Check for updates

研究论文

基于深度模型的航空拍摄视频序列实时人体检测

Nouar AlDahoul、Aznul QalidMd Sabri 和Ali Mohammed Mansoor

马来西亚吉隆坡马来亚大学计算机科学与信息技术学院

通讯事宜请联系 Nouar AlDahoul；nouar.aldahoul@live.iium.edu.my 和 Aznul Qalid Md Sabri；aznulqalid@um.edu.my

收稿日期：2017年7月18日；修订日期：2017年11月18日；接受日期：2018年1月8日；发布日期：201年2月12日 8

学术编辑：Marc Van Hulle

版权所有 © 2018 Nou ar AlDahoul 等。本文为开放获取文章，依据知识共享署名许可协议发布 ，该协议允许不受限 制地在任何媒介中使用、传播和复制，但须恰当注明原作。

视频中的人体检测在各种现实应用中发挥着重要作用。大多数传统方法依赖于使用手工设计特征，这些特征因任务而异，仅适用于特定任务。此外，这些方法极易受到光照变化、摄像机抖动以及目标尺寸变化等动态事件的影响。另一方面，所提出的特征学习方法成本更低、操作更简便，因为它们无需专家知识即可自动生成高度抽象且具有判别性的特征。本文采用自动特征学习方法，将光流与三种不同的深度模型（即监督卷积神经网络（S-CNN）、预训练CNN特征提取器和层次极限学习机）相结合，用于检测由空中平台上的非静态摄像机在不同高度拍摄的视频中的人体。我们在公开且极具挑战性的 UCF-ARG 航空数据集上对这些模型进行训练和测试，并分析它们在训练时间、测试准确率和学习速度方面的表现。性能评估涵盖五种人体动作（挖掘、挥手、投掷、行走和跑步）。实验结果表明，所提出的方法能够有效完成人体检测任务。预训练卷积神经网络的平均准确率为 98.09%。S-CNN 使用 Softmax 时的平均准确率为 95.6%，使用支持向量机（SVM）时为 91.7%。H-ELM 的平均准确率为 95.9%。使用普通中央处理器（CPU）时，H-ELM 的训练时间为 445 秒。S-CNN 在高性能图形处理器（GPU）上完成学习需要 770 秒。

1. 引言

视频（即图像序列）中的人体检测在各种现实生活应用中发挥着重要作用（例如视觉监控和自动驾驶辅助）。由于多种原因，在一系列航拍图像中进行人体检测具有挑战性。其中一个原因是视频帧中人体尺寸的变化。这是由于任务执行过程中安装摄像机的平台高度发生变化所导致的。准确率和较短的训练时间是构建稳健的人类与非人类分类系统时需要考虑的两个重要因素。

特征设计与选择过程非常敏感，因为它在提升模型（本文中的分类器）性能方面发挥着重要作用。手工设计方法依赖特定的领域知识。适用于某一应用的特征，对于其他应用可能不具备判别性。此外，这些特征还需要进行设计由专家通过特征工程流程设计。另一方面，当数据集规模庞大、类别达数千种时，特征学习方法非常有效，因为它们具有领域自适应能力。它们无需人工干预，也无需设计特定特征，便能直接从原始像素中自动学习高度抽象的特征。当缺乏专家知识时，这一优势尤为显著。此外，这些方法能够适应不同场景中的动态事件，具有很强的鲁棒性：卷积神经网络（CNN）是监督式特征学习方法之一，它通过卷积提取空间结构并提供局部表示，通过池化实现平移不变性，并通过归一化适应光照变化。层次极限学习机（H-ELM）是无监督特征学习方法之一，它利用稀疏自编码器提供更鲁棒的特征，使其能够适应数据变化，而无需

预处理。深度模型已被证明擅长人类与非人类分类。

已有多篇论文采用手工设计特征进行人体检测，并证明这些特征对特定任务十分有用且效果良好。在给定图像中，局部二值模式（LBP） [1] 或尺度不变特征变换（SIFT） [2] 被用于特征提取。研究人员利用这些特征训练了不同的分类器，例如随机森林和支持向量机。方向梯度直方图（HOG）也是一个典型示例[3]。HOG保留了已检测空间兴趣点周围的梯度信息 [3]。研究证明，HOG能够有效应对人体外观变化 [3]。研究人员利用HOG特征训练了支持向量机。西部最快行人检测器（FPDW）也是一种依赖人工设计特征的方法 [4]。它在不同尺度上使用HOG。其速度比最先进的方法快一个数量级，同时仍能保持准确率。FPDW的主要优势在于，它只需单帧即可进行人体检测，但需要大量像素。

特征学习方法主要有两种：监督式和无监督。监督式学习器需要输入/输出对来学习特征。无监督学习器仅使用输入来寻找特征。S-CNN 是一种监督式特征学习器，通过迭代微调模型参数来学习。AlexNet 是一种监督卷积神经网络模型，已在大型数据集上完成训练，可用作特征提取器。HELM 是一种无监督学习器，会随机生成输入权重，并通过解析方法计算输出权重。本文将这三种不同的模型用于人类与非人类分类。

卷积神经网络是多层感知器神经网络。它们包含多个卷积层、池化层、全连接层和归一化层。其灵感源于动物视觉皮层的生物组织过程 [5]。网络中各层的权重经过微调，以生成特定类别。网络使用随机梯度下降（SGD）算法进行训练 [5]。卷积神经网络已应用于人脸识别等不同领域 [6]。在该系统中，存在姿态变化、光照变化和部分遮挡等各种挑战因素。在 [6], 中，遗传算法用于寻找最优的卷积神经网络结构。分类采用了支持向量机集成。卷积神经网络还被用于解决自然语言处理中的各种问题，例如句子建模、分类和预测[7–9]。三维卷积神经网络被用于医学领域，以分割磁共振图像中的脑肿瘤 [10]。文献 [11]将卷积神经网络扩展至三维，用于动作识别。系统自动学习时空特征。循环神经网络（长短期记忆网络）经过训练，用于序列分类。本文展示了监督卷积神经网络能够学习判别性特征，并将这些特征应用于分类器。

不同的卷积神经网络模型已使用大型数据集进行训练。其中之一是 AlexNet。目标数据集是 ImageNet。AlexNet 用于将 120 万张高分辨率图像分类为 1000 个不同的类别 [12]。该网络取得了很高的准确率。这个大型神经网络由 6000 万个连接和 65 万个节点构成。它包含五个卷积层、最大池化层、三个全连接层以及一个 1000 类 Softmax 层。采用丢弃正则化方法来减少全连接层中的过拟合。在不同的研究中，这种架构被用作特征提取器。移除了最后一个全连接层，并将各种分类器直接连接到第七层的输出端，以进行分类。本文使用 AlexNet 模型提取特征，然后将其输入分类器。

层次极限学习机是一种高效、快速的深度模型，可用于自动学习特征[13]。HELM 已应用于多种任务，例如数字分类、汽车检测、跟踪和手势识别。研究发现，其训练速度比当前最先进的方法快一个数量级。由于无需迭代微调权重，它能够提高学习速度。在该模型中，偏置和输入权重是随机生成的，而输出权重则通过解析方式计算。本文也采用 HELM 模型，以无监督方式学习特征，然后将其应用于分类器。

多种应用中都包含人体检测这一概念，其中之一是驾驶辅助系统 [14]。系统使用摄像机代替激光雷达（LIDAR），在单帧中检测物体。激光雷达传感器通过反射激光束来测定距离。它价格昂贵、对温度敏感，而且只适用于短距离。摄像机则是更便宜的替代方案，不受距离和温度影响。在该系统中，采用快速深度网络级联进行自动特征学习，以执行人体检测。研究人员使用街道摄像机拍摄的视频，在 Caltech数据集上对其进行了测试。卷积神经网络也用于驾驶辅助系统中的人体检测 [15]。该检测系统安装在车辆上。在这种情况下，需要一种低成本、高准确率的技术。为提升泛化性能，卷积神经网络与随机丢弃和集成推理网络（EIN）相结合 [15]。本文展示了人体检测任务所面临的挑战。图像样本在活动、位置、方向、视角、服装颜色和尺度方面各不相同。摄像机的高度也会随着移动的空中平台而变化。

长期以来，研究人员一直关注利用空中平台进行人体检测。为此，人们采用了特征工程方法。研究者使用基于光流和图表示的框架，从Predator无人机（UAV）的移动摄像机拍摄的帧中提取运动区域 [16]。他们采用动态模板合并相连的图组件，以完整描述该图。另一种人体检测方法基于外观 [17]，旨在从航空图像中检测高空中的人类。该方法采用增强版 Haar特征来表征物体形状，并使用矩形特征处理颜色数据。研究者利用 AdaBoost 二分类器检测小型人类（包括其阴影）。FPDW 与运动目标检测（MOD）相结合，用于 UCF-ARG

[18] 航空数据集。MOD 基于这样一种思路：检测运动背景中的所有运动物体。在这种情况下，稳定质量对于检测运动像素簇至关重要。本研究采用特征学习方法，使用同一个 UCF-ARG 航空数据集进行人体检测。

光流模型已用于各种应用中的人体检测。移动机器人导航任务利用光流，在机器人移动 [19] 时实时检测人类。为实现这一目标，使用了热红外摄像机。文献 [20] 提出了一种结合光流与分类方法的人体运动检测方法。第一阶段采用区域分割，获得几何流后，使用 Bandelet 变换进行特征提取。最后阶段加入监督学习进行分类。本文将光流作为第一阶段，用于检测移动摄像机拍摄的运动物体。运动物体的图像块被保存为训练样本，其中包括人体和非人类图像块。利用这些样本学习特征，随后训练分类器。

已有多篇论文采用手工设计特征进行人体检测。本文不着重研究这些方法，也不尝试比较手工设计方法与基于深度模型的方法。本文旨在研究并比较不同的深度学习方法，以便在具有挑战性的场景中检测人类；在该场景中，摄像机安装在移动的空中飞行物上。本文展示了三种不同深度模型——监督卷积神经网络、预训练卷积神经网络和 HELM——在 UCF-ARG 航空数据集上的特征学习和模型构建效果。三个系统均在第一阶段加入光流模型，以获取作为深度模型输入的训练和测试样本。

我们工作的创新之处如下：

（i）据我们所知，本工作首次利用不同的深度模型对公开的 UCF-ARG 航空数据集进行人体检测。

（ii）研究表明，监督卷积神经网络能够找到对人类和非人类这两个类别具有判别性的最优特征。CNN 的最后一层采用 Softmax 和支持向量机生成分类输出。

（iii）预训练的 AlexNet 卷积神经网络模型已在 ImageNet 数据集上完成训练，可用于识别视觉对象并对 1000 个不同类别进行分类。移除全连接层后，该模型以固定参数作为特征提取器，用于寻找能够区分人类与非人类的判别性特征。

（iv）此外，还讨论了 HELM，以兼顾高准确率与较短训练时间之间的权衡。

(v) 评估了卷积神经网络作为监督式特征学习器、预训练卷积神经网络作为特征提取器，以及 HELM 作为无监督特征学习器在学习速度和准确率方面的表现，共针对五种人体动作（挖掘、挥手、投掷、行走和跑步）。

本文的组织结构如下：第 2 节介绍由光流模型和三个深度模型组成的三种所提出系统。第 3 节讨论实验结果，并从训练速度和准确率方面对其进行分析。第 4 节通过总结本工作的成果，展示所提出系统的有效性。

2. 方法

航空视觉数据首先由安装在空中平台上的移动摄像机采集。在本研究中，我们使用公开且具有挑战性的 UCF-ARG 航空数据集 [21]。视频帧用作光流稳定方法的输入。计算光流后，将生成包含人类和非人类目标的图像块。非人类图像块包括汽车、草地、工具以及背景中的其他不同区域。

随后，本研究所考察的三种深度模型（即监督卷积神经网络、预训练卷积神经网络和 HELM）将以这些图像块作为输入。两种深度模型的输出都是一种有效的目标表征。随后，使用监督卷积神经网络和预训练卷积神经网络中的 Softmax 或支持向量机（SVM），以及 HELM 中的极限学习机（ELM），将这些表征分类为二元类别（人类和非人类）。以下小节将简要介绍所提出的检测系统中使用的各个模块（光流、监督卷积神经网络、预训练卷积神经网络、ELM 和 HELM）。所提出系统的框图如图1所示。

2.1. 基于光流模型的背景稳定 *[22, 23]*。我们提出的系统的第一阶段是背景稳定，该阶段通过光流模型实现。光流用于估计连续帧之间运动矢量的速度和方向。此阶段十分重要，因为它解决了空中平台运动所导致的摄像机运动问题。对运动矢量进行阈值处理和形态学操作（闭运算）后，即可生成特征图像。随后对每幅二值特征图像进行斑点分析，以定位运动物体。接着，在检测到的物体周围叠加绿色边界框。光流在背景稳定中的质量至关重要，因为这是第一阶段；之后将通过深度模型进行特征学习，并将其结果用作分类器的输入。

为求取两帧之间的光流，需使用两个光流约束方程：

$$I_t + I_x h + I_y v = 0, \tag{1}$$

其中，I_t、I_x和 I_y 是帧的时空亮度导数，V是光流的垂直分量，h是水平分量。

当光流应用于整帧时，Horn-Schunck 方法 [23] 求得速度场

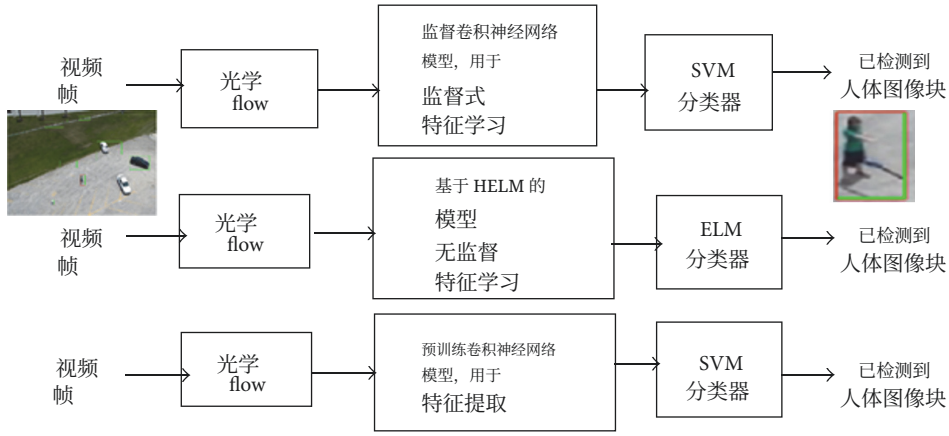


图1：利用航空数据进行人体检测的三种拟议系统框图。

估计每帧中每个像素的 $[\mathrm{h}\ \mathrm{V}]^T$，使其最小化以下方程：

$$E = \iint \left(I_t + I_x h + I_y v\right)^2 dx\,dy + \alpha \iint \left\{\left(\frac{dh}{dx}\right)^2 + \left(\frac{dh}{dy}\right)^2 + \left(\frac{dv}{dx}\right)^2 + \left(\frac{dv}{dy}\right)^2\right\}, \tag{2}$$

其中， α 是用于计算光流的缩放因子， $d\mathrm{h}/dx$ 和 $d\mathrm{h}/dy$ 是光学速度h的空间导数

$$h_{x,y}^{k+1} = \bar{h}_{x,y}^{k} - \frac{I_x\left[I_x \bar{h}_{x,y}^{k} + I_y \bar{v}_{x,y}^{k} + I_t\right]}{\alpha^2 + {I_x}^2 + {I_y}^2},$$
$$v_{x,y}^{k+1} = \bar{v}_{x,y}^{k} - \frac{I_y\left[I_x \bar{h}_{x,y}^{k} + I_y \bar{v}_{x,y}^{k} + I_t\right]}{\alpha^2 + {I_x}^2 + {I_y}^2}. \tag{3}$$

在这些方程中， $[\mathrm{h}_{x,y}^{k}\ \mathrm{V}_{x,y}^{k}]$ 是像素(x, y)处的速度估计，$[\mathrm{h}_{x,y}^{k}\ \mathrm{V}_{x,y}^{k}]$ 是$[\mathrm{h}_{x,y}^{k}\ \mathrm{V}_{x,y}^{k}]$的邻域平均值。当 $k = 0$时，初始速度为0。采用Horn-Schunck方法求解h和V。

2.2. 基于深度模型的人体检测。利用光流稳定场景后，采用深度模型方法进行人体与非人体分类，从而检测人类。

2.2.1. 卷积神经网络 *[5, 6]*。卷积神经网络（CNN）的灵感源自哺乳动物大脑视觉皮层的生物学证据。CNN 是一种具有多个层的监督神经网络，能够在训练阶段提取有效特征。因此，在本研究中，它能够从一系列包含人类的图像（即视频）中有效学习重要的判别性特征。其泛化能力使其即使面对任意一系列图像也能良好运行。CNN 对平移、旋转和尺度变化具有不变性。这得益于 CNN 中的池化层，它们能够减少局部或空间表征特征的数量，并控制过拟合问题。

CNN 对表征失真的输入数据具有鲁棒性，因为深层的结构会在最后一层生成高度抽象的表征。它也能适应类别不平衡数据，这种数据源于数据在不同类别间分配不均。局部感受野、空间下采样和权重共享是使该方法具有强鲁棒性的主要特性。权重共享特性可以减少参数数量，并提升泛化性能。

卷积神经网络结构包括两个主要模块：自动特征学习和分类器。卷积神经网络以原始图像作为输入。通过多个特征图层来学习特征。卷积滤波器应用于原始图像。随后进行下采样，以缩小每一层的尺寸。卷积神经网络有多种不同结构，它们通常在每个卷积层中特征图的数量、卷积滤波器的尺寸、层与层之间的具体连接方式以及激活函数方面有所不同。

（*1*）监督卷积神经网络模型 *[24, 25]*。在监督卷积神经网络模型中，分类层之前连接着多个全连接层。对网络中所有层的权重进行微调，以生成与任务相关的特定类别。学到的特征能够区分这些类别。这种方法称为监督式特征学习。

S-CNN 网络采用带动量的随机梯度下降（SGD）算法进行训练 [25]。在该算法中，通过一步更新参数（权重和偏置），以最小化误差函数（损失）。负梯度上会加上较小的步长。

$$Q\theta_{k+1} = \theta_k - \alpha \nabla E\left(\theta_k\right), \tag{4}$$

其中， k 表示迭代次数， α 表示学习率， θ 表示参数向量， $E(\theta)$表示损失函数， $\nabla E(\theta)$表示梯度。

整个训练数据集被划分成若干小子集，以此来计算梯度。这些子集称为小批量，每次迭代中都会通过迈出一步来更新参数。在一个训练轮次中，整个训练数据集都会以小批量的形式输入。训练

由许多训练轮次组成。训练前应确定小批量大小和训练轮次的数量。有时，梯度下降算法可能会在局部最优解附近振荡；因此，添加动量项以防止这种振荡 [24, 25]。带动量的随机梯度下降更新公式如下：

$$\theta_{k+1} = \theta_k - \alpha \nabla E(\theta_k) + \gamma(\theta_k - \theta_{k-1}), \tag{5}$$

其中，γ 表示当前迭代中上一步梯度的影响。

在误差函数的权重中加入L2正则化项，以减少过拟合 [24, 25]。带正则化的损失函数如下：

$$E_R(\theta) = E(\theta) + \lambda \nabla \Omega(w), \tag{6}$$

其中，w 是权重向量，λ 是正则化系数，$\Omega(w)$是正则化函数。

$$\Omega(w) = \frac{1}{2} w w^T. \tag{7}$$

对于 1-of-k个互斥类别，误差函数为交叉熵函数，如下式所示：

$$E(\theta) = -\sum_{i=1}^{n} \sum_{j=1}^{k} t_{ij} \ln y_j(x_i, \theta), \tag{8}$$

其中，θ 是参数向量，t_{ij} 表示第 i个样本与第 j个类别相关联，$y_j(x_i, \theta)$是第 i个样本的输出，可表示为概率。输出层的激活函数为Softmax函数：

$$y_r(x, \theta) = \frac{\exp(a_r(x, \theta))}{\sum_{j=1}^{k} \exp(a_j(x, \theta))}, \tag{9}$$

其中 $0 \le y_r \le 1, \sum_{j=1}^{k} y_j = 1$。

在本研究中，学习率固定为 0.01。小批量的大小为 300。训练轮次为 30。初始权重由均值为零、标准差为 0.01 的高斯分布生成。初始偏置值为零。动量值为 0.9。L2正则化系数为 0.0001。

(2) 预训练卷积神经网络模型 *[26, 27]*。“AlexNet”卷积神经网络包含 5 个卷积层，已在 ImageNet 数据集上完成训练。该网络在包含 120 万张图像的训练集上进行了 90 轮训练。使用两块 NVIDIA GTX 580 3GB GPU 进行训练，耗时五至六天。训练完成后，该模型用于提取判别性特征。网络中所有层的权重均保持固定。用于将对象分类为 1000 个类别的最后一个全连接层已被移除。该预训练模型能够从每张图像中提取 4096 个特征。这些特征用于训练直接连接到第七层输出端的各种分类器。

2.2.2. 极限学习机 *[28]*。极限学习机（ELM）是一种仅包含一个隐藏层的神经网络。它具有很强的泛化能力和很高的学习效率。这些特性促成了这种学习方法的成功。隐藏层的偏置和权重是随机设定的，而输出权重则通过解析方法计算得出

$$f(x) = \sum_{i=1}^{L} F_i(x, W_i, b_i) \cdot \beta_i, \quad W_i \in R^d, \; b_i, \beta_i \in R, \tag{10}$$

其中，$F_i(\cdot)$ 是第 i 个隐藏节点的激活函数，W_i 是输入权重，b_i 是偏置，β_i 是应用于输出的权重。L 个神经元用于隐藏层。

$$\begin{aligned} \beta &= U^{\dagger} T, \\ \beta &= U^T \left(\frac{1}{\lambda} + U \cdot U^T \right)^{-1} \cdot T, \end{aligned} \tag{11}$$

其中，U 是隐藏层的输出，$U^{\dagger}$ 是矩阵的摩尔－彭若斯广义逆，T 是目标值，λ 是调节系数。

2.2.3. 用于特征学习的分层极限学习机 *[13]*。在处理图像序列等视觉数据时，需要采用极限学习机的深层架构。分层极限学习机（HELM）是一种近期提出的深度模型，用于自动学习特征。它被用作分类器之前的主要模块，以提高系统的准确率。该架构通过无监督的基于 ELM 的稀疏编码器实现自学习特征学习。HELM 能够提高泛化能力并缩短学习时间。基于 ELM 的稀疏编码器采用快速迭代收缩阈值算法（FISTA）构建。该编码器是 HELM 的基本组成部分。通过堆叠多个编码器即可构建深层架构。它通过减少神经元节点的数量，确保改进数据恢复效果并缩短测试时间。有关 HELM 的更多详细信息，请参阅 [13]。HELM 不要求对编码器的权重进行迭代微调。这是其能够显著减少学习/训练时间的主要原因。

3. 实验结果

3.1. 数据集描述 *[29]*。中佛罗里达大学（UCF）发布了三种类型的数据集，分别由空中摄像机、屋顶摄像机和地面摄像机（ARG）拍摄。每个数据集都包含从多个视角拍摄的多种人体动作。这些数据集包含十二人完成的十种动作。本文聚焦于空中摄像机数据集以及五种不同的活动：挖掘、挥手、投掷、行走和跑步。

该航空数据集被认为是最具挑战性的数据集之一，因为图像样本在活动、位置、方向、视角、服装颜色和尺度方面各不相同。此外，摄像机的高度会随着空中移动平台而变化。图 2、图 3 和图 4 展示了各种情况下的人类和非人类目标样本。

本文聚焦于使用空中摄像机拍摄的视频。一台高清摄像机（1920 ×1080 像素，60fps）安装在氦气球的载荷平台上。每种活动有四十八个



图2：光流模型检测到的几个非人体目标样本。图像的实际尺寸各不相同，但已调整为相同尺寸。



图3：光流模型检测到的几个在活动、位置、方向、视角、服装颜色和尺度方面各不相同的人体目标样本。图像的实际尺寸各不相同，但已调整为相同尺寸。

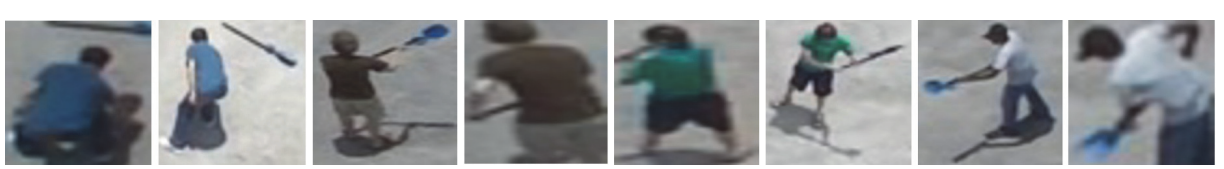


图4：由于空中运动而处于不同高度的几个非人体目标样本。图像的实际尺寸各不相同，但已调整为相同尺寸。

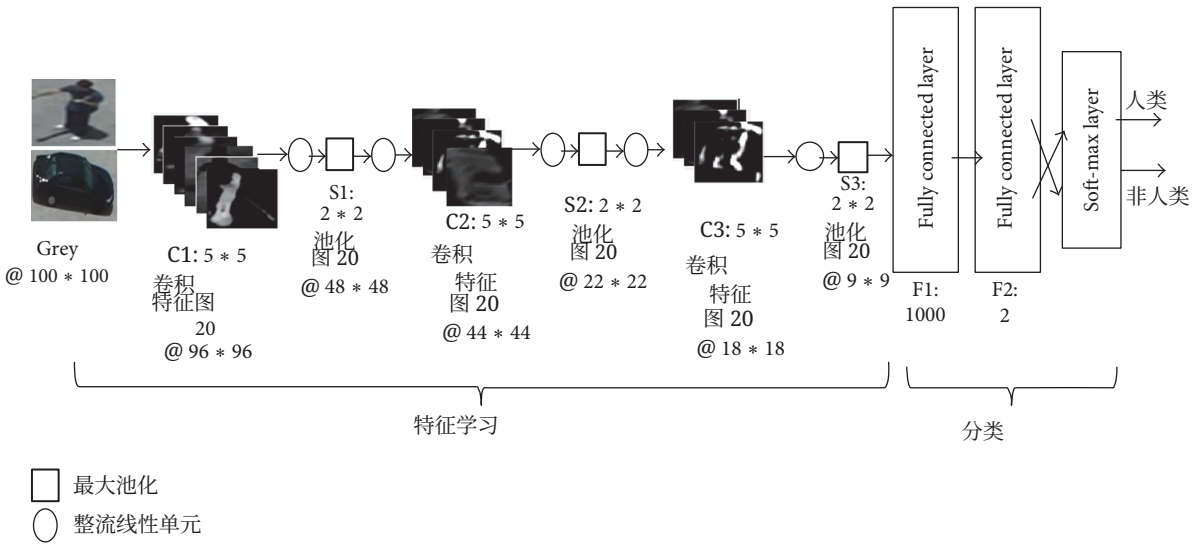


图5：用于人体检测的 S-CNN 模型结构。

视频。每个人分别进行十种动作，每种动作重复四次：拳击、搬运、拍手、挖掘、慢跑、打开和关闭后备箱、跑步、投掷、行走和挥手。动作在不同方向进行，三辆汽车也分别停放在不同方向。

为进行实验，针对五种不同动作评估检测方法的性能：挖掘、挥手、投掷、行走和跑步。这已足够，因为我们的主要目的是使用各种深度模型对航空数据集进行分类，以评估所提出方法的性能，而这些数据集极具挑战性。

本研究使用 48（每项活动的视频）$*5$（活动）$=240$ 个视频作为训练和测试数据。12 个人重复进行相同的活动 4 次。12 人中有 8 人用于训练，4 人用于测试。最终，训练集包含 160 个视频，测试集包含 80 个视频。

为减少训练所用的样本数量，并非使用视频中的所有帧，而是每隔十帧取一帧用于训练。对这些帧应用光流，以找出图像中的运动图像块，然后将这些图像块用于训练。其中一个图像块包含人类。每帧中的所有图像块均用于训练。由于飞行时高度变化，图像块的尺寸并不一致，因此在由深度模型处理并提取判别性特征之前，先将这些图像块调整为预训练卷积神经网络所需的 $227*227$ 像素大小，以及 S-CNN 和 HELM 所需的$100*100$ 像素大小。

我们的训练样本和测试样本存在的问题是，正样本（人类）与负样本（非人类）的数量不平衡。这主要是由光流阶段造成的，该阶段会检测图像中的人类以及其他区域，而这些图像中并没有人类。为进一步说明这一问题，在处理的每一帧中，只为人类提取一个图像块（通过光流），而为其他对象（即非人类）提取多个图像块。样本总数为 26541 个，其中包括 5862 个正样本和 20679 个负样本。然而，最值得注意的是，这个问题并不会影响我们的结果。即使数据集中的正负样本数量不平衡，所提出的系统依然稳健，并能够获得较高的分类准确率。

实验在一台台式计算机上使用 Matlab2016a 实现。对于 HELM，使用了 Intel Core i7 @ 3.5GHz 中央处理器。对于监督卷积神经网络和预训练卷积神经网络，使用了 NVIDIA GeForce GTX 950 图形处理器。两项实验均在 Windows 8.1（64 位）环境下进行。

3.2. 监督卷积神经网络的实现。最优卷积神经网络结构如图5所示。该架构包含17层，包括一个输入层、三个卷积层、三个最大池化层、六个线性整流单元层，最后是两个全连接层和一个 Softmax 层。输入层包含一幅 $100 * 100$ 像素的灰度图像。每个卷积层有20个特征图，包含 $5*5$个卷积滤波器。第一个卷积特征图为 $96*96$。第二个为$44*44$。然后，第三个为 $18*18$。最大池化层的参数为 $2*2$。第一个最大池化层生成 $48*48$ 个特征图。第二个生成 $22*22$个特征图。然后，第三个生成 $9*9$ 个特征图。六个线性整流单元层（ReLU）用于各层之间，以清除负值。最后几层为全连接层，各有1000个节点。2个节点连接到 Softmax 层，以输出两个类别。图6展示了第一卷积层 C1、第二卷积层 C2 和第三卷积层 C3 之后特征图的快照。使用随机梯度下降训练模型，小批量大小为300。从第八层“fc8”提取学习到的特征，该层直接连接到 Softmax 层或支持向量机分类器，以输出两个类别：人类和非人类。

监督式卷积神经网络层如下：

(1) 灰度图像输入层，包含 $100*100$ 像素的灰度图像。(2) 卷积层：20 个大小为 $5*5$ 的特征图。(3) ReLU层。(4) 最大池化层：池化区域大小为 $[2,2]$，并返回四个值中的最大值。(5) ReLU层。(6) 卷积层：20 个大小为 $5*5$ 的特征图。

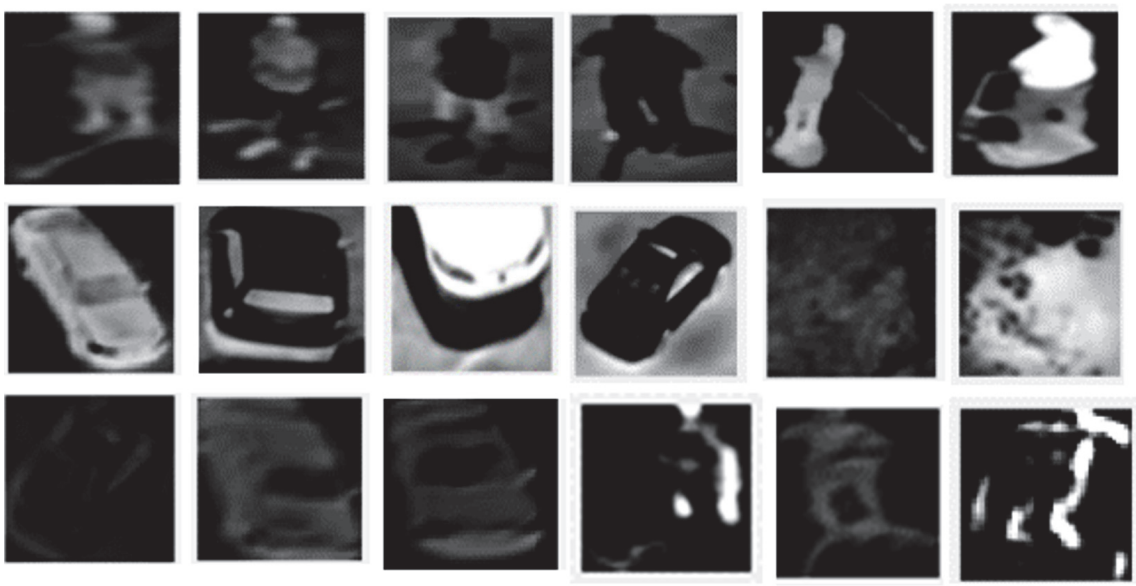


图 6：特征图快照：上面两行显示第一卷积层 C1 20 ∗ (5 ∗ 5) 之后的特征图；最下面一行显示第二卷积层 C2 20 ∗ (5 ∗ 5) 之后的特征图。

(7) ReLU层。(8) 最大池化层：池化区域大小为 $[2, 2]$，并返回四个值中的最大值。(9) ReLU层。(10) 卷积层：20 个大小为 $5 * 5$ 的特征图。(11) ReLU层。(12) 最大池化层：池化区域大小为 $[2, 2]$，并返回四个值中的最大值。(13) 全连接层：包含 1000 个节点。(14) ReLU层。(15) 包含 2 个类别的全连接层。(16) Softmax层。(17) 分类层。

3.3. 预训练卷积神经网络的实现。MatConvNet（面向 MATLAB 的卷积神经网络）是一个工具箱，旨在让卷积神经网络构建模块简单、灵活且易于使用。它提供 MATLAB 函数，用于计算滤波器组的线性卷积、特征池化等。借助 MatConvNet，可以轻松构建新的卷积神经网络架构。“AlexNet”是可从 MatConvNet 下载的模型之一 [27]。运行该模型需要支持 CUDA 的图形处理器。图 7 展示了 AlexNet 模型的架构。输入的彩色 RGB图像会被重采样至 $227 * 227 * 3$ 像素。该网络包括五个卷积层、ReLU层、最大池化层、三个全连接层、一个 Softmax层和一个分类层。最后一个全连接层被移除。从第七层“fc 7”提取特征，并将其直接连接到支持向量机分类器，以生成两个类别：人类和非人类。激活函数用于提取特征的激活函数在图形处理器上计算。使用随机梯度下降训练模型，小批量大小设为 32，以确保卷积神经网络和图像数据能够装入图形处理器内存。提取的特征数量为 4096。使用多类SVM分类器将特征分为人类和非人类类别。

3.4. HELM 实现。最优 HELM 结构如图 8 所示。该架构由三个模块组成，包括两个基于稀疏 ELM 的自编码器和一个基于 ELM 的分类器。输入包含一幅具有100∗100个像素的灰度图像。第一个稀疏自编码器包含 1000 个神经元。第二个稀疏自编码器同样包含 1000 个神经元。最后一个模块是分类器，由隐藏层中的 12,000 个节点组成，这些节点连接到输出层的 2 个节点，以生成两个类别。

3.5. 准确率分析。表1显示了十种测试配置各自的准确率。在每种配置中，采用四人留出交叉验证方法后，将四个人的数据用于测试数据。十种测试配置如下：

测试配置：{p1,p2,p3,p4}，其中 p1、p2、p3 和 p4 是 4 位测试人员。

{1, 2, 3, 4} {5, 6, 7, 8} {9, 10, 11, 12} {1, 3, 5, 7} {2, 4, 6, 8} {1, 4, 7, 10} {2, 5, 8, 11} {3, 6, 9, 12} {1, 5, 9, 12} {1, 6, 11, 12}.

该表比较了所提出的深度模型在准确率方面的平均表现。结果发现，预训练卷积神经网络优于 S-CNN 和 HELM，平均准确率为 98.09%。监督卷积神经网络使用 Softmax分类器时的平均准确率为 95.6%，使用支持向量机（SVM）分类器时为 91.7%。HELM 的平均准确率为 95.9%。

采用四人留出交叉验证模型计算平均准确率。实验重复进行

表1：卷积神经网络 Softmax、卷积神经网络支持向量机、预训练卷积神经网络和 HELM 的准确率比较。

交叉验证（12人中的4人）监督卷积神经网络 + Softmax		监督卷积神经网络 + 支持向量机	预训练卷积神经网络 + 支持向量机	HELM
1	94.0542	91.8966	97.7170	94.7065
2	97.3726	91.9131	97.6115	97.9186
3	96.1542	92.9835	98.5988	96.0008
4	97.1056	92.8199	97.8509	97.4410
5	94.6982	91.1981	97.9344	94.9851
6	95.7974	91.2816	98.4077	95.7583
7	91.7131	87.8823	98.0749	92.3971
8	96.5649	92.7828	98.5311	96.5765
9	97.0207	93.6128	98.0556	97.3239
10	95.5184	90.6524	98.1257	95.7467
平均准确率	95.5999%	91.7023%	98.0908%	95.8855%

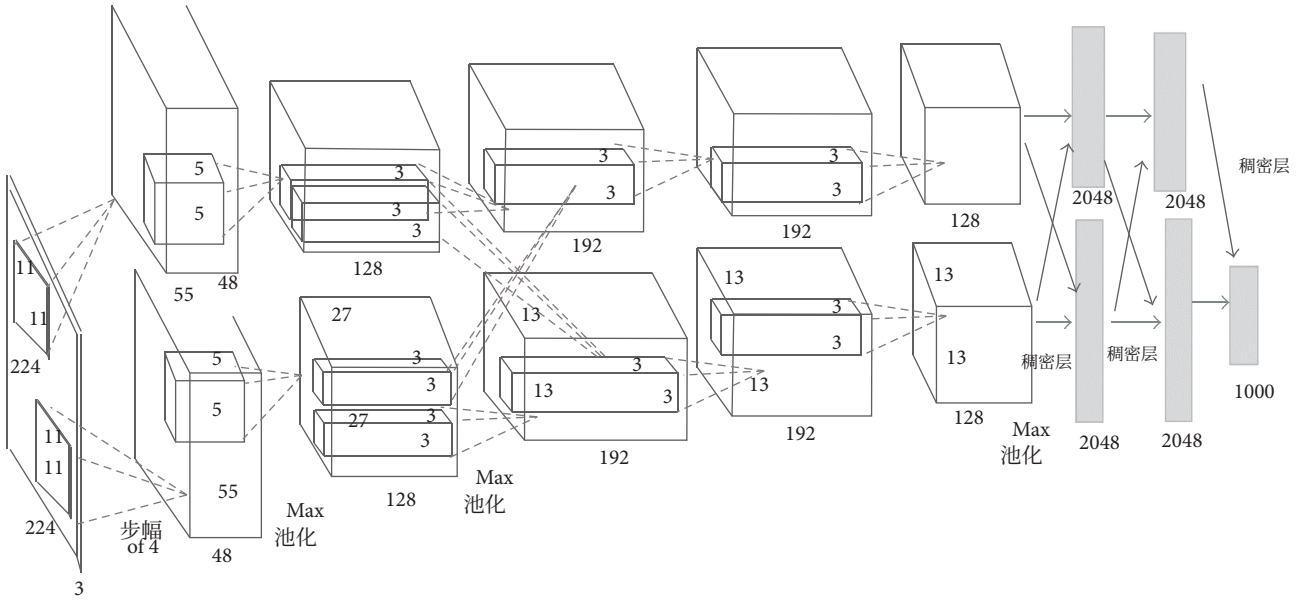


图 7：基于 AlexNet [12] 的预训练卷积神经网络模型架构。

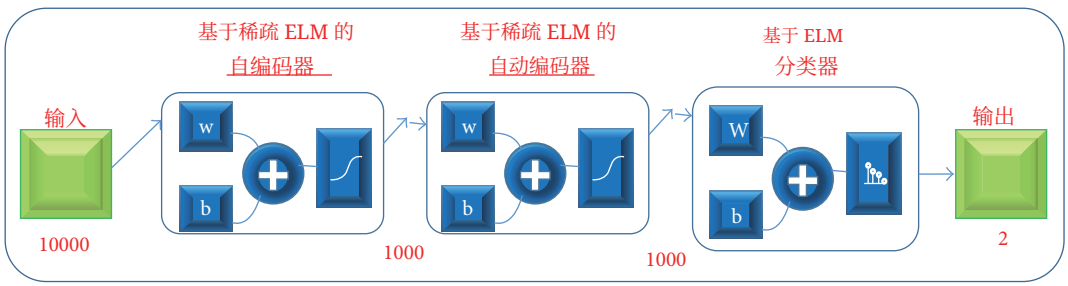


图 8：基于分层 ELM 的人体检测深度模型。

次，并将这些实验的平均值作为最终结果。

每次针对每项活动，使用32个视频作为训练数据，16个视频作为测试数据。测试视频由四名新人员组成，他们以相同方式执行该活动4次。这四名人员不会出现在训练数据中。此过程针对十种不同的配置重复进行，每种配置从十二名人员中选取四名新人员，然后计算平均值。表2展示了 UCF-ARG 航空数据集中训练样本和测试样本的数量。

最后一种配置 {1,6,11, 12} 作为测试数据时的混淆矩阵如图9所示。矩阵显示了十二人中四人的测试准确率。每个人重复五个动作四次。因此，测试视频数量为 240 个中的 80 个。

表2：UCF-ARG 航空数据集中训练数据和测试数据的数量。

交叉验证（12人中的4人）	训练数据数量	测试数据数量
1	18569	7972
2	17749	8792
3	16764	9777
4	18071	8050
5	17374	8714
6	18006	7662
7	15360	10233
8	16733	8646
9	15607	9566
10	16183	8323

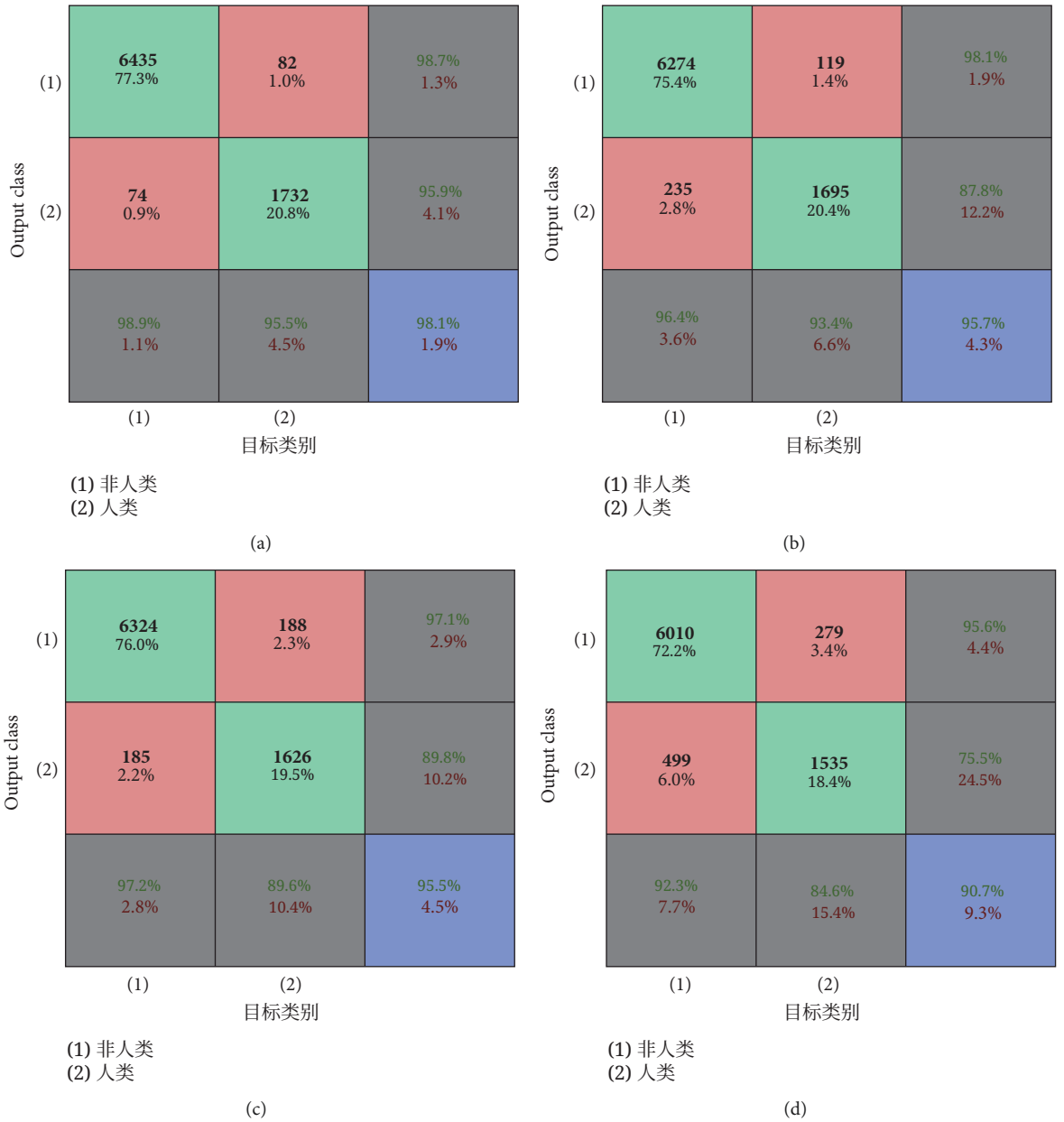


图9：（a）预训练卷积神经网络的测试准确率，（b）HELM 的测试准确率，（c）监督卷积神经网络 + Softmax 的测试准确率，以及（d）监督卷积神经网络 + 支持向量机的测试准确率。

表 3：监督卷积神经网络与 HELM 的训练和测试速度比较。

方法	训练时间 [s]	测试时间 每个 样本 [s]
S-CNN（图形处理器）	770	0.045
HELM（中央处理器）	445	0.1

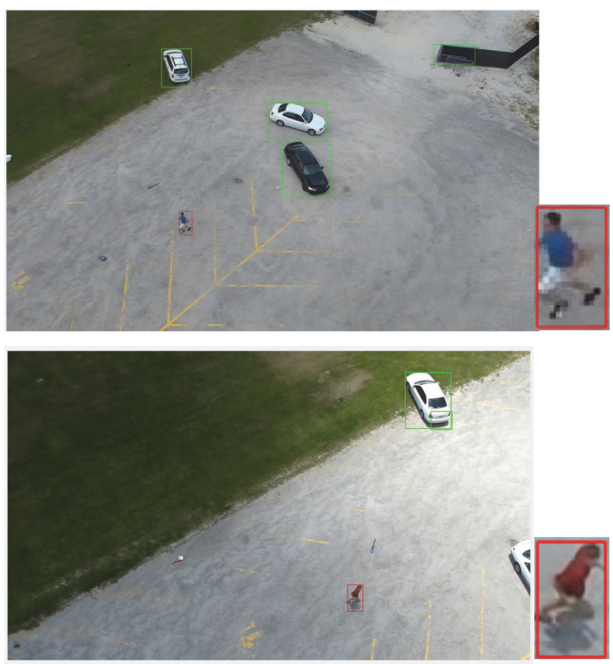


图10：UCF-ARG 数据集中的人类跑步场景：图像中的绿色边界框是使用光流后得到的结果。红色边界框是应用基于深度模型的人体分类后得到的结果。

3.6. 速度分析。表3展示了两种特征学习器（HELM 和监督卷积神经网络）在训练和测试时间方面的差异。尽管 HELM 使用中央处理器，而 S-CNN 使用图形处理器，所提出的基于 ELM 的 HELM 检测器在训练速度方面仍优于 S-CNN。这是因为它利用层次极限学习机作为一种快速深度模型，无需反复微调权重。HELM 的这一优势使其有望在低成本嵌入式系统上实现实时运行。

图10、图11和图12展示了带有绿色和红色边界框的原始帧。绿色边界框由光流模型生成。光流检测出不同的物体，并用绿色框标出。使用深度模型仅检测人类后，会绘制红色边界框。模拟视频已上传至 YouTube 网站。请参阅以下链接：

https://www.youtube.com/channel/UCvEvheiIvcV n _ _ NvK3l4vQA.

该模型针对五种活动进行训练（挖掘、挥手、投掷、行走和跑步）。在图13中，多个

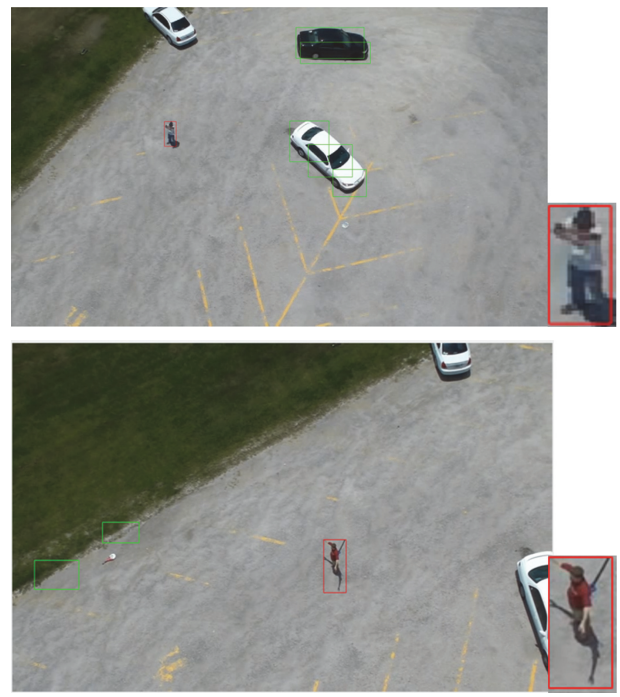


图11：UCF-ARG 数据集中的人类挥手：每幅图像中的绿色边界框是使用光流后得到的结果。红色边界框是应用基于深度模型的人体分类后得到的结果。

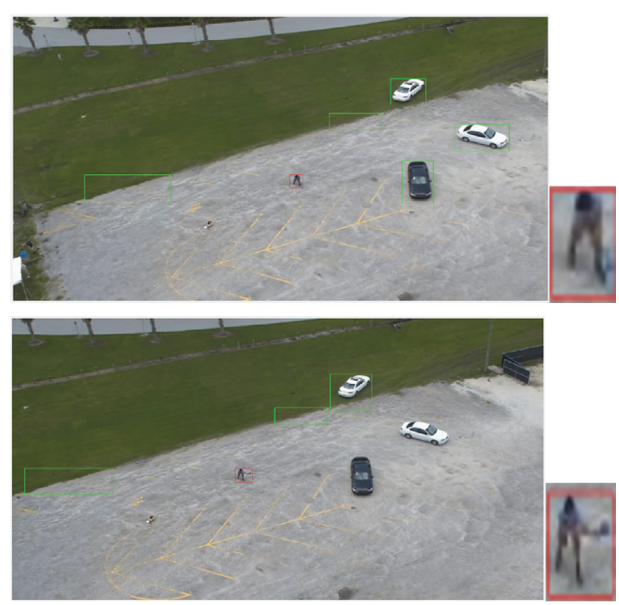
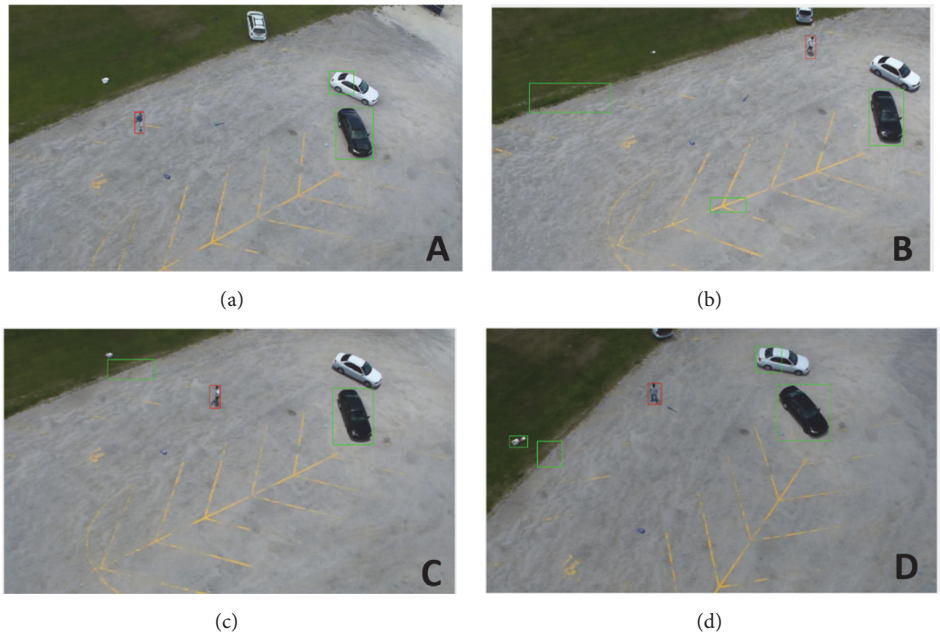


图12：UCF-ARG 数据集中的人类挖掘：每幅图像中的绿色边界框是使用光流后得到的结果。红色边界框是应用基于深度模型的人体分类后得到的结果。

对人体检测进行测试。这种泛化能力适用于检测从事各种未训练过的活动的人，例如拍手、拳击、搬运和慢跑。图14显示了多个处于



(a) (b)

(c) (d)

图13：检测 UCF-ARG 数据集中的不同未训练动作的人类（(a) 慢跑。(b) 搬运。(c) 拳击。(d) 拍手）。每幅图像中的绿色边界框为使用光流后的结果。红色边界框为应用基于深度模型的人体分类后的结果。

同一帧中的人。每个人都在进行不同的活动。所提出的模型能够检测帧中的所有人类，并在他们周围绘制边界框。

4. 讨论与结论

本文展示了三种用于人体检测的深度模型的实现，并对其在五种人体动作（挖掘、挥手、投掷、行走和跑步）上的性能进行了评估。这些模型使用公开 UCF-ARG 航空数据集进行训练和测试。该数据集的挑战在于，人体图像块的大小会随移动空中平台的高度而变化，而且同一视频中的人类具有多个视角。此外，由于使用了光流，训练数据和测试数据中的正负样本数量并不均衡。

Th本研究的结果可总结如下 ：

（1）在应用深度模型进行人类与非人类分类之前，作为第一阶段，我们提出的系统对稳定方法（光流）的质量要求很高。

(2) 所提出的系统解决了高准确率与检测速度之间的权衡问题。研究发现，预训练深度模型的表现优于 S-CNN 和 HELM，平均准确率达到 98.09%。预训练卷积神经网络和 S-CNN 模型均在图形处理器上实现。HELM 比 S-CNN 更节省时间，因为它无需对权重进行迭代微调。这是因为权重是随机生成的。由此

训练时间缩短至使用普通中央处理器时的 445 秒。同时，它取得了 95.9% 的良好平均分类准确率。对于需要高准确率且成本低廉的嵌入式系统，建议使用 HELM。

所提出系统的优势如下：

（i）所提出的系统能够自动检测人类，无需手动设置检测阈值来选择真正率最高的阈值。（ii）两个深度模型的泛化能力都能够在不属于训练数据的全部80个视频中准确检测人类。（iii）所提出的系统在测试实时采集的视频时能够实现实时性能，因为光流模型只利用连续的两帧来寻找运动。此外，深度模型仅需单帧即可将光流图像块分类为人类或非人类（即人体检测）。（iv）所提出的深度模型能够适应各种活动、位置、方向、视角、服装颜色、尺度和高度。

所提出系统的缺点在于其高度依赖光流处理阶段的质量。在人体检测的整个处理流程中加入跟踪，可以减少这种依赖并提高整体准确率。未来，我们将整合跟踪功能，利用最初提取的人体周围训练区域作为正样本，并将其他区域作为负样本。由于该结果具有很高的准确率和效率，我们将利用这些结果

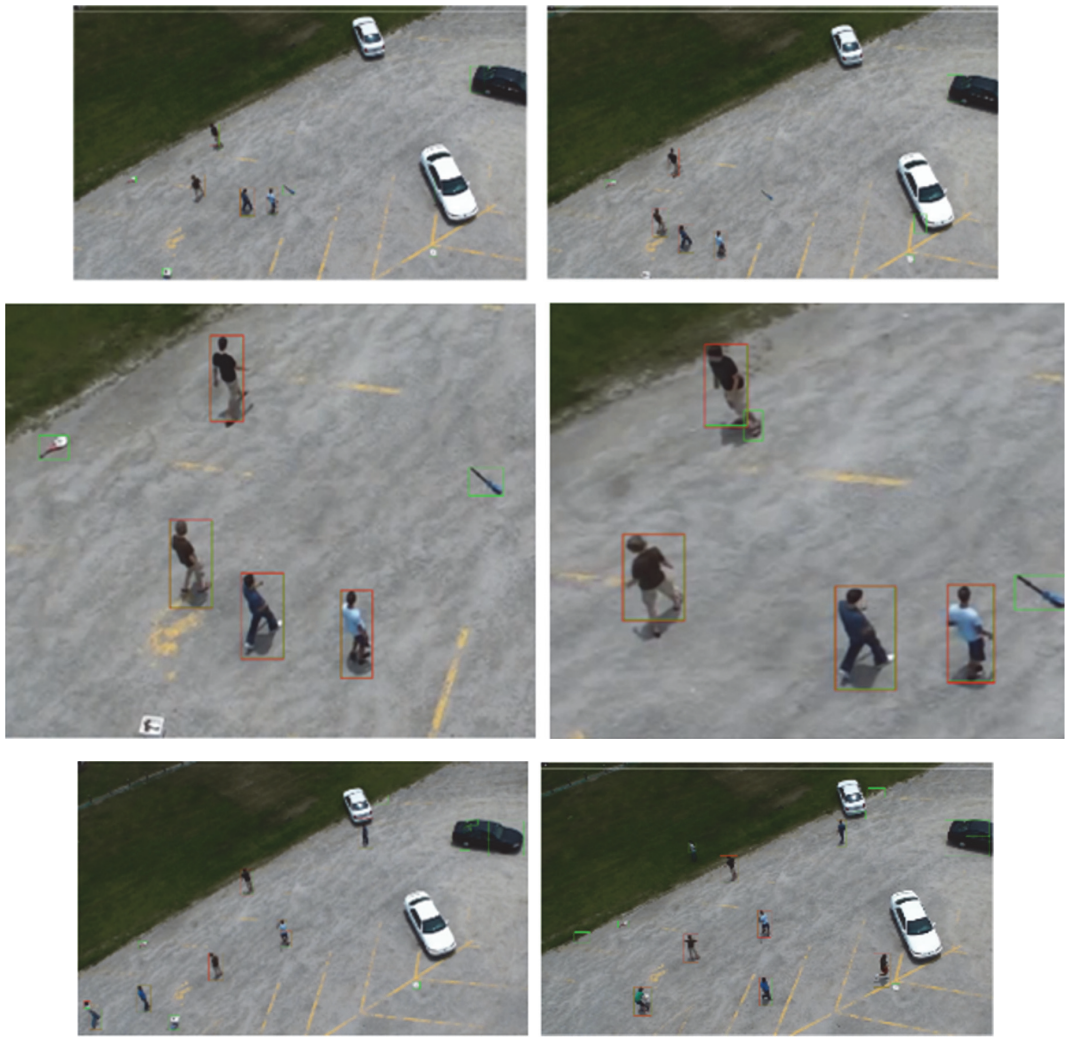


图14：使用 UCF-ARG 数据集检测具有不同动作的多个目标人物：每张图像中的绿色边界框是使用光流后的结果。红色边界框是基于深度模型进行人体分类后的结果。第一行和第三行是原始帧。第二行是第一行中人物放大后的图像。

用于本文展示的人体检测，以便在人体动作识别中将每项活动映射到特定的动作类别。

利益冲突

作者声明不存在任何利益冲突。

致谢

本研究由马来亚大学研究资助（UMRG，资助编号 RP030A-14AET）和基础研究资助（FRGS，资助编号 FP061-2014A）资助。

参考文献

[1] X. Wang、T. X. Han 和 S. Yan，“一种具备部分遮挡处理能力的 HOG-LBP 人类检测器”，收录于*IEEE* 第 *12* 届国际计算机视觉会议（*ICCV*）论文集，第 32–39 页，2009 年。

[2] D. G. Lowe，“来自尺度不变关键点的独特图像特征”，国际计算机视觉期刊，第 60 卷，第 2 期，第 91–110 页，2004 年。[3] N. Dalal 和 B. Triggs，“用于人体检测的方向梯度直方图”，载于计算机学会计算机视觉与模式识别会议论文集（*CVPR '05*），第 1 卷，第 886–893 页，2005 年 6 月。[4] P. Dollar、S. Belongie 和 A. P. Perona，“西部最快的行人检测器”，载于英国机器视觉会议论文集，第 68.1-68.11 页，BMVA 出版社，2010 年。[5] Y. Lecun、L. Bottou、Y. Bengio 和 P. Haffner，“应用于文档识别的基于梯度的学习”，载于*IEEE* 论文集，第 86 卷，第 11 期，第 2278–2324 页，2002 年。[6]A. Rikhtegar、M. Pooyan 和 M. T. Manzuri-Shalmani，“用于人脸识别应用的卷积神经网络遗传算法优化结构”，*IET* 计算机视觉，第 10 卷，第 6 期，第 559–566 页，2016 年。[7] N. Kalchbrenner、E. Grefenstette 和 P. Blunsom，“用于句子建模的卷积神经网络”，载于论文集

计算语言学协会第 *52* 届年会，*ACL 2014*，第 655–665 页，2014 年 6 月。[8] Y. Kim，“用于句子分类的卷积神经网络”，载于*2014* 年自然语言处理中的经验方法会议论文集，*EMNLP 2014*，第 1746–1751 页，qat，2014 年 10 月。[9] R. Collobert 和 J. Weston，“自然语言处理的统一架构：具有多任务学习的深度神经网络”，载于第 *25* 届国际机器学习会议论文集，第 160–167 页，ACM，2008 年 7 月。[10] A. Casamitjana、S. Puch、A. Aduriz 和 V. Vilaplana，“用于脑肿瘤分割的 3D 卷积神经网络”，载于*MICCAI* 多模态脑肿瘤图像挑战赛论文集，2016 年。[11] M. Baccouche、F. Mamalet、C. Wolf、C. Garcia 和 A. Baskurt，“用于人体动作识别的序列深度学习”，载于《人类行为理解》，第 29–39 页，施普林格，2011 年。[12] A. Krizhevsky、I. Sutskever 和 G. E. Hinton，“使用深度卷积神经网络进行 ImageNet 分类”，载于第 *26* 届神经信息处理系统年会论文集（*NIPS ’12*），第 1097–1105 页，美国内华达州太浩湖，2012 年 12 月。[13] J. Tang、C. Deng 和 G. B. Huang，“用于多层感知器的极限学习机”，*IEEE* 神经网络与学习系统汇刊，第 27 卷，第 4 期，第 809–821 页，2016 年。[14]A. Angelova、A. Krizhevsky、V. Vanhoucke、A. Ogale 和 D. Ferguson，“利用深度网络级联进行实时行人检测”，载于英国机器视觉会议论文集（*BMVC*），2015 年。[15]H. Fukui、T. Yamashita、Y. Yamauchi、H. Fujiyoshi 和 H. Murase，“基于带集成推理网络的深度卷积神经网络的行人检测”，载于*IEEE* 智能车辆研讨会论文集，*IV 2015*，第 223–228 页，kor，2015 年 7 月。[16] I. Cohen 和 G. Medioni，“用于视频监控的运动物体检测与跟踪”，载于*1999* 年 *IEEE* 计算机学会计算机视觉与模式识别会议论文集（*CVPR’99*），第 319–325 页，1999 年 6 月。[17]F. Schmidt 和 S. Hinz，“一种针对航空图像序列调优的人体检测与跟踪方案”，载于计算机科学讲义（*LNCS*）丛书中的摄影测量图像分析，第 6952 卷，*270*，第 270-257 页，2011 年。[18] A. W. Eekeren、J. Dijk 和 G. Burghouts，“从空中平台检测与跟踪人类”，载于*SPIE 9249*，《光电与红外系统：技术与应用》，第 9249 卷，2014 年。[19]A. Fern´andez-Caballeroa、J. C. Castillo、J. M. Cantos 和 R. M. Tom´as，“移动机器人红外摄像机中基于光流或图像减法的人体检测”，机器人学与自主系统，第 58 卷，第 12 期，第 1273–1281 页，2010 年。[20]H. Han 和 M. Tong，“基于光流和稀疏几何流的人体检测”，载于*2013* 年第七届图像与图形国际会议论文集，*ICIG 2013*，第 459–464 页，中国，2013 年 7 月。[21] http://spie.org/x41092.xml?ArticleID=x41092。[22] J. L. Barron、D. J. Fleet 和 S. S. Beauchemin，“光流技术性能”，国际计算机视觉期刊，第 12 卷，第 1 期，第 43–77 页。[23] B. K. P. Horn 和 B. G. Schunck，“确定光流”，人工智能，第 17 卷，第 1-3 期，第 185–203 页，1981 年。[24]C. M. Bishop，《模式识别与机器学习》，施普林格，美国纽约州纽约，2006年。[25] K. P. Murphy，《机器学习：概率视角》，麻省理工学院出版社，美国马萨诸塞州剑桥，2012年。[26] J. Deng、W. Dong 和 R. Socher，“ImageNet：一个大规模分层图像数据库”，载于*2009*年*IEEE*计算机视觉与模式识别会议（*CVPR*）论文集，第248–255页，美国佛罗里达州迈阿密，2009年6月。[27] A. Vedaldi 和 K. Lenc，“MatConvNet——用于MATLAB的卷积神经网络”，arXiv预印本 arXiv:1412.4564。[28] G. B. Huang、Q. Y. Zhu 和 C. K. Siew，“极限学习机：理论与应用”，*Neurocomputing*，第70卷，第1–3期，第489–501页，2006年。[29] K. Reddy，UCF-ARG数据集（http://crcv.ucf.edu/data/UCF- ARG.php）。