DOI:10.19651/j.cnki.emt.2417672

基于细粒度动作语境聚合的动作检测与识别*

王 峥 赵新辉 王小伟
(郑州大学体育学院 郑州 450044)

摘 要:在空域和时域上精确定位并识别视频中的人体动作对于智能体育分析等应用具有重要意义。然而,现有的分步人体动作识别方法通常受限于RoI特征的固定感受野,难以在复杂场景中进行有效建模和语义表达。为此,本文提出了一种细粒度动作语境聚合网络,利用并行的语义建模单元和动作候选单元对人物表征特征和全局时空语境特征进行有机融合。前者中采用人体定位模型从关键帧生成细粒度的人物候选特征,并通过3D视频骨干网络提取全局时空特征;后者则利用共享Transformer框架对上述多模态特征进行统一建模,捕捉人物与环境之间的复杂关联,从而获得具有高度区分能力的动作预测。进一步地,本文引入加权分数聚合策略,将多个关键帧与短时视频片段的动作分类信息整合,用于长视频片段的动作识别。在AVA-60 v2.2数据集上,本文模型在帧级mAP指标上达到了30.01%,而基于长时策略的本文模型则达到了30.74%。在Charades数据集上,本文模型的mAP提升至30.68%,而基于长时策略的本文模型结果提升至31.29%。

关键词:人体动作检测与识别;分步方法;全局时空语境特征;细粒度筛选
中图分类号:TN101 **文献标识码**:A **国家标准学科分类代码**:120.99

Motion detection and recognition model based on fine-grained motion & situation fusion

Wang Zheng Zhao Xinhui Wang Xiaowei
(College of Physical Education, Zhengzhou University, Zhengzhou 450044, China)

Abstract: Accurately localizing and recognizing human moions in both spatial and temporal dimensions is of significant importance for applications such as intelligent sports analysis. However, existing step-by-step human motion recognition methods are often limited by the fixed receptive field of RoI features, making it difficult to achieve effective modeling and semantic representation in complex scenarios. To address this issue, this paper proposes a fine-grained motion & situation fusion (FMSF) network that integrates human representation features and global spatiotemporal situation features through parallel semantic modeling and motion proposal units. The semantic modeling unit employs a human localization model to generate fine-grained human candidate features from key frames and leverages a 3D video backbone network to extract global spatiotemporal features. The motion proposal unit then uses a shared Transformer framework to jointly model these multi-modal features, capturing complex interactions between humans and their surroundings, resulting in highly discriminative motion predictions. Furthermore, a weighted score aggregation strategy is introduced to integrate the motion classification results of multiple key frames and short video segments for long-video motion recognition. On the AVA-60 v2.2 dataset, the FMSF model achieved a frame-level mAP of 30.01%, while the long-video strategy-based FMSF-Prolonged reached 30.74%. On the Charades dataset, the mAP of FMSF increased to 30.68%, and that of FMSF-Prolonged increased to 31.29%.

Keywords: human action detection and recognition; step-by-step method; global spatiotemporal situation features; fine-grained screening

0 引 言

人体动作识别(human motion recognition,HMR)指的是在空域和时域中同时定位动作实例,并对所有动作实例进行分类的任务[1]。近年来,HMR受到了大量关注,广泛应用于人机交互、自动驾驶以及体育分析等领域[2]。为了实现精细化的人体动作识别,需要从视频帧中提取深层语义信息,准确表征动作实例,从而实现视频中特定动作的精

收稿日期:2024-12-19
*基金项目:国家自然科学基金青年项目(62306284)、2024年河南省科技攻关项目(242102320282)资助

准定位和分类[3]。

目前 HMR 的方法大都利用经过预训练的骨干网络来生成时空特征，可分为两类：端到端方法和分步方法。端到端 HMR 方法基于视频骨干网络提取的时空特征，同时进行人体检测和动作分类；而分步方法则首先采用经过预训练的检测模型来定位人体区域，随后利用提取到的感兴趣区域(region of interest，RoI)特征进一步进行动作分类。

近年来，端到端方法受到了越来越多的关注。Karácsony 等[4]将人体定位头(用于检测视频中的人物位置)和动作分类头(用于分类人物的动作类型)进行结合起来，并且利用了三维卷积神经网络(three-dimensional convolutional neural network，3 D-CNN)作为骨干网络，从而获取视频的时空特征。Yu 等[5]使用了一种以人物为主体的关系网络，通过将视频中人物部分的特征与全局特征图进行成对关系编码，以此建模人物动作与周围环境的关系。Liu 等[6]提出了一种动作 tubelet 检测框架，通过跟踪视频中人物的移动轨迹来预测人物的动作实例。此方式相当于将动作的轨迹信息串联起来以进行检测。近些年来，Diba 等[7]和 Ravishankar 等[8]皆使用注意力机制将不同尺度的时空特征进行融合，作为视频的语境信息。Zheng 等[9]提出了一种 token 剪枝模块，在处理视频时去掉与人物动作和环境无关的标注信息，从而减少不必要的信息干扰，提高模型的处理效率。Lee 等[10]提出可以利用完整的 tubelet 标注或单帧的稀疏边界框，灵活地对视频中的人物和动作进行分析。相比之下，分步方法的研究相对较少。Li 等[11]通过使用目标跟踪模型生成的 tube 来构建以人物为主体的图结构，目的是捕捉视频中人物之间或人与物之间的互动关系。Akhter 等[12]对从视频中提取的目标特征之间的各种交互关系进行了建模，包括人物之间、人与物之间的交互，以及这些交互在时间上的变化。类似地，Deng 等[13]也使用了注意力机制，但其侧重于捕捉人物之间复杂的关联信息。一般而言，分步方法相较端到端方法的性能提升较为有限。究其原因，现有分步方法通常依赖于骨干网络特征图中 RoI 对齐后的特定感受野特征来进行动作检测与识别，但这常常会限制复杂环境中模型的建模能力。然而，分步方法若经过精心设计，在性能上也可能有大的突破。首先，分步方法可以充分利用目标检测模型的优势。这些目标检测模型往往通过大量数据集进行学习，性能强大。为此，最先进的目标检测模型可以无缝集成到 HMR 框架中，并提升整体性能。此外，分步方法将人体检测和动作分类分开进行，从而避免了同时优化这两个任务所带来的复杂性。基于此，本文旨在设计一种分步 HMR 模型，有效整合视频中的语境信息。

具体而言，本文提出了一种细粒度动作语境聚合网络(fine-grained motion&situation fusion，FMSF)。FMSF 首先使用人体定位模型来检测并生成大量人体候选区域，并且通过细粒度筛选方式确保每个可能的人体区域都被检测到。同时，FMSF 采用 3 D 视频骨干网络生成全局时空语境特征。通过这种并行方式，两个任务的特征可得到有效解耦，从而能够更好地表征视频动作检测和识别相关的多样化动作语境。此外，FMSF 通过共享 Transformer 框架建模人体表征与语境表征之间所有可能的关联。最终，FMSF 生成一组最终动作预测集，并使用双边匹配损失进行训练，从而优化模型性能。同时，它在最后阶段通过检测头过滤掉不可靠的预测，并利用细粒度筛选后的人体候选区域进行更精细的动作检测和人体-语境关系建模。

在传统范式中，HMR 任务要么直接从视频片段中端到端执行，要么仅使用人体定位模型的输出 RoI，然后从视频骨干网络中提取特征。前者在建模人体与周边环境之间的交互方面能力有限；而后者仅将人体定位模型作为一个简单的工具来识别视频中人物所在位置，对于提取动作或语境相关的高级特征毫无帮助。因此，以往的分步方法都是以人物为主体，通过 RoI 对齐操作从视频骨干网络中提取出视频中的人物特征，然后利用这些特征来推断动作语义。相比之下，FMSF 采用了一种非人物主体的方法，结合了细粒度筛选的人物特征和全局语境特征，通过有效捕捉它们之间的复杂关系，一起用来推导动作的语义，从而实现了精细化的动作检测和识别。FMSF 也是首个提出通过全局聚合动作语境来生成动作语义的分步 HMR 架构。此外，还将 FMSF 扩展到了处理长视频片段的任务中。具体而言，使用加权分数聚合的方法，将来自单个关键帧和多个短视频片段的动作分类得分结合起来，以获取更长时间跨度的语境信息。这种方法与传统的长期视频动作检测方法不同，后者通常依赖于特征记忆库或长期记忆编码来存储和处理视频中的长期信息。与这些传统方法相比，FMSF 模型不仅显著提高了准确性，还简化了训练过程。

1　本文方法

FMSF 的整体结构如图 1 所示，主要由两个组件构成：语义建模单元(semantic modeling unit，SMU)和动作候选单元(action proposal unit，APU)。SMU 包含两个并行分支以生成用于视频动作检测任务的嵌入特征，即人物表征分支(human representation branch，HRB)和时空表征分支(space-time representation branch，STRB)。HRB 从关键帧中生成人物候选特征，而 STRB 从连续的视频片段生成时空语境特征。此后，APU 通过建模上述两种模态特征之间的关联来推导与动作相关的语义信息，并为视频中的每个可能的人物候选区域生成与动作相关的特征，最后通过一个前馈神经网络(feedforward neural network，FNN)头来预测该人物的具体动作类别。

1.1　SMU

1) HRB

给定时刻 t 时的关键帧输入图像 I_t，使用人体检测模型并通过 RoI 对齐操作[14]从特征图提取 K 个 RoI 特征，记作

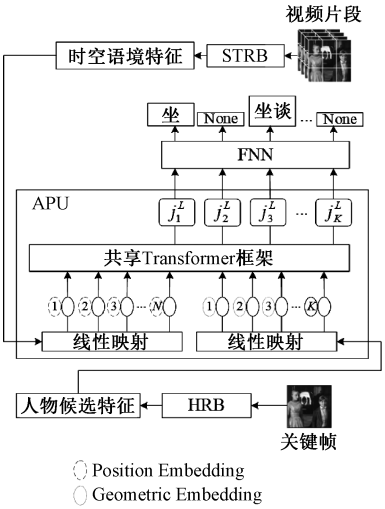

图 1　FMSF 整体框架

Fig. 1　Overall framework of FMSF

$f_a \in R^{C\times K}$，其中 C 表示人物特征的通道维度。通过对人体检测模型应用低置信度阈值，HRB 可生成细粒度的人物候选特征，以获取关于人物的多样化语境信息。接下来，将这 K 个任务特征输入至两个全连接层中，以预测人物的置信度得分 $\hat{h}=\{\hat{h}_i\}_{i=1}^{K}$ 和边界框 $\hat{b}=\{\hat{b}_i\}_{i=1}^{K}=\{(x_i^{lt},y_i^{lt},x_i^{rb},y_i^{rb})\}_{i=1}^{K}$，其中 (x_i^{lt},y_i^{lt}) 和 (x_i^{rb},y_i^{rb}) 分别表示边界框左上角和右下角的归一化坐标。然后，这些人物候选特征与时空语境特征进行交互，以便于后续 APU 处理，如图 2 所示。

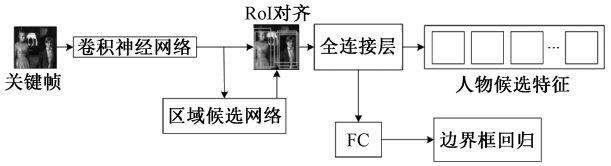

图 2　HRB 处理流程

Fig. 2　HRB processing flow

2) STRB

给定 $T=T_p+T_f+1$ 帧的输入视频片段 $I_{t-T_p:t+T_f}=\{I_{t-T_p},\dots,I_{t-T_f}\}$，STRB 使用视频骨干网络提取大小为 $H\times W\times T$ 的时空特征图。然后，时空特征图展开为语境特征 $f_v\in R^{C'\times N}$，其中 C' 是特征通道大小，$N=H\times W\times T$ 为经过展开后的特征数。

1.2　APU

APU 通过共享 Transformer 框架[15]来建模人物特征与时空语境特征之间的关系。APU 首先将人物特征线性映射到维度为 D 的嵌入向量：

$$\boldsymbol{a}=E_a f_a+E_g g \tag{1}$$

其中，$E_a\in R^{D\times C}$ 和 $E_g\in R^{D\times 6}$ 线性投影的可训练权重，$g\in R^{6\times K}$ 为由六维向量 $[x^{lt},y^{lt},x^{rb},y^{rb},w,h]^{\mathrm{T}}$ 表示的 K 个人物候选区域的几何信息，w 和 h 分别表示边界框的宽度和高度。与此同时，语境特征的嵌入向量由以下公式得到：

$$\boldsymbol{v}=E_v f_v+E_{pos} \tag{2}$$

其中，$E_v\in R^{D\times C'}$ 是特征映射的权重，$E_{pos}\in R^{D\times N}$ 表示正弦位置编码。设置嵌入向量的大小 D 为 256。构建 Transformer 编码器输入 $\boldsymbol{j}^0=[\boldsymbol{a},\boldsymbol{v}]=[a_1,\dots,a_K,v_1,\dots,v_N]\in \boldsymbol{R}^{D\times(K+N)}$。然后，这些输入嵌入经过 L 层 Transformer 编码：

$$z^l=MSA(LN(j^{l-1}))+j^{l-1} \tag{3}$$

$$j^l=FNN(LN(z^l))+z^l, l=1,2,\dots,L \tag{4}$$

其中，MSA 表示多头自注意力(multi-head self-attention)，LN 表示层归一化(layer normalization)。经过 L 层编码后，$\boldsymbol{j}^L=[j_1^L,\dots,j_{(K+N)}^L]$ 中的前 K 个嵌入向量 $[j_1^L,\dots,j_K^L]$ 对应与细粒度筛选的 K 个人物候选区域相关的特征。然后，这些 K 个嵌入向量被输入到分类头中，用于预测一组动作类别得分 $[\hat{c}_1,\hat{c}_2,\dots,\hat{c}_K]\in R^{N_{cls}\times K}$。这组动作类别得分和人物边界框构成一组动作实例 $(\hat{b},\hat{c})=\{(\hat{b}_i,\hat{c}_i)\}_{i=1}^{K}$。最后，基于置信度得分选择最终的 K' 个动作实例(其中 $K'<K$)。

1.3　加权分数聚合

本节对 FMSF 扩展以便于进行长时人体动作检测和识别。在此任务中，动作根据关键帧 I_t 和以关键帧为中心的长时视频帧 $\{I_{t-L_p},\dots,I_{t+L_f}\}(L_f+L_p+1\gg T)$ 进行检测。FMSF 在长视频片段的多个不同时间戳上应用，提取动作信息，然后将这些来自不同时间戳的信息进行聚合，以便对整个长视频的动作进行更全面的检测和理解，如图 3 所示，FMSF 在长时 HMR 任务下训练后的模型称为 FMSF-Prolonged。

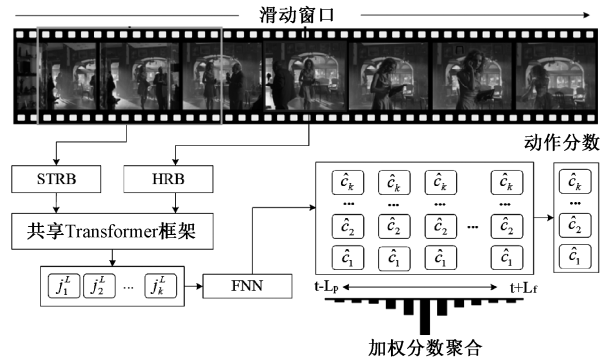

图 3　FMSF-Prolonged 的处理流程

Fig. 3　Processing flow of FMSF-Prolonged

对 $\{I_{t-L_p},\dots,I_{t+L_f}\}$ 应用一个滑动窗口，步长为 W，以提取 T 帧的短时视频片段，即 $I(n)=\{I_{t+W_n-T_p},\dots,I_{t+W_n+T_f}\}$，其中 $n\in[-\lfloor L_p/W\rfloor,\lfloor L_f/W\rfloor]$。此时得到的输出为：

$$(\hat{b}_n,\hat{c}_n)=FMSF(I(n),I_t) \tag{5}$$

其中，$FMSF(I(n),I_t)$ 表示 FMSF 将关键帧 I_t 和短

期视频片段 $I(n)$作为输入进行操作。即使关键帧 I_t 不在短期视频片段 $I(n)$内，得益于本文的时域数据增强策略，FMSF 仍能够正常检测和识别动作。基于从关键帧 I_t 获取的 K 个人物候选区域，提出的加权分数聚合方法在不同的时间窗口位置聚合这些人物的动作分数，即：

$$\hat{c}_k^{prolonged} = \sum_{n=-\lfloor L_p/W \rfloor}^{\lfloor L_f/W \rfloor} A_{n,k}\hat{c}_{n,k} \tag{6}$$

其中，$\hat{c}_{n,k}$ 和 $A_{n,k}$ 分别是第 k 类动作在第 n 个窗口位置的分数和相应的权重。假设 FMSF 模型生成了 K 个动作实例，那么 K 的值会设置为多于关键帧中可能出现的人物数量，以确保模型不会遗漏任何可能的动作区域。此外，通过二分图匹配将这些动作实例与真实的动作实例进行关联，训练过程与 DETR(detection transformer)[16]类似。

1.4　损失函数

1）二分图匹配

FMSF 模型将 K 个预测集 $\{\hat{b}_i,\hat{h}_i,\hat{c}_i\}_{i=1}^{K}$ 与关键帧中的 K^{gt} 个真实动作集 $\{b_i^{gt},h_i^{gt},c_i^{gt}\}_{i=1}^{K^{gt}}$ 进行匹配，其中 $\hat{b}_i$，$\hat{h}_i$，$\hat{c}_i$ 分别是第 i 个动作实例的预测边界框、人物检测置信度分数和动作分类得分，b_i^{gt}，h_i^{gt}，c_i^{gt} 分别是第 i 个动作实例的真实边界框、目标二分类标签和目标动作分类标签。构建真值人物目标集为：

$$\begin{cases} y^{gt} = \{y_i^{gt}\}_{i=1}^{K} \\ s.t.\ y_i^{gt} = \begin{cases} (b_i^{gt},h_i^{gt}), (1 \leqslant i \leqslant K^{gt}) \\ (\varnothing,\varnothing), 其他 \end{cases} \end{cases} \tag{7}$$

此后，使用二分图匹配方法将预测输出集 $\hat{y} = \{\hat{y}_i\}_{i=1}^{K} = \{(\hat{b}_i,\hat{h}_i)\}_{i=1}^{K}$ 与目标集 y^{gt} 进行匹配。根据 Deformable DETR[17]，使用匹配代价函数 $Loss_{match}(y_i^{gt},\hat{y}_i)$：

$$Loss_{match}(y_i^{gt},\hat{y}_i) = \delta\{h_i^{gt} \neq \varnothing\} Loss_{cls}(h_i^{gt},\hat{h}_j) + \delta\{h_i^{gt} \neq \varnothing\} Loss_{box}(b_i^{gt},\hat{b}_j) \tag{8}$$

其中，$Loss_{cls}$ 表示 sigmoid 焦点损失，$Loss_{box}$ 表示 L1 损失和 GIoU 损失的线性组合，其中 $Loss_{cls}(h_i^{gt},\hat{h}_j)$ 会受到人物检测置信度分数的影响，而动作分类分数并不会提升模型的整体性能。

最后，使用匈牙利算法[18]寻找 $\hat{y}$ 和 y^{gt} 之间的最优二分图匹配：

$$\hat{\sigma} = \underset{\sigma}{\operatorname{argmin}} \sum_{i=1}^{K} Loss_{match}(y_i^{gt},\hat{y}_{\sigma(i)}^{p}) \tag{9}$$

其中，σ 表示通过匈牙利算法获得的 K 个索引的排列，$\sigma(i)$表示 σ 中的第 i 个元素。

2）最终损失

在获得二分匹配结果 σ 后，使用动作分类损失来优化整个 FMSF 模型：

$$\begin{aligned} &Loss(c^{gt},\hat{c}) = \sum_{i=1}^{K} Loss_{cls}(c_i^{gt},\hat{c}_{\sigma(i)}) \\ &s.t.\ \hat{c} = \{\hat{c}_i\}_{i=1}^{K} \\ &s.t.\ c^{gt} = \{g_i^{gt}\}_{i=1}^{K} \\ &s.t.\ g_i^{gt} = \begin{cases} c_i^{gt}, (1 \leqslant i \leqslant K^{gt}) \\ \varnothing, 其他 \end{cases} \end{aligned} \tag{10}$$

对于长时检测与识别，首先使用短时视频片段训练模型，然后冻结参数，并使用长时视频片段优化分数聚合权重。

2　实验结果及分析

2.1　数据集与评估指标

在开源动作识别 AVA-60 数据集[19]上评估了 FMSF 模型。AVA-60 数据集包含时长为 15 min 的电影片段，其中训练集和验证集的片段个数分别为 235 和 64，动作类别数为 60。此外，采用了三折交叉验证协议，并通过对 3 次结果进行平均，评估模型的性能。最终，在所有 3 个数据集上报告了在 IoU 阈值为 0.5 时的帧级 mAP(mAP_f)。

2.2　实验设置

1）实验配置

对短时任务，使用时长为 2.4 s 的视频片段进行训练。长时任务的视频片段时长为 17 s。参数 T_p 和 T_f 设置相同，L_p 和 L_f 也设置相同。视频骨干网络的超参数设置遵循原始设置。共享 Transformer 框架包含 6 层网络，在每一层之前进行归一化处理。隐藏层的大小为 256，前馈网络的隐藏层大小为 1 024。注意力头的数量为 8，激活函数采用 SELU (scaled exponential linear unit)[20]。dropout 率和注意力 dropout 率皆设置为 0.1。

对于 FMSF-Prolonged，首先使用短视频片段进行训练，并在冻结 FMSF 模型参数的情况下，微调分数聚合权重。模型使用 RAdam (rectified adam)优化器[21]，权重衰减为 1×10^{-4}。在训练视频骨干网络时初始学习率设置为 1×10^{-5}，其他情况下设置为 1×10^{-4}，然后在第 5 轮时将学习率衰减为原来的 1/10。batch 大小设置为 32，总共训练 10 轮。对于 HRB，使用原始分辨率大小的输入图像。对于 STRB，将输入视频片段的较短空间边设置为 256 pixel (SlowFast 骨干网络)或 224 pixel(ViT 骨干网络)，同时保持纵横比不变。

此外，除了使用随机水平翻转来进行数据增强之外，还为 FMSF 设计了一种时域数据增强方法。具体而言，给定一个关键帧图像 It，在训练过程中，对输入视频片段的时间偏移 τ 进行随机化，此时视频片段的时间范围为 $I_{t-T_p+\tau},\ldots,I_{t+T_f+\tau}$。时间偏移的范围相对于关键帧的位置设置为±1.5 s。这样一来，即使视频片段没有与关键帧时间对齐，也使得模型能够完全利用视频片段中的时空信息。

2) 人体检测模型

在HRB中,使用了两种人体检测模型:一种使用EfficientNet-B7-FPN[22]作为骨干网络的Faster R-CNN[23],基于OpenMMLab进行训练;另一种使用EfficientNet-B7作为骨干网络的DETR[16]。此外,HRB首先在ImageNet[24]和MPII人物姿态图像数据集[25]上进行了预训练,然后进行微调。

3) 视频骨干网络

在STRB中,使用了基于3D-CNN和基于ViT的两种视频骨干网络。前者包含SlowFast-R50[26]和SlowFast-R101,分别在Kinetics-400 (K400)[27]和Kinetics-700 (K700)数据集上进行预训练,参数设置遵循原始设置。后者为ViT-B[28],并使用了VideoMAE[29]或VideoMAE v2的预训练权重来进行初始化。

2.3 对比实验结果及其分析

表1和2给出了FMSF在AVA-60 v2.1和AVA-60 v2.2验证集上的性能,并与基于3D-CNN骨干网络和ViT骨干网络的最新分步方法和端到端方法进行了对比。

表1 与分步方法的对比实验结果

Table 1 Compares the experimental results with the stepwise method

方法	检测模型	输入(帧数×采样率)	骨干网络	预训练集	视频长度/s	mAP_f	
						AVA-60 v2.1	AVA-60 v2.2
文献[30]	F-RCNN	32×2	I3D-R101-NL	K400	58	26.98	27.19
文献[26]	F-RCNN	32×2	SF-R101-NL	K600	2	27.49	28.50
文献[31]	F-RCNN	32×2	SF-R101	K700	61	30.85	31.85
文献[32]	F-RCNN	32×2	SF-R101	K400	19	27.47	27.58
文献[32]	F-RCNN	32×2	SF-R101	K700	19	31.96	32.09
文献[29]	F-RCNN	16×4	ViT-B	K400	2	30.50	30.76
文献[33]	F-RCNN	16×4	ViT-B	IN-1K & K400	2	33.94	34.10
FMSF	F-RCNN	32×2	SF-R101	K700	2	33.69	35.62
FMSF-Prolonged	F-RCNN	32×2	SF-R101	K700	17	35.04	36.59
FMSF	F-RCNN	16×4	ViT-B	K400	2	35.47	35.58

表2 与端到端方法的对比实验结果

Table 2 Compares the experimental results with the end-to-end method

方法	检测模型	输入(帧数×采样率)	骨干网络	预训练集	视频长度/s	mAP_f	
						AVA-60v2.1	AVA-60 v2.2
文献[34]	无	20×1	S3D-G	K400	1	16.67	17.58
文献[35]	无	64×1	I3D-VGG	K400	2	24.28	25.66
文献[36]	无	32×2	SF-R101-NL	K600	2	27.59	28.27
文献[37]	无	32×2	CSN-152	IG-65M & K400	17	31.88	33.56
文献[38]	无	32×2	SF-R101	K700	64	32.51	32.85
文献[38]	无	16×4	ViT-B	K710 & K400	2	35.96	36.08
文献[39]	无	16×4	ViT-B	K710 &K400	2	37.51	37.65
FMSF	DETR	32×2	SF-R101	K700	2	34.17	34.90
FMSF-Prolonged	DETR	32×2	SF-R101	K700	17	35.08	35.90
FMSF	DETR	16×4	ViT-B	K710 & K400	2	38.40	39.50
FMSF-Prolonged	DETR	16×4	ViT-B	K710 & K400	17	39.74	40.01

可以看到,FMSF和FMSF-Prolonged在性能上都大幅超越了其他方法。其中FMSF-Prolonged在AVA-60 v2.1和AVA-60 v2.2上相较文献[38]分别提升2.23%~2.75%和3%~3.9%。基于ViT-B骨干网络的FMSF模型,在AVA-60 v2.2上相较STMixer提升了3.4%的mAP。

2.4 消融实验结果及其分析

在消融实验中,为了加快评估,人体检测模型为基于

ResNet-50-FPN 的 Faster R-CNN，视频骨干网络为 SlowFast-R50，数据集为 AVA v2.2，并且除非特别说明，皆未使用时间数据增强方法。

1）模型组件消融

本节对 FMSF 的组件进行消融，包括分开提取人物特征与语境特征，共享 Transformer 框架，以及细粒度筛选，结果如表 3 所示。

表 3　模型组件消融实验结果

Table 3　Ablation results of model components

分开提取人物特征与语境特征	共享 Transformer 框架	细粒度筛选	mAP_f
√	√	√	30.01
×	×	×	25.84
×	√	√	27.14
√	×	√	29.67
×	√	×	26.52

可以看到，完全体 FMSF 模型的 mAP 为 30.01%；当移除 3 个组件后，性能显著下降到 25.84%；从共享的视频骨干网络中提取人物与语境特征时，性能下降了 2.87%（30.01%～27.14%）；若将共享 Transformer 框架替换为普通的编-解码器 Transformer 框架时，mAP 下滑 0.34%（30.01%～29.67%），这表明前者在建模两种嵌入特征之间的关联时的有效性。最后，细粒度筛选比直接使用人体检测模型得到的人物候选区域提升了 0.62%（27.14%～26.52%）的性能。

2）建模框架消融

本节将共享 Transformer 框架与其他两种建模框架进行了对比。表 4 为仅使用解码器的 Transformer，一般的编-解码器 Transformer，共享 Transformer 框架的性能结果，这 3 种模型的结构如图 4～6 所示。可以看到，相比前两者，本文共享 Transformer 框架的 mAP 提高了 0.54%。

表 4　建模框架消融实验结果

Table 4　Results of ablation experiment of modeling framework

Transformer 结构	mAP_f
仅使用解码器	29.47
编解码	29.47
共享框架	30.01

3）损失函数消融

本节考量 3 个动作子类别的性能，即人物姿势、人与人交互、人与物交互。具体而言，通过使用动作分类损失 $Loss_{cls}^{motion} = Loss_{cls}(c_i^{gt}, \hat{c}_j)$、人体检测损失 $Loss_{cls}^{person} = Loss_{cls}(h_i^{gt}, \hat{h}_j)$ 和边界框回归损失 $Loss_{box} = Loss_{box}(b_i^{gt}, \hat{b}_j)$ 的不同组合，构建了不同损失函数，结果如表 5 所示。

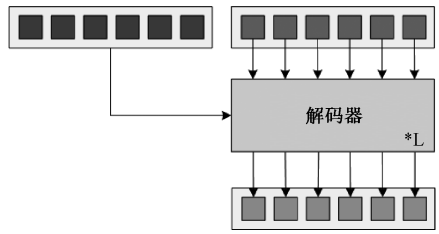

图 4　仅使用解码器的 Transformer 结构

Fig. 4　Transformer structure using only decoders

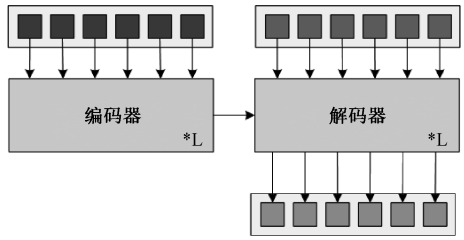

图 5　一般的编-解码器 Transformer 结构

Fig. 5　General codec Transformer structure

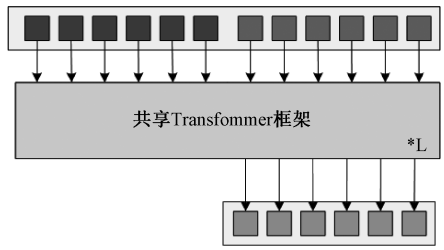

图 6　共享 Transformer 框架结构

Fig. 6　Shared Transformer framework architecture

这 3 类的 mAP 分别记为 mAP_f_p、mAP_f_{p-p} 和 mAP_f_{p-o}。可以看出，使用 $Loss_{cls}^{person}$ 和 $Loss_{box}$ 比使用 $Loss_{cls}^{motion}$ 和 $Loss_{box}$ 的性能更优，而同时使用这 3 个损失函数并不能保证最佳性能，这表明了本文设计的有效性。

表 5　建模框架消融实验结果

Table 5　Results of ablation experiments in the modeling framework

动作分类损失	人体检测损失	边界框回归损失	mAP_f	mAP_f_p	mAP_f_{p-p}	mAP_f_{p-o}
√	×	√	29.52	49.74	31.46	20.27
×	√	√	30.01	51.00	31.91	20.70
√	√	√	29.60	50.06	31.67	20.36

4）Transformer 层深消融

为了评估 FMSF 的性能随 Transformer 层数变化的情况，表 6 列出了 AVA-60 验证集中 100 个视频片段的精度-效率权衡结果。可以看到，随着 Transformer 层数增加，性能有所提高，但当层数超过 7 时，精度不变且效率降低。

表 6　Transformer 层深消融实验结果

Table 6　Experimental results of Transformer deep ablation

层数	参数量/M	GFLOPs	mAP_f
1	140.95	207.04	28.67
4	142.39	211.90	28.88
7	143.02	215.27	30.01
10	149.66	228.27	30.01

5) 细粒度筛选消融

为了评估不同细粒度筛选方案的有效性,对人物候选区域的采样率(即人体检测模型的置信度阈值)进行调整。随着置信度阈值的降低,人体检测模型生成的候选特征数量增加。表 7 为 FMSF 在不同置信度阈值下的性能。值得注意的是,在没有应用置信度阈值的细粒度筛选策略下,性能最优。这表明 FMSF 能够有效地区分人物候选区域中的正样本和负样本。

表 7　细粒度筛选消融实验结果

Table 7　Results of fine-grained screening ablation experiments

采样率	mAP
0.9	28.55
0.8	29.27
0.7	29.56
0.5	29.64
0.1	29.74
0 (top-K)	30.01

6) 聚合策略消融

本节比较了几种可用于长时动作识别的聚合方法,包括最大池化,平均池化,top-k 池化,和本文的聚合策略,结果如表 8 所示。可以看到,FMSF 使用的加权求和聚合得到了最佳的性能。

表 8　聚合策略消融实验结果

Table 8　Results of ablation experiments with polymerization strategies

聚合策略	mAP_f
最大池化	25.06
平均池化	27.36
top-K 池化	27.95
加权聚合	30.01

7)时间偏移范围消融

表 9 为 FMSF 在不同时间偏移范围下进行时间数据增强的性能。可以看到当时间偏移范围设置为[−1.5, 1.5]时,取得了最佳性能。

表 9　时间偏移范围消融实验结果

Table 9　Results of time-offset ablation experiments

偏移范围	mAP_f
[−1.0, 1.0]	29.87
[−1.5, 1.5]	30.01
[−2.0, 2.0]	29.76
[−2.5, 2.5]	29.58

2.5　性能分析

1) 加权分数聚合

图 7 为 FMSF-prolonged 中对不同时刻和动作类别优化后的分数聚合权重。关键帧位于时间步 t,因此分数聚合权重在时间步 t 处的值最大。可以看到,分数聚合权重在不同类别中分布不同。对于“触摸物品”或“跳”等瞬时动作,分数聚合权重在时间步 t 周围较为显著。而对于“游泳”或“爬山”等跨越较长时间的动态动作,考量分数的范围较大。

	坐	触碰物品	睡觉	起立	拥抱	跳	穿衣服	听音乐	爬山	工作	游泳
t-3	0.0038	0.0048	0.0088	0.0089	0.0215	0.0277	0.0283	0.0311	0.0436	0.0541	0.0526
t-2	0.0049	0.0077	0.0129	0.0107	0.0316	0.0322	0.0514	0.0479	0.0625	0.0716	0.0777
t-1	0.0317	0.0516	0.061	0.08	0.1294	0.1429	0.1744	0.1465	0.1761	0.1477	0.1624
t	0.8974	0.8501	0.7904	0.7762	0.5447	0.4619	0.3749	0.403	0.3459	0.2806	0.2133
t+1	0.0321	0.0574	0.0676	0.06	0.1475	0.161	0.1646	0.1779	0.1394	0.1555	0.1449
t+2	0.0049	0.0077	0.0129	0.0107	0.0318	0.0355	0.0525	0.051	0.0508	0.0734	0.0738
t+3	0.0038	0.0048	0.0088	0.0093	0.0209	0.0275	0.0354	0.0325	0.0388	0.0447	0.0616

图 7　分数聚合权重热力图

Fig. 7　Fractional aggregate weight heat map

2) 计算效率

表 10 中比较了 FMSF 与目前先进方法的计算效率。分步方法的总成本表示为人物检测模型与动作分类器的每秒浮点计算(floating-point operations per second, FLOPs)之和。为了公平比较,输入帧的分辨率设置相同(包括 224×224,256×256)。可以看到,像文献[37]和文献[38]这样的端到端方法相较于传统分步方法,展示了更高的效率,更适合于嵌入式部署。然而,若考虑到精度,或者采用轻量化人物检测模型进一步减少计算成本,则分步方法的优势显而易见。

2.6　可视化分析

在图 8 中,通过可视化比较了文献[38]与本文共享 Transformer 框架生成的注意力热力图。可以看到:首先,相较而言 FMSF 更能够捕捉到关键区域,以生成更优的动

作候选特征。此外，FMSF 更擅长人物之间或人与物之间的关系建模。

表 10　计算效率比较结果

Table 10　Results of calculation efficiency comparison

方法	检测模型	输入（帧数×采样率）	骨干网络	GFLOPs	mAP_f
文献[26]	F-RCNN	32×2	SF-R101-NL	119&28	29.00
文献[32]	F-RCNN	32×2	SF-R101	160&56	31.70
文献[29]	F-RCNN	16×4	ViT-B	180&73	31.80
文献[36]	无	32×2	SF-R101-NL	253	28.30
文献[37]	无	32×2	CSN-152	122	31.10
文献[38]	无	32×2	SF-R101	137	30.10
文献[38]	无	16×4	ViT-B	357	36.10
FMSF	F-RCNN	32×2	SF-R101	160&64	35.62
FMSF	F-RCNN	16×4	ViT-B	190&64	35.58
FMSF	DETR	32×2	SF-R101	160&11	34.90
FMSF	DETR	16×4	ViT-B	190&11	39.50

(a) 文献[38]
(a) Literature [38]

(b) 本文方法
(b) The proposed method

图 8　可视化比较

Fig. 8　Visual comparison

2.7　泛化性能分析

为了增强数据多样性并适配复杂场景下的人体动作检测与识别，将 Charades 数据集中与 AVA-60 高度重叠的动作类别进行映射与整合，具体映射关系如下：AVA 的“Stand”对应 Charades 的“Standing”，AVA 的“Carry (an object)”对应 Charades 的“Carrying something”，并统一归类为相应动作标签（如统一类别“Stand”、“Carry”等），共涵盖 Stand、Sit、Walk、Talk、Touch、Carry、Hold、Open、Close、Drink、Eat、Read 等 11 个交叉类别。从 Charades 标注文件中筛选只包含上述动作类别的视频，并使用 30 fps 的速率将其时间范围级标注（start_time，end_time，action_label）转换为帧级标注（frame_id，action_label），期间仅保留映射成功的类别信息以保证数据一致性。随后，将这些帧级标注与 AVA-60 的帧级标注相结合，并在视频级别随机打乱后，按照 训练集 ：验证集 ：测试集＝6∶2∶2 的比例进行重新划分，确保同一视频整体分配到训练、验证或测试子集中，以维持动作实例的时间连贯性；对映射后的动作类别，统计并对比其帧数与视频数分布，若发现类别分布不均衡，则对各子集进行了微调。在完成上述整合与划分后，得到一个更大且覆盖多样动作场景的 AVA&Charades 数据集。同样为了加快评估，人体检测模型为基于 ResNet-50-FPN 的 Faster R-CNN，视频骨干网络为 SlowFast-R50，未使用时间数据增强方法，评估长时动作视频的长度为 17 s，结果如表 11 所示。

表 11　泛化性评估结果

Table 11　Results of generalization evaluation

数据集	方法	mAP_f
AVA v2.2	FMSF	30.01
AVA&Charades	FMSF	30.68
AVA v2.2	FMSF-Prolonged	30.74
AVA&Charades	FMSF-Prolonged	31.29

从表 11 中可以看出，在引入 Charades 数据后，FMSF 的 mAP_f 由 30.01% 提升至 30.68%，增加了 0.67%；FMSF-Prolonged 则由 30.74% 增加至 31.29%，提升了 0.55%。这两组结果都体现了将 Charades 与 AVA 数据进行类别映射和整合后，能够为模型提供更丰富的场景和动作样本，弥补单一数据集在场景多样性方面的不足，从而带来一定幅度的性能增益。值得注意的是，FMSF-Prolonged 相比 FMSF 本身就更擅长处理长视频中复杂的动作转变与时序关系，因此在结合 Charades 这种多样化的日常室内场景后，也相对获得了更显著的收益。这些结果验证了长视频策略对模型性能有帮助，特别是针对跨越更长时间段的复杂动作；此外，数据扩充有助于提高模型对不同动作场景的适应能力，进一步提升识别准确率。在后续研究中，如果进一步优化数据映射策略、精细挑选 Charades 视频或改进模型结构，则有可能取得更显著的性能提升。

3　结　　论

本文提出的 FMSF 方法通过有效融合细粒度人物候选特征与全局时空语境信息，实现了对复杂视频场景中人体动作的精细化检测与识别。相较于传统分步方法在 RoI 特征上的局限性，FMSF 借助共享 Transformer 框架，充分挖掘人物与上下文之间的多级关联，从而获得更丰富、更具鲁棒性的动作表示。同时，引入的加权分数聚合策略为长时视频动作识别提供了灵活且高效的解决方案，无需依赖于特征记忆库或长期记忆编码器，即可在长时跨度上显著提升模型性能。未来的研究方向可进一步探讨轻量化骨干网络与高效检测模型的融合，以降低计算成本并提升在嵌入式设备上的应用潜力。此外，可将 FMSF 的思想扩展至更为多样化的视频理解任务，如多人交互分析、行为预测以及事件级别的视频语义理解，为构建更智能、更通

用的人机交互系统奠定坚实基础。

参考文献

[1] 许晨炀,范非易,柯冠舟,等.基于多尺度通道注意力机制的行为识别方法[J].电子测量技术,2023,46(21):114-122.
XU CH Y, FAN F Y, KE G ZH, et al. Behavior recognition method based on multi-scale channel attention mechanism [J]. Electronic Measurement Technology, 2023,46(21):114-122.

[2] 孙梓誉,顾晶.基于雷达时频变换和残差网络的人体行为检测[J].电子测量技术,2024,47(10):27-33.
SUN Z Y, GU J. Human behavior detection based on radar time-frequency transform and residual network[J]. Electronic Measurement Technology, 2024, 47(10): 27-33.

[3] 王雪,程换新,骆晓玲,等.基于改进的 YOLOv5 网络的异常行为检测算法研究[J].电子测量技术,2022,45(16):137-141.
WANG X, CHENG H X, LUO X L, et al. Research on abnormal behavior detection algorithm based on improved YOLOv5 network [J]. Electronic Measurement Technology,2022,45(16):137-141.

[4] KARÁCSONY T, JENI L A, TORRE F D L, et al. Deep learning methods for single camera based clinical in-bed movement action recognition[J]. Image and Vision Computing, 2024, 143: 104928.

[5] YU F, FANG Y Q, ZHAO ZH X, et al. CAGNet: A context-aware graph neural network for detecting social relationships in videos[J]. Visual Intelligence, 2024, 2(1): 22.

[6] LIU H, MA Y N, HU Q Y, et al. CenterTube: Tracking multiple 3D objects with 4D tubelets in dynamic point clouds [J]. IEEE Transactions on Multimedia, 2023, 25: 8793-8804.

[7] DIBA A, SHARMA V, ARZANI M, et al. Spatio-temporal convolution-attention video network [C]. IEEE/CVF International Conference on Computer Vision, 2023: 859-869.

[8] RAVISHANKAR H, ANITHAKUMARI R D, SARVAMANGALA D R, et al. Video compression through advanced video saliency aware spatial-temporal integration and attention mechanisms[J]. SN Computer Science, 2024, 5(7): 926.

[9] ZHENG Y Y, TAO F ZH, GAO ZH Y, et al. FGYOLO: An integrated feature enhancement lightweight unmanned aerial vehicle forest fire detection framework based on YOLOv8n[J]. Forests, 2024, 15(10): 1823.

[10] LEE J, KIM S, KIM S, et al. Discriminative action tubelet detector for weakly-supervised action detection[J]. Pattern Recognition, 2024, 155: 110704.

[11] LI M Y, ZHOU D, LIU X ZH, et al. Simulation of E-learning virtual interaction in Chinese language and literature multimedia teaching system based on video object tracking algorithm [J]. Entertainment Computing, 2025, 52: 100764.

[12] AKHTER I, MUDAWI N A, ALABDULLAH B I, et al. Human-based interaction analysis via automated key point detection and neural network model [J]. IEEE Access, 2023, 11: 100646-100658.

[13] DENG W B, ZHANG Y T, YU H, et al. Knowledge graph embedding based on dynamic adaptive atrous convolution and attention mechanism for link prediction[J]. Information Processing & Management, 2024, 61(3): 103642.

[14] HUANG Z P, WAN Q, CHEN J L, et al. ADATS: Adaptive roi-align based transformer for end-to-end text spotting[C]. 2023 IEEE International Conference on Multimedia and Expo(ICME). IEEE, 2023: 1403-1408.

[15] LIU Z K, CHEN S H, GUO L T, et al. Enhancing vision-language pretraining with jointly learned questioner and dense captioner [C]. 31st ACM International Conference on Multimedia, 2023: 5120-5131.

[16] ARKIN E, YADIKAR N, XU X, et al. A survey: Object detection methods from CNN to transformer[J]. Multimedia Tools and Applications, 2023, 82(14): 21353-21383.

[17] DANG M, WANG H X, NGUYEN T H, et al. CDD-TR: Automated concrete defect investigation using an improved deformable transformers [J]. Journal of Building Engineering, 2023, 75: 106976.

[18] ZHANG SH, XUE Y, ZHANG H, et al. Improved hungarian algorithm-based task scheduling optimization strategy for remote sensing big data

processing [J]. Geo-Spatial Information Science, 2024, 27(4): 1141-1154.

[19] FAY M P, FOLLMANN D A, LYNN F, et al. Anthrax vaccine-induced antibodies provide cross-species prediction of survival to aerosol challenge[J]. Science Translational Medicine, 2012, 4(151): 151126.

[20] HUANG ZH, NG T, LIU L, et al. SNDCNN: Self-normalizing deep CNNs with scaled exponential linear units for speech recognition[C]. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing, 2020: 6854-6858.

[21] KISSIEDU A N T, AGGREY G K, ASANTE-MENSAH M G, et al. Development of pneumonia identification system: A comparative analysis of some selected CNN architectures using Adam, Nadam, and RAdam optimizers [C]. 2024 IEEE Smart Block4 Africa. IEEE, 2024: 1-12.

[22] REN SH Q, HE K M, GIRSHICK R, et al. Faster R-CNN: Towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 39(6): 1137-1149.

[23] DENG J, DONG W, SOCHER R, et al. Imagenet: A large-scale hierarchical image database[C]. 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009: 248-255.

[24] ROGEZ G, WEINZAEPFEL P, SCHMID C. LCR-Net++: Multi-person 2D and 3D pose detection in natural images [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019, 42(5): 1146-1161.

[25] ZENG W, HUANG J J, ZHANG W, et al. Slowfast action recognition algorithm based on faster and more accurate detectors[J]. Electronics, 2022, 11(22): 3770.

[26] LIANG J W, ZHANG E W, ZHANG J, et al. Multi-dataset training of transformers for robust action recognition [C]. Advances in Neural Information Processing Systems, 2022, 35: 14475-14488.

[27] YANG M, GAO H, GUO P, et al. Adapting short-term transformers for action detection in untrimmed videos [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024: 18570-18579.

[28] TONG Z, SONG Y B, WANG J, et al. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training [C]. Advances in Neural Information Processing Systems, 2022, 35: 10078-10093.

[29] ZHANG R, XUE J X, LIN F, et al. Enhancing human action recognition with fine-grained body movement attention [C]. 2024 IEEE International Conference on Multimedia and Expo, 2024: 1-6.

[30] XU M Z, XIONG Y J, CHEN H, et al. Long short-term transformer for online action detection [J]. Advances in Neural Information Processing Systems, 2021, 34: 1086-1099.

[31] BABAZAKI Y, IWAMOTO K, TAKAHASHI K, et al. Heterogeneous feature fusion for improving performance of action detection [C]. Journal of Physics: Conference Series, IOP Publishing, 2024, 2759(1): 012001.

[32] ZHENG Y D, CHEN G, YUAN M L, et al. MRSN: Multi-relation support network for video action detection[C]. 2023 IEEE International Conference on Multimedia and Expo, 2023: 1026-1031.

[33] KIM H W, CHOI Y S. Fusion attention for action recognition: Integrating sparse-dense and global attention for video action recognition [J]. Sensors (Basel, Switzerland), 2024, 24(21): 6842.

[34] SU S W, GAN M G. Online spatio-temporal action detection with adaptive sampling and hierarchical modulation[J]. Multimedia Systems, 2024, 30(6): 349.

[35] YU S Q, CHEN L L, ZHANG X L, et al. VTR: Bidirectional video-textual transmission rail for clip-based video recognition[C]. 2024 IEEE International Conference on Multimedia and Expo, 2024: 1-6.

[36] SUI L, ZHANG C L, GU L X, et al. A simple and efficient pipeline to build an end-to-end spatial-temporal action detector [C]. IEEE/CVF Winter Conference on Applications of Computer Vision, 2023: 5999-6008.

[37] SINGH G, CHOUTAS V, SAHA S, et al. Spatio-temporal action detection under large motion [C]. IEEE/CVF Winter Conference on Applications of Computer Vision, 2023: 6009-6018.

[38] WU T, CAO M Q, GAO Z T, et al. Stmixer: A one-

stage sparse action detector [C]. IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023: 14720-14729.

[39] CHEN L, TONG Z, SONG Y B, et al. Efficient video action detection with token dropout and context refinement[C]. IEEE/CVF International Conference on Computer Vision, 2023: 10388-10399.

作者简介

王峥(通信作者),硕士,讲师,主要研究方向为体育信息技术。

E-mail:wangzhen1986zzu@126.com

赵新辉,博士,副教授,主要研究方向为体育信息技术。

王小伟,硕士,高级实验师,主要研究方向为视频动作识别、体育大数据分析等。