文本驱动的情绪多样化人脸动画生成研究

刘增科，殷继彬

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240100094

文本驱动的情绪多样化人脸动画生成研究

刘增科　殷继彬
昆明理工大学信息工程与自化学院　昆明 650500
(13137748978@163.com)

摘　要　文中介绍了一种新型的文本驱动人脸动画合成技术,该技术通过融合情绪模型以增强面部表情的表现力。这一技术主要由两个核心部分构成:面部情感模拟和唇形与语音的一致性。首先,通过对输入文本的深度分析,识别出其中包含的情感类型及其强度。然后,基于这些情感信息,应用三维自由变形算法(DFFD)来生成相应的面部表情。与此同时,收集人类发音时的语音音素和唇形数据,并利用强制对齐技术,将这些数据与文本中的语音音素在时间上进行精确匹配,从而产生一系列唇部关键点的变化。随后,通过线性插值方法生成中间帧,以进一步细化唇部运动的时间序列。最后,使用 DFFD 算法根据这些时间序列数据合成相应的唇形动画。通过对面部情感和唇形动画进行细致的权重配比,成功实现了高度逼真的虚拟人脸表情动画。该研究不仅解决了文本驱动面部表情合成中的信息缺失问题,而且克服了表情单一和面部表情与唇形不协调的挑战,为人机交互、游戏开发、影视制作等领域提供了一种创新的应用方案。

关键词:文本驱动动画;情绪模型;DFFD;面部动画合成;情绪强度;唇形语音一致性
中图分类号　TP315.69

Text-driven Generation of Emotionally Diverse Facial Animations

LIU Zengke and YIN Jibin
Faculty of Information Engineering and Automation,Kunming University of Science and Technology,Kunming 650500,China

Abstract　This paper presents an innovative text-driven facial animation synthesis technique,which integrates emotion models to enhance the expressiveness of facial expressions. The methodology is composed of two core components:facial emotion simulation and the consistency between lip movements and speech. Initially,a deep analysis of the input text identifies the types of emotions contained and their intensities. Subsequently,these emotional cues are utilized to generate corresponding facial expressions using the three-dimensional free-form deformation algorithm(DFFD). Concurrently,phonemes and lip movement data from human speech are collected. These are then precisely aligned with the phonemes in the text over time using forced alignment technology,resulting in a sequence of changes in lip key points. Following this,intermediate frames are generated through linear interpolation to further refine the timeline of lip movements. Finally,the DFFD algorithm synthesizes the lip animation based on this time series data. By meticulously balancing the weights between facial emotions and lip animations,this approach successfully achieves highly realistic virtual facial expressions.

Keywords　Text-driven animation,Emotion model,DFFD,Facial animation synthesis,Emotion intensity,Lip-Sync consistency

1　引言

人脸动画生成旨在利用图像、语音或文本信息生成高自然的、具有流畅自然脸部动作变化的且唇音同步的人脸动画[1]。人脸动画技术对现代电影行业、游戏行业以及教育行业等都具有重要的影响,创造自然、非机械化的人脸动画仍是一个重要的研究课题。

其中,基于图像驱动的技术虽然能生成真实的面部表情动画,但需要投入大量高端设备和人力资源。相比之下,基于文字和语音驱动的技术虽能生成较真实的唇形动画,但由于信息本身的限制,无法提供详细的面部表情信息。

为了解决在面部表情合成中遇到的各种问题,本文提出了文本与特征驱动的复合情绪语音动画研究。其主要分为面部情绪模拟和语音唇型的一致性两部分。

通过对文本的分析获取情绪类型和情绪权重,通过情绪模型[2]和真实面部的关系获取相应的面部特征形变信息。在三维模型上设计特征点,并将情绪模型的特征信息映射到三维模型上。最后通过三维自由变形算法[3-4]对三维模型上的特征点进行驱动,生成相应的情绪表情。

语音唇型的一致性实现。将文本转换为相应的语音音素,并采集音素唇型变化信息,将唇型变化信息与文本语音音素在时间上进行匹配,以此生成唇型变化的时间序列。其次,基于唇型变化与三维模型唇部特征点的映射关系,构建嘴部关键点的时间序列变化。由于该时间序列仅包含关键帧信息,采用线性插值法生成三维模型唇部的过渡帧,从而完善整个唇型变化的时间序列。最终,根据时间序列数据来生成嘴部动画。

将面部情绪与嘴部动画进行技术性拟合,通过分析情绪

通信作者:殷继彬(41868028@qq.com)

分数和音素生成音频时间序列，从而得到完整的根据文本生成的虚拟人面部表情动画，如图 1 所示。

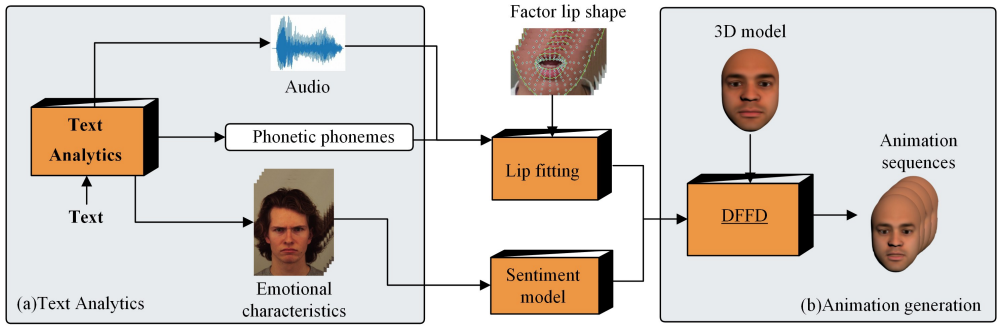


图 1 分析情绪分数和音素生成音频时间序列及其动画模拟

Fig. 1 Generating audio timeline sequence through the analysis of emotion scores and phonemes and their nimation simulation

2 相关工作

2.1 图像驱动的人脸动画生成

目前主流方法是图像驱动[5]的人脸动画方法，该方法根据用户的表演图像来驱动三维人脸模型[6]。因此，人脸运动捕捉的准确性以及实时性起着至关重要的作用。人脸运动捕捉系统主要分为：基于标志点的运动捕捉系统[7]、三维扫描系统[8]以及单摄像头系统[9]。

在基于标志点的运动捕捉系统中，需要在表演者脸部以及身体贴上标注点，实现数据的实时同步。该方法通常用于电影制作，它可以栩栩如生地还原表演者的动作和表现力。但是，该方法采集设备价格昂贵，同时需要专业人员进行专业操作，普通用户很难实施操作。

在基于三维扫描系统的方法中，需要使用三维扫描仪，如结构光扫描系统[10]、摄像机阵列[11]来捕获人脸细节动作。三维扫描仪可以获得高精度深度图，但设备价格高，数据处理时间长。此外，微软公司推出了 Kinect 相机以跟踪人脸动作，该设备也可以获取深度图但包含较大噪声，可在精度和实用性之间实现较好的折中。

由于基于单摄像头的人脸动画系统价格最低、实用性强，因此具有广阔的应用前景。在此方法中，精确且鲁棒的人脸跟踪技术为该系统的关键所在，但是单摄像头只能获得人脸的彩色图像，表演者的脸部运动只能通过跟踪视频中的人脸特征点来获取。由于受光照、姿态、表情以及遮挡等影响，人脸跟踪任务仍具有很大的挑战。

2.2 语音驱动的人脸动画生成

语音驱动的人脸动画指在给定语音输入的情况下，合成相应的唇部运动[12-13]或是发音器官运动[14-16]。考虑到语音的产生机理，语音是通过发音器官的运动而产生的。因此，通过将语音与发音器官运动进行映射，并利用这些发音器官的运动来驱动三维模型，最终可以实现人脸动画的合成。与纯粹的语音交互相比，语音驱动的人脸动画方法在人机交互、虚拟现实等领域中能够显著提高场景的真实感、增强用户的注意力，并在嘈杂的环境中提高理解性。

2.3 文本驱动

文本驱动方法包括基于视素的方法[17-18]和基于样本的方法[19]。在基于视素的方法中，视素代表了与音素对应的基本嘴部形态。这种方法需要为每个音素建立相应的视素。然而，为了确保人脸动画具有平滑的面部运动和连续的时间序列，需要通过协同发音规则来生成音素之间的过渡嘴部形态，以实现人脸动画的合成。

3 技术和方法

3.1 情绪模型

为了在三维虚拟人上实现真实的面部表情合成，需要了解情绪模型与虚拟人面部五官之间的相关性。在情绪模型的研究历程中，出现了一些著名的模型，其中包括：Pad 情感三维模型，该模型认为情绪由愉悦度、激活度和优势度 3 个维度组成[20-22]；Frijda 模型，该模型将情绪视为愉悦/不愉悦、兴奋、兴趣、社会评价、惊奇和简易/复杂的混合体[23]；Izard 模型，该模型将情绪分为愉悦维度、紧张维度、冲动维度和确信维度[24]。这些情绪模型各自在不同领域有着不同的作用。

而本文采用的五维情绪模型认为情绪由强度、愉悦度、掌控度、确定度和紧张度 5 个维度组成。图 2 给出 5 种基本情绪与五维数值关系。

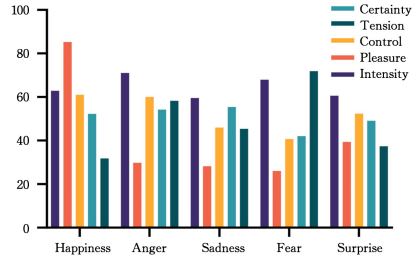


图 2 5 种基本情绪与五维数值关系

Fig. 2 Relationship between five basic emotions and five-dimensional numerical value

Ekman 等通过对面部表情进行研究，制定了“面部肌肉活动编码系统”(FACS)[25]。该系统不以肌肉作为研究单位，而是以面容活动为单位，即活动单位(AU)。然而，在常规情况下，人类大脑可以通过简短的观察轻松判断出表情所表达的情绪，无需考虑如此多的活动单位。因此，为简化五维情绪的活动单位表，将原来的 24 个活动单位合并为以下 5 个简化活动单元(SimpleAUs)：眉心的上下移动、眉心的凑近程度、眼睛的闭合、嘴角的弯曲程度以及嘴的闭合。表 1 列出了这五维情绪对应的简化活动单元系数。图 3 展示了与五维情绪模型相关的关键点。

表 1　五维对应简化活动单元系数

Table 1　Simplified activity unit coefficients corresponding to five dimensions of emotion

	嘴的闭合	嘴角的弯曲程度	眼睛的闭合	眉心的上下移动	眉心凑近程度
强度	1.00	0.395	0.000	0.56	0.218
愉悦度	0.41	1.000	0.207	0.00	0.390
掌控度	0.36	1.000	0.02	0	0.647
确定度	0.000	0.759	0.14	0.610	1.000
紧张度	1.000	0.000	0.00	0.821	0.605

图 3　五维情绪模型关键点

Fig. 3　Key points of five-dimensional emotion model

3.2　Dirichlet 自由变形算法的构建

五维情绪模型在定义简化活动单元(AU)时，为了方便计算，在嘴、眼睛、眉毛上只定义了主要区域的关键点。然而，人脸在做出表情时需要更加自然和准确的变化，因此需要对五维情绪模型的关键点数量进行一定的调整。图 4(a)展示了改进后的简化活动单元关键点。

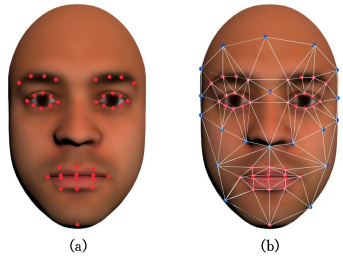
为了实现模型的自然形变，本文采用自由变形算法(FFD)[26-27]，这是计算机动画和几何建模中广泛使用的技术。与基于对象的变形方法不同，FFD 通过对对象所在的空间进行变形来定位对象，而不是直接改变对象本身。本文使用了 DFFD(Distributed Free-Form Deformation)算法，它是 FFD 的扩展，具有更高的灵活性。DFFD 算法允许任意设置控制点，而且不需要明确定义控制箱，因此增强了变形的灵活性。

DFFD 的思想是：给定控制点集合 P 和其凸包内的一点 p，可以确定 p 的 Sibson 邻居集合 $P=\{p_i \mid 0\leqslant i\leqslant n\}$，并计算出 p 相对于 Sibson 邻居的 Sibson 局部坐标 u，其中($0\leqslant i\leqslant n$)。当移动 P 中一个或多个控制点之后，假设控制点的新位置 $P_i'=p_i+\Delta p_i(0\leqslant i\leqslant n)$，则控制点 P_i 的位移 ΔP 可以由下面的公式确定：

$$\Delta P=\sum_{i}^{n}u_i\Delta P_i \tag{1}$$

$$P'=P+\Delta P \tag{2}$$

根据 DFFD 的原理，对改进后的简化活动单元关键点添加了辅助控制点，并对其进行三角划分生成 Voronoi 图，如图 4(b)所示。根据特征关键点和辅助关键点将面部划分为两个凸包，分别是眼眉区域和嘴部区域。



(a)　(b)

图 4　改进后的简化活动单元关键点和三角划分图

Fig. 4　Enhanced simplified action unit keypoints and triangulation diagram

3.3　文本情绪分析及模拟

对应的简化活动单元系数的计算方法是通过对文本的情绪进行分析，获得情绪类型，然后将五维情绪的 5 种数值相乘，最后再乘以系数以获得活动简化单元的具体数值。根据五维情绪模型简化活动单元系数表，可以计算出改进后的关键点的位置信息。

3.3.1　唇部特征关键点移动信息计算

当嘴部张开程度为 C，中性脸上嘴唇顶部到下巴的高度为 H 时，可以计算嘴的张开距离为：

$$MOD=C\times H \tag{3}$$

接着，利用嘴角的弯曲程度值 V 和嘴的张开距离 MOD，可以计算出嘴唇张开的距离为：

$$LLOD=V\times MOD \tag{4}$$

上嘴唇低的移动距离(y)等于上嘴的张开距离 MOD 减去下嘴唇张开的距离 $LLOD$：

$$ULLMD(y)=MOD-LLOD \tag{5}$$

上嘴唇顶的移动距离(y)由上嘴唇低的移动距离(y)与相对中性脸坐标的差值得出：

$$ULTMD(y)=ULLMD(y)-|ULTNF(y)-ULLNF(y)| \tag{6}$$

上嘴唇左低的移动距离(y)为上嘴唇低的移动距离(y)的 2/3：

$$ULLLMD(y)=2/3\times ULLDM(y) \tag{7}$$

上嘴唇左顶的移动距离(y)由上嘴唇左低的移动距离(y)和相对中性脸坐标的差值求得：

$$ULLTMD(y)=ULLLMD(y)+|ULLTN(y)-ULLLN(y)| \tag{8}$$

3.3.2　唇部特征关键点移动信息计算

当眼部闭合系数为 C，中性脸眼睛的宽度为 W 时，可以计算眼睛的张开距离为：

$$EOD=C\times W \tag{9}$$

上眼睑低的移动距离(y)等于眼睛的张开距离 D 的 2/9：

$$ELMD(y)=EOD\times 2/9 \tag{10}$$

上眼睑顶的移动距离(y)等于眼睛的张开距离 D 的 7/9：

$$ETMD(y)=EOD\times 7/9 \tag{11}$$

上眼睑左低的移动距离(y)为上眼睑低的移动距离(y)的 2/3：

$$ELLDM(y)=ETMD(y)\times 2/3 \tag{12}$$

上眼睑左顶的移动距离(y)为上眼睑顶的移动距离(y)的 2/3：

$$ELTDM(y)=ELMD(y)\times 2/3 \tag{13}$$

3.3.3　眉部特征关键点位移信息计算

已知眉心的上下移动系数 $C1$，眉心的凑近程度系数 $C2$，中性脸眉毛的宽度 H，可以求得眉心至眼睛眼角的距离：

$$ECTECD=C1\times H \tag{14}$$

还可以求得眉内角的移动距离：

$$EICPMD=C2\times H \tag{15}$$

通过以上步骤就得到了五维情绪模型与面部关键点之间的对应关系。其中 7/9 与 2/9 是经过实验得到的上下眼帘的变化比例。

接下来，通过对文本进行 BRET 情感分析，得到输入文本的基本情绪和情绪强度。通过五维情绪模型的计算，可以

得到简化情绪单元的系数。再计算出简化情绪单元的关键点移动信息。最后借助三维自由变形算法对原始中性脸进行变形，生成与文本相对应的三维情感表情。图 5 给出了基本情绪模拟结果。

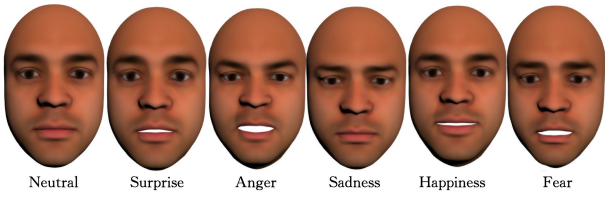


图 5 基本情绪模拟结果

Fig. 5 Simulation results of basic emotions

然而，在文字表达过程中，情绪值通常不是单一情绪，而是多种情绪的混合，只是情绪的权重不同。因此，本文以基本情绪为基础，将情绪进行权重混合，以使虚拟人物的表情更加丰富和真实。例如，对于句子“今天的一天真是美好啊，我度过了一个充实而有趣的周末”，情感分析得到的情感比例如下：喜：0.24、怒：0.10、哀：0.20、惧：0.15、惊：0.46。而对于句子“这个服务真是太差了！这种态度简直让人无法忍受！”，情感比例是：喜：0.08、怒：0.10、哀：0.28、惧：0.15、惊：0.51。这些情感比例的总和为 1。通过将这些比例与各种情感的数值相乘，得到了五维情绪模型的系数，然后将这些系数映射到简化情感活动单元上，以获得相应的值。接下来，计算简化情感活动单元上关键特征点的位移信息，从而模拟出相应的表情。如图 6 所示。

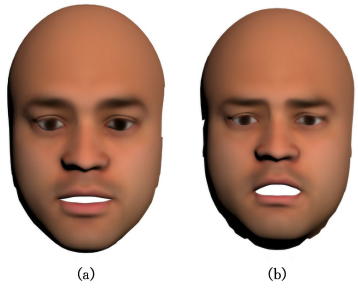


图 6 情绪权重比例表情

Fig. 6 Emotion weight ratio expression

3.4 文字驱动的唇形动画实现

文字驱动的表情动画生成具有一个非常重要的特点，即确保文字与唇型保持一致。为了生成真实的动画效果，必须确保三维模型能够准确模拟文本中的唇型。本文采用了以下步骤来实现这一目标：首先，将文本分解为音素。再将音素与已收集的音素唇型数据进行比对和拟合，使用强制对齐技术进行时间上的同步，以获取文本的唇型变化时间序列帧。最后，驱动三维模型关键点的变化，实现文本唇型动画。

3.4.1 文本处理

由于汉字的数量庞大，但是汉字的发音相对较少，因此将汉字拆分成语音单元来处理会更加方便。这些语音单元通常被称为音素。

3.4.2 中文音素唇型信息采集

与拼音侧重于汉字的发音不同，音素应该更偏向反映发音过程中口腔的运动，从而才能够将汉字转化为能被算法利用的额外信息。图 7 所示为音素唇型信息采集。

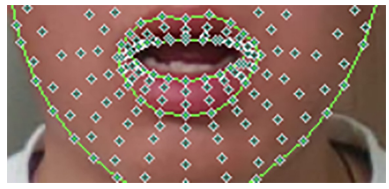


图 7 音素唇型信息采集

Fig. 7 Collection of information on lip shapes associated with phonemes

将收集的信息进行保存，如表 2 所列为音节唇形信息。时间点分别表示音节开始的时间点（Front）和结束的时间点（Rear），唇形则是唇的开合程度和唇的弯曲程度，最终获得的结果就是唇部的各种变化数值。

表 2 音素唇形信息

Table 2 Syllable lip shape information

音素	时间点	唇形	数值
a	front	mouth_open	0.55
		mouth_rad	0.65
	rear	mouth_open	0.44
		mouth_rad	0.43
b	front	mouth_open	0.12
		mouth_rad	0.21
	rear	mouth_open	0.20
		mouth_rad	0.38
c	front	mouth_open	0.24
		mouth_rad	0.36
	rear	mouth_open	0.19
		mouth_rad	0.42

不同的音素在发音时具有一定的相似性，例如 un 和 ün、u 和 ü。因此需要对音素唇形进行筛选。接下来，可以将文本转换为拼音表示，然后进一步转换为音素表示。

例如：“晚上好”——“wan3shang4hao3”。

根据文本生成音频信号，并对其进行采样，以获取各个音节在音频中的时间点。

3.4.3 生成唇形序列帧

将音频信号与音节唇形信息进行融合，形成连续的唇形。图 8 所示为唇形音素融合。

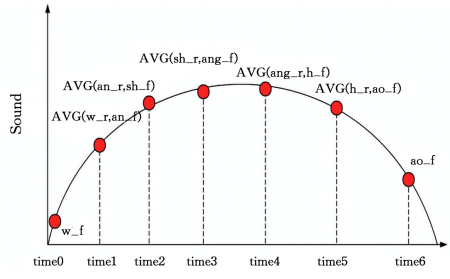


图 8 唇形音素拟合

Fig. 8 Lip shape-phoneme fitting

通过前文所述的关键点坐标计算方法，获得三维模型唇部关键点的变化时间帧数据。因为采集和计算的关键点都是关键帧所在的时间点的数据，所以需要通过线性插值对其进行关键帧补齐，公式如下：

$$result = start + (end - start) \times t \tag{16}$$

其中，$start$ 为起始坐标，end 为结束坐标，t 为插值因子。在这里，插值因子 t 是通过起始坐标与结束坐标的时间差除以帧数来计算的。

通过三维自由变形算法对补齐后的关键帧序列进行计算渲染，输出虚拟人唇形动画。如图 9 所示，对比模拟唇形与真实唇形。

图 9　模拟唇形与真实唇形对比

Fig. 9　Comparison of simulated and actual lip shapes

4　结果评估与总结

综上所述，通过将根据文本获取的表情动画和唇形动画两者融合，得到最终的动画渲染结果。在面部表情和唇形融合过程中，假设面部表情对应模型上的顶点为 S_0，而唇形对应的模型顶点为 S_i；假定唇形模型有一个权重值 W_i，其取值范围为(0,1)，这个值与发音振幅相关。最终的表情模型由这些元素融合而成，变形后的表情顶点标记为 S。描述方程如下：

$$S=S_0+W_i(S_i-S_0) \tag{17}$$

如此，即完成了唇形与面部表情的融合，作为此时刻面部动画的最终结果。如图 10 所示，展示的是“今天的一天真是美好啊，我度过了一个充实而有趣的周末”这句话的动画抽帧。

图 10　文字驱动的人脸动画

Fig. 10　Text-driven facial animation

4.1　评估

在评估生成的情绪模型时，考量其真实性和自然性是非常关键的。为了全面评估本文方法的有效性，研究采用了两种主要的评价方式：主观评分和客观分析。在主观评分中，参与者根据自己的感受和体验来评价模型生成的情绪表达的真实感和自然度；而在客观分析方面，利用具体的评估标准和量化指标来衡量情绪模型的性能。这种双重评估方法不仅提供了从不同视角对模型效果的深入理解，而且有助于识别和改进方法的潜在不足之处。

4.1.1　基本面部情绪结果评估

在主观评估方面，邀请了 136 名不同年龄的男性和女性实验参与者对方法生成的不同三维模型基本情绪集合进行打分。这些参与者的任务是对实验方法生成的各种三维模型所表现的基本情绪集合进行评分。参与者需要根据他们的主观感觉来判断所展示的情绪表现，并依据其真实性给出评分。实验设定的评分标准包括：1 分代表“非常不真实”，2 分表示“不真实”，3 分为“不确定”，4 分意味着“真实”，而 5 分则是“非常真实”。实验结果汇总如表 3 所列。这样的评分体系和结果汇总，不仅反映了每种情绪表现的受众接受度，而且提供了关于模型性能的重要反馈。

表 3　主观基础情绪打分评估

Table 3　Score assessment of subjective basic emotions

emotion	5	4	3	2	1	Σ	‘real’
Happiness	45.30%	36.3%	7.90%	10.00%	0.50%	4.10	81.6%
Anger	45.70%	39.8%	8.7%	5.40%	0.40%	4.30	85.6%
Sadness	29.30%	32.8%	9.40%	22.90%	5.70%	3.90	62.1%
fear	31.90%	25.2%	10.90%	23.90%	8.20%	3.50	57.1%
Surprise	47.00%	31.9%	9.70%	10.10%	1.40%	4.10	78.9%
AVG	39.84%	33.2%	9.32%	14.46%	3.24%	3.98	73.1%

根据表中的结果可以看出，参与者对本研究生成的基本表情的认可度相对较高，总体真实度为 73.1%。其中，悲伤和恐惧两种情绪评分较低，可能是因为在情绪模型中这两者分类较为相似，使得在转化为情绪参数时难以区分。

在客观评估部分，采用了与主观评估(基于人类评分)不同的方法。这里，客观评估主要是应用科学领域中较为先进的情绪识别算法来评价基本情绪的准确性。例如，Zadeh 等[28]提出的一种结合 Gabor 滤波器和深度学习中的 CNN(卷积神经网络)的情绪识别方法；同时也考虑了 Jiang 等[29]提出的利用 Gabor 卷积网络(GCN)进行面部表情识别的技术；此外，还有 Akhand 等[30]研究的一种方法，它结合了深度卷积神经网络和迁移学习来识别人类的面部情绪。这些不同的方法被用来客观评估实验生成的情绪模型，评估结果如表 4 所列。通过这种方法，能够从一个科学和技术的角度来验证实验模型的准确性和有效性，为进一步研究提供了更加全面和深入的理解。

表 4　客观基础情绪打分评估

Table 4　Score assessment of objective basic emotions

(%)

method	Happiness	Anger	Sadness	Fear	Surprise	result
Zadeh 等	80.4	81.2	75.3	77.6	79.5	78.8
Jiang 等	83.4	85.7	78.3	76.5	81.9	81.2
Akhand 等	85.7	84.3	77.9	78.2	83.5	81.9
AVG	83.2	83.7	77.2	77.4	81.6	84.9

图片显示，快乐和愤怒情绪的表情模型在 3 种评估方法中都显示出较高的准确性，而悲伤情绪的表情模型在各种评估方法下的识别准确性相对较低。值得注意的是，这些模型在科学评估方法下的准确率都超过了 75%。这一成绩证明了实验结果整体上达到了较高的水平。总体而言，本文方法在科学和技术层面上是有效的。

4.1.2　复合情绪结果评估

接下来，对复合情绪进行了评估和分析。复合情绪指由多种基本情绪混合而成的情绪。在本研究中，图 6 和图 10

展示了实验生成的复合情绪模型的结果。研究采用了两种方法来评估这些复合情绪模型。在第一种主观评估中，邀请了参与者根据自己的判断和感受来评价模型生成的复合情绪表达；而在第二种客观评估中，利用了先进的情绪识别算法来评价这些复合情绪的准确性和真实性。这样的双重评估方法不仅提供了从不同角度对复合情绪模型效果的深入理解，而且有助于揭示潜在的改进空间，从而提高模型的整体性能和应用价值。

在本研究中，对图 10 所展示的复合情绪动画数据集进行了详细的主观评估。实验邀请了 136 名不同年龄的参与者参与评价。这次评估的过程与基本情绪的评估略有不同：实验向参与者展示了一条文本及其相应的复合情绪。参与者的任务是评价展示的情绪是否与文本中的情绪一致，最后给出综合评分。这些具体的评估结果如表 5 所列。

表 5 主观复合情绪打分评估

Table 5 Score assessment of subjective composite emotions

Statement	5	4	3	2	1	Σ	'real'
St1	51.4%	38.20%	8.1%	2.10%	0.20%	4.50	89.60%
St2	42.9%	35.20%	7.9%	10.90%	4.20%	3.50	78.10%
St3	45.4%	36.90%	9.2%	7.70%	0.90%	4.20	82.20%
St4	43.5%	38.20%	8.9%	8.60%	0.90%	4.10	81.60%
St5	39.3%	37.80%	9.4%	7.90%	5.70%	3.90	77.10%
Main	44.5%	37.26%	8.7%	9.86%	2.38%	4.04	81.72%

综合来看，每个数据集的评分都超过了 75%，整体的平均分达到了 81.7%，这说明在实验参与者看来，本节实验结果表现出色，认为复合情绪数据集在真实性和一致性方面取得了很好的效果。这种普遍的正面反馈证明了本节方法在创建表达自然的复合情绪方面的有效性。

在研究中，对复合情绪的客观评估采用了与基本情绪评估相似的方法，即依靠科学领域中先进的复合情绪识别算法来判断情绪的准确性。但考虑到复合情绪的复杂性，实验设置了一种特定的评估标准：评估时主要关注动画中呈现的前三种主要情绪。如果所使用的算法能够正确识别出动画中占比最大的 3 种情绪之一，并且识别出的情绪权重顺序与动画中的一致，那么就认为该算法的判断是准确的。主要使用的算法包括：Swaminathan 等[31]提出的 FERCE 算法，用于使用多个数据集进行面部表情识别；Dorota 等[32]提出的一种使用 iCVMEFED 数据集进行面部表情识别的两阶段算法；以及 Heenakausar 等[33]使用 InceptionResNet-v2 架构来识别情绪复合面部表情的算法。这些算法的评估结果如表 6 所列。通过这种方法，能够从一个客观和科学的角度评估复合情绪的准确性和效果。

表 6 客观复合情绪打分评估

Table 6 Score assessment of objective composite emotions

(%)

	St1	St2	St3	St4	St5	result
Swaminathan et al.	73.2	71.7	70.8	73.4	74.9	72.8
Dorota et al.	71.0	72.4	74.9	69.9	72.4	72.1
Heenakausar et al.	70.2	69.4	78.2	71.5	68.9	71.6
main	71.5	71.2	74.6	71.6	72.0	72.2

由表 6 可以看出，Swaminathan 的方法在 St5 上的准确率最高，达到了 74.9%。相较之下，Heenakausar 的方法在 St3 上的准确率最高，为 78.2%。整体而言，这些算法在复合情绪表情模型的准确率上达到了 72.2%，这表明在总体上这些复合表情的准确性是较高的。这些数据提供了不同算法在处理复杂情绪识别方面的性能比较，同时也指出了在复合情绪表情模拟领域的总体成就。

4.1.3 结果对比

在本节实验中，采用了广泛认可的 Likert 五级量表[34]作为评价标准，使用 1～5 分的评分范围，其中 5 分表示最佳效果，1 分表示最差效果。在评估动画视频中虚拟角色模仿真人说话的效果时，主要依据 3 个指标：自然度、流利度和拟人度。基于这些维度，对系统生成的动态表情视频进行了全面评估。

为全面评估本研究方法合成的情绪动画视频效果，邀请了 25 名不同年龄的男性和女性参与者。实验对象是 25 条动画视频，这些视频是根据文本情绪分析生成的，涵盖了 5 种基本情绪：喜悦、愤怒、悲伤、恐惧和惊讶，每种情绪对应 5 条视频。在评估过程中，参与者被要求对每条视频的自然度、流利度和拟人度 3 个方面进行打分。通过对这些评分的汇总和分析，能够从各个角度综合了解视频的表现。

本文所述方法与 Fan 等[35]以及 Zeng 等[36]等研究者提出的虚拟人物动画生成技术进行了打分比较。具体的评分结果如表 7 所列。

表 7 评分结果

Table 7 Scoring results

	自然度	拟人度	流利度	平均
Fan et al.	3.77	3.65	3.57	3.66
Zeng et al.	3.89	3.74	3.63	3.75
本节研究	73.92	3.84	3.54	3.76

结果显示，本节研究在自然度(3.92 分)和拟人度(3.84 分)上均表现出色，超过了 Fan 等和 Zeng 等的方法，显示了其在创造自然和接近真人特征的虚拟人方面的优势。尽管在流利度上稍微落后(3.54 分)，本节研究的整体平均得分为 3.76 分，略高于 Zeng 等的 3.75 分和 Fan 等的 3.66 分。这一结果凸显了本节方法在综合性能上的微弱领先，尤其是在增强虚拟人物的真实感和拟人化方面表现突出，同时也提示了在流畅性方面存在进一步提升的空间。

4.2 总结

通过对文本进行情绪比例分析，应用情绪模型来计算情绪简化活动单元，进而推算出特征关键点的位移信息，最后生成三维虚拟人的情绪动画。基于情绪模型的情绪比例分析使表情模拟更加多样化且真实。

另一方面，对文本进行音素划分，并采集唇形形变信息，结合文本音素时间序列数据和唇形音素形变数据，计算出文本音素唇形的关键帧位移数据，并通过线性插值补全成为时间序列帧。使用 DFFD 算法对每帧数据进行形变计算后，实现了语音与唇形一致的虚拟人语音动画。

将情绪信息和唇语信息相结合，获得完整的音像统一虚拟人动画。本文方法相比传统动作捕捉技术，在资源消耗较少的情况下实现了多样化的人脸动画制作。利用五维情绪模型为基础的表情权重比例合成，本文方法突破了文本驱动在表情限制上的障碍，配合唇音一致的唇形动画，有效解决了语音文本驱动在人脸动画领域的天然信息缺陷，创造出更加真实的人脸动画效果。

结束语 本研究通过五维情绪模型在文本驱动的表情模

拟中实现了更广泛的表情多样性，并且在时间和设备方面相较于其他方法具有显著优势。

本研究的应用潜力广泛，涵盖了游戏领域、影视动画领域以及人机交互领域等多个方面。然而，本研究也存在一定局限性，尽管本文实现了更广泛的表情多样性，但眼神模拟仍有待提升。另外，文本唇形模拟虽然较为真实，但牙齿和舌头由于在唇形采集时的信息遮挡未被充分考虑。同时，人类说话时的下颚运动也是未来研究的重点。

参 考 文 献

[1] YANG D，LI R，YANG Q，et al. 3d head-talk：speech synthesis 3d head movement face animation[J]. Soft Computing，2024，28(1)：363-379.

[2] ZHANG H，YIN J，ZHANG X. The study of a five-dimensional emotional model for facial emotion recognition[J]. Mobile Information Systems，2020，2020(1)：8860608.

[3] ILIC S，FUA P. Using dirichlet free form deformation to fit deformable models to noisy 3-D data[C]// European Conference on Computer Vision(Springer，2002). 2002：704-717.

[4] MUZAHIDIN S，RAKUN E. Text-driven talking head using dynamic viseme and DFFD for SIBI[C]// 2020 7th International Conference on Information Technology，Computer，and Electrical Engineering(ICITACEE 2020). IEEE，2020：173-178.

[5] IGARASHI T，MOSCOVICH T，HUGHES J F. Spatial keyframing for performance-driven animation[J]. ACM SIGGRAPH 2006 Courses，2006：17-es.

[6] MAI H N，KIM J，CHOI Y H，et al. Accuracy of portable face-scanning devices for obtaining three-dimensional face models：a systematic review and meta-analysis[J]. International Journal of Environmental Research and Public Health，2021，18(1)：94.

[7] DENG Z，CHIANG P Y，FOX P，et al. Animating blendshape faces by cross-mapping motion capture data[C]// Proceedings of the 2006 symposium on Interactive 3D Graphics and Games. 2006：43-48.

[8] JAVAID M，HALEEM A，SINGH R P，et al. Industrial perspectives of 3d scanning：features，roles and it's analytical applications[J]. Sensors International，2021(2)：100114.

[9] PELEG S，BEN-EZRA M. Stereo panorama with a single camera[C]// 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition(Cat. No PR00149)，vol. 1. IEEE，1999：395-401.

[10] ZHANG L，SNAVELY N，CURLESS B，et al. Spacetime faces：high resolution capture for modeling and animation[J]. ACM Transactions on Graphics，2004，23(3)：548-558.

[11] FURUKAWA Y，PONCE J. Dense 3d motion capture for human faces[C]// 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2009：1674-1681.

[12] DOUKAS M C，SHARMANSKA V，ZAFEIRIOU S. Video-to-video translation for visual speech synthesis[J]. arXiv：1905.12043，2019.

[13] TAYLOR S，KIM T，YUE Y，et al. A deep learning approach for generalized speech animation[J]. ACM Transactions on Graphics(TOG)，2017，36(4)：1-11.

[14] LING Z H，RICHMOND K，YAMAGISHI J. An analysis of hmm-based prediction of articulatory movements[J]. Speech Communication，2010，52(10)：834-846.

[15] YU L，YU J，LING Q. Bltrcnn-based 3-d articulatory movement prediction：Learning articulatory synchronicity from both text and audio inputs[J]. IEEE Transactions on Multimedia，2018，21(7)：1621-1632.

[16] ZHU P，XIE L，CHEN Y. Articulatory movement prediction using deep bidirectional long short-term memory based recurrent neural networks and word/phone embeddings[J]. Sixteenth Annual Conference of the International Speech Communication Association，2015.

[17] KING S A，PARENT R E. Creating speech-synchronized animation[J]. IEEE Transactions on Visualization and Computer Graphics，2005，11(3)：341-352.

[18] ZHOU Y，XU Z，LANDRETH C，et al. Visemenet：Audio-driven animator-centric speech animation[J]. ACM Transactions on Graphics(TOG)，2018，37(4)：1-10.

[19] LIU K，OSTERMANN J. Realistic facial expression synthesis for an image-based talking head[C]// IEEE International Conference on Multimedia and Expo. IEEE，2011：1-6.

[20] MEHRABIAN A. Framework for a comprehensive description and measurement of emotional states[J]. Genetic，Social，and General Psychology Monographs，1995，121(3)：339-361.

[21] MEHRABIAN A. Pleasure-arousal-dominance：A general framework for describing and measuring individual differences in temperament[J]. Current Psychology，1996(14)：261-292.

[22] MEHRABIAN A，WIHARDJA C，LJUNGGREN E，Emotional correlates of preferences for situation-activity combinations in everyday life[J]. Genetic，Social，and General Psychology Monographs，1997，123(4)：461-478.

[23] FISCHER A H，VAN KLEEF G A. Where have all the people gone? a plea for including social interaction in emotion research[J]. Emotion Review，2010，2(3)：208-211.

[24] IZARD C E，ACKERMAN B P，SCHULTZ D. Independent emotions and consciousness：Self-consciousness and dependent emotions[J]. At play in the fields of consciousness：Essays in honor of Jerome L. Singer，1999，83：102.

[25] EKMAN P，FRIESEN W V. Facial Action Coding System(FACS)：a Technique for the Measurement of Facial Actions[J]. Rivista Di Psichiatria，1978，47(2)：126-138.

[26] KURENKOV A，JI J，GARG A，et al. Deformnet：Free-form deformation network for 3d shape reconstruction from a single image[C]// 2018 IEEE Winter Conference on Applications of Computer Vision(WACV). IEEE，2018：858-866.

[27] SEDERBERG T W，PARRY S R. Free-form deformation of solid geometric models[C]// Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques. 1986：151-160.

[28] ZADEH M，IMANI M，MAJIDI B. Fast facial emotion recognition using convolution alneural networks and gabor filters[C]// 2019 5th Conference on Knowledge Based Engineering and Innovation(KBEI). IEEE，2019：577-581.

[29] JIANG P，WAN B，WANG Q，et al. Fast and efficient facial expression recognition using a gabor convolutional network[J]. IEEE Signal Processing Letters，2020，27：1954-1958.

[30] AKHAND M H A，SHUVENDU R，NAZMUL S，et al. Facial emotionrecognition using transfer learning in the deep cnn[J]. Electronics，2021，10(9)：1036.

[31] SWAMINATHAN A, ADIVEL A V, AROCK M. Ferce: facial expression recognition for combined emotions using ferce algorithm[J]. IETE Journal of Research, 2022), 68(5): 3235-3250.

[32] DOROTA , KADIR A, DAVIT R, et al. Two-stage recognition and beyond for compound facial emotion recognition[J]. Electronics, 2021, 10(22): 2847.

[33] PENDHARI H, NAGDEOTE S, RATHOD S, et al. Compound emotions: a mixed emotion detection[C]// Proceedings of the International Conference on Innovative Computing & Communication(ICICC). 2022.

[34] MACEDONIA M. A bizarre virtual trainer outperforms a human trainer in foreign language word learning[J]. International Journal of Computer Science and Artificial Intelligence, 2014, 4(2): 24-34.

[35] FAN Y, LIN Z, SAITO J, et al. Faceformer: Speech-driven 3d facial animation with transformers[C]// Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 18770-18780.

[36] ZENG J, HE X, LI S, et al. Virtual Face Animation Generation Based on Conditional Generative Adversarial Networks[C]// 2022 International Conference on Image Processing, Computer Vision and Machine Learning(ICICML). IEEE, 2022: 580-583.

LIU Zengke, born in 1998, postgraduate. His main research interests include deep learning and image recognition.

YIN Jibin, born in 1976, Ph.D, associate professor. His main research interests include human-computer interaction and artificial intelligence.