文章编号：1006-3080(2020)04-0556-08 DOI：10.14135/j.cnki.1006-3080.20190619002

多特征非接触式测谎技术

魏江平， 林家骏， 陈 宁
（华东理工大学信息科学与工程学院，上海 200237）

摘要：为提高测谎准确率，提出了一种基于多模态信息融合的测谎模型。在该模型的支持下，仅需对测谎者在说话期间的视频与音频信号进行处理即可有效完成测谎评估任务。心率可以反映测谎者的情绪变化，通过光电体积描记(Photo-Plethysmography，PPG)方法及全连接网络提取心率变化特征，通过3D卷积神经网络(3D-Convolutional Neural Networks，3D-CNN)及Word2Vec+CNN提取视频与语义特征，并将特征进行融合；使用线性支持向量机(Linear Support Vector Machines，L-SVM)对融合后的特征进行分类。在开源Real-life Trial数据集上的仿真实验结果表明，与其他多模态模型相比，本文提出的测谎模型在三模态下的准确率提升了2.74%。

关键词：多模态模型；说谎检测；L-SVM；非接触式
中图分类号：TP391 **文献标志码：**A

说谎是人类通过虚假陈述、扭曲事实和遗漏等形式误导别人的一种特殊行为。自动检测说谎是计算机语言学、心理学、军事、情报机构等各学科研究的重要领域。由于人类检测说谎的能力几乎为随机猜测，因此需要科学、可靠的自动化方法检测说谎。自动检测说谎技术通过自动分析人们的行为、语言、以及各种生理指标对人们是否说谎作出判断，其在刑事案件处理、医疗以及司法等职业中起着至关重要的作用，具有广阔的应用场景。

目前的说谎检测方法大致分为两大类：一是基于言语线索的检测方法；二是基于非言语线索的检测方法。基于言语线索的检测方法主要是通过分析语法以及词性等特征来检测真话和假话[1]。文献[2]提出了基于语言探究和字数统计词典的心理语言学特征用于测谎。Newman等[3]发现说谎者使用更多的负面情绪词。还有人提出不同的语言特征(字数、词性和句子统计特征)[4-5]以及文本句法复杂性等都与说谎存在联系[6]。Pérez-Rosas等[7]基于语词计量文本分析工具LIWC(Linguistic Word Count)发现说谎者在讲述过程中比讲真话者更加自信。

基于非言语线索的检测方法虽没有基于言语线索检测方法多，但在检测说谎方面也取得了很大成功，主要分为3类：基于生理、视觉和声音线索。基于生理的检测方法包括使用测谎仪[8]、热成像方法测量面部血流量和面部皮肤温度[9-11]以及使用脑功能磁共振成像(Functional Magnetic Resonance Imaging，FMRI)测量脑血流量等[12]。这些方法都需要测试者配合且设备昂贵，另外操作人员需具备专业知识。基于声音的检测方法包括利用声压分析器(Voice Stress Analysis，VSA)和分层声音分析技术两种商业产品对人体声带进行操作来测谎[1]。有相关研究表明，音高、持续时间、能量以及说话过程中的停顿[13-16]可表明说谎信息。基于视频的检测方法近年来也越来越受到关注。Depaulo等[17]发现瞳孔扩张是一种表明说谎的行为。面部微表情如嘴唇突出翘起以及一些标志性手势也被认为是说谎的一类标志[7, 18-21]。

多模态测谎也不断受到关注。Pérez-Rosas等[7]引入了一个新的说谎数据库，包含法庭审判的真实

收稿日期：2019-06-19
基金项目：国家自然科学基金(61771196)
作者简介：魏江平(1995—)，女，四川人，硕士生，主要研究方向为机器学习和测谎技术研究。E-mail：18701729045@163.com
通信联系人：林家骏，E-mail：jjlin@ecust.edu.cn
引用本文：魏江平，林家骏，陈 宁. 多特征非接触式测谎技术[J]. 华东理工大学学报(自然科学版)，2020，46(4)：556-563.
Citation：WEI Jiangping，LIN Jiajun，CHEN Ning. Multi-feature Non-contact Deception Detection Technology[J]. Journal of East China University of Science and Technology，2020，46(4)：556-563.

审判场景视频, 并且通过提取文本、手势和面部动作等特征, 评估了不同模态特征对测谎的重要性。文献 [19, 22-24] 结合声学、视觉、文本模态提出了不同的多模态模型来检测说谎, 其中文献 [19] 除通过 Glove 对文本进行词编码来获取文本的词向量表示以及提取了音频 MFCC 等基本特征外, 还主要关注了可表明说谎信息的动态运动特征等; 文献 [23] 除提取了音频、视频、文本模态的基本特征外, 还采用了人工标注的微表情特征。文献 [24] 采用 CNN 和 LSTM 深度学习模型来提取音频和视频特征。这些模型结构相对比较复杂, 且在实际场景中不可能采用由人工标注的微表情特征, 模型的分类准确率并不理想。文献 [25-27] 提出心率变化与说谎有关, 可较好地反映说话者内心情绪的变化, 但心率等生理特征通常都是通过电子仪器设备这种接触式的方法来获得。实际的应用场合往往不允许接触式测谎, 因此开发一种易于部署的非接触式测谎系统成为必然。

本文提出了一种基于视频、心率、文本三模态融合的非接触式测谎模型, 采用线性支持向量机 (Linear Support Vector Machines, L-SVM) 作为分类器。该多模态模型未使用人工标注特征, 且引入了心率进行非接触式说谎检测, 在降低模型复杂度的同时提高测谎的准确率, 实验结果表明本文模型相较于其他模型有效地提高了测谎准确率。

1　基本理论

1.1　光电体积描记 (PPG) 技术

早期的研究工作中, 将心率用于测谎时都是通过生理传感器来测量心率, 这种接触式方法会让测试者存在防备心理。本文采用非接触式 PPG 技术从视频图像的局部区域中提取心率值, 其中正常环境光作为光源[25]。其原理是血液比周围组织吸收更多的光, 血液体积的变化影响着入射光和反射光。面部血管扩张, 入射光路径长度增加, 反射光强度也随着变化, 即血容量的变化通过反射光亮度值的变化体现出来, 反射光强度的变化反映在图像像素值的变化上。

1.2　3D 卷积神经网络 (3D-CNN)

与 2D-CNN 相比, 3D-CNN 更容易检测出视频图像中的微妙表情或者肢体动作。2D-CNN 一般是通过分析视频的每一帧信号进行识别, 没有考虑时间维度上的帧间动作信息。而 3D-CNN 通过 3D 卷积操作, 可同时获取空间和时间维度上的特征, 捕获多个连续帧编码之间的动作信息[28]。3D-CNN 中进行卷积操作的输出如式 (1) 所示。

$$z_{ij}^{xyz}=\tanh\left(b_{ij}+\sum_{m}\sum_{p=0}^{P_i-1}\sum_{q=0}^{Q_i-1}\sum_{r=0}^{R_i-1}w_{ijm}^{pqr}z_{(i-1)m}^{(x+p)(y+q)(z+r)}\right) \tag{1}$$

其中: tanh(·) 为双曲正切函数; b_{ij} 为特征映射偏差; m 表示第 $i-1$ 个卷积层连接到当前特征映射的特征映射集合; P_i、Q_i 分别为卷积核的高度、宽度; R_i 为 3D 卷积核沿着时间维度的大小; w_{ijm}^{pqr} 为 (p,q,r) 处的核连接到前一个卷积层第 m 个特征映射的值, i 表示第 i 个卷积层, j 表示第 j 个特征映射; (x,y,z) 表示进行卷积操作的坐标位置。本文通过 3D 卷积神经网络来识别视频面部表情动作, 提取基于整个面部的视频特征。

1.3　Word2Vec

由 Google 公司提出的 Word2Vec 模型在获取词向量工作方面取得了很大的成功, 它的主要优势是可使语义相接近的词语或短语之间的距离更小, 可较好地度量词与词之间的相似度。Word2Vec 主要包含 Skip-Gram 与 CBOW 两种模型。Skip-Gram 模型的输入是一个句子中的某个词语, 然后预测该词语的上下文。CBOW 模型则相反, 输入的是句子中某个词语的上下文, 然后根据上下文来预测出目标单词。本文采用基于 CBOW 的 Word2Vec 模型进行词向量映射。Word2Vec 模型是在 Google News 数据集上预训练好的, 该数据集包含一百多万个英文的短语和词语。利用 Word2Vec 模型的训练结果可得到每个词语的词向量, 将文本中每个词语映射到维度一定的向量空间中, 保存样本单词之间的语义信息。

2　特征提取

2.1　基于局部面部的心率特征提取

2.1.1　心率提取　心率反映了测谎者内心情绪的变化, 对于测谎有很大的帮助。非接触式 PPG 技术首先通过读取 AVI 格式的视频信号, 使用面部自动跟踪器检测视频帧内的人脸并定位到感兴趣的测量区域 (Regions of Interest, ROI)。借助图像处理工具包 Opencv 库和具有类似 Haar 数字图像特征的级联增强分类器来获取 ROI 的坐标以及高度、宽度。对 ROI 区域中 R、G、B 3 个通道的像素值分别进行空间平均[26] 以提高信噪比。计算 ROI 区域中 R、G、B 3 个通道的像素均值, 将每个图像帧的画面信息转变成点信息, 得到 3 个通道的脉动信号。假设在时刻 t, 脉动信号中 R、G、B 3 个通道的信号幅度分别为 $s_1(t)$、

$s_2(t)$、$s_3(t)$(感兴趣测量区域像素值的平均值),则脉动信号如式 (2) 所示。

$$x(t)=\sum_{j=1}^{3}a_j s_j(t) \tag{2}$$

其中:a_j 代表每个通道的权重。G 通道具有最强的体积描记信号[26],则 G 通道权重值最大。本文仅对 G 通道信号取灰度均值,然后进行快速傅里叶变换以获得脉动信号的功率谱密度。功率谱中最高功率对应的频率则代表了脉冲频率,即可得到每一帧视频信号对应的心率值。假设得到的频率值为 f,则根据式 (3) 可得到每一帧视频信号的心率值 (频率表示的是 1 s 内完成周期性变化的次数,心率是指每分钟心跳的次数)。

$$\text{HeartRate}=60f \tag{3}$$

文献 [27] 证明了在整个面部、前额、眼角周围这 3 个区域中,前额区域提供了更为丰富的信息,所以本文选择前额区域为 ROI。在检测到人脸后定位额头,借助非接触式 PPG 技术得到每一帧视频信号的心率值,将每一帧信号的心率值拼接组合成一维向量,此一维向量包含了一个视频样本所有帧的心率值。为防止人脸检测错误影响算法性能,若当前帧没有检测到人脸额头,则将前一帧的额头坐标返回。若检测到多个面部和额头,则选择最接近前一帧的额头坐标返回,且为了能更准确地获取心率值,只有当视频帧长大于 10 帧时才开始获取心率值。

2.1.2 心率特征提取 通过 PPG 技术获得了每个样本的心率值向量后,为获得心率变化特征,采用全连接网络来获取反映心率变化情况的心率特征。全连接网络包括 1 个输入层,4 个隐藏层以及 1 个输出层。将获取的心率值向量作为输入,然后经过隐藏层。4 个隐藏层分别具有 1 024、1 024、512、300 个神经元,都采用 ReLU 作为激活函数,且每一层都采取 Dropout 为 0.5 的措施以防止过拟合。采用随机梯度下降优化算法训练模型,训练损失函数为最小交叉熵。将训练好的模型的最后一个隐藏层的输出作为心率特征,最终得到的心率特征是长度为 300 的一维特征向量。

2.2 基于整个面部的视频特征提取

3D-CNN 模型的输入由一系列图像帧组成,在输入图像帧之间生成多个通道信息,最终的视频特征由所有通道信息组合得到。3D-CNN 网络结构如图 1 所示。首先,以原始视频图像作为输入,输入的视频维度为 (C, N, H, W),其中 C 表示信道数,实验中输入的是彩色图像帧,有 R, G, B 3 个通道,C=3;N 表示一次输入的图像帧数,取值 30;H 和 W 分别表示输入的每一帧图像的高度和宽度,取值均为 96。输入的图像帧首先通过一个硬连线层得到 5 种不同的特征,分别是灰度、X 方向的光流、Y 方向的光流、X 方向的梯度、Y 方向的梯度,即形成 5 个不同的通道。硬连线层使用一个固定 Hardwired 核对输入帧进行处理,获得多个通道信息。光流和梯度分别表明物体运动趋势和图像边沿分布,3D-CNN 模型正是通过获取光流和梯度这两种信息来识别视频行为。硬连线层的输出经过一个卷积层对 5 个通道的特征分别进行卷积操作,卷积核的大小为 (M, C, L, F_h, F_w),产生一个维度为 (M, C, N–L+1, H–F_h+1, W-F_w+1) 的输出。其中 M 表示特征映射的数量,L 表示执行一次 3D 卷积操作的图像帧数,F_h, F_w 分别表示卷积核的高度和宽度,实验中采用的卷积核大小为 (32, 3, 5, 5, 5)。卷积层的输出经过一个窗口大小为 3×3×3 的最大池化层,然后经过一个具有 300 个神经元且激活函数为 Softmax 的全连接层,全连接层的输出即为提取出的视频特征。最终得到的视频特征是长度为 300 的一

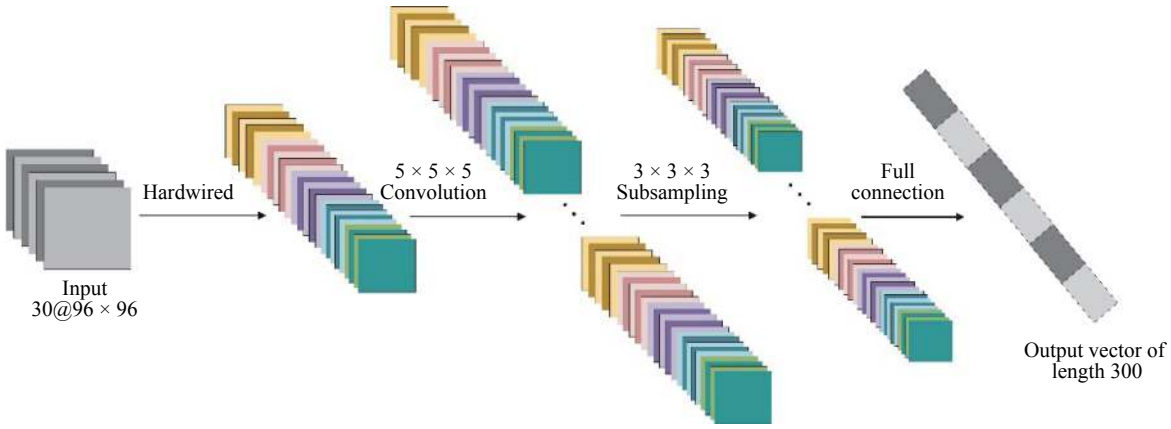

图 1 3D-CNN 网络结构

Fig. 1 Network structure of 3D-CNN

维向量。

2.3　文本语义特征提取

早期的研究工作已经证明了文本含有丰富的语义信息, 对于测谎很有帮助, 因此可以用来测谎[2-5]。受到文献 [23] 启发, 本文采用图 2 所示的模型对文本进行分析, 提取其语义特征。首先以文本作为输入, 通过 Word2Vec 模型对原始文本进行词向量映射, 得到文本中每个单词的词向量; 然后将这些词向量进行拼接得到每个样本的词向量矩阵; 最后使用 CNN 模型进一步提取词向量矩阵上下文的语义相关性信息。假设一个样本包含了 N 个单词, 经过 Word2Vec 模型进行词向量表示后得到词向量矩阵 $\boldsymbol{M}\in\mathbf{R}^{N\times L}$, L 表示词嵌入维度, 取值 300。$\boldsymbol{m}_j\in\boldsymbol{M}$ 为文本中第 j 个单词的向量表示, 即矩阵 $\boldsymbol{M}$ 的第 j 行。由于 CNN 模型的输入要求是固定大小的矩阵, 所以实验中以具有最多单词数的样本为基准, 单词数不足的样本采取补零措施, 使得每个样本得到的词向量矩阵 $\boldsymbol{M}$ 大小相同。将矩阵 $\boldsymbol{M}$ 作为 CNN 模型的输入, 通过卷积层、最大池化层以及全连接层获取语义特征向量。其中卷积层使用了 3 种不同尺寸大小的卷积核, 分别为 3×3、5×5、8×8, 卷积核个数都为 20。将卷积层的输出经过窗口大小为 2×2 的最大池化层; 然后通过 Flatten 将池化层的输出融合成一维向量。将 3 种卷积池化层的输出直接拼接成一维长向量, 该向量经过具有 300 个神经元且激活函数为 ReLU 的全连接网络层。将全连接网络层的输出表示为文本语义特征, 即得到的文本语义特征是长度为 300 的一维向量。

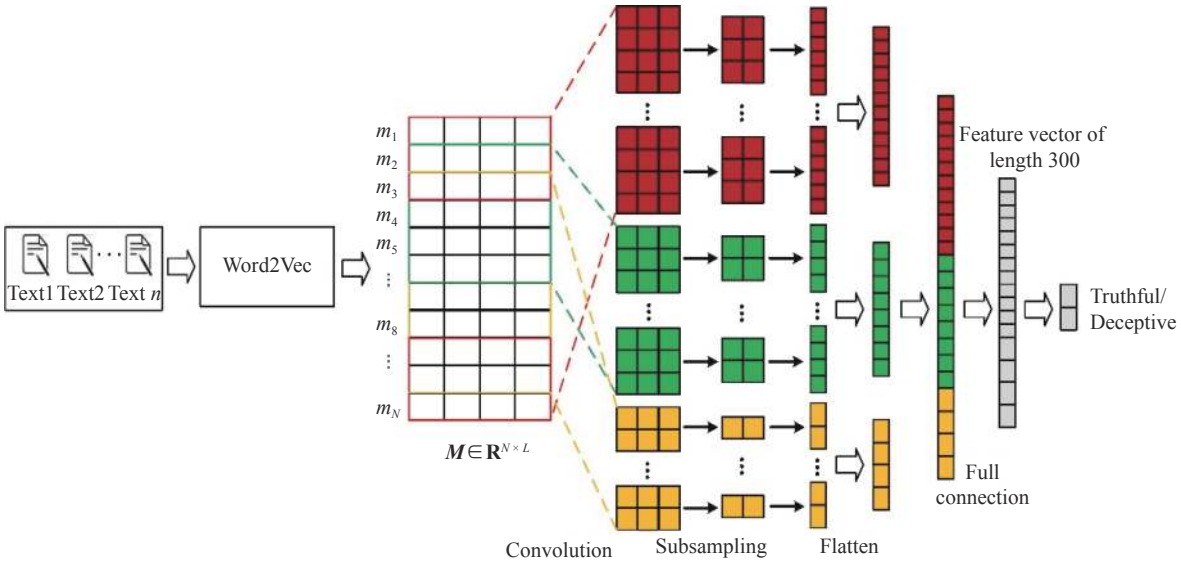

图 2　Word2Vec+CNN 网络结构

Fig. 2　Network structure of Word2Vec+CNN

2.4　模型描述与特征融合

本文采取两种不同的融合方法将各个模态的特征进行融合, 得到特征融合向量 $\boldsymbol{F}_{\mathrm{C}}$ 或 $\boldsymbol{F}_{\mathrm{H+C}}$, 不同模态之间通过信息互补能更好地检测说谎。

心率、视频、文本 3 个模态中每一个样本的特征向量都是长度为 300 的一维向量, 分别表示为 $\boldsymbol{F}_{\mathrm{h}}$、$\boldsymbol{F}_{\mathrm{v}}$、$\boldsymbol{F}_{\mathrm{t}}$。受到文献 [23] 启发, 实验采用了两种传统的特征融合方式: 一种是直接拼接, 通过直接拼接得到的特征向量是长度为 900 的一维向量, 可表示为 $\boldsymbol{F}\mathrm{c}=[\boldsymbol{F}_{\mathrm{h}}, \boldsymbol{F}_{\mathrm{v}}, \boldsymbol{F}_{\mathrm{t}}]$, 采用这种融合方式的模型称为 L-SVM$_{\mathrm{C}}$; 另一种融合方式是哈达玛积, 该方法可降低特征长度, 融合得到的特征向量 $\boldsymbol{F}_{\mathrm{H+C}}$ 如式 (4) 所示。

$$\begin{aligned}\boldsymbol{F}_{\mathrm{H+C}}&=[F_{\mathrm{h1}}, F_{\mathrm{h2}}, \cdots, F_{\mathrm{h300}}][F_{\mathrm{v1}}, F_{\mathrm{v2}}, \cdots, F_{\mathrm{v300}}][F_{\mathrm{t1}}, F_{\mathrm{t2}}, \cdots, F_{\mathrm{t300}}]=\\&[F_{\mathrm{h1}}\times F_{\mathrm{v1}}\times F_{\mathrm{t1}},\ F_{\mathrm{h2}}\times F_{\mathrm{v2}}\times F_{\mathrm{t2}}, \cdots, F_{\mathrm{h300}}\times F_{\mathrm{v300}}\times F_{\mathrm{t300}}]=\\&[F_{\mathrm{H+C_1}}, F_{\mathrm{H+C_2}}, \cdots, F_{\mathrm{H+C_{300}}}]\end{aligned}\tag{4}$$

通过哈达玛积融合方法最终得到的特征向量是长度为 300 的一维向量, 采用这种融合方式的模型称为 L-SVM$_{\mathrm{H+C}}$。

CNN 作为分类器一般需要较多的训练样本, 其优势在于多分类问题。本文采用 CNN 模型和 L-SVM 模型对三模态融合向量 $\boldsymbol{F}_{\mathrm{H+C}}$ 进行实验比较。从表 1 的实验结果中可以看出, 采用 L-SVM 可达到更高的精确度。图 3 示出了本文提出的多模态模型框架图。

表 1　不同分类模型对测试集的客观评价指标对比

Table 1　Comparison of objective evaluation indicators of different classification models on test sets

Model	Accuracy/%
CNN	84.80
L-SVM$_{\mathrm{H+C}}$	**98.88**

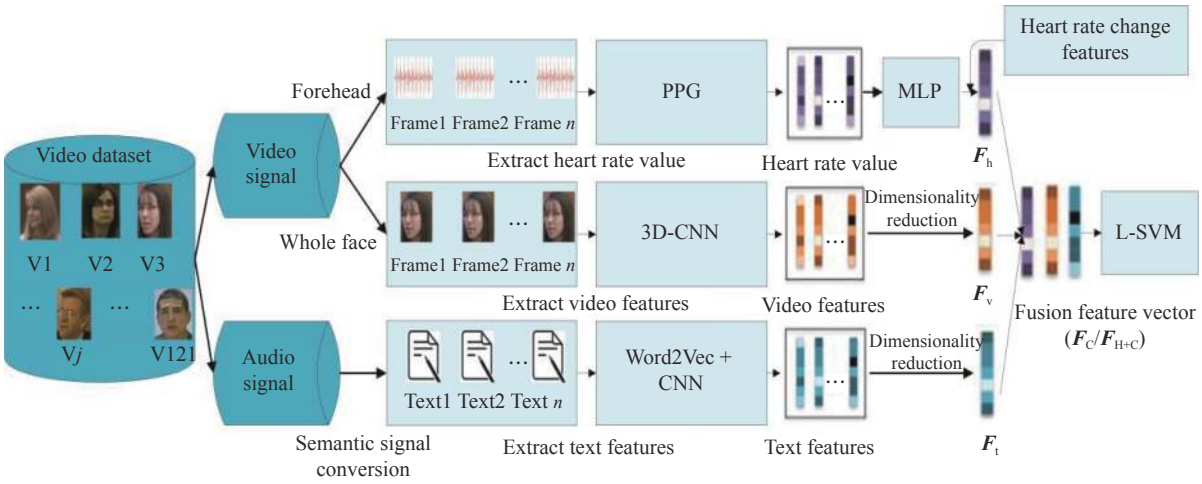

图 3 多模态模型框架图

Fig. 3 Framework of multi-modal model

3 实验结果与分析

3.1 数据集与评价标准

实验采用文献 [7] 中由真实法庭审判视频组成的 Real-life Trial 数据集，该数据集由 121 个法庭审判视频剪辑组成，其中包含 61 个说谎视频、60 个说真话视频。数据集中视频的平均长度为 28.0 s，说谎话和说真话的视频平均长度分别为 27.7 s 和 28.3 s。将 mp4 格式的视频数据转换成 WAV 格式的音频数据，即得到音频数据集；文本数据集由音频信号的手工转录得到。实验中对数据集的划分采用以下 3 种方式：

(1) 留一法交叉验证[7]。假设共有 D 个样本，每次取出一个样本作为测试集，剩余样本作为训练集，直到每个样本都作过测试集，总共计算 D 次，然后将 D 次测试结果求平均值作为最终实验结果。留一法虽然计算复杂，但其不受随机样本划分的影响，样本利用率较高，适合小样本的分类问题。

(2) 随机抽取 10 个样本作为测试集，剩余样本作为训练集，进行 10 次划分，然后取 10 次实验结果的平均值作为最终实验结果[24]。

(3) 该数据集由不同的测试者组成，为防止同一个测试者对应的样本同时被分到训练集和测试集中，实验中划分训练集和测试集时，采用了对测试者进行 10 折交叉验证而不是对样本进行交叉验证，每次将 9/10 测试者对应的样本作为训练集，1/10 测试者对应的样本作为测试集，进行 10 次实验，然后取 10 次实验结果的平均值作为最终实验结果。

采用测试分类准确度以及 AUC 值作为客观评价指标。

3.2 实验结果分析

为评估不同模态的特征组合对测谎模型性能的改善，使用各个模态的特征组合进行实验，并将其与其他多模态模型[7, 19, 23-24] 进行比较。文献 [7]、文献 [19, 23]、文献 [24] 分别采用 3.1 节中的 3 种数据集划分方式。表 2、表 3、表 4 给出了 3 种数据集划分方式的实验结果对比。表中 L-SVM$_C$ 表示采用直接拼接特征融合方法的模型，L-SVM$_{H+C}$ 表示采用哈达玛积特征融合方法的模型。针对第 3 种数据集划分方式，本文给出了不同模态特征组合矩阵相乘后的 ROC(Receiver Operating Characteristic) 图，如图 4 所示。相比之下，本文提出的多模态测谎模型的测试准确率与 AUC 值相对较高。

表 2 留一法的客观评价指标对比

Table 2 Comparison of objective evaluation indicators of leave-one-out cross-validation

Model	Accuracy/%
DT[7]	75.2
Heart rate+Video(L-SVM$_C$)	81.32
Heart rate+Video(L-SVM$_{H+C}$)	82.23
Heart rate+Text(L-SVM$_C$)	96.11
Heart rate+Text(L-SVM$_{H+C}$)	97.19
Heart rate+Text+Video(L-SVM$_{H+C}$)	**98.51**

实验结果表明，在 L-SVM$_{H+C}$ 模型基础上，本文三模态模型相较于双模态以及文献 [7, 19, 23-24] 模型的预测精度明显提升，具有更高的测试精确度以及 AUC 值。其中文本与心率模态组合的测试精确度为 96.89%~97.70%，比文献 [23] 的静态模型以及文献 [7, 19, 24] 的模型高 6%~22%，同时 AUC 值达到

表 3　随机抽取方式的客观评价指标对比

Table 3　Comparison of objective evaluation indicators of random extraction method

Model	Accuracy/%
DEV[24]	84.16
Heart rate+Video(L-SVM$_C$)	78.10
Heart rate+Video(L-SVM$_{H+C}$)	83.70
Heart rate+Text(L-SVM$_C$)	96.80
Heart rate+Text(L-SVM$_{H+C}$)	97.70
Heart rate+Text+Video(L-SVM$_{H+C}$)	**98.30**

表 4　10 折交叉验证方式的客观评价指标对比

Table 4　Comparison of objective evaluation indicators of ten-fold cross-validation method

Model	Accuracy/%	AUC
LR[19]	-	0.922 1
MLP (Static)[23]	90.99	0.934 8
MLP|(Non-static)[23]	96.14	0.979 9
Heart rate+Video(L-SVM$_C$)	71.22	0.716 0
Heart rate+Video(L-SVM$_{H+C}$)	73.46	0.730 0
Heart rate+Text(L-SVM$_C$)	95.57	0.955 6
Heart rate+Text(L-SVM$_{H+C}$)	96.89	0.968 3
Heart rate+Text+Video(L-SVM$_{H+C}$)	**98.88**	**0.988 3**

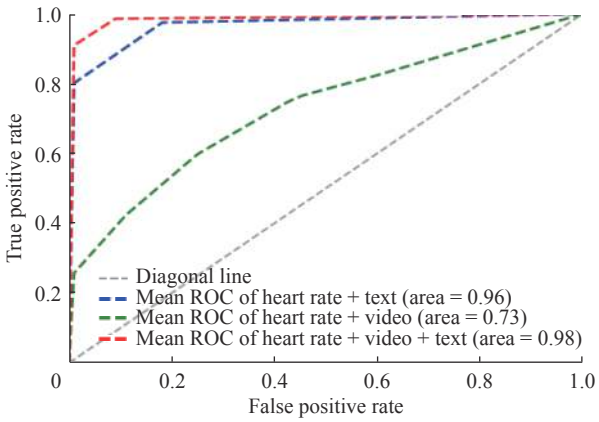

图 4　不同模态特征组合的 ROC 图

Fig. 4　ROC diagram of combination of different modal features

了 0.968 3; 在第 3 种数据集划分方式的基础上, 文本、视频、心率模态的组合取得了 98.88% 的测试精确度及 0.988 3 的 AUC 值。实验结果表明, 三模态模型都表现出了更好的性能, 比文献 [23] 的动态模型以及文献 [7, 19, 24] 模态的测试精度高出了 3%~23.3%。从 ROC 曲线图中也可以看出, 3 个模态的特征组合明显改善了模型性能, 心率、文本、视频特征组合 ROC 曲线同时包含了心率和文本的特征组合 ROC 曲线以及心率和视频特征组合 ROC 曲线。实验结果表明本文提出的多模态模型可有效提高测试准确率, 可更好地检测说谎。

4　结束语

本文提出了一种新的多模态非接触式测谎模型, 通过整合心率、视频和文本特征来检测说谎。实验结果表明从中提取的心率特征和文本特征可能是检测说谎的显著指标。此外, 整合不同的特征来检测说谎可以显著提高检测性能, 特别是心率和文本特征的组合取得了很高的精确度。与其他对数据敏感的检测模型一样, 该模型也存在局限性。使用更有效的检测特征, 收集更大规模的说谎人场景数据库以进一步提高模型的检测精度与泛化能力是未来的研究方向。

参考文献:

[1] FITZPATRICK E, BACHENKO J, FORNACIARI T. Automatic Detection of Verbal Deception[M]. USA: Morgan & Claypool Publisher, 2015.

[2] PENNEBAKER J W, FRANCIS M E, BOOTH R J. Linguis tic inquiry and word count: LIWC 2001[EB/OL]. scholar. google. cn, 2001-03-28 [2019-09-04]. http://www.depts.ttu.edu/psy/lusi/files/LIWCmanual.pdf.

[3] NEWMAN M L, PENNEBAKER J W, BERRY D S, *et al*. Lying words: Predicting deception from linguistic styles[J]. Personality and Social Psychology Bulletin, 2003, 29(5): 665-675.

[4] ZHOU L, BURGOON J K, NUNAMAKER J F, *et al*. Automating linguistics-based cues for detecting deception in text-based asynchronous computer-mediated communications[J]. Group Decision and Negotiation, 2004, 13(1): 81-106.

[5] MIHALCEA R, PULMAN S. Linguistic ethnography: Identifying dominant word classes in text[C]//International Conference on Intelligent Text Processing and Computational Linguistics. Berlin, Heidelberg: Springer, 2009: 594-602.

[6] YANCHEVA M, RUDZICZ F. Automatic detection of deception in child-produced speech using syntactic complexity features[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Bulgaria: Association for Computational Linguistics, 2013: 944-953.

[7] PÉREZ-ROSAS V, ABOUELENIEN M, MIHALCEA R,

et al. Deception detection using real-life trial data[C]//Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. USA: ACM, 2015: 59-66.

[8] ABRAMS S. The Complete Polygraph Handbook[M]. England: Lexington Books/DC Heath and Com, 1989.

[9] ABOUELENIEN M, PÉREZ-ROSAS V, MIHALCEA R, *et al*. Deception detection using a multimodal approach[C]// Proceedings of the 16th International Conference on Multimodal Interaction. USA: ACM, 2014: 58-65.

[10] PAVLIDIS I, EBERHARDT N L, LEVINE J A. Human behaviour: Seeing through the face of deception[J]. Nature, 2002, 415(6867): 35.

[11] POLLINA D A, DOLLINS A B, SENTER S M, *et al*. Facial skin surface temperature changes during a "concealed information" test[J]. Annals of Biomedical Engineering, 2006, 34(7): 1182-1189.

[12] SIMPSON J R. Functional MRI lie detection: Too good to be true?[J]. The Journal of the American Academy of Psychiatry and the Law, 2008, 36(4): 491-498.

[13] APPLE W, STREETER L A, KRAUSS R M. Effects of pitch and speech rate on personal attributions[J]. Journal of Personality and Social Psychology, 1979, 37(5): 715-727.

[14] GRACIARENA M, SHRIBERG E, STOLCKE A, *et al*. Combining prosodic lexical and cepstral systems for deceptive speech detection[C]//2006 IEEE International Conference on Acoustics Speech and Signal Processing. France: IEEE, 2006: 1033-1036.

[15] BENUS S, ENOS F, HIRSCHBERG J, *et al*. Pauses in deceptive speech[EB/OL]. scholar. google. cn, 2013-06-28 [2019-09-04]. https://academiccommons.columbia.edu/doi/10.7916/D8SQ97TG.

[16] KIRCHHUEBEL C. The acoustic and temporal characteristics of deceptive speech[D]. UK: University of York, 2013.

[17] DEPAULO B M, LINDSAY J J, MALONE B E, *et al*. Cues to deception[J]. Psychological Bulletin, 2003, 129(1): 74-118.

[18] EKMAN P. Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage[M]. USA: WW Norton & Company, 2009.

[19] WU Z, SINGH B, DAVIS L S, *et al*. Deception detection in videos[EB/OL]. scholar. google. cn, 2018-04-25 [2019-09-04].https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/viewPaper/16926.

[20] CASO L, MARICCHIOLO F, BONAIUTO M, *et al*. The impact of deception and suspicion on different hand movements[J]. Journal of Nonverbal Behavior, 2006, 30(1): 1-19.

[21] COHEN D, BEATTIE G, SHOVELTON H. Nonverbal indicators of deception: How iconic gestures reveal thoughts that cannot be suppressed[J]. Semiotica, 2010, 182: 133-174.

[22] LEVITAN S I, AN G, MA M, *et al*. Combining acoustic-prosodic, lexical, and phonotactic features for automatic deception detection [EB/OL]. scholar. google. cn, 2016-09-08 [2019-09-04]. http://dx.doi.org/10.21437/Interspeech.2016-1519.

[23] KRISHNAMURTHY G, MAJUMDER N, PORIA S, *et al*. A deep learning approach for multimodal deception detection [EB/OL].arxiv.org, 2018-03-01[2019-06-10]. https://arxiv.org/abs/1803.00344.

[24] KARIMI H, TANG J, LI Y. Toward end-to-end deception detection in videos[C]//2018 IEEE International Conference on Big Data (Big Data). USA: IEEE, 2018: 1278-1283.

[25] POH M Z, MCDUFF D J, PICARD R W. Non-contact, automated cardiac pulse measurements using video imaging and blind source separation[J]. Optics Express, 2010, 18(10): 10762-10774.

[26] VERKRUYSSE W, SVAASAND L O, NELSON J S. Remote plethysmographic imaging using ambient light[J]. Optics Express, 2008, 16(26): 21434-21445.

[27] ABOUELENIEN M, MIHALCEA R, BURZO M. Analyzing thermal and visual clues of deception for a non-contact deception detection approach[C]//Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments. Greece: ACM, 2016: 1-4.

[28] JI S, XU W, YANG M, *et al*. 3D convolutional neural networks for human action recognition[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(1): 221-231.

Multi-feature Non-contact Deception Detection Technology

WEI Jiangping, LIN Jiajun, CHEN Ning
（School of Information Science and Engineering, East China University of Science and Technology, Shanghai 200237, China）

Abstract: The non-contact deception detection technology has a lot of significant applications in some areas such as judicial. In order to improve the accuracy of deception detection model, this paper proposes a multi-modal fusion based deception detection model, by which we can effectively complete the deception detection evaluation task by only processing the video and audio signals of the deception detector during speaking. The heart rate may reflect the emotional changes of the liar. We extract the feature of heart rate via the photo-plethysmography method and the fully connected network and extract the video and text features through 3D-convolutional neural networks and Word2Vec+CNN. All of these extracted features are merged. And then, we use the linear support vector machines to classify the fused features. The simulation experiments are carried out on the open source real-life trial dataset. Compared with the latest MLP_{H+C} multi-modal model, the proposed deception detection model can increase the accuracy by 2.74%—23% in the three-mode. In order to evaluate whether the combination of different modal features could improve the performance of the deception detection model, we use the combination of features of each modality to conduct experiments. The accuracy of each category combination is over 70% and the accuracy of the combination of text and heart rate features is 96.89%. Especially, the combination of text, video and heart rate can obtain the highest accuracy, 98.88%, and the AUC values, 0.988 3. The accuracy of three-mode prediction is better than the single-mode and dual-mode. The experimental results show that the proposed multi-modal model can effectively improve the correct rate of deception detection.

Key words: multi-modal model**;** deception detection**;** L-SVM**;** non-contact