Computer Science and Application 计算机科学与应用, 2023, 13(12), 2222-2229
Published Online December 2023 in Hans. https://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2023.1312222

Hans汉斯

监控场景下基于CLIP的细粒度目标检测方法

王　齐[*]，曾卓夫[*]，黄小明[#]，费雨欣，陈逸洋，廖家俊

北京信息科技大学计算机学院，北京

收稿日期：2023年11月15日；录用日期：2023年12月14日；发布日期：2023年12月21日

摘　要

当前，随着国内摄像头数量的迅猛增长，每天所产生的庞大视频数据不仅对人力和物力资源构成了巨大的负担，而且导致了昂贵的成本开支。针对这一问题，本研究聚焦于解决细粒度目标检测领域存在的具体问题。本研究基于深度学习技术，结合Yolov4目标检测和CLIP特征分析，提出了一种综合的图像分析方法，以降低视频数据处理的成本。目前，现有的细粒度目标检测方法在处理大规模视频数据时面临着一系列挑战。这些挑战包括但不限于人工标注成本太高，而且无法保证标注的全面性，人工标注不如用户反馈及时有效；泛化能力只太弱，定制化成本太高，大多数AI任务都需要case by case实现。为了解决这些问题，本研究首先利用Yolov4模型对输入图像进行人物检测，以高效地实现目标的准确分割。随后，针对每个分割的人物，本实验采用CLIP模型进行深度特征分析，其泛化能力强且训练语料完全不需要人工标注的特点使捕捉图像和语言之间的语义精准关联。通过本研究的实验结果，本研究验证了该方法在人物检测方面的卓越表现，并展示了在基于CLIP的特征分析中显著的语义一致性。这一创新方法有望显著降低视频数据处理的成本和工作量，为细粒度目标检测领域的进一步研究提供了新的方向。

关键词

Yolov4，CLIP，深度学习，图像检测，特征分析

CLIP-Based Fine-Grained Target Detection Method in Surveillance Scenarios

Qi Wang[*], Zhuofu Zeng[*], Xiaoming Huang[#], Yuxin Fei, Yiyang Chen, Jiajun Liao

School of Computing, Beijing Information Science and Technology University, Beijing

Received: Nov. 15th, 2023; accepted: Dec. 14th, 2023; published: Dec. 21st, 2023

[*]共同第一作者。
[#]通讯作者。

文章引用: 王齐, 曾卓夫, 黄小明, 费雨欣, 陈逸洋, 廖家俊. 监控场景下基于 CLIP 的细粒度目标检测方法[J]. 计算机科学与应用, 2023, 13(12): 2222-2229. DOI: 10.12677/csa.2023.1312222

Abstract

Currently, the proliferation of cameras in the nation has resulted in an immense volume of video data being produced on a daily basis, which is not only a huge strain on human and material resources, but also comes with a hefty price tag. This paper concentrates on resolving the particular difficulties associated with precise target recognition in order to address this issue. We propose an integrated image analysis method to reduce the cost of video data processing. This method is based on deep learning techniques, combined with Yolov4 target detection and CLIP feature analysis. Currently, there are a number of challenges that current target detection methods face when working with large-scale video data. In addition to the expensive cost of manual tagging and the lack of assurance that it is comprehensive and that manual tagging is not as timely and effective as user feedback, generalization is only too weak, customization is too expensive, and most AI tasks need to be implemented on a case-by-case basis. To solve these problems, we first use the Yolov4 model to detect the characters of the input images in order to achieve accurate segmentation efficiently. The CLIP model is then used for in-depth feature analysis for each segmented character. The ability to generalize and train language materials without manual tagging makes it possible to capture semantic and precise associations between images and languages. Our findings showcase the exceptional efficacy of this method in character detection and exhibit substantial semantic coherence in CLIP-based feature analysis. This novel approach is anticipated to drastically cut down on the expense and labor of video data processing and open up fresh avenues for further exploration in the area of precise target recognition.

Keywords

Yolov4, CLIP, Deep Learning, Image Segmentation, Feature Analysis

1. 概述

据统计，中国共装有 1.76 亿个监控摄像头。如此多的摄像头，每天拍摄到海量的视频数据，需要快速分析理解视频中的场景，但当前监控视频存在数据量过大需人工标注、每种监控应用需单独训练模型等缺陷，技术仍待更新[1] [2] [3]。

CLIP 是一种使用大量图像和文本对进行训练的神经网络。通过这种多模态训练，CLIP 可以用于查找最能代表图像的文本片段，或者查找给定文本查询的最合适图像。在 image-level 的分类上，CLIP 已经取得了非常令人印象深刻的效果。考虑到其巨大的潜力，将其应用于目标检测也是非常合理的。

浙江大学的研究团队提出把视频动态识别的任务看成是视频文本检索，提出了一个 Propmt 的模块根据标签来生成本文句子，然后用 CLIP 的 Text Encoder 对生成的文本进行 encode，同样用 CLIP 的 Image Encoder 对视频的多帧图片进行编码，然后提出了几种方式将多帧图片信息变成一帧图片的信息，然后计算文本和这一帧图片的相似度[4]；有学者基于 HERO 模型，将 CLIP 的一些组件整合到 HERO 模型中，以提升在视频－文本任务上的性能表现，这些研究成果为视频动态识别和文本检索的结合提供了新的思路和方法[5]；腾讯的研究人员提出了一种名为 CLIP2Video 的模型，以端到端的方式将图像语言预训练模型转换为视频文本检索模型[6]；Google 的研究人员提出了 Vi LD 模型，这是一种通过视觉和语言知识

蒸馏的训练方法，将 CLIP 图像分类模型应用到了目标检测任务上，在新增类别推理上 Zero-Shot 超过了有监督训练的方法[7]。

因此，在国内外，基于 CLIP 预训练模型的视频监控目标检测已成为一个热门的研究领域。[8] [9] [10] [11]。本课题研究基于视觉语言预训练模型 CLIP 的视频监控，结合图像和文本数据的优势，以提高监控系统的准确性和效率为目标，通过 YOLO 算法对图像数据进行处理，实现对目标的初步检测，再通过 CLIP 模型得到每个目标的细粒度描述。本课题的研究框架如图 1 所示。

(a) 输入的图片

(b) YOLO 检测结果图

```
Category: gender of this person is
  gender of this person is: man, 预测概率: 83.58%

Category: top length of this person is
  top length of this person is: person wearing long clothes, 预测概率: 92.20%

Category: sleeve length of this person is
  sleeve length of this person is: person wearing long sleeves, 预测概率: 74.72%
```

```
Category: bottom color of this person is
  bottom color of this person is: person wearing dark pants, 预测概率: 86.24%

Category: shoes color of this person is
  shoes color of this person is: person wearing black shoes, 预测概率: 75.89%

Category: body orientation of this person is facing
  body orientation of this person is facing: a person facing forward, 预测概率: 81.61%
```

```
Category: top length of this person is
  top length of this person is: person wearing long clothes, 预测概率: 75.46%

Category: sleeve length of this person is
  sleeve length of this person is: person wearing long sleeves, 预测概率: 72.09%

Category: this person is
  this person is: person without hat, 预测概率: 62.94%

Category: accessory of this person is
  accessory of this person is: person without mask, 预测概率: 64.08%
```

(c) CLIP 检测结果图

Figure 1. System demonstration diagram
图 1. 系统演示图

本文的主要贡献包括以下三点：

(1) 实现基于 CLIP 预训练模型的图像和文本特征提取，将文本信息与图片信息结合起来，提高监控系统的准确性和效率。

(2) 探究基于 CLIP 模型的视频监控目标检测方法，进一步提高物体检测的准确性和效率。此外，使用目标检测算法 YOLO、Faster R-CNN 等对特定目标进行初步检测，并使用 CLIP 模型对目标进行细粒度分类或识别。

(3) 针对不同场景下的视频监控数据，构建、标注和优化基于视频监控的数据集，用于检测算法的学习和优化。同时探索对检测算法进行不断训练优化，提高其效率和准确性。

(4) 总结和归纳基于 CLIP 预训练模型的视频监控目标检测的优劣和应用场景，探索研究此方法的创新和发展方向。

2. 现有方法存在的问题

尽管现有目标检测算法比如 YOLO，FasterRCNN 已经实现了很好的性能，但是仍然存在两个明显不足。第一，现有目标检测算法都是在一个封闭数据集上训练的，只能检测训练集上出现过的类别，每种监控应用都需要单独训练一个模型。比如监控打斗人群和监控地铁里抽烟行为，需要分别训练模型。第二，现有目标检测方法只能输出目标的类别信息，不能输出细粒度的详细描述。比如检测人的模型，只能检测图像中是否有人的出现，无法输出人的发型、衣服颜色、体型等更详细的描述[8]。

3. 工作研究与预备知识

CLIP 是一种在各种图像、文本上训练的神经网络模型。它可以用自然语言来指示，预测图像最匹配的文本片段，而无需直接针对任务进行优化，类似于 GPT-2 和 3 的零样本功能。实验发现 CLIP 在 ImageNet 上“零样本”上与原始 ResNet50 的性能相匹配，而无需使用任何原始的 1.28M 标记示例，从而克服了计算机视觉中的几个主要挑战。

4. 监控场景下基于 YOLO 和 CLIP 的细粒度目标检测方法设计

4.1. 总体设计

在整体设计中，借用了开源项目 YOLOv4 和 CLIP，目标是实现细粒度目标检测，包括人物的外表和动作特征的获取。具体步骤如下：

1) YOLOv4 目标初步检测：使用 YOLOv4 进行目标检测，该模型具备实时性和准确性。首先，将传入的图像送入 YOLOv4 模型，得到人体的位置信息。YOLOv4 能够识别并定位多个目标，包括人物，这为后续的图像分割提供了准确的边界框信息。

2) 图像分割：基于 YOLOv4 检测到的人物位置信息，对原始图像进行分割。这一步旨在将不同目标从原始图像中提取出来，为后续的细粒度特征分析做准备。

3) CLIP 的图像细粒度目标检测：将分割后的图像传入 CLIP 模型进行细粒度的目标检测。首先使用分类标签构建每个类别的描述文本，例如，对于一个人物的图像，描述文本可以是“person wearing long clothes”。将这些文本送入 CLIP 的 Text Encoder 得到对应的文本特征。

4) 图像特征提取：将分割后的图像送入 CLIP 的 Image Encoder 得到图像特征。这一步旨在将图像转化为模型可理解的向量表示。

5) 相似度计算：对于每个类别的文本特征和图像特征，计算它们之间的缩放余弦相似度。这个相似度可以被看作是图像与描述文本之间的关联度，越大表示两者越相似。

6) 图像分类预测：选择相似度最大的文本对应的类别作为图像的分类预测结果。这一步将图像与描述文本关联，实现对图像的细粒度分类。

4.2. 基于 YOLOv4 目标初步检测

在基于 YOLOv4 的目标初步检测阶段，关键步骤包括：

1) YOLOv4 模型选择和集成：在实验中选择了 YOLOv4 作为目标检测模型，该模型以其先进的性能而闻名。在集成过程中，可能需要根据实际需求调整模型的参数，例如 anchor boxes 的数量和大小，以更好地适应细粒度目标检测任务。这些调整可以通过对训练数据进行分析和实验来确定。

2) 图像输入和预处理：将待检测的图片传入 YOLOv4 模型之前，可能需要进行适当的预处理，例如图像的缩放、归一化和其他增广技术，以确保模型能够处理不同尺寸和质量的图像。

3) 目标检测和边界框生成：YOLOv4 通过卷积神经网络检测图像中的目标，并生成相应的边界框。每个边界框都带有一个置信度分数，表示该边界框中包含目标的概率。在这一步，需要解析模型的输出，提取目标位置信息和相应的置信度分数。

4) 非极大值抑制(NMS)：由于 YOLOv4 可能在同一目标上生成多个相似的边界框，需要使用非极大值抑制来去除冗余的边界框，保留置信度最高的边界框。这确保了在后续的分割和特征提取阶段，每个目标只被处理一次。

5) 图像分割准备：使用 YOLOv4 生成的边界框信息，对原始图像进行分割。这一步旨在提取出每个检测到的目标，为后续的细粒度特征分析做准备。分割后的图像将成为送入 CLIP 模型的输入。

4.3. 基于 CLIP 的图像细粒度目标检测

在基于 CLIP 的图像细粒度目标检测中，关键步骤包括：

1) 任务标签的描述文本构建： 针对每个类别的任务标签，构建相应的描述文本，例如“person wearing long clothes”。这些文本描述了任务标签所代表的图像内容。

2) 文本特征提取：将任务描述文本送入 CLIP 的 Text Encoder，得到对应的文本特征。Text Encoder 将文本转化为高维向量表示，捕捉文本的语义信息。

3) 图像特征提取：将待预测的图像送入 CLIP 的 Image Encoder，得到图像特征。Image Encoder 将图像转化为高维向量表示，捕捉图像的语义信息。

4) 相似度计算：使用缩放的余弦相似度计算文本特征和图像特征之间的相似度。这一步反映了图像与任务描述文本之间的语义关联度。

5) 分类预测：选择相似度最大的任务描述文本对应的类别作为图像的细粒度分类预测结果。这一步通过比较相似度来决定图像所属的具体类别。进一步地，可以将这些相似度看成 logits，送入 softmax 后可以到每个类别的预测概率。

5. 方法实现

5.1. 实验流程

本实验进行了 yolov4 和 CLIP 的安装并修改了相关配置。通过 YOLOv4，将人物分割的图片储存在临时文件夹中。通过 CLIP，获取并输出图文匹配的最高细粒度分析结果。

5.2. Yolov4 图像目标检测

在模块工作流程中，本实验首先加载 YOLOv4 的权重文件和配置文件。随后，本实验传递输入图像

并执行前向传播。模型的输出包括检测框的坐标、类别标签和置信度分数。本实验仅保留类别标签为“人”(class_id = 0)并且置信度分数高于 0.5 的检测结果，以筛选出人体目标。

为了获得独立的人体目标，本实验实施了一个目标合并算法。通过计算两个检测框之间的交并比(IoU)，本实验可以确定它们是否相邻或重叠。如果 IoU 大于 0.5，本实验将多个目标合并为一个，以减少冗余检测。

5.3. CLIP 细粒度特征匹配

在细粒度特征匹配的任务中，必须对一些文件进行适度修改，以使其适应特定的研究需求。以下是进行修改的文件：

提示词文件：需创建包含多元细粒度特征描述的提示词文件，如“穿长袖”、“戴眼镜”等。务必确保文件格式与代码中的提示词一一匹配，以保证匹配的准确性。

输入文件夹：需要指明包含待进行特征匹配的图像的文件夹路径。此文件夹将作为 CLIP 模型进行匹配操作的数据源。

细粒度特征匹配的实现步骤：

加载 CLIP 模型：在研究中加载 CLIP 模型，应选用适当的预训练模型，本文使用了 ViT-B/32。该模型的主要任务是进行图像与提示词之间的语义匹配。

遍历图像文件夹：通过遍历包含待匹配图像的文件夹，获取每张图像的文件路径。

图像预处理：对每张图像进行预处理，以将其转换为适合 CLIP 模型的输入格式，从而为后续匹配操作做好准备。

特征编码：利用 CLIP 模型，分别对图像特征和一组提示词进行编码。这一过程将生成图像与每个提示词的相似性分数。

选择匹配结果：根据相似性分数，选择每个提示词类别中与图像最相似的提示词，从而确定图像的精确特征描述。

输出匹配结果：将匹配结果输出，包括每个提示词类别、其对应的匹配结果以及匹配的概率分数。

5.4. 方法测试

5.4.1. 数据集多样性和提示词集构建

本实验选择了一个多样性的图像数据集，其中包含了各种场景和不同细粒度特征描述的图像。这样的多样性为实验提供了广泛的测试基础，涵盖了不同情境下的图像，从而验证了所提出方法在各种场景下的适用性。为了引导 CLIP 模型进行特征匹配，本实验构建了一个包含各种细粒度特征描述的提示词集，包括穿着、外貌、姿势和其他特征描述。这一提示词集的设计有助于模型更精准地理解图像中的特征并进行匹配。

5.4.2. 评估指标和实验结果

在评估方法性能时，本实验关注了匹配准确性和匹配速度两个关键指标。

匹配准确性：在本次实验中，定义了 13 个类别对应的标签，其中每个标签采用 Top-k 匹配方式，其中通常选择 k 值为 1、3 和 5。从单次实验结果来看，在 Top-1 匹配中，本方法取得了高达 70%的准确性，意味着在 85%的情况下成功匹配了正确的特征。随后，本研究在此基础上扩大了该方法的测试数据量，通过尚未标注的行人数据集测试，每张图的行人特征匹配结果平均在 75%的准确性。这表明该方法在解决细粒度特征匹配问题上表现出色。此外，本实验进行了人物的身体部位检测测试，通过捕捉人物的具体部位进行细粒度测试，匹配结果的准确度有进一步的提升，这为下一步的研究确定了明确的方向。

匹配速度：考虑到实际应用的实时性需求，本研究关注了方法的匹配速度。结果表明，尽管方法涉及复杂的图像特征编码和文本描述，但平均处理时间仅为不到 100 毫秒/图像，图片处理速度远超人工标注，有效减小了获取图片信息结果时间和人力物力资源构成了巨大的负担。证明了其在实际应用中具有高效的匹配速度。

在实验过程中也存在一些挑战。对于在特征描述相似的情况下出现的不准确匹配问题，本实验增加了提示词的多样性，以更好地区分不同的特征描述。本实验还遇到了模型过拟合的情况，导致性能在不同数据集上的表现不一致。为了解决这一问题，本实验采用了数据增强技术，以增加数据集的多样性，从而提高模型的泛化能力。总的来说，本实验结果显示，该方法在细粒度特征匹配方面具有潜力，并且通过解决问题和性能优化取得了积极的进展。这为更广泛的应用提供了有力的支持和启发。

6. 结语

本文通过将 YOLOv4 目标检测和 CLIP 图文匹配相结合，提出了一种细粒度特征匹配方法，旨在实现对图像中独立人体的准确分割，并通过 CLIP 模型对其进行精细描述。实验结果表明，本方法在多样性的图像数据集上取得了令人满意的匹配准确性和速度。通过对 YOLOv4 输出的目标进行合并，本实验成功地实现了对独立人体的分割，并将其传递给 CLIP 进行语义匹配。在多个特征描述的测试中，本实验观察到了高度准确的匹配结果，验证了方法的有效性。特别是在 Top-1 匹配中，结果达到了超过 70%的准确性，其中 85%的特征内容得到了准确匹配。

未来的工作方向包括进一步优化匹配算法，尤其是在处理相似特征描述时的适用性方面。此外，本实验计划扩大实验数据集，以更全面地评估方法的性能。对于 CLIP 模型的参数调整和预训练模型的选择也是未来研究的关键方向，以进一步提升匹配的精度和效率。

综合而言，本文提出的方法在细粒度特征匹配领域取得了一定的成果，但仍有进一步改进的空间。相信这一工作为未来深入探讨图像处理和语义匹配的研究方向提供了有益的参考。

基金项目

北京信息科技大学大学生创新创业训练计划项目－计算机学院，项目名称：监控场景下基于 CLIP 的细粒度目标检测方法；项目编号：5112210832，S202311232353。

致　谢

感谢北京信息科技大学大学生创新创业训练计划项目－计算机学院基金项目的大力支持与帮助，感谢课程老师的积极指导与协助。

参考文献

[1] 龚声蓉, 曹李军, 刘纯平, 等. 面向智能监控的视频行为分析关键技术与应用[J]. 中国科技成果, 2022, 23(12): 12-13.

[2] 李景文, 韦晶闪, 姜建武, 等. 多视角监控视频中动态目标的时空信息提取方法[J]. 测绘学报, 2022, 51(3): 388-400. https://doi.org/10.11947/j.AGCS.2022.20200507

[3] 肖进胜, 申梦瑶, 江明俊, 等. 融合包注意力机制的监控视频异常行为检测[J]. 自动化学报, 2022, 48(12): 9. https://doi.org/10.16383/j.aas.c190805

[4] Wang, M., Xing, J. and Liu, Y. (2021) Actionclip: A New Paradigm for Video Action Recognition. arXiv: 2109.08472.

[5] Li, G., He, F. and Feng, Z. (2021) A CLIP-Enhanced Method for Video-Language Understanding. arXiv: 2110.07137.

[6] Fang, H., Xiong, P.F., Xu, L.H., *et al.* (2022) CLIP2Video: Mastering Video-Text Retrieval via Image CLIP. arXiv:

2106.11097.

[7] Gu, X.Y., Lin, T.Y., Kuo, W.C., *et al*. (2022) Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation. arXiv: 2104.13921.

[8] 刘立栋. 监控视频中大规模群体系统模型的研究[D]: [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2013.

[9] 赵晋巍, 刘晓鹏, 罗威, 等. 基于 CLIP 模型的军事领域图片资源多模态搜索工具研究[J]. 中华医学图书情报杂志, 2022, 31(8): 14-20.

[10] 张斌斌, 张永新, 李德光, 等. 基于 CLIP 模型的牲畜分娩监测预警系统与方法[P]. CN202210948935.7. 2023-11-01.

[11] 谭康霞, 平鹏, 秦文虎. 基于 YOLO 模型的红外图像行人检测方法[J]. 激光与红外, 2018, 48(11): 1436-1442.