



# (12) 发明专利

(10) 授权公告号 CN 112101344 B

(45) 授权公告日 2022. 09. 06

(21) 申请号 202010863241.4

G06N 3/04 (2006.01)

(22) 申请日 2020.08.25

(56) 对比文件

(65) 同一申请的已公布的文献号

CN 110147745 A, 2019.08.20

申请公布号 CN 112101344 A

审查员 刘梦飞

(43) 申请公布日 2020.12.18

(73) 专利权人 腾讯科技(深圳)有限公司

地址 518044 广东省深圳市南山区高新区

科技中一路腾讯大厦35层

(72) 发明人 宋浩 黄珊

(74) 专利代理机构 北京同达信恒知识产权代理

有限公司 11291

专利代理师 李娟

(51) Int. Cl.

G06V 30/19 (2022.01)

G06V 10/82 (2022.01)

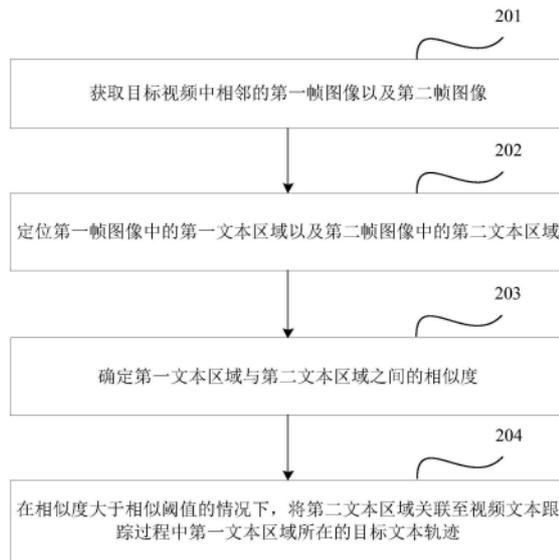
权利要求书3页 说明书13页 附图5页

(54) 发明名称

一种视频文本跟踪方法及装置

(57) 摘要

本申请涉及计算机技术领域,尤其涉及一种视频文本跟踪方法及装置,以解决相关技术中对视频进行文本识别效率低计算量大的问题。方法包括:获取目标视频中相邻的第一帧图像以及第二帧图像;定位所述第一帧图像中的第一文本区域以及所述第二帧图像中的第二文本区域;确定所述第一文本区域与所述第二文本区域之间的相似度;在相似度大于相似阈值的情况下,将所述第二文本区域关联至视频文本跟踪过程中所述第一文本区域所在的目标文本轨迹。



1. 一种视频文本跟踪方法,其特征在于,所述方法包括:
  - 获取目标视频中相邻的第一帧图像以及第二帧图像;
  - 定位所述第一帧图像中的第一文本区域以及所述第二帧图像中的第二文本区域;
  - 确定所述第一文本区域与所述第二文本区域之间的相似度;
  - 在相似度大于相似阈值的情况下,将所述第二文本区域关联至视频文本跟踪过程中所述第一文本区域所在的目标文本轨迹;
  - 其中,所述定位所述第一帧图像中的第一文本区域以及所述第二帧图像中的第二文本区域,包括:
    - 将所述第一帧图像和所述第二帧图像输入孪生跟踪网络;所述孪生跟踪网络包括定位子网络和跟踪子网络;
    - 通过所述定位子网络定位所述第一帧图像中的第一文本区域;所述定位子网络利用训练图像以及所述训练图像中文本区域的位置信息进行训练得到对应的网络模型参数;
    - 通过所述跟踪子网络定位所述第二帧图像中的第二文本区域;所述跟踪子网络采用与所述定位子网络相同的网络模型参数;
    - 所述确定所述第一文本区域与所述第二文本区域之间的相似度,包括:
      - 利用所述孪生跟踪网络中的目标对比层确定所述第一文本区域的第一特征向量;
      - 利用所述孪生跟踪网络中的跟踪检测层确定所述第二文本区域的第二特征向量;
      - 所述孪生跟踪网络根据所述第一特征向量与所述第二特征向量,确定所述第一文本区域与所述第二文本区域之间的相似度。
2. 根据权利要求1所述的方法,其特征在于,所述通过所述定位子网络定位所述第一帧图像中的第一文本区域,包括:
  - 利用所述定位子网络,将所述第一帧图像进行N次第一降采样处理,得到N个第一图像样本;所述N次第一降采样处理的倍数不相同;
  - 针对每个第一图像样本,所述定位子网络确定所述第一图像样本中第一文本区域的位置信息;
  - 所述通过所述跟踪子网络定位所述第二帧图像中的第二文本区域,包括:
    - 利用所述跟踪子网络,将所述第二帧图像进行N次第二降采样处理,得到N个第二图像样本;所述第二降采样处理的倍数与所述第一降采样处理的倍数相对应;
    - 针对每个第二图像样本,所述跟踪子网络确定所述第二图像样本中第二文本区域的位置信息;
    - 所述确定所述第一文本区域与所述第二文本区域之间的相似度,包括:
      - 分别确定相同降采样倍数下,第一图像样本与第二图像样本之间的样本相似度。
3. 根据权利要求2所述的方法,其特征在于,所述确定第一图像样本与降采样倍数相同的第二图像样本之间的相似度之后,还包括:
  - 根据不同降采样倍数对应的样本相似度,计算所述第一文本区域与所述第二文本区域之间的相似度。
4. 根据权利要求1所述的方法,其特征在于,根据以下方式训练得到所述孪生跟踪网络:
  - 对初始孪生跟踪网络中的定位子网络进行训练,得到对应的定位子网络模型参数;

获取训练样本,所述训练样本中包含有正样本和负样本,所述正样本包括文本内容相同的第一训练样本图像以及所述第一训练样本图像的标注相似度,所述负样本包括文本内容不同的第二训练样本图像以及所述第二训练样本图像的标注相似度;

将所述训练样本输入初步训练的孪生跟踪网络,并根据输出的相似度及所述标注相似度计算损失函数,所述初步训练的孪生跟踪网络包括已确定的所述定位子网络模型参数;

当所述损失函数小于预设阈值时,确定对应的参数为所述孪生跟踪网络对应的参数,得到所述孪生跟踪网络。

5. 一种视频文本识别方法,其特征在于,所述方法包括:

确定目标视频中的目标文本轨迹,其中,所述目标文本轨迹为通过如权利要求1-4中任一项所述的方法得到的;

针对每一个目标文本轨迹,将所述目标文本轨迹中的任一帧图像作为目标帧图像,并确定所述目标帧图像中的目标文本区域;

识别所述目标视频中所有目标文本区域的文字。

6. 一种视频文本跟踪装置,其特征在于,所述装置包括:

获取单元,用于获取目标视频中相邻的第一帧图像以及第二帧图像;

定位单元,用于定位所述第一帧图像中的第一文本区域以及所述第二帧图像中的第二文本区域;

对比单元,用于确定所述第一文本区域与所述第二文本区域之间的相似度;

关联单元,用于在相似度大于相似阈值的情况下,将所述第二文本区域关联至视频文本跟踪过程中所述第一文本区域所在的目标文本轨迹;

其中,所述定位所述第一帧图像中的第一文本区域以及所述第二帧图像中的第二文本区域时,所述定位单元具体用于:

将所述第一帧图像和所述第二帧图像输入孪生跟踪网络;所述孪生跟踪网络包括定位子网络和跟踪子网络;

通过所述定位子网络定位所述第一帧图像中的第一文本区域;所述定位子网络利用训练图像以及所述训练图像中文本区域的位置信息进行训练得到对应的网络模型参数;

通过所述跟踪子网络定位所述第二帧图像中的第二文本区域;所述跟踪子网络采用与所述定位子网络相同的网络模型参数;

所述确定所述第一文本区域与所述第二文本区域之间的相似度,包括:

利用所述孪生跟踪网络中的目标对比层确定所述第一文本区域的第一特征向量;

利用所述孪生跟踪网络中的跟踪检测层确定所述第二文本区域的第二特征向量;

所述孪生跟踪网络根据所述第一特征向量与所述第二特征向量,确定所述第一文本区域与所述第二文本区域之间的相似度。

7. 一种视频文本识别装置,其特征在于,所述装置包括:

确定单元,用于确定目标视频中的目标文本轨迹,其中,所述目标文本轨迹为通过如权利要求1-4中任一项所述的方法得到的;

选取单元,用于针对每一个目标文本轨迹,将所述目标文本轨迹中的任一帧图像作为目标帧图像,并确定所述目标帧图像中的目标文本区域;

识别单元,用于识别所述目标视频中所有目标文本区域的文字。

8. 一种计算机设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,其特征在于,所述处理器执行所述程序时实现权利要求1~4任一权利要求所述方法的步骤。

9. 一种计算机可读存储介质,其特征在于,其存储有可由计算机设备执行的计算机程序,当所述程序在计算机设备上运行时,使得所述计算机设备执行权利要求1~4任一所述方法的步骤。

## 一种视频文本跟踪方法及装置

### 技术领域

[0001] 本申请涉及计算机技术领域,尤其涉一种视频文本跟踪方法及装置。

### 背景技术

[0002] 随着视频技术发展,用户越来越多的通过视频获取外界信息,例如通过视频进行远程学习、观看电子书等;在日常使用中,用户往往会希望能够获取视频内的文本进行保存或者记录等。

[0003] 为了获取视频内的文本,一般可以利用算法模型对视频帧图像进行文本识别。这种文字识别方式需要对视频的每一帧图像进行文本识别,计算量大,效率较低。

### 发明内容

[0004] 本申请实施例提供一种视频文本跟踪方法及装置,以解决相关技术中对视频进行文本识别效率低计算量大的问题。

[0005] 一方面,本申请实施例提供了一种视频文本跟踪方法,包括:

[0006] 获取目标视频中相邻的第一帧图像以及第二帧图像;

[0007] 定位所述第一帧图像中的第一文本区域以及所述第二帧图像中的第二文本区域;

[0008] 确定所述第一文本区域与所述第二文本区域之间的相似度;

[0009] 在相似度大于相似阈值的情况下,将所述第二文本区域关联至视频文本跟踪过程中所述第一文本区域所在的目标文本轨迹。

[0010] 一方面,本申请实施例提供了一种视频文本跟踪装置,包括:

[0011] 获取单元,用于获取目标视频中相邻的第一帧图像以及第二帧图像;

[0012] 定位单元,用于定位所述第一帧图像中的第一文本区域以及所述第二帧图像中的第二文本区域;

[0013] 对比单元,用于确定所述第一文本区域与所述第二文本区域之间的相似度;

[0014] 关联单元,用于在相似度大于相似阈值的情况下,将所述第二文本区域关联至视频文本跟踪过程中所述第一文本区域所在的目标文本轨迹。

[0015] 可选的,所述定位单元,具体用于:

[0016] 将所述第一帧图像和所述第二帧图像输入孪生跟踪网络;所述孪生跟踪网络包括定位子网络和跟踪子网络;

[0017] 通过所述定位子网络定位所述第一帧图像中的第一文本区域;所述定位子网络利用训练图像以及所述训练图像中文本区域的位置信息进行训练得到对应的网络模型参数;

[0018] 通过所述跟踪子网络定位所述第二帧图像中的第二文本区域;所述跟踪子网络采用与所述定位子网络相同的网络模型参数;

[0019] 所述对比单元,具体用于:

[0020] 利用所述孪生跟踪网络中的目标对比层确定所述第一文本区域的第一特征向量;

[0021] 利用所述孪生跟踪网络中的跟踪检测层确定所述第二文本区域的第二特征向量;

[0022] 所述孪生跟踪网络根据所述第一特征向量与所述第二特征向量,确定所述第一文本区域与所述第二文本区域之间的相似度。

[0023] 可选的,所述定位单元,具体用于:

[0024] 利用所述定位子网络,将所述第一帧图像进行N次第一降采样处理,得到N个第一图像样本;所述N次第一降采样处理的倍数不相同;

[0025] 针对每个第一图像样本,所述定位子网络确定所述第一图像样本中第一文本区域的位置信息;

[0026] 利用所述跟踪子网络,将所述第二帧图像进行N次第二降采样处理,得到N个第二图像样本;所述第二降采样处理的倍数与所述第一降采样处理的倍数相对应;

[0027] 针对每个第二图像样本,所述跟踪子网络确定所述第二图像样本中第二文本区域的位置信息;

[0028] 所述对比单元,具体用于:

[0029] 分别确定相同降采样倍数下,第一图像样本与第二图像样本之间的样本相似度。

[0030] 可选的,还包括计算单元,用于:

[0031] 根据不同降采样倍数对应的样本相似度,计算所述第一文本区域与所述第二文本区域之间的相似度。

[0032] 可选的,还包括训练单元,用于根据以下方式训练得到所述孪生跟踪网络:

[0033] 对初始孪生跟踪网络中的定位子网络进行训练,得到对应的定位子网络模型参数;

[0034] 获取训练样本,所述训练样本中包含有正样本和负样本,所述正样本包括文本内容相同的第一训练样本图像以及所述第一训练样本图像的标注相似度,所述负样本包括文本内容不同的第二训练样本图像以及所述第二训练样本图像的标注相似度;

[0035] 将所述训练样本输入初步训练的孪生跟踪网络,并根据输出的相似度及所述标注相似度计算损失函数,所述初步训练的孪生跟踪网络包括已确定的所述定位子网络模型参数;

[0036] 当所述损失函数小于预设阈值时,确定对应的参数为所述孪生跟踪网络对应的参数,得到所述孪生跟踪网络。

[0037] 一方面,本申请实施例提供了一种视频文本识别装置,所述装置包括:

[0038] 确定单元,用于确定目标视频中的目标文本轨迹,其中,所述目标文本轨迹为通过如上所述的方法得到的;

[0039] 选取单元,用于针对每一个目标文本轨迹,将所述目标文本轨迹中的任一帧图像作为目标帧图像,并确定所述目标帧图像中的目标文本区域;

[0040] 识别单元,用于识别所述目标视频中所有目标文本区域的文字。

[0041] 一方面,本申请实施例提供了一种计算机设备,包括存储器、处理器及存储在存储器上并可在处理器上运行的计算机程序,所述处理器执行所述视频文本跟踪方法的步骤。

[0042] 一方面,本申请实施例提供了一种计算机可读存储介质,其存储有可由计算机设备执行的计算机程序,当所述程序在计算机设备上运行时,使得所述计算机设备执行所述视频文本跟踪方法的步骤。

[0043] 本申请实施例对目标视频进行视频文本跟踪过程中,获取目标视频中的第一帧图

像和第二帧图像,其中,第一帧图像与第二帧图像为相邻的两帧图像。在第一帧图像中定位出文本区域,作为第一文本区域;在第二帧图像中定位出文本区域,作为第二文本区域。确定第一文本区域与第二文本区域之间的相似度。根据相似度,确定第一文本区域与第二文本区域之间的关联关系,在相似度大于相似阈值时,将第二文本区域与第一文本区域相关联,也就是将第二文本区域关联至第一文本区域所在的目标文本轨迹。当相似度小于或等于相似阈值时,则认为第一文本区域与第二文本区域不关联。从而目标视频中的各帧图像可以形成多个文本轨迹,同一个文本轨迹中,图像内的文本区域相似度大于相似阈值,即可以认为文本内容相同。因此在后续对目标视频进行文本识别的过程中,仅针对同一个文本轨迹中的第一帧图像或任选一帧图像进行文字识别处理,即可确定目标视频中包含的文字。这样,无需对目标视频中的每一帧图像进行文字识别,降低了计算量,减少了文字识别耗费的时间,提高了效率。

### 附图说明

- [0044] 图1为本申请实施例提供的一种应用场景架构图;
- [0045] 图2为本申请实施例提供的一种视频文本跟踪方法的流程示意图;
- [0046] 图3为本申请实施例提供的一种孪生跟踪网络的结构示意图;
- [0047] 图4为本申请实施例提供的一种YOLO网络的结构示意图;
- [0048] 图5为本申请实施例提供的一种进行降采样处理的孪生跟踪网络的结构示意图;
- [0049] 图6为本申请实施例提供的一种视频文本跟踪装置的结构示意图;
- [0050] 图7为本申请实施例提供的一种电子设备结构示意图。

### 具体实施方式

[0051] 为了使本发明的目的、技术方案和优点更加清楚,下面将结合附图对本发明作进一步地详细描述,显然,所描述的实施例仅仅是本发明一部分实施例,而不是全部的实施例。基于本发明中的实施例,本领域普通技术人员在没有做出创造性劳动前提下所获得的所有其它实施例,都属于本发明保护的范围。

[0052] 下文中所用的词语“示例性”的意思为“用作例子、实施例或说明性”。作为“示例性”所说明的任何实施例不必解释为优于或好于其它实施例。

[0053] 文中的术语“第一”、“第二”仅用于描述目的,而不能理解为明示或暗示相对重要性或者隐含指明所指示的技术特征的数量。由此,限定有“第一”、“第二”的特征可以明示或者隐含地包括一个或者更多个该特征,在本发明实施例的描述中,除非另有说明,“多个”的含义是两个或两个以上。此外,术语“包括”以及它们任何变形,意图在于覆盖不排他的保护。例如包含了一系列步骤或单元的过程、方法、系统、产品或设备没有限定于已列出的步骤或单元,而是可选地还包括没有列出的步骤或单元,或可选地还包括对于这些过程、方法、产品或设备固有的其它步骤或单元。

[0054] 以下对本发明实施例中的部分用语进行解释说明,以便于本领域技术人员理解。

[0055] 人工智能:(Artificial Intelligence, AI) 是利用数字计算机或者数字计算机控制的机器模拟、延伸和扩展人的智能,感知环境、获取知识并使用知识获得最佳结果的理论、方法、技术及应用系统。换句话说,人工智能是计算机科学的一个综合技术,它企图了解

智能的实质,并生产出一种新的能以人类智能相似的方式做出反应的智能机器。人工智能也就是研究各种智能机器的设计原理与实现方法,使机器具有感知、推理与决策的功能。

[0056] 人工智能技术是一门综合学科,涉及领域广泛,既有硬件层面的技术也有软件层面的技术。人工智能基础技术一般包括如传感器、专用人工智能芯片、云计算、分布式存储、大数据处理技术、操作/交互系统、机电一体化等技术。人工智能软件技术主要包括计算机视觉技术、语音处理技术、自然语言处理技术以及机器学习/深度学习等几大方向。

[0057] 计算机视觉技术:(Computer Vision,CV) 计算机视觉是一门研究如何使机器“看”的科学,更进一步的说,就是指用摄影机和电脑代替人眼对目标进行识别、跟踪和测量等机器视觉,并进一步做图形处理,使电脑处理成为更适合人眼观察或传送给仪器检测的图像。作为一个科学学科,计算机视觉研究相关的理论和技术,试图建立能够从图像或者多维数据中获取信息的人工智能系统。计算机视觉技术通常包括图像处理、图像识别、图像语义理解、图像检索、OCR、视频处理、视频语义理解、视频内容/行为识别、三维物体重建、3D技术、虚拟现实、增强现实、同步定位与地图构建等技术,还包括常见的人脸识别、指纹识别等生物特征识别技术。

[0058] 机器学习:(Machine Learning,ML) 是一门多领域交叉学科,涉及概率论、统计学、逼近论、凸分析、算法复杂度理论等多门学科。专门研究计算机怎样模拟或实现人类的学习行为,以获取新的知识或技能,重新组织已有的知识结构使之不断改善自身的性能。机器学习是人工智能的核心,是使计算机具有智能的根本途径,其应用遍及人工智能的各个领域。机器学习和深度学习通常包括人工神经网络、置信网络、强化学习、迁移学习、归纳学习、式教学习等技术。

[0059] 卷积神经网络:(Convolutional Neural Networks,CNN) 是一类包含卷积计算且具有深度结构的前馈神经网络(Feedforward Neural Networks),是深度学习(deep learning)的代表算法之一。卷积神经网络具有表征学习(representation learning)能力,能够按其阶层结构对输入信息进行平移不变分类(shift-invariant classification),因此也被称为“平移不变人工神经网络”。

[0060] 帧:就是影像动画中最小单位的单幅影像画面。一帧就是一副静止的图像,连续的帧就形成动画,如电视图像等。通常说帧数,简单地说,就是在1秒钟时间里传输的图像的帧数,也可以理解为图形处理器每秒钟能够刷新几次,通常用FPS(Frames Per Second,每秒钟帧数)表示。每一帧都是静止的图像,快速连续地显示帧便形成了运动的假象。高的帧率可以得到更流畅、更逼真的动画,FPS越大,所显示的动作就会越流畅。

[0061] 像素点:是指在由一个数字序列表示的图像中的一个最小单位,也称为像素。像素是整个图像中不可分割的单位或者是元素。每一个点阵图像包含了一定量的像素,这些像素决定图像在屏幕上所呈现的大小。一张图片由好多的像素点组成。例如图片尺寸是 $500 \times 338$ 的,表示图片是由一个 $500 \times 338$ 的像素点矩阵构成的,这张图片的宽度是500个像素点的长度,高度是338个像素点的长度,共有 $500 \times 338 = 149000$ 个像素点。把鼠标放在一个图片上,这个时候会显示尺寸和大小,这里的尺寸就是像素。

[0062] 孪生网络:(Siamese network)即“连体的神经网络”,神经网络的“连体”是通过共享权值来实现的。孪生网络是一种网络结构,包括两个子网络,这两个子网络的结构一般是相同的,并且参数是共享的,即参数是一致的。孪生网络的目的一般是比较两幅图片是否相

似,或者说相似度是多少。

[0063] NMS (Non-Maximum Suppression,非极大值抑制) 算法,即抑制不是极大值的元素,可以理解为局部最大搜索。这个局部代表的是一个邻域,邻域有两个参数可变,一是邻域的维数,二是邻域的大小。例如在行人检测中,滑动窗口经提取特征,经分类器分类识别后,每个窗口都会得到一个分数。但是滑动窗口会导致很多窗口与其他窗口存在包含或者大部分交叉的情况。这时就需要用到NMS来选取那些邻域里分数最高(是行人的概率最大),并且抑制那些分数低的窗口。NMS在计算机视觉领域有着非常重要的应用,如视频目标跟踪、数据挖掘、3D重建、目标识别以及纹理分析等。

[0064] 在具体实践过程中,随着互联网中的视频数量及质量的不断提高,视频跟踪技术也得到了广泛的关注。视频目标跟踪包括单目标跟踪和多目标跟踪。单目标跟踪是在视频的初始帧中对某一个物体进行人为框选,随着物体位置的改变而生成对应的物体运动轨迹。多目标跟踪是使用检测算法对视频中的每一帧进行目标检测,根据数据关联(data association)方法将检测结果与跟踪结果进行匹配,生成视频中多个物体的跟踪轨迹。

[0065] 但是相关技术中,视频跟踪的目标对象为视频中的行人或物体,需要考虑被遮挡或者发生变形等问题。而对于视频中的文字,由于文字一般不会被遮挡且很少发生形变,因此,本申请实施例中对视频中文字跟踪的方法与相关技术中的视频跟踪算法不同。

[0066] 由于文本跟踪的准确性严格依赖检测结果,因此本申请实施例利用孪生跟踪网络对视频中的文本进行跟踪。其中,通过定位子网络快速准确地定位视频帧图像中的文本,跟踪子网络对相邻两帧图像中的文本区域进行相似度评估,根据相似度确定相邻两帧图像中文本区域的关联关系,进而生成文本轨迹。

[0067] 本申请实施例提供的文本跟踪方法可以应用于视频文本识别场景等。例如通过本申请实施例中的文本跟踪方法,可以将目标视频的各帧图像中的文本区域形成多个文本轨迹,这样在对目标视频进行文本识别时,对于同一个文本轨迹中的文本区域,仅选取一个进行文本识别,从而降低了文本识别的计算量。

[0068] 本申请实施例提供的视频文本跟踪方法的一种应用场景可以参见图1所示,该应用场景中包括终端设备101、服务器102和数据库103。

[0069] 其中,终端设备101为可以安装各类客户端,并且能够将已安装的客户端的运行界面进行显示的电子设备,该电子设备可以是移动的,也可以是固定的。例如,手机、平板电脑、笔记本电脑、台式电脑、各类可穿戴设备、智能电视、车载设备或其它能够实现上述功能的电子设备等。客户端可以是视频客户端或浏览器客户端等。各终端设备101通过通信网络与服务器102连接,该通信网络可以是有线网络或无线网络。服务器102可以是客户端对应的服务器,可以是一台服务器或由若干台服务器组成的服务器集群或云计算中心,或者是一个虚拟化平台。

[0070] 其中,图1是以数据库103独立于所述服务器102存在进行说明的,在其他可能的实现方式中,数据库103也可以位于服务器102中。

[0071] 服务器102与数据库103连接,数据库103中存储有训练样本,服务器102接收终端设备101发送的目标视频,将目标视频中相邻帧图像输入孪生跟踪网络中,通过定位子网络定位第一帧图像中的第一文本区域,通过跟踪子网络定位第二帧图像中的第二文本区域,对比第一文本区域与第二文本区域的相似度,在相似度大于阈值的情况下,确定第一文本

区域与第二文本区域之间的关联关系。服务器102针对目标视频中的每两个相邻帧图像均执行上述操作后,针对目标视频中的文本区域可以生成多个文本轨迹,实现目标视频中的文本跟踪。

[0072] 进一步地,服务器102还可以对目标视频进行文本识别。即利用目标视频的多个文本轨迹,从每一个文本轨迹中选取一帧图像,识别其中文本区域对应的文字,实现目标视频的文本识别。

[0073] 训练上述孪生跟踪网络的训练样本可以存储在数据库103中,服务器102从数据库103中获取训练样本进行网络模型的训练,得到模型参数。此外,目标视频也可以为存储在数据库103中,服务器102从数据库103中获取目标视频,生成文本轨迹并进行文本识别。

[0074] 基于图1所示的应用场景图,本申请实施例提供了一种视频文本跟踪方法,该方法的流程可以由图1中的服务器执行,如图2所示,包括以下步骤:

[0075] 步骤S201,获取目标视频中相邻的第一帧图像以及第二帧图像。

[0076] 具体实施过程中,目标视频可以是服务器接收终端设备上传的视频,也可以是服务器从数据库中获取的视频,或者为服务器从网络上下载的视频。由于视频是由视频帧按序播放形成的,本申请实施例通过将目标视频解码为连续的视频帧,得到一系列帧图像。具体可以通过FFmpeg (Fast Forward Mpeg,一种记录、转换数字音频、视频,并能将其转化为流的开源计算机程序)等拆解视频帧,具体的拆解方式这里不做限定。

[0077] 其中,第一帧图像与第二帧图像为目标视频的一系列帧图像中任意相邻的两帧图像。例如,目标视频是由10000个视频帧组成的,将目标视频进行解析,即得到其包含的10000个视频帧,为了便于分辨,可以将这些视频帧进行编号,依次为0000-9999。则本申请实施例中的第一帧图像可以为视频帧0000,则第二帧图像为视频帧0001;或者第一帧图像为视频帧0100,则第二帧图像为0101。

[0078] 步骤S202,定位第一帧图像中的第一文本区域以及第二帧图像中的第二文本区域。

[0079] 服务器获取视频帧图像后,确定视频帧图像中各个像素点的像素值,并根据像素值定位视频帧图像中的文本区域。像素值是图像被数字化时由计算机赋予的值,它代表了一个像素点的平均亮度信息,或者说是该像素点的平均反射(透射)密度信息。本发明实施例中,像素点的像素值可以是RGB色彩模式的颜色值,也可以是HSV (Hue-Saturation-Value,色调-饱和度-明度)颜色模型的颜色值,还可以是像素点的灰度值。

[0080] 具体实施过程中,视频帧图像中的像素点可以分为文本像素点和非文本像素点,根据像素点的像素值可以将目标图像中的所有像素点进行分类,确定每一个像素点是文本像素点还是非文本像素点。具体的定位文本区域的方式可以通过采用卷积神经网络、长短期记忆网络(Long Short-Term Memory,LSTM)等算法,也可以采用其他的文本区域定位方法,本实施例对所采用的文本区域定位方法不作限定。

[0081] 步骤S203,确定第一文本区域与第二文本区域之间的相似度。

[0082] 具体实施过程中,确定第一帧图像中的第一文本区域以及第二帧图像中的第二文本区域之后,可以根据文本区域的特征向量进行相似度对比。文本区域的特征向量也可以通过神经网络模型等算法计算得出,这里不做限制。通过对比两个文本区域的特征向量,计算得出文本区域之间的相似度。

[0083] 需要说明的是,在实际应用中,并非所有视频帧图像中都存在文本,若视频帧图像中不存在文本,则进行文本定位时,直接提示视频帧图像中不存在文本区域,则可以不对比文本区域之间的相似度。

[0084] 步骤S204,在相似度大于相似阈值的情况下,将第二文本区域关联至视频文本跟踪过程中第一文本区域所在的目标文本轨迹。

[0085] 具体的,若第一文本区域与第二文本区域之间的相似度大于相似阈值,则认为第一文本区域与第二文本区域相关联,即第一文本区域与第二文本区域为同一目标文本轨迹中的文本区域。

[0086] 举例来说,若连续三帧或三帧以上图像的文本区域均相关联,则这些帧图像中的文本区域作为一个文字轨迹;即若某文本区域仅出现在一帧图像或两帧图像中,则该文本区域不作为文字轨迹。若连续三帧图像中的文本区域均无法与上一帧图像相关联,则认为文本轨迹停止。

[0087] 本申请实施例对目标视频进行视频文本跟踪过程中,获取目标视频中的第一帧图像和第二帧图像,其中,第一帧图像与第二帧图像为相邻的两帧图像。在第一帧图像中定位出文本区域,作为第一文本区域;在第二帧图像中定位出文本区域,作为第二文本区域。确定第一文本区域与第二文本区域之间的相似度。根据相似度,确定第一文本区域与第二文本区域之间的关联关系,在相似度大于相似阈值时,将第二文本区域与第一文本区域相关联,也就是将第二文本区域关联至第一文本区域所在的目标文本轨迹。当相似度小于或等于相似阈值时,则认为第一文本区域与第二文本区域不关联。从而目标视频中的各帧图像可以形成多个文本轨迹,同一个文本轨迹中,图像内的文本区域相似度大于相似阈值,即可以认为文本内容相同。因此在后续对目标视频进行文本识别的过程中,仅针对同一个文本轨迹中的第一帧图像或任选一帧图像进行文字识别处理,即可确定目标视频中包含的文字。这样,无需对目标视频中的每一帧图像进行文字识别,降低了计算量,减少了文字识别耗费的时间,提高了效率。

[0088] 进一步地,本申请实施例中利用孪生跟踪网络对第一帧图像中的第一文本区域以及第二帧图像中的第二文本区域进行定位以及相似度对比。其中,孪生跟踪网络包括定位子网络和跟踪子网络。

[0089] 则上述步骤202,定位第一帧图像中的第一文本区域以及第二帧图像中的第二文本区域,包括:

[0090] 将第一帧图像和第二帧图像输入孪生跟踪网络;

[0091] 通过定位子网络定位第一帧图像中的第一文本区域;定位子网络利用训练图像以及训练图像中文本区域的位置信息进行训练得到对应的网络模型参数;

[0092] 通过跟踪子网络定位第二帧图像中的第二文本区域;跟踪子网络采用与定位子网络相同的网络模型参数。

[0093] 其中,定位子网络与跟踪子网络为孪生网络中的两个子网络,则这两个子网络的结构相同,并且参数一致。定位子网络是利用训练图像以及训练图像中文本区域的位置信息进行训练得到对应的网络模型参数,跟踪子网络与定位子网络是相同的结构,且与定位子网络的网络模型参数相同,因此,训练得出定位子网络的网络模型参数后,直接将网络模型参数赋予跟踪子网络即可。

[0094] 图3示出了本申请实施提供的一种孪生跟踪网络的结构示意图。如图3所示,将第一帧图像输入定位子网络中,确定出第一帧图像中的第一文本区域的位置,定位子网络可以输出第一文本区域的文本框。同时,将第二帧图像输入跟踪子网络中,确定出第二帧图像中的第二文本区域的位置,跟踪子网络可以输出第二文本区域的文本框。

[0095] 之后,为了对第一文本区域和第二文本区域进行相似度对比,本申请实施例在孪生跟踪网络中引入目标对比层,利用目标对比层确定第一文本区域的特征向量以及第二文本区域的特征向量,并利用特征向量进行对比。

[0096] 则,上述步骤203,确定第一文本区域与第二文本区域之间的相似度,包括:

[0097] 利用孪生跟踪网络中的目标对比层确定第一文本区域的第一特征向量;

[0098] 利用孪生跟踪网络中的跟踪检测层确定第二文本区域的第二特征向量;

[0099] 孪生跟踪网络根据第一特征向量与第二特征向量,确定第一文本区域与第二文本区域之间的相似度。

[0100] 举例来说,目标对比层可以为ROI Align(region of interest align,感兴趣区域校准)层,定位子网络对应第一ROI Align层,跟踪子网络对应第二ROI Align层,其中,第一ROI Align层和第二ROI Align层也是结构相同且参数相同。

[0101] 具体地,如图3所示,将第一文本区域的文本框输入第一ROI Align层,根据第一文本区域生成大小为 $14*14$ 的特征图,再通过平均池化层和连续2个1024维的全连接层,生成1024维的第一特征向量。同样的,将第二文本区域的文本框输入第二ROI Align层,根据第二文本区域生成大小为 $14*14$ 的特征图,再通过平均池化层和连续2个1024维的全连接层,生成1024维的第二特征向量。将第一特征向量与第二特征向量进行相似度对比,得到第一文本区域与第二文本区域之间的相似度。

[0102] 一种具体的实施例中,本申请实施例中的定位子网络和跟踪子网络均采用YOLO (You Only Look Once,基于单个神经网络的目标检测系统)网络。进一步地,为了增加计算的准确性,可以为YOLO-v3网络。

[0103] 具体地,一般YOLO网络结构如图4所示,包括24个卷积层和2个全连接层。其中,卷积层用来提取图像特征,全连接层用来预测图像位置和类别概率值。YOLO网络将输入图像分成 $S*S$ 个格子,每个格子负责检测落入该格子的物体。若某个物体的中心位置的坐标落入到某个格子,那么这个格子就负责检测出这个物体。每个格子输出 $B$ 个Bounding box (包含物体的矩形区域)信息,以及 $C$ 个物体属于某种类别的概率信息。Bounding box信息包含5个数据值,分别是 $x$ 、 $y$ 、 $w$ 、 $h$ 和confidence。其中 $x$ 、 $y$ 是指当前格子预测得到的物体的bounding box的中心位置的坐标; $w$ 、 $h$ 是bounding box的宽度和高度。因此,YOLO网络最终的全连接层的输出维度是 $S*S*(B*5+C)$ 。

[0104] 为了增加计算准确度,本申请实施中,针对同一帧图像采用YOLO-v3网络,在三种尺度上进行计算,分别为降采样8倍、16倍和32倍。降采样又名下采样或缩小图像,即减少采样点数。对于一幅 $N*M$ 的图像来说,如果降采样系数为 $k$ ,则即是在原图中每行每列每隔 $k$ 个点取一个点组成一幅新的图像。降采样的目的一般为使得图像符合显示区域的大小,生成对应图像的缩略图。

[0105] 图5示出了进行降采样处理的孪生跟踪网络的示意图。如图5所示,将第一帧图像输入YOLO-v3网络中,yolo-v3在三种尺度上进行定位,分别为降采样8倍、16倍和32倍。根据

三个不同的降采样倍数,输出三个不同的文本边缘框,结合NMS算法,产生最终的文本定位结果,即输出三个不同降采样倍数下第一文本区域的位置信息以及对应的置信度。

[0106] 同样的,将第二帧图像输入YOLO-v3网络中,分别降采样8倍、16倍和32倍。根据三个不同的降采样倍数,输出三个不同的文本边缘框,结合NMS算法,产生最终的文本定位结果,即输出三个不同降采样倍数下第二文本区域的位置信息以及对应的置信度。

[0107] 对于第一帧图像,定位子网络将YOLO-v3网络输出的特征图,输入对应的目标对比层中,生成降采样倍数分别为8倍、16倍和32倍的情况下,第一文本区域的特征向量,即定位子网络对应的ROI Align层输出第一文本区域的三个特征向量,其中一个特征向量对应一个降采样倍数。

[0108] 同理,对于第二帧图像,跟踪子网络将YOLO-v3网络输出的特征图,输入对应的目标对比层中,生成降采样倍数分别为8倍、16倍和32倍的情况下,第二文本区域的特征向量,即跟踪子网络对应的ROI Align层输出第二文本区域的三个特征向量,其中一个特征向量对应一个降采样倍数。

[0109] 之后,确定第一文本区域与第二文本区域之间的相似度,包括:分别确定相同降采样倍数下,第一图像样本与第二图像样本之间的样本相似度。

[0110] 即在降采样倍数为8倍的情况下,第一文本区域的特征向量与第二文本区域的特征向量进行相似度对比;在降采样倍数为16倍的情况下,第一文本区域的特征向量与第二文本区域的特征向量进行相似度对比;在降采样倍数为32倍的情况下,第一文本区域的特征向量与第二文本区域的特征向量进行相似度对比。这样,分别计算出三个相似度值,每个相似度值对应一个降采样倍数。

[0111] 在进行不同倍数的降采样处理后,确定第一图像样本与降采样倍数相同的第二图像样本之间的相似度之后,还包括:

[0112] 根据不同降采样倍数对应的样本相似度,计算第一文本区域与第二文本区域之间的相似度。具体根据以下公式计算第一文本区域与第二文本区域之间的相似度:

$$[0113] \quad Sim_{b_1, b_2} = \sum_i (c_{i, b_1} + c_{i, b_2}) / 2 \cdot sim_{b_1, b_2} \dots \dots \text{公式 1}$$

[0114] 其中,  $c_{i, b_1}$  为定位子网络中第i个分支的文本区域 $b_1$ 的置信度;  $sim_{b_1, b_2}$  为文本区域 $b_1, b_2$ 在对应的第i个分支的相似度结果,  $Sim_{b_1, b_2}$  为文本区域 $b_1, b_2$ 之间的相似度。

[0115] 根据计算得到的第一文本区域与第二文本区域的相似度,以及文本区域的IOU (intersection-over-union, 重叠度),生成第一帧图像与第二帧图像之间的相似度矩阵。一种可选的实施例中,本申请实施例利用Kuhn-Munkres算法,发现存在关联关系的文本区域。

[0116] 进一步地,本申请实施例根据以下方式训练得到所述孪生跟踪网络:

[0117] 对初始孪生跟踪网络中的定位子网络进行训练,得到对应的定位子网络模型参数;

[0118] 获取训练样本,所述训练样本中包含有正样本和负样本,所述正样本包括文本内容相同的第一训练样本图像以及所述第一训练样本图像的标注相似度,所述负样本包括文本内容不同的第二训练样本图像以及所述第二训练样本图像的标注相似度;

[0119] 将所述训练样本输入初步训练的孪生跟踪网络,并根据输出的相似度及所述标注

相似度计算损失函数,所述初步训练的孪生跟踪网络包括已确定的所述定位子网络模型参数;

[0120] 当所述损失函数小于预设阈值时,确定对应的参数为所述孪生跟踪网络对应的参数,得到所述孪生跟踪网络。

[0121] 具体实施过程中,为了更有效地训练孪生跟踪网络,因为跟踪子网络的网络模型参数与定位子网络的网络模型参数相同即可,因此本申请实施例先训练定位子网络。在训练得到定位子网络的网络模型参数后,固定定位子网络的参数,对整个孪生跟踪网络进行训练。在训练过程中,使用Adam优化器,学习率设置为0.0001,mini-batch的大小设置为4。在训练过程中,动量和权重衰减分别设置为0.9和0.0001。

[0122] 以下通过具体实例说明本发明实施例提供的视频文本的跟踪方法以及文本识别的实现过程。

[0123] 首先,获取目标视频中相邻的第一帧图像以及第二帧图像。

[0124] 将第一帧图像和第二帧图像输入孪生跟踪网络;所述孪生跟踪网络包括定位子网络和跟踪子网络。

[0125] 利用定位子网络,将第一帧图像进行三个倍数的降采样处理,得到三个第一图像样本。

[0126] 针对每个第一图像样本,定位子网络确定第一图像样本中第一文本区域的位置信息。

[0127] 同时,利用跟踪子网络,将第二帧图像进行三个倍数的降采样处理,得到三个第二图像样本。其中,跟踪子网络的降采样倍数与定位子网络的降采样倍数相对应。

[0128] 针对每个第二图像样本,跟踪子网络确定第二图像样本中第二文本区域的位置信息。

[0129] 分别确定相同降采样倍数下,第一图像样本与第二图像样本之间的样本相似度。

[0130] 根据不同降采样倍数对应的样本相似度,计算第一文本区域与第二文本区域之间的相似度。

[0131] 之后,本发明提出的方法可以提取视频中的文字,对生成文字轨迹的第一帧或任选一帧,进行文字识别处理,即可获得视频中包含的文字。

[0132] 在相似度大于相似阈值的情况下,将第二文本区域关联至视频文本跟踪过程中第一文本区域所在的目标文本轨迹。

[0133] 按照上述过程,针对目标视频生成多个文本轨迹。

[0134] 之后基于目标视频的文本轨迹,进行文本识别。

[0135] 确定目标视频中的目标文本轨迹,其中,目标文本轨迹为通过上述方法得到的。

[0136] 针对每一个目标文本轨迹,将目标文本轨迹中的任一帧图像作为目标帧图像,并确定目标帧图像中的目标文本区域。

[0137] 识别目标视频中所有目标文本区域的文字。

[0138] 在对目标视频进行文本识别的过程中,仅针对同一个文本轨迹中的第一帧图像或任选一帧图像进行文字识别处理,即可确定目标视频中包含的文字。这样,无需对目标视频中的每一帧图像进行文字识别,降低了计算量,减少了文字识别耗费的时间,提高了效率。

[0139] 基于上述实施例,参阅图6所示,本申请实施例提供一种视频文本跟踪装置600,包

括：

[0140] 获取单元601,用于获取目标视频中相邻的第一帧图像以及第二帧图像；

[0141] 定位单元602,用于定位所述第一帧图像中的第一文本区域以及所述第二帧图像中的第二文本区域；

[0142] 对比单元603,用于确定所述第一文本区域与所述第二文本区域之间的相似度；

[0143] 关联单元604,用于在相似度大于相似阈值的情况下,将所述第二文本区域关联至视频文本跟踪过程中所述第一文本区域所在的目标文本轨迹。

[0144] 可选的,所述定位单元602,具体用于：

[0145] 将所述第一帧图像和所述第二帧图像输入孪生跟踪网络；所述孪生跟踪网络包括定位子网络和跟踪子网络；

[0146] 通过所述定位子网络定位所述第一帧图像中的第一文本区域；所述定位子网络利用训练图像以及所述训练图像中文本区域的位置信息进行训练得到对应的网络模型参数；

[0147] 通过所述跟踪子网络定位所述第二帧图像中的第二文本区域；所述跟踪子网络采用与所述定位子网络相同的网络模型参数；

[0148] 所述对比单元603,具体用于：

[0149] 利用所述孪生跟踪网络中的目标对比层确定所述第一文本区域的第一特征向量；

[0150] 利用所述孪生跟踪网络中的跟踪检测层确定所述第二文本区域的第二特征向量；

[0151] 所述孪生跟踪网络根据所述第一特征向量与所述第二特征向量,确定所述第一文本区域与所述第二文本区域之间的相似度。

[0152] 可选的,所述定位单元602,具体用于：

[0153] 利用所述定位子网络,将所述第一帧图像进行N次第一降采样处理,得到N个第一图像样本；所述N次第一降采样处理的倍数不相同；

[0154] 针对每个第一图像样本,所述定位子网络确定所述第一图像样本中第一文本区域的位置信息；

[0155] 利用所述跟踪子网络,将所述第二帧图像进行N次第二降采样处理,得到N个第二图像样本；所述第二降采样处理的倍数与所述第一降采样处理的倍数相对应；

[0156] 针对每个第二图像样本,所述跟踪子网络确定所述第二图像样本中第二文本区域的位置信息；

[0157] 所述对比单元603,具体用于：

[0158] 分别确定相同降采样倍数下,第一图像样本与第二图像样本之间的样本相似度。

[0159] 可选的,还包括计算单元605,用于：

[0160] 根据不同降采样倍数对应的样本相似度,计算所述第一文本区域与所述第二文本区域之间的相似度。

[0161] 可选的,还包括训练单元606,用于根据以下方式训练得到所述孪生跟踪网络：

[0162] 对初始孪生跟踪网络中的定位子网络进行训练,得到对应的定位子网络模型参数；

[0163] 获取训练样本,所述训练样本中包含有正样本和负样本,所述正样本包括文本内容相同的第一训练样本图像以及所述第一训练样本图像的标注相似度,所述负样本包括文本内容不同的第二训练样本图像以及所述第二训练样本图像的标注相似度；

[0164] 将所述训练样本输入初步训练的孪生跟踪网络,并根据输出的相似度及所述标注相似度计算损失函数,所述初步训练的孪生跟踪网络包括已确定的所述定位子网络模型参数;

[0165] 当所述损失函数小于预设阈值时,确定对应的参数为所述孪生跟踪网络对应的参数,得到所述孪生跟踪网络。

[0166] 基于相同的技术构思,本申请实施例提供了一种计算机设备,如图7所示,包括至少一个处理器701,以及与至少一个处理器连接的存储器702,本申请实施例中不限定处理器701与存储器702之间的具体连接介质,图7中处理器701和存储器702之间通过总线连接为例。总线可以分为地址总线、数据总线、控制总线等。

[0167] 在本申请实施例中,存储器702存储有可被至少一个处理器701执行的指令,至少一个处理器701通过执行存储器702存储的指令,可以执行前述的视频文本跟踪方法中所包括的步骤。

[0168] 其中,处理器701是计算机设备的控制中心,可以利用各种接口和线路连接终端设备的各个部分,通过运行或执行存储在存储器702内的指令以及调用存储在存储器702内的数据,从而获得客户端地址。可选的,处理器701可包括一个或多个处理单元,处理器701可集成应用处理器和调制解调处理器,其中,应用处理器主要处理操作系统、用户界面和应用程序等,调制解调处理器主要处理无线通信。可以理解的是,上述调制解调处理器也可以不集成到处理器701中。在一些实施例中,处理器701和存储器702可以在同一芯片上实现,在一些实施例中,它们也可以在独立的芯片上分别实现。

[0169] 处理器701可以是通用处理器,例如中央处理器(CPU)、数字信号处理器、专用集成电路(Application Specific Integrated Circuit,ASIC)、现场可编程门阵列或者其他可编程逻辑器件、分立门或者晶体管逻辑器件、分立硬件组件,可以实现或者执行本申请实施例中公开的各方法、步骤及逻辑框图。通用处理器可以是微处理器或者任何常规的处理器等。结合本申请实施例所公开的方法的步骤可以直接体现为硬件处理器执行完成,或者用处理器中的硬件及软件模块组合执行完成。

[0170] 存储器702作为一种非易失性计算机可读存储介质,可用于存储非易失性软件程序、非易失性计算机可执行程序以及模块。存储器702可以包括至少一种类型的存储介质,例如可以包括闪存、硬盘、多媒体卡、卡型存储器、随机访问存储器(Random Access Memory,RAM)、静态随机访问存储器(Static Random Access Memory,SRAM)、可编程只读存储器(Programmable Read Only Memory,PROM)、只读存储器(ReadOnly Memory,ROM)、带电可擦除可编程只读存储器(Electrically Erasable Programmable Read-Only Memory,EEPROM)、磁性存储器、磁盘、光盘等等。存储器702是能够用于携带或存储具有指令或数据结构形式的期望的程序代码并能够由计算机存取的任何其他介质,但不限于此。本申请实施例中的存储器702还可以是电路或者其它任意能够实现存储功能的装置,用于存储程序指令和/或数据。

[0171] 基于相同的技术构思,本申请实施例提供了一种计算机可读存储介质,其存储有可由计算机设备执行的计算机程序,当所述程序在计算机设备上运行时,使得所述计算机设备执行视频文本跟踪方法的步骤。

[0172] 本领域内的技术人员应明白,本申请的实施例可提供为方法、系统、或计算机程序

产品。因此,本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且,本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0173] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器,使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于产现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0174] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中,使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品,该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0175] 尽管已描述了本申请的优选实施例,但本领域内的技术人员一旦得知了基本创造性概念,则可对这些实施例作出另外的变更和修改。所以,所附权利要求意欲解释为包括优选实施例以及落入本申请范围的所有变更和修改。

[0176] 显然,本领域的技术人员可以对本申请实施例进行各种改动和变型而不脱离本申请实施例的精神和范围。这样,倘若本申请实施例的这些修改和变型属于本申请权利要求及其等同技术的范围之内,则本申请也意图包含这些改动和变型在内。

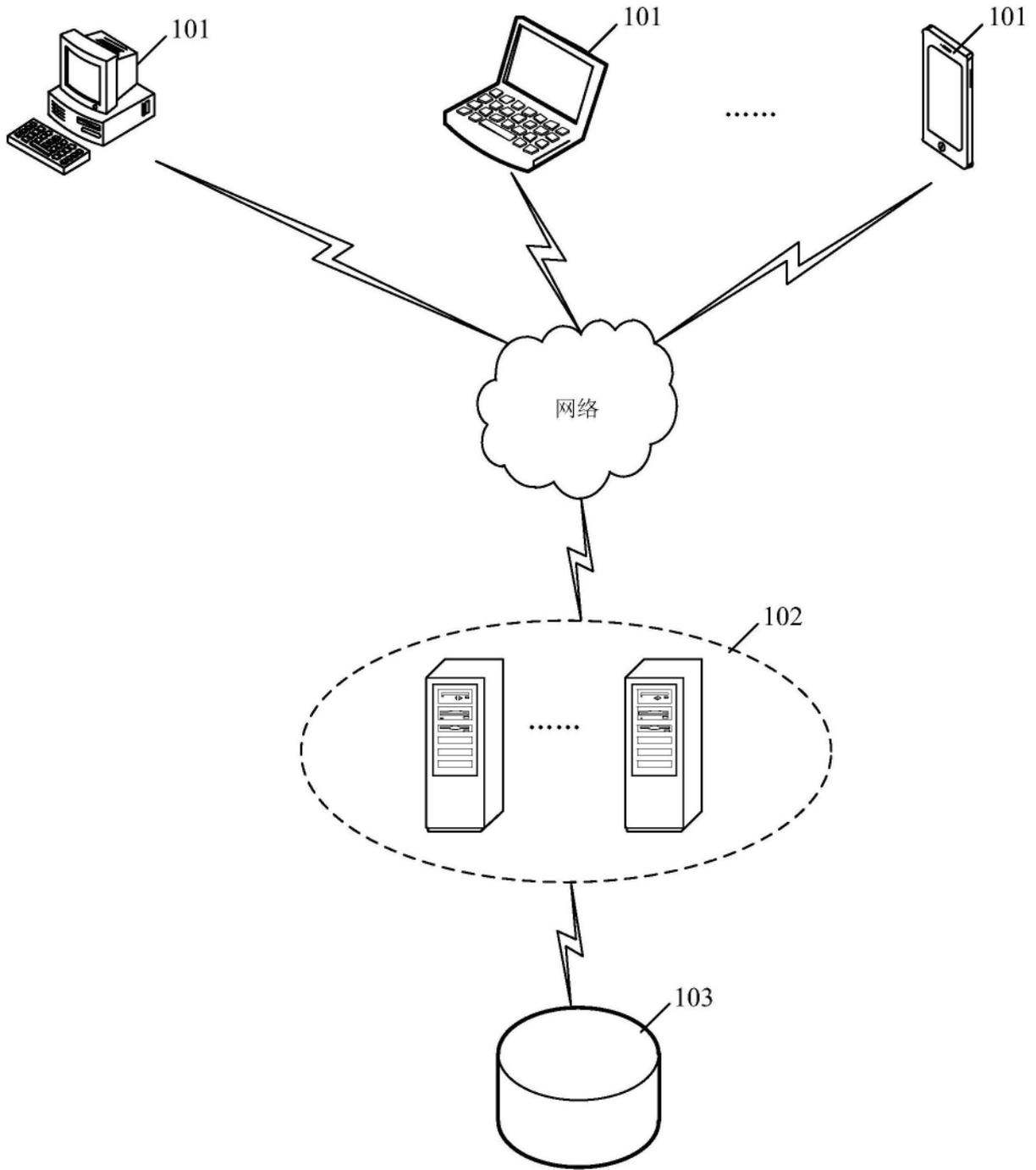


图1

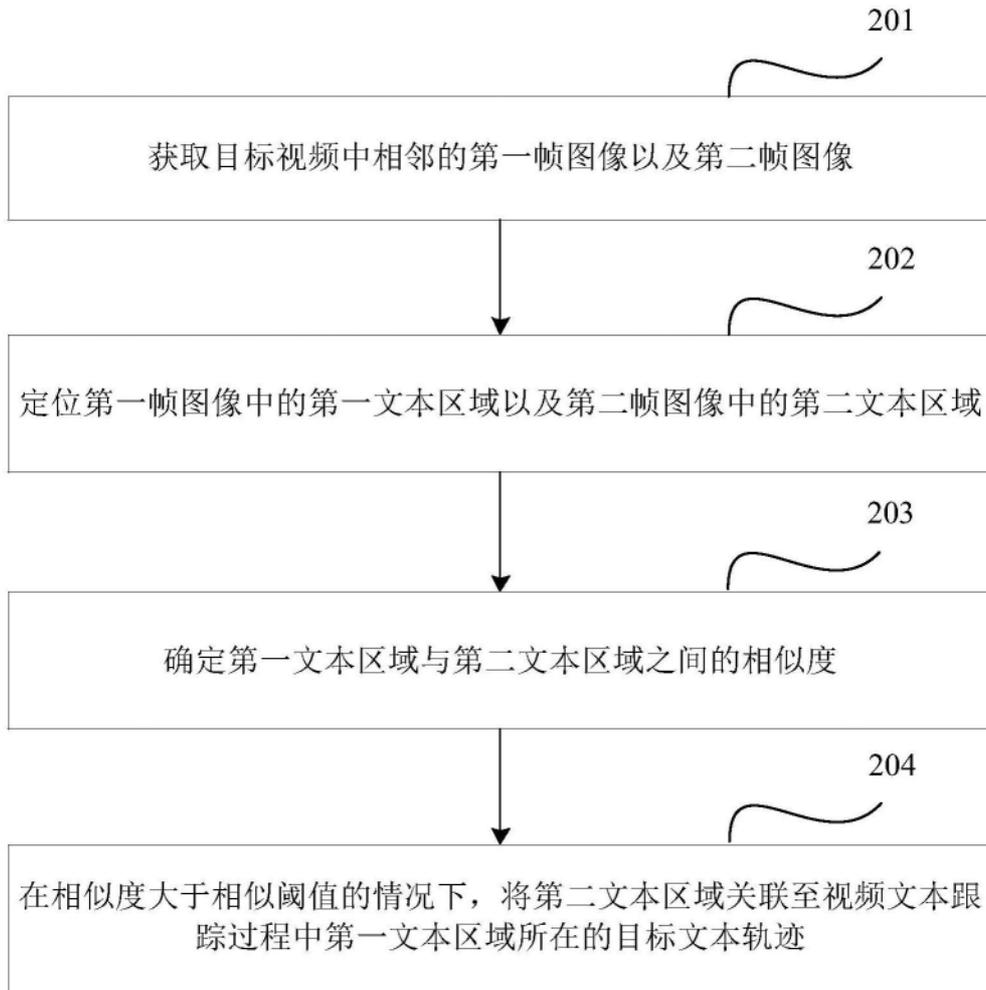


图2

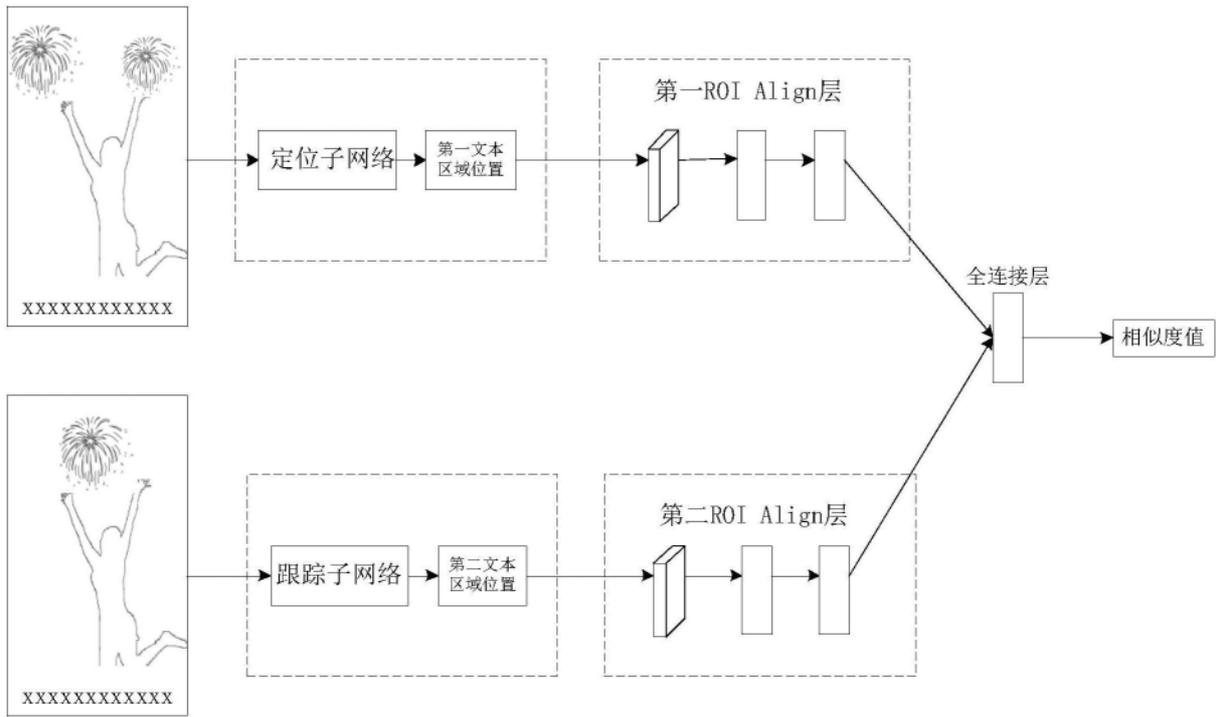


图3

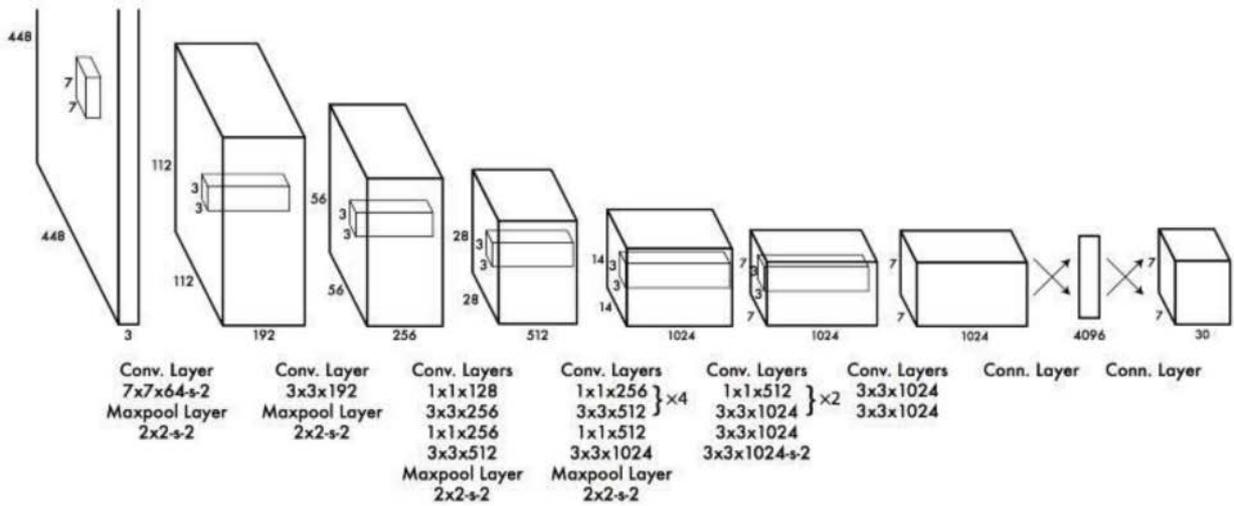


图4

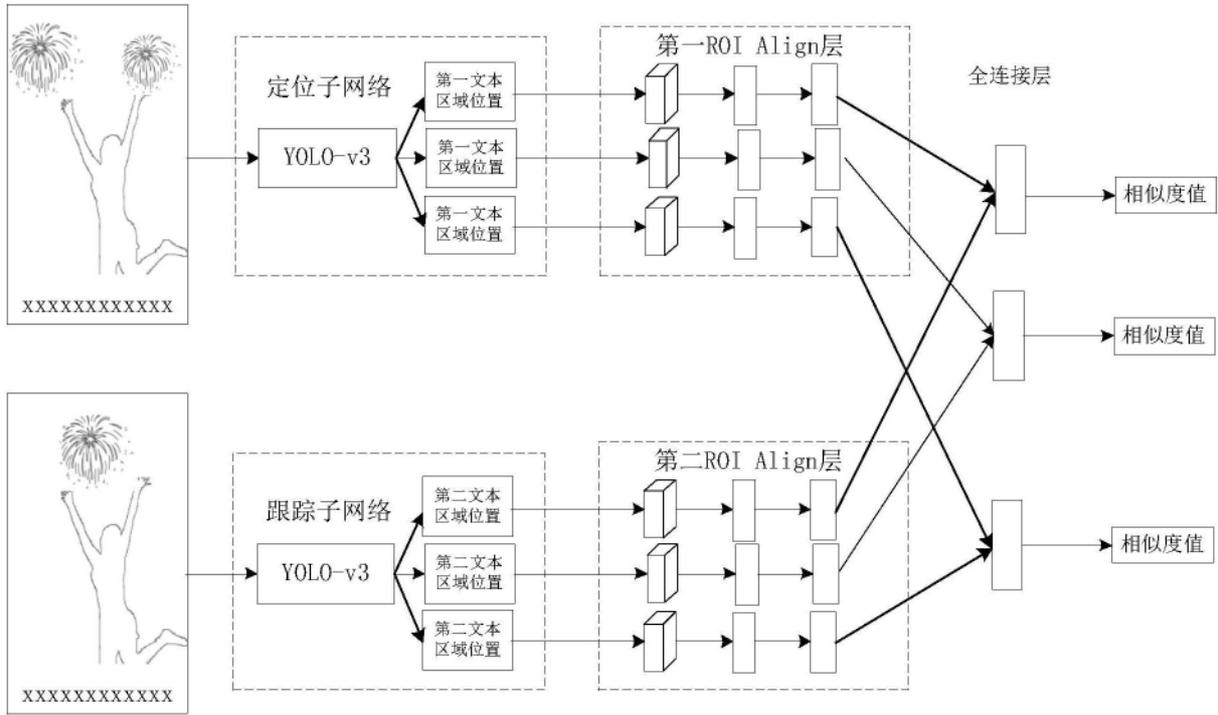


图5

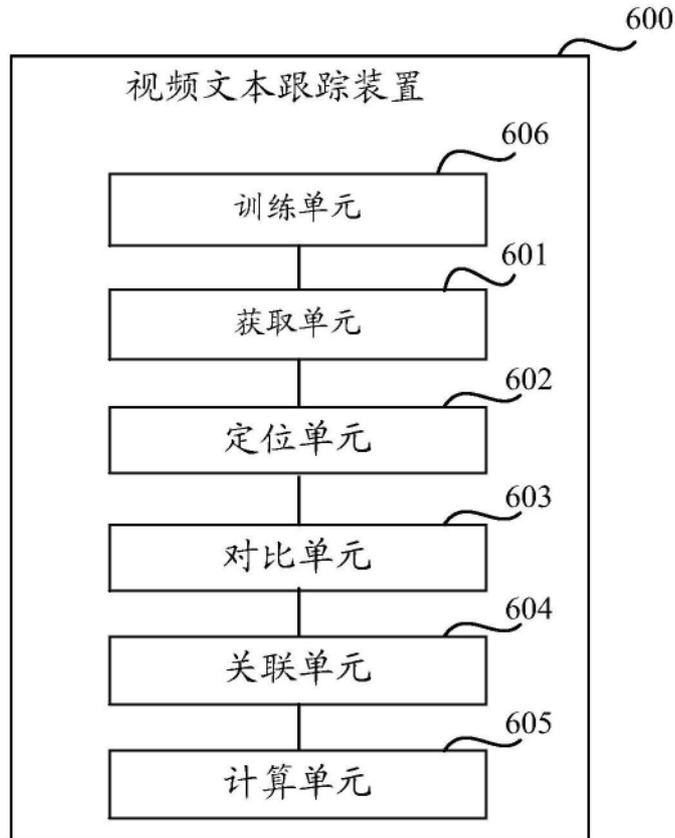


图6

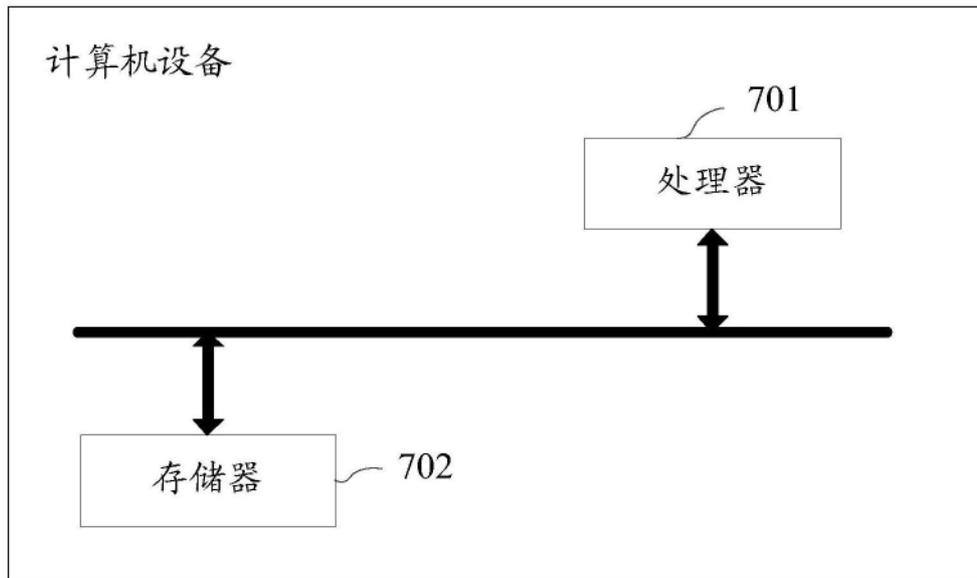


图7