Journal of System Simulation

Volume 34 | Issue 4 Article 10

4-20-2022

Multi-modality Affective Computing Model Based on Personality and Memory Mechanism

Sijin Zhou
1.School of Engineering, Shantou University, Shantou 515063, China;, 19sjzhou@stu.edu.cn

Dicheng Chen
1.School of Engineering, Shantou University, Shantou 515063, China;

Geng Tu
1.School of Engineering, Shantou University, Shantou 515063, China;

Dazhi Jiang
1.School of Engineering, Shantou University, Shantou 515063, China;2.Intelligent Manufacturing Key Laboratory of Ministry of Education, Shantou University, Shantou 515063, China;, dzjiang@stu.edu.cn

Multi-modality Affective Computing Model Based on Personality and Memory Mechanism

Abstract

Abstract: With the development of affective computing, the correlation of memory, individuation and emotion is more and more important. *Focus on the machine emotion shortcomings in the perception, understanding and expression, an emotion computing model integrating the emotion perception, understanding and expression is proposed. The model is a memory-oriented deep network perception model that accepts multiple modal inputs (visual, auditory, lexical) and applies a fuzzy emotion integration decision to realize the understanding of uncertain emotions*. The simulation experiments prove that the model has a good performance in all kinds of multimodal affective computing.

Keywords

affective computing, computational model of emotion, deep neural network, integrated decision, personalization

Recommended Citation

Sijin Zhou, Dicheng Chen, Geng Tu, Dazhi Jiang. Multi-modality Affective Computing Model Based on Personality and Memory Mechanism[J]. Journal of System Simulation, 2022, 34(4): 745-758.

第 34 卷第 4 期 系统仿真学报© Vol. 34 No. 4
2022 年 4 月 Journal of System Simulation Apr. 2022

基于个性化和记忆机制的多模态情感计算模型

周思锦[1]，陈棣成[1]，涂耿[1]，姜大志[1,2*]
(1. 汕头大学 工学院，广东 汕头 515063；2. 汕头大学 智能制造技术教育部重点实验室，广东 汕头 515063)

摘要：随着情感计算研究的不断深入，记忆、个性化、情感之间的密切联系逐渐引起研究者的重视。已有方法在机器情感的感知、理解和表达方面仍存在着诸多不足。提出一种集情感感知、理解和表达于一体的情感计算模型。该模型是一个面向记忆机制的、接受多模态输入（视觉、听觉、词汇）的深度网络感知模型，并应用一种模糊化的情感集成决策来实现不确定性情感的理解。实验证明：该模型在各类多模态的情感计算中都有着较好的表现。

关键词：情感计算；计算情感模型；深度神经网络；情感集成决策；个性化
中图分类号：TP311 文献标志码：A 文章编号：1004-731X(2022)04-0745-14
DOI: 10.16182/j.issn1004731x.joss.20-0899

Multi-modality Affective Computing Model Based on Personality and Memory Mechanism

Zhou Sijin[1], *Chen Dicheng*[1], *Tu Geng*[1], *Jiang Dazhi*[1,2*]
(1. School of Engineering, Shantou University, Shantou 515063, China; 2. Intelligent Manufacturing Key Laboratory of Ministry of Education, Shantou University, Shantou 515063, China)

Abstract: With the development of affective computing, the correlation of memory, individuation and emotion is more and more important. *Focus on the machine emotion shortcomings in the perception, understanding and expression, an emotion computing model integrating the emotion perception, understanding and expression is proposed. The model is a memory-oriented deep network perception model that accepts multiple modal inputs (visual, auditory, lexical) and applies a fuzzy emotion integration decision to realize the understanding of uncertain emotions*. The simulation experiments prove that the model has a good performance in all kinds of multimodal affective computing.

Keywords: affective computing; computational model of emotion; deep neural network; integrated decision; personalization

引言

没有情感的机器不可能是智能的[1]，而要使机器人具有更人性化的情感智能，需要建立友好的人机情感交互模型，并赋予机器人深度的感知、理解和表达能力。这是从弱人工智能发展到强人工智能的必经之路，亦是情感计算的远景目标[2]。要全面提升机器人的感知、理解与表达能力，让面向机器人的情感计算更具现实可操作性，记忆与个性化是两个重要的组成元素[3-4]。

在情感研究中，Oatley等[5]提到，情感是人与人之间的沟通，它由行为、生理变化和由思想和外部事件引起的主观体验组成。Davidson等[6]认为情感是一种短暂的心理和生理现象，它代表了人

收稿日期：2020-11-15 修回日期：2021-01-05
基金项目：国家自然科学基金(61902232，61902231)；广东省自然科学基金(2019A1515010943)；广东省普通高校基础研究与应用基础研究重点项目(2018KZDXM035)；广东省普通高校基础研究与应用基础研究人工智能重点领域专项(2019KZDZX1030)
第一作者：周思锦(1995-)，男，硕士生，研究方向为情感计算。E-mail：19sjzhou@stu.edu.cn
通讯作者：姜大志(1982-)，男，博士，教授，研究方向为人工智能、情感计算。E-mail：dzjiang@stu.edu.cn

体对环境变化的适应模式。Scherer 等[7]认为情感反映了环境造成的心理和生理状态。情感计算可以分为 2 类：①基础性情感分析；②人工情感模拟。

基础性情感分析，主要面向的是图像、视频、文本、生理信号等情感分析。这一类工作的重点主要是特征提取。目前常见的特征类别包括音频特征、视觉特征、唤醒特征和价值特征。Irie 等[8]提取了音高、短时能量、梅尔频率倒谱系数(melfrequency cepstral coefficients, MFCCs)作为音频特征，颜色重心、图像亮度、拍摄持续时间作为视觉特征，利用了线性判别分析(linear discriminant analysis, LDA)将情感分为 9 类，模型准确率可达 85.5%。Kang[9-10]提取了颜色、运动、镜头切割率等视觉特征，并利用隐马尔可夫模型(hidden Markov model, HMMs)将情感分为了恐惧、悲伤和欢乐 3 类，模型准确率为 87.6%。Zhang[11-12]等选取了运动强度、短切换率、过零率、节奏和拍子强度作为唤醒特征，亮度、饱和度、色彩能量、节奏规则性和音高作为价值特征，也获得了较好的情感分类效果。

情感分类的另一个主要分支是模拟人的情绪，其可以应用于各种智能化的人机交互产品，例如，社交机器人[13]可以感知和模拟情绪并改善在人机交互的表现。情感的激活需要通过一系列外在因素与内在因素相互作用来实现，是一个复杂的情感状态转移过程[14]。部分研究者借助认知心理学相关知识构建计算情感模型，该计算情感模型受情绪事件、情感影响因素和人格特征的影响[15-16]。文献[17]研究了大五人格(five-factor model, FFM) 对状态转换矩阵的具体影响以及人格表达与情绪事件之间的关系，并根据情感能量理论[18]建立了记忆机制。此外，与人类有关的情感特征，如：感知、记忆、情感信息，以及个性影响，都被纳入了机器人的情感模型[19-20]。

在情感计算领域，大多数研究者较为关注情感分类的准确率，忽视情感计算的难点在于面向机器的情感感知、理解与表达上。本文结合现在阶段性需求，情感计算尚待挖掘的问题至少包括：

(1) 情感个性化作用：对于情感分析问题，现有研究太过依赖样本的标签，忽视情感产生过程中客体的个性化因素及其对情感产生带来的不确定性影响[21]。

(2) 情感识别的机械化：目前情感理解以机械化方式为主，先进行特征工程，再进行情感辨识，最终以类标签预测的准确性为目标。这些均是标准的模式识别过程，缺乏考虑人类情感产生过程中记忆与个性化等特性所带来的影响，极大地限制了情感理解的现实可用性[22]。

(3) 人工情感缺乏现实可操作性。人工情感目前的研究主要关注情感的表达。现有的情感状态转移模型具有一定的开创性，但相对基础，尤其是在刺激物这一基本要素的处理上较为抽象，不具有现实可操作性。另外，许多模拟实验不是采用真实数据，这势必导致仿真效果缺乏良好的现实意义。

一个更为合理、可靠的计算情感模型应该从现实可用性和个性化出发。诚如 Scherer K. R. R. 所说：计算情感模型不能为了艺术而艺术，为了情感而情感，计算情感模型的设计需要明确具体的目标[23]。Frijda N H 曾概述了情感建模中的一些基础性的问题[24]，其中 2 个较为重要的问题分别是：①如何从过程的角度来理解情绪？②如果给定某个具体情境或事件，那么人产生各种相关情绪的条件是什么？

基于上述问题，本文针对情感感知、理解和表达提出了 3 种建模方法并将其综合成一个整体的情感研究框架，并经过一系列实验证明了模型的合理性与有效性。在情感分类中，运用模糊数学的相关理论，在保持一定分类精度的前提下，适当地使分类结果模糊化，以此来体现情感的不确定性。

DOI: 10.16182/j.issn1004731x.joss.20-0899

1 方法

1.1 长记忆体结合双向全模态，LSTM 网络

1.1.1 最小情感单位与数据集构建

本文分析的主要素材是音乐视频。音乐视频是由音频、片段和歌词组成的，在做情感分析的时候，需要对音乐视频进行切分，而切分的结果应该是一系列有意义的单元。文献[25]提出了最小情感单位 MSU(minimal sentiment unit)的概念，如图 1 所示，MSU 被用来定义一个最小的情感判别单位。

实验选取了 30 个音乐视频，包括 15 个表达积极情感(激励、温暖、浪漫、快乐等)、15 个消极主题(失恋、思乡、沮丧等)，每个音乐视频的时长最小为 230 s，最长为 340 s。然后，通过手工操作，将 30 个音乐原声带切割成 498 个 MSUs。在解决问题之前，给出以下假设：

(1) 每一个镜头都包含情感内容，这会让测试者产生一些特定的情感。

(2) 测试员情绪只由当前镜头决定。

(3) 对于每一个镜头，所有测试者标记的情感标签都是近似正态分布的。

(4) 每个测试人员都专心于测试，没有任何干扰。

为了对 498 个 MSUs 进行人工情感标注。招募 60 名志愿者(30 名男性，30 名女性)进行情感标注测试。在年龄组成上，23 岁(不含)以下的被试者 11 人(占 18%)，23~30 岁的被试者 35 人(占 59%)，30 岁(不含)以上的被试者 14 人(占 23%)，被试者主要集中在青年，符合音乐原声带流行活跃的年龄层面。在学历分布上，专科及以下的被试 15 人(占 25%)，本科生被试 23 人(占 39%)，研究生被试 22 人(占 36%)，被试者知识层面分布较为平均，满足实验假设情感激发与客观知识储备无关。尽量避免选取音乐艺术专业有关的测试者，因为他们往往会用专业的角度判断情感类别，而不是真切地通过视频引起情感共鸣。

将每个测试者单独安排到不同的静谧房间，每个测试者带上耳机后将专心于测试，没有受到任何场外干扰。电脑会每隔 10 s 播放一个 MSUs，测试者观看后(可选择回放)，将会有共 20 s 的标注时间与平缓情绪间隔，以保证测试者的情感只由当前镜头决定，对后测 MSUs 无影响。为了防止审美疲劳，每个测试者只需随机完成 100 个 MSUs 视频情感备注。最后，对标注时间过短、填写不全及大量填写相同选项的标注表进行筛选和剔除，有效标注表份数为 54 份(回收率为 0.90)。

这一数据集其实在规模上并不足够庞大，为此，采用滑动窗口的技术(滑动窗口的长度为 17)对每个 MSU 进行再处理，形成更多的样本数据，最终数据集大小为 498×17。随机抽取 3/4 的样本作为训练集。另外，剩下的 1/4 是测试集。为了数据集分类标签的分布均衡，在实际划分训练集和测试集后，通过复制部分序列数据对 2 个数据集进行配平。

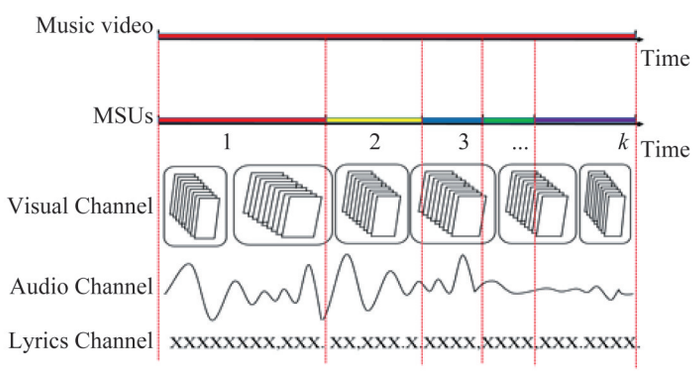


图 1　音乐原声带分解成若干个 MSU 的过程概述

Fig. 1　An process overview of music soundtrack being decomposed into several MSU

1.1.2 特征提取

因为其包含了视觉特征、听觉特征和文本特征 3 种类别，所以视频的特征提取工作比其他数据要复杂得多。

在听觉特征方面，本文整合了文献[26-29]的观点，分别提取了每个视频分段内，频谱带宽、过零率、MFCC 的均值和方差，通过恒 Q 变换(constant Q transform, CQT) 得到的音色特征值的均值和方差，频谱滚降点(spectral roll-off)的均值和方差，以及节拍、音调、频谱带宽等音乐特征。

在视觉特征方面，本文整合了文献[30-32]的观点，主要提取了 3 个部分的特征：①镜头特征，包括视频分段里面的镜头切换率(即有多少个镜头)和镜头的平均时长；②色彩特征，包括视频分段、每两个相邻镜头之间关键帧亮度的均值和方差、色温的均值和方差、平均色调、平均饱和度、平均明度，以及按 RGB 统计的累计彩色直方图的最高峰和次高峰的下标；③帧间的动作强度特征，这部分计算的是关键帧的后一帧与前后两帧之间绝对差值的平均值。

在语言特征方面，利用了自然语言处理中常用的 Word2Vector，把所有歌词集中起来构建 Word2Vector 形成属于语言上的特征[33]。

1.1.3 LMFDB-LSTM 网络结构设计

在提取了各个多模态(视觉、听觉和语言)的特征以后，使用双向长短期记忆网络(bidirectional long short-term memory，Bi-LSTM)作为基础模型，如图 2 所示为双向单模态 LSTM。

定义正序 LSTM(S_0~S_n)为正向短时记忆，定义逆序 LSTM(S'_0~S'_n)为反向短时回忆 LSTM。对于每一个 MSU，假设分片成 n 份，每一份进入双向单模态 LSTM 进行训练时都会得到 4 个预测值，分别是正向短时记忆预测、正向短时记忆分类、反向短时回忆预测和反向短时回忆分类。假设当前分片为 X_i 输入网络后，从正向 LSTM 中(蓝色线)得到了正向短时记忆分类 Y_i 和正向短时记忆预测 Y_{i+1}；从反向 LSTM 中(红色线)得到了反向短时回忆分类 Y'_i 和反向短时回忆预测 Y'_{i-1}；利用预测值与真实值的差，可重构 Loss 中的 Y 分类输出，计算方法为

$$\hat{Y}_{i,\text{ final}}=\left(1-\frac{\hat{Y}'_{i-1}-Y_{i-1}}{(\hat{Y}'_{i-1}-Y_{i-1})+(\hat{Y}_{i+1}-Y_{i+1})}\right)\hat{Y}_i'+\left(1-\frac{\hat{Y}_{i+1}-Y_{i+1}}{(\hat{Y}'_{i-1}-Y_{i-1})+(\hat{Y}_{i+1}-Y_{i+1})}\right)\hat{Y}_i \tag{1}$$

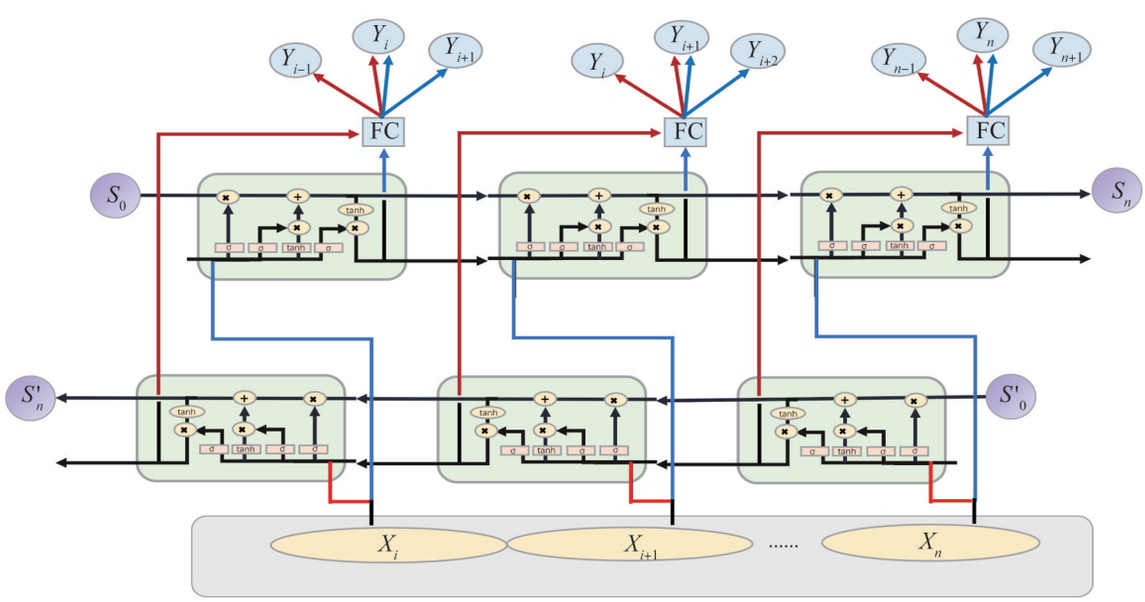


图2 双向单模态 LSTM 结构示意图

Fig. 2 Schematic diagram of bidirectional single mode LSTM structure

从式(1)可以得知，最后的输出分类 $\hat{Y}_{i,\text{final}}$ 是正向分类 $\hat{Y}_i$ 与其 i+1 预测的可信度分量 $\left(1-\frac{\hat{Y}_{i+1}-Y_{i+1}}{(\hat{Y}'_{i-1}-Y_{i-1})+(\hat{Y}_{i+1}-Y_{i+1})}\right)$ 的乘积，加上反向分类 $\hat{Y}'_i$ 与其 i–1 预测的可信度分量 $\left(1-\frac{\hat{Y}_{i+1}-Y_{i+1}}{(\hat{Y}'_{i-1}-Y_{i-1})+(\hat{Y}_{i+1}-Y_{i+1})}\right)$ 的乘积，其中 2 个分量是动量因子。

类似地，还可以设计两两组合的双向双模态 LSTM 模型和双向三(全)模态的 LSTM，为后续的情感集成决策作准备，如图 3~4 所示。双向双模态 LSTM 模型是 2 个不同模态的双向单模态 LSTM 的结合体，将各自的隐层状态 h_A 与 h_B 共享合成一个独有的 h_{tot} 再返还各自网络进行分类，其更新公式为

$$\text{Bi\_LSTM}_A\begin{cases}h_{t-1}^{tot}=(h_{t-1}^{A}+h_{t-1}^{B})/2\\ i_t=\tan h(W_{xi}x_t+W_{hi}h_{t-1}^{tot}+b_i)\\ j_t=\text{sigm}(W_{xj}x_t+W_{hj}h_{t-1}^{tot}+b_j)\\ f_t=\text{sigm}(W_{xf}x_t+W_{hf}h_{t-1}^{tot}+b_f)\\ o_t=\tan h(W_{oi}x_t+W_{ho}h_{t-1}^{tot}+b_o)\\ c_t=c_{t-1}\odot f_t+i_t\odot j_t\\ h_t^{A}=\tan h(c_t)\odot j_t\end{cases}\tag{2}$$

$$\text{Bi\_LSTM}_B\begin{cases}h_{t-1}^{tot}=(h_{t-1}^{A}+h_{t-1}^{B})/2\\ i_t=\tanh(W_{xi}x_t+W_{hi}h_{t-1}^{tot}+b_i)\\ j_t=\text{sigm}(W_{xj}x_t+W_{hj}h_{t-1}^{tot}+b_j)\\ f_t=\text{sigm}(W_{xf}x_t+W_{hf}h_{t-1}^{tot}+b_f)\\ o_t=\tanh(W_{oi}x_t+W_{ho}h_{t-1}^{tot}+b_o)\\ c_t=c_{t-1}\odot f_t+i_t\odot j_t\\ h_t^{B}=\tanh(c_t)\odot j_t\end{cases}\tag{3}$$

原理是在更新或者传递信息前进行一次隐状态权值共享，然后再各自回到自身的双向单模态 LSTM 中进行下一步操作，直到需要下一趟更新各自隐状态，双向三(全)模态的 LSTM 原理也一样：

$$\text{Bi\_LSTM}_{\text{Full}}\begin{cases}h_{t-1}^{tot}=(h_{t-1}^{A}+h_{t-1}^{B}+h_{t-1}^{B})/3\\ \text{respectively upgrade}\\ h_t^{A}=\tanh(c_t^{A})\odot j_t^{A}\\ h_t^{B}=\tanh(c_t^{B})\odot j_t^{B}\\ h_t^{C}=\tanh(c_t^{C})\odot j_t^{C}\end{cases}\tag{4}$$

值得注意的是，这里都是双向的 LSTM，逆向的隐层状态也是需要同样的共享机制，并且更新方式完全与正向一样。

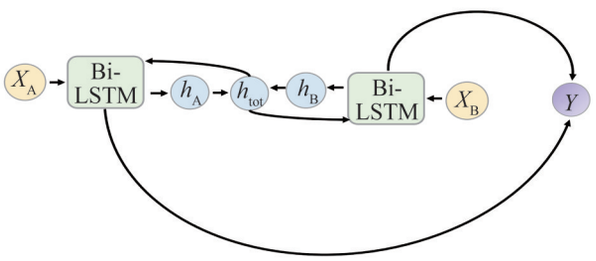


图 3 两两组合的双向双模态 LSTM 结构示意图
Fig. 3 Schematic diagram of bidirectional bimodal LSTM structure in pair-pair-combination

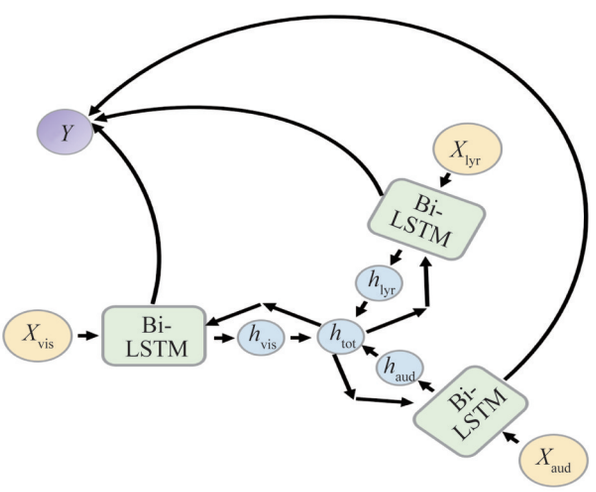


图 4 双向三(全)模态的 LSTM 结构示意图
Fig. 4 Schematic diagram of LSTM structure with bidirectional tri-modal (full)

本文选取双向全模态 LSTM 网络进行下一步 LMFDB-LSTM (long memory combined with bidirectional full-mode LSTM) 的构建。如图 5 所示，LMFDB-LSTM 需要 2 个元件，一个是双向全模态 LSTM 网络，一个是长期记忆体。本文把长记忆体作为一种影响因子去作用到双向全模态 LSTM 网络的情感分类输出 Y，从而得到 LMFDB-LSTM 的输出。这种影响因子是多模态各自信息熵的均值函数：

$$factorLM=\sigma\left[\frac{1}{3}(E(MSU_{\text{simular\_frames}})+E(MSU_{\text{simular\_Music}})+E(MSU_{\text{simular\_Lyrics}}))\right]\tag{5}$$

映射采用sigmoid函数。作用到Y上的影响：

$$Y^{`}=Y\pm\Omega\cdot factorLM,\ \Omega\in[0,0.1] \tag{6}$$

式中：Ω用来控制影响的程度；±取决于该分类Y的正负性，若是正性情绪则用减法，反之用加法，目的越混乱越让情绪趋近于中性0。

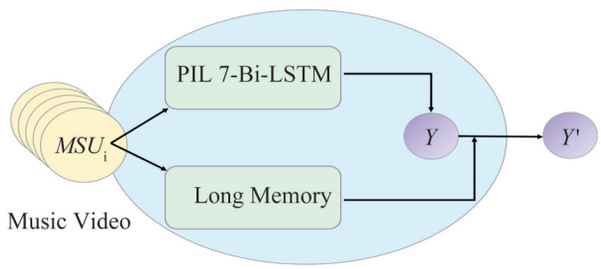


图5 LMFDB-LSTM结构示意图
Fig. 5 LMFDB-LSTM schematic diagram

在图6中，每给定一个MSU，长期记忆体都会通过相似性评价方法，找到与该MSU类似的图像、音频与歌词内容。假设这些相似性的多模态信息是记忆体里面早已存储的数据，而且已经得到情感标注。假设记忆体容量大小为n，分配到各个模态上的容量是$n/3$，对各个模态上的标注进行自信息熵后得：

$$E(MSU_{\text{simular}})=\sum_{i=1}^{3}-\frac{n_i}{n/3}\left(\text{lb}\frac{n_i}{n/3}\right) \tag{7}$$

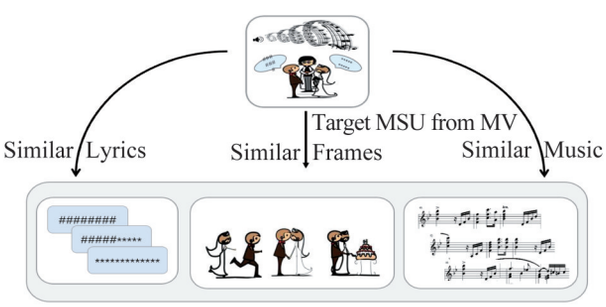


图6 长期记忆体结构与运作机制示意图
Fig.6 Schematic diagram of long-term memory structure and operating mechanism

1.2 情感的集成化决策

在情感的集成化决策中，本文应用了集成决策的思想。首先构建了多模态的决策拓扑，该拓扑中包含了7个分类器，实现对刺激的综合决策。引入多模态综合置信决策的概念，在决策中加入模糊的思想，以便让本文提出的情感理解模型能够更好地体现人的真实的、带有不确定性质的情感产生机制。结合综合置信决策，设计了情感管道，情感管道的宽窄取决于置信决策中的置信区间，情感管道能够让确定性的刺激物输入，转化为一种不确定性的情感输入，为后续的情感表达奠定个性化基础。

1.2.1 模糊化情感集成决策

本文所用的数据集，包含3个模态，分别是A，V和L。根据实验心理学的相关结论，人对客观事物的敏感程度不同，不同的人对图像、声音和文字各具有敏感偏好，而且还存在一个组合模式的敏感偏好。不考虑这种差异，很难构建面向个性化的机器情感计算模型。本文从A，V和L出发，根据上文所提到的双向三(全)模态的LSTM、双向双模态的LSTM和双向单模态的LSTM，把它们组织起来可形成7个不同的分类器，分别是AVL，AV，AL，VL，V，L，A，分别代表着视听觉语义综合分类器、视听觉综合分类器、听觉语义综合分类器、视觉语义综合分类器、视觉单分类器、听觉语义单分类器和听觉语义单分类器。

7种分类器可以得到7种输出，记为$\beta_i=$[A, L, V, VL, AL, AV, AVL]，具体如图7所示。依据这7种输出，再构建基于多模态的模糊化情感集成决策。模糊化情感集成决策的目的就是从个性化的不确定出发，来综合化评判输入数据可能带来的情感不确定性以及不确定性的程度。根据程度，刻画出情感管道。管道中的通道大小就直观的刻画输入数据可能引发的不确定性大小。

1.2.2 多模态综合置信决策

文献[26]提出了多模态综合置信决策的思路。该思路定义了积极情感区域、中性情感区域、消极情感区域。同时，将消极情感与中性情感的交集、积极情感与中性情感的交集命名为模糊情感区域。借助这种划分手段，可以更好地模拟人的真实的、带有不确定性质的情感状态。

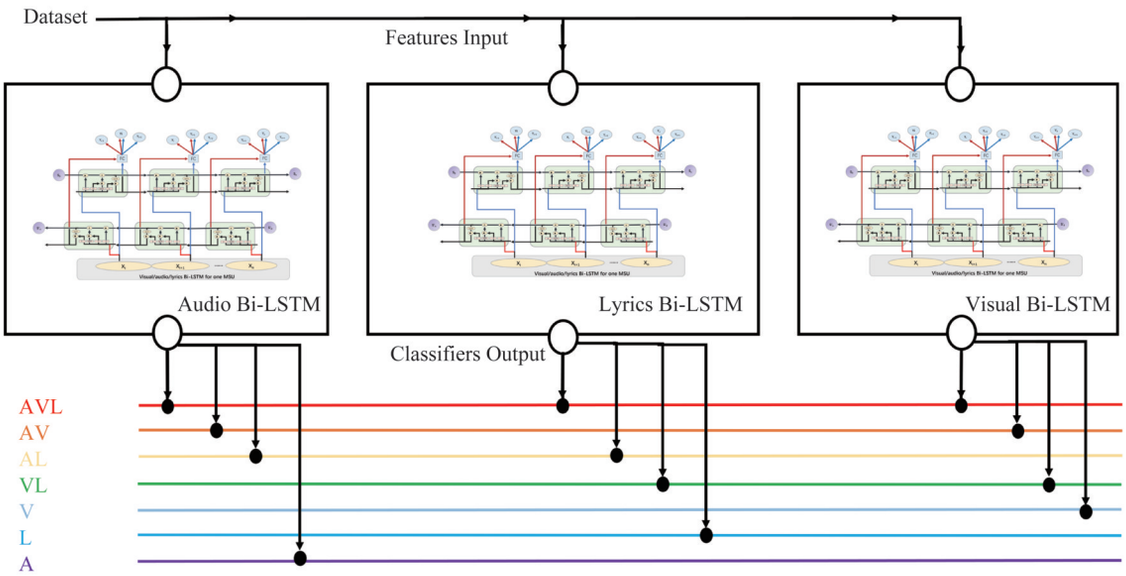


图 7　模糊化情感集成决策框架拓朴图结构

Fig. 7　Fuzzy emotion integration decision framework topology diagram structure

(1) 假设上述的 7 个网络的输出 Out_i 都服从某个均值与方差的正态分布，即：

$$Out_i \sim Gaussian\ distribution(\mu, \sigma^2)$$

(2) 运用式(8)，计算 t 统计量：

$$\frac{Out_{i\bar{X}} - \mu}{Out_{iS}/\sqrt{n}} \sim t(n-1) \tag{8}$$

式中：$Out_{i\bar{X}}$ 与 Out_{iS} 分别为 Out_i 的样本均值与方差；n 为样本数 7。

(3) 对于给定一个超参数 $\phi, 0<\phi<1$，都会满足：

$$P\left(\left|\frac{Out_{i\bar{X}} - \mu}{Out_{iS}/\sqrt{n}}\right| < t_{\phi/2}\right) = 1-\phi \tag{9}$$

式中：$t_{\phi/2}$ 为 $t-distribution(n-1)$ 的接受域面积，最后可得 Out_i 的置信区间：

$$\mu = Out_{i\bar{X}} \pm t_{\phi/2}(n-1)\cdot Out_{iS}/\sqrt{n} \tag{10}$$

(4) 假设一维数轴线段区域为 X，$|X|$ 表示该一维区域 X 的长度，$\max(X)$ 为一维区域的上确边界，$\min(X)$ 为一维区域的下确边界，$P(X)$ 为一维子区域在总区域下的几何概率。综合决策输出阈值函数为 $\theta(x,\varepsilon,z)$，如果 $x>\varepsilon$，就输出 Z 的值，否则输出 0。本文的 ε 超参数为 50%。有用 7 个分类器的 μ 来代表当前 MSU_i 的模糊输出。μ 同时也可以看成一个一维子区域，上确边界为 $\max(\mu)$，下确边界为 $\min(\mu)$。通过 $\max(\mu)$ 来判断情绪基调，通过 $\min(\mu)$ 来判断情感强度。同时满足 $\min(\mu)<0$ 和 $\max(\mu)>0$ 时，最终的情感置信区间决策直接判断为中性否则需要与中性情绪、消极情绪和积极情绪进行占比归属计算，其中：

$$|\text{neutral emotion}| = |\text{positive emotion}| = |\text{negative emotion}| = 1$$

$$\min(\text{neutral emotion}) = -0.5$$

$$\max(\text{neutral emotion}) = 0.5$$

根据上述定义，置信区间决策过程如图 8 所示。首先，确定置信区间两端边界满足的具体条件。然后计算对应于具体边界条件的情绪概率。最后，根据函数 $\theta(x,\varepsilon,z)$ 与计算后的各个情绪概率，确定需要激活哪些情绪(输出 1)。

1.2.3 情感管道

在 ID 的置信区间判定时，$\boldsymbol{E}_{\bar{x}}$ 称为点估计，可用于确定 MSU 的情绪分类输出。每个 MSU 可以得到一个 $\boldsymbol{E}_{\bar{x}}$，然后通过音乐视频的时间索引绘制到一起，就可以构成情感管道(情感管道的上下界就是置信区间的上下确界)，如图 9 所示。情感管

道的较宽部分是情感比较不确定性的位置，其宽度设定的主要因素是置信区间，置信区间越大，不确定越大，管道越宽，反之越窄。为简便起见，把开始和结束都设置为0。从图9中可以发现，情感管道提供了一种更细致、更合理的以视频为刺激物的视频片段情感描述。情感管道将作为下一步情感状态转移模型的模糊刺激物输入到模型中，以实现近似化的、接近真实的刺激物输入。

1.2.4 个性化因子

基于具有有限离散状态的马尔可夫链，构建调节情感状态转移的计算模型。鉴于情感与外界刺激之间的对应关系也是离散的，并且情感状态的转移也是在有限的情感状态空间中，从某一个特定的状态到另一状态的转变，所以该计算模型可以对情感状态转移过程进行描述。借鉴文献[17, 34]，引入个性化因子对情感的产生过程进行个性化的调节。

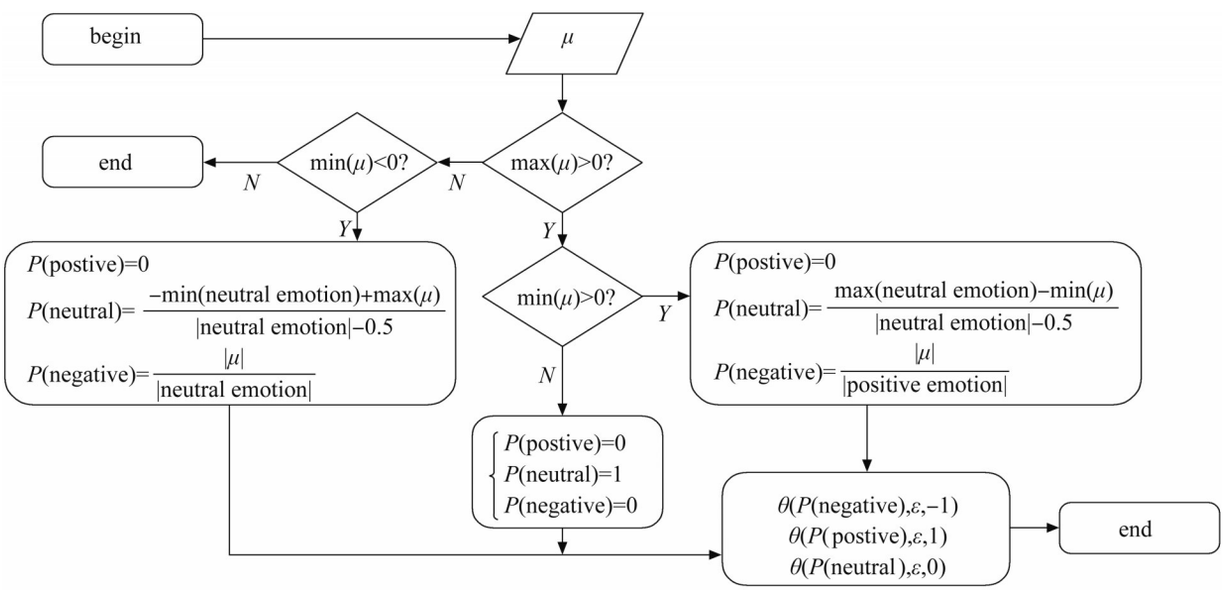


图8 综合置信区间决策框架概述

Fig. 8 Overview of integrated confidence interval decision framework

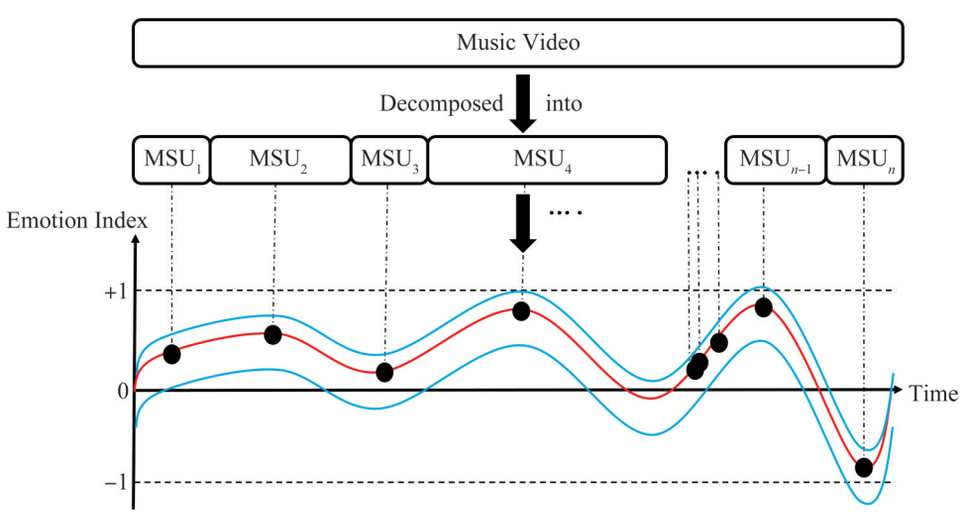


图9 情感管道示意图

Fig. 9 Emotional channel schematic

DOI: 10.16182/j.issn1004731x.joss.20-0899

为了定义个体对同一情绪的响应程度，定义一个个性化因子α，$\alpha\in[-10,10]$。α的值越大，说明该个体对情绪事件的理解越积极，同时，对于自身的消极情绪抑制的越好，甚至将其转换为积极的情绪，同理，α的值越小，说明该个体对情绪事件的理解越消极，越容易触发负面情绪。

以人为参考对象，假设人类的情感状态转移模型是基于具有有限离散状态的非齐次马尔可夫链。将有限情感状态放入集合S(S是离散的，由情感状态转移过程中的所有情感状态组成)。则马尔可夫过程可表示为

$$P(S^{t+1}=u)=\sum_{j\in S}P(S^{t+1}=u|S^{t}=v)\cdot P(S^{t}=v) \tag{11}$$

该情感转移过程的矩阵：

$$S_{n\times 1}^{t+1}=\boldsymbol{M}(\boldsymbol{x})_{n\times n}\times S_{n\times 1}^{t}$$

式中：n为情感状态数，本文取3；$\boldsymbol{M}(\boldsymbol{x})_{n\times n}$为反映个体受到外界刺激后的状态转移矩阵，$\boldsymbol{M}(\boldsymbol{x})_{n\times n}$的组成元素为$m_{uv}$；具体的含义为个体情感状态从$S_u$转移到$S_v$的概率。

将上述提到的个性化因子α与情感转移矩阵相结合，当$\alpha>0$时：

$$m'_{uv}=\begin{cases}M_1\ , & v\in[1,n_1]\\ M_2\ , & v\in[n_1+1,n_1+n_2]\\ M_3\ , & v\in[n_1+n_2+1,n_1+n_2+n_3]\end{cases} \tag{12}$$

其中：

$$M_1=m_{uv}+\frac{1}{2(\mathrm{e}-1)}[(1+\frac{1}{t})^t-1]\frac{m_{uv}}{\sum_{k=1}^{n_1}m_{uv}}\cdot\sum_{k=n_1+1}^{n_1+n_2}m_{uv}$$

$$M_2=m_{uv}+\frac{1}{2(\mathrm{e}-1)}[(1+\frac{1}{t})^t-1]\frac{m_{uv}}{\sum_{k=n_1+1}^{n_1+n_2}m_{uv}}\cdot\sum_{k=n_1+n_2+1}^{n_1+n_2+n_3}m_{uv}-\frac{1}{2(\mathrm{e}-1)}[(1+\frac{1}{t})^t-1]m_{uv}$$

$$M_3=m_{uv}-\frac{1}{2(\mathrm{e}-1)}[(1+\frac{1}{t})^t-1]m_{uv} \tag{13}$$

同理，当$\alpha<0$时，

$$m_{uv}'=\begin{cases}N_1\ , & v\in[1,n_1]\\ N_2\ , & v\in[n_1+1,n_1+n_2]\\ N_3\ , & v\in[n_1+n_2+1,n_1+n_2+n_3]\end{cases} \tag{14}$$

其中：

$$N_1=m_{uv}-\frac{1}{2(\mathrm{e}-1)}[(1+\frac{1}{t})^t-1]m_{uv}$$

$$N_2=m_{uv}+\frac{1}{2(\mathrm{e}-1)}[(1+\frac{1}{t})^t-1]\frac{m_{uv}}{\sum_{k=n_1+1}^{n_1+n_2}m_{uv}}\cdot\sum_{k=1}^{n_1}m_{uv}-\frac{1}{2(\mathrm{e}-1)}[(1+\frac{1}{t})^t-1]m_{uv}$$

$$N_3=m_{uv}+\frac{1}{2(\mathrm{e}-1)}[(1+\frac{1}{t})^t-1]\frac{m_{uv}}{\sum_{k=n_1+n_2+1}^{n_1+n_2+n_3}m_{uv}}\cdot\sum_{k=n_1+1}^{n_1+n_2}m_{uv} \tag{15}$$

至此，构建了一个引入了个性化因子的情感转移状态模型，模型的整体流程为：当前情感状态向量作为输入，作用在情感状态转移矩阵上，得到下一时刻的情感状态向量。

2 实验

2.1 算法有效性验证

为了验证本文所算法的有效性，本文和2种典型的算法进行对比，分别是LSTM[35]和Bi-LSTM[36]。采用2个国际公认数据集IEMOCAP (interactive emotional dyadic motion capture) 和CMU-MOSEI (CMU multimodal opinion sentiment and emotion intensity)。

IEMOCAP[37]是包括6个分类类别：悲伤、中性、兴奋、愤怒、快乐和沮丧。而CMU-MOSEI[38]选择了一组23 400多个句子，1 000人在YouTube上表达出来作为最终样本。样本的情绪状态有3类：积极的、消极的和中性的。在各个通道和各个组合通道上计算各种可能的情况，分别是AVL，AV，AL，VL，V，L，A，运行结果如表1所示。通过运行结果表格发现，我们提出的LMFDB-LSTM算法，在多模态情绪分类上的结果要优于LSTM算法和Bi-LSTM算法。

2.2 情感不确定性验证

在情感的不确定性验证中，做了5组实验，实验结果见表2~6。对比表2~4，可以看出采取视觉

听觉双向双模态时准确率得到了显著的提升，证明了多模态带来的影响是显著与积极的。从表5可以看出，在双向三(全)模态的LSTM中，由于模型的复杂性的提高，导致了分类深度网络出现了过拟合的情况，这也体现了情感分类问题中的矛盾性。根据本文的立意，需要找到个性化带来的不确定性与分类准确率的平衡，而不是一味追求精度，所以本文降低了模型的复杂度，用线性融合的方法构造三单通道(听/视觉/文字)结合的LSTM，实验结果如表6所示。如果考虑到错误判断($0 \rightleftarrows 1$ 或 $0 \rightleftarrows -1$)是正确的，那么三单通道(听/视觉/文字)线性融合的LSTM在测试期间的准确率将达到100%，这意味着组合分类器没有犯什么大错误($1 \rightleftarrows -1$)的同时，具有较为良好的不确定性，能够达到模糊性与准确性的平衡，因此后面的实验将会沿用此模型。

表1 在IEMOCAP和CMU-MOSEI公开数据集上的性能(准确率)对比

Table 1 Comparison of performance (accuracy) on IEMOCAP and CMU-MOSEI public datasets

分类器	IEMOCAP			CMU-MOSEI		
	LSTM	Bi-LSTM	LMFDB-LSTM	LSTM	Bi-LSTM	LMFDB-LSTM
A	0.412 5	0.452 7	**0.474 0**	0.545 8	**0.574 6**	0.573 9
L	0.582 1	0.602 2	**0.618 3**	0.579 5	0. 5 861	**0.601 4**
V	0.321 5	**0.339 7**	0.325 2	0.535 7	0.526 2	**0.558 1**
AL	0.600 5	0.612 4	**0.621 6**	0.585 8	**0.586 6**	0.583 7
AV	0.402 2	0.425 3	**0.428 9**	0.570 2	0.575 4	**0.576 6**
TV	0.603 2	0.614 3	**0.627 6**	0.593 6	0.600 3	**0.604 9**
AVL	0.608 4	0.618 0	**0.627 5**	0.583 1	0.599 7	**0.604 1**

表2 双向听觉单模态LSTM

Table 2 Bidirectional auditory single mode LSTM

情绪	精确率	召回率	F1值
积极	0.887	0.812	0.846
消极	0.690	0.785	0.733
中性	0.793	0.720	0.759

表3 双向视觉单模态LSTM

Table 3 Bidirectional vision single mode LSTM

情绪	精确率	召回率	F1值
积极	0.805	0.534	0.647
消极	0.655	0.913	0.752
中性	0.856	0.841	0.850

表4 视觉听觉双向双模态LSTM

Table 4 Visual and auditory bimodal LSTM

情绪	精确率	召回率	F1值
积极	0.905	0.872	0.887
消极	0.882	0.880	0.880
中性	0.867	0.900	0.882

表5 双向三(全)模态的LSTM

Table 5 Bidirectional three (full) mode LSTM

情绪	精确率	召回率	F1值
积极	0.899	0.910	0.903
消极	0.966	0.873	0.914
中性	0.861	0.930	0.900

表6 三单通道(听/视觉/文字)线性融合的LSTM

Table 6 Linear fusion of three single channels (audio / visual / text) LSTM

情绪	准确率	精确率	召回率	F1值
积极	0.869	0.733	0.924	0.895
消极	0.869	0.835	0.913	0.891
中性	0.869	0.659	0.666	0.754

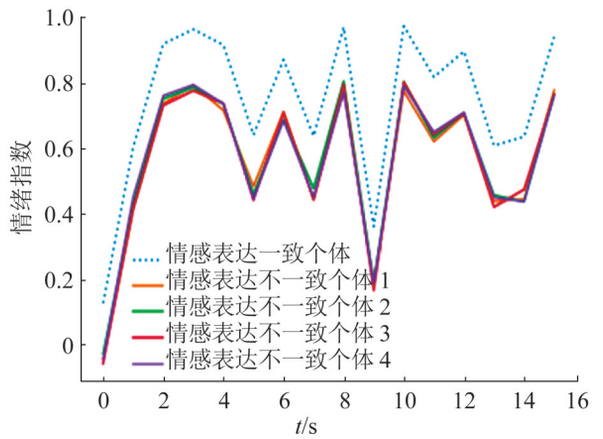
2.3 LMFMB-LSTM 长记忆体对抗仿真

把长记忆体看作是一个对外的接口，对各场景相似度进行匹配，如图10所示。Match person是指长记忆体内部情感分类与刺激物所表达的情感很一致的人，再次将其看成是参考基准。然后随机选择4个对比测试者，他们的长记忆体内部情感分类与刺激物所表达的情感相对不一致，把5个人随时间的情感状态转移仿真结果绘制到统一坐标系后发现：在相同的刺激物下，对此测试者的情感状态转移结果受到了一定程度的抑制，在不同的刺激下，情感状态都有往中性情感区域靠近的趋势。

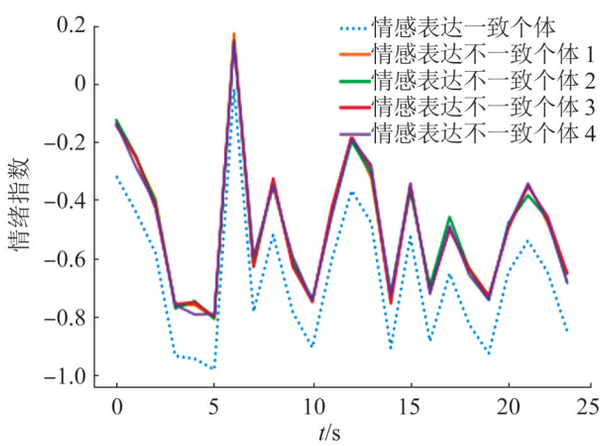
2.4 情感状态转移分析

利用ID分类的真实模糊刺激输入和情感管道进行了一系列仿真测试，其主要目的是为了验证在相同的音乐视频刺激下，探究不同的个性对该模型的影响。图11~12分别是2组对比实验。

DOI: 10.16182/j.issn1004731x.joss.20-0899



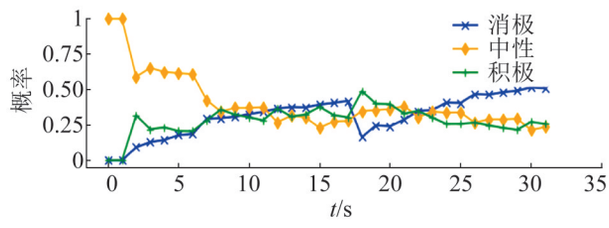
(a) 积极



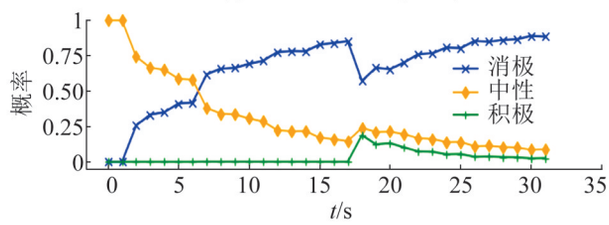
(b) 消极

图 10　两种不同刺激下的 LMFMB-LSTM 长记忆体对抗仿真结果

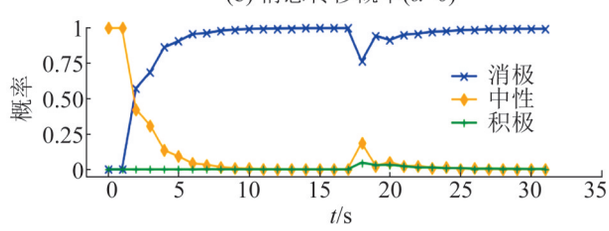
Fig. 10　Simulation results of LMFMB-LSTM long memory confrontation under two different stimuli



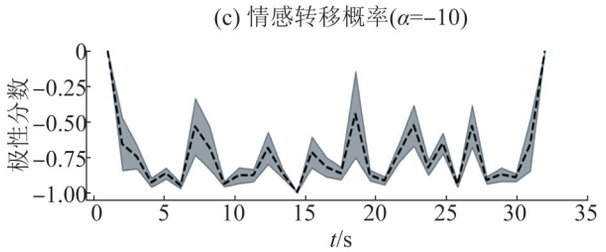
(a) 情感转移概率(α=10)



(b) 情感转移概率(α=0)



(c) 情感转移概率(α=−10)



(d) 原始刺激

图 11　不同个性的情感曲线在同一个悲伤的视频中的表现

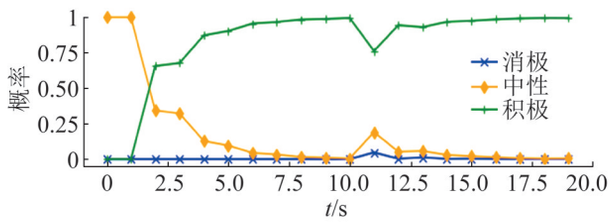
Fig.11　Emotional curves of different personalities in same sad video

实验中，随机选择一个消极的音乐视频和一个积极的音乐视频作为目标，通过 ID 绘制出对应的情感管道。通道的宽度就是不确定性的量化指标，通道宽度越大，说明对应该视频片段的情感表达越模糊。

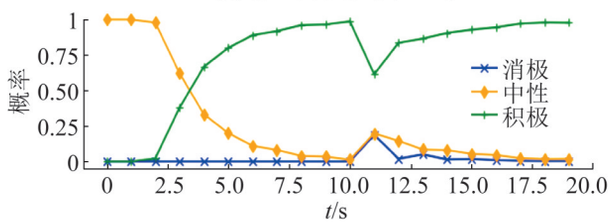
在图 11 中，随机选择了一个消极的音乐视频刺激物，构建了它的情感管道，它不断地将模糊刺激输入到模型中，分别获得了本视频中乐观者、中立者和悲观者的情绪概率状态转移曲线。研究发现，对于一个消极的视频刺激物，积极乐观的人不容易产生消极的情绪。相反，中立的人和消极的人更容易引起他们的消极情感，这基本上与现实情况是相符的。

同样地，随机选择一个积极音乐视频来描述不同性格的人的情绪状态转移曲线，如图 12 所示，也模拟出了合理的结果。在一个积极的视频刺激物下，消极的人不容易产生积极的情感。相反，中立的人和积极的人则更容易引起他们的积极情感。

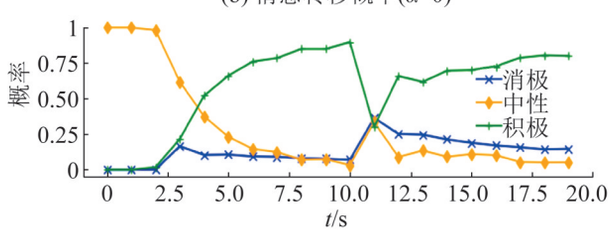
在情感状态转移分析中，通过模拟仿真的方式，重点阐述在情感状态转移中，由于情感的连续性，后者称为情绪的一致性原理[39]，在不同的个性因子调节下，模拟不同个性的人，在受到不同的刺激时所体现的情感变化情况是不同的。这为机器类人情感的表达提供了一条可借鉴的思路。



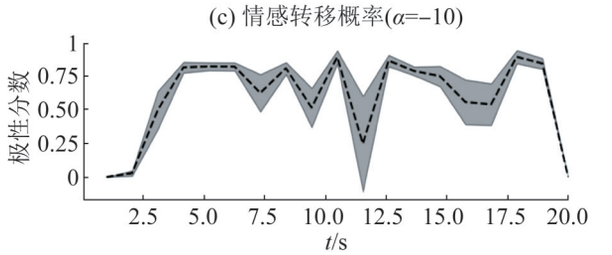
(a) 情感转移概率(α=10)



(b) 情感转移概率(α=0)



(c) 情感转移概率(α=−10)



(d) 原始刺激

图12　不同个性的情感曲线在同一个快乐的视频中的表现

Fig.12　Emotional curves of different personalities in same happy video

3　结论

本文主要提出了一个多模态情感计算模型，并设计了相关实验来验证该模型的有效性。实验仿真结果表明，该模型能更好地反映情感的个性化，特别是在刺激物情感与人格形成强烈对比的时候，对情感个性化的刻画将更为准确。与现有方法相比，本文通过构建LMFDB-LSTM网络来实现音乐视频的个性化情感分类，随后提出了一种多模态的综合情感置信决策来加深对情感的理解，最后设计了一种情绪转移模型来反映个性化的情感表达。不过，受包含硬件在内的各种客观条件所限，本文的情感分类模型以及情感状态表达模型只利用了数据集的低阶特征，而且也没有在一些大规模的视频情感分析数据集上进行实验。

本文对人工情感模拟的构建，仅仅是做了一些建设性的探索，还有诸多工作需要进一步完善。对于情感集成决策方面，它的泛化能力还需要通过大量的数据或更多的新问题来验证。本文目前研究的是一个相对简单的模糊情感区间，若要把ID算法扩展到更高维的V-A空间，以便它能够识别更复杂和更具体的情感类别，还需要大量的研究与论证工作。此外，记忆对情感产生的关系与意义重大，但是本文所采用的记忆机制相对简单，后续需要开展强化记忆机制的研究，以及围绕记忆进行记忆检索机制的研究。

参考文献：

[1] Minsky M. Society of Mind[M]. New York: Simon and Schuster, 1988.

[2] Picard R W. Affective Computing[M]. London: MIT Press, 1997.

[3] Richards J M, Gross J J. Personality and Emotional Memory: How Regulating Emotion Impairs Memory for Emotional Events[J]. Journal of Research in Personality (S0092-6566), 2006, 40(5): 631-651.

[4] Armony J. Computational Models of Emotion[C]//IEEE International Joint Conference on Neural Networks. Piscataway: IEEE, 2005: 1598-1602.

[5] Keltner D, Oatley K, Jenkins J M. Understanding Emotions[M]. London: Wiley Global Education, 2018.

[6] Davidson R J, Abercrombie H, Nitschke J B, et al. Regional Brain Function, Emotion and Disorders of Emotion[J]. Current Opinion in Neurobiology(S0959-4388), 1999, 9 (2): 228-234.

[7] Scherer K R, Bänziger T. Emotional Expression in Prosody: A Review and an Agenda for Future Research [C]//Speech Prosody. Lisbon: International Conference, 2004: 359-366.

[8] Irie G, Satou T, Kojima A, et al. Affective Audio-visual Words and Latent Topic Driving Model for Realizing Movie Affective Scene Classification[J]. IEEE Transactions on Multimedia(S1520-9210), 2010, 12(6): 523-535.

[9] Kang H B. Affective Content Detection Using HMMs

DOI: 10.16182/j.issn1004731x.joss.20-0899

[C]//The 11th ACM International Conference on Multimedia. New York: ACM, 2003: 259-262.

[10] Kang H B. Emotional Event Detection Using Relevance Feedback[C]//International Conference on Image Processing (Cat. No. 03CH37429). Piscataway: IEEE, 2003: I-721.

[11] Zhang S, Huang Q, Jiang S, et al. Affective Visualization and Retrieval for Music Video[J]. IEEE Transactions on Multimedia(S1520-9210), 2010, 12(6): 510-522.

[12] Zhang S, Tian Q, Jiang S, et al. Affective MTV Analysis Based on Arousal and Valence Features[C]// IEEE International Conference on Multimedia and Expo. Piscataway: IEEE, 2008: 1369-1372.

[13] Spaulding S, Breazeal C. Towards Affect-Awareness for Social Robots[C]//AAAI Fall Symposia. Palo Alto: AAAI, 2015: 128-130.

[14] Reisenzein R, Hudlicka E, Dastani M, et al. Computational Modeling of Emotion: Toward Improving the Inter-and Intradisciplinary Exchange[J]. IEEE Transactions on Affective Computing(S1949-3045), 2013, 4(3): 246-266.

[15] Teng S, Wang Z, Wang L, et al. Affective Computing Model Based on Markov Chain[J]. Computer Engineering(S1000-3428), 2005, 31(5): 17-19.

[16] 滕少冬. 应用于个人机器人的人工情感模型研究[D]. 北京: 北京科技大学, 2006.
Teng Shaodong. Research on Artificial Psychology Model Applied in Personal Robot[D]. Beijing: Beijing University of Science and Technology, 2006.

[17] Xiaolan P, Lun X, Xin L, et al. Emotional State Transition Model Based on Stimulus and Personality Characteristics[J]. China Communications(S1637-5447), 2013, 10(6): 146-155.

[18] Yi W, Zhi-Liang W, Wei W. Research on Associative Memory Models of Emotional Robots[J]. Advances in Mechanical Engineering(S1687-8132), 2014, 6(1): 208153.

[19] Masuyama N, Loo C K, Seera M. Personality Affected Robotic Emotional Model with Associative Memory for Human-robot Interaction[J]. Neurocomputing(S0925-2312), 2018, 272(10): 213-225.

[20] Ibanez R V, Keysermann M U, Vargas P A. Emotional Memories in Autonomous Robots[C]// The 23rd IEEE International Symposium on Robot and Human Interactive Communication. Piscataway: IEEE, 2014: 405-410.

[21] Tiedens L Z, Linton S. Judgment Under Emotional Certainty and Uncertainty: the Effects of Specific Emotions on Information Processing[J]. Journal of Personality and Social Psychology(S0022-3514), 2001, 81(6): 973-988.

[22] Eich E, Metcalfe J. Mood Dependent Memory for Internal Versus External Events[J]. Journal of Experimental Psychology: Learning, Memory, and Cognition(S0278-7393), 1989, 15(3): 443-455.

[23] Scherer K R. Towards a Prediction and Data Driven Computational Process Model of Emotion[J]. IEEE Transactions on Affective Computing(S1949-3045), 2019, 12(2): 279-292.

[24] Frijda N H, Bower G H, Hamilton V. Cognitive Perspectives on Emotion and Motivation[M]. Abingdon: Taylor & Francis, 1988.

[25] Jiang D, Wu K, Chen D, et al. A Probability and Integrated Learning Based Classification Algorithm for High-Level Human Emotion Recognition Problems[J]. Measurement(S0263-2241), 2020, 150(1): 107049.

[26] Salway A, Graham M. Extracting Information about Emotions in Films[C]//The 11th ACM International Conference on Multimedia. New York: ACM, 2003: 299-302.

[27] Ekman P, Dalgleish T, Power M. Handbook of Cognition and Emotion[M]. New Jersey: Basic Emotions Chapter, John Wiley & Sons, Ltd, 1999.

[28] Chan C H, Jones G J F. Affect-based Indexing and Retrieval of Films[C]//The 13th Annual ACM International Conference on Multimedia. New York: ACM, 2005: 427-430.

[29] Huang S J, Gao N, Chen S. Multi-Instance Multi-Label Active Learning[C]// IJCAI. San Francisco: Margan Kaufmann, 2017: 1886-1892.

[30] Arifin S, Cheung P Y K. A Computation Method for Video Segmentation Utilizing the Pleasure-Arousal-Dominance Emotional Information[C]//The 15th ACM International Conference on Multimedia. New York: ACM, 2007: 68-77.

[31] Arifin S, Cheung P Y K. Affective Level Video Segmentation by Utilizing the Pleasure-Arousal-Dominance Information[J]. IEEE Transactions on Multimedia(S1520-9210), 2008, 10(7): 1325-1341.

[32] Canini L, Benini S, Migliorati P, et al. Emotional Identity of Movies[C]// The 16th IEEE International Conference on Image Processing (ICIP). Piscataway: IEEE, 2009: 1821-1824.

[33] Lev G, Klein B, Wolf L. In Defense of Word Embedding for Generic Text Representation[C]//International Conference on Applications of Natural Language to Information Systems. Berlin: Springer, 2015: 35-50.

[34] 韩晶, 解仑, 刘欣, 等. 基于Gross认知重评的机器人认知情感交互模型[J]. 东南大学学报(自然科学版), 2015, 45(2): 270-274.

Han Jing, Xie Lun, Liu Xin, et al. Robot Cognitive Emotional Interaction Model Based on Gross Cognitive Reevaluation[J]. Journal of Southeast University (Natural Science), 2015, 45(2): 270-274.

[35] Boleda G, Gulordava K, Aina L. Putting Words in Context: LSTM Language Models and Lexical Ambiguity [C]// The 57th Annual Meeting of the Association for Computational Linguistics. Pennsylvania: ACL, 2019: 3342-3349.

[36] Li H, Xu H. Video-based Sentiment Analysis with hvnLBP-TOP Feature and bi-LSTM[C]//The AAAI Conference on Artificial Intelligence. Palo Alto: AAAI, 2019: 9963-9964.

[37] Busso C, Bulut M, Lee C C, et al. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database[J]. Language Resources and Evaluation(S1574-020X), 2008, 42(4): 335-339.

[38] Zadeh A A B, Liang P P, Poria S, et al. Multimodal Language Analysis in the Wild: Cmu-mosei Dataset and Interpretable Dynamic Fusion Graph[C]//The 56th Annual Meeting of the Association for Computational Linguistics. Pennsylvania: ACL, 2018: 2236-2246.

[39] Masuyama N, Loo C K, Seera M. Personality Affected Robotic Emotional Model with Associative Memory for Human-Robot Interaction[J]. Neurocomputing (S0925-2312), 2018, 272(19): 213-225.

DOI: 10.16182/j.issn1004731x.joss.20-0899