DOI:10.3969/j.issn.1005-202X.2024.09.006

医学影像物理

基于双重注意力机制增强的复合型眼震分类框架

王卓然，方志军，王海玲，高永彬，李玉霞
上海工程技术大学电子电气工程学院，上海 201600

【摘要】针对现有研究仅能识别水平、垂直或轴向上某一方向是否有眼震发生，且未能考虑临床上具有强度变化、由多方向组成的复合型眼震的问题，提出一种基于双重注意力机制增强的复合型眼震分类框架。首先，提出一种眼震视频时空浓缩算法，结合卷积神经网络与霍夫变换，去除无效帧和无效区域的干扰，提高眼震视频质量。然后，采用密集光流算法提取眼球运动光流场。最后，构建一种基于双重注意力机制增强的复合型眼震分类网络，提出一种改进高效通道注意力模块，有效获取光流图不同通道中眼球震颤的方向、强度信息；在Bi-LSTM网络末端添加时间注意力模块，实现不同时序特征对分类结果的显著性表达。在合作医院提供的眼震数据集上，本文方法对复合型眼震分类准确率达到83.17%，在单独的水平、垂直、轴向上眼震分类准确率达到91.03%、89.74%、86.05%。本文方法实现复合型眼震的智能分类，具有一定的临床应用价值。

【关键词】医学图像处理；视频眼震电图；良性阵发性位置性眩晕；深度学习；注意力机制
【中图分类号】R318；TP391 【文献标志码】A 【文章编号】1005-202X(2024)09-1093-11

Composite nystagmus classification framework enhanced by dual attention mechanism

WANG Zhuoran, FANG Zhijun, WANG Hailing, GAO Yongbin, LI Yuxia
School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201600, China

Abstract: A composite nystagmus classification framework enhanced by dual attention mechanism is proposed to address the problem that the existing researches only identify whether nystagmus occurs in a horizontal, vertical, or axial direction, but fail to consider the issue of composite nystagmus composed of multiple directions with various intensities in clinical practice. A spatiotemporal concentration algorithm for nystagmus videos is presented, and it combines convolutional neural networks and Hough transform to remove interference from invalid frames and regions and to improve the quality of nystagmus videos. Then, a dense optical flow algorithm is used to extract the optical flow field of eye movement. Finally, a composite nystagmus classification network based on dual attention mechanism enhancement is constructed. An improved efficient channel attention module is used to effectively obtain the direction and intensity of nystagmus in different channels of the optical flow map; and a temporal attention module is added at the end of the bidirectional long short-term memory network to achieve significant expression of classification results based on different temporal features. On the nystagmus dataset provided by the cooperating hospital, the proposed method has an accuracy rate of 83.17% for composite nystagmus classification, and achieved accuracy rates of 91.03%, 89.74%, and 86.05% for individual horizontal, vertical, and axial nystagmus classifications. The proposed method realizes the intelligent classification of composite nystagmus and is valuable in clinic.

Keywords: medical image processing; videonystagmography; benign paroxysmal positional vertigo; deep learning; attention mechanism

前 言

良性阵发性位置性眩晕(Benign Paroxysmal Positional Vertigo, BPPV)是一种特发性、由头部位置改变激发、伴有眼球震颤的短暂阵发性眩晕的前庭器官疾病[1]。相关研究表明，近年来我国的耳科和神经科门诊中，BPPV的发病率高，且呈逐年上升趋势，终身患病率为2.4%，给初级卫生保健带来巨大的经

【收稿日期】2024-05-11
【基金项目】国家自然科学基金(62001284)；上海市科委“科技创新行动计划”社会发展科技攻关项目(21DZ1204900)
【作者简介】王卓然，硕士，研究方向：计算机视觉、医学图像处理，E-mail: zrwang@sues.edu.cn
【通信作者】方志军，博士，教授，博士生导师，研究方向：机器视觉、大数据分析，E-mail: zjfang@dhu.edu.cn

济负担[2]。在发病时，患者可能会经历严重的头晕和其他症状。在严重的情况下，患者可能感到恶心和呕吐，甚至无法睁开眼睛或行走[3]。这种情况还可能引发紧张和焦虑，严重威胁人们的健康生活和工作。因此，BPPV的相关研究愈发受到关注，尤其是对BPPV的诊断方面。在临床检查中，眼球震颤是BPPV最敏感、最具特异性的体征。BPPV的临床检查方法主要分为肉眼检查法、眼震电图（Electronystagmography, ENG）和视频眼震电图（Videonystagmography, VNG）技术3种类型。传统诊断方式只能由专科医生完成，对医生水平要求较高。即使是经验丰富的医生，在诊断时也容易受到主观判断、疲劳程度和外界因素的影响[4]。近年来，随着人工智能技术的飞速发展，智能诊断在智慧医疗中发挥越来越重要的作用[5]。VNG通过红外摄像头直接记录眼球震颤，为BPPV的识别提供全面信息，可有效区分其他前庭障碍，在临床上得到广泛应用[6]。基于VNG的智能诊断愈发受到研究者的关注，通过临床采集的眼球运动视频识别眼球震颤模式，可为BPPV的诊断提供有价值的科学依据。

现有研究主要从不同方向分别提取视频中眼球的运动特征，如水平、垂直、轴向，因此仅能区分水平、垂直或轴向上某一方向是否有眼震发生。但是在临床上，眼震并非是仅发生在单一方向，更多的是具有强度变化、由多方向组成的复合型眼震，例如“水平向右+垂直向上+顺时针旋转+由强变弱”。文献[7]通过对瞳孔中心坐标在水平/垂直运动方向的时间序列进行差分运算，得到速度曲线；在扭转维度上，使用相位相关技术作为扭转测量方法，定义两个相邻帧之间的相似度度量为扭转速度，从而分别在3个方向上识别是否有眼震发生。文献[8]提出基于扭转感知的双流识别网络，提取眼球扭转时产生的光流信息，可以对扭转眼震进行识别，但无法区分其它眼震类型。此类方法未能将不同维度上的运动信息有效融合，所提特征没有包含眼球在多个维度上的运动变化情况，从而导致对于临床上具有强度变化且由多个方向组成的复合型眼震的识别尚存在困难。

针对以上问题，本研究设计一个分类框架，能够更好地从患者红外眼球震颤视频中捕获运动信息，并可以对具有水平运动、垂直运动、轴向运动、强度变化4种标签的复合型眼震进行分类，提高临床诊断效率。首先，设计一种眼震视频时空浓缩算法，能够有效去除眼震视频录制时产生的光照、眨眼等噪声干扰，减少医生诊断时浏览视频的工作量，同时降低视频的存储成本。其次，采用密集光流估计算法提取眼震视频的光流场，在此基础上构建一种基于双重注意力机制增强的复合型眼震分类网络。在空间特征提取部分，利用所提改进高效通道注意力模块（Improved Efficient Channel Attention, IECA），增强不同颜色通道的特征表达，从而获取多个维度上眼震的方向与强度信息；在时间特征提取部分，引入时间注意力模块，增强眼震视频帧序列中关键变化节点的时序特征，提升不同时序特征对分类结果的显著性表达，从而更好地区分变化差异较小的眼震类型。该分类框架更贴合实际临床诊断情况，能够对BPPV复合型眼震类型进行区分，具有更高的临床应用价值。

1 相关工作

BPPV的主要临床特征是由头部或身体运动引起的短暂眩晕，并伴随特征性的眼球震颤[9]。常用的传统诊断方法是肉眼检查，通过观察特定头部姿势触发的眼震特征来诊断BPPV。尽管该方法相对简单且成本效益高，但在检测微弱眼震方面可能缺乏客观性和敏感性，进而导致误诊或漏诊。诊断BPPV的另一种工具是ENG，通过在眼周围放置电极来测量和记录眼动[10]。然而ENG需要特殊设备和受过培训的技术人员，限制其在某些临床环境中的可用性。目前，通过临床采集的红外视频，使用VNG技术进行诊断成为BPPV诊断的主流方案。VNG可准确观察BPPV患者眼震发生情况，显著提高其眼震检出率[11]。

近年来，随着计算机技术的进一步发展，其在眼震分类领域得到广泛应用。文献[12]使用多种信号处理技术来提取患者的生物信号，包括ENG、头部加速度和姿势等，再利用Fuzzy C-Means聚类算法进行分类，该方法需要临床采集多种数据，诊断流程不够简便。为了捕捉瞳孔的运动特征，文献[13]使用连续波动静脉评估系统采集患者数据，通过眼球产生的角膜-视网膜电位记录水平和垂直眼动。文献[14]通过增强眼白的血管图像的对比度，以小型蓝色LED光线辅助照射，实现对旋转眼球运动的高精度测量。其通过捕捉旋转眼球运动的图像，并通过视觉测量和系统获取血管位置的数据，以评估估计误差和处理速度，能够在几乎黑色眼球图像的情况下，测量旋转眼球运动，平均估计误差不超过0.24°。但是这种方法不能应用于普通设备收集的常见红外眼球震颤视频。文献[15]提出一种扭转计算算法，通过在两张图像间进行匹配，计算旋转角度。该类方法将水平、垂直和轴向上的眼球运动信息单独提取并进行分类。然而，其仅能诊断单一方向上是否存

在眼震，并未能直接识别具有强度变化和由多个方向组成的复合型眼震，因此具有一定的局限性。

随着人工智能的飞速发展，基于深度学习的医疗辅助智能诊断在临床上的应用越来越广泛[16]，取得良好的效果，如医学图像处理、医学自然语言处理等[17-19]。VNG技术采集的数据格式为红外视频，而深度学习在视频理解领域取得巨大的成就，因此利用深度学习进行眼震类型的智能区分引起广泛的关注。文献[20]通过分析患者眼震视频，训练基于Inception V4的深度学习模型定位瞳孔中心，从而提取出时间、频率等特征，再将提取到的特征输入卷积神经网络进行分类，从而判断患者疾病类型。文献[21]从每个视频片段中提取的个体眼球运动被量化为水平、垂直和扭转方向上的三维旋转的矢量和，将眼球震颤的振幅和方向的手工特征表示为网格图像并进行分类。该方法通过计算不同方向上速度的矢量和，融合了运动信息，但在分类时并未考虑强度变化。文献[22]提出一种基于深度学习的垂直眼球震颤识别方法，能够很好地识别垂直方向上发生的眼球震颤，但无法区分其它眼震类型。此类研究多数仅能在水平、垂直、旋转方向上分别提取运动信息，判断某一方向上是否有眼震发生，进而区分患者和正常人群。部分研究将眼球在不同方向上的运动信息进行融合，能够识别简单眼震类型，但难以检测临床中实际出现的具有强度变化、由多方向组成的复合型眼震。

2 方 法

本研究提出的基于双重注意力机制增强的复合型眼震分类框架如图1所示，输入是医院录制的眼球震颤红外视频，该框架由4部分组成：数据预处理、视频时空浓缩、光流提取、基于双重注意力机制增强的复合型眼震分类网络。

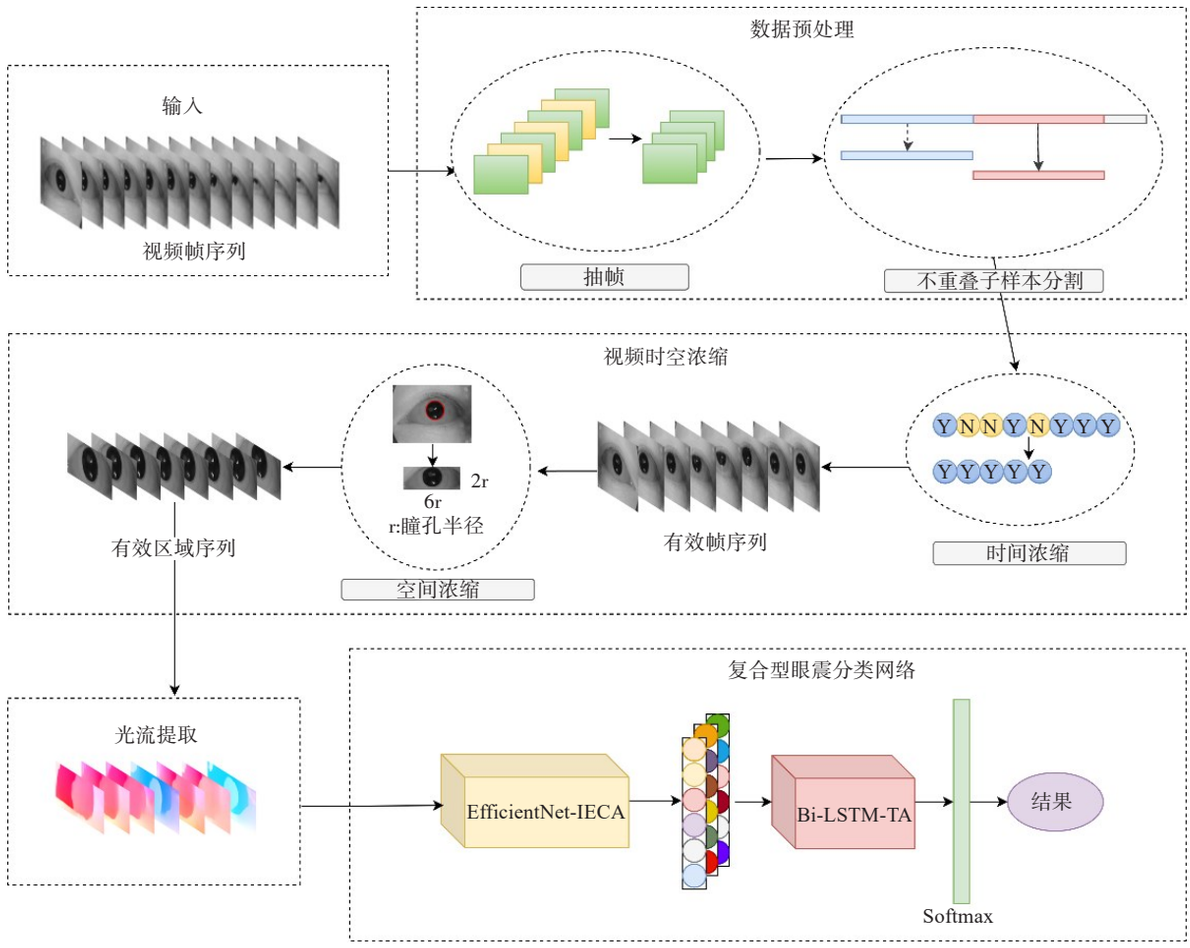

图1 基于双重注意力机制增强的复合型眼震分类框架

Figure 1 Composite nystagmus classification framework enhanced by dual attention mechanism

2.1 数据预处理

医院采集的原始视频数据帧速率为60帧/s，相邻帧之间变化幅度较为微弱。为强化眼球的运动趋势，首先对原始视频进行抽帧处理，将原始视频解帧后间隔1帧保存，有利于后续模型捕捉眼球运动信息。由于深度学习模型训练的视频数量不足，本研究将医院采集到的眼震视频片段，分割成固定长度的不重叠子样本。对于抽帧处理后的视频，每100帧切割为一个子样本，从而获取更多样本片段，扩充模型训练可用数据，该方法能够在保证信息不重复的情况下增加数据量，分割过程如图2所示。同时，医院采集的原始视频数据存在数据不平衡问题。不同类型的BPPV发病率本身存在较大区别，导致不同类别的样本数量存在显著差异，模型可能会倾向于预测样本数量较多的类别，而对于样本数量较少的类别则表现较差[23]。在进行子样本分割之后，采用随机过采样技术来平衡不同类型的眼震数据，针对样本量较少的眼震类型，研究通过随机复制该类型的子样本，增加该类型数据量，使不同类别样本量达到平衡，避免模型对发病率较高的类型产生倾向性。然后对视频片段中的图像帧进行缩放和平移以及随机裁剪，提高模型的泛化能力。

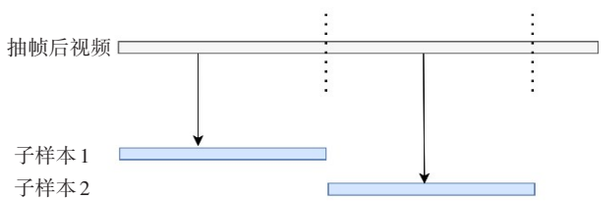

图2 不重叠子样本分割过程

Figure 2 Non-overlapping subsample segmentation process

2.2 视频时空浓缩

医院采集的患者眼部视频中存在噪声干扰，例如眼睑和眼睫毛的遮挡、反射引起的光斑等。对于瞳孔和虹膜区域被遮挡的帧，如闭眼帧、半闭眼帧等，难以提取到眼球运动的有效信息，故认为此类帧为无效帧。同时，在医生临床诊断和模型训练过程中，仅需要聚焦瞳孔及虹膜附近区域，其他区域为无效区域。为减少无效帧和无效区域带来的噪声干扰，提出一种面向眼震视频的视频时空浓缩算法，用来剔除视频中的无效帧和无效区域。该算法共包含时间浓缩和空间浓缩两个阶段。首先结合卷积神经网络和霍夫圆变换，去除视频中瞳孔被遮挡或者面积过小的无效帧，实现眼震视频时间浓缩；在此基础上，对帧图像进行处理，利用随机椭圆拟合方法定位有效区域并进行分割，剔除与眼震运动不相干的无效区域，实现眼震视频的空间浓缩。每一阶段具体处理过程如下：眼震视频时间浓缩阶段将无效帧剔除问题细化为图像二分类任务，首先将视频解帧，有效帧作为模型训练中的正样本，无效帧为负样本。在标记的数据集上训练一个卷积神经网络。将输入视频的每一帧，识别为无效帧或有效帧。考虑到瞳孔被部分遮盖的眨眼帧可能被卷积神经网络识别为正样本，本研究在CNN网络末端加入霍夫圆变换，来计算每一帧中瞳孔的面积S，如果该面积小于预设阈值T，则同样视为无效帧。最后，通过将标记为有效帧的所有视频帧按序组合在一起，完成对原视频的时间浓缩，处理流程如图3所示。

在视频时间浓缩后的基础上，进行眼震视频空间浓缩。本研究设计一套眼震视频空间浓缩处理流程，如图4所示。对于原始视频帧，首先将利用高斯滤波处理图片，之后对图片进行二值化操作，再使用形态学中的开操作消除瞳孔中因光反射而产生的白斑，然后使用闭操作消除眼睫毛等细微的黑色无关物；随机生成一定数量的椭圆，对每个椭圆进行适应度评估，筛选出最佳的椭圆拟合结果，确定瞳孔中心位置和半径。以瞳孔为中心，在不超过边界的情况下，裁取长为6倍半径，宽为2倍半径的矩形区域，能够覆盖瞳孔、虹膜及附近位置，得到粗略分割的有效区域。矩形区域面积公式如下所示：

$$S_R = 6r \times 2r \tag{1}$$

其中，S_R为矩形区域面积，r为瞳孔的半径。

2.3 光流提取

本研究所用数据是合作医院在临床诊断过程中，由红外视频眼动记录仪采集的眼震视频。因为红外视频亮度和对比度较低，故难以定位到准确的特征点，并以此来跟踪眼球的运动。基于深度学习的递归全对场变换网络（Recurrent All-Pairs Field Transforms, RAFT）[24]能够有效地处理光流场中的错配和异常情况，从而在复杂场景下表现更加稳定可靠，能有效减少眼震运动估计中睫毛、光斑等带来的负面影响。同时，RAFT算法还具有更低的计算成本和更快的速度，这使得其在眼震运动估计领域具有更好的实用性和效率。近年来随着深度学习技术的发展，基于深度学习的光流方案已超越传统方法，在准确性和运行速度上表现出色[25-26]。深度学习方法通过多层架构更高效地提取图像特征，避免传统方案中的复杂优化问题。通过选取几种最先进的基于深度学习的光流估计算法进行实验对比，最终选择

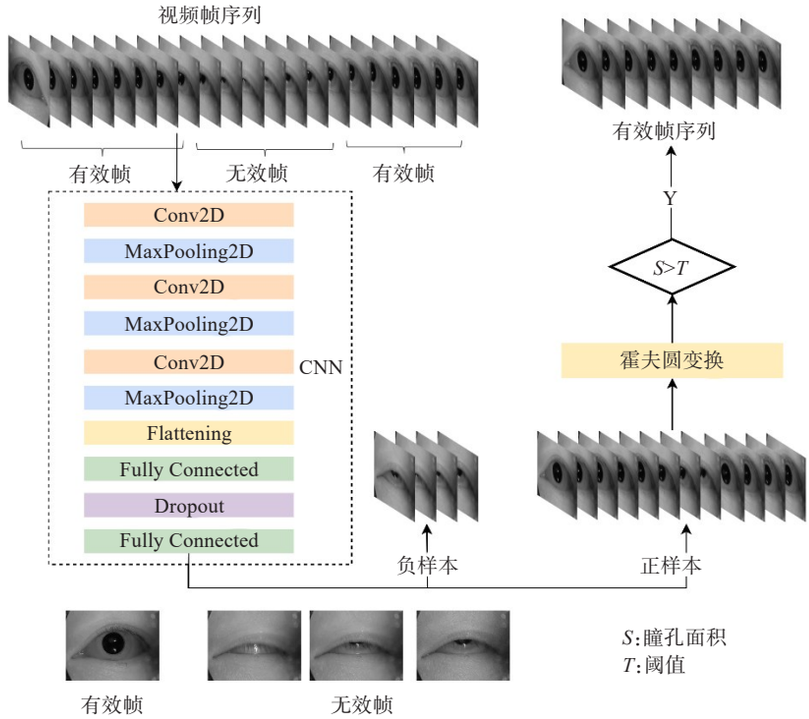

图3 眼震视频时间浓缩

Figure 3 Temporal concentration of nystagmus video

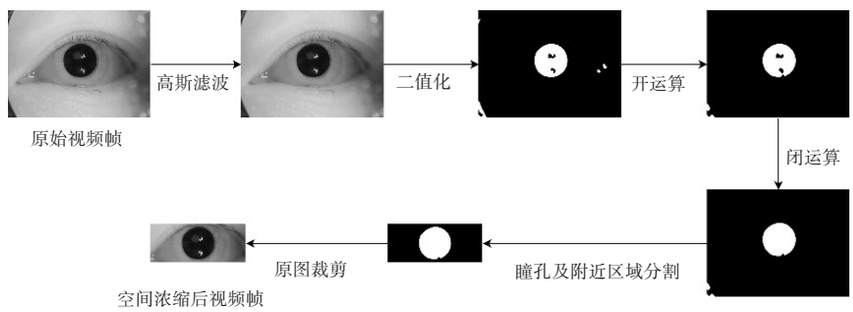

图4 眼震视频空间浓缩

Figure 4 Spatial concentration of nystagmus video

使用RAFT算法从时空浓缩后的眼震视频中提取光流信息。首先使用RAFT算法对每个眼震视频处理，提取光流信息，然后以HSV颜色空间的方式保存为光流图像[27]。在HSV颜色空间中，光流的方向表示为色调（Hue），而光流的强度表示为饱和度（Saturation）和明度（Value）。最后，将保存的光流图像组合得到眼震数据的光流视频。

2.4 基于双重注意力机制增强的复合型眼震分类网络

为了有效地提取眼球震动信息，实现对复合型眼震的有效识别，构建一种基于双重注意力机制增强的复合型眼震分类网络，其整体架构如图5所示。网络的输入为提取的光流图序列；空间特征提取部分采用EfficientNet-IECA网络，该网络考虑到光流图以不同通道上的颜色信息代表眼球震颤的方向、强度信息，在基础网络EfficientNetV2上加入改进高效通道注意力机制，利用较少的参数量，提升不同通道间的信息交互，增强对眼球震颤运动信息的全面捕捉；时序特征由Bi-LSTM网络进行提取，利用Bi-LSTM获取时间特征向量；为强化不同时序特征对分类结果的显著性影响，提高关键位置的权重，将提取到的时间特征向量输入时间注意力模块，对不同时序特征进行自适应加权，最终通过Softmax层得到眼震分类结果。

EfficientNetV2是一种采用神经结构搜索技术的CNN网络，通过平衡网络深度、宽度和图像分辨率，优化运算量和准确率。首次将EfficientNetV2引入复合型眼震分类领域，其在EfficientNet的基础上提出Fused-MBConv模块，提高模型训练速度和识别准确率，各模块参数如表1所示。

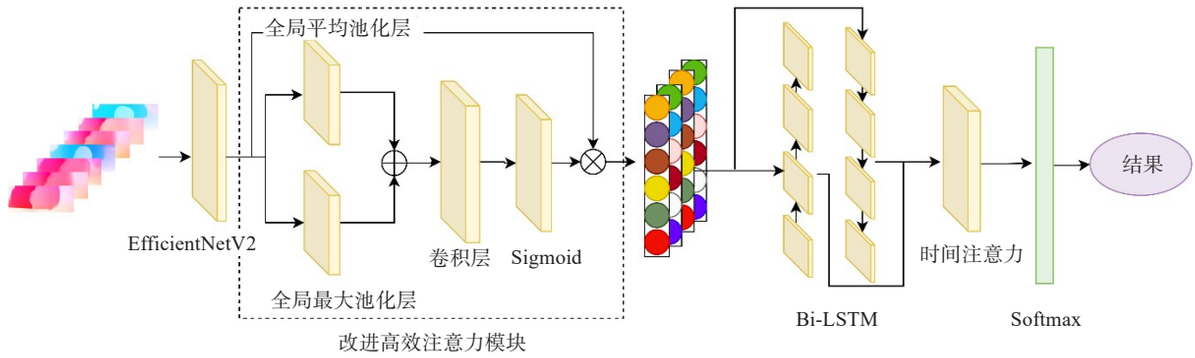

图5 基于双重注意力机制增强的复合型眼震分类网络

Figure 5 Composite nystagmus classification network enhanced by dual attention mechanism

表1 EfficientNetV2基本模块

Table 1 Basic modules of EfficientNetV2

阶段	模块	步长	层数	通道数
0	Conv 3×3	2	1	24
1	Fused-MBConv1,k3×3	1	2	24
2	Fused-MBConv4,k3×3	2	4	48
3	Fused-MBConv4,k3×3	2	4	64
4	MBConv4,k3×3	2	6	128
5	MBConv6,k3×3	1	9	160
6	MBConv6,k3×3	2	15	272

ECA是在不降维的基础上,采取局部跨通道方案的高效通道注意力模块,其原始结构如图6所示。X_i为输入的特征光流图,Y_i为特征图与注意力权重相乘得到注意后的特征图。相较于ECA原始结构,IECA模块增加了全局最大池化层,其结构如图5所示。通过全局最大池化层获取最突出的特征信息,再将全局最大池化层和全局平均池化层提取特征信息进行融合,在有效获取通道注意信息的同时,强化关键特征信息的差异性。在输入的特征图中,其不同通道上包含眼球震颤的方向、强度信息,利用全局最大池化层和全局平均池化层在空间维度上获取特征信息,将所得特征向量进行相加;再使用一维卷积代替常见的全连接运算,获取不同通道间的信息,避免全连接运算导致的通道维度降低,减少信息丢失,从而捕捉到更多眼球震颤信息。最后通过Sigmoid激活函数生成新的特征向量。

Bi-LSTM网络由一个前向LSTM和一个后向LSTM构成,包含若干个LSTM单元。每个LSTM单元由输入门i_t、遗忘门f_t、输出门o_t组成,通过3个门控制信息的保留和遗忘,实现对时间流中信息的获取,其计算公式如下所示:

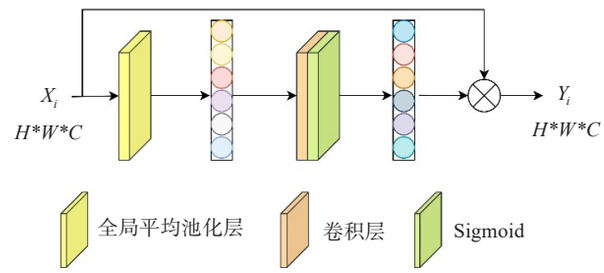

图6 ECA模型结构图

Figure 6 ECA model structure

$$f_t = \sigma(w_f x_t + w_f h_{t-1} + b_f) \quad (2)$$

$$i_t = \sigma(w_i x_t + w_i h_{t-1} + b_i) \quad (3)$$

$$u_t = \tanh(w_u x_t + w_u h_{t-1} + b_u) \quad (4)$$

$$o_t = \sigma(w_o x_t + w_o h_{t-1} + b_o) \quad (5)$$

其中,x_t表示t时刻的输入,h_{t-1}表示t-1时刻的隐层状态值,u_t为状态更新,σ为Sigmoid激活函数;w_f、w_i、w_o和w_u分别表示遗忘门、输入门、输出门和状态更新过程中的权重系数;b_f、b_i、b_o和b_u分别表示遗忘门、输入门、输出门和状态更新过程中的偏置值。单元状态c_t根据输入门和遗忘门的结果计算得出:

$$c_t = i_t \times u_t + f_t \times c_{t-1} \quad (6)$$

最终隐藏层的输出结果h_t为:

$$h_t = o_t \times \tanh(c_t) \quad (7)$$

相较于传统LSTM只能单向学习,Bi-LSTM可充分考虑眼震视频前后帧的时序信息,能够增强模型对过去和未来信息的捕捉,从而提高模型的准确性和鲁棒性。Bi-LSTM层集成眼震视频连续帧图像中的特征,并获得时间特征向量。考虑到眼震视频序列中,各时刻的特征与眼震类型预测值的关联程度不同,而Bi-LSTM对所有信息的重视程度是相同的,容易使结果产生误差。为增强时间序列中重要特征对预测结果的显著性表达,在Bi-LSTM后添加时间注意力模型,实现对Bi-LSTM提取到的时间序列特

征进行自适应加权。将经过Bi-LSTM后的输出特征h_t作为时间注意力模型的输入特征，利用tanh函数激活，之后与注意力参数矩阵$\boldsymbol{\omega}^{\mathrm{T}}$相乘，与偏置权重$\boldsymbol{b}$相加后，再经过Softmax函数激活得到$t$时刻权重得分$S_t$。最终原输入特征与权重得分相乘，得到时间注意力输出结果A_t，其公式如下所示：

$$S_t = \mathrm{Softmax}\left(\boldsymbol{\omega}^{\mathrm{T}} \times \tanh\left(h_t\right) + \boldsymbol{b}\right) \tag{8}$$

$$A_t = h_t \times S_t \tag{9}$$

时间注意力机制在Bi-LSTM提取的时间特征向量基础上，通过计算不同时刻的权重得分，对提取到的时间特征进行加权，实现自适应增强不同时刻对眼震类型预测的显著性表达。

3 结果与分析

本研究实验在Windows 10操作系统下进行，所用计算机的CPU为Intel(R) Xeon(R) CPU E3-1220 v6 @ 3.00 GHz，3.00 GHz，GPU为NVIDIA 1080ti（11 GB），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。表2展示了本研究具体的实验参数设置。

表2 实验参数设置

Table 2 Experimental parameter settings

名称	设置参数
损失函数	交叉熵损失函数
优化器	AdamW
学习率	5e-4
批量大小	16
训练周期	120

3.1 数据介绍

本文使用的眼球震颤视频数据来源于上海市复旦大学附属眼耳鼻喉科医院。研究整理从2020年10月到2021年6月期间BPPV患者的诊断记录，使用红外视频眼动记录仪采集患者产生的眼动视频，从中创建一个由604个眼震视频组成的数据集，视频格式是MP4，视频帧大小为640×480，帧率为60，原始数据集中每一类别眼震样本数量如表3所示。该数据集由医院专家进行标注，标出每个视频中眼球震颤在水平、垂直、轴向的运动方向、强度变化，起止帧位置。每一类眼震标签由4位数字组成：第1位代表水平方向眼震情况；第2位代表垂直方向眼震情况；第3位代表轴向眼震情况；第4位代表眼震强度变化情况，标签中数字具体含义如表4所示。

表3 数据预处理后每一类眼震样本量

Table 3 Sample size for each category of nystagmus after data preprocessing

数据集	标签					
	0012	0221	1012	1102	1122	1211
原始	77	113	76	143	83	112
预处理后	200	200	200	200	200	200

表4 视频眼震标签含义

Table 4 Video nystagmus label meanings

标签	水平方向	垂直方向	轴向	强度信息
0	向左	向上	顺时针旋转	从弱到强
1	向右	向下	逆时针旋转	从强到弱
2	无	无	无	无

为了避免数据不平衡及增加数据量，首先将原始视频分割成相同长度的不重叠子样本，按照患者为单位进行随机划分，训练集、测试集、验证集的比例设置为3:1:1。再对数量较少的类别进行随机过采样，最终得到一个包含六类眼震类型，共计1 200个子样本的数据集，处理前后不同眼震类别视频数量如表3所示。

3.2 评价指标

为了衡量模型的分类准确性和泛化能力，采用多个性能指标对模型的分类结果进行评估。首先，使用准确率（Accuracy）作为主要评价指标。准确率是指模型在整个测试集上正确分类的样本比例，它能够直观反映模型分类的整体准确性。为了更全面地评估模型的性能，引入精确率（Precision）、召回率（Recall）和F1值（F1-score）。精确率是指模型预测为正类别的样本中真正为正类别的比例，召回率是指真正为正类别的样本中被模型正确预测为正类别的比例，F1值是精确率和召回率的调和平均值。计算公式如下所示：

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{FP} + \mathrm{FN} + \mathrm{TN}} \tag{10}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{11}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{12}$$

$$\text{F1-score} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}} \tag{13}$$

其中，假设存在眼震类型A：TP代表真阳性的数量，即A类型眼震被正确识别为A类型的数量；FP代表假阳性的数量，即其他类型眼震被识别为A类型的数量；TN代表真阴性的数量，即其他类型眼震未被识别

为A类型的数量；FN分别代表假阴性的数量，即A类型眼震被识别为其他类型眼震的数量。

3.3 实验结果

首先对不同算法的光流估计效果进行对比，然后通过实验展示采用不同基准网络的复合型眼震分类网络的分类性能，并展示其在整体和每类眼震上的分类效果。考虑到眼震分类属于动作识别领域中的特定场景问题，选取动作分类领域经典网络模型在眼震数据集上进行对比实验。最后，对所提出的框架进行消融实验，以验证所提模块的必要性，并在此基础上对所提双重注意力机制进行消融实验，验证其对所提方法的增强效果。

3.3.1 光流估计效果分析 为了研究基于深度学习的不同光流估计模型对BPPV眼震运动的效果，本研究选择几种最先进的算法进行实验对比。在图7中展示了相同视频帧序列使用不同光流估计算法得到的眼球震颤运动估计效果。由图7可知，PWC-Net算法难以准确捕捉眼球的具体位置信息、运动方向和强度变化。相比之下，LiteFlowNet算法提取的光流场虽然勉强可见眼球位置，但其仍然无法满足对单个像素运动趋势的准确估计，在LiteFlowNet算法中，眼球被分割成多个部分，每部分被赋予不同的运动趋势，导致色彩较为混杂，与实际情况不符。而采用RAFT算法的部分则呈现出更好的效果，该算法清晰地显示眼球的位置和边缘轮廓信息，同时对于眼球边缘部分运动最为显著的像素变化有着显著的表现，能够更好地捕捉到眼球震动的运动信息。从整体效果来看，基于RAFT的光流估计算法在提高眼球运动估计的准确性和可靠性方面表现出色，为BPPV眼震运动研究提供可靠的实验基础。

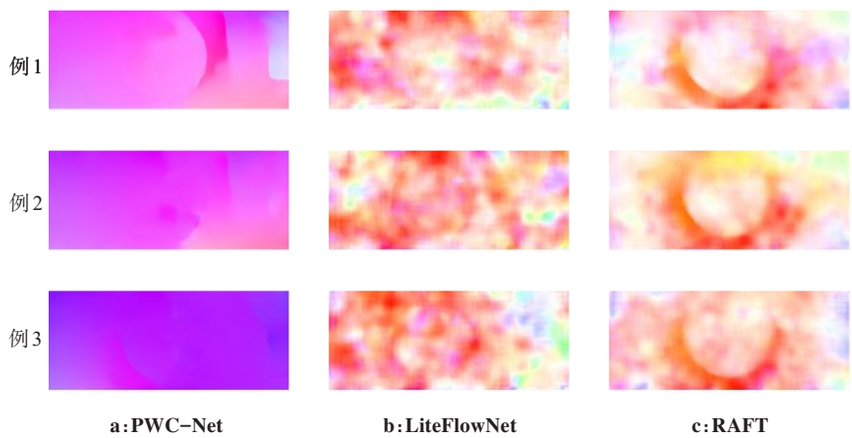

图7 不同光流估计算法效果图

Figure 7 Visual results of different optical flow estimation algorithms

3.3.2 基于双重注意力机制增强的复合型眼震分类网络结果 研究使用在ImageNet预训练的ResNet152-Bi-LSTM网络和EfficientNetV2_S-Bi-LSTM网络作为预选基准网络，使用预训练模型，能够加快模型的训练速度，并且提升模型性能。使用ResNet152和EfficientNetV2_S分别作为基准网络，准确率达到77.92%和78.26%。ResNet152的参数量为60.2 M，而EfficientNetV2_S参数量仅为21.5 M。考虑到模型实际应用场景中计算机硬件配置较低，故选用参数量更少的EfficientNetV2_S模型作为基准网络。为了使用预训练模型，在网络训练输入之前，调整所有的数据大小为256×256。

表5展示了以EfficientNetV2_S-Bi-LSTM为基准网络构建的基于双重注意力机制增强的复合型眼震分类网络，在不同诊断指标下全面评估的结果。其中，水平、垂直、轴向是指在单一方向上进行分类，水平方向分为向左震颤和向右震颤，垂直方向分为向上震颤和向下震颤，轴向方向分为顺时针方向旋转和逆时针方向旋转，复合型眼震为同时包含水平、垂直、轴向、强度变化4种标签的眼震类别。在单一方向上进行分类时，由于仅使用该方向上眼球运动特征，不需考虑与其它方向信息进行融合，水平方向准确率为91.03%，垂直方向准确率为89.74%，轴向方向准确率为86.05%。而复合型眼震的分类，需要融合上述3个方向上的运动信息，并且考虑强度变化。相较于以上单标签的二分类问题，针对复合型眼震的多标签六分类问题难度较大，但其在临床诊断中更为常见，临床应用价值更高，分类准确率达到83.17%。

表5 不同诊断标准下的分类准确率

Table 5 Classification accuracy under different diagnostic criteria

诊断标准	准确率	强度信息	类别数
水平方向	0.910 3	×	2
垂直方向	0.897 4	×	2
轴向方向	0.860 5	×	2
复合型眼震	0.831 7	√	6

本文研究的复合型眼震分类属于多分类问题，故选用交叉熵函数作为损失函数。同时，使用基于Adam优化算法改进的AdamW算法，在Adam的基础上增加权重衰减的正则化项，以减少模型的过拟合。权重衰减会使权重参数变小，从而降低模型的复杂度。表6展示所提模型在测试集不同眼震类别上的分类效果，每一类的精确率、召回率和F1值结果的平均值为0.836 9、0.832 1、0.827 8。在包含多个方向眼球震颤的1012类别，即“水平向右+垂直向上+逆时针旋转+强度无明显变化”类型眼震上，精确率指标达到0.921 1，其余类别眼震也取得较好的效果，验证本文方法的有效性。

表6 所提模型在每类眼震上的分类结果

Table 6 Classification results of the proposed model on each type of nystagmus

标签	精确率	召回率	F1值
0012	0.880 0	0.628 6	0.733 3
0221	0.720 9	0.968 8	0.826 7
1012	0.921 1	0.897 4	0.909 1
1102	0.827 6	0.774 2	0.800 0
1122	0.838 7	0.866 7	0.852 5
1211	0.833 3	0.857 1	0.845 1
平均值	0.836 9	0.832 1	0.827 8

3.3.3 对比实验　复合型眼震分类问题可以视为动作识别的一类特定场景问题。目前3D卷积在动作识别领域取得良好的效果，通过在空间维度和时间维度上进行3D卷积，捕捉视频中包含的运动信息。因此，本研究使用动作识别中经典3D卷积网络3D ResNet[28]和基于轻量级神经网络的3D MobileNetV2[29]、3D EfficientNet[30]以及混合模型LRCN[31]进行实验方法对比。表7是本文提出的眼震分类网络与经典动作视频分类网络的准确率比较结果。从表中可看出，3D ResNet50分类准确率为33.45%，随着层数的增多，分类准确率略有提升，但3D ResNet101分类准确率仅达到39.77%。基于轻量级神经网络改进来的3D MobileNetV2与3D EfficientNet分类准确率分别为34.24%与36.75%。根据实验结果分析，仅通过简单的3D卷积，可能无法有效捕捉复合型眼震的运动信息，从而进行类型判断。因为眼震运动不同于普通行为动作，其是发生在短时间内的快速颤动，不同类型眼震之间图像特征较为相似，加深特征提取层次对眼震类型分类准确率提升并不明显，并且训练3D卷积神经网络往往需要大量的数据，研究所用眼震数据集受临床采集限制，规模较小，在3D卷积网络训练过程中训练集准确率较高，验证集与测试集准确率较低，出现过拟合现象。混合模型LRCN相较于3D卷积模型，利用LSTM模型加强了对时序信息的捕捉，分类准确率达到60.62%，说明通过重点关注不同帧间眼球的运动变化，可以有效提升对不同类型眼震的区分。而本文方法准确率达到83.17%，表明在不同类别眼震视频中目标和背景较为相似、目标运动差异较小的情况下，本文方法利用IECA模块完成不同通道间信息交互，高效利用光流图以颜色保存运动信息的特性，对眼球震颤的方向、强度信息进行有效融合，同时时序特征提取部分利用时间注意力机制，增加关键时序特征的权重。两者互为补充，提取到更完整的时空特征，增强模型的整体性能。

表7 不同深度学习网络在复合型眼震分类数据集上的表现

Table 7 Performance of different deep learning networks on a composite nystagmus classification dataset

网络模型	准确率
3D ResNet50[28]	0.334 5
3D ResNet101[28]	0.397 7
3D MobileNetV2[29]	0.342 4
3D EfficientNet[30]	0.367 5
LRCN[31]	0.606 2
本文方法	0.831 7

3.3.4 消融实验　为了验证所提框架的每个模块对模型整体性能的影响，对所提模块进行消融实验。将不同模块处理后的数据，输入基准的EfficientNet+Bi-LSTM网络。实验结果如表8所示，其中M1模块代表数据预处理，M2模块代表视频时空浓缩，M3模块代表光流提取。实验1为将原始数据直接输入分类网络，准确率仅有36.09%。实验2是原始数据经过缺

失数据预处理模块的模型进行分类，准确率为74.35%，较实验1提升38.26%。实验3是缺少视频时空浓缩模块的模型，分类准确率为66.96%，较实验1提升30.87%。实验4是缺少光流提取模块的模型，准确率较实验1提升14.78%。实验5是包含所有模块的模型，准确率达到78.26%。从实验结果可以看出，每个模块都有效提高模型分类的准确率，其精确率、召回率、F1值也有明显提升。相较于其它模块，光流提取模块由于将眼球震颤的方向、强度信息通过不同颜色进行加强表示，更有利于分类网络提取运动特征，对模型准确率贡献较大。

表8 不同模块对模型性能的影响

Table 8 Effect of different modules on model performance

实验编号	包含模块			准确率	精确率	召回率	F1值
	M1	M2	M3				
实验1	×	×	×	0.3609	0.2308	0.3502	0.2507
实验2	×	√	√	0.7435	0.7529	0.7452	0.7389
实验3	√	×	√	0.6696	0.6631	0.6740	0.6564
实验4	√	√	×	0.5087	0.5572	0.5030	0.4208
实验5	√	√	√	0.7826	0.7853	0.7834	0.7814

同时，在前置模块的基础上，为验证研究在基准EfficientNet+Bi-LSTM网络上添加双重注意力机制的有效性，对改进后的分类网络进行消融实验。从表9可知，在所提框架基础上添加IECA模块后，准确率提升3.3%，验证IECA模块通过通道间局部交互，更高效融合分布在不同通道上眼球震颤的方向、强度信息；将TA模块加入网络后，准确率提升1.61%，验证时间注意力通过自适应的权重分配，增强关键时序特征对分类结果的影响度，分类准确率达到83.17%。将IECA模块替换为原始ECA模块，准确率下降0.62%，说明IECA模块强化类别特征，进而提高分类准确率。以上结果验证了使用IECA模块和TA模块的双重注意力机制的有效性。

表9 不同注意力机制对分类网络性能的影响

Table 9 Effect of different attention mechanisms on network performance

模型结构	准确率
EfficientNet + Bi-LSTM	0.7826
EfficientNet-IECA + Bi-LSTM	0.8156
EfficientNet-ECA + Bi-LSTM-TA	0.8255
EfficientNet-IECA + Bi-LSTM-TA	0.8317

4 结 论

从提高诊断过程中BPPV复合型眼震的分类准确率需求出发，提出一种基于双重注意力机制增强的复合型眼震分类框架，该框架由数据预处理、视频时空浓缩、光流提取、基于双重注意力机制增强的复合型眼震分类网络4个部分组成。首先通过不重叠子样本分割增大数据量，通过随机过采样实现数据平衡，用于防止模型过拟合，然后提出一种面向眼震视频的时空浓缩方法。通过剔除无效帧和无效区域，高效获取关键的瞳孔及附近区域信息，并减少原视频中的噪声干扰。再利用RAFT光流估计算法提取眼球运动光流场。最后构建一种基于双重注意力机制增强的复合型眼震分类网络，提出改进高效通道注意力提升不同通道上眼震方向、强度信息的交互，引入时间注意力增强不同时序特征的显著性表达，进而预测复合型眼震的类别。实验结果表明，本文方法在BPPV复合型眼震分类任务上表现良好，在真实的临床BPPV眼震数据集上，针对复合型眼震分类准确率达到83.17%，同时在单一方向眼震分类任务中表现出色。与现有研究相比，本研究能够对具有强度变化、由多方向组成的复合型眼震进行类型区分，具有更高的临床应用价值。

【参考文献】

[1] von Brevern M, Radtke A, Lezius F, et al. Epidemiology of benign paroxysmal positional vertigo: a population based study[J]. J Neurol Neurosurg Psychiatry, 2007, 78(7): 710-715.
[2] Kovacs E, Wang XT, Grill E. Economic burden of vertigo: a systematic review[J]. Health Econ Rev, 2019, 9(1): 37.
[3] You P, Instrum R, Parnes L. Benign paroxysmal positional vertigo[J]. Laryngoscope Investig Otolaryngol, 2018, 4(1): 116-123.
[4] Pham TX, Choi JW, Mina RJ, et al. LAD: a hybrid deep learning

system for benign paroxysmal positional vertigo disorders diagnostic [J]. IEEE Access, 2022, 10: 113995-114007.
[5] Huang SG, Yang J, Shen N, et al. Artificial intelligence in lung cancer diagnosis and prognosis: current application and future perspective[J]. Semin Cancer Biol, 2023, 89: 30-37.
[6] Mekki S. The role of videonystagmography (VNG) in assessment of dizzy patient[J]. Egypt J Otolaryngol, 2014, 30(2): 69-72.
[7] Lu W, Li ZZ, Li YN, et al. A deep learning model for three-dimensional nystagmus detection and its preliminary application [J]. Front Neurosci, 2022, 16: 930028.
[8] Zhang WL, Wu HY, Liu Y, et al. Deep learning based torsional nystagmus detection for dizziness and vertigo diagnosis[J]. Biomed Signal Process Control, 2021, 68: 102616.
[9] 曹莉梅, 蒋宾, 廖远高. 良性阵发性位置性眩晕患者临床特征和危险因素分析[J]. 临床耳鼻咽喉头颈外科杂志, 2021, 35(10): 905-909.
Cao LM, Jiang B, Liao YG. Analysis of clinical characteristics and risk factors in patients with benign paroxysmal positional vertigo [J]. Journal of Clinical Otorhinolaryngology Head and Neck Surgery, 2021, 35(10): 905-909.
[10] Bojrab DI, Lai WD, Bojrab DI. Electronystagmography and videonystagmography[M]//Babu S, Schutt CA, Bojrab DI. Diagnosis and Treatment of Vestibular Disorders. Cham: Springer International Publishing, 2019: 45-65.
[11] 惠晶, 訾定京, 范秀博. 视频眼震电图在良性阵发性位置性眩晕患者诊断中的临床应用价值研究[J]. 陕西医学杂志, 2019, 48(5): 636-638.
Hui J, Zi DJ, Fan XB. Clinical effects of video nystagmus with benign paroxysmal positional vertigo[J]. Shaanxi Medical Journal, 2019, 48 (5): 636-638.
[12] Ben Slama A, Mouelhi A, Manoubi S, et al. An enhanced approach for vestibular disorder assessment [C]//2018 IEEE 4th Middle East Conference on Biomedical Engineering (MECBME). Piscataway, NJ, USA: IEEE, 2018: 243-246.
[13] Newman JL, Phillips JS, Cox SJ. 1D convolutional neural networks for detecting nystagmus[J]. IEEE J Biomed Health Inform, 2021, 25 (5): 1814-1823.
[14] Hoshino K, Ono N. Measurement of eyeball rotational movements in the dark environment [C]//2017 IEEE Winter Applications of Computer Vision Workshops (WACVW). Piscataway, NJ, USA: IEEE, 2017: 75-80.
[15] Jansen SM, Kingma H, Peeters RL, et al. A torsional eye movement calculation algorithm for low contrast images in video-oculography [C]//2010 Annual International Conference of the IEEE Engineering in Medicine and Biology. Piscataway, NJ, USA: IEEE, 2010: 5628-5631.
[16] Khan A, Brouwer N, Blank A, et al. Computer-assisted diagnosis of lymph node metastases in colorectal cancers using transfer learning with an ensemble model[J]. Mod Pathol, 2023, 36(5): 100118.
[17] 张帅, 张俊忠, 曹慧, 等. 基于ConvNeXt网络的新冠肺炎X射线图像诊断方法[J]. 激光与光电子学进展, 2023, 60(14): 20-28.
Zhang S, Zhang JZ, Cao H, et al. Coronavirus disease X-ray image diagnosis method based on ConvNeXt network [J]. Laser & Optoelectronics Progress, 2023, 60(14): 20-28.
[18] Khurana D, Koli A, Khatter K, et al. Natural language processing: state of the art, current trends and challenges[J]. Multimed Tools Appl, 2023, 82(3): 3713-3744.
[19] Dhar T, Dey N, Borra S, et al. Challenges of deep learning in medical image analysis-improving explainability and trust[J]. IEEE Trans Technol Soc, 2023, 4(1): 68-75.
[20] Ben Slama A, Mouelhi A, Sahli H, et al. A deep convolutional neural network for automated vestibular disorder classification using VNG analysis[J]. Comput Methods Biomech Biomed Eng Imaging Vis, 2020, 8(3): 334-342.
[21] Lim EC, Park JH, Jeon HJ, et al. Developing a diagnostic decision support system for benign paroxysmal positional vertigo using a deep-learning model[J]. J Clin Med, 2019, 8(5): 633.
[22] Li HB, Yang ZF. Vertical nystagmus recognition based on deep learning [J]. Sensors, 2023, 23(3): 1592.
[23] Sen S, Singh KP, Chakraborty P. Dealing with imbalanced regression problem for large dataset using scalable artificial neural network[J]. New Astron, 2023, 99: 101959.
[24] Teed Z, Deng J. RAFT: recurrent all-pairs field transforms for optical flow[C]//Computer Vision-ECCV 2020. Cham: Springer International Publishing, 2020: 402-419.
[25] 崔毅博, 汤仁东, 邢大军, 等. 视觉光流计算技术及其应用[J]. 电子与信息学报, 2023, 45(8): 2710-2721.
Cui YB, Tang RD, Xing DJ, et al. Visual optical flow computing: algorithms and applications[J]. Journal of Electronics & Information Technology, 2023, 45(8): 2710-2721.
[26] Cao YJ, Zhang XS, Luo FY, et al. Learning generalized visual odometry using position-aware optical flow and geometric bundle adjustment[J]. Pattern Recognit, 2023, 136: 109262.
[27] Sural S, Qian G, Pramanik S. Segmentation and histogram generation using the HSV color space for image retrieval [C]//Proceedings. International Conference on Image Processing. Piscataway, NJ, USA: IEEE, 2002: II-589-II-592.
[28] Anvarov F, Kim DH, Song BC. Action recognition using deep 3D CNNs with sequential feature aggregation and attention [J]. Electronics, 2020, 9(1): 147.
[29] Wei DF, Tian Y, Wei LQ, et al. Efficient dual attention SlowFast networks for video action recognition[J]. Comput Vis Image Underst, 2022, 222: 103484.
[30] Song YF, Zhang Z, Shan CF, et al. Constructing stronger and faster baselines for skeleton-based action recognition[J]. IEEE Trans Pattern Anal Mach Intell, 2023, 45(2): 1474-1488.
[31] Dewar SKS, Hiremath A, Patil S, et al. Human activity recognition using LRCN[C]//2022 IEEE North Karnataka Subsection Flagship International Conference (NKCon). Piscataway, NJ, USA: IEEE, 2022: 1-4.

（编辑：陈丽霞）