(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 106845374 B
(45)授权公告日 2020.03.27

(21)申请号 201710010476.7

(22)申请日 2017.01.06

(65)同一申请的已公布的文献号
申请公布号 CN 106845374 A

(43)申请公布日 2017.06.13

(73)专利权人 清华大学
地址 100084 北京市海淀区清华园

(72)发明人 丁贵广 郝晖 陈仕江

(74)专利代理机构 北京清亦华知识产权代理事务所(普通合伙) 11201
代理人 张润

(51)Int.Cl.
G06K 9/00(2006.01)
G06N 3/08(2006.01)

(56)对比文件
CN 106250863 A,2016.12.21,
CN 102542289 A,2012.07.04,
CN 106022237 A,2016.10.12,
WO 2016095117 A1,2016.06.23,
CN 106203506 A,2016.12.07,
ShaoqingRen等.“Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks”.《NIPS’15 Proceedings of the 28th International Conference on Neural Information Processing Systems》.2015,
王斌.“基于深度学习的行人检测”.《中国优秀硕士学位论文全文数据库 信息科技辑》.2015,(第10期),
Girshick等.“Fast R-CNN”.《Proceedings of the IEEE International Conference on Computer Vision(ICCV)》.2015,

审查员 李亚楠

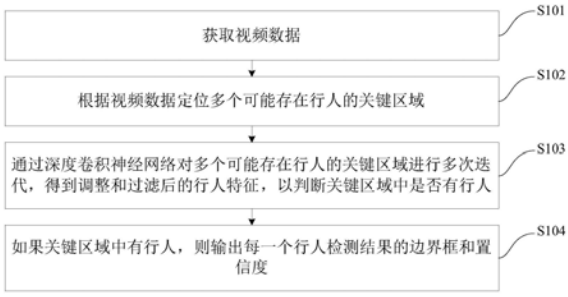
权利要求书2页 说明书11页 附图2页

(54)发明名称

基于深度学习的行人检测方法及检测装置

(57)摘要

本发明公开了一种基于深度学习的行人检测方法及检测装置,其中,方法包括:获取视频数据;根据视频数据定位多个可能存在行人的关键区域;通过深度卷积神经网络对多个可能存在行人的关键区域进行多次迭代,得到调整和过滤后的行人特征,以判断关键区域中是否有行人;如果关键区域中有行人,则输出每一个行人检测结果的边界框和置信度。该方法在行人检测中,可以提升关键区域检测效果和关键区域中行人检测效果,实现满足真实应用场景需要的高清视频实时行人检测的目的,不但提高检测的精确度,而且提高检测效率,简单易实现。



CN 106845374 B

1.一种基于深度学习的行人检测方法,其特征在于,包括以下步骤:

获取视频数据;

根据所述视频数据定位多个可能存在行人的关键区域;

通过深度卷积神经网络对所述多个可能存在行人的关键区域进行多次迭代,得到调整和过滤后的行人特征,以判断关键区域中是否有行人,其中,所述深度卷积神经网络包括多个卷积层、Roi采样层、全连接层和回归拟合层,以对所述视频数据中输入图片的多个关键区域进行统一采样和规范化的特征表示,并且对预测区域与标注数据进行回归拟合,以得到用于区域边界框的调整偏置,所述深度卷积神经网络的损失函数为:

$L(p,u,t^u,v)=L_{cls}(p,u)+\lambda[u\geq 1]L_{loc}(t^u,v)$

其中,L_{cls}表示关于行人及辅助类类别的损失函数,L_{loc}表示关于判断边界框位置的损失函数,u表示对象类别,v表示对象边界框,p表示辅助类别的离散概率分布,t^u表示行人对象边界框的预测结果,λ表示损失函数中的超参数;

在定位所述多个关键区域时,将每一个关键区域赋予一个初始类标,所述类标确定方式为:

$$G(t_i^u(s))=arg\max_{g\in G_i}IoU(t_i^u(1),g)$$

其中,$t_i^u(s)$表示在第s次迭代中第i个训练区域的位置,G_i表示与$t_i^u(s)$所在的图片上所有标注的目标检测区域,$t_i^u(1)$表示图像上原始划分的第i个训练区域,其中,在每一次迭代s中,将调整$t_i^u(s)$拟合回归其被确定的类标$G(t_i^u(s))$,在每一次迭代中的拟合目标为:

$$\Phi(t_i^u(s),G(t_i^u(s)),s)=t_i^u(s)+\frac{G(t_i^u(s))-t_i^u(s)}{S^*-s}$$

其中,S^*表示总的迭代次数;以及

如果所述关键区域中有行人,则输出每一个行人检测结果的边界框和置信度。

2.根据权利要求1所述的基于深度学习的行人检测方法,其特征在于,所述深度卷积神经网络在多次迭代训练过程中的目标函数为:

$$L(\{B_i\}_{i=1}^N)=\sum_{S=1}^{S^*}\left(\frac{1}{N_{cls}}\sum_i L_{cls}(p_i(s),p_i^*(s))+\lambda\frac{1}{N_{loc}}[u\geq 1]\sum_i p_i^*(s)L_{loc}(t_i(s),t_i^*(s))\right),$$

其中,$t_i(s)$表示第i次迭代中关于B_i的区域预测结果,$t_i^*(s)$表示拟合目标。

3.一种基于深度学习的行人检测装置,其特征在于,包括:

获取模块,用于获取视频数据;

定位模块,用于根据所述视频数据定位多个可能存在行人的关键区域;

判断模块,用于通过深度卷积神经网络对所述多个可能存在行人的关键区域进行多次迭代,得到调整和过滤后的行人特征,以判断关键区域中是否有行人,其中,所述深度卷积神经网络包括多个卷积层、Roi采样层、全连接层和回归拟合层,以对所述视频数据中输入图片的多个关键区域进行统一采样和规范化的特征表示,并且对预测区域与标注数据进行回归拟合,以得到用于区域边界框的调整偏置;

所述深度卷积神经网络的损失函数为:

$L(p,u,t^u,v)=L_{cls}(p,u)+\lambda[u\geq 1]L_{loc}(t^u,v)$

其中，L_{cls}表示关于行人及辅助类类别的损失函数，L_{loc}表示关于判断边界框位置的损失函数，u表示对象类别，v表示对象边界框，p表示辅助类别的离散概率分布，t^u表示行人对象边界框的预测结果，λ表示损失函数中的超参数；

在定位所述多个关键区域时，将每一个关键区域赋予一个初始类标，所述类标确定方式为：

$$G(t_i^u(s)) = arg\max_{g \in G_i} IoU(t_i^u(1), g)$$

其中，$t_i^u(s)$表示在第s次迭代中第i个训练区域的位置，G_i表示与$t_i^u(s)$所在的图片上所有标注的目标检测区域，$t_i^u(1)$表示图像上原始划分的第i个训练区域，其中，在每一次迭代s中，将调整$t_i^u(s)$拟合回归其被确定的类标$G(t_i^u(s))$，在每一次迭代中的拟合目标为：

$$\Phi(t_i^u(s), G(t_i^u(s)), s) = t_i^u(s) + \frac{G(t_i^u(s)) - t_i^u(s)}{S^* - s}$$

其中，S^*表示总的迭代次数；以及

输出模块，在所述关键区域中有行人时，用于输出每一个行人检测结果的边界框和置信度。

4.根据权利要求3所述的基于深度学习的行人检测装置，其特征在于，所述深度卷积神经网络在多次迭代训练过程中的目标函数为：

$$L(\{B_i\}_{i=1}^N) = \sum_{s=1}^{S^*} \left(\frac{1}{N_{cls}} \sum_i L_{cls}(p_i(s), p_i^*(s)) + \lambda \frac{1}{N_{loc}} [u \geq 1] \sum_i p_i^*(s) L_{loc}(t_i(s), t_i^*(s)) \right),$$

其中，$t_i(s)$表示第i次迭代中关于B_i的区域预测结果，$t_i^*(s)$表示拟合目标。

基于深度学习的行人检测方法及检测装置

技术领域

[0001] 本发明涉及计算机多媒体技术领域，特别涉及一种基于深度学习的行人检测方法及检测装置。

背景技术

[0002] 相关技术中，利用背景建模和统计学习的行人检测方法在特定条件下可能取得较好的行人检测效率或精确度，但这两种方法都还不能满足实际应用中的要求。其中，背景建模方法普遍比较复杂，导致不能满足实际应用中实时检测的需要，而基于统计学习的方法由于分类器训练比较复杂，尤其是样本量大时难以训练出通用的行人检测分类器，且分类器的训练时间开销极大，如果能提前检测出视频内容中的一些关键区域，然后提高在这些关键区域上行人检测的准确度，将能够在时间效率和检测精度上均获得提升。

[0003] 因此，如何更好地利用视频数据本身特征，在行人检测过程中减少对视频数据的重复处理，提升关键区域检测精度和关键区域上行人检测精度，仍需要进一步的研究。

发明内容

[0004] 本发明旨在至少在一定程度上解决相关技术中的技术问题之一。

[0005] 为此，本发明的一个目的在于提出一种基于深度学习的行人检测方法，该方法可以提高检测的精确度，且提高检测效率，简单易实现。

[0006] 本发明的另一个目的在于提出一种基于深度学习的行人检测装置。

[0007] 为达到上述目的，本发明一方面实施例提出了一种基于深度学习的行人检测方法，包括以下步骤：获取视频数据；根据所述视频数据定位多个可能存在行人的关键区域；通过深度卷积神经网络对所述多个可能存在行人的关键区域进行多次迭代，得到调整和过滤后的行人特征，以判断关键区域中是否有行人；如果所述关键区域中有行人，则输出每一个行人检测结果的边界框和置信度。

[0008] 本发明实施例的基于深度学习的行人检测方法，通过深度卷积神经网络对多个可能存在行人的关键区域进行多次迭代，从而得到调整和过滤后的行人特征，实现提升关键区域检测效果和关键区域中行人检测效果，有效满足真实应用场景需要的高清视频实时行人检测的目的，不但提高检测的精确度，而且提高检测效率，简单易实现。

[0009] 另外，根据本发明上述实施例的基于深度学习的行人检测方法还可以具有以下附加的技术特征：

[0010] 进一步地，在本发明的一个实施例中，所述深度卷积神经网络包括多个卷积层、Roi采样层、全连接层和回归拟合层，以对所述视频数据中输入图片的多个关键区域进行统一采样和规范化的特征表示，并且对预测区域与标注数据进行回归拟合，以得到用于区域边界框的调整偏置。

[0011] 进一步地，在本发明的一个实施例中，所述深度卷积神经网络的损失函数为：

[0012] $L(p,u,t^u,v)=L_{cls}(p,u)+\lambda[u\geq 1]L_{loc}(t^u,v)$，

[0013] 其中，L_{cls}表示关于行人及辅助类类别的损失函数，L_{loc}表示关于判断边界框位置的损失函数，u表示对象类别，v表示对象边界框，p表示辅助类别的离散概率分布，t^u表示行人对象边界框的预测结果，λ表示损失函数中的超参数。

[0014] 进一步地，在本发明的一个实施例中，在定位所述多个关键区域时，将每一个关键区域赋予一个初始类别，所述类标确定方式为：

[0015] $$G(t_i^u(s)) = arg\max_{g\in G_i} IoU(t_i^u(1), g),$$

[0016] 其中，$t_i^u(s)$表示在第s次迭代中第i个训练区域的位置，G_i表示与$t_i^u(s)$所在的图片上所有标注的目标检测区域，$t_i^u(1)$表示图像上原始划分的第i个训练区域，其中，在每一次迭代s中，将调整$t_i^u(s)$拟合回归其被确定的类标$G(t_i^u(s))$，在每一次迭代中的拟合目标为：

[0017] $$\Phi(t_i^u(s), G(t_i^u(s)), s) = t_i^u(s) + \frac{G(t_i^u(s)) - t_i^u(s)}{S^* - s},$$

[0018] 其中，S^*表示总的迭代次数。

[0019] 进一步地，在本发明的一个实施例中，所述深度卷积神经网络在多次迭代训练过程中的目标函数为：

[0020] $$L(\{B_i\}_{i=1}^N) = \sum_{s=1}^{S^*}\left(\frac{1}{N_{cls}}\sum_i L_{cls}(p_i(s), p_i^*(s)) + \lambda\frac{1}{N_{loc}}[u \geq 1]\sum_i p_i^*(s) L_{loc}(t_i(s), t_i^*(s))\right),$$

[0021] 其中，$t_i(s)$表示第i次迭代中关于B_i的区域预测结果，$t_i^*(s)$表示拟合目标。

[0022] 为达到上述目的，本发明另一方面实施例提出了一种基于深度学习的行人检测装置，包括：获取模块，用于获取视频数据；定位模块，用于根据所述视频数据定位多个可能存在行人的关键区域；判断模块，用于通过深度卷积神经网络对所述多个可能存在行人的关键区域进行多次迭代，得到调整和过滤后的行人特征，以判断关键区域中是否有行人；输出模块，在所述关键区域中有行人时，用于输出每一个行人检测结果的边界框和置信度。

[0023] 本发明实施例的基于深度学习的行人检测装置，通过深度卷积神经网络对多个可能存在行人的关键区域进行多次迭代，从而得到调整和过滤后的行人特征，实现提升关键区域检测效果和关键区域中行人检测效果，有效满足真实应用场景需要的高清视频实时行人检 测的目的，不但提高检测的精确度，而且提高检测效率，简单易实现。

[0024] 另外，根据本发明上述实施例的基于深度学习的行人检测装置还可以具有以下附加的技术特征：

[0025] 进一步地，在本发明的一个实施例中，所述深度卷积神经网络包括多个卷积层、Roi采样层、全连接层和回归拟合层，以对所述视频数据中输入图片的多个关键区域进行统一采样和规范化的特征表示，并且对预测区域与标注数据进行回归拟合，以得到用于区域边界框的调整偏置。

[0026] 进一步地，在本发明的一个实施例中，所述深度卷积神经网络的损失函数为：

[0027] $L(p,u,t^u,v) = L_{cls}(p,u) + \lambda[u \geq 1]L_{loc}(t^u,v)$，

[0028] 其中，L_{cls}表示关于行人及辅助类类别的损失函数，L_{loc}表示关于判断边界框位置的损失函数，u表示对象类别，v表示对象边界框，p表示辅助类别的离散概率分布，t^u表示行人对象边界框的预测结果，λ表示损失函数中的超参数。

[0029] 进一步地，在本发明的一个实施例中，在定位所述多个关键区域时，将每一个关键区域赋予一个初始类别，所述类标确定方式为：

[0030] $G(t_i^u(s)) = arg\max_{g\in G_i} IoU(t_i^u(1), g)$，

[0031] 其中，$t_i^u(s)$表示在第s次迭代中第i个训练区域的位置，G_i表示与$t_i^u(s)$所在的图片上所有标注的目标检测区域，$t_i^u(1)$表示图像上原始划分的第i个训练区域，其中，在每一次迭代s中，将调整$t_i^u(s)$拟合回归其被确定的类标$G(t_i^u(s))$，在每一次迭代中的拟合目标为：

[0032] $\Phi(t_i^u(s), G(t_i^u(s)), s) = t_i^u(s) + \frac{G(t_i^u(s)) - t_i^u(s)}{S^* - s}$，

[0033] 其中，S^*表示总的迭代次数。

[0034] 进一步地，在本发明的一个实施例中，所述深度卷积神经网络在多次迭代训练过程中的目标函数为：

[0035] $L(\{B_i\}_{i=1}^N) = \sum_{s=1}^{S^*}\left(\frac{1}{N_{cls}}\sum_i L_{cls}(p_i(s), p_i^*(s)) + \lambda\frac{1}{N_{loc}}[u \geq 1]\sum_i p_i^*(s) L_{loc}(t_i(s), t_i^*(s))\right)$，

[0036] 其中，$t_i(s)$表示第i次迭代中关于B_i的区域预测结果，$t_i^*(s)$表示拟合目标。

[0037] 本发明附加的方面和优点将在下面的描述中部分给出，部分将从下面的描述中变得明显，或通过本发明的实践了解到。

附图说明

[0038] 本发明上述的和/或附加的方面和优点从下面结合附图对实施例的描述中将变得明显和容易理解，其中：

[0039] 图1为根据本发明实施例的基于深度学习的行人检测方法的流程图；

[0040] 图2为根据本发明一个实施例的基于深度学习的行人检测方法的原理示意图；

[0041] 图3为根据本发明一个实施例的卷积神经网络的结构示意图；

[0042] 图4为根据本发明一个实施例的基于深度学习的行人检测方法的检测结果示意图；

[0043] 图5为根据本发明实施例的基于深度学习的行人检测装置的结构示意图。

具体实施方式

[0044] 下面详细描述本发明的实施例，所述实施例的示例在附图中示出，其中自始至终相同或类似的标号表示相同或类似的元件或具有相同或类似功能的元件。下面通过参考附图描述的实施例是示例性的，旨在用于解释本发明，而不能理解为对本发明的限制。

[0045] 下面在描述根据本发明实施例提出的基于深度学习的行人检测方法及检测装置之前，先来简单描述一下准确检测行人的重要性。

[0046] 目前，行人检测技术在诸多现实场景中有着广泛的应用：智能辅助驾驶，智能监控，行人分析以及智能机器人等领域。随着智能辅助驾驶及智能机器人的飞速发展，行人检测技术近年来进入了一个快速的发展阶段，但也有很多问题还有待解决，这其中主要是大规模视频数据，尤其是大规模高清监控视频数据或行车记录视频中关于行人检测的效率和

精度之间的权衡。

[0047] 比较经典的行人检测方法大致可以分为两类:基于背景建模的行人检测方法和基于统计学习的行人检测方法。基于背景建模的行人检测方法是首先利用背景建模的方法,提取视频画面中前景运动的目标,在目标区域内进行特征抽取及分类器学习,进而判断其中是否包含行人。仅对背景基本能保持不变的监控视频而言,其中会出现由于光照的变化导致的图像色度等的变化,基于背景建模的方法很难处理这种由于环境变化而带来的视频内容变化对行人检测所产生的影响。而在手持摄像机拍摄的视频或者行车记录设备等拍摄的视频中,背景可能会随摄像机的移动产生变化,这种情况下基于背景建模的行人检测方法极容易失效。此外,当视频中行人或一些干扰对象,如树叶等出现比较密集时,会对背景造成较大的遮挡,为背景建模带来极大的困难,而且背景中可能会出现一些物体的改变,比如停车场中车辆的移动等。如果这些情况下的背景信息不能及时被校正,那么背景建模的失效会直接导致行人检测的低准确率。基于统计学习的行人检测方法是另外一类常用的行人检测方法,具体来说这类方法依据大量的样本构建行人检测分类器,通过对样本的特征提取和分类器训练来实现行人检测。常用的特征主要包括目标的颜色直方图、边缘纹理信息等,常用的分类器主要包括神经网络、支持向量机,其中目前图像识别与分类领域运用得最广泛的是卷积神经网络。同基于背景建模的方法一样,基于统计学习的行人检测方法也还存在着一些难以解决的问题,比如在视频内容中,行人距离摄像机距离的远近造成行 人尺度变化很大;且行人在视频数据中所表现出来的姿势、穿着也各不一样;视频拍摄时光照条件等环境因素的变化也会给行人检测带来极大的不方便。而基于统计学习的方法在通过训练分类器达到较高的检测精度的同时,也受到自身固有弊端的不利影响,比如在视频数据中抽取的特征的有效性在很大程度上影响着后面的分类器训练和最终的行人检测效果,而分类器的训练也受到训练样本的极大影响。在实际应用中,分类器训练一般采用离线训练的方式,即先选取实际应用中的部分样本训练出一个分类器,然后应用到后续的检测任务中,而样本选择本身就是一个难以评估与优化的问题,离线分类器训练中使用的样本基本无法涵盖到真实应用场景中的所有情况,在遇到新的场景时分类器的应用可能会失效。近年来,随着神经网络在图像、音频分类与识别领域的良好表现,针对图像分类与识别任务进行优化的卷积神经网络方法开始被广泛地应用到图像分类、视频事件检测等任务中,卷积神经网络在一定程度上克服了深度神经网络方法中网络参数过多,训练过程漫长且训练不易收敛的问题,但针对具体任务如何设计高效且简洁的卷积神经网络结构仍是当前大规模多媒体数据检索与识别中的一个重要问题。

[0048] 本发明正是基于上述问题,而提出了一种基于深度学习的行人检测方法与一种基于深度学习的行人检测装置。

[0049] 下面参照附图描述根据本发明实施例提出的基于深度学习的行人检测方法及检测装置,首先将参照附图描述根据本发明实施例提出的基于深度学习的行人检测方法。

[0050] 图1是本发明实施例的基于深度学习的行人检测方法。

[0051] 如图1所示,该基于深度学习的行人检测方法包括以下步骤:

[0052] 在步骤S101中,获取视频数据。

[0053] 在步骤S102中,根据视频数据定位多个可能存在行人的关键区域。

[0054] 可以理解的是,如图2所示,由于一般进行行人检测时视野中出现的行人数目有

限，为了在保证检测精度的条件下大大提高检测过程运行效率，以满足实际应用需求，本发明实施例的方法首先将图片划分为若干区域(数目可视具体应用场景而定)，每一个区域当成一个可能出现行人的感兴趣区域。

[0055] 在步骤S103中，通过深度卷积神经网络对多个可能存在行人的关键区域进行多次迭代，得到调整和过滤后的行人特征，以判断关键区域中是否有行人。

[0056] 也就是说，如图2所示，通过设计针对行人检测应用场景的卷积神经网络结构，并利用合适的训练数据以及对应的损失函数(Loss Function)，对卷积神经网络模型参数进行训练，实现从视频图像帧到行人检测感兴趣区域(Region of Interest,RoI)及行人边界框(Bounding Box)的直接输出，下面将进行详细赘述。

[0057] 具体地，卷积神经网络由于其较传统神经网络参数数目更少，特征抽取更为完备，因而被大量用于图像视频等可视数据的处理中，本发明实施例同样采用的是针对行人检测设计的卷积神经网络。通过对这一网络进行训练，可以由视频的图像帧直接得到行人检测结果，包括对视频图像帧中是否出现行人进行判断，且在有行人时，输出关于每一个行人检测结果的边界框及置信度等相关信息。

[0058] 需要说明的是，如图2所示，在本发明实施例中，在保证行人检测和相关输出结果精确度的同时，本发明实施例通过预先在图片上划分感兴趣区域避免使用不同尺度的滑动窗口对图像进行遍历，考虑到主要使用的卷积神经网络的运行效率，保证了行人检测过程的准确与高效。

[0059] 因此，本发明实施例有效地借助深度学习的思想，并使用深度学习领域适合图像处理的卷积神经网络，通过精心设计卷积神经网络结构并学习网络参数，可以由视频图像帧直接获得行人检测的输出结果。同时，通过预先选取行人检测感兴趣区域减少对图像区域的重复处理，保证了网络运行的效率，提高行人检测处理速度。从标准数据集实验结果来看，本发明实施例的基于深度学习的行人检测方法具有精确度高、实时性强、易于移植到其它应用场景等特点，能满足实际应用场景的需求。

[0060] 下面对深度卷积神经网络进行详细描述。

[0061] 其中，在本发明的一个实施例中，深度卷积神经网络包括多个卷积层、Roi采样层、全连接层和回归拟合层，以对视频数据中输入图片的多个关键区域进行统一采样和规范化的特征表示，并且对预测区域与标注数据进行回归拟合，以得到用于区域边界框的调整偏置。

[0062] 可以理解的是，如图3所示，第一步，构造网络结构。其中，由于在对视频进行处理时，往往相当于直接对视频的图像帧进行处理，而图像一般被表示为像素的向量，如今随着视频采集过程中清晰度的提升，高清视频图像帧会被表示成一个很长的向量。在传统深度学习方法使用的神经网络结构中，由于网络各层之间的节点采用全连接的方式，如果直接用来进行图像处理会导致参数数目过多，无法对网络参数进行训练，因而为了将深度学习方法应用到图像处理中，必须减少神经网络结构中的参数个数以加快速度，这就推动了卷积神经网络的发展。

[0063] 卷积神经网络主要通过两种方式来减少参数数目，其一是局部感知野，在图像处理研究中可以发现，图像的空间联系表现为局部像素联系较为紧密，而距离较远的像素可能表现出的相关性较弱。因而，在设计神经网络结构时，网络中每个神经元没必要对全局图

像进行感知，只需要对图像的某一局部区域进行感知，然后在神经网络后期，即比较高层的网络结构中将图像的这些局部信息综合起来得到图像的全局信息。其二是参数共享，又称权值共享，在传统的神经网络中，每个神经元的参数需要分别进行训练，而卷积神经网络中引入了权值共享的思想，这样更进一步地压缩了网络中的参数个数。权值共享是指在某 一个具体的网络层，每一个神经元对应的网络参数都是统一的，基于局部感知的设定，每个神经元对应的参数都可以认为是该层对应的特征抽取方式，且特征抽取网与神经元对应的局部感知野无关，在卷积神经网络中，每个神经元对应的参数被称为卷积核。卷积神经网络中局部感知野和参数共享的设定大大减少了网络训练过程中的参数，保证了网络训练和运行的效率，同时为了保证较完备的特征抽取，在卷积神经网络结构中一般采取多卷积核及多层卷积的设计。多卷积核是为了保证在每一个卷积层中特征抽取尽可能充分，而由于一个单独的卷积层学习到的特征是局部的，因此会采用多层卷积的方式来获得关于图像更加全局的特征，且多层卷积后一般会接上全连接层将特征变换为向量的形式。

[0064] 如图3所示，在本发明的实施例中，使用如图所示的卷积神经网络结构，将输入图片划分为若干区域后，经过若干卷积层，通过对这些感兴趣区域进行统一采样，然后经过全连接层得到规范化的特征表示，之后通过对预测区域与标注数据进行回归拟合，得到对行人边界框的调整偏置，且调整偏置向量将会应用到输入的区域边界框上。

[0065] 进一步地，第二步，获取训练数据。其中，虽然卷积神经网络与传统深度神经网络相比参数数量已大大减少，但由于其网络层数多，结构复杂，其中的参数规模仍然十分庞大。因此，如果从随机初值开始对深层卷积神经网络进行训练时，需要有准确可靠且规模庞大的训练数据。针对行人检测任务，图像训练数据中不仅需要包含每一幅图像中是否包含行人的判断，还需要在标注确定有行人时，同时给出行人的边界框，这样给行人检测的卷积神经网络训练的数据准备带来了极大的挑战。为了提高训练数据准备的效率，减少在网络训练过程中的开销，本发明主要采用两种方法来实现以较小规模的训练数据集对网络参数进行训练。其一是借助现有公开的网络结构及参数，在ImageNet LSVRC和MicrosoftCOCO等大型比赛中，都设定了对象检测的任务，即需要在大规模图像数据中识别出指定的若干类别的对象。许多参赛队伍公开发布其在对象检测任务中取得较好效果的卷积神经网络结构及对应网络参数，本发明通过借助这些公开发布的网络结构并针对行人检测任务对网络结构进行修改，然后使用预先训练好的网络参数对未做修改的层进行初始化，接着使用针对行人检测任务准备的训练数据对修改后的网络参数进行调整和重新训练，减少在训练过程中需要重新修改的网络参数，加快网络训练过程。其二是对现有的训练数据进行变化补充训练数据，具体而言，对每一个具有行人标注的样本，通过对其进行相关的平移、旋转等几何变换及改变亮度、色度等模拟环境变化来产生新的标注样本，这样可以用来扩充训练样本数据，在训练网络参数时保证卷积神经网络可以对各种不同环境条件和姿势的行人特征进行检测。

[0066] 例如，使用的训练数据可以来源于CaltechPedestrian公开数据集，其中包含了六个不同的训练数据集合，每个训练集合包含6-13个一分钟时长的视频序列。此外，在设计卷积神 经网络结构时，为了减少在行人检测过程中对外观类似对象的误检，在最终网络输出中添加了若干辅助类别，这部分的训练数据可以来源于ImageNetLSVRC比赛中的训练数据。

[0067] 进一步地，在本发明的一个实施例中，深度卷积神经网络的损失函数为：

[0068] $L(p,u,t^u,v)=L_{cls}(p,u)+\lambda[u\geq 1]L_{loc}(tu,v)$，

[0069] 其中，L_{cls}表示关于行人及辅助类类别的损失函数，L_{loc}表示关于判断边界框位置的损失函数，u表示对象类别，v表示对象边界框，p表示辅助类别的离散概率分布，t^u表示行人对象边界框的预测结果，λ表示损失函数中的超参数。

[0070] 进一步地，在本发明的一个实施例中，在定位多个关键区域时，将每一个关键区域赋予一个初始类别，类标确定方式为：

[0071] $$G(t_i^u(s))=arg\max_{g\in G_i} IoU(t_i^u(1),g),$$

[0072] 其中，$t_i^u(s)$表示在第s次迭代中第i个训练区域的位置，G_i表示与$t_i^u(s)$所在的图片上所有标注的目标检测区域，$t_i^u(1)$表示图像上原始划分的第i个训练区域，其中，在每一次迭代s中，将调整$t_i^u(s)$拟合回归其被确定的类标$G(t_i^u(s))$，在每一次迭代中的拟合目标为：

[0073] $$\Phi(t_i^u(s),G(t_i^u(s)),s)=t_i^u(s)+\frac{G(t_i^u(s))-t_i^u(s)}{S^*-s},$$

[0074] 其中，S^*表示总的迭代次数。

[0075] 进一步地，在本发明的一个实施例中，深度卷积神经网络在多次迭代训练过程中的目标函数为：

[0076] $$L(\{B_i\}_{i=1}^N)=\sum_{s=1}^{S^*}\left(\frac{1}{N_{cls}}\sum_i L_{cls}(p_i(s),p_i^*(s))+\lambda\frac{1}{N_{loc}}[u\geq 1]\sum_i p_i^*(s)L_{loc}(t_i(s),t_i^*(s))\right),$$

[0077] 其中，$t_i(s)$表示第i次迭代中关于B_i的区域预测结果，$t_i^*(s)$表示拟合目标。

[0078] 具体地，第三步，构造损失函数。针对行人检测设计的卷积神经网络结构包含两个并列的输出，其中一个直接输出每一个感兴趣区域(RoI)关于行人、背景及若干辅助类别的离散概率分布$p=(p_0,\cdots,p_K)$，其中K为除背景以外类别个数；另外一个输出是对于检测出的行人对象边界框的预测结果，$t^u=\left(t_x^u,t_y^u,t_w^u,t_h^u\right)$，其中x，y，w，h分别表示预测出的边界框在图片上的横向位置、纵向位置、边界框宽度、边界框高度，均以像素作为计量单位。对于训练数据集中的每一个图像帧，其中包含多个对象区域的标注，每一个对象区域均包含对象类别u和对象边界框v。为了对网络参数进行训练，本发明针对行人检测任务使用如下损失函数：

[0079] $L(p,u,t^u,v)=L_{cls}(p,u)+\lambda[u\geq 1]L_{loc}(t^u,v)$，

[0080] 其中，L_{cls}是关于行人及辅助类类别的损失函数，L_{loc}是关于判断边界框位置的损失函数，对网络参数进行训练的目标是最小化损失函数值。本发明实施例在训练过程中，对行人类别及每一个辅助类别，L_{cls}使用对数损失函数：

[0081] $L_{cls}(p,u)=-\log p_u$，

[0082] L_{loc}对于每一个类别u，其定义在关于每一个类别u，边界框标注$v=(v_x,v_y,v_w,v_h)$及边界框预测结果$t^u=\left(t_x^u,t_y^u,t_w^u,t_h^u\right)$上。当u＝0时，其表示的是背景类别，故[u≥1]当u≥1时值为1，否则为0，表示仅考虑除背景以外的类别。L_{loc}定义为：

[0083] $$L_{loc}(t^u,v)=\sum_{i\in\{x,y,w,h\}}\text{smooth}_{L_1}(t_i^u-v_i),$$

[0084] 其中，

[0085] $$\mathrm{smooth}_{L_1}(x)=\begin{cases}0.5x^2 \text{if } |x|<1\\ |x|-0.5 \text{ otherwise}\end{cases},$$

[0086] 其是一种L_1归一化方法，比在一些其它的卷积神经网络方法，如R-CNN和SPPnet中使用的L_2归一化方法对于离群值能保持更强的不变性，更加鲁棒。

[0087] 损失函数中的超参数λ用来均衡类别预测损失函数和对象位置预测损失函数，本发明在对卷积神经网络进行训练时，首先将所有边界框标注v_i进行标准正态分布归一化，即均值为0标准差为1。然后在所有不同配置的实验中均使用λ＝1的设置。

[0088] 为了去除在网络训练过程中需要人为确定感兴趣区域这一耗时步骤，本发明实施例在将图片放入卷积神经网络进行训练前，首先将图片按一定规则划分为若干不重叠的矩形区域。在划分区域时可以使用任意的方式，比如按尺寸比例划分等，如图2中就是将图片划分为四个同样大小的区域。每一个划分出的区域会被赋予一个初始类标，类标确定方式为：

[0089] $$G(t_i^u(s))=arg\max_{g\in G_i} IoU(t_i^u(1),g),$$

[0090] 其中，$t_i^u(s)$是在第s次迭代中第i个训练区域的位置，G_i是与$t_i^u(s)$所在的图片上所有标注的目标检测区域，$t_i^u(1)$即该图像上原始划分的第i个训练区域。在每一次迭代s中，本发明实施例提出的网络都将调整$t_i^u(s)$拟合回归其被确定的类标$G(t_i^u(s))$，在每一次迭代中的拟合目标为：

[0091] $$\Phi(t_i^u(s),G(t_i^u(s)),s)=t_i^u(s)+\frac{G(t_i^u(s))-t_i^u(s)}{S^*-s},$$

[0092] 其中，S^*为总的迭代次数。

[0093] 因此，在本发明实施例提出的多次迭代的卷积神经网络训练过程中的目标函数可以表示为：

[0094] $$L(\{B_i\}_{i=1}^N)=\sum_{s=1}^{S^*}\left(\frac{1}{N_{cls}}\sum_i L_{cls}(p_i(s),p_i^*(s))+\lambda\frac{1}{N_{loc}}[u\geq 1]\sum_i p_i^*(s)L_{loc}(t_i(s),t_i^*(s))\right),$$

[0095] 其中，$t_i(s)$为第i次迭代中关于B_i的区域预测结果，$t_i^*(s)$为其拟合目标Φ($t_i(s)$，$G(t_i(s))$,s)。

[0096] 综上，整个卷积神经网络的损失函数已经被确定，可以采用合适的方式对卷积神经网络进行参数调整或重新训练。

[0097] 进一步地，第四步，训练网络参数。其中，在确定网络损失函数之后，可以使用现有深度学习工具Caffe对网络进行参数调整与重新训练，在训练过程中可以交替训练预测网络和类别预测网络或直接使用端对端的方式，即从图片到最终的行人检测结果，进行训练，训练后的网络可以直接用于对视频图像帧中的行人进行检测。

[0098] 最后，第五步，产生检测结果。综上所述，训练好的卷积神经网络可以从图像帧直接输出行人检测结果及检测出的行人所在区域。由于在网络中直接添加了感兴趣区域的预测，避免了滑动窗口或人为指定感兴趣区域所产生的大量重复运算，网络运行效率大大提升，可以满足实际行人检测应用需要，如图4所示，其为某行车拍摄视频图像帧中的行人检

测结果，其中，行人检测示例中包含辅助类别。

[0099] 在步骤S104中，如果关键区域中有行人，则输出每一个行人检测结果的边界框和置信度。

[0100] 即言，通过利用卷积神经网络对这些区域进行不断调整和过滤，最终实现行人检测结果的输出，具有如下优点：

[0101] 1)通过使用多次迭代的卷积神经网络直接从视频图像帧中抽取行人特征，对行人特征进行有效理解，能够识别不同姿势、不同尺度的行人，极大程度减少传统方法中环境变化等因素对检测效果带来的负面影响，提升行人检测的精确度。

[0102] 2)通过对图片进行区域划分以及在网络运行过程中对可能存在行人的感兴趣区域进行筛选，避免传统方法需要预先获取感兴趣区域或者通过在图像上使用滑动窗口对每个窗口进行判断等带来的大量重复计算，极大提高了网络运行效率，可以满足实际应用场景的需求。

[0103] 3)通过在网络训练时使用辅助类标的思想，对一些传统方法极难分辨的对象进行单独建模，减少训练过程中难以判断的负例带来的干扰，进一步提升了网络训练的有效性和行人检测的精确度。

[0104] 举例而言，通过在行人检测领域的标准数据集Caltech Pedestrian、ETH上的实验，本发明实施例的方法表现出了有效性。具体来说，在Caltech Pedestrian数据集中，包含6个训练数据集和5个测试数据集，每个包含6-13个一分钟时长的视频序列，视频序列每一帧上都有关于每一个行人的标注。在ETH数据集中包含三个视频数据集，按实验标准配置划分训练集和数据集。在多组不同实验配置取得的实验结果中取效果最佳的输出作为最终实验输出，本发明提出的方法最终在两个数据集的标准配置下分别 取得了32.82％和38.17％的准确率，效果优异，且在Caltech Pedestrian和ETH上平均能达到18帧每秒的处理速度，可以满足实际应用需求。

[0105] 根据本发明实施例的基于深度学习的行人检测方法，通过深度卷积神经网络对多个可能存在行人的关键区域进行多次迭代，从而得到调整和过滤后的行人特征，实现提升关键区域检测效果和关键区域中行人检测效果，其中，利用深度学习的方法，主要是通过构造并训练针对行人检测应用场景的深度卷积神经网络，在大规模的监控或行车记录等视频数据中，首先高效率且较精准地定位出可能存在行人的关键区域，然后精确判断关键区域中是否存在行人，来取得高精确度的行人检测效果，以满足现实场景中的行人检测应用对时间效率和检测精度的要求，有效满足真实应用场景需要的高清视频实时行人检测的目的，不但提高检测的精确度，而且提高检测效率，简单易实现。

[0106] 其次参照附图描述根据本发明实施例提出的基于深度学习的行人检测装置。

[0107] 图5是本发明实施例的基于深度学习的行人检测装置的结构示意图。

[0108] 如图5所示，该基于深度学习的行人检测装置10包括：获取模块100、定位模块200、判断模块300和输出模块400。

[0109] 其中，获取模块100用于获取视频数据。定位模块200用于根据视频数据定位多个可能存在行人的关键区域。判断模块300用于通过深度卷积神经网络对多个可能存在行人的关键区域进行多次迭代，得到调整和过滤后的行人特征，以判断关键区域中是否有行人。在关键区域中有行人时，输出模块400用于输出每一个行人检测结果的边界框和置信度。本

发明实施例的装置10可以提升关键区域检测效果和关键区域中行人检测效果，实现满足真实应用场景需要的高清视频实时行人检测的目的，不但提高检测的精确度，而且提高检测效率，简单易实现。

[0110] 进一步地，在本发明的一个实施例中，深度卷积神经网络包括多个卷积层、Roi采样层、全连接层和回归拟合层，以对视频数据中输入图片的多个关键区域进行统一采样和规范化的特征表示，并且对预测区域与标注数据进行回归拟合，以得到用于区域边界框的调整偏置。

[0111] 进一步地，在本发明的一个实施例中，深度卷积神经网络的损失函数为：

[0112] $L(p,u,t^u,v)=L_{cls}(p,u)+\lambda[u\geq 1]L_{loc}(t^u,v)$，

[0113] 其中，L_{cls}表示关于行人及辅助类类别的损失函数，L_{loc}表示关于判断边界框位置的损失函数，u表示对象类别，v表示对象边界框，p表示辅助类别的离散概率分布，t^u表示行人对象边界框的预测结果，λ表示损失函数中的超参数。

[0114] 进一步地，在本发明的一个实施例中，在定位多个关键区域时，将每一个关键区域赋予一个初始类别，类标确定方式为：

[0115] $$G(t_i^u(s))=arg\max_{g\in G_i}IoU(t_i^u(1),g),$$

[0116] 其中，$t_i^u(s)$表示在第s次迭代中第i个训练区域的位置，G_i表示与$t_i^u(s)$所在的图片上所有标注的目标检测区域，$t_i^u(1)$表示图像上原始划分的第i个训练区域，其中，在每一次迭代s中，将调整$t_i^u(s)$拟合回归其被确定的类标$G(t_i^u(s))$，在每一次迭代中的拟合目标为：

[0117] $$\Phi(t_i^u(s),G(t_i^u(s)),s)=t_i^u(s)+\frac{G(t_i^u(s))-t_i^u(s)}{S^*-s},$$

[0118] 其中，S^*表示总的迭代次数。

[0119] 进一步地，在本发明的一个实施例中，深度卷积神经网络在多次迭代训练过程中的目标函数为：

[0120] $$L(\{B_i\}_{i=1}^N)=\sum_{s=1}^{S^*}\left(\frac{1}{N_{cls}}\sum_i L_{cls}(p_i(s),p_i^*(s))+\lambda\frac{1}{N_{loc}}[u\geq 1]\sum_i p_i^*(s)L_{loc}(t_i(s),t_i^*(s))\right),$$

[0121] 其中，$t_i(s)$表示第i次迭代中关于B_i的区域预测结果，$t_i^*(s)$表示拟合目标。

[0122] 需要说明的是，前述对基于深度学习的行人检测方法实施例的解释说明也适用于该实施例的基于深度学习的行人检测装置，此处不再赘述。

[0123] 根据本发明实施例的基于深度学习的行人检测装置，通过深度卷积神经网络对多个可能存在行人的关键区域进行多次迭代，从而得到调整和过滤后的行人特征，实现提升关键区域检测效果和关键区域中行人检测效果，其中，利用深度学习的方法，主要是通过构造并训练针对行人检测应用场景的深度卷积神经网络，在大规模的监控或行车记录等视频数据中，首先高效率且较精准地定位出可能存在行人的关键区域，然后精确判断关键区域中是否存在行人，来取得高精确度的行人检测效果，以满足现实场景中的行人检测应用对时间效率和检测精度的要求，有效满足真实应用场景需要的高清视频实时行人检测的目的，不但提高检测的精确度，而且提高检测效率，简单易实现。

[0124] 在本发明的描述中，需要理解的是，术语“中心”、“纵向”、“横向”、“长度”、“宽度”、“厚度”、“上”、“下”、“前”、“后”、“左”、“右”、“竖直”、“水平”、“顶”、“底”“内”、“外”、“顺时

针”、“逆时针”、“轴向”、“径向”、“周向”等指示的方位或位置关系为基于附图所示的方位或位置关系，仅是为了便于描述本发明和简化描述，而不是指示或暗示所指的装置或元件必须具有特定的方位、以特定的方位构造和操作，因此不能理解为对本发明的限制。

[0125] 此外，术语“第一”、“第二”仅用于描述目的，而不能理解为指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至少一个该特征。在本发明的描述中，“多个”的含义是至少两个，例如两个，三个等，除非另有明确具体的限定。

[0126] 在本发明中，除非另有明确的规定和限定，术语“安装”、“相连”、“连接”、“固定”等术语应做广义理解，例如，可以是固定连接，也可以是可拆卸连接，或成一体；可以是机械连接，也可以是电连接；可以是直接相连，也可以通过中间媒介间接相连，可以是两个元件内部的连通或两个元件的相互作用关系，除非另有明确的限定。对于本领域的普通技术人员而言，可以根据具体情况理解上述术语在本发明中的具体含义。

[0127] 在本发明中，除非另有明确的规定和限定，第一特征在第二特征“上”或“下”可以是第一和第二特征直接接触，或第一和第二特征通过中间媒介间接接触。而且，第一特征在第二特征“之上”、“上方”和“上面”可是第一特征在第二特征正上方或斜上方，或仅仅表示第一特征水平高度高于第二特征。第一特征在第二特征“之下”、“下方”和“下面”可以是第一特征在第二特征正下方或斜下方，或仅仅表示第一特征水平高度小于第二特征。

[0128] 在本说明书的描述中，参考术语“一个实施例”、“一些实施例”、“示例”、“具体示例”、或“一些示例”等的描述意指结合该实施例或示例描述的具体特征、结构、材料或者特点包含于本发明的至少一个实施例或示例中。在本说明书中，对上述术语的示意性表述不必须针对的是相同的实施例或示例。而且，描述的具体特征、结构、材料或者特点可以在任一个或多个实施例或示例中以合适的方式结合。此外，在不相互矛盾的情况下，本领域的技术人员可以将本说明书中描述的不同实施例或示例以及不同实施例或示例的特征进行结合和组合。

[0129] 尽管上面已经示出和描述了本发明的实施例，可以理解的是，上述实施例是示例性的，不能理解为对本发明的限制，本领域的普通技术人员在本发明的范围内可以对上述实施例进行变化、修改、替换和变型。

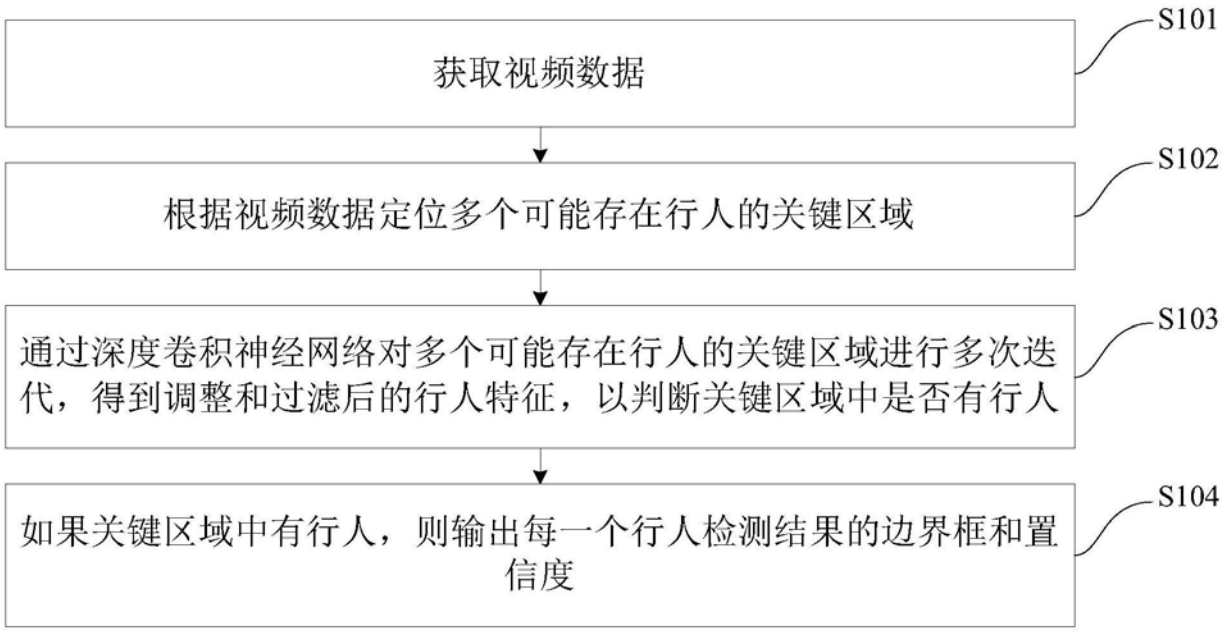

获取视频数据
S101
根据视频数据定位多个可能存在行人的关键区域
S102
通过深度卷积神经网络对多个可能存在行人的关键区域进行多次迭代，得到调整和过滤后的行人特征，以判断关键区域中是否有行人
S103
如果关键区域中有行人，则输出每一个行人检测结果的边界框和置信度
S104

图1

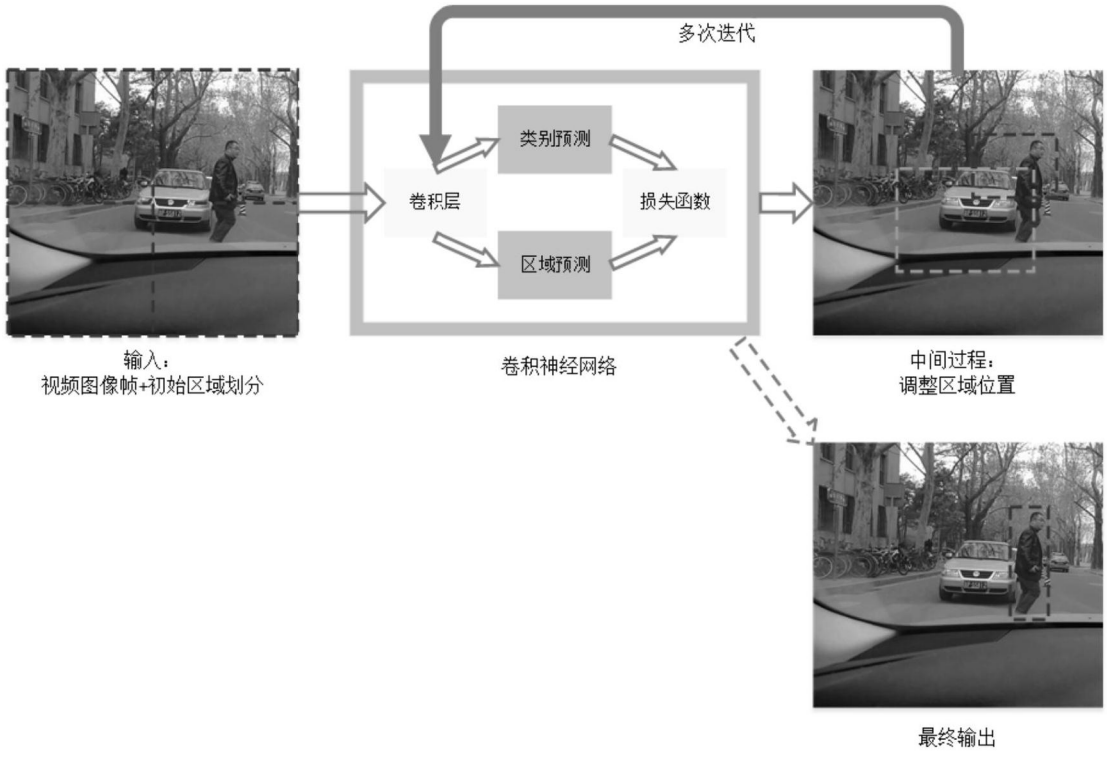

多次迭代
类别预测
卷积层
损失函数
区域预测
输入：
视频图像帧+初始区域划分
卷积神经网络
中间过程：
调整区域位置
最终输出

图2

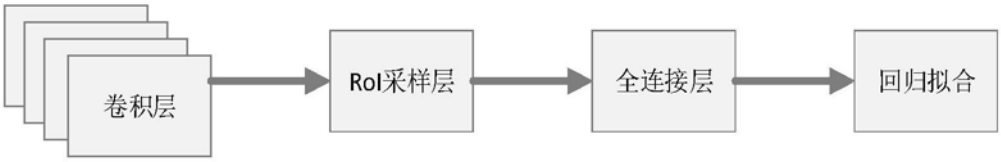

卷积层
RoI采样层
全连接层
回归拟合

图3


person 0.998
car 1.000
car 0.947
person 0.872
person 1.000
person 0.990

图4

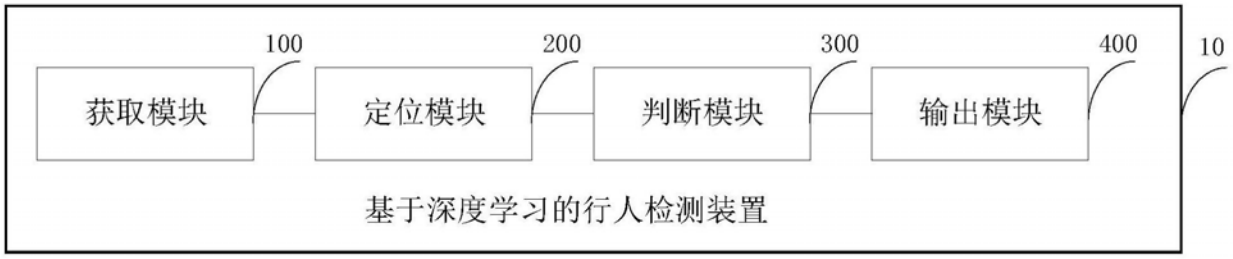

100
200
300
400
10
获取模块
定位模块
判断模块
输出模块
基于深度学习的行人检测装置

图5