CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 40 卷第 5 期 2018 年 5 月
Vol. 40, No. 5, May. 2018

文章编号:1007-130X(2018)05-0872-08

基于计算机视觉及深度学习的无人机手势控制系统*

马乐乐,李照洋,董嘉蓉,侯永宏
(天津大学电子信息工程学院,天津 300072)

摘　要:传统的无人机人机交互需要专门的设备和专业的训练,便捷新颖的交互方式往往更令人青睐。利用普通相机,对基于计算机视觉以及深度学习的无人机手势控制系统进行了研究。该系统首先利用快速跟踪算法在视频序列中提取出操作者所在区域,大大减少后续视频处理压力的同时去除了复杂背景以及相机漂移的影响。其次,根据动作的时间信息,用不同颜色编码光流特征,叠加在一张图片上,将视频转换为同时包含时间特征以及空间特征的彩色纹理图。最后,利用卷积神经网络对彩色纹理图进行学习及分类,根据分类结果生成控制无人机的指令。该系统每 0.4 s 对 1.6 s 内的动作进行一次判定,利用卷积神经网络对图片的分类实现实时性的人机交互,系统在 60 m 范围内的识别准确率在 93%以上,在室内和室外环境下,操作者可以通过模仿指令动作方便地控制无人机。

关键词:人机交互;深度学习;卷积神经网络;无人机;手势控制
中图分类号:TP391.4　　**文献标志码**:A
doi:10.3969/j.issn.1007-130X.2018.05.016

UAV gesture control system based on computer vision and deep learning

MA Le-le, LI Zhao-yang, DONG Jia-rong, HOU Yong-hong
(School of Electric Information Engineering, Tianjin University, Tianjin 300072, China)

Abstract: The traditional Unmanned Aerial Vehicle (UAV) human-machine interaction requires specialized equipment and professional training, and convenient and innovative ways of interaction are often more popular. In this paper, with ordinary cameras, we study the UAV gesture control system based on computer vision and deep learning. The system first uses the fast tracking algorithm to extract the operator's region in the video sequence, greatly reducing the pressure of subsequent video processing while removing the influence of complex background and camera drift. Secondly, according to the time information of the actions, the optical flow features are encoded in different colors and superimposed on a picture, and the video is converted into a color texture map that contains both temporal features and spatial features. Finally, colored texture images are well learned and classified by a deep Convolutional Neural Network (CNN) and UAV controlling commands are generated according to the classified results. The proposed system estimates actions within 1.6s every 0.4s and uses CNN to classify pictures so as to achieve real-time human-computer interaction. The system has a recognition accuracy of over 93% within 60 meters. In indoor and outdoor environments, the operator can conveniently control the UAV by imitating command actions.

Key words: human machine interface; deep learning; CNN; UAV; gesture control

* **收稿日期**:2016-08-16;**修回日期**:2016-12-07
通信地址:300072 天津市卫津路 92 号天津大学电子信息工程学院
Address: School of Electric Information Engineering, Tianjin University, 92 Weijin Rd, Tianjin 300072, P. R. China

1 引言

近年来,人机交互得到了越来越多的关注。在追求更自然的交互模式过程中,陈一新[1]的研究展示了基于计算机视觉的手势控制人机交互模式具有广阔的应用前景。

随着低成本无人机 UAV(Unmanned Aerial Vehicle)的兴起,无人机的新颖交互方式是研究者关注的热点。传统无人机控制方法使用遥控器、摇杆等专用设备,新颖的无人机控制方法[2-4]是通过让人穿戴特殊辅助设备,简化了无人机的控制。Téllez-Guzmán 等[2]提出让人穿戴装有惯性测量单元 IMU(Intertial Measurement Unit)的头盔来控制小型无人机。Vincenzi 等[3]提出了为残障人士设计的人机交互系统。Lupashin 等[4]建立了 FMA(Flying Machine Arena)平台,使得控制者可以用专用的控制棒控制无人机的飞行。相比于依赖特殊辅助设备的控制方法,基于计算机视觉的无人机交互模式更具有普适性及发展前景。Mantecón 等[5]基于 Kinect 平台探索了基于动作识别的人机交互方法,操纵者在距离固定于地面的 Kinect 传感器 4.5 m 处,通过识别动作来控制飞机。不同于识别整个人身体动作,Pfeil 等[6]基于 Kinect 平台,仅通过识别手掌的动作,对无人机进行手势控制。文献[5,6]中的方法虽然有较高的识别准确率,但操作者无法离开固定于地面的 Kinect 摄像机。Naseer 等[7]把 Kinect 挂载在无人机上,但因 RGB-D 传感器对距离的限制,该方法只能对操控者进行近距离的跟踪,若跟踪失败则存有较大安全隐患。为了扩大控制者与无人机交互的范围,Monajjemi 等[8]在无人机上挂载普通摄像机,通过对特定使用者进行动作采集并预先建立相应的数据库,从而实现个性化定义人机交互动作。然而,该系统仅限于在室内环境下运行,在复杂的室外场景应用中有较大的困难。综上,对应用于室外环境的无人机手势识别算法,需要在识别精度、控制距离和识别速度上有更大的提升。

传统的人类动作识别主要基于手工特征。为了能够学习到更高层次的动作特征,较为普遍的方式是利用深度学习。Taylor 等[9]提出了一个基于卷积神经网络的动作识别框架,该框架可直接以连续的图片序列作为输入。Ji 等[10]通过一个三维的卷积神经网络挖掘视频中人类动作的时间特征与空间特征。然而 Karpathy 等[11]在对比研究了包括文献[9,10]在内的多种卷积神经网络后指出,采用卷积神经网络单独处理视频序列中的每一帧与同时处理一系列帧的结果相近,即时间与空间特征在网络模型中并没有很好地结合。为了能够更好地结合动作序列中的空间特征与时间特征,Simonyan 等[12]利用两个卷积神经网络来分别处理空间数据流与时间数据流,并在最后通过特定的方法对不同特征的学习结果进行融合。然而,空间特征与时间特征分开计算以及两个网络的并行,大大增加了计算开销,使得该方法无法满足人机交互这种实时性强的应用需求。

本文中的动作识别算法主要受 Wang 等[13]和 Shi 等[14]的启发。Wang 等[13]对动作序列的深度信息进行了压缩,将一段深度视频序列有效地压缩为三张图片,最后利用卷积神经网络来完成对动作序列的学习分类。Shi 等[14]则基于普通视频序列,通过计算稠密的运动轨迹,将一段视频的运动轨迹进行累积并叠加成图片,在利用卷积神经网络对图片学习的同时,结合多种其余视频特征来弥补时间特征的丢失。虽然以上两篇文献所提算法均在当下知名的数据集中取得了远高于传统手工特征的识别准确率,但是由于计算的复杂性,其均不适用于基于普通相机的实时人类动作识别任务。

本文基于普通相机,通过嵌入式平台,构建了一个基于计算机视觉及深度学习的无人机人机交互系统。该系统利用跟踪算法对视频进行预处理,有效地克服了复杂背景以及相机漂移问题。该系统每 0.4 s 对 1.6 s 内的动作进行一次判定,利用卷积神经网络对图片的分类实现实时性的人机交互,创新性地将卷积神经网络扩展到了实时应用上。系统在 60 m 范围内的识别准确率在 93% 以上,在室内和室外环境下,操作者可以通过模仿指令动作方便地控制无人机。鲁棒的识别算法不仅使该系统成为首个能应用于室外环境的无人机手势控制系统,也为手势识别这种新颖的人机交互走进千家万户提供了研究基础。

2 动作识别框架

在本文提出的无人机人机交互系统中,动作识别框架主要由两个部分组成:视频预处理并生成彩色纹理图、卷积神经网络模型的训练及分类。第一部分首先人为指定无人机操作者,利用跟踪算法对操作者在视频中进行跟踪,并在高分辨率的视频中截取只包含操作者的部分用于后续计算。其次,在

相邻帧之间计算光流特征。最后,利用不同颜色在光流基础上编码时间特征,将光流叠加成彩色纹理图。第二部分主要应用卷积神经网络进行彩色纹理图的学习,并通过分类彩色纹理图实现对动作的识别。本文提出的动作识别框架借鉴了 Wang 等[13]将视频序列压缩成图片来进行动作分类的思想,但是不同于其基于深度视频序列的识别框架,本文致力于普通视频序列的动作识别。同时,避免了 Shi 等[14]所提方法对时间特征计算的额外开销,本文系统极大地提高了识别速度。

2.1 视频预处理并生成彩色纹理图

为了让无人机手势控制系统鲁棒地应用于户外环境,系统在启动时,根据摄像机当前的显示内容,在地面站通过鼠标点击的方式人为指定唯一的操作者,并将操作者人脸所在范围设置为跟踪区域。利用快速的视觉跟踪算法[15]来对操作者进行跟踪的同时,根据跟踪结果在较高分辨率的视频中裁剪出以操作者为中心的低分辨率视频序列(详细跟踪算法请看文献[15],本文在此不作描述)。通过这样的视频预处理,一方面能够很好应对相机漂移,有效去除操作者附近以外的复杂背景;另一方面,只处理剪裁后的视频大大加快了无人机系统中算法的处理速度。对操作者的跟踪是本文动作识别的基础,唯有良好的跟踪才能在后续取得良好的动作识别结果,在无人机搭载的相机中,操作者的移动速度相对缓慢且不存在镜头切换等场景,文献[15]中的跟踪算法能够保证对操作者的良好跟踪。

视频经过裁剪后,先对裁剪后视频的每相邻两帧求取含有动作空间特征的光流图,并通过彩色信息对光流进行编码,将动作视频序列叠加成彩色纹理图片。

具体细节如下:

假设动作由 n 帧视频序列构成:$f_1,f_2,\cdots,f_n$,其中 f_i 表示视频序列中的第 i 帧。我们利用光流来描述人物的动作信息。在相邻两帧 f_{i-1},f_i 计算出的光流中,像素位置 (x,y) 处的光流矢量用 $\langle u_{x,y,i},v_{x,y,i}\rangle$ 来表示。则该处光流的幅值可表示为 $M_{x,y,i}$:

$$M_{x,y,i}=\sqrt{u_{x,y,i}^2+v_{x,y,i}^2} \tag{1}$$

通过光流有效捕捉动作的空间特征后,本文提出了一种利用颜色来编码光流的方法,该方法有效地捕捉了动作的时间特征。基于 HSV 彩色空间,动作序列中不同时刻的光流赋予不同的色度值。另 $h_{\max}$ 和 $h_{\min}$ 表示实验中 HSV 彩色空间中色度的取值范围。相邻帧 f_{i-1},f_i 之间计算出的光流图中,所有计算出光流的像素位置均用色度 H_i 来进行编码:

$$H_i=\frac{i}{n}\times(h_{\max}-h_{\min})+h_{\min} \tag{2}$$

计算光流并用颜色进行编码后,在整个动作序列中计算出的多张光流图上,可求出每一像素位置 (x,y) 处光流幅度的极大值 $M_{x,y,k}$:

$$M_{x,y,k}=\max\{M_{x,y,1},M_{x,y,2},\cdots,M_{x,y,n}\} \tag{3}$$

由于光流幅度大,即运动幅度大的地方,往往包含更多动作信息,因此,取光流幅值 $M_{x,y,k}$ 对应的色度 H_k 作为该像素位置经过叠加后的色度值更为合理。由此,在对所有像素位置进行以上计算后,视频序列中的动作特征被压缩成一张色彩丰富的彩色纹理图。彩色的位置充分表达了动作的空间特征,而色度充分表达了动作的时间特征。

动作视频序列的叠加过程如图 1 所示,图中动作取自 Keck 数据集[16](逆时针画圆),左侧为该动作中相邻帧间计算出的光流图样例,右侧为光流叠加成的纹理图。从图 1 中可看出,对光流直接叠加的结果仅具有动作的空间特征,而经过颜色的编码,图片信息更为丰富,对动作信息的表达更加全面。

2.2 数据扩展以及模型训练

视频序列中的动作信息被有效压缩成彩色纹理图后,便可通过卷积神经网络 CNN(Convolu-

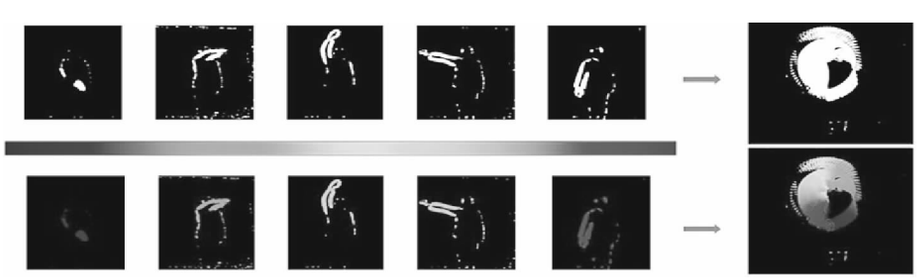


Figure 1 Color coded texture images

图 1 彩色纹理图

tional Neural Network)对图片的充分学习以及分类能力来完成动作的识别。然而,过多的学习参数使卷积神经网络往往在训练数据不充足时陷入过拟合的困境。

为了有效避免过拟合且令算法更适应于无人机系统,我们首先采用了旋转图片以及分辨率变换的方法对训练数据集进行了扩展。图片的小角度旋转用于模拟无人机飞行中的不同姿态,而分辨率的变换用于模拟无人机与操作者间不同的距离,后续实验表明该方法简单而有效。

我们采用了在 ImageNet 数据集上取得了令人瞩目成绩的 AlexNet[17] 网络结构。为使得卷积神经网络能够更快地取得良好结果,AlexNet 网络在 ImageNet 上训练好的模型被应用在本文中进行初始化网络参数。笔者通过显卡(NVIDIA TITAN X)对训练过程进行加速,实现过程采用了开源的 Caffe 工具箱[18]。网络中的各层权重通过梯度下降法求解,每一次迭代(Iteration)使用 256 张训练图片(Batch Size),我们一共训练了约 90 个周期(Epoch)。初始的学习率被设置为 0.001(Base Learning Rate),该学习率在进行了 60 个周期后下降为 0.000 1。其余训练参数参考 Krizhevsky 等[17]的默认设置。

具体地:本文 4.1 节中,训练数据为根据 Keck 数据集的 126 个训练数据生成的彩色纹理图,每一个训练数据序列均生成一张彩色纹理图。故未经扩展的训练数据共包含 126 张图片,而经过数据扩展后(具体扩展配置:(1)在 $-20°$到 $20°$范围内,以 $4°$为步长对图片做旋转;(2)对旋转后的所有图片,在 40%~100%的缩放比例范围内,以 10%为步长进行缩放变换。训练数据扩充至 77 倍),训练数据包含 9 702 张图片。在未经过扩展的数据集上训练耗时约 2 min,数据存在较为严重的过拟合现象,在经过扩展的数据集上训练耗时约 1 h,大量的数据有效地改善了过拟合现象。本文 4.2 节中,动作指令的训练数据由 Keck 数据集中选定的 5 类动作(同时包含 Keck 的训练集与测试集数据)生成,共 105 张彩色纹理图,经过与 4.1 节中相同配置的数据扩展后,动作指令训练数据包含 8 085 张图片。而非动作指令为 Keck 数据集中其余动作经过 3.2 节所介绍的滑动窗口选取的数据,经统计共包含 7 927 张图片(由于数目较多故不做扩展),故 4.2 节中无人机系统上的训练数据(5 个操作指令动作,1 个非操作指令动作)共包含 16 012 张图片,训练耗时约 1 h 35 min。

3 系统结构及运行时间分析

3.1 系统结构

本系统基于四旋翼无人机平台搭建。无人机通过飞行控制器和 GPS 模块,实现了在室外自主悬停。飞机搭载嵌入式平台进行在线的图像处理和动作识别。无人机平台用电压为 22.2 V、容量为 5 200 mAH 的锂离子电池供电。另外,我们用电压为 11.1 V、容量为 2 300 mAH 的锂离子电池单独为嵌入式平台供电。

嵌入式平台选用 Jetson TK1 平台,该平台搭载的处理器包含有 4 个主频为 2.3 GHz ARM-a15 内核的 CPU 和一个 GK20 GPU,可以给图像处理提供足够的运算能力。图像的采集处理、卷积神经网络对动作的分类均在该平台上进行。为提高识别距离,本系统采用的摄像机能够捕捉具有 300 万像素的高清图片,视角为 60°。同时嵌入式处理平台作为飞行控制器与地面站数据传输的中继,通过串口连接飞行控制器,通过 Wi-Fi 连接地面站。地面站可以监测多轴飞行器的状态,用于指定操作者和查看实时运算的结果。出于安全考虑,我们在传统遥控器上设置了应急开关,紧急情况下可以切换到手动控制的状态。

本系统基于机器人操作系统 ROS(Robot Operation System)[19]。视频采集、动作特征提取及压缩、卷积神经网络的分类和指令生成分别以独立的进程并行地执行。图 2 为系统硬件结构。

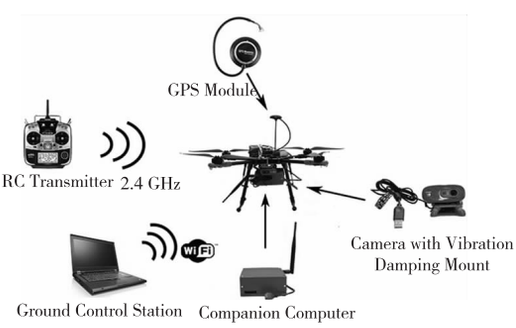


Figure 2 System hardware structure

图 2 系统硬件结构

3.2 算法运行时间分析

实时的人机交互系统对算法的运行时间具有严格的要求。然而,对视频的处理以及利用卷积神经网络进行分类是一项耗时的工作。在本系统中,相对耗时的算法逻辑包含:视频中人物跟踪、计算光流、叠加成彩色纹理图以及利用卷积神经网络分

类。我们巧妙地利用了摄像机捕捉视频时的时间间隔,以及嵌入式系统的并行处理能力,在捕捉视频的同时,并行地对图像进行计算。在利用卷积神经网络分类过程中,又有效地在 Caffe 平台借助 GPU 进行加速。最终对动作判决的平均延时为 354 ms。

本文所采用的跟踪算法运行时间与跟踪范围成正比。Zhang 等[15]提出的快速跟踪算法在基于 Intel i7 处理器的电脑上,每秒可处理 350 帧图像,平均延时低于 3 ms。经测试,该算法的跟踪速度与跟踪范围成反比,且跟踪范围增加会导致跟踪速度显著下降。为保证运行速度,本系统中的跟踪范围只限于操作者人脸部分,后续图像处理中,再根据跟踪区域在周围截取更大区域。经统计,视频中完成人物跟踪以及剪裁的耗时为 16 ms(t_1)。

光流计算方面,本文并没有以速度为最重要因素,在众多光流算法中,本系统采用了速度较快,且较为稠密的 HS 光流算法,对于 640 * 480 大小的图像,该算法的平均运行时间为 15～60 ms。然而得益于基于跟踪的视频裁剪,实际计算的图片为只以操作者为中心的图片,其尺寸更小。经统计,本系统中光流计算的平均时间约为 41 ms(t_2)。

在光流进行叠加成彩色纹理图的过程中,理论上首先用颜色对光流进行编码,然后再通过统一比较光流幅度值来选取像素颜色。在系统实现过程中,采用迭代的方式,将像素位置光流取幅值的过程转化为不断取最大值的过程,颜色选取也在迭代的比较中一并实现。每次迭代以帧为单位,平均每次迭代耗时 14 ms(t_3)。

对彩色纹理图的分类过程是采用线下训练好的卷积神经网络进行一次正向流动得到的结果。出于准确率考虑,本系统采用了较为复杂的 AlexNet 网络。依赖于嵌入式系统的 GPU 支持,我们在 Caffe 平台上利用了 NVIDIA 提供的 cuDNN 对其进行了加速,网络的参数在系统运行初期初始化,使得平均每次分类时间约为 283 ms(t_4)。

至此本算法的每一步平均耗时被列出后,假设捕捉视频中,每两帧之间的时间差为 t_0 ,对于整体的运行延时计算如图 3 所示。由图 3 可知,假设 $t_0 > t_n = t_1 + t_2 + t_3$,则无论多久进行一次判定,延时始终为 $t_m = t_1 + t_2 + t_3 + t_4$,经计算 t_n =71 ms,t_m =354 ms。因此,较为保守地将系统设置为每秒 10 帧:t_0 =100 ms> t_n ,故每次判定延时可有效控制在 354 ms 以内。

出于嵌入式平台承载能力考虑,仅设有 1 个进程用于卷积神经网络的分类,故平均每 354 ms 才能进行一次动作判定。经验证,在操作手势平均在 1～2 s 内完成时,本系统采用平均每 0.4 s 进行一次动作判定,已可流畅反馈当前人机交互控制指令。据此,本系统中动作识别模块共启用了 4 个生成彩色纹理图的线程,生成间隔为 400 ms,均压缩 1.6 s 内的动作。相当于 1.6 s 宽度的滑动窗口,以 0.4 s 为步长不断压缩当前窗口内的动作。一块固定的内存被用于刷新存储生成的彩色纹理图,以供识别进程方便地读取神经网络的输入。

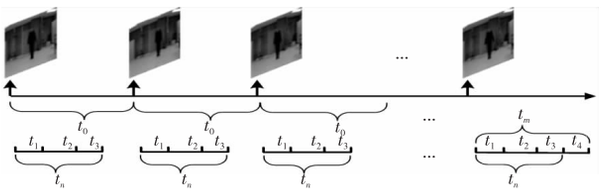


Figure 3 Algorithm runtime analysis

图 3 算法运行时间分析

4 实验结果

本节基于 Keck 数据集[16],对算法步骤的有效性以及算法的可靠性进行充分验证。其次,在该数据集上取部分动作进行训练后,直接应用于室外场景,展现了算法的普适性以及系统的鲁棒性。最后根据室外实验结果,对识别距离以及识别准确度进行了分析。

4.1 识别算法评估

Keck 数据集[20]是用分辨率为 640 * 480 的传统摄像机录制的数据集。该数据集包含 14 类手势动作,分别标识为类别:左转、右转、注意向左、注意向右、双侧注意、左侧停止、右侧停止、双侧停止、左右移动、开始、后腿、缩小距离、加速、靠近。训练数据集中相机稳定,背景简单且无干扰,共包含 126 个视频序列。测试数据集的录制背景相对复杂,同时伴有相机漂移、人物走动以及其余人物干扰,共包含 168 个视频序列。本系统所面临的识别难点与该数据集相似,无人机在飞行过程中往往伴有相机漂移以及包含人员走动的复杂背景,因此选择该数据集进行算法验证。

由于本算法在实际应用中首先采用了跟踪算法进行视频预处理,为了方便与其他算法进行比较,在跟踪上避免手动指定跟踪区域,以下针对该数据集的实验中,默认采用跟踪视频中央 96 * 72 像素区域,相应地进行视频处理的划定范围为跟踪区域附近的 400 * 300 像素范围。

为验证算法的有效性,本文对光流叠加中有无颜色编码、普通分类算法与卷积神经网络、卷积神经网络训练是否经过 ImageNet 模型初始化网络、数据扩展以及视频序列是否经过跟踪预处理等五个方面设置以下对比实验:(1)跟踪+黑白纹理图+HOG+SVM;(2)跟踪+彩色纹理图+HOG+SVM;(3)跟踪+彩色纹理图+无 ImageNet 模型初始化 CNN;(4)跟踪+彩色纹理图+CNN;(5)无跟踪+数据扩展+彩色纹理图+CNN;(6)跟踪+数据扩展+彩色纹理图+CNN。其中 HOG+SVM 算法[20]是图片分类的经典方法,通过提取 HOG 特征,利用线性 SVM 作为分类器进行分类,其广泛应用于行人检测,速度与效果综合平衡性能较好(详细请参考文献[20])。数据扩展所采取的配置(将数据集扩大到 77 倍):(1)在-20°到 20°范围内,以 4°为步长对图片做旋转;(2)对旋转后的所有图片,在 40%~100%的缩放比例范围内,以 10%为步长进行缩放变换。实验结果如表 1 所示。其中,跟踪:Tracking,黑白纹理图:BW,彩色纹理图:Color,数据扩展:Data Augmentation,无 ImageNet 模型初始化网络:No-Pretrain-CNN,有 ImageNet 模型初始化网络:CNN。

Table 1 Testing results on Keck dataset with different algorithm settings

表 1 Keck 数据集上的对比实验识别结果

Methods	Accuracy/%
Tracking,BW,HOG+SVM	66.07
Tracking,Color,HOG+SVM	75.60
Tracking,Color,No-Pretrain-CNN	85.12
Tracking,Color,CNN	89.28
Two-Stream Model (fusion by average)[12]	88.69
Shape-Motion Prototype Trees[16]	91.07
No Tracking,Data Augmentation,Color,CNN	82.14
Tracking,Data Augmentation,Color,CNN	94.05

从黑白纹理图与彩色纹理图的对比实验中能够看出,在不改变分类方式时,光流叠加采用颜色编码能够更有效地捕捉动作的时间特征,从而提高识别精度;在采用相同的彩色纹理图时,得益于卷积神经网络优秀的学习以及识别能力,利用卷积神经网络进行分类的效果明显优于 2005 年 Dalal[20]提出的 HOG+SVM 分类方法。在采用卷积神经网络进行训练时,预先用 AlexNet 网络在 ImageNet 数据集上训练好的模型来初始化网络参数,能够将识别准确率从 85.12%提升至 89.28%。初始化带来的效果提升主要是由于本文生成的彩色纹理图与 ImageNet 中的图像有很大的差异,而本文的数据集相比 ImageNet 数据集而言仍然相对较小,单纯在本数据集进行训练所得的网络参数对图像的表示能力还不够强,泛化性能稍差,而经过 ImageNet 上的模型初始化后,网络参数丰富,尤其第一、二层卷积层已经是适于描述图像的基本单元,对图像的表达能力更强,不仅能够更快地学习至收敛,还能够具有更强的泛化特性。此外,本算法与 Lin 等[16]在 2009 年提出 Keck 数据集时所采用的运动模型原型树算法(Shape-Motion Prototype Trees)进行了对比,该算法旨在鲁棒地应对相机漂移、繁杂背景以及行人干扰,本文算法的识别准确率比其高出约 3%。同时,本算法与 Simonyan 等[12]在 2014 年提出的基于深度学习的动作识别算法进行了对比,该算法分别针对动作的时间特征以及空间特征,采用两个卷积神经网络进行学习,并在业内多个数据集上取得了良好的表现。然而该算法对 Keck 数据集中相机大尺度的漂移、人员走动等复杂场景下的动作识别效果并不理想,准确率仅为 88.69%。最后,在跟踪的基础上进行了经过数据扩展的实验中,本文算法分类准确率达到 94.05%,远高于未经跟踪预处理的 82.14%以及未经过数据扩展的 89.28%。跟踪带来的优势主要是测试数据集中有约 1/4 的视频序列包含镜头漂移以及附近范围人员走动,跟踪并剪裁能够鲁棒应对该问题,数据扩展的优势来自于训练数据集的规模较小,未扩展的数据在训练时伴有较为严重的过拟合现象。

4.2 系统整体实验

Keck 数据集与本系统实际应用环境极其相似,即数据集中存在相机漂移以及复杂环境下人物走动,同时该数据集中的动作手势适合作为操作指令。本系统在进行室外实验时,采用了极具挑战的交叉实验方法,即利用 Keck 数据集进行数据训练,未包含在数据集中的操作者直接在室外模仿相应的动作来进行实验。在 Keck 数据集中,我们抽取了 5 个动作类别并分别指定为无人机操作手势(括号中为 Keck 数据集中指令名称):向左飞行(左转)、向右飞行(右转)、拍摄照片(开始)、向下飞行(左右移动)、向上飞行(加速)。此五类动作依据官方的序列分隔作为动作开始结束的界限,生成了动作标准的训练数据集。此外,对于无人机可分类的动作中,需要包含非控制指令。因此,利用 Keck 数据集中其余 9 类动作序列,根据 3.2 节的设计方案,以 1.6 s 宽度的滑动窗口平均每 0.4 s 生成彩

色纹理图,将这些纹理图作为非控制指令类别的训练数据。基于 Keck 数据集生成了原始训练数据后,将代表 6 类控制手势的彩色纹理图(5 个控制指令与 1 个非控制指令)进行以下数据扩展:(1)在-20°到 20°范围内,以 4°为步长对图片做旋转;(2)对旋转后的所有图片,在 40%~100%的缩放比例范围内,以 10%为步长进行缩放变换。该数据扩展后,训练数据被扩充了 77 倍,可以有效改善卷积神经网络的过拟合问题。

本实验所选取的训练数据全部来自于 Keck 数据集,实际场景与数据集场景差异过大,为尽可能地降低环境差异,本系统在室外环境测试中,操作者身着深色纯色衣裤(与 Keck 数据集相近),并在距离无人机 5~90 m 范围内每间隔 5 m 对每一控制指令分别做 20 次,共 100 个操作指令,期间伴随左右走动以及干扰动作。图 4 中操作者距离无人机 60 m,左下方为根据跟踪算法剪裁出的视频,方框为跟踪范围,右下图为 1.6 s 内的动作特征计算出的彩色纹理图。

Figure 4 Sample of colored texture image and video frame

图 4 实验视频及彩色纹理图样例

图 5 为识别准确率与距离的对应关系。从图 5 中能够看出,操作者在距离无人机 60 m 范围内均保持较高的识别准确率。在该范围内,少量的错误识别主要集中在"向上飞行"指令中,原因为该动作与非控制指令中的一些纹理图有些相似,造成了一定干扰。在距离超过 60 m 后,识别准确率骤降,其主要受分辨率影响,操作者在图片中的分辨率太差以至于无法有效地计算光流。

得益于 Keck 数据集跟实际应用场景非常相似,且在实际环境中,即便相机随着无人机运动产生漂移,其漂移的尺度也远小于 Keck 数据集中的漂移尺度,操作者附近的人员走动并不会影响手势

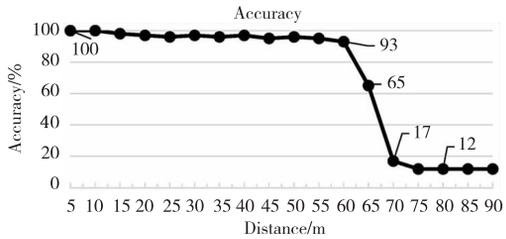


Figure 5 Recognization accuracy in different distances

图 5 不同距离上的识别准确率

控制,是因为无人机相机的视野宽广,通过跟踪对操作者附近视频进行剪裁避免了这些干扰。实际环境下,操作者对无人机控制手势模仿时,会随着无人机的飞行,伴有操作者倾斜以及不同距离导致的分辨率差异,但得益于训练过程的旋转以及分辨率变换的数据扩展,使得卷积神经网络模型能够有效地对实际环境的动作进行区分。对于操作者的衣着以及环境选取,当前较为理想的是身着深色的纯色衣裤,并在几乎不含遮挡的环境下操作,因为对于该系统而言,跟踪算法是动作识别的基础。本文出于运行速度以及跟踪准确度考虑,暂没有针对过复杂的遮挡环境,而操作者的衣着在目前尽可能简单是由于训练数据集的限制。交叉实验本身的难度巨大,要求算法具有强大的泛化性能,对于神经网络而言,后续可通过扩大训练数据集的方式来弥补这一缺陷。

5 结束语

本文基于计算机视觉及深度学习,开发了基于普通相机的动作识别算法,同时基于该算法构建了一套无人机手势控制交互系统。利用该系统,操作者可在 60 m 范围内通过手势控制无人机飞行。识别算法在动作特征的提取及压缩方面,创新性地将一段视频中动作的时间特征以及空间特征同时压缩到一张图片中;在特征学习及分类方面,我们将卷积神经网络扩展到了实时性的应用上。该系统能够有效应对人物走动以及相机的漂移,识别速度快、准确率高且范围广。针对无人机手势控制这一应用,后续拟进行如下研究:在控制动作的选取方面,尽可能选取差异更大的动作来保证识别准确,并自主构建室外场景下的数据集;在当前跟踪算法的基础上,考虑遮挡问题并加之人脸识别,同时,更多地挖掘跟踪系统的优势,逐步实现无人机的跟随以及相对位移的控制。

参考文献:

[1] Chen Yi-xin. Application of Kinect based hand gesture recog-

nition interactions[D]. Chengdu:Southwest Jiaotong University,2015. (in Chinese)

[2] Téllez-Guzmán J J,Gomez-Balderas J E,Marchand N,et al. Velocity control of mini-UAV using a helmet system[C]// Proc of Workshop on Research,Education and Development of Unmanned Aerial Systems (RED-UAS),2016:329-335.

[3] Vincenzi D A,Terwilliger B A,Ison D C. Unmanned Aerial System (UAS) human-machine interfaces:New paradigms in command and control[J]. Procedia Manufacturing,2015(3):920-927.

[4] Lupashin S,Hehn M,Mueller M W,et al. A platform for aerial robotics research and demonstration:The flying machine Arena[J]. Mechatronics,2014,24(1):41-54.

[5] Mantecón T,del Blanco C R,Jaureguizar F,et al. New generation of human machine interfaces for controlling UAV through depth-based gesture recognition[C]// Proc of SPIE Defense+ Security,2014,9084(29):139-142.

[6] Pfeil K,Koh S L,LaViola J. Exploring 3d gesture metaphors for interaction with unmanned aerial vehicles[C]// Proc of 2013 International Conference on Intelligent User Interfaces,2013:257-266.

[7] Naseer T, Sturm J,Cremers D. Followme:Person following and gesture recognition with a quadrocopter[C]// Proc of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),2013:624-630.

[8] Monajjemi V M, Wawerla J, Vaughan R, et al. Hri in the sky:Creating and commanding teams of uavs with a vision-mediated gestural interface[C]// Proc of 2013 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS),2013:617-623.

[9] Taylor G W,Fergus R,LeCun Y,et al. Convolutional learning of spatio-temporal features[C]// Proc of Eruopean Conference on Computer Vision(ECCV 2010),2010:140-153.

[10] Ji S, Xu W, Yang M, et al. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence,2013,35(1):221-231.

[11] Karpathy A, Toderici G,Shetty S,et al. Large-scale video classification with convolutional neural networks[C]// Proc of IEEE Conference on Computer Vision and Pattern Recognition, 2014:1725-1732.

[12] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos[C]// Proc of Advances in Neural Information Processing Systems,2014:568-576.

[13] Wang P,Li W,Gao Z,et al. Action recognition from depth maps using deep convolutional neural networks[J]. IEEE Transactions on Human-Machine Systems, 2016,46(4):498-509.

[14] Shi Y,Zeng W,Huang T,et al. Learning deep trajectory descriptor for action recognition in videos using deep neural networks[C]// Proc of 2015 IEEE International Conference on Multimedia and Expo (ICME),2015:1-6.

[15] Zhang K, Zhang L, Liu Q, et al. Fast visual tracking via dense spatio-temporal context learning[C]// Proc of European Conference on Computer Vision(ECCV 2014), 2014:127-141.

[16] Lin Zhe,Jiang Zhuo-lin,Davis L S. Recognizing actions by shape-motion prototype trees[C]// Proc of International Conference on Computer Vision (ICCV),2009:444-451.

[17] Krizhevsky A,Sutskever I,Hinton G E. Imagenet classification with deep convolutional neural networks[C]// Proc of International Conference on Neural Information Processing Systems, 2012:1097-1105.

[18] Jia Y,Shelhamer E,Donahue J,et al. Caffe:Convolutional architecture for fast feature embedding[C]// Proc of the ACM International Conference on Multimedia,2014:675-678.

[19] Quigley M,Conley K,Gerkey B,et al. ROS:An open-source robot operating system[C]// Proc of ICRA Workshop on Open Source Software,2009:1-6.

[20] Dalal N,Triggs B. Histograms of oriented gradients for human detection[C]// Proc of 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05),2005:886-893.

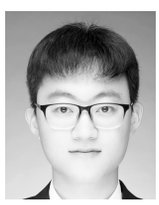
附中文参考文献:

[1] 陈一新. 基于 Kinect 的手势识别技术在人机交互中的应用研究[D]. 成都:西南交通大学,2015.

作者简介:

马乐乐(1993 -),女,陕西榆林人,硕士生,研究方向为无人机和嵌入式系统。**E-mail**:malele@tju.edu.cn

MA Le-le, born in 1993, MS candidate, her research interests include UAV, and embedded system.



李照洋(1991 -),男,吉林通化人,硕士生,研究方向为深度学习、计算机视觉和动作识别。**E-mail**:lizhaoyang@tju.edu.cn

LI Zhao-yang, born in 1991, MS candidate, his research interests include deep learning, computer vision, and action recognition.

董嘉蓉(1995 -),女,福建邵武人,研究方向为深度学习和实时动作检测识别。**E-mail**:dongjr@tju.edu.cn

DONG Jia-rong, born in 1995, her research interests include deep learning, and online action recognition.