DOI: 10.13382/j.jemi.B1902824

融合边缘检测和递归神经网络的视频表情识别*

胡 敏[1] 高 永[1] 吴 昊[1] 王晓华[1] 黄 忠[2]
(1. 合肥工业大学 计算机与信息学院 合肥 230601; 2. 安庆师范大学 物理与电气工程学院 安庆 246011)

摘 要:为有效解决传统视频人脸表情识别通常只关注单张视频帧的空间特征,而忽略了相邻帧之间隐藏的时间特征的问题,提出一种结合边缘检测和递归神经网络的视频表情识别方法,利用梯度边缘检测准确地提取输入图像的纹理信息,同时提出一种分片交叉 LSTM 结构,提取出图像序列中隐藏的时空特征。实验在 CK+和 MMI 视频库上进行,在 OCNN-RNN 网络中分别取得 88.4%和 69.7%的识别率,在 GCNN-RNN 网络中分别取得 89.8%和 73.6%的识别率,最终使用提出的加权随机搜索方法融合 GCNN-RNN 和 OCNN-RNN 两个网络之后,分别取得了 94.6%和 79.9%的识别率,均优于单流网络算法,证明了所提算法的有效性。

关键词: 时空特征;边缘检测;递归神经网络;随机搜索
中图分类号: TP391.4;TN0 **文献标识码**:A **国家标准学科分类代码**:510.4050

Video facial emotion recognition based on edge detection and recurrent neural network

Hu Min[1] Gao Yong[1] Wu Hao[1] Wang Xiaohua[1] Huang Zhong[2]
(1. School of Computer and Information of Hefei University of Technology, Hefei 230601, China;
2. School of Physics and Electronic Engineering, Anqing Normal University, Anqing 246011, China)

Abstract: In view of the existing algorithms, traditional video emotion-based facial expression recognition method only pays attention to the spatial features of a single video frame, and ignores the hidden temporal features between adjacent frames. Therefore, this paper proposes a novel method to extract features using edge detection and improved recurrent neural network. Gradient edge detection can extract texture information of video frame in a more accurate way, at the same time, a kind of overlapping LSTM structure is proposed, and the recurrent neural network can acquire the hidden spatio-temporal information from the input frames. The experiments in this paper are carried out on the CK+ and MMI video database. the result of 88.4% and 69.7% are obtained in the OCNN-RNN network respectively, and the outcome of 89.8% and 73.6% are acquired in the GCNN-RNN network from each database. and finally the random search is used to weight the fusion of the results of the GCNN-RNN network and the OCNN-RNN network. After the two networks are finally merged, the average recognition rate of the integrated model is 94.6% and 79.9% respectively, and the accuracy is better than other algorithms, the effectiveness of the proposed algorithm is proved.

Keywords: spatio-temporalfeatures; edge detection; recurrent neural network; random search

0 引 言

情感交流是人们日常生活中必不可少的一个部分,自从 20 世纪 Picard 提出情感计算以来,人脸表情识别、语音情感分析、人机交互、医疗陪护等基于情感研究的各个领域都有了长足发展,其中人脸表情识别[1-2]是非语言的情感表达中非常重要的一个部分。

使用人脸表情来分析人类情感状态相比于其他情感信号来说更加切实可行,作为人机交互的重要组成部分,关于人脸表情的研究已成为人工智能的一个重要研究方向。目前,传统的表情识别方法,如局部二值模式(local

收稿日期:2019-12-12 Received Date: 2019-12-12
*基金项目:国家自然科学基金(61672202,61673156)、国家自然科学基金-深圳联合基金重点项目(U1613217)资助

binary pattern, LBP), 方向梯度直方图(histogram of oriented gradient, HOG)[3], 韦伯局部描述子(Weber local descriptor, WLD)[4]等均取得了不错的实验效果,但传统方法提取的均是手工特征,提取出来的特征局限于预先设定的特征空间,在没有额外训练阶段的情况下,难以形成其他特征来覆盖人脸图像的所有变化。同时,为了解决图片光照、人脸姿态、拍摄角度和不同人种等各类外界因素的变化对识别结果造成干扰,需要在实验中引入更多数据,以从中提取较充分的信息。

随着计算机技术的持续发展,大量数据的获取变得更加容易,计算机硬件的性能也稳步提高,这为深度学习提供了充分的成长空间,并因此诞生了众多优秀的深度学习模型。卷积神经网络(convolutional neural networks, CNN)便是深度学习领域最为经典的模型之一,通过对卷积层和池化层的组合,CNN 相较于传统的空间特征提取方法,能够在保持图片空间不变性的基础上,非常高效地提取图像的空间特征,并达到优异的效果。许多学者在 CNN 的基础上提出了自己的方法[5-8]。由于传统的 2D 卷积核不能直接应用在视频特征提取,所以为了能够处理视频序列的分类任务,Tran 等[9]提出了 3D 卷积,能够更加充分的提取时空特征,并在行为识别中得到了有效的验证。在引入 3D 卷积之后,Hosseini 等[10]发现找到给定问题的适当特征可能仍然很重要,它们可以增强基于 CNN 的算法性能。

如在行为识别中,光流经常被提取出来作为一种特征输入网络训练。Sevilla-Lara 等[11]证明光流之所以能够取得比较好的识别结果,原因在于其对于图像表观的不变性,受到文献[11]的启发,考虑到 Sobel 算子在提取图像边缘特征方面表现的优越性能,利用文献[12]的改进 Sobel 算子,先对人脸图片进行边缘检测处理,然后再送入网络进行训练。

相较于传统单输入流的卷积神经网络,提取空间特征,但针对视频序列,卷积网络并不能提取重要的时间维度信息,而双流神经网络通过效仿人体视觉过程,在处理视频图像中的环境空间信息的基础上,对视频帧序列中的时序信息进行理解,Fenchtenhofer 等[13]最先在行为识别上使用双流网络,取得了不错的实验成果,所以本文借鉴文献[13]使用双流网络作为基础模型架构,同时为了充分提取视频序列间隐藏的时间特征,遵循 Ma 等对 LSTM(long short-term memory)的研究工作[14],在此基础上本文提出了基于视频情感识别的交叉分片 LSTM 网络(overlapping segment LSTM, OS-LSTM),对视频中的人脸表情进行分类,在实验中取得了很好的效果。

1　Sobel 算子

Sobel 算子的原理是对于连续图像函数 $f(x,y)$ 在点 (x,y) 处梯度 $f'(x,y)$ 是一个具有方向和大小的矢量,即:

$$f'(x,y)=\frac{\partial f}{\partial x}\boldsymbol{i}+\frac{\partial f}{\partial y}\boldsymbol{j} \tag{1}$$

式中:$\boldsymbol{i}$ 和 $\boldsymbol{j}$ 分别代表 X 轴和 Y 轴方向的单位矢量,点 (x,y) 的梯度幅值,$grad[f(x,y)]$ 和梯度方向 θ 如式(2)、(3)所示,其中式(2)计算出的梯度值就是图像在 (x,y) 处的边缘数据。

$$grad[f(x,y)]=\left[\left(\frac{\partial f}{\partial x}\right)^2+\left(\frac{\partial f}{\partial y}\right)^2\right]^{\frac{1}{2}} \tag{2}$$

$$\theta=\arctan\left[\frac{\partial f}{\partial y}\Big/\frac{\partial f}{\partial x}\right] \tag{3}$$

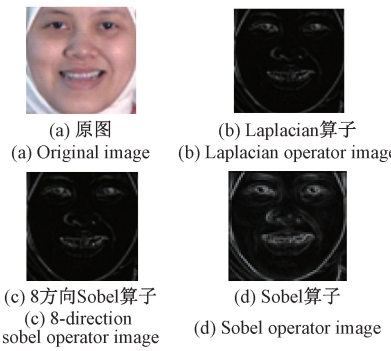
原始的 Sobel 算子模板只包含 0°和 90°两个方向,为了能得到整幅图片全方位的更加全面的边缘信息,文献[12]将模板从两个方向扩展到 8 个方向,分别是 0°、45°、90°、135°、180°、225°、270°和 315°,具体算子的模板如图 1 所示。

$$\begin{pmatrix}1&0&-1\\2&0&-2\\1&0&-1\end{pmatrix}\begin{pmatrix}2&1&0\\1&0&-1\\0&-1&-2\end{pmatrix}\begin{pmatrix}0&1&2\\0&0&0\\-1&-2&-1\end{pmatrix}\begin{pmatrix}0&1&2\\-1&0&1\\-2&-1&0\end{pmatrix}$$
$$\begin{pmatrix}-1&0&1\\-2&0&2\\-1&0&1\end{pmatrix}\begin{pmatrix}-2&-1&0\\-1&0&1\\0&1&2\end{pmatrix}\begin{pmatrix}0&-1&-2\\0&0&0\\1&2&1\end{pmatrix}\begin{pmatrix}0&-1&-2\\0&0&0\\1&2&1\end{pmatrix}$$

图 1　8 方向 Sobel 算子

Fig. 1　8-direction sobel operator graph

图 2 所示为人脸图片分别经 Sobel 算子、改进后的 8 方向 Sobel 算子和 Laplacian 算子处理之后的结果,可以看出,8 方向的 Sobel 算子在表现人脸轮廓方面效果更好。



(a) 原图 (a) Original image　(b) Laplacian算子 (b) Laplacian operator image

(c) 8方向Sobel算子 (c) 8-direction sobel operator image　(d) Sobel算子 (d) Sobel operator image

图 2　人脸图像边缘检测效果

Fig. 2　Edge detection of face image

2　提出方法

2.1　交叉分片 LSTM 神经网络 OS-LSTM

长短记忆神经网络，它通过计算网络单元激活，图 3 所示为 LSTM 网络单元的示意图。

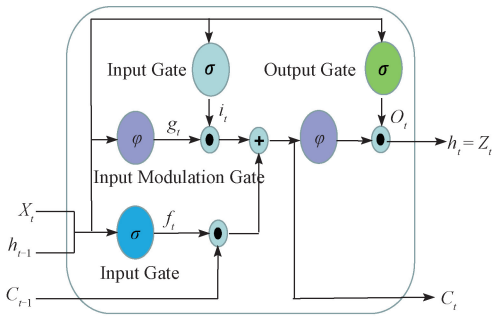


图 3　LSTM 网络单元示意图

Fig. 3　The diagram of LSTM network unit

计算从输入序列 $\boldsymbol{x}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_s)$ 到输出序列 $\boldsymbol{y}=(\boldsymbol{y}_1,\boldsymbol{y}_2,\cdots,\boldsymbol{y}_s)$ 的映射，通过迭代计算 $t=1\cdots s$ 的网络中的激活单元。LSTM 网络其中输入门 i_t、遗忘门 f_t、和输出门 o_t 的定义如式(4)~(8)所示。

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{xi}\boldsymbol{x}_t+\boldsymbol{W}_{hi}\boldsymbol{h}_{t-1}+\boldsymbol{b}_i) \tag{4}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{xf}\boldsymbol{x}_t+\boldsymbol{W}_{hf}\boldsymbol{h}_{t-1}+\boldsymbol{b}_f) \tag{5}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t+\boldsymbol{W}_{ho}\boldsymbol{h}_{t-1}+\boldsymbol{b}_o) \tag{6}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\varphi(\boldsymbol{W}_{xc}\boldsymbol{x}_t+\boldsymbol{W}_{hc}\boldsymbol{h}_{t-1}+\boldsymbol{b}_c) \tag{7}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\varphi(\boldsymbol{c}_t) \tag{8}$$

式中：$\boldsymbol{W}$ 表权重矩阵；$\boldsymbol{b}$ 表偏置向量；σ 是 *sigmoid* 函数；i、f、o 和 c 分别代表输入门、遗忘门、输出门和存储单元。LSTM 包括的 h_t、i_t、f_t 和 o_t 的大小都是 $\boldsymbol{R}^N$。$\odot$ 表示向量按元素相乘，φ 是细胞单元的输入和输出激活函数，一般为 tanh。

由于 Sobel 算子重点关注面部五官信息，忽略了皮肤纹理变化对表情识别的贡献，所以为了减少提取人脸轮廓信息时所丢失的部分特征，并提取出隐藏在图像帧序列间的时空特征，本文提出一种交叉分片 LSTM 神经网络，网络结构如图 4 所示。

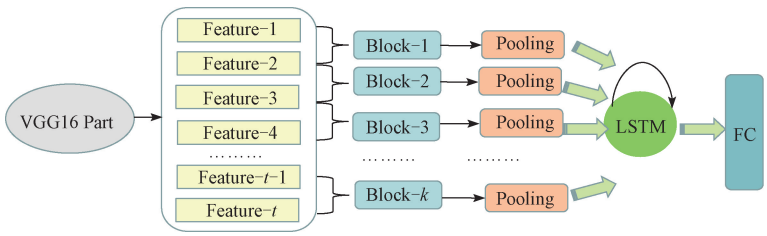


图 4　LSTM 交叉分片示意图

Fig. 4　The diagram of LSTM overlapping segment

首先对给定的视频序列 F，从序列中抽取 t 帧，记作 $\boldsymbol{X}_t=\{\boldsymbol{f}_1,\boldsymbol{f}_2,\boldsymbol{f}_3,\cdots,\boldsymbol{f}_t\}$，利用 VGG16 卷积神经网络逐帧提取空间特征，转换成特定长度的空间向量，接着将提取出来的空间特征堆叠分成 k 片 $\{\boldsymbol{T}_1,\boldsymbol{T}_2,\boldsymbol{T}_3,\cdots,\boldsymbol{T}_k\}$ 分片采用步长为 1 的交叉，每片包含固定帧数，然后对分片后的空间特征矩阵进行池化操作，接着将提取出的特征再送入到 LSTM 中进行学习，最后经过 softmax 层进行分类。本文的视频序列时空特征用公式如下表示：

$$\boldsymbol{h}_k=D(\boldsymbol{h}_{k-1},\boldsymbol{p}_k;\boldsymbol{w}_r) \tag{9}$$

$$\boldsymbol{p}_k=p(F_{conv}(\boldsymbol{f}_i;\boldsymbol{w}_c),\cdots,F_{conv}(\boldsymbol{f}_j;\boldsymbol{w}_c)),\quad f_i,\cdots,f_j\in T_k \tag{10}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{xo}\boldsymbol{x}_t+\boldsymbol{W}_{ho}\boldsymbol{h}_{t-1}+\boldsymbol{b}_o) \tag{11}$$

式中：$F_{conv}(\boldsymbol{f}_i;\boldsymbol{w}_c)$ 表示对 $\boldsymbol{f}_i$ 做参数为 $\boldsymbol{w}_c$ 的卷积操作；$p()$ 表示由 CNN 提取出的一组特征进行池化操作，代表对帧序列特征进行池化操作；D 是代表 LSTM 网络遗忘更新函数；最终输出 h_k 是表情分类中每种类别的得分，经过 Softmax 函数后得到每类表情的概率值。

2.2　基于梯度边缘检测的加权网络融合模型

本文利用 GCNN-RNN(gradient-CNN-RNN)网络对梯度边缘检测图进行处理，提取面部纹理信息，OCNN-RNN(original-CNN-RNN)网络对原图提取人脸面部五官信息，Kaya 等[15]，在多模型任务中，对不同模型的预测结果进行加权融合，对于提升实验预测结果有很大帮助，鉴于 OCNN-RNN 与 GCNN-RNN 具有一定程度的互补性，为此本文构建了一个基于上述两个网络加权融合的深度学习网络模型框架，本文的网络框架如图 5 所示。

整个网络分为两个部分。1)输入的是 RGB 三通道的原图，利用预训练后的 VGG16 模块用来提取空间特征，然后将空间特征分片送入到分片结构中，提取时间特征。2)输入处理后的边缘检测图片，经由 VGG16 提取空间特征后，再送入分片网络，最终的预测结果是由这两部分加权融合得到的，如式(12)所示。

$$\boldsymbol{R}_i=\theta\boldsymbol{P}_i+(1-\theta)\boldsymbol{Q}_i,0\leqslant\theta\leqslant1 \tag{12}$$

式中：i 的值是从 $1\sim c$，c 代表情感类别数目。从 OCNN-RNN 中得到的预测向量是 $\boldsymbol{P}_i$；从 GCNN-RNN 获得的预测向量是 $\boldsymbol{Q}_i$，$\boldsymbol{R}_i$ 代表最终的预测结果。

本文使用随机搜索来筛选不同权值，选择范围在[0,1]之间，并且对于每一类来说，所有模型的权值之

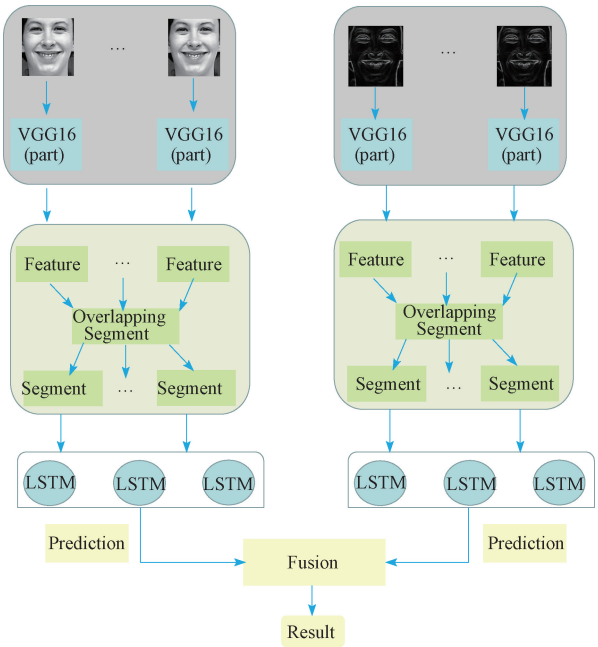


图5　基于梯度边缘检测的双流网络结构

Fig. 5　The overview of the integrated framework

和为1。首先,网络权重经过10 000次初始迭代,得到在验证集上的一个最优的数值 θ, 然后再对这个权值进行局部随机搜索。局部随机搜索就是缩小搜索的区域,把搜索范围缩小到以 θ 为平均值,标准差 δ 为0.5的高斯分布范围内,一旦找到比目前更好的权值,就替换掉原来的权值 θ, 每经过10000次搜索迭代,对应的标准差 δ 会乘上衰退系数0.9,如果标准差 δ 小于0.000 1,则会停止迭代。步骤如表1所示。

表1　求取最佳权重

Table 1　The method of getting the best weight

输入:实际标签,预测结果
输出:最佳权重 θ 和最佳识别率 r
变量:权重 w,最佳权重 θ, 准确度 p
标准差 δ, 最佳识别率 r
for w in linspace (0,1,10 000)
计算准确度 p,如果 $p>r$
更新 θ 和 r
while $\delta>0.000\ 1$
for w in normal(r, δ, 10 000)
计算准确度 p,如果 $p>r$
更新 θ 和 r
$\delta = \delta \cdot 0.9$

3　实验结果与分析

3.1　表情训练数据库

本文的实验主要针对视频情感识别,选取MMI库和CK+用来作为模型评估的数据集。

CK+数据库是在实验室环境下拍摄的人脸表情库,包括123个主体,593段表情序列,其中有118个主体,327段表情序列具有情感标签,每张图片大小是640×496。本文实验选取6种表情,分别是生气、厌恶、恐惧、高兴、悲伤、惊讶。

MMI数据库包括30个年龄在19~62岁的不同性别的主体,MMI库中包含有不同的种族(欧洲、亚洲和南美),共有213个已被标记的6类人脸表情序列,分别是生气、厌恶、恐惧、高兴、悲伤和惊讶,每张图片大小为720×576。本文在实验时剔除了8个侧面拍摄的表情序列,选取剩余的205个表情序列作为实验数据。

CK+和MMI库中的每类表情,均是选择从平静到峰值的变化序列作为实验数据。

3.2　预处理与网络模型训练

1)数据预处理

对CK+和MMI数据库中所选取的表情序列,按照固定间隔从中选取16帧作为本文实验的表情序列。此外在实际训练中,如果直接将所选取的表情序列输入到网络中,由于光照变化、人脸姿态等问题,会对实验结果造成不必要的影响,所以对输入人脸数据进行预处理是有必要的。首先为了减少光照变化对实验的影响,采用文献[16]的方法,使用直方图均衡化和线性映射相结合的方式,既避免了直方图均衡化可能过分强调局部对比度,又解决了当图像已经具有较大的全局对比度时,线性映射不能很好地工作的问题。

为了去除无关背景的干扰,并达到较为理想的人脸检测效果,本文采用文献[17]的Viola Jones检测方法,检测效果如图6所示,它是一种经典并且被广泛采用的方法。

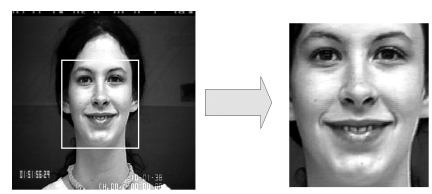


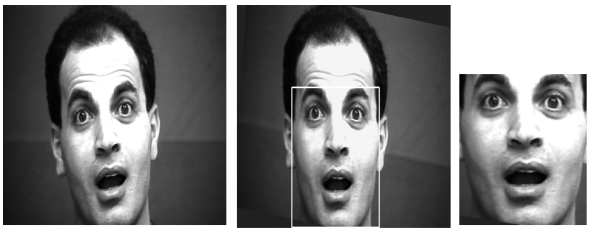
图6　人脸检测

Fig. 6　Face detection

为了纠正面部姿态问题,本文采用仿射变换的方法,对齐结果如图7所示,将人脸图像调整到统一角度,并最终将图片裁剪至196×196大小。

2)网络模型训练

深度学习在训练过程中,如果训练的数据量比较小,采用过深的网络如VGG19更容易造成过拟合,从而导致模型的泛化能力比较弱,因此,采用相对较浅的卷积神经网络结构模型是一个比较好的选择。



（a）原图 (a) Original image　（b）人脸对齐 (b) Face alignment　（c）人脸裁剪 (c) Cropped face

图 7　仿射变换

Fig. 7　Affine transformation

本文借鉴了 Fan 等[18]在 EmotiW2016 大赛中的成功经验，首先使用 FER2013 数据集对 VGG16 模型进行预训练，然后在数据集上进行微调。具体网络细节如图 8 所示，VGG16 模型中包含 13 个卷积层，5 个池化层，3 个全连接层和一个 sofamax 层。本文所有实验是在 Windows10 系统下，搭配 GTX1080 显卡使用 Keras 框架完成，采用学习率为 0.000 01 的 Adam 算法训练网络模型，每批次输入 6 个数据的批处理方式训练网络。

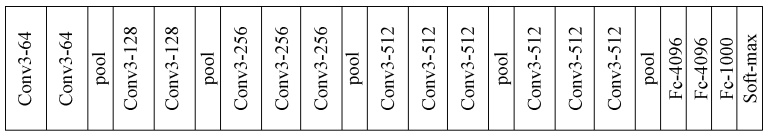


图 8　VGG16 模型结构

Fig. 8　The diagram of VGG model

3.3　实验结果

在 GCNN-RNN 模型中，本文选取上述经过预训练的 VGG16 网络作为模型，将梯度边缘检测算子提取出的特征图作为输入送入 VGG16 网络，然后 LSTM 层学习时空特征，最后输出结果。由于最终的测试结果是人脸无关的，即保证测试集中主体不会在训练集中出现，因此每个数据集都会被分为训练集和测试集，最终的识别结果采用五折交叉验证，然后求取平均值。

使用不同算子检测结果整理如图 9 所示。实验结果表明，在 CK+和 MMI 上，改进的 Sobel 算子均要优于 laplacian 算子和 Sobel 算子，能够更充分的体现人脸纹理信息。

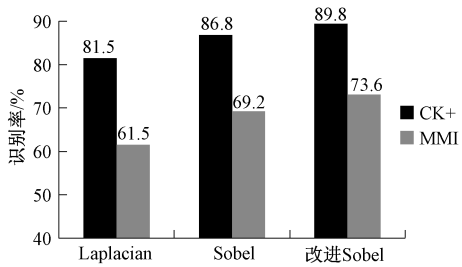


图 9　不同边缘检测算法在 GCNN-RNN 网络中识别率

Fig. 9　Recognition rate of different edge detection algorithms in GCNN-RNN network

在 OCNN-RNN 模型中，CNN 依旧采用是上述预训练好的 VGG16 网络，RNN 采用本文 2.1 节的 OS-LSTM 模型。实验时，在数据集中等间隔选择 16 帧，依次作为输入然后送到 VGG16 网络，再将经 VGG16 网络提取后的特征送入到 RNN 模型中提取分片，输入数据是维度为(16,196,196,3)的张量，其中 16 表示连续 16 帧图片作为一个序列送入到网络中，两个 196 表示输入图片的大小是 196×196，最后的 3 表示 3 图像的 RGB 通道，如图 10 所示，在 CK+和 MMI 库上每类表情的实验结果以混淆矩阵的形式来展现。

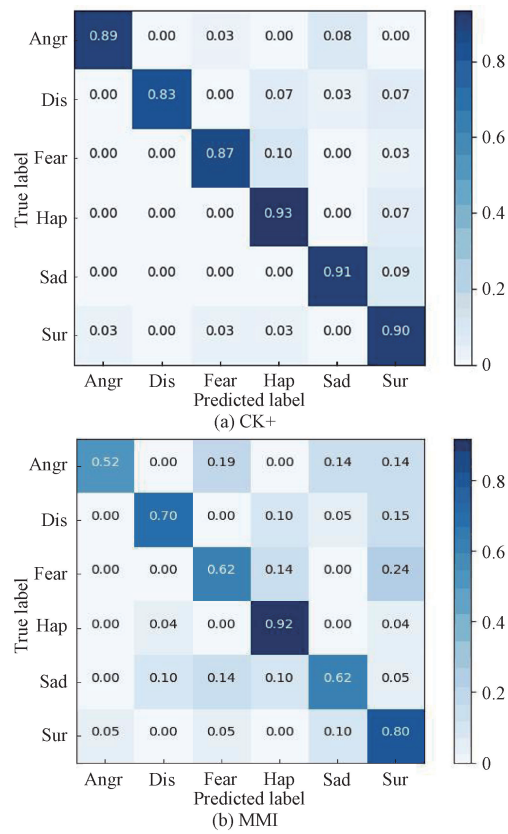


图 10　在 CK+和 MMI 库上各类表情识别率

Fig. 10　Recognition rate of expression on CK+ and MMI database

从图10可以看出,高兴和惊讶表情的识别效果较好,但害怕和悲伤表情的识别效果则较低,产生这种结果的原因在于,高兴、惊讶表情的变化比较明显,反应到特征上即相互之间差异大,容易区分,而害怕、悲伤的面部变化较小,容易产生误识别的情况。

此外,由于不同的分片数对应着不同的时间段划分,也即是包含了不同的时空特征,因此为了探讨不同分片数对实验结果的影响,本文将分片数分为4、5、7、15,对应的每片包含的特征数目为5、4、3、2。

实验结果如表2所示,其中特征帧数目是指一个分片中所包含的帧数。由于片与片之间帧的交叉,因此表格中分片数与特征帧数目的乘积并不等于输入序列的长度。

表2　分片数对识别率的影响

Table 2　The effect of the number of fragments on the recognition rate

分片个数	特征帧数目	MMI,CK+/%
4	5	66.2,83.8
5	4	67.6,86.3
7	3	69.7,88.4
15	2	68.2,84.2

实验结果表明,识别率先随着分片数的增加而提高,达到峰值后逐渐下降。这是因为在一个分片中,随着分片数的增加,它拥有帧的片段数越少,保留了更多空间特征,从而提高了识别率;但当分片数目过多时,每片包含的帧数目不足以提供足够的时间信息,从而影响最终的识别效果。

为了更进一步地了解神经网络内部所学习到的特征,本文利用神经网络可视化的方法,将神经网络部分层的学习结果取出,以探究网络到底在学习什么样的特征。如图11所示,可以看出同一个人的相同表情,经过学习之后的特征图,原图更加关注五官轮廓特征,而用Sobel算子处理之后的图像经过网络处理之后,更加关注面部纹理特征,两个网络学习到的特征具有互补性。

最终将两个网络的结果用2.2节提出的方法融合,融合后的结果如图12所示。从图12可以看出,无论是CK+还是MMI,融合后的结果均优于单流网络。

为了探究网络权重选择对识别结果的影响,两个网络的融合权重如图13所示,在CK+和MMI两个库中的权重集中在0.56左右。

为了探究OS-LSTM分片的方法对本文实验的影响,图14所示为不采用OS-LSTM分片,而是直接使用LSTM进行实验对比,可以看出无论是CK+还是MMI,OS-LSTM方法都是优于LSTM的。

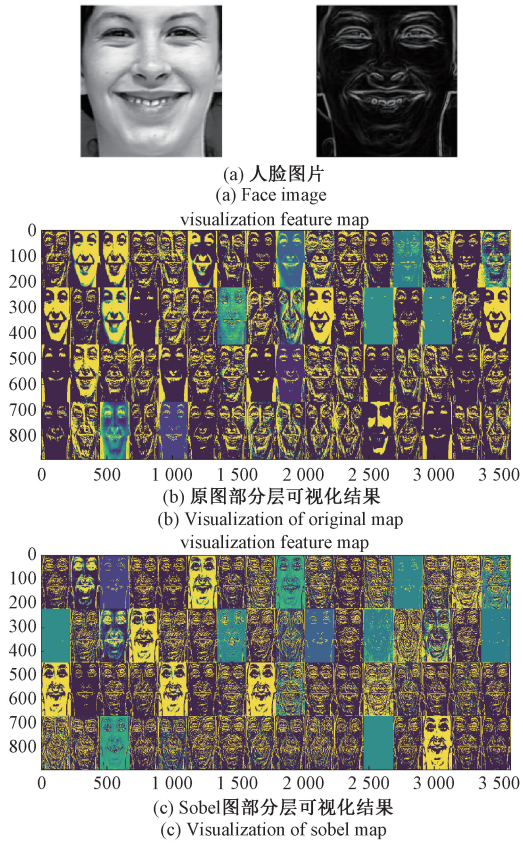


图11　可视化结果展示

Fig. 11　Visualization results display

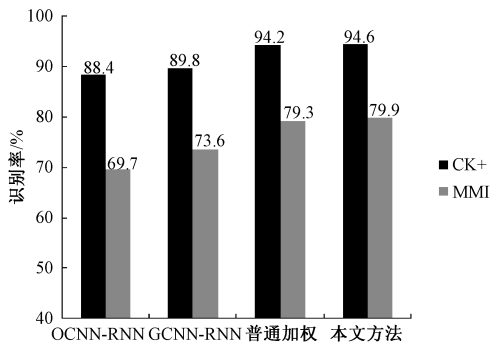


图12　完整模型融合结果

Fig. 12　Fusion results of the integrated network

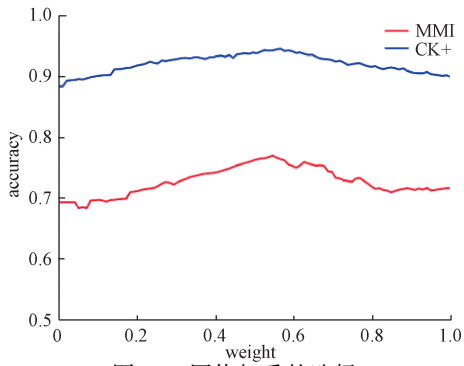


图13　网络权重的选择

Fig. 13　The choice of the network weight selection

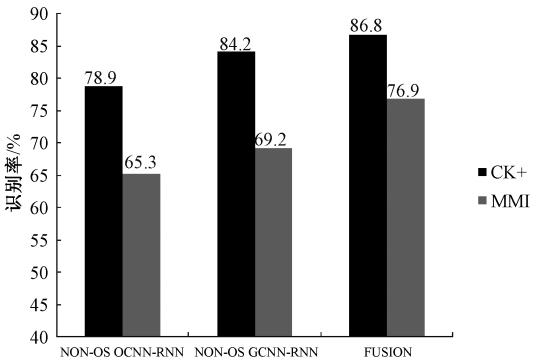


图 14　无分片结构实验结果

Fig. 14　The result of Non overlapping segment

为了验证本文算法的有效性,将本文方法与其他文献所提方法进行了对比,如表 3 和 4 所示。文献[20]融合了 GoogLeNet 和 AlexNet,提出了一种新的网络结构,前面两层是卷积层,中间是 Inception 结构,最后是全连接层,但这种网络结构忽略了样本之间的内在相关性,限制了模型的辨别能力。文献[22]提出了一种基于词典的面部表情分析方法,将表情按运动单元进行分解,但在提取特征时并没有考虑到视频序列中的时间信息,影响了最终实验效果。从表中对比可以看出本文的实验结果在 CK+和 MMI 数据集上取得了一定的优势,体现了本文所提方法的可靠性。

表 3　CK+库上不同文献识别率的比较

Table 3　Comparison of recognition rate on CK+

算法	识别率/%
文献[19]	93.4
文献[20]	93.2
文献[21]	94.0
文献[22]	88.5
文献[23]	93.2
文献[24]	92.5
本文算法	94.6

表 4　MMI 库上不同文献识别率的比较

Table 4　Comparison of the recognition rate on MMI

算法	识别率/%
文献[22]	63.4
文献[25]	75.1
文献[26]	78.5
文献[23]	77.5
文献[24]	74.6
本文算法	79.9

4　结　论

本文提出了一种融合梯度边缘检测和改进的循环神经网络的视频表情识别方法,相较于其他人脸表情识别方法,本文方法表现出来一定的优越性。

一方面利用梯度边缘检测算子处理后的梯度图,能够更清晰地表达人脸的五官信息,更加有效的获取人脸表情的纹理特征。

另一方面,利用交叉分片的方式将 CNN 提取的人脸表情序列送入 LSTM 网络中,避免了在边缘检测时,对视频人脸表情序列时空信息的丢失。最后将两种方法的结果使用所提的算法加权融合,在 CK+和 MMI 视频库上均达到了较高的准确率。

本文选取 CK+和 MMI 人脸表情库作为所提方法的实验评估数据集,因为 CK+和 MMI 人脸表情库均是在实验室特定环境下获得的,视频帧序列受到光照、头部偏转等影响较小,所以如何在类似于 AFEW 等自然条件下获取的数据集取得良好的实验效果,是下一步要进行的研究。

参考文献

[1] 王飞,张莹,张东波,等. 基于捷径的卷积神经网络在人脸表情识别中的应用研究[J]. 电子测量与仪器学报,2018,32(4):80-86.
WANG F, ZHANG Y, ZHANG D B, et al. Research on application of convolutional neural networks in face recognition based on shortcut connection[J]. Journal of Electronic Measurement and Instrumentation, 2018, 32(4): 80-86.

[2] 成翔昊,达飞鹏,邓星. 基于关键点的由粗到精三维人脸特征点定位[J]. 仪器仪表学报,2018,39(10):256-264.
CHENG X H, DA F P, DENG X. Coarse-to-fine 3D facial landmark localization based on keypoints[J]. Chinese Journal of Scientific Instrument, 2018, 39(10): 256-264.

[3] CHEN J, TAKIGUCHI T, ARIKI Y. Rotation-reversal invariant HOG cascade for facial expression recognition[J]. Signal, Image and Video Processing, 2017, 11(8): 1485-1492

[4] CHEN J, SHAN S, HE C, et al. WLD: A Robust Local Image Descriptor[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2010, 32(9): 1705-1720.

[5] ZEILER M D, FERGUS R. Visualzing and understanding convolutional networks[C]. European Conference on Computer Vision, Springer, 2014: 818-833.

[6] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]. Advances in Neural Information

Processing Systems,2012:1097-1105.

[7] SZEGEDY C, LIU N W, JIA N Y, et al. Going deeper with convolutions [C]. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE,2015:1-9.

[8] HE K, ZHANG X, REN S, et al. Deep residual Learning for image recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[9] TRAN D, BOURDEYY L, FERGUS R, et al. Learning spatiotemporal features with 3d convolutional networks [C]. Proceedings of the IEEE International Conference on Computer Vision,2015: 4489-4497.

[10] HOSSEINI S, LEE S H, CHO N I. Feeding hand-crafted features for enhancing the performance of convolutional neural networks [J]. 2018, arXiv: 1801.07848, 2018.

[11] SEVILLA-LARA L, LIAO Y, GUNEY F, et al. On the integration of optical flow and action recognition [C]. German Conference on Pattern Recognition, Springer, 2018:281-297.

[12] 夏清, 张振鑫, 王婷婷, 等. 基于改进Sobel算子的红外图像边缘提取算法[J]. 激光与红外, 2013, 43(10):1158-1161.
XIA Q, ZHANG ZH X, WANG T T, et al. Infrared image edge extraction algorithm based on improved Sobel operator [J]. Laser & Infrared, 2013, 43(10): 1158-1161.

[13] FEICHTENHOFER C, PINZ A, ZISSERMAN A. Convolutional two-stream network fusion for video action recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition,2016:1933-1941

[14] MA C Y, CHEN M H, KIRA Z, et al. TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition [J]. Signal Processing: Image Communication, 2019,71:76-87.

[15] KAYA H, GURPMAR F, SALAH A A. Video-based emotion recognition in the wild using deep transfer learning and score fusion [J]. Image and Vision Computing,2017, 65: 66-75.

[16] KUO C M, LAI S H, SARKIS M. A compact deep learning model for robust facial expression recognition[C]. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2018: 2121-2129.

[17] MARTINEZ B, VALSTAR M F, JIANG B, et al. Automatic analysis of facial actions: A survey[J]. IEEE Transactions on Affective Computing, 2017,13(9):1-22.

[18] FAN Y, LU X, LI D, et al. Video-based emotion recognition using CNN-RNN and C3D hybrid networks[C]. 18th ACM International Conference on Multimodal Interaction, 2016: 445-450.

[19] XIE S, HU H. Facial expression recognition using hierarchical features with deep comprehensive multipatches aggregation convolutional neural networks[J]. IEEE Transactions on Multimedia, 2019, 21(1): 211-220.

[20] MOLLAHOSSENI A, CHAM D, MAHOOR M H. Going deeper in facial expression recognition using deep neural networks[C]. 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), 2016: 1-10.

[21] 徐琳琳,张树美,赵俊莉.构建并行卷积神经网络的表情识别算法[J].中国图象图形学报,2019,24(2):0227-0236.
XU L L, ZHANG SH M, ZHAO J L. Expression recognition algorithm for parallel convolutional neural networks[J]. Journal of Image and Graphics, 2019, 24(2): 0227-0236.

[22] TAHERI S, QIU Q, CHELLAPPA R. Structure-preserving sparse decomposition for facial expression analysis [J]. IEEE Transactions on Image Processing, 2014, 23(8): 3590-3603.

[23] BEHZAD H, MOHAMMAD H. Facial expression recognition using enhanced deep 3D convolutional neural networks[C]. IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017:30-40.

[24] 胡敏,张柯柯,王晓华,等. 结合滑动窗口动态时间规整和CNN的视频人脸表情识别[J]. 中国图象图形学报,2018,23(8): 1144-1153.
HU M, ZHANG K K, WANG X H, et al. Video facial expression recognition combined with sliding window dynamic time warping and CNN[J]. Journal of Image and Graphics,2018,23(8): 1144-1153.

[25] LIU M, SHAM S, WANG R, et al. Learning express-sionlets onspatio-temporal manifold for dynamic facial expression [C]. IEEE Conference on Computer Vision and Pattern Recognition,2014:1749-1756.

[26] MOHAMMADI M R, FATEMIZADEH E, MAHOOR M H. PCA-based dictionary building for accurate facial expression recognition via sparse representation [J]. Journal of Visual CommunicationandImage Representation,2014,25(5):1082-1092.

作者简介

胡敏，2004 年于合肥工业大学获得博士学位，现为合肥工业大学教授、博士生导师，主要研究方向为计算机视觉、情感计算等。

E-mail：uhnim@ 163. com

Hu Min received Ph. D. from HeFei University of Technology in 2004. Now she is a professor and Ph. D. supervisor at Hefei University of Technology. Her main research interest is computer vision, affective computing and so on.

高永，2017 年于安徽理工大学获得学士学位，现为合肥工业大学硕士研究生，主要研究方向是计算机视觉，深度学习等。

E-mail：ygaohfut@ 163. com

Gao Yong received his B. Sc. degree from AnHui University of Scienct and Technology in 2017. He is currently a M. Sc. candidate at Hefei University of Technology. His main research interests include computer vision and deep learning.