Computer Science and Application 计算机科学与应用, 2020, 10(5), 990-1001
Published Online May 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.105102

Hans汉斯

A Review of Vision Based Dynamic Hand Gestures Recognition

Jie Sha[1], Jian Ma[1], Haijun Mou[2], Jianhua Hou[1*]

[1]College of Electronics and Information Engineering, South-Central University for Nationalities, Wuhan Hubei
[2]WuhanJiemu Technology Co., Ltd. Wuhan Hubei
Email: *hou8781@126.com

Received: May 3rd, 2020; accepted: May 18th, 2020; published: May 25th, 2020

Abstract

With the rapid advancement of computer vision, dynamic gesture recognition has been applied in a range of applications, such as human-computer interaction, intelligent driving, virtual reality, etc. In order to understand the current status of dynamic gesture recognition, the paper provides a comprehensive survey from three aspects: multi-mode input, hand detection and dynamic gesture modeling. The challenges faced by dynamic gesture recognition and application scenarios are also introduced.

Keywords

Dynamic Gesture Recognition, Multi-Mode Input, Hand Detection, Gesture Modeling

基于视觉的动态手势识别综述

沙　洁[1]，麻　建[1]，牟海军[2]，侯建华[1*]

[1]中南民族大学电子信息工程学院，湖北　武汉
[2]武汉杰目科技有限公司，湖北　武汉
Email: *hou8781@126.com

收稿日期：2020年5月3日；录用日期：2020年5月18日；发布日期：2020年5月25日

摘　要

随着计算机视觉研究的迅速发展，动态手势识别在人机交互、智能驾驶、虚拟现实等领域得到了广泛的应用。为了解动态手势识别的发展现状，通过对近年来动态手势识别算法的研究，从多模态输入、手部

*通讯作者。

文章引用: 沙洁, 麻建, 牟海军, 侯建华. 基于视觉的动态手势识别综述[J]. 计算机科学与应用, 2020, 10(5): 990-1001.
DOI: 10.12677/csa.2020.105102

检测、手势建模三个方面进行梳理总结，介绍当前动态手势识别算法研究进展、面临的难题和应用场景。

关键词

动态手势识别，多模态输入，手部检测，手势建模

1. 引言

肢体动作是一种有效的交流手段，追根溯源，人们最早是通过肢体动作进行沟通的。后来，随着语言的出现，语言成为人们主要的交流方式，但是肢体动作在传达信息方面仍然有着举足轻重的作用。如裁判通过动作发出判决结果、失语者通过动作传递信息。过去，肢体动作的识别一般只发生在人与人之间，可是随着科技的发展以及人们对智能设备有着很大的应用需求，目前，通过机器识别出肢体动作是一个热门的研究方向[1] [2] [3] [4]。

肢体动作识别可以分成手势识别[5] [6]和行为识别[7] [8] [9]，手势识别与行为识别的原理是相通的，都是操作者做出动作后，算法经过分析、理解，识别出动作的种类。手势主要发生在手部，行为可能发生在人体的各个部位。因为手部是人体的一部分，所以行为识别算法与手势识别算法可以互相进行迁移学习。但是，又因为手部结构与人体结构的不同特性，手势与行为仍然存在着较大的差异。具体差异见表 1。

手势识别是人机交互(HCI) [10]领域的一个研究热点。根据一个手势是发生在某一时刻还是一个时间段，手势识别可以分为两大类：静态手势识别[11] [12]和动态手势识别[13] [14] [15]。静态手势识别的研究对象是某个时间点上的手势图像，其识别结果与图片中手部的外观特征，如位置、轮廓、纹理有关；动态手势识别的研究对象是连续时间段上的图像序列，其识别结果与图片中手部的外观特征及序列中描述手部运动轨迹的时序特征有关。相比于静态手势，动态手势种类更多、表达能力更强、更具有实用性。

动态手势识别算法的任务是当操作者做出一个手势后，计算机能够捕捉并准确地识别出该手势。如何从外界将手势输入到计算机中？解决方法分为两种。一种是基于有线技术将计算机与操作者连接在一起。如数据手套[16]能够将操作者的手部信息传送到计算机中，并通过算法对其进行手势识别；虽然识别效果较好，但是繁琐的穿戴方式、要求严格的操作环境影响了用户体验的自然性和易用性。另一种是基于视觉的方法[17] [18]，不需要手部与计算机有任何接触，只需普通的摄像机就可以直接捕获手势的 RGB 模态数据。后来推出的 Kinect [19]摄像机可以同时捕获手势的 RGB 数据和 Depth 数据，与基于数据手套、基于普通摄像头的方式相比，这种方法更加方便快捷，能够满足操作者更多的应用需求，成为计算机获取手势的首选方法。

本文对近年来基于视觉的动态手势识别算法进行了研究，其基本流程如图 1 所示。本文着重从输入数据模态、手部检测、动态手势建模三方面进行梳理。本文第 2 节、第 3 节、第 4 节分别介绍多模态数据在动态手势识别中的应用、手部检测器对动态手势识别的预处理作用、动态手势识别算法的建模方式；其中动态手势建模包括基于手工特征和基于深度学习的方法；第 5 节简要指出了当前动态手势识别算法面临的挑战和具体应用场景；最后是全文总结。

Table 1. The difference between gesture recognition and action recognition
表 1. 手势识别与行为识别的差异

	手势识别	行为识别
动作发生部位	手部/手臂	人的整个躯体
图像中的占比	手部在人体中的占比较小，当一个人出现在摄像头中时，手部在图像中的占比很小	正常情况下，如果人体不离摄像头非常远，人体在图像中的占比不会太小
自遮挡	手指间距离较小、手指运动幅度较小，手部做运动时，容易发生自遮挡	四肢较长，且运动幅度大，不易发生自遮挡
复杂度	手部凭借着复杂的手部结构可以做出成千成万种手势。但是因为手指和关节的运动幅度较小，不同类手势间的差异可能很小，增加了手势识别的复杂度	行为发生在人的四肢、躯干和头部的共同作用下。相较于手势，其发生部位结构较简单，虽然行为种类间也存在着较小的差异，但是其复杂度小于手势的
背景干扰	制约了手势识别精确度的提高	一些背景会有助于行为的正确识别，如“打篮球时”，背景中应该出现篮球

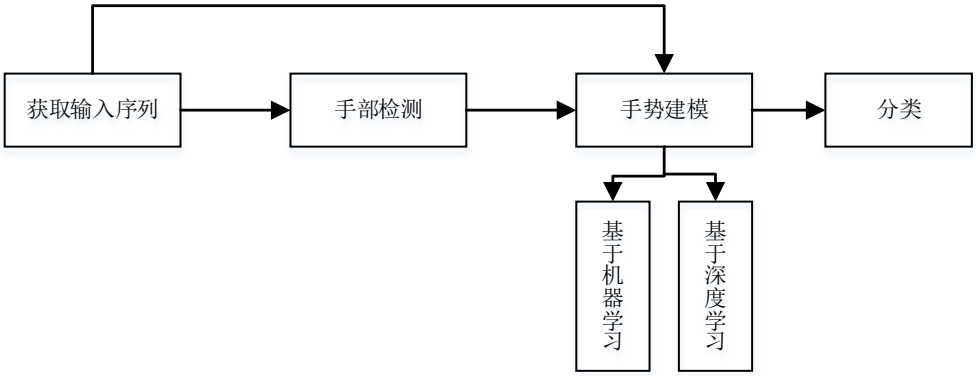
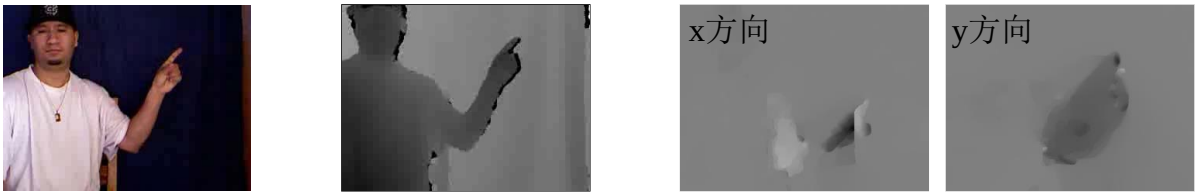


Figure 1. The basic flow of dynamic gesture recognition algorithm
图 1. 动态手势识别算法的基本流程

2. 多模态输入

一个序列不同的模态携带不同的信息，如图 2 所示，RGB 图片包含丰富的颜色和纹理信息，深度图片包含物体的轮廓和深度信息，光流分别描述了像素在 x 方向与 y 方向的运动信息。这些数据在信息描述上具有互补性，将它们融合起来，能够有效弥补单一模态数据的局限性。例如“up-down”手势用 RGB 模态序列表示时，每帧图片对手部外观描述的差异很小，很难判别序列中发生的手势类别，但用 Depth 模态表示该手势，由于深度图片中描述手部的像素值会随着手部与摄像头之间的距离变化而变化，可以根据像素值的差异很容易判别出手势类别。“turn-around”手势用 Depth 模态序列表示时，每帧图片上只有手部轮廓，且手部的细节描述非常模糊；但使用 RGB 模态序列表示，每帧图片上的手部外观清晰可见，大大提高了手势的辨识度。Wang 等[20]提出基于异构网络的框架来处理手势识别：用 3D ConvLSTMs [21]识别视频序列中的动态手势，用 ConvNet 识别由 Rank Pooling 构建的动态图像中的手势，将这两种网络分别应用于 RGB 与 Depth 序列及各自对应的只含有手部区域的序列，最后取各模态输出的均值得到识别结果。Köpüklü 等[22]提出运动融合框架(Motion Fused Frames, MFFs)，通过堆叠 RGB 图片与光流图片描述手势运动的外观及运动信息。文献[21] [23] [24]利用 RGB 序列和 Depth 序列共同描述手部的运动过程。文献[13] [14] [15]等基于 RGB 序列和 Depth 序列和 Flow 序列的融合特征进行动态手势识别。这些文献证实了多模输入是增强特征鲁棒性的有效手段，能显著提高手势识别的精度。



(a) RGB图片 (b) Depth图片 (c) Flow图片

Figure 2. Description information of each mode picture
图 2. 各模态图片的描述信息

如何融合多个模态数据的信息是一个值得思考的问题，恰当的融合方式是提高特征鲁棒性的有效手段，可以更有效地利用各通道的数据信息。常用的融合方式包括数据级融合[10]、特征级融合[25] [26] [27]和分数级融合[21] [24] [28]。1) 数据级融合发生在数据输入之前，这种方法在多模态数据输入情况下只训练一个网络，可以大大减少参数量；在调用网络时，只需要简单的修改输入端参数即可，可以自动建立各模态数据之间的像素对应关系。不过由于输入过于复杂，特征提取器很难从中同时学习到各通道携带的最优化信息，会导致识别率较低；而且当多模数据来自不同机器时，该方案很难实现。2018 年，Köpüklü 等人[22]第一次把数据级别的融合应用到行为和手势识别中，他们将若干帧光流图像连接在单张彩色图像的通道维度上，将其作为深度网络的输入，使得输入数据中既包含空间信息又携带时序信息，以此来实现动态手势识别。在 isoGD [29]测试集上，准确率只有 56.7%。2) 特征级融合发生在输入与输出之间的任何位置。文献[30]首先将 RGB 序列与 Depth 序列通过双流循环神经网络(two streams Recurrent Neural Networks, 2S-RNN)提取出时序特征，然后采用一个 LSTM 层将两种特征融合在一起。文献[25] [27]分别将特征提取器输出的各通道数据特征拉伸为向量，并将这些向量串联成一个向量，该融合向量经过 SVM 分类器处理后得到对动态手势的预测结果。Miao 等[26]提出基于典型关联分析(canonical correlation analysis, CCA)的融合方法，以最大化各模态特征间的成对相关性为目的，融合由 RGB 序列、Depth 序列以及 Flow 序列生成的时空特征，融合特征经过 SVM 分类器处理后得到最后的手势识别结果。特征级融合虽然能独立的学习各通道空间时序特征，但其融合特征容易丢失各通道信息的优势及各通道之间的关联信息，而且该方式需要训练的参数量会增加。3) 与数据级和特征级融合相比，分数级融合是最简单的，它既解决了数据级融合和特征级融合存在的问题，又充分利用了各模态数据携带的信息且不需修改网络结构。在训练时，多个输入数据相互独立，每个数据在网络的最后通过 softmax 函数得到对每类手势的预测置信度，然后对所有的输出进行取最大值或均值处理便可得到融合结果，既最终的输出结果。其问题在于最大值处理直接忽略了非最大值对应数据提供的信息，均值处理容易削弱关键信息的作用强度，限制了融合效果的发挥。为解决该问题，Narayana 等[31]提出通过训练一个神经网络，为每一类动态手势学习出在各个输入数据上的权重，即每类动态手势对各个输入数据的依赖程度。

3. 手部检测器

手部检测[31]-[40]是手势识别中重要的预处理环节，它可以挖掘出重要的手势信息，去除背景干扰。早期的手部检测方法主要利用人工提取的特征来获取图片中手部的区域[32] [33] [34]，如基于肤色的方法[32]，基于形状的方法[33]，基于运动信息的方法[34]等。这些方法易受光照变化、肤色差异、背景干扰、姿态变化、手指间自遮挡等影响，检测效果不理想且不稳定，而且计算量大、计算速度慢，难以满足实际场景中的检测要求。随着深度学习的发展，基于深度学习的检测器自动从图片中学习手部特征[31] [35]-[40]，而且该特征对手部信息有更强的表达能力。Zimmermann 等[35]利用 HandSegNet 检测手部，

HandSegNet 是一个 16 层的卷积神经网络，输入一帧 RGB 图片，返回一个 2 通道图片：一个通道是手模图片，另一个是背景模图片，将手模图片与原图合并，可以裁剪出手部区域。近年来，基于候选区域的检测器也被应用于手部检测。Ren 等[36]通过将 Faster R-CNN 微调后去检测每帧图片手部区域的边界框，以消除不在边框内的场景。为了利用 Depth 深度信息增强特征的鲁棒性、提高手部检测的精确度，Liu 等[37]提出了融合彩色和深度数据的双通道 Faster R-CNN 手部检测框架，该方法在原有 Faster R-CNN 检测框架基础上，增加了 Depth 通道信息，并在特征层面上将其与 RGB 通道的信息进行融合；将融合后的特征通过 RPN 产生可能包含手部区域的候选框，通过对候选框对应的感兴趣区域进行池化、分类和回归，得到最终的手部检测结果，显著提升了手势检测性能。但是对于一些运动幅度大或双手的手势，边界框的大小接近整个图像的尺寸，难以消除背景干扰。为解决该问题，Narayana 等[31]在用 two stream Faster R-CNN 检测出每帧图片手部区域的基础上，使用多人姿态估计器从 RGB 图片中提取人的骨架信息，用于区分人的左右手，此时手部检测任务为分别检测出左右手。

由于手部区域在全景中占比和位置具有未知性，文献[31]基于手部检测器采用全局信息与局部信息相结合的方式描述一个手势的发生过程。全局信息表示整个视频序列描述的信息，包括背景变化、手部运动细节和非手部的其他人体部件的运动过程；局部信息仅表示手部的运动过程，由检测器通过在每帧图片中检测出手部区域并组成一个新的序列而获得。针对人体或环境对手部运动提供标定作用或产生很小干扰的动态手势来讲，利用全局信息可以高效率的识别出手势类别；但是，对于一些相对于身体的偏移手部仅发生微小运动的手势来讲，背景或噪音会干扰甚至覆盖掉手部的运动信息，如“勾食指”，当人的身体发生大幅度的晃动，相比于身体大区域的移动，食指小区域的位移很可能被忽略掉，进而影响手势的判别结果。利用全局信息与局部信息在信息描述上的互补性，可以显著提高算法的鲁棒性。

4. 动态手势建模

如何从输入中获取关键信息作为手势的判别依据，是动态手势识别算法的研究核心。根据关键信息获取方式的不同，动态手势识别算法可以分成两大类：基于人工特征的机器学习算法[41]、基于自主学习特征的深度学习算法[42] [43]。

4.1. 基于人工特征的机器学习算法

经典的方法有规整动态时间(Dynamic Time Warping, DTW)算法[44] [45]、隐马尔科夫模型(Hidden Markov Model, HMM) [46] [47]。DTW 是一种衡量两个不同长度时间序列相似度的方法，最早用于语音识别，Corradinni 等[48]第一次将其用于识别动态手势。基于 DTW 进行动态手势识别时，首先需要对训练集中的每个手势进行预处理，然后提取特征并将其归一化为序列模板，再对待测手势进行与训练手势一样的操作，最后将生成的结果与训练集中的每一个模板进行匹配分析，将距离最小模板的类别作为手势识别结果，具体流程如图 3 所示。DTW 方法没有采用统计模型框架进行训练，也很难将上下文的各种知识用于图像识别算法中，因此在解决大数据量、复杂手势等问题时存在劣势。为了解决该问题，很多文献对 DTW 进行了改进。如张建荣[49]提出了一种改进的动态时间规整(Improve Dynamic Time Warping, iDTW)动态手势识别算法，它在 DTW 的基础上采用点和线相结合的范围来约束搜索路径以防止搜索不合理，并利用节点在运动序列中的距离方差对各节点进行权值动态分配。与 DTW 相比，iDTW 在公开数据集上的精确度整体提高，并且运算量显著减少。HMM 是一个经典的模型，它主要用于处理基于时间序列或状态序列的问题；在基于 HMM 的算法中存在两类数据，一类是可以观测到的数据，即观测序列；另一类是隐藏数据，即状态序列。在动态手势识别中，手部所做的一系列动作是观测序列，而手部所发生的动作为隐藏序列，HMM 的任务就是从一系列可观测数据中确定隐藏数据中的内容。在动态手势识

别中，每类手势对应一个隐马尔科夫模型；在训练时，首先将每个手势样本按类别划分，然后利用前向与后向算法为每类手势训练出对应的 HMM 模型。在测试时，待测样本利用前向算法与所有的 HMM 模型进行遍历，计算出每个 HMM 模型产生该手势序列的概率值，数值最大的 HMM 模型对应的类别便是待测样本的识别结果。Saha 等[50]基于 HMM 算法完成了对 60 个不同主题下的 12 种动态手势的识别，每类手势都获得了将近 90%的识别精度。

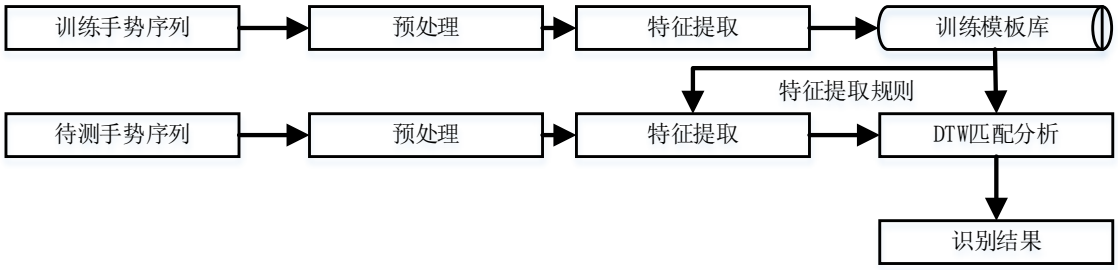


Figure 3. Dynamic gesture recognition algorithm flow based on DTW algorithm
图 3. 基于 DTW 算法的动态手势识别算法流程

基于人工特征的动态手势识别算法已经取得了较好的识别效果，但针对不同的环境、不同的运动角度，研究者很难设计出一种对所有样本都适用的特征提取方法，基于人工特征的动态手势识别算法仍然存在很大的挑战。

4.2. 基于深度学习的动态手势识别算法

近年来，随着深度学习的兴起，基于深度学习的动态手势识别得到了广泛的关注。其核心思想是首先搭建神经网络，对模型初始化，然后利用神经网络的前向传播、损失计算、后向反馈，从数据中学习网络参数。不同于人工特征已经固定好对输入的处理方式，神经模型的参数完全由数据的特性和神经网络的结构共同决定，直接影响手势识别的性能。因为输入数据是固定的，所以搭建鲁棒的神经网络是整个基于深度学习算法的关键。根据空间信息与时序信息的编码方式，当前常用的神经网络可以分成两大类：一类是基于 2DCNNs 的 two stream 网络[51] [52] [53] [54]，另一类是 3DCNNs [55]。

4.2.1. 基于 Two Stream 的算法

这种网络由空间网络(spatial network)和时序网络(temporal network)两个子网络组成，分别负责从 RGB 图片中挖掘出手部的空间信息和从堆叠的光流中挖掘出手部的运动信息，再将两种信息融合后构成时空信息用于视频分析任务。spatial network 和 temporal network 结构相同，一般可以采用 AlexNet、VGG、ResNet、Inception 等经典的 2DCNNs 构成。2014 年，Simonyan 等[51]第一次提出 two stream 网络结构，并将其成功的应用于行为识别任务。该算法中子网络由 VGG 搭建而成，其工作流程如下：首先从视频序列中随机采样一帧 RGB 图片作为 spatial network 的输入，并取该帧随后的 5 帧光流图片作为 temporal network 的输入；单帧 RGB 图片和堆叠的光流图片经过两个子网络分别学习对输入类别的预测概率向量，最后将这两个概率向量在类别的维度上取均值，得到最终的识别结果。在公开数据集 UCF-101 [56]和 HMDB-51 [57]上与其它算法相比，该算法取得了当时最优效果。但由于该算法仅从序列中选取一帧图片和几帧光流图片为两个子网络的输入，只能理解行为的短期内容。然而在视频的理解中长期信息起着关键的作用，这些信息的丢失会大大影响识别效果。为解决该问题， Wang 等[52]提出了时域分割网络(Temporal Segment Network, TSN)，该算法引入稀疏采样策略对长时间序列信息进行建模；为了获得更鲁棒的特征表达，TSN 使用 Inception v2 [58]搭建子网络，其工作流程如下：首先将视频片段均匀分成 N 段，

对每段进行如[51]的操作，得到算法对每个短期小片段的预测概率向量(RGB 图片和堆叠的光流分别对两个子网络权值共享)；然后再将所有片段的概率得分取均值得到长期视频的预测结果。为了进一步提高 two stream 算法的性能，Feichtenhofer 等[53]在 TSN 的基础上研究空间信息与时序信息的融合方式。实验证明，在卷积层融合两类信息不仅不会降低性能还会减少参数的数量；在网络高层卷积层上进行融合比在低层卷积层上效果要好；对类别得分进行融合可以提高性能，文献[51] [52]采用的是此种方法。在 two stream 算法中，手势的运动信息来自于光流图片，但是该图片的获取需要对邻近的 RGB 图片进行像素级别的计算，需占用大量内存空间，浪费了时间和存储资源。为解决该问题，Zhu 等[54]提出通过训练一个卷积神经网络 MotionNet 做光流估计来代替光流计算。在该算法中，temporal network 和 MotionNet 级联后以堆叠的数张 RGB 图片为输入。动态手势识别和行为识别的原理是相同的，都是利用算法理解出视频序列中人的肢体动作所要传达的语义信息。随着 two stream 算法在行为识别上的成功应用，国内外大量的研究者提出了基于 two stream 做动态手势识别。如 Okan Köpük 等[22]提出将若干帧光流图与单张彩色图像堆叠在一起后送入 TSN 的 temporal network 做动态手势识别。

4.2.2. 基于 3DCNNs 的算法

3DCNNs 可以从序列中同时提取出空间与时序信息，它由多个 3 维卷积层、3 维池化层以及激活函数组成。3 维卷积层和 3 维池化层对特征图的操作分别与 2 维情形类似，仅有的差异在于 2 维卷积层和 2 维池化层只对一个特征图在宽和高的维度上进行操作，但 3 维卷积层和 3 维池化层对多个特征图同时在宽、高和时间维度上进行操作。因此 3DCNNs 可以从序列中同时捕捉空间与时序信息。最经典的 3DCNNs 是深度 3 维卷积神经网络(Deep 3 Dimensional Convolutional Network, C3D) [59]，其网络结构简单，只有 8 个卷积层、5 个池化层、2 个全连接层以及一个 sotfmax 层；每个卷积层后接一个 relu 激活函数实现对网络的非线性变换，除了第一个和第二个卷积层后面都连接一个池化层外，其余都采用每两个卷积层连接一个池化层，这种结构可以在不增加参数的同时增大神经网络的感受野，两个全连接层都设置 Dropout = 0.3。C3D [39]是第一个被验证具有有效性及高效性的空间时序特征提取器，被很多动态手势识别算法采用。如 Liu 等[37]提出基于 C3D 的动态手势识别。Zhu 等[28]提出将金字塔输入和金字塔融合策略嵌入 C3D 模型，在该算法中，金字塔输入对每个手势序列进行金字塔式划分，并利用时间抖动对每个金字塔段进行均匀采样，该输入能够保持手势序列的多尺度上下文信息，金字塔融合放置在 C3D 最后一个全连接层的后面，在将金字塔段的特征融合后连接一个 softmax 层得到对手势的预测结果。Li 等[23]提出基于 C3D 从一系列梯度图与深度图中识别驾驶员手势。Gupta [24]提出一种具有连续时间分类(Connectionist Temporal Classification, CTC)的 R3DCNN 网络框架来进行手势识别，将 3D 卷积神经网络与循环卷积神经网络(Recurrent Neural Network, RNN)级联在一起，先用 C3D 学习序列短期的时序信息，在此基础上使用循环神经网络学习序列的长期时序信息。RNN 的使用增强了时序信息前后文之间的联系，提高了整个视频序列时序特征的表达能力；但是由于梯度膨胀和梯度消失等问题，RNN 容易丧失远距离学习信息的能力。为解决该问题，Zhu 等[21]提出 C3D 和长短期记忆网络(Long Short Term Memory network, LSTM)级联的框架来提取动态手势的时序特征。Zhang 等[25]提出使用 C3D 和双向长短期记忆网络(Bidirectional Convolutional Long-Short-Term-Memory Networks)来学习 2D 时空特征图，在此基础上进一步利用 2DCNNS 从 2D 特征图中学习更高级别的时空特征以用于最终手势识别。虽然基于 C3D 识别动态手势取得了较好的效果，但是简单的网络结构制约了其特征表达能力；加深网络深度是一种解决思路，可是无限制增加深度又会引起退化问题。为解决该问题，Tran 等[60]提出 ResC3D 网络结构，将残差的概念引入到 C3D 中；ResC3D 中的核心是残差块，它通过学习恒等函数使得网络深度加深时不至于降低性能，甚至能够提高性能。Li 等[26]先是直接基于 ResC3D 做动态手势识别算法，后来为了能够充分利用关键信息，

Li 等[27]提出在对输入序列进行特征提取前，预先引入注意力机制，使网络的注意力集中在能够有效描述手势的视频帧上和执行者发生运动的区域。

5. 动态手势识别的研究难点和应用场景

研究难点：动态手势识别作为计算机视觉的研究热点，近年来取得了长足的发展。对于简单的、区分度较大的动态手势数据集，一些主流算法已经取得了优秀的效果；但是对于含有复杂动态手势的数据集，当前最先进算法的识别效果仍然差强人意。分析其原因如下：

1) 动态手势复杂度高。手是一个非常灵活的人体部位，凭借五根手指上的关节以及两个手的任意搭配，操作者可以做出成百上千种动态手势，虽然大大增加了动态手势的类别，但是却容易引起某些问题，如类间差距小、类内差距大。

2) 环境干扰。嘈杂的运动环境、变化的光照条件和人为的干扰都对动态手势的识别带来很大的影响。

3) 不规范性。每个动态手势并没有规范的动作指标，如手指的幅度、运动速度，同一个人在做同一个手势时，手势也存在较大差异。

4) 关键区域的大小不确定。只发生在手指上的动态手势，如“勾食指”、“弹指”，在每帧图片上占比很小；而需要手部和手臂共同参与的动态手势，如“摆手”、“鼓掌”，在每帧图片上占比较大。如何提取出关键区域，是一个很重要的问题。

应用场景：近年来，由于基于视觉的动态手势识别技术能够自然地实现人机交互，动态手势识别在众多场景得到了广泛的应用[61] [62] [63] [64]。例如手语识别：聋哑人的手语与动态手势识别技术相结合，开辟了聋哑人与正常人之间沟通的新方式，大大方便了聋哑人士的生活。智能驾驶：智能驾驶是汽车科技推新的一个热点，车内摄像头捕捉到驾驶员的动态手势后，经过智能分析、理解和识别，可以实现对汽车导航、娱乐系统的控制。虚拟现实：这种方式多用于娱乐游戏中，通过虚拟设备，将操作者的手势直接投递到虚拟场景中，实现现实场景与虚拟场景的互动。

6. 展望

当前，动态手势识别算法已经取得了不错的识别效果，但是仍然存有进步的空间。为了提高识别效果，后续的研究者可能从以下几方面进行研究。

1) 制作动态手势数据集。当前动态手势数据集较少，而且所包括的手势种类也比较少。增加在实际场景中的动态手势数据，可以提高动态手势识别算法在实际场景中的应用能力。

2) 提高手部检测器的检测精度。在动态手势识别算法中，常使用手部检测器对输入进行预处理，处理结果直接影响着动态手势识别的效果。好的检测结果可以提高动态手势识别的精确度。

3) 搭建更加鲁棒的特征提取器。特征提取器是动态手势识别算法中的核心环节，鲁棒的特征提取器可以从序列中提取出更加高级、更加抽象的空间时序特征，直接决定了算法的性能。

4) 实现对连续动态手势的分割、识别。当前动态手势识别算法对单一动态手势已经取得了较好的识别效果，但是不适用于实际场景。为了解决该问题，在连续的动态手势算法中，实现对每个动态手势的开始点与结束点的精准定位、对每个分割出来的单一手势进行精确识别，是一个值得关注的热点问题。

7. 总结

本文对近年来基于视觉的动态手势识别方法进行了回顾，着重从输入数据模态、手部检测、动态手势建模三个方面分别介绍了研究现状，简要概述了该领域研究难点。虽然近年来动态手势识别算法已经取得了长足的发展，但仍存在较大的提升空间，包括以下三个方面：提高数据对动态手势细节的表达能力、采用手部检测器提高手部的精确检测、增强神经网络对动态手势的特征表达能力。

基金项目

企业委托项目资助(编号 HZY18010)。

参考文献

[1] Chen, L., *et al*. (2020) Survey of Pedestrian Action Recognition Techniques for Autonomous Driving. *Tsinghua Science and Technology*, **25**, 458-470.

[2] D'Sa, A.G. and Prasad, B.G. (2019) A Survey on Vision Based Activity Recognition, Its Applications and Challenges. 2019 *Second International Conference on Advanced Computational and Communication Paradigms* (*ICACCP*), Gangtok, 25-28 February 2019, 1-8. https://doi.org/10.1109/ICACCP.2019.8882896

[3] Wang, Z., *et al*. (2019) Hand Gesture Recognition Based on Active Ultrasonic Sensing of Smartphone: A Survey. *IEEE Access*, **7**, 111897-111922. https://doi.org/10.1109/ACCESS.2019.2933987

[4] Xia, Z., *et al*. (2019) Vision-Based Hand Gesture Recognition for Human-Robot Collaboration: A Survey. 2019 5*th International Conference on Control, Automation and Robotics* (*ICCAR*), Beijing, 19-22 April 2019, 198-205. https://doi.org/10.1109/ICCAR.2019.8813509

[5] Moin, A., Zhou, A., Benatti, S., Rahimi, A., Alexandrov, G., Menon, A., Tamakloe, S., Ting, J., Yamamoto, N., Khan, Y., Burghardt, F., Arias, A.C., Benini, L. and Rabaey, J.M. (2019) Adaptive EMG-Based Hand Gesture Recognition Using Hyperdimensional Computing.

[6] Liu, X., Shi, H., Hong, X., Chen, H., Tao, D. and Zhao, G. (2020) 3D Skeletal Gesture Recognition via Hidden States Exploration. *IEEE Transactions on Image Processing*, **29**, 4583-4597. https://doi.org/10.1109/TIP.2020.2974061

[7] Martínez, B.M., Modolo, D., Xiong, Y. and Tighe, J. (2019) Action Recognition with Spatial-Temporal Discriminative Filter Banks. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 5481-5490. https://doi.org/10.1109/ICCV.2019.00558

[8] Diba, A., Sharma, V., Van Gool, L. and Stiefelhagen, R. (2019) DynamoNet: Dynamic Action and Motion Network. 2019 *IEEE/CVF International Conference on Computer Vision* (ICCV), Seoul, 27-28 October 2019, 6191-6200. https://doi.org/10.1109/ICCV.2019.00629

[9] Feichtenhofer, C., Fan, H., Malik, J. and He, K. (2019) SlowFast Networks for Video Recognition. 2019 *IEEE/CVF International Conference on Computer Vision* (*ICCV*), Seoul, 27-28 October 2019, 6201-6210. https://doi.org/10.1109/ICCV.2019.00630

[10] Sun, J., Ji, T., Zhang, S., Yang, J. and Ji, G. (2018) Research on the Hand Gesture Recognition Based on Deep Learning. 2018 12*th International Symposium on Antennas, Propagation and EM Theory* (*ISAPE*), Hangzhou, 3-6 December 2018, 1-4. https://doi.org/10.1109/ISAPE.2018.8634348

[11] Guo, X., Xu, W., Tang, W.Q. and Wen, C. (2019) Research on Optimization of Static Gesture Recognition Based on Convolution Neural Network. 2019 4*th International Conference on Mechanical, Control and Computer Engineering* (*ICMCCE*), Hohhot, 25-27 October 2019, 398-3982. https://doi.org/10.1109/ICMCCE48743.2019.00095

[12] Sharma, P. and Anand, R.S. (2020) Depth Data and Fusion of Feature Descriptors for Static Gesture Recognition. *IET Image Processing*, **14**, 909-920. https://doi.org/10.1049/iet-ipr.2019.0230

[13] Lai, K. and Yanushkevich, S.N. (2018) CNN + RNN Depth and Skeleton Based Dynamic Hand Gesture Recognition. 2018 24*th International Conference on Pattern Recognition* (*ICPR*), Beijing, 20-24 August 2018, 3451-3456. https://doi.org/10.1109/ICPR.2018.8545718

[14] Kajan, S., Goga, J. and Zsíros, O. (2020) Comparison of Algorithms for Dynamic Hand Gesture Recognition. 2020 *Cybernetics & Informatics* (*K & I*), Velke Karlovice, 29 January-1 February 2020, 1-5. https://doi.org/10.1109/KI48306.2020.9039850

[15] Li, G., Wu, H., Jiang, G., Xu, S. and Liu, H. (2019) Dynamic Gesture Recognition in the Internet of Things. *IEEE Access*, **7**, 23713-23724. https://doi.org/10.1109/ACCESS.2018.2887223

[16] Bhowmick, S., Talukdar, A.K. and Sarma, K.K. (2015) Continuous Hand Gesture Recognition for English Alphabets. *IEEE Signal Processing and Integrated Networks* (*SPIN*-15), Noida, 19-20 February 2015, 443-446. https://doi.org/10.1109/SPIN.2015.7095264

[17] 陈甜甜, 姚璜, 左明章, 等. 基于深度信息的动态手势识别综述[J]. 计算机科学, 2018, 45(12): 49-58 + 83.

[18] 马力, 冯瑾. 基于 Leap Motion 的动态手势识别研究[J]. 计算机与数字工程, 2019, 47(1): 206-210.

[19] Krisandria, K.N., Dewantara, B.S.B. and Pramadihanto, D. (2019) HOG-Based Hand Gesture Recognition Using Kinect. 2019 *International Electronics Symposium* (*IES*), Surabaya, 27-28 September 2019, 254-259. https://doi.org/10.1109/ELECSYM.2019.8901607

[20] Wang, H., Wang, P., Song, Z., *et al*. (2017) Large-Scale Multimodal Gesture Recognition Using Heterogeneous Networks. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 3129-3137. https://doi.org/10.1109/ICCVW.2017.370

[21] Zhu, G., Zhang, L., Shen, P., *et al*. (2017) Multimodal Gesture Recognition Using 3-D Convolution and Convolutional LSTM. *IEEE Access*, **5**, 4517-4524. https://doi.org/10.1109/ACCESS.2017.2684186

[22] Köpüklü, O., Köse, N. and Rigoll, G. (2018) Motion Fused Frames: Data Level Fusion Strategy for Hand Gesture Recognition. *Proceedings of the CVPR Workshop on Analysis and Modeling of Faces and Gestures*, Salt Lake City, 18-22 June 2018, 2184-21848. https://doi.org/10.1109/CVPRW.2018.00284

[23] Li, Y., Miao, Q., Tian, K., *et al*. (2016) Large-Scale Gesture Recognition with a Fusion of RGB-D Data Based on the C3D Model. 2016 23*rd International Conference on Pattern Recognition* (*ICPR*), Cancún, 4-8 December 2016, 25-30. https://doi.org/10.1109/ICPR.2016.7899602

[24] Molchanov, P., Yang, X.D., Gupta, S., *et al*. (2016) Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks. *IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 4207-4215. https://doi.org/10.1109/CVPR.2016.456

[25] Zhang, L., Zhu, G., Shen, P., *et al*. (2017) Learning Spatiotemporal Features Using 3DCNN and Convolutional LSTM for Gesture Recognition. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 3120-3128. https://doi.org/10.1109/ICCVW.2017.369

[26] Miao, Q., Li, Y., Ouyang, W., *et al*. (2017) Multimodal Gesture Recognition Based on the ResC3D Network. 2017 *IEEE International Conference on Computer Vision Workshop*, Venice, 22-29 October 2017, 3047-3055. https://doi.org/10.1109/ICCVW.2017.360

[27] Li, Y., Miao, Q., Qi, X., *et al*. (2018) A Spatiotemporal Attention-Based ResC3D Model for Large-Scale Gesture Recognition. *Machine Vision and Applications*, **30**, 875-888. https://doi.org/10.1007/s00138-018-0996-x

[28] Zhu, G., Zhang, L., Mei, L., Shao, J., Song, J. and Shen, P. (2016) Large-Scale Isolated Gesture Recognition Using Pyramidal 3D Convolutional Networks. 23*rd International Conference on Pattern Recognition*, Cancún, 4-8 December 2016, 19-24. https://doi.org/10.1109/ICPR.2016.7899601

[29] Wan, J., Zhao, Y.B., Zhou, S., Guyon, I., Escalera, S. and Li, S.Z. (2016) ChaLearn Looking at People RGB-D Isolated and Continuous Datasets for Gesture Recognition. *CVPR Workshop*, Las Vegas, 26 June-1 July 2016, 761-769. https://doi.org/10.1109/CVPRW.2016.100

[30] Chai, X.J., Liu, Z.P., Yin, F., *et al*. (2016) Two Streams Recurrent Neural Networks for Large-Scale Continuous Gesture Recognition. 2016 23*rd International Conference on Pattern Recognition* (*ICPR*), Cancún, 4-8 December 2016, 31-36. https://doi.org/10.1109/ICPR.2016.7899603

[31] Narayana, P., Beveridge, J.R. and Draper, B.A. (2018) Gesture Recognition: Focus on the Hands. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Salt Lake City, 18-22 June 2018, 5235-5244. https://doi.org/10.1109/CVPR.2018.00549

[32] Mcbride, T.J., Vandayar, N. and Nixon, K.J. (2019) A Comparison of Skin Detection Algorithms for Hand Gesture Recognition. *Southern African Universities Power Engineering Conference/Robotics & Mechatronics/Pattern Recognition Association of South Africa*, Bloemfontein, 28-30 January 2019, 211-216. https://doi.org/10.1109/RoboMech.2019.8704839

[33] Belongie, S.J., Malik, J. and Puzicha, J. (2010) Shape Matching and Object Recognition Using Shape Contexts. *IEEE International Conference on Computer Science and Information Technology*, Chengdu, 9-11 July 2010, 483-507.

[34] 何胜皎. 视频序列中运动目标检测算法的研究[D]: [硕士学位论文]. 兰州: 兰州理工大学, 2018.

[35] Zimmermann, C. and Brox, T. (2017) Learning to Estimate 3D Hand Pose from Single RGB Images. 2017 *IEEE International Conference on Computer Vision* (*ICCV*), Venice, 22-29 October 2017, 4913-4921. https://doi.org/10.1109/ICCV.2017.525

[36] Ren, S.Q., He, K.M., Girshick, R. and Sun, J. (2016) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **39**, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031

[37] Liu, Z.P., *et al*. (2017) Continuous Gesture Recognition with Hand-Oriented Spatiotemporal Feature. *Proceedings of the IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 3056-3064. https://doi.org/10.1109/ICCVW.2017.361

[38] Gulati, S. and Bhogal, R.K. (2018) Comprehensive Review of Various Hand Detection Approaches. 2018 *International Conference on Circuits and Systems in Digital Enterprise Technology* (*ICCSDET*), Kottayam, 21-22 December 2018, 1-5. https://doi.org/10.1109/ICCSDET.2018.8821238

[39] Zhao, S., Yang, W. and Wang, Y. (2018) A New Hand Segmentation Method Based on Fully Convolutional Network.

2018 *Chinese Control and Decision Conference* (*CCDC*), Shenyang, 9-11 June 2018, 5966-5970. https://doi.org/10.1109/CCDC.2018.8408176

[40] Le, T., Jaw, D., Lin, I., Liu, H. and Huang, S. (2018) An Efficient Hand Detection Method Based on Convolutional Neural Network. 2018 7*th International Symposium on Next Generation Electronics* (*ISNE*), Taipei, 7-9 May 2018, 1-2. https://doi.org/10.1109/ISNE.2018.8394651

[41] Huu, P.N. and The, H.L. (2019) Proposing Recognition Algorithms for Hand Gestures Based on Machine Learning Model. 2019 19*th International Symposium on Communications and Information Technologies* (*ISCIT*), Ho Chi Minh City, 25-27 September 2019, 496-501. https://doi.org/10.1109/ISCIT.2019.8905194

[42] Zhan, F. (2019) Hand Gesture Recognition with Convolution Neural Networks. 2019 *IEEE* 20*th International Conference on Information Reuse and Integration for Data Science* (*IRI*), Los Angeles, CA, 30 July-1 August 2019, 295-298. https://doi.org/10.1109/IRI.2019.00054

[43] Du, T., Ren, X. and Li, H. (2018) Gesture Recognition Method Based on Deep Learning. 2018 33*rd Youth Academic Annual Conference of Chinese Association of Automation* (*YAC*), Nanjing, 18-20 May 2018, 782-787. https://doi.org/10.1109/YAC.2018.8406477

[44] Hong, J.Y., Park, S.H. and Baek, J. (2019) Segmented Dynamic Time Warping Based Signal Pattern Classification. 2019 *IEEE International Conference on Computational Science and Engineering* (*CSE*) *and IEEE International Conference on Embedded and Ubiquitous Computing* (*EUC*), New York, 1-3 August 2019, 263-265. https://doi.org/10.1109/CSE/EUC.2019.00058

[45] Plouffe, G. and Cretu, A.-M. (2015) Static and Dynamic Hand Gesture Recognition in Depth Data Using Dynamic Time Warping. *IEEE Transactions on Instrumentation & Measurement*, **65**, 305-316. https://doi.org/10.1109/TIM.2015.2498560

[46] Fine, S., Singer, Y. and Tishby, N. (1998) The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine Learning*, **32**, 41-62. https://doi.org/10.1023/A:1007469218079

[47] Haid, M., Budaker, B., Geiger, M., Husfeldt, D., Hartmann, M. and Berezowski, N. (2019) Inertial-Based Gesture Recognition for Artificial Intelligent Cockpit Control Using Hidden Markov Models. 2019 *IEEE International Conference on Consumer Electronics* (*ICCE*), Las Vegas, 11-13 January 2019, 1-4. https://doi.org/10.1109/ICCE.2019.8662036

[48] Corradini, A. (2001) Dynamic Time Warping for Off-Line Recognition of a Small Gesture Vocabulary. *Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems*, Vancouver, 13 July 2001, 82-89. https://doi.org/10.1109/RATFG.2001.938914

[49] 张建荣. 基于 Kinect 手势识别的虚拟环境体感交互技术研究[D]: [硕士学位论文]. 重庆: 重庆邮电大学, 2016.

[50] Saha, S., Lahiri, R., Konar, A., *et al*. (2017) HMM-Based Gesture Recognition System Using Kinect Sensor for Improvised Human-Computer Interaction. 2017 *International Joint Conference on Neural Networks* (*IJCNN*), Anchorage, 14-19 May 2017, 2776-2783. https://doi.org/10.1109/IJCNN.2017.7966198

[51] Simonyan, K. and Zisserman, A. (2014) Two-Stream Convolutional Networks for Action Recognition in Videos. *Advances in Neural Information Processing Systems*, Vol. 1, 568-576.

[52] Wang, L., Xiong, Y., Wang, Z., *et al*. (2016) Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. *European Conference on Computer Vision*, Springer, Cham, 20-36. https://doi.org/10.1007/978-3-319-46484-8_2

[53] Feichtenhofer, C., Pinz, A. and Zisserman, A. (2016) Convolutional Two-Stream Network Fusion for Video Action Recognition. *CVPR* 2016, Las Vegas, 27-30 June 2016, 1933-1941. https://doi.org/10.1109/CVPR.2016.213

[54] Zhu, Y., Lan, Z., Newsam, S., *et al*. (2017) Hidden Two-Stream Convolutional Networks for Action Recognition. *ACCV* 2018, Perth, 2-6 December 2018, 363-378.

[55] Ji, S., Xu, W., Yang, M., *et al*. (2013) 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 221-231. https://doi.org/10.1109/TPAMI.2012.59

[56] Soomro, K., Zamir, A.R. and Shah, M. (2012) UCF101: A Dataset of 101 Human Actions Classes from Videos in the Wild. *CoRR*, **12**, 1-7. https://arxiv.org/abs/1212.0402

[57] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T. and Serre, T. (2011) HMDB: A Large Video Database for Human Motion Recognition. *Proceedings of the International Conference on Computer Vision* (*ICCV*), Barcelona, 6-13 November 2011, 2556-2563. https://doi.org/10.1109/ICCV.2011.6126543

[58] Szegedy, C., Vanhoucke, V., Ioffe, S., *et al*. (2016) Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 2818-2826. https://doi.org/10.1109/CVPR.2016.308

[59] Tran, D., Bourdev, L., Fergus, R., *et al*. (2015) Learning Spatiotemporal Features with 3D Convolutional Networks.

Proceedings of the IEEE International Conference on Computer Vision, Santiago, 7-13 December 2015, 4489-4497. https://doi.org/10.1109/ICCV.2015.510

[60] Tran, D., Ray, J., Shou, Z., *et al*. (2017) Convnet Architecture Search for Spatiotemporal Feature Learning. arXiv preprint arXiv:1708.05038

[61] Kabir, R., Ahmed, N., Roy, N. and Islam, M.R. (2019) A Novel Dynamic Hand Gesture and Movement Trajectory Recognition model for Non-Touch HRI Interface. 2019 *IEEE Eurasia Conference on IOT, Communication and Engineering* (*ECICE*), Yunlin, 3-6 October 2019, 505-508. https://doi.org/10.1109/ECICE47484.2019.8942691

[62] Rahman, A., Mahmud, J.A. and Hasanuzzaman, M. (2018) Pointing and Commanding Gesture Recognition in 3D for Human-Robot Interaction. 2018 *International Conference on Innovation in Engineering and Technology* (*ICIET*), Dhaka, 6-8 January 2018, 1-10. https://doi.org/10.1109/CIET.2018.8660913

[63] Zhang, X. and Wu, X. (2019) Robotic Control of Dynamic and Static Gesture Recognition. 2019 2*nd World Conference on Mechanical Engineering and Intelligent Manufacturing* (*WCMEIM*), Shanghai, 22-24 November 2019, 474-478. https://doi.org/10.1109/WCMEIM48965.2019.00100

[64] Hu, K., Yin, L. and Wang, T. (2019) Temporal Interframe Pattern Analysis for Static and Dynamic Hand Gesture Recognition. 2019 *IEEE International Conference on Image Processing* (*ICIP*), Taipei, 22-25 September 2019, 3422-3426. https://doi.org/10.1109/ICIP.2019.8803472