智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于人类视觉皮层双通道模型的驾驶员眼动行为识别

申天啸， 韩怡园， 韩冰， 高新波

引用本文:
申天啸, 韩怡园, 韩冰, 等. 基于人类视觉皮层双通道模型的驾驶员眼动行为识别[J]. 智能系统学报, 2022, 17(1): 41–49.
SHEN Tianxiao, HAN Yiyuan, HAN Bing, et al. Recognition of driver’s eye movement based on the human visual cortex two-stream model[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(1): 41–49.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202106051

DOI: 10.11992/tis.202106051
网络出版地址: http://kns.cnki.net/kcms/detail/23.1538.TP.20211222.0843.002.html.

基于人类视觉皮层双通道模型的驾驶员眼动行为识别

申天啸[1], 韩怡园[1], 韩冰[1], 高新波[2]
（1. 西安电子科技大学 电子工程学院, 陕西 西安 710071; 2. 重庆邮电大学 重庆市图像认知重点实验室, 重庆 400065）

摘　要: 驾驶员的危险行为会增加交通事故的发生率, 目前对驾驶员行为的研究中大多通过面部识别等方法对异常行为如疲劳驾驶、接电话等进行识别。这种方法仅客观地对驾驶员行为进行分类, 而忽略了他们在驾驶过程中的主观心理。眼动仪是记录和分析驾驶员眼动数据的有效工具, 可以清晰地了解驾驶员的想法并总结其视觉认知模式。因为目前还没有针对驾驶员眼动行为的数据库, 首先构建了真实道路场景下的眼动视频数据集 VIPDAR_5, 与传统数据相比, 它存在更多的摄像机运动、光照变化、视线遮挡等情况。针对这些问题提出了一个基于人类视觉皮层双通路的模型 TWNet, 通过模拟视觉机制, 提高了驾驶员眼动行为的识别性能。另一方面, 通过自适应最大池化层和通道权重设置, 减少参数, 提高准确率。在 VIPDAR_5 数据集上的实验结果表明, 与现有方法相比, 该模型能有效识别驾驶员眼动行为。

关键词: 眼动视频数据库; 行为识别; 深度学习; 道路安全; 辅助驾驶; 眼动追踪; 人类视觉系统; 行为研究
中图分类号: TP391　　**文献标志码:** A　　**文章编号:** 1673−4785(2022)01−0041−09

中文引用格式: 申天啸, 韩怡园, 韩冰, 等. 基于人类视觉皮层双通道模型的驾驶员眼动行为识别 [J]. 智能系统学报, 2022, 17(1): 41–49.
英文引用格式: SHEN Tianxiao, HAN Yiyuan, HAN Bing, et al. Recognition of driver's eye movement based on the human visual cortex two-stream model[J]. CAAI transactions on intelligent systems, 2022, 17(1): 41–49.

Recognition of driver's eye movement based on the human visual cortex two-stream model

SHEN Tianxiao[1], HAN Yiyuan[1], HAN Bing[1], GAO Xinbo[2]
(1. School of Electronic Engineering, Xidian University, Xi'an 710071, China; 2. Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing 400065, China)

Abstract: Drivers' dangerous actions will increase the incidence of traffic accidents. The current researches on driver's action are based on facial recognition to recognize abnormal actions, such as fatigue driving, cell phone usage. These methods only classify drivers' actions objectively and ignore their subjective thoughts during driving. The eye tracker is a device that can record and analyze driver's eye movement effectively, understand their thoughts clearly and summarize their visual cognition patterns. There is no dataset for driver's eye movement currently. Therefore, this paper first builds a eye movement video dataset named VIPDAR_5 applicable in real road scenes. Compared with traditional dataset, it contains more camera motion, illumination change, and sight occlusion situations. Therefore, the TWNet model based on two channels of the human visual cortex is built in this paper, which can improve recognition performance by simulating human visual mechanisms. On the other hand, adaptive max-pooling layer and channel weight setting are added to reduce parameters and improve recognition accuracy. Experimental results on the VIPDAR_5 dataset indicate that the model proposed in this paper can effectively recognize drivers' eye movement in comparison with existing methods.

Keywords: eye movement video dataset; action recognition; deep learning; road safety; aided driving; eye tracking; human visual system; behavioral research

随着全球工业和经济的快速发展, 各国汽车保有量和机动车驾驶人员逐年上升, 其中我国驾驶员已超 4 亿人, 约占全国总人口的三分之一[1]。

收稿日期: 2021−07−01.　**网络出版日期:** 2021−12−22.
基金项目: 国家自然科学基金项目（61572384, 62076190, 41831072）; 西安电子科技大学研究生创新基金项目.
通信作者: 韩冰, E-mail: bhan@xidian.edu.cn.

与此同时，车祸发生数目也逐年上升，根据统计显示我国 2019 年交通事故发生数超 20 万次，死亡和受伤人数则高达 30 多万人，车祸带来了巨大的生命财产损失。

交通事故发生是由多种因素引起的，而驾驶员接收的大部分外界信息经由视觉获得，所以很多交通事故都由驾驶员的危险行为造成。目前，针对驾驶员行为的研究多集中于异常检测，如分心行为检测[2-3]、疲劳驾驶[4] 等。这些方法多采用人脸关键点检测，通过眼睛、鼻子、嘴巴等区域位置的变化来识别驾驶员的状态，或者直接利用支持向量机或一些简单的人工智能方法如神经网络来对这些行为进行检测和分类。然而它们都仅从客观上对驾驶员行为进行检测识别，而没有从主观上分析驾驶员的心理和视觉认知模式。

眼动仪作为一种能有效采集佩戴者的主观视觉信息的工具，能真实地记录佩戴者正在注意的区域以及正在交互的对象。它在多个计算机视觉领域取得了较大的应用，其中也有应用于驾驶行为分析，如将驾驶员视野区域分为多个部分，根据眼动仪记录注视点在不同区域间的变化[5]，通过数学分析、马尔科夫链等传统方法，得到驾驶员变道行为时的注视转移特性[6]。但上述方法使用的眼动数据较少且多采集于封闭路段或汽车模拟软件。

针对目前驾驶员眼动数据分析方法对真实道路场景下识别能力弱的问题，我们收集了 10 名驾驶员约 11 h 的眼动驾驶视频数据，经过数据筛选与标注，最终构建了一个有 350 段视频片段，超 9 万视频帧，包含多场景下真实路况的驾驶员眼动视频数据库 VIPDAR_5。此外本文参考了人类大脑皮层视觉系统中定位和识别结构[7]，提出了一种基于三维卷积的双通路动作识别网络 TWNet。网络的 What 通路提取图像信息，Where 通路经过自适应池化层，降低图像分辨率，使其更加关注于捕获视频序列中的运动信息。最后通过权重超参数设置来融合两个通道信息并得到最终结果。

本文的主要贡献如下：

1) 构建了新的驾驶员眼动行为数据库 VIPDAR_5，包含不同路况、气候、时间等情况下约 11 h 的原始视频序列，共 350 个视频片段。

2) 针对驾驶员眼动行为识别任务，模拟人类大脑皮层视觉双通路机制，提出了基于三维卷积的双通路网络 TWNet。在网络通道中加入自适应最大池化层降低输入图像分辨率，减少网络参数，提高对运动信息的捕获能力。在网络输出部分设置超参数，控制两个通道的输出权重。

3) TWNet 在 VIPDAR_5 数据库上取得良好表现，性能超过了现有行为识别方法。

1　相关工作

1.1　相关数据库

目前常用行为识别数据库有 KTH[8]、UCF101[9]、Kinetics[10] 等。KTH 是最早的动作识别数据库之一，包含 4 个场景下的 6 类人体行为，如：走路、跑步、拳击等，这些视频片段中存在尺度和光照变化，但背景较为单一且拍摄位置固定。UCF101 是目前最广泛使用的动作识别数据集，通常用于评估模型性能。其中视频主要采集于互联网，由人类行为、人人交互、人机交互、乐器演奏和体育运动这五类构成，包含 101 个子种类共计 13 320 个视频序列。这些序列持续时间都在 10 s 左右且每个片段中包含一个完整的动作。由于来源自网络，这些片段的背景较为杂乱且是任务驱动的，不具有普适性。Kinetics 包含人物互动、人人交互、仅人体行为三大类。因为它包含大量的标注信息，所以广泛用于预训练。

虽然基于自然场景下的行为识别数据库已较为完善，但它们仅记录动作的客观表示而无法分析行为发生时的主观信息。近年来越来越多的以人为中心的主观数据库被提出，如 Charades-ego[11] 和 Epic-Kitchens[12]，通过录制第一视角视频，记录被试与物体间的交互，但由于仅使用头戴式摄像头，无法记录受试者的实时注视位置。

眼动仪是记录人类视觉认知变化最有效的工具。它可以获取设备佩戴者的瞳孔状态、注视点位置、注视时间等相关信息，已应用于显著性检测、行为识别等计算机视觉任务，表 1 比较了 6 个眼动数据库。LEDOV[13] 针对视频显著性预测任务，采集了来自不同网站的 158 类视频共 538 段，并提供了 32 个被试观看时的关注点。EGTEA Gaze+[14] 与 GTEA Gaze+[15] 主要用于行为识别，它们记录了被试在室内烹饪时的动作。EGTEA Gaze+更是在 GTEA Gaze+基础上扩大了数据量，并提供多模态信息如音频、被试手部掩膜等。Hollywood[16] 由 Hollywood-2[17] 和 UCF sports[18] 这两个动作识别数据库组成。其中 Hollywood-2 挑选自 69 部电影中的 12 类行为，如开车、吃饭和握手等，UCF sports 取自体育资料片，分为 9 个动作共 150 个视频。Hollywood 捕捉了 19 名被试观看这些视频时的关注区域，用于视频的显著性预

测。MIT[19] 是最早建立的眼动追踪数据库之一, 针对大多数方法不符合实际眼球运动情况的问题, Judd 采集了 15 位被试在 1 003 幅图像上的眼球轨迹数据。POET[20] 是一个开源的物体检测数据库, 由 Pascal VOC 2012[21] 中 10 类图像组成, 并记录了 5 位被试观察这些图像时的眼动信息。

表 1　眼动数据集对比

Table 1　Comparison of eye movement datasets

数据库	任务驱动	被试数目	提出时间	视频/图片数目	分辨率	片段时间/s
LEDOV	×	32	2018	538	1 280×720	5-60
EGTEA Gaze+	√	32	2018	86	1 280×960	1 500
GTEA Gaze+	√	26	2012	35	1 280×960	750
Hollywood	√	19	2015	1857	720×480	2-90
POET	√	5	2014	6 270	1 680×1 020	—
MIT	×	15	2009	1 003	1 024×1 024	—
VIPDAR_5(Ours)	×	10	2021	350	1920×1 080	8-15

1.2　相关方法

视频序列相较于静态图像, 不仅包含空间语义信息, 还包含时间运动信息。目前, 基于深度学习的动作识别方法主要分为二维卷积和三维卷积[22] 两类。

Simonyan 等[23] 针对卷积神经网络时间信息建模能力弱的问题, 设计了双流法。它由空间与时间流组成, 分别输入单帧 RGB 图像和帧间光流图来提取图像特征与运动特征, 但通道间没有信息交换, 无法学习特征间的对应关系。Ng 等[24] 提出将长短时记忆网络 (long short-term memory, LSTM)[25] 用于聚合视频序列的特征, 获得视频的时序信息。Wang 等[26] 为解决长时间行为识别以及数据量较少产生的过拟合问题, 在双流法基础上提出了 temporal segment networks (TSN) 网络。通过稀疏采样, 将长时序列分割为若干片段, 再从各片段中随机采样一帧, 使得输入包含序列的各时间段, 具有了提取全局特征的能力。Lin 等[27] 提出了 temporal shift module (TSM) 来处理时序信息, 通过移位时域通道, 完成了帧间信息交换。

Tran 等[28] 在三维卷积基础上提出了 C3D, 它同时捕获时空间信息, 获得的特征更加紧凑。Tran 等[29] 因残差网络在图像分类等任务上表现出色, 将其应用于动作识别并提出 Res3D[30]。又通过将三维卷积解耦成二维空间和一维时间卷积, 提出了 R(2+1)D[31]。虽然参数量不变, 但由于卷积块中存在额外的 ReLU 函数, 所以它具有更小的误差, 便于优化。Feichtenhofer 针对行为识别中空间信息变化慢而动作信息变化快, 根据人眼不同细胞对时空信息的敏感度差异, 设计由两路卷积神经网络组成的 SlowFast[32] 网络, 该网络用低帧率和高帧率通道来分别捕捉空间信息和动作信息。然而, 三维卷积引入大量的参数, 造成较高的计算量和内存消耗。

针对驾驶员眼动数据, 文献 [33] 按里程数将 12 名被试分为熟练与非熟练两类, 并通过数学方法分析不同区域的注视次数以及心率的变化, 得到了驾驶经验和通行方式对视觉特性的影响。文献 [34] 用眼动仪测录了 7 名受试者的视角分布、注视时间等眼动信息, 分析在不同车速下路侧标志信息对驾驶员视觉搜索模式的影响。但这些驾驶员眼动数据多采集于封闭道路或模拟机器, 难以应对真实复杂的道路场景。

2　数据库 VIPDAR_5

本节介绍眼动驾驶行为数据库 VIPDAR_5 的构建过程, 包括数据收集、数据筛选及标注等。

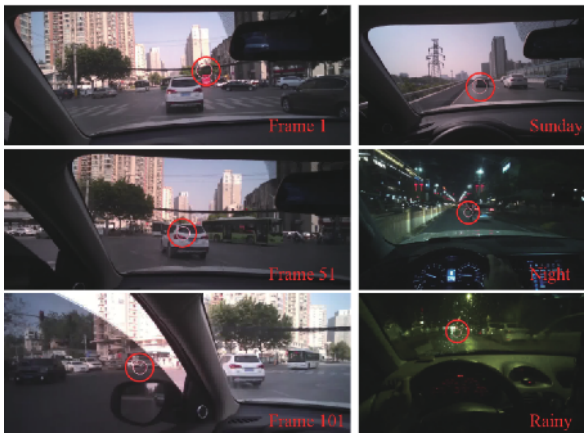
2.1　数据采集

我们使用便携式眼动仪 Tobii Glasses pro 2 记录被试的眼动数据, 该眼动设备质量仅 45 g, 能确保驾驶员的自由感和舒适性, 且不会影响正常的驾驶操作。目前已记录 10 位不同性别、车辆和驾驶经验的驾驶员的眼动视频。

数据采集的具体流程: (1) 在数据采集前对被试佩戴的眼动仪进行校正, 以确保设备能够准确地跟踪被试的眼睛和瞳孔, 设置眼动视频的帧率为 30 f/s, 分辨率为 1920 像素×1080 像素; (2) 告知参与者按照平时的驾驶习惯在道路上自由驾驶, 从而得到他们驾驶过程中的真实意图而不是完成特定任务的行为; (3) 被试开始驾驶 5 min 后开始录制, 并在录制过程中记录驾驶员动作起止的时间戳, 以便提取数据时更方便地将这些长序列中

剪切成短片段，且这些动作在录制中也不会告知驾驶员，以防打扰到被试，与其平时的驾驶习惯产生差异；(4) 在行驶约 20 min 后，被试者会停下来休息一段时间，这是由于较长时间记录可能会导致眼动仪对瞳孔的跟踪产生误差，也避免因被试者的疲劳造成驾驶行为变化。因此，每次短暂休息后需要对眼动仪重新校正。每个被试者每次将记录 2~3 个序列，单次总记录时间控制在 1 h 左右。

数据库中记录的眼动视频图像如图 1 所示，其中红色圆圈表示驾驶员当前的注视区域。图 1(a) 中给出了一个左转驾驶行为片段中的第 1 帧、第 51 帧与第 101 帧的示例图，可以看出该片段光照变化强烈且存在视线受限或遮挡等情况。图 1(b) 中分别是 VIPDAR_5 数据库中不同时间与天气下录制的驾驶行为视频帧，时间不同因此光照条件区别很大，且雨天前挡风玻璃上雨水也会增加驾驶员注视的难度。因此该数据库较现有动作识别数据库更具挑战性。



(a) 左转视频序例示例图　　(b) 不同时间天气示例

图 1　左转视频片段中的示例图以及不同环境的视频帧

Fig. 1　Example of left turn and frames in different conditions

2.2　数据筛选及标注

在获得所有被试眼动驾驶视频后，将对这些数据进行筛选与标注。首先去除采样正确率在 80% 以下的数据以及一些明显受其他事物影响的驾驶行为片段，接着将眼动数据与车辆驾驶视频数据进行时间上进行对齐，最后在不改变帧速率和分辨率的情况下，根据记录的时间戳对这些长视频序列进行分割并按照动作类型进行分类整理和排序，每个视频片段的长度在 10 s 左右。

2.3　VIPDAR_5

考虑到数据多样性，VIPDAR_5 中包含了不同场景下不同路况、环境、天气和时间，数据库具体内容如表 2 所示。

表 2　VIPDAR_5 中视频片段多样性

Table 2　Diversity of video clips in VIPDAR_5

属性	内容
驾驶员	10人
录制时长	约11 h
视频数目	25段长序列
视频片段数	350段
场景	校园、高速公路、闹市区、小巷、停车场
动作种类	左转、右转、直行、左变道、右变道
时间	早晨、中午、傍晚、夜晚、凌晨
天气	晴天、多云、雨天
背景	轿车、公交车、摩托车、自行车、行人

根据动作类型将驾驶员行为分为 5 个主要类别。与其他行为识别数据库不同，本数据库根据驾驶员眼动关注点与眼动轨迹对视频片段进行分类：左转、右转、左变道、右变道、直行，选择这几类驾驶行为是因为它们在日常驾驶较为常见，方便采集。

图 2 是 VIPDAR_5 与 UCF101 视频图像对比，图 2(a) 是本数据库中直行片段的某帧示例图，图 2(b) 和 (c) 分别是 UCF101 中 JumpRope 和 HorseRiding 类中某片段的一帧。从图中可以发现我们的数据库图像分辨率更高且包含驾驶员注视点信息。

(a) VIPDAR_5 视频图像　　(b) UCF101 中 JumpRope 视频图像　　(c) UCF101 中 HorseRiding 视频图像

图 2　VIPDAR_5 与 UCF101 视频图像对比

Fig. 2　Video image comparison between VIPDAR_5 and UCF101

图 3 给出了 VIPDAR_5 数据库中每个类的视频片段数量，两种不同的颜色分别表示数据库中白天和夜晚的视频片段。表 3 中给出了不同驾驶行为的帧数、总时长、平均时长以及片段分布情况，从中可以看到，这五类中左转和右转的时间比其他类稍长，且 10~15 秒片段的数量比例较大，这是因为驾驶员在左转或右转时会考虑更多的交通路况信息。

这些视频片段中有的是同一条路上的不同时间，有的是同一时间下的不同道路，以及不同天气情况下记录的。由于雨水对驾驶者视线的

影响, 在雨天录制的驾驶员眼动视频数据更具挑战性。

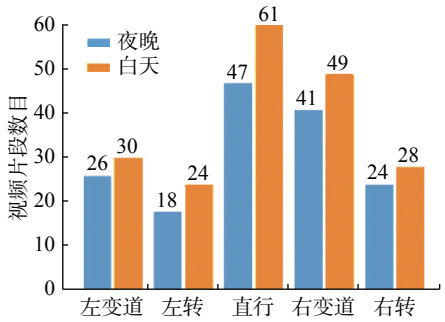


图 3　白天与夜晚的视频片段在各类中的分布

Fig. 3　Number of daytime and night video clips per action class

表 3　VIPDAR_5 数据库各类数据统计

Table 3　Summary of per action class on VIPDAR_5

类名	总帧数	总时长/s	片段平均时长/s	5~10 s 片段数	10~15 s 片段数
左变道	14 154	472	8.43	39	17
左转	13 227	441	10.50	15	27
直行	28 339	944	8.74	71	37
右变道	22 321	744	8.27	62	28
右转	14 495	483	9.29	25	27

3　双通路模型 TWNet

在日常生活中, 若是要准确地描述一个物体, 可能需要一个清晰的图像。然而, 如果只是为了识别它的动作, 只需要给出几副连续的低分辨率图像甚至是剪影, 通过分析物轮廓和运动信息, 就可以得出它的行为类别。

根据神经学与生理学研究, 根据神经元种类和连接方式, 人脑视觉皮层包括纹状皮层和纹外皮层两类, 纹状皮层通常也称为初级视皮层即视觉第一区域 V1, 纹外皮层包含视觉第二三四五区域即 V2、V3、V4、V5。初级视皮层接受到来自视网膜经外侧膝状体的信息, 再通过两个通道输出, 这两个通道分别称为背侧流 Dorsal stream 和腹侧流 Ventral stream。背侧流常被称为空间通路, 通常认为由 V1、V2 和 V5 等组成, 参与处理物体的空间位置信息, 确定物体的空间位置。而腹侧流常被称为内容通路, 通常认为由 V2 和 V4 等组成, 参与物体识别, 确定物体的形状、颜色等, 该通路也与长期记忆有关。

根据人类大脑对事物认知的视觉皮层结构, 本文设计了 TWNet(What-Where Network), 它具有双通路结构, 分别为 What 通路和 Where 通路, 用于捕获驾驶员眼动视频数据的图像语义信息和空间运动信息。如图 4 所示, 其中蓝色部分为 Where 通路, 黄色部分为 What 通路, 参考 Res3D_18 模型的设置, 具体的网络结构如表 4 所示。

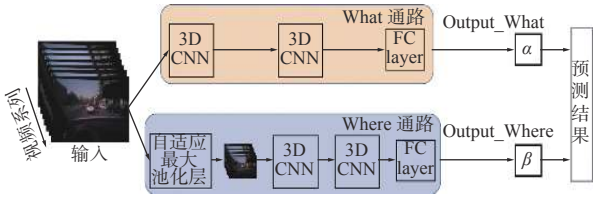


图 4　TWNet 网络结构示意

Fig. 4　Example of the proposed architecture TWNet

表 4　TWNet 的网络结构

Table 4　The TWNet architecture

层名	What通路	Where通路	输出尺寸/像素×像素
Raw clip	—	自适应最大池化层	What: 32×64² Where: 32×8²
Conv	1×7×7, 64	3×7×7, 64	What: 32×64² Where: 32×8²
Conv_1	$\begin{bmatrix} 3\times3\times3,64 \\ 3\times3\times3,64 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3\times3,64 \\ 3\times3\times3,64 \end{bmatrix}\times2$	What: 32×64² Where: 32×8²
Conv_2	$\begin{bmatrix} 3\times3\times3,128 \\ 3\times3\times3,128 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3\times3,128 \\ 3\times3\times3,128 \end{bmatrix}\times2$	What: 16×32² Where: 16×4²
Conv_3	$\begin{bmatrix} 3\times3\times3,256 \\ 3\times3\times3,256 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3\times3,256 \\ 3\times3\times3,256 \end{bmatrix}\times2$	What: 8×16² Where: 8×2²
Conv_4	$\begin{bmatrix} 3\times3\times3,512 \\ 3\times3\times3,512 \end{bmatrix}\times2$	$\begin{bmatrix} 3\times3\times3,512 \\ 3\times3\times3,512 \end{bmatrix}\times2$	What: 4×8² Where: 4×1²
Fc	α	β	—

由上述分析可知，图像的清晰度并不会对物体动作的识别造成很大的影响。如图 4 所示，TWNet 中 Where 通道内第一个自适应最大池化层模块对输入的图像进行处理，降低图像分辨率，在不影响输入帧数的情况下，使得整体网络的参数量下降，降低网络复杂性，同时因为图像分辨率下降，该通道能更加关注于帧与帧之间的运动信息，从而更好地捕获眼动视频中的时序信息。

本方法主要利用自建的眼动数据库 VIPDAR_5 中的眼动信息，如视点轨迹，注视位置等来对这些驾驶行为进行分类。而背景信息主要用于判断眼动信息相似的类别，如左转和左变道，右转和右变道。我们在两个通道的输出各设置了一个超参数 α 和 β，通过调整这两个超参数的值来控制最终输出结果中图像信息与运动信息的权重，从而更好的让网络分类驾驶员眼动视频数据。

$$\begin{aligned}&\text{Output\_mix} = \alpha \times \text{Output\_What} + \beta \times \text{Output\_Where}\\&\text{s.t.}\alpha + \beta = 1\end{aligned} \tag{1}$$

式中：Output_What 是 What 通路的预测输出；Output_Where指眼动视频数据经 Where 通路得到的结果；Output_mix是在 α 和 β 超参数控制下得到的最终输出，两参数 α 与 β 之和为 1。由于在 Where 通路中加入了最大池化层降低输入分辨率，所以单独通路识别准确率并不高，需要通过设置权重，使得两通道获得的图像和运动信息融合起来而不影响最终输出准确率。针对双通路的消融实验以及超参数的具体设置将在下一章实验与结果中详细描述。

4　实验和结果

在本节中，首先简要地介绍了 5 种基于卷积神经网络的行为识别方法。接着给出了本文实验环境以及对这五种方法的调整，最后通过消融实验证明了 TWNet 网络双通道的优越性，对超参数进行设置说明 Where 通路捕获视频动作信息的有效性。并将这五种相关方法在眼动驾驶行为视频数据库 VIPDAR_5 上进行训练和测试，比较它们的 Top-1 和 Top-3 性能，对实验结果进行分析。

4.1　对比方法

本文将对比近年来行为识别领域中的五种常见模型包括：C3D[28]，Conv+LSTM[24]，Res3D[30]，R(2+1)D[31]，SlowFast[32]。下面将简要介绍方法设置：C3D 包含 8 个卷积层、5 个池化层、2 个全连接层和一个 Softmax 层；Conv+LSTM 中使用了 ResNet50 作为底层网络，同时在全连接层输出后添加了一个有 300 个隐藏层节点的 LSTM 用来存储视频时序信息，并添加 2 个全连接层用来进行视频分类。对于 Res3D 和 R(2+1)D 方法，考虑到数据量大小，选择使用网络深度较浅的 Res3D_18 和 R(2+1)D_18。SlowFast 使用 ResNet101 作为底层网络。同时在对比实验中不改变这些方法中的其他模块结构。

4.2　实验环境与实现细节

VIPDAR_5 中将训练集和验证集的视频片段数目划分比例设置为 2:1，同时保证它们包含不同时间和天气的的视频片段且数目比例基本相同。

上述对比方法直接在 VIPDAR_5 的训练集上训练，并在数据读取阶段保持输入图像尺寸为 64×64，训练采用 Adam 优化算法，批处理（batchsize）大小设置为 4，学习率设置为 0.000 1，且设置阶梯学习率，每迭代 20 次，学习率下降 10 倍。若验证集的损失若在迭代 10 次之后仍不下降，那么训练结束。所有的实验都是基于开源框架 Pytorch，操作系统为 Windows 7，处理器型号为 Intel i5-9400F CPU，显卡型号为 NVIDIA RTX2070 GPU。

4.3　实验结果与分析

对本文所提出的 TWNet 方法的两通道性能进行了验证，通过将超参数 α 和 β 分别设置为 0 和 1 以及 1 和 0，得到了 TWNet 单独使用某一通道时的性能表现。

消融实验结果如表 5 所示，前两行数据展示了仅单独使用某通路结构时的分类准确率，可以看到单独使用 What 和 Where 通道时的 Top-1 和 Top-3 准确率都低于同时使用双通道时的结果。

表 5　TWNet 消融实验结果

Table 5　Results of the ablation experiment　%

通路	Top-1	Top-3
What-only	63.4	90.0
Where-only	52.5	82.3
What-Where	**66.9**	**93.3**

表 6 给出了 VIPDAR_5 中各行为类准确率在不同超参数设置下的结果对比。当 α=0.7，β=0.3 时，What-Where 的 Top-1 和 Top-3 准确率达到了 66.9% 和 93.3%，相较于单独使用某一通路分别提高了 3.5% 和 3.3%。通过调整两通道的权重，最终准确率得到了提高并超过了单独使用 What 的

结果, 这是因为 Where 通路弥补了 What 通路对时序运动信息不敏感的弱点。经过两通路输出的结合, 识别这些动作持续时间短、幅度大类别的能力增强, 所以准确率提升了。

表 6　VIPDAR_5 中各行为类准确率在不同超参数情况下的结果对比

Table 6　Accuracy comparison of each action class under different hyperparameters on VIPDAR_5

α	β	Top-1/%	Top-3/%	左变道/%	左转/%	右变道/%	右转/%	直行/%
0	1	52.5	82.3	39.0	54.4	35.5	52.0	65.8
1	0	63.4	90.0	45.2	62.2	53.4	58.5	79.0
0.5	0.5	64.3	92.8	46.3	62.0	54.3	59.6	77.1
0.6	0.4	65.0	92.8	50.7	65.6	57.0	59.7	78.0
0.7	0.3	**66.9**	**93.3**	**52.8**	**67.0**	57.3	**61.6**	**80.0**
0.8	0.2	64.6	92.9	51.7	65.1	**60.0**	59.6	79.4
0.9	0.1	64.5	92.4	48.4	63.6	57.0	58.7	78.7

表 7 中给出了 5 种对比方法在 VIPDAR_5 上的实验结果, TWNet 相较于其他对比算法, Top-1 和 Top-3 分别提高了 8.0% 和 1.7%。另外, SlowFast 算法在眼动驾驶行为数据集中的性能表现并不好, 这可能是由于 VIPDAR_5 数据集与 UCF101 等其他数据集之间的差异, 如数据集中左右变道类的动作持续时间较短, 眼动轨迹点变化剧烈, 并可能存在遮挡的问题。

表 7　不同方法在 VIPDAR_5 上的实验结果对比

Table 7　Comparison of the results of different methods on VIPDAR_5　%

方法	Top-1	Top-3
C3D[28]	52.9	83.4
Conv+LSTM[24]	40.9	81.8
Res3D[30]	53.7	85.8
R(2+1)D[31]	58.9	91.6
SlowFast[32]	43.1	80.7
TWNet(Ours)	**66.9**	**93.3**

TWNet 网络中加入的时空信息权重模块, 通过设置两通道的不同权重控制了运动信息对最终结果的影响, 使得分类准确率获得了提升。同时针对不同的数据库, 也可以通过调整时空信息权重及时适配。

5　结束语

本文首先构建了基于眼动的驾驶员行为识别视频数据集 VIPDAR_5。它包含多种路况、天气、时间情况的共超 9 万帧的 350 个视频片段。针对眼动行为识别任务, 提出了基于人类视觉的双通道模型 TWNet, 实验结果表明该模型具有良好的性能。同时本文还有一些不足, 因为存在两个通路的网络, 所以训练时间较长参数量较多, 在后续研究中将会针对该问题进一步优化。本文希望通过分析驾驶员的眼动数据, 理解驾驶心理, 总结驾驶习惯, 在后续研究中根据这些信息, 预测驾驶员的注视区域及行为, 提前对进行预警, 辅助驾驶, 提高道路安全。

参考文献:

[1] 国家统计局. 中华人民共和国 2019 年国民经济和社会发展统计公报 [N]. 人民日报, 2020-02-29(5).

[2] JAIN D K, JAIN R, LAN Xiangyuan, et al. Driver distraction detection using capsule network[J]. Neural computing and applications, 2021, 33(11): 6183–6196.

[3] LE T H N, ZHENG Yutong, ZHU Chenchen, et al. Multiple scale faster-RCNN approach to driver's cell-phone usage and hands on steering wheel detection[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops. New York, USA: IEEE, 2016: 46−53.

[4] 王荣本, 郭克友, 储江伟, 等. 适用驾驶员疲劳状态监测的人眼定位方法研究 [J]. 公路交通科技, 2003(5): 111–114.
WANG Rongben, GUO Keyou, CHU Jiangwei, et al. Study on the eye location method in driver fatigue state surveillance[J]. Journal of highway and transportation research and development, 2003(5): 111–114.

[5] 张杰. 基于眼动仪的驾驶员视点分布特性研究 [J]. 湖南交通科技, 2012, 38(4): 153–155,170.
ZHANG Jie. Driver's viewpoint distribution based on the

eye tracker[J]. Hunan communication science and technology, 2012, 38(4): 153–155,170.

[6] 袁伟, 徐远新, 郭应时, 等. 车道变换与直行时的驾驶人注视转移特性 [J]. 长安大学学报(自然科学版), 2015, 35(5): 124–130.
YUAN Wei, XU Yuanxin, GUO Yingshi, et al. Fixation transfer characteristics of drivers during lane change and straight drive[J]. Journal of chang'an university (natural science edition), 2015, 35(5): 124–130.

[7] MISHKIN M, UNGERLEIDER L G, MACKO K A. Object vision and spatial vision: two cortical pathways[J]. Trends in neurosciences, 1983, 6: 414–417.

[8] KOOTSTRA G, DE BOER B, SCHOMAKER L R B. Predicting eye fixations on complex visual stimuli using local symmetry[J]. Cognitive computation, 2011, 3(1): 223–240.

[9] SOOMRO K, ZAMIR A R, SHAH M. UCF101: a dataset of 101 human actions classes from videos in the wild [EB/OL]. (2012-12-1) [2021-05-30]. https://arxiv.org/abs/1212.0402.

[10] KAY W, CARREIRA J, SIMONYAN K, et al. The kinetics human action video dataset[EB/OL]. (2017-05-19) [2021-05-30]. https://arxiv.org/abs/1705.06950.

[11] SIGURDSSON G A, GUPTA A, SCHMID C, et al. Charades-ego: a large-scale dataset of paired third and first person videos[EB/OL]. (2018-04-30) [2021-05-30]. https://arxiv.org/abs/1804.09626.

[12] DAMEN Dima, DOUGHTY H, FARINELLA G M, et al. Scaling egocentric vision: the dataset[M]//Computer Vision – ECCV 2018. Cham: Springer International Publishing, 2018: 753−771.

[13] JIANG Lai, XU Mai, WANG Zulin. Predicting video saliency with object-to-motion CNN and two-layer convolutional LSTM[EB/OL]. (2017-09-19) [2021-06-30]. https://arxiv.org/abs/1709.06316.

[14] Li Y, Liu M, Rehg J M. In the eye of beholder: Joint learning of gaze and actions in first person video[C]// 2018 European Conference on Computer Vision. Berlin, German: Springer, 2018: 619−635.

[15] YIN Li, YE Zhefan, REHG J M. Delving into egocentric actions[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2015: 287−295.

[16] MATHE S, SMINCHISESCU C. Actions in the eye: dynamic gaze datasets and learnt saliency models for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 37(7): 1408–1424.

[17] MARSZALEK M, LAPTEV I, SCHMID C. Actions in context[C]//2009 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2009: 2929−2936.

[18] RODRIGUEZ M. Spatio-temporal maximum average correlation height templates in action recognition and video summarization[EB/OL]. (2013-12-10) [2021-06-30]. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.221.5006.

[19] JUDD T, EHINGER K, DURAND F, et al. Learning to predict where humans look[C]//2009 IEEE 12th International Conference on Computer Vision. New York, USA: IEEE, 2009: 2106−2113.

[20] PAPADOPOULOS D P, CLARKE A D F, KELLER F, et al. Training object class detectors from eye tracking data[C]//Computer vision–ECCV 2014. Berlin, German: Springer, 2014: 361−376.

[21] EVERINGHAM M, GOOL L V, WILLIAMS C K I, et al. The pascal visual object classes (VOC) challenge[J]. International journal of computer vision, 2010, 88(2): 303–338.

[22] JI Shuiwang, XU Wei, YANG Ming, et al. 3D convolutional neural networks for human action recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2013, 35(1): 221–231.

[23] SIMONYAN K, ZISSERMAN A. Two-stream convolutional networks for action recognition in videos[EB/OL]. (2014-11-12) [2021-06-30]. https://arxiv.org/abs/1406.2199.

[24] NG Joey H, HAUSKNECHT M, VIJAYANARASIMHAN S, et al. Beyond short snippets: Deep networks for video classification[C]//2015 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2015: 4694−4702.

[25] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735–1780.

[26] WANG Limin, XIONG Yuanjun, WANG Zhe, et al. Temporal segment networks: towards good practices for deep action recognition[C]//Computer vision–ECCV 2016. Berlin, German: Springer, 2016: 20−36.

[27] LIN Ji, GAN Chuang, HAN Song. TSM: temporal shift module for efficient video understanding[C]//2019 IEEE/CVF International Conference on Computer Vision. New York, USA: IEEE, 2019: 7082−7092.

[28] TRAN D, BOURDEV L, FERGUS R, et al. Learning spatiotemporal features with 3D convolutional networks [C]//2015 IEEE International Conference on Computer

Vision. New York, USA: IEEE, 2015: 4489−4497.

[29] HE Kaiming, ZHANG Xiangyu, REN Shaoqing, et al. Deep residual learning for image recognition[C]//2016 IEEE Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2016: 770−778.

[30] TRAN D, RAY J, SHOU ZHENG, et al. ConvNet architecture search for spatiotemporal feature learning [EB/OL]. (2017-8-16) [2021-06-30]. https://arxiv.org/abs/1708.05038.

[31] TRAN D, WANG Heng, TORRESANI L, et al. A closer look at spatiotemporal convolutions for action recognition[C]//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. New York, USA: IEEE, 2018: 6450−6459.

[32] FEICHTENHOFER C, FAN Haoqi, MALIK J, et al. SlowFast networks for video recognition[C]//2019 IEEE/CVF International Conference on Computer Vision. New York, USA: IEEE, 2019: 6201−6210.

[33] 彭金栓, 高翠翠, 郭应时. 基于熵率值的驾驶人视觉与心理负荷特性分析 [J]. 重庆交通大学学报(自然科学版), 2014, 33(2): 118–121.
PENG Jinshuan, GAO Cuicui, GUO Yingshi. Drivers' visual characteristics and mental load based on entropy rates[J]. Journal of Chongqing Jiaotong University (natural science edition), 2014, 33(2): 118–121.

[34] 袁伟, 付锐, 马勇, 等. 车速与标志文字高度对驾驶人视觉搜索模式的影响 [J]. 交通运输工程学报, 2011, 11(1): 119–126.
YUAN Wei, FU Rui, MA Yong, et al. Effects of vehicle speed and traffic sign text height on drivers' visual search patterns[J]. Journal of traffic and transportation engineering, 2011, 11(1): 119–126.

作者简介:

申天啸, 硕士研究生, 主要研究方向为深度学习、人类眼动行为、行为识别。

韩怡园, 博士研究生, 主要研究方向为深度学习、人类视觉注意和人类眼动行为。

韩冰, 教授, 博士生导师, 主要研究方向为模式识别、计算机视觉和极光影像分析。主持和参与国家自然科学基金重点项目、国家自然科学基金面上项目、中国博士后一等资助项目、海洋公益项目和青年项目等, 发表论文 30 余篇, 授权国家发明专利 13 项, 其中成果转化 1 项。获省科学技术进步奖 2 项、省高等学校科学技术一等奖 1 项。