用于抑郁症分析的行为原语的频谱表示

吴诗玲

摘要

抑郁症是一种严重的精神障碍，影响着全世界数百万人。传统的临床诊断方法主观、复杂，需要临床医生的广泛参与。自动抑郁症分析系统的最新进展有望在未来通过客观、可重复且现成可用的诊断工具来解决这些缺点，以帮助专业人员开展工作。然而，现有的此类工具仍然存在许多问题。一个问题是现有的自动抑郁症分析算法基于非常简短的连续片段进行预测，有时甚至只有一帧。另一个是现有方法没有考虑到被测到的行为是什么，缺少可解释性。在本文中，对视频中的面部行为原语进行检测，并以此作为视频的低维描述符，后再利用傅里叶变换，将时间序列行为信号转换为频域作为频谱信号，其中频谱信号中的每个分量编码整个视频的不同频率信息。因此，产生的光谱信号包含多尺度视频级时间信息。然而，由于原始视频长度的变化，其对应的时间序列行为信号和频谱信号的长度也是可变的。为了让标准神经网络模型能够轻松处理频谱信号，文中还提出了两种频率对齐方法，使得光谱表示可以被送入神经网络中来进行后续的抑郁症分析。复现该方法后在 AVEC 2014 和 AVEC 2019 基准数据集上进行了实验。

关键词：自动抑郁症分析；傅里叶变换；频谱表示；时频分析；神经网络

1 引言

标准的临床抑郁症评估技术可能是主观的，因为这些技术几乎完全取决于卫生专业人员自己对个人口头心理报告的理解，例如临床访谈和患者或护理人员完成的问卷调查。此外，这通常是一个冗长的过程，阻碍了早期治疗的进行，从而导致许多患者在抑郁症的早期阶段错过了预防或治疗抑郁症的最佳时机。为了使患者得到正确及时的早期治疗，近年来广泛探索了用于辅助监测和诊断的自动客观评估方法。

有一致的心理学证据表明利用表情和注视[1-2]，无需临床医生干预即可进行自动检测和分析。基于此类线索构建自动系统不仅可以提供客观且可重复的评估，还有助于缓解成本和时间要求的问题。当前大多数基于视觉的自动抑郁症分析方法是基于参与者在采访中的非语言面部行为进行的[3-7]。但要实现更加准确便捷的自动抑郁症分析仍然存在一些挑战，该论文提出的方法主要集中于解决以下挑战。第一个挑战是采访视频的长度通常是可变的，最长视频的持续时间有时比最短视频的持续时间长几倍。然而，大多数机器学习模型需要固定大小的输入。因此需要解决的第一个问题是如何将可变长度视频的信息编码为固定大小的视频表示，同时保留尽可能多的相关信息。第二个挑战是，虽然许多研究表明面部表情[8-10]、头部运动[11-12]等特征对抑郁症分析很有价值，但还不知道如何最好地编码这些特征的时间模式。因此，该论文要解决的第二个研究问题是：用什么方法提取此类特征可以尽可能地保留时间信息。

关于第一个挑战，一个通用的解决方案是预测每个帧或短片段的抑郁症分数，然后使用简单平均[4,13-15]、线性回归[16]来融合预测。但这些方法忽略了参与者的长期时间行为模式，因为从单个帧或短片段中提取的行为可以模棱两可，可以用各种原因来解释，例如，微笑可能是由于感到高兴或感到

无助而引起的。此外，具有不同抑郁程度的受试者可能会表现出相同的短期行为。换句话说，使用整个视频而不是视频的短片段可以更可靠地描述抑郁程度。另外，一些研究通过融合帧/片段级表示来构建视频级描述符。对于这种方法，可以考虑通过使用插值、动态时间扭曲 (DTW) 等将视频的每帧表示重新采样到固定长度。然而，这方法会扭曲原始信号。为了避免失真，其他研究采用固定大小的直方图或其他统计数据来总结表示的分布。

为了应对第二个挑战，即保留多尺度时间动态，最近的研究通常将每个视频分成一系列短片段（范围从 5 帧到几秒），然后从中提取时间特征[6,15,17]。然而，决定时间尺度的片段的最佳持续时间很难确定。这种方法只编码单一尺度或可能是少量的时间尺度，而忽略了长期的时间动态。

在本文中采用很容易被人和机器解释多种客观非语言的人类行为属性，例如面部动作单元 (AU)、头部姿势和注视方向，在论文中将这些称为行为原语。通过连接这些逐帧的行为原语，我们获得了人类行为信号的多通道时间序列。为了获得多尺度、与长度无关的表示，论文提出了用光谱表示，对整个视频的人类行为信号进行编码。所提出的频谱表示包含频域中的视频级行为信息，其中每个频率分量代表一个独特的动态尺度。进一步采用两种频率对齐方法来创建大小和频率覆盖范围相等的频谱表示，而不管输入视频长度的变化如何。最后，将光谱表示提供给标准神经网络模型，允许从多个渠道获得的人类行为动态被联合学习以预测抑郁症的严重程度。所提出方法的概述如图 1所示。总而言之，该方法的主要创新点和贡献如下：

1. 提出了一种新颖的基于傅里叶变换的方法，该方法将长且不定长度的时间序列视频数据转换为短且固定大小的频谱表示，可以很容易地与标准机器学习技术一起使用。；

2. 所提出的对面部行为的多尺度视频级时间动态进行编码的频谱表示被证明可用于自动抑郁分析；

3. 研究了访谈内容对抑郁分析的影响，发现不同的访谈任务会导致完全不同的抑郁预测。

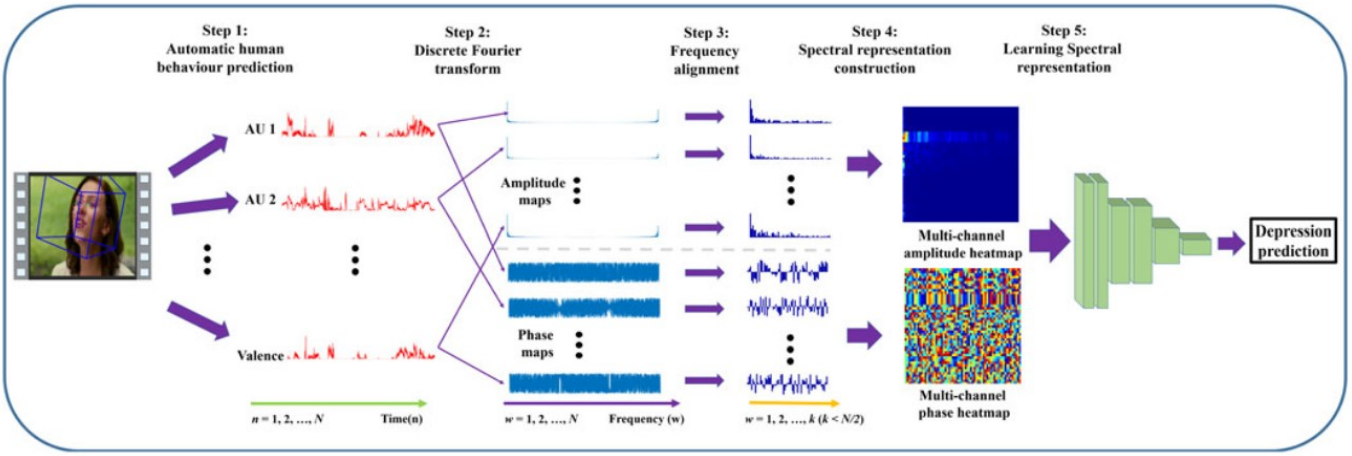


图 1: 方法示意图。首先使用低维多通道人类行为时间序列数据来表示视频（第 1 步，第 3.1 节），然后将它们转换为由所有帧的多个频率信息组成的频谱信号（第 2 步，第 3.2.1 节）. 由于光谱信号是对称的，只需保留它们的前半部分。然后，通过去除高频分量并获得所有视频的人类行为的公共频率来实现频率对齐（第 3 步，第 3.2.2 节）。最后，将所有人类行为（第 4 步，第 3.2.3 节）的对齐光谱信号组合到机器学习模型中，用于抑郁症分析（第 5 步，第 3.3 节）。

2 相关工作

2.1 非语言线索与抑郁症之间的关系

在过去的十年中，许多心理学研究都研究了人类非语言行为与抑郁症之间的关系。在这些研究中，经常得出这样的结论：抑郁症通常伴随着积极面部表情的减少[9-10,18-21]。还有一些证据表明，抑郁症与一般面部表情[19,22]和头部运动[11-12]的减少有关。因此，一些研究试图应用这种非语言线索来识别抑郁症。Cohn 等人[23]探讨了使用音频和视觉非语言线索进行抑郁症分类的可行性。他们将三种不同类型的非语言行为特征，即手动注释的面部动作单元、主动外观模型特征和声音韵律特征，分别提供给支持向量机模型。结果表明，所有这些都为检测抑郁症提供了信息，面部 AU 达到了 88% 的最佳准确度。上述发现表明，自动面部行为分析可用于自动抑郁症分析。Girard 等人[24]使用人工和自动系统专门研究了抑郁症与非语言面部行为（例如 AU 和头部姿势）之间的关系。两个系统的结果表明，抑郁严重程度高的参与者表现出更少的从属面部表情（AU 12 和 15），更多的非从属面部表情（AU 14）和减少头部运动。

2.2 自动抑郁症分析

由于深度学习的最新进展，大多数当前方法都建立在卷积神经网络（CNN）和递归神经网络（RNN）的基础上。Ma 等人[25]提出了基于音频的抑郁症分类 DeepAudioNet，它将 CNN 与长短期记忆网络 (LSTM) 相结合。大多数基于视觉的方法将视频分成几个等长的片段，并从每个片段中独立提取深度学习特征。Al Jazaery 等人[17]采用 C3D 网络从短片段中提取短期动态抑郁相关特征。然后，这些特征被送到 RNN 以进行片段级预测。最终预测是通过对所有片段的预测进行平均而获得的。为了识别抑郁症患者的面部显着区域，Zhou 等人[4]提出了 DepressNet 来学习具有视觉解释的抑郁表征。在这种方法中，最能提供抑郁症信息的面部区域被突出显示并用于在帧级别预测抑郁症。视频级别的抑郁分数是通过对所有帧的分数求平均来计算的。最近，Haque 等人[26]采用因果卷积网络从音频、文本和 3D 面部标志中学习以预测抑郁症的严重程度。

除了直接从图像中进行抑郁症分析外，一些方法还尝试从更高级别的视频表示中学习抑郁症的严重程度。Yang 等人[7]建议在每个视频中选择几个等长的片段来平衡抑郁和非抑郁训练示例的数量。他们还提出了一种位移范围直方图 (HDR) 方法，可以记录视频片段中面部特征点的动态。同时使用 CNN 从手工制作的音频和视频描述符中学习深层特征，并通过使用决策树融合音频、视频和文本特征的预测来做出最终决定。

3 本文方法

在本节中，描述了一种新颖的基于视频的自动抑郁症分析方法，可以从可变长度视频中提取固定大小的描述符。首先提取一组自动检测到的人类行为基元来表示视频，从而使高维视频显着减少为低维多通道时间序列信号（第 3.1 节）。在 3.2 节中，提出了两种频谱表示作为多通道行为信号的视频级描述符。最后，将生成的光谱表示应用于抑郁症分析（第 3.3 节）。

3.1 人类行为原语提取

为了构建视频级描述符，首要任务是降维。当前的研究要么提取手工制作的特征[27-29]或深度学习的特征[4,6,17]来表示每个帧或短视频片段。传统的手工特征，例如 HOG、LBP 等，并不是专门为面部

行为应用设计的，因此，它们不是抑郁症应用的最佳表示。另一方面，如第 2 节所述，先前的心理学和计算机视觉研究表明，非语言视觉线索是抑郁症的特征。受此启发，使用面部行为属性，包括 AU、注视方向和头部姿势作为逐帧描述符。在实验中使用的是 OpenFace 2.0[30]自动检测 17 个不同 AU 的强度、6 个注视方向和 6 个头部姿势，从而为每个视频生成 29 通道人类行为时间序列数据。

与以前使用的手工制作和深度学习的特征相比，这些人类行为描述符有几个优点。首先，它们更易于解释，因为它们具有明确的含义并且是低维的；其次，它们的提取是模块化的，因为标准的面部属性检测软件经常在非常大的数据库上训练，可以用于不同场景下的不同人；第三，客观性，其价值观独立于主体身份，避免了最终预测受到性别、年龄、种族等偏见的影响；第四，所提出的行为描述符的维数 (31-D) 比传统的手工制作特征和深度学习表征低得多。

3.2 人类行为原语的光谱表示

为了构建多通道时间序列数据的频谱表示，首先将每个时间序列转换为频域。然后进一步提出了两种频率对齐方法，以便每个视频（可能具有不同长度）的频谱表示表示相同的频率。最后，将所有行为基元的光谱表示串联成为向量，以生成给定视频的单一表示。在本文中，将 $f_c^m(n)$ 定义为第 m 个视频中的第 c 个行为时间序列信号。

3.2.1 编码多尺度视频级动态

使用傅立叶变换 (FT) 将表示每个行为原语的时间序列信号转换为频域。由此产生的频谱表示是将原始时间序列分解为其组成频率。令 $f(x)$ 为对应于行为基元的时间序列信号，然后傅里叶变换可以将其转换为频谱表示 $F(w)$。

$$F(w) = \int_{-\infty}^{\infty} f(x) e^{-(2\pi i x w)/N} dx \tag{1}$$

其中 w 可以是任何实数，$F(w)$ 是一个复杂的函数，可以重写为

$$\begin{aligned} F(w) &= \int_{-\infty}^{\infty} f(x)(\cos((2\pi i x w)/N) - i\sin((2\pi i x w)/N))dx \\ &= \int_{-\infty}^{\infty} (\mathrm{Re}(fc(x)) + i\,\mathrm{Im}(fs(x))) \\ &= \mathrm{Re}(F(w)) + i\,\mathrm{Im}(F(w)) \end{aligned} \tag{2}$$

其中 $fc(x)$ 和 $fs(x)$ 分别表示 $f(x)\cos((2\pi i x w)/N)$ 和 $-f(x)\sin((2\pi i x w)/N)$。$Re(F(w))$ 是 $F(w)$ 的实部，$\mathrm{Im}(F(w))$ 是 $F(w)$ 的相应虚部。这里，w 决定了 $F(w)$ 所代表的频率 $(2\pi w)/N$。因此，频谱表示 $F(w)$，$w \in [-\infty, \infty]$ 包含来自 $f(x)$ 中存在的所有频率的信息。

在实际应用中，每个视频都由一系列帧组成，从而为每个行为原语产生一个离散的时间序列信号。因此，将离散傅立叶变换 (DFT) 应用于行为信号 $f_c(n)$，其中 $c = 1, 2, \ldots, C$ 表示行为原语的索引，

$n = 1, 2, \ldots, N$ 表示帧索引，如下所示：

$$
\begin{aligned}
F_c(w) &= \sum_{n=0}^{N-1} f_c(n) e^{-\frac{2\pi i}{N} wn} \\
&= \sum_{n=0}^{N-1} f_c(n)[\cos(2\pi wn/N) - i\sin(2\pi wn/N)] \\
&= \sum_{n=0}^{N-1} (\mathrm{Re}\,(fc_c(n)) + i\,\mathrm{Im}\,(fs_c(n))) \\
&= \mathrm{Re}\,(F_c(w)) + i\,\mathrm{Im}\,(F_c(w))\,,
\end{aligned}
\tag{3}
$$

$f_c(n)$ 是由 N 个帧构成的第 c 个行为的时间序列信号,$F_c(w)$ 是 $f_c(n)$ 在频率 w 下的离散傅里叶变换, 其中 $w = 0, 1, 2 \ldots, W - 1$。

正如公式 3，每个频率分量是根据 $f_c(n)$ 的所有帧计算的。也就是说，频谱信号中的每个分量都概括了整个视频中存在的单个频率信息。因此，频谱信号包含了通过 $2\pi w/N, w = 0, 1, 2, \ldots W - 1$ 给定的 W 个频率对应的信息。这些成分编码不同类型的行为动态，即高频成分代表急剧的行为变化，低频成分代表行为中更渐进的变化。因此，可以说产生的频谱信号总结了整个视频的多尺度时间信息。在这里，我们将 $F_c(w)$ 中的离散频率分量 W 的数量设置为与 N 相同，以便完整地提取离散时间序列数据 $f_c(n)$ 中包含的信息（众所周知，$f_c(n)$ 可以从 $F_c(w)$ 完全重构，如果 $W = N$）。

3.2.2 频率对齐

如上所述，N 帧的时间序列行为信号可以转换为具有 $W = N$ 个频率分量的频谱信号。因此，可变长度视频的频谱信号将具有不同数量的分量，这将再次导致不同维度的特征表示。为了使它们相等，我们首先注意到时间序列数据的频谱信号始终围绕其中心频率 W/2 对称，即，如果 $F(w) = \mathrm{Re}(w) + i\,\mathrm{Im}(w)$ 和 $F(W - w) = \mathrm{Re}(W - w) + i\,\mathrm{Im}(W - w)$, 那么 $\mathrm{Re}(w) = \mathrm{Re}(W - w), \mathrm{Im}(w) = -\mathrm{Im}(W - w)$。这意味着频谱信号的前 $W/2$ 个分量可以完全代表 $f_c(n)$ 中包含的信息。此外，由于面部动作是连续且平滑的过程，高频信息通常代表噪声或异常值，例如人脸检测错误、面部点定位错误或 AU 强度估计错误等。在实践中，去除高频信息后，减少后的频谱信号仍然可以很好地代表原始时间序列数据，因为将逆 DFT 应用于频谱信号可以恢复原始时间序列数据中存在的大部分信息。作者在图 2和图 3中进行了说明, 其中用零替换所有未使用的频率分量, 可以观察到，即使去除了 90% 以上的高频成分，重构后的信号仍然与原始信号具有显着的相关性。

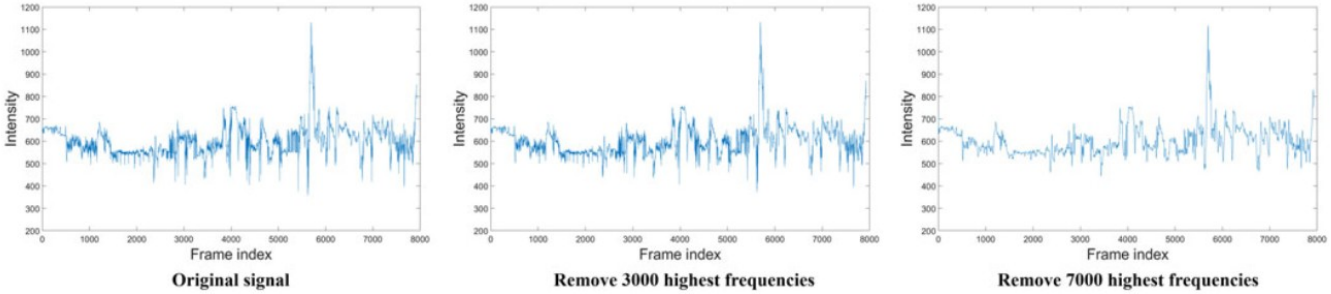


图 2: 去除高频分量后重建的时间序列信号。原始信号有 7923 帧，其频谱信号也有 7923 个频率。

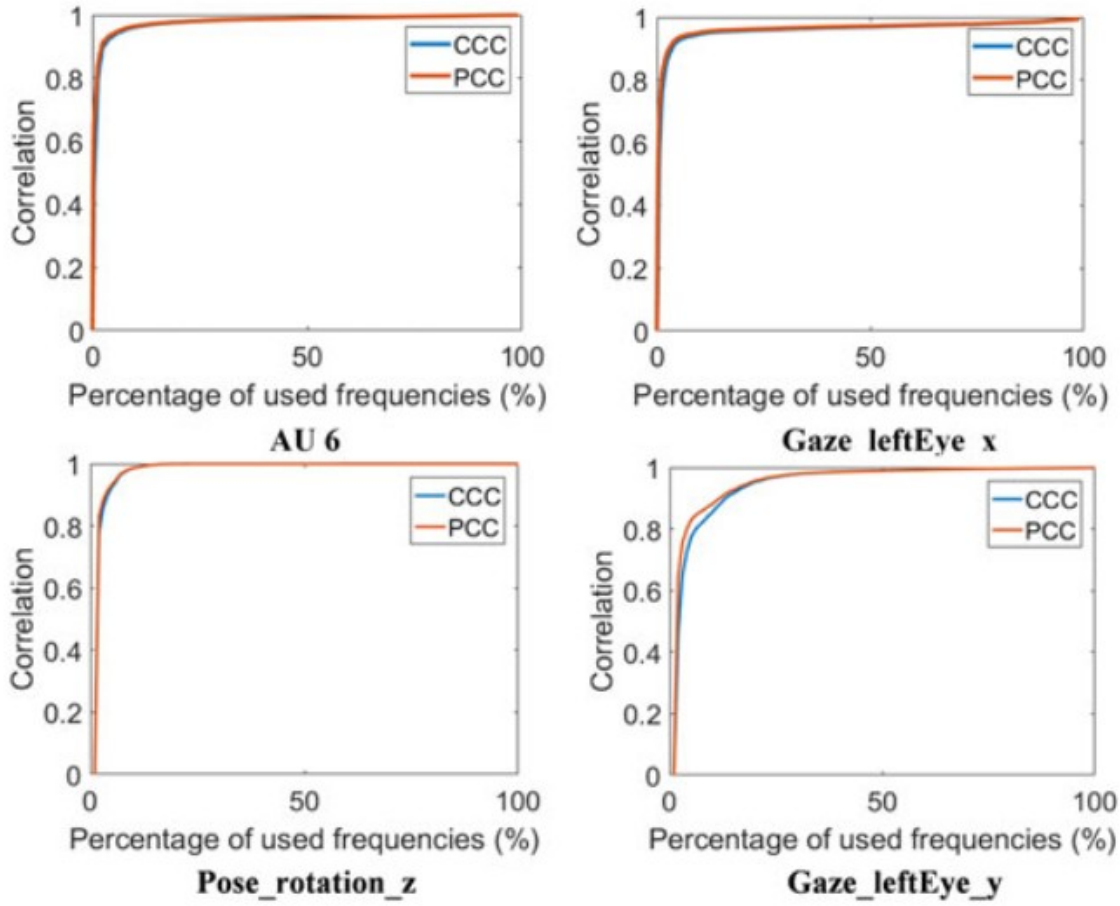


图 3: 重建的行为信号和原始行为信号之间的平均相关性作为所用频率百分比的函数的示例。

受此启发，该方法仅保留光谱信号的前 $W/2$ 个分量。然后，对应于高频的分量也被去除。由于目标是为可变长度的时间序列数据生成相同大小的视频级频谱表示，因此可以考虑为所有视频保留前 K 个最低频率的频谱信号，其中 $K < W/2$。然而，不同长度的视频中的 w_th 分量将代表不同的频率。考虑长度分别为 N_1 和 N_2 的两个时间序列信号 $f^1(n)$ 和 $f^2(n)$，相应的光谱表示分别表示为 $F^1(w)$ 和 $F^2(w)$。若 $N_1 \neq N_2$，则频谱信号 $F^1(w)$ 的第 w 个分量（$0 < w < N_1/2 \square N_2/2$）表示频率 $(2\pi w)/N_1$ 时的 DFT 值，而 $F^2(w)$ 的第 w 个分量表示频率 $(2\pi w)/N_2$ 时的 DFT 值。显然，$2\pi w)/N_1 \neq 2\pi w)/N_2$，因此频谱信号的第 w 个分量 $F^1(w)$ 和 $F^2(w)$ 不代表相同的频率。为了解决上述频率错位问题，论文提出以下两种解决方案：

第一种零填充是一种常用方法，通常用于离散时间序列提高傅里叶变换后频率分辨率。在这种方法中，将零附加到时间序列数据以增加其长度，从而使该时间序列数据的 DFT 具有更多的频率分量。特别地，频谱信号的频率分辨率 W 等于原始时间序列数据中的帧数 N。通过用零填充，我们在原始时间序列的末尾添加 N_add 零以创建长度为 $N + N_add$ 的新时间序列。因此，新时间序列的频谱信号将具有 $W + N_add$ 频率分量。设置一个固定的帧数后，将所有行为原语的时间序列都填充到该固定大小，则其频谱信号将具有相同的分辨率，然后再取前 K 个频率。

第二种通过从每个视频获得的频谱信号中选择 k 个公共频率，从可变长度的时间序列数据中提取固定大小的频谱信号。在这种情况下，k 个所选频率的值是从原始信号而不是扩展信号中获得的。因此，生成的表示中的每个分量代表准确值而不是相应频率的估计值。假设有 M 个时间序列信号 $f^1, f^2, \ldots f^M$ 对应 M 个可变长度的视频，提出的解决方案遵循以下步骤：

(1) 选择一个固定的频率分辨率 R，即用于表示每个时间序列数据的频率分量的数量，然后缩短

时间序列，将原始时间序列信号 $f_m(n)$ 中的总帧数从 N_m 减少到 $N_m - (N_m \bmod R)$ 帧，它是 R 的倍数。在实际应用上，从每个视频中从头删除 $(N_m \bmod R)/2$ 帧和从末尾删除 $(N_m \bmod R)/2$ 帧。

(2) 每个时间序列 $S(f_m(n))$ 使用公式3转换为频谱信号 $S(F_m(w))$。由于频率分量的个数等于帧数，所以 $S(F_m(w))$ 中的频率分量的个数也将是 R 的倍数，可以定义为 $W_m = (t_m \times R), m = 1, 2, \ldots, M$。因此，每个频谱信号中表示的频率可以表示为 $2\pi w_m/(t_m \times R), w_m = 0, 1, 2, \ldots, t_m \times (R-1)$。

(3) 由于每个频谱信号中的频率数是 R 的倍数，因此它们都包含相同的 R 个分量，其频率由下式给出：

$$\begin{aligned} n_f(m) &= 2\pi w_m(r)/W_m \\ &= 2\pi r \times t_m/(R \times t_m) \\ &= 2\pi r/R, \end{aligned} \tag{4}$$

其中 $r = 0, 1, 2, \ldots (R-1)$，显然,$R$ 的选择频率与 t_m 无关, 且这 R 个频率，$2\pi \times 0/R, 2\pi \times 1/R, 2\pi \times 2/R, \ldots, 2\pi \times (R-1)/R$，被编码在所有的光谱信号中，整个过程如图 4所示。最后一样选择前 K 个分量

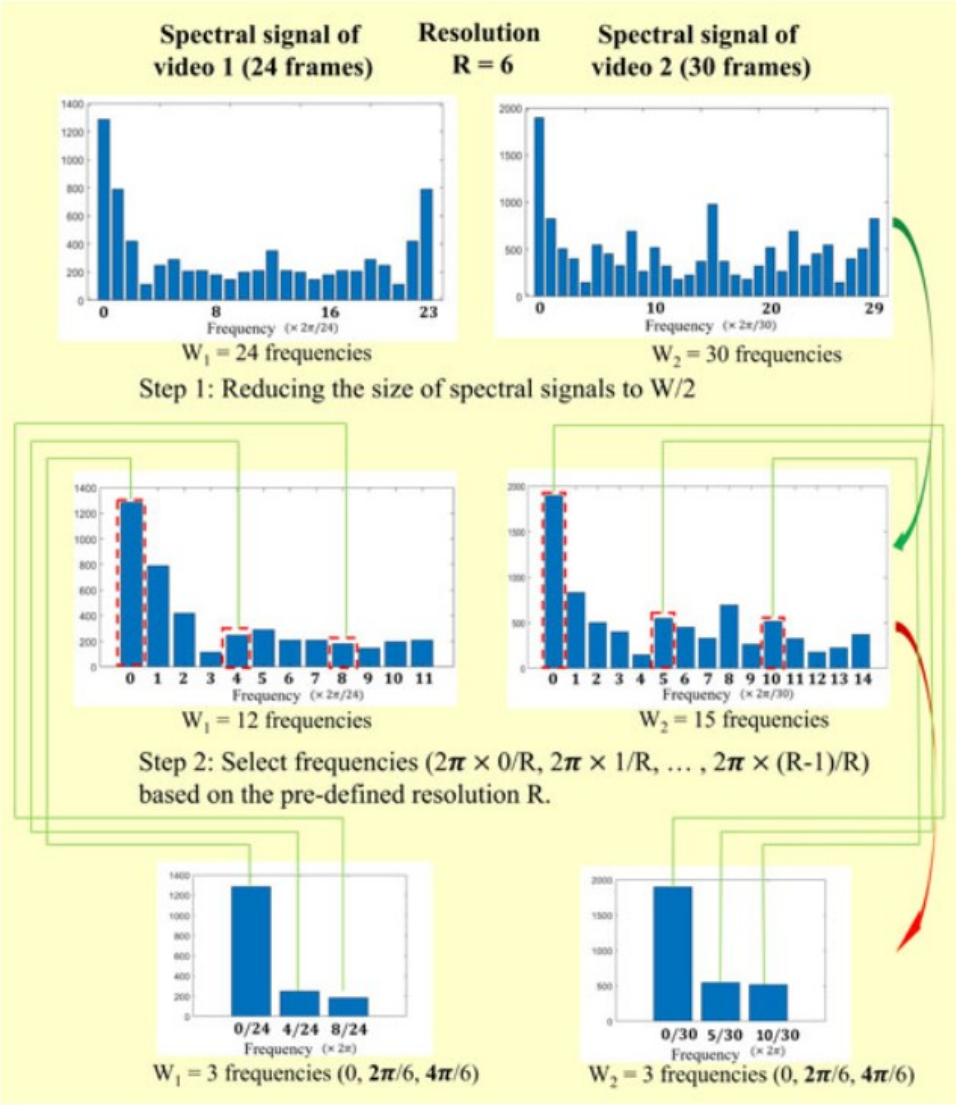


图 4: 第二种频率对齐方法

3.2.3 频谱表示

在获得对应于每个行为原语的对齐频谱信号后，使用了两种不同的方法来构建固定大小的联合表示，以便所有行为频谱信号都可以轻松地作为标准机器学习模型的输入特征。

假设从每个帧中提取 C 个行为基元，每个视频转化为由 K 个频率组成的 C 个对齐的频谱信号。由于频谱信号中的值是复数，我们将它们中的每一个转换为实域中的幅度图和相位图，其中幅度图和相位图分别可以通过以下方式计算。

$$\left|F_c^m(w)\right| / N = \sqrt{\mathrm{Re}_c^m(w)^2 + \mathrm{Im}_c^m(w)^2}/N \tag{5}$$

$$\arg\left(F_c^m(w)\right) = \arctan \frac{\mathrm{Im}_c^m(w)}{\mathrm{Re}_c^m(w)} \tag{6}$$

进一步提出以下两种方法来组合它们：

光谱热图。将一个 $C \times K$ 多通道幅度谱图和一个 $C \times K$ 多通道相位谱图组合为双通道光谱热图。

光谱向量。将一个 $C \times K$ 多通道幅度谱图和一个 $C \times K$ 多通道相位谱图连接为长 $C \times K \times 2$ 的向量。

获得上述表示后，即可利用机器学习模型来进行抑郁症分析。

4 复现细节

4.1 与已有开源代码对比

该论文所提的方法作者在 github 有用 matlab 实现的 demo，但只包含傅里叶变换与使用从每个视频的频谱信号中选择 k 个公共频率实现频谱对齐的方法。在本次复现中，重新使用 python 复现了整个过程。包括使用 openface 来对原始的视频数据提取行为原语，再将未检测到人脸的帧进行删除，进行完数据的预处理工作后，再将 29 个行为原语的时间序列进行傅里叶变换，在频谱对齐中，分别尝试了对原始时间序列补 0 以达到一致长度的和提取公共频率两种方法来实现。因为数据集中的视频数量较少，为了防止训练过拟合，因此需要对已获得的频谱表示进行降维，在论文中作者使用的是基于相关性的特征选择，本次复现也使用该特征选择方法，来对原始的频谱表示进行降维，获得降维后的频谱表示后，输入由全连接层 +Dropout 层 + 激活函数层构成的神经网络中，来进行抑郁症分析。

4.2 数据集

本次复现使用了两个抑郁症数据集，分别是原论文中使用到的 AVEC2014 数据集，和新发布的 AVEC2019 数据集。AVEC2014 数据集包含每个参与者的两个视听文件，分别对应 Northwind 和 Freeform 任务，每个视频都标有贝克抑郁量表 (BDI-II) 分数，该表的分数范围从 0 到 63，数值越大抑郁严重程度越高，实验中使用已划分的 test 数据集进行评估。AVEC2019 数据集里记录的是参与者与 AI 间的交互视听视频，每个参与者仅对应单一视频, 使用 PHQ-8 标记分数，其范围从 0 到 24，实验中分别评估在 development 和 test 数据集上的效果。

4.3 性能指标

为了更好地评估模型能力，本次实验采用四个性能指标，分别为均方根误差（RMSE）、平均绝对误差（MAE）、皮尔逊相关系数（PCC）、相关系数（CCC）。各指标的计算公式，如下所示。

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(f_i - y_i)^2}, \tag{7}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|f_i - y_i| \tag{8}$$

$$\text{PCC} = \frac{\text{cov}(f, y)}{\sigma_f \sigma_y} \tag{9}$$

$$\text{CCC} = \frac{2\rho_{f,y}\sigma_f\sigma_y}{\sigma_f^2 + \sigma_y^2 + (\mu_f - \mu_y)^2} \tag{10}$$

其中 f_i 是预测的抑郁严重程度，y_i 是相应的标签值，cov 是方差，σ_f, σ_y 是相应的标准差，μ_f 和 μ_y 是相应的均值。

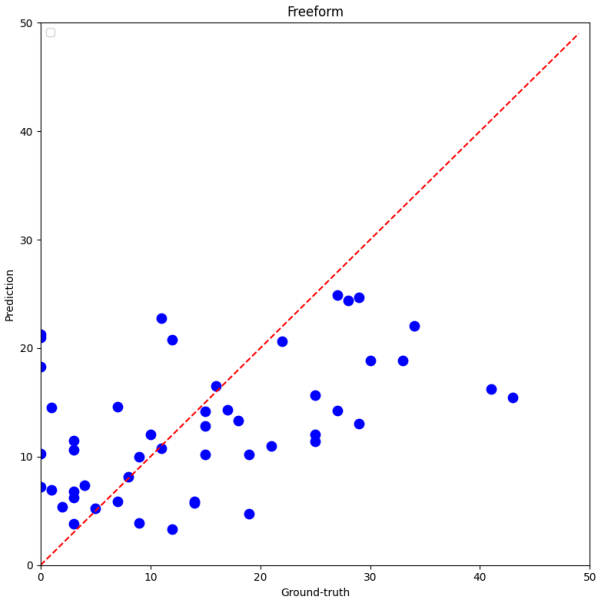
5 实验结果分析

在实验过程中，发现使用提取公共频率的方式来实现频率对齐效果更好，因此以下结果均是使用该对齐方法得到的。在 AVEC2014 的 Freeform 的抑郁症分析结果如图 5(a)所示，可以看到整体预测结果还是存在一定程度偏差的，但也可以看到有部分人的结果与实际值较为接近。在 AVEC2014 的 Northwind 的抑郁症分析结果如图 5(b)所示，整体结果与 Freeform 相近，可以说明该方法的泛化能力还是客观的。平均两个任务的性能，得到了模型在该个任务上的综合效果，如表 1所示，其中 Baseline 为 AVEC2014 数据集报告的 Baseline 性能，可以看到该方法的性能是超过了 Baseline 的，但是复现效果没能达到论文中的性能，具体原因分析可能有以下 3 点。（1）数据预处理阶段：因为视频中存在未检测到人脸的情况，复现过程中采取的方法是直接删除这些帧，但不清楚论文中对该类情况的处理方式；（2）实验中超参数的设定：在设置频率 N 及选取 K 个公共频率，均是对实验结果影响较大的超参；（3）神经网络训练策略的差异：因数据集较小，不同的训练策略可能对模型的训练影响较大。

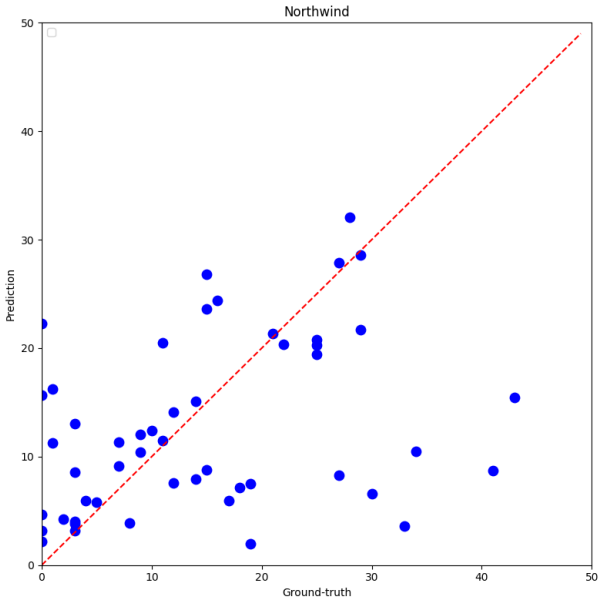
在 AVEC2019 的数据集上的分析结果如图 6所示，可以看到在验证集上的预测准确性是远高于测试集的，从表 2也可以看到，但应 AVEC2019 没有提供具体的视频数据，只提供了所提取的特征，无法显现地观察到验证集与测试集数据间的差异，推测可能是测试集上的数据与训练集中数据分布差距较大所导致的。但也可以看到无论是在验证集还是测试集上，该论文所使用的方法性能是显著优于该数据集的 Baseline 结果的。

表 1: 在 AVEC2014 的复现性能与报告结果的比较

	MAE	RMSE	PCC	CCC
复现效果	8.12	11.02	0.40	0.34
论文报告	7.18	9.27	0.56	0.42

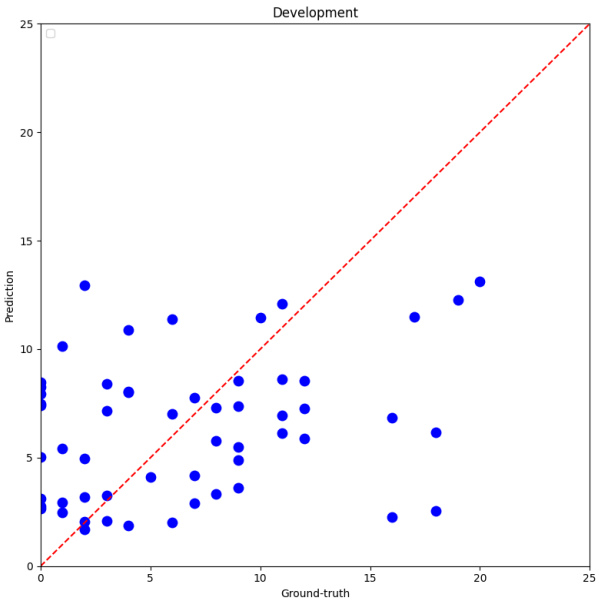


(a) Freeform 任务上的抑郁症预测结果

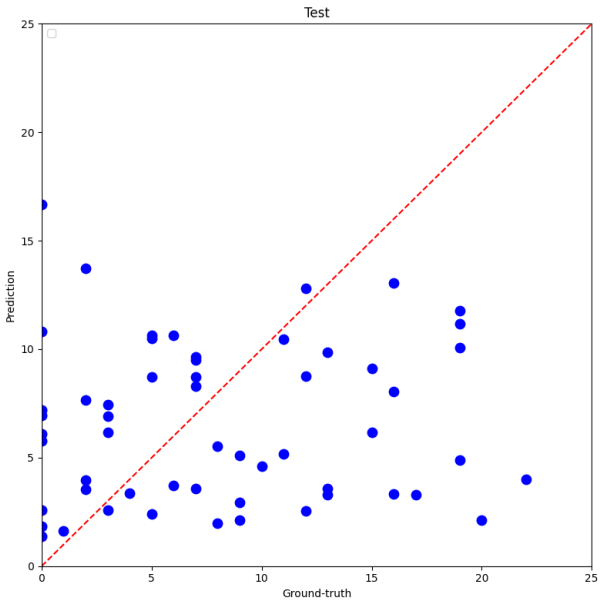


(b) Northwind 任务上的抑郁症预测结果

图 5: 在 AVEC2014 上的抑郁症预测结果



(a) 验证集上抑郁症预测结果



(b) 测试集上的抑郁症预测结果

图 6: 在 AVEC2019 上的抑郁症预测结果

表 2: 在 AVEC2019 的性能与 Baseline 的比较

		MAE	RMSE	PCC	CCC
Dev	Baseline	–	7.02	–	0.115
	Spectral	4.51	5.65	0.29	0.245
	Baseline	–	10.0	–	0.019
Test	Spectral	5.86	7.33	0.06	0.049

6 总结与展望

本报告基于原论文介绍了整个方法提出的思路过程，并展示了相关复现结果。从结果中可以看到，其对于参与者抑郁水平的分析距离真实情况仍是存在一定差距。在实验中也观察到所采用的特征选择方式会对结果有较大影响，当测试集中的数据也一起参与特征选择，但不参与后续的神经网络训练，仍会对最后的结果有很大程度的提高，可知当前使用的基于相关性的特征选择方法仍有待优化也可尝试不同特征选择方法，因此如何在训练集上选出同样能在其他数据下也仍具有较大的区分度的特征是一个可以继续探索的问题。另一方面，在该论文中，是直接将 29 个行为原语直接拼接成向量表示，并没有探究是否有更优的联合方式，在后续中也可探究各个行为原语之间的关系，找出更有效的联合方式。

参考文献

[1] SCHERER S, STRATOU G, MORENCY L P. Audiovisual behavior descriptors for depression assessment[C]//Proceedings of the 15th ACM on International conference on multimodal interaction. 2013: 135-140.

[2] GOLDSTEIN I B. Role of muscle tension in personality theory.[J]. Psychological Bulletin, 1964, 61(6): 413.

[3] WEN L, LI X, GUO G, et al. Automated depression diagnosis based on facial dynamic analysis and sparse coding[J]. IEEE Transactions on Information Forensics and Security, 2015, 10(7): 1432-1441.

[4] ZHOU X, JIN K, SHANG Y, et al. Visually interpretable representation learning for depression recognition from facial images[J]. IEEE Transactions on Affective Computing, 2018, 11(3): 542-552.

[5] ZHU Y, SHANG Y, SHAO Z, et al. Automated depression diagnosis based on deep networks to encode facial appearance and dynamics[J]. IEEE Transactions on Affective Computing, 2017, 9(4): 578-584.

[6] JAN A, MENG H, GAUS Y F B A, et al. Artificial intelligent system for automatic depression level analysis through visual and vocal expressions[J]. IEEE Transactions on Cognitive and Developmental Systems, 2017, 10(3): 668-680.

[7] YANG L, JIANG D, SAHLI H. Integrating deep and shallow models for multi-modal depression analysis—Hybrid architectures[J]. IEEE Transactions on Affective Computing, 2018, 12(1): 239-253.

[8] ELLGRING H. Non-verbal communication in depression[M]. Cambridge University Press, 2007.

[9] CHENTSOVA-DUTTON Y E, TSAI J L, GOTLIB I H. Further evidence for the cultural norm hypothesis: positive emotion in depressed and control European American and Asian American women.[J]. Cultural Diversity and Ethnic Minority Psychology, 2010, 16(2): 284.

[10] TSAI J L, POLE N, LEVENSON R W, et al. The effects of depression on the emotional responses of Spanish-speaking Latinas.[J]. Cultural Diversity and Ethnic Minority Psychology, 2003, 9(1): 49.

[11] FISCH H U, FREY S, HIRSBRUNNER H P. Analyzing nonverbal behavior in depression.[J]. Journal of abnormal psychology, 1983, 92(3): 307.

[12] JOSHI J, GOECKE R, PARKER G, et al. Can body expressions contribute to automatic depression analysis?[C]//2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG). 2013: 1-7.

[13] VALSTAR M, SCHULLER B, SMITH K, et al. Avec 2013: the continuous audio/visual emotion and depression recognition challenge[C]//Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. 2013: 3-10.

[14] VALSTAR M, SCHULLER B, SMITH K, et al. Avec 2014: 3d dimensional affect and depression recognition challenge[C]//Proceedings of the 4th international workshop on audio/visual emotion challenge. 2014: 3-10.

[15] De MELO W C, GRANGER E, HADID A. Combining global and local convolutional 3d networks for detecting depression from facial expressions[C]//2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019). 2019: 1-8.

[16] MENG H, HUANG D, WANG H, et al. Depression recognition based on dynamic facial and vocal expression features using partial least square regression[C]//Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge. 2013: 21-30.

[17] AL JAZAERY M, GUO G. Video-based depression level analysis by encoding deep spatiotemporal features[J]. IEEE Transactions on Affective Computing, 2018, 12(1): 262-268.

[18] GEHRICKE J G, SHAPIRO D. Reduced facial expression and social context in major depression: discrepancies between facial muscle activity and self-reported emotion[J]. Psychiatry Research, 2000, 95(2): 157-167.

[19] RENNEBERG B, HEYN K, GEBHARD R, et al. Facial expression of emotions in borderline personality disorder and depression[J]. Journal of behavior therapy and experimental psychiatry, 2005, 36(3): 183-196.

[20] SLOAN D M, STRAUSS M E, WISNER K L. Diminished response to pleasant stimuli by depressed women.[J]. Journal of abnormal psychology, 2001, 110(3): 488.

[21] ROTTENBERG J, KASCH K L, GROSS J J, et al. Sadness and amusement reactivity differentially predict concurrent and prospective functioning in major depressive disorder.[J]. Emotion, 2002, 2(2): 135.

[22] GAEBEL W, WÖLWER W. Facial expressivity in the course of schizophrenia and depression[J]. European archives of psychiatry and clinical neuroscience, 2004, 254(5): 335-342.

[23] COHN J F, KRUEZ T S, MATTHEWS I, et al. Detecting depression from facial actions and vocal prosody[C]//2009 3rd International Conference on Affective Computing and Intelligent Interaction

and Workshops. 2009: 1-7.

[24] GIRARD J M, COHN J F, MAHOOR M H, et al. Nonverbal social withdrawal in depression: Evidence from manual and automatic analyses[J]. Image and vision computing, 2014, 32(10): 641-647.

[25] MA X, YANG H, CHEN Q, et al. Depaudionet: An efficient deep model for audio based depression classification[C]//Proceedings of the 6th international workshop on audio/visual emotion challenge. 2016: 35-42.

[26] HAQUE A, GUO M, MINER A S, et al. Measuring depression symptom severity from spoken language and 3D facial expressions[J]. arXiv preprint arXiv:1811.08592, 2018.

[27] DHALL A, GOECKE R. A temporally piece-wise fisher vector approach for depression analysis[C]// 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). 2015: 255-259.

[28] HE L, JIANG D, SAHLI H. Automatic depression analysis using dynamic facial appearance descriptor and dirichlet process fisher encoding[J]. IEEE Transactions on Multimedia, 2018, 21(6): 1476-1486.

[29] SONG S, SÁNCHEZ-LOZANO E, KUMAR TELLAMEKALA M, et al. Dynamic facial models for video-based dimensional affect estimation[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. 2019: 0–0.

[30] BALTRUSAITIS T, ZADEH A, LIM Y C, et al. Openface 2.0: Facial behavior analysis toolkit[C]// 2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018). 2018: 59-66.