大数据生物特征识别技术研究进展

刘琦[1],于汉超[2*],蔡剑成[3],韩琥[3]

1. 河南警察学院,郑州 450000
2. 中国科学院前沿科学与教育局,北京 100864
3. 鹏城国家实验室,深圳 518055

摘要 随着机器学习领域研究的持续发展,特别是深度学习方面的进步及图像处理器(GPU)等算力的持续提高,利用生物特征大数据的识别技术获得广泛关注,并在人证比对、智能监控以及疫情防控等多个领域取得了很好的应用。分析了大数据生物特征识别技术的发展态势,总结了生物特征类型以及大数据驱动的生物特征识别技术发展与应用,探讨了大数据生物特征识别技术的未来发展趋势。

关键词 生物特征识别;大数据驱动;自监督学习;领域通用特征;人工智能

生物特征识别是指利用人体固有的生理或行为特征进行身份识别,如人脸识别、指纹识别和虹膜识别等。早期的生物特征识别技术大多基于人工设计特征与传统机器学习方法,且受限于数据与算力获取的难度,因此,相关技术的研究和应用往往局限于实验室等严格控制的场景,很难有效利用大规模数据构建强大的生物识别模型。早在50多年前,好莱坞电影就将生物特征识别技术用于人物身份识别的黑科技搬上了电影银幕,吸引了大量观众的眼球。例如,在1968年上映的美国科幻电影《2001太空漫遊》中,人工智能机器人HAL 9000能够利用人脸识别技术进行身份识别,此外还能进行语音、表情及唇语识别;在1987年上映的电影《机械战警》中,墨菲通过人脸识别技术对人群中的人脸进行扫描,抓捕逃逸多年的通缉犯;在从1996年开始上映的电影《碟中谍》系列中,汤姆·克鲁斯每次在解锁新任务前都需要进行人脸扫描,通过虹膜识别技术确认其身份(图1)。《碟中谍》《机械战警》等好莱坞科幻电影中的身份识别技术其实都是不同类型的生物特征识别技术[1]。如今,人脸识别、指

收稿日期:2020-06-23;修回日期:2021-09-07
基金项目:国家自然科学基金项目(61732004,62176249)
作者简介:刘琦,副教授,研究方向为公安信息技术,电子信箱:lq@hnp.edu.cn;于汉超(通信作者),博士,研究方向为信息科学及科研管理,电子信箱:hcyu@cashq.ac.cn
引用格式:刘琦, 于汉超, 蔡剑成, 等. 大数据生物特征识别技术研究进展[J]. 科技导报, 2021, 39(19): 74-82; doi: 10.3981/j.issn.1000-7857.2021.19.009

(a) 电影《碟中谍》中的掌纹识别技术

(b) 电影《机械战警》中的人脸和虹膜识别技术

图1　科幻电影中经常出现的基于人脸、掌纹及虹膜的生物特征识别技术(图片来源于互联网)

纹识别、虹膜识别等生物特征识别技术已经非常成熟,并被广泛应用于人们的日常生活和工作中。

生物特征识别技术最早被应用于对罪犯的身份鉴别任务。19世纪末期,法国警察Alphonse Bertillon设计并制造了第1套身份鉴定系统(Bertillon system,https://en.wikipedia.org/wiki/Alphonse_Bertillon)。该系统是一种基于身体测量指标描述个人身份信息的系统,即通过站立高度、坐姿高度(躯干和头部的长度)、伸直双臂到指尖之间的距离以及头部、右耳、左脚、手指和前臂大小等特征确定身份信息;此外,该系统基于这些特征将人分为243种类别,支持通过手工方式对人体进行检索。该系统以其易用性和有效性随即被大多数欧洲国家所采用,但在实际使用中,存在少数不同个体身体测量指标十分相近的情况,导致该系统不能精准区分用户身份信息。

随着20世纪八九十年代人工智能的第2次热潮,基于统计学习模型的生物特征识别方法研究兴起,主要针对不同的生物特征,手工设计适宜的特征表示用于身份识别,例如局部二值编码方法。此类方法对于指纹、虹膜等特征明显、纹路清晰的生物特征,能够取得较好的识别结果;而对于人脸、步态、声纹等特征不清晰的生物特征,其识别效果尚未达到实用化要求。因此,在这一阶段,指纹识别、虹膜识别等技术被公安部门、移民局等大量使用,而人脸识别等生物特征识别技术最终由于系统性能无法满足实际应用的入门需求,未被广泛使用。

进入第3次人工智能热潮以来,特别是随着深度学习技术的发展,机器在计算机视觉标准数据集ImageNet等图像分类数据集上的分类准确率不断被刷新,生物特征识别技术再次成为人工智能领域研究的热点之一,其性能也随着大数据与高算力的不断获取而日渐提高。例如,在人脸评测基准LFW(labled faces in the wild)人脸数据库[2]上,传统人脸识别方法的准确率最高只能达到96.33%,但基于深度学习算法的DeepID2方法[3]将LFW上的识别准确率刷新到99.15%,成为首个超过人类识别准确率(97.53%)的方法,性能趋于饱和。此外,指纹、虹膜、步态、掌纹等生物特征识别技术也取得了飞速发展,例如在步态识别标准数据库CASIA-B上,传统方法在90°步态上最高可达到60.4%的识别准确率,而采用深度学习的方法识别准确率最高可达到95.1%[4]。

中国具有发展人工智能技术的良好基础,经过多年持续积累,语音识别、视觉识别等生物特征识别技术达到世界领先水平并逐步进入实际应用。同时,国家也高度重视生物特征识别技术,在2017年国务院发布的《新一代人工智能发展规划》[5]中将发展生物特征识别技术列为重要内容。本研究将重点综述生物特征的特点、大数据驱动的生物特征识别技术及其发展趋势。

1　生物特征

人类具有多种可用于区分不同个体的生理或行为特征,利用这些标识性特征进行个体身份识别的过程称为生物特征识别。一般认为,用于身份识别的生物特征应当具备以下特点。

1) 普遍性(universality):该特征是否普遍存在于每个个体中。

2) 唯一性(uniqueness):该特征是否针对每个个体都是不同的。

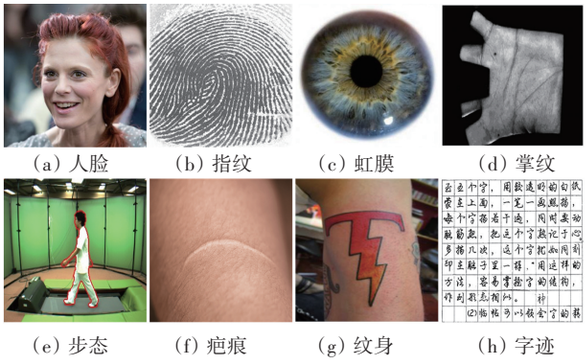
3) 持久性(permanence):该特征是否长时间不发生变化。

4) 可获取(collectability):该特征是否方便收集。

5) 防伪性(circumvention):该特征是否容易被伪造。

根据不同生物特征满足条件的不同,可将生物

特征大致分为两类:一类是主要生物特征(primary biometric traits),另一类是辅助生物特征(soft biometric traits)[1]。辅助生物特征与主要生物特征的区别在于"唯一性"方面,即辅助生物特征可能无法唯一确定一个人,但通常能有助于将身份限定在某个范围内。图2展示了人脸、指纹、虹膜、掌纹、步态、疤痕、纹身和字迹8种常见的生物特征,其中,人脸、指纹、虹膜、掌纹、步态和字迹一般被归为主要生物特征,而疤痕和纹身一般被归为辅助生物特征。



(a) 人脸 (b) 指纹 (c) 虹膜 (d) 掌纹

(e) 步态 (f) 疤痕 (g) 纹身 (h) 字迹

图2 常见的8种生物特征

1.1 主要生物特征

主要生物特征是具有较强区分能力的一类生物特征,可分为生理特征和行为特征2类。

1) 生理特征:人与生俱来的特征,如人脸、指纹、虹膜、掌纹等。

人脸是指人体的面部信息,其在胚胎发育时开始形成,人成年后面部信息基本不发生改变。面部特征相对容易采集,且每个人各不相同,可作为生物特征对身份进行准确认证。

指纹是指人手指表面凸起或凹陷的纹路,是在遗传和环境的作用下产生的,能在手指接触物体时增加摩擦力,且指纹重复率极低,出现相同指纹的概率小于10^{-11}[6-7]。

虹膜是指位于角膜和晶状体之间的扁圆形环状薄膜,由结缔组织构成,在光照下有明显的视觉特征,虹膜在人出生8个月后形成稳定的形态,终生不再发生改变,且不易被损坏[8]。

掌纹是指人手掌皮肤表面凸起或凹陷的纹路信息,人的掌纹信息也各不相同,可用于身份信息的鉴定。

2) 行为特征:人在后天形成的特征,如步态、笔迹、声纹等。

步态是指人行走的姿态,相比其他特征,步态无需用户配合,不用担心用户故意遮挡生理特征而拒绝特征采集[9]。

笔迹是指个人所写的字体带有的特有形体特点,每个人由于生理和心理因素不同,导致书写的字体形态不同,通过对字体细微处的区分,可以对身份进行鉴别[10]。

声纹是指用电声学仪器显示的携带言语信息的声波频谱,人在成年之后,声音长期保持不变,且不易被人模仿,可用于对身份的确认[11]。

在实际应用中,可通过普遍性、唯一性、持久性、可获取性、防伪性、可识别性和可接受性7种特性(每种特性均分为高、中、低3个等级),对上述人脸、指纹、虹膜、掌纹、步态、笔迹和声纹7类常见的主要生物特征的特性进行比较评价(表1)。由比较结果可知,当前常见的主要生物特征普遍存在于每个个体中,且是现阶段人们普遍接受的可用于身份识别的生物特征。此外,相比行为特征,生理特

表1 主要生物特征的特性比较

生物特征		普遍性	唯一性	持久性	可获取性	防伪性	可接受性	可识别性
生理特征	人脸	高	高	中	高	高	高	高
	指纹	高	高	高	中	高	高	高
	虹膜	高	高	高	中	高	高	高
	掌纹	高	高	高	中	中	高	高
行为特征	步态	高	中	中	高	中	中	高
	笔迹	高	中	中	高	低	高	高
	声纹	高	中	中	高	低	低	高

征具有更大的区分度和更好的防伪性,更易于大规模使用。

1.2　辅助生物特征

辅助生物特征是具有一定区分能力的生物特征,其主要包括属性、纹身、疤痕、斑痣和行为等。

属性是一种具有区别性的客观特征,包含统计性属性(例如年龄、性别、种族)和描述性视觉属性(例如衣服、发型),用于确定主体的身份。

纹身是指通过使用带墨的针刺入皮肤底层,在皮肤上绘制特定图案来标记身体,从而表达个人信念或表示团体关联的一种方式。

疤痕是指由特定因素带来的皮肤损害,因皮肤软组织无法正常修复导致创伤处呈现短时间内难以消除的异常外观(例如皮肤颜色加深、表面不光滑等)。产生疤痕的因素多种多样,例如利器割伤、烧伤、烫伤等。这些因素从侧面反映了主体的生活经历,且相较于人脸随时间变化程度,疤痕外观不易改变,因此在已知一定的事件信息下(例如火灾),疤痕可用于准确标识主体身份。

斑痣是指由黑色素细胞增生引发的局部皮肤颜色异常,通常呈现斑点状。与疤痕不同,斑痣一般不会对皮肤表面带来创伤,引发的原因可以是身体内部自发病变,也可以是外部环境因素(例如强紫外线),且引发的因素难以准确判断。

行为(击键)是指主体在使用智能设备时所显示的个人习惯,例如点击按键的频率、力度、手势等。类似主要生物特征,可通过普遍性、唯一性、持久性、可获取性、防伪性、可接受性6种特性(每种特性均分为高、中、低3个等级),对上述属性、纹身、疤痕、斑痣和行为(击键)5类常见的辅助生物特征的特性进行比较评价(表2)。由比较结果可知,当前常见的辅助生物特征的防伪性普遍较差,难以作为唯一的特征用于身份信息鉴定。

表2　辅助生物特征的特性比较

生物特征	普遍性	唯一性	持久性	可获取性	防伪性	可接受性
属性	高	中	低	高	低	高
纹身	低	高	高	中	低	中
疤痕	中	高	高	中	低	中
斑痣	高	高	高	中	低	中
行为(击键)	高	中	低	高	低	高

2　大数据驱动的生物特征识别技术

2.1　生物特征大数据

现阶段,国内存在数量众多、不同类型的生物特征采集和识别系统,例如,智慧城市监控系统、人脸安检系统、人手智能解锁系统、智能门禁系统等。这些系统每天采集大量的生物特征数据,以智慧城市监控系统为例,截至2019年,北京、上海、广州和深圳等城市安装的摄像监控设备超过了6000万(数据来源:https://www.comparitech.com/vpn-privacy/the-worlds-most-surveilled-cities),每天产生大量包含人脸、人体、步态等不同生物特征的视频数据,若按照每秒钟产生24张图像计算,每天产生超过5000亿张图像,而且未来随着中国智慧城市建设的推进,摄像头数目还将进一步增加(图3)。此外,2019年上半年,中国智能手机出货量为1.8亿台,这些手机几乎全部搭载了指纹识别或人脸识别系统,据统计机构Strategy Analytics调查显示,用户日均解锁手机60次以上,因此全世界智能手机用户每天采集和识别的生物特征数据高达百亿条以上。

图3　现代城市中随处可见各种视频监控摄像头

生物特征大数据通常具有多模态、多样性、非结构化、大容量、难标注和可复制6类特点。

1)多模态:由于生物特征的采集设备和采集方式不同,对于相同的生物特征,所采集到的生物数据往往具有差异性,每一种采集到的生物信息都可称之为模态,生物特征往往具有多模态的特点。以人脸为例,普通摄像机采集到的是RGB图像,深度摄像机采集到的是深度图像,近红外摄像机采集到的是近红外图像,在部分场景中还会采集人脸的视频数据。在一些安全等级要求较高的场景中,通

常会选用2种或2种以上的模态,以多模态融合的方式进行身份信息认证,但这种验证方式会增加计算量,对现有的生物识别系统造成额外负担。

2)多样性:在开放场景下,生物特征的采集会受多种因素干扰,采集到的特征通常具有多样性。以指纹采集为例,在受限场景下,指纹采集会受传感器的分辨率、信噪比等因素影响;而在开放场景下,采集到的指纹可能会出现部分缺失或被油污干扰等问题,通常会对现有的指纹识别系统造成极大挑战。此外,生物特征的多样性还体现在不同类型生物特征的表现形态上,如人脸、指纹、虹膜、掌纹等,对于不同类型的生物特征,需要有针对性地设计不同类型的算法,这也对现有的生物识别系统造成了一定挑战。

3)非结构化:开放场景中,生物特征数据大多数是非结构化的。例如,同一张图像中可能混杂了多个人的多种生物特征,并且不同图像混杂的人数和生物特征类型、数目也不一致;每个人的生物特征在画面中可能存在不同程度的遮挡,遮挡物可能是他人的生物特征、自身衣物或场景物体等;同一个人的生物特征可能会连续或离散的出现在不同数据中,包括不同时刻的画面、画面中的不同位置、在画面中不同的缩放程度和旋转角度等。因此,从非结构化数据中直接提取特定的生物特征信息是一件十分困难的事。

4)大容量:生物特征数据的规模很大,例如,一个分辨率为1080 P的监控摄像头,以20帧/s的速度采集图像,24 h可以产生约172万张图像,数据存储量约8 GB。大城市摄像头的数量已经达到百万级别,而一百万摄像头每天产生的图像量约17200亿张,如此万亿级别的海量图像数据,给存储、识别、检索等分析任务都带来了巨大挑战。

5)难标注:生物特征数据同样面临标注难的问题。生物特征识别常常需要标注每个人的身份信息,而从大量数据中筛选出包含某个人的数据非常困难。例如,将单个监控摄像头一天拍摄的图像按照不同的人所出现的画面归类,需要逐张筛选,非常耗时耗力;如果将多个摄像头中出现的同一个人的画面全部归类到一起,难度更高。除了单模态数据,特定任务有时还需要另一个模态的成对数据,如人脸三维重建、人脸去遮挡、人脸风格转换等,收集这类成对数据也很难。此外,许多任务很难准确标注数据,如人脸特征点、人体关节关键点、手势关键点、人体遮挡标注等。

6)可复制:生物特征数据虽然由本人直接携带,但也存在可复制的问题,即其他人可以使用生物采集设备轻易采集到本人的生物特征。对于不同类型的生物特征,有不同的复制方式,例如,采用人脸照片或人脸面具复制人脸、通过制作指模复制指纹信息、利用录音复制声纹等。生物特征复制会对当前生物特征识别系统的安全性造成极大挑战,因此,在进行生物特征识别前,需要对生物特征进行活体检测,判断该特征是否由人直接携带。

2.2 大数据生物特征识别技术与应用

2.2.1 多模态生物特征识别

多模态生物特征识别包含对同一生物特征由不同传感器采集得到的多模态数据进行识别,和对从同一个体获取的人脸、指纹等多种生物特征的多模态数据进行识别。对于相同生物特征的多模态数据,以人脸为例,包含可见光、近红外、深度等不同模态,对这些模态进行组合,可以获取更多有效信息,有利于人脸识别的效果。Zhang等[12]提出了一种多模态人脸识别方法,将人脸RGB图像与人脸深度图像在特征层面进行融合,相比于单一模态的人脸识别,有效提升了人脸识别的准确率和可靠性。对于不同生物特征的多模态数据,以人脸和声纹为例,常用基于注意力机制的人脸和语音特征相融合的方法进行身份识别,以提高身份识别的准确率和识别系统的安全性,防止他人刻意使用照片或面具对系统进行攻击。

2.2.2 生物特征数据增广

在开放场景下,生物特征采集设备可能采集到低质量的生物特征,往往存在特征部分缺失、低分辨率、模糊等问题。现有的生物特征识别模型大多采用高质量的生物特征进行训练,对于低质量的生物特征并不鲁棒,因此需要对生物特征进行数据增强,进而提升生物特征的质量和识别效果。以人脸为例,Cai等[13]提出了基于多任务学习的人脸超分

辨与遮挡补全算法,能够同时对低分辨率、有遮挡的人脸图像进行超分辨和遮挡修复,有效提升了人脸识别的准确率。类似地,在开放场景指纹识别中,指纹图像数据增强也展现出了重要作用[7]。

2.2.3　生物特征数据结构化

非结构化数据是大数据分析中的重要挑战之一。为了从非结构化数据中挖掘有用的信息,通常需要借助数据处理方法,将非结构化数据转化为结构化数据,进而进行数据分析和模型学习。具体来说,针对非结构化的生物特征数据,可以通过检测、识别等方法,从非结构化的视频或图像中抽取不同维度的个体特征,形成结构化的个体表达。早期的结构化方法主要通过单任务方式实现,近年来基于多任务学习的数据结构化方法获得越来越多的重视。例如,通过并行多任务学习或级联多任务学习,挖掘不同任务间的相关性和异质性[14],将目标检测与识别任务进行高效建模,从而实现对视频图像中生物特征的高效结构化。

2.2.4　大规模生物特征检索

在刑侦、司法和公共安全领域,往往需要在数亿甚至数十亿的视频图像数据库中进行生物特征检索,因此生物特征检索任务中除了关注精准率与召回率外,还特别关注检索方法的高效性。在生物特征的特征表示方面,除了要求特征表示具有良好判别力,还要求特征表示具有紧凑性,但判别力和紧凑性之间往往存在平衡关系。此外,现有的面向生物特征检索的特征表示往往忽略了上下文任务之间的相关性,例如特征表示学习没有考虑前段生物特征结构化,特别是生物特征检测之间的相关性[15],因而可能存在计算冗余及特征表示不是最优的问题。如何借鉴人在视觉感知中的定位与识别协同进行机理,建立高效的生物特征检索技术,仍然是一个需要进一步研究的问题。

2.2.5　自监督与弱监督生物识别建模

自监督与弱监督学习是近年来机器学习领域的一个重要发展方向,其优势在于降低模型对大规模准确标注数据的依赖。其中,自监督学习可以完全不使用数据标注,而弱监督学习使用不完全、不准确或不确切的数据标注。不完全是指数据只含有少量准确数据标注;不准确是指数据的标注可以存在部分错误标注,也称为噪声数据;不确切是指数据标注是正确的,但是与所面临的任务并不直接相关(比如粗粒度图像类别标注)。自监督与弱监督建模在自然图像物体分类、目标检测等领域已取得飞速发展,但在大数据生物特征识别中,基于自监督与弱监督学习的建模方法还相对较少。此外,在视频监控、社交媒体等实际应用场景中,往往存在海量的无身份标注生物特征数据图像,且每类生物特征图像都是具有共性的一大类图像,例如,不同的人脸都具有相似的三维结构(比如都有五官)和形状。因此,如何利用自监督与弱监督学习,实现更有效和更鲁棒的特征学习,进而服务于不同的下游任务,例如生物特征的检测、识别和检索,对于构建具有领域通用性的生物特征识别模型具有重要意义。

2.2.6　生物特征欺骗及其防护

现阶段生物特征识别技术运用越来越广泛,但该技术本身还存在易被攻击的安全隐患。目前,生物特征识别系统的攻击类型主要有3种:第1种是伪造生物特征对系统进行攻击,常见的攻击手段有伪造指膜对指纹识别系统进行攻击,采用人脸面具或人脸照片对人脸识别系统进行攻击等;第2种是使用对抗样本对系统进行攻击,对抗样本是指对输入系统的样本故意添加一些干扰项,导致系统以较高的置信度给出一个错误的输出;第3种是对生物特征的模板数据集进行攻击,现阶段生物特征数据大多被保存在大型的数据库中,对数据库进行攻击可以对模板进行修改和替换,进而可造成生物特征识别系统出错。

为了提升生物特征识别系统的安全性,系统在进行识别之前应先对生物特征进行检测:针对第1种类型的攻击,系统应加装生物特征活体检测模块,检测该特征是由人体携带的,还是后期复制得来的;针对第2种类型的攻击,应该从加强神经网络的鲁棒性着手,防止此类异常样本对系统造成影响;针对第3种类型的攻击,系统应加装安全防护模块,并对所有生物特征进行加密,防止模板被恶意破坏。

3 大数据生物特征识别技术发展趋势

生物特征识别技术的广泛应用一方面得益于人工智能技术的迅速发展,另一方面是由于存储和计算能力的飞速提升使得获取和利用生物特征大数据成为可能。虽然在深度学习与大数据的共同加持下,生物特征识别技术取得了突飞猛进的发展,但大数据生物特征识别技术与整个人工智能领域研究的发展趋势是一致的,也面临一些共性的挑战[16-18]。未来大数据生物特征识别技术发展主要呈现以下5方面的趋势。

3.1 开放场景生物特征识别

现有的生物特征识别技术,在受限场景下表现出了良好性能并已大规模运用于生产生活中,但在开放场景下的性能尚不尽如人意,其原因在于开放场景下存在大量低质量的生物特征,且开放场景下的用户通常不配合生物特征识别系统。此外,现有的生物特征数据集大多是高质量且在用户配合下采集的,仅基于当前的数据集训练模型很难泛化到开放场景,因此开放场景下的生物特征识别技术已经成为当前研究的重点和难点。在该问题的研究上,针对低质量的生物特征,未来将向生物特征增强的研究方向发展;针对用户不配合的问题,未来将向多种生物特征融合的研究方向发展。

3.2 低资源场景模型迁移

实际应用中,很难为每个任务都构建一个大规模精确标注的数据集用于模型训练。由于收集数据困难或者降低收集成本等原因,新场景中可用的数据往往数量有限,且可能存在标注缺失、错误、不准确的情况。比如,一张含有多个人的图像只标注了一部分人的身份信息,表情属性标注时错误将嘴微微张开的图像标注为微笑,标注人体关键点位置时偏离真实位置等。此外,在一些边缘计算场景中,算力往往是有限的,如何对这些低资源场景进行快速鲁棒建模,是亟需研究的一类问题。因此,研究将原有模型向不同目标场景和任务进行高效迁移具有重要意义。特别是随着自然语言处理领域和视觉领域大模型的涌现,如何在低资源情况下将这些强大的模型用于某一具体任务的建模,具有重要意义。

3.3 未知生物特征攻击防范

现阶段生物特征识别技术已经深入人们衣食住行的方方面面,针对生物特征识别系统的欺骗和攻击也随之成为大数据生物特征技术研究的一个重要方面。生物特征欺骗的方式越来越多样化,例如,针对人脸图像有照片、视频播放、人脸面具等多种呈现式的攻击方式(图4),针对指纹有指纹膜的攻击方式。除了这些传感器前端的呈现式攻击方式,还有针对生物特征识别模型设计的对抗样本攻击,图5[19]所示为深度伪造技术(DeepFake),可以将一个人的表情动作等迁移到另一个人物身上,几乎达到以假乱真的效果。目前针对这些已知的呈现攻击和对抗样本攻击已经有了较大的研究进展,然而在模型学习工程桩攻击的方式难以穷举,因此在未来的研究上,需要针对精心设计的各种未知的欺骗方式设计出更可靠的防护方法,加强生物特征识别系统的安全性。

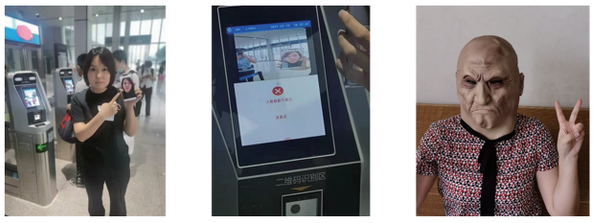


图4 常见的人脸照片欺骗和更具挑战的面具欺骗

图5 几乎达到以假乱真的人脸深度伪造技术(DeepFake)

3.4 个人隐私与数据保护

生物特征是人体具有的生理或行为特征,每个人的生物特征在一段时间会保持稳定,不能像密码口令那样随时改变,个人的生物特征数据一旦被非授权用户非法获取,则会面临重大的身份信息泄露风险,理应受到法律保护。如果将生物特征大数据

(人脸、指纹、虹膜等)与各类网站、App等获取的个人地理位置、行为轨迹等相结合,则可能掌握个人居住地址、工作场所、出行路线与爱好、甚至人格等方面的个人隐私信息。因此,在大数据生物特征识别技术使用的同时,应当做好用户生物特征使用范围的授权管理和隐私保护,避免个人的生物特征被非授权使用和滥用。2021年7月,中华人民共和国最高人民法院发布了《最高人民法院关于审理使用人脸识别技术处理个人信息相关民事案件适用法律若干问题的规定》,认定在宾馆、商场、银行、车站、机场、体育场馆、娱乐场所等经营场所、公共场所,违反法律、行政法规的规定使用人脸识别技术进行人脸验证、辨识或者分析,属于侵害自然人人格权益的行为,这一司法规定对于保护敏感个人信息、促进数字经济健康发展具有积极的作用。

3.5　生物特征识别模型可解释性

目前的生物特征识别方法大多数是基于"深度模型+大数据"的方式实现,深度学习模型更多的时候被当作一个进行特征抽取的"黑盒子"来使用,缺乏像基于物理模型的传统方法那样的可解释性[20],导致在很多情况下难以判断模型为什么能成功和为什么会失败。这会导致在一些关键场景中,错误不可控的问题,甚至生物特征识别系统会出现一些在人看起来很"愚蠢"的错误。因此,未来也亟需从模型可解释性的角度研究生物特征识别模型的构建方法,避免因为数据密集型模型建模带来的偏差(bias)问题。此外,未来大数据生物特征识别技术的研究,也可以将人类知识与机器知识相结合,提升模型的可解释性和鲁棒性。

4　结论

对生物特征识别技术,特别是进入大数据与高算力时代以来的技术发展进行了总结,并基于国内外发展现状,分析了未来提升中国大数据生物特征识别技术竞争力需要关注的方面:开放场景生物特征识别、低资源场景模型迁移、未知生物特征攻击防范、个人隐私与数据保护以及生物特征识别模型可解释性。期望业界能够对上述几个方面给予更多的关注和研发投入,进一步提升中国在生物特征识别领域的核心竞争力,推动相关技术在更广泛的场景落地应用。

参考文献(References)

[1] Jain A K, Flynn P, Ross A A. Handbook of biometrics [M]. Boston: Springer, 2008.

[2] Huang G B, Ramesh M, Berg T, et al. Labeled faces in the wild: A database for studying face recognition in unconstrained environments[R]. Amherst: University of Massachusetts, Amherst, 2007.

[3] Sun Y, Chen Y H, Wang X G, et al. Deep learning face representation by joint identification-verification[C]//Proceedings of the 21th International Conference on Neural Information Processing Systems. Heidelberg: Springer, 2014: 1988-1996.

[4] Zhang Z Y, Tran L, Liu F, et al. On learning disentangled representations for gait recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, doi: 10.1109/TPAMI.2020.2998790.

[5] 国务院关于印发新一代人工智能发展规划的通知[A/OL]. (2017-07-08)[2018-06-30]. http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm.

[6] Maltoni D, Maio D, Jain A K, et al. Handbook of fingerprint recognition[M]. London: Springer London, 2009.

[7] Feng J J, Zhou J, Jain A K. Orientation field estimation for latent fingerprint enhancement[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(4): 925-940.

[8] Sun Z N, Tan T N. Ordinal measures for iris recognition [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2009, 31(12): 2211-2226.

[9] Wang L, Tan T N, Ning H Z, et al. Silhouette analysis-based gait recognition for human identification[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2003, 25(12): 1505-1518.

[10] 刘成林, 刘迎建, 戴汝为. 基于多通道分解与匹配的笔迹鉴别研究[J]. 自动化学报, 1997, 23(1): 56-63.

[11] Atal B S. Automatic recognition of speakers from their voices[J]. Proceedings of the IEEE, 1976, 64(4): 460-475.

[12] Zhang H, Han H, Cui J Y, et al. RGB-D face recognition via deep complementary and common feature learning[C]//2018 13th IEEE International Conference on Au-

tomatic Face & Gesture Recognition (FG 2018). Piscataway: IEEE, 2018: 8-15.
[13] Cai J C, Han H, Shan S G, et al. FCSR-GAN: Joint face completion and super-resolution via multi-task learning [J]. IEEE Transactions on Biometrics, Behavior, and Identity Science, 2019, 2(2): 109-121.
[14] Han H, Jain A K, Wang F, et al. Heterogeneous face attribute estimation: A deep multi-task learning approach [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(11): 2597-2609.
[15] Han H, Li J, Jain A K, et al. Tattoo image search at scale: Joint detection and compact representation learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 41(10): 2333-2348.
[16] Tan T N. Recent advances and future directions of biometric recognition[R]. Beijing: CAAI, 2019.
[17] Nixon M. A future of biometrics[R]. Beijing: CAAI, 2019.
[18] 于汉超, 汪峰, 蒋树强. 中国人工智能发展的若干紧要问题[J]. 科技导报, 2018, 36(17): 40-44.
[19] Güera D, Delp E J. Deepfake video detection using recurrent neural networks[C]//2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS). Piscataway: IEEE, 2018: 1-6.
[20] Kuo J C C. Towards effective and explainable biometrics [R]. Beijing: CAAI, 2019.

Research progress and trend of big-data based biometrics

LIU Qi[1], YU Hanchao[2*], CAI Jiancheng[3], HAN Hu[3]

1. Henan Police College, Zhengzhou 450000, China
2. Bureau of Frontier Sciences and Education, Chinese Academy of Sciences, Beijing 100864, China
3. Pengcheng National Laboratory, Shenzhen 518055, China

Abstract Biometric recognition refers to the use of inherent physiological or behavioral characteristics of human body for personal identification, such as face recognition, fingerprint recognition, and iris recognition. Early technologies are usually based on handcrafted features and traditional machine learning methods, and constrained by data and computing power acquisition ability. Thus, they are generally limited to controlled environment and difficult to build powerful biometric recognition models by making use of large-scale biometric data effectively. In recent years, with the advancement of machine learning, especially deep learning and increasing computing power of GPU, big data based biometric recognition technologies have received widespread attention and have been widely used in the areas such as personal authentication, intelligent video surveillance, and prevention and control of epidemic. This paper summarizes the development of big-data based biometrics technologies, covering different types of biometrics and their applications, and then discusses the trend of big-data based biometrics technologies in the future.

Keywords biometric recognition; big data driven; self-supervised learning; universal domain representation; artificial intelligence

（责任编辑　刘志远）