南京师范大学

毕业设计（论文）开题报告

姓　　名： 仇思宇　　**学　　号：** 21150611

学　　院： 计算机科学与技术学院

专　　业： 计算机科学与技术

题　　目： 基于 Kinect 体感信息的动作及行为识别技术研究

指导教师： 宋凤义

2019 年 1 月 4 日

开题报告填写要求

1．开题报告作为毕业设计（论文）答辩委员会对学生答辩资格审查的依据材料之一。此报告应在指导教师指导下，由学生在毕业设计（论文）工作前期内完成，经指导教师签署意见及院、系审查后生效；

2.开题报告内容必须用黑墨水笔工整书写或按教务处统一设计的电子文档标准格式（可从教务处网址上下载）打印，禁止打印在其它纸上后剪贴，完成后应及时交给指导教师签署意见；

3．有关年月日等日期的填写，应当按照国标 GB/T 7408—94《数据元和交换格式、信息交换、日期和时间表示法》规定的要求，一律用阿拉伯数字书写。如“2005 年 4 月 26 日”或“2005-04-26”。

毕 业 设 计（论 文）开 题 报 告

1. 本课题的目的及研究意义

研究目的：

图像传感设备在过去十年取得了重大进步，并在各种学科中得到越来越多的应用。使用彩色数字摄像设备进行实时监控已经成为保障公共空间安全、犯罪取证、交通疏导的重要手段。基于彩色摄像设备实现的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：从行为监视，视频分析，人机交互[7]，到辅助生活，健康监控，危险行为预警等相关技术[6]。但是这其中的视频监控技术却不具备分析行为的能力，在无人工参与的条件下一般只能作为事故发生后的证据。

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体感数据[8]，尤其是低成本的深度传感器，进而利用这些数据进行人体动作和行为识别和分析，为实现无人工参与的监控技术提供了无限可能。

为了保障某种特定工作人员人身安全和工作顺利进行，避免威胁员工人身安全的突发行为，如工人突然晕倒、群体打架斗殴、特大灾害的发生，实时监控技术就显得尤为重要。而实时监控技术需要投入大量的人力物力，同时需要监控工作者大量重复简单而枯燥的工作，与当下人工智能和计算机视觉技术飞速发展的时代显得十分矛盾。

研究意义：

本课题列举和比较了现有动作和行为识别技术，通过现有的动作识别方法评价体系，系统地总结了当前这些技术的优点和存在的问题。在探讨各技术的适应场景后，提出了未来人类行为识别研究的特点和方向。

2．本课题的国内外的研究现状

原始图像类型及其优劣势：

1) 彩色图像代表技术及其劣势

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，它们被广泛用于视频序列中的人体动作识别。如[9]中，局部特征与支持向量机分类器的结合，证明了可以通过度量局部特征实现动作识别。在[10]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，[11]中介绍了直接运动识别方法：使用时空特征包(BoF) ，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) 和运动历史图像(MHI) 在[12]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 。

2) 深度图像特点及其优势

虽然基于彩色图像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。

深度相机被大量应用于人体动作识别及其相关领域。基于深度图像的动作识别技术的研究方向主要是人体姿态和手势信息提取与识别等，如基于 Kinect 深度传感器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。在手势识别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]。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。

动作的表示方法：

1) 动作的骨架关节表示方法

利用在[13]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作为特征提取出姿态信息、姿态运动信息，然后使用朴素贝叶斯最邻近分类器[14]进行动作分类是最为简单的方法。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[15]中，使用三维关节位置直方图（Histograms Of 3D Joint, 简称 HOJ3D）表示姿势，通过对深度图像序列的每一帧计算三维关节位置直方图并使用隐含狄利克雷分布（Latent Dirichlet Allocation，简称 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。

2) 动作的三维模型表示方法

在[16]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（Depth Motion Map，简称 DMM）。从三个运动图中提取定向梯度柱状图并将其构成的完整时间序列表示动作。

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[17]（Random Occupancy Pattern，简称 ROP）。在使用稀疏编码对该特征进行编码后，使用支持向量机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[18]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。

3) 动作的空-时特征表示方法：

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。

将特征关节按相同时间尺度组合，作为朴素贝叶斯最邻近分类器（NBNN）

的输入实现动作分类是最为简单的方法。对 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[18]。空-时占用模式（Space-Time Occupancy Pattern，简称 STOP）[19]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。

3. 本课题的研究内容

本论文的主要研究内容分为以下几个方面：

1) 简要介绍基于深度图像的动作特征

简要介绍当前主流文献中使用动作特征的发展过程。通过详细论述各种动作特征的计算方法，讨论其在各种具有噪声、遮挡、背景变换环境下鲁棒性和适用性。本文中将从动作的关节位置特征、动作的三维模型特征、动作的空-时特征以及动作的习得特征，介绍基于深度图像的动作识别方法。

2) 介绍当前动作识别技术评价体系

在介绍当前用于动作识别的深度图像数据集与其包含数据的噪声、遮挡、背景变换环境等特殊性后，本文将使用不同数据集测试同一识别方法的性能，并提供动作识别的评价标准以验证本文提出的关于不同环境下不同识别动作识别方法适用性。

3) 介绍基于不同动作特征的动作识别技术的原理

基于骨架关节特征动作识别方法、基于三维模型特征动作识别方法和基于空-时特征的动作识别方法在工作流程上具有相似性。他们使用不同动作表示方法，包含对深度图像视频序列的单帧处理和多帧处理，并以此为基础，选择不同的分类器对动作样本和动作标签进行拟合的机器学习技术。一般通过机器学习技术实现图像识别的过程如图 2 卷积神经网络的结构图所示，其中包含数据预处理、特征提取、模型训练等。很明显，如果原始训练集中的错误、异常值和噪声（错误测量引入的）太多，系统检测出潜在规律的难度就会变大，性能就会降低，而不相关的特征则会使分类环节变得复杂低效。因此需要在数据预处理阶段进行数据清洗，使用降维算法减少冗余数据。模型训练由分类器实现，本文中所介绍的分类器包含：朴素贝叶斯分类器、随机决策树与随机森林、支持向量机等。

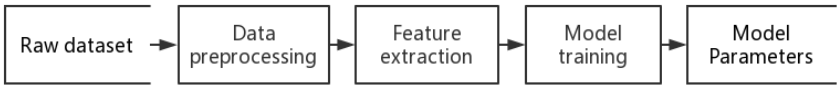

图 1 机器学习图像识别流程

与以上三种方法略有不同，基于学习特征的动作识别方法使用深度卷积神经网络（Convolutional Neural Network，简称 CNN），这是一种专为图像识别设计的深度神经网络。典型的深度神经网络有 AlexNet，VGG，GoogLeNet，ResNet，和 DenseNet[21]等。在深度神经网络处理深度学习任务时，通常使用原始输入上训练和运行模型，而无需先手动提取任何特征。这样做的原因是，对原始输入进行过培训的网络可以学会自己提取这些功能，但与使用预设好的特征相比，它还能够在网络改进时进一步优化特征提取[22]。

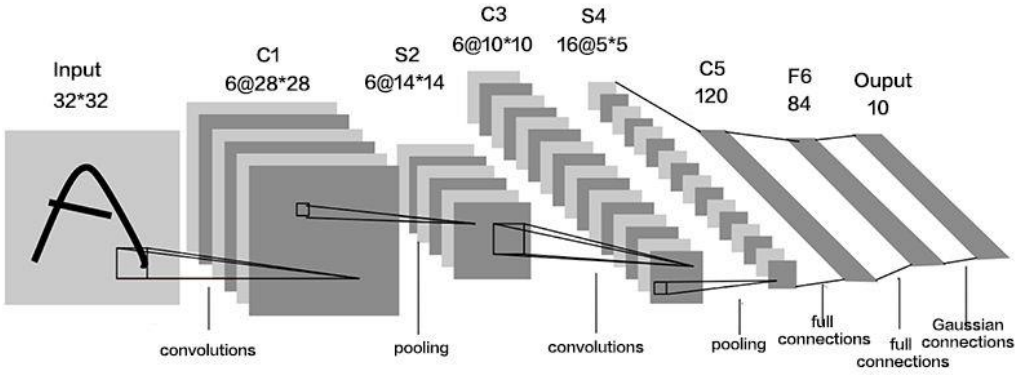

图 2 卷积神经网络的结构图

一个典型的深度卷积神经网络如图 2 卷积神经网络的结构图所示，其内部结构表现为多层结构，包含：卷积层、池化层、全连接层、输出层。卷积层和池化层的作用是降低卷积神经网络的复杂性，提取图像识别对象的特征，同时去除原始图像中的噪声。全连接层、输出层则进行分类。

4. 本课题的实行方案、进度及预期效果

实行方案：

1) 深度图像数据集整理

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。从综述文献[20]中引用的 14 个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 1 深度数据集汇总所示。这些数据集中的绝大多数均采用微软的 Kinect 相机作为采集工具，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 1 深度数据集汇总

数 据 集	类别数	样本数	特 性
CMU MoCap	144 个大类	2605	目前规模最大的 MoCap 数据集；没有按照特点的脚本执行动作，而实随机采样执行的动作，类内、类间差异巨大；提供低分辨率的 RGB 视频和 3 种格式的关节点数据：tvd，c3d，amc
HDM05 Mocap	130	2601	5 个演员，每类动作每个演员执行若干次；类内差异大，每个动作执行实例有 10~50 次
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
UCF Kinect	16	1280	16 个演员，每类动作每个演员执行 5 次；提供 15 个关节点的 3D 坐标及方向数据
MSRC-12 KinectGesture	12	594	30 个演员，每类动作每个演员执行 3 次；提供 20 个关节点的 3D 坐标及方向数据
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
MSR Gesture3D	12	336	10 个演员，每类动作每个演员执行 2~3 次；提供深度图像与 RGB 图像，样本存在相当普遍的自遮挡
ACT4 Dataset	12	6844	24 个演员，由 4 个相机从不同视角采集日生活活动视频；提供深度图像与 RGB 图像
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

2) 多种方法的评价

对不同的动作识别方法分别采用跨目标测试和交叉验证的方法，得出其对应的识别率。跨目标测试的思想是：训练样本与测试样本分别来自不同演员执行的动作序列。即使是同类型的动作，由于个体在执行时的差异性，往往使得采集的样本具有较大的类内方差。该类验证机制可以有效评估算法的泛化性能和鲁棒性。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，以此来作为评价分类器的性能指标。

同时，在一些较特殊的数据集上测试其对不同背景、自遮挡、噪声的敏感程度，如：RGBD-HuDaAct、MSR DailyActivities3D 等。

3) 总结

总结当前技术存在的问题，并提出在未来的研究工作中，一方面要从深度和骨架数据中设计更具判别力和紧致的特征来描述人体动作，另一方面是拓展当前已有的方法来应对更加复杂的人体动作，如交互活动和群体活动等。具体来说，将涉及到在交互动作与群体活动识别、多视角与跨视角动作识别、低延时动作识别等问题。

进度安排：

1) 2019.2-2019.3 机器学习与深度图像处理等相关文献的研读
2) 2019.3-2019.4 搜集不同数据集，研究动作识别技术的评价体系
3) 2019.4-2019.5 对多个动作识别方法实现并测试
4) 2019.5- 毕业论文的撰写与答辩

预期效果：

实现对动作识别方法的评价系统，并通过此评价系统对测试结果进行可视化，以便对不同方法的优缺点及适用环境进行评价和总结。

5. 已查阅参考文献

[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] 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.

[10] 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.

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

[12] 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.

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

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] 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.

[21] Jing L , Tian Y . Self-supervised Visual Feature Learning with Deep Neural Networks: A Survey[J]. 2019.

[22] Liu W , Wang Z , Liu X , et al. A survey of deep neural network architectures and their applications[J]. Neurocomputing, 2017, 234:11-26.

指导教师意见

该课题具有重要的理论研究价值和实际应用价值。课题调研方向清晰合理，能够概括总结典型的解决思路以及代表性方法，参考文献比较全面，可多增加2018、2019 最新研究工作。文字组织上仍需要加强逻辑性，做到贯通流畅，同时应该注重实验分析与验证，完善论文。 指导教师：宋凤义 2019 年 2 月 27 日
院（系）审查意见
同意开题。 学院领导（公章）： 2019 年 3 月 1 日