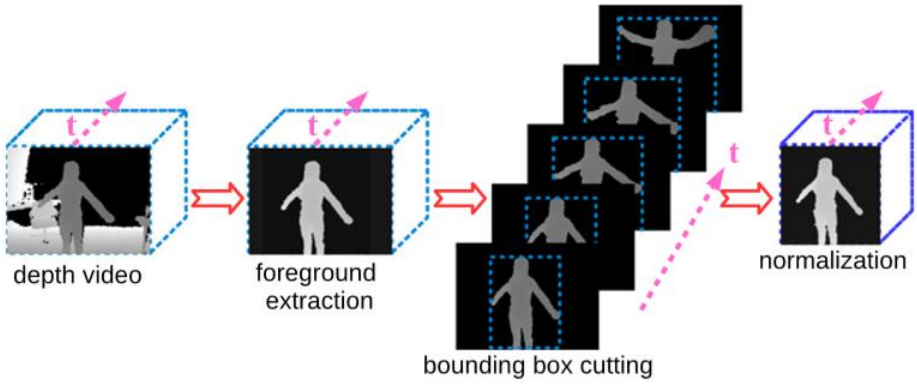


图 6-1 深度数据集图像预处理流程示意图[25]

为了使神经网络更快地拟合并且具有更高的精确度，本文使用前景提取、边框切割和标准化等方法对视频的每一帧图像进行简单的预处理。

前景提取通过把人体以外的像素深度值设为 0，达到去除背景干扰的目的。部分动作识别数据集没有复杂的背景变换，出于简化实验的考虑，本文省去了对这些数据集（如：MSR Action3D）进行前景提取的过程。

通过寻找人体的最大边框，边框切割可以减少图像中人体的位移信息，从而避免人体运动对动作及行为识别造成的干扰。

标准化过程涉及图像大小的标准化与深度值的标准化。在经过图像大小标准化后，每一帧图像的大小均为32×32，而深度数据在经过标准化后将被映射到$[-1,1]$之间。

6.1.2 数据增强

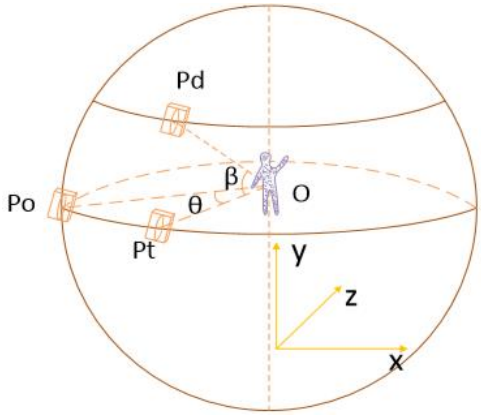


图 6-2 提取多视角点云图像的原理示意图[27]

对于较小的深度动作数据集，若进一步提高识别正确率，可考虑加大数据量。例如，使用三个不同视角的点云图像作为神经网络的输入，对每个视角的数据训练三维卷积神经网络，将获得的后验概率作为决策依据[27]。在如图 6-2 所示的世界坐标系中，视角P_d处的点云图像可以通过两步旋转得到：先从P_o处向P_t处旋转，再从P_t处向P_d处旋转。P_d处的坐标R可以通过公式 14 计算得到：

$$R = R_y R_z [X, Y, Z, 1]^{\mathrm{T}} \tag{14}$$

在公式 14 中使用的是与图 6-2 不同的右手坐标系，其原点位于视点P_o处。其中R_y代表围绕y轴旋转，R_z代表围绕z轴旋转。R_y和R_z分别为：

$$R_y = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & cos(\theta) & -sin(\theta) & Z \cdot sin(\theta) \\ 0 & sin(\theta) & cos(\theta) & Z \cdot (1 - cos(\theta)) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{15}$$

$$R_z = \begin{bmatrix} cos(\beta) & 0 & sin(\beta) & -Z \cdot sin(\beta) \\ 0 & 1 & 0 & 0 \\ -sin(\beta) & 0 & cos(\beta) & Z \cdot (1 - cos(\beta)) \\ 0 & 0 & 0 & 1 \end{bmatrix} \tag{16}$$

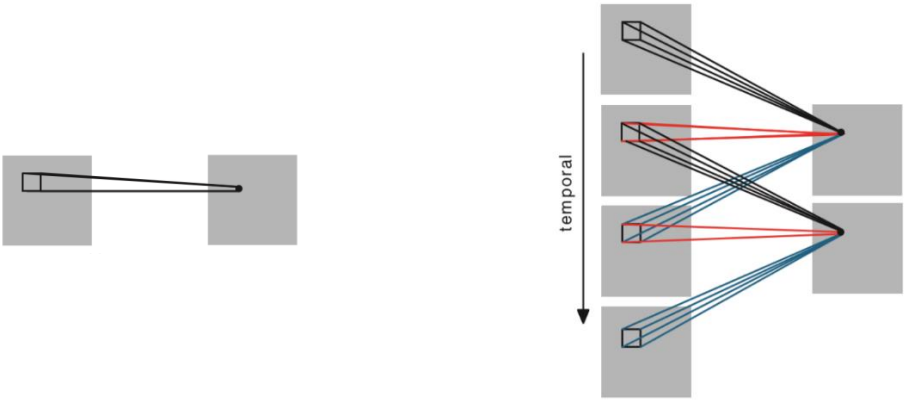
6.2 识别框架的设计

基于学习特征的动作识别框架有典型的两种设计思路，一是直接使用纯神经网络结构对原始数据进行学习，或者使用神经网络与其他分类器组合使用。本小节将对两种思路进行对比。

6.2.1 神经网络参数的设定

如图 6-3 (a)所示，对于单张图片内容的识别，通常使用二维卷积核进行卷积操作。由于每次卷积运算只涉及单张图片，所以无法表示动作在时间维度上的变化。使用三维卷积运算可以更好的捕获视频中动作的时间特征，因为三维卷积往往涉及视频中相邻的若干帧。

如图 6-3 (b)所示的卷积过程在时间维度上使用相邻的 3 帧进行卷积运算。运算形成的卷积层是一个“三维数据块”，卷积层中的每一个值都来自原始图像对应局部位置相邻 3 帧组合成的“数据块”。



(a) 二维卷积过程　　(b) 三维卷积过程

图 6-3 二维卷积(a)和三维卷积(b)的对比示意图[24]

本文在实验中使用文献[25]中对三维卷积神经网络的设置参数，测试数据集为 MSR Action3D。其中原始图像序列的大小为height × width × time（32 × 32 × 38），卷积层 CL1 大小为5 × 5 × 7（5 × 5为空间维度，7是时间维度），CL2 大小为5 × 5 × 5。池化层 MP1 和 MP2 均使用2 × 2 × 2的下采样。完整的神经网络结构如图 6-4 所示，图中用虚线围起来的部分可用于高级特征的提取。

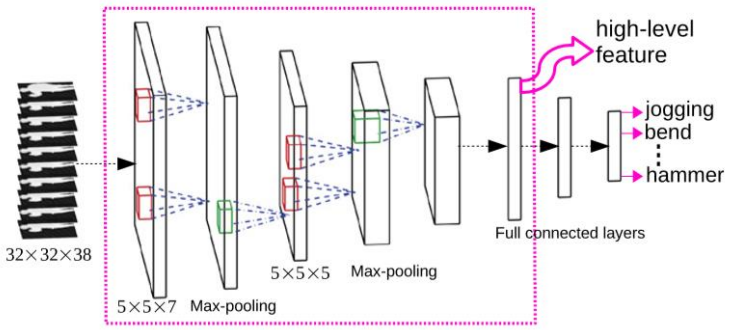


图 6-4 三维卷积神经网络结构示意图[25]

使用完整神经网络进行动作识别的混淆矩阵如图 6-5 和表 6-1 所示，可见使用两层卷积层的卷积神经网络对深度动作序列进行直接识别结果并不理想，

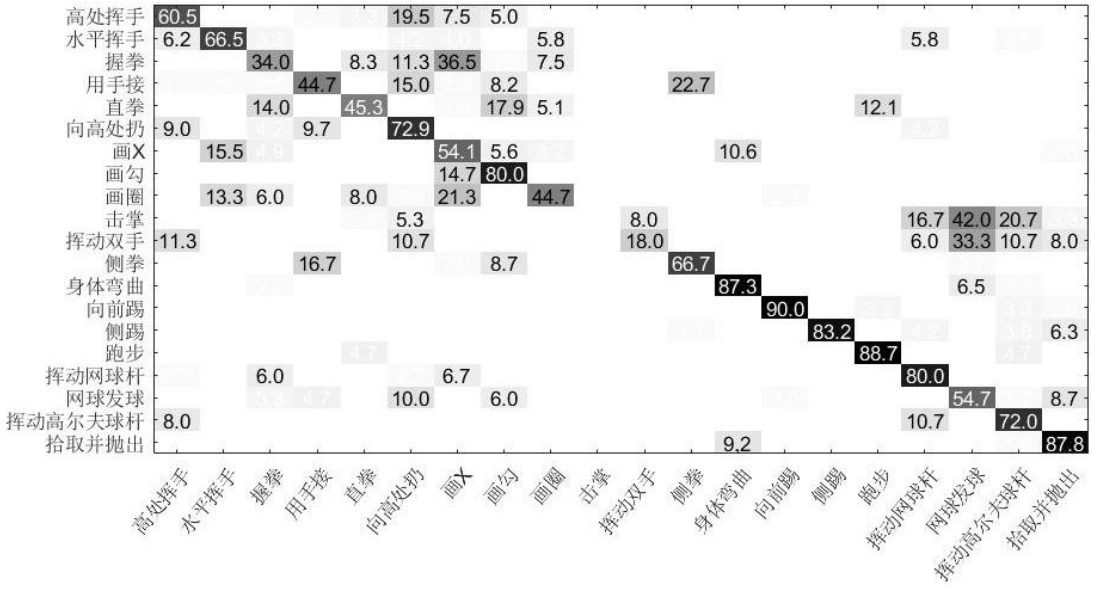


图 6-5 三维卷积神经网络识别结果的混淆矩阵

表 6-1 使用三维卷积神经网络和学习特征的动作识别准确率

动作识别模型名称	三维卷积神经网络	学习特征-SVM 分类器决策融合
准确率	0.61	0.74

6.2.2 分类器决策的融合

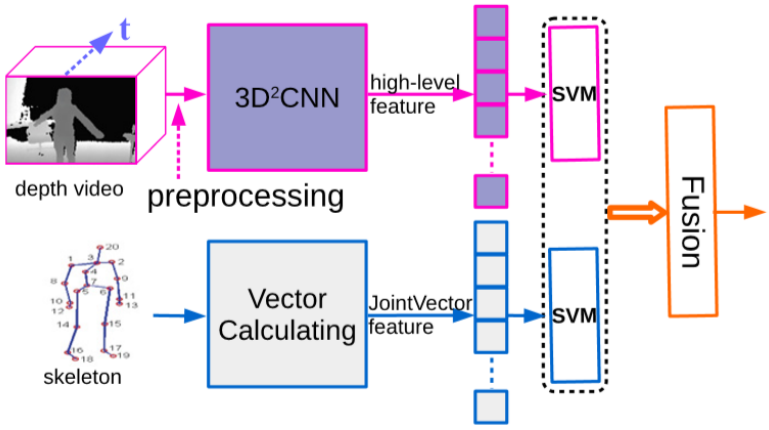


图 6-6 人体动作识别框架的结构及原理示意图[25]

本章参照文献[25]使用的人体动作识别框架对动作及行为进行识别。如图 6-6 所示，该框架使用三维卷积神经网络提取高级学习特征，并和骨架关节特征（特征关节）

一同作为输入，训练 SVM 模型。最终，采取使用多个 SVM 模型进行决策融合的方法，对得到的后验概率直接相加取最小值，判断样本所属分类。

如图 6-7 和表 6-1 所示，相对完全使用卷积神经网络进行动作识别，使用 SVM 分类器进行决策融合后，与纯三维神经网络模型相比，识别性能和正确率有了明显的改善。本章实验环境使用 Google 公司免费提供的 Colaboratory 环境。该环境配备了 Tesla T4 GPU，16GB 显存，同时还有 12.7GB 的内存。实验模型在此环境下平均每批数据仅需计算 353 毫秒。

从实验结果上，可以得出使用神经网络或多模型决策融合的方式可以获得一定的准确率。但是由于神经网络的搭建需要一定的硬件资源，在其无法得到满足的环境下，实时性和鲁棒性均存在挑战。

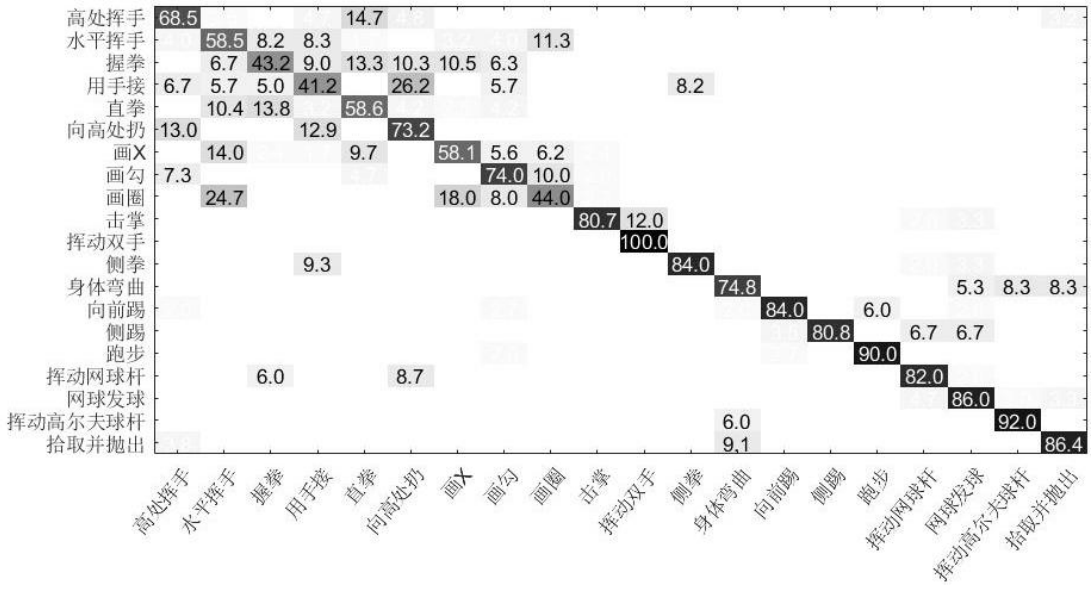


图 6-7 分类器决策融合识别结果的混淆矩阵

参考文献

[1] Vinh T Q , Tri N T . Hand gesture recognition based on depth image using kinect sensor[C]// Information & Computer Science. IEEE, 2015.

[2] Chuan C H , Chen Y N , Fan K C . Human Action Recognition Based on Action Forests Model Using Kinect Camera[C]// 2016 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA). IEEE, 2016.

[3] Fujino M , Zin T T . Action Recognition System with the Microsoft KinectV2 Using a Hidden Markov Model[C]// Third International Conference on Computing Measurement Control & Sensor Network. IEEE, 2017.

[4] Wang Z, Mirbozorgi S A, Ghovanloo M. Towards a Kinect-based behavior recognition and analysis system for small animals[C]//Biomedical Circuits and Systems Conference (BioCAS), 2015 IEEE. IEEE, 2015: 1-4.

[5] Jonguk L , Long J , Daihee P , et al. Automatic Recognition of Aggressive Behavior in Pigs Using a Kinect Depth Sensor[J]. Sensors, 2016, 16(5):631-.

[6] Banerjee T, Yefimova M, Keller J M, et al. Exploratory analysis of older adults' sedentary behavior in the primary living area using kinect depth data[J]. Journal of Ambient Intelligence and Smart Environments, 2017, 9(2): 163-179.

[7] Dawar N, Kehtarnavaz N. Real-Time Continuous Detection and Recognition of Subject-Specific Smart TV Gestures via Fusion of Depth and Inertial Sensing[J]. IEEE Access, 2018:1-1.

[8] Chen C, Jafari R, Kehtarnavaz N. A survey of depth and inertial sensor fusion for human action recognition[J]. Multimedia Tools and Applications, 2017, 76(3): 4405-4425.

[9] Real-time human pose recognition in parts from single depth images[J]. Communications of the ACM, 2013, 56(1):116.

[10] Schuldt C , Laptev I , Caputo B . Recognizing human actions: a local SVM approach[C]// Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. IEEE, 2004.

[11] Dollar P , Rabaud V , Cottrell G , et al. Behavior recognition via sparse spatio-temporal

features[C]// Joint IEEE International Workshop on Visual Surveillance & Performance Evaluation of Tracking & Surveillance. IEEE, 2006.

[12] Laptev I , Marszalek M , Schmid C , et al. Learning realistic human actions from movies[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2008.

[13] 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.

[14] Real-time human pose recognition in parts from single depth images[J]. Communications of the ACM, 2013, 56(1):116.

[15] Yang X , Tian Y L . EigenJoints-based action recognition using Naïve-Bayes-Nearest-Neighbor[C]// 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPR Workshops). IEEE Computer Society, 2012.

[16] Xia L , Chen C C , Aggarwal J K . View invariant human action recognition using histograms of 3D joints[C]// Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on. IEEE, 2012.

[17] Yang X , Zhang C , Tian Y L . Recognizing actions using depth motion maps-based histograms of oriented gradients[C]// Acm International Conference on Multimedia. ACM, 2012.

[18] Wang J , Liu Z , Chorowski J , et al. Robust 3D Action Recognition with Random Occupancy Patterns[M]// Computer Vision – ECCV 2012. 2012.

[19] Wang J , Liu Z , Wu Y , et al. Learning Actionlet Ensemble for 3D Human Action Recognition[J]. IEEE Transactions on Software Engineering, 2013, 36(5):914-927.

[20] Chen L , Wei H , Ferryman J . A survey of human motion analysis using depth imagery[J]. Pattern Recognition Letters, 2013, 34(15):1995-2006.

[21] Vieira A W, Nascimento E R, Oliveira G L, et al. STOP: Space-Time Occupancy Patterns for 3D Action Recognition from Depth Map Sequences[J]. 2012.

[22] https://blog.csdn.net/xierhacker/article/details/70903617

[23] 陈万军, 张二虎. 基于深度信息的人体动作识别研究综述[J]. 西安理工大学学报, 2015(3):253-264.

[24] 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.

[25] Liu Z, Zhang C, Tian Y. 3D-based deep convolutional neural network for action recognition with depth sequences[J]. Image and Vision Computing, 2016, 55: 93-100.

[26] Vemulapalli R , Arrate F , Chellappa R . Human Action Recognition by Representing 3D Skeletons as Points in a Lie Group[C]// 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2014.

[27] Wang P , Li W , Gao Z , et al. Deep Convolutional Neural Networks for Action Recognition Using Depth Map Sequences[J]. Computer Science, 2015.

致　　谢

大学生活行将结束，在此感谢每一位在成长路上帮助我的老师和同学，感谢你们对我的栽培和帮助。其中，特别感谢王琼老师对大创的指导和宋凤义老师对毕业论文的指导。

感谢父母四年来的对我生活和学业的支持。感谢漫威电影和进击的巨人为我的课余时光带来快乐。感谢女朋友陪我度过考研日日夜夜，没有你真的难以坚持。

本科期间主要研究成果

申请软件著作权：

[1] 王琼，**仇思宇**，钟婷，姚悦. 基于 Kinect 的手腕三维鼠标驱动程序，V1.0，2018SR580468

完成项目：

[1] 姚悦，钟婷，**仇思宇**. 大学生创新训练项目. 南京师范大学，2017-2018，0.2 万.