



(12) 发明专利申请

(10) 申请公布号 CN 103793697 A

(43) 申请公布日 2014. 05. 14

(21) 申请号 201410053879. 6

(22) 申请日 2014. 02. 17

(71) 申请人 北京旷视科技有限公司
地址 100080 北京市海淀区海淀大街 3 号 1 幢 1001-011 室

(72) 发明人 曹志敏 印奇 姜宇宁

(74) 专利代理机构 北京君尚知识产权代理事务
所(普通合伙) 11200
代理人 余长江

(51) Int. Cl.
G06K 9/00(2006. 01)
G06K 9/66(2006. 01)
G06F 21/32(2013. 01)

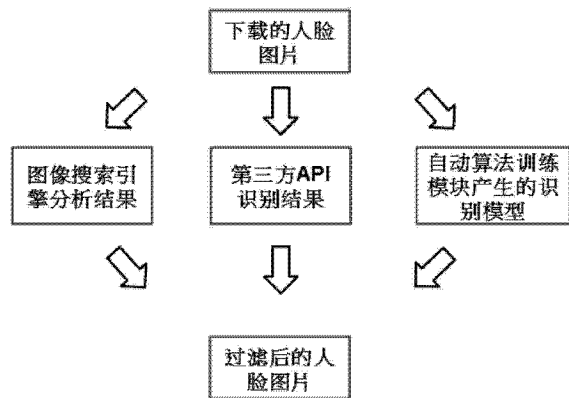
权利要求书1页 说明书5页 附图2页

(54) 发明名称

一种人脸图像的身份标注方法及人脸身份识别方法

(57) 摘要

本发明公开了一种人脸图像的身份标注方法及人脸身份识别方法。本识别方法为:1)对每一待标注人脸图片的身份进行标注:搜索与该图片相似的人脸图像及对应的网页;根据返回网页中所出现人名的频率确定该图片的身份;分别采用人脸技术平台和人脸身份识别模型检测该图片的身份;综合上述识别结果确定该图片的最终身份并标注;2)将属于同一人名的一组图片与1)中标注结果为该人名的图片进行匹配过滤;3)提取过滤后的每一身份标注图片的特征向量,利用机器学习算法对标注后的人脸图片进行训练,生成人脸身份识别模型;4)对于待检测的两人脸图像,提取其特征向量利用人脸身份识别模型判断其是否属于同一人。本发明大大提升标注效率和识别效果。



1. 一种人脸图像的身份标注方法,其步骤为:
 - 1) 从图像搜索引擎中搜索与待标注人脸图片相似的人脸图像及对应的网页;
 - 2) 统计所返回网页中所出现人名的频率,并根据该频率初步确定该待标注人脸图片的身份;
 - 3) 分别采用人脸技术平台和人脸身份识别模型检测该待标注人脸图片的身份;
 - 4) 根据步骤 2)、3) 的识别结果确定该待标注人脸图片的最终身份,标注该待标注人脸图片的身份。
2. 如权利要求 1 所述的方法,其特征在于首先提取待标注人脸图片的人脸位置和关键点信息,将人脸位置标准化到一标准格式脸上。
3. 如权利要求 1 或 2 所述的方法,其特征在于所述人脸身份识别模型将搜索的属于同一人名的 N 张图像两两比对图片中人脸的身份属于同一个人的置信度;然后根据置信度确定待标注人脸图片身份是否为该人名。
4. 一种人脸图像的人脸身份识别方法,其步骤为:
 - 1) 数据自动采集系统从服务器获取人脸图片及其上下文信息;
 - 2) 数据自动标注系统对获取的每一待标注人脸图片的身份进行标注;其中标注方法为:
 - 21) 从图像搜索引擎中搜索与待标注人脸图片相似的人脸图像及对应的网页;
 - 22) 统计所返回网页中所出现人名的频率,并根据该频率初步确定该待标注人脸图片的身份;
 - 23) 分别采用人脸技术平台和人脸身份识别模型检测该待标注人脸图片的身份;
 - 24) 根据步骤 22)、23) 的识别结果确定该待标注人脸图片的最终身份,标注该待标注人脸图片的身份;
 - 3) 将数据自动采集系统中采集的属于同一人名的一组图片与步骤 2) 中标注结果为该人名的图片进行匹配过滤,去除该组中不属于该人名的人脸图片;
 - 4) 提取步骤 3) 过滤后保留的每一身份标注图片的特征向量,自动算法训练系统利用机器学习算法对身份标注后的人脸图片进行训练,生成一人脸身份识别模型;
 - 5) 对于待检测的两人脸图像,提取其特征向量利用所述人脸身份识别模型判断其是否属于同一人。
5. 如权利要求 4 所述的方法,其特征在于所述数据自动采集系统从服务器获取人脸图片及其上下文信息的方法为:
 - 51) 所述服务器根据输入的人名关键字搜索相应的人脸图片文件并保存;
 - 52) 计算每一人脸图片文件的哈希码、颜色直方图、上下文和标签信息;
 - 53) 将每一人脸图片与已存人脸图片进行哈希码和颜色直方图比对,去除重复的图像;
 - 54) 使用人脸检测算法模块检测步骤 53) 处理后保留的每一人脸图片,将人脸位置信息保存到数据库;使用人脸关键点定位算法定位人脸上的关键点信息并保存到数据库。
6. 如权利要求 4 或 5 所述的方法,其特征在于步骤 21) 之前先提取待标注人脸图片的人脸位置和关键点信息,将人脸位置标准化到一标准格式脸上。

一种人脸图像的身份标注方法及人脸身份识别方法

技术领域

[0001] 本发明涉及一种人脸识别方法,尤其涉及一种人脸图像的身份标注方法及人脸身份识别方法,属于图像识别技术领域。

背景技术

[0002] 目前人脸识别技术在各领域得到广泛应用,成为当前的一项研究热点,比如申请号 201210313721.9、名称“人脸识别方法”的专利文献,申请号 201210310643.7、名称“一种人脸识别方法及其系统”的专利文献。

[0003] 其中,人脸检测识别方法中人脸特征点的提取和标注是一项必不可少的工作,比如申请号 201310115471.2、名称“一种人脸自动标注方法及系统”首先从截取的视频中检测出人脸,获取人脸图片集合,然后过滤出人脸图片集合,同时,获取相邻帧图片的 HSV 颜色直方图差值,采用空间颜色直方图的镜头边缘检测算法进行镜头分割,对来自相邻帧的人脸,检测第一帧的目标区域内角点,并使用局部匹配的方法将这些角点递延给下一帧,并进行相应的更新,并统计匹配个数,依据匹配个数的阈值,依此进行下去获取人脸序列。然后通过唇动检测模块根据人脸序列中说话人的唇动检测出说话人和未说话人,将说话人、说话内容及说话时间三者融合进行标注;最后,读入每个序列上的人脸,逐个定位,再根据定位结果进行仿射变换,并提取变换后特征点附近固定大小圆形区域内的像素灰度值,作为该人脸特征。

[0004] 申请号 200610096709.1,名称“人脸识别系统中的人脸特征点定位方法”也涉及人脸识别系统中的人脸特征点定位方法,利用图像梯度方向信息的统计模型,通过统计推理的方法确定人脸特征点,包括以下步骤:(1)定义及定位人脸特征点,即利用图像梯度的方向定义和定位候选人的人脸特征点;(2)提取步骤(1)中人脸特征点的特征向量(3)利用一个考虑了人脸特征点的特征及相对关系的统计模型,采用统计推理的方法,标注人脸特征点,从而确定需要的人脸特征点的位置。

[0005] 人脸技术属于机器学习范畴,技术和系统都需要经历数据训练过程,即把大量人脸图像和相应的标注一起作为输入给到算法,算法会根据这些训练数据自动学习出相应的模型从而用于实际应用。由于目前的人脸检测方法所要求检测的特征属性信息要求越来越丰富,一般通过有标注的人脸图像利用机器学习算法进行训练得到识别模型,从而对众多未标注的人脸图像进行标注和识别。然而关于人脸身份的标注方法一直未得到有效解决,如果简单的通过人工方法去逐一筛选标注,非常耗时。

发明内容

[0006] 针对现有技术中存在的问题,本发明的目的在于提供一种人脸图像的身份标注方法及基于人脸图片大数据循环的识别方法。

[0007] 本发明的技术方案为:

[0008] 一种人脸图像的身份标注方法,其步骤为:

- [0009] 1) 从图像搜索引擎中搜索与待标注人脸图片相似的人脸图像及对应的网页；
- [0010] 2) 统计所返回网页中所出现人名的频率,并根据该频率初步确定该待标注人脸图片的身份；
- [0011] 3) 分别采用人脸技术平台和人脸身份识别模型检测该待标注人脸图片的身份；
- [0012] 4) 根据步骤 2)、3) 的识别结果确定该待标注人脸图片的最终身份,标注该待标注人脸图片的身份。
- [0013] 进一步的,首先提取待标注人脸图片的人脸位置和关键点信息,将人脸位置标准化到一标准格式脸上。
- [0014] 进一步的,所述人脸身份识别模型将搜索的属于同一人名的人脸图像两两比对图片中人脸的身份属于同一个人的置信度;然后根据置信度确定待标注人脸图片身份是否为该人名。
- [0015] 一种人脸图像的人脸身份识别方法,其步骤为:
- [0016] 1) 数据自动采集系统从服务器获取人脸图片及其上下文信息;
- [0017] 2) 数据自动标注系统对获取的每一待标注人脸图片的身份进行标注;其中标注方法为:
- [0018] 21) 从图像搜索引擎中搜索与待标注人脸图片相似的人脸图像及对应的网页;
- [0019] 22) 统计所返回网页中所出现人名的频率,并根据该频率初步确定该待标注人脸图片的身份;
- [0020] 23) 分别采用人脸技术平台和人脸身份识别模型检测该待标注人脸图片的身份;
- [0021] 24) 根据步骤 22)、23) 的识别结果确定该待标注人脸图片的最终身份,标注该待标注人脸图片的身份;
- [0022] 3) 将数据自动采集系统中采集的属于同一人名的一组图片与步骤 2) 中标注结果为该人名的图片进行匹配过滤,去除该组中不属于该人名的人脸图片;
- [0023] 4) 提取步骤 3) 过滤后保留的每一身份标注图片的特征向量,自动算法训练系统利用机器学习算法对身份标注后的人脸图片进行训练,生成一人脸身份识别模型;
- [0024] 5) 对于待检测的两人脸图像,提取其特征向量利用所述人脸身份识别模型判断其是否属于同一人。
- [0025] 进一步的,所述数据自动采集系统从服务器获取人脸图片及其上下文信息的方法为:
- [0026] 51) 所述服务器根据输入的人名关键字搜索相应的人脸图片文件并保存;
- [0027] 52) 计算每一人脸图片文件的哈希码、颜色直方图、上下文和标签信息;
- [0028] 53) 将每一人脸图片与已存人脸图片进行哈希码和颜色直方图比对,去除重复的图像;
- [0029] 54) 使用人脸检测算法模块检测步骤 53) 处理后保留的每一人脸图片,将人脸位置信息
- [0030] 保存到数据库;使用人脸关键点定位算法定位人脸上的关键点信息并保存到数据库。
- [0031] 进一步的,步骤 21) 之前先提取待标注人脸图片的人脸位置和关键点信息,将人脸位置标准化到一标准格式脸上。

[0032] 本发明检测系统如图 1 所述,其检测方法包含如下步骤:

[0033] 1) 数据自动采集系统,自动从搜索引擎,社交网络,和拍照片册类应用后台服务器不断挖掘学习算法所需要的人脸数据和相关上下文信息;

[0034] 2) 数据自动标注系统,通过少量人工干预,自动过滤采集数据中的噪声,并利用上下文信息自动挖掘学习算法所需要的人脸身份标注信息;

[0035] 3) 自动算法训练系统,在获得了自动挖掘出的人脸数据和身份标注信息,该系统定期自动的将数据送入算法学习系统进行算法训练,待训练完成后自动构建可执行算法模块;

[0036] 4) 3) 中所得到的最新的算法模块会循环进入 2) 的子系统,从而帮助更好地的人脸身份标注。

[0037] 与现有技术相比,本发明的积极效果为:

[0038] 本发明可以实现对人脸图像身份进行自动标注,大大提高了人脸图像标注的效率;本发明的人脸识别方法可以帮助充分利用大数据的优势,大大提升识别效果。

附图说明

[0039] 图 1. 总体系统示意图;

[0040] 图 2. 数据自动采集方法示意图;

[0041] 图 3. 数据自动标注方法示意图;

[0042] 图 4. 自动算法训练示意图。

具体实施方式

[0043] 下面结合附图对本发明的技术进行进一步详细描述。

[0044] 1) 数据自动采集系统(如图 2 所示)

[0045] 提升人脸技术各个技术环节算法性能的一个关键条件是获得较高质量的大规模人脸数据。传统方法是人工搭建采集环境,组织志愿者采集人脸图像,人工标注采集的人脸数据,比如人脸的图像位置,人脸关键点的图像坐标,人脸的年龄,身份等等。传统方法采集耗时,采集到的数据也非常单调,比如都是在一个地区的,或者某个年龄段的,某种光照条件下,某种人脸姿态的图像数据,其多样性的缺乏无法满足高性能的人脸技术的算法训练要求。搜索引擎和互联网的出现提供了大数据挖掘和利用的可能性,大量的名人数据可以非常快捷的通过人名关键字检索得到。同时社交网络和相册上也有大量的同一个人的人脸图像数据,这些都为提升人脸识别算法提供了丰富的人脸身份数据来源。如何利用这些数据提升算法性能也是一个目前有待研究的问题。

[0046] 针对上述问题,本方法使用如下步骤自动化的采集挖掘人脸数据:

[0047] 1. 系统在搜索引擎上搜索人名关键字,关键字库从各类百科数据中获得,比如体育,演艺明星的名字等等。

[0048] 2. 系统自动下载搜索引擎提供的结果图像文件,保存到一个临时文件系统中,文件中部分人脸图像数据是和检索的人名匹配的人脸图像,剩余的人脸图像并不属于该人名,需要采取数据自动标注系统,即系统 2) 中描述的步骤进行过滤。

[0049] 3. 计算步骤 2 中下载的图像文件的哈希码(例如使用 MD5 算法)和颜色直方图数

据和上下文和标签信息(如数据来源网站,时间戳,上下文关键字等),存入数据库,并建立索引。

[0050] 4. 对步骤3中得到的数据进行去重处理:每一张图片都要和数据库中的已经入库的图片进行哈希码和颜色直方图比对,去除重复的图像。

[0051] 5. 将步骤4中筛选后剩下的图片保存入一个持续的分布式文件系统。

[0052] 6. 使用人脸检测算法模块检测步骤5中保存的图像中的人脸,将人脸位置信息保存到数据库;使用人脸关键点定位算法模块定位人脸上的关键点信息并保存到数据库。

[0053] 7. 最终该系统产生一个存储了图像文件数据的分布式文件系统和一个保存有各种人脸和图像元信息的分布式数据库。

[0054] 2) 数据自动标注系统(如图3所示)

[0055] 1. 对于采集系统中产生的人脸图片,利用采集系统步骤6产生的人脸位置和关键点信息将人脸位置标准化到一个固定大小和位置保存成图像方便后续步骤4)中提取特征向量。

[0056] 2. 在第三方搜索图像搜索引擎中上传步骤1中产生的人脸图像,搜索得到相似的人脸图片和其对应的源网页连接。在得到的结果网页中分析人名关键字,统计各个人名关键字出现的频率,将其频率统计结果用如下公式转化为置信度(假设有M个人名,其中 p_i 表示第i个人名出现的频率, X_i 表示第i个人名是该人脸图像身份的置信度)。

[0057] $X_i = p_i / (p_1 + p_2 + \dots + p_M)$,

[0058] 3. 在第三方人脸技术API平台(参考<http://www.skybiometry.com/Demo>; <http://www.lambdal.com/>)中上传步骤1中产生的人脸图像,获取身份识别结果,提取返回结果列出的可能的人名身份(假设API返回有K个可能人名)和第三方人脸技术API平台提供的置信度。该置信度同样也是特定人名是该人脸图像身份的置信度。

[0059] 4. 将2,3的结果使用加权平均得到采集系统中下载的数据哪些和下载所用的人名关键字是匹配的,从而得到过滤后的图片数据。

[0060] 实验表明,本方法可以得到极为精准的人脸身份标注数据。性能结果见表1。我们列出了传统手动标注方法的性能数字作为对比。传统方法逐一检查下载图片中的每一个人脸图片是否匹配下载所用的人名关键字从而过滤出符合的图片。

[0061] 表1为身份标注效果对比表

[0062]

	手动标注	本发明提出的方法
标注准确度	99.4%	99.2%
平均每张图片标注时间	12秒	0.8秒

[0063] 3) 自动算法训练系统(如图4所示)

[0064] 在获得了标注系统2)过滤后的人脸图像数据,本系统提取每一身份标注图片的特征向量(可以使用公开领域中的LBP, Gabor, HOG等任意一种特征向量),自动算法训练系统利用机器学习算法定期对身份标注后的人脸图片进行训练,生成一人脸身份识别模型;然后将符合筛选条件的数据导入算法训练系统从而检测所输入的图片是否属于同一个人。其

具体步骤如下：

[0065] 1. 用户定期根据需求将需要的人脸数据量和筛选条件(比如图像均来源于 2013 年的互联网 10000 个名人搜索数据,每个人 50 张图片)录入一个任务队列数据库。

[0066] 2. 自动算法训练系统定时从任务队列数据库中读取任务。

[0067] 3. 系统将 2 中的图像和数据根据任务中的目标算法规格化成该算法训练所需要的存储格式。

[0068] 4. 系统将 3 中的规格化后的数据上传到学习训练服务器进行训练,训练目标是给定一对人脸图像数据算法模块输出这一对人是否是同一个人,生成一人脸身份识别模型;对于同一人名检索的人脸图像,提取其特征向量;然后利用所述人脸身份识别模型对其进行检测,识别出其是否属于同一人。

[0069] 本发明描述的基于大数据的自适应人脸机器学习算法训练系统可以用于人脸技术的各个模块,包含但不限于人脸检测,人脸关键点定位,人脸属性分性(性别,年龄,种族,表情等),和人脸身份识别。

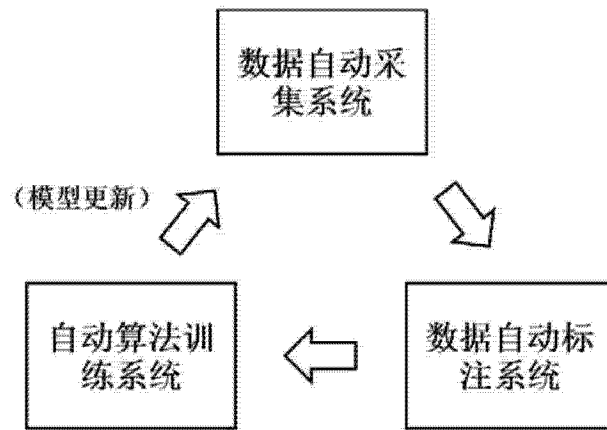


图 1

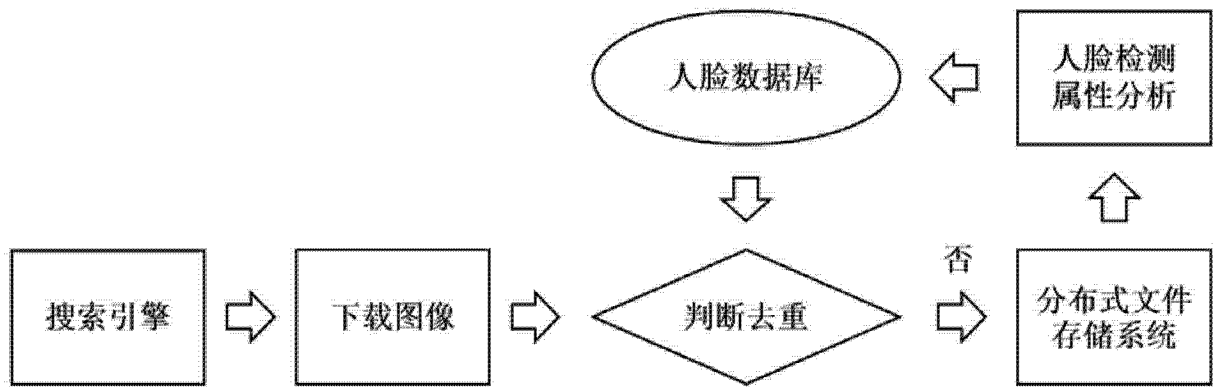


图 2

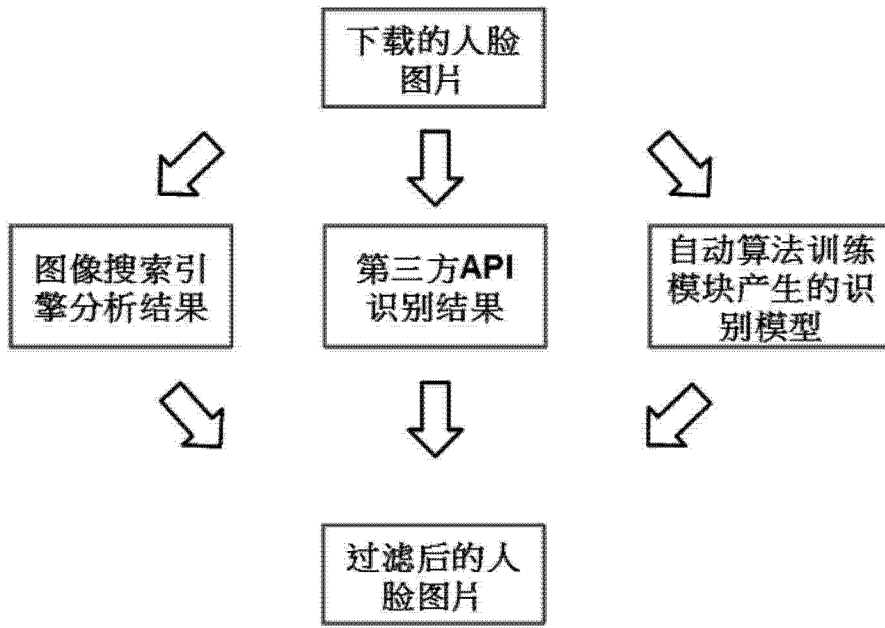


图 3

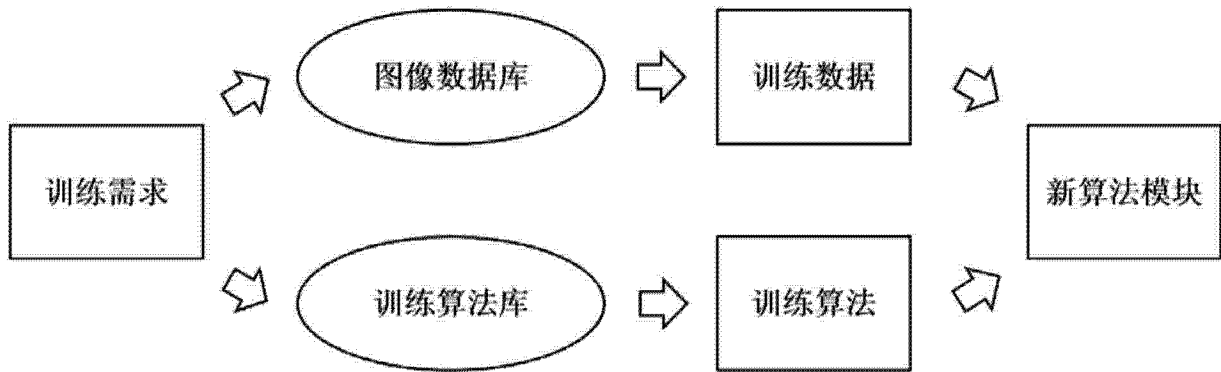


图 4