(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 108734289 A
(43)申请公布日 2018.11.02

(21)申请号 201810348276.7

(22)申请日 2018.04.18

(30)优先权数据
102017206555.9 2017.04.19 DE
102018200724.1 2018.01.17 DE

(71)申请人 罗伯特·博世有限公司
地址 德国斯图加特

(72)发明人 C.K.穆马迪 J.H.梅岑 V.菲舍尔

(74)专利代理机构 中国专利代理(香港)有限公司 72001
代理人 臧永杰 申屠伟进

(51)Int.Cl.
G06N 3/08(2006.01)
G06N 99/00(2010.01)

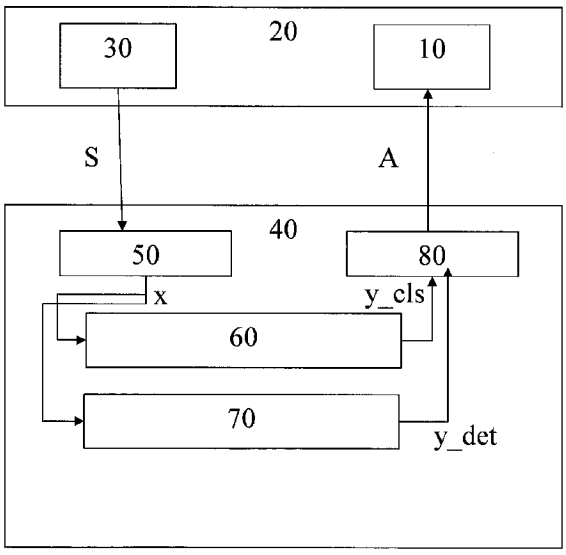
权利要求书3页 说明书9页 附图5页

(54)发明名称

用于改善相对于“对抗样本”的稳健性的方法和设备

(57)摘要

本发明涉及用于改善相对于“对抗样本”的稳健性的方法和设备。用于产生通用数据信号干扰(ξ)以生成被操纵的数据信号(x^{adv})来迷惑第一机器学习系统(60)的方法,所述第一机器学习系统被设立用于确定所接收的一维或多维数据信号(x)的语义分段(y_cls),所述方法具有以下步骤:a)确定训练数据组(D^{train}),所述训练数据组包括数据信号($x^{(k)}$)和所属的所期望的语义分段($y^{target,k}$)的对,b)根据所述训练数据组(D^{train})的数据信号($x^{(k)}$)、所述所属的所期望的语义分段($y^{target,k}$)以及加载有数据信号干扰(ξ)的数据信号($x^{(k)}$)的所估计的语义分段(f_θ)来生成所述数据信号干扰(ξ)。



CN 108734289 A

1.一种用于产生通用数据信号干扰(Ξ)以生成被操纵的数据信号(x^{adv})来迷惑第一机器学习系统(60)的方法,所述第一机器学习系统被设立用于确定所接收的一维或多维数据信号(x)的语义分段(y_cls),所述方法具有以下步骤:

a)确定训练数据组(D^{train}),所述训练数据组包括数据信号($x^{(k)}$)和所属的所期望的语义分段($y^{target,k}$)的对,

b)根据所述训练数据组(D^{train})的数据信号($x^{(k)}$)、所述所属的所期望的语义分段($y^{target,k}$)以及加载有数据信号干扰(Ξ)的数据信号($x^{(k)}$)的所估计的语义分段(f_θ)来生成所述数据信号干扰(Ξ)。

2.根据权利要求1所述的方法,其中,进行所述数据信号干扰(Ξ)的生成,使得所述数据信号干扰(Ξ)是周期性的。

3.根据以上权利要求中任一项所述的方法,其中,根据所述所属的数据信号($x^{(k)}$)的所估计的语义分段($y^{pred,k}$)来选择所述所期望的语义分段($y^{target,k}$)。

4.根据权利要求3所述的方法,其中,根据所估计的语义值($y^{pred}{}_{ij}$)是否采用可预给定的值(o)来选择所期望的语义值($y^{target,k}{}_{ij}$)。

5.根据权利要求4所述的方法,其中,通过替代值替代所述所估计的语义值($y^{pred}{}_{ij}$)采用所述可预给定的值(o)的位置(i,j)处的所期望的语义值($y^{target,k}{}_{ij}$)。

6.根据权利要求5所述的方法,其中,所述替代值是替代位置(I',j')的所估计的语义值($y^{pred}{}_{i',j'}$),所述替代位置在所述位置(i,j)附近。

7.根据权利要求3至6中任一项所述的方法,其中,当所述所估计的语义值($y^{pred}{}_{ij}$)不采用所述可预给定的值(o)时,与所述所估计的语义值($y^{pred}{}_{ij}$)相同地选择所述所期望的语义值($y^{target,k}{}_{ij}$)。

8.根据以上权利要求中任一项所述的方法,其中,根据所述第一机器学习系统(60)的经加权的成本函数(J_{ss}^{α})的函数值来进行所述数据信号干扰(Ξ)的生成,所述成本函数(J_{ss}^{α})在加载有所述数据信号干扰(Ξ)的数据信号($x^{(k)}$)的情况下并且对于所述所期望的语义分段($y^{target,k}$)采用所述函数值。

9.根据权利要求8所述的方法,其中,从所允许的数据信号干扰(Ξ)的集合中选择所述数据信号干扰(Ξ),使得所述数据信号干扰(Ξ)在加载有所述数据信号干扰(Ξ)的数据信号($x^{(k)}$)的情况下对于所述所期望的语义值($y^{target,k}$)使所述成本函数(J_{ss}^{α})在数值上最小化。

10.根据权利要求8或9和权利要求4所述的方法,其中,所述成本函数(J_{ss}^{α})包括两个部分,所述两个部分中的一个部分表征在所述成本函数(J_{ss}^{α})上的其所估计的语义值($y^{pred}{}_{ij}$)采用所述可预给定的值(o)的位置(i,j)的部分,并且另一部分表征在所述成本函数(J_{ss}^{α})上的其所估计的语义值($y^{pred}{}_{ij}$)不采用所述可预给定的值(o)的位置(i,j)的部分,其中,可预给定地相互对所述两个部分加权。

11.一种用于生成被操纵的数据信号(x^{adv})来迷惑第一机器学习系统(60)的方法,所述第一机器学习系统被设立用于确定所接收的一维或多维数据信号(x)的语义分段(y_cls),其中,利用根据权利要求1至10中任一项所述的方法产生通用数据信号干扰(Ξ),并且根据所述数据信号(x)和所生成的数据信号干扰(Ξ)来生成所述被操纵的数据信号(x^{adv})。

12.一种用于评价执行器控制系统(40)的稳健性的方法,所述执行器控制系统用于操控执行器(10),所述执行器控制系统包括第一机器学习系统(60),所述第一机器学习系统被设立用于确定所接收的一维或多维数据信号(x)的语义分段(y_cls),所述数据信号表征包括所述执行器(10)和所述执行器(10)的环境(20)的执行器系统的状态,

其中,所述执行器控制系统(40)被设立用于,根据所确定的语义分段(y_cls)确定操控信号(A)并且根据所确定的操控信号(A)操控所述执行器(10),

其中,根据所述数据信号(x)和利用根据权利要求1至10中任一项所述的方法确定的数据信号干扰(☒)产生被操纵的数据信号(x^{adv}),并且根据所确定的第二语义分段(y_cls^{adv})确定第二操控信号(A^{adv}),所述第二语义分段借助于所述第一机器学习系统(60)当通过被操纵的数据信号(x^{adv})更换所述数据信号(x)时来确定,

其中,根据所述操控信号(A)和第二操控信号(A^{adv})判定,所述执行器控制系统(40)是否是稳健的。

13.一种用于运行执行器控制系统(40)的方法,所述执行器控制系统用于操控执行器(10),其中,借助于根据权利要求12所述的方法评价:所述执行器控制系统(40)是否是稳健的,并且所述执行器控制系统(40)根据所述评价操控所述执行器(10),其中,如果所述评价得出,所述执行器控制系统(40)是不稳健的,则可以使所述执行器控制系统(40)尤其转变到防卫模式中。

14.一种用于训练执行器控制系统(40)的方法,所述执行器控制系统尤其包括第一机器学习系统(60),所述第一机器学习系统被设立用于确定所接收的一维或多维数据信号(x)的语义分段(y_cls),其中,所述执行器控制系统(40)此外被设立用于确定第二机器学习系统(70)的输出值(y_det),所述方法具有以下步骤:

a)从用于所述第二机器学习系统(70)的训练集合中选择数据信号(x)的子集,所述训练集合包括数据信号(x),

b)判定,是否应操纵所述子集的数据信号(x),

c)只要判定出应操纵所述数据信号(x),就根据各自的数据信号(x)以及根据利用根据权利要求1至10中任一项所述的方法确定的数据信号干扰(☒)来确定被操纵的数据信号(x^{adv})并且通过所述各自的所确定的被操纵的数据信号(x^{adv})来替代所述数据信号(x),

d)如果不操纵所述数据信号(x),则将所期望的输出值设置为可预给定的第一数值的值,或者如果操纵所述数据信号(x),则将所期望的输出值设置为第二、不同于第一数值的可预给定的第二数值,以及

e)利用数据信号(x,x^{adv})的(必要时被操纵的)子集和所属的所期望的输出值训练所述第二机器学习系统(70)的参数。

15.一种计算机程序,所述计算机程序被设立用于实施根据权利要求1至14中任一项所述的方法。

16.一种机器可读存储介质,在其上存储有根据权利要求15所述的计算机程序。

17.一种计算机,所述计算机被设立用于实施根据权利要求1至14中任一项所述的方法。

18.一种执行器控制系统(40),所述执行器控制系统包括第一机器学习系统(60)和第二机器学习系统(70),所述第一机器学习系统被设立用于确定所接收的一维或多维数据信

号(x)的语义分段(y_cls),所述第二机器学习系统利用根据权利要求14所述的训练方法得以训练。

用于改善相对于“对抗样本”的稳健性的方法和设备

技术领域

[0001] 本发明涉及用于产生数据信号干扰的方法并且基于此涉及用于产生被操纵的数据信号的方法、用于评价执行器控制系统的稳健性(Robustheit)的方法、用于运行执行器控制系统的方法、用于训练执行器控制系统的方法、利用所述方法所训练的执行器控制系统、计算机程序、机器可读存储介质和计算机，其中所述计算机程序包括指令，所述指令被设立用于当在计算机上实施所述计算机程序时实施所述方法之一，在所述机器可读存储介质上存储有计算机程序，所述计算机被设立用于实施所述方法之一。

背景技术

[0002] 由DE 10 2005 050 577 A1已知一种用于控制设备的神经元网络。本发明检验用于控制设备的神经元网络1。神经元网络具有在第一层中的多个第一神经元N1、N2、...、Nn和在跟随第一层的第二层中的第二神经元M。从预给定的多个测试信号组合中选择每个测试信号组合。每个测试信号组合给每个第一神经元N1、N2、...、Nn分配测试输入信号向量ut1、ut2、...、utk，所述测试输入信号向量或者是零信号或者使所属的第一神经元N1、N2、...、Nn饱和(sättigt)，使得第一神经元N1、N2、...、Nn输出下饱和值φ_{min}，或者使所属的第一神经元(Neuron)N1、N2、...、Nn饱和，使得第一神经元N1、N2、...、Nn输出上饱和值。将测试信号组合施加到第一神经元N1、N2、...、Nn上并且检测第二神经元M的输出信号p。当输出信号p大于预给定的阈值时，存储部分检验信号。在施加所述测试信号组合中的每一个之后，并且当不存储部分检验信号时，输出正的总检验信号，其中对于所述部分检验信号(für die)存储预给定的多个测试信号组合。

[0003] 发明优点

与此相对，具有独立权利要求1的特征的方法具有以下优点：所述方法使得能够使借助机器学习方法确定的语义分段(semantische Segmentierungen)相对于误导示例(Irreführungsbeispielen)(英语“Adversarial Examples”：对抗样本)特别稳健。对抗样本是被轻微操纵的输入数据(其在图像数据情况下与未被操纵的输入数据类似，使得它们对于人类专家而言实际上不可区分，所述被轻微操纵的输入数据可能导致所确定的语义分段的显著改变。例如可能可设想的是，恶意攻击者使用这样的误导示例，以便将自主机器人引入歧途，其方式是，该攻击者例如抑制将实际上存在的流标记为“流(Fluss)”的语义分段，这可能会导致危害自主机器人，所述机器人基于所述语义分段执行其路线规划(Routenplanung)。利用根据本发明的方法可以减少这样的攻击的有效性。

[0004] 有利的改进方案是独立权利要求的主题。

发明内容

[0005] 在第一方面，本发明涉及用于产生通用数据信号干扰以生成被操纵的数据信号来迷惑(täuschen)第一机器学习系统的方法，所述第一机器学习系统被设立用于确定所接收的一维或多维数据信号的语义分段，所述数据信号表征代理系统(Agentensystem)、尤其包

括执行器(10)和所述执行器(10)的环境(20)的执行器系统的状态。

[0006] 数据信号的语义分段在此情况下如常见的那样意味着，给所述数据信号的范围(Bereichen)分配来自可能的语义值的预给定的选择中的语义值。

[0007] 所述方法包括以下步骤：

a)确定训练数据组，所述训练数据组包括数据信号和所属的所期望的语义分段的对，

b)根据所述训练数据组的数据信号、所属的所期望的语义分段以及加载有数据信号干扰的数据信号的所估计的语义分段来生成数据信号干扰。

[0008] 如此确定的数据信号干扰在以下意义上是通用的，即通过所述数据信号干扰(也即加载数据信号干扰)可以将不包含在训练数据组中的多个数据信号变换成误导示例。如果利用第一机器学习系统确定所述误导示例的语义分段，则根据所期望的语义分段得出相对于各自的相应的未经加载的数据信号的相应语义分段的明显差别。

[0009] 所估计的语义分段在此情况下可以通过第一机器学习系统来确定或者通过第一机器学习系统的模拟(Nachbildung)来确定。

[0010] 语义分段在此情况下优选地“逐像素地”进行，也即语义分段具有与一维或多维数据信号相同的维数并且优选地给一维或多维数据信号中的每个数据点分派语义值。

[0011] 利用这样的数据信号干扰尤其可以生成被操纵的数据信号，其方式是，给数据信号加载数据信号干扰(例如通过相加)。通过生成如此被操纵的数据信号，尤其可能的是，相对于可能的攻击者改善并且可靠地评价第一机器学习系统的用于语义分段的稳健性。尤其由于所接收的数据信号和训练数据组的数据信号的独立性而可能的是，借助所生成的数据信号干扰探测误导示例的宽广的类别。

[0012] 在一种改进方案中可以规定，进行所述数据信号干扰的生成，使得所述数据信号干扰是周期性的。通过这种方式可以有效地抑制被操纵的(manipulierten)数据信号与训练数据组的过度匹配。否则基于数据信号干扰的维度可能预期这样的过度匹配，其中所述数据信号干扰具有与数据信号相同的维度。

[0013] 可替代地或附加地可以规定，根据所属的数据信号的所估计的语义分段来选择所期望的语义分段。

[0014] 通过这种方式可以以特别少的耗费生成所期望的语义分段，因为不需要训练数据组的数据信号的实际上“真正的”语义分段。

[0015] 在该方面的一种改进方案中可以规定，根据所估计的语义值是否采用可预给定的值来选择所期望的语义值。通过这种方式可以特别好地模拟攻击，所述攻击旨在抑制以下数据点的信息：所述数据点的语义值具有可预给定的值。

[0016] 在此情况下可以规定，通过替代值来替代所估计的语义值采用可预给定的值的位置处的所期望的语义值。该替代值在此情况下是不同于所估计的语义值的值。

[0017] 有利地，所述替代值是在所述位置附近的替代位置处的那个所估计的语义值。该替代位置在此情况下可以是例如最近的位置，该最近的位置的所估计的语义值不同于可预给定的值。

[0018] 此外可以规定，当所估计的语义值不采用所述可预给定的值时，与所估计的语义值相同地选择所期望的语义值。

[0019] 利用先前提到的两个措施中的每一个可能的是，模拟攻击者，所述攻击者使被抑

制的数据点特别协调地过渡到背景的语义分段中并且因此执行特别有针对性的攻击。

[0020] 这可以通过以下方式来发生，即根据第一机器学习系统的成本函数的函数值来进行所述数据信号干扰的生成，其中所述成本函数在加载有所述数据信号干扰的数据信号的情况下并且对于所期望的语义分段采用所述函数值。也即，成本函数通常具有第一机器学习系统的输出值(所确定的语义分段)作为自变量，所述第一机器学习系统根据其输入信号(数据信号)确定了所述输出值。此外，成本函数如常见的那样具有所期望的语义分段(即所期望的语义值)作为另外的自变量。在这里现在规定，作为输入信号，通过加载有数据信号干扰的数据信号替代所述数据信号。

[0021] 具体地，可以从所允许的数据信号干扰的集合中选择所述数据信号干扰，使得所述数据信号干扰在加载有所述数据信号干扰的数据信号的情况下对于所期望的语义值使所述成本函数在数值上最小化。这样的在数值上最小化例如可以通过迭代的方法进行，其中，如果达到可预给定的中止标准(例如在可预给定数目的迭代步骤之后)，则结束所述迭代的方法。对于本领域技术人员可以理解的是，在此情况下，理论上可达到的最小值通常将是不准确地而是仅仅近似地达到的。

[0022] 在一种灵活的改进方案中可以规定，所述成本函数包括两个部分，所述两个部分中的一个部分表征在所述成本函数上的其所估计的语义值采用可预给定的值的位置的部分，并且另一部分表征在所述成本函数上的其所估计的语义值不采用可预给定的值的位置的部分，其中，可预给定地相互对所述两个部分进行加权。

[0023] 通过这种方式，通过特别简单的方式可能的是，灵活地不仅模拟这样的攻击，即所述攻击的目标是尽可能高效地抑制其语义分段具有可预给定的值的数据点的信息(这可以通过对第一部分高度地加权的方式来发生)，而且模拟这样的攻击，即所述攻击的目标是，使语义背景尽可能完好，以便因此特别难以识别(这可以通过对第二部分高度地加权的方式来发生)。

[0024] 在另一种改进方面中规定，将本发明应用于用于评价执行器控制系统的稳健性的方法中，所述执行器控制系统用于操控执行器，所述执行器控制系统包括第一机器学习系统，所述第一机器学习系统被设立用于确定所接收的一维或多维数据信号的语义分段，所述数据信号表征包括所述执行器和所述执行器的环境的执行器系统的状态，其中，所述执行器控制系统被设立用于，根据所确定的语义分段确定操控信号并且根据所确定的操控信号操控所述执行器，其中，根据所述数据信号和数据信号干扰产生被操纵的数据信号，(其中，利用先前提到的方法之一确定了数据信号干扰)，并且根据所确定的第二语义分段确定第二操控信号，所述第二语义分段借助于所述第一机器学习系统通过由被操纵的数据信号更换所述数据信号来确定(也即给第一机器学习系统输送被操纵的数据信号而不是所述数据信号)。在此情况下，于是根据所述操控信号并且也根据第二操控信号判定，所述执行器控制系统是否是稳健的。

[0025] 这样的方法尤其也可以在运行执行器控制系统期间运行，以便连续地或者以有规律的时间区间来评价：执行器控制系统相对于可能的攻击是否将会表现为稳健的。

[0026] 基于此可以设置一种用于运行执行器控制系统的方法，所述执行器控制系统用于操控执行器，其中，借助根据先前提到的方法评价：所述执行器控制系统是否是稳健的，并且所述执行器控制系统根据所述评价操控所述执行器，其中，如果所述评价得出，所述执行

器控制系统是不稳健的，则可以使所述执行器控制系统尤其转变到防卫模式中。

[0027] 在又另一方面中，本发明涉及一种用于训练执行器控制系统的方法，所述执行器控制系统尤其包括第一机器学习系统，所述第一机器学习系统被设立用于确定所接收的一维或多维数据信号的语义分段，所述数据信号表征代理系统、尤其包括执行器和执行器的环境的执行器系统的状态，其中，所述执行器控制系统此外被设立用于确定第二机器学习系统的输出值，其中，所述输出值表征语义分段是错误的(falsch)、尤其是数据信号是否是误导示例的概率，其中，根据所确定的语义分段可以进行执行器的操控，并且如果判定出存在错误，则可以根据错误防御反应进行执行器的操控。

[0028] 所述训练方法包括以下步骤：

a)从用于所述第二机器学习系统的训练集合中选择数据信号的子集，所述训练集合包括数据信号，

b)判定，是否应操纵所述子集的数据信号，

c)只要判定出应操纵所述数据信号，就根据相应的数据信号以及利用先前提到的方法之一确定的数据信号干扰来确定被操纵的数据信号并且通过所述各自的所确定的被操纵的数据信号来替代所述数据信号，

d)如果不操纵所述数据信号，则将所期望的输出值设置为可预给定的第一数值的值，或者如果操纵所述数据信号，则将所期望的输出值设置为第二、不同于第一数值的可预给定的第二数值，以及

e)利用观察值的(必要时被操纵的)子集和所属的所期望的输出值训练所述第二机器学习系统的参数。

[0029] 在又一个方面中，本发明涉及执行器控制系统，所述执行器控制系统包括第一机器学习系统和第二机器学习系统，所述第一机器学习系统被设立用于确定所接收的一维或多维数据信号的语义分段，所述第二机器学习系统利用所述训练方法得以训练。

附图说明

[0030] 下面参考附图更详细地阐述本发明的实施方式。在附图中：

图1示意性地示出在执行器与执行器控制系统之间的交互；

图2示意性地示出在训练系统与执行器控制系统之间的交互；

图3示出训练方法的一种实施方式；

图4示出用于确定数据信号干扰的方法的一种实施方式；

图5示出用于确定被操纵的数据信号的方法的一种实施方式；

图6示出用于评价执行器控制系统的稳健性的方法的一种实施方式；

图7示出用于运行执行器控制系统的方法的一种实施方式。

具体实施方式

[0031] 图1示出执行器10，所述执行器在其环境20中与执行器控制系统40处于交互中。执行器10和环境20以下共同地也被称作执行器系统。执行器系统的状态利用一个传感器30来检测，所述状态也可以通过多个传感器来给出。传感器30的输出信号S被传送给执行器控制系统40。执行器控制系统40从中确定操控信号A，其中执行器10接收所述操控信号。

[0032] 执行器10可以例如是(半)自主机器人，例如(半)自主机动车。传感器30可以例如是一个或多个视频传感器和/或一个或多个雷达传感器和/或一个或多个超声传感器和/或一个或多个位置传感器(例如GPS)。可替代地或附加地，传感器30也可以包括信息系统，所述信息系统确定关于执行器系统的状态的信息，诸如天气信息系统，所述天气信息系统确定环境20中的天气的当前的或将来的状态。

[0033] 在另一个实施例中，执行器10可以是制造机器人，并且传感器30于是例如是光学传感器，其检测制造机器人的制造产品的特性。

[0034] 在另一个实施例中，执行器10可以是准许系统(Freigabesystem)，其被设立用于准许或不准许设备的活动。传感器30可以例如是光学传感器(例如用于检测图像数据或视频数据)，所述光学传感器被设立用于检测脸部。执行器10根据操控信号A确定准许信号，所述准许信号可以被用于根据准许信号的值来准许所述设备。所述设备可以例如是物理的或逻辑的访问控制装置。根据操控信号A的值，访问控制装置于是可以设置，准予或不准予访问。

[0035] 执行器控制系统40在可选的接收单元50中接收传感器的输出信号S，所述接收单元将输出信号S转换成数据信号x(可替代地，也可以直接地作为数据信号x接管输出信号S)。数据信号x可以例如是输出信号S的片段或进一步处理。输出信号x被输送给第一机器学习系统60，例如神经元网络。

[0036] 在下面描述的优选实施例中，数据信号x是二维图像信号，所述图像信号的像素通过一个或三个数值来表征。但数据信号x也可以是其他的一维或多维数据信号。

[0037] 第一机器学习系统60从图像信号x中确定所属的语义分段y_cls。在此情况下，给图像信号x的每个范围分配语义值。在该实施例中，执行逐像素的分段，也即语义分段y_cls具有与数据信号x相同的维数。

[0038] 在此情况下，语义分段y_cls表征以下概率：即该像素被分类为属于(来自多个语义类别中的)一个语义类别。语义分段y_cls对于每个像素可以例如是向量值参量，所述向量值参量对于语义类别中的每一个借助值区间[0;1]中的所分配的数说明：可以将该像素分配给相应的语义类别的概率是多高。语义分段y_cls对于每个像素也可以例如是标量参量(skalare Größe)，例如以下语义类别的标识符(Bezeichner)：所述语义类别的上述概率采用最大的值。

[0039] 执行器控制系统40此外包括第二机器学习系统70，例如神经元网络。第二机器学习系统70同样接收图像信号x并且从中确定输出值y_det，所述输出值可以是例如在值范围[0;1]中的数并且可以表征以下概率：即观察值B已被操纵，使得语义分段y_cls不正确地表征图像数据x。

[0040] 在该实施例中，这通过以下方式来实现：即第二机器学习系统70被设立，使得输出值y_det表征以下概率：图像信号x是误导示例。

[0041] 语义分段y_cls和输出值y_det被传送给输出单元80，所述输出单元从中确定操控信号A。例如可能的是，输出单元首先检验，输出值y_det是否小于可预给定的阈值。如果这是这种情况，则根据语义分段y_cls确定操控信号A。这是正常情况。相反如果确定出输出值y_det不小于可预给定的阈值，则可以规定，操控信号A被构造，使得所述操控信号使执行器A转变到防卫模式(abgesicherten Modus)中。

[0042] 在一种实施方式中，执行器控制系统40包括计算机和机器可读存储介质(未示出)，在所述机器可读存储介质上存储有计算机程序，当由计算机实施所述计算机程序时，所述计算机程序促使所述计算机实施执行器控制系统40的所描述的功能性。第一机器学习系统60和第二机器学习系统70在这里尤其可以作为单独的或共同的计算机程序来实施。

[0043] 图2图解在训练系统90、第一机器学习系统60和第二机器学习系统70之间的交互。训练系统90暂时提供(vorhalten)训练数据组，所述训练数据组包括图像信号x。从训练数据组中选择一个或多个图像信号x并且将所述一个或多个图像信号提供给第一机器学习系统60。在此情况下可以涉及各个图像信号x，即即使在执行器控制系统40与执行器10和传感器30交互时也被输送给第一机器学习系统60的这样的图像信号。但也可以涉及堆叠(Stapel)(英语“batch(批量)”)，即多个这样的图像数据x。

[0044] 第一机器学习系统60从给其所输送的图像数据x中分别确定语义分段y_cls。同样地，第二机器学习系统60类似于图1地确定输出值y_det。语义分段y_cls和输出值y_det又被输送给训练系统90。训练系统90从中确定参数匹配信号P，所述参数匹配信号编码：应改变第二机器学习系统70的哪些参数，例如其值。该所期望的匹配例如通过为输出值y_det预给定所期望的值和反向传播(Backpropagation)来进行。为此目的，训练系统90将参数匹配信号P输送给匹配块95，所述匹配块相应地匹配第一机器学习系统60中的和第二机器学习系统70中的参数。在此情况下可能的是，仅仅匹配第一机器学习系统60的参数或仅仅匹配第二机器学习系统70的参数。

[0045] 在一种实施方式中，训练系统90包括计算机和机器可读存储介质(未示出)，在所述机器可读存储介质上存储有计算机程序，当所述计算机程序由计算机实施时，所述计算机程序促使所述计算机实施学习系统90的所描述的功能性。

[0046] 图3以流程图示出用于通过训练系统90训练第二机器学习系统70的参数的方法的一种实施方式。

[0047] 在此情况下，冻结(eingefroren)第一机器学习系统60的参数并且训练第二机器学习系统60的参数。

[0048] 首先(1110)，对于在第一训练阶段中输送给第一机器学习系统60的每个图像信号x判定，所述图像信号是否被操纵。这可以例如随机地以可预给定的概率、例如50%进行。

[0049] 接着(1120)，将应被操纵的图像值x的所期望的第二输出值y_det设置为值“1”，否则设置为值“0”。

[0050] 在跟着的步骤1130中，提供通用数据干扰信号 。可以或者读入所述通用数据干扰信号，或者可以生成所述通用数据干扰信号。用于生成通用数据干扰信号的方法在图4中予以图解。

[0051] 然后(1140)，将应被操纵的图像值x通过其被操纵的形式x^{adv}替代，其方式，将通用数据干扰信号加到x上。

[0052] 在现在跟着的步骤1150中，冻结第一机器学习系统60的所训练的参数并且训练第二机器学习系统70的参数。

[0053] 借助第二机器学习系统70的所描述的成本函数J_{det}和反向传播进行所述参数的训练。

[0054] 第二阶段1100由此结束。

[0055] 图4以流程图示出用于生成通用数据信号干扰Ξ的方法。

[0056] 在第一步骤2000中，提供由图像信号$x^{(k)}$组成的训练数据组D^{train}，其中，k=1...m，具有固定的数m。

[0057] 接着是步骤2010，在该步骤中，对于图像信号$x^{(k)}$中的每一个，借助第一机器学习系统60确定所属的语义分段y_cls(替代第一机器学习系统60，可以应用等效的系统)。所述语义分段y_cls下面也用符号$f_\theta(x^{(k)})$表示，其中，θ表示第一机器学习系统60的参数，并且f表示成像准则，机器学习系统60实施所述成像准则。因此对于每个点(i，j)估计语义值$y^{pred}_{ij}=f_\theta(x^{(k)})_{ij}$。所属的所估计的语义分段也用符号$y^{pred}$来表示。

[0058] 在步骤2020中，定义可预给定的类别o，也即o是可预给定的值，其相应于允许的语义值中的一个。现在，将前景集合(Vordergrund-Menge)$I_o^{(k)}$定义为相应的图像信号$x^{(k)}$的所有以下点(i，j)的集合：即所述点的所属的所估计的语义值$y^{pred}{}_{ij}$采用所述可预给定的值o。此外，确定背景集合$I_{bg}^{(k)}$，所述背景集合包括所有以下点：即所述点的所属的所估计的语义值$y^{pred}{}_{ij}$不采用所述可预给定的值o。由前景集合$I^{(k)}{}_o$和背景集合$I_{bg}^{(k)}$组成的并集(Vereinigung)对于k的每一个值是相同的并且也被称作总集合I。

[0059] 在步骤2030中，对于图像信号$x^{(k)}$中的每一个确定所期望的语义分段$y^{target,k}$，所述所期望的语义分段的所属的语义值$y^{target,k}{}_{ij}$对于包含在背景集合$I_{bg}^{(k)}$中的每一个点(i，j)与相应的所估计的语义值$y^{pred}{}_{ij}$相同地来选择。此外，对于包含在前景集合$I_o^{(k)}$中的每一个点(i，j)确定替代值。为此，首先确定图像信号$x^{(k)}$中的替代位置(Ersatzstelle)(i'，j')，更确切地说，借助

$$i',j'=\arg\min_{i,j\in I_{bg}}(i'-i)^2+(j'-j)^2$$ 来确定。

[0060] 然后，作为替代值，选择在该替代位置(i'，j')处的语义值。在所期望的语义分段$y^{target,k}$中对于点(i，j)选择所述语义值，即$y^{target,k}{}_{ij}=y^{pred}{}_{i'j'}$。

[0061] 现在(2040)，预给定数值h、w。h、w被选择，使得图像信号$x^{(k)}$中的每一个的高度是h的整数倍，并且图像信号$x^{(k)}$中的每一个的宽度是w的整数倍，从而图像信号$x^{(k)}$的维数分别是Rh×Sw，其中R、S为整数。以$\xi^{(0)}=0$初始化具有h×w个点的维数的原型数据信号干扰$\xi^{(0)}$。数n被设置为值n=0。

[0062] 然后(2050)，计算参量

$$\nabla^{(n)}=\frac{1}{mRS}\sum_{r=1}^{R}\sum_{s=1}^{S}\sum_{k=1}^{m}\nabla_\xi J_{ss}^{(o)}\left(f_\theta\left(x_{[r,s]}^{(k)}+\xi^{(n)}\right),y_{[r,s]}^{target,k}\right)。$$

[0063] 在此情况下，[r，s]表示根据

$$[r,s]=\{i,j:[rh\le i<(r+1)h]\wedge[[sw\le j<(s+1)w]]\}$$

的点的集合。

[0064] $x^{(k)}{}_{[r,s]}$表示图像信号$x^{(k)}$的相应的片段，$y^{target,k}{}_{[r,s]}$表示所属的所期望的语义分段$y^{target,k}$的相应的片段。

[0065] 参量$\nabla_\xi J_{ss}^{(o)}$表示根据

$$J_{ss}^{\omega}(f_\theta(x), y^{target}) = \frac{1}{|I|}\left\{\omega \sum_{(i,j)\in I_o} J_{cls}\left(f_\theta(x)_{ij}, y_{ij}^{target}\right) + (1-\omega) \sum_{(i,j)\in I_{bg}} J_{cls}\left(f_\theta(x)_{ij}, y_{ij}^{target}\right)\right\}$$

的第一机器学习系统(60)的经加权的成本函数。在此情况下，J_{cls}是第一机器学习系统60的成本函数，例如交叉熵。

[0066] 参量ω是可预给定的参量，其可以在值范围[0;1]中被预给定，例如被预给定到值0.5上。

[0067] 在步骤2050中，因此确定(经加权的)成本函数J_{ss}^{ω}的在训练数据组上求平均的梯度。

[0068] 接着(2060)，根据公式

$$\xi^{(n+1)} = Clip_\varepsilon\left\{\xi^{(n)} - \alpha\, sgn\left(\nabla^{\alpha}(\xi^{(n)})\right)\right\}$$

计算原型数据信号干扰$\xi^{(n+1)}$的下一迭代。在此情况下，函数$Clip_\varepsilon(x)$将变量x的值规范化(normiert)到围绕原点的ε球上。该规范(Norm)在此情况下可以是L^2规范，或者也可以是L^∞规范。将α确定到可预给定的值上，例如α=0.25。变量n的值被递增1(um 1 inkrementiert)。

[0069] 现在(2070)检验，是否满足收敛标准。例如可以检验，n是否超出了可预给定的极限值nmax，例如nmax=10。如果这不是这种情况，则分支回步骤2050。否则在步骤2080中从原型数据信号干扰$\xi^{(n)}$(其中，对于i<h，j<w，$\Xi_{ij} = \xi^{(n)}_{ij}$)和周期性边缘条件$\Xi_{i+n_i,j} = \Xi_{i,j+w} = \Xi_{ij}$中迭代地确定数据信号干扰Ξ。所述方法由此结束。

[0070] 图5以流程图示出用于确定被操纵的数据信号x^{adv}的方法的一个实施例。在第一步骤3000中，接收数据信号x。在下一步骤3010中，可选地借助在图4中图解的方法确定数据信号干扰Ξ。可替代地，所述数据信号干扰已经在更早的时刻借助在图4中图解的方法被确定，被存储在存储介质中，并且在步骤3010中现在被调用。接着(3020)，作为数据信号x和数据信号干扰Ξ的相加来确定被操纵的数据信号x^{adv}。所述方法由此结束。所述方法可以作为计算机程序来实施，可以存储在执行器控制系统40和/或训练系统90的存储介质上或者由该执行器控制系统和/或训练系统实施。计算机程序也可以存储在一般性计算机上并且可由该计算机实施。

[0071] 图6以流程图示出用于评价执行器控制系统40的稳健性的方法的一种实施例。在第一步骤4000中，接收数据信号x。接着(4010)，所述执行器控制系统40从中确定所属的操控信号A，以便操控执行器10。然后(4020)，利用在图5中图解的方法从数据信号x中确定被操纵的数据信号x^{adv}。同样，将该被操纵的数据信号x^{adv}输送给第一机器学习系统60并且确定第二语义分段y_cls^{adv}。类似于操控信号A的确定，从所确定的第二语义分段y_cls^{adv}中来确定第二操控信号A^{adv}。现在(4030)，确定操控信号A和第二操控信号A^{adv}之间的差并且检验(4040)，该差的量值(Betrag)是否超出可预给定的极限值。如果这是这种情况(4050)，则判定，执行器控制系统40不稳健，否则(4060)判定，执行器控制系统40是稳健的。所述方法由

此结束。所述方法可以作为计算机程序来实施，可以存储在执行器控制系统40的存储介质上或者由所述执行器控制系统实施。计算机程序也可以存储在一般性计算机上并且可由该计算机实施，所述一般性计算机可耦合到执行器控制系统40上。

[0072] 图7以流程图示出用于运行执行器控制系统的方法的一个实施例。在第一步骤5000中，利用在图6中图解的方法确定：执行器控制系统40是否是稳健的。然后在步骤5010中检验，是否判定出执行器控制系统40是稳健的。如果这是这种情况(5020)，则正常地运行执行器控制系统40并且正常地操控执行器10。如果这不是这种情况(5030)，则检验，执行器控制系统40转变到防卫模式中，其方式是，例如可以设置，将操控信号A的绝对量值限制到小于临界极限的值上。所述方法由此结束。所述方法可以作为计算机程序来实施，可以存储在执行器控制系统40的存储介质上或者由该执行器控制系统实施。

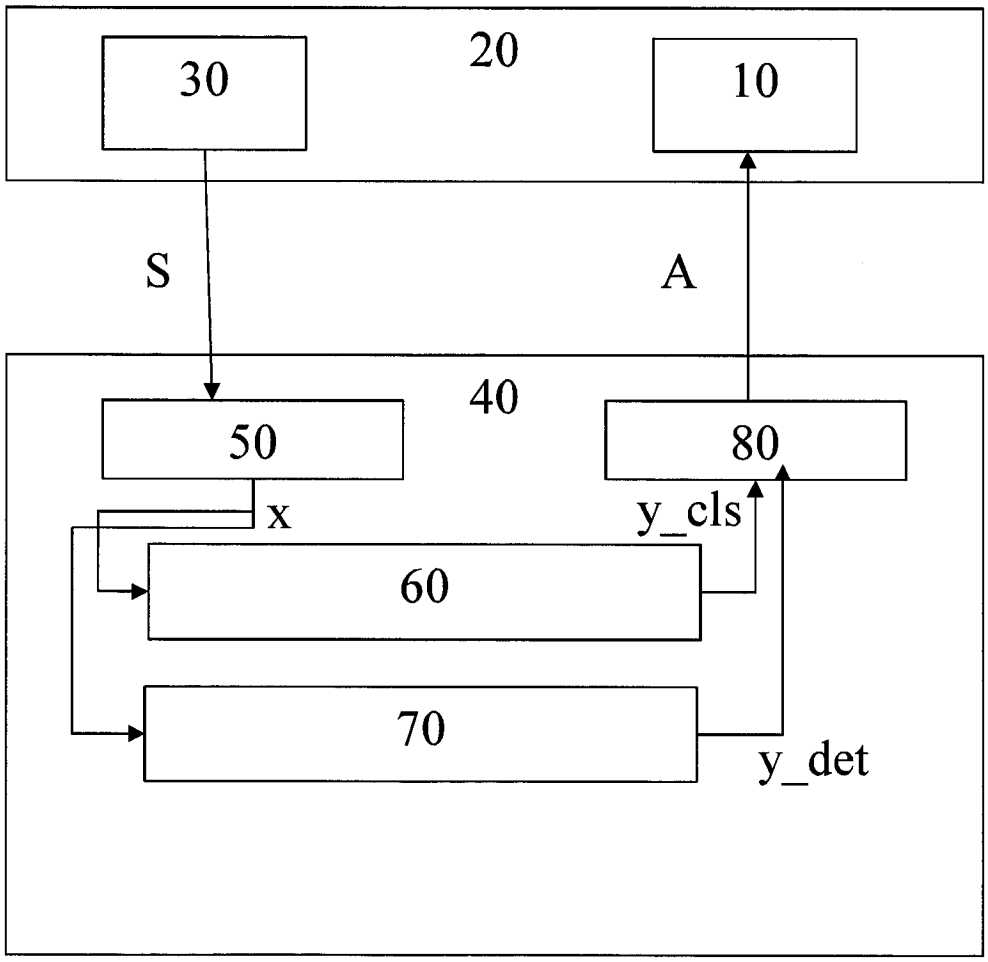


图 1

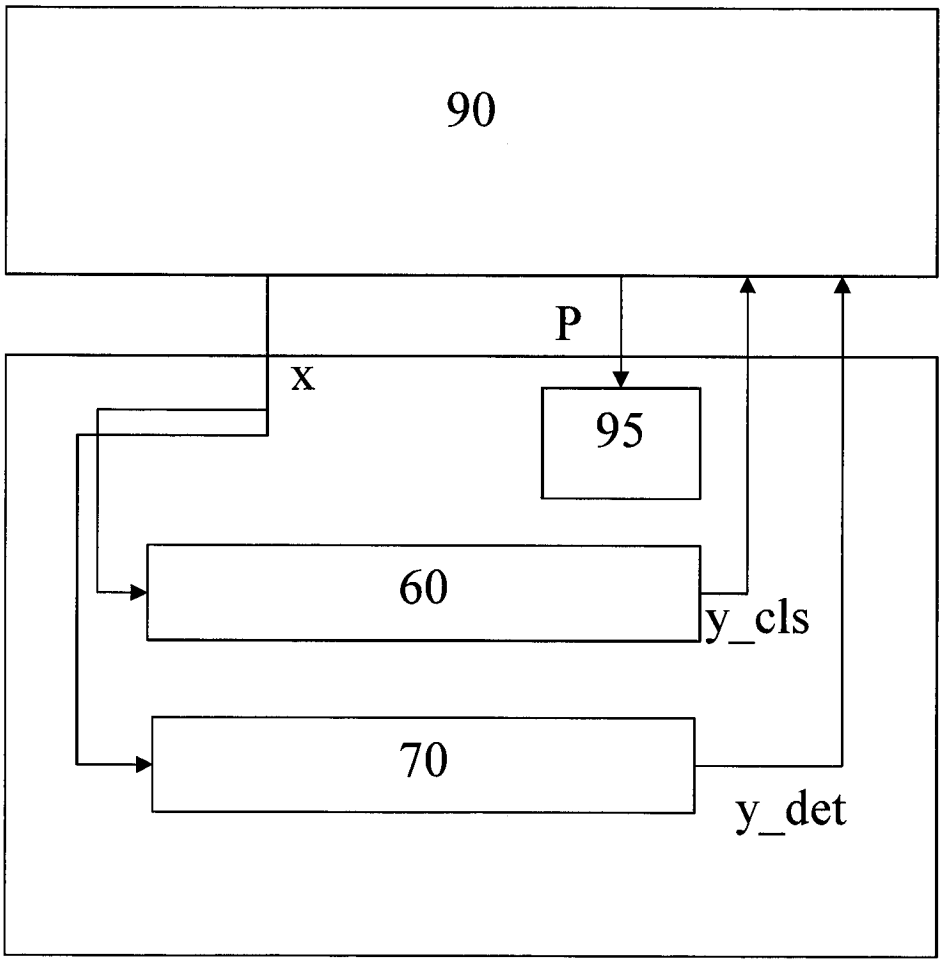


图 2

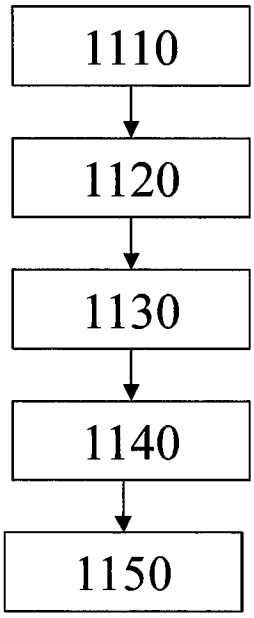


图 3

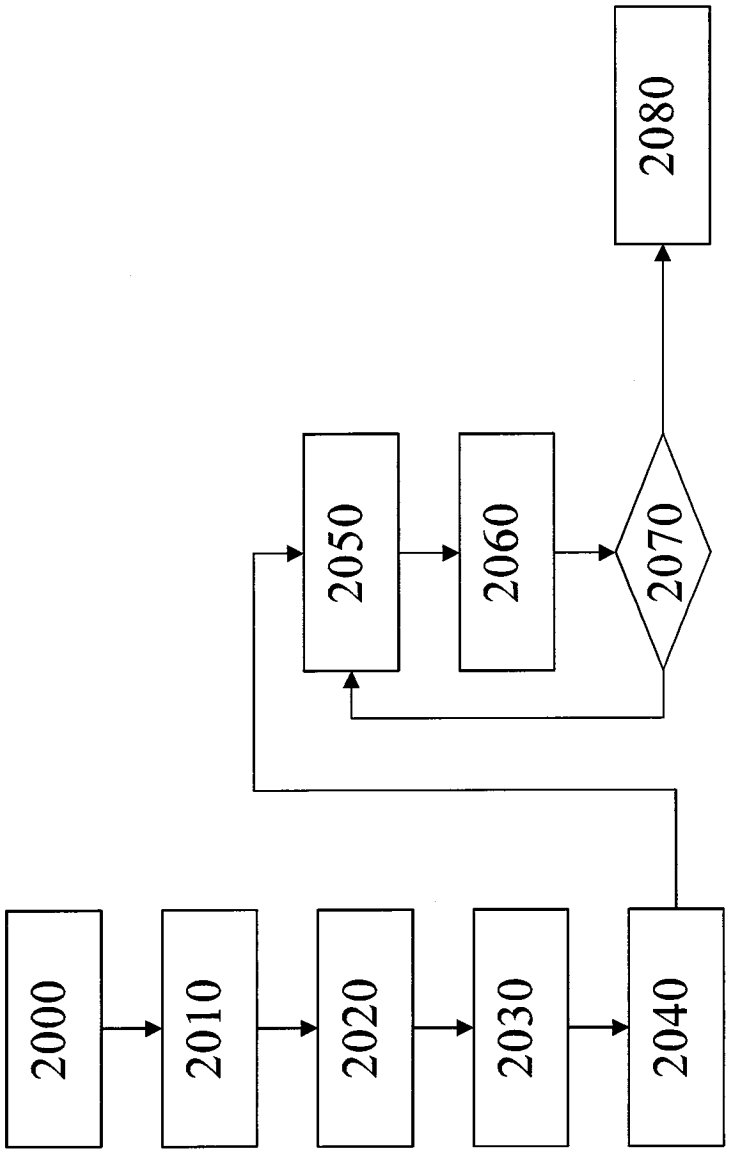


图 4

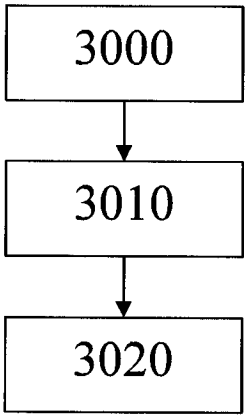


图 5

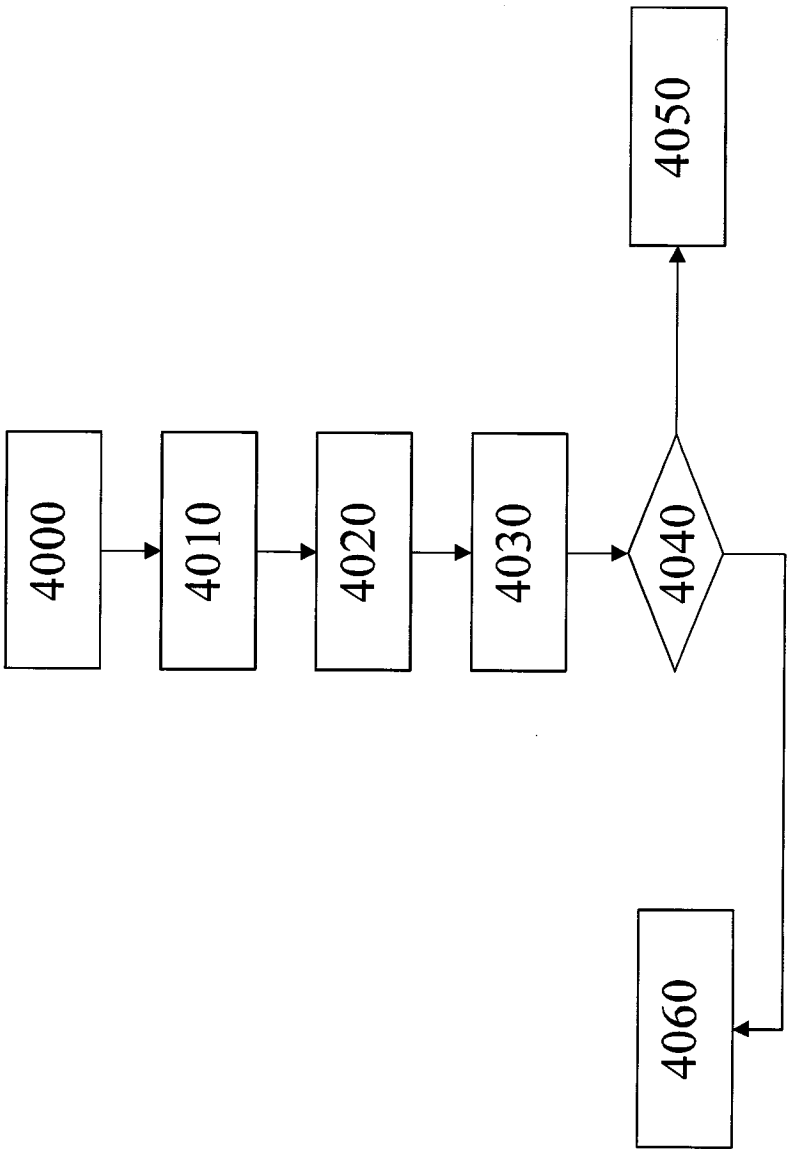


图 6

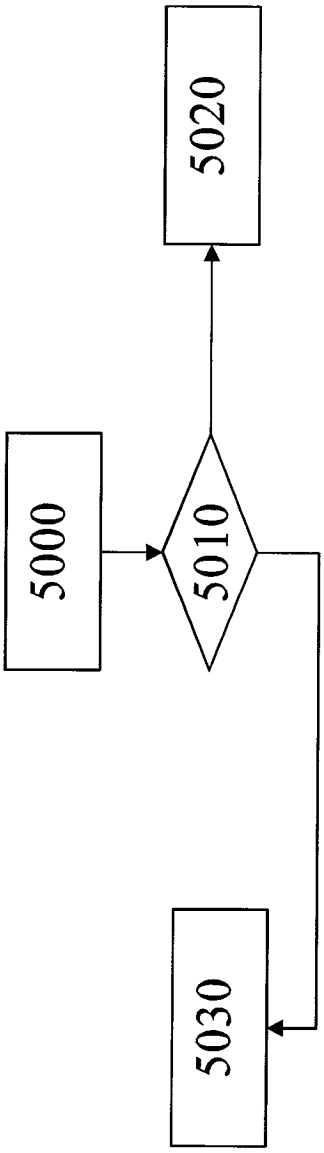


图 7