**融合音频情感和视频表情分析的欺诈识别系统专利解答**

1. **对于语音流数据，通过特定算法提取情感高阶特征，特定算法是什么？如何提取？什么叫高阶特征，低阶特征？**

答：原始的语音流数据，通过Opensmile中的config配置，通过调用公开的算法模块进行音频的特征提取。提取特征的目的是最大限度地从原始数据中提取特征以供算法和模型使用。低阶特征指的对原始数据进行提取，构造，选择数据等基础操作得到特征。高阶特征指的是，对低阶特征进行组合、变换等操作后，得到的一系列更具代表性的特征。对于图像处理之类的领域来说，像素点就可以作为低阶特征输入（代表性较弱）。而通过将像素点组合变换后形成的像素块（较具代表性），便可理解为高阶特征。

1. **将两组特征数据分别代入离线预训练的语音欺诈识别模型和视频表情欺诈识别模型计算欺诈概率。具体模型，如何计算？**

答：视频按照某个问题本身持续的时间分为3类，分别为0-3秒，3-7秒和7秒以上。根据三类时间长度不同已标注的样本数据，分别训练三个SVM分类器。语音直接放个LinearSVM的模型。欺诈概率按照距离分类函数距离的远近来计算，距离分类函数越远的测试样本，欺诈概率越高。

1. **通过加权模块融合两组模型的输出结果，返回欺诈识别结果，如何融合?**

答：通过准确率的高低衡量权重。比如语音准确率为60%，视频准确率80%。则语音权重为60/(80+60)，视频权重为80/(80+60)。以下公式中，P(Video)为视频的准确率，P(Audio)为音频的准确率。W(Video)和W(Audio)分别表示两组模型加权时，对应视频和音频的权重。

a与b为准确率的具体值。

P(Video) = a

P(Audio) = b

W(Video) =

W(Audio) =

1. **通过音频分析软件，什么软件？提取音频低阶特征后续有何作用？**

答：通过opensmile软件，对语音信号进行预处理和特征提取。提取音频低阶特征后可通过特征提取和特征选择等方式，提取出高阶更具有代表性的特用用以数据建模。此处特征提取指，将opensmile提取道的原始特征转换为一组更具物理意义或是统计意义或核的特征；此处特征选择指，从原始特征集合中挑选一组最具统计意义的特征，达到降维。进行该两种操作的目的是：减少数据存储和输入数据带宽；减少冗余信息；发现更具有代表性的潜在变量。

1. **从特征中提取进一步的动态回归系数，****动态回归系数的意思？**

答：回归系数指的是，在回归方程中，自变量对因变量影响大小的参数。回归系数越大，则表示自变量对因变量的影响越大。在回归方程式Y=bX+a中，斜率b称为回归系数，表示X每变动一单位，平均而言Y将变动b单位。动态回归，强调自变量对应变量的连锁反应，所以其结果被认为更可信。此处的动态回归系数，包含变量的滞后值，强调[自变量](http://wiki.mbalib.com/wiki/%E8%87%AA%E5%8F%98%E9%87%8F)对应变量的[连锁反应](http://wiki.mbalib.com/wiki/%E8%BF%9E%E9%94%81%E5%8F%8D%E5%BA%94)。

1. **运用大量统计函数，如max、min、kurtosis等进行高阶特征提取，举例说明如何提取？**

答： 高阶特征是通过对低阶特征进行组合、变换而得。如通过最大值、最小值、峰度和偏度等统计量，可以将原始低阶特征处理为更具有代表性的高阶特征。偏度这一指标，又称偏斜系数、偏态系数，是用来帮助判断数据序列的分布规律性的指标。峰度（系数）是一个用于衡量离群数据离群度的指标。峰度（系数）越大，说明该数据系列中的极端值越多。

通过计算语音mfcc,pitch, zero crossing rate等指标后，利用统计函数，将样本特征的更深层含义提取。方便后期的数据建模。

1. **将****序列化数据转化为标量特征，序列化数据和****标量特征的意思？**

答：序列化数据是指，为了方便数据在网络中传输或可存储在内存或者文件中等，将对象或者数据结构转化成特定的格式。对象序列化后的数据格式可以是二进制，可以是XML，也可以是JSON等任何格式。标量特征指将原始序列化特征转化为只包含大小的物理量，更方便建模与数值计算。

1. **采用何种分类算法，对筛选特征后的数据进行训练，如何****交叉验证确定最优模型参数？**

答：通过训练SVM分类器，对于筛选后的特征进行训练，得到SVM的相关参数。为了保障最大化利用训练样本与训练出分类器的有效性，此处采用交叉验证法确定最优模型参数。将所有样本分成K份，一般每份样本的数量相等或相差不多。取一份作为测试样本，剩余K-1份作为训练样本。这个过程重复K次，最后的平均测试结果可以将SVM的模型参数调整到较优状态。

1. **对收集的视频数据进行切割和人工标注，如何切割和标注，标注什么？**

答：视频切割是通过人工切割的方式。根据被试者根据要求演绎的不同的表情，人为将不同的视频切割为对应到具体某种表情的短视频，并为该视频标注所演绎表情的标签。如被试表演了“接受”的表情，则标注“Acceptance”；如表演“希望”，则标注“Desire”。

1. **高维特征构造，如对多种表情同时出现持续时间，次数等高阶特征进行构造，标准化处理。举例说明构造后、标准化处理后的结果是怎样的？**

答：从三类视频样本中，提取出共计4170个特征，分别为：眼睛转动的8个方向、微表情（来自于actionUnit 中抽取的17个表情特征）、头转动的方向。特征计算有两种，其一为该特征在该视频中出现的次数；其二为该特征在该视频中持续的时长。0-3以及3-7秒时间内的视频特征，不需要除以视频时长。7秒以上的视频处理时需要除以视频的时长。

1. **特征筛选，选出最优特征组合如何操作**

答：通过采用mlxtend库中的Sequential Feature Selector(SFS)贪婪搜索算法，将原始d维度的特征降到k维，此处k<d。在SFS中选用的分类器为LogisticRegression分类器。

1. **根据预测精度分配权重，权重怎么分配，如何融合两个模型的预测结果？**

答：具体参考第三题的问题，根据语音分析和视频分析的准确率进行权重计算。计算公式为：

P(Video) = a

P(Audio) = b

W(Video) =

W(Audio) =

1. **以一个具体的例子贯彻各个步骤，要实现的目的是什么？为实现这个目的，执行了哪些步骤、每一步做了什么操作、操作如何实现的（应用的原理、模型、算法以例子show）**

答：本专利意在保护“一种融合音频情感和视频表情分析的实时欺诈识别系统”。整个系统可以拆分为三部分，分别为视频欺诈模块，语音欺诈模块以及加权融合模块。此问题将结合上面的所有问题，对整个专利进行完整说明。

* **视频欺诈模块。**

·视频按照某个问题本身持续的时间分为3类，分别为0-3秒，3-7秒和7秒以上。从三类视频样本中，提取出共计4170个特征，分别为：眼睛转动的8个方向、微表情（来自于actionUnit 中抽取的17个表情特征）、头转动的方向。特征计算有两种，其一为该特征在该视频中出现的次数；其二为该特征在该视频中持续的时长。0-3以及3-7秒时间内的视频特征，不需要除以视频时长。7秒以上的视频处理时需要除以视频的时长。

根据三类时间长度不同已标注的样本数据，分别训练三个SVM分类器。通过采用mlxtend库中的Sequential Feature Selector(SFS)贪婪搜索算法，将原始d维度的特征降到k维，此处k<d。在SFS中选用的分类器为LogisticRegression分类器。

为了保障最大化利用训练样本与训练出分类器的有效性，此处采用交叉验证法确定最优模型参数。将所有样本分成K份，一般每份样本的数量相等或相差不多。取一份作为测试样本，剩余K-1份作为训练样本。这个过程重复K次，最后的平均测试结果可以将SVM的模型参数调整到较优状态。语音直接放个LinearSVM的模型。欺诈概率按照距离分类函数距离的远近来计算，距离分类函数越远的测试样本，欺诈概率越高。

* 语音欺诈模块

通过opensmile软件，对语音信号进行预处理和特征提取。提取音频低阶特征后可通过特征提取和特征选择等方式，提取出高阶更具有代表性的特用用以数据建模。此处特征提取指，将opensmile提取道的原始特征转换为一组更具物理意义或是统计意义或核的特征；此处特征选择指，从原始特征集合中挑选一组最具统计意义的特征，达到降维。进行该两种操作的目的是：减少数据存储和输入数据带宽；减少冗余信息；发现更具有代表性的潜在变量。

* 加权融合模块

此模块结合之前的视频欺诈模块与语音欺诈模块，通过各模块准确率的高低其权重。比如语音准确率为60%，视频准确率80%。则语音权重为60/(80+60)，视频权重为80/(80+60)。以下公式中，P(Video)为视频的准确率，P(Audio)为音频的准确率。W(Video)和W(Audio)分别表示两组模型加权时，对应视频和音频的权重。a与b为准确率的具体值。

P(Video) = a

P(Audio) = b

W(Video) =

W(Audio) =