面向演播室多机位跨镜头人物跟踪与导播决策方法

刘春雷

南京传媒学院 江苏 南京 210000

【摘　要】：演播室多机位节目制作长期依赖人工导播完成镜头选取与切换，人力成本高且质量难以量化。为此提出一种融合身份感知与时空约束的跨镜头人物跟踪及导播决策方法，构建由感知、单镜头跟踪、身份确认、跨镜头关联与导播决策组成的五层框架：以限制对比度自适应直方图均衡抑制强补光造成的照度失衡；以人脸质量筛选、局部二值模式粗筛与轻量卷积嵌入精判构成的级联结构建立稳定身份；将地面单应投影一致性、运动时空门控与再识别外观相似度整合为自适应加权代价，经两级匈牙利分配实现弱重叠及无重叠视域下的全局身份传递；并以占屏比、构图偏移、人脸朝向、清晰度与运动平稳度五要素构图评分结合最小驻留时长与迟滞阈值的切换规则输出播出画面。经实践检验，该方法具备实时辅助导播能力。

【关键词】：多摄像头多目标跟踪；跨镜头身份关联；时空约束；自动导播；镜头切换决策

DOI:10.12417/2705-1358.26.16.074

1 引言

一场双人访谈正在录制，六台摄像机同时开机：一台全景、两台对切中景、两台分别锁定主持人与嘉宾的近景、一台游动补位。此刻嘉宾突然转向提问，画面该切给谁、什么时候切——这个在资深导播看来"下意识"的判断，恰恰是自动导播最难复制的环节。

传统上，这个决策完全依赖人；而随着直播与常态化微课录制需求激增，专业导播资源严重供不应求，切换质量还高度依赖个人经验、难以量化考核。让机器接手这项工作的呼声由此而来。

然而，机器要完成"切给谁、何时切"两个动作，必须先回答两个更底层的问题：其一，此刻哪个画面里是谁——这是跨镜头的身份关联问题；其二，此刻哪个画面"最好看"——这是镜头选择与切换时机的决策问题。这是现有研究的两块空白。

一方面，多摄像头多目标跟踪的研究与导播追求的画面美学、叙事节奏背道而驰；且该领域的跨镜头数据关联至今仍未被闭合——无重叠视域、光照突变、长时间遮挡下的身份碎片化与轨迹断裂依旧严重。另一方面，面向演播室的自动导播研究凤毛麟角[1]，已有工作多采用"人脸匹配＋关键词命中"的规则式方案，既缺乏跨镜头的身份连续性建模，也少见把跟踪结果与导播质量建立量化关联的工作。

针对固定机位布局的演播室场景，本文提出一种融合身份感知与时空约束的跨镜头人物跟踪及导播决策方法：（1）提出身份感知的跨镜头关联方法，将地面单应性约束、时空一致性门控与外观相似度整合为自适应加权代价，以两级匈牙利分配实现弱重叠及无重叠视域下的全局身份传递；（2）构建面向导播任务的镜头质量评价体系，定义占屏比合格率、构图偏移、切换时机一致率与冗余切换率四项可量化指标，并建立带迟滞约束的切换决策规则，使"画面好不好看"具备可量化的客观判据；（3）给出演播室多机位同步数据集的构建方案与实验协议，为该方向提供可复现的评价基础。

2 方法

2.1 场景假设与五层框架

设演播室内布置 C 台已标定的固定摄像机，全部视频流经硬件时钟同步并统一重采样至帧率 f_0。定义三个基本概念：局部轨迹，即单镜头内由检测器与关联器维持的目标轨迹，记为 L(c,i)；全局身份，即跨机位统一的主体标识，记为 G_k；候选画面，即任一时刻各机位输出的一帧。本文的目标为：对任意时刻，输出各候选画面中主体的全局身份，并给出应当播出的画面。

作如下假设：（1）参与跨镜几何关联的机位位置固定且光轴在录制周期内不漂移；（2）地面可近似为平面，人员主要在该平面活动；（3）允许相邻机位视场重叠，但不要求全部机位两两重叠；（4）同一主体同一时刻只能处于一个物理位置。其中第三、四条假设分别保证了跨镜几何映射的可行性与身份解的唯一性。

方法的总体框架如图 1 所示，由五层构成[2]，各层的功能定位与数据接口见表 1。感知层完成光照补偿、人物检测与包围盒裁切；单镜头跟踪层在各机位内部独立维持帧间轨迹；身份确认层以级联结构为局部轨迹给出确定身份；跨镜头关联层计算三类证据并输出局部轨迹到全局身份的最优分配；导播决策层对每个候选画面计算构图质量分，经角色加权与迟滞判断输出播出画面序列。

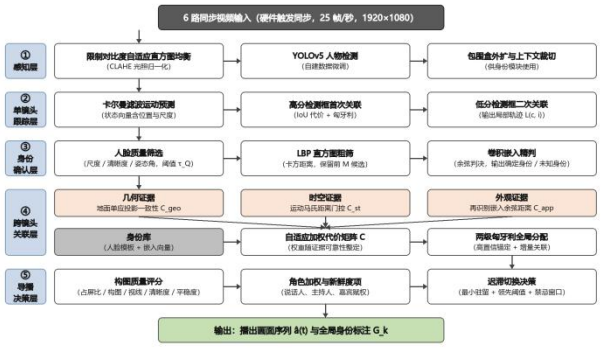


图 1 方法总体框架

表 1 五层框架的功能定位与数据接口

层次	主要功能	输入	输出	关键方法
①感知层	光照归一化、人物检测与裁切	同步视频帧	检测框与置信度	CLAHE、YOLOv5
②单镜头跟踪层	帧间轨迹关联	检测框序列	局部轨迹 L(c,i)	卡尔曼滤波、二次关联
③身份确认层	人脸质量筛选与级联身份判定	局部轨迹、身份库	确定全局身份/未知身份	人脸质量分、LBP 粗筛、卷积嵌入精判
④跨镜头关联层	三类证据计算与全局身份分配	局部轨迹、身份库	全局身份 G_k	单应投影、马氏门控、匈牙利分配
⑤导播决策层	构图评分与镜头选取	候选画面、全局身份	播出画面序列	加权评分、迟滞决策

2.2 光照自适应的人物检测

演播室采用三点布光与正面补光，人物面部常出现高光溢出与局部欠曝并存。直接调用在通用数据上训练的检测器，在低照度区域容易漏检。为此在检测前引入限制对比度自适应直方图均衡（CLAHE）：将图像划分为若干子块并逐块做直方图均衡，同时对频数施加裁剪限幅、超出部分均匀回填，从而避免噪声被局部过度放大；子块边界以双线性插值消除，处理后灰度动态范围被压缩到检测器训练分布附近。

检测器采用 YOLOv5 轻量分支，在自建数据上对人物类别微调。为提升后续身份模块的输入质量，检测包围盒在裁切前适度外扩（宽高分别乘以 1.15 至 1.30 的系数），使裁剪区域保留部分背景上下文，并保证窗口不超出图像边界。

2.3 级联式身份确认

演播室中同类服装与相近体型并不少见，单纯依赖行人再识别嵌入易混淆；而人脸作为强判别线索在此场景可获得性较高，主持人面部基本正对镜头且补光充足。因此设计两级级联的身份确认流程。

前端筛选对每个局部轨迹按固定间隔采样若干帧，定位人脸区域后计算质量分。质量分综合人脸尺度、清晰度与姿态角三项因素，仅当质量分超过阈值 τ_Q 的帧进入下一级，其余帧丢弃，从而把大量侧脸、运动模糊与远距离小脸过滤掉，显著降低后端计算量并提高特征纯净度。

后端确认先做与前述一致的光照归一化，再提取两类特征：一是对光照稳健的局部二值模式（LBP）直方图，用于粗筛；二是轻量卷积网络输出的嵌入向量，用于精判。第一阶段以卡方距离衡量待确认人脸直方图与身份库模板的相似度，保留距离最小的前 M 个候选；第二阶段对 M 个候选逐一计算嵌入向量与待确认人脸嵌入 f 的余弦相似度，取其中最高者 cos_max 与次高者 cos_2nd，判决规则为：

若 cos_max ≥ τ_s 且(cos_max−cos_2nd) ≥ τ_m，则确认为该身份；否则记为未知身份。（1）

其中 τ_s 为相似度阈值 τ_m 为间隔阈值，用于在近邻身份之间保持判别余量。

2.4 时空与外观联合的跨镜头关联

本层是核心，输入为各机位在时刻 t 的局部轨迹集合，输出为局部轨迹到全局身份的分配结果[3]，处理流程如图 2 所示。

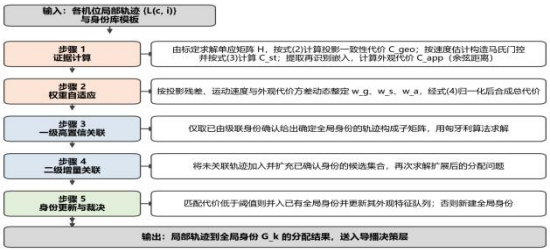


图 2 跨镜头身份关联流程

（1）几何证据，录制前在地面铺设已知尺寸的标定图案，使存在共视地面区域的机位均能观测，据此标定任意两机位 V_a、V_b 之间的单应矩阵 H(a→b)。对站立人物取其足底点（包围盒底边中点）作为地面接触点的近似，记为齐次坐标 u_a；将机位 a 中第 i 条轨迹的足底点经单应变换投影到机位 b，得到预测点 $\hat{p}_i = \pi(H(a\to b)\cdot u_a)$；再以 $\hat{p}_i$ 为中心、按人物在地面的典型占位宽度生成预测框 $\hat{B}_i$，与机位 b 中候选目标 j 的足底点邻域框 B_j 计算交并比（IoU），定义投影一致性代价：

C_geo(i,j)=1−IoU($\hat{B}$_i,B_j)（2）

其中 u_a 为足底点的齐次坐标 π 为齐次到非齐次的映射，B_j 为候选目标 j 的足底点邻域框，按人物在地面的典型占位宽度生成；$\hat{B}$_i 采用与 B_j 相同的宽度参数生成，以保证二者尺度可比。

（2）时空证据，静态投影无法处理运动维度的歧义：多人在邻近站位时投影框可能相互重叠。为此对每个全局身份维护最近若干时刻的地面速度估计，由滑动窗口内的位置差分得到并以指数衰减加权突出近期运动，据此预测下一时刻的中心 μ 与不确定度协方差 Σ，构造马氏距离搜索门：

Ω={j|(p_j−μ)^{T}Σ^{-1}(p_j−μ)≤γ}，C_st(i,j)=d_M^{2}(p_j,μ)/γ（3）

其中 γ 为卡方分布在给定置信水平下的分位数。位于门内的候选保留，门外候选本轮不予考虑。人物快速走动时速度估计方差增大、门自动放宽；静止时门收缩，实现搜索范围的自适应控制。此时门内候选满足 C_st≤1，代价与门限处于同一量纲。

（3）外观证据，对每条局部轨迹按间隔采样有效裁剪图，提取再识别嵌入并以指数滑动平均维护代表嵌入ē，两条轨迹间的外观代价由余弦距离 C_app(i,j)=1−cos(ē_i,ē_j)给出；对长时间未刷新的历史嵌入按时间衰减降低其贡献，避免换装与道具变更造成的特征漂移。

（4）自适应加权，总代价为三类证据代价的线性组合，其中权重依据当前帧的可用证据动态整定。对第 k 类证据（k∈{g,s,a}分别表示几何、时空与外观）定义可靠性得分 r_k：几何项取投影残差的负归一化值，时空项取目标运动平稳度的归一化值，外观项取候选外观代价分布的方差归一化值（方差越小说明外观越缺乏判别力）。权重按下式归一化[3]：

C(i,j)=w_g·C_geo(i,j)+w_s·C_st(i,j)+w_a·C_app(i,j)，

w_k=e^(λ·r_k)/Σ_{l∈{g,s,a}}e^(λ·r_l)（4）

其中 λ 为温度系数，控制权重对可靠性差异的敏感程度。

（5）两级匈牙利分配，求得代价矩阵后进行两级分配：一级仅对已由级联身份确认给出确定全局身份的局部轨迹子集 S_1 求解最小代价匹配，二级在此基础上将未关联轨迹集合 S_2 加入并扩充候选、再次求解扩展后的分配问题。一级以身份明确者为锚，避免错误在全局扩散；二级负责吸收不确定性较高的剩余关联。对于两级的匹配代价仍高于阈值 τ_c 的新轨迹，则新建全局身份 G_{K+1}；否则并入代价最小者对应的身份并更新其外观特征队列。已被某机位确认占用的全局身份，同一时刻在其他机位的重复占用由互斥规则压制。两级分配的整体时间复杂度为 O(n_1^{3}+(n_1+n_2)^{3})，其中 n_1、n_2 分别为两级参与分配的轨迹规模；由于 n_1 仅覆盖已确认身份的轨迹，实际计算量显著低于一次性全局匹配。

2.5 面向导播的构图质量评价与切换决策

对任意机位 c 在时刻 t 的画面定义质量分 S(c,t)，由五项要素加权构成，依次记为 r_1 至 r_5。

（1）主体占屏比 r_1，取目标包围盒面积与画面面积之比，以理想区间[a,b]为参照：落在区间内得满分，低于下界或高于上界则按幂指数 p 衰减。访谈类节目取偏大的理想区间，讲授类取偏小值以保留板书与演示画面。

（2）构图中心偏移 r_2，同时考虑居中与三分两种可接受构图，取主体中心到画面中心、最近三分线交点两者归一化距离的较小者作为偏差。

（3）人脸朝向与视线 r_3，由头部姿态估计给出的偏转与俯仰角与正对机位方向的一致程度确定，是区分同一时刻哪个机位拍到正脸的关键判据，也是人工导播最常用的依据。

（4）清晰度 r_4，取包围盒区域拉普拉斯响应方差的归一化值。

（5）运动平稳度 r_5，取包围盒中心位移一阶差分波动的负归一化值，用于避免切到尚未稳定的画面。

综合评分与角色加权按下式进行[4]：

S(c,t)=Σ_{m=1}^{5}β_m·r_m(c,t)，

S̃(c,t)=η(c,t)·S(c,t)+ν·novelty(c,t)（5）

其中 β_m 为第 m 项要素的权重、且 Σ_{m=1}^{5}β_m=1；η 为角色加分系数，依据全局身份把主体区分为当前说话人、主持人、主要嘉宾、次要人物与无人五种情形分别取值；novelty 为新鲜度项，长时间未被使用的机位获轻微加分，长时间定格于同一机位施加轻微惩罚，ν 为该项权重。η 与 ν 按节目形态整定，其中 η 依据角色分档取值、novelty 按机位最近未使用时长归一化至[0,1]，具体取值见 3.2 节。

直接逐帧取最大值会导致相邻帧评分接近时画面来回跳动，观感比不切换更差。为此引入三条约束：(1)最小驻留时长 t_min，一次切换后至少经过该时长才允许再次切换；(2)领先幅度阈值 δ，仅当最高分领先当前播出画面评分达该阈值才允许切换；(3)禁忌窗口 Ξ，在检测到明显画面扰动或现场调度行为的时间窗内禁止切换。切换规则为[4]：

â(t)=arg max_c S̃(c,t)，

当且仅当 S̃_max−S̃_cur>δ 且 t−t_last>t_min 且 t∉Ξ；否则â(t)=â(t−1)（6）

该规则使输出在时间维上平滑，同时保留对显著事件的即时响应能力，例如新嘉宾入场与发言权转移。

3 实验与结果分析

3.1 数据集与评价指标

通常的多摄像头数据集以交通路口、校园广场等室外监控为主，缺乏演播室特有的强补光、固定布景与正面主持等特征。为此构建 Studio-MC 数据集：使用 6 台同型号网络摄像机，分辨率 1920×1080、帧率 25 帧/秒，通过统一硬件触发信号同步；机位配置为全景 1 台、左右中景各 1 台、主持人近景与嘉宾近景各 1 台、游动 1 台。节目形态覆盖 A 类双人面对面访谈、B 类单人讲授型录制（含走动与板书）、C 类多人访谈走位（含人员交叉与部分遮挡）三类，共录制 24 段、总时长 186 分钟，标注轨迹 412 条、全局身份 68 个。切换时刻真值由两名具备三年以上经验的导播独立标注，共 2140 个切换点，二者标注一致的占 91.2%；不一致的切换点经第三人仲裁后确定最终真值，标注过程采用插值后人工校正完成。

跟踪侧采用 HOTA（Higher Order Tracking Accuracy）、IDF1（ID F1 Score）、MOTA（Multiple Object Tracking Accuracy）与身份切换次数（IDS），以 IDF1 与身份切换次数为主要判据。跨镜头指标的计算口径统一为：先将各机位的局部轨迹按关联结果合并为全局轨迹，再在全局轨迹上按标准多目标跟踪评测协议计算上述指标，各基线方法采用同一口径处理，以保证可比性。

导播侧定义四项指标：(1)占屏比合格率，即播出帧中主体占屏比落在理想区间内的帧数占比；(2)构图偏移，即播出帧主体中心到理想构图位置的平均归一化距离；(3)切换时机一致率，即在人工给定切换点容差窗内机器发生切换的比例；(4)冗余切换率，即相邻两次机器切换间隔小于最小驻留时长的比例。需说明的是，切换时机一致率属召回型指标，须与冗余切换率联合解读。

3.2 实现细节

实验平台为单张 RTX 3060 显卡，CUDA 11.8。检测器采用 YOLOv5s，输入分辨率 640×640。为避免评测数据泄漏，检测器的微调与评测按段划分：24 段素材按四折交叉验证，每折 18 段用于微调、6 段用于评测切分，评测段不参与任何训练与微调，表 2、表 4 报告各折平均结果。检测框外扩系数取 1.20；身份确认中 τ_s 取 0.82、τ_m 取 0.06；式（3）中 γ 取置信水平 0.95 对应的分位数；式（4）中 λ 取 2.5、新建身份阈值 τ_c 取 0.45。再识别网络采用公开行人数据集上预训练的轻量骨干，未在自建数据上微调以检验泛化性。单应矩阵在每段录制前重新标定一次。占屏比理想区间[a,b]在 A、B、C 三类场景分别取[0.14,0.32]、[0.06,0.18]、[0.10,0.26]，衰减指数 p 取 1.5；式（5）中 β_1 至 β_5 分别取 0.30、0.24、0.26、0.10、0.10；t_{min} 在三类场景分别取 4.0、3.0、2.0 秒，δ 取 0.08。式（5）中角色加分系数 η 对当前说话人、主持人、主要嘉宾、次要人物与无人五种情形分别取 1.20、1.10、1.00、0.90、0.70，新鲜度项权重 ν 取 0.05，novelty 以机位最近未使用时长与参考时长之比归一化至[0,1]；上述取值经经验整定确定。

3.3 跟踪性能对比

选取四组基线：B1 为逐机位独立 DeepSORT 加帧级外观跨镜匹配，B2 为逐机位 ByteTrack 加离线匈牙利跨镜关联，B3 为仅使用外观嵌入的跨镜关联，B4 为仅使用几何投影的关联。结果如表 2。

表 2 跟踪性能对比（Studio-MC 全集）

方法	HOTA ↑	IDF1 ↑	MOTA ↑	身份切换次数 ↓
B1 DeepSORT+外观跨镜	62.4	68.7	74.2	143
B2 ByteTrack+匈牙利	66.8	73.5	78.6	108
B3 仅外观	60.1	65.2	76.9	176
B4 仅几何	68.2	75.1	79.4	96
本文方法	74.6	81.3	83.5	57

注：↑表示数值越大越好，↓表示数值越小越好；B1—B4 为按主流范式搭建的对比方法，其中 B3、B4 亦可视作本文方法的退化变体。

由表 2 可见，本文方法 IDF1 较最优基线 B4 提升 6.2 个百分点，身份切换次数由 96 降至 57，降幅 40.6%。B3 表现最弱说明单靠外观不足以支撑跨镜关联，演播室中同色系着装与相近体型普遍存在；B4 弱于本文则说明几何投影在人员交叉、短时遮挡时仍会产生歧义，需要时间维度的连续性约束加以消解。分场景观察，A 类场景主体少、站位稳定，几何投影最准，IDF1 达 86.4%；C 类场景人员交叉频繁，IDF1 降至 76.8%，是各方法差距最大的部分。

3.4 导播质量评价

导播侧结果如表 3，其中随机切换作为下限参照，人工导播作为上限参照。

表 3 导播质量对比

方法	占屏比合格率↑	构图偏移↓	切换时机一致率↑	冗余切换率↓	主观评分↑
随机切换	41.2	0.42	26.5	68.3	1.8
B2+逐帧取最大	68.5	0.24	51.7	42.6	2.9
本文（无迟滞）	82.4	0.16	74.3	33.8	3.7
本文方法	88.6	0.13	78.9	6.4	4.2
人工导播（参照）	92.3	0.11	—	2.1	4.6

注：↑表示数值越大越好，↓表示数值越小越好；表中「—」表示该项不适用；随机切换与人工导播分别作为下限与上限参照，不参与方法间比较。

迟滞决策的作用是决定性的：去掉迟滞后冗余切换率由 6.4%升至 33.8%，主观评分由 4.2 降至 3.7，而跟踪指标几乎不变，说明式(6)专门作用于观感层面。本文方法与人工导播在占屏比合格率上相差 3.7 个百分点，主观评分相差 0.4 分，差距集中在需要现场经验的即兴处理上，例如嘉宾突然插入发言时的优先给镜。以同一口径衡量，两名人工标注者之间的切换时机一致率为 91.2%，可作为该指标的人类参照，本文方法的 78.9%与之仍有明显差距。占屏比合格率与主观评分的 Pearson 相关系数为 0.81、切换时机一致率与主观评分的相关系数为 0.76（以 24 段节目为样本、按段计算，n=24），说明客观指标较好地反映了主观感受；受样本量所限，相关系数的置信区间较宽，结论仍需更大规模主观实验支持。

4 结束语

本文针对演播室多机位节目录制中镜头选择依赖人工、缺少量化标准的问题，提出了一种融合身份感知与时空约束的跨镜头人物跟踪及导播决策方法。方法以三类证据的自适应加权与两级匈牙利分配改善了弱重叠及无重叠视域下的身份碎片化，并以构图质量评分与迟滞切换规则使导出画面兼具构图合理性与时间平滑性。实验表明，该方法在身份一致性与导播质量两方面均优于四类基线，且具备实时处理能力。

需指出的是，本文存在四方面局限。其一，数据集规模与多样性有限，24 段素材难以覆盖全部节目形态，结论的泛化性有待在更大规模、多场地的数据上检验。其二，导播侧指标尚属探索阶段，主观评分仅由三名评分者给出，未报告评分者间一致性。其三，从方法结构推断，残余的身份错误最可能集中于多人长时间相互遮挡导致足底点不可见、以及人员快速换位造成投影框重叠两类情形，其实际占比有待进一步统计。其四，音频线索尚未纳入决策，而说话人身份与发言节奏恰是人工导播最常用的判据之一，是下一步的扩展方向。

参考文献：

[1] Zhang Y F,Sun P Z,Jiang Y,et al.ByteTrack:multi-object tracking by associating every detection box[C]//Proceedings of the European Conference on Computer Vision.Cham:Springer,2022:1-21.

[2] Luiten J,Ošep A,Dendorfer P,et al.HOTA:a higher order metric for evaluating multi-object tracking[J].International Journal of Computer Vision,2021,129(2):548-578.

[3] Tang Z,Naphade M,Liu M Y,et al.CityFlow:a city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.Piscataway:IEEE,2019:8797-8806.

[4] Tanabe Y,Nakanishi T.Learning-based scene selection for autonomous multi-camera video editing[C]//Proceedings of the 19th IIAI International Congress on Advanced Applied Informatics.Piscataway:IEEE,2025:155-160.