基于视觉的人脸+手势识别原型研究

罗强

1. 背景

近年来，随着计算机视觉和机器学习的进步，人机交互技术也逐渐成熟，在人们的日常生活特别是金融领域逐步发挥着越来越多的作用。在这些交互信息中，人脸和手势具有直观性、自然性和丰富性等特点，其识别研究引起了人们的极大关注。

人脸识别是一种利用机器自动识别人脸图像以判断用户身份的技术，其研究涉及模式识别、计算机视觉、人工智能、图像处理、心理学等，与计算机人机交互领域有着密切的联系，是模式识别和计算机领域一个非常活跃的研究热点。手势识别是一种肢体智能感知技术，该技术以计算机视觉为基础，识别出手势表示的含义，从而触发机器智能产生下一步响应，实现人机自然交互。人脸识别和手势识别分属于生物识别技术的不同应用领域，前者主要作用于身份识别，后者主要作用于智能感知。

在金融领域，人脸+手势识别技术能够实现基于身份认证的人机交互，具有广泛的市场需求。例如：在转账支付场景中，银行系统可以为用户提供个性化的人脸+手势注册服务，用户只要在智能终端经过人脸识别认证并预留表示特定交易含义的手势信息，后续再次转账时，就可以降低过多的交易信息录入和人机交互，为用户带来安全而

快捷的体验。

为此，本文提出了一种基于视觉的人脸+手势识别技术，该技术捕捉视频中的人脸和手势图像，精准分割人脸和手势图像，然后采用基于 CNN（卷积神经网络）的深度学习网络训练和识别方法，提取图像中人脸身份及手势特征信息，在完成人脸认证的同时实现手势识别，有效提升图像中感知信息比对效率。基于本研究搭建的视觉人脸+手势原型为人机交互研究打下基础，能够广泛应用于金融、社会治安、教育、医疗、娱乐等多个领域，具有巨大的市场价值和社会价值。

2. 视觉人脸和手势识别概述

视觉手势识 别（ＶｉｓｉｏｎＧｅｓｔｕｒｅＲｅｃｏｇｎｉｔｉｏｎ）是人 机 交 互技术研究的一个分支领域，是机器视觉领域的一个重要的研究内容，其发展状况与人机交互技术紧密相连。成熟 的 视觉手势识别产品已在世界范围内被广泛应用，典型的消费类电子产 品 有：２００３年，Ｓｏｎｙ公司推出了一款名为 ＥｙｅＴｏｙ的手势识别设备，这种设备能将玩家的动作传输到游戏画面，使玩家互动；２０１０年１１月，微软公司推出的 Ｋｉｎｅｃｔ体感设备在手势跟踪与识别方面有着出色的表现，它能实时识别用户手势，结合 Ｘｂｏｘ，使用户完成对游戏的控制指令；２０１２年，三星推出的智能电视 ＥＳ８０００，结合用户的手势，可以 对 电视进行换台、搜台以及音量调节等操作。2017 年，大疆推出带人脸+手势识别的无人机 Spark，能够让用户体验手势控制 Spark 起飞、悬

停、下降等功能。

国内对手势识别的研究较晚，但取得的成果较显著，比较具有代表性的主要有：清华大学祝远新等提出了一种基于表观的新的手势识别技术。该课题组通过结合手势的运动表观、形状表观和时序信息建立了动态手势的时空表观模型。为抽取时空表观模型的参数，提出了基于运动、形状和颜色等多模式信息分层融合的策略。而且建立的实验系统可对１２种手势进行在线识别，识别率超过９０％。北方交通大学的王延江等人提取手势轨迹中关键点的运动方向，将之与标准手势中所有可能的特征码进行匹配，从而实现识别手势轨迹。中科院软件所的王西颖等结合ＨＭＭ与模糊神经网络提出了一种基于ＨＭＭ－ＦＮＮ模型的结构，能够识别出复杂背景下的动态手势。上海交通大学的刘江华等通过跟踪双手的运动识别的动态手势，采用光流法和耦合隐马尔科夫模型，所能达到的识别率为９６．７％。张习文和王西颖等利用一组二维手势模型来替代三维模型，其过程是先利用贝叶斯分类器对静态手势进行识别，然后动态跟踪图像中的手指和指尖。由于该方法结合了基于模型和表观方法的特点，因此大大地减少了计算量。中科院自动化所的方亦凯提出一种快速的尺度空间特征检测方法，通过对手势图像中的Ｂｌｏｂ和Ｒｉｄｇｅ结构的检测，得到手掌与手指的结构描述，进而完成手势识别。

综上，随着人机交互和人工智能的发展和不断成熟，人脸+手势识别正从2D技术走向3D技术，从模式识别技术迈向深度学习技术，

并大量采用基于神经网络的深度学习算法，多模态多因子相互融合的趋势也越来越明显，未来人机交互的发展方向已然显现。

3. 基于视觉的人脸+手势识别设计

本文构建了一种基于 CNN（卷积神经网络）的视觉人脸+手势识别原型系统，该系统从视频中截取人脸和手势图片，然后采用图像分割技术分离人脸和手势图像，接着分别送入人脸处理模块和手势处理模块进行识别处理，最后输出用户身份和手势信息。人脸和手势特征分析和比对的过程运用了基于 CNN 的机器学习训练模型，具有识别精度高、识别速度快、使用方便的特点。

3.1. 总体思路

基于视觉的人脸+手势识别分为图像处理、特征分析、特征识别 3 个步骤：首先，输入图像经人脸分割和手势分割进行分离，前者定位出人脸，后者定位出动态手势；然 后，根 据 需 求 选 择人脸模型和手势模型进行手势分析，并依据模型提取手势参数；最后，根据模型参数对手势选择合适的算法进行人脸识别和手势识别，具体流程如图 1 所示。

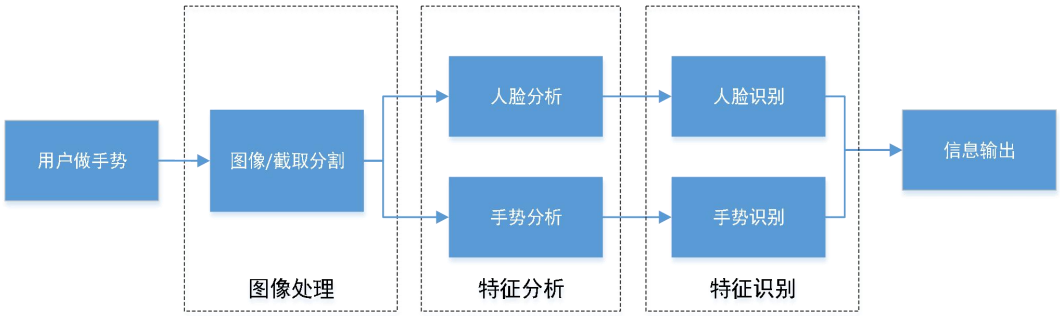
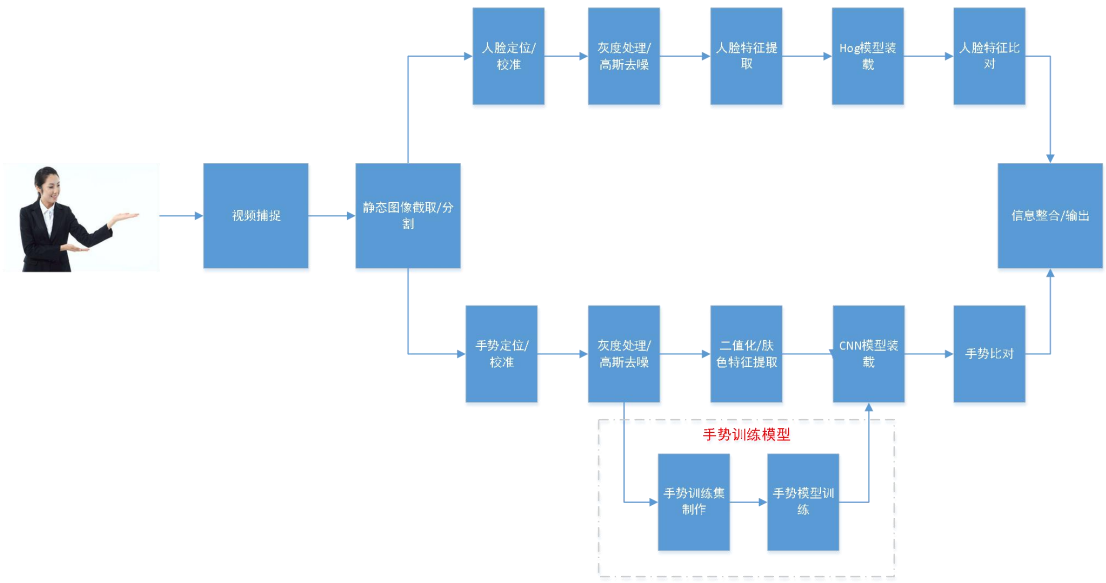


图 1 基于视觉人脸+手势总体结构图

图 1 中，图像处理步骤主要实现对视频的图片进行截取，并分割出人脸区域和手势区域；特征分析步骤是人脸和手势识别的核心流程，主要实现人脸和手势的建模，包括图像校准、特征提取、训练、模型构建等；特征识别步骤是模型的应用流程，主要是基于训练好的 CNN 模型对特征进行分类，从而获得匹配度最高的手势的过程。

3.2. 系统设计

基于视觉的人脸+手势识别原型的系统框架如图 2 所示，图像的采集采取视频实时截取方式，获取的场景中深度信息不受物体自身的颜色、纹理特征以及背景环境光线强弱的影响。本文选用笔记本电脑（Dell vostro 5471）内置机载摄像头进行摄制，视频录制质量为 720p，比例 16：9，30fps，刷新频率 60Hz。



基于视觉的人脸+手势系统结构

3.2.1. 人脸手势区域截取分割

在人机交互过程中，手势一般放置于身体位置之前，为了避免与身体其它部位（如人脸）的肤色重合，影响手势分割的效果，在视频显示区域专门开辟一个边长为 200 像素的正方形区域，用户需要在此区域展现手势以进行手势识别和手势训练。如下左图所示。

同一深度的像素点在深度图像中灰度值相同，但是每次手势展现和摄像头的之间的距离都不完全相同，无法用固定深度阈值实现区域的分割。本文提供了高斯模糊和肤色模型两种方法来寻找手势区域和背景的区分阈值。

（1）**高斯模糊**：高斯滤波的模板如下[2]：

$$G_{ij}=\alpha*e^{-\frac{(i-u_i)^2}{2*\sigma_1^2}-\frac{(j-u_j)^2}{2*\sigma_2^2}}$$

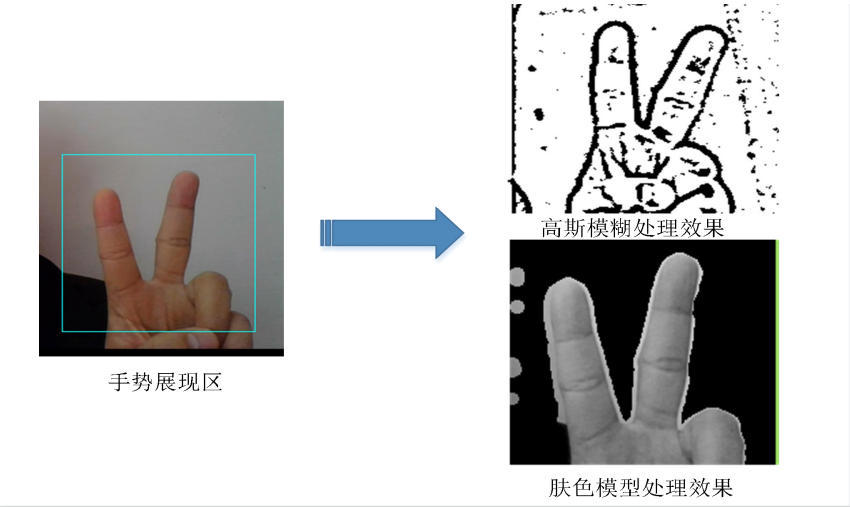
其中，i，j 表示 卷积核的坐标，$G_{i,j}$ 表示（i，j）位置的像数值，$\sigma_1\sigma_2$ 分别表示高斯滤波在横轴和纵轴的滤波参数。为了计算方便，设置 $\sigma_1\sigma_2$ 相同，且 $u_i=u_j=0$，公式简化为：

$$G_{ij}=\alpha*e^{-\frac{i^2+j^2}{2*\sigma}}$$

本研究对图像进行灰度处理后，设置高斯模糊的卷积核为 5*5，σ=2 ，处理效果如图右上图所示。

（2）肤色模型：通过定义肤色取值的 HSV（H：色调 S：饱和度 V：明度）范围，过滤复杂背景的干扰色，以达到提取手势的效果。

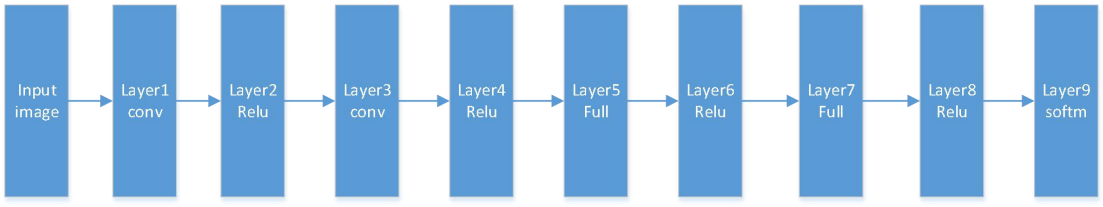
本研究取 HSV 的范围参数为：（0，50，80）~（30，200，255）。具体处理步骤如下：首先对原始图像 roi 进行腐蚀和膨胀操作，得到处理结果 mask，然后对 mask 采用高斯模糊，设置高斯模糊的卷积核为 15*15，σ=1，最后运用 mask 对 roi 进行与操作，实现背景干扰色的过滤。处理效果如图右下图所示。



图：手势展现区和手势预览区

3.2.2. CNN 模型结构

卷积神经网络是一个层级结构，主要包含输入层、卷积层、池化层、全连接层以及输出层。网络一般使用多个卷积层和池化层组合，在末端使用多层全连接的前馈神经网络，训练过程使用反向传播算法，本研究搭建的 CNN 模型结构如下图所示。



图：手势识别的卷积神经网络结构

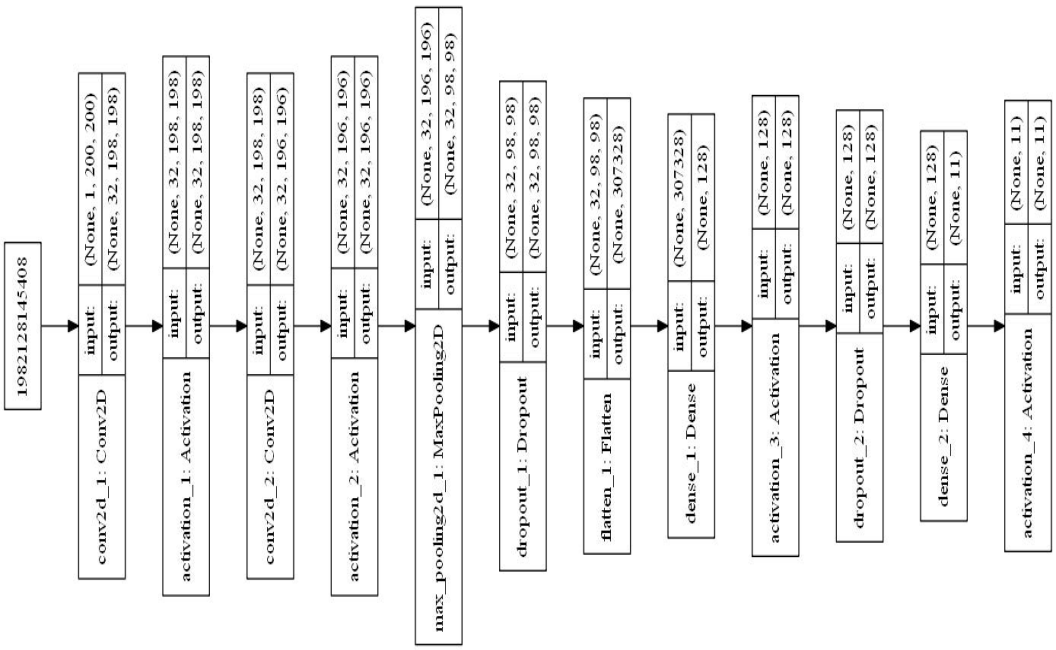
该结构共有 10 层，Input Image 为输入层，具体为 200×200 的手势特征融合图像，Layer1－Layer4 为卷积层，Layer5－Layer8 为全连接层，Layer9 softmax 为分类和输出层，输出层神经元有 3 个，分别代表手势类别: one（数字 1）、two（数字 2）、five（数字 5）、good（数字 10）。卷积核和各偏置等参数的初始值均随机产生，输入样本后通过前向传播和反向传播算法对网络进行训练来更新参数。

卷积滤波实质就是用卷积核在图像矩阵中滑动遍历，卷积核与图像上相对位置的元素作乘积，将所得结果相加得到一个结果值，最后通过激活函数获得卷积结果。当卷积核滑动遍历整张图像，结束特征提取，获取一个新的图像特征矩阵(feature map) 。同时卷积核滑动的步幅也和最后获取的特征矩阵存在以下关系[1]:

$$a_{i,j} = f\left(\sum_{m=0}^{M} \sum_{n=0}^{N} w_{m,n} x_{i+m,j+n} + w_b \right)$$

$$f(x) = \max(0, x)$$

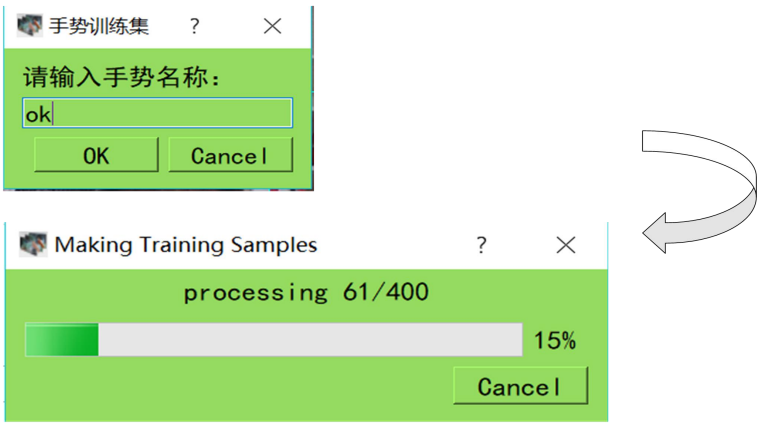
式 1 为卷积计算，式 2 为激活函数。$x_{i,j}$ 为图像第 i 行 j 列的像素值，$w_{m,n}$ 为卷积核中第 m 行第 n 列权重，w_b 为卷积核的偏置项常数;f 为激活函数，即 relu 函数。卷积滤波后再通过下采样图像特征矩阵进行降维，减少计算量，同时避免特征过多导致出现过拟合，增强网络结构对位移的鲁棒性。



图： 本研究之 CNN 模型结构图

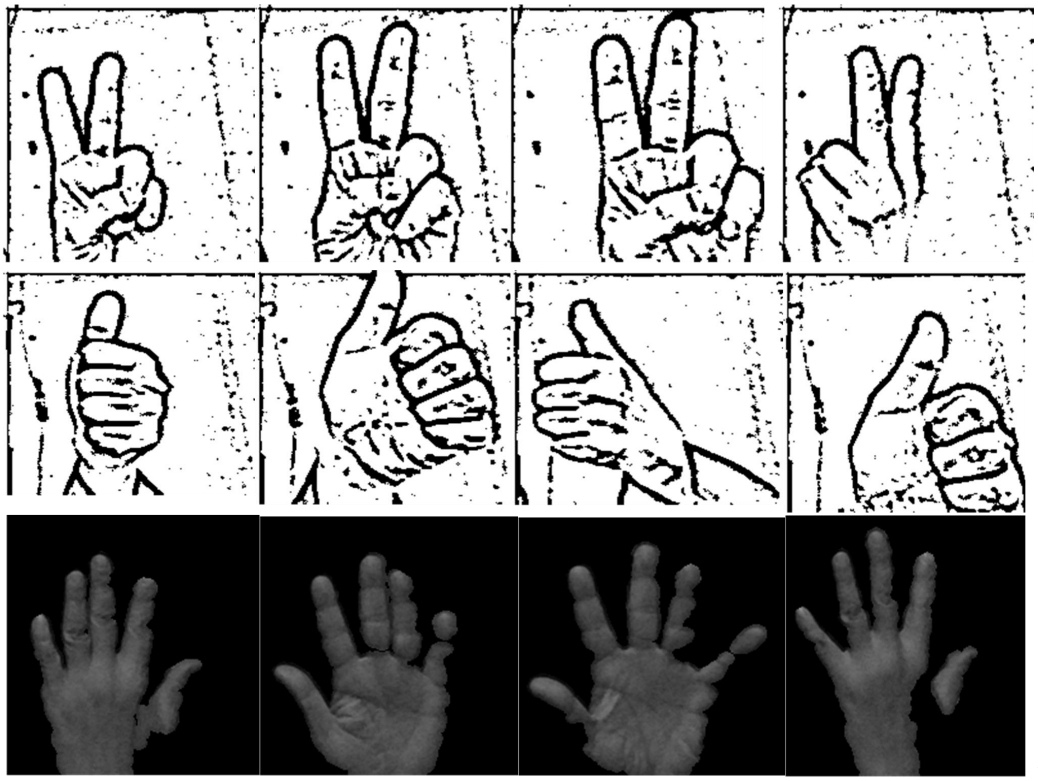
3.2.3. 训练库构建

本模型除了支持引入外部训练库，还支持用户自定义训练库的制作。用户只需要点击“制作样本”按钮，设置手势的含义，就可以根据手势展示区的手型以 0.04s 的间隔进行抽帧处理，并保存到训练库中。本研究设置每次抽帧的次数不超过 400，样本以*.png 的格式保存，样本分辨率为 200*200。



图：手势训练集制作过程

图 二值化手势训练图像（1、2 行）、肤



色模型手势训练图像（3 行）

3.2.4. 模型训练

CNN 模型训练在 Keras 框架实现，采用交叉熵损失函数（categorical_crossentropy loss），交叉熵是用来评估当前训练得到的概率分布与真实分布的差异情况。具体公式为：

$$C = -\frac{1}{n}\sum_{x}[y\ln a + (1-y)\ln(1-a)]$$

其中，a 为神经元实际输出。当误差大的时候，权重更新就快，当误差小的时候，权重的更新就慢。

3.3. 实验效果

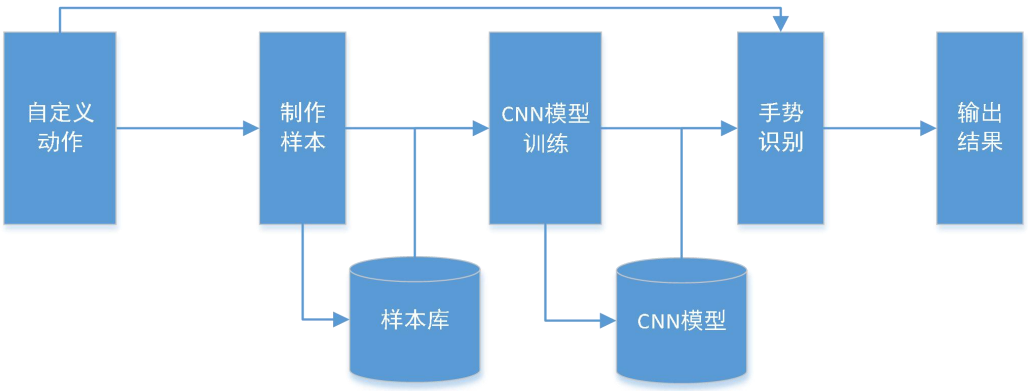
为检验基于视觉的人脸+手势模型效果，我们采用自定义训练样本的功能构建 4 种手势（one：1，two：2，five：5，good：10）的样本库，样本总数为 1581 个，其中 80%（1264 个）用于训练，20%（317 个）用于测试，具体分布如下：

表：手势训练和测试情况一览

手势名称	含义	训练样本（个）	测试样本（个）	合计（个）
One	1	318	79	397
Two	2	313	78	391
Five	5	318	79	397
Good	10	317	79	396

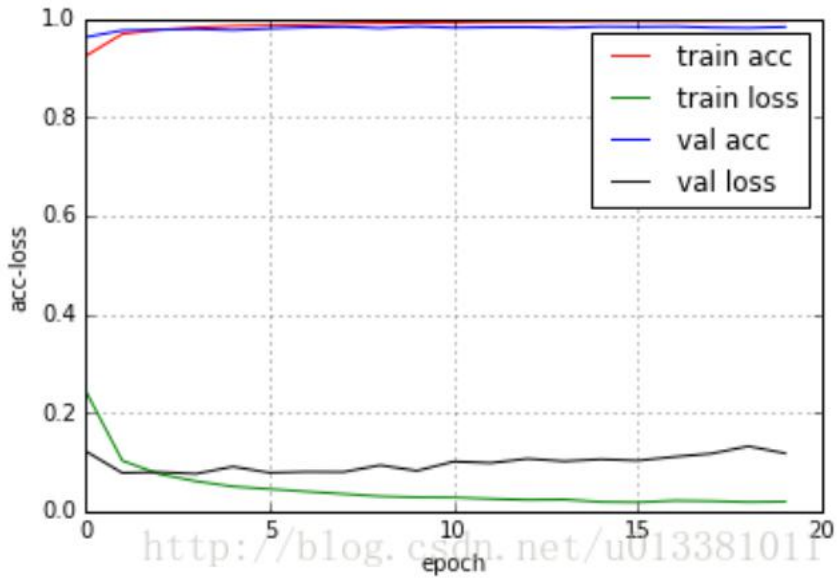
实验在是在 Dell Vostro 5471 便携式计算机上开展的，处理器为 Intel Core i7-8550U CPU 1.80GHz，内存（RAM）为 8GB，GPU 处理器为 AMD RANDON 8GB，采用 win10 x64 操作系统。

实验过程主要包括 3 个步骤：1、制作样本、2、CNN 模型训练、3、手势/人脸识别：制作样本步骤接收自定义动作视频，抽帧后按照手势分类存储原图，输出样本库；CNN 模型训练步骤接收样本库，根据比例随机抽取样本库进行 CNN 模型训练和验证，经过 25 次迭代训练，输出 CNN 模型文件；手势/人脸识别步骤导入 CNN 模型文件，接收用户自定义动作的图像，对图像中的手势和人脸进行分析，输出识别结果。具体步骤如下图所示：



图：实验步骤一览

CNN 模型训练过程的 acc-loss 图如下图所示，横轴表示迭代的次数，纵轴表示 acc-loss 的变化情况。Train loss 在 15 次迭代（epoch=15）后逐渐趋于平稳，小于 0.000047，Val acc 和 Train acc 在 10 次迭代（epoch=10）以后逐渐趋于稳定， 大于 0.99；Val loss 则表现小范围不稳定，始终在 0.1 左右摆动，表明测试样本与训练样本仍存在一定程度的偏差，可能和迭代次数较小有关。



图：acc-loss 曲线

为了测试手势比对的准确性，分别对每个手势进行 10 组测试，并记录识别正确和错误的个数、平均执行时间。从测试情况看，手势

识别的平均正确率为 80%，平均执行时间为 68.3ms。

表：手势比对结果一览

手势	测试总数	正确（个数）	错误（个数）	平均执行时间
One(1)	10	9	1	68.9ms
Two(2)	10	8	2	53.1ms
Five(5)	10	10	0	82.3ms
Ok(10)	10	4	6	69ms

人脸比对实验主要进行了基准照和现场照比对，本实验采用了开源的人脸识别算法（face-recognition），选取 6 组名人（OBAMA，JIM, YAOM, YIJIANLAN, ZHAOBS, SHENT）及本人基准照作为比对的基础，进行视频抓取和 1：1 比对，现场展现识别结果，结果表明识别准确率接近 100%。

图：手势识别系统

基于视觉的人脸+手势识别原型要求实时识别人脸、检测手势的状态，对系统的性能有一定的要求，本实验的 CPU 消耗为 18.4%，内

存占用为 963.4MB。

4. 结论

随着计算机视觉和机器学习的进步，人机交互技术也逐渐成熟，人脸和手势具有直观性、自然性和丰富性等特点，其识别研究引起了人们的极大关注。本文提出了一种基于视觉的人脸+手势原型，该原型采用基于 CNN 的机器学习训练模型，具有识别精度高、识别速度快、使用方便的特点。该原型能够应用于金融领域，为客户提供定制化的推荐服务。此研究也为我行的个性化人机交互研究建立基础，便于推广到社会治安、教育、医疗、娱乐等多个泛金融领域，具有巨大的市场价值和社会价值。

文献

[1] 杨红玲、宣士斌、莫愿斌，“基于卷积神经网络的手势识别”，《计算机科学与发展》，vol28，No7，July 2018；

[2] Hohong. "高斯模糊与图像卷积滤波一些知识点". 《简书》，https://www.jianshu.com/p/8d2d93c4229b，2017 年 7 月 17 日.