·图形图像处理· 文章编号：1000-3428(2022)01-0228-08 文献标志码：A 中图分类号：TP391.41

结合迁移学习与可分离三维卷积的微表情识别方法

梁正友[1,2]，刘德志[1]，孙 宇[1]
（1.广西大学 计算机与电子信息学院，南宁 530004；2.广西多媒体通信与网络技术重点实验室，南宁 530004）

摘 要：针对现有微表情自动识别方法准确率较低及微表情样本数量不足的问题，提出一种融合迁移学习技术与可分离三维卷积神经网络(S3D CNN)的微表情识别方法。通过光流法提取宏表情和微表情视频样本的光流特征帧序列，利用宏表情样本的光流特征帧序列对S3D CNN进行预训练，并采用微表情样本的光流特征帧序列微调模型参数。S3D CNN网络由二维空域卷积层及添加一维时域卷积层的可分离三维卷积层构成，比传统的三维卷积神经网络具有更好的学习能力，且减少了模型所需的训练参数和计算量。在此基础上，采用迁移学习的方式对模型进行训练，以缓解微表情样本数量过少造成的模型过拟合问题，提升模型的学习效率。实验结果表明，所提方法在CASME II微表情数据集上的识别准确率为67.58%，高于MagGA、C3DEvol等前沿的微表情识别算法。

关键词：微表情识别；深度学习；卷积神经网络；迁移学习；光流法

开放科学（资源服务）标志码（OSID）：

中文引用格式：梁正友，刘德志，孙宇.结合迁移学习与可分离三维卷积的微表情识别方法[J].计算机工程，2022，48(1)：228-235.

英文引用格式：LIANG Z Y，LIU D Z，SUN Y.Micro-expression recognition method combining transfer learning and separable 3D convolution[J].Computer Engineering，2022，48(1)：228-235.

Micro-Expression Recognition Method Combining Transfer Learning and Separable 3D Convolution

LIANG Zhengyou[1,2]，LIU Dezhi[1]，SUN Yu[1]
（1.School of Computer and Electronics Information，Guangxi University，Nanning 530004，China；
2.Guangxi Key Laboratory of Multimedia Communications and Network Technology，Nanning 530004，China）

【Abstract】The existing automatic micro-expression recognition methods are limited in accuracy，and suffer from inadequate micro-expression samples.To address the problem，a micro-expression recognition method that combines transfer learning and a Separable 3D Convolutional Neural Network（S3D CNN）is proposed.The optical flow method is used to extract the feature frame sequences of optical flow from macro-expression and micro-expression video samples.The sequence extracted from macro-expression samples is used to pre-train the S3D CNN，and the sequence extracted from micro-expression samples is used to tune the model parameters.S3D CNN consists of separable 3D convolutional layers，which are composed by 2D spatial convolutional layers and 1D time-domain convolutional layers，so S3D CNN can provide better learning ability than traditional 3D CNN with fewer required parameters and calculations for model training.Furthermore，transfer learning is used to train the model，so the over-fitting problem of the model caused by inadequate micro-expression samples can be alleviated，and the learning efficiency of the model can be improved.Experimental results on the CASME II micro-expression dataset show that the recognition accuracy of the proposed method reaches 67.58%，higher than MagGA，C3DEvol and other advanced algorithms.

【Key words】micro-expression recognition；deep learning；convolutional neural networks；transfer learning；optical flow method

DOI：10. 19678/j. issn. 1000-3428. 0059771

基金项目：国家自然科学基金（61763002）。

作者简介：梁正友（1968—），男，教授、博士，主研方向为计算机视觉、无线传感器网络、人工智能；刘德志，硕士研究生；孙 宇，讲师、博士。

收稿日期：2020-10-20 **修回日期：**2021-01-15 **E-mail：**zhyliang@gxu.edu.cn

0　概述

根据表情持续时间的长短和运动强度的大小，可以将表情分成宏表情和微表情两类。宏表情的持续时间约为2~3 s，运动涉及整个面部区域。目前，研究人员已经利用计算机实现了接近100%的宏表情识别率[1]。与宏表情相比，微表情的运动时间相对较短，仅为0.5 s左右[2]，且运动强度非常微弱，通常只涉及局部的面部区域。这些特点导致微表情特征提取较为困难，使当前微表情自动识别的准确率远低于宏表情。但微表情是由内心真实情绪激发所产生，难以抑制或伪造，比宏表情更能准确地反映人内心的真情实感，因此能够作为测谎的重要依据。

近年来，随着深度学习相关技术的迅速发展，具有时空特征提取能力的三维卷积神经网络（3D Convolutional Neural Networks，3D CNN）[3-4]在视频分类任务中的效果优于仅能提取空域特征的二维卷积神经网络。受到这些成果的鼓励，一些研究人员开始尝试利用3D CNN来提取微表情的时空特征，从而提高识别准确率。文献[5]设计的3D-FCNN通过三流结构的3D CNN同时提取微表情原始视频帧序列和光流帧序列的时空特征，在全连接层对三流提取到的时空特征进行融合。但由于样本数量过少导致的过拟合问题，准确率的提升有限。文献[6]利用3D CNN提取眼睛、嘴部等微表情运动较为频繁部位的时空特征，减少了无关区域对算法的影响。文献[7]提出一种双流结构的3D CNN，使模型能够提取包含在微表情光流帧序列中的时空特征，同时增强了模型对不同帧率样本的适应性，与STCLQ[8]、MDMO[9]等手工特征方法相比，准确率提高了约10%。文献[10]利用具有全局搜索及优化能力的遗传算法对3D CNN的结构和参数进行编码、选择、交叉、变异等操作，从而得到适用于微表情识别任务的最佳参数组合和模型结构，提高了模型的识别能力。但微表情运动强度较弱，相邻两帧之间的差异非常小，在原始的微表情视频帧序列中提取用于分类的时空特征难度较大。

迁移学习是一种常用于解决由于样本数量过少导致模型在训练过程当中出现过拟合现象的方法。迁移学习首先在拥有大量样本的源任务上对模型进行预训练，然后用目标任务的少量样本对预训练获得的模型参数进行微调，以找到源任务与目标任务间能够共享的模型参数，并使模型更加适用于目标任务[11]。近年来，一些研究人员开始采用迁移学习的方法解决当前微表情样本数量过少导致的模型过拟合问题。文献[12]利用遗传算法，从VGG-Net学习到的宏表情分类特征中筛选出适用于微表情分类的特征进行识别。文献[13]在利用ResNet10提取微表情特征前，先用大量的宏表情样本对模型进行预训练，有效地提升了模型在小规模的微表情数据集上的表现。为解决微表情运动强度较弱的问题，文献[14]首先采用欧拉视频运动放大算法（Eulerian Video Magnification，EVM）[15]对微表情进行运动放大，然后利用能够进行人脸识别的模型VGG-Face来提取微表情运动强度峰值帧的特征，并用于分类。但如果EVM算法的放大倍数过大容易产生伪影，对算法的识别准确率造成一定影响。

由于自发式微表情的采集难度较大，可用于研究的样本数量较少，导致当前采用3D CNN来提取时空特征进行微表情识别的研究普遍存在样本数量过少造成的过拟合问题，对识别准确率造成一定的影响，而当前采用迁移学习技术进行微表情识别的研究，通常仅利用二维卷积神经网络来提取静态微表情图像的空域特征，并未考虑微表情变化过程中的动态时域特征，准确率提升较为有限。

针对上述问题，本文提出一种结合迁移学习和可分离三维卷积神经网络（Separable 3D Convolutional Neural Networks，S3D CNN）的微表情自动识别方法。利用光流法提取每一个宏表情和微表情视频帧序列相邻两帧的水平、垂直方向光流图，并导出对应的光流应变模式图。将3个光流图以通道叠加的方式构成光流特征图后，按时间顺序将光流特征图连接成光流特征帧序列。此外，利用宏表情样本的光流特征帧序列对S3D CNN进行预训练，使模型获得与表情分类相关的时空特征，从而缓解传统的3D CNN训练参数较多、所需计算量较大的问题。在此基础上，将预训练得到的模型参数迁移至用于微表情识别的模型中，利用微表情样本的光流特征帧序列对模型参数进行微调，从而使模型更加适用于微表情识别任务。

1　可分离三维卷积

可分离三维卷积[16-18]是近年来出现的一种介于二维卷积与三维卷积之间的轻量化时空特征提取方法，其原理是利用二维空域卷积加一维时域卷积来模拟三维卷积的时空特征提取过程。传统3D CNN的三维卷积核大小为$k\times k\times t$，其中：k为卷积核空域维度的高度和宽度；t为卷积核时域维度的长度。设3D CNN的三维卷积层输出的特征映射(i,j,z)处的值为$y_{i,j,z}$则：

$$y_{i,j,z}=f\left[\sum_{n}\sum_{a=1}^{k}\sum_{b=1}^{k}\sum_{c=1}^{t}w_{a,b,c}^{l}x_{(i+a),(j+b),(z+c)}^{l}+b\right] \tag{1}$$

其中：$f(x)$为三维卷积层的激活函数；n为上层传入的特征映射中包含的帧数；$w_{a,b,c}^{l}$为第l个卷积层的三维卷积核(a,b,c)处的权重值；$x_{i,j,z}^{l}$为输入第l个卷积层的特征映射(i,j,z)处的元素值；b为偏置值。

可分离三维卷积将三维卷积拆分成了二维的空域卷积和一维的时域卷积2个独立的过程，如图1所

示。利用卷积核大小为$k\times k\times 1$的二维空域卷积层来提取输入帧序列的空域特征，计算公式如下：

$$y_{i,j,z}=f_s\left[\sum_{n}\sum_{a=1}^{k}\sum_{b=1}^{k}w^{l}_{a,b,1}x^{l}_{(i+a)(j+b),z}+b_s\right] \tag{2}$$

其中：$y_{i,j,z}$为空域卷积层输出的特征映射(i,j,z)处的元素值；$f_s(x)$为空域卷积层的激活函数；b_s为空域卷积层的偏置值。

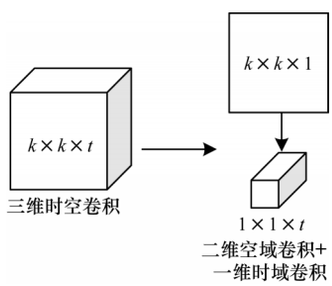


图1 可分离三维卷积原理

Fig.1 Separable 3D convolution principle

将卷积结果输入卷积核大小为$1\times 1\times t$的一维时域卷层中，提取帧与帧之间的时域特征，计算公式如下：

$$a_{i,j,z}=f_t\left[\sum_{n}\sum_{c=1}^{t}w^{l}_{1,1,c}y^{l}_{i,j,(z+c)}+b_t\right] \tag{3}$$

其中：$f_t(x)$为时域卷积层的激活函数；b_t为时域卷积层的偏置值；$a_{i,j,z}$为输出特征映射(i,j,z)处的时域卷积结果。

与2D CNN相比，由可分离三维卷积层构建的S3D CNN只在每个二维空域卷积层之后增加了一个一维时域卷积层，用于提取视频帧序列的时域特征，但模型的训练参数与所需计算量并未显著增加。与3D CNN相比，将三维卷积层拆分成二维空域卷积层和一维时域卷积层后，2个卷积层间增加了1个额外的激活函数，使模型比3D CNN能更好地拟合非线性函数，增强了模型的学习能力。

2 本文方法

本文所提方法首先需要对原始的宏表情与微表情视频帧序列进行预处理；然后利用光流法对每个视频帧序列的相邻两帧进行运动估计，提取相邻2帧的光流特征图来组成光流特征帧序列；最后采用迁移学习的方法对S3D CNN进行训练。主要流程如图2所示。

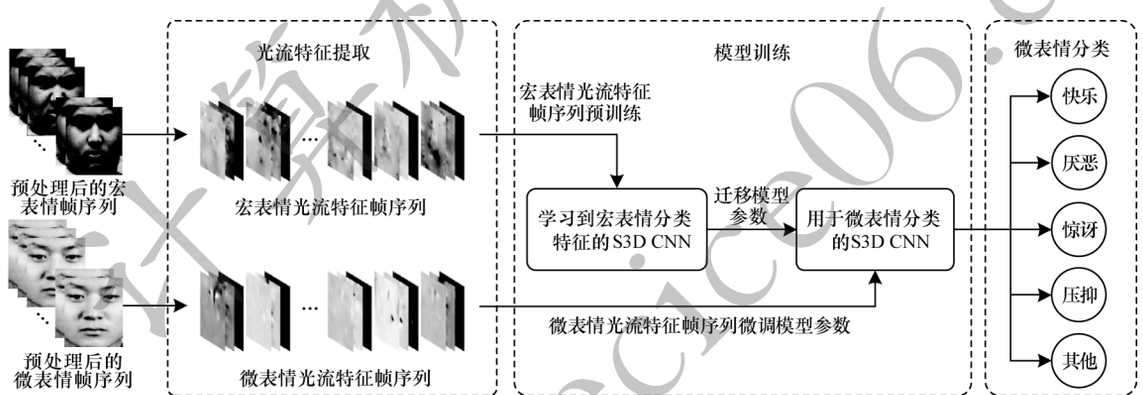


图2 本文所提微表情识别方法流程

Fig.2 Flowchart of micro-expression recognition method proposed in this paper

2.1 预处理

为减少微表情运动过于微弱对光流特征提取和模型学习效果的影响，在预处理过程中，首先通过EVM算法将每个微表情样本的面部运动放大10倍。然后利用时域插值模型(Temporal Interpolation Model，TIM)[19]将每个宏表情和微表情视频帧序列归一化为11帧，以满足输入S3D CNN的样本帧数必须一致的要求。最后，采用平移裁剪和随机旋转等数据增强方式获取更多的宏表情样本，增强模型的鲁棒性，并通过类别重采样来避免宏表情和微表情数据集的样本类别分布不均衡对模型学习效果的影响，具体步骤如下：

1)利用OpenCV的Dlib库来检测每个宏表情和微表情样本第1帧的68个面部特征点，根据最左侧、最顶部、最右侧和最下侧共4个特征点的坐标确定面部矩形区域。

2)将步骤1)中确定的面部矩形区域分别向上、下、左、右、左上、右上、左下、右下共8个方向平移10个像素，并按照平移前后的面部矩形区将样本第1帧和剩余帧的面部区域均裁剪下来，使每个宏表情和微表情样本能够获得9个视频帧序列，从而将样本数量扩充9倍。

3)对数据增强后的宏表情数据集进行类别重采样，从每个宏表情类别中随机抽取1 500个样本，共7×1 500=10 500个样本组成新的宏表情数据集，并将抽取到的样本随机旋转0°、90°、180°或270°，以增加样本的多样性。

4)将按照步骤1)确定的面部矩形区域裁剪的微表情样本作为测试集，并从步骤2)平移后的微表情样本中以类别重采样的方式随机抽取训练集样本。在类别重采样过程中，每个微表情类别分别抽取50个样本，共5×50=250个微表情样本构成新的训练

集，与原数据集的样本量近似。

2.2　光流特征提取

光流特征提取是对每个宏表情和微表情样本的相邻2帧进行运动估计，提取高层次的面部表情运动特征。根据文献[20]的实验结果可知，相对于其他光流法，TL-V1光流法[21]的鲁棒性较好，更加适用于微表情识别任务，因此本文采用TL-V1光流法对宏表情和微表情进行运动估计。光流法基于亮度恒定原则估计视频中的运动物体。设(d_x, d_y)为图像上某个像素点在d_t时间后的下一帧移动的距离，由亮度恒定原则可以认为这2个像素的值不变，即：

$$I(x,y,t)=I(x+d_x,y+d_y,t+d_t) \tag{4}$$

上述方程称为光流方程。光流法的目的是利用光流方程求出图像上每个像素运动的大小和方向矢量$\boldsymbol{p}=\left[p=\frac{d_x}{d_t},q=\frac{d_y}{d_t}\right]^{\mathrm{T}}$。

此外，本文进一步利用宏表情和微表情相邻2帧的光流场来导出对应的光流应变模式。应变模式用于衡量物体在外力作用下的形变程度。设$\boldsymbol{u}=[u,v]^{\mathrm{T}}$表示三维空间中面部表情形变导致的位移在二维图像上的投影向量，则可用柯西张量来表示表情发生过程中面部肌肉组织的形变程度：

$$\boldsymbol{\varepsilon}=\frac{1}{2}[\nabla\boldsymbol{u}+(\nabla\boldsymbol{u})^{\mathrm{T}}] \tag{5}$$

其中：∇表示对$\boldsymbol{u}$进行求导。可以将式(5)的二维应变张量展开成矩阵形式：

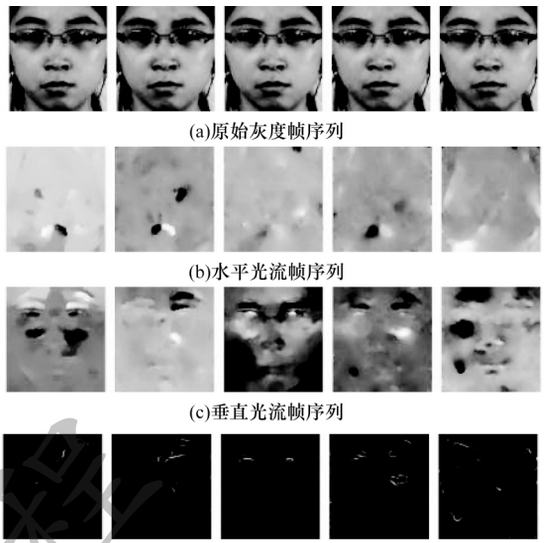
$$\boldsymbol{\varepsilon}=\begin{bmatrix}\varepsilon_{xx}=\dfrac{\partial u}{\partial x} & \varepsilon_{xy}=\dfrac{1}{2}\left(\dfrac{\partial u}{\partial y}+\dfrac{\partial v}{\partial x}\right)\\ \varepsilon_{yx}=\dfrac{1}{2}\left(\dfrac{\partial v}{\partial x}+\dfrac{\partial u}{\partial y}\right) & \varepsilon_{yy}=\dfrac{\partial v}{\partial y}\end{bmatrix} \tag{6}$$

由于表情发生过程中的肌肉运动可能包含多个方向，因此采用应变模式的4个分量来计算每个像素的应变大小，如式(7)所示：

$$\varepsilon_m=\sqrt{\varepsilon_{xx}^2+\varepsilon_{yy}^2+\varepsilon_{xy}^2+\varepsilon_{yx}^2} \tag{7}$$

应变模式具有仅与物体表面的形变有关，不易受到光照条件等外部因素影响的优点[22]，鲁棒性较强，在微表情识别任务中有较好的表现[23]。

在提取光流特征之后，每个宏表情和微表情样本能得到3个光流帧序列，即水平、垂直方向光流帧序列和光流应变模式帧序列。图3展示了CASME II微表情数据集其中1个样本的原始灰度帧序列和对应的3个光流帧序列。将3个光流帧序列中相对应的每一帧以通道叠加的方式连接起来，构成三通道的光流特征帧序列。在预处理过程中每个宏表情和微表情视频帧序列均用TIM算法归一化为11帧，因此1个光流特征帧序列共包含10帧反映原样本相邻两帧之间面部运动和形变情况的光流特征图。将光流特征帧序列的空域维度调整至96×96并进行标准化处理后，输入模型进行训练。



(a)原始灰度帧序列　(b)水平光流帧序列　(c)垂直光流帧序列　(d)光流应变模式帧序列

图3　微表情原始灰度帧序列和光流帧序列

Fig.3　Original gray frame sequence and optical flow frame sequence of micro-expression

2.3　模型设计

本文用于特征提取和分类的S3D CNN主要由8个可分离三维卷积层、4个池化层和1个全连接层组成，模型结构和主要参数如表1所示。表中的Conv_s_i和Conv_t_i分别表示第i个可分离三维卷积层的空域卷积层和时域卷积层，空域卷积层用于提取视频帧序列的静态空域特征，而时域卷积层则对帧与帧之间的动态时域特征进行编码。

表1　S3D CNN参数设置

Table1　S3D CNN parameter settings

层名	输出大小	核大小	卷积核数量	是否参与微调
Conv_s_1	96×96×10×16	(3,3,1)	16	否
Conv_t_1	96×96×10×16	(1,1,3)	16	否
Conv_s_2	96×96×10×16	(3,3,1)	16	否
Conv_t_2	96×96×10×16	(1,1,3)	16	否
Max Pooling 1	48×48×10×16	(2,2,1)	—	—
Conv_s_3	48×48×10×32	(3,3,1)	32	否
Conv_t_3	48×48×10×32	(1,1,3)	32	否
Conv_s_4	48×48×10×32	(3,3,1)	32	否
Conv_t_4	48×48×10×32	(1,1,3)	32	否
Max Pooling 2	24×24×10×32	(2,2,1)	—	—
Conv_s_5	24×24×10×64	(3,3,1)	64	否
Conv_t_5	24×24×10×64	(1,1,3)	64	否
Conv_s_6	24×24×10×64	(3,3,1)	64	否
Conv_t_6	24×24×10×64	(1,1,3)	64	否
Max Pooling 3	12×12×10×64	(2,2,1)	—	—
Conv_s_7	12×12×10×128	(3,3,1)	128	是
Conv_t_7	12×12×10×128	(1,1,3)	128	是
Conv_s_8	12×12×10×128	(3,3,1)	128	是
Conv_t_8	12×12×10×128	(1,1,3)	128	是
Average Pooling	6×6×5×128	(2,2,2)	—	—
Dense Layer	128×1	—	—	是
Dropout Layer	128×1	—	—	—

采用4个池化层对特征映射进行特征降维,以减少冗余信息。其中前3个池化层采用最大池化,即通过保留池化窗口内最大元素的方式进行特征降维,从而突出重要的特征。最后1个池化层采用平均池化,使池化窗口内的每个元素均对降维结果产生影响,防止损失过多的高维特征。由于样本的帧数较少,为更好地保留时域特征,仅在平均池化层采用三维时空池化,而在最大池化层采用二维的空域池化。为充分利用从预训练过程中学习到的宏表情分类特征,本文在微调过程中冻结了部分卷积层的参数,被冻结的卷积层参数在微调过程中保持不变,仅以较低的学习率对余下的卷积层和全连接层进行调整,使模型更加适用于微表情识别任务。表1中"是否参与微调"一列表示模型中对应的卷积层是否参与了微表情的微调训练,即该层在微调训练中是否被冻结。

本文在全连接层后加入了丢弃率为0.2的Dropout层。Dropout层能以一定概率使某个神经元的激活值失效,避免模型依赖某些局部特征,以增强模型的泛化性,并缓解模型的过拟合问题。最后,将Dropout层输出的特征送入Softmax层中完成分类。

3 实验

3.1 表情数据集

3.1.1 Cohn-Kanade扩展数据集

Cohn-Kanade扩展数据集(CK+)[24]常用于人脸宏表情识别研究,样本形式为动态的视频帧序列。CK+收集了123个受试者的593个宏表情样本,其中327个样本带有情感类型标签。该数据集将宏表情分为7个类别,各个类别的样本数量分别为愤怒45、蔑视18、厌恶59、恐惧25、快乐69、悲伤28和惊讶83。

3.1.2 CASME II微表情数据集

CASME II微表情数据集[25]由来自26名受试者的246段视频样本组成。样本的帧率为200 frame/s,分辨率为640×480,平均帧长为68帧。每个样本根据诱导材料内容、受试者的自我报告等信息被分成5个类别,各个类别的样本数量分别为快乐32、厌恶60、惊讶25、压抑27和其他102。CASME II还提供了每个样本的起始帧、峰值帧和结束帧位置。

3.2 模型训练

本文采用迁移学习的方法对所设计的S3D CNN进行训练,具体步骤如下:

1)利用从宏表情样本中提取的光流特征帧序列对S3D CNN进行预训练。预训练的学习率初始化为0.000 1,迭代周期为80,每迭代20个周期学习率下降10倍,batch_size=20。迭代80个周期后,模型在训练集上对7种宏表情的识别准确率达到95.73%。

2)将预训练获得的卷积层和全连接层参数迁移至用于微表情分类任务的模型中,并按照表1中"是否参与微调"一列所示冻结部分卷积层参数后,利用微表情样本提取到的光流特征帧序列对模型参数进行微调。此外,为使模型输出的判别向量维度与CASME II数据集的类别数相同,还需将预训练模型中具有7个输出单元的Softmax层替换成一个新的具有5个输出单元的Softmax层。微调的学习率初始化为0.000 01,迭代周期为40,每迭代10个周期学习率下降10倍,batch_size=10。

3.3 实验环境与评价指标

本文主要采用留一受试交叉验证(Leave-One-Subject-Out,LOSO)对算法性能进行评估。每一轮交叉验证将1名受试者的样本作为测试集,通过式(8)计算LOSO准确率:

$$A_{\text{acc}}=\frac{1}{k}\sum_{i=1}^{k}A_{\text{acc}_i} \tag{8}$$

其中:k为受试者数量。CSAME II微表情数据集包含26名受试者的微表情样本,因此需要执行26轮验证,即k=26。A_{acc_i}为第i轮验证的准确率。

实验的操作系统环境为Centos6.5,利用Keras2.3.1完成模型的搭建,编程语言为Python3.6,模型训练的主要硬件设备为NVIDIA TESLA T4。

3.4 实验结果与分析

3.4.1 与前沿方法的对比

将所提方法的LOSO准确率与现有的手工特征识别方法及深度学习识别方法进行对比,如表2所示。与当前较为前沿的STLBP-IP[26]、LBP-TOP[25]等手工特征识别方法相比,深度学习识别方法能避免繁琐的手工特征提取过程,直接从原始的微表情视频帧序列中提取特征,并通过学习的方式不断调整模型参数,以优化所提取的分类特征,在简化特征提取步骤的基础上取得更好的识别效果。本文所提S3D CNN-transfer微表情识别方法结合了近年来新兴的可分离三维卷积和迁移学习技术,使模型能够同时提取光流特征帧序列中的微表情静态空域特征和动态时域特征。此外,通过迁移学习技术避免微表情样本数量过少造成的过拟合问题,使利用深度学习的方法进行微表情识别的准确率有了进一步的提升。

表2 不同微表情识别方法准确率对比

Table 2 Accuracy comparison of different micro-expression recognition methods %

方法类别	方法	准确率
手工特征识别方法	STCLQP[8]	58.39
	LBP-TOP[25]	63.41
	STLBP-IP[26]	59.51
深度学习识别方法	3D-FCNN[5]	59.11
	C3DEvol[10]	63.71
	SDF[12]	47.30
	MagGA[14]	63.30
	S3D CNN-transfer(本文)	67.58

3.4.2　迁移学习对模型学习效果的影响

为了探索迁移学习对模型学习效果的影响,本节对迁移学习与非迁移学习模型的识别准确率进行了对比,如图4所示。S3D CNN通过随机赋值的方式对模型参数进行初始化。由于需要从0开始学习与表情分类相关的特征,因此该模型与迁移学习的预训练一样从学习率为0.000 1开始训练,迭代周期为40,每迭代10个周期学习率下降10倍。S3D CNN-transfer则遵循了3.2小节的迁移学习方法和参数进行训练。

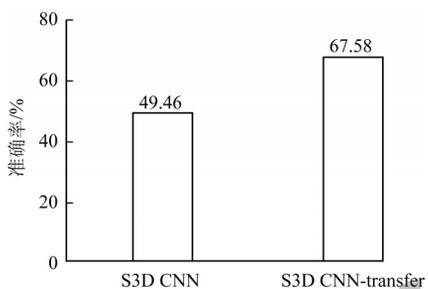


图4　S3D CNN与S3D CNN-transfer方法识别准确率比较

Fig.4　Comparison of recognition accuracy between S3D CNN and S3D CNN-transfer method

由图4可知,S3D CNN的识别准确率仅为49.46%,而采用迁移学习方法进行训练的S3D CNN-transfer则有效避免了直接利用小规模数据集训练深度神经网络出现的过拟合问题,准确率达到67.58%,提升了18.12个百分点。

本节还探究了迁移学习对模型学习效率的影响。如图5所示为S3D CNN和S3D CNN-transfer两个模型在LOSO验证中的其中2轮验证结果的对比。在图5(a)所示的第6轮验证中可以看出,S3D CNN-transfer在迭代10个周期左右训练准确率趋于稳定,模型开始收敛。而S3D CNN需要迭代20~25个周期才收敛,且波动相对较大。在图5(b)所示的第16轮验证中,S3D CNN-transfer在迭代11个周期左右就开始达到收敛状态,而S3D CNN在迭代15~20个周期左右才逐渐达到收敛状态。在其它轮次的验证中也存在类似的情况。从图中还可以看出,S3D CNN-transfer的训练准确率远远高于采用随机赋值进行初始化的S3D CNN。表3对2个模型在26轮交叉验证中的平均收敛周期、训练40个周期后的平均训练准确率和平均训练损失进行了对比。从表3中可以看出,由于采用迁移学习方法进行训练的S3D CNN-transfer在预训练阶段已经学习到了一些与表情分类相关的特征,因此在微调阶段实现了更好的学习效果,在26轮交叉验证中的平均收敛周期和平均训练损失分别降低了9.27和0.36;平均训练准确率提高了27.81个百分点。由表3可知,采用迁移学习的方法后,模型能够利用在宏表情分类任务中学习到的模型参数来提高在微表情分类任务中的学习效率,加快了模型的收敛速度,提升了模型的学习效果。

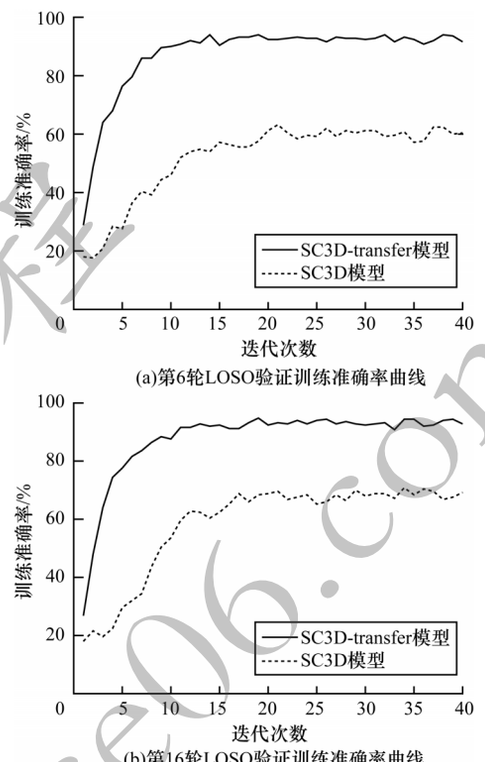


图5　S3D CNN与S3D CNN-transfer模型训练准确率对比

Fig.5　Comparison of training accuracy between S3D CNN and S3D CNN-transfer model

表3　S3D CNN与S3D CNN-transfer模型训练效果对比

Table 3　Comparison of training effect between S3D CNN and S3D CNN-transfer model

模型	平均收敛周期	平均训练准确率/%	平均训练损失
S3D CNN	20.73	64.35	0.818
S3D CNN-transfer	11.46	92.16	0.458

3.4.3　不同模型识别效果比较

本节将所提出的S3D CNN-transfer与2D CNN-transfer和3D CNN-transfer两个模型的训练参数数量、每秒浮点数运算次数(Floating Point Operations,FLOPs)和LOSO准确率进行了比较,如表4所示。2D CNN-transfer将S3D CNN-transfer中的一维时域卷积层全部删除,仅保留了二维空域卷积层用于提取光流特征帧序列的空域特征。3D CNN-transfer将S3D CNN-transfer中的可分离三维卷积层全部替换成了传统的三维卷积层。3个模型均采用了3.2节的迁移学习方法和参数进行训练。

表4 2D CNN-transfer、3D CNN-transfer与S3D CNN-transfer方法识别效果对比

Table 4 Comparison of recognition effect between 2D CNN-Transfer, 3D CNN-Transfer and S3D CNN-Transfer method

模 型	参数数量/10^6	浮点运算数/10^9 FLOPs	准确率/%
2D CNN-transfer	3.24	2.43	58.27
3D CNN-transfer	3.83	7.26	64.68
S3D CNN-transfer	3.37	3.56	67.58

通过对比表中的数据可知,由于2D CNN-transfer无法捕获表示微表情动态变化的时域特征,仅能通过空域特征识别,准确率低于能够同时提取空域特征和时域特征的3D CNN-transfer和S3D CNN-transfer。3D CNN-transfer虽然能够提取微表情的时空特征,实现了比2D CNN-transfer更好的识别效果,但增加了较多的训练参数,且模型所需的计算量也大幅增加。S3D CNN-transfer利用二维空域卷积加一维时域卷积的方式来模拟3D CNN-transfer的三维卷积过程,使模型与3D CNN-transfer一样具有时空特征提取能力,而训练参数和计算量比采用传统三维卷积的3D CNN-transfer更少。此外,在S3D CNN-transfer的二维空域卷积层和一维时域卷积层之间增加的激活函数有效提升了模型的学习能力,因此准确率稍高于3D CNN-transfer。

4 结束语

现有的微表情识别方法准确率较低,且由于微表情样本数量不足导致了过拟合问题。本文提出一种结合迁移学习与S3D CNN的微表情自动识别方法,提取包含宏表情和微表情运动与形变特征的光流特征帧序列,并根据迁移学习的方法,利用宏表情的光流特征帧序列对S3D CNN进行预训练。在此基础上,通过使用微表情的光流特征帧序列微调预训练后的模型参数,有效提升微表情自动识别的准确率。实验结果表明,所提方法相比于MagGA、C3DEvol等前沿微表情识别算法,具有更高的识别准确率。但光流法仍然存在计算量较大、算法较为复杂、实时性和实用性较差等问题。下一步将在保证识别准确率的前提下,通过降低算法的复杂度、减少运行所需时间和计算资源,使该方法能更好地满足实时应用及在复杂场景下的应用需求。

参考文献

[1] KHARAT G U, DUDUL S V. Emotion recognition from facial expression using neural networks[J]. Springer, 2009, 60(3): 207-219.

[2] YAN W J, WU Q, LIANG J, et al. How fast are the leaked facial expressions: the duration of micro-expressions[J]. Journal of Nonverbal Behavior, 2013, 37(4): 217-230.

[3] CARREIRA J, ZISSERMAN A. Quo vadis, action recognition? a new model and the kinetics dataset[C]// Proceedings of IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C., USA: IEEE Press, 2017: 6299-6308.

[4] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Washington D. C., USA: IEEE Press, 2015: 4489-4497.

[5] LI J, WANG Y D, SEE J, et al. Micro-expression recognition based on 3D flow convolutional neural network[J]. Pattern Analysis and Applications, 2019, 22(4): 1331-1339.

[6] REDDY S P T, KARRI S T, DUBEY S R, et al. Spontaneous facial micro-expression recognition using 3D spatiotemporal convolutional neural networks[C]// Proceedings of 2019 International Joint Conference on Neural Networks. Washington D. C., USA: IEEE Press, 2019: 1-8.

[7] PENG M, WANG C Y, CHEN T, et al. Dual temporal scale convolutional neural network for micro-expression recognition[J]. Frontiers in Psychology, 2017, 8(3): 1745-1745.

[8] HUANG X H, ZHAO G Y, HONG X P, et al. Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns[J]. Neurocomputing, 2016, 175(12): 564-578.

[9] 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.

[10] 梁正友,何景琳,孙宇. 一种用于微表情自动识别的三维卷积神经网络进化方法[J]. 计算机科学,2020,47(8):227-232.
LIANG Z Y, HE J L, SUN Y. Three-dimensional convolutional neural network evolution method for facial micro-expression auto recognition[J]. Computer Science, 2020, 47(8): 227-232. (in Chinese)

[11] WEISS K, KHOSHGOFTAAR T M, WANG D D. A survey of transfer learning[J]. Journal of Big Data, 2016, 3(1): 1-40.

[12] PATEL D, HONG X P, ZHAO G Y. Selective deep features for micro-expression recognition[C]//Proceedings of the 23rd International Conference on Pattern Recognition. Washington D. C., USA: IEEE Press, 2016: 2258-2263.

[13] PENG M, WU Z, ZHANG Z H, et al. From macro to micro expression recognition: deep learning on small datasets using transfer learning[C]//Proceedings of the13th IEEE International Conference on Automatic Face & Gesture Recognition. Washington D. C., USA: IEEE Press, 2018: 657-661.

[14] LI Y T, HUANG X H, ZHAO G Y. Can micro-expression be recognized based on single apex frame?[C]// Proceedings of the 25th IEEE International Conference on Image Processing. Washington D. C., USA: IEEE Press, 2018: 3094-3098.

[15] 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):1-8.

[16] TRAN D,WANG H,TORRESANI L,et al. A closer look at spatiotemporal convolutions for action recognition[C]// Proceedings of 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2018:6450-6459.

[17] QIU Z F, YAO T, MEI T. Learning spatio-temporal representation with pseudo-3D residual networks[C]// Proceedings of IEEE International Conference on Computer Vision. Washington D. C. ,USA:IEEE Press,2017:5533-5541.

[18] XIE S N,SUN C,HUANG JONATHAN,et al. Rethinking spatiotemporal feature learning:speed-accuracy trade-offs in video classification [C]//Proceedings of European Conference on Computer Vision. Berlin, Germany: Springer,2018:318-335.

[19] ZHOU Z H,ZHAO G Y,PIETIKÄINEN M. Towards a practical lipreading system[C]//Proceedings of 2011 IEEE Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA:IEEE Press,2011:137-144.

[20] LIONG S T,GAN Y S,ZHENG D N,et al. Evaluation of the spatio-temporal features and GAN for micro-expression recognition system [J]. Journal of Signal Processing Systems,2020,23(3):1-21.

[21] ZACH C,POCK T,BISCHOF H. A duality based approach for realtime TV-L1 optical flow[C]///Proceedings of Joint Pattern Recognition Symposium. Berlin, Germany: Springer,2007:214-223.

[22] SHREVE M,GODAVARTHY S,MANOHAR V,et al. Towards macro-and micro-expression spotting in video using strain patterns[C]//Proceedings of Workshop on Applications of Computer Vision. Washington D. C. ,USA: IEEE Press,2009:1-6.

[23] LIONG S T,SEE J,PHAN R C,et al. Subtle expression recognition using optical strain weighted features[C]// Proceedings of the 12th Asian Conference on Computer Vision. Berlin,Germany:Springer,2014:644-657.

[24] LUCEY P,COHN J F,KANADE T,et al. The extended Cohn-Kanade dataset (CK+):a complete dataset for action unit and emotion-specified expression[C]//Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Washington D. C. ,USA: IEEE Press,2010:94-10.

[25] YAN W J,LI X B,WANG S J,et al. CASME II:an improved spontaneous micro-expression database and the baseline evaluation[J]. Plos One,2014,9(4):1-8.

[26] HUANG X H,WANG S J,ZHAO G Y,et al. Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection[C]//Proceedings of 2015 IEEE International Conference on Computer Vision. Washington D. C. ,USA:IEEE Press,2015:1-9.

编辑　赖玉玲