基于深度学习的隐私摄像安全防护方案

刘田田
（江苏开放大学信息工程学院，江苏 南京 210017）

摘 要：随着网络技术的发展，视频的采集与获取变得越来越便捷，如何安全、可控的采集视频成为使用者不得不面对的问题。为了解决上述问题，提出一种基于深度学习的隐私摄像安全防护的设计思路，解决视频内容的安全与可控问题。该系统通过人体轮廓分割解决视频内容隐私问题，使用手势识别解决视频可控问题，使用人脸识别技术进一步保护视频隐私。

关键词：深度学习；人脸识别；人工神经网络；视频安全
中图分类号：TP391 **文献标识码**：A **文章编号**：2096-4390(2020)31-0102-02

现代社会，安全隐私是每个人甚至每个企业不得不面对的问题，安全隐私涉及到的问题，在生活中随处可见，如教育、医疗、交通等领域。近年来，随着网络技术的发展，视频作为信息传递载体具有诸多优势，而视频来源的重要设备——摄像头则备受关注。如何安全、可控的采集视频成为使用者不得不面对的问题。现有技术的重点仅仅为了保护视频的安全，然而内容的安全却少有关注。2010 年新一代信息技术变革，深度学习技术也随之迅猛发展，因此利用深度学习来进行研究探索视频内容的安全与可控问题，不失为一种有效的方法。

1 研究现状

1956 年美国汉诺斯小镇的达特茅斯“用机器来模仿人类学习以及其他方面的智能”的会议上，“人工智能”首次被提出，经历了繁荣、低谷的轮回期，于 2010 年新一代信息技术引发的海量信息与数据的变革中迎来了增长爆发期。深度学习是机器学习研究领域目前发展势头最好的一个新的领域，由 Hinton 等人于 2006 年，在顶级期刊《科学》上的一篇论文中提出[1]，核心是模拟人脑的机制来解释数据，例如图像、声音和文本。对人工神经网络进行学习训练，试图寻找最优解。语义分割，是计算机视觉中的基本任务，在语义分割中我们需要将视觉输入分为不同的语义可解释类别，也就是像素级图像分类任务[2]。视频动作识别也是深度学习领域一个较新的研究方向，潘陈听等人研究了复杂背景下的视频动作识别[3]。

2 技术分析

U-Net[4]是 Olaf Ronneberger 等人参加 ISBI Challenge 提出的一种分割网络，能够适应很小的训练集（大约 30 张图）。U-Net 是很小的分割网络，既没有使用空洞卷积，也没有后接 CRF（随机场），结构简单。整个 U-Net 网络结类似于一个大大的 U 字母：首先进行 Conv+Pooling 下采样；然后 Deconv 反卷积进行上采样，crop 之前的低层 feature map，进行融合；然后再次上采样。重复这个过程，直到获得输出 388x388x2 的 feature map，最后经过 softmax 获得分割图。总体来说与 FCN 思路非常类似。U-Net 采用将特征在通道维度拼接在一起，形成更“厚”的特征。

MTCNN 网络是 Kaipeng Zhang 等人于 2016 年发表的“基于多任务级联卷积神经网络的人脸检测和对齐”一文中提出[5]，主要作用主要可以实现特定目标检测与对齐，其网络结构为三层网络。第一层 PNet 网络的结果经过 bounding boxes regression 和 NMS 处理之后变为 24*24 的图像大小放入第二层处理；第二层 RNet 处理后的结果同样经过 bounding boxes regression 和 NMS 处理变成 48*48 大小图像放入第三层处理；结果同样经过 bounding boxes regression 和 NMS 处理输出目标框与类别信息。

3 系统分析与设计

本方案所应用的语义分割深度网络 U-NET 是一种经典网络，最初用来处理医学影像问题，经过改进后用来处理分割人体前景与背景的问题。基于深度学习的图像分类技术，是输入图像对该图像内容分类的描述的问题。本方案所应用的手势分类深度网络 MTCNN-P 为较浅网络，最初用来处理人脸识别定位问题，经过改进后用来处理手势识别的问题。基于深度学习的人脸识别技术，是当下人脸识别的主要方向，以数据作为驱动引擎，解决诸多传统算法的弊端。本方案所应用的人脸识别网络为 IsightFace 网络，用来解决视频中人脸识别的问题。

3.1 功能分析

本方案采用改进 MTCNN 网络，即 MTCNN-P 网络。MTCNN 网络模型尺寸足够小，使得其可以应用于嵌入式，满足系统性能要求。MTCNN 网络主要作用主要可以实现特定目标检测与对齐，其网络结构为三层网络。微调后 MTCNN-P 其基本的构造是一个简单分类网络，去除原有的框回归，输出二值信息，判断类别。基于 MTCNN-P 的手势分类采用的是基于深度学习的普通分类算法，该网络用来检测人脸，可以胜任简单的分类任务。

IsightFace 网络核心部分损失函数（Centre loss）主要惩罚了深层特征与其相应的欧几里得空间类中心之间的距离，以实现类内紧凑性。假设在最后一个完全连接的层中的线性变换矩阵可以用角空间中的类中心来表示，并且以乘法方式惩罚深度特征与其相应的权重之间的角度。特征和最后一个完全连接的层之间的点积等于特征和权重归一化之后的余弦距离。利用余弦函数（arc-cosine function）计算人脸特征和目标权重之间的夹角。然后，在目标角度上增加一个附加的角余量，通过余弦函数

基金项目：“基于云计算的大数据管理与分析研究团队”成果、江苏省高等学校自然科学研究项目资助（18KJB520008）。
作者简介：刘田田（1990-），女，山东济宁人，助教，硕士，研究方向：算法分析与设计。

再次得到目标 logit。最后，用一个固定的特征范数重新缩放所有 logits，并且后续步骤与 softmax loss 中的步骤完全相同。传统的 softmax loss 损失函数为：

$$L_1 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{W_{y_i}^T x_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T x_i + b_i}}$$

这里做了一个变换，将 $W_j^T x_i$ 替换为下式：

$$W_j^T x_i = \left|W_j\right|\left|x_i\right|\cos\theta_j$$

也就是向量内积的结果是向量各自的模相乘，在乘上向量夹角的余弦值。那么向量相乘得到的结果其实就是 x_i 对应在第 j 类的夹角。然后使用 L2 正则化处理 W_j 使得 $|W_j|=1$，L2 正则化就是将 W_j 向量中的每个值都分别除以 W_j 的模，从而得到新的 W_j，新的 W_j 的模就是 1，实际上是个方向向量进而获得概率。

$$L_2 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s\cos\theta_{y_i}}}{e^{s\cos\theta_{y_i}} + \sum_{j=1,\ j\neq y_i}^{n} e^{s\cos\theta_{y_j}}}$$

并且将 $\cos\theta_{y_i}$ 用 $\cos\theta_{y_i} + m$ 替代：

$$L_3 = -\frac{1}{N}\sum_{i=1}^{N}\log\frac{e^{s(\cos(\theta_{y_i}+m))}}{e^{s(\cos(\theta_{y_i}+m))} + \sum_{j=1,\ j\neq y_i}^{n} e^{s\cos\theta_j}}$$

3.2 系统设计

集成三种深度神经网络，分别实现人体轮廓分割、手势识别、人脸识别三大功能。人体轮廓分割为主要处理任务，手势识别与人脸识别相当于外层逻辑，实现“隐私”控制。整套系统架构如图 1 系统架构图所示。

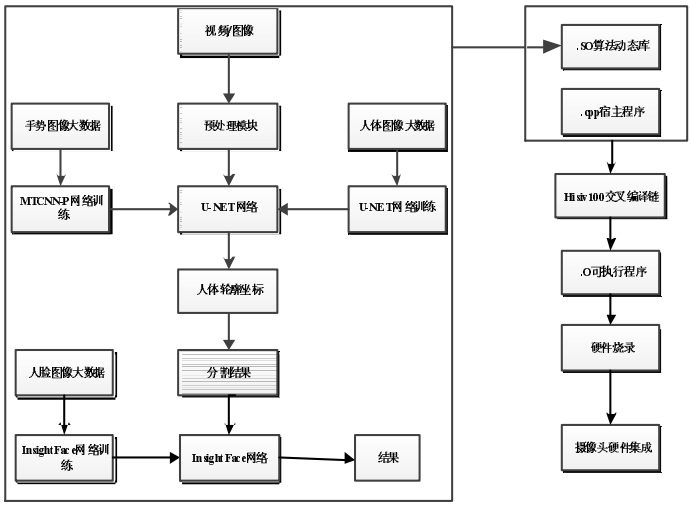

图 1 系统架构图

整体代码为 C++ 程序，便于后续集成宿主程序。深度学习模型代码文件经过特定平台编译器，生成.SO 算法动态库，这个动态库与宿主程序经过 Hisiv100 交叉编译工具生成.o 可执行程序，烧录进摄像头，实现最终软硬件结合。

集成到摄像头终端的三个深度学习模型，为提前训练好的模型。为了满足在嵌入式设备上运行深度学习模型，需要进一步优化。本方案使用了常见的 int8 量化方法，进一步压缩模型，提升性能。原始图像经过预处理模块简单进行噪声过滤处理，消除常见噪声对图像质量的影响。图像在进入 U-NET 网络之前，会进行手势判断，检测手部区域并定位手部关节点，根据手部关键节点的形状判断属于哪种手势。这个手势为人的手掌“OK”造型时，表示验证通过，视频流可以进入 U-NET 网络。这样做的目的就是录像的自主可控，在不想要录制的时候可以“示意”摄像头“拳头”造型，表示终止视频流。视频流进入 U-NET 网络，实现人体轮廓分割，得到轮廓坐标，进一步提取人体前景与背景信息，并对背景部分进行遮挡，实现视频流隐私的保护。在进行最终结果输出的时候，会进行人脸识别判断，如果非设定人员，则不会输出最终结果，实现视频流的自主控制。

本方案组合新颖，核心部分均采用以数据为驱动的深度学习网络，对原创视频（直播）数据进行多层防加密护，真正做到数据的安全自主可控。系统架构清晰，可轻松移植到嵌入式、服务器中，而且不需要过多代码。整个架构鲁棒性较强，应对人为破坏能力较强，安全性和稳定性较高。

结束语

本方案为了解决视频内容的安全与可控问题，提出集成三种深度神经网络。通过人体轮廓分割处理视频内容任务，通过手势识别与人脸识别，实现“隐私”控制。本方案中集成到摄像头终端的三个深度学习模型，是提前训练好的模型，若将该模型应用到嵌入式设备上，后续需要进一步优化。

参考文献

[1] Hinton G E, Osindero S, Teh Y W. A fast learning algorithm for deep belief nets [J]. Neural computation, 2006, 18 (7): 1527-1554.

[2]马煜,杜慧敏,毛智礼,张霞.深度语义分割人群密度检测技术[J/OL]. 计算机科学与探索:1-11 [2020-09-15].http://kns.cnki.net/kcms/detail/11.5602.tp.20200910.1530.006.html.

[3]潘陈听,谭晓阳.复杂背景下基于深度学习的视频动作识别[J].计算机与现代化,2020(07):97-103.

[4]Olaf Ronneberger, Philipp Fischer, Thomas Brox. U -Net: Convolutional Networks for Biomedical Image Segmentation [J]. 2015,arXiv:1505.04597 [cs.CV]

[5]Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li,Yu Qiao. Joint Face Detection and Alignment using Multi -task Cascaded Convolutional Networks[J]. IEEE Signal Processing Letters (SPL), vol. 23, no. 10, pp. 1499-1503, 2016.