文章编号:1671-4598(2025)01-0009-11 DOI:10.16526/j.cnki.11-4762/tp.2025.01.002 中图分类号:TP391.41 文献标识码:A

基于视觉的动态手势识别技术综述

付智凯，李文新，罗新奎
（兰州空间技术物理研究所，兰州 730000）

摘要：动态手势识别是计算机视觉领域较为热门的任务之一，受到了研究者的广泛关注；动态手势识别技术在自动驾驶、虚拟现实和人机交互等诸多领域展现出很高的应用潜力；手势是在虚拟空间中与其他人交换信息、指导机器人在恶劣环境中执行特定任务或与计算机交互的一种直观而理想的方式；调研归纳了一些常用的动态手势数据集，对动态手势数据集的模态、数据量、应用场景进行了总结与分析；从使用方法的网络类别出发，综述了基于视觉的动态手势识别技术研究进展，重点介绍归纳了基于深度学习的方法，对基于卷积神经网络、循环神经网络以及图神经网络的方法进行了整理总结与性能比较；最后对基于视觉的动态手势识别的研究方向进行了展望。

关键词：计算机视觉；人机交互；动态手势识别；深度学习网络；手势数据集

Review of Vision-Based Dynamic Gesture Recognition Techniques

FU Zhikai，LI Wenxin，LUO Xinkui
（Lanzhou Institute of Physics，Lanzhou 730000，China）

Abstract：Dynamic gesture recognition is one of the popular tasks in the field of computer vision，which is widely concerned by researchers. Dynamic gesture recognition technology has high potential features in many fields such as automatic driving，virtual reality and human-computer interaction. Gestures are an intuitive and ideal way to exchange information with others in a virtual space，and direct robots to perform specific tasks in hostile environments，or to interact with a computer；Some commonly used dynamic gesture data sets are investigated and generalized，and the modes，data and application scenarios of dynamic gesture data sets are summarized and analyzed. From the network category of usage methods，this paper summarizes the research progress of vision-based dynamic gesture recognition technology，focuses on introducing and concluding the methods based on deep learning，and concludes and compares the methods based on convolutional neural network，recurrent neural network and graph neural network. Finally，the research direction of dynamic gesture recognition based on vision is prospected.

Keywords：computer vision；human-computer interaction；dynamic gesture recognition；deep learning network；gesture datasets

0 引言

手势作为人们在日常生活中除语言外最常用的交流方式，具有直观、自然、蕴含信息量大的特点，因此，手势也自然而然地成为了人机交互的一种重要方式。

手势识别技术根据手势模态大体上可以分类为静态手势识别与动态手势识别，动态手势在复杂程度和包含的信息量方面要远高于静态手势，并且，随着识别方法的不断革新与硬件算力的不断进步，动态手势识别已经成为了近些年手势识别技术领域的主流研究方向。

根据手部外观信息与运动信息的获取方式，可以将动态手势识别技术分类为基于机器视觉的方法与基于可穿戴设备的方法，二者各有优势，基于机器视觉的方法相较于基于可穿戴设备的方法更加便利，使用者不必被臃肿的数据采集设备束缚手脚，在该技术的民用市场如智慧家居与体感游戏等领域，应用前景更为广阔；而可穿戴设备能够进行更加精确的数据采集并且受外部环境干扰极小，在工业控制领域与复杂环境下精准控制领域更受欢迎。

本文就基于视觉的动态手势识别技术进行综述，对该

收稿日期:2023-11-28； **修回日期**:2024-01-03。
基金项目:中国载人航天工程重大专项(RWZY640601)。
作者简介:付智凯(1998-),硕士研究生,CCF 学生会员。
通讯作者:李文新(1966-),博士,研究员,博士生导师。
引用格式:付智凯,李文新,罗新奎.基于视觉的动态手势识别技术综述[J].计算机测量与控制,2025,33(1):9-19.

领域的技术发展进程进行介绍与总结，具体而言，本文对动态手势的视觉采集设备、动态手势数据集、动态手势识别方法进行了深入的介绍与总结，其中重点对基于视觉的动态手势识别方法进行了介绍与总结，如图 1 所示。

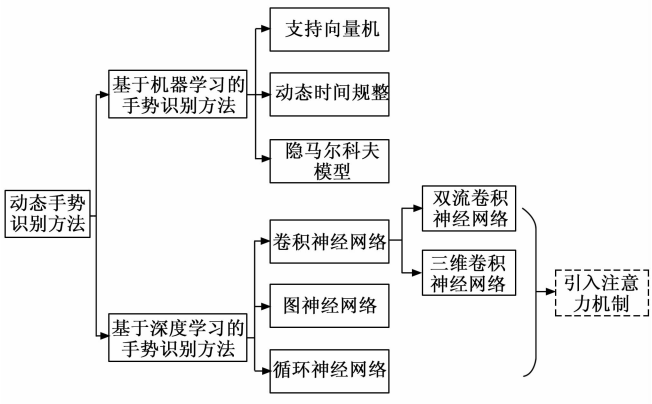


图 1 动态手势识别方法分类图

1 动态手势数据集

随着深度学习在手势识别领域的不断演进，自 2012 年开始，用于对手势识别进行训练、验证的数据集的数量开始爆发式地增长，其中以 RGB 模态与骨架模态为主的数据集增长尤为迅速，本节中，将重点讨论基于 RGB、骨架、深度 3 个模态的数据集发展情况，基于红外数据的数据集应用较少，在此不做讨论。

本文收集整理了 2012 年以后公开发布的 24 个动态手势数据集，包括 NVgesture[1]，HaGRID[2]，MSRC-12[3]，ChaLearn Lap ConGD[4]，Jester[5]，DHG14/28[6]，EgoGesture[7]，Kinetics-600[8]，FHPA[9]，ASL[10]，Montalbano[11]，SKIG[12]，GRIT[13]，UCF101[14]，AMI[15]，SHREC-2017[16]，KSU-SSL[17]，NTU-RGBD[18]，LeapGesture DB[19]，DVS-128[20]，UT-Kinect[21]，LD-ConGR[22]，CSL[23]，HandNET[24]。

针对上述 24 个数据集，从模态、数据量、公布时间、使用场景进行归纳总结，见表 1。

由表 1 中显示的数据可以看出，自 2012 年起除 2014 与 2021 年外，每年都至少有一个动态手势识别领域公共数据集被发布，而仅 2016 年一年就发布了 6 个数据集，这也从侧面印证了 2016 年是深度学习应用于图像处理领域的高潮时期。数据集模态方面，如图 2 所示，基于 RGB 模态的数据集占到了总数的 54.3%，包含深度模态的数据集占到了总数的 34.3%，包含骨架模态的数据集占到了总数的 11.4%，大多数数据集是由输出 RGB 数据的设备收集而来，这也说明了当前输出 RGB 格式数据的机器视觉设备仍是主流，而深度数据的应用也在不断地加强。从图 2（c）中可以得出结论，由于硬件设备算力的限制，单一模态的数据集仍然是动态手势识别算法训练、测试的主流，而随着硬件设备算力的提升，混合模态的数据集也开始被广泛应用。从数据规模来看，大多数数据集的数据量都位于1 000 到 10 000 区间内，原因是该规模的数据集已经可以满足大部分方法对训练、测试准确度与拟合度的要求。

表 1 2012 年以来发布的动态手势数据集

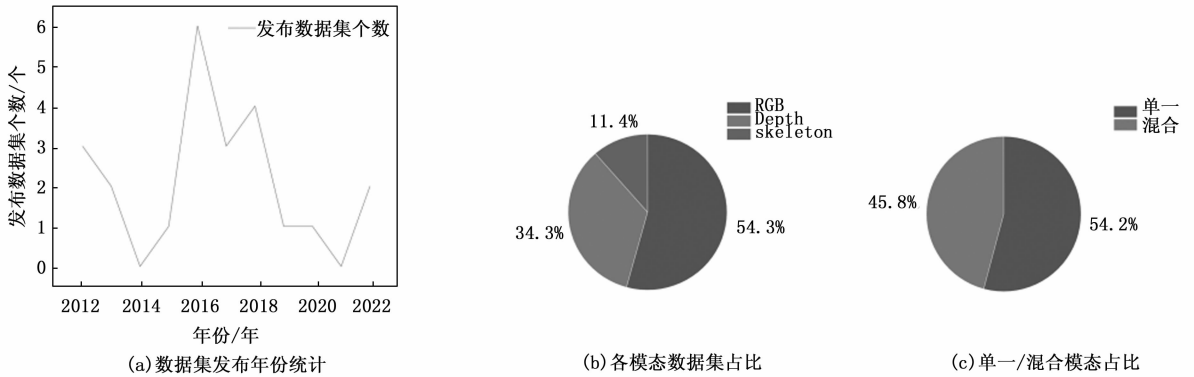
序号	数据集	年份	模态	数据量	应用场景
1	NVgesture	2016	RGB	1532	智能驾驶领域
2	HaGRID	2022	RGB	552992	智慧家居,智能驾驶,智慧会议
3	MSRC-12	2012	RGB	6244	日常动作识别
4	SKIG	2013	RGBD	2160	变化光照背景下的手势采集
5	CharlearnLAP	2016	RGBD	47933	日常动作识别
6	Jester	2019	RGB	148092	日常动作识别
7	HandNet[87]	2015	RGBD	24071	日常动作识别
8	FPHA[88]	2018	RGBD	1175	生活常用手势动作识别
9	CSL[89]	2015	骨架	25000	中国手语识别
10	LD-ConGR	2022	RGBD	44887	长距离连续手势识别
11	DHG14/28	2016	深度	2800	日常动作识别
12	EgoGesture	2017	RGBD	24161	与可穿戴设备进行交互
13	Kinetics600	2018	RGB	480000	日常动作识别
14	ASL	2018	骨架	1200	美国手语识别
15	Montalbano	2013	RGBD	956	意大利文化语境下的手势识别
16	GRIT	2016	RGB	543	人机交互指令
17	UCF101	2012	RGB	13320	日常动作识别
18	AMI	2018	RGB	120000	可穿戴设备,AR 眼镜
19	SHREC2017	2017	深度＋骨架	2800	虚拟现实、增强现实
20	KSU-SSL	2020	RGB	8000	日常动作识别
21	NTU-RGBD	2016	RGBD	56880	日常动作识别
22	LeapGestureDB	2016	骨架	550	医疗领域
23	DVS-128	2017	RGBD	1342	变化光照背景下的手势采集
24	UT-Kinect	2012	RGBD＋骨架	200	室内光照条件下采集

2 动态手势识别方法

手势识别的目标是将从背景中分割出的手势进行特征提取，并按照提取出的特征进行分类，而动态手势识别则是按照预定的语义对一定长度帧序列的手势组合，结合时间与空间特征进行分类。近几年流行的动态手势识别方法按照特征的获取方式可以分为机器学习方法与深度学习方法[25]，而近年来，以端到端的深度学习的方法在动态手势识别领域应用进展飞速，深度学习强大的高级特征提取能力与强大的分类能力为动态目标识别任务提供了强大的发展动力。

2.1 传统的机器学习方法

经典的动态手势识别方法包括支持向量机（SVM，support vector machine）、动态时间规整（DTW，dynamic time warping）以及隐马尔可夫模型（HMM，



(a)数据集发布年份统计　　(b)各模态数据集占比　　(c)单一/混合模态占比

图 2　数据集相关信息统计

hidden Markov model）等。

2.1.1　基于支持向量机的方法

支持向量机是一种典型的二分类模型方法，其核心思想是在特征空间内寻找一条最优的“线”使得两类对象尽可能地分布在这条“线”的两侧[26]。在面对多类问题时，SVM 一般通过构建多层次特征进行二分类迭代将样本分为多类或者将样本映射到更高维度的空间，寻找超平面进行多分类[27]，如图 3 所示。一些研究者利用人手的肤色与纹理特征对视频帧内的人手进行连续检测，截取检测到的帧序列，在这些帧序列中利用方向梯度直方图提取帧内的手势颜色与纹理特征，利用光流法提取帧间手势区域像素点运动的矢量特征，集合作为一个样本，再结合最大平均概率集成方法与 SVM 的分类器进行分类[28]。

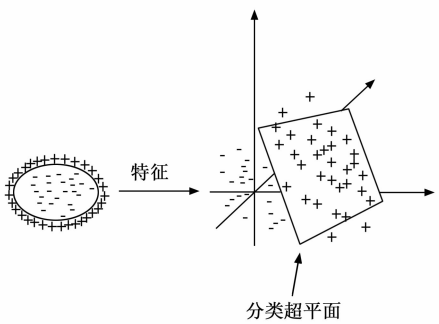


图 3　超平面决策域

此外，一些研究者将卷积神经网络（CNN，convolutional neural network）与 SVM 结合，使用 SVM 对 CNN 的训练样本进行迭代分类并标注，利用 SVM 能够平衡样本区间的特点，减少了训练误差[29]，该方法与单独使用 CNN 的识别方法在 LMDI 数据集上进行验证，其识别准确率高达 92.22%，单独使用 CNN 的准确率只有 80%。

文献［28］使用自建的小规模数据集上完成训练与测试，而文献［29］则在大规模数据集上进行训练测试，这也能表明单纯的 SVM 方法更适用于小样本分类场景，而近些年为了追求识别的准确率与良好的鲁棒性，大规模样本数据集上的训练与测试是必要的，因而 SVM 往往与 CNN 等深度学习算法结合使用，其效果往往要优于单独使用 SVM 或深度学习算法。

2.1.2　基于动态时间规整的方法

动态时间规整是一种时间相关序列相似度比对度量方法，在相似度比较、模式识别等多种问题中有广泛的应用[30-31]。DTW 方法具有时间上下文处理能力以及可扩展性，这些特点让其较为适合处理动态手势识别问题[32-33]。随着近些年基于大型数据集的机器学习算法兴起，对动态手势识别的精度与速度都提出了更高的标准，DTW 方法也在逐渐被改良以适应新的任务需求。大多数基于 DTW 的方法都采用欧式距离表征特征的相似程度，有研究者提出了一种用于手语识别的 Procrustes-DTW 方法，该方法用 Procrustes 距离（普氏距离）代替 DTW 方法中的欧氏距离，普氏距离对特征相似度的描述更加合理[34]，应用该方法在自建的数据集上进行测试，识别率达到了 97.02%，远高于使用欧式距离的 DTW 方法 64.37%的识别率。文献［35］对 DTW 方法添加了首帧约束、特征约束以及全局搜索路径约束，相较于传统 DTW 方法减少了匹配运算量，减少了识别时间，提高了手势序列的识别率。

2.1.3　基于隐马尔可夫模型的方法

隐马尔可夫模型是一种典型的关于时间可变序列的概率统计模型，适合用作动态目标的识别[36-37]，其典型流程图如图 4 所示。

在提取特征形成 HMM 观测序列时，往往需要耗费大量的资源，而且在手动提取一些复杂的特征时，会丢失许多有用的信息，为了解决这一问题，有研究者将卷积神经网络与 HMM 相结合，在手语和手势识别背景下，将 CNN 嵌入到 HMM 模型中，使用 CNN 提取输入模型的手势序列进行静态的帧内手势特征向量与动态的帧间手势特征向量，将这些提取的特征作为 HMM

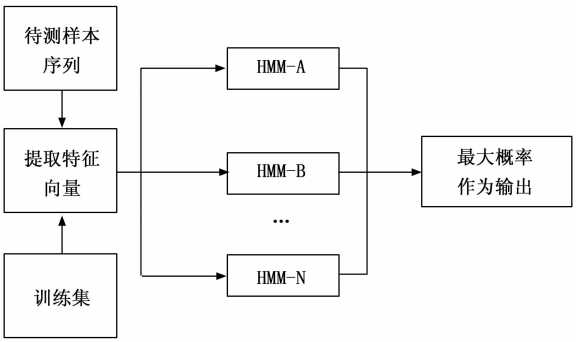


图 4 基于 HMM 方法的动态手势识别流程图

模型的观察序列，并使用一个额外的 CNN 模型用于降低特征维度[38]。该方法在 Montalbano 数据集上进行测试对孤立的手语词汇手势达到了 90.15%的识别率，该方法以 HMM 作为分类器，其识别率几乎与使用长短期记忆网络（LSTM，long short term memory）作为分类器的模型相当，在使用较少参数且不使用 GPU 加速的情况下，CNN-HMM 方法优势明显。文献[39]基于复杂动态手势的特点，将手势图像分解为手部形状变化、手部在二维平面上的位置变化和手部在 z 轴方向上的运动 3 个特征序列，进行特征提取。根据 3 个子序列分别建立 HMM 模型，并通过模糊推理连接模糊神经网络（FNN，fuzzy neural network）进行手势语义判断。该方法可以快速有效地识别动态手势，效果优于一般的 HMM 方法。

综上所述，基于机器学习的动态手势识别方法在小样本场景下，能够发挥其体量小，结构相对简单的特点，对动态手势进行快速有效的识别，然而这些方法普遍需要对模型训练的样本进行手工标注，在样本量较大的情况下，标注成本十分巨大，在一些较为复杂的场景，外界环境变化频繁，如遮挡、光照条件变化等，基于机器学习的方法在识别率与稳定性方面都表现一般。深度学习技术的不断发展为解决这些问题提供了新的方向。

2.2 基于深度学习的方法

随着深度学习技术的不断发展，在图像分类领域应用深度学习技术受到越来越多研究者的关注。卷积神经网络作为深度学习的代表性方法，在历经几十年的发展以后，其已经成为了图像特征提取与分类任务的首选方法，其出色的特征自动提取和分类能力，使得其能够在大规模图像数据集中取得优异的表现[40-41]。根据用于动态手势识别的神经网络类型，常用的深度学习架构分为卷积神经网络、图神经网络以及循环神经网络的方法。此外，基于注意力机制的深度学习方法也值得关注。

2.2.1 基于卷积神经网络的方法

卷积神经网络在图像处理领域的应用已经非常成熟，在基于深度学习方法的手势识别领域，CNN 占据着主导地位，而对于动态手势的识别而言，神经网络需要考虑丰富的帧间信息，传统的二维神经网络在这方面有所欠缺，近几年来能够将时间信息纳入网络结构的方法大体上有双流网络与三维卷积神经网络。

1）基于双流卷积神经网络的方法：

传统的卷积神经网络甚至二维的卷积神经网络大都忽略了动作序列中时间信息的应用，对于动态目标的行为识别不甚理想[42]，在此基础上，双流卷积神经网络被提出[43]，如图 5 所示，对视频分类任务有着良好的效果。

网络首先提取一帧 RGB 图像作为表述空间信息的载体，包括目标与背景环境，进入空间流卷积神经网络，再提取之后的数帧图像作为时序信息的载体进入时间流卷积神经网络，通过卷积神经网络对光流图像的学习，可以提取变化动作的特征，以实现动态手势的识别。然而，由于双流网络结构仅提取 RGB 图像后的数帧光流图像进入时间自网络，导致了视频长期信息的丢失。一些研究者对双流网络的结构进行了改进，以解决该问题，使双流网络达到更好的性能，一些研究者用

表 2 基于机器学习的动态手势识别方法对比

类别	文献	方法	数据集	识别准确度/%	特点
SVM	[28]	MMPE+SVM	自建，10 个类别，944 个样本	89.83	HOG 提取静态特征，光流法提取动态特征
	[29]	CNN+SVM	LMDI	92.22	使用 SVM 对 CNN 的训练样本进行迭代分类并标注
DTW	[34]	Procrustes-DTW	自建，145 个类别，559 个样本	97.02	使用普式距离代替欧氏距离描述特征相似度
	[35]	CDTW	MSRC-12	——	对 DTW 方法添加全局搜索路径限制、特征限制以及首帧限制
HMM	[38]	CNN+HMM	Montalbano	90.15	使用 CNN 提取手势序列的静态与动态特征，将之作为 HMM 的观察序列
	[39]	FNN+HMM	UESTC-ASL	99.5	为手形变化、手部位置变化和手部在 z 轴方向上的运动 3 个特征序列分别建立 HMM 模型，通过模糊推理连接 FNN 进行手势语义判断

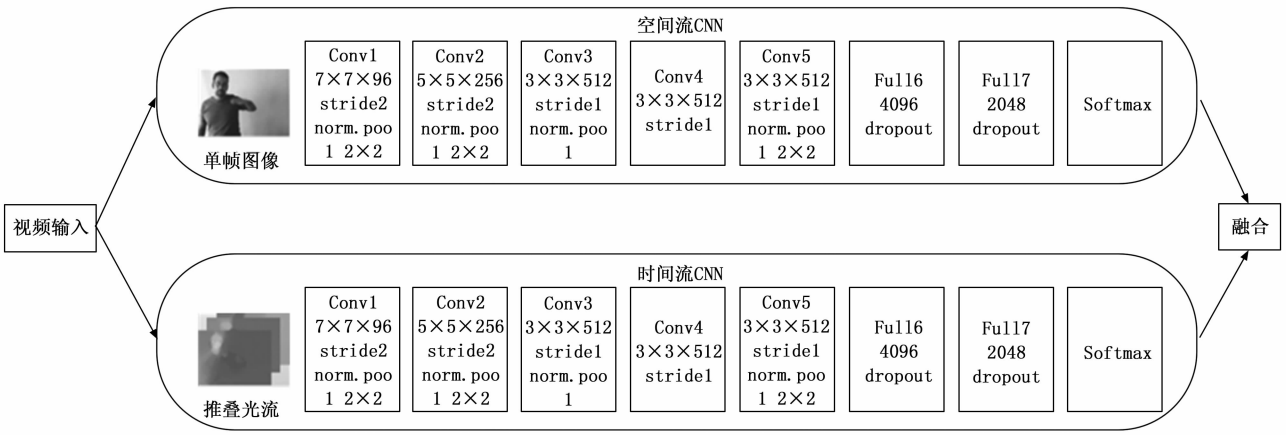


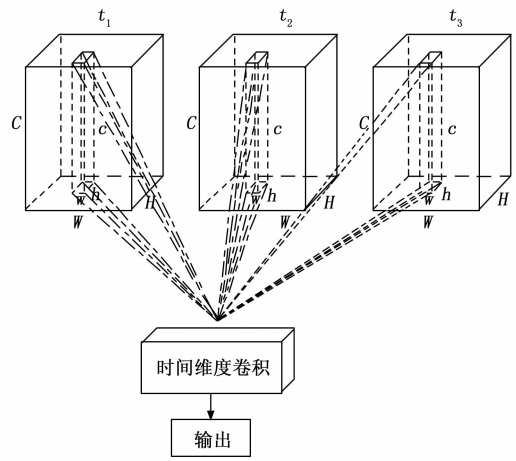
图 5　双流网络架构

Mobile Net v2 来替换提取时空流特征的 CNN 网络，并在融合双流网络的空间流特征与时间流特征之后，加入一个 LSTM 网络[44]，对长期手势信息的时空依赖关系进行学习，该方法在 20BN－Jester 数据集上进行训练测试，准确率达到了 91.25%，优于原始的双流网络，另外，由于使用 Mobile Net v2 作为时空子网络，减小了子网络的尺寸，该方法对动态手势的识别速度相较于原始双流网络也有了较大的提升。

2）基于三维卷积神经网络的方法：

三维卷积神经网络（C3D，3D convolutional neural network）可以从视频中提取手势的空间和时间信息，其卷积操作与池化操作同时在具有宽度、高度和时间维度的多个特征图上运行，可以同时提取手部的静态特征（颜色、纹理、深度以及骨架信息等）与动态特征。三维卷积神经网络相比于二维卷积神经网络，其卷积核上多了对时间维度的卷积，卷积核大小由［T，C，W，H］4 个参数决定，多帧图像堆叠形成的立方体与 3D 内核进行 3D 卷积，3D 卷积核如图 6 所示，很好地解决了二维卷积神经网络忽略时间信息与动态特征提取的问题。2015 年，C3D 在二维卷积神经网络的基础上被提出，该网络可以很好地利用动作帧序列中的时间信息[45]。此外，一些研究者将 C3D 应用于拉伯手语的手语识别中[17]，随后，众多研究者都在 C3D 的基础上进行了研究改进。

近年来发表的论文中新提出的基于 C3D 网络的方法，大多是对子网络进行改进与优化或者将一些在静态手势识别任务中表现良好的二维网络扩展到三维结构，如 3D Mobile Net、3D Squeeze Net、引入残差块的 ResC3D 网络，这些网络都是通过网络宽度或者深度的增加来提高对于高维特征的表达能力。2018 年，Mobile Net v2 网络在 CVPR 会议中被正式提出[46]，其特点是分类准确率高，模型体积小，随后一些研究者将该

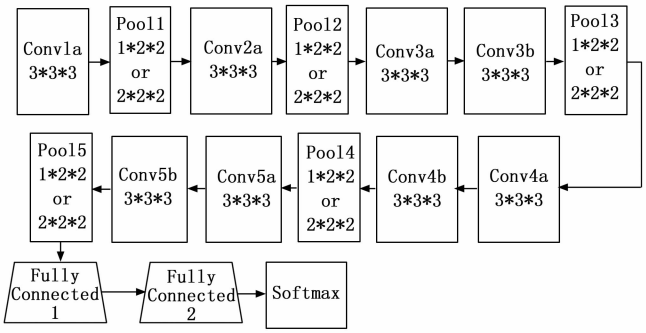


C 为卷积核通道数，W 是卷积核的宽度，H 是卷积核的高度，T 是时间长度。

图 6　3D 卷积核

网络扩展为 C3D 架构并且应用于动态手势识别，由于加深了 C3D 的网络深度，使得该方法在对高维复杂特征的表达上面优于基准的 C3D 方法[47]，然而，网络深度的不断增加会引起不可逆的时空信息损失导致网络退化，为此，残差神经网络被提出[48]，该网络将残差块引入 C3D，构成了 ResC3D 网络结构，残差块的引入，使得网络拥有了恒等映射的能力，至少能够保证加深后的子网络在特征提取于分类方面的能力与浅层网络持平。文献［49］将 ResC3D 网络用于手势识别，结合了 C3D 与残差网络的优点，在使用深度网络提取手势序列时空特征的同时，避免了网络退化，同时将 ResC3D 网络与卷积 LSTM 级联形成融合模型架构，利用一种选择性时空特征学习的动态选择机制，使得网络能够自适应地调整 ResC3D 和卷积 LSTM 对分类的贡献。除增加网络深度的方法外，增加网络的宽度也是提高对高维特征表达能力的有效方法，例如将 Squeeze Net 扩展为 C3D 网络[50]，在 Kopuklu 等人的研究中，该方法也取

得了不错的效果。三维卷积网络架构如图 7 所示。



由 8 个卷积层，5 个池化层，2 个全连接层，1 个 Softmax 层构成

图 7 三维卷积网络架构

综上所述，对于动态手势的特征提取与分类方面，无论是基于双流网络的方法还是基于 C3D 的方法，都较为重视与时间信息相关的动态特征的提取与利用。而对于这两类方法的改进，大多从子网络的结构与深度着手进行设计，如何加快训练速度、寻找识别的实时性与准确率之间的平衡点仍然是一个重要的问题。

2.2.2 基于图神经网络的方法

图神经网络（GNN，graph neural networks）是一种用于处理图结构数据的深度学习架构，手掌的骨架图是天然的拓扑图，非常契合图神经网络的结构[51]。近几年跟随深度学习演进的脚步，图神经网络也有了长足的发展。图神经网络发展至今，大多数方法结合骨架模态数据应用于人体行为识别，而后随着骨架模态数据的精细化发展，人体手部的骨架建模趋于完善，越来越多的研究者意识到图神经网络在手势识别方面应用潜力巨大。一些研究者首先将图神经网络应用于人体行为识别[52]，将人体关节作为时空图顶点，连通性骨骼作为边建立图神经网络提取空间特征，其工作极大地启发了后来的研究者，自此图神经网络开始在人体行为识别与手势识别领域得到广泛应用。一些研究者使用图卷积神经网络，如图 10 所示，对人手骨架数据的空间关联特征进行表征，并利用循环门控单元（GRU，gate recurrent unit）如图 8 所示对手势的时序特征进行提取，能够很好地解决对长期手势数据的依赖问题[53]。该文章的创新之处在于将循环门控单元加入了 GNN 结构中，采用门控机构控制输入与记忆信息，解决了在较长的手势序列识别过程中，当前手势的识别依赖开始时输入信息的问题。

此外，文献［54］提出了基于双空间网络和变换图编码器的动态手势识别方法，该方法的特点是利用图卷积神经网络对关键帧内手势的空间特征进行提取，利用一个图变形解码器提取帧间手势的运动变化特征，如图 9 所示，并且，该方法引入了注意力机制，与文献

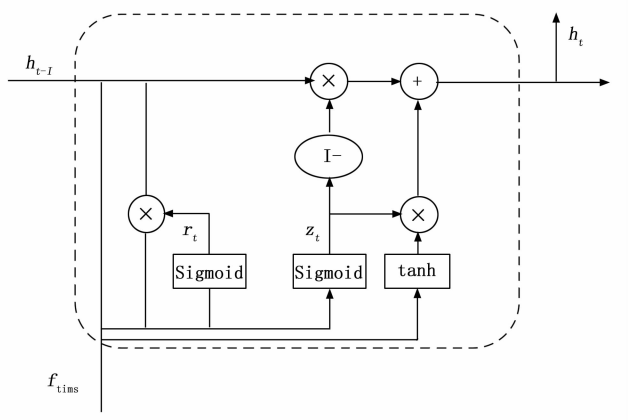


图 8 门控循环单元结构图

［53］提取时序特征的方法不同，文献［54］用到的解码器是基于多头自注意力机制的，通过使用多个独立的注意力头，分别计算注意力权重，并将它们的结果进行拼接或加权求和，从而获得更丰富的表示，该模块由多个相同的编码器层堆叠在一起组成，每个编码器层利用一个多头自注意模块和一个简单的全连接前馈网络将其输出馈送到下一个编码器层以提取时序特征。

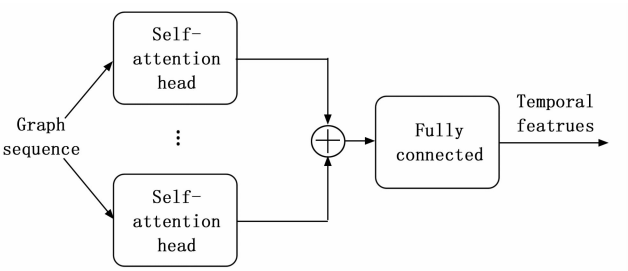


图 9 基于时间自注意模块的解码器

2.2.3 基于循环神经网络的方法

循环神经网络（RNN，recurrent neural network）也是动态手势识别常用的网络架构，RNN 最显著的特点在于其内部的环状结构，这些在连接隐藏层的环状结构使得信息能够在网络中循环，能够对序列中紧密相连数据的相关性进行处理，从而实现对序列信息的存储与处理，与动态手势的多帧序列特征相契合。基于此，有研究者提出了一种基于双流架构的 RNN 网络[55]，与 2.2.1 节所述的双流架构类似，该方法利用两个并行的 RNN 网络分别提取手势的空间特征与手势序列的时间特征。然而，由于较长的序列存在梯度消失的问题，使得 RNN 网络无法支持长时间序列，为此，一些研究者使用门控结构将长、短期记忆结合，改进了 RNN，构建了长短期记忆网络。如双层双向的长短期记忆网络结构（BiLSTM，bidirection-LSTM），使用两个独立、反向的 LSTM 网络，在每个时间点，将输入提供给两个 LSTM 网络，并根据隐藏状态组合它们的结果，该网络

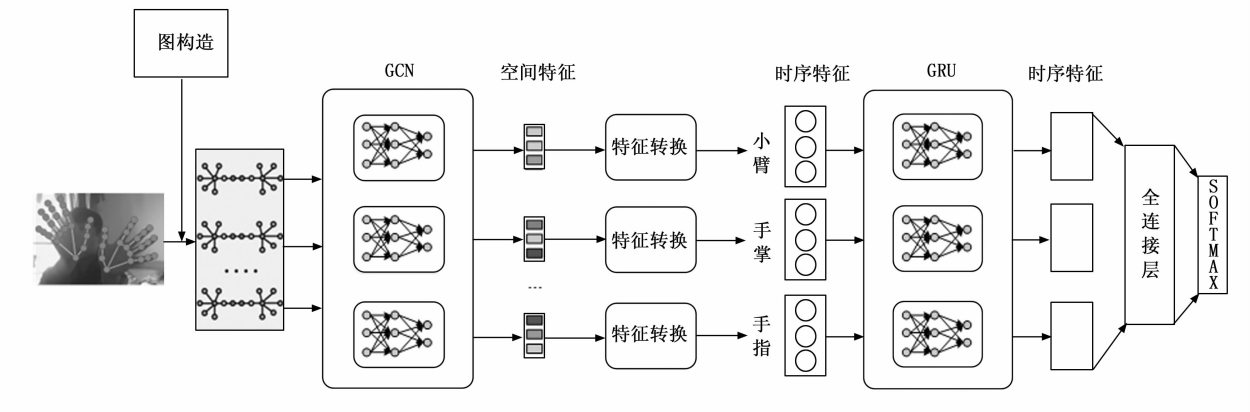


图 10　STGNN 框架

拥有处理同时依赖过去和未来手势信息的能力[56]。此外，另一些研究者也在各自文章中用到了双层双向的 BiLSTM 结构[57-58]。一些研究者还将注意力机制引入了 LSTM，对 LSTM 的门结构做出了改进，根据注意力机制重构了输入输出门与遗忘门，提高了 LSTM 单元处理手势序列数据的效率[59]。在近几年提出的基于 RNN 的动态手势识别方法中，绝大多数都使用了 LSTM 网络结构，由于其出色的融合长短期信息能力，为许多缺乏长期信息的手势识别模型提供了解决方案。最为常见的是 LSTM 作为后置网络，为 CNN 等前置特征提取网络提供长短期时间特征。如文献［60］构建的 CNN-LSTM 网络结构，该网络利用 CNN 提取手势的空间特征，之后将特征送入两层 LSTM 网络中，分别进行时间特征的提取与长期时间特征的融合。此外，结合 Slow-Fast 路径结构与 LSTM 网络也是一种较为新颖的思路，该网络基于双流架构，由两路提取不同类别特征的 ResC3D 网络组成，Fast 路径提取高分辨率的手势空间特征，Slow 路径用于捕获手势帧序列的时间特征，两路特征在之后的卷积 LSTM 网络中，并与长期手势信息融合[61]。

基本的 LSTM 网络单元如图 11 所示，图中，i_t、o_t、f_t 分别式 LSTM 单元的输入、输出、遗忘门，x_t 为 t 时刻输入向量，h_t 为 t 时刻隐藏层的状态向量，C_t 为 t 时刻整个单元的细胞状态，是指 LSTM 中的长期记忆，与遗忘门 f_t 共同作用，控制前一个单元的状态，若其一为 0 则遗忘状态，同为 1 则保持状态[56]。

以上介绍归纳了主流的基于深度学习的动态手势识别网络结构，近年来该领域发表论文中所提出的基于深度学习的方法大多都是对 CNN、RNN 以及 GNN 的组合或延伸，此外，一些研究者将注意力机制引入到了深度学习网络，文献［54］将注意力机制引入到了 GNN 中取得了良好的识别效果，注意力机制受到越来越多研究者的关注。

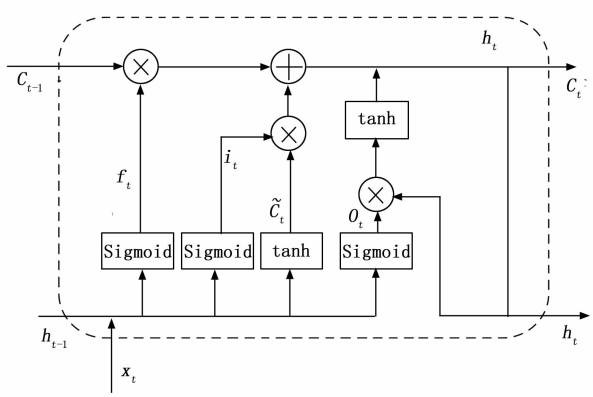


图 11　LSTM 网络单元

典型的注意力机制包括自注意力机制、时间注意力机制和空间注意力机制，注意力机制允许模型对输入序列的不同部分分配不等的权重，以便于在处理数据时专注于最相关的部分。

基于这一思想，引入了通道、空间混合注意力机制（CBAM，convolutional block attention module）的 C3D-LSTM 架构的动态手势识别方法被提出，该模型在 CNN 的最后一个池化层后、LSTM 网络前加入了 CBAM 模块，如图 12 所示，该模块根据前级网络的输入特征图生成通道与空间注意力特征图，两种通道特征图与原特征图相乘，将通道特征图产生的权重应用于原特征图上的每个位置，使模型专注于运动的手部图像[62]。

给定一个中间特征映射 $F \in \mathbf{R}^{C\times H\times W}$ 作为输入，CBAM 推导出一维通道注意力图 M_C 与二维通道注意力图 M_S，整个注意力过程可以概括为：

$$\begin{aligned} F' &= M_C(F) \otimes F, \\ F'' &= M_S(F') \otimes F' \end{aligned} \tag{1}$$

F'' 是最终输出的混合注意力图。通道注意力机制通过特征内部的关系来计算通道注意力值，将特征图的每个通道都作为特征检测器，并通过平均池化与最大池化为特征图降维，使用多层感知机（MLP，multi layer

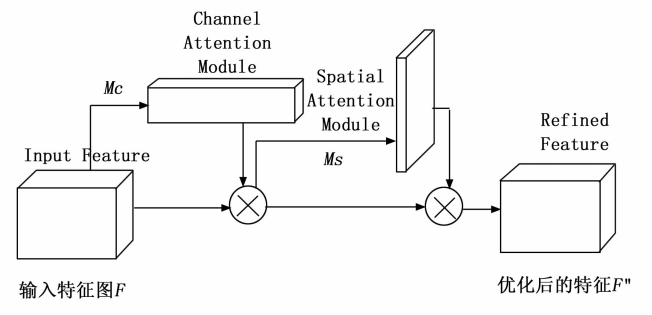


图 12 混合注意力机制结构示意图

perceptron）生成通道注意力图 M_C[62]，计算过程如式（2）所示：

$$M_C(F) = sigmoid(MLP(AvgPool(F)) + MLP(MaxPool(F))) = sigmoid(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \quad (2)$$

其中：W_1、W_0 为 MLP 权重，F^c_{avg}、F^c_{max} 为平均池化特征与最大池化特征。

空间注意力机制由特征图内部的空间关系产生，对平均池化特征与最大池化特征拼接后进行卷积，如式（3）所示。

$$M_S(F) = sigmoid(conv^{n\times n}(MLP(AvgPool(F)), MLP(MaxPool(F)))) = sigmoid(conv^{n\times n}(W_1(W_0(F^c_{avg})), W_1(W_0(F^c_{max})))) \quad (3)$$

其中：$conv^{n\times n}$ 代表 $n \times n$ 大小的卷积操作。

此外，文献［63］在 Transformer 模型中嵌入了空间注意卷积与时间注意卷积，使得模型能够将计算资源分配到手部出现的帧序列和帧位置中。文献［64］提出了一种动态－静态区域注意力模块，该模块导出高斯热图和动态运动图，分别在空间和时间域中突出手或手臂区域和运动信息，为后续特征提取网络减少了冗余信息的输入，极大地提高了模型的训练速度。

根据已在网上公开的代码，本文将普通的双流网络与使用 MobileNetV2 的双流网络置于同一数据集 Jester 下进行训练与测试以便更好的进行横向比对，硬件平台采用 Intel Core i7 11390H、RTX3070Ti，迭代更新 100 次。如表 3 所示，在采用 MobileNetV2 作为双流网络的子网络进行时间空间信息提取后，识别准确率有了很大的提升，引入了倒残差模块的轻量级 MobileNetV2 在增加了模型深度的情况下依然能够保持较小的运算规模，且准确度相较于标准的 CNN 子网络有了很大的提高。在基于 C3D 的识别模型中，得益于残差块的引入，卷积神经网络可以被设计得更深，使得运算规模可以更大，这也切实带来了准确率的提升，而 C3D 模型引入 MobileNetV2 后，在参数量不到基准模型十分之一的条件下，识别准确率能够只相差百分之五左右，这无疑是对 C3D 模型巨大的改进，SqueezeNet 也在较小规模的计算量下，达到了良好的识别率，这也表明轻量化的识别模型是未来深度学习领域的一个方向。得益于骨架数据在复杂场景中相较于 RGB 数据更好的鲁棒性，应用图神经网络的方法在各自的训练、测试数据集上都取得了不错的成绩。动态手势识别方法中所用到的 RNN 网络主要是 LSTM 网络以及少量双向 RNN 网络。LSTM 网络在最近几年提出的模型网络中使用率非常高，一般用作提取动态手势的时间维度特征或融合长短期信息。

表 3 基于深度学习的方法对比

类别	文献	方法	参数量	数据集	模态	准确率/%
基于卷积神经网络	[43]	双流网络	2 874 650	Jester V1	RGB	78
	[44]	双流网络＋MobileNetV2	3 538 984	Jester V1	RGB	91.25
	[45]	C3D	32 513 307	Jester V1	RGB	87.70
	[49]	ResC3D	46 307 419	Jester V1	RGB	88.34
	[47]	C3D＋ MobileNetV2	2 429 651	Jester V1	RGB	82.27
	[50]	C3D＋SqueezeNet	5 345 990	Jester V1	RGB	85.95
	[62]	C3D＋CBAM 机制	—	Jester V1	RGB	95.58
	[63]	Transformer＋注意力机制	—	Jester V1	RGB	96.72
基于循环神经网络	[55]	双流架构 RNN	—	Chalearn LAP	RGBD	91.7
	[56]	双层双向 LSTM	—	Handi-CASL	— 骨架	96.67 95.298
	[59]	重构 LSTM 单元	—	Jester V1 SKIG	RGB RGBD	95.54 96.63
	[60]	CNN＋LSTM	—	*	RGB	99.14
基于图神经网络	[53]	STGNN	—	CSL	骨架	97.27
	[54]	STGCN＋GRU	—	FPHA	骨架	91.16

“*”表中文献[60]所使用的方法，在自建的数据集上进行训练与测试，该数据集包含 66 930 帧双手训练图集与 12 195 帧测试图集。“－”表示未公开参数量或数据集模态。

此外，文献［54］、［62］、［63］、［64］提出的方法引入了注意力机制，可以发现引入注意力机制后，模型的识别准确率有了非常大的提升，注意力机制对于输入数据的权重分配处理，让模型能够将资源集中在相对而言更加重要的部分，提高了模型处理数据的效率，因此注意力机制应当引起相关研究者的关注。

3 挑战与展望

3.1 动态手势追踪与匹配

在进行动态手势识别的过程当中，由于手势所具有的高自由度、多变性以及背景复杂的特点，长时间地追踪手势和提高手势精度变得极具挑战，而在一些实际应用场景中，如体感游戏与 VR 中，长时间的追踪手势是不可避免

的，因此如何提高动态手势匹配与跟踪精度以及进行长时间有效追踪，将会是未来一个重要的研究方向。

3.2 轻量化的深度学习模型

在第三节的叙述中，可以看到轻量化模型如 SqueezeNet 与 MobileNet 在较小规模的计算中仍然能够取得良好的识别效果，并且，动态手势识别功能需要嵌入到不同应用场景的硬件平台中，受限于应用场景与硬件平台的算力，识别模型不能过于庞大，算法设计不能过于复杂，还要将算法的实时性能考虑在内，不能一位地追求高精度而忽略了实际应用场景与平台的限制，如何平衡精度的要求与模型体量的大小，将是未来研究的重要课题之一。

3.3 非受控环境下的手势识别

尽管在过去几年中大规模的数据集的收集有了不错的进展，但由于环境设置受到限制，包括光照变化等因素可能对手的颜色产生的影响，在受控环境中收集的数据集与真实世界环境之间存在很大差距。例如，大多数可用的数据集不涉及遮挡情况。然而，在实际场景中，遮挡是不可避免的。从多模态数据中恢复或找到用于此类识别任务的线索将是一个重要的研究方向。

4 结束语

本文从动态手势数据集、动态手势识别方法两个方面全面梳理了动态手势识别的相关技术，详细地对各种动态手势识别方法进行了讨论与归纳。大量的研究结果表明，基于深度学习的方法性能总体上要优于传统方法，未来动态手势识别的方向应该更偏向基于深度学习的方法，但是基于深度学习的方法往往需要耗费大量的资源去训练、推理模型，在一些实时性要求较高的场景下，如何提高模型的训练与推理效率是未来研究工作中的重点。

参考文献：

[1] YANG Y, SHAWN N. Bag-of-visual-words and spatial extensions for land-use classification [C] //Proc of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems. New York: ACM Press, 2010: 270 - 279.

[2] KAPITANOV A, MAKHLYARCHUK A, KVANCHIANI K. Hagrid-hand gesture recognition image dataset [EB/OL]. (2022-06-16) [2023-08-28]. ArXiv Preprint.

[3] FOTHERGILL S, MENTIS H, KOHLI P, et al. Instructing people for training gestural interactive systems [C] // Proc of the SIGCHI Conference on Human Factors in Computing Systems. New York: ACM Press, 2012: 1737 - 1746.

[4] WAN J, ZHAO Y B, ZHOU S, et al. ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition [C] //Proc of 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops. Washington DC: IEEE Computer Society, 2016: 56 - 64.

[5] MATERZYNSKA J, BERGER G, BAX I, et al. The Jester Dataset: A Large-Scale Video Dataset of Human Gestures [C] //Proc of 2019 IEEE/CVF International Conference on Computer Vision Workshop. Piscataway, NJ: IEEE Press, 2019: 0 - 0.

[6] DE S Q, WANNOUS H, VANDEBORRE J-P. Skeleton-based dynamic hand gesture recognition [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2016: 1 - 9.

[7] ZHANG Y, CAO C, CHENG J, et al. EgoGesture: A new dataset and benchmark for egocentric hand gesture recognition [J]. IEEE Transactions on Multimedia, 2018, 20 (5): 1038 - 1050.

[8] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? A new model and the kinetics dataset [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 6299 - 6308.

[9] GARCIA H G, YUAN S, BAEK S, et al. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 409 - 419.

[10] AVOLA D, BERNARDI M, CINQUE L, et al. Exploiting recurrent neural networks and leap motion controller for the recognition of sign language and semaphoric hand gestures [J]. IEEE Transactions on Multimedia, 2018, 21 (1): 234 - 245.

[11] ESCALERA S, BARó X, GONZALEZ J, et al. Chalearn looking at people challenge 2014: Dataset and results [C] //European Conference on Computer Vision, 2014: 459 - 473.

[12] LIU L, SHAO L. Learning discriminative representations from RGB-D video data [C] // Twenty-third International Joint Conference on Artificial Intelligence, 2013.

[13] TSIRONI E, BARROS P V, WERMTER S. Gesture recognition with a convolutional long short-term memory recurrent neural network [C] // Proceedings of ESANN, 2016: 213 - 218.

[14] SOOMRO K, ZAMIR A R, SHAH M. UCF101: A dataset of 101 human actions classes from videos in the wild [EB/OL]. ArXiv preprint arXiv: 1212.0402, 2012.

[15] HU Z, HU Y, LIU J, et al. 3D separable convolutional neural network for dynamic hand gesture recognition [J]. Neurocomputing, 2018, 318: 151 - 161.

[16] DE S Q, WANNOUS H, VANDEBORRE J-P, et al.

Shrec'17 track: 3d hand gesture recognition using a depth and skeletal dataset [C] //3DOR-10th Euro Graphics Workshop on 3D Object Retrieval, 2017: 1 - 6.

[17] AL-HAMMADI M, MUHAMMAD G, ABDUL W, et al. Hand gesture recognition for sign language using 3DCNN [J]. IEEE Access, 2020, 8: 79491 - 79509.

[18] SHAHROUDY A, LIU J, NG T-T, et al. NTU RGB+D: A large scale dataset for 3d human activity analysis [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 1010 - 1019.

[19] AMEUR S, KHALIFA A B, BOUHLEL M S. A comprehensive leap motion database for hand gesture recognition [C] // 2016 7th International Conference on Sciences of Electronics, Technologies of Information and Telecommunications (SETIT), 2016: 514 - 519.

[20] AMIR A, TABA B, BERG D, et al. A low power, fully event-based gesture recognition system [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017: 7243 - 7252.

[21] WU Z, WANG X, JIANG Y G, et al. Modeling spatial-temporal clues in a hybrid deep learning framework for video classification [C] //Proceedings of the 23rd ACM International Conference on Multimedia, 2015: 461 - 470.

[22] LIU D, ZHANG L B, WU Y J. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) [C] // 2022: 3304 - 3312.

[23] PU J, ZHOU W, LI H. Iterative alignment network for continuous sign language recognition [C] //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019: 4165 - 4174.

[24] WETZLER A, SLOSSBERG R, KIMMEL R. Rule of thumb: Deep derotation for improved fingertip detection [J]. ArXiv Preprint ArXiv: 1507.05726, 2015.

[26] LI J, LI C, HAN J, et al. Robust hand gesture recognition using hog-9ulbp features and svm model [J]. Electronics, 2022, 11 (7): 988.

[27] MAHARANI D A, FAKHRURROJA H, MACHBUB C. Hand gesture recognition using K-means clustering and support vector machine [C] //2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), IEEE, 2018: 1 - 6.

[28] SRUTHI C J, LIJIYA A. Double-handed dynamic gesture recognition using contour-based hand tracking and maximum mean probability ensembling (MMPE) for Indian Sign language [J]. The Visual Computer, 2022: 1 - 21.

[29] IKRAM A, LIU Y. Real time hand gesture recognition using leap motion controller based on CNN-SVM architechture [C] //2021 IEEE 7th International Conference on Virtual Reality (ICVR), IEEE, 2021: 5 - 9.

[30] CAI S, LU Z, CHEN B, et al. Dynamic gesture recognition of A-mode ultrasonic based on the DTW algorithm [J]. IEEE Sensors Journal, 2022, 22 (18): 17924 - 17931.

[31] ZHOU Z, DAI Y, LI W. Gesture recognition based on global template DTW for Chinese sign language [J]. Journal of Intelligent & Fuzzy Systems, 2018, 35 (2): 1969 - 1978.

[32] CARMONA J M, CLIMENT J. A performance evaluation of HMM and DTW for gesture recognition [C] //Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications: 17th Iberoamerican Congress, CIARP 2012, Buenos Aires, Argentina, Springer. 2012: 236 - 243.

[33] HISHAM B, HAMOUDA A. Arabic static and dynamic gestures recognition using leap motion [J]. J. Comput. Sci., 2017, 13 (8): 337 - 354.

[34] ARVANITIS N, SARTINAS E, KOSMOPOULOS D. Procrustes-DTW: Dynamic time warping variant for the recognition of sign language utterances [C] //2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW). IEEE, 2023: 1 - 5.

[35] LIU J, LU X, YIN H, et al. A constraints-based dynamic time warping method for gesture recognition with kinect [C] //Proceedings of the 2023 3rd International Conference on Robotics and Control Engineering. 2023: 178 - 183.

[36] PARCHETA Z, MARTiNEZ-HINAREJOS C D. Sign language gesture recognition using HMM [C] //Pattern Recognition and Image Analysis: 8th Iberian Conference, IbPRIA 2017, Faro, Portugal. Springer International Publishing, 2017: 419 - 426.

[37] SINHA K, KUMARI R, PRIYA A, et al. A computer vision-based gesture recognition using hidden markov model [C] //Innovations in Soft Computing and Information Technology: Proceedings of ICEMIT 2017, Springer Singapore, 2019: 55 - 67.

[38] TUR A O, KELES H Y. Evaluation of hidden Markov models using deep CNN features in isolated sign recognition [J]. Multimedia Tools and Applications, 2021, 80: 19137 - 19155.

[39] GUO X L, YANG T T. Gesture recognition based on HMM-FNN model using a Kinect [J]. Journal on Multimodal User Interfaces, 2017, 11: 1 - 7.

[40] HUSSAIN S, SAXENA R, HAN X, et al. Hand gesture recognition using deep learning [C] //2017 International SoC Design Conference (ISOCC). IEEE, 2017: 48 - 49.

[41] DEVINEAU G, MOUTARDE F, XI W, et al. Deep learning for hand gesture recognition on skeletal data

[C] //2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). IEEE, 2018: 106 - 113.

[42] XIE J, ZHANG B, Lü Q. A dynamic head gesture recognition method based on 3D convolutional two stream network fusion [J]. Acta Electonica Sinica, 2021, 49 (7): 1363.

[43] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos [J]. Advances in Neural Information Processing Systems, 2014, 27.

[44] HUU P N, NGOC T L. Two-stream convolutional network for dynamic hand gesture recognition using convolutional long short-term memory networks [J]. Vietnam Journal of Science and Technology, 2020, 58 (4): 514 - 523.

[45] 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.

[46] SANDLER M, HOWARD A, ZHU M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks [C] // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018: 4510 - 4520.

[47] KOPUKLU O, GUNDUZ A, KOSE N, et al. Real-time hand gesture detection and classification using convolutional neural networks [C] // 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), 2019: 1 - 8.

[48] TRAN D, RAY J, SHOU Z, et al. Convnet architecture search for spatiotemporal feature learning [J]. ArXiv Preprint ArXiv: 1708.05038, 2017: 1 - 12.

[49] LI Y, MIAO Q, QI X, et al. A spatiotemporal attention-based ResC3D model for large-scale gesture recognition [J]. Machine Vision and Applications, 2019, 30: 875 - 888.

[50] LIU S, REN Y, LI L, et al. Micro-expression recognition based on SqueezeNet and C3D [J]. Multimedia Systems, 2022, 28 (6): 2227 - 2236.

[51] GORI M, GABRIELE M, FRANCO S. A new model for learning in graph domains [C] // Proceedings 2005 IEEE International Joint Conference on Neural Networks. Montreal: IEEE, 2005: 729 - 734.

[52] YAN S, XIONG Y, LIN D. Spatial temporal graph convolutional networks for skeleton-based action recognition [C] //Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32 (1).

[53] YUAN G, BING R, LIU X, et al. CAI Zhuo. Spatial-temporal graph neural network based hand gesture recognition [J]. Acta Electonica Sinica, 2022, 50 (4): 921 - 931.

[54] SLAMA R, RABAH W, WANNOUS H. STR-GCN: Dual spatial graph convolutional network and transformer graph encoder for 3D hand gesture recognition [C] // 2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG), Waikoloa Beach, HI, USA, 2023: 1 - 6

[55] WANG H S, WANG L. Modeling temporal dynamics and spatial configurations of actions using two-stream recurrent neural networks [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 499 - 508.

[56] YANG L, CHEN J, ZHU W. Dynamic hand gesture recognition based on a leap motion controller and two-layer bidirectional recurrent neural network [J]. Sensors. 2020; 20 (7): 2106.

[57] LEFEBVRE, GRéGOIRE, et al. Inertial gesture recognition with blstm-rnn [J]. Artificial Neural Networks: Methods and Applications in Bio-/Neuroinformatics. Cham: Springer International Publishing, 2015: 393 - 410.

[58] AMEUR S, KHALIFA A B, BOUHLEL M S. A novel hybrid bidirectional unidirectional LSTM network for dynamic hand gesture recognition with leap motion [J]. Entertainment Computing, 2020 (35): 100373.

[59] PENG Y, TAO H, YUAN H, et al. Dynamic gesture recognition based on feature fusion network and variant ConvLSTM [J]. IET Image Processing, 2020, 14. 11: 2480 - 2486.

[60] SHARMA V, JAISWAL M, Sharma A, et al. Dynamic two hand gesture recognition using CNN-LSTM based networks [C] //Proceedings of 2021 IEEE International Symposium on Smart Electronic Systems (iSES), Jaipur, India, 2021: 224 - 229.

[61] ZHANG X L, TIE Y, QI L. SlowFast convolution LSTM networks for dynamic gesture recognition [C] // Proceedings of the 2021 3rd Asia Pacific Information Technology Conference. 2021: 59 - 63.

[62] 黄　圣，茅　健. 基于注意力机制的动态手势识别方法 [J]. 智能计算机与应用，2023，13 (9)：111 - 115.

[63] ZHANG Y, WANG F P. HandFormer: A dynamic hand gesture recognition method based on attention mechanism [J]. Applied Sciences, 2023 (13) 7: 4558.

[64] ZHOU B, LI Y, WAN J. Regional attention with architecture-rebuilt 3D network for RGB-D gesture recognition [J]. Proceedings of the AAAI Conference on Artificial Intelligence, 35 (4): 3563 - 3571.