# 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

随着高分辨率彩色传感器的进步，实现具备行为分析能力的视频监控系统已经不再是难点，但是其在动作识别精度上没有足够的保证。而低成本深度传感器（如Kinect）的出现，为更加精确的动作识别提供条件，也为可靠的视频实时动作监测技术提供了无限可能。

本文在对最近几年提出的基于深度信息识别动作的相关方法进行回顾的基础上，总结了当前动作识别具有代表性的方法，并且以骨架关节特征、三维模型特征、空-时特征和学习特征作为思路对现有动作表示方法进行分类，并以此讨论不同方法在不同环境（背景变化、视角变换、噪声和遮挡）中的适用性。最后，为了应对实时视频序列所产生的大量数据，本文使用在线学习的方式对动作识别框架进行优化，并提出对未来研究方向的相关建议。

**关键词：**视频监控；动作识别；在线学习；……

# Abstract

With the advancement of high-resolution color sensors, it is no longer difficult to implement a video surveillance system with behavior analysis capabilities, but it does not have sufficient guarantee for action recognition accuracy. The emergence of low-cost depth sensors (such as Kinect) provides conditions for more accurate action recognition and unlimited possibilities for reliable video real-time action monitoring technology.

By reviewing the related methods based on depth information for action recognition in recent years, this paper summarizes the representative current methods of action recognition, with skeletal joint features, 3D model features, space-time features and learning features. The idea is to classify existing action representation methods and discuss the applicability of different methods in different environments (background changes, perspective transformation, noise, and occlusion). Finally, in order to cope with the large amount of data generated by real-time video sequences, this paper uses the online learning method to optimize the action recognition framework and propose relevant suggestions for future research directions.

**Key words:** video surveillance, action recognition, online learning

# 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

……

### 1.2.1 彩色图像特点及其劣势

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。如[10]中，局部特征与 SVM 分类器的结合，证明了可以通过度量局部特征[1]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，这种算法使用空时空特征[2]点（单张图像上的局部特征）来表征行为。为了降低动作分类结果对背景杂乱，遮挡和比例变化的敏感度，[13]中介绍了直接运动识别方法：使用时空特征包(BoF)[3]，判断人体运动特征（判断局部图像块的运动如何进行），而不是通过恢复人的身体二维模型或三维模型，以其局部结构特征实现动作分类。动态能量图像(MEI)[4]和运动历史图像(MHI)[5]在[13]中作为运动模板被引入，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[8]因素影响识别结果。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。

[1] 局部特征：图像中的图案或不同结构，例如点，边缘或小图像块。 它们通常与图像贴片相关联，这些贴片在其纹理，颜色或强度上与其周围环境不同。

[2] 时空特征：短的局部视频序列，例如眼睛张开或膝盖弯曲，或者用于快速前后移动的爪子。 然后根据存在的特征点的类型和位置充分描述行为。

[3] 时空特征包(BoF)：一组时空特征的集合，反应了局部特征的运动特征如时空轨迹、周期性等。

[4] 动态能量图像（Motion-Energy Images，简称 MEI）：表示图像序列中发生运动位置灰度图像。

[5] 运动历史图像（Motion-History Images，简称 MHI）：标量值图像，其中每个像素的值是运动新近度的函数。

### 1.2.2 深度图像特点及其优势

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。

与由摄像机捕获的传统 RGB 图像相比，深度相机生成的深度图像显示出对照明变化不敏感并且在人类动作识别中具有高性能、实时性强等特点。同时，人体骨骼信息也可以从深度图像中获得[9]。微软的 Kinect 设备的原理就是利用了深度或距离传感器，进行人体骨骼和动作的识别。

### 1.2.3 动作的骨架关节表示方法

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（AF）[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和自由度。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，将其分类为若干已知动作。

### 1.2.4 动作的三维模型表示方法

在[17]中，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相关的运动能量，组合为深度运动图（DMM）。从三个 DMM 中提取定向梯度柱状图（HOG）并将其连接为 DMM-HOG 表示动作。

与投影方法将三维图像转变为二维图像的思路不同，将空间划分为若干子空间，并计算落入子空间中的占有体积的特征被称为随机占用模式[18]（Random

Occupancy Pattern，简称 ROP）。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。

在 ROP 特征的基础上，在文献[19]中提出一种新的人体动作特征和一种新的动作识别方法：局部占用模式（Local Occupancy Pattern，简称 LOP）和动作类集合模型（Actionlet Ensemble Model），并明确了动作类是关节子集的特征的特定组合。新的动作模型对于特征中的错误更加健壮，并且可以更好地表征动作中的类内变化。

### 1.2.5 动作的时-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。同时，利用滤波技术对连续的动作信息进行滤波可以达到去除噪声的效果，实现更加精确的动作预测。

将 1.2.3 阐述的特征关节按相同时间顺序组合，作为朴素贝叶斯最邻近分类器（Naïve Bayes Nearest Neighbor，简称 NBNN）的输入实现动作分类是最为简单的方法。对[18]中提到的 ROP 特征进行稀疏编码，其编码系数按时间顺序组合后，使用支持向量机的实现动作识别[19]。空-时占用模式（Space-Time Occupancy Pattern，简称 STOP）[21]也与之类似，但他们略有不同。STOP 特征使用相同尺寸的测量空-时体积。

# 第 2 章　动作识别方法评价体系

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 |
| MSR DailyActivities3D | 16 | 320 | 10 个演员；大部分样本涉及到人和物体的交互；捕获的 3D 关节点坐标受噪声污染严重 |
| 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 |
| RGBD-HuDaAct | 12 | 1189 | 30 个演员，每类动作每个演员执行 2~4 次；提供深度图像与 RGB 图像，样本中混有随机背景动作 |
| MSR ActionPairs | 6 | 180 | 10 个演员，每类动作每个演员执行 3 次；每个动作对有相似的运动和形状 |

| | | | |
|---|---|---|---|
| UWA3D Multiview Activity | 30 | 720 | 10 个演员，每类动作每个演员执行 2~3 次；存在自遮挡和高度相似性；具有视角和尺度变化；提供关节点的 3D 坐标数据、深度图像、深度的前景分割图像与 RGB 图像 |
| 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……

## 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和交叉验证的方法。

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。

交叉验证是用来验证分类器性能的一种常用统计分析方法，基本思想是按照一定的划分方式将原始数据集进行分组，一部分作为训练集，另一部分作为验证集。首先用训练集对分类器进行训练，再利用验证集来测试训练得到的模型。评价分类器的性能指标将使用验证集中的测试结果得出。

在本文中将使用跨目标验证和交叉验证相结合的方法对多用方法进行评估。

### 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 2-2 四种情况之一：

表 2-2 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 2-3 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 2-3 分类准确率的二维分布表**

| | | 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$M \times M$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下：

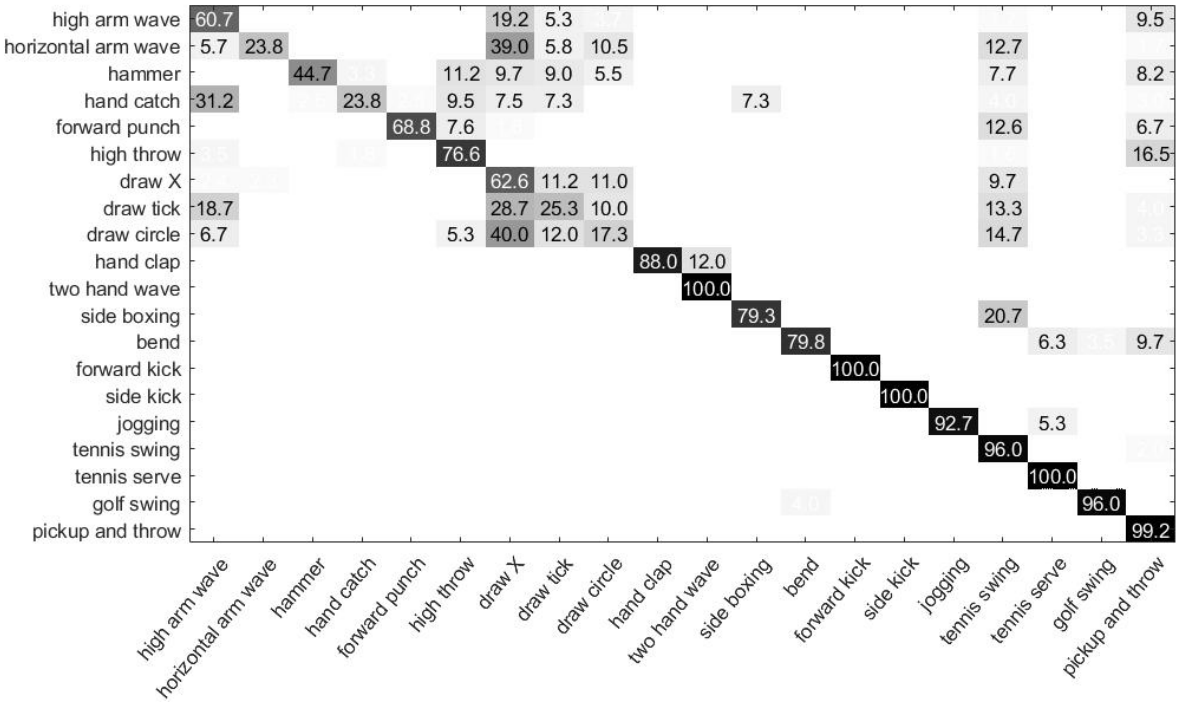



**图 2-1 评估动作分类模型的混淆矩阵示意图**

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，high throw、draw X、draw tick、draw circle、tennis swing 和 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

在 20 世纪 70 年代进行了一项着名的实验[20]，其中一个人穿着黑色且各关节附有光源。另一人在观察移动灯光的同时，可以识别正在执行的动作。这项工作激发了后来的研究人员，使人们认识到可以从主要的关节运动中感知人类行为。早期利用多个相机并配以关节点标记来进行可靠的关节点位置估计，从而实现人体的动作识别，且具有较高的精确性。但是，在仅仅使用单一相机的情况下会出现自身遮挡和互相遮挡。而且，基于标记点的动作识别只能在特殊的室内环境使用，且价格昂贵。

## 3.1 身体部位判断和关节点获取

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

对任意关节点的平面位置$x$，$f_\theta$表示该关节点的深度特征。

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

其中$d_I(x)$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位置的向量$u$和$v$。通过$\frac{1}{d_I(x)}$使偏移规范化，确保特征不随深度相机的位置变化而变化。因此，这些特征具有平移不变的特性。如果偏移像素位于背景上或图像边界之外，则$d_I(x)$被给予大的正数常量值。

### 3.1.2 关节位置预测

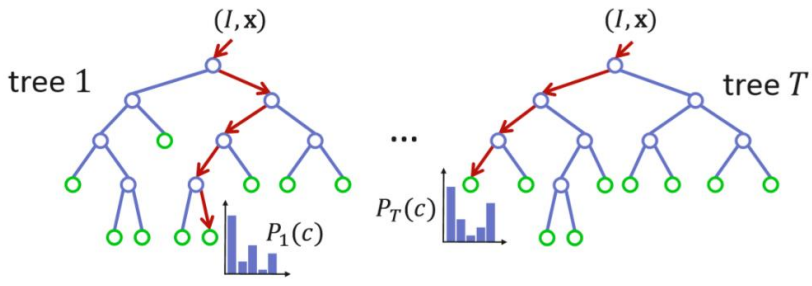



图 3-1 随机决策森林原理示意图[14]

分类器选择随机决策森林模型，该模型是决策树的集合。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。

如图 3-1 所示，森林是$T$个决策树的集合，每棵树由中间节点和叶节点组成。每个中间节点包含两个元素：特征$f_\theta$和阈值$\tau$。为了对图像$I$的像素$x$进行分类，从根处开始重复使用公式 1 并与阈值$\tau$比较，进而向左或向右访问。在树 t 中到达的叶节点处，存储由身体部位标签$c$得到分布$P_t(c|I,x)$，即各个偏移坐标所占的权重。对森林中的所有树的分布求平均值，以给出最终的分布。

我们采用基于加权高斯核的均值移位[]的局部模式发现方法，将每个身体部位的密度估计量定义为：

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

其中$\hat{x}$是为世界坐标，$N$是图像像素的数量，$w_{ic}$是像素加权，$\hat{x}_i$是图像像素$x_i$在给定深度$d_I(x_i)$下重投影到世界坐标空间中的位置，并且$b_c$是学习的每部分带宽。像素加权$w_{ic}$为像素$x_i$处所占的权重和像素深度平方的乘积：

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

估计量高于学习概率阈值$\lambda_c$的像素被用作部位 c 的起始点。给出最终的置信估计值作为到达每个模式的像素权重的总和。

## 3.2 动作的表示和动作的识别

### 3.2.1 姿态特征获取

对于不包含骨架关节点坐标的数据集，在使用以上算法完成关节点的识别后，对于深度动作序列的每一帧$c$，都包含$N$个关节的空间坐标：$X=\{x_1,x_2,\dots,x_N\}$。以关节的空间位置差异表示动作信息，需其包含三个特征：姿态特征$f_{cc}$，运动特征$f_{cp}$和偏移特征$f_{ci}$。

其中$f_{cc}$为当前帧$c$中的成对关节差异，表征了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:

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

$f_{ci}$捕获了帧$c$和初始帧$i$之间的成对关节差异，表征当前帧$c$中的偏移特征和整体位移：

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

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。

### 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：

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

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$C^* = \underset{C}{\text{argmin}}\, \text{dist}(v_C, v^*)$。

不同的姿态特征与分类器得到的正确率和混淆矩阵如图 3-2 所示，可以看到，骨架关节特征中，特征关节和关节位置直方图作具有较强的识别力。NBNN 分类器和 HMM 模式动作分类的性能好于 SVM 分类器。

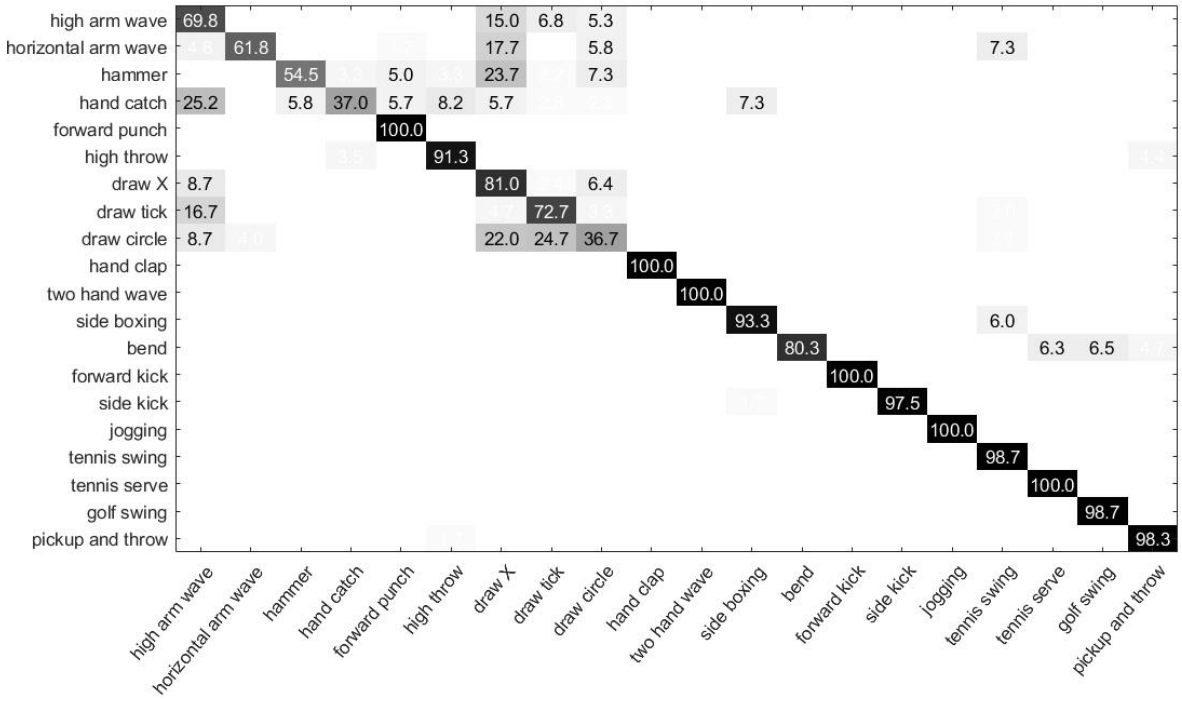



图 3-2 使用特征关节和 NBNN 分类器的分类结果

# 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。而三维模型则可以将与人体相接触的物体相互融合，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：

$$\begin{cases} x_w = z_c \cdot (u - u_0) \cdot C_x \\ y_w = z_c \cdot (v - v_0) \cdot C_y \\ z_w = z_c \end{cases} \tag{8}$$

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形成的特征被称为局部占用模式（Local Occupancy Patterns，简称 LOP）。

### 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标准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，其局部占用信息为：

$$o_{xyz} = \delta(\sum_{q \in b_{xyz}} I_q) \tag{9}$$

其中，如果点云在像素$q$中存在坐标点，则$I_q = 1$，否则$I_q = 0$。$\delta(.)$是sigmoid标准化函数：$\delta(x) = \frac{1}{1+e^{-\beta x}}$。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。

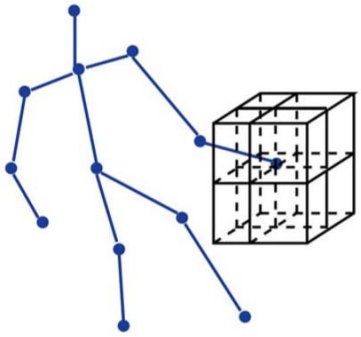


图 4-1 局部占用模式计算方法示意图[19]

## 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，需要使用数据挖掘算法。

### 4.2.1 动作的关节关系的挖掘

对于训练集合中的第$i$个样本，其内容为$(x^{(i)}, y^{(i)})$， $x^{(i)}$是第$i$个样本的特征，$y^{(i)}$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，其分类标签$y^{(i)}$等于动作标签$c$的概率表示为$P_j(y^{(i)}=c|x^{(i)})$，该概率可以通过成对耦合方法从成对概率估计出。

Actionlet 判断动作类型使用并集运算，即：对于每个的 Actionlet，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 才预测$y^{(i)}=c$。且每个关节预测$y^{(i)}=c$的概率是独立的，分类标签$y^{(i)}$等于 Actionlet $S$中的样本$x^{(i)}$的动作分类$c$的概率为：

$$P_S\left(y^{(i)}=c\middle|x^{(i)}\right)=\prod_{j\in S}P_j(y^{(i)}=c|x^{(i)}) \tag{10}$$

将$\chi_c$定义为具有类标签$c$：$\{i:t^{(i)}=c\}$的训练集合。对于有判别力的 Actionlet，$\chi_c$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

$$\mathrm{Conf}_S=\max_{i\in\chi_c}\log P_S\left(y^{(i)}=c\middle|x^{(i)}\right) \tag{11}$$

Actionlet $S$歧义评分

$$\mathrm{Amb}_S=\frac{\sum_{i\notin\chi_c}\log P_S\left(y^{(i)}=c\middle|x^{(i)}\right)}{\sum_{i\notin\chi_c}1} \tag{12}$$

Actionlet $S$的辨别力可以通过其置信评分$\mathrm{Conf}_S$和歧义评分$\mathrm{Amb}_S$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动作，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 $S'$。

# 第 5 章　基于时-空特征的动作识别方法

## 5.1 视频序列的数据冗余度降低

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]提出傅立叶时间金字塔模型，来表示这些帧级特征的时间模式。

傅立叶时空金字塔是一种能够良好展示动作时间结构的描述性特征。为了捕捉动作的时间结构，除了全局傅里叶系数之外，我们递归地将动作划分为金字塔，并对所有分段使用短时傅里叶变换，如图所示。最终特征是来自所有段的傅里叶系数的组合。

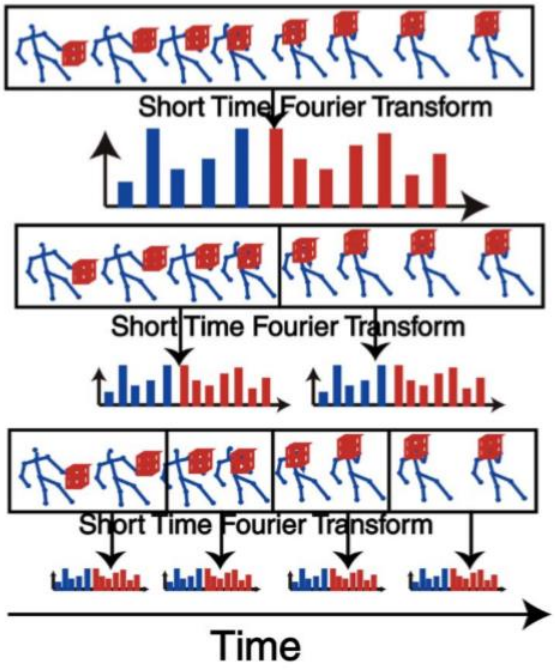



图 5-1 傅里叶时间金字塔原理示意图

对于每个关节$j$，令$g_j(p_j, o_j)$表示其整体特征向量，其中$p_j$是其 3D 成对位置向量而$o_j$是其 LOP 向量。令$N_j$表示$g_j$的维数，即$g_j = (g_1, g_2, \dots, g_{N_j})$。每个元素$g_n$是时间的函数，我们可以将其写为$g_j[t]$。对于每个金字塔等级的每个时间段，我们将短傅里叶变换应用于元素$g_n[t]$，并获得其傅里叶系数，并将它们用作慢速频率系数

作为特征。关节$j$处的傅立叶时间金字塔特征是定义为金字塔所有层次的低频系数，并表示为$G_j$。

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，所提出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。最后，动作的时间结构以金字塔结构为特征

## 5.2 视频序列特征的分批处理

由于视频序列涉及大量实时更新的数据，将来自硬盘或者网络的数据一次性读入并提取特征是不现实的。因此，对视频序列进行分配处理，就要借助增量学习的方法。增量学习思想可以描述为：每当新增数据（读取数据）时，并不需要重建已有的模型，而是原有模型的基础上，仅对由于新增数据所引起的变化进行更新。增量训练如图 5-2 所示。

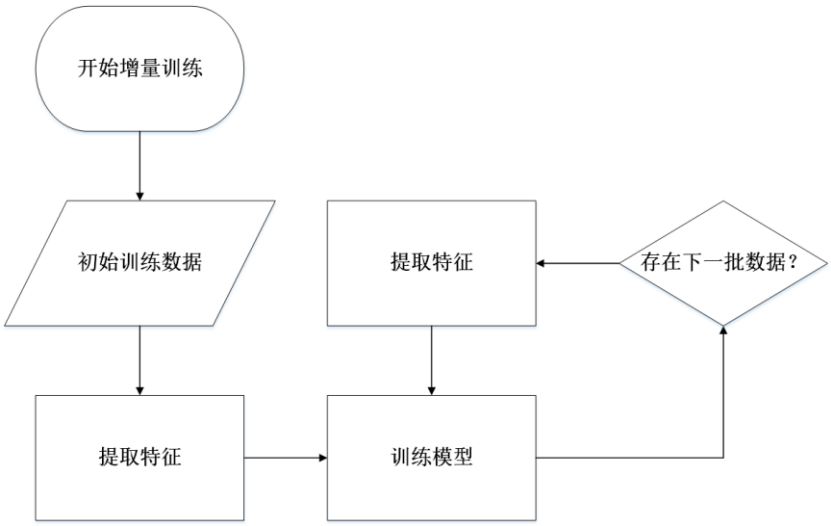



图 5-2 增量训练流程图

增量学习的作用体现在两个方面：一方面不需要一次性从外部存储设备中读取所有数据，减少了内存资源的占用；另一方面充分利用了已经部署用于训练的数据集，减少了后续训练的时间，符合视频序列识别的特点。

### 5.2.1 识别精度与数据冗余的平衡

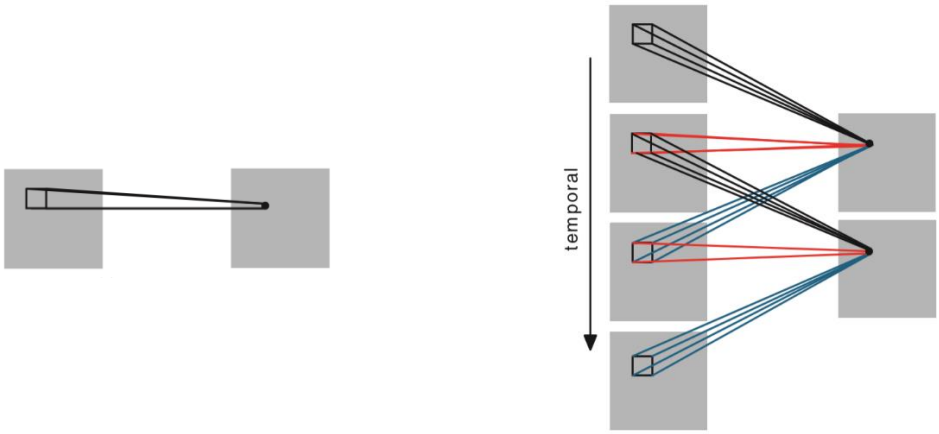
降低数据冗余度的方法依赖样本特征与所属分类的相关性，而往往样本容量越大越容易正确地分析出特征和分类的相关性。同时，样本容量大又必将导致数据冗

余度上升，模型运行效率变低。反之，样本容量减小使模型运行效率增高的同时，识别精度将会下降。因此需要在识别精度和数据冗余之间找到一个平衡点，达到我们需要的效果。

# 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，这种特征与之前章节所提到的骨架关节特征、三维模型特征和时-空特征均有不同。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。如图 6-1 (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。




(a) 二维卷积过程　　(b) 三维卷积过程

图 6-1 二维卷积(a)和三维卷积(b)的对比[24]

如图 6-1 (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”。

## 6.1 高级学习特征的提取

本文中使用文献[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$2 \times 2 \times 2$的下采样。完整的神经网络结构如图 6-2 所示，图中用虚线围起来的部分可用于高级特征的提取。

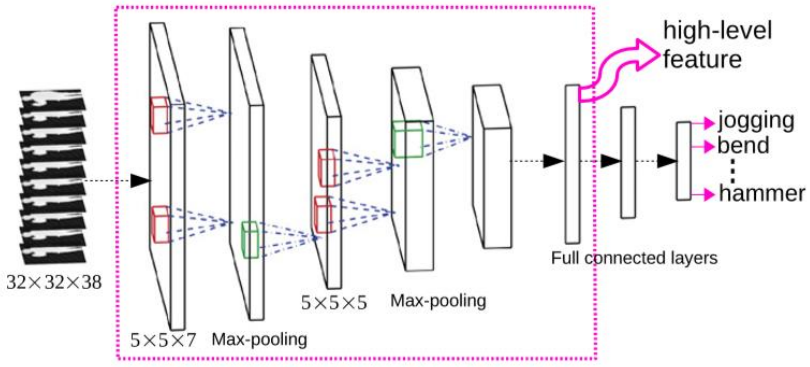
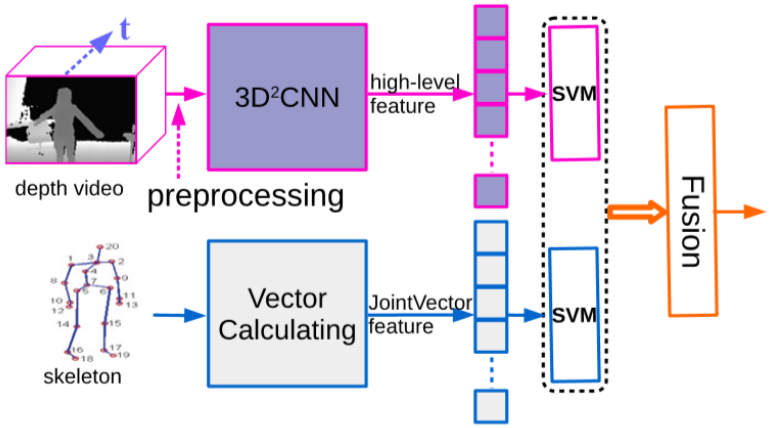



图 6-2 三维卷积神经网络结构示意图[25]

## 6.2 多种特征的融合



# 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.

# 致　　谢

一级标题，三号，宋体，加粗，段前 0.5 行，段后 0.5 行

四年的时间飞逝，……

正文，小四，宋体 + Times New roman，1.5 倍行间距，
首行缩进 2 字符

本页可选

# 本科期间主要研究成果

一级标题，三号，宋体，加粗，段前 0.5 行，段后 0.5 行

**发表论文：**

[1] **……**

[2]

**完成项目：**

[1] **高晨**，***，***. 大学生创新训练项目. 南京师范大，2016-2017，0.3 万.

小四，宋体 + Times New roman，1.5 倍行间距