doi:10.11959/j.issn.1000-436x.2017294

基于 SAE 和 LSTM RNN 的多模态生理信号融合和情感识别研究

李幼军[1,2,3]，黄佳进[1,2,3]，王海渊[1,2,3]，钟宁[1,2,3,4]

（1. 北京工业大学国际 WIC 研究院，北京 100124；2. 磁共振成像脑信息学北京市重点实验室，北京 100124；
3. 脑信息智慧服务北京市国际科技合作基地，北京 100124；4. 北京未来网络科技高精尖创新中心，北京 100124 ）

摘 要：为了提高情感识别的分类准确率，提出一种将栈式自编码神经网络（SAE）和长短周期记忆单元循环神经网络（LSTM RNN）融合的多模态融合特征情感识别方法。该方法通过 SAE 对不同模态的生理特征进行信息融合和压缩，随后用 LSTM RNN 对长时间周期的融合进行情感分类识别。通过将该方法用到开源数据集中进行验证，得到情感分类准确率达到 0.792 6。实验结果表明，SAE 对多模态生理特征进行了有效融合，LSTM RNN 能够有效地对长时间周期中的关键特征进行识别。

关键词：多模态生理信号情感识别；栈式自编码神经网络；长短周期记忆循环神经网络；多模态生理信号融合
中图分类号：TP181, TP183 **文献标识码**：A

Study of emotion recognition based on fusion multi-modal bio-signal with SAE and LSTM recurrent neural network

LI You-jun[1, 2, 3], HUANG Jia-jin[1, 2, 3], WANG Hai-yuan[1, 2, 3], ZHONG Ning[1,2,3,4]

(1. Institute of International WIC, Beijing University of Technology, Beijing 100124, China;
2. Beijing Key Laboratory of Magnetic Resonance Imaging and Brain Informatics, Beijing 100124, China;
3. Beijing International Collaboration Base on Brain Informatics Wisdom and Services, Beijing 100124, China;
4. Beijing Advanced Innovation Center for Future Internet Technology, Beijing 100124, China)

Abstract: In order to achieve more accurate emotion recognition accuracy from multi-modal bio-signal features, a novel method to extract and fuse the signal with the stacked auto-encoder and LSTM recurrent neural networks was proposed. The stacked auto-encoder neural network was used to compress and fuse the features. The deep LSTM recurrent neural network was employed to classify the emotion states. The results present that the fused multi-modal features provide more useful information than single-modal features. The deep LSTM recurrent neural network achieves more accurate emotion classification results than other method. The highest accuracy rate is 0.792 6

Key words: multi-modal bio-signal emotion recognition, stacked auto-encoder neural network, LSTM recurrent neural network, multi-modal bio-signals fusion

1 引言

情感是人们对客观事物是否满足自身需要而产生的综合状态，不同的情感状态影响了人们的学习、记忆与决策等。对于不同情感状态的识别在远程教育、医疗、智能系统以及人机交互等领域均有着广泛的应用前景，因此，情感识别近期受到研究者的高度重视，成为研究的热点[1]。

情感识别可以通过面部表情、语音语调和身体姿态等外部特征进行识别，也可以通过神经系统和内分泌系统的变化进行识别。然而外部特征由于种种原因，容易被人为加以掩饰，从而用外

收稿日期：2017-05-31；**修回日期**：2017-11-23
通信作者：钟宁，zhong@maebashi-it.ac.jp
基金项目：国家自然科学基金资助项目（No.61420106005）；国家重点基础研究发展计划基金资助项目（No.2014CB744600）；国家国际科技合作专项基金资助项目（No.2013DFA32180）
Foundation Items: The National Natural Science Foundation of China (No.61420106005), The National Basic Research Program of China (No.2014CB744600),The International Science & Technology Cooperation Program of China (No.2013DFA32180)

部特征进行情感识别其客观性不强；而通过神经系统和内分泌系统的变化进行的情感识别，由于生理信号的不可伪装性，其分析结果相对客观[2]。以前的研究大多数都是通过一种生理信号进行情感识别，即所谓的单模态数据分析。例如，文献[3~5]通过人们的语音信号进行情感识别，文献[6~12]从脑电信号进行情感识别等。根据文献[13]的研究，不同的情感状态有可能引发生理信号产生同样的变化规律。但是通过多种模态的生理信号进行融合分析能更加准确地识别出具体的情感状态。例如，“生气”情感能够引起心率的加快、心率变异性的降低和皮肤导电性的升高；而“吃惊”情感除了能够引起心率的加快之外，在心率变异性和皮肤导电性方面却没有显著变化。因此，在进行情感识别过程中通过将多模态生理信号特征进行融合识别，能使识别结果更加客观和准确。在生理信号情感识别中，另外一个需要注意的问题就是情感状态的时间特征。有的情感会维持很长一段时间，而在实验中被测试应激性的情感往往是随着对被测试的刺激而产生的，其维持时间不会太长，因此从整个实验周期中，以多长的时间粒度去识别情感状态，这也关系到最终情感识别准确率的高低。

本文提出了一种对多模态生理数据进行融合，然后依据融合的生理特征进行情感分类识别的方法。该方法的创新点主要有 2 点：1) 针对多模态生理信号的融合问题，通过建立栈式自编码神经网络（SAE, stacked auto-encoder neural network）对前期独立提取的多模态生理特征信息进行信息融合和信息压缩，达到组合多模态生理信号的目的，为情感识别做好准备；2) 针对情感识别的长周期时间窗口问题，通过建立带有长短周期记忆单元的循环神经网络（LSTM RNN, long-short term memory recurrent neural network）对融合后的多模态生理融合特征进行情感识别分析，从而达到提高分类准确率的目的。本文通过将本方法应用到开源数据集 DEAP[14]上进行了验证分析，并将该实验得到的结果与相关研究进行对比后，表明本文的方法能够有效提高情感识别的准确率。

2 相关工作

2.1 DEAP 数据集

DEAP 是由伦敦玛丽皇后大学的相关研究团队开发的开源数据集，其内容主要记录了 32 个被测试在所选音乐视频（MV）刺激下的多模态生理电信号（多导脑电信号、眼电信号、肌电信号和皮肤电信号）和面部表情，主要用于研究被测试情感的变化。用于实验的 1 000 余首 MV，实验人员通过被测试在线投票方式，筛选出了 40 首 MV 作为刺激媒体进行实验，在实验过程中除了生理信号之外，还记录了被测试对 MV 的效价、唤醒度、喜好程度和熟悉度等进行的评价。

DEAP 中的每个实验时长为 63 s，前 3 s 是实验准备阶段，随后的 60 s 中被测试观看 MV，在观看的同时通过可穿戴设备对被测试者的生理信号进行采集。每个被测试者共观看 40 个 MV。

本文的实验数据是从 DEAP 数据集中提取的多模态生理信号和被测试对 MV 的评价指标 2 个部分数据。

2.2 情感分类模型

本文采用的混合神经网络是一种有监督的机器学习方法。因此，在对生理信号进行学习之前需要将对应的多模态生理数据打上情感分类标签。本文中情感标签的生成主要基于被测试在实验过程中对于 MV 的情感评价数据。在生成情感分类标签之前，涉及如何对情感进行分类的问题及情感分类模型问题。现有的情感分类模型主要有 2 类：基本情感模型和维度情感模型。基本情感模型认为人的情感由基本情感构成。Paul 等[15]提出人的情感由 6 种基本情感组成，即生气、厌恶、害怕、高兴、悲伤和吃惊。也有其他的学者提出了 8 种甚至 22 种基本情感构成的情感模型，表 1 总结了 5 种被人们接受的基本情感模型，其中，文献[15]提出的基本情感模型被世人广泛接受。基本情感模型在情感表述方面存在着较大的局限性。例如，基本情感模型是基于情感表述词汇的，然而相同的词汇在不同的语言具有二义性；其次，在基本情感模型中的词汇有可能在非英语语系中不一定存在；还有一些混合情感通过单一词汇无法表达。

另外一种情感分类模型是 Rosner 等[16]于 1979 年左右提出的维度情感模型（二维情感模型）。二维情感模型中，Rosner 等引入了效价和唤醒度 2 个指标来对人类的情感状态进行度量。效价指标按照引起人们情感的愉悦程度将情感分为正性情感和负性情感，正性情感是指使人们感到快乐和积极的情

表 1　　基本情感模型

提出者	情感种类	基本情感
Paul	6 种	生气、厌恶、害怕、高兴、悲伤、吃惊
Parrot	6 种	生气、害怕、高兴、喜爱、悲伤、吃惊
Frijda	6 种	渴望、高兴、感兴趣、吃惊、惊奇、悲伤
Plutchik	8 种	接纳、生气、期望、厌恶、快乐、害怕、悲伤、吃惊
Tomkins	9 种	生气、感兴趣、轻蔑、厌恶、悲痛、害怕、高兴、害羞、吃惊

感，负性情感是指使人感到悲伤和消极的情感。唤醒度指标反映了人们感受到情感的强度，唤醒度值越大感受的情感的刺激程度就越高，唤醒度值越小对应的情感越不易察觉。

在 DEAP 数据集中，被测试对于 MV 的评价也正是基于二维情感模型的，其中，包含了效价和唤醒度的值，2 个指标的评价值是从 1~9 的连续数字，分别代表了由负到正，由弱到强。本文为了简化问题，将被测试对 MV 连续性的评价指数按照二维情感模型的 4 个象限分为 4 类，如图 1 所示，分别为 HVHA（high valence high arousal）、LVHA（low valence high arousal）、LVLA（low valence low arousal）、HVLA（high valence low arousal）。具体的做法为针对每个被测试者分别计算效价和唤醒度的平均值，随后针对每次实验计算 2 个指标各自的均差，按照两者的正负属性，将具体实验对应的情感评价标签映射到 4 个维度中。DEAP 中一共有 32 个被测试者，1 280 个实验样本。通过映射后，各个象限中样本的个数如表 2 所示。从表 2 可以看出，各类样本量之间基本上是相互平衡的，这对于随后通过神经网络训练进行情感识别也是有益的。

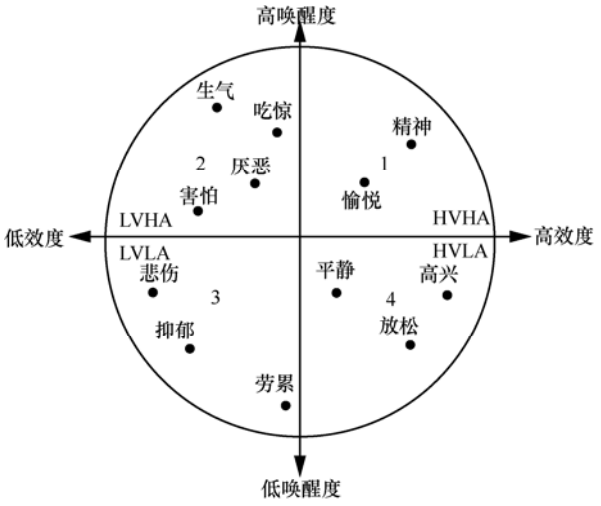


图 1　效价/唤醒度二维情感模型

表 2　　情感分类象限中包含的样本个数

情感分类标签	样本个数
HVHA	348
LVHA	298
LVLA	282
HVLA	352
合计	1 280

为了检测 4 个情感象限中被测试情感评价指标的可区分度，本文用 *K*-means 方法对 4 个象限中的情感评价值做了聚类分析，以求得各个分类的分类中心点。表 3 展示的是各个中心点的具体值。

表 3　通过 *K*-means 聚类得到的 4 个情感分类中心值

分类	HVHA	LVHA	LVLA	HVLA
效价	1.634 35	−1.108 59	−2.324 97	1.185 36
唤醒度	2.127 02	1.267 37	−1.596 66	−1.896 65

为了更加直观地展示聚类中心的离散度，本文以散点图的方式将表 3 的内容展示到图 2 中，图 2 中横坐标为唤醒度的均差值，纵坐标为效价的均差值，由于是情感评价值，所以没有量纲。从图 2 中可以看到，每个象限的聚类中心距离分类边界是比较远的，各个类别之间有较好的区分度。

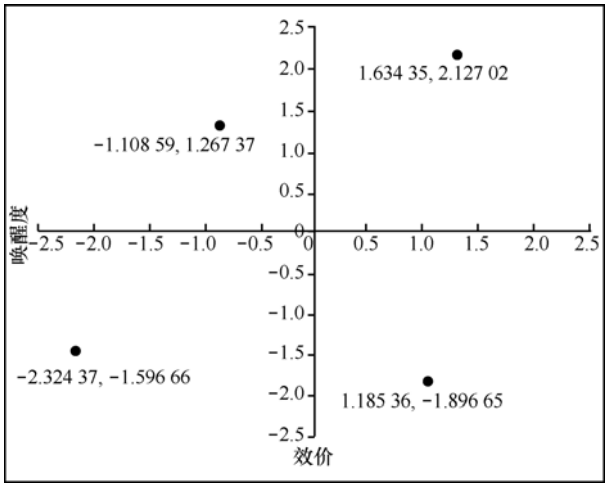


图 2 通过 *K*-means 方法得到的情感分类中心散点

需要说明的是，在将被测试个体的情感评价指标映射到 4 个情感象限的过程中，确实有少量的指标比较接近情感分界边缘，这就意味着被测试者在对某些 MV 进行评价的时候，对于自身的感受是比较模糊的。同时，不同的被测试者对相同的 MV 的评价也存在很大的差别，这个现象反映了被测试者在情感的感受和评价方面的差异，对应到基于被测试个体的情感识别准确率的时候，就会显示有的被测试者情感识别准确率高，有的被测试者对应的情感识别准确率低。

3 多模态生理信号特征融合及情感识别模型的构建

本文使用 SAE 和 LSTM RNN 进行多模态生理信号情感识别的原因。然后从混合网络构建的角度，详述了混合神经网络的构建。

3.1 SAE 在本研究中的作用

SAE 经常被用来进行数据压缩和数据融合[17]。它是一种无监督的学习算法，使用反向传播算法进行反馈，最终的目标是让输出值无限接近于输入值。假设 SAE 的编码过程为 ϕ，解码过程为 φ，则 SAE 可形式化表示为

$$\phi: X \to F$$
$$\varphi: F \to X'$$
$$\phi, \varphi = \arg\min_{\phi,\varphi} \left\| X - X' \right\|^2 \tag{1}$$

式(1)中表达了 SAE 网络在进行信息编码和信息解码的 2 个过程，即 X 到 F 为信息编码过程 ϕ，而从 F 到 X' 为信息的解码过程 φ，而式(1)的总体目标是使经过编码和解码 2 个过程后，通过调整网络参数，达到输出信息 X 与输入信息 X' 两者之间的差最小。对 SAE 网络结构的调整，主要是对神经元之间的连接权重矩阵以及偏置向量的训练和调整，具体过程可表示为

$$\boldsymbol{f} = \sigma(\boldsymbol{W}\boldsymbol{x} + \boldsymbol{b})$$
$$\boldsymbol{x}' = \sigma'(\boldsymbol{W}'\boldsymbol{f} + \boldsymbol{b}')$$
$$L(\boldsymbol{x}, \boldsymbol{x}') = \left\| \boldsymbol{x} - \boldsymbol{x}' \right\|^2 \tag{2}$$

其中，$x \in X, f \in F$，通常 F 的维度要比 X 的维度小，SAE 网络的训练目的就是让 $L(x, x')$ 函数取得最小值。其中，σ, σ' 为激活函数，本文取激活函数为 sigmoid 函数，可表示为

$$f(x) = \frac{1}{1 + \mathrm{e}^{-x}} \tag{3}$$

采用 SAE 的原因在于本文涉及多模态生理信号，SAE 能够提供一种有效的特征融合机制对多模态信号特征进行融合；同时，通过 SAE 网络结构可以对不同模态生理信号的内部特征进行信息压缩，以减少后续分类网络的结构复杂性和计算量。

2017294-4

通过 SAE 进行特征信息无损压缩的原理在于，SAE 的网络结构中包含了多个隐藏层，将多模态特征输入 SAE 进行训练，当网络结构达到稳定之后，取 SAE 最后一个隐藏层的输出作为接下来 LSTM RNN 的输入，进行分类训练，由于隐藏层的神经元个数远远小于输入层的神经元个数，从客观上就达到了信息压缩的效果，从而简化了后续分类网络的结构。图 3 展示了一个自编码神经网络的结构，可以看到自编码神经网络由输入层、隐藏层和输出层 3 个网络结构构成，其中，输入层与输出层完全相同，而隐藏层的变量个数要小于输入层（$m < n$），这样从输入层到隐藏层等效于建立了一个输入数据的压缩编码器，而从隐藏层到输出层是压缩编码器的解码过程。经过训练之后，把解码器去掉，仅取输入层到隐藏层的神经网络，就可以达到特征融合和信息压缩的目的。

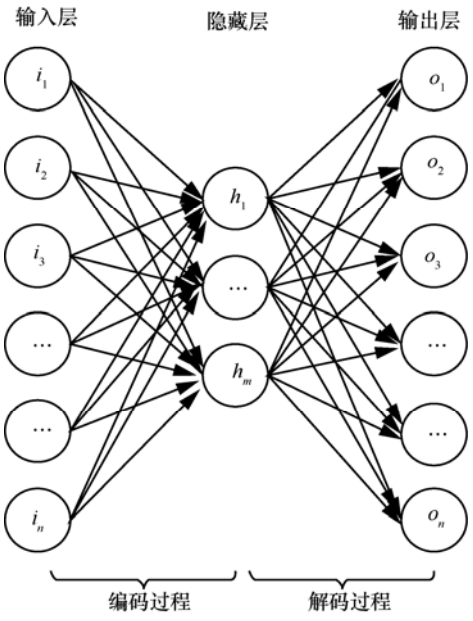


图 3 自编码神经网络结构示意

3.2 LSTM RNN 在本研究中的作用

LSTM RNN 比传统 RNN 在对时序相关的信号进行分类分析的过程中更占优势，这是因为 LSTM

记忆单元将时序数据中的关键特征在整个分类网络的长周期计算过程中进行了有效的保持和传递[18]。

LSTM 单元的结构如图 4 所示，为了进行对比，图 4 中包含了简单循环神经网络单元。图 4 中 f_i，f_g 和 f_o 分别代表 3 种类型的激活函数。f_i 是输入激活函数，f_o 是输出激活函数，f_g 是门激活函数，在本文的具体代码设置中，激活函数都取 sigmoid 函数。

假设 x_t 定义为 t 时刻的输入，h_t 定义为 t 时刻的隐藏层状态，i_t 定义为 t 时刻输入门的输出状态，f_t 定义为 t 时刻遗忘门的输出状态，o_t 定义为 t 时刻输出门的输出状态，那么其定义可分别表示为

$$i_t = f_g(\boldsymbol{w}_{x_i} x_t + \boldsymbol{w}_{h_i} h_{t-1} + \boldsymbol{b}_i) \tag{4}$$

$$f_t = f_g(\boldsymbol{w}_{x_f} x_t + \boldsymbol{w}_{h_f} h_{t-1} + \boldsymbol{b}_f) \tag{5}$$

$$o_t = f_g(\boldsymbol{w}_{x_o} x_t + \boldsymbol{w}_{h_o} h_{t-1} + \boldsymbol{b}_o) \tag{6}$$

在式(4)~式(6)中，$\boldsymbol{w}_{x_i}$、$\boldsymbol{w}_{x_f}$、$\boldsymbol{w}_{x_o}$ 为对应每个门的输入权重矩阵，$\boldsymbol{w}_{h_i}$、$\boldsymbol{w}_{h_f}$、$\boldsymbol{w}_{h_o}$ 为对应每个门的反馈权重矩阵，$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$ 为每个门对应的偏置向量。图 4 中 LSTM 的中间状态分别为 t 时刻的输入函数所对应的输出状态 C_{in_t}，输出函数所对应的输出状态 C_t 和隐藏层对应的输出状态 h_t，中间状态可以表示为

$$C_{in_t} = f_i(\boldsymbol{w}_{x_c} x_t + \boldsymbol{w}_{h_c} h_{t-1} + \boldsymbol{b}_{c_{in}}) \tag{7}$$

在式(7)中，C_{in_t} 作为输入函数的 t 时刻的输出状态将与输入门 t 时刻的输出状态 i_t 共同参与 t 时刻输入状态的整体更新。其中，$\boldsymbol{w}_{x_c}$、$\boldsymbol{w}_{h_c}$ 和 $\boldsymbol{b}_{c_{in}}$ 分别为输入权重矩阵和对应的偏置向量。在 t 时刻通过新的输入和上一时刻的状态反馈，对整个 LSTM 单元进行更新。信息更新包含 2 部分，分别对应图 4 中的 C_t 和 h_t 状态的更新，如式(8)和式(9)所示

$$C_t = f_t C_{t-1} + i_t C_{in_t} \tag{8}$$

$$h_t = o_t f_o(C_t) \tag{9}$$

通过 LSTM 的各个门函数和整个单元输出状态的更新函数和过程，输入数据通过遗忘门函数以及状态的传递将输入特征中的关键信息进行保留和传递。LSTM 单元的这一特性对数据集 DEAP 来说，其意义在于音乐视频（MV）在 60 s 的时间中对被测试的刺激是变化的，通常 60 s 的 MV 中总有

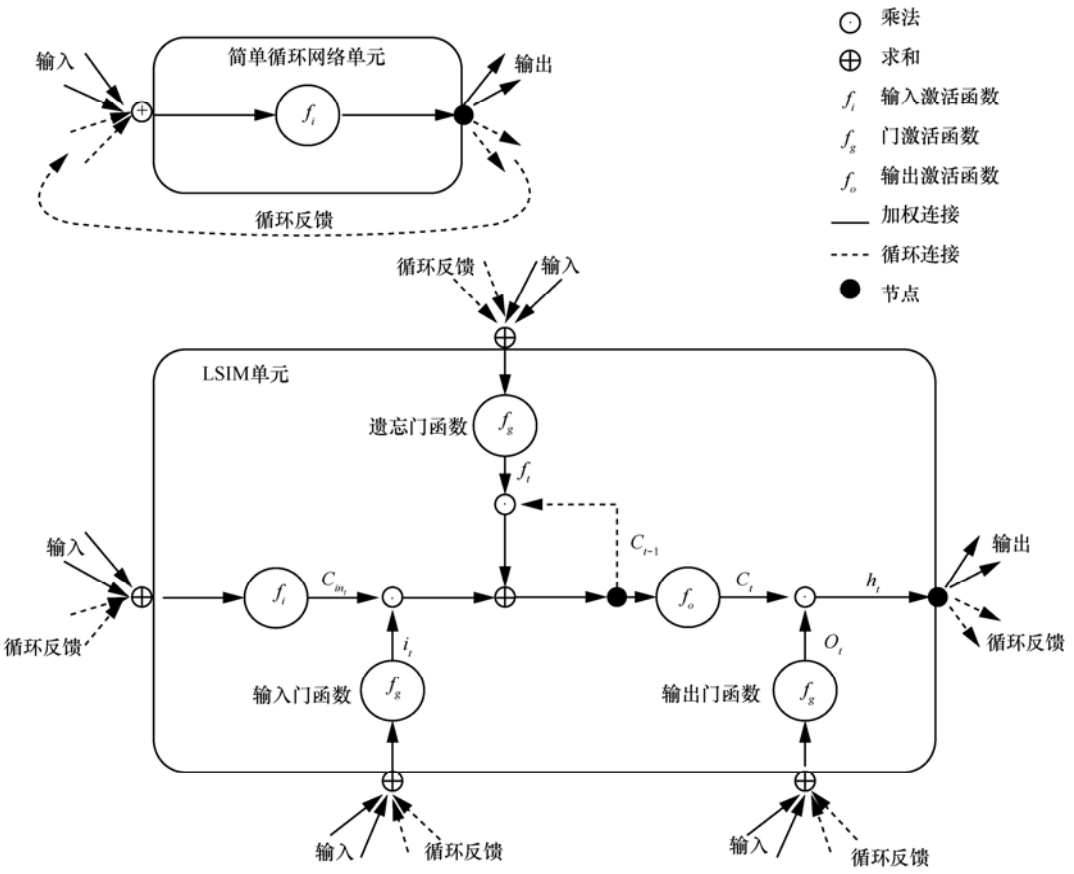


图 4　简单循环网络单元和 LSTM 单元结构对比

一段情节最能引起被测试情感的变化，而被测试在实验后整个 MV 的打分也是基于视频中最能引起情感的片段进行的。因此，通过 LSTM RNN 可以抓取持续 60 s 生理特征序列的关键信息。本文涉及的各种生理信号，都在前期进行了初步特征提取，提取过程中按照不同的时间窗口对原始信号进行了时间窗口划分，然后对时间窗口中的数据进行特征提取。因此，在情感识别过程中需要情感识别分类器具有从时间序列中识别变化的能力。

以前的研究，大部分都是将生理信号按照一个整体进行特征提取之后再识别，或划分之间窗口之后，将不同窗口的特征进行平均之后进行识别，从而基本上忽略了生理特征在时间维度上前后的变化，以及分类方法对于关键特征的记忆特性[13~16]，本文采用 LSTM RNN，正是因为它在对长周期信号进行训练的过程中，可以有别效识别时间序列特征中的关键特征，从而提高情感识别准确率。

3.3　多模态生理信号特征融合及情感识别模型

本节整体介绍了 2 种神经网络生成情感识别模型的过程，方法的结构如图 5 所示。

图 5 展示的是结合 SAE 和 LSTM RNN 搭建的情感识别混合深度神经网络结构。

图 5 中按照纵向的虚线可分为 4 个部分，按照从左到右的次序，各部分第一层和第二层为 SAE 的网络结构，第一层结构用来进行信息压缩，第二层结构是用来进行数据融合；第三层结构是 LSTM RNN 网络结构，主要用来进行情感分类识别；最后一层是网络的输出层，用来输出情感分类识别结果。

最左边的 SAE 网络一共有 4 类输入变量，这 4 类输入变量是提前从生理信号中提取出的特征值。由于本文涉及的生理信号都是与时序相关的，因此，在前期的特征提取中，用不同长度的时间窗口对原始信号进行了划分，并从不同时长窗口的生理信号

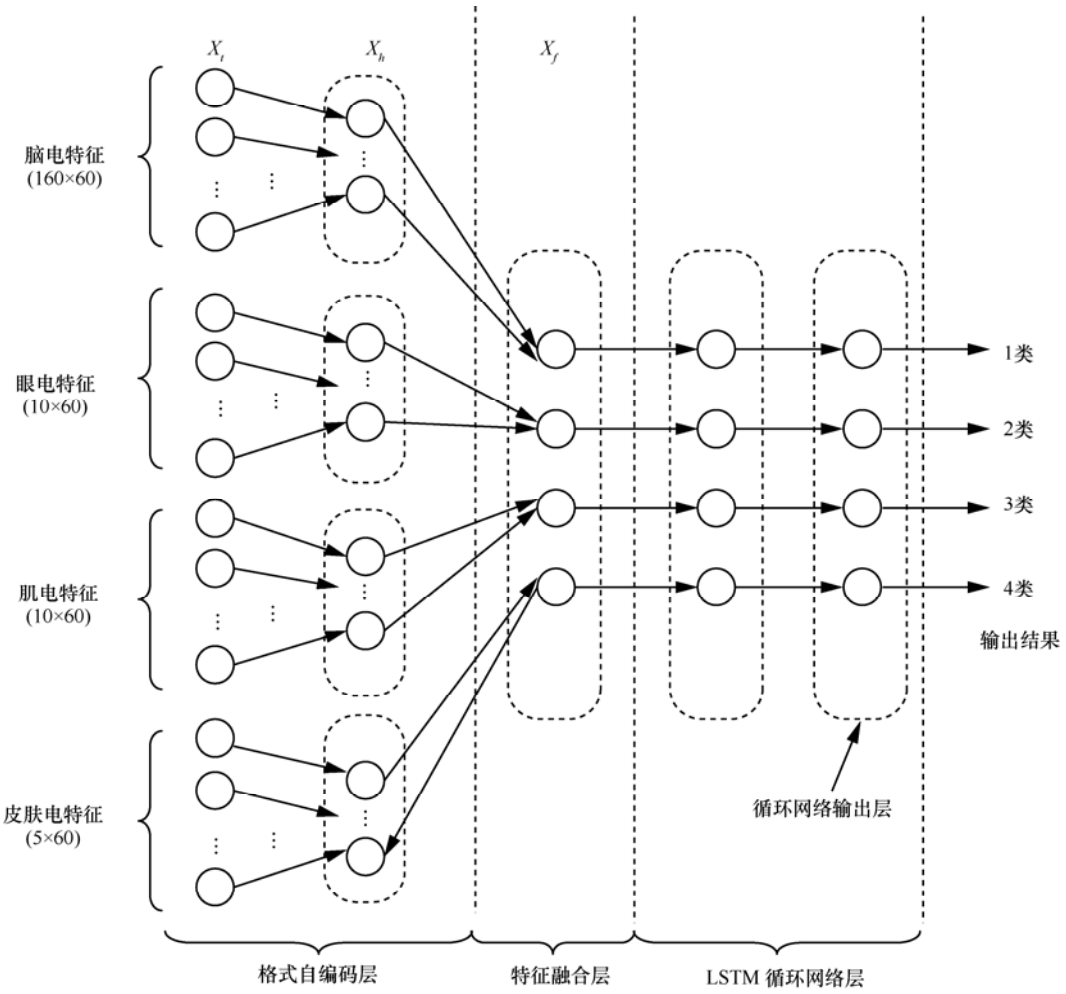


图 5　结合 SAE 与 LSTM RNN 的情感识别混合深度神经网络结构

中进行特征提取，最终得到的特征向量长度也是不同的，以 1 s 时长窗口为例，特征向量的长度为 60。而输入SAE的特征向量的宽度分别是160维度的脑电信号特征，10 维度的眼电信号特征，10 维度的肌电信号特征和 1 维度的皮肤电信号特征。表 4 对上述的特征值进行了归纳。

在生理信号特征提取过程中，主要是提取生理信号的能量特征。数据集中的多模态生理信号可以简单地划分为脑电信号和其他生理信号。对于脑电信号来说，根据之前的研究，人类在清醒时段大脑的活动信息主要集中在频率较高的脑电信号中，由于DEAP实验数据中的脑电信号是在被测试者清醒状态下获取的，因此，脑电中与情感相关的有用信息主要集中于频率较高的脑电波段。以前对于脑电信号的特征提取方法有时域提取法、频域提取法以及时域频域结合提取法。本文为了得到高频段的脑电信息同时又不丢失脑电信号的时域特征，提取脑电信号特征的过程中采用了希尔伯特黄变换的方法对脑电等时序信号进行处理。具体的方法是：首先，将脑电信号通过经验模态分解[19]（EMD, empirical mode decomposition）方法分解为本征模函数（IMF, intrinsic mode function），然后对前 5 个频率较高，能量分布比较集中的前 5 个本征模函数（IMF_1~IMF_5）进行希尔伯特变换，然后从中提取功率谱密度（PSD, power spectral density）作为脑电特征。EMD 方法分解出的不同 IMF 代表了从生理电信号中分解出的能量分量，越靠前的 IMF 包含的能量越多，频率越高。

数据集中其他的生理电信号来源于末梢神经系统，对情感分析起到了信息补充的作用。其中，眼电信号分别采集了左眼角、右眼角、右眼上和右眼下的 4 个眼电信号，对于眼电信号若以 1 s 时长为时间窗口，每个被测试最终得到的特征矩阵的大小为 4×60。肌电原始信号包含了斜方肌和颧肌的肌电信号，跟眼电信号一样提取信号能量作为特征，对信号同样采取 1 s 为时间窗口计算信号能量，每个被测试得到 2×60 维度的肌电信号特征。皮肤电反应采集位置为左手中指和无名指，若以 1 s 时间窗口为例，每个被测试最终得到 1×60 的特征向量。表 4 展示了不同的生理电信号以 1 s 时长为时间窗口提取的特征信息。末梢神经系统的生理信号，在进行特征提取的时候，没有用 EMD 方法分解，而是在对原始信号进行去除噪音信号之后，进行时间窗口的划分，随后在时间窗口内，计算信号的功率谱密度作为信号特征。

实际分析过程中，为了验证 LSTM RNN 在情感识别过程中的时间敏感性，本文对生理信号按照不同时间窗口进行分割，随后在时间窗口内提取特征值形成特征向量。对不同时间窗口对应的特征向量，用同一种识别方法进行情感识别，这样就能对比出时间划分对于分类精确度的影响，分类方法对时间的敏感度以及情感分类精确度提高的原因。时间窗口的时长设置方面，本文按自然时间单位的 1 ~5 s 对原始信号进行划分。

可以预见的是，时间窗口越大，所得到的生物信号特征就越概括，但是其对于情感变化的描述细度就会相对的下降；时间窗口越小，所得到的生理特征对于情感变化的细节描述性就越好，但是对应的信息量也随之变大。

表 4 中涉及 167 种生理特征，将特征输入 SAE 进行的信息压缩过程可用式(10)和式(11)表示为

$$x_{h_n}^{p_n} = \varphi(\boldsymbol{W} x_{i_n}^{p_n} + b) \tag{10}$$

$$x_{h_{n+1}}^{p_{n+1}} = \varphi(\boldsymbol{W} x_{i_{n+1}}^{p_{n+1}} + b) \tag{11}$$

其中，$x_{i_n}^{p_n}$，$x_{h_n}^{p_n}$ 分别为自编码神经网络的输入层变量和隐藏层变量。$x_{i_n}^{p_n}$ 为第 n 层网络的第 i 个输入变量，$x_{h_n}^{p_n}$ 为第 h 个输出变量，p 为输入变量对应的生物信号类型，$\boldsymbol{W}$ 和 b 为自编码神经网络中的权重矩阵和偏离率（$i_n, h_n, p_n \in N$）。式(11)中 $x_{i_{n+1}}^{p_{n+1}} = x_{h_n}^{p_n}$，即 SAE 不同的隐藏层之间相连接。$\varphi$ 是激活函数，本文激活函数取 sigmoid 函数，如式(3)所示。

表 4　　DEAP 多模态生理信号对应 1s 时长窗口提取的特征矩阵维数

生理信号类型（维数）	特征值及描述
脑电信号（EEG）特征值（32×5×60）	32 导脑电数据× 60 s×5 层 IMF 提取的 PSD 特征值
眼电信号（EOG）特征值（4×60）	4 个测量点×60 信号能量特征值
肌电信号（EMG）特征值（2×60）	斜方肌肌电 1×60、颧肌肌电 1×60 信号能量特征值
皮肤电信号 （GSR）特征值（1×60）	1×60　0~2.4 Hz 功率谱能量特征值

假设 x' 为自编码神经网络的输出，那么 $\boldsymbol{W}$ 和 b 通过输出与输入之间的均方差代价函数通过反向传播来确定，均方差代价函数可表达为

$$C(x,x')=\sum_{i=1}^{N}\left\|x^{i}-x^{i'}\right\|^{2} \tag{12}$$

其中，N 为输入变量的个数。依据本文图 5 中的设计，将 SAE 的隐藏层数设为 2 层，第一层为各个模态特征向量内部的信息压缩，第二层为各个模态特征向量之间的信息融合。表 5 中列出了每层的数据宽度，假设所有生理信号都按照 1 s 的时间窗口计算特征值，那么对应于 60 s 的实验时长，提取的特征向量长度为 60，而数据输入层每种生理信号的数据宽度分别为 160、4、2 和 1，第一隐藏层神经元个数的设置表示了进行生理信号信息压缩后特征向量的宽度，脑电信号特征的宽度降为 80，眼电信号降为 2，肌电信号降为 1，皮肤电信号原值输出。第二隐藏层的作用是对 4 类生理信号特征进行信息压缩和信息融合，其总体输入为第一隐藏层的总体输出，即 84×60 的向量矩阵，经过第二隐藏层的信息压缩和融合，输出的特征数据宽度为 64。

表 5　　SAE 每层包含神经元个数

生物信号类别	输入层宽度	第一隐藏层宽度	第二隐藏层宽度
脑电信号	160	80	64
眼电信号	4	2	
肌电信号	2	1	
皮肤电信号	1	1	

每个隐藏层的神经元个数是通过循环训练 SAE，随后对比最终的结果而确定的，在循环训练 SAE 的过程中，通过设定最终 SAE 的损失函数阈值，从输入数据宽度开始逐步减 1，设定隐藏层神经元个数，随后训练 SAE，在总体信息损失不超过 5%的前提下，取隐藏层的最小获取值，作为最终隐藏层宽度。SAE 的具体实现和网络训练操作在 Matlab 中进行。最终 SAE 的输出特征向量存储于 mapped $\boldsymbol{X}$ 向量中，每次模型的训练迭代在判断均方差大于设定阈值（5%）的时候，就停止训练，随后将原特征向量对应的 mapped $\boldsymbol{X}$ 值存入对应的输出特征向量中，特征向量的维度为 64×60。实验的信息压缩和融合步骤中，每个被测试对应的 40 个特征向量被存储为特征文件，作为将要训练的 LSTM RNN 的输入矩阵。

本文采用的 LSTM RNN 的结构如图 5 所示，也由 2 层网络结构组成。第一层是 LSTM 层，然后连接一个普通的循环网络输出层作为整个网络的输出，输出情感识别的最终结果。整个 LSTM RNN 采用双向反馈机制，用随机梯度下降法作为网络的优化算法，循环网络输出层的损失函数为 MCSENT。网络的训练过程中，采用了不同的网络学习率和网络优化算法，网络训练样本按照 3:1:1 的比例分成训练集、验证集和预测集。

4　实验结果

本节主要将上述方法应用到 DEAP 数据集上，进行多模态融合生理特征情感识别的实验结果。结果主要从 2 个方面进行展示：1) 对比以脑电为代表的单模态生理信号和多模态生理融合信号在进行情感分类时的分类精确度结果，目标是展示和分析多模态融合信号在情感分类中的优势；2) 从分类方法出发，对比 LSTM RNN 与其他 4 种分类方法在对多模态生理信号进行情感分类分析中的优劣。

为了说明第一点结果，本文将脑电信号和末梢神经系统生理信号分别进行压缩和融合后、输入 LSTM RNN 进行情感识别，同时也按照前面的表述，将两者之间进行融合之后、输入 LSTM RNN 进行情感识别，最终以情感识别准确率和 F1 得分来展示多模态生理信号识别和单模态生理信号识别之间的差别。实验的结果记录在表 6 中。表 6 中显示将不同模态的生理信号通过 LSTM RNN 网络进行训练，脑电信号的最好分类准确率为 0.762 3，对应的 F1 得分为 0.744 8；通过皮肤电、眼电和肌电信号等末梢神经生理信号进行情感识别，最好分类精确度为 0.543 7，对应的 F1 得分为 0.528 7；而将两者结合后生成的多模态生理信号通过 LSTM RNN 进行分类识别研究，其最终的分类正确率得到了提升，最好的分类精确度为 0.792 6，F1 得分为 0.768 2。从表 6 的结果，可以说明末梢神经系统的生理信号在进行情感分类识别的过程中对产生于大脑的脑电信号是有补充作用的，通过融合过后的多模态生理信号进行情感分类识别，能够达到更高的分类识别精度。

为了说明 LSTM RNN 方法在对多模态融合生理信号进行情感分类分析时的优势，尤其 LSTM 在对长周期数据分析过程中的优势，本文在特征提取阶段按照不同时长窗口提取了生理信号特征，随后

表 6　单模态信号与多模态融合信号通过 LSTM RNN 进行情感分类训练得到的最高精确度以及 F1 对比

信号类型	准确率	F1 得分
皮肤电信号、眼电信号和肌电信号特征	0.543 7	0.528 7
脑电信号特征	0.762 3	0.744 8
多模态融合信号	0.792 6	0.768 2

将信号特征按照同样的信息压缩和融合方法进行了处理。然后将融合过的特征向量用不同的分类方法求出每种分类方法得到的最优分类精确度，并将结果在表 7 中给出了展示。

表 7 中，前 3 种分类方法的实验过程是在 Matlab 平台中调用相应的分类函数进行分析得出的结果。后 2 种方法是调用 DL4J 开源框架中的 RNN 和 LSTM RNN 学习方法进行分类训练得到的分类结果。为了更加直观地展示表 7 中不同方法对时间窗口的敏感性，以及各种方法分类精度的特征，本文以箱形图的形式将表 7 中获得的分类精度展示到图 6 中。从图 6 可以看出，LSTM RNN、RNN 和 SVM 获得了较好的分类结果。同时，在对时间窗口的敏感性方面，LSTM RNN 对于用不同时长窗口的特征值进行情感分类是敏感的。从表 7 中得到的结果可以看出，LSTM RNN 在对以 3 s 为时间窗口的特征进行分类时，得到了分类精确度的最高值；不带有 LSTM 单元的 RNN 方法与 SVM 方法对时长窗口的长短在分类过程中敏感性不强。对于不带有 LSTM 单元的 RNN 来说，正如第 3.2 节中对于普通 RNN 和 LSTM RNN 的对比分析一样，LSTM 单元通过记忆门能够将长时间序列中的间隔较长的特征变化保持到整个网络结构中，而 RNN 仅仅反馈了相邻连续时间窗口之间的变化，相隔更远的关键特征变化在进行网络迭代训练时容易被微小特征变化最终平均了。

本文 SVM 的分类方法，主要用高斯核函数的 SVM 分类法，通过构造“一对一”的多分类器，对多得到的特征矩阵进行分类。通过图 6 中 SVM 的箱线图可以看出，SVM 对于不同时长窗口对应的特征值进行情感分类，分类精确度变化不大。表 7 中 LSTM RNN 得到对应于 7 种时长窗口的特征分类精度，其中，3 s 时间窗口对应的分类精确度最高为 0.792 6，而以 60 s 为时间窗口得到的特征值对应的分类精确度最低。1~3 s 分类精确度是上升的趋势。3~60 s 时长窗口对应的分类精确度都是下降趋势。这说明，在通过 LSTM RNN 方法对融合过的多模态情感进行分类时，3 s 时长窗口对应得到的特征值具有良好的情感分类特征。

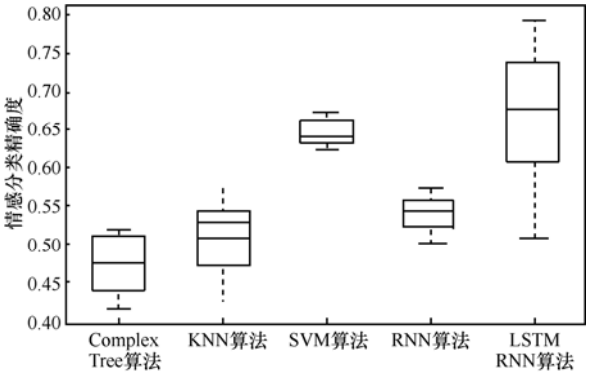


图 6　不同时长窗口对应的融合特征通过不同分类算法得到的分类结果

如表 8 所示，本文对比了相关的研究结果。对比研究选择的依据如下。首先，研究内容都是通过生理信号对情感进行分类识别研究；其次，在实验中，所选的刺激媒体都是视频片段。而采集的生理信号有单模态生理信号也有多模态生理信号，表 8 中被识别的情感类型表示了研究中能

表 7　用不同的分类方法对不同时长窗口的特征进行情感分类所得情感识别精度

分类方法	特征提取时间窗口时长						
	1 s	2 s	3 s	4 s	5 s	10 s	60 s
Complex Tree	0.413 2	0.453 4	0.432 3	0.477 2	0.519 1	0.512 3	0.511 2
KNN	0.423 2	0.463 4	0.482 3	0.527 2	0.549 1	0.572 3	0.531 2
SVM	0.637 2	0.651 0	0.672 1	0.641 0	0.628 1	0.665 7	0.626 0
RNN	0.572 1	0.558 2	0.543 9	0.551 3	0.533 7	0.521 7	0.501 0
LSTM RNN	0.594 3	0.648 9	0.792 6	0.755 6	0.683 4	0.678 1	0.507 5

表 8　　**情感分类识别相关研究识别正确率情况**

主要研究者	情感刺激媒体	识别情感类型	被测试个数	生理信号种类	数据分析方法	正确率
Agrafioti	视频，游戏	正负唤醒度和效价	44	心电信号	EMD	正向唤醒度：0.784 3 负向唤醒度：0.524 1 留一交叉验证
Bailenson	视频	有趣、悲伤、平静	41	脸部表情、心电、皮肤电传导、体细胞活性	WEKA 数据分析平台中的相关分析方法	对于有趣情感类型，结合脸部表情和生理信号识别率达到 0.90 二折交叉验证
Lisetti	电影片段 数学难题	悲伤、气氛、害怕、吃惊、挫败、有趣	29	皮肤电反应、心率、体温	KNN	对于不同的情感类型识别率 0.704~0.809 留一交叉验证
Fleureau	视频片段 声音片段	事件判定、效价判定	10	皮肤电反应、肌电	高斯函数	对效价的最好分类达到 0.854 1 二折交叉验证
Chung	视频片段	唤醒度、效价、喜欢程度	32	脑电、肌电、皮肤电、血压、体温、呼吸	Bayes classifier	0.666 效价 0.664 唤醒度
Li	视频片段	唤醒度、效价、喜欢程度	32	脑电、肌电、皮肤电、血压、体温、呼吸	CNN+RNN	0.720 6 效价 0.741 2 唤醒度
Koelstra	视频片段	唤醒度、效价、喜欢程度	32	脑电、肌电、皮肤电、血压、体温、呼吸	Single-trial Classification	0.616~0.647

够识别的情感类型。其中，最后的 3 个研究，都是对 DEAP 数据集的研究。文献[20]通过单一的心电信号对 44 个被测试的情感状态进行了分析，其情感识别类型为将唤醒度和效价分开进行衡量，根据 2 个评价指标得分的中位数将唤醒度和效价分成正负，随后通过对心电信号作 EMD 分解，提取特征向量进行分析，最终得到的正确率如下。正向唤醒度为 0.784 3，负向唤醒度为 0.524 1。将本文的研究结果与文献[20]进行对比，除去实验数据的不同，文献[20]仅采用了一种生理信号，并且对情感的分类仅为通过正负唤醒度或效价进行分类，分类情况简单，而本文采用了多模态生理信号，并且对情感分类过程中，将唤醒度和效价进行综合评价后分 4 类，最终得到的分类精确度最高为 0.792 6；文献[21]通过 41 个被测试观看电影片段的方式来诱发被测试的悲伤和有趣 2 种情感，在实验中记录了被测试的脸部表情、心电、皮肤电传导特性等生理信号，然后从被测试个体角度，性别角度以及整体被测试角度对被测试的情感类型进行了分析，文中结合脸部表情信号和生理信号对被测试的情感进行了识别，其最高分类精度是对于有趣情感，最高识别率为 0.90，文献[21]中较高的情感识别率，来自于表情信号而非生理信号的贡献；文献[22]通过电影片段和数学难题对 29 名被测试进行了实验，然后用 KNN 方法对不同的情感类别进行了分析，对不同的情感分类精确率从 0.704 到 0.809；文献[23]通过视频和声音片段，对 10 名被测试进行了刺激，记录了被测试的皮肤电反应信号和肌电等信号，通过高斯函数方法对信号进行分析后得到，对单个被测试效价平均分类准确率最高为 0.851 2，对于多用户效价平均分类准确率为 0.50，效价最好的分类准确率达到 0.854 1；文献[14, 24, 25]是对同一个数据集 DEAP 的研究，其中，文献[14]是开发 DEAP 的团队对数据集的分析研究，文献[14]通过单次分类方法对数据集中的数据进行了融合的多模态数据分析，其最好的分类结果如下。唤醒度识别准确率为 0.616，效价识别准确率为 0.647，喜欢程度识别准确率为 0.618；文献[24]是通过贝叶斯分类方法分别对效价和唤醒度的分类进行了识别，两者单独进行分类识别为 0.666 和 0.664；文献[25]中通过卷积神经网络和循环网络的方式，对 DEAP 数据提取了特征并进行了分类分析，其中，效价的分类精确率达到 0.720 6，而唤醒度的分类准确率达到了 0.741 2。本文通过融合的多模态数据对唤醒度和效价分 4 类进行识别，综合的分类识别率达到 0.792 6，比上述的相似性研究和在同一数据集上的研究都有了相应的提高。

5　结束语

本文通过提出一种对多模态生理信号进行融合分析的方法，以期有效提高从生理信号进行情感

的识别准确率。具体的做法是，首先，将开源数据集 DEAP 中脑电信号、皮肤电信号、肌电信号和眼电信号在时域进行分段，随后各自提取频域的特征，将提取出的特征采用栈式自编码神经网络的方法，对不同模态的特征进行信息压缩和融合，形成融合过的特征矩阵。然后，将融合的多模态特征作为带有长短周期记忆单元的循环神经网络的输入数据进行情感分类识别分析。本文将得到的实验结果从 3 个方面进行了对比分析。首先，本文对比了单模态生理信号与多模态生理信号在相同的分类方法进行情感分类识别，比较发现在单模态生理信号中脑电信号的分类率最高，达到了 0.762 3，而通过融合的多模态生理信号进行情感识别时，情感分类率达到 0.792 6，说明皮肤电信号、眼电信号以及肌电信号在进行情感分类识别过程中，对于脑电信号来说是有益的补充，能够提高情感分类的准确率；其次，本文在进行情感分类识别时，对于多模态生理信号进行特征提取时，时长窗口长短的设置，进行了对比，通过将不同时长窗口对应的特征值，输入 LSTM RNN 网络进行训练，发现不同的时长窗口对应的情感分类准确率是不同的，这个结果说明：1) LSTM RNN 在进行分类的过程中对于时间窗口的长短是敏感的，2) 对比 1~60 s 的时长窗口对应的情感分类准确率的时候，发现 LSTM RNN 对应的 3 s 的时长窗口对应的分类准确率最高；同时，在将 LSTM RNN 与其他的分类方法进行对比时发现，LSTM RNN 的方法在进行分类时得到了较高的分类准确率。再次，将本文所提方法与现有的多模态生理信号情感识别研究进行了对比，通过本文所提方法进行情感分类分析所得到的分类精度达到 0.792 6，与在同一数据集上的研究对比，分类准确率也得到了相应提高。

需要注意的是本文在进行多模态生理信号进行分类识别的过程中，LSTM RNN 损失函数的选择以及网络层数的设置仍需要进一步的研究，这也是下一阶段研究工作的主要内容。

参考文献：

[1] 聂聃, 王晓韡, 段若男,等. 基于脑电的情绪识别研究综述[J]. 中国生物医学工程学报, 2012, 31(4):595-606.
NIE D, WANG X H, DUAN R N, et al. A survey on EEG based emotion recognition[J]. Journal of Biomedical Engineering, 2012, 31(4): 595-606.

[2] JONGHWA K, ANDRE E. Emotion recognition based on physiological changes in music listening[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30: 2067-2083.

[3] 赵力, 钱向民, 邹采荣, 等. 语音信号中的情感识别研究[J]. 软件学报, 2001, 12(7):1050-1055.
ZHAO L, QIAN X M, ZOU C R, et al. A study on emotional recognition in speech signal[J]. Journal of Software, 2001, 12(7):1050-1055.

[4] 林奕琳, 韦岗, 杨康才. 语音情感识别的研究进展[J]. 新能源进展, 2007, 12(1):90-98.
LIN Y L, WEI G, YANG K C. A survey of emotion recognition in speech[J]. Journal of Circuits and Systems, 2007, 12(1):90-98.

[5] 赵腊生, 张强, 魏小鹏. 语音情感识别研究进展[J]. 计算机应用研究, 2009, 26(2):34-38.
ZHAO L S, ZHANG Q, WEI X P. Survey on speech emotion recognition[J]. Application Research of Computers, 2009, 26(2):34-38.

[6] OTHMAN M, WAHAB A, KARIM I, et al. EEG emotion recognition based on the dimensional models of emotions[J]. Procedia-Social and Behavioral Sciences, 2013, 97(2):30-37.

[7] 陈曾, 刘光远. 脑电信号在情感识别中的应用[J]. 计算机工程, 2010, 36(9):168-170.
CHEN Z, LIU G Y. Application of EEG signal in emotion recognition[J]. Computer Engineering, 2010, 36(9):168-170.

[8] 张栋, 陈东伟, 游雅, 等. 基于自适应 Lempel-Ziv 复杂度的情感脑电信号特征分析[J]. 计算机应用与软件, 2014(9):162-165.
ZHANG D, CHEN D W, YOU Y, et al. Analyzing emotional EEG signals feature based on adaptive LEMPEL-ZIV complexity[J]. Computer Applications and Software, 2014(9):162-165.

[9] UPASANA T, SHYAMANTA M H. Estimation of mental fatigue during EEG based motor imagery[C]//IHCI 2016: Intelligent Human Computer Interaction. 2016:122-132.

[10] BAJAJ V, PACHORI R B. Detection of human emotions using features based on the multiwavelet transform of EEG signals[M]. Springer International Publishing, 2015:215-240.

[11] HOSSEINI S A, NAGHIBISISTAN M B. Emotion recognition method using entropy analysis of EEG signals[J]. International Journal of Image Graphics & Signal Processing, 2011, 3(5):30-36.

[12] 王凯明, 钟宁, 周海燕. 基于改进功率谱熵的抑郁症脑电信号活跃性研究[J]. 物理学报, 2014, 63(17):178701-178701.
WANG K M, ZHONG N, ZHOU H Y. Activity analysis of depression electroencephalogram based on modified power spectral entropy[J]. Acta Phys Sin, 2014, 63(17):178701-178701.

[13] KREIBIG S D. Autonomic nervous system activity in emotion: a review[J]. Biological Psychology, 2010, 84(3):394-421.

[14] KOELSTRA S, MUHL C, SOLEYMANI M, et al. DEAP: a database for emotion analysis; using physiological signals[J]. IEEE

Transactions on Affective Computing, 2012, 3(1):18-31.

[15] PAUL E. An argument for basic emotions[J]. Cognition and emotion, 1992, 6(3/4):169-200.

[16] POSNER J, RUSSELL J A, PETERSON B S. The circumplex model of affect: an integrative approach to affective neuroscience, cognitive development, and psychopathology[J]. Development and Psychopathology, 2005, 17(3):715-734.

[17] ZHANG P, MA X, ZHANG W, et al. Multimodal fusion for sensor data using stacked autoencoders[C]//IEEE Tenth International Conference on Intelligent Sensors, Sensor Networks and Information Processing. 2015.

[18] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[19] HUANG N E, SHEN Z, LONG S R, et al. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. The Royal Society, 1998, 454(1971): 903-995.

[20] AGRAFIOTI F, HATZINAKOS D, ANDERSON A K. ECG pattern analysis for emotion detection[J]. IEEE Transactions on Affective Computing, 2012, 3(1):102-115.

[21] BAILENSON J N, PONTIKAKIS E D, MAUSS I B, et al. Real-time classification of evoked emotions using facial feature tracking and physiological responses[J]. International Journal of Human-Computer Studies, 2008, 66(5):303-317.

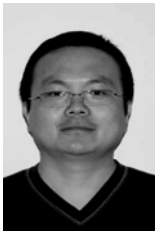
[22] LISETTI C L, NASOZ F. Using noninvasive wearable computers to recognize human emotions from physiological signals[J]. Eurasip Journal on Advances in Signal Processing, 2004, 2004(11): 1672-1687.

[23] FLEUREAU J, GUILLOTEL P, QUAN H T. Physiological-based affect event detector for entertainment video applications[J]. IEEE Transactions on Affective Computing, 2012, 3(3):379-385.

[24] CHUNG S Y, YOON H J. Affective classification using Bayesian classifier and supervised learning[C]//International Conference on Control, Automation and Systems. 2012:1768-1771.

[25] LI X, SONG D, ZHANG P, et al. Emotion recognition from multi-channel EEG data through Convolutional Recurrent Neural Network[C]//IEEE International Conference on Bioinformatics and Biomedicine. 2017:352-359.

作者简介：



李幼军（1978-），男，河南栾川人，北京工业大学博士生，主要研究方向为生物信号分析、机器学习及情感计算等。

黄佳进（1977-），男，贵州遵义人，博士，北京工业大学助理研究员，主要研究方向为人工智能、推荐系统等。

王海渊（1981-），男，山西朔州人，博士，北京工业大学工程师，主要研究方向智能传感器、人工智能、智慧医疗系统的开发等。

钟宁（1956-），男，北京人，北京工业大学教授、博士生导师，主要研究方向为人工智能、Web 智能、脑信息学、知识发现与数据挖掘、粒计算等。