生物特征识别方法、服务器以及客户端

技术领域

本申请要求中国专利申请（申请号：202210108922.9 ）的优先权，并且本申请涉及身份识别方案，更具体地，涉及一种生物特征识别方法、生物特征识别服务器、客户端、计算机存储介质以及计算机程序产品。

背景技术

近年来，生物特征（例如人脸、指纹等）由于其与个体身份的紧密关联性逐渐成为主流的身份识别方法，取款、支付、手机银行登录等重要的金融场景有不少的应用。

近期人民银行科技司发布《金融科技应用风险提示》，就媒体报道的“多款安卓手机人脸识别系统被破解”，对人脸技术在金融支付领域应用存在的潜在风险进行提示，其中攻破人脸识别系统所用到的方法就是基于生成对抗网络的对抗样本攻击技术。

因而，期望一种基于生物特征的身份识别方案，该身份识别方案能有效防止该对抗样本攻击技术。

发明内容

根据本发明的一方面，提供了一种生物特征识别方法，所述方法包括：步骤 A：接收生物特征数据；步骤 B：对所接收的生物特征数据中的低位数据进行随机变换；步骤 C：随机选择一种或多种滤波算法，对经随机变换的生物特征数据进行预处理；以及步骤 D：基于预处理后的生物特征数据进行身份识别。

作为上述方案的补充或替换，在上述方法中，所述生物特征数据包括人脸图像数据。

作为上述方案的补充或替换，在上述方法中，对所接收的生物特

征数据中的低位数据进行随机变换包括：对所述生物特征数据中的最低位进行随机变换。

作为上述方案的补充或替换，在上述方法中，对所述生物特征数据中的最低位进行随机变换包括下列所示任一种：将所述生物特征数据中最低位全部置为 0；将所述生物特征数据中最低位全部置为 1；将所述生物特征数据中最低位随机重排；以及将所述生物特征数据中最低位随机生成。

作为上述方案的补充或替换，在上述方法中，在所述步骤 D 中，基于预处理后的生物特征数据进行身份识别以便获得第一比对结果，所述方法还包括：重复执行上述步骤 B 至步骤 D，以便获得第二比对结果；以及基于所接收的生物特征数据直接进行身份识别，以便获得第三比对结果。

作为上述方案的补充或替换，所述方法还包括：将所述第一比对结果、所述第二比对结果以及所述第三比对结果进行一致性比对；以及在所述第一比对结果、所述第二比对结果以及所述第三比对结果均一致时反馈所述第一比对结果。

作为上述方案的补充或替换，上述方法还包括：在所述第一比对结果、所述第二比对结果以及所述第三比对结果存在不一致时，反馈识别过程异常；以及将对应的数据采集终端加入黑名单。

作为上述方案的补充或替换，在上述方法中，所述生物特征数据还包括时间戳，并且在所述步骤 A 和所述步骤 B 之间，所述方法还包括：对所述生物特征数据进行解密，以获得所述人脸图像数据以及所述时间戳；以及在系统时间与所述时间戳之间的差值不符合要求时，反馈异常。

作为上述方案的补充或替换，在所述步骤 A 和所述步骤 B 之间，所述方法还包括：在所述系统时间与所述时间戳之间的差值符合要求时，检测所述人脸图像数据中的人脸位置。

作为上述方案的补充或替换，在所述步骤 A 和所述步骤 B 之间，

所述方法还包括：在所述人脸位置加入对抗样本，该对抗样本用于抵消在数据采集终端处所加入的负向对抗样本的作用，其中所述负向对抗样本是所述对抗样本的负值。

作为上述方案的补充或替换，在上述方法中，所述对抗样本依据人脸识别结果对生成器进行反向传导训练而预先生成。

根据本发明的另一个方面，提供了一种生物特征识别服务器，所述服务器包括：接收装置，用于接收生物特征数据；低位随机滤波器，用于对所接收的生物特征数据中的低位数据进行随机变换；预处理器，用于随机选择一种或多种滤波算法，对经随机变换的生物特征数据进行预处理；以及身份识别装置，用于基于预处理后的生物特征数据进行身份识别。

作为上述方案的补充或替换，在上述服务器中，所述生物特征数据包括人脸图像数据。

作为上述方案的补充或替换，在上述服务器中，所述低位随机滤波器配置成对所述生物特征数据中的最低位进行随机变换。

作为上述方案的补充或替换，在上述服务器中，所述低位随机滤波器配置成执行下列所示任一种：将所述生物特征数据中最低位全部置为 0；将所述生物特征数据中最低位全部置为 1；将所述生物特征数据中最低位随机重排；以及将所述生物特征数据中最低位随机生成。

作为上述方案的补充或替换，上述服务器还可包括：识别决策装置，用于将多次识别结果进行一致性比对；以及在识别结果均一致时反馈该结果。

作为上述方案的补充或替换，在上述服务器中，所述识别决策装置还配置成：在所述识别结果存在不一致时，反馈识别过程异常；以及将对应的数据采集终端加入黑名单。

作为上述方案的补充或替换，在上述服务器中，所述生物特征数据还包括时间戳。

作为上述方案的补充或替换，上述服务器还可包括：解密装置，用于对所述生物特征数据进行解密，以获得所述人脸图像数据以及所述时间戳；以及验证装置，用于在系统时间与所述时间戳之间的差值不符合要求时，反馈异常。

作为上述方案的补充或替换，在上述服务器中，所述验证装置还配置成：在所述系统时间与所述时间戳之间的差值符合要求时，检测所述人脸图像数据中的人脸位置。

作为上述方案的补充或替换，在上述服务器中，所述验证装置还配置成：在所述人脸位置加入对抗样本，该对抗样本用于抵消在数据采集终端处所加入的负向对抗样本的作用，其中所述负向对抗样本是所述对抗样本的负值。

作为上述方案的补充或替换，在上述服务器中，所述对抗样本依据人脸识别结果对生成器进行反向传导训练而预先生成。

根据本发明的又一个方面，提供了一种与上述服务器配合使用的客户端，所述客户端包括：人脸采集装置，用于采集人脸图像数据；以及嵌入装置，用于在人脸位置嵌入负向对抗样本，其中所述负向对抗样本是对抗样本的相反数，并且所述对抗样本是叠加在真实人脸图像上能使人脸识别模型产生错误识别结果的加性噪声。

作为上述方案的补充或替换，在上述客户端中，所述嵌入装置配置成：基于所述人脸图像数据检测所述人脸位置；在所述人脸位置嵌入所述负向对抗样本；以及将处理后的人脸图像以及时间戳以加密方式发送。

根据本发明的又一个方面，提供了一种利用如前所述的生物特征识别服务器来生成对抗样本的方法，所述方法包括：步骤 A：构建对抗生成网络，所述生成网络包括生成器以及鉴别器；步骤 B：利用所述生成器生成噪声信息，将噪声信息与真实人脸图像相加从而得到生成的人脸数据；步骤 C：利用人脸识别模型对生成的人脸数据进行识别；步骤 D：根据识别结果对所述生成器进行反向传导训练；步骤 E：

重复步骤B至步骤D直至所述人脸识别模型产生错误识别结果；以及

步骤F：将与所述错误识别结果对应的噪声信息确定为对抗样本。

根据本发明的又一个方面，提供了一种计算机存储介质，所述介质包括指令，所述指令在运行时执行如前所述的方法。

根据本发明的又一个方面，提供了一种计算机程序产品，包括计算机程序，该计算机程序被处理器执行时实现如前所述的方法。

与现有的生物特征识别方案相比，根据本发明的一个或多个实施例的生物特征识别方案通过对所接收的生物特征数据中的低位数据进行随机变换，并基于预处理后的生物特征数据进行身份识别。该方案对图像中的对抗噪声破坏性极大，会使得对抗攻击失灵。这样，即使攻击者获取生物特征识别模型的参数，对抗攻击模型的训练过程无法稳定收敛，因而无法形成有效的攻击模型。

附图说明

从结合附图的以下详细说明中，将会使本发明的上述和其他目的及优点更加完整清楚，其中，相同或相似的要素采用相同的标号表示。

图1示出了根据本发明的一个实施例的生物特征识别方法的流程示意图；

图2示出了对抗样本攻击的原理；

图3示出了根据本发明的一个实施例的生物特征识别服务器的结构示意图；

图4示出了根据本发明的一个实施例的人脸识别的流程图；

图5示出了根据本发明的一个实施例的包含数据采集终端以及人脸认证服务平台的人脸识别系统的示意图；以及

图6示出了根据本发明的另一个实施例的人脸识别的流程图。

具体实施方式

下面结合附图和实施例对本发明的实施方式作进一步详细描述。以下实施例用于说明本发明，但不能用来限制本发明的范围。

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实施例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，术语“第一”、“第二”、“第三”仅用于描述目的，而不能理解为指示或暗示相对重要性。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。

在本发明的上下文中，术语“客户端”也称为用户端，是指与服务器相对应，为客户提供本地服务的设备或装置。在一个或多个实施例中，该客户端具有一些基本功能，包括采集用户的生物特征信息的采集功能等。该客户端可以是用户的智能设备，包括但不限于，用户的手机、笔记本电脑以及头戴式设备。

术语“服务器”也称为远端服务器或云端服务器，是指与“客户端”相对应，为客户端提供远程服务的设备或装置。在一个或多个实施例中，服务器可为客户端提供防对抗样本攻击的生物特征识别（例如人脸识别）功能，即将从客户端接收的生物特征数据在经过一些预处理（例如去噪）步骤后进行生物特征识别。

图 1 示出了根据本发明的一个实施例的生物特征识别方法 1000 的流程示意图。如图 1 所示，方法 1000 包括如下步骤：

在步骤 S110 中，接收生物特征数据；

在步骤 S120 中，对所接收的生物特征数据中的低位数据进行随机变换；

在步骤 S130 中，随机选择一种或多种滤波算法，对经随机变换的生物特征数据进行预处理；以及

在步骤 S140 中，基于预处理后的生物特征数据进行身份识别。

在本发明的上下文中，术语“生物特征数据”是指任何可用于进行个人身份鉴定的人体所固有的数据以及与该数据相关的辅助信息。例如，生物特征数据包括但不限于，生理特征（例如，指纹、虹膜、面相、DNA 等）和行为特征（步态、击键习惯等）。

在一个实施例中，所述生物特征数据包括人脸图像数据。在一个实施例中，除了人脸图像数据，所述生物特征数据还包括与该人脸图像数据相关联的时间戳。该时间戳可以是记录该人脸图像数据的时间，也可以是上传该人脸图像数据的时间。

参考图 2，它示出了对抗样本攻击的原理。所谓“对抗样本”是指能在保证不影响人眼的识别的情况下误导分类器的产生错误预测的样本，例如对抗样本可以是由攻击者通过轻微地扰动正常样本所产生的异常样本。在一个实施例中，对抗样本可以使 CNN 模型输出高置信度的错误输出。图 2 中给出了人脸比对算法中的对抗样本。叠加对抗噪声后，人脸的比对分从 100 分急剧下降到了 60 以下。也就是说，人脸比对算法此时已经无法成功识别出这两个人了。

关于对抗样本攻击可以分为白盒攻击和黑盒攻击。白盒攻击是指攻击者知道识别模型的所有信息，包括训练数据、模型架构、超参数、层数、激活函数以及模型权重，通过计算模型梯度生成对抗样本进行攻击。黑盒攻击是指攻击者在不知道以上信息的情况下进行攻击。

从对抗样本的攻击原理可知，导致生物特征识别系统（例如人脸识别系统）发生误识别的根本原因是攻击者在生物特征数据（例如人脸图像数据）中增加了加性对抗噪声。

本申请的发明人发现，对抗噪声或对抗样本的大部分数据数值极低，例如仅改变了 8bit 图像数据的低位。因而，对生物特征数据中的低位数据进行随机变换可破坏对抗噪声的大部分数据。在这里，“低位数据”是指可用于进行个人身份鉴定的人体所固有的数据中的最低位数据、倒数两位数据、或倒数三位数据。

在一个实施例中，步骤 S120 包括：对所述生物特征数据中的最

低位进行随机变换。当然，本领域技术人员可以理解，在一个实施例中，可对生物特征数据中的倒数两位数据或倒数三位数据进行随机变换。

在一个实施例中，对所述生物特征数据中的最低位进行随机变换包括下列所示任一种：将所述生物特征数据中最低位全部置为 0；将所述生物特征数据中最低位全部置为 1；将所述生物特征数据中最低位随机重排；以及将所述生物特征数据中最低位随机生成。例如，对于 8bit 图像数据，最低位随机变换可包括如下四种算法：（1）将 8bit 最低位全部置 0；（2）将 8bit 最低位全部置 1；（3）将 8bit 最低位随机重排；以及（4）将 8bit 最低位随机生成。在一个或多个实施例中，最低位随机变换的算法并不固定，而是从多种算法（例如上述四种算法）中进行随机选择，从而对输入图像进行预处理。

在对生物特征数据中的最低位进行随机变换后，有利的是对经随机变换的生物特征数据再次进行滤波（预处理），从而使生物特征数据平滑（例如使人脸图像数据平滑）。例如，通过随机选择一种或多种滤波算法，对经随机变换的生物特征数据进行预处理。

以图像数据（例如人脸图像数据）为例，存在多种图像去噪/滤波算法，包括但不限，BM3D 降噪；DCT 降噪；PCA 降噪；K-SVD 降噪；非局部均值降噪；WNNM 降噪；基于主成分分析和双边滤波的图像降噪算法；小波变换；小波阈值降噪；Contourlet 变换；基于平移不变 Contourlet 变换的 SAR 图像降噪等。在一个或多个实施例中，从上述 11 种滤波算法中进行随机选择一种或几种，对输入图像进行预处理。正常情况下，去除噪声对人脸识别系统的识别结果不会产生影响，却对图像中的对抗噪声破坏性极大，会使得对抗攻击失灵。

在一个实施例中，在所述步骤 S140 中，基于预处理后的生物特征数据进行身份识别以便获得第一比对结果。尽管图 1 中未示出，上述方法 1000 还可包括：重复执行上述步骤 S120 至步骤 S140，以便获得第二比对结果(其中，重新执行步骤 S120 和 S130 所采用的滤波算法

与第一次执行时所采用的滤波算法不同或部分不同)；以及基于所接收的生物特征数据直接进行身份识别，以便获得第三比对结果。

在上述实施例中，所述方法 1000 还包括：将所述第一比对结果、所述第二比对结果以及所述第三比对结果进行一致性比对；以及在所述第一比对结果、所述第二比对结果以及所述第三比对结果均一致时反馈所述第一比对结果。而在所述第一比对结果、所述第二比对结果以及所述第三比对结果存在不一致时，反馈识别过程异常；以及将对应的数据采集终端加入黑名单。

可见，第三次识别是针对原始生物特征数据（即，未经低位随机变换）进行身份识别，第一次识别和第二次识别则均是对经低位随机变换（可采用不同的滤波算法）的生物特征数据进行身份识别。若这三次身份识别的结果一致，则可直接反馈该身份识别结果。而在多次身份识别的结果（例如三次身份识别的结果）不同时，则可反馈识别过程异常（很有可能是对抗样本攻击）并将采集到异常数据的终端加入黑名单，后续将拒绝为该终端提供身份识别服务。

这样，通过增加多图校验策略和黑名单拒绝策略，能够有效阻止攻击者进行黑盒模式的攻击训练。

目前，影响人脸识别最大的问题就是摄像头劫持攻击严重破坏人脸识别系统的安全性。其基本原理如下：首先，攻击者通过对支付应用 APP 进行“破壳”，实现对摄像头数据的劫持；接着，攻击者配合完成人脸识别的活体检测过程；最后，攻击者将摄像头上传的视频数据替换为事先伪造的被攻击者视频数据，将伪造数据上传至后台人脸识别系统，从而完成整个攻击过程。目前绝大多数的人脸识别系统，对此类摄像头劫持攻击都没有防范能力。

针对该问题，在一个实施例中，在步骤 S110 与步骤 S120 之间，方法 1000 还可包括：对所述生物特征数据进行解密，以获得所述人脸图像数据以及所述时间戳；以及在系统时间与所述时间戳之间的差值不符合要求时，反馈异常。在所述系统时间与所述时间戳之间的差

值符合要求时，检测所述人脸图像数据中的人脸位置，并在所述人脸位置加入对抗样本以用于人脸识别。

此处，在人脸位置加入对抗样本是为了抵消人脸图像采集过程中嵌入负向对抗样本的影响。也就是说，在人脸图像采集过程中，人脸采集终端（例如客户端）需检测人脸位置，在人脸检测区域嵌入负向对抗样本，将处理后的人脸图像、人脸位置和时间戳以加密形式传送至人脸识别认证服务器。这样，即使发生摄像头劫持，未在采集时嵌入负向对抗样本的图像也会在认证服务进行人脸识别前被加入对抗样本，从而无法在人脸识别过程得出正确结果。

在一个实施例中，对抗样本依据人脸识别结果对生成器进行反向传导训练而预先生成。例如，对抗样本可通过如下步骤来生成：（1）构建对抗生成网络，生成模型负责生成伪造人脸图像，鉴别器为人脸识别模型，固定人脸识别模型参数，仅对生成器进行参数优化；（2）生成器生成噪声信息，将噪声信息与真实人脸图像相加形成伪数据和伪标签；（3）人脸识别模型对生成的人脸数据进行识别；（4）依据人脸识别结果对生成器进行反向传导训练；（5）重复步骤（2）~（4）直至人脸识别模型产生错误识别结果。此时生成器生成的图像为对抗样本 A，其相反数-A 即为负向对抗样本。可以理解的是，对抗样本属于加性噪声，负向对抗样本是对抗样本的负值，用于抵消对抗样本的作用。

在一个实施例中，负向对抗样本只需进行一次生成，后续在图像中嵌入和去除的过程计算简便、效率高。而且，嵌入负向对抗样本的人脸数据对人眼不产生影响，即便数据在传输过程中被破解，也无法进行数据去除。

转到图 3，它示出了根据本发明的一个实施例的生物特征识别服务器 3000 的结构示意图。如图 3 所示，生物特征识别服务器 3000 包括：接收装置 310、低位随机滤波器 320、预处理器 330 以及身份识别装置 340。其中，接收装置 310 用于接收生物特征数据；低位随机

滤波器 320 用于对所接收的生物特征数据中的低位数据进行随机变换；预处理器 330 用于随机选择一种或多种滤波算法，对经随机变换的生物特征数据进行预处理；以及身份识别装置 340 用于基于预处理后的生物特征数据进行身份识别。

在本发明的上下文中，术语“生物特征数据”是指任何可用于进行个人身份鉴定的人体所固有的数据以及与该数据相关的辅助信息。例如，生物特征数据包括但不限于，生理特征（例如，指纹、虹膜、面相、DNA 等）和行为特征（步态、击键习惯等）。

在一个实施例中，所述生物特征数据包括人脸图像数据。在一个实施例中，除了人脸图像数据，所述生物特征数据还包括与该人脸图像数据相关联的时间戳。该时间戳可以是记录该人脸图像数据的时间，也可以是上传该人脸图像数据的时间。

所谓“对抗样本”是指能在保证不影响人眼的识别的情况下误导分类器的产生错误预测的样本，例如对抗样本可以是由攻击者通过轻微地扰动正常样本所产生的异常样本。在一个实施例中，对抗样本可以使 CNN 模型输出高置信度的错误输出。如图 2 所示，它给出了人脸比对算法中的对抗样本。叠加对抗噪声后，人脸的比对分从 100 分急剧下降到了 60 以下。也就是说，人脸比对算法此时已经无法成功识别出这两个人了。关于对抗样本攻击可以分为白盒攻击和黑盒攻击。白盒攻击是指攻击者知道识别模型的所有信息，包括训练数据、模型架构、超参数、层数、激活函数以及模型权重，通过计算模型梯度生成对抗样本进行攻击。黑盒攻击是指攻击者在不知道以上信息的情况下进行攻击。

从对抗样本的攻击原理可知，导致生物特征识别系统（例如人脸识别系统）发生误识别的根本原因是攻击者在生物特征数据（例如人脸图像数据）中增加了加性对抗噪声。

本申请的发明人发现，对抗噪声或对抗样本的大部分数据数值极低，例如仅改变了 8bit 图像数据的低位。因而，对生物特征数据中的

低位数据进行随机变换可破坏对抗噪声的大部分数据。在这里，“低位数据”是指可用于进行个人身份鉴定的人体所固有的数据中的最低位数据、倒数两位数据、或倒数三位数据。

在一个实施例中，低位随机滤波器 320 配置成对所述生物特征数据中的最低位进行随机变换。当然，本领域技术人员可以理解，在一个实施例中，随机滤波装置 320 可对生物特征数据中的倒数两位数据或倒数三位数据进行随机变换，而不限于最低位。

在一个实施例中，低位随机滤波器 320 配置成执行下列所示任一种：将所述生物特征数据中最低位全部置为 0；将所述生物特征数据中最低位全部置为 1；将所述生物特征数据中最低位随机重排；以及将所述生物特征数据中最低位随机生成。例如，对于 8bit 图像数据，最低位随机变换可包括如下四种算法：（1）将 8bit 最低位全部置 0；（2）将 8bit 最低位全部置 1；（3）将 8bit 最低位随机重排；以及（4）将 8bit 最低位随机生成。在一个或多个实施例中，低位随机滤波器 320 所采用的算法并不固定，而是从多种算法（例如上述四种算法）中进行随机选择，从而对输入图像进行预处理。

在低位随机滤波器 320 对生物特征数据中的最低位进行随机变换后，有利的是预处理器 330 对经随机变换的生物特征数据再次进行滤波（预处理），从而使生物特征数据平滑（例如使人脸图像数据平滑）。例如，预处理器 330 配置成通过随机选择一种或多种滤波算法，对经随机变换的生物特征数据进行预处理。

以图像数据（例如人脸图像数据）为例，存在多种图像去噪/滤波算法，包括但不限，BM3D 降噪；DCT 降噪；PCA 降噪；K-SVD 降噪；非局部均值降噪；WNNM 降噪；基于主成分分析和双边滤波的图像降噪算法；小波变换；小波阈值降噪；Contourlet 变换；基于平移不变 Contourlet 变换的 SAR 图像降噪等。在一个或多个实施例中，预处理器配置成从上述 11 种滤波算法中进行随机选择一种或几种，对输入图像进行预处理。正常情况下，去除噪声对人脸识别系统

的识别结果不会产生影响，却对图像中的对抗噪声破坏性极大，会使得对抗攻击失灵。

在一个实施例中，上述服务器 3000 还可包括：识别决策装置 350，用于将多次识别结果进行一致性比对；以及在识别结果均一致时反馈该结果。例如，第一次识别是针对原始生物特征数据（即，未经低位随机变换）进行身份识别，第二次识别和第三次识别则均是对经低位随机变换（可采用不同的滤波算法）的生物特征数据进行身份识别。若这三次身份识别的结果一致，则识别决策装置可直接反馈该身份识别结果。

在一个实施例中，识别决策装置 350 还配置成：在所述识别结果存在不一致时，反馈识别过程异常；以及将对应的数据采集终端加入黑名单。也就是说，在多次身份识别的结果（例如三次身份识别的结果）不同时，则可反馈识别过程异常（很有可能是对抗样本攻击）并将采集到异常数据的终端加入黑名单，后续将拒绝为该终端提供身份识别服务。这样，通过增加多图校验策略和黑名单拒绝策略，能够有效阻止攻击者进行黑盒模式的攻击训练。

目前，影响人脸识别最大的问题就是摄像头劫持攻击严重破坏人脸识别系统的安全性。其基本原理如下：首先，攻击者通过对支付应用 APP 进行“破壳”，实现对摄像头数据的劫持；接着，攻击者配合完成人脸识别的活体检测过程；最后，攻击者将摄像头上传的视频数据替换为事先伪造的被攻击者视频数据，将伪造数据上传至后台人脸识别系统，从而完成整个攻击过程。目前绝大多数的人脸识别系统，对此类摄像头劫持攻击都没有防范能力。

针对该问题，在一个实施例中，上述服务器 3000 还可包括：解密装置 360，用于对所述生物特征数据进行解密，以获得所述人脸图像数据以及所述时间戳；以及验证装置 370，用于在系统时间与所述时间戳之间的差值不符合要求时，反馈异常。在一个实施例中，验证装置 370 还配置成在所述系统时间与所述时间戳之间的差值符合要求

时，检测所述人脸图像数据中的人脸位置，并在所述人脸位置加入对抗样本以用于人脸识别。

验证装置370在人脸位置加入对抗样本是为了抵消人脸图像采集过程中嵌入负向对抗样本的影响。也就是说，在人脸图像采集过程中，人脸采集终端（例如客户端）需检测人脸位置，在人脸检测区域嵌入负向对抗样本，将处理后的人脸图像、人脸位置和时间戳以加密形式传送至服务端（例如人脸识别认证服务器）。这样，即使发生摄像头劫持，未在采集时嵌入负向对抗样本的图像也会在认证服务进行人脸识别前被加入对抗样本，从而无法在人脸识别过程得出正确结果。

在一个实施例中，对抗样本依据人脸识别结果对生成器进行反向传导训练而预先生成。例如，对抗样本可由生物特征识别服务器来生成，具体的生成方法包括：步骤A：构建对抗生成网络，所述生成网络包括生成器以及鉴别器；步骤B：利用所述生成器生成噪声信息，将噪声信息与真实人脸图像相加从而得到生成的人脸数据；步骤C：利用人脸识别模型对生成的人脸数据进行识别；步骤D：根据识别结果对所述生成器进行反向传导训练；步骤E：重复步骤B至步骤D直至所述人脸识别模型产生错误识别结果；以及步骤F：将与所述错误识别结果对应的噪声信息确定为对抗样本A。对抗样本A的相反数-A即为负向对抗样本。可以理解的是，对抗样本属于加性噪声，负向对抗样本是对抗样本的负值，用于抵消对抗样本的作用。

在一个实施例中，负向对抗样本只需进行一次生成，后续在图像中嵌入和去除的过程计算简便、效率高。而且，嵌入负向对抗样本的人脸数据对人眼不产生影响，即便数据在传输过程中被破解，也无法进行数据去除。

图4示出了根据本发明的一个实施例的人脸识别的流程图。如图4所示，人脸识别系统可包括多个装置或模块，如图4中所示的人脸采集模块、像素低位随机滤波器、随机噪声滤波器、人脸识别功能模块以及识别决策模块。在一个实施例中，人脸采集模块可位于客户端

侧，例如包含在用户智能手机中。像素低位随机滤波器、随机噪声滤波器、人脸识别功能模块以及识别决策模块可位于服务器侧，例如位于人脸认证服务平台中。

人脸采集模块可用于负责采集人脸数据。由于对抗噪声的大部分数据数值极低仅改变了 8bit 图像数据的低位，像素低位随机滤波器将图像数据的最低位进行随机变换，从而破坏对抗噪声的大部分数据，像素低位随机滤波器包含的滤波算法主要包括：将 8bit 最低位全部置 0；将 8bit 最低位全部置 1；将 8bit 最低位随机重排；将 8bit 最低位随机生成。在滤波过程中像素低位随机滤波器可从上述 4 种算法中进行随机选择，对输入图像进行预处理。在随机噪声滤波器中，图像去噪的算法包括但不限于以下方法：BM3D 降噪；DCT 降噪；PCA 降噪；K-SVD 降噪；NL-Means 非局部均值降噪；WNNM 降噪；基于主成分分析和双边滤波的图像降噪算法；小波变换；小波阈值降噪；Contourlet 变换；基于平移不变 Contourlet 变换的 SAR 图像降噪。随机噪声滤波器可从上述 11 种滤波算法中进行随机选择一种或几种，对输入图像进行预处理。

人脸识别功能模块用于负责完成图像中人脸识别的过程。识别决策模块用于对多次人脸比对的结果进行一致性比对，若检测结果一致则反馈该结果，若检测结果不一致则反馈识别过程异常，并将对应的数据采集终端加入黑名单。

在一个实施例中，人脸采集模块完成数据采集后，人脸识别功能模块对原始人脸数据进行第一次人脸识别，将识别结果反馈至识别决策模块。接着，将原始人脸数据进行第一次像素低位随机滤波和第一次随机噪声滤波，对滤波后的数据进行第二次人脸比对。然后，将原始人脸数据进行第二次像素低位随机滤波和第二次随机噪声滤波（确保第二次滤波方法与第一次滤波方法不同），对滤波后的数据进行第三次人脸比对。最后，识别决策模块对三次滤波的结果进行比对，若三次识别结果一致则反馈人脸识别的正常值；若三次识别结果不一致

则反馈识别存在异常（异常值为 S），并将采集到异常数据的终端加入黑名单，后续将拒绝为该终端提供人脸识别服务。

图 5 示出了根据本发明的一个实施例的包含数据采集终端以及人脸认证服务平台的人脸识别系统的示意图。数据采集终端可位于用户侧，而人脸认证服务平台位于远端服务器侧。

在图 5 的实施例中，数据采集终端用于为摄像头和可信执行环境提供运行资源，其中摄像头负责采集人脸数据，可信执行环境（TEE）用于为数据采集终端的应用程序和驱动程序提供安全运行、存储环境。如图 5 所示，数据采集终端还包括负向对抗样本嵌入模块，该嵌入模块用于检测人脸位置，在人脸检测区域嵌入负向对抗样本，并将处理后的人脸图像、人脸位置和时间戳以加密形式传送至位于远端的人脸认证服务平台。

在一个实施例中，人脸认证服务平台用于完成人脸识别认证服务，负向对抗样本验证模块用于对人脸图像、人脸位置和时间戳进行解密，而人脸识别模型用于对验证过的人脸识别图像进行识别，反馈被识别人的身份信息。

图 6 示出了利用图 5 所示的人脸识别系统进行人脸识别的流程图。如图 6 所示，在进行人脸识别前，数据采集终端进行应用初始化，将人脸检测模型和负向对抗样本-A 存储于可信执行环境。接着，摄像头采集人脸图像。负向对抗样本嵌入模块检测人脸图像中的人脸位置 P，并在 P 区域嵌入负向对抗样本-A 后将人脸图像和嵌入时间戳 t 加密上传至人脸认证服务平台。然后，服务器端的负向对抗样本验证模块解密人脸数据和嵌入时间戳 t，对嵌入时间戳 t 进行验证，若系统时间与 t 的差值不符合要求，则向平台反馈异常，否则将检测人俩位置 P，并在 P 位置嵌入对抗样本 A。最后，人脸识别模型执行人脸识别流程，并反馈识别结果。

在一个实施例中，负向对抗嵌入过程可包括如下步骤：在输入图像 F 中检测人脸位置 P（x1, y1, x2, y2），其中 x 和 y 分别为人脸框的

横纵坐标；图像中位置 P 确定的像素块内容为 S，令 S1 = （S - A）mod 256，将 S 值替换为 S1，此时图像转换为 F1；记录负向样本嵌入时间 t，将（F1，t）进行加密。

在一个实施例中，负向对抗样本验证流程可包括如下步骤：将（F1，t）进行解密；验证系统时间 t1 和时间戳 t 的差值在（-a, a）范围内；在输入图像 F 中检测人脸位置 P（x1, y1, x2, y2）；图像中位置 P 确定的像素块内容为 S1，令 S = （S1 + A）mod 256，将 S1 值替换为 S，此时图像转换为 F。

需要说明的是，尽管在图 6 中，负向对抗样本嵌入模块以及负向对抗样本验证模块等被示出为单个模块，但本领域技术人员可以理解，这些模块可以根据功能或需要被进一步细分，例如负向对抗样本验证模块可进一步细分为解密模块以及验证模块。

另外，以上所描述的装置实施例仅仅是示意性的，其中所述作为分离部件说明的模块也可以不是物理上分开的，即可以位于一个地方，或者也可以分布到多个网络模块上。可以根据实际的需要选择其中的部分或者全部模块来实现本实施例方案的目的。本领域普通技术人员在不付出创造性的劳动的情况下，即可以理解并实施。

通过以上各种实施方式的描述，本领域的技术人员可以清楚地了解到各实施方式可借助软件加必需的通用硬件平台的方式来实现，当然也可以通过硬件。基于这样的理解，上述技术方案本质上或者说对现有技术做出贡献的部分可以以软件产品的形式体现出来，该计算机软件产品可以存储在计算机可读存储介质中，所述计算机可读记录介质包括用于以计算机(例如计算机)可读的形式存储或传送信息的任何机制。例如，机器可读介质包括只读存储器(ROM)、随机存取存储器(RAM)、磁盘存储介质、光存储介质、闪速存储介质、电、光、声或其他形式的传播信号(例如，载波、红外信号、数字信号等)等，该计算机软件产品包括若干指令用以使得一台计算机设备(可以是个人计算机，服务器，或者网络设备等)执行各个实施例或者实施例的某些部

分所述的方法。

本申请是参照根据本申请实施例的方法、装置(设备)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的装置。

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。

这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于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的步骤。

综上，根据本发明的一个或多个实施例的生物特征识别方案通过对所接收的生物特征数据中的低位数据进行随机变换，并基于预处理后的生物特征数据进行身份识别。该方案对图像中的对抗噪声破坏性极大，会使得对抗攻击失灵。这样，即使攻击者获取生物特征识别模型的参数，对抗攻击模型的训练过程无法稳定收敛，因而无法形成有效的攻击模型。

另外，在一个或多个实施例中，通过增加多图校验策略和黑名单拒绝策略，能够有效阻止攻击者进行黑盒模式的攻击训练。此外，在人脸图像采集过程中加入负向对抗样本嵌入流程，在人脸识别系统端

加入负向对抗样本验证（去除）过程，这一设计使得未在采集时嵌入负向对抗样本的图像会在认证服务进行人脸识别前被加入对抗样本，从而无法在人脸识别过程得出正确结果。负向对抗样本只需进行一次生成，后续在图像中嵌入和去除的过程计算简便、效率高，而且嵌入负向对抗样本的人脸数据对人眼不产生影响，即便数据在传输过程中被破解，也无法进行数据去除。

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。显然，本领域的技术人员可以对本申请进行各种改动和变型而不脱离本申请的精神和范围。这样，倘若本申请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。

权 利 要 求 书

1. 一种生物特征识别方法，所述方法包括：

步骤 A：接收生物特征数据；

步骤 B：对所接收的生物特征数据中的低位数据进行随机变换；

步骤 C：随机选择一种或多种滤波算法，对经随机变换的生物特征数据进行预处理；以及

步骤 D：基于预处理后的生物特征数据进行身份识别。

2. 如权利要求 1 所述的方法，其中，所述生物特征数据包括人脸图像数据。

3. 如权利要求 1 所述的方法，其中，对所接收的生物特征数据中的低位数据进行随机变换包括：

对所述生物特征数据中的最低位进行随机变换。

4. 如权利要求 3 所述的方法，其中，对所述生物特征数据中的最低位进行随机变换包括下列所示任一种：

将所述生物特征数据中最低位全部置为 0；

将所述生物特征数据中最低位全部置为 1；

将所述生物特征数据中最低位随机重排；以及

将所述生物特征数据中最低位随机生成。

5. 如权利要求 1 所述的方法，其中，在所述步骤 D 中，基于预处理后的生物特征数据进行身份识别以便获得第一比对结果，所述方法还包括：

重复执行上述步骤 B 至步骤 D，以便获得第二比对结果；以及

基于所接收的生物特征数据直接进行身份识别，以便获得第三比对结果。

6. 如权利要求 5 所述的方法，所述方法还包括：

将所述第一比对结果、所述第二比对结果以及所述第三比对结果进行一致性比对；以及

在所述第一比对结果、所述第二比对结果以及所述第三比对结果均一致时反馈所述第一比对结果。

7. 如权利要求 6 所述的方法，还包括：

在所述第一比对结果、所述第二比对结果以及所述第三比对结果存在不一致时，反馈识别过程异常；以及

将对应的数据采集终端加入黑名单。

8. 如权利要求 2 所述的方法，其中，所述生物特征数据还包括时间戳，并且在所述步骤 A 和所述步骤 B 之间，所述方法还包括：

对所述生物特征数据进行解密，以获得所述人脸图像数据以及所述时间戳；以及

在系统时间与所述时间戳之间的差值不符合要求时，反馈异常。

9. 如权利要求 8 所述的方法，在所述步骤 A 和所述步骤 B 之间，所述方法还包括：

在所述系统时间与所述时间戳之间的差值符合要求时，检测所述人脸图像数据中的人脸位置。

10. 如权利要求 9 所述的方法，在所述步骤 A 和所述步骤 B 之间，所述方法还包括：

在所述人脸位置加入对抗样本，该对抗样本用于抵消在数据采集终端处所加入的负向对抗样本的作用，其中所述负向对抗样本是所述对抗样本的负值。

11. 如权利要求 10 所述的方法，其中，所述对抗样本依据人脸识别结果对生成器进行反向传导训练而预先生成。

12. 一种生物特征识别服务器，所述服务器包括：

接收装置，用于接收生物特征数据；

低位随机滤波器，用于对所接收的生物特征数据中的低位数据进行随机变换；

预处理器，用于随机选择一种或多种滤波算法，对经随机变换的生物特征数据进行预处理；以及

身份识别装置，用于基于预处理后的生物特征数据进行身份识别。

13. 如权利要求 12 所述的服务器，其中，所述生物特征数据包括人脸图像数据。

14. 如权利要求 12 所述的服务器，其中，所述低位随机滤波器配置成对所述生物特征数据中的最低位进行随机变换。

15. 如权利要求 14 所述的服务器，其中，所述低位随机滤波器配置成执行下列所示任一种：

将所述生物特征数据中最低位全部置为 0；

将所述生物特征数据中最低位全部置为 1；

将所述生物特征数据中最低位随机重排；以及

将所述生物特征数据中最低位随机生成。

16. 如权利要求 12 所述的服务器，还包括：

识别决策装置，用于将多次识别结果进行一致性比对；以及在识别结果均一致时反馈该结果。

17. 如权利要求 16 所述的服务器，其中，所述识别决策装置还配置成：

在所述识别结果存在不一致时，反馈识别过程异常；以及

将对应的数据采集终端加入黑名单。

18. 如权利要求 13 所述的服务器，其中，所述生物特征数据还包括时间戳。

19. 如权利要求 18 所述的服务器，还包括：

解密装置，用于对所述生物特征数据进行解密，以获得所述人脸图像数据以及所述时间戳；以及

验证装置，用于在系统时间与所述时间戳之间的差值不符合要求时，反馈异常。

20. 如权利要求 19 所述的服务器，其中，所述验证装置还配置成：

在所述系统时间与所述时间戳之间的差值符合要求时，检测所述人脸图像数据中的人脸位置。

21. 如权利要求 20 所述的服务器，其中，所述验证装置还配置成：在所述人脸位置加入对抗样本，该对抗样本用于抵消在数据采集终端处所加入的负向对抗样本的作用，其中所述负向对抗样本是所述对抗样本的负值。

22. 如权利要求 21 所述的服务器，其中，所述对抗样本依据人脸识别结果对生成器进行反向传导训练而预先生成。

23. 一种与如权利要求 21 所述的服务器配合使用的客户端，所述客户端包括：

人脸采集装置，用于采集人脸图像数据；以及

嵌入装置，用于在人脸位置嵌入负向对抗样本。

24. 如权利要求 23 所述的客户端，其中，所述嵌入装置配置成：

基于所述人脸图像数据检测所述人脸位置；

在所述人脸位置嵌入所述负向对抗样本；以及

将处理后的人脸图像以及时间戳以加密方式发送。

25. 一种利用如权利要求 12 至 22 中任一项所述的生物特征识别服务器来生成对抗样本的方法，所述方法包括：

步骤 A：构建对抗生成网络，所述生成网络包括生成器以及鉴别器；

步骤 B：利用所述生成器生成噪声信息，将噪声信息与真实人脸图像相加从而得到生成的人脸数据；

步骤 C：利用人脸识别模型对生成的人脸数据进行识别；

步骤 D：根据识别结果对所述生成器进行反向传导训练；

步骤 E：重复步骤 B 至步骤 D 直至所述人脸识别模型产生错误识别结果；以及

步骤 F：将与所述错误识别结果对应的噪声信息确定为对抗样本。

26. 一种计算机存储介质，所述介质包括指令，所述指令在运行时执行如权利要求 1 至 11 中任一项所述的方法。

27. 一种计算机程序产品，包括计算机程序，该计算机程序被处理

器执行时实现如权利要求 1 至 11 中任一项所述的方法。