2021 年 6 月 第 6 期
Vol. 42 No. 6 2021

用于微表情识别的改进双流浅层卷积神经网络

李昆仑,陈 栋,王 珺,王怡辉

(河北大学 电子信息工程学院,河北 保定 071000)
E-mail: likunlun@ hbu. edu. cn

摘 要: 在微表情自动识别任务中,浅层卷积神经网络和深层网络相比更好地改善了网络训练过拟合的情况,但是多数浅层卷积神经网络存在输入特征单一和提取高维有效特征能力不足的问题. 针对上述问题本文同时使用图像的灰度特征和运动特征表征原图像,并且提出了一种改进双流浅层卷积神经网络(Enhanced Dual-stream Shallow Convolutional Neural Network,EDSS-Net)用于微表情的识别. 本文首先使用欧拉视频放大算法和 TV-L1 光流法对视频关键帧处理,提取图像的灰度特征和运动特征,然后用空洞卷积和注意力模块改进双流浅层卷积网络模型,提高网络提取有效特征的能力,最后将两种特征输入网络训练后进行分类. 理论分析及在 CASME Ⅱ、SMIC-HS 和 SAMM 微表情数据库上的实验结果均表明了改进模型的有效性.

关 键 词: 微表情识别; 双流卷积神经网络; 欧拉视频放大算法; TV-L1 光流法; 空洞卷积; 注意力机制
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2021)06-1219-08

Enhanced Dual-stream Shallow Convolutional Neural Network for Facial Micro-expression Recognition

LI Kun-lun,CHEN Dong,WANG Jun,WANG Yi-hui
(Electronic Information Engineering College,Hebei University,Baoding 071000,China)

Abstract: In micro-expression recognition,comparing with deep neural networks,shallow networks show a better performance in solving the overfitting problems,but most shallow convolutional neural networks have the problems of single input features and insufficient ability to extract high-level effective features. In response to above problems,the algorithm in this paper uses both the grayscale feature and the motion feature of the image to represent the original image. This paper proposes an enhanced dual-stream shallow convolutional neural network(EDSSNet) to identify micro-expression. The algorithm in this paper first uses Euler video magnification algorithm and Total Variation-L1 optical flow method to obtain the gray-scale features and motion features of the image. Then,we use the atrous convolution and attention module to improve the dual-stream shallow convolution network model,which can enhance the ability of the network to extract effective features. Finally,two kinds of features are input into the network for classification. Theoretical analysis and experimental results on the CASME Ⅱ,SMIC-HS and SAMM micro-expression databases prove the effectiveness of the improved model.

Key words: micro-expression recognition; dual-stream convolutional neural network; euler video magnification algorithm; total variation-L1 optical flow method; atrous convolution; attention module

1 引 言

人类面部表情是人们日常生活交流中传递情感的主要手段. 心理学家的研究表明,面部表情所传达的信息量可达人们情感信息传递总量的一半[1]. 微表情是一种短暂的自发式表情,通常持续时间在 1/25s-1/5s 之间,它本质上是人的一种无意识的举动,具有无法隐藏和无法改变的特性,因此它能真实地反应人的情感状态.

微表情的概念是由 Haggard 和 Isaacs 于 1966 年在一次临床访谈中首次提出[2]. 1969 年,Ekman 和 Friesen 在研究中发现了同样的特殊表情,他们将其命名为微表情[3]. 微表情自动识别因其在临床诊断、测谎、教学辅助和商务谈判等领域的良好表现而受到越来越多的关注.

特征提取是微表情自动识别的主要研究内容之一,现有的特征提取方法主要分为基于传统算法的特征提取和基于深度学习的特征提取. 微表情是一段连续的面部运动,对微表情识别需要从三维空间提取有效信息. 文献[4]首先提出了 LBP-TOP(Local Binary Patterns from Three Orthogonal Planes)算法,将 LBP 算法从二维平面扩展到了三维空间. 为了能够更加准确的描述微表情视频中图像间的运动信息,一些研究者开始将光流特征应用于微表情识别,如 Liu 等提出的主方向平均光流特征(Main Directional Mean Optical Flow Feature,MD-MO)和 Liong 等人提出的双权重光流算法(Bi-Weighted Oriented Optical Flow,Bi-WOOF)[5,6]. 还有一些研

收稿日期: 2020-12-08 收修改稿日期: 2021-02-05 基金项目: 国家自然科学基金项目(61672205)资助. 作者简介: 李昆仑,男,1962 年生,博士,教授,CCF 会员,研究方向为模式识别、图像处理、计算机网络等; 陈 栋,男,1996 年生,硕士研究生,研究方向为模式识别、人脸表情识别; 王 珺,男,1988 年生,博士,讲师,研究方向为图像处理、计算机视觉、视频目标跟踪和模式识别; 王怡辉,男,1995 年生,硕士研究生,研究方向为模式识别、图像分类等.

究者针对微表情运动幅度小的问题,如 Li 等则使用欧拉视频放大(Eulerian Video Magnification,EVM)算法将微表情运动放大后再用传统识别方法对微表情进行识别[7].

近期,深度学习在图像分类领域取得了显著的成就.越来越多的学者也将目光从传统算法转移到了深度学习上.2016 年开始有学者用深度学习来解决微表情识别问题,其中有代表性的是 Patel D 等人发表在 ICPR 会议上的文章[8].他们使用在 ImageNet 图像数据库和 CK + 表情数据库上训练好的卷积神经网络模型来提取微表情的深度特征,对提取的深度特征进行特征选择后用以分类,分类准确率高于基线方法 LBP-TOP 算法.不久之后,类似 VGG-16,AlexNet 和 GoogleNet 这些在图像分类中表现较好的深度学习框架被用在微表情自动识别任务中,并且取得了优于传统算法的识别准确率[9-11].但是这些卷积神经网络只能处理单张图片,无法利用视频中图像之间的信息[12].而且传统深度学习网络需要大量的数据用于训练模型,微表情数据量小训练时过拟合现象严重,无法直接用于深度网络训练模型.所以在微表情识别任务中深度学习往往和迁移学习一起使用,但迁移学习不但需要额外在别的数据库上先进行训练,而且训练使用的数据库和微表情数据库间的特征差异大小对最后的训练结果也会产生影响.针对以上问题一些研究者先对微表情数据库原始视频进行预处理提取视频帧间的关联信息,再将处理后的图像直接用于训练浅层网络,获取适合微表情分类的特征.如文献[13]采用了光流图像和双流浅层卷积神经网络的组合,在文献[14]中提出了一种关键帧框架,即只使用视频中的起始帧,关键帧和结束帧获取图像特征,使用光流图像和浅层卷积神经网络对微表情分类,文献[15]中则先用动态成像算法集中原始图像中的运动信息在一张图像上,再使用加入了融合层的浅层网络进行训练分类.以上方法中浅层的卷积神经网络虽然改善了网络训练产生过拟合的情况,但同时网络提取有效特征的能力也下降了,文献[15]注意到了这一点,因此在浅层网络中加入了融合层以提高网络性能.但是在对原始图像提取特征时,以上方法只使用一种特征很难完整的表达原始图像中的有效信息,而直接输入原始图像又会含有大量的冗余信息.

综上所述浅层网络虽然比深度网络可以更好的改善过拟合的情况,并且可以使用微表情数据库直接训练,更有利于提取适合微表情分类的特征,但是浅层网络存在输入网络的特征单一和浅层网络难以有效提取高维特征的问题.针对这些问题:1)本文在微表情预处理阶段,采用关键帧框架,使用欧拉视频放大算法和 TV-L1 光流法生成灰度图像和光流图像替代原图像,预处理后的灰度图像和光流图像和原图像相比不仅含有更少的冗余特征,突出了细节特征,而且包含了视频的时空信息;2)在特征提取阶段,结合空洞卷积和注意力机制,在增强网络对全局特征提取能力的同时尽可能的保留图像中的细节信息.

本文的具体工作为首先利用欧拉视频放大算法和 TV-L1 光流算法对微表情视频的起始帧和关键帧进行预处理,提取放大图像的灰度特征和运动特征;然后利用双流卷积神经网络和空洞卷积核(Atrous Convolutions)组建了新的网络结构,提高网络获取高维信息的能力;在此基础上加入了注意力模块,帮助网络从输入特征中筛选出有效的信息.最后在数据增强后的微表情数据库 CASME Ⅱ(Chinese Academy of Sciences Micro-expression Database Ⅱ),SMIC-HS(Spontaneous Micro-expression Database,High Speed)和 SAMM(Spontaneous Activity and Micro-Movements Dataset)上进行了实验,并将提出的方法与现有的几种方法进行了比较,证明了算法的有效性[16-18].

本文其余部分的内容安排如下:第 2 节介绍了相关的算法包括 EVM 算法,TV-L1 光流算法,空洞卷积原理,注意力机制模块和卷积神经网络结构等内容.第 3 节介绍了本文提出的 EDSSNet(Enhanced Dual-stream Shallow Convolutional Neural Network).第 4 节介绍了实验设置和实验结果对比.第 5 节对论文进行了总结和展望.

2 相关工作

目前微表情自动识别还无法准确识别复杂环境下的微表情,识别的微表情也大多都是单一的表情类别,即每个微表情视频只对应一种情感状态.微表情自动识别任务一般包括图像预处理,特征提取,分类识别 3 部分,本文在图像预处理阶段采用了欧拉视频放大算法和 TV-L1 光流法,前者用于放大微表情的动作幅度,后者用于提取微表情视频中的运动信息.在特征提取和分类阶段使用结合空洞卷积和注意力机制的改进双流卷积神经网络进行特征提取和分类.下面是这些理论的详细介绍.

2.1 欧拉视频放大算法

2012 年,Wu 等人基于流体欧拉法提出了一种欧拉视频放大算法(Eulerian Video Magnification,EVM)[19].流体欧拉法(Euler Method)是描述流场中各空间点流体质点流经时运动状态的一种方法.EVM 算法将视频中的像素看作空间和时间的函数来研究像素和时间的变化关系.EVM 算法可以分为以下 3 个步骤:空间滤波,时域滤波,放大与合成图像.

在空间滤波阶段,EVM 算法使用多分辨率的图像金字塔把目标视频分解成不同的空间频率,然后进行时域滤波提取每个尺度图像中的感兴趣频带,人脸微表情的脸部运动主要包括眉毛、眼睛、鼻子和嘴巴的微小运动,面部的大部分区域都保持不变,这些面部的微小运动大多数都集中在低频区域.接着用泰勒级数近似滤波后的频带,给滤波后的频带乘以一个放大因子 e,放大滤波结果.最后把放大后的信号与原信号重构合成图像.

2.2 TV-L1 光流

光流法是研究动态物体的常用方法,光流图像也适合描述面部肌肉的微观运动[20].TV-L1 光流(Total Variation-L1 Optical Flow)算法适合相邻两帧图像位移量小的运动分析,它保留了图像的边缘特征信息[21].

对于相邻的图片 I_0 和 I_1,$X=(x,y)$ 是帧 I_0 上的像素值.则 TV-L1 光流模型的能量函数如式(1)所示:

$$\begin{aligned} I_1' &= \mu\nabla I_1 + I_1(X+U_0) - U_0\nabla I_1 - I_0 \\ E &= \int_\Omega \{\lambda |I_1'| + |\nabla U|\}\,dx \end{aligned} \tag{1}$$

公式中 $U=(u,v)$ 表示光流场,∇u 和 ∇v 是 u 和 v 的二阶导数,参数 λ 是权重常数.公式(1)中的第 1 项表示相邻两帧对应像素的灰度值差,第 2 项是对运动的正则化约束项.

TV-L1 光流采用双向求解的数值分析机制(Two-way Numerical Analysis Mechanism)最小化光流能量函数,使用一阶泰勒展开公式来近似相邻帧灰度差公式,采用双向求解的方法交替更新变量来减少线性逼近带来的误差:

$$\begin{aligned} I''_1 &= U\nabla I_1 + I_1(X+U_0) - U_0\nabla I_1 - I_0 \\ \mathrm{E} &= \int_\Omega \{\lambda |I''_1| + |\nabla U|\} dx \end{aligned} \tag{2}$$

将 $\rho U = I_1(X+U_0) + (U-U_0)\nabla I_1 - I_0$ 和变量 U' 代入公式(2)中可得公式(3):

$$\mathrm{E} = \int_\Omega \left\{\lambda|\nabla U| + \frac{1}{2\theta}(U-U')^2 + \rho U\right\} dx \tag{3}$$

公式中 θ 是一个常量,U' 随着迭代次数的增加无限逼近 U,公式通过迭代更新 U' 和 U 来优化函数. 它们的阈值函数如下式(4)所示:

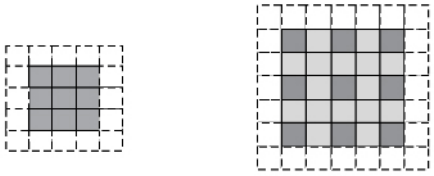
$$U' = U + \begin{cases} \lambda\theta I_1^X, \rho(U) < -\lambda\theta(I_1^X)^2 \\ -\lambda\theta I_1^X, \rho(U) > -\lambda\theta(I_1^X)^2 \\ -\rho U/I_1^X, |\rho(U)| \leqslant \lambda\theta(I_1^X)^2 \end{cases} \tag{4}$$

2.3　空洞卷积

空洞卷积(Atrous Convolutions)又被称为扩张卷积(Dilated Convolutions),它和传统卷积方式相比,在扩大了卷积核的感受野范围同时保证了获取图像的分辨率在要求范围之内,具体操作是在传统的卷积核中增加了权重为 0 的卷积参数,在卷积层加入了新的参数扩张率(Dilation Rate),本文记为 r,该参数表示卷积核进行卷积操作时权重值之间的间距,和传统 $n\times n$ 的卷积核相比,扩张率为 r 的空洞卷积第 i 层卷积层的空洞卷积核大小计算公式如式(5)所示[22]:

$$R = [n + (n-1)\times(r_i - 1)]^2 \tag{5}$$

其中 R 表示实际卷积核的大小,r_i 表示第 i 层的扩张率,n 表示原始卷积核的大小. 本文使用扩张率为 2 的卷积核代替普通的卷积核. 普通卷积感受野和空洞卷积感受野对比如图 1 所示.



(a) 普通卷积核大小为3×3　　(b) 扩张率r=2的空洞卷积核大小为5×5

图 1　卷积核对比图

Fig. 1　Convolution kernel comparison chart

2.4　卷积神经网络注意力模型

卷积神经网络注意力模型(Convolutional Block Attention Module,CBAM)是 Sanghyun Woo 等人提出的一种注意力机制模块[23]. 它在通道和空间两个维度生成注意力映射图,用以提高网络对有效特征的获取能力. 注意力模型总体结构如图 2 所示.

注意力运算过程可用下式表示:

$$\begin{aligned} F' &= M_c(F)\otimes F \\ F'' &= M_s(F')\otimes F' \end{aligned} \tag{6}$$

式(6)中 $F\in R^{C\times H\times W}$ 表示输入特征,$M_C\in R^{C\times 1\times 1}$ 表示通道注意力映射,$M_s\in R^{1\times H\times W}$ 表示空间注意力映射,$\otimes$ 表示对应元素相乘.

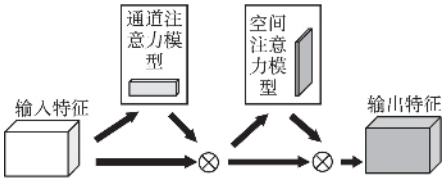


图 2　注意力模型图

Fig. 2　Attention module diagram

通道注意力模块如图 3 所示.

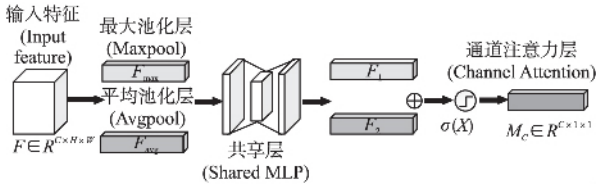


图 3　通道注意力模型图

Fig. 3　Channel attention model diagram

图 3 中共享层由含有一个隐藏层的多层感知器(MLP)组成. 为了简化运算过程,隐藏层的激活大小设置为 $R^{C/a\times 1\times 1}$,其中 a 是缩减率,通道注意力具体计算公式为式(7):

$$\begin{aligned} F_1 &= MLP(F^c_{avg}) = W_1(W_0(F^c_{avg})) \\ F_2 &= MLP(F^c_{max}) = W_1(W_0(F^c_{max})) \\ M_c(F) &= \sigma(F_1 + F_2) \end{aligned} \tag{7}$$

式(7)中 σ 表示 sigmoid 函数,$w_0\in R^{C/a\times C}$,$w_1\in R^{C\times C/a}$. F^c_{avg} 和 F^c_{max} 表示平均池化特征和最大池化特征,多层感知器权重 W_0 后使用 ReLU 作为激活函数.

空间注意力模型计算如式(8):

$$M_s(F) = \sigma(f^{7\times 7}([F^s_{avg}; F^s_{max}])) \tag{8}$$

其中 σ 表示 sigmoid 函数,$f^{7\times 7}$ 表示滤波器尺寸为 7×7 的卷积运算,$F^s_{avg}\in R^{1\times H\times W}$ 和 $F^s_{max}\in R^{1\times H\times W}$,分别表示通道上的平均融合特征和最大融合特征. 空间注意力模型具体运算流程如图 4 所示.

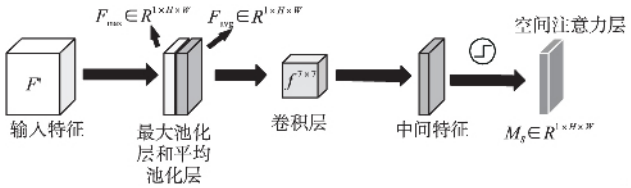


图 4　空间注意力模型图

Fig. 4　Spatial attention model diagram

2.5　双流卷积神经网络

视频识别和静止图像的分类相比,视频提供了图像间随时间变化的信息. 为了能够更好的利用这些信息,文献[24]提出了一种用于视频中动作识别的双流卷积网络(Two-stream Convolutional Networks). 该网络使用了两个独立的卷积网络分别提取图像的空间信息和时间信息,随后融合两个

网络的分类结果作为最终的识别结果. 空间卷积网络(Spatial Convolutional Networks) 使用视频帧进行训练,时间卷积网络(Temporal Convolutional Networks) 使用帧间提取的光流图训练. 网络模型如图 5 所示.

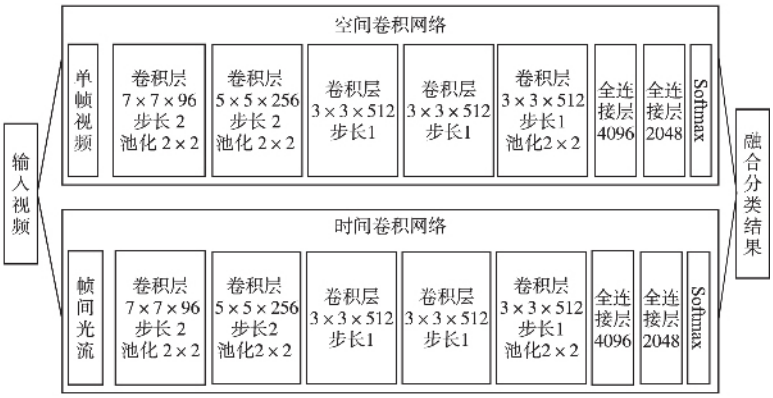


图 5 双流卷积网络

Fig. 5 Two-stream convolutional networks

3 改进双流浅层卷积神经网络

本文提出的改进双流浅层卷积神经网络(Enhanced Dual-stream Shallow Convolutional Neural Network,EDSSNet) 采用了放大灰度图像和光流图像双输入的方式,融合图像的灰度信息和运动信息,在网络中使用了融合层,通过结合两个通道对图像的卷积响应来改善卷积通道中图像细微特征丢失的情况. 利用空洞卷积和注意力机制增加网络提取高维特征和筛选有效特征的能力,最终通过 Softmax 层对图像进行分类. 本文所构建的网络中并没有添加池化层,因为微表情的运动是微小的,池化层的下采样操作容易丢失这些微小的运动信息,而这些信息的丢失会影响分类精度. 本文采用步长为2的卷积来代替池化层的下采样操作,因为使用步长较长的卷积来代替相对应的池化层可以增加提取特征之间的相关性. 同时为了防止网络过拟合,本文在网络中加入了 L2 正则化. EDSSNet 结构如图 6 所示. EDSSNet 用到的网络层介绍如下:

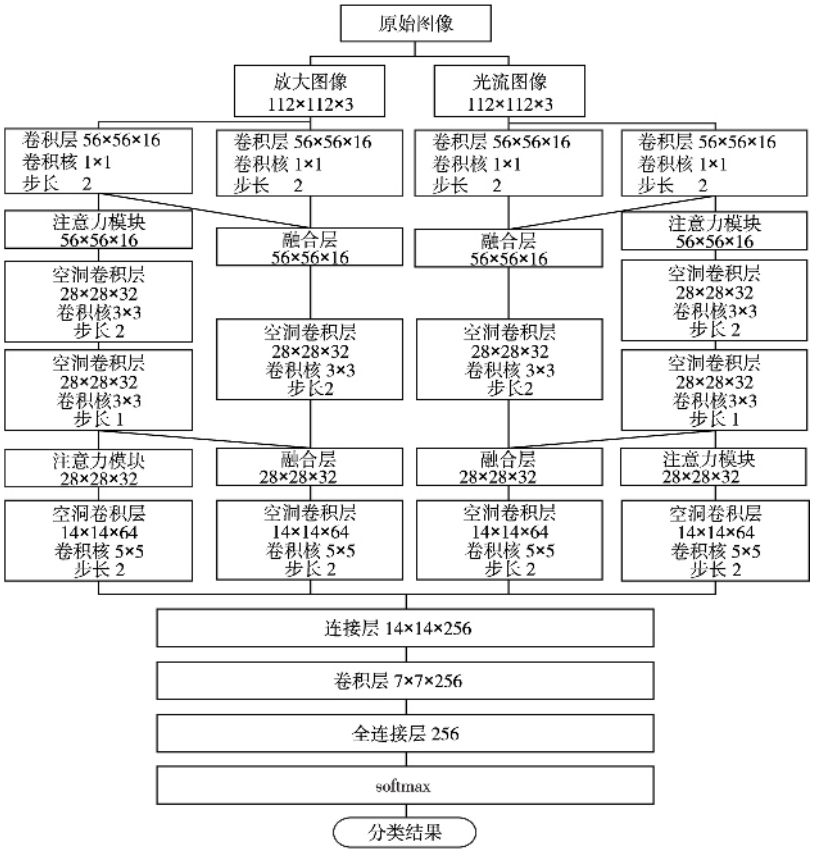


图 6 EDSSNet 框架

Fig. 6 Framework of EDSSNet

卷积层(Convolution Layer) : 卷积层采用卷积核对输入特征进行卷积操作提取特征,通过反馈更新卷积核的学习权重和偏差. 卷积层的计算过程如式(9) 所示:

$$R_M^n = f_w^k \otimes F_M^n + b^k \tag{9}$$

式中 F_M^n 表示输入大小为 $M \times M$ 的图像,用 f_w^k 表示大小为 $k \times k$ 的卷积核,b^k 表示偏差.

激活单元(Rectified Linear Unit) : 激活单元使用单调函数,将线性输入转换为非线性的输入. 本文的网络结构使用 ReLU 激活单元,ReLU 激活函数表示如式(10) :

$$f(\varphi) = max(0, \varphi) \tag{10}$$

ReLU 激活函数与 tanh 和 sigmoid 激活函数相比,改善了梯度消失的情况,而且具有良好的收敛性.

融合层(Accretion Layer) : 融合层结合前面两层的特征,加强网络提取特征的能力,具体公式如式(11) :

$$A_{acc}(x) = A^1(x) + A^2(x) \tag{11}$$

式中 A^1 和 A^2 表示输入卷积层的特征.

连接层(Concatenation Layer) : 本网络中使用连接层将 4 个通道的特征在通道维度融合,保留微表情运动的信息,将其送入到下一层.

批量归一化(Batch Normalization) : 将每层网络的输入值规范化到均值为 0 方差为 1 的标准分布,加快数据的收敛和训练速度.

全连接层(Fully Connected Layer) : 全连接层中的每个神经元都包含上一层全部神经元的信息,神经元的激活是由上一层的输出乘以权重系数 W 再加上偏置决定. 神经元的激活公式如式(12) 所示:

$$f(\alpha) = \psi(W \times \alpha) \tag{12}$$

式(12) 中 α 表示输入的神经元特征,ψ 表示激活函数.

L2 正则化: 正则化是以增大训练误差为代价来减少测试误差的一种策略. 它可以防止模型过拟合. 在神经网络学习中往往决策面越复杂对应的权重值就越大,而 L2 正则化通过减小权重值使得到的决策面更加平缓. 其定义如式(13) 所示:

$$\|x\|_2 = \sqrt{\sum_i x_i^2} \tag{13}$$

L2 正则化便是在使用的损失函数后面加上 L2 范数.

丢失层(Dropout Layer) : 主要用来缓解卷积神经网络训练中出现的过拟合情况,它每次会按照设定好的概率随机的屏蔽神经元,从而降低神经网络对某些特定权重的依赖性,提高网络的泛化能力.

以下是本文算法的详细步骤:

输入: 微表情样本 F,放大系数 e,学习率 η,最大训练次数 n_epochs,每次处理的样本个数 $batch\_size$,检测次数 x,训练停止条件: 达到最大训练次数 n_epochs 或者经过 x 次训练验证集损失不再发生变化.

输出: 微表情样本的平均预测准确率 $\bar{p}$.

Step 1. 计算 EVM 放大后的视频关键帧 G;

Step 2. 计算视频起始帧和关键帧间的光流特征 O;

Step 3. 训练网络模型;

1. for i in range(k) :
2. $k-1$ 份数据分为训练集和验证集,剩余一份为测试集;
3. for *epoch* in range(n_epochs) : or while 经过 x 次训练验证集损失不再变化:
4. 计算误差,用随机梯度下降法更新网络参数;
5. end
6. 对测试集进行类别预测,计算准确率 p_i;
7. end
8. 计算平均准确率 $\bar{p} = \frac{1}{k}\sum_{i=0}^{k-1} p_i$.

4　实验与结果

4.1　数据库介绍

本文选用 CASME Ⅱ,SMIC-HS 和 SAMM 3 个公开的数据库进行实验. CASME Ⅱ数据库是由中国科学院心理研究所建立,本文使用数据库样本中“快乐”、“惊讶”、“厌恶”、“压抑”和“其它”共 5 类进行分类,对于样本数量较少的“恐惧”和“悲伤”两类数据没有采用. SMIC-HS 数据库是由芬兰奥卢大学的赵国英等人建立,主要分为“积极”、“消极”和“惊讶”3 类. SAMM 数据库由曼彻斯特城市大学和情绪智力学院(Manchester Metropolitan University and the Emotional Intelligence Academy) 共同采集. 图像主要分为“高兴”、“惊讶”、“生气”、“厌恶”、“伤心”、“害怕”、“轻蔑”和“其它”共 8 类. 数据库详细参数如表 1 所示.

表 1　用于实验的数据库基本信息

Table 1　Basic information of databases used in experiment

	CASME Ⅱ	SMIC-HS	SAMM
帧速率(fps)	200	100	200
表情类别	7	3	8
视频总数	255	164	159
高兴(Happy)	32	–	26
惊讶(Surprise)	28	43	15
生气(Anger)	–	–	57
厌恶(Disgust)	63	–	9
悲伤(Sad)	4	–	6
恐惧(Fear)	2	–	8
轻蔑(Contempt)	–	–	12
压抑(Repression)	27	–	–
其它(Others)	99	–	26
积极(Positive)	–	51	–
消极(Negative)	–	70	–
起始帧	有标记	有标记	有标记
关键帧	有标记	无标记	有标记
结束帧	有标记	有标记	有标记

4.2　实验设置

在 EVM 算法中,本文在空间滤波阶段采用的是拉普拉斯金字塔,金字塔的层数设为 3 层. 进行时域滤波时本文选用的是巴特沃斯带通滤波器,放大频率段为 1HZ-5HZ. 在信号放大阶段本文根据不同放大系数对图像的影响进行比较,由图 7 中观察可以发现当放大倍数 e 为 10 时,图像的放大效果并不明显,无法清晰地观察到面部的运动信息. 当放大倍数 e 超过 20 时,图像中噪声的影响较大,面部表情甚至会产生失真现象. 本文设置放大倍数 e 为 20.

本文将输入网络的图像统一裁剪为 112×112 大小,选用放大倍数为 20 的图像. 对于 SMIC-HS 数据库中没有标记关

键帧的情况,本文选取视频的中间帧作为关键帧来使用. 卷积神经网络的具体参数如表 2 和表 3 所示,梯度下降采用随机梯度下降算法,初始的学习率设为 0.001,训练次数为 1000,丢失层的概率设为 0.5,L2 正则化的超参数设为 0.01,检查次数为 20. 因为样本数量少,本文采用了数据增强技术对样本进行扩增,采用图像镜像反转和用关键帧前后两帧图像代

(a) 原图 (b) e=10 (c) e=20 (d) e=30 (f) e=40

图 7 不同放大倍数下的图像比较

Fig. 7 Comparison of images at different magnification factors

替关键帧的方法增加数据量,扩增后 CASME Ⅱ 数据库样本数为 996,SMIC-HS 数据库样本数为 656,SAMM 数据库样本数为 636. 本文采用 10 折交叉验证法,即将实验数据分成 10 份,每次将其中的 9 份用于训练,训练的数据以 8∶2 的比例分为训练集和验证集,剩下的 1 份作为测试集,取 10 次试验的准确率平均值和 F1-Score 为最终评价结果.

表 2 注意力模块通道网络结构

Table 2 Attention module channel network structure

	卷积核大小	步长	输出特征大小	扩张率
输入层	–	–	112×112×3	–
卷积层	1×1	2	56×56×16	–
注意力模块	–	–	56×56×16	–
空洞卷积层	3×3	2	28×28×32	2
空洞卷积层	3×3	1	28×28×32	2
注意力模块	–	–	28×28×32	–
空洞卷积层	5×5	2	14×14×64	2
连接层	–	–	14×14×256	–
卷积层	3×3	2	7×7×256	–
全连接层	–	–	1×1×256	–

表 3 融合层通道网络结构

Table 3 Accretion Layer network structure

	卷积核大小	步长	输出特征大小	扩张率
输入层	–	–	112×112×3	
卷积层	1×1	2	56×56×16	–
融合层	–	–	56×56×16	–
空洞卷积层	3×3	2	28×28×32	2
融合层	–	–	28×28×32	–
空洞卷积层	5×5	2	14×14×64	2
连接层	–	–	14×14×256	–
卷积层	3×3	2	7×7×256	–
全连接层	–	–	1×1×256	–

4.3 对比算法简介

本文将提出的算法和 9 种主流算法进行了对比,以下是对比算法的简单介绍:

LBP-TOP[4]:LBP 算法使用中心像素和周围像素的对比来反映图像的局部纹理特征,LBP-TOP 将其扩展到了三维空间,丰富了特征信息,使用 SVM 分类器进行分类.

Bi-WOOF[7]:算法使用视频的起始帧和关键帧计算光流,并将图像分成不重叠的区域,提取不同区域的光流直方图并进行两次加权后连接起来,使用 SVM 分类器进行分类.

AlexNet[10]:输入的图像大小为 227×227×3,网络结构为 8 层,含有局部响应归一化层(Local Response Normalization,LRN)和重叠池化层,使用 Softmax 层分类.

VGG-16[9]:输入图像大小为 224×224×3,网络结构为 16 层,网络采用 3×3 卷积核来代替大的如 5×5 和 7×7 的卷积核,使用 Softmax 层分类.

GoogleNet[11]:输入图像大小为 224×224×3,网络结构为 22 层,网络中采用了 inception 结构来捕捉不同尺度的信息,使用 Softmax 层分类.

SSSN[13]:使用光学图像作为输入特征,用 AlexNet 的一部分网络结构组成浅层网络提取特征信息,使用 Softmax 层分类.

DSSN[13]:将 SSSN 结构由单通道拓展为了双通道,融合两个通道的结果对微表情识别分类.

CapsuleNet[25]:使用胶囊网络对微表情视频的关键帧处理,提取图片整体和部件间的关系特征识别微表情.

GACN[26]:算法将人脸划分为不同的兴趣单元,使用图卷积网络对不同情感建立模型识别微表情.

4.4 实验结果分析

从表 4 中可以看出传统算法中识别率最高的是 Bi-WOOF 算法,本文算法和其相比在 CASME Ⅱ 数据库,SMIC-HS 数据库和 SAMM 数据库识别准确率分别提高了18.12%,5.33% 和 13.6%.

表 4 准确率对比

Table 4 Comparison of accuracy

方法	CASME Ⅱ	SMIC-HS	SAMM
LBP-TOP	0.3968	0.4373	0.3556
Bi-WOOF	0.5789	0.6159	0.5139
VGG-16	0.7100	0.5964	0.4793
GoogleNet	0.6414	0.5511	0.5992
AlexNet	0.7415	0.6373	0.6643
SSSN	0.7119	0.6329	0.5662
DSSN	0.7078	0.6341	0.5735
CapsuleNet	0.7018	0.5877	0.5989
GACN	0.7120	0.6120	0.5231
EDSSNet	0.7601	0.6692	0.6499

卷积神经网络中 VGG-16 是经典的深度学习网络,也是深度学习中经常作为对比基线算法的卷积网络. 本文算法和其相比在 CASME Ⅱ 数据库,SMIC-HS 数据库和 SAMM 数据库识别准确率分别提高了 5.01%,7.28% 和 17.06%.

DSSN 使用了和本文相似的多通道卷积结构,也是基于双流卷积网络的改进网络. 本文算法和其相比在 3 个数据库识别准确率分别提高了 5.23%,3.51% 和 7.64%.

对比算法中识别准确率最高的是 AlexNet,本文算法在 CASMEⅡ数据库和 SMIC-HS 数据库上的表现均优于 AlexNet,在 SAMM 数据库识别准确率略低于 AlexNet. 这是因为 SAMM 数据库数据量最少,类别最多,训练容易产生过拟合现象. AlexNet 和本文算法卷积层数相近,本文方法为了保留更多的

图像细节信息并没有采用池化操作,AlexNet 则在网络中使用了重叠池化和最大池化,网络受过拟合问题影响小一些,所以对 SAMM 数据库识别准确率高于本文算法,但池化操作丢失图像细节信息使得 AlexNet 对数据量相对丰富的 CASMEⅡ数据库和 SMIC-HS 数据库识别准确率低于本文算法.

从实验数据中可以看出,传统算法的平均识别准确率要低于卷积神经网络的平均识别精度,相对于深层的网络结构 VGG-16 和 GoogleNet,浅层的网络 AlexNet,SSSN,DSSN 和本文算法识别准确率更高. 浅层网络比深层网络更适合数据量小的微表情分类.

3 个数据库中算法普遍对 SAMM 数据库的分类准确率是最低的,这是因为 SAMM 数据库数据量少,分类任务又是最复杂的. CASMEⅡ数据库和 SMIC-HS 数据库传统算法和卷积神经网络却有着不同的表现. 对于传统算法,SMIC-HS 数据库是三分类任务,CASMEⅡ数据库是五分类任务. 显然 SMIC-HS 数据库分类任务更简单,对该数据库的识别准确率也就更高. 但是对于卷积神经网络,SMIC-HS 数据库数据量的不足和简单的三分类任务,使得网络在训练过程中容易训练出比实际更复杂的模型,模型的泛化能力下降,从而导致在测试集识别准确率下降. 因此,卷积神经网络在 CASMEⅡ数据库的识别准确率最高.

表 5　F1-Score 对比

Table 5　Comparison of F1-Score

方法	CASMEⅡ	SMIC-HS	SAMM
LBP-TOP	0.5100	0.6025	0.3640
Bi-WOOF	0.6125	0.6110	0.3970
VGG-16	0.4862	0.5025	0.2911
GoogleNet	0.5367	0.5123	0.4921
AlexNet	0.6601	0.6013	0.4200
SSSN	0.7100	0.6329	0.4500
DSSN	0.7300	0.6461	0.4640
CapsuleNet	0.7068	0.5820	0.5909
GACN	0.7255	0.6023	0.5210
EDSSNet	0.7347	0.6591	0.6084

从表 5 中可以看出本文算法在 F1-Score 指标上比传统算法 Bi-WOOF 在 CASMEⅡ数据库,SMIC-HS 数据库和 SAMM 数据库上分别高了 0.1222,0.0481 和 0.2177. 本文算法与传统卷积神经网络算法 VGG-16,GoogleNet 和 AlexNet 相比,比它们在 CASMEⅡ数据库,SMIC-HS 数据库和 SAMM 数据库上的 F1-Score 最高值分别高了 0.0746,0.0578 和0.1163. 本文算法和 2019 年后发表的新的微表情识别算法 SSSN,DSSN,CapsuleNet 与 GACN 相比,本文算法也取得了比较好的效果. 因为本文采用的浅层网络结构和可以使用微表情数据库直接训练的优势使得本文算法面对类别数量少的数据也有一定的分类准确率,从而提高了 F1-Score 评价指标.

本文还给出了 EDSSNet 在各个数据库的分类混淆矩阵,在图 8 中以 CASMEⅡ的混淆矩阵为例,第 1 行第 1 列数据为样本标签类别为“高兴”,同时被本文方法分类为“高兴”的样本所占数据比例,第 1 行第 2 列为样本标签类别为“高兴”,但被本文方法分类为“厌恶”的样本数据比例.

从图 8 中可以看到在 SAMM 数据库上算法的准确率最低,因为 SAMM 数据库的平均类别数据量是最少的,这些数据难以训练出有效的模型. 在 SAMM 数据库中,对“生气”的识别率是最高的,这和“生气”类别的样本在数据库中占的比例最高有一定的关系. 而且在对别的类别进行分类时,如果分错类别,分为“生气”类别的可能性也更高. “伤心”类别的样本少,对该类别的识别率也最低. SMIC-HS 数据库和 CASMEⅡ数据库的分类结果也存在相同的问题.

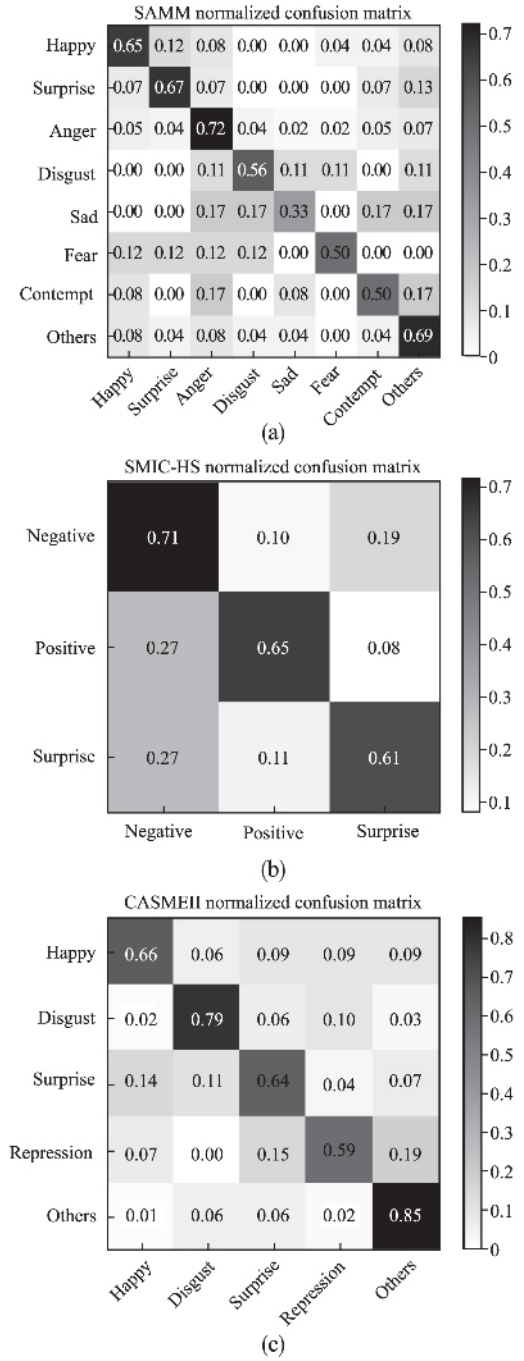


图 8　数据库 SAMM,SMIC-HS 和 CASMEⅡ的混淆矩阵

Fig. 8　Confusion matrix of SAMM,SMIC-HS and CASMEⅡ

综上所述,可以看出样本数据的不平衡对分类精度会产生影响,样本数据多的,网络越有可能学到有效特征,分类的准确率就高.

5 总 结

本文提出了一种用于微表情识别的改进双流卷积神经网络 EDSSNet,并且针对单一特征难以表达微表情运动和浅层网络提取特征能力不足的问题,提出以下改进: 1) 采用欧拉视频放大算法对图像进行放大并提取关键帧特征,同时采用 TV-L1 光流法提取视频关键帧和起始帧的光流特征,用放大后的灰度图像和光流图像替代原始图像; 2) 在双流浅层卷积神经网络中引入空洞卷积和注意力机制,提高网络获取有效特征的能力.

本文算法将图像的灰度信息和运动信息相结合,通过注意力机制引导网络获取有效的信息,空洞卷积和融合层提高网络对高维信息的获取能力. 理论分析和实验结果表明了改进模型的有效性. 本文下一步将针对微表情数据库样本数量的不均衡问题作进一步的研究.

References:

[1] Darwin C. The expression of the emotions in man and animals: frontmatter [J]. Portable Darwin, 2013, 123(1): 146-147.

[2] E A Haggard, K S Isaacs. Micromomentary facial expressions as indicators of ego mechanisms in psychotherapy [M]. Methods of Research in Psychotherapy, Springer US, 1966: 154-165.

[3] Ekman P, Friesen W V. Nonverbal leakage and clues to deception [J]. Psychiatry-interpersonal & Biological Processes, 1969, 32(1): 88-106.

[4] Pfister T, Li X B, Zhao G Y, et al. Recognising spontaneous facial micro-expressions [C] //Proceedings of the 2011 IEEE International Conference on Computer Vision, Barcelona, Spain, IEEE, 2011: 1449-1456.

[5] Liu Y J, Zhang J K, Yan W J, et al. A main directional mean optical flow feature for spontaneous micro-expression recognition [J]. IEEE Transactions on Affective Computing, 2016, 7(4): 299-310.

[6] Liong S T, See J, Wong K S, et al. Less is more: micro-expression recognition from video using apex frame [J]. Signal Processing: Image Communication, 2018, 62(3): 82-92.

[7] Li X, Hong X, Moilanen A, et al. Towards reading hidden emotions: a comparative study of spontaneous micro-expression spotting and recognition methods [J]. IEEE Transactions on Affective Computing, 2017, 9(4): 563-577.

[8] Patel D, Hong X, Zhao G. Selective deep features for micro-expression recognition [C] //International Conference on Pattern Recognition, IEEE, 2017: 2258-2263.

[9] Li Y, Huang X, Zhao G. Can micro-expression be recognized based on single apex frame [C] //2018 25th IEEE International Conference on Image Processing(ICIP), IEEE, 2018: 3094-3098.

[10] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks [J]. Communications of the ACM, 2017, 60(6): 84-90.

[11] Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions [C] //2015 IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2015: 1-9.

[12] Su-Jing W, Bing-Jun L, Yong-Jin L, et al. Micro-expression recognition with small sample size by transferring long-term convolutional neural network [J]. Neurocomputing, 2018, 312(42): 251-262.

[13] Khor H Q, See J, Liong S T, et al. Dual-stream shallow networks for facial micro-expression recognition [C] //2019 IEEE International Conference on Image Processing(ICIP), IEEE, 2019: 36-40.

[14] Liong S T, Gan Y S, Yau W C, et al. Off-apexnet on micro-expression recognition system [J]. arXiv preprint arXiv: 1805.08699, 2018.

[15] Verma M, Vipparthi S K, Singh G, et al. LEARNet: dynamic imaging network for micro expression recognition [J]. IEEE Transactions on Image Processing, 2019, 29(9): 1618-1627.

[16] Yan W J, Li X, Wang S J, et al. CASME II: an improved spontaneous micro-expression database and the baseline evaluation [J]. PloS one, 2014, 9(1): 1-8.

[17] Li X, Pfister T, Huang X, et al. A spontaneous micro-expression data-base: inducement, collection and baseline [C] //10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition(FG), IEEE, 2013: 1-6.

[18] Davison A K, Lansley C, Costen N, et al. SAMM: a spontaneous micro-facial movement dataset [J]. IEEE Transactions on Affect, Computing, 2018, 9(1): 116-129.

[19] Wu H Y, Rubinstein M, Shih E, et al. Eulerian video magnification for revealing subtle changes in the world [J]. Acm Transactions on Graphics, 2012, 31(4): 65-83.

[20] Li Q, Yu J, Kurihara T, et al. Deep convolutional neural network with optical flow for facial micro-expression recognition [J]. Journal of Circuits, Systems, and Computers, 2020, 29(1): 1-7.

[21] Zach C, Pock T, Bischof H. A duality based approach for realtime TV-L1 optical flow [C] //Proceedings of the 29th DAGM Conference on Pattern Recognition, 2007, 9: 214-223.

[22] Qiao S, Chen L C, Yuille A. DetectoRS: detecting objects with recursive feature pyramid and switchable atrous convolution [J]. arXiv preprint arXiv: 2006.02334, 2020.

[23] Woo S, Park J, Lee J Y, et al. Cbam: convolutional block attention module [C] //Proceedings of the European Conference on Computer Vision(ECCV), 2018: 3-19.

[24] Simonyan K, Zisserman A. Two-stream convolutional networks for action recognition in videos [C] //Advances in Neural Information Processing Systems, 2014: 568-576.

[25] Nguyen Van Quang, Jinhee Chun, Takeshi Tokuyama. Capsulenet for micro-expression recognition [C] //2019 14th IEEE International Conference on Automatic Face & Gesture Recognition(FG 2019), Lille, France, IEEE, 2019: 1-7.

[26] Xie H X, Lo L, Shuai H H, et al. AU-assisted graph attention convolutional network for micro-expression recognition [C] //Proceedings of the 28th ACM International Conference on Multimedia, 2020: 2871-2880.