第 59 卷 第 16 期/2022 年 8 月/激光与光电子学进展 研究论文

激光与光电子学进展 先进成像

基于注意力机制的视频眼震图分类算法研究

周浩军，赵晓丽*，高永彬，李海波，程若然

上海工程技术大学电子电气工程学院，上海 201600

摘要 现有的良性阵发性位置性眩晕视频眼震图分类算法存在以下不足：人工提取的特征主观性和局限性强；眼球的轴向转动特征提取困难；仅能区分正常人群和患者，或对简单的眼震进行分类。针对上述问题，提出了一种基于注意力机制的视频眼震图分类算法。以轻量级模型三维 MobileNet V2 为基础网络进行特征提取，在全局细节特征、时空信息丰富的网络低层引入全局时空注意力模块，融合眼球震颤空间信息和帧间时序信息；在网络高层引入时空通道注意力机制，筛选高级语义特征；采用带有类别调制系数的交叉熵损失函数对网络进行训练，有效缓解了类别数量不平衡的问题。在复旦大学附属眼耳鼻喉科医院提供的包括 66 种类别的视频眼震图数据集上进行了实验，所提算法的分类准确度达到 90.08%，各类别的平均精准度、召回率、F1-score 分别为 90.50%，92.00%，90.40%，表明了所提算法的优越性。

关键词 医用光学；图像处理；医学图像处理；视频眼震图分类；时空注意力机制；良性阵发性位置性眩晕；三维卷积神经网络

中图分类号 TP391.4 **文献标志码** A **DOI**：10.3788/LOP202259.1617001

Video Nystagmus Classification Algorithm Based on Attention Mechanism

Zhou Haojun, Zhao Xiaoli*, Gao Yongbin, Li Haibo, Cheng Ruoran

School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201600, *China*

Abstract The existing classification algorithms for benign paroxysmal positional vertigo video nystagmus have the following shortcomings. The features extracted manually are subjective and limited; the feature extraction of axial rotation of eyeballs is difficult; it can only distinguish between normal people and patients or classify simple nystagmus. To overcome the above shortcomings, a video nystagmus classification algorithm based on attention mechanism is proposed. Based on the lightweight model three-dimensional MobileNet V2, a network is used for feature extraction, and the global spatiotemporal attention module is introduced at the lower level of the network with rich global detail features and spatiotemporal information to integrate the spatial information of nystagmus and the temporal information between frames. The attention mechanism of the spatiotemporal channel is introduced to the high-level network to screen high-level semantic features. The cross entropy loss function with category modulation coefficient is used to train the network, which effectively alleviates the problem of imbalance in several categories. Experiments were conducted on 66 types of video nystagmus datasets provided by the Eye and ENT Hospital of Fudan University. The classification accuracy of the proposed algorithm reached 90.08%, and the average accuracy, recall, and F1-score of each category were 90.50%, 92.00%, and 90.40%, respectively, indicating the superiority of the proposed algorithm.

Key words medical optics; image processing; medical image processing; video nystagmus classification; spatiotemporal attention mechanism; benign paroxysmal positional vertigo; three-dimensional convolutional neural network

1 引 言

相关研究表明，近年来我国的耳科和神经科门诊中，良性阵发性位置性眩晕（BPPV）的发病率高，且呈逐年上升趋势，终身患病率为 2.4%[1-3]。持续的未经治疗的 BPPV 会严重影响患者的日常生活，BPPV 患者眼球震颤的特征分析对 BPPV 的诊断至关重要，临床诊治中一般通过位置实验来获取眼震信息[4-5]。视

收稿日期：2021-05-07；**修回日期**：2021-06-09；**录用日期**：2021-07-13

基金项目：国家自然科学基金（61772328）

通信作者：*evawhy@163.com

频眼震图技术(VNG)可以显著提升BPPV眼震的检出率[6-8]。诊断时,医疗人员需逐个观看VNG视频,分析眼震的方向、频率、强度、持续时间和强弱变化等特征,这对医生的要求较高,非常容易受到医生经验、主观判断和疲劳程度的影响。随着计算机技术、深度学习的兴起[9],对比人工诊断方式,计算机辅助的诊断判断客观、效率高、临床应用价值高[10]。

为提取眼球运动特征,文献[11]和文献[12]采用圆形算子来检测瞳孔位置,但是对理想瞳孔边缘和噪声点之间的分割精度较低。文献[13]采用马氏距离技术消除噪声,得到较好的瞳孔提取结果,但对于眼球轴向转动的检测效果不佳。机器学习出现以后,针对传统算法的弊端,文献[14]从患者眼震视频中提取时间、频率特征和旋转角波形等特征,将其输入卷积神经网络(CNN)进行特征分类,并判断受试者所患疾病。文献[15]提出了一种基于机器学习的眼球震颤角位移矢量特征分析方法,该方法通过Fuzzy C-Means (FCM)聚类算法对前庭疾病进行分类。为了选择更好的特征,文献[16]使用Fisher线性判别器分析法选择更有效果的特征,并输入多层感知机获得患者和健康测试者的分类结果。文献[17]将传统算法提取出来的尺寸为3×10的特征矩阵作为CNN的输入,该方法可以区分在水平、竖直和轴向方向的典型眼球震颤,但对于轴向转动的判断误差敏感性较大。文献[18]在分割瞳孔并提取眼球旋转角度后,基于Empirical Mode Decomposition (EMD)方法进行特征提取,最后使用深度神经网络对正常测试者和患者进行分类。文献[19]使用四个测试程序和一个瞳孔跟踪程序来实现眼震特征的选择和分类,再通过线性判别分析进行缩减,最后使用稀疏表示法将其分为三种前庭疾病和正常病例。文献[20]提出了一种基于CNN的眼震识别方法,该方法引入Hough变换和轨迹追踪提升网络的鲁棒性,为了利用运动眼球的光流信息,还采用了基于扭转感知的双流识别网络,可以对扭转眼震进行识别,但识别精度不高。

传统方法只能向医生提供精度不高的瞳孔跟踪数据,现有的基于机器学习和深度学习的方法也存在着以下不足:1)均依靠人工提取特征,再将特征输入算法中进行分类,具有较强的主观性和局限性;2)对眼球轴向旋转矢量特征提取困难;3)仅可以区分患者和正常人群,或者对简单的眼球震颤模式进行分类,但通常情况下临床诊断中存在的是多种类型复合的球震。

针对以上问题,本文提出了一种基于改进的3D MobileNet V2[21]的良性阵发性位置性眩晕视频眼震图分类算法。首先,用特征提取能力强大的3D CNN提取特征,解决了人工提取的特征具有主观性和局限性的问题;其次,引入全局时空注意力机制对全局时空特征施加权重,采用时空通道注意力机制对网络高层的特征通道进行重标定,有效增强了对眼球运动特征的提取能力;最后,用带有类别调制系数的交叉熵损失函数对算法进行优化,从而更好地对临床诊断中存在的多种类型复合的眼震进行分类。

2 相关工作

2.1 MobileNet V2

MobileNet V2是在MobileNet V1[22]基础上改进的轻量级卷积神经网络模型,MobileNet V1将传统卷积方式改为深度可分离卷积,大幅减小了模型参数和卷积过程的运算量,卷积核深度为1,输入特征矩阵的通道数等于卷积核数量和输出特征矩阵通道数;同时,增加控制卷积层中卷积核数量的超参数α和控制输入图像尺寸的超参数β。MobileNet V2在初代的基础上,引入了Inverted Residual Block,如图1所示,其中左图表示卷积步长为1时的卷积过程,一般会有输入和输出相加的残差结构,右图表示卷积步长为2时的卷积过程,没有残差结构,直接输出。卷积核尺寸的变化为1×1→3×3→1×1,两端使用1×1卷积对特征图进行升维,中间采用3×3的深度可分离卷积对特征图进行降维,大幅减小了模型参数。为了减少Relu激活函数对低维特征信息造成的损失,前2次卷积使用Relu6[23]激活函数代替Relu激活函数,第3次卷积使用Linear激活函数。在卷积步长为1时,跳跃联接可以有效缓解因网络层数过深导致的网络退化现象,提高了分类的准确率。

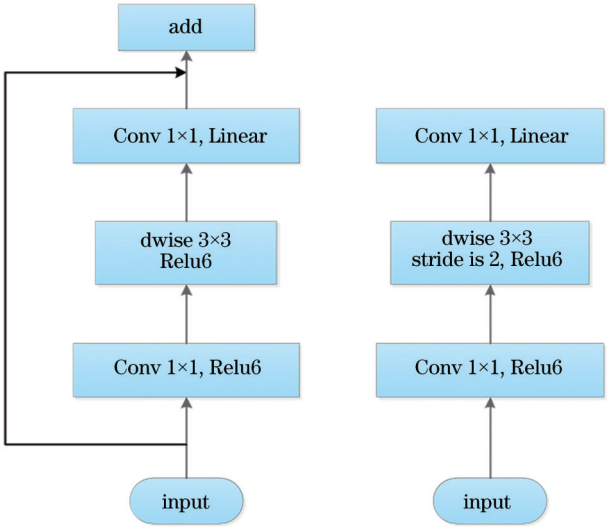

图1 MobileNet V2不同步长下的卷积过程

Fig. 1 Convolution process of Mobilenet V2 under different strides

2.2 全局时空注意力机制

注意力机制[24]来源于人类的视觉注意力,在深度学习领域有着非常重大的影响,其主要思想是对不同的特征施加不同的权重从而得到新的特征组合,简而言之就是能够让模型更关注对当前任务重要的特征信

息。文献[25]认为卷积和循环网络都是对局部区域进行操作，不能够捕捉长距离依赖关系，受计算机视觉中经典的非局部均值(non-local means)的启发，提出了一种非局部操作(non-local operation)，该操作可以建立图像上某个像素点和其他所有的像素点之间的联系，也可以建立视频中的同一帧中不同像素点之间的联系、不同帧中的所有像素点之间的联系。

Non-local operation可以描述为

$$\boldsymbol{y}_i=\frac{1}{C(\boldsymbol{x})}\sum_{\forall j}f(\boldsymbol{x}_i,\boldsymbol{x}_j)g(\boldsymbol{x}_j),\tag{1}$$

$$f(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left[\theta(\boldsymbol{x}_i)^{\mathrm{T}}\phi(\boldsymbol{x}_j)\right],\tag{2}$$

$$C(\boldsymbol{x})=\sum_{\forall j}f(\boldsymbol{x}_i,\boldsymbol{x}_j),\tag{3}$$

式中：$\boldsymbol{x}$是输入数据；$\boldsymbol{y}$是输出响应；i和j表示位置索引；函数f用来计算$\boldsymbol{x}_i$和$\boldsymbol{x}_j$之间的相互联系，其中$\theta(\boldsymbol{x}_i)$和$\phi(\boldsymbol{x}_j)$表示线性变换，T表示转置；线性函数g则用来计算在位置j处输入数据的表示；$C(\boldsymbol{x})$表示标准化因子。计算i处响应时，考虑了所有位置特征$\boldsymbol{x}_j$的加权——这些位置可以是有空间联系、时间联系、时空联系的像素点。

图2是非局部块在卷积神经网络中的结构示意图[24]。对于输入特征向量$\boldsymbol{X}$，高和宽分别是H和W，T表示视频帧数，1024代表通道数量。最左边的分支是跳跃联接，右边三条分支分别是Query、Key和Value支路，分别通过三个卷积变换后将通道数降为512，再进行维度重塑。Query和Key支路通过相似度函数计算得到每帧中每个像素点对其他所有帧中所有像素点的权重关系。接着使用Softmax函数对这些权重进行归一化，然后利用归一化后的权重对Value支路进行权重分配，得到的特征再经过一次卷积操作后融合原始输入$\boldsymbol{X}$，得到最终的输出$\boldsymbol{Z}$。

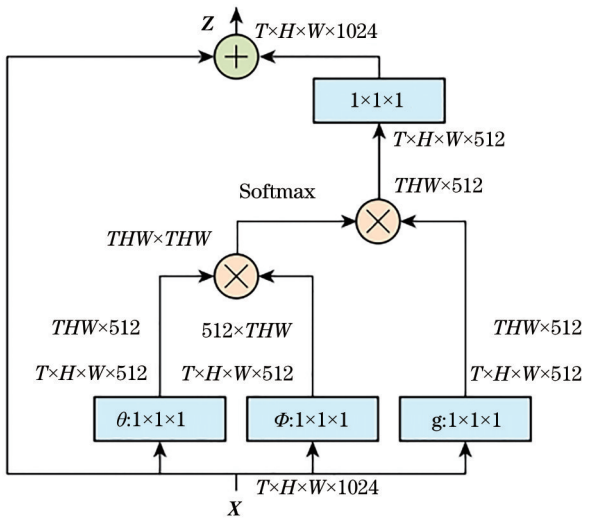

图2　非局部块[24]

Fig. 2　Non-local Block[24]

2.3　通道注意力机制

图3是Hu等[26]提出的Squeeze and Excitation Block(SE Block)的结构示意图，SE Block从通道维度引入注意力机制，获取每个特征通道的重要程度权重后，将权重分别赋予每个特征通道，从而让神经网络重点关注某些特征通道，即提升对当前任务重要的特征通道，抑制对当前任务作用不大的特征通道。图3中，输入特征向量$\boldsymbol{X}$通过深度卷积网络F_{tr}之后得到特征向量$\boldsymbol{U}$。

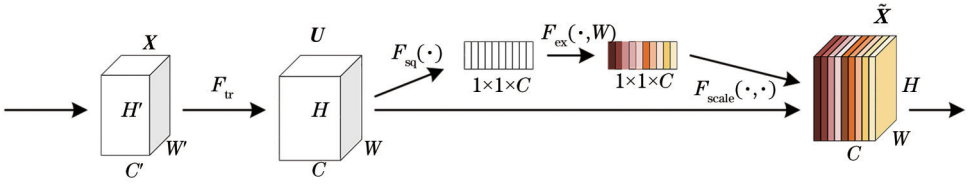

图3　SE Block[25]

Fig. 3　SE Block[25]

SE Block首先通过全局平均池化将每个通道的空间二维特征降维成具有全局感受野的一个实数，即

$$z_c=\frac{1}{H\times W}\sum_{i=1}^{H}\sum_{j=1}^{W}\boldsymbol{U}_c(i,j),\tag{4}$$

式中：$\boldsymbol{z}$为全局平均池化的结果；c表示的是$\boldsymbol{z}$和$\boldsymbol{U}$的通道编号；H和W为输入特征图的高和宽；i和j表示的是像素点在某个特征通道的坐标。在此过程中，通道数保持不变，得到$1\times1\times C$的向量；然后通过全连接网络和非线性激活函数学习得到每个通道的权重值。充分体现通道注意力思想的是最后一步，将归一化权重看作是通过特征选择后的每个通道的重要程度，通过和$\boldsymbol{U}$相乘把权重施加到每个特征通道上，完成通道特征重标定。

3　所提算法

3.1　3D卷积神经网络提取眼震特征

现有的传统方法或者深度学习的方法均采用人工提取特征，效率低下的同时容易遗漏某些关键信息，具有较强的主观性和局限性。本实验组将眼震视频解帧后得到的带有时序信息的图片序列直接作为算法的输入，由深度神经网络进行特征提取，弥补了人工提取特征方式存在的不足。2D卷积网络可以有效提取图像或视频空间信息，但帧间时序信息往往得不到保留，而3D卷积神经网络在进行特征提取时，会同时对时空信

息进行建模。因此，本实验组采用 3D MobileNet V2 作为基础网络框架对眼球震颤视频进行分类。

图 4 中，左图和右图分别表示在不同卷积步长时的网络结构，左图卷积步长为 1，右图卷积步长为 2。该网络的基础模块 3D Inverted Residual Block 是将 Inverted Residual Block 中的 2D 卷积改为 3D 卷积，其他 2D 操作也更改为相应的 3D 操作，节约了计算资源、加快模型推断速度的同时，特征提取能力强，低维信息的损失少。

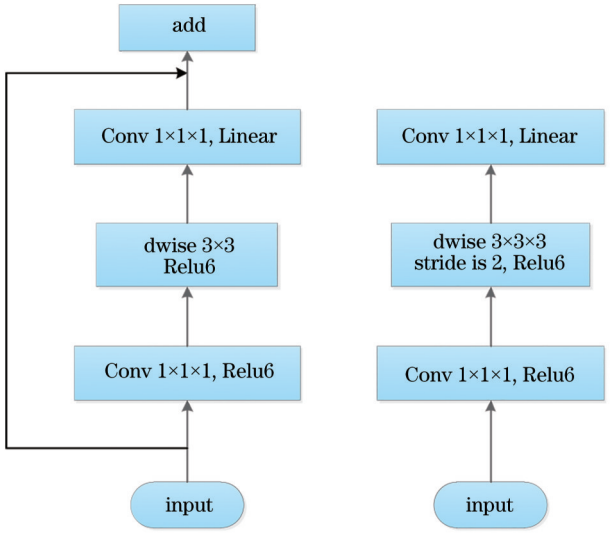

图 4　3D Inverted Residual Block

Fig. 4　3D Inverted Residual Block

3.2　时空通道注意力机制

将 SE Block 中的通道注意力思想扩展到三维卷积中，提出了一种时空通道注意力机制，该机制可以对各特征通道的空间信息和时序信息进行更好的筛选。

首先，将二维特征图的全局平均池化改为三维全局时空平均池化：

$$z_c=\frac{1}{H\times W\times T}\sum_{i=1}^{H}\sum_{j=1}^{W}\sum_{k=1}^{T}\boldsymbol{U}_c(i,j,k),\tag{5}$$

式中：T 代表特征图的时间序列；k 表示特征图的帧间位置索引。经过三维时空全局平均池化及后续的全连接层操作对时空信息进行整合，并且将整合后的时空信息作为通道权重对各特征通道进行重标定。

结合 3D Inverted Residual Block 和时空通道注意力机制，本实验组提出的 3D SE Inverted Residual Block 如图 5 所示。对于输入 Input，最左边的支路表示跳跃连接，中间支路表示 3D Inverted Residual Block 分支，最右边的支路表示时空通道注意力机制，通过时空通道注意力机制对 3D Inverted Residual Block 输出的各特征通道进行重标定，并将得到的结果与 Input 相加，得到最后的输出。该模块既保持了 3D SE Inverted Residual Block 的时空特征提取能力，又增加了网络对时空特征的辨识度，可以更有针对性的利用某些特征来增加分类精度。

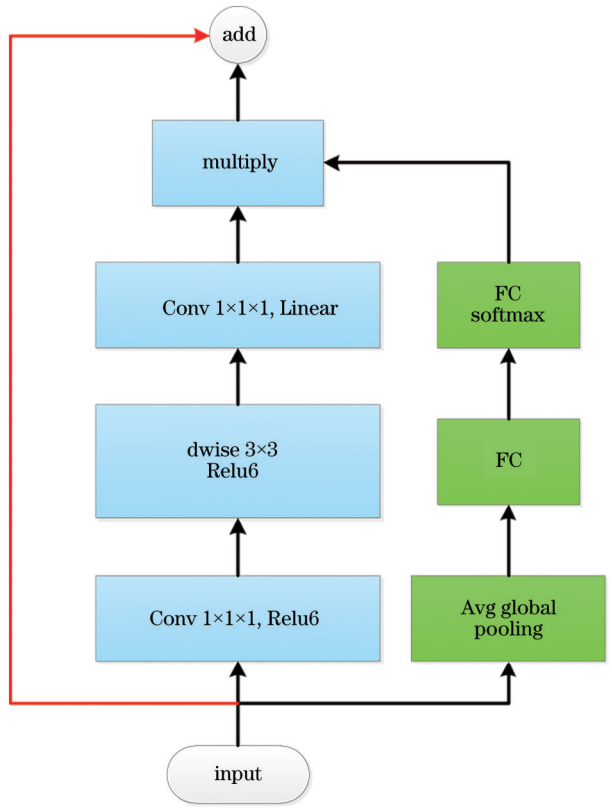

图 5　3D SE Inverted Residual Block

Fig. 5　3D SE Inverted Residual Block

3.3　基于注意力机制的 BPPV 眼震视频分类算法

Non-Local Block 的设计初衷是为了获取全局信息，捕捉长距离时空依赖关系。SE Block 则可以使网络重点关注对当前任务重要的特征通道，传统的卷积操作则能较好地获取局部信息。本实验组在全局特征、时空信息丰富的网络低层引入 Non-Local Block，通过全局时空注意力机制对全局时空特征赋予权重，可以更好地提取并融合眼球震颤空间信息和帧间时序信息；在网络中层，使用 3D Inverted Residual Block 进一步进行特征提取，融合全局和局部特征；在网络高层，特征通道数多，高级语义特征丰富，基于时空通道注意力的 3D SE Inverted Residual Block 可以调节眼球运动语义特征的权重，对更重要的特征施加更大的权重，增强模型和某些重要特征的相关性，减少冗余特征。

所提出的基于注意力机制的 BPPV 视频眼震图分类算法如图 6 所示。低层引入全局时空注意力机制，采用轻量型模块 3D Inverted Residual Block 进行特征提取和融合（图中的“…”代表该模块重复 12 次），网络高层使用 3D SE Inverted Residual Block 进行特征筛选，三者相辅相成，提升了网络的特征提取能力。

每个模块的具体位置、步距、重复次数和每层输出特征图的尺寸如表 1 所示。若某个模块重复多次，Stride 参数表示第 1 次时的步距，后续重复中步距均为 1。

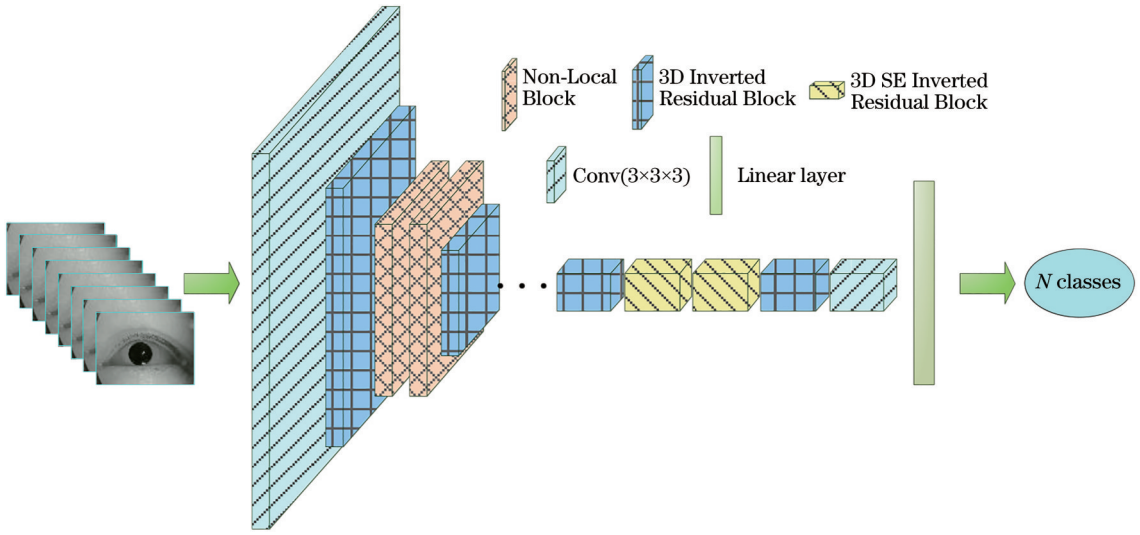

图 6　所提 BPPV 视频眼震图分类算法

Fig. 6　Proposed BPPV nystagmus video classification algorithm

表 1　所提 BPPV 视频眼震图分类算法框架

Table 1　Proposed BPPV nystagmus video classification algorithm framework

Layer/Stride	Repeat	Output size
Input		3×16×224×224
Conv (3×3×3)/2	1	32×16×112×112
Inverted Residual Block/2	1	16×16×56×56
NL Block/1	2	16×16×56×56
Inverted Residual Block/2	2	24×8×28×28
Inverted Residual Block/2	3	32×8×14×14
Inverted Residual Block/2	4	64×2×7×7
Inverted Residual Block/1	3	96×2×7×7
Inverted Residual Block/2	2	160×1×4×4
SE Inverted Residual Block/1	2	160×1×4×4
Inverted Residual Block/2	1	320×1×4×4
Conv (3×3×3)/1	1	1280×1×4×4
AvgPool/1	1	1280×1×1×1
Linear	1	*N* Classes

3.4　损失函数

在深度学习分类任务中，最常用的是交叉熵损失函数，不同的损失函数对模型性能的影响不同[27]。为了使目标检测算法更专注于难分类的样本，文献[28]提出了调整正负样本损失权重的研究思路。受此思想启发，本实验组使用 Pytorch 框架提供的 weighted cross entropy (WCE)损失函数，在交叉熵损失函数的基础上加入调制系数 ∂，用于对模型进行优化训练。所用自适应交叉熵损失的表达式为

$$L_{\text{weighted}}=-\sum_{i=1}^{N}\partial_i y_i\cdot\log\hat{y}_i, \tag{6}$$

$$\partial_i=1-\frac{N_i}{\sum_{k=1}^{K}N_k}, \tag{7}$$

式中：y_i 为第 i 个视频的类别；$\hat{y}_i$ 为模型输出的预测标签；N_i 表示第 i 类的样本在数据集中的数量；K 为样本总类别数。样本数量较少的类别对应的类别调制系数更大，属于这些类别的训练样本产生的损失值相对更大，在反向传播的优化过程中，模型参数会更多地朝着某个方向优化，这种优化会使样本数量较少的类别产生的损失值减少，从而更有效地降低损失值。该函数的引入可以有效降低数据样本不平衡带来的训练精度损失，增加模型对数量少的样本类别的识别能力。

4　实验结果与分析

所有的实验均在 Windows 10 上使用 CUDA 并行计算架构，并在 Cudnn 加速计算库的基础上搭建 PyTorch 框架，然后进行加速计算。显卡为 NVIDIA GEForce GTX3090(24 GB)，内存为 64.0 GB，CPU 为 Intel(R) Xeon(R) CPU E5-2660 v4 @2.00 GHZ。迭代次数为 100，优化器选择 Adam[29]，优化参数选择如下：冲量为(0.9，0.999)，前 50 次迭代学习率 0.001，第 51～100 次迭代学习率为 0.0001，权值衰减率为 1×10^{-5}，批大小为 32。无预训练和其他前置任务。每进行一次迭代训练，进行一次模型测试，将结果最好的模型参数保存，最后在验证集上进行验证。

4.1　数据介绍与预处理

4.1.1　数据介绍

建立数据集所用的 VNG 视频数据来自复旦大学附属眼耳鼻喉科医院。在位置实验中，使用来自上海志听医疗科技有限公司型号为 VertiGoggles R ZT-VNG-II 的红外视频眼动记录仪记录和保存患者的在收到外部刺激时真实准确的眼球运动视频，视频格式是 mp4，视频帧大小为 640×480，帧率为 60。

数据集由 1328 名患者的 27852 段眼震视频组成，除去异常和干扰数据，剩余 22193 个视频。所有的数

据均由四位耳科专家根据眼球震颤的四种运动特征(水平、竖直、轴向、强弱变化)进行标注,每个样本的标签表示在此视频剪辑中患者的眼球震颤模式。

表 2 为数据集标签描述,针对某个患者的眼震视频片段,其对应的标签标注格式如下。眼球的水平震颤方向(向左:0,向右:1,无水平眼震:2);眼球的竖直震颤方向(向上:0,向下:1,无竖直眼震:2);眼球的轴向震颤方向(顺时针:0,逆时针:1,无轴向眼震:2);眼震强度变化(由弱变强:0,由强变弱:1,强度无变化:2)。

表 2 数据集标签描述

Table 2 Label description of data set

Mode	0	1	2
Horizontal	Left	Right	None
Vertical	Up	Down	None
Axial	Clockwise	Counterclockwise	None
Intensity	From weak to strong	From strong to weak	None

按照眼球运动规律,将其分为不同的类别,如类别标签为“0020”的样本表示其眼球震颤模式为水平向左、竖直向上且无轴向颤动,强度由弱变强。

理论上存在 88 种不同的复合眼震类型,但由于某些复合眼震类型在患者中出现的情况极少,在现有的数据集中只存在着 66 种不同的眼震类型。本实验组以 3∶1∶1 的比例设置训练集、测试集、验证集。

4.1.2 数据预处理

为了减小深度神经网络的运算量,裁剪与眼球运动无关的区域。首先对数据集中的每个视频片段使用霍夫圆变换算法[30]提取瞳孔中心,得到瞳孔直径,如图 7 所示。

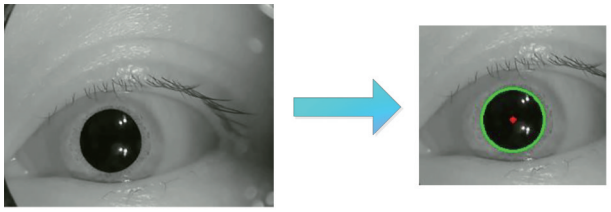

图 7 视频裁剪示意图

Fig. 7 Schematic diagram of video cropping

与此同时,可以甄别和删除眨眼和干扰时长较多的视频。然后以每个视频第 1 帧的瞳孔中心为中心点,在不超过边界的情况下,将视频裁剪为长宽均为六倍瞳孔半径的大小,若瞳孔半径小于 35 pixel,则将视频大小裁剪为 224×224。

裁剪之后,对视频进行解帧,根据视频的宽、高和帧率将视频解帧成连续的保留时序信息的图片序列,图片格式为 jpg。

最后,对于每个输入的数据样本,随机选取连续的 16 帧输入网络模型。

4.2 评价指标

为了评估所提算法性能,采用 One vs Rest 的多分类评价策略,即模型将某个数据样本分为某一类别,则其他类别对于该样本来说则是错误的。采用准确率(accuracy)、精确率(precision)、召回率(recall)作为评价指标,对模型进行整体评估。

准确率是预测正确的样本数量占总样本的比例,精确率是预测为正样本的样本中正样本所占的比例,召回率是正样本中被成功预测为正样本的比例,其表达式分别为

$$R_{\text{accuracy}}=\frac{N_{\text{TP}}+N_{\text{TN}}}{N_{\text{TP}}+N_{\text{TN}}+N_{\text{FN}}+N_{\text{FP}}},\tag{8}$$

$$R_{\text{precision}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FP}}},\tag{9}$$

$$R_{\text{recall}}=\frac{N_{\text{TP}}}{N_{\text{TP}}+N_{\text{FN}}},\tag{10}$$

式中:N_{TP}、N_{FP}、N_{TN}、N_{FN} 分别代表真阳性、假阳性、真阴性、假阴性的数量。

F1-score 是平均精确度和平均召回率的加权调和平均,其表达式为

$$S_{\text{F1}}=\frac{(1+\beta^2)\times R_{\text{precision}}\times R_{\text{recall}}}{\beta^2\times R_{\text{precision}}+R_{\text{recall}}},\tag{11}$$

式中:平衡系数 $\beta=1$。

4.3 实验

4.3.1 对比实验

为了验证所提眼球震颤分类算法的有效性,将该算法与一些主流的 3D 卷积神经网络算法进行了对比实验,结果如表 3 所示。

表 3 主流 3D 卷积神经网络在眼震视频分类数据集上的表现

Table 3 Performance of mainstream 3D convolutional neural networks on nystagmus video classification dataset

Algorithm	Number of parameters /MB	Accuracy
C3D[31]	34.80	0.8443
3D ResNet18[32]	33.24	0.8518
3D ResNet34[32]	63.55	0.8717
3D SqueezeNet[33]	1.87	0.8625
3D ShuffleNetV2[34]	1.37	0.8502
3D MobileNetV2	2.44	0.8791
Proposed algorithm	2.65	0.9085

从表 3 可以看出,与其他同类的 3D 卷积神经网络算法相比,所提算法具有明显的优势。C3D 算法由于只是简单的 3D 卷积层和池化层的堆叠,分类效果一般。3D ResNet 系列算法是经典卷积神经网络的代表,但实验结果表明,与 3D ResNet18 网络层数相近的轻量级模型表现更好,特别是 3D MobileNet V2,分类准确度超越了层数更深的 3D ResNet34。在综合模型性能和参数数量的前提下,选择 3D

MobileNet V2作为基础网络。所提算法引入了两种注意力机制后，在3D MobileNet V2的基础上分类准确度提升了将近3个百分点，表明了所提算法的优越性。

4.3.2　消融实验

为了更好说明所提算法的每个模块对于模型整体性能的影响，对在3D MobileNet V2基础上添加的两种注意力模块进行消融实验，实验结果如表4所示。在3D MobileNet V2基础上，对以下几种情况进行了对比：既不引入Non-Local Block（NL Block）也不引入SE Inverted Residual Block；仅引入NL Block；仅引入SE Inverted Residual Block；同时引入NL Block和SE Inverted Residual Block。

从表4可以看出：两种注意力机制模块的引入对此任务均有不同程度的提升；单独引入NL Block比单独引入3D SE Inverted Residual Block带来的分类准确度提升大；当同时引入NL Block和3D SE Inverted Residual Block时，分类准确度提升了0.0294，提升幅度明显。这说明在网络低层引入全局时空注意力机制，可以更好地提取眼震运动特征以及时空信息，高层的时空通道注意力机制在提升分类准确度上也有较好的效果，二者结合后效果最佳。

表4　不同模块对于模型的影响

Table 4　Influence of different modules on the model

Condition	Accuracy
3D MobileNet V2	0.8791
3D MobileNet V2 ＋NL Block	0.8922
3D MobileNet V2 ＋3D SE Inverted Residual Block	0.8853
3D MobileNet V2 ＋NL Block ＋3D SE Inverted Residual Block	0.9085

针对选用的WCE损失函数，通过实验与原始交叉熵损失函数（cross entropy）进行了对比，结果如图8所示。从[图8(a)]可以看出，虽然在训练前期，原始交叉熵函数的损失值收敛较快，但是当迭代次数越大，特别是在第51次迭代学习率降低至0.0001后，原始交叉熵函数收敛效果显然不如WCE。[图8(b)]的两种损失函数对应的准确度曲线表明，引入类别调制系数的交叉熵函数在此任务中的表现更加优秀。

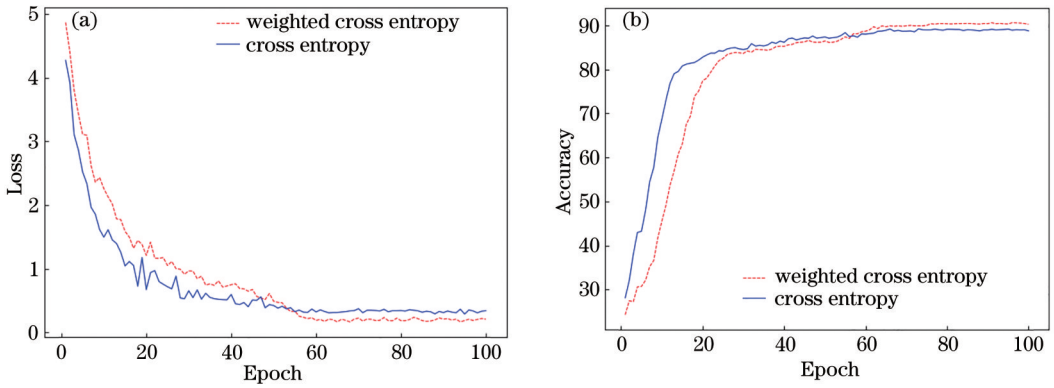

图8　不同损失函数的损失值和精确率与迭代次数的关系。(a)损失值；(b)精确率

Fig. 8　Relationship between loss value and accuracy of different loss functions and number of iterations. (a) Loss value; (b) accuracy

4.3.3　算法性能

表5是所提算法在验证集上的66个类别的表现，其中label表示的眼震震颤模式。从表5数据可以看出：每一类的precision、recall、F1-score值的平均值分别是0.905，0.920，0.904；某些类别（没有数据的）因数量N极少，参与模型训练的样本数严重不足，导致无法保证其分类精度，这些都是在临床诊断中极少出现的眼震类型；其余63类即使在各样本数量偏差较大的情况下，也能够取得较好的效果，表明所提算法的性能优异。

表5　所提算法在每一类上的表现

Table 5　The performance of the proposed algorithm in each category

Label	Precision	Recall	F1-score	N	Label	Precision	Recall	F1-score	N
0000	0.500	1.000	0.667	27	1111	1.000	0.933	0.966	71
0001	0.810	0.708	0.756	123	1112	1.000	1.000	1.000	251
0002	0.854	0.875	0.864	150	1120	0.864	1.000	0.927	74
0010	1.000	1.000	1.000	50	1121	1.000	0.927	0.962	190

表5（续）

Label	Precision	Recall	F1-score	N	Label	Precision	Recall	F1-score	N
0011	0.953	0.968	0.961	371	1122	0.972	0.977	0.975	782
0012	0.955	0.955	0.955	350	1200	0.700	1.000	0.824	45
0020	0.833	0.833	0.833	30	1201	0.867	0.929	0.897	128
0021	1.000	1.000	1.000	100	1202	0.984	0.918	0.950	673
0022	0.976	0.953	0.965	226	1210	0.500	0.400	0.444	13
0101				2	1212	0.857	0.750	0.800	21
0110	0.909	1.000	0.952	84	1220	0.800	0.821	0.810	1280
0111	0.968	0.989	0.978	426	1221	0.830	0.855	0.843	1742
0112	0.947	0.957	0.952	455	1222	0.907	0.874	0.890	2387
0120	1.000	0.933	0.966	60	2000	1.000	0.667	0.800	6
0121	0.952	1.000	0.976	145	2001	1.000	1.000	1.000	29
0122	0.981	0.963	0.972	877	2002	1.000	1.000	1.000	7
0210	0.857	0.857	0.857	51	2010	0.250	1.000	0.400	10
0211	1.000	0.933	0.966	237	2011	1.000	0.818	0.900	32
0212	0.955	0.980	0.967	767	2012	0.500	0.667	0.571	21
0220	0.853	0.871	0.862	1240	2021	1.000	1.000	1.000	29
0221	0.901	0.856	0.878	1746	2022	0.977	1.000	0.988	169
0222	0.878	0.904	0.891	2434	2101	1.000	1.000	1.000	11
1001	1.000	0.979	0.989	249	2102				3
1002	0.953	0.943	0.948	398	2110	1.000	1.000	1.000	12
1010	1.000	0.636	0.778	51	2111	1.000	1.000	1.000	45
1011	0.921	0.946	0.933	114	2112	1.000	1.000	1.000	190
1012	0.889	0.930	0.909	259	2120	1.000	1.000	1.000	6
1020	1.000	1.000	1.000	13	2122	1.000	1.000	1.000	276
1021	1.000	0.862	0.926	136	2201				5
1022	0.986	0.986	0.986	395	2202	1.000	0.857	0.923	32
1100	0.881	0.952	0.915	263	2211	1.000	1.000	1.000	8
1101	0.904	0.881	0.893	540	2212	0.500	1.000	0.667	5
1102	0.921	0.925	0.923	1071	2222	0.976	1.000	0.988	200

5 结　论

提出了一种基于改进的3D MobileNet V2的视频眼震图分类算法。该算法与现有算法的人工提取特征方式不同，直接采用深度卷积神经网络对眼震视频进行特征提取，通过引入全局时空注意力机制和时空通道注意力机制提升模型的特征提取能力，并使用改进的损失函数更好地对模型进行优化训练。实验结果表明：所提算法较其他算法在分类准确度上提升明显，达到90.85%；相比现有眼震分类算法，可以较为准确地对多类型眼球复合震颤进行分类，对各类的识别能力较强，在临床应用上价值更高。

参 考 文 献

[1] 头晕诊断流程建议专家组. 头晕的诊断流程建议[J]. 中华内科杂志, 2009, 48(5): 435-437.
Dizziness diagnosis process recommendation expert group. Dizziness diagnosis process recommendations[J]. Chinese Journal of Internal Medicine, 2009, 48(5): 435-437.

[2] 戴春富. 前庭医学发展现状[J]. 中国眼耳鼻喉科杂志, 2014, 14(3): 137-141.
Dai C F. Current status of vestibular medicine[J]. Chinese Journal of Ophthalmology and Otorhinolaryngology, 2014, 14(3): 137-141.

[3] von Brevern M, Radtke A, Lezius F, et al. Epidemiology of benign paroxysmal positional vertigo: a population based study[J]. Journal of Neurology, Neurosurgery, and Psychiatry, 2007, 78(7): 710-715.

[4] Schmal F, Stoll W. Diagnosis and management of benign paroxysmal positional vertigo[J]. Laryngo- Rhino-Otologie, 2002, 81(5): 368-380.

[5] 周国庆, 孔玉, 高志强, 等. 后半规管和水平半规管BPPV变位实验时眼震特点初步分析[J]. 中国现代医学杂志, 2017, 27(25): 92-94.
Zhou G Q, Kong Y, Gao Z Q, et al. Characteristics of nystagmus in position test for posterior canal and horizontal canal benign paroxysmal positional vertigo[J]. China Journal of Modern Medicine, 2017, 27(25): 92-94.

[6] Yülek F, Konukseven O Ö, Çakmak H B, et al. Comparison of the pupillometry during videonystagmography in asymmetric pseudoexfoliation patients[J]. Current Eye Research, 2008, 33(3): 263-267.

[7] Mekki S. The role of videonystagmography (VNG) in assessment of dizzy patient[J]. The Egyptian Journal of Otolaryngology, 2014, 30(2): 69-72.

[8] 张波, 孙敬武. 良性阵发性位置性眩晕患者裸眼及视频眼震图下眼震特征及定位诊断分析[J]. 听力学及言语疾病杂志, 2012, 20(3): 235-237.
Zhang B, Sun J W. The observation and diagnosis of 108 patients with benign positional paroxysmal vertigo with naked eyes and VNG[J]. Journal of Audiology and Speech Pathology, 2012, 20(3): 235-237.

[9] 亢超, 李文祥, 黄岫, 等. 基于深度学习的主动光学校正算法研究[J]. 光学学报, 2021, 41(6): 0611004.
Kang C, Li W X, Huang S, et al. Research on Active Optical Correction Algorithm Based on Deep Learning [J]. Acta Optica Sinica, 2021, 41(6): 0611004.

[10] 刘中法, 杨艺哲, 方宇, 等. 基于深度学习的虚拟相衬成像方法[J]. 光学学报, 2021, 41(22): 2217001.
Liu Z F, Yang Y Z, Fang Y, et al. Deep Learning-Based Virtual Phase Contrast Imaging Method[J]. Acta Optica Sinica, 2021, 41(22): 2217001.

[11] Daugman J. Probing the uniqueness and randomness of IrisCodes: results from 200 billion iris pair comparisons [J]. Proceedings of the IEEE, 2006, 94(11): 1927-1935.

[12] Yamada Y, Kobayashi M. Detecting mental fatigue from eye-tracking data gathered while watching video[M]// Teije A T, Popow C, Holmes J H, et al. Artificial intelligence in medicine. Lecture notes in computer science. Cham: Springer, 2017, 10259: 295-304.

[13] Charoenpong T, Thewsuwan S, Chanwimalueang T, et al. Pupil extraction system for Nystagmus diagnosis by using K-mean clustering and Mahalanobis distance technique [C]//Knowledge and Smart Technology (KST), July 7-8, 2012, Chonburi, Thailand. New York: IEEE Press, 2012: 24-29.

[14] Slama A B, Mouelhi A, Sahli H, et al. A deep convolutional neural network for automated vestibular disorder classification using VNG analysis[J]. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2020, 8(3): 334-342.

[15] Slama A B, Mouelhi A, Manoubi S, et al. An enhanced approach for vestibular disorder assessment[C]//2018 IEEE 4th Middle East Conference on Biomedical Engineering, March 28-30, 2018, Tunis, Tunisia. New York: IEEE Press, 2018: 243-246.

[16] Sayadi M, Lahiani R, Salah M B, et al. A new neural network method for peripheral vestibular disorder recognition using VNG parameter optimisation[J]. International Journal of Biomedical Engineering and Technology, 2018, 27(4): 321-336.

[17] Lim E C, Park J H, Jeon H J, et al. Developing a diagnostic decision support system for benign paroxysmal positional vertigo using a deep-learning model[J]. Journal of Clinical Medicine, 2019, 8(5): 633.

[18] Slama A B, Sahli H, Mouelhi A, et al. DBN-DNN: discrimination and classification of VNG sequence using deep neural network framework in the EMD domain[J]. Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2020, 8(6): 681-690.

[19] Mouelhi A, Slama A B, Marrakchi J, et al. Sparse classification of discriminant nystagmus features using combined video-oculography tests and pupil tracking for common vestibular disorder recognition[J]. Computer Methods in Biomechanics and Biomedical Engineering, 2021, 24(4): 400-418.

[20] Zhang W L, Wu H Y, Liu Y, et al. Deep learning based torsional nystagmus detection for dizziness and vertigo diagnosis[J]. Biomedical Signal Processing and Control, 2021, 68(10): 102616.

[21] Sandler M, Howard A, Zhu M L, et al. MobileNetV2: inverted residuals and linear bottlenecks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 4510-4520.

[22] Howard A G, Zhu M L, Chen B, et al. MobileNets: efficient convolutional neural networks for mobile vision applications[EB/OL]. (2017-04-17) [2021-06-08]. https://arxiv.org/abs/1704.04861.

[23] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//the Fourteenth International Conference on Artificial Intelligence and Statistics, {AISTATS} 2011, April 11-13, 2011, Fort Lauderdale, USA. Cambridge: JMLR, 2011, 15: 315-323.

[24] Kastner S, Ungerleider L G. Mechanisms of visual attention in the human cortex[J]. Annual Review of Neuroscience, 2000, 23: 315-341.

[25] Wang X L, Girshick R, Gupta A, et al. Non-local neural networks[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7794-7803.

[26] Hu J, Shen L, Sun G. Squeeze-and-excitation networks [C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE Press, 2018: 7132-7141.

[27] 王一同, 周宏强, 闫景逍, 等. 基于深度学习算法的计算光学研究进展[J]. 中国激光, 2021, 48(19): 1918004.
Wang Y T, Zhou H Q, Yan J X, et al. Advances in Computational Optics Based on Deep Learning[J]. Chinese Journal of Lasers, 2021, 48(19): 1918004.

[28] Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE Press, 2017: 2999-3007.

[29] Kingma D P, Ba J. Adam: A method for stochastic optimiza-tion[EB/OL]. (2017-01-30) [2021-06-08]. https://arXiv.org/abs/1412.6980.

[30] Duda R O, Hart P E. Use of the Hough transformation to detect lines and curves in pictures[J]. Communications of the ACM, 1972, 15(1): 11-15.

[31] Tran D, Bourdev L, Fergus R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//2015 IEEE International Conference on Computer Vision, December 7-13, 2015, Santiago, Chile. New York: IEEE Press, 2015: 4489-4497.

[32] He K M, Zhang X Y, Ren S Q, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE Press, 2016: 770-778.

[33] Iandola F N, Han S, Moskewicz M W, et al. Squeezenet: Alexnet-level accuracy with 50x fewer parameters and <0.5 MB model size[EB/OL]. (2016-11-04)[2021-06-08]. https://arxiv.org/abs/1602.07360.

[34] Ma N N, Zhang X Y, Zheng H T, et al. ShuffleNet V2: practical guidelines for efficient CNN architecture design [EB/OL]. (2018-07-30)[2021-06-08]. https://arxiv.org/abs/1807.11164.