计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2021,30(5):59−65 [doi: 10.15888/j.cnki.csa.007889]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于唇语识别的身份认证研究及系统设计①

胡中坚[1], 冯　晗[1], 陈飞宇[2], 张文强[2]

[1](复旦大学 软件学院, 上海 201203)
[2](复旦大学 计算机科学技术学院, 上海 201203)
通讯作者: 张文强, E-mail: 805934132@qq.com

摘　要: 随着人脸识别身份认证技术的广泛应用, 各类针对人脸识别系统的攻击手段逐渐出现. 为了应对这类安全性问题, 提出了基于唇语识别的身份认证方法. 基于唇语识别的身份认证系统要求用户在进行人脸识别认证的同时读出验证码, 系统既要对人脸进行比对, 还要通过唇语识别技术识别出说话内容并与验证码进行比对, 只有两部分比对都通过才能通过系统的身份认证. 最后设计了基于唇语识别的身份认证系统, 主要包括前端、网关和后端.

关键词: 身份认证; 唇语识别; 人脸识别系统; CNN; LSTM

引用格式: 胡中坚,冯晗,陈飞宇,张文强.基于唇语识别的身份认证研究及系统设计.计算机系统应用,2021,30(5):59–65. http://www.c-s-a.org.cn/1003-3254/7889.html

Identity Authentication Research and System Design Based on Lip Reading Recognition

HU Zhong-Jian[1], FENG Han[1], CHEN Fei-Yu[2], ZHANG Wen-Qiang[2]

[1](School of Software, Fudan University, Shanghai 201203, China)
[2](School of Computer Science, Fudan University, Shanghai 201203, China)

Abstract: Amid the widespread application of face recognition technology for authentication, various attack methods against face recognition systems have emerged over time. An identity authentication system based on lip reading recognition is proposed to cope with such security issues. It requires users to read the verification code during face recognition for authentication. The system not only checks the face, but recognizes the content of the speech through the lip reading recognition technology to compare it with the verification code. The user's identity can be authenticated only when both the two links are passed. Finally, an identity authentication system based on lip reading recognition is designed, mainly including front-end, gateway, and back-end.

Key words: identity authentication; lip reading recognition; face recognition system; CNN; LSTM

随着科技的发展, 当今社会用户身份认证技术已经被广泛使用, 当下如何有效并且准确地对用户进行身份验证变得十分重要. 传统身份认证一般包括: 一通过密码等来确认访问者身份; 二通过其拥有的东西对访问者进行认证, 例如钥匙等. 然而随着技术的飞速发展, 传统身份认证的弊端逐渐显现出来. 例如, 密码容易遗忘, 钥匙容易丢失.

为了解决传统身份认证方法中存在的问题, 基于生物特征的身份认证方法逐渐被人们关注. 生物特征是人们与生俱来的特征, 其具有唯一性、稳定性、方便性等优点. 人脸识别是当今较为成熟的生物识别技术, 在门禁、安检等诸多领域广泛使用, 具有方便快捷、不易丢失等优势.

随着人脸识别技术的广泛应用, 人脸识别系统存在的一些弊端逐渐显露出来. 典型的欺骗手段包括照片欺骗、视频欺骗等. 为了解决这些问题, 可以在人脸

① 收稿时间: 2020-09-01; 修改时间: 2020-09-25; 采用时间: 2020-10-09; csa 在线出版时间: 2021-04-28

识别系统中引入唇语识别技术, 打造高安全性的身份认证系统, 以防不法分子的攻击.

1 身份认证相关研究

随着技术的发展, 国内外出现了大量关于身份认证方法的研究. 广义上的身份认证在日常生活中十分常见, 包括在车站内查身份证等其实都是属于身份认证. 计算机领域的身份认证常见的包括基于口令的身份认证和基于生物特征的身份认证等[1,2].

文献 [3] 中提到了一种基于口令的身份认证方案. 该方案提到口令加盐的方法增强安全性, 盐就是一个字母数字组合的字符串. 数据库中存储用户标识 userid 以及对应的盐值 salt, 加盐 hash 后的口令 hash (passwd+salt). 验证阶段, 接收到用户标识 userid 和密码 passwd, 先到数据库中找到 userid 对应的盐值 salt, 如果存在该用户, 接下来计算加盐口令 hash (passwd+salt), 将计算得到的加盐口令与数据库里存储的加盐口令进行比较, 如果相同, 则认证成功, 否则, 认证失败[3].

基于口令的身份认证方案是较为常见的一种身份认证方案. 但该方案有一定的缺点, 比如口令容易遗忘, 被盗用等.

文献 [4] 提到了一种人脸反欺诈认证系统架构. 主要包括人脸检测对齐, 活体检测和人脸比对匹配. 如图 1 所示.

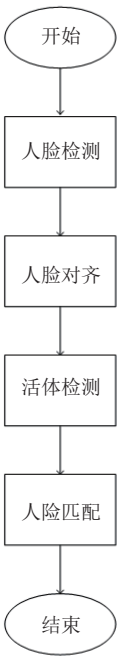

图 1 人脸识别身份认证的流程

通过摄像头获取用户人脸, 接下来人脸检测和对齐, 输入到活体检测模块进行检验, 最后是人脸比对匹配. 人脸检测对齐使用 MTCNN 网络, 该网络包括: P-Net (Proposal Network)、R-Net (Refine Network)、O-Net (Output Network). 活体检测讨论了基于色彩纹理分析和轻量级卷积神经网络两种方法[4]. 人脸识别比对利用 FaceNet 网络[5,6].

随着人脸识别和活体检测技术在诸多领域被广泛使用, 其逐渐暴露出安全性问题. 照片和视频是常见的攻击手段[7]. 照片欺骗是指非法获取合法用户的人脸照片攻击人脸识别系统. 为了解决照片欺骗等问题, 研究人员在人脸识别系统中引入了活体检测. 然而不法分子又通过录制脸部视频攻击人脸识别系统. 除了偷拍, 也可以在各类短视频 APP 里找到用户本人上传的视频, 还可以在各类直播平台里获取. 识别视频攻击比识别照片攻击更加困难, 现在有些软件甚至可以将人脸照片转换成活体视频! 还有一种是合成人脸的三维模型, 这种难度和成本很高, 一般很少见.

2 基于唇语识别的身份认证

2.1 基于唇语识别的身份认证方案

基于唇语识别的身份认证系统要求用户在对着摄像头进行人脸识别的同时读出指定的验证码. 系统在验证用户身份时, 不仅要对人脸识别比对, 还要通过唇语识别技术识别出视频里用户说话的内容, 将唇语识别结果和验证码进行比较. 只有人脸识别比对和唇语识别比对都通过才能通过系统的认证. 验证码是随机生成, 不法分子很难及时获得唇动特征与验证码唇语一致的视频[8,9].

如图 2 所示, 基于唇语识别的身份认证方案主要包括两大模块. 人脸识别模块负责用人脸识别 1:1 技术比对, 判断人脸照片和身份证照片是否匹配. 活体检测模块负责用唇语识别技术识别出视频里用户的说话内容, 并比对唇语识别结果和系统的验证码. 只有两大模块的比对验证都通过, 该用户才能通过身份认证.

2.2 基于深度学习的唇语识别

如图 3 所示, 基于深度学习 CNN+LSTM 的唇语识别技术[10], 首先用 CNN 卷积神经网络提取基于关键点的嘴部特征, 然后将 CNN 提取的特征信息输入 LSTM, 输出是唇语视频中的内容.

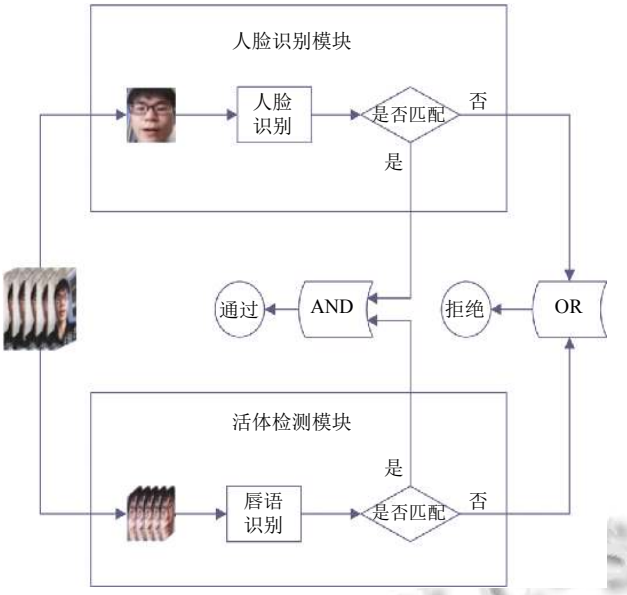

图 2 基于唇语识别的身份认证流程图

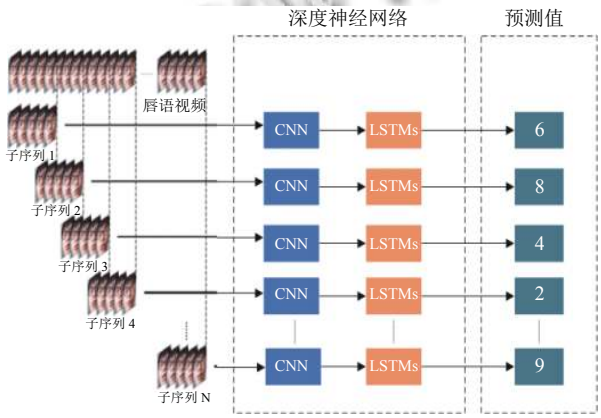

图 3 基于深度学习的唇语识别

在自然语言处理中很少会对单独的单词来进行识别和处理, 一般要考虑句子中的单词有上下文的语义关联. 本文的唇语识别研究的是识别随机验证码, 验证码通常是单独的一串数字或者字母等, 并不存在上下文的语义关联. 本文的唇语识别中需要先将唇语视频序列切分成若干个子序列, 然后将这些子序列再输入到 CNN 卷积神经网络中进行处理. CNN 提取出唇语视频中这些子序列的唇动特征, 再将特征信息输入到 LSTM 网络中, 对提取出的唇动特征进行编码, 最后还会用 Softmax 来对唇语视频里的说话内容进行预判[11].

2.2.1 基于 CNN 的关键点检测

唇语识别技术依赖说话者的嘴部动作变化, 唇动变化可以通过嘴部关键点的唇动特征来表示. 如图 4 所示, 将关键点提高到 68 个, 嘴唇关键点数量就是 20 个.

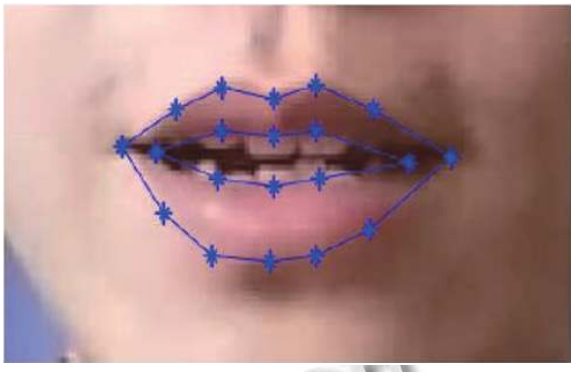

图 4 嘴唇关键点

在嘴唇上检测 20 个关键点, 其中外嘴唇关键点有 12 个, 内嘴唇关键点有 8 个. 通过每帧图像里的唇部关键点信息记录视频序列中的嘴部运动[11].

随着深度学习的发展, AlexNet[12] 网络在 ILSVRC12 比赛上获得了冠军, 文中使用 AlexNet 网络检测嘴唇关键点. AlexNet 有 3 层全连接层和 5 层卷积层. 采用 CASIA_WebFace 数据集预训练该模型, CASIA_WebFace 数据集[13] 包含 10 575 个人, 共 494 414 张人脸图片. 完成预训练后, 将 AlexNet 网络最后一层中 10 575 类的分类器替换为 40 分类的分类器, 用数据集再次训练. 再次训练的数据集, 照片的人脸有 68 个关键点, 嘴唇有 20 个关键点. 欧氏距离用作训练的损失函数[14].

训练人脸关键点模型时, 根据检测的区域信息 $\{x, y, w, h\}$, 规范化关键点坐标[11]:

$$(\hat{x}_i, \hat{y}_i) = \left(\frac{x_i - x}{w}, \frac{y_i - y}{h}\right) \tag{1}$$

其中, (x, y) 是检测到的人脸中心点坐标, w 是人脸区域的宽度, h 是人脸区域的高度, (x_i, y_i), $i = 1, 2, \cdots, 20$ 是标准关键点坐标, $(\hat{x}_i, \hat{y}_i)$, $i = 1, 2, \cdots, 20$ 为规范化后的关键点坐标[11].

2.2.2 基于 LSTM 的唇语识别

用 CNN 提取出特征向量输入 LSTM 网络, 输出唇语视频中的内容. 已知输入 x_t, t 时刻的记忆状态向量 C_t, 前一个隐藏层状态 h_{t-1}, 可计算当前隐藏层状态 h_t, 计算如下:

$$i_t = sigm(W_{xi}x_t + W_{hi}h_{t-1} + b_i) \tag{2}$$

$$f_t = sigm\left(W_{xf}x_t + W_{hf}h_{t-1} + b_f\right) \tag{3}$$

$$o_t = sigm(W_{xo}x_t + W_{ho}h_{t-1} + b_o) \tag{4}$$

$$g_t = \tanh\left(W_{xg}x_t + W_{hg}h_{t-1} + b_g\right) \tag{5}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes g_t \tag{6}$$

$$h_t = o_t \otimes \tanh(c_t) \tag{7}$$

其中, *sigm* 表示非线性的 Sigmoid 函数, 符号⊗表示点乘运算, 非线性的双曲正切函数是tanh, w_{ij}与b_j是训练得到的参数[14].

每个序列输出的条件概率都有对应的输入序列, 概率计算如下[14]:

$$p(y \mid x_1, x_2, \cdots, x_n) = p\left(y \mid h_n^l\right) \tag{8}$$

$$p\left(y|h_n^l\right) = \frac{\exp\left(W_y h_n^l\right)}{\sum_{y' \in v} \exp\left(W_{y'} h_n^l\right)} \tag{9}$$

在该模型中, 分类层用的是 Softmax, $p\left(y \mid h_n^l\right)$是 Softmax 回归模型的概率分布, 其用来表示h_n^l属于哪一类的概率, h_n^l是最后一个 LSTM 层对第 n 帧编码后的隐藏状态[14].

如图 5 所示, 在训练的阶段, 将 CNN 的学习率置为 0, 用 CNN 提取出特征向量, 输入 LSTM 网络, 预测结果由 LSTM 输出. 损失函数如下[14]:

$$J(\theta) = -\sum_{j=1}^{k} l\{y = j\} \log p\left(y \mid h_n^l\right) \tag{10}$$

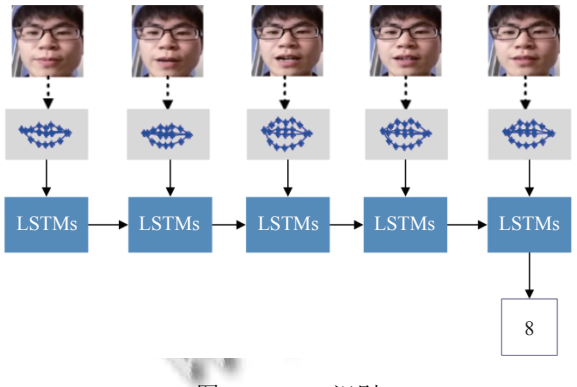

图 5 LSTM 识别

最后在 LSTM 网络的最后一层引入 dropout 层可以避免过度拟合问题, 提高识别效果[11,15].

2.2.3 实验分析

(1) 数字串实验

本文提出了一种基于深度学习 CNN+LSTM 的唇语识别技术. 由于身份认证系统的验证码一般为随机生成的数字串, 为了评估该方法在系统中的实际应用价值, 将对其在数字串数据集上进行实验. 而目前公开的唇语数据集大部分是单词短语, 并且是外文发音, 无法满足我们的需求. 故实验数据集使用公司自主创建的数字串数据集, 该数据集中的唇语视频都是收录的说话人录制的 6 个数字的组合, 其中 80% 作为训练集, 其余作为测试集.

实验结果如表 1 所示. 表中 pre *a* 和 err *b* 代表选取前 *a* 个预测结果的时候, 识别错误的个数小于等于 *b* 个时的准确率. 观察表中数据, 取 pre 1 且全部识别正确时的准确率只有 49.6%. 分析数据发现, 表中 err 1 比 err 0 的准确率要高, 这说明当允许出错的个数增多时, 准确率会有一定提升, 同时 pre 1 到 pre 4 的准确率也有所提升, 这说明取多个预测结果往往比单纯的取一个预测结果更加准确. 在实际应用场景中, 根据实际需求的不同, 可以通过设置不同的 *a* 和 *b* 以提高准确率.

表 1 数字串唇语识别准确率 (%)

参数	pre 1	pre 2	pre 3	pre 4
err 0	49.6	51.1	84.2	89.2
err 1	79.1	81.3	95.7	96.4
err 2	92.8	93.5	99.3	99.3

(2) OuIuVS 数据集上的对比

为了验证本文 CNN+LSTM 方法的泛化能力, 继续在公开的数据集 OuIuVS 上进行验证. OuIuVS 数据集中的视频共有 20 个人. 每个人读 10 个短语, 每个短语读 5 遍.

将本文中的方法与其他有关文献里的唇语识别方法在 OuIuVS 上进行比较. 观察表 2 中的数据, 我们可以发现本文中使用的方法效果更好. 本文方法在短语识别的平均准确率约为 81.9%.

表 2 OuIuVS 上的结果对比

方法	短语平均准确率(%)
文献[16]	58.6
文献[17]	62.3
本文	81.9

虽然高精度唇语识别很困难, 但唇语识别仍具有较大的研究价值, 尤其在特定的应用场景, 通过训练和调整, 唇语识别能有较好的表现.

3 基于唇语识别的身份认证系统分析设计

本系统需要完成从采集用户身份证信息和用户认证视频到用户身份认证的全流程. 采集用户身份证信息和用户认证视频的功能一般由前端负责提供, 显然

前端需要部署在外网, 以便用户可以使用. 后端负责对采集的信息和视频进行处理, 即后端需要进行人脸识别验证和唇语识别验证, 出于安全性考虑, 后端一般部署在内网. 而网关作为前后端的桥梁, 也是不可或缺的.

3.1 需求分析

(1) 前端

前端一般是 APP 端或者网页端形式. 如图 6 所示, 用户打开 APP, 输入自己的姓名, 身份证号, 再对着摄像头读出验证码, 即可提交身份认证请求.

图 6 用户操作流程

前端包括信息录入和视频录制, 信息和视频存储, 验证码生成等功能. 用户输入信息后, 前端负责保存到相应的数据库里. 然后需要随机生成验证码供用户录制视频时读, 并将用户录制的视频保存到文件存储平台. 在用户点提交按钮时向后端发送请求, 获取身份认证结果展现给用户.

(2) 网关

网关是连接前后端的桥梁, 为了使前后端能正常配合提供完整功能, 并且出于外网与内网交互的安全性考虑, 提供网关这一关口是非常有必要的.

网关包括限流、加解密、鉴权验签、请求路由等功能. 为保护后端系统, 网关需提供限流功能, 防止大并发. 由于前端处于外网, 故数据需加密通信. 为了确保请求未被篡改并且是合法的, 需要进行验签鉴权. 同时网关需要将合法请求转发给后端服务.

(3) 后端

如图 7, 后端包括人脸识别和唇语识别两大能力.

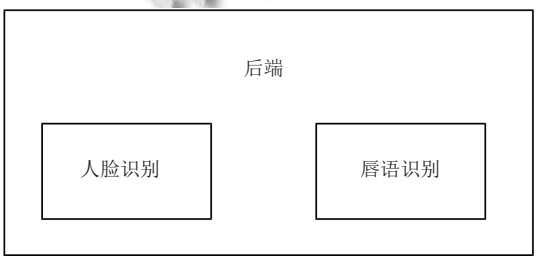

图 7 后端功能

后端接收到前端的身份认证请求后, 需要对视频里的用户进行人脸识别比对和唇语识别比对, 以判断是否允许该用户通过系统的身份认证, 并将身份认证的结果通过网关返回给前端.

3.2 系统设计

如图 8, 本系统主要包含前端、网关、后端 3 个子系统组成, 前端和后端通过网关进行交互.

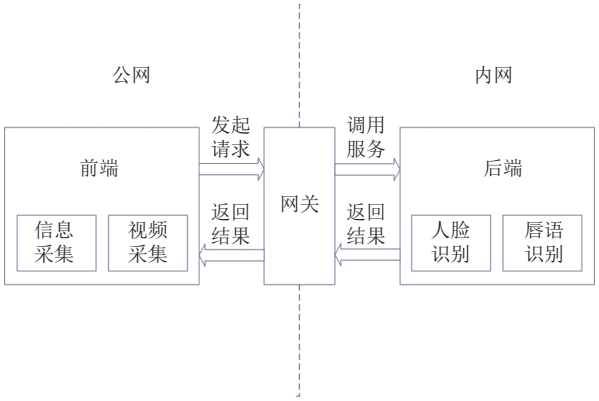

图 8 系统部署

前端一般是 APP 端和网页端, 负责采集用户信息和用户视频. 用户信息包括用户的身份证信息, 例如姓名、身份证号等, 采集用户视频是指采集用户录制的用于身份认证的视频. 前端需要部署在公网, 即用户可以在外网直接使用. 后端负责接收前端身份认证的请求并对请求进行处理响应, 包括人脸识别 1:1 比对及唇语识别结果比对, 后端通常在内网部署. 网关可用于前端和后端之间的交互.

(1) 前端

前端主要是负责采集用户的身份证信息和用户录制的视频. 采集包括姓名和身份证号, 以及包含人脸用于身份认证的视频, 然后保存, 通过网关往后端发送身份认证请求, 获取后端身份认证处理的返回结果. 前端一般是 APP 端或者网页端的形式. 系统采用前后端分离的架构.

图 9 是前端流程图. 前端提供的信息输入功能允许用户填写个人身份证信息. 用户填写完毕后, 单击界面按钮, 系统会将用户身份证信息存储在数据库中. 前端提供的视频捕获功能是捕获用户通过摄像头录制的视频, 并将用户视频保存到文件存储平台. 文件存储平台将返回与该文件相对应的文件标识 id. 通常, 视频文件相对较大, 并且在网络上的传输占用带宽.

因此, 通常将视频存储在文件存储平台中, 并获得对应的文件标识 id. 需要获取此文件时, 只需传递相应

的文件标识 id 即可下载. 在与文件存储平台的交互过程中, 一般将文件进行 Base64 编码后传输. 收集用户信息和视频后, 通过网关将身份认证请求发送给后端服务, 请求参数包括验证码, 用户身份证信息和用户视频文件标识. 最后获取后端身份认证处理的返回结果.

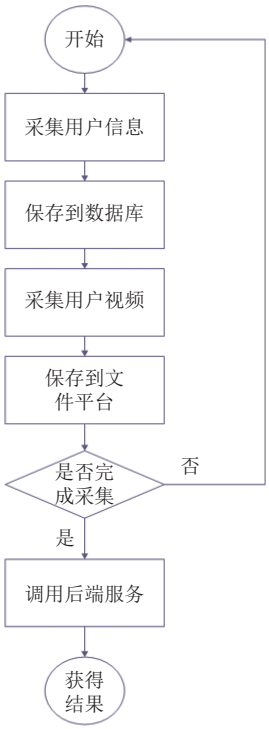

图 9 前端流程

(2) 网关

网关是公网系统和内网系统之间的关口. 前端通过网关访问后端. 网关提供一系列功能, 例如限流、鉴权、验签、加密解密以及请求路由等.

图 10 是网关系统的流程图. 前端系统将服务请求发送到网关系统. 首先, 网关系统执行限流判断, 例如当前请求限制为 100 个/s, 则 1 秒钟内超过 100 个请求之后的剩余请求将被拒绝. 如果满足当前的限流拒绝条件, 则拒绝请求. 如果不满足当前的限流拒绝条件, 则网关需要处理请求. 网关解密请求的数据后, 首先需要验证签名. 采用 MD5 加盐验签提高安全性. 然后是鉴权, 需要验证请求是否具有调用相关服务的权限. 判断后, 如果是合法请求, 则处理该请求, 例如请求相应的后端系统处理, 并获取后端系统处理的返回结果. 最后, 对获得的处理结果进行加密和签名, 返回到前端系统.

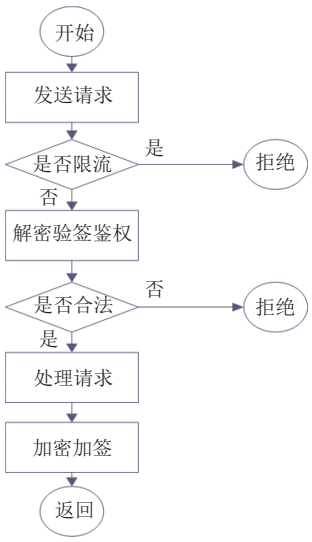

图 10 网关流程

(3) 后端

后端是实际执行身份认证处理的子系统, 主要包括两个模块: 人脸比对和基于唇语识别的活体检测. 只有两项验证均通过, 才能通过身份验证. 图 11 是后端的流程图.

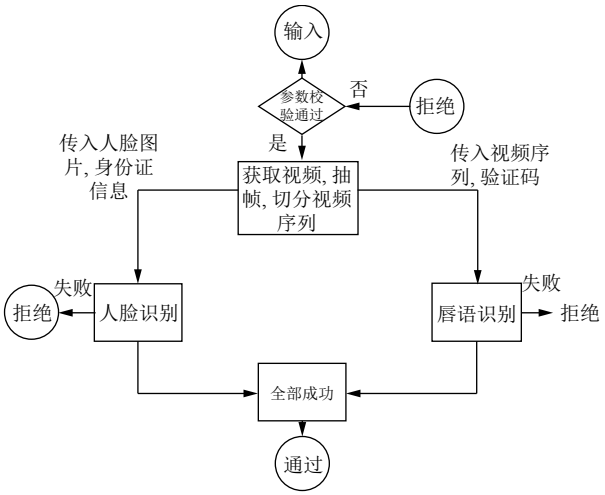

图 11 后端流程

前端发送身份认证请求给后端. 输入参数包括用户身份证信息, 随机唇语验证码, 用户视频文件标识 id 等. 随机验证码一般由前端自行生成. 后端收到请求后, 检查必需的参数, 如果参数检验失败, 则返回参数错误. 如果通过了参数校验, 则进行后续处理. 通过视频文件标识 id 从文件存储平台获取视频文件. 从视频中提取帧, 并将视频划分为若干子序列. 将抽取的人脸

图片和用户身份证信息传输到人脸识别模块, 该模块对接公民身份证信息系统或其服务提供商系统[9], 通过传入身份证信息获得用户的身份证照片, 再使用人脸识别比对技术, 以确定人脸图像是否与用户的身份证照片一致. 将视频分割后的若干子序列传给唇语识别模块, 使用唇语识别技术识别视频中用户说话的内容, 然后比对唇语识别结果和验证码以确定用户是否通过唇语验证. 如果人脸识别比对和唇语结果比对均通过, 则用户通过身份认证, 否则用户无法通过身份认证.

4 结束语

文中首先分析了常见的身份认证方案的不足之处, 然后提出基于唇语识别的身份认证方案. 基于唇语识别的身份认证方案包括人脸识别和唇语识别两大部分, 只有这两大部分的比对校验都通过才能通过身份认证. 其中唇语识别采用基于深度学习 CNN+LSTM 的唇语识别技术. 最后分别设计了基于唇语识别的身份认证系统的前端、网关和后端.

参考文献

1 郭志达, 王鑫, 李金宇. 基于区块链应用模式的铁路旅客身份认证系统. 计算机系统应用, 2019, 28(11): 63–71. [doi: 10.15888/j.cnki.csa.007136]

2 黄君浩, 贺辉. 基于 LSTM 的眼动行为识别及人机交互应用. 计算机系统应用, 2020, 29(3): 206–212. [doi: 10.15888/j.cnki.csa.007388]

3 徐军. 基于口令的身份认证方案安全性分析及其改进. 山东理工大学学报 (自然科学版), 2019, 33(3): 19–22. [doi: 10.13367/j.cnki.sdgc.2019.03.004]

4 陈放, 刘晓瑞, 杨明业. 基于活体检测和身份认证的人脸识别安防系统. 计算机应用, 2020, 40(12): 3666–3672.

5 黑富郁, 王景中, 赵林浩. 基于 CNN 和 LSTM 的异构数据舆情分类方法. 计算机系统应用, 2019, 28(6): 141–147. [doi: 10.15888/j.cnki.csa.006900]

6 何伟鑫, 邓建球, 方轶, 等. 基于改进 CNN 的部队门禁系统. 计算机系统应用, 2020, 29(6): 126–131. [doi: 10.15888/j.cnki.csa.007453]

7 曹瑜, 涂玲, 毋立芳. 身份认证中灰度共生矩阵和小波分析的活体人脸检测算法. 信号处理, 2014, 30(7): 830–835. [doi: 10.3969/j.issn.1003-0530.2014.07.011]

8 Zhou ZH, Zhao GY, Pietikäinen M. Towards a practical lipreading system. Proceedings of 2011 Conference on Computer Vision and Pattern Recognition. Providence, RI, USA. 2011. 137–144. [doi: 10.1109/CVPR.2011.5995345]

9 Venugopalan S, Rohrbach M, Donahue J, *et al*. Sequence to sequence —video to text. Proceedings of 2015 IEEE International Conference on Computer Vision. Santiago, Chile. 2015. 4534–4542.

10 任玉强, 田国栋, 周祥东, 等. 高安全性人脸识别系统中的唇语识别算法研究. 计算机应用研究, 2017, 34(4): 1221–1225, 1230. [doi: 10.3969/j.issn.1001-3695.2017.04.060]

11 任玉强. 高安全性人脸识别身份认证系统中的唇语识别算法研究 [硕士学位论文]. 重庆: 中国科学院重庆绿色智能技术研究院, 2016.

12 Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Proceedings of the 25th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2012. 1097–1105.

13 Yi D, Lei Z, Liao SC, *et al*. Learning face representation from scratch. arXiv: 1411.7923, 2014.

14 吴伟. 基于深度学习的唇语识别研究 [硕士学位论文]. 哈尔滨: 哈尔滨理工大学, 2019.

15 刘坤. 基于人脸识别的身份认证系统的设计与开发 [硕士学位论文]. 保定: 河北大学, 2017.

16 Zhao GY, Barnard M, Pietikainen M. Lipreading with local spatiotemporal descriptors. IEEE Transactions on Multimedia, 2009, 11(7): 1254–1265. [doi: 10.1109/TMM.2009.2030637]

17 Bakry A, Elgammal A. MKPLS: Manifold kernel partial least squares for lipreading and speaker identification. Proceedings of 2013 IEEE Conference on Computer Vision and Pattern Recognition. Portland, OR, USA. 2013. 684–691. [doi: 10.1109/CVPR.2013.94]