(19)国家知识产权局

(12)发明专利

(10)授权公告号 CN 115223239 B
(45)授权公告日 2024.05.07

(21)申请号 202210717218.3
(22)申请日 2022.06.23
(65)同一申请的已公布的文献号
申请公布号 CN 115223239 A
(43)申请公布日 2022.10.21
(73)专利权人 山东科技大学
地址 266590 山东省青岛市黄岛区前湾港路579号
(72)发明人 曾庆田 宋戈 王通 段华 曲祥雯
(74)专利代理机构 青岛锦佳专利代理事务所(普通合伙) 37283
专利代理师 朱玉建
(51)Int.Cl.
G06V 40/20(2022.01)
G06N 3/0464(2023.01)
G06N 3/08(2023.01)
G06V 10/40(2022.01)
G06V 10/774(2022.01)
G06V 10/80(2022.01)
G06V 10/764(2022.01)

(56)对比文件
CN 109033978 A,2018.12.18
CN 112836597 A,2021.05.25
CN 113255602 A,2021.08.13
CN 114120350 A,2022.03.01
CN 114360067 A,2022.04.15

审查员 周庆成

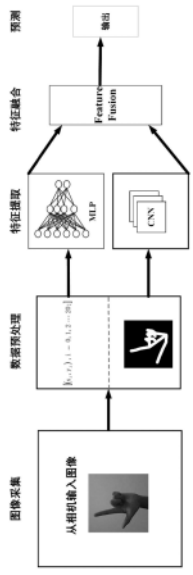
权利要求书4页 说明书11页 附图4页

(54)发明名称

一种手势识别方法、系统、计算机设备以及可读存储介质

(57)摘要

本发明属于人机交互技术领域,具体公开了一种手势识别方法、系统、计算机设备以及可读存储介质。该方法通过创建一个基于MLP和CNN的手势识别模型,便于使用手势图片以及手部关键点特征数据作为混合输入,使得机器学习模型能够从手势图片以及手部关键点数据中获取和识别特征,本发明模型为通过输入手势图像和对应手部关键点特征数据来识别手势的多输入融合深度网络模型,该手势识别模型充分结合了MLP和CNN两种不同网络以及手势图片和手部关键点特征数据两种数据的优点,以提高手势识别网络的整体性能,有效地解决了当前手势识别中精度低、实时性差、鲁棒性差的问题,在模型中同时输入手势图片和手部关键点特征数据,获得了较高的手势识别精度。



CN 115223239 B

1.一种基于MLP和CNN的多输入融合深度网络的手势识别方法,其特征在于,

包括如下步骤:

步骤1.获取原始手势图像数据,并构建原始手势图像数据集;

步骤2.对原始手势图像数据集中各幅原始手势图像数据进行预处理,分别提取每幅原始手势图像中所包含的21个手部关键点的特征数据以及手势图片数据;

将从每幅原始手势图像中提取到的21个手部关键点的特征数据和手势图片数据,以及每幅原始手势图像对应的标签,共同组成一组样本数据;

将所有原始手势图像对应的样本数据组成样本数据集,并分为训练数据集和测试数据集;

步骤3.搭建多输入融合深度网络模型;

多输入融合深度网络包括特征提取模块、特征融合模块以及分类模块;

所述特征提取模块包括两个分支网络,分别是针对21个手部关键点特征提取的MLP分支网络以及针对手势图片特征提取的CNN分支网络;

两个分支网络的输出分别与特征融合模块相连,特征融合模块与分类模块相连;

步骤4.训练及测试多输入融合深度网络模型;

利用步骤2中训练数据集中的样本数据训练多输入融合深度网络;

其中,MLP分支网络的输入为21个手部关键点的特征数据,MLP分支网络的输出为对应于手部关键点的特征数据的第一特征向量;

CNN分支网络的输入为手势图片,CNN分支网络的输出为第二特征向量;

特征融合模块用于将第一、第二特征向量组合起来,并经过分类模块预测输出预测结果;

利用测试数据集中的样本数据对训练好的多输入融合深度网络进行测试;

步骤5.对于待识别的手势图像,提取图像包含的21个手部关键点的特征数据以及手势图片数据,利用训练及测试好的多输入融合深度网络进行手势识别,得到识别结果。

2.根据权利要求1所述的多输入融合深度网络的手势识别方法,其特征在于,

所述步骤1具体为:

获取不同光照强度以及不同背景下捕捉的手势图像,剔除其中模糊不清的手势图像,将收集好的手势图像进行分类打标签,建立原始手势图像数据集。

3.根据权利要求1所述的多输入融合深度网络的手势识别方法,其特征在于,

所述步骤2中,手部关键点的特征数据的提取方法如下:

首先利用目标检测框架从原始手势图像中检测识别手部的21个手部关键点像素坐标;

对原始数据集采用欧几里得距离归一化处理,具体步骤如下:

定义手腕位置对应的手部关键点为基准手部关键点并将其设为原点,其余20个手部关键点与原点间的横、纵轴方向的距离绝对值作为对应手部关键点的新坐标;

分别计算各个手部关键点的新坐标到原点坐标的欧几里得距离,如公式(1)所示;

$$\rho_i = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2} \qquad (1)$$

其中,i=0,1,…,19,20;

ρ_i表示第i个手部关键点的新坐标(x_i, y_i)与原点坐标(x_0, y_0)之间的欧几里得距离;

根据公式(1)中得到的ρ_i,由公式(2)进一步归一化处理;

$k_i=(\rho_i-\mu)/\sigma$ (2)

其中,k_i为手部第i个手部关键点经过欧几里得归一化处理后的数值,即手部关键点特征数据;μ、σ分别表示21个手部关键点经欧式距离处理后的均值和标准差;

μ、σ的计算方式如公式(3)、公式(4)所示;

$$\mu=\sum_{i=0}^{n}\rho_i/(n+1) \quad (3)$$

$$\sigma=\sum_{i=0}^{n}(\rho_i-\mu)^2/(n+1) \quad (4)$$

其中,n取值为20。

4.根据权利要求1所述的多输入融合深度网络的手势识别方法,其特征在于,

所述步骤2中,手势图片数据的提取过程如下:

利用图像分割技术显示目标检测框架检测的手势关键点及轮廓,然后进行膨胀操作,接着去除杂乱背景,并将图片调整为统一尺寸大小,完成对原始手势图像的预处理。

5.根据权利要求1所述的多输入融合深度网络的手势识别方法,其特征在于,

所述步骤3中,CNN分支网络的结构如下:

CNN分支网络包含九层网络结构,分别是一个输入层、三个卷积层、三个最大池化层以及两个全连接层;其中,各层网络结构的连接结构分别如下:

定义三个卷积层分别为第一卷积层、第二卷积层以及第三卷积层;

定义三个最大池化层分别为第一最大池化层、第二最大池化层以及第三最大池化层;

定义两个全连接层分别为第一全连接层以及第二全连接层;

其中,输入层、第一卷积层、第一最大池化层、第二卷积层、第二最大池化层、第三卷积层、第三最大池化层、第一全连接层以及第二全连接层依次连接;

输入层的输入为预处理后的手势图片,输入尺寸大小为64×64×3;

第一卷积层、第二卷积层以及第三卷积层分别包含16、32、64个滤波器,第一卷积层、第二卷积层以及第三卷积层的卷积核的大小均为3×3;

第一最大池化层、第二最大池化层以及第三最大池化层采用最大池化,设置步长为2;

第一全连接层的神经元个数为32,第二全连接层的神经元个数为类别的数量;

所述步骤3中,MLP分支网络的结构如下:

MLP分支网络由三层全连接层构成;

定义三层全连接层分别为第三全连接层、第四全连接层以及第五全连接层;则第三全连接层、第四全连接层以及第五全连接层依次连接;

第三全连接层作为MLP分支网络的输入层,包含21个神经元,输入为预处理后得到的21个手部关键点特征数据;第四全连接层为隐藏层,包含16个神经元;

第五全连接层为MLP分支网络的输出层,神经元的个数设置为类别的数量。

6.根据权利要求5所述的多输入融合深度网络的手势识别方法,其特征在于,

所述步骤3中,定义第一特征向量为T_{out},第二特征向量为J_{out};

特征融合模块对两个分支网络提取的特征信息合理融合,引入自适应的特征权重ω_1、ω_2,使模型根据数据的特征分布来自行决定权重参数,在特征融合模块以不同的权重来融合特征;

融合的手势特征C_f使用公式(5)计算得来：

$$C_f = \omega_1 * T_{out} \oplus \omega_2 * J_{out} \quad (5)$$

其中,⊕代表Sum Fusion融合方式,权重ω_1、ω_2由公式(6)得到：

$$\omega_i = \frac{e^{\alpha_i}}{\sum_j e^{\alpha_j}} \quad (6)$$

其中,i＝1,2,j＝1,2,ω_i为归一化的权重,且$\Sigma\omega_i$＝1,α_i为初始化的权重参数；

分类模块包括两个全连接层；

定义两个全连接层分别为第六全连接层以及第七全连接层；

最终经过特征融合模块融合后的手势特征C_f经过两层全连接层完成最终的分类；

其中,第六全连接层包含32个神经元；

第七全连接层作为输出层,使用Softmax分类函数,神经元的个数设置为类别的数量。

7.根据权利要求1所述的多输入融合深度网络的手势识别方法,其特征在于,

所述步骤4中,多输入融合深度网络的训练过程如下：

将步骤2得到的21个手部关键点的特征数据作为分支网络MLP的输入,经过三层全连接层进行特征提取,得到一个第一特征向量,标记为特征向量T_{out}；

将步骤2得到的手势图片数据作为分支网络CNN的输入,经过CNN网络提取特征后,得到一个第二特征向量,标记为输出特征向量J_{out}；

通过特征融合模块使用自适应特征融合的方法将两个分支网络的输出特征向量组合起来,然后进一步经过全连接神经网络使用softmax分类器进行预测分类；

在训练过程中使用Dropout防止过拟合,使得模型收敛速度加快；

使用分类交叉熵损失函数,计算方法如公式(2)所示：

$$LOSS = \sum_{i=1}^{m} y_i \cdot log\hat{y}_i \quad (2)$$

其中,m是手势类别的数量,$\hat{y}_i$表示模型的预测输出,y_i表示真实的标签；

设定模型训练次数,模型使用Adam优化器进行训练；

根据预测输出与对应的分类标签进行计算,得出分类损失函数Loss值;当Loss值不再下降时停止训练更新,保存模型以及权重参数；

最后从保存的模型权重中读取参数,得到训练好的多输入融合深度网络模型。

8.一种基于MLP和CNN的多输入融合深度网络的手势识别系统,其特征在于,包括：

图像采集模块,用于获取原始手势图像数据并构建原始手势图像数据集；

数据预处理模块,用于对原始手势图像数据集中各幅原始手势图像数据进行预处理,分别提取每幅原始手势图像中所包含的21个手部关键点的特征数据以及手势图片数据；

将从每幅原始手势图像中提取到的21个手部关键点的特征数据和手势图片数据,以及每幅原始手势图像对应的标签,共同组成一组样本数据；

将所有原始手势图像对应的样本数据组成样本数据集,并分为训练数据集和测试数据集；

模型搭建及训练测试模块,用于搭建、训练以及测试多输入融合深度网络模型；

多输入融合深度网络包括特征提取模块、特征融合模块以及分类模块；

所述特征提取模块包括两个分支网络,分别是针对21个手部关键点特征提取的MLP分

支网络以及针对手势图片特征提取的CNN分支网络;

两个分支网络的输出分别与特征融合模块相连,特征融合模块与分类模块相连;

利用训练数据集中的样本数据训练多输入融合深度网络;

其中,MLP分支网络的输入为21个手部关键点的特征数据,MLP分支网络的输出为对应于手部关键点的特征数据的第一特征向量;

CNN分支网络的输入为手势图片,CNN分支网络的输出为第二特征向量;

特征融合模块用于将第一、第二特征向量组合起来,并经过分类模块预测输出预测结果;

利用测试数据集中的样本数据对训练好的多输入融合深度网络进行测试;

预测模块,对于待识别的手势图像,用于提取图像包含的21个手部关键点的特征数据以及手势图片数据,利用训练及测试好的多输入融合深度网络进行手势识别得到识别结果。

9.一种计算机设备,包括存储器和一个或多个处理器,所述存储器中存储有可执行代码,其特征在于,所述处理器执行所述可执行代码时,

实现如权利要求1至7任一项所述的多输入融合深度网络的手势识别方法。

10.一种计算机可读存储介质,其上存储有程序,其特征在于,该程序被处理器执行时,实现如权利要求1至7任一项所述的多输入融合深度网络的手势识别方法。

一种手势识别方法、系统、计算机设备以及可读存储介质

技术领域

[0001] 本发明属于手势识别技术领域,涉及一种手势识别方法、系统、计算机设备以及可读存储介质。

背景技术

[0002] 手势识别是人机交互和计算机视觉领域研究的热点,如虚拟现实、智能控制、娱乐游戏以及手语翻译等。手势识别应用领域的一个基本特征是实时性,因此,手势识别系统必须在用户输入手势的情况下提供实时结果。然而,由于设备条件、光照效果和背景的复杂程度不同,手势识别仍具挑战性。为了更好地实现人机交互,手势识别算法应该在各种光照强度、背景等复杂环境中具有良好的实时识别能力。目前手势识别方法主要分为两大类:

[0003] 一是基于传感器的手势识别方法。此类方法的优点是手势识别不会被不同的背景分散注意力,但会造成佩戴笨重、不灵活且成本高的问题,违背了人机自然交互的初衷。

[0004] 二是基于视觉的手势识别方法,此类方法需要通过摄像头获取手势的图像或视频。相比基于传感器的手势识别,基于视觉的手势识别系统,能够使用较低成本的摄像头可以让用户更自然地与计算机设备进行交互。在基于视觉的手势识别方法中,最常用的手势提取方法包括肤色检测、背景减法、边界建模、轮廓、手势分割以及手形估计等。

[0005] 然而,这些传统的识别方法在进行手势识别过程中存在一些不足之处,如算法的鲁棒性不强,模型对数据集的依赖性大,样本数据受环境等因素影响,例如光照变化、背景问题、距离范围和多手势等问题,导致手势特征不明显,神经网络模型识别率低。

[0006] 可见,基于视觉的手势识别方法,手势图像的预处理成为一个需要解决的问题。

[0007] 随着深度学习算法的飞速发展,如YOLO(you only look once)(Redmon等,2016;Redmon和Farhadi,2017,2018)、SSD(single shot multibox detector)(Liu等,2016)、RCNN(region convolutional neural network)(Girshick等,2014)和Faster R-CNN(Ren等,2015)等算法在目标检测和分类问题中取得了较高的准确率,然而这些算法往往通过设计更深层次的网络结构来提取更多的深度特征,对硬件的计算能力和存储能力的要求很高,这些检测模型普遍存在模型较大和检测时间长等问题,难以在嵌入式设备中普及,也不能满足许多场合中对于实时性的要求。

发明内容

[0008] 本发明的目的之一在于提出一种基于MLP和CNN的多输入融合深度网络的手势识别方法,以提高各种光照强度、背景等复杂环境下手势识别的准确性。

[0009] 本发明为了实现上述目的,采用如下技术方案:

[0010] 一种基于MLP和CNN的多输入融合深度网络的手势识别方法,包括如下步骤:

[0011] 步骤1.获取原始手势图像数据,并构建原始手势图像数据集;

[0012] 步骤2.对原始手势图像数据集中各幅原始手势图像数据进行预处理,分别提取每幅原始手势图像中所包含的21个手部关键点的特征数据以及手势图片数据;

[0013] 将从每幅原始手势图像中提取到的21个手部关键点的特征数据和手势图片数据,以及每幅原始手势图像对应的标签,共同组成一组样本数据;

[0014] 将所有原始手势图像对应的样本数据组成样本数据集,并分为训练数据集和测试数据集;

[0015] 步骤3.搭建多输入融合深度网络模型;

[0016] 多输入融合深度网络包括特征提取模块、特征融合模块以及分类模块;

[0017] 所述特征提取模块包括两个分支网络,分别是针对21个手部关键点特征提取的MLP分支网络以及针对手势图片特征提取的CNN分支网络;

[0018] 其中,两个分支网络的输出分别与特征融合模块相连,特征融合模块与分类模块相连;

[0019] 步骤4.训练及测试多输入融合深度网络模型;

[0020] 利用步骤2中训练数据集中的样本数据训练多输入融合深度网络;

[0021] 其中,MLP分支网络的输入为21个手部关键点的特征数据,MLP分支网络的输出为对应于手部关键点的特征数据的第一特征向量;

[0022] CNN分支网络的输入为手势图片,CNN分支网络的输出为第二特征向量;

[0023] 特征融合模块用于将第一、第二特征向量组合起来,并经过分类模块预测输出预测结果;

[0024] 利用测试数据集中的样本数据对训练好的多输入融合深度网络进行测试;

[0025] 步骤5.对于待识别的手势图像,提取图像包含的21个手部关键点的特征数据以及手势图片数据,利用训练及测试好的多输入融合深度网络进行手势识别,得到识别结果。

[0026] 此外,本发明还提出了一种与上述基于MLP和CNN的多输入融合深度网络的手势识别方法相对应的基于MLP和CNN的多输入融合深度网络的手势识别系统,其技术方案如下:

[0027] 一种基于MLP和CNN的多输入融合深度网络的手势识别系统,包括:

[0028] 图像采集模块,用于获取原始手势图像数据并构建原始手势图像数据集;

[0029] 数据预处理模块,用于对原始手势图像数据集中各幅原始手势图像数据进行预处理,分别提取每幅原始手势图像中所包含的21个手部关键点的特征数据以及手势图片数据;

[0030] 将从每幅原始手势图像中提取到的21个手部关键点的特征数据和手势图片数据,以及每幅原始手势图像对应的标签,共同组成一组样本数据;

[0031] 将所有原始手势图像对应的样本数据组成样本数据集,并分为训练数据集和测试数据集;

[0032] 模型搭建及训练测试模块,用于搭建、训练以及测试多输入融合深度网络模型;

[0033] 多输入融合深度网络包括特征提取模块、特征融合模块以及分类模块;

[0034] 所述特征提取模块包括两个分支网络,分别是针对21个手部关键点特征提取的MLP分支网络以及针对手势图片特征提取的CNN分支网络;

[0035] 其中,两个分支网络的输出分别与特征融合模块相连,特征融合模块与分类模块相连;

[0036] 利用训练数据集中的样本数据训练多输入融合深度网络;

[0037] 其中,MLP分支网络的输入为21个手部关键点的特征数据,MLP分支网络的输出为

对应于手部关键点的特征数据的第一特征向量；

[0038] CNN分支网络的输入为手势图片,CNN分支网络的输出为第二特征向量；

[0039] 特征融合模块用于将第一、第二特征向量组合起来,并经过分类模块预测输出预测结果；

[0040] 利用测试数据集中的样本数据对训练好的多输入融合深度网络进行测试；

[0041] 预测模块,对于待识别的手势图像,用于提取图像包含的21个手部关键点的特征数据以及手势图片数据,利用训练及测试好的多输入融合深度网络进行手势识别得到识别结果。

[0042] 此外,本发明还提出了一种与上述多输入融合深度网络的手势识别方法相对应的计算机设备,该计算机设备包括存储器和一个或多个处理器。

[0043] 所述存储器中存储有可执行代码,所述处理器执行所述可执行代码时,用于实现上面述及的基于MLP和CNN的多输入融合深度网络的手势识别方法。

[0044] 此外,本发明还提出了一种与上述多输入融合深度网络的手势识别方法相对应的计算机可读存储介质,其上存储有程序;该程序被处理器执行时,

[0045] 用于实现上面述及的基于MLP和CNN的多输入融合深度网络的手势识别方法。

[0046] 本发明具有如下优点：

[0047] 如上所述,本发明述及了一种多输入融合深度网络的手势识别方法,该方法通过创建一个基于MLP和CNN的手势识别模型,便于使用手势图片以及手部关键点特征数据作为混合输入,使得机器学习模型能够从手势图片以及手部关键点数据中获取和识别特征,本发明手势识别模型为通过输入手势图像和一些手部关键点特征数据来识别手势的多输入融合深度网络模型,该手势识别模型充分结合了MLP和CNN两种不同网络的优点,以提高手势识别网络的整体性能,有效地解决了当前手势识别中精度低、实时性差、鲁棒性差的问题,通过在模型中同时输入手势图片和手部关键点特征数据,获得了较高的手势识别精度。从应用对象看,本发明方法可以应用在单目相机采集的图像,所需设备简单方便,应用场景更为广泛。

附图说明

[0048] 图1为本发明实施例中基于MLP和CNN双分支特征融合的手势识别方法的流程图；

[0049] 图2为本发明实施例中基于MLP和CNN双分支特征融合的手势识别方法的模型框图；

[0050] 图3为本发明实施例中提取的21个手部关键点位置分布图；

[0051] 图4为本发明实施例中提取的手势图像示意图；

[0052] 图5为本发明实施例中特征融合模块进行特征融合的示意图。

具体实施方式

[0053] 下面结合附图以及具体实施方式对本发明作进一步详细说明：

[0054] 本实施例述及了一种基于MLP和CNN的多输入融合深度网络的手势识别方法,以解决当前手势识别中存在的精度低、实时性差、鲁棒性差的技术问题。

[0055] 如图1所示,基于MLP和CNN的多输入融合深度网络的手势识别方法,包括如下步

骤：

[0056] 步骤1.获取原始手势图像数据，并构建原始手势图像数据集。

[0057] 本实施例中所使用的原始手势图像数据集从实际人机交互场景中收集而来。

[0058] 在实际生活场景中使用摄像头获取不同光照强度以及不同背景下捕捉的手势图像，剔除模糊不清的手势图像，将收集好的手势图像进行分类打标签，建立原始手势图像数据集。

[0059] 本实施例中总共收集了12种不同的手势，即类别总共有12种，例如握拳手势、伸大拇指手势、OK手势、五指全部伸开手势等等，代表了平时常用到的几种手势。

[0060] 其中，每种手势包含1000张不同背景下的图片。

[0061] 基于此，本实施例中建立了一个拥有12000张手势图片的原始手势图像数据集。

[0062] 当然，本实施例中标签的数量或类别的数量并不局限于以上12种，例如，还可以根据适用场景的不同灵活增加一些手势，即扩充标签或类别的数量。

[0063] 步骤2.对原始手势图像数据集中各幅原始手势图像数据进行预处理，分别提取每幅原始手势图像中所包含的21个手部关键点的特征数据以及手势图片数据。

[0064] 表1示出了从一幅原始手势图像中提取21个手部关键点的特征数据的过程。

[0065] 其中，手部关键点的特征数据的提取方法如下：

[0066] 首先利用目标检测框架(如openpose，mediapipe)，从原始手势图像中实时准确地检测识别手部的21个手部关键点像素坐标，并输出每个关键点的像素坐标，如表1第2行所示。

[0067] 识别出的21个手部关键点的具体位置如图3所示，各个手部关键点分别用0-20替代。

[0068] 其中，0表示手腕，1表示拇指的CMC关节位，2表示拇指的MCP关节位，3表示拇指的IP关节位，4表示拇指的TIP关节位，5表示食指的MCP关节位，6表示食指的PIP关节位，7表示食指的DIP关节位，8表示食指的TIP关节位，9表示中指的MCP关节位，10表示中指的PIP关节位，11表示中指的DIP关节位，12表示中指的TIP关节位，13表示无名指的MCP关节位，14表示无名指的PIP关节位，15表示无名指的DIP关节位，16表示无名指的TIP关节位，17表示小指的MCP关节位，18表示小指的PIP关节位，19表示小指的DIP关节位，20表示小指的TIP关节位。

[0069] 由于同一手势在手势图像中的不同位置、不同距离等因素影响下，得到的像素坐标是不同的，因此为了消除此影响，对原始手势图像采用欧几里得距离归一化处理。

[0070] 欧几里得距离归一化处理的具体步骤如下：

[0071] 定义手腕位置对应的手部关键点为基准手部关键点并将其设为原点，其余20个手部关键点与原点间的横、纵轴方向的距离绝对值作为对应手部关键点的新坐标，如表2第3行所示。

[0072] 分别计算各个手部关键点的新坐标到原点坐标的欧几里得距离，如公式(1)所示。

[0073] $$\rho_i = \sqrt{(x_i - x_0)^2 + (y_i - y_0)^2} \qquad (1)$$

[0074] 其中，i＝0,1,…,19,20。

[0075] ρ_i表示第i个手部关键点的新坐标(x_i,y_i)与原点坐标(x_0,y_0)之间的欧几里得距离。

[0076] 计算出来的各个手部关键点的ρ_i值如表1中第4行所示。

[0077] 根据公式(1)中得到的ρ_i,由公式(2)进一步归一化处理。

[0078] $k_i=(\rho_i-\mu)/\sigma$ (2)

[0079] 其中,k_i为手部第i个手部关键点经过欧几里得归一化处理后的数值,即手部关键点特征数据,各个手部关键点的ρ_i值如表1中第5行所示。

[0080] μ、σ分别表示21个手部关键点经欧式距离处理后的均值和标准差。

[0081] μ、σ的计算方式如公式(3)、公式(4)所示。

[0082] $$\mu=\sum_{i=0}^{n}\rho_i/(n+1) \quad (3)$$

[0083] $$\sigma=\sum_{i=0}^{n}(\rho_i-\mu)^2/(n+1) \quad (4)$$

[0084] 其中,由于手部关键点的数量为21个,因此此处n取值为20。

[0085] 表1

[0086]

手部关键点	0	1	2	…	18	19	20
原始坐标	[74,322]	[140,312]	[190,278]	…	[60,142]	[54,110]	[47,82]
①	[0,0]	[66,-10]	[116,-44]	…	[-14,-180]	[-20,-212]	[-27,-240]
②	0.00	66.75	124.07	…	180.54	212.94	241.51
k_i	-2.74	-1.75	-0.90	…	-0.06	0.42	0.84

[0087] 通过欧几里得距离归一化处理,将相同手势在不同位置、不同距离下的关键点像素坐标转换为无单位的数值,使得数据标准统一化,提高了数据可比性。

[0088] 接着从同一幅原始手势图像中提取手势图片数据,提取过程如下:

[0089] 利用图像分割显示目标检测框架(如openpose,mediapipe)检测的手势关键点及轮廓,然后进行膨胀操作,接着去除杂乱背景,并将手势图片调整为统一尺寸大小,例如64×64。

[0090] 通过以上过程完成了对手势图片的提取,提取后的手势图像如图4所示。

[0091] 将从每幅原始手势图像中提取到的21个手部关键点的特征数据和手势图片数据,以及每幅原始手势图像对应的标签,共同组成一组样本数据。

[0092] 将所有原始手势图像对应的样本数据组成样本数据集,并分为训练数据集和测试数据集;其中训练数据集用于模型的训练,测试数据集用于模型的性能测试。

[0093] 本实施例中的手势识别方法,用户无需佩戴任何辅助性设备或者其他标记物,将目标检测框架识别出来的手部关键点特征数据和对应手势图片数据结合起来,以获得更多更准确的特征信息,本发明方法在不同的人机交互场景下具有良好的识别鲁棒性和实时性。

[0094] 步骤3.搭建多输入融合深度网络模型MIFD-Net(Multi-input fusion deep network)。如图2所示,多输入融合深度网络包括特征提取模块、特征融合模块以及分类模块。

[0095] 本实施例中特征提取模块包括两个分支网络,分别是针对21个手部关键点特征提取的MLP分支网络以及针对手势图片特征提取的CNN分支网络。

[0096] 在设计的MIFD-Net模型中,使用CNN提取手势图片特征信息的共包含九层网络:

[0097] 本实施例CNN分支网络的输入是预处理后的手势图片,输入尺寸大小为64×64×

3,包含一个输入层、三个卷积层、三个最大池化层以及两个全连接层。

[0098] 定义三个卷积层分别为第一卷积层Conv1、第二卷积层Conv2以及第三卷积层Conv3。

[0099] 定义三个最大池化层分别为第一最大池化层Pool1、第二最大池化层Pool2以及第三最大池化层Pool3,两个全连接层分别为第一全连接层Fc1以及第二全连接层Fc2。

[0100] 输入层、第一卷积层、第一最大池化层、第二卷积层、第二最大池化层、第三卷积层、第三最大池化层、第一全连接层以及第二全连接层依次连接。

[0101] 第一卷积层Conv1、第二卷积层Conv2以及第三卷积层Conv3分别包含16、32、64个滤波器,第一卷积层、第二卷积层以及第三卷积层的卷积核的大小均为3×3。

[0102] 其中,第一卷积层、第二卷积层以及第三卷积层后均设置一个ReLU激活函数。

[0103] 第一最大池化层、第二最大池化层以及第三最大池化层采用最大池化,设置步长为2。

[0104] 第一全连接层Fc1的神经元个数为32。

[0105] 第二全连接层Fc2的神经元个数为类别的数量,在本实施例中例如为12个。

[0106] 以上CNN分支网络的设计,可保证本实施例在针对预处理后的手势图像,所设计的CNN分支网络能够在保证准确率的同时拥有着更少的模型参数量,降低了计算量。

[0107] 本实施例设计的MIFD-Net模型中,使用MLP提取手部关键点特征信息。

[0108] 其中,MLP分支网络由三层全连接层构成。

[0109] 定义三层全连接层分别为第三全连接层Fc3、第四全连接层Fc4以及第五全连接层Fc5;则第三全连接层、第四全连接层以及第五全连接层依次连接。

[0110] 第三全连接层Fc3作为MLP分支网络的输入层,包含21个神经元,输入为预处理后得到的21个手部关键点特征数据;第四全连接层Fc4为隐藏层,包含16个神经元。

[0111] 第五全连接层Fc5为输出层,第五全连接层中神经元的个数设置为类别的数量。在本实施例中类别的数量设定为12个,此处第五全连接层中神经元的数量也为12个。

[0112] 本实施例中使用Relu函数作为各层全连接层之后的激活函数。

[0113] 两个分支网络的输出分别与特征融合模块相连,特征融合模块与分类模块相连。

[0114] 本实施例设计的MIFD-Net模型中,分支网络MLP提取的特征向量T_{out}包含较多的手部关键点位置信息,分支网络CNN提取的特征向量J_{out}包含较多的语义信息。

[0115] 图5展示了本发明设计的自适应权重特征融合模块C1。

[0116] 特征融合模块对两个分支网络提取的特征信息合理融合,引入自适应的特征权重ω_1、ω_2,使模型根据数据的特征分布来自行决定权重参数,在特征融合模块以不同的权重来融合特征。

[0117] 融合的手势特征C_f使用公式(5)计算得来:

[0118] $$C_f = \omega_1 * T_{out} \oplus \omega_2 * J_{out} \qquad (5)$$

[0119] 其中,$\oplus$代表Sum Fusion融合方式,权重ω_1、ω_2由公式(6)得到。

[0120] $$\omega_i = \frac{e^{\alpha_i}}{\sum_j e^{\alpha_j}} \qquad (6)$$

[0121] 其中,i=1,2,j=1,2,ω_i为归一化的权重,且$\sum \omega_i = 1$,α_i为初始化的权重参数,将α_i添加到优化器更新的参数中,使α_i向损失函数最小化的方向进行优化。

[0122] MIFD-Net中的特征融合模块引入了两个可训练的权重参数ω_1、ω_2。

[0123] 伴随着迭代过程的进行,网络的Loss逐渐趋于稳定,ω_1、ω_2的值逐渐适应其特征,充分考虑不同特征间的互补信息,获得更好的融合效果,模型泛化能力增强。

[0124] 最终经过特征融合模块融合后的手势特征信息经过两层全连接层完成最终的分类,分类模块包括两个全连接层,分别为第六全连接层Fc6以及第七全连接层Fc7。

[0125] 第六全连接层Fc6包含32个神经元,第七全连接层Fc7作为输出层,使用Softmax分类函数,神经元的个数设置为类别的数量,在本实施例中例如为12个。

[0126] 本实施例中模型的具体参数如表2所示,其中:

[0127] Input1为手势图片,Input2为对应的手势关键点的特征数据,Flatten层将数据展平。

[0128] 一般将Flatten层放置在卷积层和全连接层中间,起到一个转换的作用。

[0129] 卷积层的输出结果是二维张量,经过卷积层后会输出多个特征图,需要将这些特征图转换成向量序列的形式,才能与全连接层一一对应。

[0130] 表2

[0131]

层	卷积核大小	卷积核个数	步长	输出大小	参数数量
Input1	-	-	-	64×64×3	-
Conv1	3×3	16	1	64×64×16	448
Pool1	2×2	-	2	32×32×16	-
Conv2	3×3	32	1	32×32×32	4640
Pool2	2×2	-	2	16×16×32	-
Conv3	3×3	64	1	16×16×64	18496
Pool3	2×2	-	2	8×8×64	-
Flatten				4096	-
FC1	-	-	-	32	131104
FC2	-	-	-	12	396
Input2	-	-	-	21	
FC3	-	-	-	21	462
FC4	-	-	-	16	352
FC5	-	-	-	12	204
C1	-	-	-	12	2
FC6	-	-	-	32	416
FC7	-	-	-	12	396

[0132] 本发明搭建的多输入融合深度网络模型具有如下优势:

[0133] 使用手势图片和对应的手部关键点特征作为混合输入,使用CNN和MLP分支网络模型分别提取其特征,通过进一步特征融合,从而获取更多信息,进而提高模型的准确率。

[0134] 本发明同时采用手势图片数据以及手部关键点特征数据,是因为不同数据的表现方式不一样,看待事物的角度也会不一样,因此存在一些互补(所以比单特征更优秀)的现象。

[0135] 其中,手部关键点特征数据中包含更精确的手部关键点位置信息,手势图片数据

中则包含了更全面的手势全局信息,MIFD-Net联合手部关键点信息和图像视觉信息共同推理,通过使用自适应的权重,合理利用了不同信息的间的互补性,使得模型更具有普适性。

[0136] 步骤4.训练及测试多输入融合深度网络模型。

[0137] 利用步骤2中训练数据集中的样本数据训练多输入融合深度网络。

[0138] 多输入融合深度网络的训练过程如下:

[0139] 将步骤2得到的21个手部关键点的特征数据作为分支网络MLP的输入,经过隐藏层进行特征提取,在输出层得到一个第一特征向量,标记为特征向量T_{out}。

[0140] 将步骤2得到的手势图片数据作为分支网络CNN的输入,经过CNN网络提取特征后,得到一个第二特征向量,标记为输出特征向量J_{out}。

[0141] 为了获得更多特征信息增加识别准确率,本发明通过特征融合模块C1使用自适应特征融合的方法将两个分支模型的输出向量组合起来,然后进一步经过全连接神经网络使用softmax分类器进行预测分类。

[0142] 在训练过程中使用了Dropout防止过拟合,使得模型收敛速度加快,本实施例中使用了分类交叉熵损失函数,计算方法如公式(2)所示:

[0143] $LOSS = \sum_{i=1}^{m} y_i \cdot log\hat{y}_i$。

[0144] 其中,m是手势类别的数量,$\hat{y}_i$表示模型的预测输出,y_i表示真实的标签。

[0145] 本发明将模型训练的epoch设置为200,batchsize设置为32。该模型使用了Adam优化器进行训练,其中初始学习率设置为0.001,decay设置为1e-3/200。

[0146] 根据预测输出与对应的分类标签进行计算,得出分类损失函数Loss值。

[0147] 本发明通过使用EarlyStopping方法来监测模型训练的精度,在网络模型训练的过程中,记录每一次epoch的验证集Loss值,并记录其中的最小Loss值。

[0148] 当连续20次Epoch验证集Loss值一直大于这个最小Loss值时,则认为验证集Loss值不再下降,停止模型训练更新,取整个训练过程中最小Loss值的epoch训练结果作为最终的模型权值,进而保存最优的模型以及权重参数。

[0149] 从保存的模型权重中读取参数,得到训练好的多输入融合深度网络模型。

[0150] 在训练过程中,本发明方法将原始手势图像数据集按照训练集:测试集:验证集=7:2:1的比例切分,在训练过程中epoch设置为200,batchsize设置为32。

[0151] 该模型使用了Adam优化器进行训练,初始学习率设置为0.001,decay设置为1e-3/200。经过100次迭代训练得到的网络模型在测试集上的平均准确率可以达到99.65%。

[0152] 步骤5.对于待识别的手势图像,提取图像包含的21个手部关键点的特征数据以及手势图片数据,利用训练及测试好的多输入融合深度网络进行手势识别,得到识别结果。

[0153] 为了验证本发明方法的有效性,本发明还进行了如下实验。

[0154] 为了方便比较不同模型,将数据集按训练集:测试集:验证集=7:2:1的比例切分,保持输入到各模型的数据一致,在训练完模型后,本发明对模型的效果进行评价。

[0155] 本发明采用的模型评价指标有准确率(Accuracy)、召回率(Recall)、精确率(Precision)、F值(F-Measure)。其中,各个模型评价指标的计算方法为:

[0156] $$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

[0157] $$Recall = \frac{TP}{TP+FN} \quad (4)$$

[0158] $$Precision = \frac{TP}{TP+TN} \quad (5)$$

[0159] $$F-Measure = \frac{2\times(Precision\times Recall)}{(Precision+Recall)} \quad (6)$$

[0160] 其中,TP表示正例样本被标记为正例;FP表示假例样本被标记为正例;TN表示假例样本被标记为假例;FN表示正例样本被标记为假例。

[0161] 针对21个手部关键点特征提取,本实施例将MLP分支网络模型与其他处理方式进行了对比。其中,手部关键点数据在不同模型下的实验结果对比如表3所示。

[0162] 表3

[0163]

名称	精确率	召回率	F值	准确率
Logistic回归模型	0.96	0.96	0.96	0.9634
支持向量机模型	0.97	0.96	0.96	0.9631
1D-CNN模型	0.97	0.97	0.97	0.9705
本发明MLP模型	0.98	0.97	0.97	0.9725

[0164] 由上述表3对比能够看出,本实施例采用的MLP获得了更好的分类性能。

[0165] 此外,在针对预处理后的手势图像,本实施例与其他方式进行了对比,三种CNN模型在图片预处理后的数据集下的性能表现对比结果如表4所示。

[0166] 表4

[0167]

模型	精确率	召回率	F 值	准确率	参数量
LeNet	0.96	0.95	0.96	0.9580	697,216
AlexNet	0.98	0.98	0.98	0.9835	61,100,840
本发明CNN	0.98	0.98	0.98	0.9810	155,084

[0168] 由表4能够看出,CNN网络在保证准确率的同时拥有着更少的模型参数量。

[0169] 此外,本实施例还比较了MIFD-Net模型与其他用于手势识别的CNN的模型复杂度指标,即Parameters、FLOPs,如表5所示。其中,表5为手势识别模型复杂度对比。

[0170] 模型的Parameters和FLOPs是衡量模型大小的主要指标,Parameters衡量神经网络中所包含参数的数量,参数数量越小,模型体积越小,更容易部署。

[0171] 每个卷积层的参数量可以用公式(8)计算。

[0172] $Paras_{conv}=n\times(k\times k\times c+1)$ (8)

[0173] 全连接层的参数量可用公式(9)计算。

[0174] $Paras_{fc}=n\times(c+1)$ (9)

[0175] 其中,$Paras_{conv}$、$Paras_{fc}$分别代表卷积层、全连接层的参数量。

[0176] FLOPs衡量神经网络中前向传播的运算次数,FLOPs越小,则计算速度越快。每个卷积层的FLOPs可以用公式(10)计算。

[0177] $FLs_{conv}=2\times h\times w\times(k\times k\times c+1)\times n$ (10)

[0178] 全连接层的参数量FLs_{fc}可用公式(11)计算。

[0179] $FLs_{fc}=(2c-1)\times n$ (11)

[0180] 其中,FLs_{conv}、FLs_{fc}分别代表代表卷积层、全连接层的FLOPs。

[0181] 卷积核的大小为k×k,c、n分别代表该层输入特征图和输出特征图的通道数。

[0182] 表5

[0183]

模型	Parameters	FLOPs
MIFD-Net	0.157×10^6	0.84×10^6
MobileNet V2	2.3×10^6	320×10^6
AlexNet	62×10^6	700×106
GoogleNet	7×10^6	1510×10^6
VGG16	138×10^6	15500×10^6
SqueezeNet	0.7×10^6	830×10^6
InceptionV3	23.9×10^6	2850×10^6

[0184] 由上述表5能够看出,本实施例中的MIFD-Net的模型参数量和FLOPs更少。

[0185] 通过在本实施例自建的12种手势共12000张图片的数据集中可以达到99.65%的准确率,识别距离最远可达到500cm,识别速度为32帧/秒。

[0186] 在公开数据集The NUS hand posture datasets II中,本发明算法达到了98.89%的准确率。经过测试,在实际应用场景中可以达到很好的实时准确率。

[0187] 本发明方法通过图像采集模块检测到手势后,经过设计的手势预处理模型,得到手部关键点数据和消除背景后的手势图像数据,送入MIFD-Net模型,得到预测结果。

[0188] 本发明方法能够减少复杂背景的干扰,在强光、复杂背景中取得了良好的实时效果,实现在摄像头不同背景、不同距离下的实时手势识别。

[0189] 本发明述及的多输入融合深度网络的手势识别方法,应用于静态手势识别中。

[0190] 基于同样的发明构思,本发明实施例还提供了一种用于实现上述基于MLP和CNN的多输入融合深度网络的手势识别方法的识别系统,其包括如下几个模块:

[0191] 图像采集模块,用于获取原始手势图像数据并构建原始手势图像数据集;

[0192] 数据预处理模块,用于对原始手势图像数据集中各幅原始手势图像数据进行预处理,分别提取每幅原始手势图像中所包含的21个手部关键点的特征数据以及手势图片数据;

[0193] 将从每幅原始手势图像中提取到的21个手部关键点的特征数据和手势图片数据,以及每幅原始手势图像对应的标签,共同组成一组样本数据;

[0194] 将所有原始手势图像对应的样本数据组成样本数据集,并分为训练数据集和测试数据集;

[0195] 模型搭建及训练测试模块,用于搭建、训练以及测试多输入融合深度网络模型;

[0196] 多输入融合深度网络包括特征提取模块、特征融合模块以及分类模块;

[0197] 所述特征提取模块包括两个分支网络,分别是针对21个手部关键点特征提取的MLP分支网络以及针对手势图片特征提取的CNN分支网络;

[0198] 其中,两个分支网络的输出分别与特征融合模块相连,特征融合模块与分类模块相连;

[0199] 利用训练数据集中的样本数据训练多输入融合深度网络;

[0200] 其中,MLP分支网络的输入为21个手部关键点的特征数据,MLP分支网络的输出为对应于手部关键点的特征数据的第一特征向量;

[0201] CNN分支网络的输入为手势图片,CNN分支网络的输出为第二特征向量;

[0202] 特征融合模块用于将第一、第二特征向量组合起来,并经过分类模块预测输出预测结果;

[0203] 利用测试数据集中的样本数据对训练好的多输入融合深度网络进行测试;

[0204] 预测模块,对于待识别的手势图像,用于提取图像包含的21个手部关键点的特征数据以及手势图片数据,利用训练及测试好的多输入融合深度网络进行手势识别得到识别结果。

[0205] 需要说明的是,基于MLP和CNN的多输入融合深度网络的手势识别系统中,各个模块的功能和作用的实现过程具体详见上述方法中对应步骤的实现过程,在此不再赘述。

[0206] 本发明提出的MIFD-Net模型在保持精度的同时较少了模型参数,因此,在各种情况下,包括户外活动,可以使用配备该系统的便携式终端轻松识别常用手势。

[0207] 此外,本发明还提出了一种用于实现上述多输入融合深度网络的手势识别方法的计算机设备。该计算机设备包括存储器和一个或多个处理器。

[0208] 其中,在存储器中存储有可执行代码,处理器执行可执行代码时,用于实现上述多输入融合深度网络的手势识别方法。

[0209] 本实施例中计算机设备为任意具备数据数据处理能力的设备或装置,此处不再赘述。

[0210] 此外,本发明实施例还提供一种计算机可读存储介质,其上存储有程序,该程序被处理器执行时,用于实现上述多输入融合深度网络的手势识别方法。

[0211] 该计算机可读存储介质可以是任意具备数据处理能力的设备或装置的内部存储单元,例如硬盘或内存,也可以是任意具备数据处理能力的设备的外部存储设备,例如设备上配备的插接式硬盘、智能存储卡(Smart Media Card,SMC)、SD卡、闪存卡(Flash Card)等。

[0212] 当然,以上说明仅仅为本发明的较佳实施例,本发明并不限于列举上述实施例,应当说明的是,任何熟悉本领域的技术人员在本说明书的教导下,所做出的所有等同替代、明显变形形式,均落在本说明书的实质范围之内,理应受到本发明的保护。

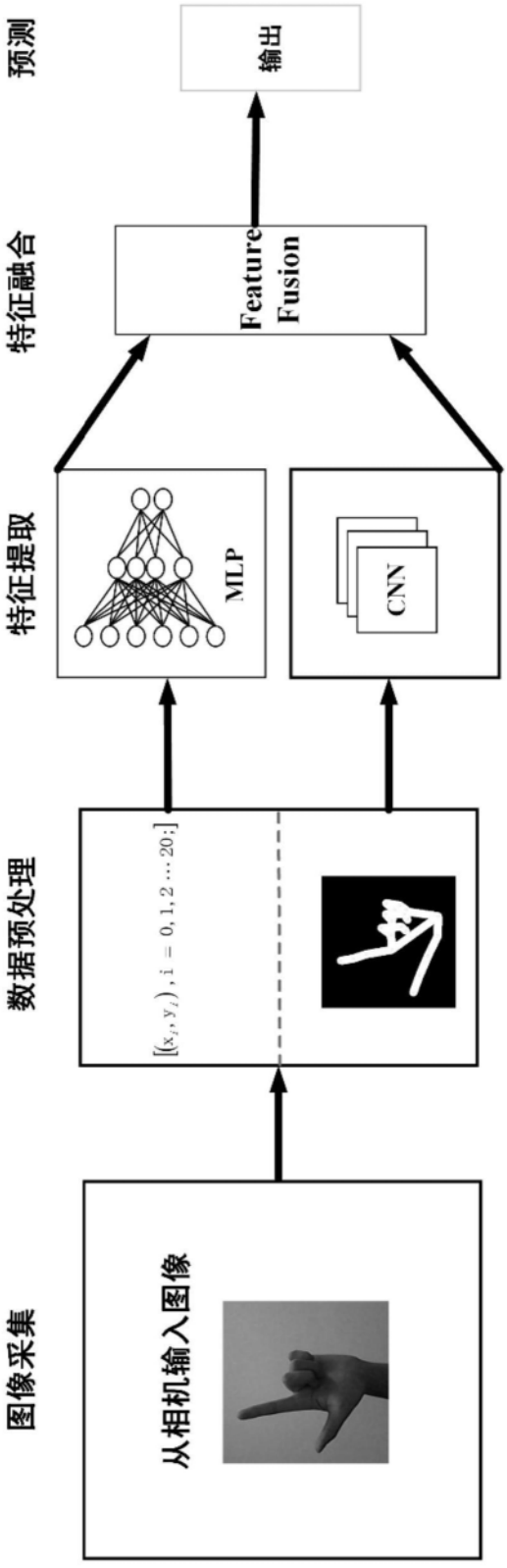

预测
输出
特征融合
Feature Fusion
特征提取
MLP
CNN
数据预处理
$[(x_i, y_i), i = 0, 1, 2 \cdots 20;]$
图像采集
从相机输入图像

图1

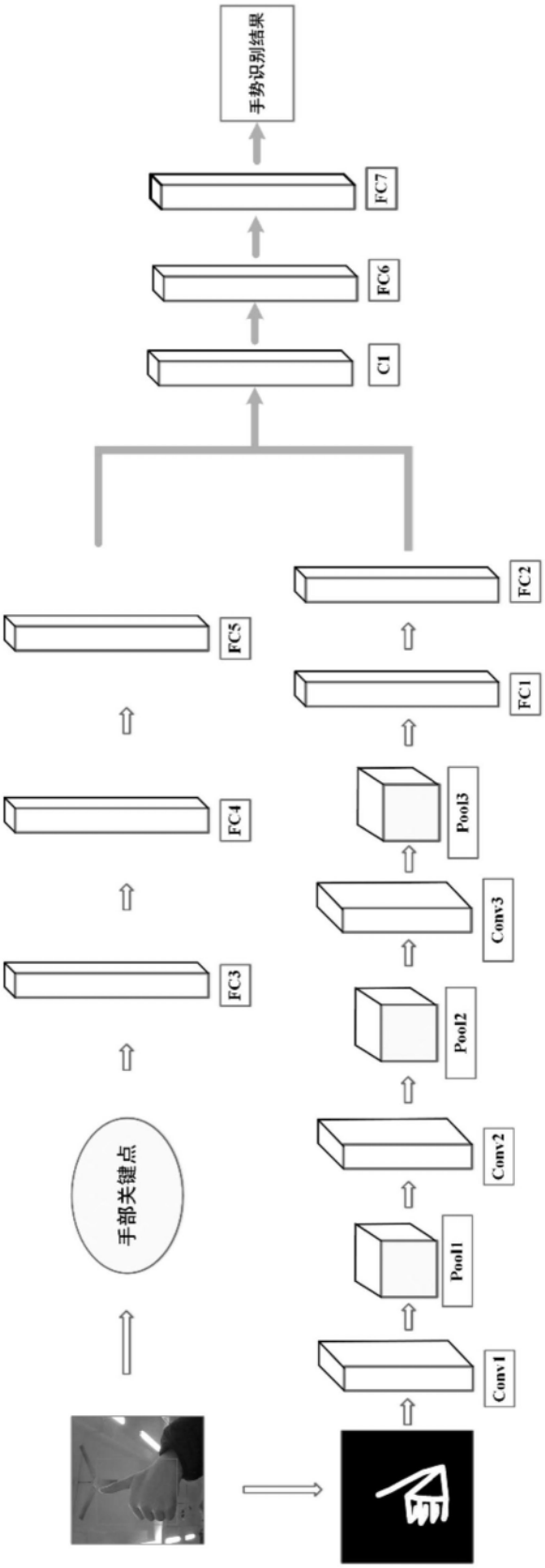

手势识别结果
FC7
FC6
C1
FC2
FC5
FC1
Pool3
FC4
Conv3
FC3
Pool2
手部关键点
Conv2
Pool1
Conv1

图2

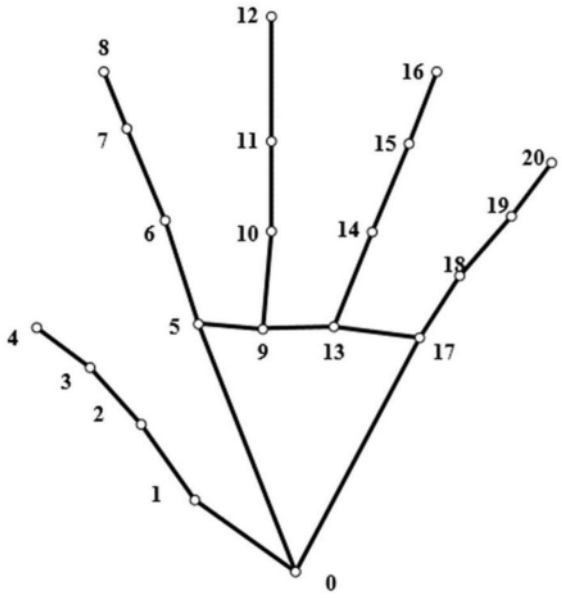


图3

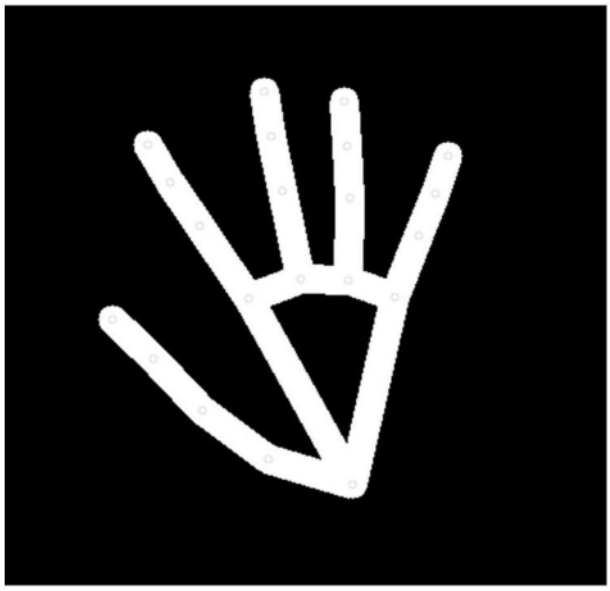


图4

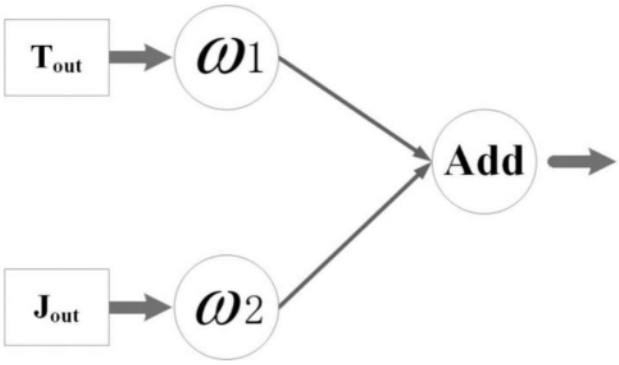


图5