

摘要

智能家居场景中的人类活动识别方法研究

近年来,随着老龄化趋势愈发明显,老年人口逐渐增多。此外,由于子女外出务工以及人力成本的增高,老年人将面临无人照顾的困境。为使老年人在家独立和健康地生活,智能家居是一种目前被广泛探索的家庭服务解决方案。智能家居被广泛应用于健康保健监测、日常任务协助、能源管理以及安全管控。为了提供这些服务,智能家居系统必须能够理解和识别用户的日常人类活动。

智能家居场景中的日常人类活动识别对生活辅助至关重要。尽管活动识别技术已经取得了显著的进步,但是仍然存在着许多挑战:

首先,在对日常活动识别的时候,由于大部分数据集是不平衡的,以及部分活动存在类间差异小类内差异大的问题,这些问题导致整体活动识别的准确率高,但细化到单一活动识别性能表现糟糕,部分活动误分类严重。

其次,除了使用环境传感器以外,目前越来越多的工作开始引入物体传感器以及可穿戴传感器来进行活动识别,这使得多模态数据的使用变得更加普遍。然而,需要注意的是,只有某些模态的信号能够为活动识别提供有价值的信息,而冗余或不相关的信息往往会降低模型的识别能力。

本文旨在研究智能家居场景中的人类活动识别,具体研究工作包括以下几个方面:

(1) 本文设计了一个基于活动级上下文的全局信息融合活动识别方法 GIF-ALC,通过分析传感器事件流来识别人类的日常活动。该方法堆叠了三个全局信息融合模块,每一个模块采用带残差连接的自注意力机制进行全局信息抽取。同时, GIF-ALC 引入了活动级别上下文,以此增强模型对于相似活动的区分度。

通过在 ARUBA 和 MILAN 数据集上进行实验验证, GIF-ALC 能够在一定程度上提高活动识别的准确性,并减少部分活动误分类的情况,从而缓解类间差异小,类内差异大的问题。特别是在 MILAN 数据集上, GIF-ALC 的表现要优于其他方法。具体而言,在 MILAN 数据集上, GIF-ALC 准确率比 Casc-LSTM 提高了 2.95%, 平衡准确率提高了 9.33%, 加权 F1 值提高 3.42%。

(2)本文设计了一个基于多模态传感器数据融合的活动识别方法 MultiGIF。该方法首先利用挤压与激励操作将每个模态的信号进行加权，再利用 GIF 提取各模态的特征，其中 GIF 使用自注意力机制，以捕获不同时间戳对活动识别的贡献。当所有模态的特征提取完毕后，特征融合部分采用挤压与激励操作对不同模态进行加权，然后输入到分类器进行最终的决策分类。

通过在 OPPORTUNITY 数据集的运动模式以及手势这两个任务的识别上对所提方法进行实验验证。尤其在手势任务上，MultiGIF 取得了最佳的效果，在无 null 类的情况下，比第二名的 DeepConvLSTM 平衡准确率高出了 9.01%，加权 F1 值高出了 8.52%。在有 null 类的情况下，平衡准确率比第二名的 MLP 高出了 13.64%，加权 F1 值比第二名的 ResNet 高出了 4.96%。

此外，本文探究了不同模态组合对模型性能的影响。实验发现，在仅使用单一模态下的情况下，可穿戴传感器的模态要优于环境传感器、物体传感器这两种模态。在多种模态组合的情况下，环境传感器与物体传感器组合的模态要低于有可穿戴传感器存在的组合模态。通过上述分析，我们可以发现，单独采用环境传感器或物体传感器虽然能够有效地保护隐私，但却不能够捕捉到人类肢体运动所产生的信号。引入可穿戴传感器模态可以在智能家居场景中提供更为有力的日常活动识别支持。另外，在手势任务中，MultiGIF 在对组合模态的识别中，能获得比仅使用单一模态更高的识别性能，它相比于别的方法，更能从多模态中获取有益的增益信息，削弱不相关以及冗余信息。

关键词：

智能家居，人类活动识别，深度学习，数据不平衡，多模态融合

Abstract

Research on Human Activity Recognition Methods in Smart Homes

In recent years, as the aging trend becomes more and more obvious, the elderly population is gradually increasing. In addition, due to children going out to work and the increase in labor cost, the elderly will face the plight of no one to take care of them. To enable the elderly to live independently and healthily at home, smart home is a home service solution that is now being widely explored. Smart homes are widely used for health care monitoring, daily task assistance, energy management, and security control. In order to provide these services, smart home systems must be able to understand and recognize the daily human activities of the user.

Daily human activity recognition in smart home scenarios is critical to assisted living. Although significant advances have been made in activity recognition technology, many challenges remain.

First, when identifying daily activities, as most of the dataset is unbalanced, as well as some activities have the problem of small inter-class differences and large intra-class differences, these problems lead to high accuracy in overall activity identification, but poor performance in refining to single activity identification, and serious misclassification of some activities.

Second, in addition to the use of environmental sensors, more and more work is now introducing object sensors as well as wearable sensors for activity recognition, which makes the use of multimodal data more common. However, it is important to note that only signals from certain modalities can provide valuable information for activity recognition, while redundant or irrelevant information tends to reduce the recognition capability of the model.

The purpose of this paper is to study human activity recognition in smart home scenarios, and the specific research work includes the following aspects.

(1) In this paper, we design a global information fusion activity recognition method GIF-ALC based on activity-level contexts to recognize human daily activities by analyzing sensor event streams. The method stacks three global information fusion modules, each of which uses a self-attention mechanism with residual connections for global information extraction. At the same time, GIF-ALC introduces activity-level contexts as a way to enhance the model's discrimination of similar activities.

Through experimental validation on the ARUBA and MILAN datasets, GIF-ALC can improve the accuracy of activity identification to some extent and reduce the cases of partial activity misclassification, thus alleviating the problem of small inter-class differences and large intra-class differences. In particular, GIF-ALC outperforms other methods on the MILAN dataset. Specifically, on the MILAN dataset, GIF-ALC improved the accuracy by 2.95%, the balance accuracy by 9.33%, and the weighted F1 value by 3.42% over Casc-LSTM.

(2) In this paper, we design MultiGIF, an activity recognition method based on multimodal sensor data fusion. the method first uses squeeze and excitation operations to weight the signals of each modality, and then extracts the features of each modality using GIF, where GIF uses a self-attention mechanism to capture the contribution of different timestamps to activity recognition. When the features of all modalities are extracted, the feature fusion part uses the squeeze and excitation operation to weight the different modalities, which are then fed to the classifier for final decision classification.

MultiGIF is able to enhance the signals in the modalities that are relevant for activity recognition and weaken the irrelevant ones by using the squeeze and excitation operation. At the same time, the same squeeze and excitation operations are used in the modal fusion to enhance the relevant modal features and weaken the irrelevant ones.

The proposed method is experimentally validated by recognizing the motion patterns on the OPPORTUNITY dataset as well as on the two tasks of hand gestures. In particular, MultiGIF achieves the best results on the gesture task, with 9.01% higher

balance accuracy and 8.52% higher weighted F1 value than the second-place DeepConvLSTM without null classes. In the presence of null classes, the balance accuracy was 13.64% higher than the second-place MLP, and the weighted F1 value was 4.96% higher than that of the second-place ResNet.

In addition, this paper explores the effect of different modal combinations on model performance. It is found that the modality of the wearable sensor is better than the two modalities of the environmental sensor and the object sensor in the case where only a single modality is used. In the case of multiple modal combinations, the modal of the combination of environmental sensor and object sensor is lower than the combined modal with the presence of the wearable sensor. From the above analysis, we can find that the use of environmental or object sensors alone, although effective in protecting privacy, is not able to capture the signals generated by human limb movements. The introduction of wearable sensor modalities can provide more powerful daily activity recognition support in smart home scenarios. Moreover, in gesture tasks, MultiGIF can achieve higher recognition performance than using only a single modality in the recognition of combined modalities, and it is better than other methods to obtain beneficial gain information from multiple modalities and weaken irrelevant and redundant information.

Key words:

smart homes, human activity recognition, deep learning, data imbalance, multimodal fusion

目 录

第 1 章 绪论	1
1.1 研究的背景与意义	1
1.2 研究现状	4
1.2.1 基于视觉的活动识别	4
1.2.2 基于传感器的活动识别	5
1.2.3 多模态的活动识别	6
1.3 研究内容	7
1.4 论文组织结构	9
第 2 章 活动识别相关原理及关键技术	10
2.1 引言	10
2.2 相关活动识别方法	10
2.2.1 多层感知机	10
2.2.2 全卷积网络	12
2.2.3 残差网络	13
2.2.4 Casc-LSTM	14
2.2.5 DeepCovLSTM	14
2.3 评估指标	14
2.4 本章小结	15
第 3 章 基于活动级上下文的全局信息融合活动识别方法	16
3.1 引言	16
3.2 GIF-ALC 模型	19

3.2.1	活动级上下文	20
3.2.2	词嵌入	21
3.2.3	全局信息融合	21
3.2.4	决策分类	24
3.3	实验分析	24
3.3.1	数据集描述	24
3.3.2	数据集预处理	25
3.3.3	实验环境与参数设置	25
3.3.4	数值结果	26
3.4	本章小结	33
第4章	基于多模态传感器数据融合的活动识别	34
4.1	引言	34
4.2	MultiGIF 模型	35
4.2.1	模态信号加权	36
4.2.2	模态特征抽取	37
4.2.3	多模态特征加权融合	37
4.2.4	决策分类	38
4.3	实验分析	38
4.3.1	数据集描述	39
4.3.2	实验环境与参数设置	40
4.3.3	数值结果	40
4.4	本章小结	45

第 5 章 结论	47
5.1 工作总结	47
5.2 未来展望	48
参考文献.....	50
作者简介以及在学期间取得的科研成果	56
致谢.....	57

第 1 章 绪论

1.1 研究的背景与意义

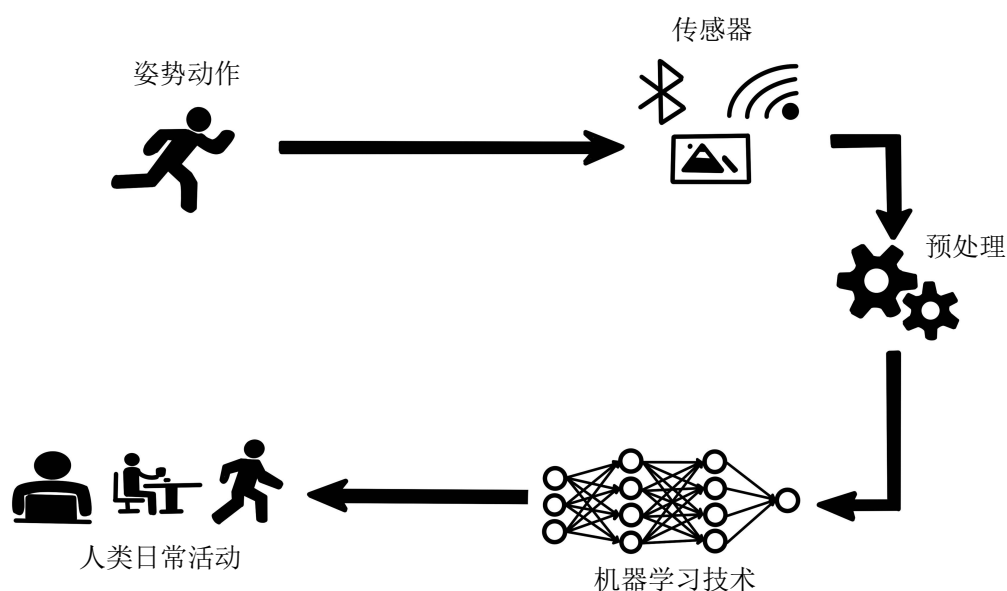
人口老龄化是年龄结构随社会经济发展而出现的一种必然趋势^[1]，我国人口老龄化趋势蔓延迅猛，1997 年我国 65 岁以上人口占比超过 7%，正式步入老龄化社会。据统计 2009 年我国人口老龄化系数已经达到 9.72%，部分省份该系数已达到 11% 左右。2020 年第七次人口普查数据显示，我国 60 岁及以上老年人口占比已达到 18.73%，80 岁及以上高龄老年人口占总人口比重达到 2.54%^[2]。人口老龄化不仅严重制约我国经济的发展，还带来了一系列的社会问题。随着人力成本的增高，越来越多的老年人将不得不面临失去独自生活的能力以及无人照顾的困境。因此，如何辅助老人更好地生活、减少护理成本成为当今社会面临的巨大挑战之一。为使人们能够在家独立和健康地生活，智能家居是一种目前正在探索的用于提供家庭服务的解决方案。

随着无线传感设备的进步，大量的低成本、低功耗的传感器被部署到环境中，智能家居系统的研究重点也从底层数据收集转移到高层信息集成、活动识别等问题上。智能家居是配备了许多传感器和执行器的房子，可以检测门的开度、房间的亮度、温度和湿度等，另外也可以控制日常生活中的一些设备，如暖气，灯，百叶窗等^[3]。智能家居目前被广泛地应用于健康保健监测^[4,5]、日常任务协助、能源管理以及安全管控。为了提供上述这些服务，智能家居必须理解和识别用户的行为活动，用户的日常行为反映着居民的认知情况和身体状况，是了解居民、照顾居民和为他们提供合适服务的重要依据^[6]。目前，在智能家居场景中的人类活动识别(Human Activity Recognition, HAR)已经成为一个重要的研究领域，此外，越来越多的智能家居系统的研究项目，如 CASAS、SWEETHOME、OPPORTUNITY 等也将人类活动识别作为其研究的关键问题。

智能家居系统的人类活动识别是通过使用传感器和机器学习技术来识别人类在家庭环境中的各种活动，如行走、睡觉、看电视、使用厨房等。HAR 需要采用多种传感器以收集人类活动的各种信息，同时需要使用机器学习算法对这些信息进行分析和分类。然而，HAR 仍然存在一些挑战，如准确性、鲁棒性、隐私保

护等问题，这些问题需要得到充分的关注和解决。因此，研究和开发智能家居的人类活动识别技术是智能家居领域的一个重要方向，它将为未来智能家居的发展提供更为可靠和高效的技术支持。

HAR 的典型流程如图 1.1 所示，最初，使用数据采集设备，如传感器和相机记录行动。从这些设备获得的数据大多是以原始形式获得的冗余信息，此外，这些数据可能不符合管道中其他步骤的要求，因此，需要进行预处理。预处理涉及不同类型的过滤、平滑等技术。一旦数据被预处理，机器学习技术就会被应用到上面，以便在下一步对不同的人类活动进行识别分类。



HAR 方法主要分为基于视觉和传感器两大类，其中基于传感器的方法可以细分为基于可穿戴传感器（wearable sensor-based）、环境传感器（ambient sensor-based）和物体传感器（object sensor-based）三种模式，见图 1.2。活动识别背景下的环境传感器是一种被动的传感器，它被整合到环境本身。由于物体可以被看作是环境的一部分，一些研究者认为物体传感器属于这个类别。当然，也有许多研究者倾向于对传感器类型进行更细化的分类，因此为物体传感器单独划分了一个类别。

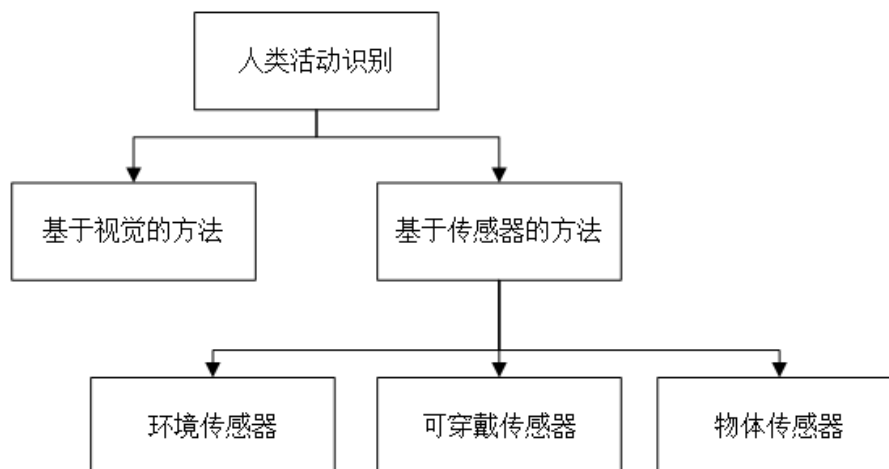


图 1.2 HAR 方法

基于视觉的 HAR 利用摄像头跟踪人类行为和环境变化。该方法使用计算机视觉技术，如标记提取、结构模型、运动分割、动作提取和运动跟踪。一些基于视觉活动识别的综述论文已经被发表^[7,8]，研究人员使用各种各样的相机，从简单的 RGB 相机到更复杂的相机系统。然而这些系统往往会涉及到用户隐私暴露的问题，其次视频设备在获取信息时容易受到非目标检测体的干扰，对图像处理后期识别带来不便。一般的基于计算机视觉的 HAR 方法流程包括数据采集、数据预处理、人体检测、姿态估计、特征提取、活动识别等步骤。数据采集阶段使用摄像头采集图像或者视频数据，可以是单个摄像头或多个摄像头也可以是深度摄像头。数据预处理阶段，对采集的图像或视频数据进行预处理，如图像去噪、图像增强等，以提高后续分析的准确性。人体检测阶段使用人体检测技术对视频中的人体进行检测和定位。姿态估计阶段使用姿态估计技术如 OpenPose^[9]对检测到的人体进行姿态估计，得到人体的骨骼关键点信息。特征提取阶段可以从人体关键点信息中提取特征，如身体朝向、动作幅度、运动速度等。活动识别阶段使用机器学习算法，如决策树、支持向量机等，对提取到的特征进行分类和识别，以确定人类正在进行的活动。

基于传感器的 HAR 通过传感器网络和被连接的设备来跟踪居民的活动。传感器被放置在人，物体或环境中，广泛被使用的传感器包括门磁传感器、光照传感器、红外运动传感器、无线电频率识别器标签（Radio Frequency Identification, RFID）、温湿度传感器、压力传感器、加速度计（accelerometer）、陀螺仪（gyroscope）、磁力计（magnetometer）等。一般的基于传感器的 HAR 方法流程包括传感器数据采

集、数据预处理、特征提取、特征选择和数据分类。将采集到的传感器信号进行滤波去除高频或低频噪声，再对采集到的数据提取特征，如平均值、标准差等时域特征，功率谱密度、能量谱密度等频域特征，然后使用分类器对特征进行分类。相比于基于视觉的方法，基于传感器的人类活动识别方法对居民的隐私保护更安全，更易被居民接受。另外相比于使用单一模态数据，有研究者开始在智能家居系统中引入多模态数据，并提出多模态融合的活动识别框架进一步提高活动识别的精度^[10-12]。

1.2 研究现状

监测用户行为和环境的變化是人类活动识别的重要前提，这个监测过程能为活动识别系统提供能够推断用户行为的上下文信息。本节主要介绍基于视觉的活动识别、基于传感器的活动识别以及多模态的活动识别方法。

1.2.1 基于视觉的活动识别

基于视觉的 HAR 的主要目的是调查和解释来自视频或直接来自图像的人类活动。用合适的相机来捕捉活动对识别系统的整体功能有很大影响。基于视觉的 HAR 是一个具有挑战性的问题，因为人类肢体的变化、遮挡、杂乱的背景、固定或移动的相机、不同的照明条件、光照强度和视点变化等都会影响到 HAR 的识别精度^[13]。

目前，已经有众多学者对基于视觉的 HAR 方法展开研究。Jalal 等人利用深度视频传感器技术，将收集到的数据利用隐马尔科夫模型进行活动识别，首先，训练系统包括使用深度相机进行数据采集，通过隐马尔可夫模型进行特征提取和活动训练。其次，训练结束后，识别引擎开始识别所学的活动并产生生活日志。该系统使用生活日志特征对主成分和独立成分特征进行了评估，并在与传统方法的比较中取得了令人满意的识别率^[14]。Kushwaha 等人提出了一种有效的视图不变框架，用于从输入视频序列中识别人类活动。所提出的框架由三个连续的模块组成：首先通过背景减法检测和定位人的位置；其次为不同的活动创建视图不变的时空模板；最后，为视图不变的活动识别执行模板匹配^[15]。Bouchard 等人以及 TRAN 等人将多个摄像头得到的图像合成三维图像，从而进行活动识别研究^[16, 17]。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/207105113111006053>