引用格式：吴晓雨，闵静萱，邱驹成，吴建琴.基于机位计算的云演艺智能虚拟拍摄系统[J].中国传媒大学学报（自然科学版），2023，30（04）：17-26.
文章编号：1673-4793（2023）04-0017-10

基于机位计算的云演艺智能虚拟拍摄系统

吴晓雨*，闵静萱，邱驹成，吴建琴
（中国传媒大学信息与通信工程学院，北京 100024）

摘要：云演艺智能虚拟拍摄对于降低拍摄专业门槛和空间人力等成本具有重要意义。本文提出了以机位计算为核心、美学评估为辅助的基于数据驱动的智能拍摄仿真技术，构建虚拟环境下的云演艺智能拍摄系统。首先提取真实视频中的人物关节点信息，然后判断主要演员并构建复曲面坐标系。接着用特征估计网络提取出相机和人物特征，通过相机运动提取门控网络和相机轨迹预测网络计算得到相机的运动参数。然后在Unity3D环境下，控制虚拟相机并生成相机轨迹和相应画面。再通过美学评价模型对画面进行整体美学评分和主要构图模式判断，选取较优机位。最终，用户可以在小程序端得到较优机位下的运镜轨迹和虚拟环境的仿真画面视频。以中国传媒大学礼堂和威海荣成西霞口剧场的虚拟场景为例展示了智能拍摄系统的效果。
关键词：智能拍摄；机位计算；美学评估；云演艺
中图分类号：TP391.9　**文献标识码：**A

Virtual automatic cinematography system based on virtual camera position calculation for cloud performing arts

WU Xiaoyu*, MIN Jingxuan, QIU Jucheng, WU Jianqin
(School of Information and Communication Engineering, Communication University of China, Beijing 100024, China)

Abstract: Virtual automatic cinematography is of great importance to lower the professional threshold for filming and reduce space cost and labor cost of shooting. In this paper it was proposed to use data-driven automatic cinematography technology with aesthetic assessment technology to build an automatic cinematography system for cloud performing arts. Firstly, extracted character joint point in the real video, and then the main actors were judged and the toric space was constructed. Secondly, the feature estimation network was used to extract the camera features and character features, and the motion parameters of the camera were calculated through camera motion extraction gating network and camera trajectory prediction network. After that, these parameters controlled the virtual camera in Unity3D to generate camera trajectory and video. Then through the aesthetic assessment model, the overall aesthetic score of the picture was marked and the main composition pattern was judged. And the system selected the camera position with the best performance. Finally, the user could get the track of camera and the video captured by the chosen camera on Wechat MiniProgram. The virtual environment based on the auditorium of Communication University of China and Xixiakou Theater of Weihai, Rongcheng was taken as an example to demonstrate the effect of the intelligent shooting system.

基金项目：国家重点研发计划课题（2021YFF0900701）
作者简介（*为通讯作者）：吴晓雨（1979-），女，博士，教授，主要从事视频智能分析技术研究。Email：wuxiaoyu@cuc.edu.cn；闵静萱（2001-），女，本科生，主要从事视频智能分析技术研究。Email：3198672528@qq.com

Keywords: automatic cinematography; camera position calculation; aesthetic assessment; cloud performing arts

1 引言

伴随前沿技术的迭代升级,传统拍摄对于智能拍摄的需求日益强烈,机位计算的关键技术也有了更高的要求。演艺场景真实拍摄过程中,面临着空间局限、人为干扰和成本较高等问题,如何通过AI技术实现节目拍摄的高效低成本轻量化制作,无疑是非常重要的。此外,高质量拍摄离不开专业镜头语言知识,一定程度上限制了非专业人士参与,而智能化拍摄的研究将有助于降低数字演艺内容的拍摄门槛。与此同时,近年来随着虚拟世界以及元宇宙等理念的出现,利用Unity和Unreal等虚拟引擎能够较为真实地还原现实世界,使用虚拟仿真环境也可避免现实条件下时间、空间及成本的限制。

本文主要研究了面向云演艺的智能拍摄仿真技术,在基于真实视频学习的数据驱动方式基础上加入了美学评估,进而反映镜头的艺术表征。通过美学评估技术对数据驱动下计算的机位进行微调,构建以机位计算为核心、美学评估为辅助的基于数据驱动的智能拍摄系统,从而有效验证和辅助指导真实拍摄环境下的高质量拍摄,降低拍摄的专业门槛,提升拍摄智能化水平。

2 相关研究

目前对于一个演艺节目而言,真实视频拍摄过程一般包括场景设计、表演安排、情节设计和相机位置设定等多个步骤。每个拍摄的镜头都受到上述因素的影响,由于客观条件的限制,视频制作者可能无法拍摄出最好的镜头。同时,这类节目由于其不确定性,导演往往会选择提前进行一次预演的拍摄,这需要耗费大量的人力物力。所以高成本和低可控性是目前演艺节目提前预演的主要问题,如果能够在正式拍摄前在虚拟场景中进行模拟并由AI提供几种较好的相机拍摄方案,将有助于导演快速找到最佳拍摄方案,提高节目的制作效率并且节省大量的成本消耗。

在过去的时间里,许多研究人员提出了不同的自动化相机拍摄的方法。

2.1 传统的智能拍摄方法

传统的方法大多数使用了机器学习和脚本套用的方法。从视频中提取镜头知识,从而构造了一个镜头库,使用镜头时,通过脚本直接调用该镜头的相关参数应用到相机上,但是该方法过于简单,使用起来有诸多限制。Wang等[1]建立了视频素材库并标注了关键词,用户输入脚本后,将相符的素材库中的视频合成输出。Xiong等[2]提出一个弱监督的框架,使用脚本作为输入,从广泛的镜头集合中自动创建视频序列。Chen等[3-5]使用循环决策树网络训练了一个三自由度的相机位姿预测器来自动拍摄篮球和足球比赛,可以通过运动对象和当前相机的位姿来预测下一个相机的最佳拍摄角度,但该种方式简化了相机的参数,只能应用于室内固定机位的拍摄任务。Jia等[6]则是使用具有目标玩家运动行为数据的决策树网络在空中自动拍摄虚拟开放游戏中建筑物的视频。

2.2 采用深度学习的智能拍摄方法

随着深度学习的兴起,其在自动摄影中也被广泛应用,Huang等[7]采用序列到序列(seq2seq)的结构来进行单人室外运动视频自动拍摄,结合时间和空间信息,根据当前的位置和运动状态预测下一帧中相机的光流,再根据光流和相机的参数矩阵来算出相机的坐标,但该种方式较为复杂且误差较大。在最新的虚拟相机研究中,Jiang等[8-9]使用复曲面空间坐标[10]来代替传统的六自由度空间坐标,减少了因坐标系产生的误差,然后使用真实电影视频提取出它们的拍摄风格,再应用到虚拟场景中来驱动相机的拍摄,并且加入了关键帧技术使得用户可以进行精细化的相机控制,然而由于坐标系的原因,只能应用到两个人的场景中,十分受限。Yu等[11]提出了一个自动动画电影拍摄的框架T2A,对于给定的虚拟场景和脚本,该框架可以自动拍摄出符合脚本内容的镜头,其中自动拍摄优化使用逼真度和美学模型来进行联合优化,在优化过程中可以共同考虑输入脚本的视觉呈现以及生成的视频与给定电影技术规范的合规性。

近期最新的研究中,模仿学习开始被应用到自动摄影中,而RT2A[12]则是在T2A[11]的基础上加入了强化学习的内容,提出了一个奖励函数来指导算法找到最佳拍摄策略并模仿导演对每个场景的相机选择的决策过程,其实验结果表明,所提出的RT2A可以有效地模仿导演对镜头语言模式的使用。文献[13]提出了

一个基于强化学习的无人机的自动拍摄程序，可以实时跟随移动的演员，同时根据镜头设计做出实时的决策，这些决策是基于通过强化学习得到的经验得到的。Dang等[14]为无人机摄影系统提出了一个端到端的模仿学习框架，提出了基于路径分析的强化学习(PABRL)算法，由人物运动信息、图像构图特征和相机运动矩阵得到人物运动相关的美学拍摄策略，同时使用了一种注意力机制和一种长短期奖励函数，分别增强运动特征空间和生成轨迹的完整性。文献[15]提出了一个集成的航拍系统，用于自动捕捉动作场景的电影镜头，通过模仿观看主体运动的演示来学习预测下一个相机的最佳视点。

然而对于演艺场景而言，其运镜要求更为专业，所得到的拍摄效果美学要求更高，场景也较为复杂。基于脚本驱动的智能拍摄仿真技术缺乏对于环境和摄像机的控制，虚拟环境下生成内容实用性较差；基于真实视频学习的数据驱动方式，忽略了镜头语言美学属性的学习，得到的内容无法有效体现构图和拍摄意图。因此，本文在基于数据驱动的智能拍摄基础上，通过美学评估来反应镜头的艺术表征，并基于此技术研发了面向云演艺的智能拍摄系统。

3 智能拍摄系统设计与实现

3.1 系统整体框架设计

系统的框架图如图1所示，前端为用户操作的微信小程序，设计常用的按钮、界面，用户可以通过上传运镜参考视频来得到理想的运镜轨迹和虚拟仿真的效果视频。后端分为核心算法、数据库以及Unity虚拟场景，前后端通过API接口传递信息。后端算法模型部分是系统的核心，包含关节点提取、特征提取、相机运动提取、相机轨迹预测和美学评价这五个模块。前4个模块根据用户上传参考视频来预测合适的相机参数，并在虚拟环境中仿真，得到轨迹及输出视频，最后的美学评估网络模块则是对相机参数的仿真画面进行美学评分预测和主要构图模式判断，并进行拍摄指导，进一步选择最合适的机位；内嵌数据库用于存储用户信息、视频数据；Unity虚拟环境则是把算法模型的参数结果进行仿真，得到更用户友好的运镜轨迹和仿真视频。最终，将经过美学指导的仿真结果（相机运动轨迹图及虚拟相机拍摄画面）返回前端小程序并呈现给用户。

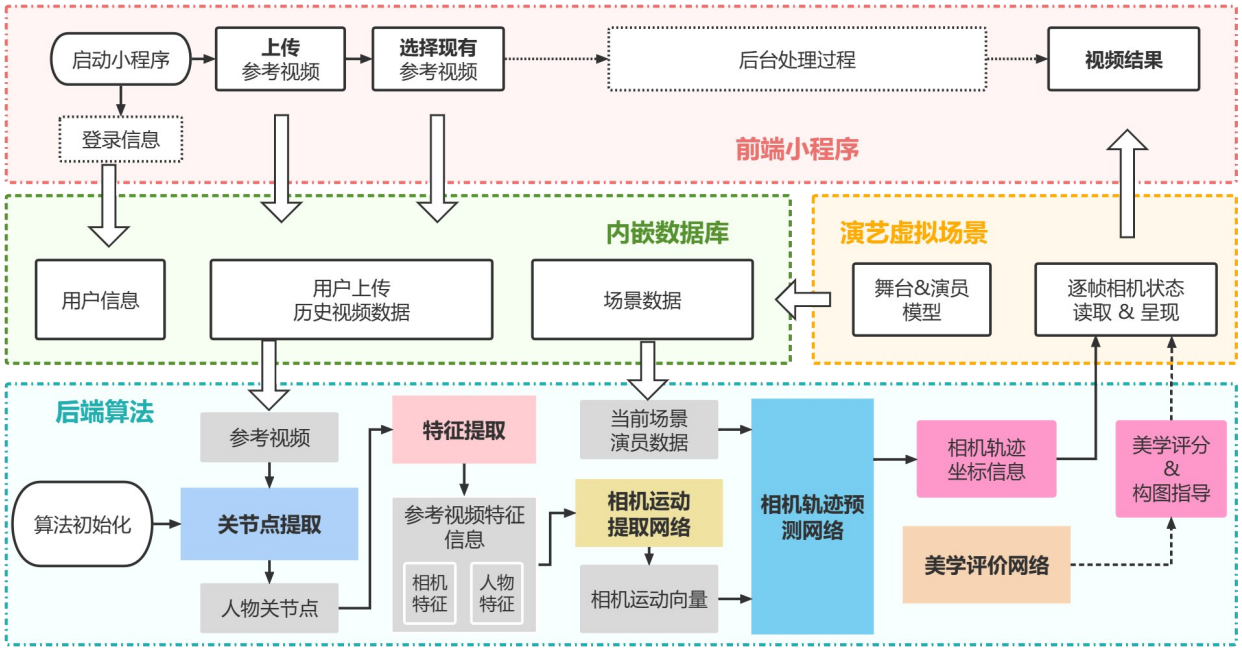

图1 系统框架图

3.2 后端机位计算算法

本节主要介绍智能拍摄的核心算法模型——基于Unity虚拟环境的相机参数计算模型。如前文所述，该模型包括关节点提取、特征提取网络、相机运动提取、相机轨迹预测和美学评估等五个模块。机位计算模型的结构图如图2所示。

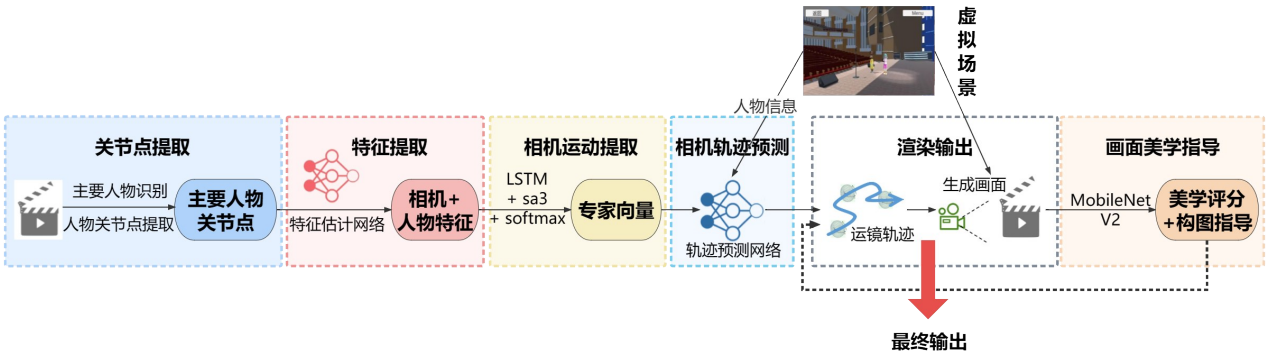

图2 机位计算模型结构图

3.2.1 关节点提取

该模块首先将根据用户所选择的特定镜头切换方式或影片拍摄风格在数据库中选取合适的影片片段,或者直接使用用户上传的运镜参考视频,然后使用DEKR模型[16]提取其中的人物关节点信息。该模型采用了自底向上的方法,使用自适应卷积激活关键区域的像素,进而能够聚焦人物关节点区域,返回其空间信息。

在构建坐标系时,本系统采用了复曲面坐标系,即Toric空间[10]。该空间是以两个演员为参考点构建的曲面坐标系,相比起传统的直角坐标系,Toric空间可以将相机和演员构建在一个坐标系中,避免了坐标估计和转换带来的误差,同时也更容易理解相机的运动控制。但是缺点是只能受限于两个人,不能多也不能少。

为了解除视频人数的限制,我们将上述方法拓展到了多人,使得网络能接受多人的视频输入,从中提取出两个主要人物的关节点信息。通过人脸到屏幕中心距离和人脸大小来判断出主要演员[17],同时考虑人物的大小、距离屏幕中心的距离以及人物的清晰度,最终得到每个人的重要性分数,选取分数最高的两名人物提取其关节点。

3.2.2 特征提取网络

得到两个主要人物(A,B)关节点后,我们使用一个回归网络,即特征提取网络[8]来提取相机特征 $\mathbf{c}$ 和人物特征 $\mathbf{v}$,在Toric空间中的相机、人物特征分别如下:

$$\mathbf{c} = \{\mathbf{p}_A, \mathbf{p}_B, \theta, \varphi\} \in \mathbb{R}^6 \quad (1)$$

$$\mathbf{v} = \{d_{AB}, s_A, s_B, s_{AB}, M\} \in \mathbb{R}^5 \quad (2)$$

其中,$\mathbf{p}_A$ 和 $\mathbf{p}_B$ 表示人物的屏幕位置,为二维信息,θ 和 φ 是两个参数角,直观地表示两个主要人物之间偏航角和俯仰角;d_{AB} 是两个演员之间的3D距离,s_A 是直线AB和演员A肩膀正交向量的夹角,$s_{AB} = s_A + s_B$ 代表了A和B的朝向的不同,M 代表是否为主要人物,$\mathbb{R}^5$、$\mathbb{R}^6$ 分别表示5维和6维实数空间。。

将演员的骨骼关节点输入特征提取网络,即可提取出画面中对应的人物特征 θ、φ 和相机特征 d_{AB}、s_A、s_B、s_{AB}。

3.2.3 相机运动提取

将运镜参考视频的一段序列输入相机运动提取门控网络,其中每一帧都包含之前提取的相机参数和人物信息,然后使用长短期记忆(Long Short-Term Memory,LSTM)网络将参考视频逐帧输入。网络在LSTM的基础上还加入了多头自注意力机制(head=3),在提升模型泛化能力的同时,也能更好地捕捉长距离的上下文信息。

选取最后一帧输出,经过Softmax和全连接层得到一个和为1的专家向量,即成功将相机运动特征整合成一个专家向量,该专家向量为用户选取的2D视频相机行为的压缩表示。相机运动提取模型结构如图3所示。

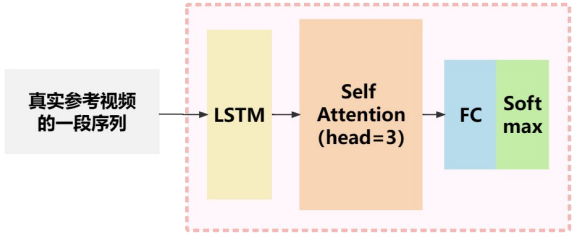

图3 相机运动提取模型结构图

3.2.4 相机轨迹预测

相机轨迹预测模块的设计参考文献[9],将相机运动特征整合成一个专家向量(即相机行为压缩向量),送入到混合专家(MOE-Mixture of Experts)网络中,网络尝试把样例视频识别为训练数据集中的一种或几种相机行为的组合,赋予不同权重后送入预测网络进

行相机运动轨迹预测。

相机轨迹预测网络由三个全连接层构成，采用ELU激活函数，输入新场景的人物信息后，在专家向量和人物信息的控制下逐帧生成新场景的相机参数信息。最终输出信息为Toric坐标下5维参数：

$$\mathrm{x}_i^{\mathrm{c}} = \left\{ x_{\mathrm{A}}, x_{\mathrm{B}}, mean\left(y_{\mathrm{A}}, y_{\mathrm{B}}\right), \theta, \varphi \right\} \in \mathbb{R}^5 \tag{3}$$

其中，$\left(x_{\mathrm{A}}, y_{\mathrm{A}}\right)$、$\left(x_{\mathrm{B}}, y_{\mathrm{B}}\right)$表示两个演员头部在帧中的位置坐标，取y坐标的平均值确保相机在坐标系中保持水平。θ和φ分别表示相机在空间中的偏航角和俯仰角。将这些相机参数$\mathrm{x}_i^{\mathrm{c}}$传入Unity3D虚拟环境中，通过脚本控制相机位置及运动，仿真生成相应的运镜轨迹图和虚拟相机拍摄画面。

3.2.5 图像美学质量评价模型

图像美学质量评价模型可以辅助优化摄像机参数计算，是系统中重要的一部分，其整体框架图如图4所示。虚拟场景运镜轨迹可能会有一些小误差，带来不好的观演体验。因此，在计算生成摄像机参数后，在该参数下生成相应的画面，将该画面输入图像美学质量评价模型，对虚拟场景中各机位画面进行整体美学评分预测，选取平均得分最高、标准差较小的机位作为参考机位，以这种方式取代了以个人主观为导向的美学判断。其次，对参考机位画面进行构图评分预测和主要构图模式判断。当收到构图评分较低的反馈时，我们可以适当调整摄像方向使画面构图更加贴近主要构图模式，直到构图评分和画面美学达到预期。

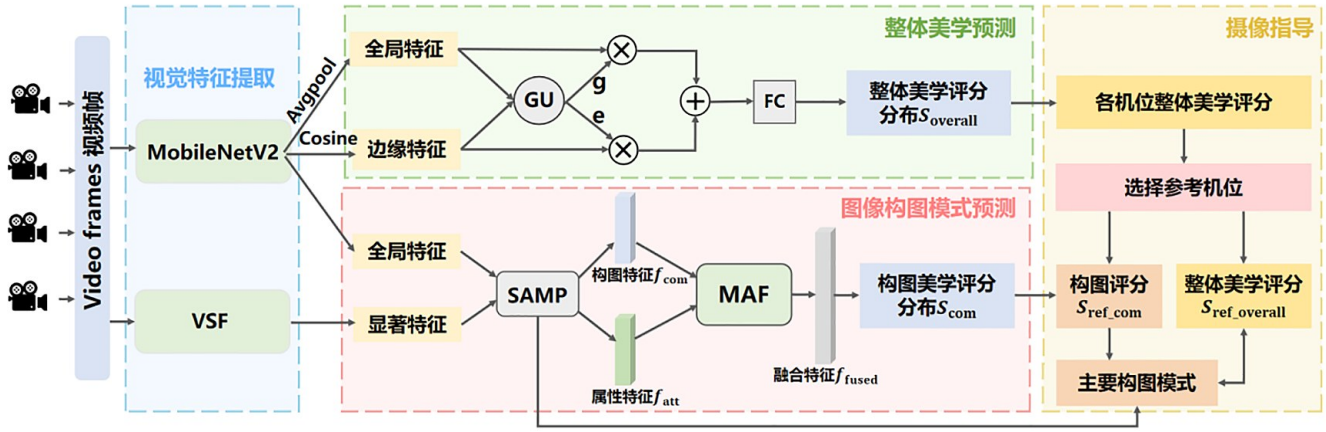

图4 美学评估算法框架图

本模块使用MobileNetV2[18]作为提取视觉特征的骨干网络。在整体美学预测模块，参考了文献[19]的网络结构。首先基于视觉特征提取网络的输出，构建一个全连通图来表示图像的组成。在连通图中，每个位置都被视为一个节点，将每两个节点特征向量之间的余弦距离表示为图像边缘特征。由于在高维特征中可以捕捉到各种视觉特征，如锐度、色调、几何形状等，因此边缘特征被认为能够表征图像的构图特征。所有节点特征向量的平均值代表全局特征，全局特征可以描述各种各样的美学特征。然后使用门单元GU将美学特征f_{aes}和构图特征f_{com}结合，赋予它们不同权重u,v并进行拼接操作，得到融合特征S_{aes}。最后通过一个全连接层获取图像整体美学评分分布，根据评分分布可以计算图像分数均值，用于选取合适的参考机位。

$$\mathrm{u} = \mathrm{GU}\left(f_{\mathrm{aes}}\right) \tag{4}$$

$$\mathrm{v} = \mathrm{GU}\left(f_{\mathrm{com}}\right) \tag{5}$$

$$S_{\mathrm{aes}} = \mathrm{u} \cdot f_{\mathrm{aes}} + \mathrm{v} \cdot f_{\mathrm{com}} \tag{6}$$

在图像构图模式预测模块，输入上一模块选取的机位画面，获取全局特征图，并对该画面进行显著图提取和最大值池化下采样。然后，将全局特征图和显著图送入显著增强多模式池化（Saliency-augmented Multi-pattern Pooling, SAMP）[20]，得到图像多模式（八种基本构图模式）权重和聚合特征f_{samp}。引入三分法、整体构图的平衡程度(Balancing Elements)、是否有主体物体(Object Emphasis)、重复与对称这五个属性作为属性特征来补充构图特征[20]，将聚合特征f_{samp}分解成构图特征f_{com}和属性特征f_{att}。再动态权衡f_{com}和f_{att}对构图评估的贡献，得到融合特征f_{fused}。

$$f_{\mathrm{fused}} = f'_{\mathrm{com}} + f'_{\mathrm{att}} \tag{7}$$

f'_{com}和f'_{att}分别代表动态权衡后的构图特征以及属性特征。最后获取图像构图评分分布计算平均分。

当评分较低时,我们可以对机位方向进行适当地调整,使图像更贴合由最大模式权重对应的基本构图模式。该操作使机位画面构图评分增大,即更符合大众审美。

其中八种构图模式如图5所示[20],每个构图模式包含两个或多个不重叠的分区,并为评估构图中质量提供了单独的视角。其中,模式1、2、6、7、8与考虑对称或径向视觉平衡的对称构图有关;模式3、4涉及对角线构图;模式5与中心构图有关,该构图的主要对象被放置在图像的中心。

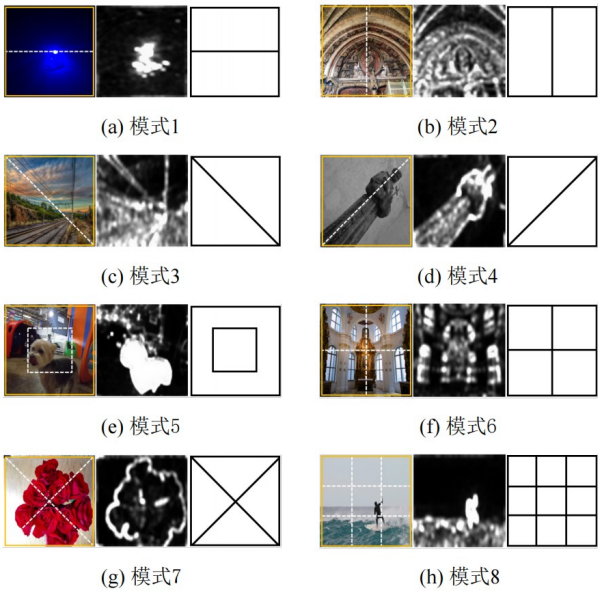

图5 八种基本构图模式示意图

3.3 前端开发设计

前端小程序具体用户交互逻辑结构图如图6所示。

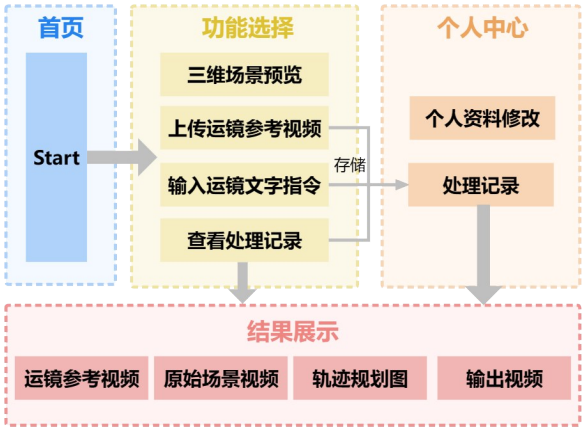

图6 前端用户交互结构图

前端部分主要采用首页、功能选择页面、个人中心界面、结果展示页面。其中包括视频、图像、文字等UI设计和显示。各页面UI原型设计图如图7所示。

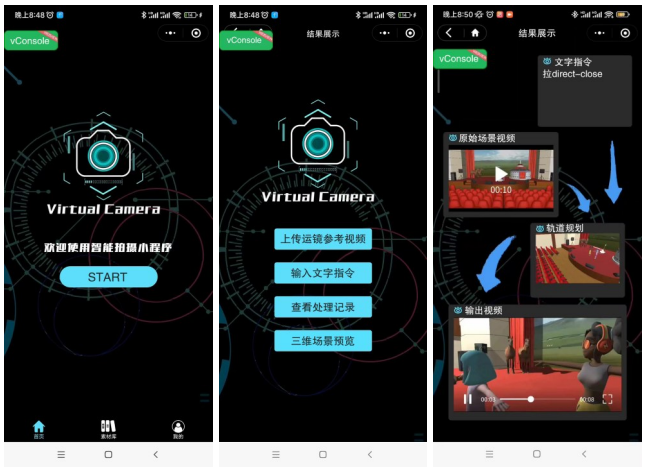

图7 前端小程序原型图

首页欢迎界面清晰地展示小程序的logo,底部栏包含三个页面定位:首页、素材库以及用户信息管理,在欢迎界面点击START按钮,即可跳转到功能选择页面。在功能选择页面中,用户可以选择上传视频、输入文字指令、查看处理记录或者查看三维场景。结果展示界面则清晰展示了4项内容:用户上传的运镜参考视频/用户输入的文字指令、原始虚拟场景视频、运镜轨迹图和输出的仿真视频,便于用户进行对比。

3.4 前后端的信息交互

3.4.1 数据库设计

为了便于用户查看任务进程状态以及之前的处理记录及结果,生成自己的视频素材库,采用了MySQL数据库,设计了两张表分别存储视频数据、任务数据,具体字段设置如表1、表2所示。

表1 视频数据表

字段	含义	数据类型
id	序号	int
vid	视频ID	string
video_link	视频的对象存储url	char
size	视频大小	float
duration	视频持续时间	float
thumb	视频缩略图的对象存储链接	char
upload_time	上传时间	timestamp
openid	微信用户唯一标识	string

表2 任务数据表

字段	含义	数据类型
id	序号	int
task_id	任务ID	string
task_type	任务类型	string
status	任务状态	int
return_msg	返回的结果（当任务完成时才有）JSON格式	string
vid	任务要处理的视频ID	string
start_time	任务的开始时间	timestamp
end_time	任务的结束时间（当任务完成时才有）	timestamp
openid	微信用户标识	string
camera_parameter	用户设置的相机参数	json
command	用户文字需求指令	string

3.4.2 前端–后端接口设计

采用RESTful API来设计前后端的接口，API接口设计如图8所示。

前端：用户在小程序前端上传视频（先小程序端将视频上传对象存储，然后将cloud://开头的链接通过POST /v1/videos接口上传），即可在用户数据库中的获取已有视频（GET /v1/videos），用户选择要处理的视频（POST /v1/tasks）。等待后端算法模型处理完后，或者点击了刷新按钮的流程，即可查看目前的处理情况和结果（结果视频也是cloud://开头的文件 GET /v1/tasks）以及服务器在线情况（GET /v1/servers）。

后端：定时携带“AFS-SERVER-ID”header获取小程序端传来的任务（GET /v1/tasks all=0），获取到后就占据想要处理的任务（POST /v1/tasks/<task_id>）并获取任务所需资源（GET /v1/tasks/<task_id>），接着通过资源的COS链接进行下载，在经过机位计算算法处理完后，就可以向前端小程序返回任务结果。

图8 API接口示意图

4 实验与分析

4.1 系统实验开发环境

系统的开发环境如表3所示。

表3 智能拍摄系统开发环境表

开发项	开发工具
操作系统	Linux
开发平台	Unity3D、Pycharm
开发语言	C#、Python
CPU型号	Intel Xeon Gold 6240
显卡配置	NVIDIA A40

4.2 机位计算模型对比实验

将机位计算模型相机运动提取模块中所采用的门控网络(LSTM加多头自注意力机制(LSTM+sa3))与其他模型进行性能的对比。计算预测的5维参数$x_i^c=\{x_A,x_B,mean(y_A,y_B),\theta,\varphi\}$的整体均方误差(Mean-Square Error,MSE),将其作为评价指标来对比各模型性能。MSE的值越小,代表模型的性能越好,实验的结果见表4。为方便对比,各模型Epoch均取300。

表4 门控网络模型性能比较表

模型	MSE
GRU	0.098
LSTM	0.060
LSTM+se(reduction=2)	0.156
LSTM+sa1	0.043
LSTM+sa3	**0.041**
LSTM+sa3+skip	0.082
LSTM+sa9+skip	0.181

表格中的se表示在LSTM之后加上SE(Squeeze Excitation)模块,其中参数reduction设为2;sa1表示单头自注意力机制,sa3表示3头自注意力机制,sa9表示9头自注意力机制;skip表示skip connection。

由表4所示,将LSTM替换成GRU模型后模型性能反而下降,因此在本系统中采用LSTM更为合适。而加入SE模块或skip也对模型性能提升并无帮助。但加入自注意力机制后,MSE降低,模型性能有所提升,故本系统的相机运动提取门控网络采取LSTM加多头自注意力机制(head=3),其误差最小、性能较优。

4.3 美学评估模型对比实验

4.3.1 美学评价指标

为了评估不同模型的性能,采用了4个广泛使用的指标,即分类准确率(Accuarcy,ACC)、推土距离(Earth Mover's Distance,EMD)损失、皮尔逊线性相关系数(Pearson Linear Correlation Coefficient,PLCC)和斯皮尔曼等级相关系数(Spearman's Rank-order Correlation Coefficient,SRCC)。其中,EMD[21]度量预测得分分布与真实得分分布之间的相似程度;PLCC和SRCC[22]表示预测值和真实值之间的线性相关性。ACC、PLCC和SRCC值越大,性能越好;EMD和MSE值越小,性能越好。

4.3.2 对比实验结果

基于AVA数据集,将本系统所采用的美学评价模型与A-Lamp[23]和NIMA[21]模型进行了ACC、EMD、PLCC、SRCC四个指标的对比。算法性能对比实验的结果如表5所示。

表5 模型在AVA数据集上的性能比较表

方法	ACC	EMD	PLCC	SRCC
A-Lamp(VGG16)	**82.50**			
NIMA(MobileNet)	80.36	0.081	0.518	0.510
NIMA(VGG16)	80.60	0.052	0.610	0.592
NIMA(inception V2)	81.51	**0.050**	0.636	0.612
Ours(MobileNetV2)	82.48	0.071	**0.774**	**0.760**

可以看到,A-Lamp在美学上的性能突出,而代价是大量的计算复杂度,因为它需要额外的对象检测器,而且输入复杂,需要提取原始图像的特征来附加输入。而NIMA的美学性能略逊一筹,这是因为它在美学评价时没有体现图像的构图,而只提取了全局特征。

本系统所采用的美学评价模型在分类准确率上看,和A-Lamp预测的效果比较接近,而实现过程的复杂度却要更小;在损失函数上看,我们的模型虽然不如使用VGG16的NIMA和用inceptionV2的NIMA,但是相比起这二者,我们的视觉特征提取网络需要较少的参数,有着更精简的训练过程。并且在其他三个指标上,我们所使用的美学算法模型表现都优于NIMA。

4.4 模型应用测试

基于Unity中传大礼堂的虚拟环境,选定了多个不同的参考镜头视频,通过机位参数计算模型,可以得到如图9所示的实验结果(包含预测的运镜轨迹和相应生成的画面)。

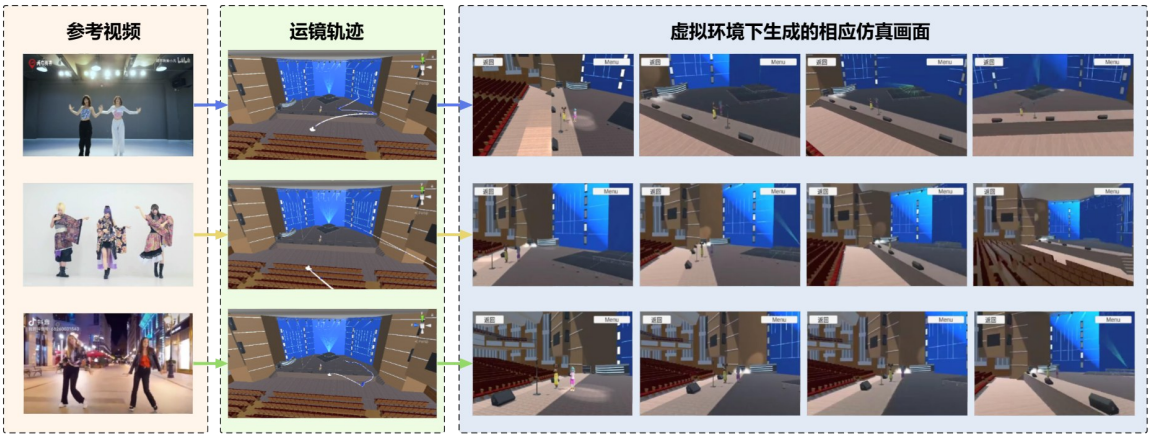

图9 机位计算模型实验效果图

同时对得到的输出视频进行构图评分预测和主要构图模式判断并进行拍摄指导，部分指导结果如图10、图11所示。图11中，左侧(a)画面的最主要构图为模式1(纵向对称构图，该图中地面和墙体刚好在中间部位有明显的分界线，橙色线标注)，权重达到了0.6889，修改空间不大。因而，选择权重占比第二的模式5(中心构图)给出调整意见：在该模式下将画面中人物主体移到中心部位可能会提高构图质量。结果显示，模式5权重占比提升，预测构图评分也增大了0.354，调整所得画面更符合大众审美。所以在摄像机机位参数计算模型(虚拟智能拍摄)的基础上，可以使用美学评估的反馈来进一步优化。

图10 不同拍摄效果和美学评分(均值±标准差)

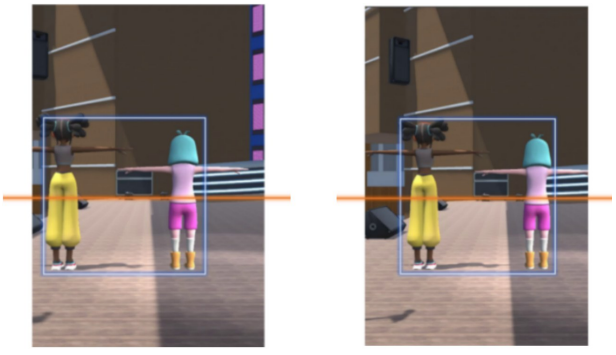

图11 美学评估摄像指导结果图

4.5 系统前后端测试

4.5.1 后端接口测试

使用Postman进行后端接口测试，选择请求方式并且输入服务的外网地址、请求的参数或post的body内容，点击“send”发送请求，最后返回请求结果。后端API接口测试结果如表6所示，所有端口均可正常响应请求。

表6 后端API接口测试表

请求方法	请求链接	请求结果
GET	/v1/tasks/<task_id>	success
POST	/v1/tasks/<task_id>	success
GET	/v1/deltasks	success
GET	/v1/delvideos	success
GET	/v1/servers	success

4.5.2 前端小程序测试

由图12小程序结果展示页面所示的前后端联动测试结果可以看出，用户上传运镜参考视频后，小程序均能获取到当前任务且“处理成功”，并可以跳转到结果显示页面，直观清晰地向用户展示最终效果。

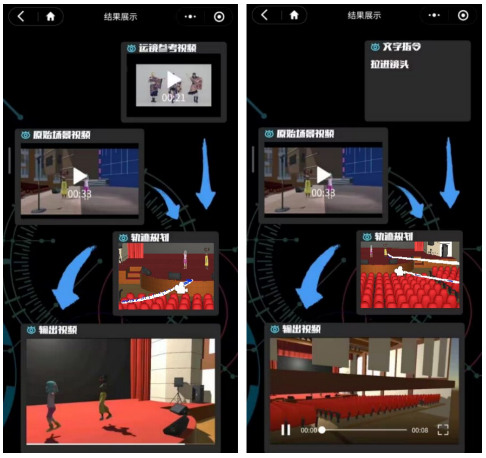

图12 小程序测试效果图

5 结论

本文研发了一种基于机位计算的云演艺智能虚拟拍摄系统,仅通过真实参考视频即可学习到相机运动参数及风格,并可将该运镜方式迁移到新的虚拟场景中,预测得到机位参数和运镜轨迹。本系统在采用基于数据驱动的智能拍摄(机位计算)基础上,加入了美学评估技术,提升了镜头的艺术表现力。本文的研究可用于云演艺场景下的自动运镜,可为其提供实用性高的虚拟拍摄工具。此外,本文还对前端、后端、算法模型分别进行了部署测试,测试结果验证了模型和算法的可行性和先进性,且系统各部分较为独立、有利于后续功能拓展。

参考文献(References):

[1] Wang M, Yang G W , Hu S M , et al. Write-a-video: computational video montage from themed text [J]. ACM Transactions on Graphics, 2019, 38(6): 177.1-177.13.

[2] Xiong Y, Heilbron F C, Lin D. Transcript to video: efficient clip sequencing from texts [C]//Proceedings of the 30th ACM International Conference on Multimedia, 2022: 5407-5416.

[3] Chen J, Carr P. Mimicking human camera operators[C]// 2015 IEEE Winter Conference on Applications of Computer Vision, 2015: 215-222.

[4] Chen J , Le H M , Carr P , et al. Learning online smooth predictors for realtime camera planning using recurrent decision trees[C]// 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 4688-4696.

[5] Chen J, Little J J. Where should cameras look at soccer games: improving smoothness using the overlapped hidden Markov model [J]. Computer Vision and Image Understanding, 2017, 159: 59-73.

[6] Jia H, Thawonmas R, Paliyawan P. An aerial cinematographer AI for settlements in minecraft – toward their crowd assessment[C]//2021 IEEE 10th Global Conference on Consumer Electronics (GCCE), 2021: 853-854.

[7] Huang C, Lin C E, Yang Z, et al. Learning to film from professional human motion videos [C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2019: 4239-4248.

[8] Jiang H, Wang B, Wang X, et al. Example-driven virtual cinematography by learning camera behaviors [J]. ACM Transactions on Graphics (TOG), 2020, 39(4): 45. 1-45. 14.

[9] Jiang H, Christie M, Wang X, et al. Camera keyframing with style and control[J]. ACM Transactions on Graphics, 2021, 40(6): 209.1-209.13.

[10] Lino C, Christie M. Intuitive and efficient camera control with the toric space[J]. ACM Transactions on Graphics, 2015, 34(4):82.1-82.12.

[11] Yu Z, Guo E, Wang H, et al. Bridging script and animation utilizing a new automatic cinematography model[C]//2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR), 2022: 268-273.

[12] Yu Z, Yu C, Wang H, et al. Enabling automatic cinematography with reinforcement learning[C]//2022 IEEE 5th International Conference on Multimedia Information Processing and Retrieval (MIPR), 2022: 103-108.

[13] Gschwindt M, Camci E, Bonatti R, et al. Can a robot become a movie director? learning artistic principles for aerial cinematography[C]//2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), 2019: 1107-1114.

[14] Dang Y, Huang C, Chen P, et al. Path-analysis-based reinforcement learning algorithm for imitation filming[J]. IEEE Transactions on Multimedia, 2023,25: 2812-2824.

[15] Dang Y, Huang C, Chen P, et al. Imitation learning-based algorithm for drone cinematography system [J]. IEEE Transactions on Cognitive and Developmental Systems, 2022, 14(2): 403-413.

[16] Geng Z, Sun K, Xiao B, et al. Bottom-up human pose estimation via disentangled keypoint regression[C]// 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021: 14671-14681.

[17] Seker M, Männistö A, Iosifidis A, et al. Automatic main character recognition for photographic studies[C]//2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP), 2021: 1-6.

[18] Sandler M, Howard A, Zhu M, et al. Mobilenetv2: inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018: 4510-4520.

[19] Zhao L, Shang M, Gao F, et al. Representation learning of image composition for aesthetic prediction [J]. Computer Vision and Image Understanding, 2020, 199:103024.

[20] Zhang B, Niu L, Zhang L. Image composition assessment with saliency-augmented multi-pattern pooling [DB/OL]. arXiv:2104.03133, 2021.

[21] Talebi H, Milanfar P. NIMA: neural image assessment[J]. IEEE Transactions on Image Processing, 2018, 27(8): 3998-4011.

[22] Antkowiak J, Baina T J. Final report from the video quality experts group on the validation of objective models of video quality assessment March 2000 [J]. ITU-T Standards Contribution COM, 2000.

[23] Ma S, Liu J, Chen C W. A-lamp: adaptive layout-aware multi-patch deep convolutional neural network for photo aesthetic assessment [C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(CVPR), 2017: 722-731.

编辑:赵志军,龙学锋