计算机工程与科学
Computer Engineering & Science
ISSN 1007-130X,CN 43-1258/TP

《计算机工程与科学》网络首发论文

题目：基于深度学习和证据理论的表情识别模型
作者：徐其华，孙波
收稿日期：2019-11-04
网络首发日期：2020-09-27
引用格式：徐其华，孙波．基于深度学习和证据理论的表情识别模型[J/OL]．计算机工程与科学. https://kns.cnki.net/kcms/detail/43.1258.TP.20200925.1107.002.html

基于深度学习和证据理论的表情识别模型*

徐其华[1,2]，孙 波[2]

（1：西北师范大学商学院，甘肃 兰州 730070；2：北京师范大学人工智能学院，北京 100875）

摘 要：表情识别是在人脸检测基础之上的更进一步研究，是计算机视觉领域的一个重要研究方向。将研究的目标定位于基于微视频的表情自动识别，研究在大数据环境下，如何使用深度学习技术来辅助和促进表情识别技术的发展。针对表情智能识别过程中存在的一些关键性技术难题，设计了一个全自动表情识别模型。该模型结合深度自编码网络和自注意力机制，构建了一个人脸表情特征自动提取子模型，然后结合证据理论对多特征分类结果进行有效融合。实验证明，该模型能显著提升表情识别的准确率，具有重要的理论意义和研究价值。

关键词：深度学习；表情识别；证据理论；自编码网络；自注意力

中图分类号：TP391.41 文献标志码：A

Expression recognition model based on deep learning and evidence theory

XU Qi-hua[1,2], SUN Bo[2]

(1: School of Business, Northwest Normal University, Lanzhou, 730070, China;

2: School of Artificial Intelligence, Beijing Normal University, Beijing, 100875, China)

Abstract: Facial expression recognition is a further research based on face detection, which is an important research direction in the field of computer vision. The goal of the research is to automatically recognize facial expressions based on micro video, and to study how to use deep learning technology to assist and promote the development of facial expression recognition technology in a big data environment. A fully automated expression recognition model has been designed to address some of the key technical challenges in the expression intelligence recognition process. The model combines a deep auto-encoding network and a self-attention mechanism to construct a sub-model for

* 收稿日期：2019-11-04
基金项目：甘肃省高等学校创新能力提升项目（2019B-043）；国家自然科学基金资助项目（71861031）
通信作者：孙波(tosunbo@bnu.edu.cn)
通信地址：730070 甘肃省兰州市西北师范大学商学院
Address: School of Business, Northwest Normal University, Lanzhou 100875, Gansu, P.R.China

automatic extraction of facial expression features, and then the evidence theory is used to fuse the results of multi-feature classification. Experimental results show that the model can significantly improve the accuracy of expression recognition, which has important theoretical significance and research value.

Key words: deep learning; expression recognition; evidence theory; auto-encoding network; self-attention

1 引言

表情是人类在进行社会活动时心理感受和精神状态的自然流露，通过观察一个人的面部细微变化，就能判断出他此时的内心情感。根据心理学家 Mehrabian[1]的研究，一个人想要表达出来的全部信息，口头语言只占到 7%，语言辅助（如语调、语速等）占到 38%，而面部表情却占到了 55%，因此大量有价值的信息，都可以从面部表情上获取到。而且相对于生理信号，面部表情的数据更加容易获得，因此受到更多人的关注。

随着计算机技术、传感技术以及通讯技术的发展，高清摄像头的使用已经越来越普遍，特别是智能手机的广泛应用，获取一小段带有人脸的高清视频已经是非常容易的事情。通过深度学习技术，对带有人脸的高清视频片段进行自动分析，识别出视频中人类的表情，识别结果不仅能在各种系统中帮助人机进行高效交互，而且能应用在现实生活中的不同领域。

面部表情是指通过眼部肌肉、颜面肌肉和口部肌肉的变化来表现各种情绪状态，是体现人类内心情感比较直接的一种表达方式。根据科学家们的研究，人类有 7 种基本情感，即快乐、悲伤、愤怒、厌恶、惊讶、恐惧和中性。表情识别的研究，实际上可以认为是在这七类情感上的模式分类问题。随着人工智能的发展以及实际应用需求的推动，基于微视频的自发性表情识别已经取得了不错的研究进展，涌现出了各种各样的表情自动识别模型，如 EmoNets[2]、VGGNet[3]、HoloNet[4]、VGG-LSTM[5]、C3Ds[6]等，这些模型在各大表情识别竞赛中，都取得了不错的成绩。但总体来说，这些模型识别的精确度还不尽人意，相比于人类的识别能力以及在这些领域上的应用，表情的智能识别还有很长的一段路要走。本文针对表情智能识别过程中存在的一些关键性问题，设计了一个全自动表情识别模型，并在该模型中构建了一个深度自编码网络来自动学习人脸表情特征，并结合证据理论对多分类结果进行有效融合。

2 研究现状

表情识别是在人脸检测的基础上发展起来的，和人脸识别一样，也需要经历人脸检测、图像预处理、面部特征提取、分类识别等过程。随着深度学习技术的广泛应用，表情识别方法也逐渐由传统的浅层学习方法向深度学习方法过渡。近些年来，表情识别技术的研究得到了学术界持续的重视，与之相关的情感识别竞赛也吸引着越来越多的人参加。其中由国际计算机协会多模态人机交互国际会议（ACM ICMI）主办的情感识别大赛 EmotiW 是世界范围内情感识别领域最高级别、最具权威性的竞赛，长期吸引了世界顶尖科研机构和院校的参与，包括微软美国研究院、Intel 研究院、IBM 研究院、美国密西根大学、美国波士顿大学、新加坡国立大学、北京大学、东南大学、爱奇艺等均参加了比赛。该

赛事每年举办一次，从 2013 年开始，迄今已连续举办了 6 届。国内举办的情感识别竞赛起步比较晚，由中国科学院自动化研究所领头举办的多模态情感竞赛 MEC(Multimodal Emotion Recognition)，迄今只举办了两次[7,8]。这些竞赛的定期举办，吸引了情感识别研究领域大部分研究机构参加，对该领域的交流和发展起到了巨大的推动作用。

面部表情特征提取在整个表情识别过程具有非常重要的作用，特征提取的好坏直接影响着最终的识别精度。在广泛使用深度学习技术来提取表情特征之前，研究者们主要提取一些传统的手工特征，如基于纹理信息变化的 Gabor 特征[9,10]和局部二值模式(LBP，Local Binary Pattern）特征[11]，以及在两者基础上扩展的 LGBP（Local Gabor Binary Pattern）特征[12]和 LBP-TOP（Local Binary Patterns from Three Orthogonal Planes）特征[13]；基于梯度信息变化的尺度不变性特征变换特征(SIFT，Scale Invariant Feature Transform，)[14]、方向梯度直方图(HOG，Histogram of Oriented Gradient)特征[15,16]和局部相位量化（LPQ，Local Phase Quantization）特征[17]等，以及在这三种特征上的扩展，如 Dense SIFT、MDSF(Multi-scale Dense SIFT Features)[18]、PHOG(Pyramid of Histogram of Gradients)[19]等。这些传统的手工特征在刚提出时，都取得了不错的效果。但这些特征在提取时容易受到干扰，对光照强度、局部遮挡和个体差异都非常敏感，而且提取的特征向量维度一般比较大，需要和其它的特征降维方法结合使用。

随着深度学习技术的应用，基于深度神经网络的面部特征自动学习方法逐渐成为热门。这类方法从局部到整体对面部信息进行统计，得到一些面部特征的统计描述，这类方法也可以简称为深度学习方法。深度学习方法本本质就是研究者们首先构建一个深度神经网络，然后利用大量样本进行训练，让机器自动统计其中的变化规律，从而学习出有效的特征表示。深度学习方法不同于浅层学习方法，它将特征学习和分类识别结合在了一起，并不需要单独提取出特征之后再进行分类。这种集特征提取和分类识别的深度神经网络模型，近些年发展比较快，比较典型的模型方法如表 1 所示。基于深度学习技术的特征学习方法虽然对旋转、平移和尺度变换都有着很强的鲁棒性，但也有着所有特征提取方法共同的缺陷：易受到噪声干扰。而且深度学习还需要大量的样本进行训练，如果样本量太少，效果并不如别的方法好。

Table 1 Facial expression recognition model

表 1 面部表情识别模型

作者	模型方法
Kahou[2]	EmoNets
白雪飞[20]，Kim[21]	Deep CNNs
Chan[22]	PCANet
李校林[23]，Parkhi[3]	VGG-Net
Fan[24]	CNN-RNN,C3D
Yao[4]	HoloNet
Hu[25]	SSE
Vielzeuf[5]	VGG-LSTM,C3D
Li[26]	DLP-CNN
Nguyen[6]	C3Ds,DBNs
Zhang[27],Li[28]	Hybrid CNNs

基于深度神经网络的特征学习方法虽然是现在主流使用的特征提取方法，但它也不能完全替代传统的手工提取方法，大部分研究者的做法是同时使用多种方法提取出特征，然后进行特征级融合，或者先对每个特征进行分类识别，再进行决策级的融合。也有研究者先提取传统的手工特征，再将这些特征融入到深度神经网络进行特征再学习[29-31]。本文的研究方法也提取了多种特征，并使用

证据理论方法进行决策级融合。

3 面部表情特征提取

每一张面部表情图像都来自于视频中的一帧，在这帧图像中，除了人类的面部信息，还有大量的背景信息。在进行特征提取时，需要先进行面部检测，只提取人物面部的特征。背景信息对人物情感识别没有太大的帮助作用，需要剔除掉。人脸检测本文采用开源的人脸检测算法 DSFD(Dual Shot Face Detector)[32]来完成，通过该算法，可以将视频转换成面部表情图片序列。

3.1 SA-DAE网络模型

基于微视频的表情识别，都是一个视频对应一个表情标签，单视频帧并不进行标注。大部分研究者在进行面部表情特征提取时，通常的做法是将整个视频的表情标签默认为每个帧的标签，再进行深度神经网络模型训练。这样做有很大的缺陷，会造成大量的图像样本标注错误。针对此种情况，本文将自适应注意力模型与自编码网络相结合，研究构建一个 SA-DAE(Self-attention Deep AutoEncoder)模型，该模型不仅可以以非监督方式提取面部表情特征，还能对传统的卷积神经网络进行改进，在不增加参数规模的前提下，最大可能地去获取全局信息。

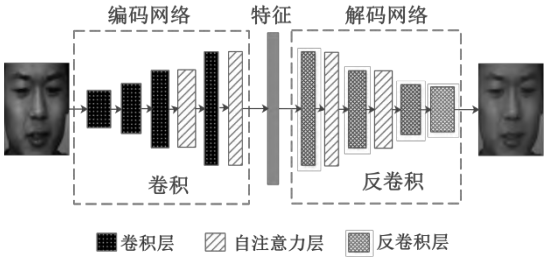

Figure 1 SA-DAE network model

图 1　SA-DAE 网络模型

通过人脸检测后，每个微视频就转换成了一个人脸图片序列，然后本文将序列中每一张人脸图片输入到已经训练好的 SA-DAE 网络中，根据自编码网络的特性，对每帧图片进行非监督特征提取。本文构建的 SA-DAE 网络如图 1 所示，该模型是对原始的自编码网络的一种改进，将原来的全连接层全部改成了卷积层或反卷积层，并在其中加入了自注意力层。模型训练好后，输入一张新的人脸图片，经过编码网络就能提取出该人脸的面部行为特征。

3.2 自注意力机制

卷积神经网络的核心之处是卷积操作，不同于全连接，它以局部感受野和权值共享为特点，对某个区域进行卷积操作时，默认为只与周围小范围内区域有关，与其它部分无关。这种卷积操作的特性大大减少了参数量，加快了整个模型的运行过程，因此卷积层一直是深度神经网络中的首选。但就因为这些特性，导致了卷积操作的弊端：会丢失一些空间上的关联信息。如果一张图片中两个区域离得比较远，但却是相互关联的，比如人脸具有对称性，左右眼角、左右嘴角在进行表情识别时，是有空间联系的。卷积操作忽略了这一个问题，默认为这两个区域无关联，从而丢失一些至关重要的空间关联信息。解决方法就是扩大卷积核，但卷积核太大时，参数量又会呈直线上升。如何在参数量和卷积范围之间找到一个平衡，本研究结果引入自注意力机制，该机制既考虑到了非局部卷积问题，又考虑到了参数量问题，具体实现如图 2 所示。

经过前一层的卷积操作后，会得到很多的卷积特征图（Convolutional Feature Maps），在进行下一层的卷积操作之前，模型将这些卷积特征图输入到一个自注意力层中，提取这些图中包含的全局空间信息。主要步骤包括：

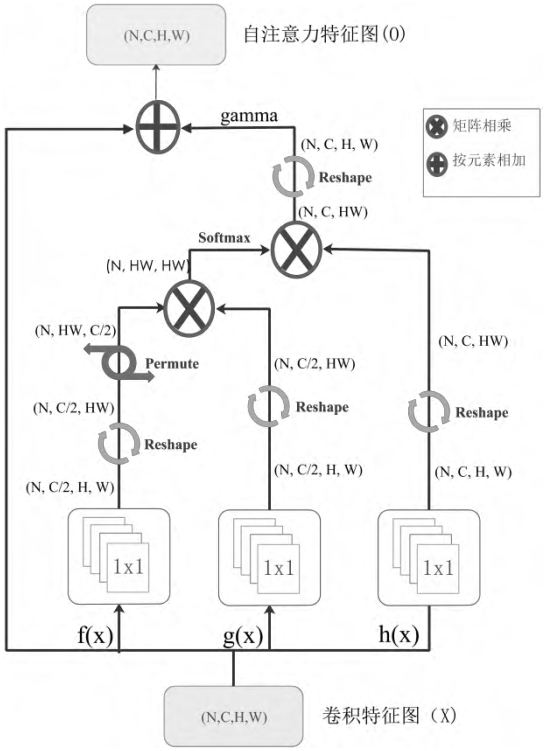

Figure 2 The flow chart of Self-Attention

图 2　Self-Attention 层实现流程

(1)将每个特征图分别进行 $f(x)$，$g(x)$和 $h(x)$ 变换，这三种变换都是普通的 1x1 卷积，差别只在于输出通道数量不同。变换之后再分别进行 reshape 操作，以便于后继的矩阵运算。

$$\begin{aligned} f(x) &= W_f x \\ g(x) &= W_g x \\ h(x) &= W_h x \end{aligned} \quad , \qquad (1)$$

(2)将 f(x)变换后的输出进行转置，并和 g(x)的输出进行矩阵相乘，再经过 softmax 进行归一化。其中下标 i 和 j 分别表示图像不同的区域，矩阵相乘表示模型生成 j 区域图像内容时 i 区域的参与程度，即两个区域间的空间相关性，S 为归一化后的相关性矩阵。

$$S_{j,i} = softmax(f(x_i)^T \cdot g(x_j)) \quad , \qquad (2)$$

(3)归一化后的结果和 h(x)变换后的矩阵相乘，得到自注意力特征图（Self-Attention Feature Maps）。最终把全局空间信息和局部信息整合到一起，融合得到加入了注意力机制的特征图。

$$o = x + \gamma(\sum_{i=1}^{N} S_{j,i} h_i(x)) \quad , \qquad (3)$$

注意力层的最终输出兼顾了邻域信息和远距离空间相关性，这里引入了一个参数 γ，它从 0 开始初始化，为的是让网络首先关注邻域信息，之后再慢慢把权重分配到其他远距离特征上。

4 表情自动识别模型

4.1 Dempster-Shafer证据理论融合策略

不同的特征表征着不同的辨别信息，将这些信息的分类结果进行融合，可以有效地互补。本文除了使用 SA-DAE 网络自动提取面部表情特征，还通过其它成熟的特征提取算法，提取了一些传统的手工特征，如 LBP-TOP、HOG、DSIFT 等，不同的特征有不同的分类结果，这就需要采用信息融合方法对不同的分类结果进行融合。

某一个样本应该分在哪一类，这是不确定的；同一个样本，通过不同的特征信息进行分类，也有可能分在完全不同的类。这种模式分类的不确定性和模糊性，刚好与不确定性推理方法相吻合，因此在本文中，不确定推理方法中的 DS 证据理论被引入到分类结果融合策略中。

在经典的 D-S 证据理论中，Θ 表示识别框架，它包含了 n 个不相容的命题，数学符号表示为 $\Theta = \{A_j | 1 \le j \le n\}$，$\Omega = 2^{\Theta}$ 是 Θ 的幂集，函数 $m: 2^{\Theta} \to [0,1]$ 将所有命题的幂集全部映射到一个概率值（取值介于 0 和 1 之间），满足下列两个条件：

$$m(\Phi) = 0, \qquad (4)$$

$$\sum_{A \subseteq \Theta} m(A) = 1, \qquad (5)$$

公式中的函数 $m(.)$ 称为基本概率分配(BPA，basic probability assignment)，也称之为 mass 函数。A 代表命题，$m(A)$ 表示在识别框架中证据对某个命题 A 的精确信任度。D-S 证据理论的融合规则如下：

$$(m_1 \oplus m_2)(A) = \frac{1}{1-k} \sum_{B \cap C = A} m_1(B) \times m_2(C), \quad (6)$$

其中，

$$k = \sum_{B \cap C = \phi} m_1(B) \times m_2(C), \quad (7)$$

称之为归一化因子，反应了证据之间的冲突程序。B 和 C 为任意两种命题，如果两种命题间无交集(相互独立)，则二者的 mass 函数值乘积就是一个冲突的衡量。当 k 趋近于 0 时，表示两证据之间无冲突，可以完全融合；反之，当 k 趋近于 1 时，表示两证据之间高度冲突，融合效果会很差。

在具体的表情识别模型中，每个命题即是一种表情类别，每个特征即为一个证据。mass 函数则代表某个特征对某种表情的信任度，即在某种特征情况下，视频被分类为该表情的概率。在本文提出的表情自动识别模型中，先利用随机森林算法对每个特征分别进行分类，每个特征的分类结果为一个 7 维的概率向量，向量中的每个值表示视频在该特征情况下分类为某种表情类别的概率。如果有 m 个特征，则最终的分类结果为 m x 7 的一个矩阵。模型再通过 D-S 证据理论的融合规则，把多个不同的分类结果向量融合成一个概率向量。

4.2 表情识别总体模型

表情的自动识别，需要经过人脸检测、特征提取、特征聚合、分类识别、结果融合等流程，本文将这些分散的模块结合在一起，就构成了一个全自动表情识别模型，模型结构如图 3 所示。该模型在微视频经过人脸检测得到人脸图片序列后，能自动学习深度神经网络特征，也能提取一些传统的手工特征，随后通过一个长短期记忆网络(LSTM，Long Short-Term Memory)将多个帧级特征聚合成视频级特征，再分别经过随机森林分类得到不同特征的分类结果，最后经过 DS 证据理论进行融合后，即得到最终的面部表情识别结果。

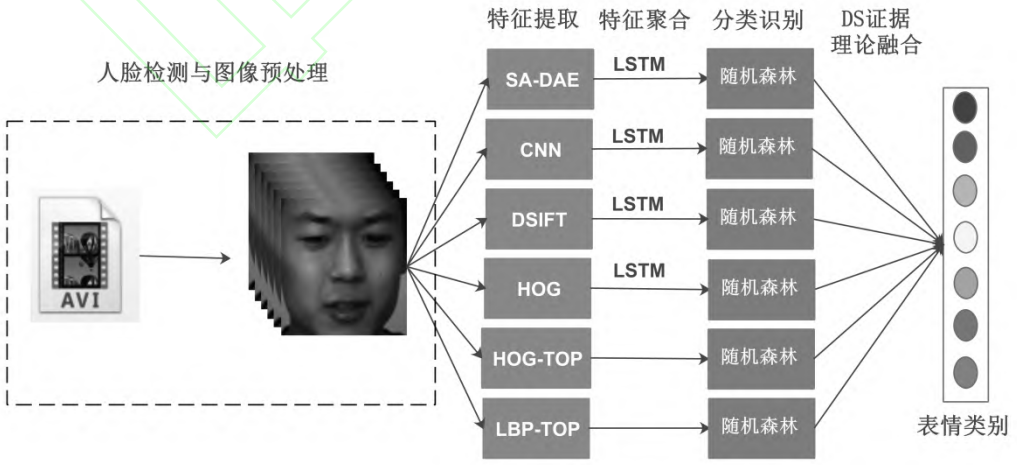

Figure 3 The model for multi-feature facial expression recognition

图 3　多特征面部表情识别模型

5 实验结果及分析

5.1 数据集

本文在中国科学院自动化研究所构建的CHEAVD2.0 数据库上进行了实验，实验结果与第二届多模态情感识别竞赛（MEC 2017）的参赛结果进行了对比。CHEAVD2.0 数据库的数据来源于影视剧中所截取的音视频片段，每一个音视频片段分别标注为一些常见情感（高兴、悲伤、生气、惊讶、厌恶、担心、焦虑）及中性情感中的一种。整个数据库将被分为训练集、验证集和测试集三部分，由于本文未收集到测试集的标签，因此本文用训练集来进行整个表情识别模型的训练，用验证集来验证模型的性能。在进行 SA-DAE 模型训练时，本文使用了迁移学习方法，先用大型人脸库 CeleA 进行初步训练，训练出来的模型参数再用 CHEAVD2.0 数据库进行微调。

5.2 评价指标

考虑到样本数据分布的不均衡性，本文以 MAP（macro average precision）作为模型的第一评价指标，以识别精确度作为第二评价指标。两个评价指标的公式如式(8)-(10)：

$$MAP = \frac{1}{s}\sum_{i=1}^{s} P_i \,, \tag{8}$$

$$P_i = \frac{TP_i}{TP_i + FP_i} \,, \tag{9}$$

$$ACC = \frac{\sum_{i=1}^{s} TP_i}{\sum_{i=1}^{s} (TP_i + FP_i)} \,, \tag{10}$$

其中， s 表示表情的类别，P_i 表示第 i 类表情的识别准确率，TP_i 和 FP_i 分别表示在第 i 类表情上分类正确的数量和分类错误的数量。

5.3 实验结果分析

针对每一个视频，本文分别提取了 SA-DAE、CNN、DSIFT、HOG、HOG-LBP、LBP-TOP 六种特征。其中 CNN 特征是采用 VGG 网络模型经有监督训练提取出来的特征，HOG-LBP 特征是仿照 LBP-TOP 算法提取出来的特征，由 xy 面的 HOG 特征与 yz、xz 的 LBP 特征串联而成。各特征的在验证集上的分类结果如表 2 所示：

Table 2 Feature recognition result on the validation set

表 2　在验证集上各特征的分类识别结果

特征	参数［a , b］	MAP(%)	ACC(%)
SA-DAE	[100,10]	**48.6**	41.1
CNN	[50,10]	42.5	36.9
DSIFT	[100,9]	47.3	41.2
HOG	[100,10]	43.2	**41.5**
HOG-LBP	[100,None]	37.8	39.7
LBP-TOP	[100.10]	33.3	35.2

a ： 随机森林算法中树的数量参数

b ： 随机森林算法中树的深度参数

根据结果显示，在宏观平均精确度(MAP)评价指标上，SA-DAE 特征的分类效果优于其它特征，但在总体分类精确度(ACC)的评价指标上，SA-DAE 特征和传统的 DSIFT、HOG 特征，分类效果没有太多的区别。

在决策级融合阶段，本文先将所有的特征按照分类识别准确度从高到低进行了排序，然后将准确度最高的 SA-DAE 特征作为基础，按照顺序将其它特征逐项融合进来。SA-DAE、DSIFT 和 HOG 三个特征融合之后，分类效果有了较大的提升，但融合进第四个特征时，分类效果出现了下降，因此本文又以 SA-DAE+DSFIT+HOG 的融合特征作为基础，与剩下的特征进行穷举组合，最终不同特征融

合的分类结果如表 3 所示。在宏观平均精确度（MAP）评价指标上，SA-DAE、DSIFT、HOG、HOG-LBP 四种特征的证据理论融合结果效果最好，达到了 53.39%，在总体分类精确度(ACC)的评价指标上，SA-DAE 、DSIFT、HOG、HOG-LBP、CNN 五种特征融合效果优于其它特征融合策略。

Table 3 Feature fusion recognition result on the validation set

表 3 在验证集上不同特征融合后的分类识别结果

特征融合策略	MAP(%)	ACC(%)
SA-DAE+DSIFT	50.22	41.38
SA-DAE+DSIFT+HOG	50.57	41.67
SA-DAE+DSIFT+HOG+CNN	39.96	45.98
SA-DAE+DSIFT+HOG+HOG-LBP	**53.39**	42.24
SA-DAE+DSIFT+HOG+LBP-TOP	40.04	42.09
SA-DAE+DSIFT+HOG+HOG-LBP+CNN	42.74	**46.55**
SA-DAE+DSIFT+HOG+HOG-LBP+LBP-TOP	45.43	42.24

实验最后，本文将提出的表情识别模型也应用到了数据库的测试集上，并根据数据库提供方反馈的识别结果，与数据库的分类识别基线水平进行了对比（如表 4），本文提出的模型不管是在验证集上还是在测试集上，识别精度都取得了不错的效果，远远超过了基线水平。

Table 4 Recognition result on the validation set and testing set

表 4 在验证集和测试集上的分类识别结果

	验证集		测试集	
	MAP(%)	ACC(%)	MAP(%)	ACC(%)
本文的方法	53.39	46.55	59.68	44.81
基线水平[8]	34.1	36.5	21.7	35.3

6 结束语

本文结合深度自编码网络、自注意力模型和 D-S 证据理论，构建了一个表情自动识别模型。根据实验结果显示，该模型提取的非监督深度学习特征，分类效果优于其它特征。在多特征分类结果融合方面，该模型也取得了不错的成绩，识别效果远远高于数据库的基线水平。但是，模型识别的精确度还远远落后于人类的识别能力，表情自动识别在现实生活中的应用，还有很长的一段路要走。

参考文献:

[1] Mehrabian A. Communication without words[J]. Psychology Today, 1968, 2(9): 52-55.

[2] Kahou S E, Bouthillier X, Lamblin P, et al. EmoNets: Multimodal deep learning approaches for emotion recognition in video[J]. Journal on Multimodal User Interfaces, 2015, 10(2):1-13.

[3] Parkhi O M, Vedaldi A, Zisserman A. Deep face recognition[J]. bmvc. 2015, 1(3): 6.

[4] Yao A, Cai D, Hu P, et al. HoloNet: towards robust emotion recognition in the wild[C]//Proc of the 18th ACM International Conference on Multimodal Interaction, ACM, 2016: 472-478.

[5] Vielzeuf V, Pateux S, Jurie F. Temporal multimodal fusion for video emotion classification in the wild[C]//Proc of the 19th ACM International Conference on Multimodal Interaction, ACM, 2017: 569-576.

[6] Nguyen D, Nguyen K, Sridharan S, et al. Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition[J]. Computer Vision and Image Understanding, 2018, 174: 33-42.

[7] Li Y, Tao J, Schuller B, Shan S, et al. MEC 2016: the multimodal emotion recognition challenge of CCPR 2016[C]//Porc of the 7th Chinese Conference on Pattern Recognition, 2016:667-678.

[8] Li Y, Tao J, Schuller B, et al. Mec 2017: Multimodal emotion recognition challenge[C]//Proc of 1th Asian Conference on Affective Computing and Intelligent Interaction

(ACII Asia),IEEE, 2018: 1-5.

[9] Lyons M J, Budynek J, Akamatsu S, Automatic Classification of Single Facial Images[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2002, 21(12):1357-1362.

[10] Zhang Z, Lyons M, Schuster M, et al. Comparison between geometry-based and gabor-wavelets-based facial expression recognition using multi-layer perceptron[C]//Proc of 3th IEEE International Conference on Automatic face and gesture recognition, IEEE, 1998:454-459.

[11] Shan C, Gong S, Mcowan P W. Robust facial expression recognition using local binary patterns[C]// Proc of IEEE International Conference on Image Processing, 2005,(2):II-370.

[12] Zhang W, Shan S, Chen X, et al. Local gabor binary patterns based on mutual information for face recognition[J]. International Journal of Image and Graphics,2007, 7(04):777-793.

[13] Zhao G, Pietikinen M. Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2007. 29(6): 915-28.

[14] Mo Xiu-fei. The recognition of facial expression based on SIFT algorithm [D]. Xidian University, 2012.

[15] Tong Ying. Facial expression recognition algorithm based on spatial multi-scaled HOG feature[J]. Computer engineering and design, 2014, 35(11):3918-3922.

[16] Zhong Wei, Huang Yuan-liang. A facial expression recognition algorithm based on feature fusion and hierarchical decision tree technology[J]. Computer Engineering & Science, 2017,39(02): 393-398.

[17] Zhang B., Liu G, Xie G. Facial expression recognition using LBP and LPQ based on Gabor wavelet transform[C]Proc of the 2nd IEEE International Conference on Computer and Communications, IEEE, 2016:365-369.

[18] Sun B, Li L, Zuo T, et al. Combining multimodal features with hierarchical classifier fusion for emotion recognition in the wild[C]Proc of the 16th international conference on multimodal interaction , 2014:481-486.

[19] Dhall A, Asthana A, Goecke R, et al. Emotion recognition using PHOG and LPQ features[C]Proc of International Conference on Face and Gesture, IEEE, 2011 :878-883.

[20] Bai Xue-fei, Li Ru. Neural network ensemble based expression invariant face recognition[J].Computer Engineering and Applications, 2010,46(04):145-148.

[21] Kim B K, Lee H, Roh J, et al. Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition[C]//Proc of the 2015 ACM on International Conference on Multimodal Interaction. ACM, 2015: 427-434.

[22] Chan T H, Jia K, Gao S, et al. PCANet: A simple deep learning baseline for image classification[J]. IEEE transactions on image processing, 2015, 24(12): 5017-5032.

[23] Li Xiao-ling, Niu Hai-tao. Facial expression recognition using feature fusion based on VGG-NET [J]. Computer Engineering & Science, 2020,42(3): 500-509.

[24] Fan Y, Lu X, Li D, et al. Video-based emotion recognition using CNN-RNN and C3D hybrid networks[C]//Proc of the 2016 ACM on International conference on multimodal interaction. ACM, 2016:445-450.

[25] Hu P, Cai D, Wang S, et al. Learning supervised scoring ensemble for emotion recognition in the wild[C]//Proc of the 2017 ACM International Conference on Multimodal Interaction. ACM, 2017: 553-560.

[26] Li S, Deng W, Du J. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild[C]//Proc of the IEEE conference on computer vision and pattern recognition. CVPR, 2017: 2852-2861.

[27] Zhang S, Pan X, Cui Y, et al. Learning affective video features for facial expression recognition via hybrid deep learning[J]. IEEE Access, 2019: 32297-32304.

[28] Li S, Zheng W, Zong Y, et al. Bi-modality Fusion for Emotion Recognition in the Wild[C]//Proc of the 2019 ACM International Conference on Multimodal Interaction. ACM, 2019: 589-594.

[29] Sun Xiao, Pan Ting, Ren Fu-ji. Facial expression recognition using ROI-KNN deep convolutional neural networks [J]. Acta automatica sinica, 2016,42(06):883-891.

[30] Jiang Da-peng, Yang Biao, Zou Lin. Facial expression recognition based on local binary mode convolution neural network[J]. Computer engineering and design, 2018, 39(07): 1971-1977.

[31] Zhang Yu-qing, He Ning, Wei Run-chen. Face expression recognition based on convolutional neural network fusion SIFT features [J]. Computer applications and software, 2019, 36(11): 161-167.

[32] Li J, Wang Y, Wang C, et al. DSFD: dual shot face detector[C]//Proc of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2019: 5060-5069.

附中文参考文献：

[14] 莫修飞. 基于SIFT算法的人脸表情识别[D].西安电子科技大学,2012.

[15] 童莹. 基于空间多尺度HOG特征的人脸表情识别方法[J]. 计算机工程与设计, 2014, 35(11): 3918-3922.

[16] 钟伟,黄元亮. 基于特征融合与决策树技术的表情识别方法[J]. 计算机工程与科学, 2017,39(02): 393-398.

[20] 白雪飞,李茹.神经网络集成的多表情人脸识别方法[J]. 计算机工程与应用,2010,46(04): 145-148.

[23] 李校林,钮海涛. 基于VGG-NET的特征融合面部表情识别[J]. 计算机工程与科学, 2020,42(3): 500-509.

[29] 孙晓,潘汀,任福继.基于ROI-KNN卷积神经网络的面部表情识别[J].自动化学报,2016,42(06): 883-891.

[30] 江大鹏,杨彪,邹凌. 基于LBP卷积神经网络的面部表情识别[J].计算机工程与设计, 2018, 39(07): 1971-1977.

[31] 张俞晴,何宁,魏润辰.基于卷积神经网络融合SIFT特征的人脸表情识别[J].计算机应用与软件, 2019, 36(11): 161-167.

作者简介：

徐其华（1979-），男，四川宜宾人，博士生，副教授，研究方向为视频与图像智能处理，**E-mail**:marco@nwnu.edu.cn

XU Qi-hua, born in 1979, Ph.D. candidate, associate professor, his research interest includes video and image intelligent processing.

孙波（1968-），男，湖南常德人，博士，教授，研究方向为图像处理与模式识别，**E-mail:**tosunbo@bnu.edu.cn

SUN Bo, born in 1968, Ph.D, professor, his research interest includes image processing and pattern recognition.