doi:10.3772/j.issn.1002-0470.2021.02.003

基于 Kinect 多生物识别技术的智能视频播放器交互系统①

李国友② 王维江③ 李晨光 杭丙鹏 杨梦琪

(燕山大学电气工程学院 秦皇岛 066004)

摘 要 为了实现观看者无接触操作情况下的视频播放器智能控制,系统利用 Kinect 传感器采集彩色图像,使用 FaceNet 提取人脸特征向量,经支持向量机(SVM)训练后进行人脸识别,该过程在计算机中央处理器(CPU)运行环境下,利用 OpenVINO 实现人脸检测与识别实时运行,用于视频播放器的登录验证。系统采集音频数据使用 Speech Platform Runtime v11 进行中文命令识别,使用 Kinect Speech Language 进行英文命令识别,进而实现语音控制。采集骨骼数据,计算骨骼点之间的距离与角度进行人体姿态和手势识别,将识别结果转换为控制命令,进而实现播放器的快进、切换视频、加减音量等常用的控制功能。实验结果表明,该交互系统实现了使用者无接触全自动人体控制,为视频播放器提供了一种自然便捷的交互方式。

关键词 Kinect V2 传感器; OpenVINO; 人脸识别; 播放器控制; 语音识别; 手势识别

0 引 言

随着科技的不断进步和发展,人机交互技术逐步由传统的键盘鼠标接触式控制方式转向基于动作识别的体感交互模式[1-2]。人机交互的变革使人们的生活方式更加丰富多彩,语音和手势作为日常生活中人们习惯的交流方式,有着自然、方便、灵活的特点。因此,基于人体多生物特征识别技术也逐渐成为人机交互领域的研究热点,如语音、人脸、人体姿态和手势等。这些识别技术应用于智能电视、监控系统、医疗康复系统、机器人操作等领域,使交互更加自然便捷[3]。

近年来,李健等人[4]提出运用人脸识别及手势识别的 PPT 全自动控制系统,其人脸识别主要使用主成分分析算法。此算法虽未摒弃图像信息,但是其计算耗时较长,当人脸图像姿态、尺度变化较大时,人脸识别性能一般。陈一新[5]提出了基于手势与智能电视图形用户界面进行交互,其手势识别采用位置相识度权重改进动态时间归划(dynamic time warping, DTW)算法,提高了分类率。李佳怡和刘东旭[6]提出了体感识别的自衡车交互系统,通过 Kinect 识别人体肢体动作或手势,进而实现对自衡车常用功能的实现。Ma 等人[7]提出了利用骨骼追踪技术,基于 Hu 矩手势识别,进而控制机器人手臂,实现人与机器人的交互。上述的交互系统都使用了 Kinect 传感器,Kinect 设备单一简单,在游戏、医学、教育、农业等众多领域都有 Kinect 的身影[8]。

目前对于 Kinect 的开发大多数只局限于利用其一种数据开发一种实现方式,如 Alzahrani 等人[9]使用 Kinect 获取人体骨骼信息进行人体跌倒检测。Liu 等人[10]利用 Kinect 获取人脸的红外图像并运用多任务卷积神经网络(multi-task convolutional neural network, MTCNN)和 FaceNet 模型进行人脸识别,有效杜绝了人脸照片对人脸识别系统的欺骗。而本文的优势在于使用 Kinect 传感器采集音频数据、彩色图像、深度数据、骨骼数据进行语音识别、人脸识别、

① 国家自然科学基金(F2012203111)和河北省高等学校科学技术研究青年基金(2011139)资助项目。

② 男,1972 年生,博士,教授;研究方向:工业控制,图像处理,机器视觉,智能控制;E-mail: lgyysu@163.com

③ 通信作者,E-mail: 392307856@qq.com

(收稿日期:2020-02-13)

手势识别和人体动作姿态识别，利用一个 Kinect 传感器实现多功能的交互系统。

目前视频播放器交互系统还在使用键盘鼠标进行相应的控制，比如爱奇艺、腾讯视频播放器等。针对传统交互方式，本文提出了运用人脸识别、语音识别、手势与人体姿态识别相结合的视频播放器交互系统，摆脱了传统的交互方式，并且该智能交互系统可以应用于任何一个视频播放器，使交互方式更加灵活多样。本文交互系统以爱奇艺播放器为对象，利用 Kinect 采集各类数据，采用多生物识别技术，应用深度学习、机器学习、图像处理对数据进行识别，并且将识别结果转化为控制命令，从而实现无接触情况下的视频播放器控制。实验结果表明，本文所设计的系统具有较高的稳定性、实时性和准确性。

1　人机交互系统的总体设计

本文人机交互系统由 3 个部分组成，即人脸识别模块、语音识别模块和手势与人体姿态识别模块。人脸识别模块主要负责交互系统中人员的认证与登录，主要包括人脸检测、特征提取、模型训练、人脸识别。语音识别模块主要负责交互系统语音控制功能的实现，主要包括音频处理、中英文辨别、语音命令识别。手势与人体姿态识别模块主要负责交互系统人体动作控制功能的实现，其主要包括骨骼追踪、人体关节点之间角度及距离的计算、体感识别。语音识别和手势与人体姿态识别是整个交互系统的核心，实现了使用者无接触全自动人体控制。

交互系统利用 Kinect 收集图像数据后，首先对使用者进行人脸识别、认证登录，然后根据使用者个人偏好来确定使用语音识别模块或人体姿态与手势识别模块，系统将识别结果转换为控制命令，进而实现播放器的快进、切换视频、加减音量等常用的控制功能。人机交互系统的整体框架如图 1 所示。

2　人脸识别

人脸识别模块首先采用基于卷积神经网络单步多框检测器（single shot multibox detector，SSD）的人脸检测器进行人脸检测并获取人脸图像，对人脸图像进行预处理，后经 FaceNet 提取人脸特征向量，对特征向量使用余弦相似度与数据库比较，进行阈值预筛选，未通过限定阈值则重新采集图像，通过则采用支持向量机（support vector machine，SVM）进行分类识别。人脸识别的流程如图 2 所示。

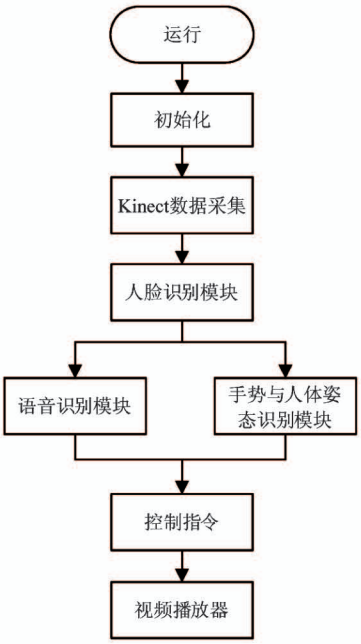


图 1　系统总体流程图

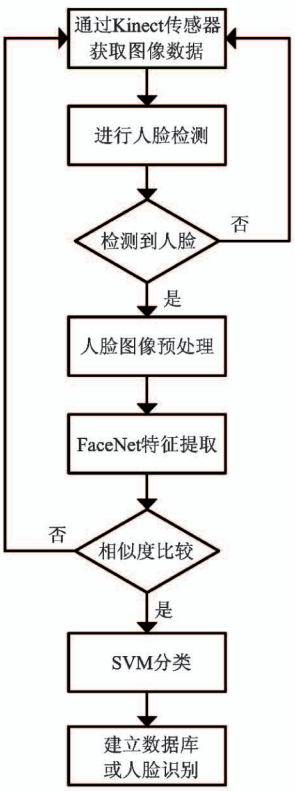


图 2　人脸识别流程图

2.1 人脸检测与数据库建立

人脸检测就是在给定的图片中查看是否有人脸存在,并进行人脸的定位,是人脸识别的前提。早期的人脸检测使用 HOG/LBP 算法[11],该算法很不稳定,无法商用,并且计算量比较大,一直开窗检测。之后人脸检测采用 HAAR 级联检测器[12],算法比较稳定,但是对遮挡、不同角度、光线等影响很敏感。本文采用的人脸检测器是 OpenCV 所提供的基于卷积神经网络 SSD[13] 的人脸检测器。该人脸检测器可以实时运行,对各种角度人脸均可以做到正确检测,具有很强的抗干扰能力。

SSD 的基本思想是采用一个 CNN 网络来进行检测,其网络使用 VGG16 作为基础模型,然后在 VGG16 的基础上新增了卷积层来获取更多的特征图用于检测。SSD 基本架构如图 3 所示。SSD 算法的整体流程为:输入一张 300 × 300 像素、通道为 3 的图片,图片经过 SSD 特征提取网络的卷积运算生成一系列特征图,SSD 选择其中的 6 个(Conv4 _ 3、Conv7、Conv8 _ 2、Conv 9 _ 2、Conv10 _ 2、Pool 11)作为"有效特征层"。接下来 SSD 对"有效特征层"进行密集采样,采样时使用不同尺度和长宽比,然后卷积运算获得目标的位置信息与类别信息,并将各个"有效特征层"获得的信息合并;最后使用非极大值抑制法(non-maximum suppression, NMS)筛选出最终的检测结果。

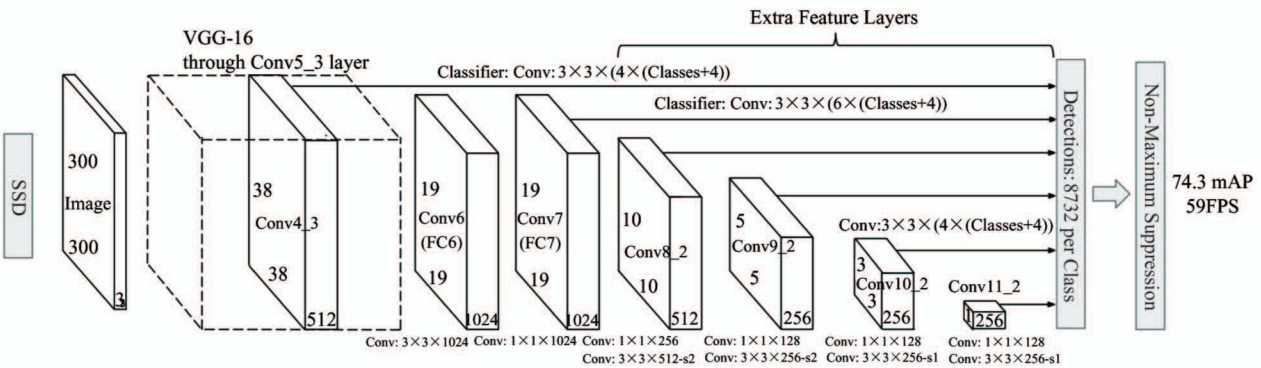
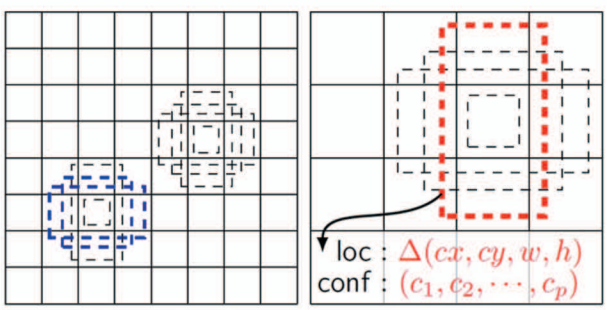


图 3 SSD 检测模型框架

多尺度特征图检测是 SSD 的核心,也是其精度更高、更擅长检测小物体的重要原因。浅层分辨率大的特征图检测小目标,深层分辨率大的特征图检测大目标,充分体现了各层特征图的优点。候选框选取图如图 4 所示。



(a)8 ×8 特征图　　(b)4 ×4 特征图

图 4 候选框选取图

如图 4 所示,SSD 在各个"有效特征层"的每一个网格上设置了一组尺度和长宽比不同的 Default Box(候选框)以匹配不同尺寸和形状的物体。图 4(a)中 8 ×8 的特征图用于预测小目标,图 4(b)中 4 × 4 的特征图用于预测大目标。图中,loc:$\Delta(cx, xy, w, h)$ 表示预测候选框的位置信息,分别表示候选框的中心横、纵坐标及长和宽;conf:$(c_1, c_2, \cdots, c_p)$ 表示类别置信度预测值。SSD 的网络预测值实际上就是每个 Default Box 的调整情况,有 Default Box 作为预测的基准,在一定程度上降低了模型训练的难度。

基于卷积神经网络 SSD 的人脸检测如图 5 所示,在处于光线不充足的场景中,并且对人脸进行遮挡的情况下,基于卷积神经网络 SSD 的人脸检测器有良好的检测效果。在进行人脸检测后,首先需要对图像进行预处理,其中包括人脸裁剪,将图片进行裁剪,使图片只包含人脸,即人脸的感兴趣区域(region of interest, ROI)。其次对人脸图片进行格式转化与缩放,以便后续提取人脸特征向量。

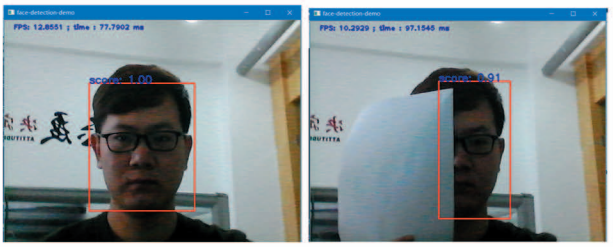


图 5　人脸检测图

为了建立人脸数据库，文中共对 4 个人进行人脸图像的采集，在不同光照、不同角度的情况下，对每个人采集 100 张图片，其图片都经过人脸图像预处理得到。并在此基础上加入了 ORL 人脸数据集，以提高样本数据量。

2.2　特征提取

在人脸特征提取部分，使用 FaceNet 模型。FaceNet 是一种人脸识别模型，它对遮挡、模糊、光照和角度具有很强的鲁棒性[14]，因此使用 FaceNet 模型进行人脸特征提取。

FaceNet 网络模型结构如图 6 所示。人脸图像批量输入层（Batch）是指人脸图像样本通过面部检测并裁剪为 160 × 160 像素的图像，作为 FaceNet 的输入。Facenet 的深度学习框架（deep architecture）是采用了 GoogleNet，在传统卷积神经网络基础上加入了多个 Inception 结构，并在此基础上进行优化，将 GoogleNet 的 softmax 分类器去掉，替换成一个 L2 特征归一化的嵌入层，经过 L2 的归一化后，输出一个 128 维的特征向量，并用三元损失函数（triplet loss）进行优化。

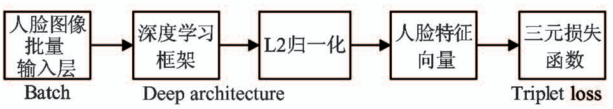


图 6　FaceNet 结构图

FaceNet 力求嵌入 $f(x)$，从人脸图像 x 映射到特征空间 R^d 中，即 $f(x) \in R^d$，通过 $\| f(x) \|_2 = 1$ 限制 d 维超平面，这样使得相同身份的所有人脸的平方距离都很小，而来自不同身份的人脸图像之间的平方距离都很大。FaceNet 使用三元损失函数基于最大间隔近邻分类（large margin nearest neighbor，LMNN）训练输出 128 位连续向量来表示人脸。

FaceNet 使用三元损失函数，其三元损失函数包含了两个匹配人脸图像和一个不匹配的人脸图像，损失函数的训练目标是使其在匹配人脸和不匹配人脸之间距离足够大为止。三元损失函数如图 7 所示。

图 7　三元损失函数示意图

Anchor（固定标签）是一个特定的人脸图像，Positive（正例图像）是来自同一个人的其他人脸图像，Negative（反例图像）是来自其他人的任意人脸图像。在模型学习的过程中，确保特定人脸的图像（Anchor）在欧氏空间中尽可能接近同一个人的所有其他人脸图像（Positive），并且尽可能地远离其他人的任意人脸图像（Negative），公式化表示如式（1）所示：

$$\| f(x_i^a) - f(x_i^p) \|_2^2 + \alpha < \| f(x_i^a) - f(x_i^n) \|_2^2$$
$$\forall f(x_i^a), f(x_i^p), f(x_i^n) \in T \qquad (1)$$

其中，x_i^a（Anchor）为某个人的人脸图像，x_i^p（Positive）为这个人的其他人脸图像，x_i^n（Negative）为不属于这个人的其他任何人的人脸图像，α 是强制区分正负样本的距离，$f(x) \in R^d$ 表示人脸图像 x 映射到 d 维欧式空间中，其中集合 T 表示所有可能的三元组。

最终的损失函数计算公式如式（2）所示：

$$L = \sum_i^N \left[\| f(x_i^a) - f(x_i^p) \|_2^2 - \| f(x_i^a) - f(x_i^n) \|_2^2 + \alpha \right]_+ \qquad (2)$$

其中，中括号左侧的欧几里得范数表示同一人脸不同图像间的距离，右侧欧几里得距离表示不同人脸图像间的距离，α 是强制区分正负样本的距离，$+$ 表示中括号内的值大于 0 时，取该值为损失；小于 0 时，损失为 0。N 为三元组的个数。

三元损失函数试图在一个人的每张人脸与其他人的人脸之间建立一个界限，这个过程允许同一个身份的人脸存在多种形式，同时仍然建立距离关系，从而增强了与其他身份的人脸的区别性。因此，

FaceNet 模型对姿态、光照和表情[15]具有较强的鲁棒性,这在以前被认为是人脸验证系统的难点。

2.3 OpenVINO 加速计算

为了提高实时性,本文提出在运行卷积神经网络 SSD 的人脸检测器和人脸识别模型 FaceNet 加入了 OpenVINO。OpenVINO[16]是面向 OpenCV 和 OpenVX 的优化计算机视觉库,支持加速高性能计算机视觉应用和深度学习推理,增强视觉系统功能和性能。OpenVINO 可以使系统在中央处理器(central processing unit, CPU)环境下达到实时性运行,其运行工作流程如图 8 所示。

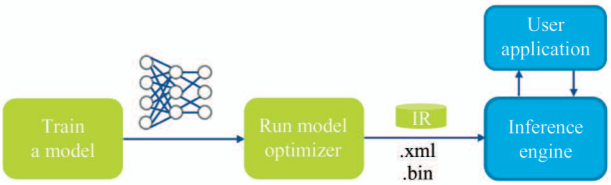
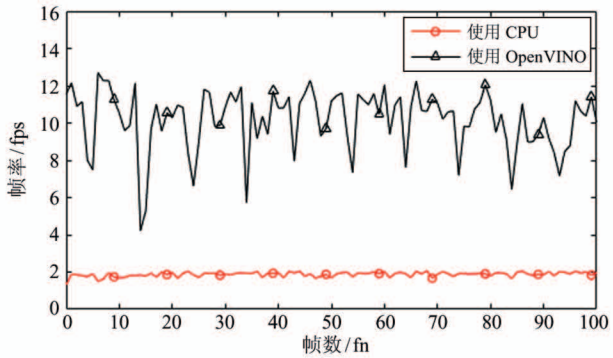


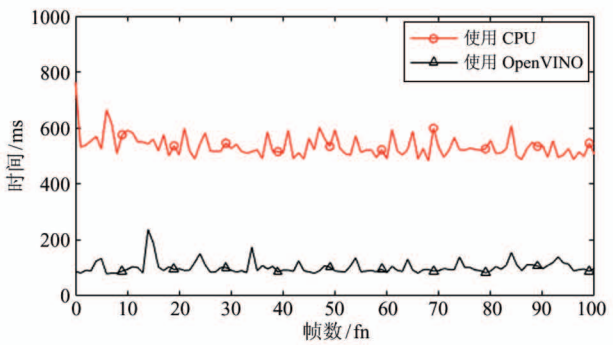
图 8 OpenVINO 工作流程图

Train a model 是经过深度学习框架训练的网络模型,经过 run model optimizer(模型优化器)将模型生成 IR(中间表示),inference engine(推理机)将 IR 在 Intel 平台上提供了统一的 API 用于用户程序。从而在 Intel 平台上提升计算机视觉相关深度学习性能。本文在人脸检测器和 FaceNet 模型运行时使用 OpenVINO,其帧率较没有使用 OpenVINO 显著提高,模型执行推理用时明显减少,使系统能够实时性运行。实验图如图 9 所示。

如图 9 所示,系统帧率及模型执行推理用时较没有使用 OpenVINO 性能大约均提升了 6 倍。



(a) 帧率性能提升图



(b) 时间性能提升图

图 9 OpenVINO 提高性能效果图

2.4 训练模型和人脸识别

在人脸特征提取完成之后,若直接使用特征向量进行余弦相似度计算,设置阈值后进行人脸识别[17],其识别结果准确度不是很高,原因是人脸识别对人种非常敏感。FaceNet 的训练集是采用西方人脸,本文对其进行改进,在此基础上加入 SVM 模型。若人脸相似度计算结果通过设置的阈值,再进行 SVM 分类识别;没有通过阈值则重新采集图像并识别,以增加准确度。

SVM 作为样本分类器主要分为 3 个步骤,即数据准备、特征提取和 SVM 参数配置。其中数据准备和特征提取在 1.1 节和 1.2 节已经完成。参数配置最重要的是核函数的选择,共有 3 种核函数,分别为线性核函数、多项式核函数和高斯核函数。根据本文样本量和特征量,并且因为人脸样本数据集所设置的标签线性可分,故本文选用了简单高效的线性核函数,其公式如式(3)所示:

$$K(x_i, x_j) = x_i^T x_j \tag{3}$$

人脸识别中余弦相似度计算公式为

$$\cos\theta = \frac{\vec{a}\cdot\vec{b}}{\|\vec{a}\cdot\vec{b}\|} = \frac{\sum_1^n a_i b_i}{\sqrt{\sum_1^n a_i^2}\sqrt{\sum_1^n b_i^2}} \tag{4}$$

其中,向量 $\boldsymbol{a}$ 与向量 $\boldsymbol{b}$ 表示人脸的特征向量。余弦值结果越接近 1,即向量夹角越接近 0°则表示向量相似,结果为相似人脸;若两向量余弦值结果越接近 0,即夹角越接近 90°,则表示向量正交,结果为不同人脸。经过实验,本文的阈值设为 0.4。因为若阈值大于 0.4,会导致不同身份的人脸更容易通过筛

选;而阈值小于 0.4,会使本身份的人脸不通过筛选,所以文本选用阈值为 0.4。

在人脸识别完成后,系统会通过验证,并且自动打开播放器。人脸识别不仅可以提供视频播放器的登录,还可以设置少儿防沉迷模式,设置规定的时间,超时则退出交互系统,并关闭视频播放器。

3　语音识别

交互系统的语音控制部分,针对 Kinect 语音识别技术[18]中语音交互控制功能没有得到有效开发,系统采集音频数据使用 Speech Platform Runtime v11 进行中文命令识别,使用 Kinect Speech Language 进行英文命令识别,本文语音识别模块可离线运行,实现了中文和英文双重语音控制。

3.1　音频处理技术

本系统语音识别硬件是使用 Kinect 的麦克风阵列,Kinect 四元麦克风阵列位于 Kinect 设备的下方,4 个麦克风是不对称分布的,每一个麦克风都捕获相同的音频信号,麦克风阵列可以探测到声音的来源方向,其捕获的音频数据流经过复杂的音频增强算法处理,用来移除不相关的背景噪声。这些复杂的操作在 Kinect 硬件和 Kinect SDK 之间进行处理,这些处理包括降噪、自动增益控制和回声消除。Kinect 麦克风阵列如图 10 所示。

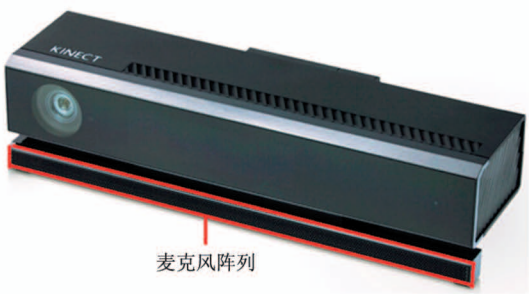


图 10　Kinect 麦克风阵列

对于同一段音频数据,在音量大小相同的条件下,使音源与 Kinect 传感器的距离以及与 Microsoft 声音映射器(PC 机默认的录制和播放驱动器)的距离保持相同,Microsoft 声音映射器和 Kinect 传感器同时对音频数据进行采集,使用音频处理软件 Audacity 对其结果进行频谱图分析,其分析结果如图 11、图 12 所示。

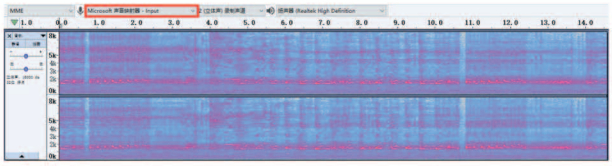


图 11　声音映射器采集结果频谱图

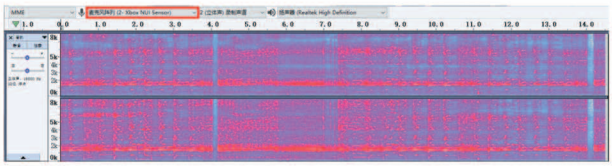


图 12　Kinect 传感器采集结果频谱图

图 11、图 12 中频谱图中空白部分表示音调低于20 dB,阴影部分表示为音频调高于 20 dB,其中阴影部分较为规律的横线和竖线表示音调为 20 ~ 60 dB,其余表示音调为 60 dB 以上。人正常通话的声音音调为 20 ~ 60 dB,60 dB 以上被认为是噪音。由图 11 可以看出,只存在少量横线和竖线并存在大量空缺区域,即有大量的噪声以及部分音频数据缺失,而图 12 中阴影区域密集,空缺部分相对较少,充分证明 Kinect 传感器在音频采集上可以保真并且可以对声音进行一系列处理,以便后续语音命令识别。

3.2　语音命令识别

语音识别分为两类:对特定命令的识别和对自由形式的语音识别。在人机交互中,并不需要对所有的语音进行识别,只需识别特定的命令词汇。考虑到在播放视频时,视频背景声音对识别产生干扰,因此本系统采用命令识别的方式,限制了命令词汇的范围,排除了视频播放背景声音的干扰。

Kinect SDK 支持 7 种语言的语音识别,其中并不支持中文的语音识别。为实现中文语音识别,系统使用微软的 Speech Platform Runtime v11 进行中文语音命令识别。本系统的语音识别流程如图 13 所示。

其中初始化语音识别组件相应的流程为:加载语音识别引擎,加载语音识别器(中文、英文),加载中文和英文的语音包,加载语法文件。其语法文件使用W3C的语法规范1.0标准,简称SRGS1.0,支

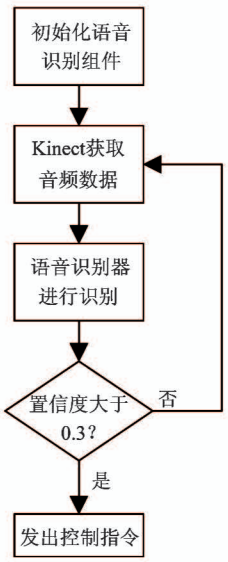


图 13　语音识别流程图

持英文与中文。下面以播放器“全屏”命令为例,考虑到用户对“全屏”可能说的是“full screen”、“full”、“FS”、“全屏”,在编写语法文件时就要把这4个子项包含于“全屏”命令中。一旦语音识别引擎接收到这 4 个子项的任意一项,就识别为“全屏”命令。主要语句如下所示。

```
<item>
    <tag>FULL</tag>
    <one-of>
    <item> full screen </item>
    <item> FS </item>
    <item> full </item>
    <item> 全屏 </item>
    </one-of>
</item>
```

在系统运行过程中,Kinect 持续采集音频数据,将数据传送给语音识别器,语音识别器进行语音命令识别,识别结果与预先定义好的语法文件命令进行匹配,若置信度大于 0.3,则语音命令识别成功。系统对视频播放器发出相应的控制命令,并且刷新控制面板,使其相应的命令字体变为蓝色。其控制面板如图 14 所示。

(a) 英文控制面板

(b) 中文控制面板

图 14　视频播放器控制面板

4　手势与人体姿态识别

交互系统中手势与人体姿态控制部分,系统采集人体骨骼点的位置坐标,并计算骨骼点之间的距离和角度,经算法识别后转化为控制命令。本文并没有采用常用的手势识别算法,如概率统计的方法隐马尔科夫模型(hidden Markov model, HMM)和模板匹配的动态时间规划(DTW)算法等。因为基于 HMM 的手势识别在本系统无法实时运行;基于 DTW 的手势识别识别率高,但需要大量的路径且路径中所有节点都需进行匹配计算,导致计算量大,而本系统需要持续统计人体各个骨骼节点的数据,所以 DTW 并不适用于本系统。此外,系统还需对人体姿态动作进行识别,因此并未采用传统的手势识别算法。

本文在动态手势和人体姿态识别时,采用模板匹配算法,即系统采集 t 与 $t+\Delta t$ 两个时刻的指定关节点位置坐标,计算关节角度与距离,然后同设定的阈值相比较,判定与规定的手势或姿态是否相匹配,避免了对手势轨迹的识别,简化了识别的难度。

在静态手势识别时,使用 Kinect SDK 对静态手势识别,共有 3 种静态手势(张开、闭合、半张开),避开了未定义静态手势的干扰,提高了识别的准确性。

4.1　骨骼追踪技术

Kinect SDK 中提供了人体的骨骼数据,其中包括人体的各个关节点的空间坐标位置分布信息,及其相对位置的变化。骨骼节点的数据类型以骨骼帧的形式表示,每一帧都是由 25 个骨骼节点组成的 JointType 类型的结合[19],每一个关节点都有 3 种状态,分别为 Tracked、Not Tracked 和 Inferred,用于表示骨骼关节点的信息状态。人体骨骼节点的位置分布如图 15 所示。

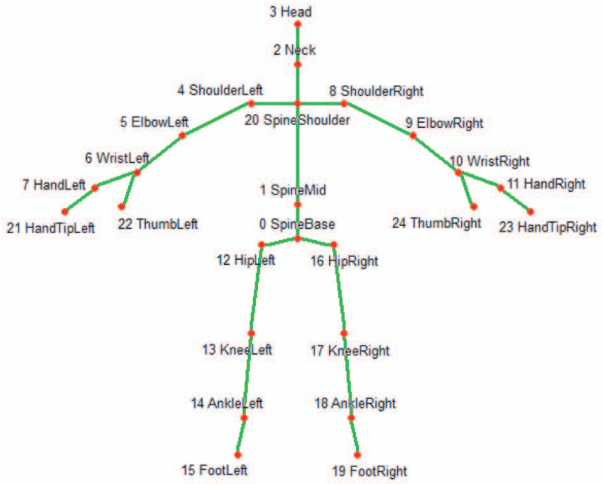


图 15　人体骨骼节点分布图

Kinect 骨骼坐标描述都以传感器坐标系为参考，传感器向前是 Z 轴正向，向上和向右分别为 Y 轴正向和 X 轴正向。

Kinect SDK 使用 HandState 来描述手势，共有 5 种手势状态，该手势指的是手掌的状态，并不是整个手臂的肢体动作。其5种状态分别为Unknown（未知），NotTracked（未跟踪），Open（张开），Closed（闭合），Lasso（半张开）。使用 Kinect 采集彩色图像，进行深度阈值前景分割[20]，得到含有手臂及手的图像，再进行高斯肤色模型手势分割与图像二值化得到手势图像，如图 16所示。

(a) Lasso　(b) Open　(c) Closed

图 16　手势状态图

4.2　空间向量法计算关节角度及距离

本文选取 25 个骨骼节点中的 8 个骨骼点，分别为头、左肩、右肩、左手、右手、左肘、右肘和脊椎中点。假设在某一帧骨骼图像中，取右手关节点坐标、右肘关节坐标、右肩关节坐标，分别以 $HR(x_0, y_0, z_0)$，$ER(x_1, y_1, z_1)$，$SR(x_2, y_2, z_2)$ 表示，3 个关节点示意图如图 17 所示。

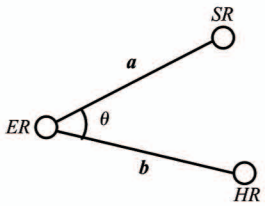


图 17　关节夹角示意图

则向量 $\boldsymbol{a}$ 以 ER 为起点，SR 为终点，向量 $\boldsymbol{b}$ 以 ER 为起点，HR 为终点。两个向量的夹角为 θ，则有：

$$\boldsymbol{a} = (x_2 - x_1,\ y_2 - y_1,\ z_2 - z_1) \tag{5}$$

$$\boldsymbol{b} = (x_0 - x_1,\ y_0 - y_1,\ z_0 - z_1) \tag{6}$$

$$|\boldsymbol{a}| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{7}$$

$$|\boldsymbol{b}| = \sqrt{(x_0 - x_1)^2 + (y_0 - y_1)^2 + (z_0 - z_1)^2} \tag{8}$$

$$\boldsymbol{a} \cdot \boldsymbol{b} = (x_2 - x_1)(x_0 - x_1) + (y_2 - y_1)(y_0 - y_1) + (z_2 - z_1)(z_0 - z_1) \tag{9}$$

$$\cos\theta = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{|\boldsymbol{a}| \cdot |\boldsymbol{b}|} \tag{10}$$

将上述中的式(7) ~ (9)代入式(10)即可求出 3 个骨骼关节点两两连成的向量夹角。式(7)、式(8)为计算两个骨骼关节点之间的距离。

4.3　体感识别

交互系统的动态手势识别过程，例如右手向左挥，首先采集右手在 t 与 $t + \Delta t$ 两个时刻的骨骼位置坐标并计算角度与距离，然后同设定的阈值相比较，判定与规定的动态手势是否相匹配。

手势识别部分定义了左手左滑，右手左滑的动态手势和手部张开、闭合、半张开的静态手势；人体姿态部分定义了左臂高于头部且呈 90 °弯曲，右臂高于头部且呈 90 °弯曲，左臂与身体呈 90 °并且右臂与身体呈 45 °。将动态手势、静态手势、人体姿态相结合，可以实现对视频播放器常用功能的控制。动作与控制指令对照表如表 1 所示。

本系统还解决了在多人场景的情况下，控制者的判定问题。解决方法是使用 Kinect 采集所有在场人的骨骼数据及深度数据，计算每一人距离 Kinect 传感器的距离。距离最近的认为控制者，其他人为观看者，若距离相同，则处于左侧的人为控制者。

表 1　动作与控制指令对照表

人体动作	功能定义
左手向左滑	切换下一视频
右手向左滑	切换上一视频
双手闭合且手位于身体前方	播放/暂停
双手张开且手位于身体前方	切换语音控制
左手闭合,右手张开,左手左滑	快进
左手闭合,右手张开,右手向左滑	快退
左手半张开,右手张开,左手左滑	增加音量
左手半张开,右手张开,右手左滑	降低音量
右臂高于头部且呈 90 °弯曲	屏幕锁定
左臂高于头部且呈 90 °弯曲	解除锁定
左臂与身体呈 90 °且右臂呈 45 °	退出交互系统

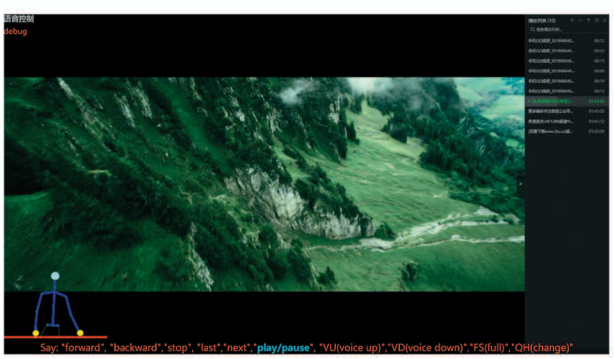
5　实　验

5.1　实验环境

本次实验中硬件部分主要使用的处理器为 Intel(R) Core(TM) i5-4200H CPU @ 2.80 GHz,Win10 64 位计算机和微软 Kinect 2.0 设备,其软件部分主要是基于 OpenCV 视觉库、Kinect V2 开源数据库 Kinect SDK 2.0 和 Intel(R) Distribution of OpenVINO toolkit 2019 R1.1 for Windows,以及在 Visual Studio 2017 编程环境下使用 C#与 C + +编程语言进行编程,视频播放器使用爱奇艺万能播放器作为控制对象。

5.2　实验结果

通过构建实验环境,先后使 3 名实验者依次操作视频播放器交互系统,并分别验证人脸识别的成功率、手势与语音控制相应功能的成功率。系统整体使用效果图如图 18 所示,图 8(a)为视频播放界面,图 8(b)为“播放”功能操作。



(a) 视频播放界面

(b)“播放”功能操作

图 18　系统使用效果图

5.3　实验结果分析

5.3.1　人脸识别结果分析

在人脸识别测试的过程中,实验者进行头部摆动、走动等一系列的动作,以此来检验算法的容错性及鲁棒性。Kinect 传感器采集彩色图像为 30 fps,人脸总类别为 44 种。在不同光照的情况下,选取实验对象 3 人,每人识别 200 次,进行实验测试,其测试结果如表 2 所示。

表 2　人脸识别结果

人脸	光照条件		黑暗条件	
	误检次数	识别率/%	误检次数	识别率/%
Wang	5	97.5	8	96.0
Li	6	97.0	7	96.5
Xu	8	96.0	9	95.5

从表 2 可以看出,在黑暗条件下,本文的人脸识别算法识别率没有明显的波动。3 人识别的准确性均达到较高的水平,其平均识别率达到了 96.41%。

针对本文人脸识别算法的识别率,选取 3 种算法作为对比对象,即算法 1 为主成分分析算法[4]、算法 2为基于 FaceNet 模型人脸识别[14]、算法 3 为神经网络与多哈希相似度加权结合算法[21]。其对比结果如表 3 所示。

实验结果表明,本文的人脸识别算法能够保证识别结果的准确率,相比于其他算法,识别率提高了几个百分点,并且本文算法可以很好地适应黑暗的环境,满足了智能视频播放器的人脸登录系统的应用需求。

表 3　人脸识别算法识别率对比

算法	识别率/%
算法 1	91.20
算法 2	95.12
算法 3	96.20
本文算法	96.41

5.3.2　功能操作结果分析

本次测试中，分别由 3 名实验者根据自己的意图触发视频播放器的某一项功能，其中包括手势与人体姿态操作及语音操作，记录下符合自己意图的操作次数，人体动作操作识别结果如表 4 所示，语音操作识别结果如表 5 所示。

表 4　手势与人体姿态识别结果

动作操作	切换操作		播放/暂停		快进/快退		加减音量		锁定/解锁		关闭系统	
	成功	失败	成功	失败	成功	失败	成功	失败	成功	失败	成功	失败
Wang	50	0	49	1	47	3	45	5	48	2	48	1
Li	48	2	48	2	49	1	47	3	49	1	46	4
Xu	50	0	48	2	44	6	46	4	50	0	49	1
次数	148	2	145	5	140	10	138	12	147	3	144	6
操作成功率/%	98.67		96.67		93.33		92.00		98.00		96.00	

表 5　语音识别结果

语音操作	切换操作		播放/暂停		快进/快退		加减音量		锁定/解锁		关闭系统	
	成功	失败	成功	失败	成功	失败	成功	失败	成功	失败	成功	失败
Wang	43	7	48	1	44	6	41	9	45	5	47	3
Li	47	3	46	4	45	5	45	5	46	4	45	5
Xu	45	5	45	5	47	3	43	7	44	6	46	4
次数	135	15	140	10	136	14	129	21	135	15	138	12
操作成功率/%	90.00		93.33		90.67		86.00		90.00		92.00	

由表 4 可知，对 3 个人识别结果的准确率均达到了较高的水平，其平均识别率为 95.78%。在日常使用中，播放/暂停，切换视频，快进/快退使用频率最高，其识别率均达到 93% 以上。

由表 5 可知，语音控制识别的准确率比手势识别的准确率略低，其原因是该语音识别系统为离线识别，采用标准的英文和普通话，系统对带方言的语音识别效果较差，而且视频播放声音对控制语音产生干扰，其平均识别率为 90.33%。

针对本文手势与人体姿态识别算法的识别率，选取 3 种算法作为对比对象，即算法 4 为基于多特征组合的动态手势识别[22]、算法 5 为利用有限状态机及 DTW 算法的动态手势识别[23]、算法 6 为融合表面肌电和加速度的手势动作识别[24]。其对比结果如表 6 所示。由表可知，本文算法较其他算法提高了对手势的识别率，实验方法可行有效。

表 6　手势与人体姿态识别算法识别率对比

算法	识别率/%
算法 4	94.13
算法 5	95.00
算法 6	91.20
本文算法	95.78

针对本文语音识别算法的识别率，选取 3 种算法作为对比对象，即算法 7 为基于改进 CNN 的中文语音识别[25]、算法 8 为端到端的深度卷积神经网络语音识别[26]、算法 9 为神经网络-隐马尔可夫混合系统中使用 DBLSTM（深度双向 LSTM 模型）语音识别[27]。其对比结果如表 7 所示。由表可知，本文算法提高了语音识别效果。因为本文只涉及到语音命令的识别，并不对自由形式的语音进行识别，所以识别率较高。

表 7　语音识别算法识别率对比

算法	识别率/%
算法 7	89.57
算法 8	82.60
算法 9	86.90
本文算法	90.33

总体来说,本文手势与人体姿态识别和语音识别成功率已经满足基本要求,符合日常对视频播放器的操作需求。

6　结　论

本文利用 Kinect 传感器获取数据,对视频播放器建立了人机交互系统,系统使用 FaceNet 与 SVM 结合算法提高了人脸识别率,并使用 OpenVINO 实现系统实时运行,可对使用者进行人脸识别,实现播放器的登录验证。系统在语音控制部分,可识别中文和英文的语音命令,支持普通话及标准英语,并且可离线运行。在手势与人体姿态控制部分,静态手势、动态手势和人体姿态相结合,即使用户动作过快,系统仍可以及时作出响应,可完成所有的视频播放器预定功能,并且避免了光照和复杂背景的影响。系统稳定性良好,运行流畅,基本满足应用需求。本交互系统可以应用到智能家居、虚拟现实、机器人操作等领域,具有实际推广性。

本文的人脸识别,并没有考虑照片欺骗检测,在语音识别部分,没有加入方言语音识别,当使用方言进行语音控制时,识别率较低,以上问题将是本文下一步的研究方向。

参考文献

[1] Lee C H, Kim J H, Cho S B, et al. Development of real-time hand gesture recognition for tabletop holographic display interaction using Azure Kinect[J]. *Sensors*, 2020, 20(16): 4566

[2] 齐帅,潘克刚,齐宝峰, 等. 交互机器人技术与发展[J]. 通信技术, 2020,53(6):1449-1453

[3] 吴晓雨,杨成,冯琦. 基于 Kinect 的手势识别算法研究及应用[J]. 计算机应用与软件,2015,32(7):173-176,276

[4] 李健,路飞,田国会,等. 基于 Kinect 的 PPT 全自动控制系统研究[J]. 计算机工程与应用,2013,49(17):133-138

[5] 陈一新. 基于 Kinect 的手势识别技术在人机交互中的应用研究[D]. 成都:西南交通大学信息科学与技术学院,2015:36-42

[6] 李佳怡,刘东旭. 基于 Kinect 体感识别的自衡车交互系统[J]. 实验室研究与探索,2019,38(7):77-79,144

[7] Ma F, Sun Z, Wang H. The design and implementation of natural human-robot interaction system based on Kinect sensor[C]// International Conference on Logistics Engineering, Shenyang, China, 2015:242-246

[8] 张诗潮,钱冬明. 体感技术现状和发展研究[J]. 华东师范大学学报(自然科学版),2014(2):40-49,126

[9] Alzahrani M S, Jarraya S K, Abdallah H B, et al. Comprehensive evaluation of skeleton features based fall detection from Microsoft Kinect v2[J]. *Signal, Image and Video Processing*, 2019,13(7):1431-1439

[10] Liu S, Song Y, Zhang M Y, et al. An identity authentication method combining liveness detection and face recognition[J]. *Sensors*,2019,19(21):4733

[11] 卢晓静,陈华华. 基于 HOG-LBP 特征提取的人脸识别研究[J]. 杭州电子科技大学学报,2013,33(3):25-28

[12] 李文娜. 基于 Haar 特征级联强分类器和肤色模型的人脸检测[J]. 辽宁石油化工大学学报,2010,30(3):61-64

[13] Liu W, Anguelov D, Erhan D, et al. SSD: single shot multiBox detector[C]//European Conference on Computer Vision, Amsterdam, Netherlands, 2016:21-37

[14] Schroff F, Kalenichenko D, Philbin J. Facenet: a unified embedding for face recognition and clustering[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Santiago, Chile, 2015:815-823

[15] Bsat M, Sim T, Baker S. The CMU pose, illumination, and expression database[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2003, 25(12): 1615-1618

[16] 孙峰毅. 医学图像伪影消除的图像增强的神经网络算法[D]. 厦门: 华侨大学信息科学与工程学院, 2019: 23-28

[17] 李淑. 基于卷积神经网络的视频人脸检测与识别[J]. 电脑知识与技术,2018,14(21):210-211,216

[18] 朱荣, 李小映. 基于 Kinect 的语音识别技术研究[J]. 计算机与数字工程,2017,45(6):1211-1215

[19] 党宏社,侯金良,强华,等. 基于 Kinect 的人体动作识别算法研究[J]. 电子器件,2017,40(5):1309-1313

[20] Ding I J, Chang Y J. HMM with improved feature extraction-based feature parameters for identity recognition of gesture command operators by using a sensed Kinect-data stream[J]. *Neurocomputing*, 2017, 262:108-119

[21] 邓良,许庚林,李梦杰,等. 基于深度学习与多哈希相似度加权实现快速人脸识别[J]. 计算机科学,2020,47(9):163-168

[22] 曹海婷,戎海龙,焦竹青,等. 基于多特征组合的动态手势识别[J]. 计算机工程与设计,2018,39(6):1727-1732

[23] 千承辉,邵晶雅,夏涛,等. 基于 Kinect 的手语识别方法[J]. 传感器与微系统,2019,38(6):31-34,38

[24] 鲍磊,罗志增,席旭刚,等. 融合表面肌电和加速度的手势动作识别[J]. 传感技术学报,2019,32(12):1843-1848,1863

[25] 查兴兴,陈恳. 基于改进 CNN 的中文语音识别研究[J]. 无线通信技术,2019,28(4):40-44

[26] 刘娟宏,胡彧,黄鹤宇. 端到端的深度卷积神经网络语音识别[J]. 计算机应用与软件,2020,37(4):192-196

[27] Graves A, Jaitly N, Mohamed A R. Hybrid speech recognition with deep bidirectional LSTM[C]// 2013 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU 2013), Olomouc, Czech Republic, 2013:273-278

Interactive system of intelligent video player based on Kinect multi-biometrics

Li Guoyou, Wang Weijiang, Li Chenguang, Hang Bingpeng, Yang Mengqi

(School of Electrical Engineering, Yanshan University, Qinhuangdao 066004)

Abstract

In order to realize the intelligent control of the video player under the non-contact operation of the viewer, the system uses Kinect sensors to collect color images, uses FaceNet to extract facial feature vectors, and performs face recognition after support vector machine (SVM) training. This process is under the computer central processing unit (CPU) operating environment. OpenVINO is used to realize real-time operation of face detection and recognition, which is used for login verification of video players. The audio data collected by the system uses Speech Platform Runtime v11 for Chinese command recognition, and Kinect Speech Language for English command recognition, thereby realizing voice control. Bone data is collected, the distance and angle between the bone points are calculated to recognize human postures and gestures, and the recognition results are converted into control commands to realize the player's common control functions such as fast forwarding, switching videos, and adding and subtracting volume. The experimental results show that the interactive system realizes the user's contactless full-automatic human body control, and provides a natural and convenient way of interaction for the video player.

Key words: Kinect V2 sensor, OpenVINO, face recognition, player control, speech recognition, gesture recognition