Journal of Security and Safety Technology 安防技术, 2021, 9(4), 19-25
Published Online December 2021 in Hans. http://www.hanspub.org/journal/jsst
https://doi.org/10.12677/jsst.2021.94004

Hans汉斯

一种基于SSD目标检测算法的安全帽识别视频监控系统

范恒搏，马士伟，乐　文，王　建

中建地下空间有限公司，四川 成都

收稿日期：2021年9月17日；录用日期：2021年10月19日；发布日期：2021年11月2日

摘　要

视频监控技术源自人工智能的一个分支。它可以使用一台计算机，来自动分析视频图像源，从中识别并提取有用的关键信息，并自动控制机器执行相应的操作。本文是一种基于SSD目标检测算法的视频监视系统，用于识别安全帽的特定数据。目的是能够以更智能的方式进行巡逻和对事件做出响应，而无需投入大量人力资源停留在显示器前。

关键词

视频监控技术，SSD目标检测算法，安全帽

Safety Helmet Recognition Video Surveillance System Based on SSD Target Detection Algorithm

Hengbo Fan, Shiwei Ma, Wen Le, Jian Wang

China Construction Underground Space CO., LTD., Chengdu Sichuan

Received: Sep. 17th, 2021; accepted: Oct. 19th, 2021; published: Nov. 2nd, 2021

Abstract

Video surveillance technology comes from a branch of artificial intelligence. It can use a computer to automatically analyze the video image source, identify and extract useful key information, and automatically control the machine to perform corresponding operations. This paper is a video surveillance system based on SSD target detection algorithm, which is used to identify the specific

文章引用: 范恒搏, 马士伟, 乐文, 王建. 一种基于 SSD 目标检测算法的安全帽识别视频监控系统[J]. 安防技术, 2021, 9(4): 19-25. DOI: 10.12677/jsst.2021.94004

data of safety helmet. The goal is to patrol and respond to events in a more intelligent way without investing a lot of human resources in front of the display.

Keywords

Video Surveillance Technology, SSD Target Detection Algorithm, Safety Helmet

1. 引言

近年来，数字和网络视频监视系统相对于传统系统的优势已变得越来越明显。它们的高度标准化，开放性，集成性和灵活性为安防建筑等其他行业的发展提供了更广阔的发展空间，其中，智能视频监控是网络视频监控领域的主要应用开发方向之一。

在传统系统甚至网络视频监视系统中，已经使用了诸如运动检测和警报之类的图像技术，但是误报率和误报率很高，使得该技术在实际使用中无效。但是，智能视频监控系统能够识别不同的对象，在监控屏幕中发现异常情况，并以最快和最佳的方式发出警报并提供有用的信息，从而可以更有效地帮助安全人员处理危机，并最大程度地减少误报和误报。

近年来，深度学习在各个领域取得了巨大的进展，目标检测领域也不例外。当前的基于深度学习的目标检测方法存在两种方式，一种是以 SSD [1]和 YOLO [2]为代表的 one-stage，另一种是以 R-CNN [3]和 Fast RCNN [4]以及 Faster RCNN [5]为代表的 two-stage。与 Faster RCNN 相比，SSD 这类 one-stage 方式虽然准确率低一些，但是其可以在大部分场景下达到实时的结果，因此非常适用于安全帽识别视频监控系统。

2. 基于 SSD 目标检测的安全帽识别的视频系统

安全帽识别的视频监视系统模型具有三个模块：神经网络模块，识别器模块和后处理模块。神经网络模块中的基本网络使用 VGG-16 [6]网络，然后在 VGG-16 的基础上添加卷积层以获得更多的特征图进行检测。主要功能是用作特征提取器来提取图像的特征。识别器模块的主要功能是基于神经网络模块提取的特征，生成包含商品位置和类别信息的候选框(此处使用卷积实现)；最后的后处理模块的主要功能是对识别器提取的候选帧进行解码和过滤，以输出最终的候选帧。

该系统的整个过程也分为三个步骤：首先，输入一组数据流(图片或视频)，然后将它们输入到预先训练的[7] [8]分类网络中，以获得不同的特征图大小。传统的 VGG-16 是修改后的网络，主要将 VGG-16 的 FC6 和 FC7 层转换为卷积层(conv6 和 conv7)，删除所有 conv8 层，并添加 ATROUS 算法[9] (空洞算法)。然后提取层 con4_3，conv7，conv8_2，conv9_2，conv10_2 和 conv11_2 的特征图，然后在这些特征图层上方的每个点构造 6 个不同比例的 boxes，然后分别对其进行检测和分类以生成多个 boxes。最后，将从不同特征图获得的 boxes 进行组合，然后使用 NMS [10] (非极大值抑制)方法抑制一些重叠或不正确的 boxes，以生成最终的 boxes (检测结果)。

3. SSD 目标检测模型

在我们的网络模型中：输入图像的大小标准化为 300 × 300，系统的基本网络结构是 VGG-16 网络。其结构图如图 1 所示。

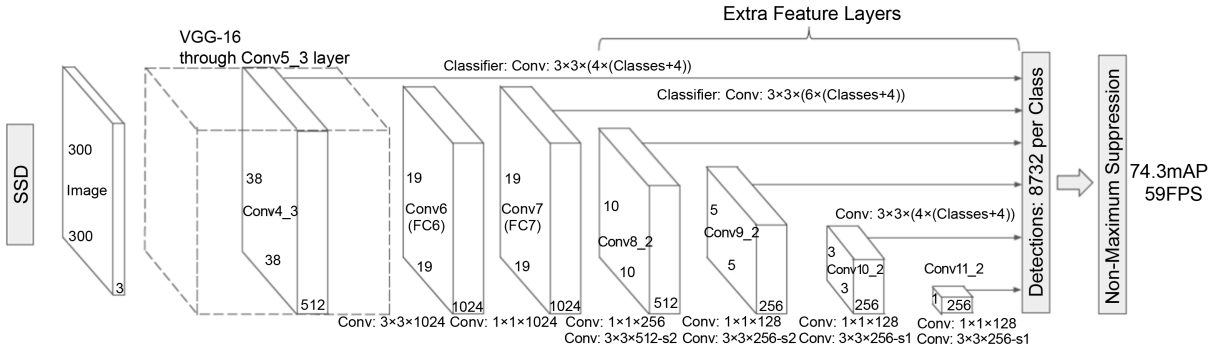

Figure 1. SSD network model structure
图 1. SSD 网络模型结构

VGG-16 网络由一系列 3 × 3 卷积顺序连接组成。在 conv5_3 层卷积之前，将合并 stride = 2 的最大值，因此该层的输出长度和宽度比原始输入减少 16 倍。因此该层的输出长度和宽度为[300/2] = [18.75] = 19，通道为 512，即基本网络 VGG-16 的输出大小为 512 × 19 × 19。其中，是否输出标记为 Y 的列被输出并发送到识别器，即最终识别器接受不同大小(5 + 1) = 6 (5 个附加输出层和 1 个基本网络输出)的特征图，分别是 10 × 10、5 × 5、3 × 3、1 × 1 和两个 19 × 19。

conv4_3，conv_7，conv8_2，conv7_2，conv8_2，conv9_2，conv10_2 和 conv11_2 具有不同大小的功能图。目的是能够准确地检测出不同尺度的物体，因为在低级特征图中，接收场相对较小，而高级接收场较大，在不同特征图中的卷积可以达到多目标的目的。

在识别器模块中：该系统的识别器是使用卷积层构造的，并且卷积层的大小为(box numbers * (4 + class numbers)) * channel * 3 * 3，其中 box numbers 生成特征图上的网格点识别帧的数量，class numbers 是类别的数量，包括背景类别，例如，一个 4 × 4 的特征图，总共 4 × 4 = 16 个网格点，每个上有 3 个候选点格，即 box numbers = 3，类别中的信息中有 p 个数据，class numbers = p。识别器处理完 C × W × H 特征图后，它变成 (box numbers * (4 + class numbers)) * W * H，其中包含 W * H * box numbers 候选框。

在后处理模块中：第一个后处理是解析候选帧中的数据。每个候选帧由 4 + class numbers 数据组成：四个位置信息 x，y，w，h 和 class numbers 类别信息。如下：

$$\begin{aligned} G_x &= P_w \times \frac{P_x + sigmoid(x)}{F_x} \\ G_y &= P_h \times \frac{P_y + sigmoid(y)}{F_y} \\ G_w &= P_w \times \left(D_w \times \mathrm{e}^{sigmoid(w)}\right) \\ G_h &= P_h \times \left(D_h \times \mathrm{e}^{sigmoid(h)}\right) \end{aligned} \tag{1}$$

其中 G_x，G_y，G_w，G_h：所标识物品的宽度坐标，中心点的高度坐标以及物品的高度和宽度。P_w，P_h 是输入图像的宽度和高度，P_x，P_y 是候选帧所在的网格的坐标。候选框所在的要素地图的宽度和高度。D_w，D_h 是相应默认框的默认标准化宽度和高度。有关类别信息，则选择最大的类别：

$$\begin{aligned} cls &= \arg\max_{i \in class\_id} \left(class\_feature_i\right) \\ conf &= \max\left(class\_feature_i\right) \end{aligned} \tag{2}$$

Default box 生成规则为：

以要素地图上每个点的中点为中心(offset = 0.5)，生成一系列同心的默认框(然后，中心点的坐标将被逐步乘以 steps，这等效于从要素地图位置进行映射回到原始地图位置)。

使用不同大小的 m (在 SSD300 中为 $m = 6$)特征图进行预测。最低特征图的比例值是 $s_{\min} = 0.2$，最高是 $s_{\max} = 0.95$。其他层通过以下公式计算：

$$s_k = s_{\min} + \frac{s_{\max} - s_{\min}}{m-1}(k-1), k \in [1, m] \tag{3}$$

使用不同的比率值[1, 2, 3, 1/2, 1/3]，通过以下公式计算默认框的宽度 w 和高度 h：

$$\text{width}\left(w_k^a = s_k\sqrt{a_r}\right) \text{ and } \text{height}\left(h_k^a = s_k / \sqrt{a_r}\right) \tag{4}$$

对于 ratio = 0，指定的比例如下，即有 6 个不同的默认框：

$$s_k' = \sqrt{s_k s_k + 1} \tag{5}$$

我们的模型的损失函数分为两部分：计算相应的默认框和目标类别的置信损失以及相应的位置回归。

从匹配到地面真相的默认框的数量在哪里；alpha 参数用于调整置信度损失和位置损失之间的比率。默认 alpha = 1。

位置回归使用平滑损失，损失函数为：

$$\begin{aligned} L_{loc}(x,l,g) &= \sum_{i\in Pos}^{N}\sum_{m\in\{cx,cy,w,h\}} x_{ij}^k smooth_{L1}\left(l_i^m - \hat{g}_j^m\right) \\ \hat{g}_j^{cx} &= \left(g_j^{cx} - d_i^{cx}\right)/d_i^w \\ \hat{g}_j^{cy} &= \left(g_j^{cy} - d_i^{cy}\right)/d_i^h \\ \hat{g}_j^w &= \log\left(\frac{g_j^w}{d_i^w}\right) \\ \hat{g}_j^h &= \log\left(\frac{g_j^h}{d_i^h}\right) \end{aligned} \tag{6}$$

置信度损失是典型的 softmax 损失：

$$\begin{aligned} L_{conf}(x,c) &= -\sum_{i\in Pos}^{N} x_{ij}^p \log\left(\hat{c}_i^p\right) - \sum_{i\in N}\log\left(\hat{c}_i^0\right) \\ \text{where } \hat{c}_i^p &= \frac{\exp\left(c_i^p\right)}{\sum_p \exp\left(c_i^p\right)} \end{aligned} \tag{7}$$

第二步是在处理后使用 NMS (非最大抑制)过滤候选框：当两个候选框的 IOU 超过框阈值时，将丢弃置信度低的候选框。

4. 实验结果与分析

在本文中，对给定的数据集进行了实验。为了证明该方法的有效性，本文选择了几种当前主流的深度学习对象检测算法进行比较，包括 Fast RCNN 和 YOLO。在精度比较方面，我们选择 *mAP*，即均值平均精度。计算公式如下：

$$mAP = \frac{1}{|Q_R|}\sum_{q\in Q_R} AP(q) \tag{8}$$

其中 *AP* 代表平均精度，这是通过精度和召回率获得的。精度代表预测的 *n* 个阳性样本中真实阳性的数量与预测的 *n* 个阳性样本的比例。召回率表示预测的 *n* 个阳性样品中真实阳性与模型测试的所有样品中真实阳性的比例。在多标签图像分类中，首先将训练后的模型获得的所有测试样本的置信度得分按降序排序，然后逐步选择 *n* 张图片，计算每张图片的精度和召回率，并计算出与各召回率相对应的精度。通过平均值即 *AP* 求出的平均精度，即对应于一个类别的平均精度，他表示某个类别中与模型对应的 *AP* 的平均值，然后可以得到 *mAP*。

从表 1、表 2 中可以看出，与其他方法相比，SSD 目标检测算法的 mAP 提高了约 5%。

Table 1. Compared mean average precision table
表 1. 均值平均精度比较表

Method	number of train images	number of test images	mean average precision
Fast R-CNN	2493	200	0.48
YOLO	2493	200	0.38
SSD	2493	200	0.49

Table 2. Compared average precision (no security) table
表 2. 平均精度(无异常)比较表

Method	number of train images	number of test images	average precision (no security)
Fast R-CNN	2493	200	0.45
YOLO	2493	200	0.41
SSD	2493	200	0.47

细节点：

估计的运行时间：从接收输入到启动算法的正常运行时间分为两部分。第一部分是五秒钟的视频流质量判断。如果视频流正常，则算法开始运行。从打开到返回总共需要大约二十秒钟。

输入：该算法有两个主要输入。第一个是 RTSP 流的地址；第二个是返回预警信息时的安全令牌。

输入数据来自从工地摄像机实时传输的视频流数据。传输数据的质量与摄像机的质量，网络速度和视频流的设置有关。

数据格式为 RTSP 视频流格式。RTSP 的一般格式如图 2：

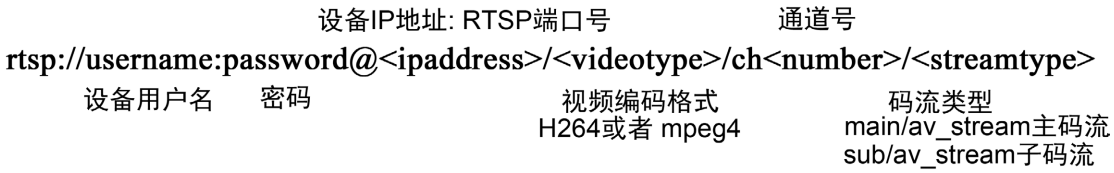

Figure 2. Sample RTSP format
图 2. RTSP 样本格式

该算法的输出为警报信息，主要由警报图像组成。开启算法后，输出的返回值采用 JSON 格式。返回参数为：

1) 状态 0 成功打开，其他状态关闭；

2) “ ERRORMSG”是发生错误时的错误消息，否则为空。

当算法正常运行时，识别目标时的输出为警报图像，警报时间，警报类型等。

检测效果图如图 3，可以看出无论是白天还是夜晚，我们的模型都可以正确检测出未带安全帽的人员。

Figure 3. Algorithm recognition results
图 3. 算法识别结果

5. 结论与未来工作

通常，智能视频监控系统大致由两部分组成，即智能视频分析处理子系统和网络视频监控系统。其中，智能视频分析子系统负责实时分析由关键摄像机捕获的连续视频图像内容，分析并记录图片中的各种目标情况。一旦目标运动情况违反预设规则，就会通过网络将其发送到系统。告警信息。网络视频监控子系统的功能是在接收到来自智能视频分析子系统的告警信息后，按照预设的规则，将告警输出，录像，云台等功能链接在一起。该视频监控系统也由这两部分组成，其基本功能主要分为三个主要部分：读取由摄像头发给视频分析服务器的 RTSP 视频流；实时分析视频流的算法模型；确定目标将警报图像发送到平台，并将其保存在视频分析服务器中。该系统具有以下优点：

1) 在安全帽算法中，由于工地上的人员基本上都带有安全帽，导致缺乏负样本。通过互联网上开放的环境进行搜索，获取负样本与工地的正样本进行混合来对模型进行训练和优化。

2) 在数据集的制作中，我们结合了网上先进的预训练模型，再使用我们标记的工地数据集进行微调。再通过长时间的收集容易出错的场景的数据，对模型进行针对性的优化来确保模型可以胜任日常检测的使用。

基金项目

四川省青年科技创新研究团队专项计划项目(2019JDTD0023)。

参考文献

[1] Liu, W., Anguelov, D., Erhan, D., *et al.* (2016) SSD: Single Shot Multibox Detector. *European Conference on Com-*

puter Vision, Springer, Cham, 21-37. https://doi.org/10.1007/978-3-319-46448-0_2

[2] Redmon, J., Divvala, S., Girshick, R., *et al.* (2016) You Only Look Once: Unified, Real-Time Object Detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 779-788. https://doi.org/10.1109/CVPR.2016.91

[3] Girshick, R., Donahue, J., Darrell, T., *et al.* (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[4] Li, W.S., *et al.* (2019) Improved Faster R-CNN Coal Mine Descent Detection Algorithm, 2019. Chelsea Finn. 2018. Learning to Learn with Gradients. Ph.D. Thesis, EECS Department, University of Berkeley.

[5] Faster, R. (2015) Towards Real-Time Object Detection with Region Proposal Networks. *Advances in Neural Information Processing Systems*, **2015**, 9199.

[6] Simonyan, K. and Zisserman, A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. https://arxiv.org/pdf/1409.1556.pdf

[7] Deng, J., Dong, W., Socher, R., *et al.* (2009) ImageNet: A Large-Scale Hierarchical Image Database. 2009 *IEEE Conference on Computer Vision and Pattern Recognition*, Miami, 20-25 June 2009, 248-255. https://doi.org/10.1109/CVPR.2009.5206848

[8] Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 1097-1105.

[9] 张灏, 王素珍, 郑宇, 王伟, 等. 一种组合 GAUSS-filter、SOBEL、NMS、OTSU 4 种算法的图像边缘检测的 FPGA 实现[J]. 液晶与显示, 2020, 35(3): 250-261.

[10] 温捷文, 战荫伟, 李楚宏, 卢剑彪. 一种加强 SSD 小目标检测能力的 Atrous 滤波器设计[J]. 计算机应用研究, 2019, 36(3): 861-865.